Deep Speaker Embeddings with Convolutional Neural Network on Supervector for Text-Independent Speaker Recognition

Danwei Cai, Zexin Cai and Ming Li*
Data Science Research Center, Duke Kunshan University, China
E-mail: ming.li369@dukekunshan.edu.cn
* corresponding author

***Abstract*—Lexical content variability in different utterances is the key challenge for text-independent speaker verification. In this paper, we investigate using supervector which has ability to reduce the impact of lexical content mismatch among different utterances for supervised speaker embedding learning. A DNN acoustic model is used to align a feature sequence to a set of senones and generate centered and normalized first order statistics supervector. Statistics vectors from similar senones are placed together and reshaped to an image to maintain the local continuity and correlation. The supervector image is then fed into residual convolutional neural network. The deep speaker embedding features are the outputs of the last hidden layer of the network and we employ a PLDA back-end for the subsequent modeling. Experimental results show that the proposed method outperforms the conventional GMM-UBM i-vector system and is complementary to the DNN-UBM i-vector system. The score level fusion system achieves 1.26% ERR and 0.260 DCF10 cost on the NIST SRE 10 extended core condition 5 task.**

Index Terms: Speaker verification, text-independent, CNN, supervector, deep speaker embedding

I. Introduction

Speaker verification, as a biometric technology, authenticate a speaker's identity given a speech signal. Generally, speech contains not only lexical information but also paralinguistic speech attributes, e.g. speaker, language, channel, emotion and so on. Reducing these variability factors unrelated with speaker information has been a key challenge in speaker verification systems. Based on the constraints of the lexical contents, speaker verification can be categorized into text-dependent and text-independent one. Text-dependent speaker verification employs the same set of phrases for enrollment and verification, so all utterances are with nearly the same word sequence. In contrary, text-independent one is a context-free task. Text-dependent speaker verification usually outperforms the text-independent one [1] for the constraint of the linguistic contents.

To reduce the impact of lexical content mismatch, most text-independent speaker verification systems adopt a high- and fixed-dimensional supervector to represent a speech utterance [2]. In this approach, acoustic features of an utterance are aligned to a set of universal background model (UBM) tokens which can be Gaussian mixture model (GMM) components trained in an unsupervised manner [3] or phonetic states of a deep neural network (DNN) acoustic model [4]. The aligned features can be seen as the statistical acoustic patterns of the given utterance on every token and are stacked together to generate a supervector. As supervector is a high-dimensional feature, current speaker verification systems use i-vector modeling to perform dimension reduction [5]. In i-vector modeling, a single factor analysis is used to generate a low dimensional total variability space (i.e. i-vector space) which jointly models language, speaker and channel variabilities.

Motivated by the success of deep learning, researchers in speaker verification have been working on learning discriminative deep speaker embedding features. Variani et al. train a DNN to classify speakers and extract d-vector at the frame level for text-dependent speaker verification [6]. Heigold et al. train an end-to-end system to discriminate speaker pairs [7] and achieved better performance than i-vector baseline system. Zhang et al. learn speaker embedding features with an attention based convolutional neural network (CNN) framework for the text-dependent task [8]. In text-independent speaker verification task, DNN based methods outperform i-vector baseline under short duration [9], [10], [11] or large amount of data [12] condition. Recently, end-to-end deep learning based approaches with different encoding layer designs have also been proposed for speaker verification and language identification [13], [14], [15].

All the DNN based systems mentioned above use frequency domain features (i.e. MFCC, spectrogram, mel-filter bank energy) as inputs. Since the input utterance may have any arbitrary length, existing end-to-end approaches usually take a fixed length input and then perform average pooling at different layers or score level. This may not be ideal since the variability of lexical contents may require a finer statistics calculation on different phonetic units for text-independent tasks. Motivated by the success of DNN i-vector [4] and tandem i-vector [16], [17], the acoustic model generated phoneme posterior probabilities is a natural 0^{th} order occupancy probability for different phonetic tokens. Therefore, we perform speaker embedding learning directly on the DNN phonetic-aware supervector which already serves as a sequence-to-

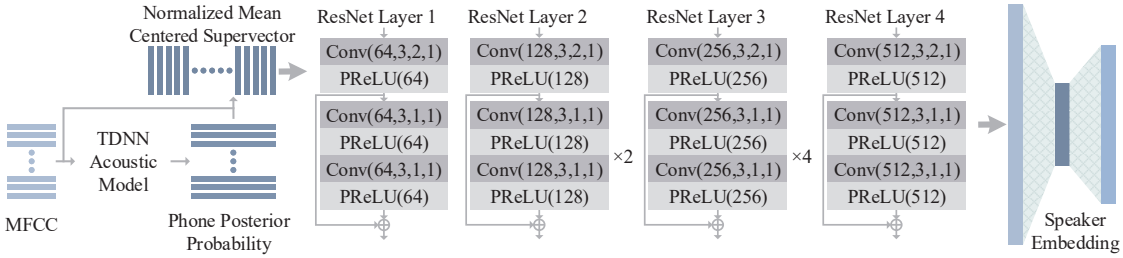

Fig. 1. The residual CNN architecture on supervector image for text-independent speaker recognition. The parameters in convolutional layer is Conv(*output channel, kernel size, stride, padding*), the number in PReLU stands for learnable parameters.

vector pooling module. Moreover, the existing end-to-end methods usually employ fully connected layers to model the pooled fixed dimensional activations or features. In this paper, we use CNN to explore the local correlation of features on different tokens. Using residual network [18], we could achieve a very deep network with good robustness against overfitting.

The contributions of the proposed method are as follows. Firstly, we utilize the power of DNN acoustic model to robustly perform phonetic aware 1^{st} order statistical pooling to get supervectors and reshape the high-dimensional supervector as a image. Secondly, we feed supervector images into deep CNN to learn speaker discriminative embeddings. Convolutional layers have much less parameters comparing to fully connected layers, this allows us to utilize deeper network architecture and helps to prevent overfitting. Thirdly, we explore the local continuity and correlation of similar phonetic tokens in supervectors. As the phonetic tokens tied with DNN acoustic model may have phonetic similarities among different granularity levels, we use hierarchical clustering and phonetic decision tree to reorder the token indexes and form a supervector image, in which similar phonetic tokens are grouped together to highlight local correlation. This reorder operation can help CNN to learn more discriminative speaker embeddings. Finally, the speaker embedding features are modeled and scored by cosine similarity and PLDA back-end [19]. Experimental results show that the new supervector image and ResNet based speaker embedding outperforms the DNN or fully connected layer learnt embedding from supervectors. Furthermore, the proposed method achieves comparable performance with the start-of-art DNN i-vector speaker verification systems, and score level fusion can further boost the performance.

The rest of the paper is organized as follows. Section 2 presents the proposed framework for text-independent speaker verification. The experimental results are presented in Section 3 while conclusions are future works are provided in Section 4.

II. METHOD

Fig. 1 illustrates the network architecture in this work. Details are shown in the following sections.

A. Feature extraction

Given a MFCC sequence $\{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_L\}$ from a L frames utterance, the 0^{th} and centered 1^{st} order Baum-Welch statistics on the UBM are calculated as follows:

$$N_i = \sum_{t=1}^{L} P(c_i|\mathbf{y}_t) \tag{1}$$

$$\mathbf{F}_i = \sum_{t=1}^{L} P(c_i|\mathbf{y}_t)(\mathbf{y}_t - \boldsymbol{\mu}_i) \tag{2}$$

where c_i stands for each of the senone in time delayed neural network (TDNN) acoustic model and $\boldsymbol{\mu}_i$ is the mean vector of the corresponding senone, $P(c_i|\mathbf{y}_t)$ is the frame level phone posterior probability (PPP) stands for the i^{th} senone extracted from a TDNN [20]. The corresponding centered mean supervector $\tilde{\mathbf{F}}$ is generated by concatenating all the $\tilde{\mathbf{F}}_i$ together:

$$\tilde{\mathbf{F}}_i = \frac{\sum_{t=1}^{L} P(c_i|\mathbf{y}_t)(\mathbf{y}_t - \boldsymbol{\mu}_i)}{\sum_{t=1}^{L} P(c_i|\mathbf{y}_t)} \tag{3}$$

The MFCC features are 60 dimensional feature vectors consisting 20 MFCC coefficients and their first & second order deltas. The TDNN acoustic model outputs a 5515 dimensional PPP and the final supervector as input to residual network can be seen as a 5515×60 image.

B. Data pre-processing

The 5515×60 dimensional supervector image consists of 3 5515×20 blocks corresponding to MFCC coefficients, first and second derivatives, respectively. We perform mean and variance normalization for each block separately.

Furthermore, since each senone may have different occupancy probability in different utterances, we use the 0^{th} Baum-Welch statistics to re-weight the supervector as follows,

$$\tilde{\mathbf{F}}_i = \sqrt{\sum_{t=1}^{L} P(c_i|\mathbf{y}_t)} \frac{\sum_{t=1}^{L} P(c_i|\mathbf{y}_t)(\mathbf{y}_t - \boldsymbol{\mu}_i)}{\sum_{t=1}^{L} P(c_i|\mathbf{y}_t)} \tag{4}$$

This normalized supervector gives more weight on those 'confident' senone and less on others, thus we hope the DNN may concentrate more on $\tilde{\mathbf{F}}_i$ with 'confident' senone.

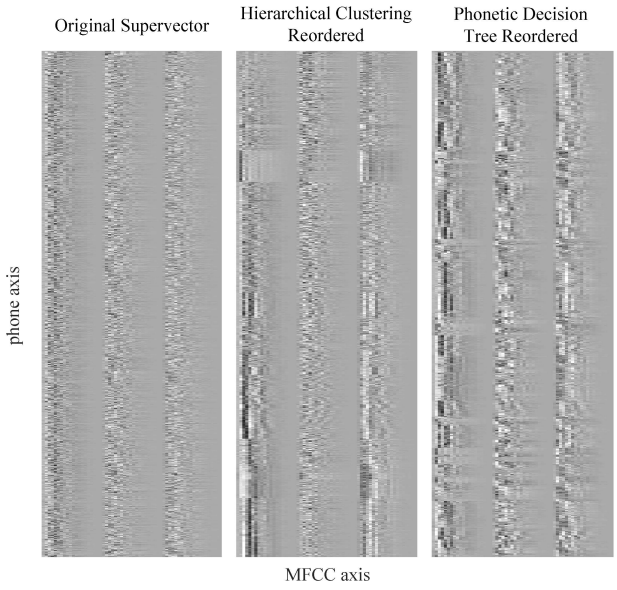

Fig. 2. Original, hierarchical clustering reordered, phonetic decision tree reordered DNN-UBM mean supervector. The three supervectors above are normalized as mentioned in section II-B.

C. Explore continuity in supervector

The 5515 senones of TDNN output layer have phonetic similarity among different granularity level such as tri-phone level, mono-phone level and vowel & consonant level. Grouping the similar senones together may offer continuity and correlation in the supervector image and make CNN to learn the speaker embedding more effectively. We explore two methods to group similar senones together in this work.

1) Hierarchical clustering index: The first method used to generate reordering indexes is hierarchical clustering. Firstly, the distances between each of the senone's mean vector $\boldsymbol{\mu}_i$ are calculated. We try cosine distance and correlation distance in this work. Secondly, a hierarchy tree of clusters is built based on the distances. The leaf nodes of the hierarchy tree are associated with 5515 senones. Finally, going through all the leaf nodes with the same direction forms the reordering indexes.

2) Phonetic decision tree clustering index: For the second method, we extract reordering indexes from the phonetic decision tree which is generated during TDNN acoustic model training phase. In decision tree, a question relates to the phonetic context is attached to each node. The tree is constructed for each state of each phone to cluster all of the corresponding states of all of the associated tri-phones [21]. We thus use the indexes of the leaf nodes in phonetic decision tree to reorder the senones.

Figure 2 shows the differences between the original supervector image without reordering and the two reordered supervector image. We can see that the supervector image without reordering is like a 'noisy' picture to some extent. The hierarchical clustering reordered supervector shows some continuity in the MFCC coefficients block and second derivatives block, but remains the first derivatives part in a noisy style. Phonetic decision tree groups the similar senone together, the third supervector image shows its continuity in all three blocks.

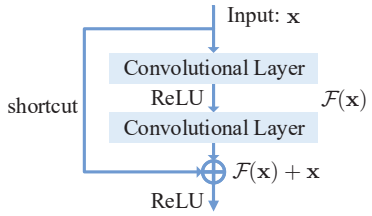

Fig. 3. A ResNet building block

D. Neural network architecture

Deep residual convolutional network (ResNet) [18] is a kind of very deep network architectures showing competitive accuracy and nice convergence behaviours in many computer vision tasks such as object recognition, face identification, emotion recognition. Residual network has a series of residual block in which a bypassing shortcut connects the residual block's outputs and inputs. A typical ResNet block is shown in figure 3. It is defined as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \mathbf{W}_i) + \mathbf{x} \tag{5}$$

where $\mathbf{x}$ and $\mathbf{y}$ are input and output of vectors of the ResNet block, the function $\mathcal{F}(\cdot)$ can be one or more convolutional layers with parameters $\mathbf{W}_i$.

In this work, we have 4 residual network layers with 1, 2, 4, 1 residual block(s). Each residual block contains two convolutional layers with one additional down-sampling layer for the first block of each ResNet layer. The total number of convolutional layer is 20. Pooling layer is not used and the size of the supervector images change whenever a down-sampling convolutional layer works. Each convolutional layer is followed by a parametric rectified linear unit (PReLU) [22].

The output of ResNet is then sent into a feed-forward neural network with two fully connected layers. The first is a bottleneck layer and is used to learn speaker embeddings. The final layer is used to classify speaker labels when training ResNet. The loss function for ResNet training is multi-class cross entropy.

E. Deep speaker embeddings

The last hidden layer of the network serves as a bottleneck layer with 512 nodes. The final fully connected layer is performed on the last hidden layer to classify speaker labels during the training phase. Thus the outputs of the last hidden layer can be seen as a deep speaker embedding feature. We use simple cosine similarity scoring and probabilistic linear discriminant analysis (PLDA) back-end [19] to assess the learned deep speaker embeddings. Pairs of speaker embeddings are compared on the PLDA model to generate final verification scores.

III. Experiments

A. Dataset

We conduct our speaker verification experiments on NIST SRE 2010 extended core condition 5 [23]. Training data are conversational telephone speech from datasets released through the Linguistic Data Consortium. The SRE portion consists of NIST SRE data from 2004 through 2008. The Switchboard portion consists of Switchboard 2 Phase 2, 3 and Switchboard Cellular. We employ a energy based voice activity detector (VAD) to drop all frames that are decoded as silence or speaker noises. This result in a total number of 57350 utterances from 5756 speakers. In the evaluation set, there are 4267 enrol models and 767 test segments. In total there are 416119 trials with 7169 target trials including both male and female trials.

B. i-vector baseline system

We first train a GMM-UBM based ivector system using Kaldi toolkit. The front-end features are 20 MFCCs with a frame length of 30ms that are normalized over a sliding window of up to 3 seconds. First and second derivatives are appended to create 60 dimensional features. The UBM is a 2048 component GMM and the dimension of i-vector is 600. I-vectors are centered and length normalized before PLDA scoring. GMM and i-vector extractor are trained on all the training data and PLDA is trained on SRE portion.

We also train a DNN-UBM ivector system. It uses supervector described in section II-A. The TDNN acoustic model used as a DNN-UBM is trained on about 1,800 hours of the English portion of Fisher dataset [24]. The other configurations on i-vector extractor and PLDA model are same as GMM-UBM based i-vector system.

C. ResNet system setup

When training the ResNet based speaker verification system, we refine the training data by removing speakers with less than 4 utterances. Then we have a total of 4352 speakers which matches the size of final output layer. All the training data including Switchboard and SRE portion are used to train the PLDA back-end.

D. Fully connected DNN system setup

To verify the effectiveness of the deep residual convolutional neural network on supervectors, we also train a fully connected neural network (FC Net) on the same supervector as a comparison system. This network has two hidden fully connected layers with 512 neurons each and the activation function is rectified linear unit (ReLU). The input is a 330900×1 high-dimensional supervector which is preprocessed as described in section II-B, and the output layer has 4352 nodes which matches with the 4352 speakers in the training set. The speaker features are read from the outputs of the last hidden layer.

TABLE I
Performance comparison on NIST SRE 2010

ID	System Description	PLDA		Cosine	
		EER(%)	DCF10	EER(%)	DCF10
1	GMM-UBM i-vector	2.28	0.489	6.93	0.799
2	DNN-UBM i-vector	1.45	0.255	3.42	0.461
3	FC Net / supervector	3.84	0.673	9.54	0.892
4	ResNet / supervector	2.22	0.402	8.90	0.811
5	ResNet / cosine HC	2.16	0.383	8.26	0.808
6	ResNet / correlation HC	2.09	0.391	8.84	0.821
7	ResNet / decision tree	2.16	0.407	7.85	0.788
8	Fusion 4 5 6 7	1.74	0.329	7.57	0.772
9	Fusion 1 8	1.51	0.293	5.61	0.664
10	Fusion 2 8	1.34	0.246	3.21	0.455
11	Fusion 1 2 8	1.26	0.260	3.29	0.442

E. Results

Table I shows the results obtained with experimental setup presented above. Performance metrics for the experiments are Equal Error Rate (EER) as well as the minimum of the normalized detection cost function (DCF) with $P_{\texttt{Target}} = 10^{-3}$ for SRE10 [23].

During implementing the systems, we observe that the deep speaker embedding trained by ResNet does not achieve the best performance on both cosine and PLDA back-ends. As the training epoch increasing, cosine EER decreases from 9.5% to 7.5% and the PLDA EER increases from 2.1% to 3.0%. The main reason for this may be overfitting. Overfitting may cause the mismatch distribution between training data and testing data. So this might explain that the performance of PLDA back-end which requires the same training data degrades. We believe that when training data is sufficiently large, cosine distance back-end may be good enough to model deep speaker embeddings just as in face verification tasks.

It can be observed from table I that all the four deep residual convolutional networks outperform the fully connected DNN on both cosine back-end and PLDA back-end. The performance gain mainly comes from the power of ResNet which is trained on supervector images. Comparing the four ResNet systems, we see that the reordered supervector images can boost the performance of ResNet. The most useful reordering methods are correlation distance based hierarchical clustering for EER and cosine distance based hierarchical clustering for minDEF10. Moreover, different reordered methods are complementary for each other, score level fusion can greatly improve the performance as as shown in Table I system 8. The combination of these systems was obtained by linear fusion of their scores. We use BOSARIS toolkit [25] to fuse the system scores. The fusion results show that the supervector image ResNet system and the i-vector system are complementary to each other. The final fusion result achieves 1.26% EER and 0.260 normalized minDCF10. Figure 4 shows the Detection Error Trade-off (DET) curves of the PLDA back-end system wit ID 1, 2, 8 and 11 in table I.

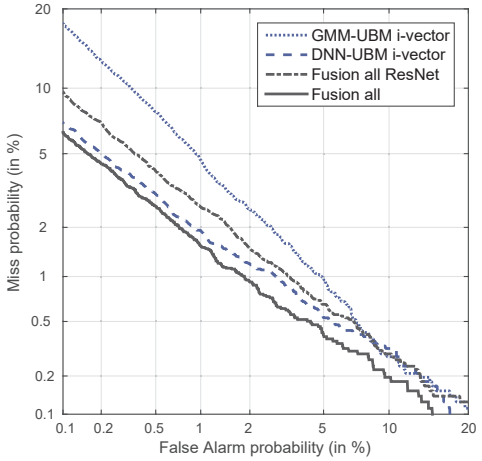

Fig. 4. DET curve for NIST SRE 2010

IV. Conclusion

In this paper, we investigate learning speaker embeddings with deep residual CNN on supervector images for text-independent speaker recognition system. A DNN acoustic model is used to align a feature sequence to a set of senones and generate the phonetic aware supervector. Statistics vectors from similar senones are placed together to maintain the local continuity and correlation using hierarchical clustering or phonetic decision tree. Experimental results show that training deep residual CNN on reordered supervector images outperforms applying fully connected DNN directly on high-dimensional supervectors.

Generating supervector can be seen as an encoding process in which the feature sequence is encoded into a set of senone tokens tied with ASR acoustic model. In the further work, we will extend the idea of using supervector for CNN. We will explore a encoding layer in DNN architecture to automatic learn speaker discriminative tokens and align a feature sequence to the learned tokens to get supervector. Then feature reordering and deep CNN can be applied on top of the learned supervector in an unified end-to-end architecture.

V. Acknowledgement

This research was funded in part by the National Natural Science Foundation of China (61773413), Natural Science Foundation of Guangzhou City (201707010363), National Key Research and Development Program (2016YFC0103905) and CCF-Tecent Open Fund.

References

[1] T. Matsui and S. Furui, "A text-independent speaker recognition method robust against utterance variations," in *Proc. of ICASSP*, 1991, pp. 377–380 vol.1.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[4] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. of ICASSP*, May 2014, pp. 1695–1699.

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

[6] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. of ICASSP*, May 2014, pp. 4052–4056.

[7] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. of ICASSP*, March 2016, pp. 5115–5119.

[8] S. X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *Proc. of SLT*, Dec 2016, pp. 171–178.

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

[10] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *Proc. of Interspeech*, Aug 2017, pp. 1487–1491.

[11] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang, "Deep speaker feature learning for text-independent speaker verification," in *Proc. of Interspeech*, Aug 2017, pp. 1542–1546.

[12] 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 verification," in *Proc. of SLT*, 2016, pp. 165–170.

[13] Weicheng Cai, Jinkun Chen, and Ming Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Proc. of Odyssey*, 2018, pp. 74–81.

[14] Weicheng Cai, Zexin Cai, Xiang Zhang, Xiaoqi Wang, and Ming Li, "A novel learnable dictionary encoding layer for end-to-end language identification," in *Proc. of ICASSP*, 2018.

[15] Weicheng Cai, Zexin Cai, Wenbo Liu, Xiaoqi Wang, and Ming Li, "Insights into end-to-end learning scheme for language identification," in *Proc. ICASSP*, 2018.

[16] M. Li and W. Liu, "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features," in *Proc. of Interspeech*, Sept 2014, pp. 1120–1124.

[17] M. Li, L. Liu, W. Cai, and W. Liu, "Generalized i-vector representation with phonetic tokenizations and tandem features for both text independent and text dependent speaker verification," *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 207–215, Feb 2016.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of CVPR*, June 2016, pp. 770–778.

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

[20] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *Proc. of ASRU*, Dec 2015, pp. 92–97.

[21] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *The Workshop on Human Language Technology*, 1994, pp. 307–312.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. of ICCV*, Dec 2015, pp. 1026–1034.

[23] National Institute of Standards and Technology, "The NIST year 2010 speaker recognition evaluation plan," *www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplan-r6.pdf*.

[24] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text," in *International Conference on Language Resources & Evaluation*, 2004, vol. 4, pp. 69–71.

[25] N. Brmmer and E. D. Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new DCF," in *NIST SRE Analysis Workshop*, 2011.