Interspeech 2018
2-6 September 2018, Hyderabad

Angular Softmax for Short-Duration Text-independent Speaker Verification

Zili Huang†*, Shuai Wang*†*, Kai Yu*

Key Lab of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering
Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China

{huangziliandy,feixiang121976,kai.yu}@sjtu.edu.cn

Abstract

Recently, researchers propose to build deep learning based end-to-end speaker verification (SV) systems and achieve competitive results compared with the standard *i*-vector approach. In addition to deep learning architectures, optimization metric such as softmax loss or triplet loss, is important for extracting speaker embeddings which are discriminative and generalizable to unseen speakers. In this paper, angular softmax (A-softmax) loss is introduced to improve speaker embedding quality. It is investigated in two SV frameworks: a CNN based end-to-end SV framework and an *i*-vector SV framework where deep discriminant analysis is used for channel compensation. Experimental results on a short-duration text-independent speaker verification dataset generated from SRE reveal that A-softmax achieves significant performance improvement compared with other metrics in both frameworks.

Index Terms: text-independent speaker verification, metric learning, A-softmax

1. Introduction

Speaker recognition aims to recognize or verify one's identity through the given speech segment. It can be classified into text-dependent and text-independent according to the lexicon constraint on the spoken content. Research interests in speaker recognition include acoustic features, speaker modeling and noise robustness, among which most researchers pay their attention to speaker modeling.

Gaussian Mixture Model-Universal Background Model (GMM-UBM) system dominated the speaker recognition field for one decade since proposed in [1]. Inspired by Joint Factor Analysis in [2], *i*-vector[3] represents the state-of-the-art speaker modeling framework. By modeling the speaker factors and channel factors in a single total variability subspace, *i*-vector provides a low-dimensional embedding representation of the speaker identity.

Deep neural networks (DNN) achieve incredible performance in many tasks such as image recognition[4], machine translation [5] and speech recognition[6, 7, 8], which also inspires researchers to apply this powerful tool to speaker recognition. In previous works, DNN is usually utilized in two ways. The first is similar to the speech recognition task[6], in which DNN substitutes the GMM in the *i*-vector framework[9, 10]. In this framework, sufficient statistics are computed against a pre-trained speech recognition DNN instead of the original GMM-UBM. The second approach is to extract bottleneck features [11, 12, 13, 14] or speaker representations [15, 16, 17] with DNN, among which *d*-vector is the most typical one. By averaging the frame-level extracted deep features, the utterance-level representation *d*-vector is obtained. Some researchers follow and extend this work by replacing the simple neural network with complicated architectures such as Convolutional Neural Network (CNN) and Time-Delay Neural Network (TDNN)[18], or redesign the optimization metric and propose new embeddings such as *j*-vector [19]. Recently, instead of training the DNN on the frame level, researchers in [20] add a temporal pooling layer and train the model on the utterance level.

Standard speaker verification tasks are open-set problems, which means speakers in the training set and the evaluation set have no overlap. We expect the DNN to learn a discriminative speaker embedding space which is generalizable enough to unseen speakers. Good speaker embeddings should have small intra-speaker variations and large inter-speaker differences. More and more researchers are regarding the speaker embedding learning as a metric learning problem. Metrics including triplet loss[21, 22] and the generalized end-to-end loss[23] are adopted. In these two frameworks, the training criterion of the DNN is to reduce the variations of embeddings from the same speaker and enlarge the distances between embeddings from different speakers. However, the performances of these frameworks are sensitive to the sampling strategy. According to our experience with the triplet loss based systems, careful design of the triplets plays a critical role in the training procedure. Tricks such as hard trial selection are adopted to achieve better accuracy.

The angular softmax (A-softmax) was first proposed in face recognition [24], which shares many properties with the speaker recognition task. A-softmax loss modifies the softmax loss function to learn angularly discriminative embeddings and adds a controllable parameter m to pose constraints on the intra-speaker variation of the learned embeddings. In this paper, we investigate the application of A-softmax loss in two SV frameworks. First, A-softmax loss is adopted in the same way as other metrics such as softmax and triplet loss to directly learn speaker embeddings from cepstral features. Second, A-softmax loss is used to train a simple neural network as a compensation method in the *i*-vector space, which is termed as deep discriminant analysis (DDA) in our previous work. A-softmax loss achieves significant performance improvement in both frameworks.

The rest of this paper is organized as follows. Section 2 reviews some works on deep speaker embeddings. Two detailed types, softmax and triplet loss based speaker embeddings are introduced. Section 3 introduces the A-softmax loss and its ap-

†:These authors have contributed equally to this work

The corresponding author is Kai Yu

This work has been supported by the National Key Research and Development Program of China under Grant No.2017YFB1002102 and the Major Program of Science and Technology Commission of Shanghai Municipality (STCSM) (No.17JC1404104). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

10.21437/Interspeech.2018-1545

plication for deep speaker embedding learning. Section 4 introduces an intra-speaker variability compensation method named DDA and how we incorporate A-softmax loss into this framework. We discuss the experiments and analyze the results in Section 5. Section 6 concludes this paper.

2. Deep Speaker Embeddings

One of the most common applications of DNN in speaker verification is to learn speaker representations (extract speaker embeddings). Researchers investigated different deep learning architectures[18, 25] and optimization metrics[23, 25, 26] to learn compact and discriminative speaker embeddings. In this section, two frameworks to extract deep speaker embeddings are introduced.

2.1. Softmax loss based speaker embeddings

Softmax loss is the most commonly used classification loss function, which is formulated as

$$\mathcal{L}_{\text{softmax}} = -\frac{1}{N}\sum_i \log\left(\frac{e^{\mathbf{w}_{y_i}^\top \mathbf{x}_i + b_{y_i}}}{\sum_j e^{\mathbf{w}_j^\top \mathbf{x}_i + b_j}}\right) \quad (1)$$

where N is the number of samples, $\mathbf{x}_i$ is the deep feature of the i-th sample and y_i is the corresponding label index. $\mathbf{W}$ is the parameter of the last fully connected layer and $\mathbf{b}$ is the bias term. The softmax loss based SV system is shown in Figure 1. The deep neural network takes the cepstral features as the input. After several frame-level layers, a temporal pooling layer aggregates the frame-level features from the same utterance to a single utterance representation. Compared with the classical *d*-vector[9], the DNN is trained on the utterance level. It should be noted that our implementation also differs from the one in [20, 27], because we didn't take the covariance statistics into consideration in the temporal pooling layer.

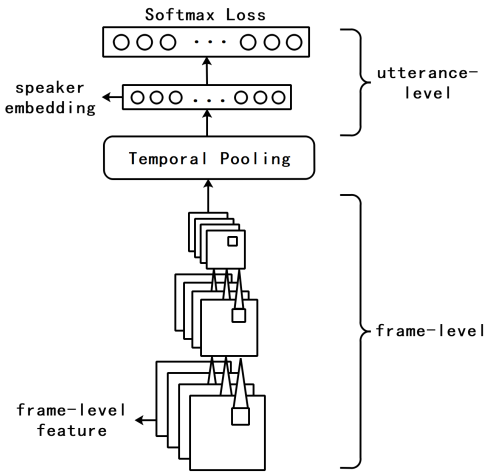

Figure 1: *Softmax loss based speaker verification*

2.2. Triplet loss based speaker embeddings

Triplet loss explicitly reduces intra-class variations and enlarges inter-class differences[21]. The architecture of triplet loss based system is depicted in Figure 2[28]. In the training stage, we first organize the samples into triplets. Each triplet consists of an anchor (an utterance from a specific speaker), a positive sample (an utterance from the same speaker) and a negative sample (an utterance from a different speaker). Similar to the architecture in Section 2.1, the deep neural network also derives utterance-level embeddings from frame-level features with a temporal pooling layer. The triplet loss is calculated with the utterance embeddings in the same triplet and back-propagation algorithm is performed to update parameters.

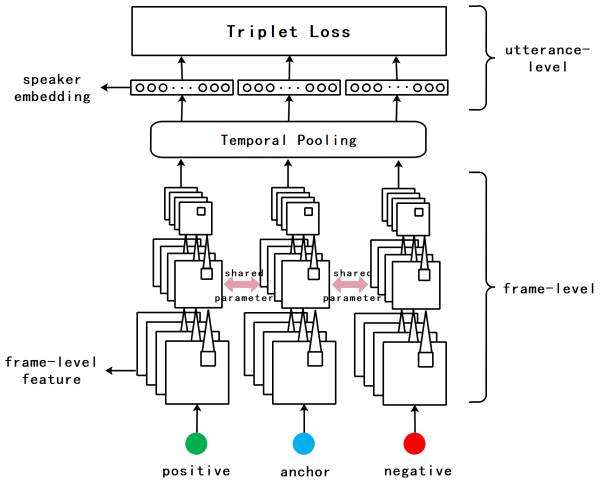

Figure 2: *Triplet loss based speaker verification*

3. Angular softmax (A-softmax) loss based speaker embeddings

Although researchers have obtained promising results using the deep neural networks supervised by aforementioned metrics, some problems also exist. In the softmax loss based systems, there is no explicit constraint on the intra-speaker variation. As a result, the generalization ability of the model is doubtful. Although triplet loss supervised speaker embeddings exhibit good properties on discriminative ability and robustness[21, 28], complex sample mining is required, which is time-consuming and performance-sensitive. In this section, we introduce the angular softmax (A-softmax) loss[24] and its application for deep speaker embedding learning.

3.1. Angular softmax

The motivation of A-softmax loss comes from the observation shown in Figure 3. The embeddings supervised by softmax loss have great discrimination ability in the angular space (which is also mentioned in the center loss related research [29]).

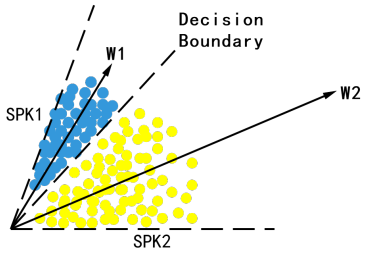

Figure 3: *Decision boundary learned by Softmax loss*

If we further constrain that $||\mathbf{w}_j|| = 1$ (this is accomplished by normalizing the weight matrix every time it is updated) and $b_j = 0$, the softmax function becomes the modified softmax loss,

$$\mathcal{L}_{\text{modified}} = -\frac{1}{N}\sum_i \log\left(\frac{e^{||\mathbf{x}_i||\cos(\theta_{y_i,i})}}{\sum_j e^{||\mathbf{x}_i||\cos(\theta_{j,i})}}\right) \quad (2)$$

where $\theta_{j,i}$ is the angle between $\mathbf{w}_j$ and $\mathbf{x}_i$. This formula shows the probability of a sample i belonging to a class j is only determined by the angle $\theta_{j,i}$ between them. The training process aims to reduce the angle between the sample and the corresponding class and enlarge the angle with other classes.

Different from the softmax and modified softmax loss, the A-softmax loss not only separates samples in the angular space, but also enforces an angular margin between classes. Traditional softmax function classifies sample i into its corresponding class y_i if $\forall k \neq y_i, \mathbf{w}_{y_i}\mathbf{x}_i + b_{y_i} > \mathbf{w}_k\mathbf{x}_i + b_k$, and the modified softmax loss requires $\forall k \neq y_i, \cos(\theta_{y_i,i}) > \cos(\theta_{k,i})$. A-softmax loss makes it more stringent to classify a sample into the corresponding class. It requires $\forall k \neq y_i, \cos(m\theta_{y_i,i}) > \cos(\theta_{k,i})$ where m is an integer and $m \geq 2$. By directly formulating this idea into the modified softmax loss, we derive

$$\mathcal{L}_{\text{angular}} = -\frac{1}{N}\sum_i \log\left(\frac{e^{||\mathbf{x}_i||\cos(m\theta_{y_i,i})}}{Z}\right) \quad (3)$$

$$Z = e^{||\mathbf{x}_i||\cos(m\theta_{y_i,i})} + \sum_{j\neq y_i} e^{||\mathbf{x}_i||\cos(\theta_{j,i})} \quad (4)$$

where $\theta_{y_i,i} \in [0, \frac{\pi}{m}]$. This constraint can be removed by substituting the cosine function with a monotonic function $\psi(\theta_{y_i,i}) = (-1)^k\cos(m\theta_{y_i,i}) - 2k$, where $\theta_{y_i,i} \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}]$ and $k \in [0, m-1]$. $m \geq 1$ is an integer that controls the size of angular margin. (When $m = 1$, A-softmax loss becomes the modified softmax loss.) Therefore, the A-softmax loss is formulated as

$$\mathcal{L}_{\text{angular}} = -\frac{1}{N}\sum_i \log\left(\frac{e^{||\mathbf{x}_i||\psi(\theta_{y_i,i})}}{Z}\right) \quad (5)$$

$$Z = e^{||\mathbf{x}_i||\psi(\theta_{y_i,i})} + \sum_{j\neq y_i} e^{(||\mathbf{x}_i||\cos(\theta_{j,i}))} \quad (6)$$

The learned decision boundary is depicted in Figure 4. Compared with softmax loss in Figure 3, A-softmax loss greatly enlarges the angular margin of the deep features.

3.2. A-softmax loss based speaker embeddings

As introduced in Section 3.1, A-softmax enlarges the angular margin between different classes and forces embeddings from the same speaker to approach their corresponding $\mathbf{w}$-vector. This effect is quite similar to softmax loss combined with center loss[29], despite the distance measurement differs. The architecture of A-softmax loss based speaker verification system is similar to the one depicted in Figure 1, and the only difference is the training criterion.

4. Deep Discriminant Analysis

Although *i*-vector is the state-of-the-art method for text-independent speaker verification, it models the speaker factors and channel factors in the same variability space. Channel compensation methods such as Linear Discriminant Analysis (LDA)

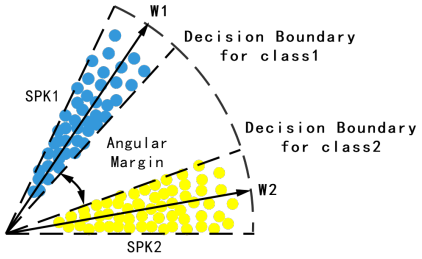

Figure 4: *Decision boundary learned by A-softmax loss*

and Probability Linear Discriminant Analysis (PLDA) are usually applied to the raw *i*-vectors. Through reducing the intra-speaker variations and enlarging the inter-speaker differences, LDA projects *i*-vectors onto a more discriminative space. We proposed a neural network based compensation scheme (termed as deep discriminant analysis, DDA) in [30], which shares the same spirit with LDA. In [30], the DDA is trained with the joint supervision of softmax loss and center loss. The center loss is formulated as

$$\mathcal{L}_{\text{center}} = \frac{1}{2}\sum_i ||\mathbf{x}_i - \mathbf{c}_{y_i}||_2^2 \quad (7)$$

where $\mathbf{c}_{y_i}$ denotes the y_ith class center of deep features. Experiments reveal DDA's superiority over traditional compensation methods such as LDA and PLDA. In this paper, we extend our previous work by adopting the A-softmax loss as the optimization criterion.

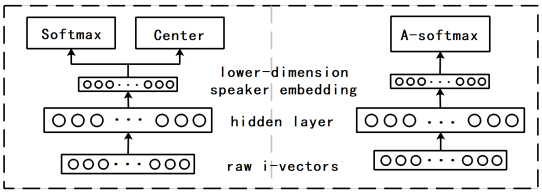

Figure 5: *Deep discriminant analysis*

As shown in Figure 5, the compensation neural network can be trained against center loss or A-softmax loss. Given the raw *i*-vector $\mathbf{x}$, the compensated lower-dimension speaker embedding $\mathbf{y}$ is represented as $\mathbf{y} = \mathcal{G}(\mathbf{x})$, where $\mathcal{G}()$ denotes the nonlinear transformation function learned by the NN with the training data.

5. Experiments

5.1. Dataset

Following our previous works[28, 30], we evaluate the performance of our methods on a short-duration text-independent dataset generated from the NIST SRE corpus. This short duration text-independent task is more challenging for speaker verification. The training set consists of selected data from SRE04-08, Switchboard II phase 2, 3 and Switchboard Cellular Part1, Part2. The utterances are chopped into short segments ranging from 3-5s after we remove silence frames with an energy-based VAD. The training set contains 4000 speakers and each speaker has 40 short utterances. The enrollment set and test set are selected from NIST SRE 2010 following a similar procedure. The enrollment set contains 300 speaker models (150

males and 150 females) and each model is enrolled by 5 utterances. The test set consists of 4500 utterances from the 300 models in the enrollment set. There are 392660 trials in the trial list, with 15 positive samples and 1294 negative samples on average for each model. No cross-gender trial exists. The detailed segmentation files and trial list will be released at `https://github.com/wsstriving/DEL_Segments.git`.

5.2. Implementation Details

Our baseline system is a standard *i*-vector system based on Kaldi SRE10 V1 recipe[31]. 20-dimension MFCCs with a frame-length of 25ms are extracted as front-end features, which are then extended to 60 dimensions with delta and acceleration. The UBM is a 2048 component full covariance GMM and the dimension of extracted *i*-vectors is 400. PLDA serves as the scoring back-end. The UBM, T-maxtrix and PLDA are trained with the training set mentioned in Section 5.1.

The softmax loss, triplet loss and A-softmax loss based systems adopt the same neural network architecture in Figure 6. It is a VGG-style CNN with 4 convolution layers, 2 max pooling layers and 1 fully-connected layer to extract the frame-level features. The frame-level features are averaged to utterance embeddings via a temporal pooling layer. The embedding dimension is set to 400 in all experiments.

The initial learning rate is set as 0.01 and is gradually reduced according to the validation accuracy. For triplet loss based system, we adopt the same configuration and strategy in our previous work[28]. For the A-softmax loss based system, to make training easier and more stable, we initialize the parameters with pretrained softmax models. 36-dimension Fbank features are extracted as front-end features for all the three systems and we extend 8 frames on each side to form the 17×36 time-frequency feature maps for each frame.

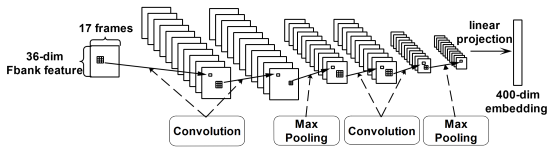

Figure 6: *VGG-style CNN architecture in our end-to-end system*

5.3. Results and Analysis

The A-softmax loss based embeddings are evaluated on the dataset described in Section 5.1 and compared with other speaker embeddings. Cosine distance scoring (CDS) is used as the back-end for neural embeddings. As shown in Table 1, softmax loss based speaker embeddings slightly outperform the *i*-vector/PLDA framework, which exhibits the effectiveness of utterance-level training (Classical *d*-vector is also experimented but cannot achieve comparable performance therefore not listed here). Through careful triplets designing and the "hard trial selection" trick, the Euclidean margin based triplet loss achieve better performance than the softmax loss. The best result is obtained from A-softmax loss, which outperforms *i*-vector/PLDA and traditional softmax by 24.4% and 19.4%, respectively.

5.3.1. Impact of the hyper-parameter

Intuitively, the hyper-parameter m controls the size of the angular margin. Larger m gives more stringent constraint on the distribution of the deep embeddings and enforces a larger angular margin between classes. However, larger m also leads to slower convergence. In our experiment, the performance of modified softmax loss($m = 1$) is close to the traditional softmax. With more stringent constraint posed by a larger m, the performance will be enhanced. The best result is obtained when $m = 3$, outperforming *i*-vector/PLDA framework by 24.4%. No further performance improvement is observed with a larger m.

Table 1: *EER comparison of different speaker embeddings, CDS as the scoring back-end*

Embeddings	EER (%)
i-vector (PLDA)	4.96
Softmax Embeddings	4.65
Triplet Embeddings	4.33
A-softmax Embeddings	**3.75**

Table 2: *Impact of the hyper-parameter* m

m	1	2	3	4
EER(%)	4.51	4.1	**3.75**	3.82

5.3.2. Compensation in the i-vector space

As described in Section 4, we incorporated A-softmax loss into the DDA compensation framework in the *i*-vector space. In Table 3, we illustrate the superiority of A-softmax loss based DDA as a scoring backend. The dimension of the raw *i*-vector is 400, and the transformed embedding dimension is set to 300 for both LDA and DDA. m is set to 3 in the A-softmax loss based DDA.

Table 3: *EER (%) of different compensation methods*

Methods	CDS	PLDA
Baseline	6.8	4.96
LDA	5.67	-
DDA(Center)	4.44	-
DDA(A-softmax)	**4.27**	-

As shown in Table 3, our proposed DDA clearly outperforms traditional LDA and PLDA. This performance is further enhanced if we substitute A-softmax loss for center loss. However, as shown in [30], both LDA and the proposed compensation methods are not compatible with PLDA on this dataset, which are not listed here.

6. Conclusions

In this paper, we investigate the application of angular softmax (A-softmax) in speaker verification. Inspired by the fact that the features learned by softmax loss have intrinsic angular distribution, A-softmax loss makes more stringent requirements during training to enforce an angular margin between different classes. Two A-softmax loss based SV frameworks are investigated, 1) a CNN based end-to-end SV framework 2) an *i*-vector SV framework where DDA is used for channel compensation. The proposed methods are evaluated on a short-duration text-independent speaker verification dataset generated from the SRE corpus. Relative improvements of 24.4% and 13.9% against the *i*-vector/PLDA baseline have been achieved in the two proposed A-softmax loss based frameworks, respectively.

7. References

[1] 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.

[2] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[3] 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, 2011.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[6] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[7] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[8] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2012.

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

[10] 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.

[11] T. Fu, Y. Qian, Y. Liu, and K. Yu, "Tandem deep features for text-dependent speaker verification." in *Interspeech*, 2014, pp. 1327–1331.

[12] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[13] Y. Tian, M. Cai, L. He, and J. Liu, "Investigation of bottleneck features and multilingual deep neural networks for speaker verification." in *Interspeech*, 2015, pp. 1151–1155.

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

[15] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*. IEEE, 2014, pp. 4052–4056.

[16] L. Li, Y. Lin, Z. Zhang, and D. Wang, "Improved deep speaker feature learning for text-dependent speaker recognition," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*. IEEE, 2015, pp. 426–429.

[17] S. Wang, Y. Qian, and K. Yu, "What does the speaker embedding encode?" in *Interspeech*, vol. 2017, 2017, pp. 1497–1501.

[18] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang, "Deep speaker feature learning for text-independent speaker verification," *arXiv preprint arXiv:1705.03670*, 2017.

[19] N. Chen, Y. Qian, and K. Yu, "Multi-task learning for text-dependent speaker verification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

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

[22] H. Bredin, "Tristounet: triplet loss for speaker turn embedding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017*. IEEE, 2017, pp. 5430–5434.

[23] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," *arXiv preprint arXiv:1710.10467*, 2017.

[24] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2017.

[25] 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 *Spoken Language Technology Workshop (SLT), 2016*. IEEE, 2016, pp. 165–170.

[26] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *Spoken Language Technology Workshop (SLT),2016*. IEEE, 2016, pp. 171–178.

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

[28] Z. Huang, S. Wang, and Y. Qian, "Joint i-vector with end-to-end system for short duration text-independent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018*. IEEE, 2018.

[29] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[30] S. Wang, Z. Huang, Y. Qian, and K. Yu, "Deep discriminant analysis for i-vector based robust speaker recognition," *arXiv preprint arXiv:1805.01344*, 2018.

[31] 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.