INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Shortcut Connections based Deep Speaker Embeddings for End-to-End Speaker Verification System

*Soonshin Seo, Daniel Jun Rim, Minkyu Lim, Donghyun Lee, Hosung Park, Junseok Oh, Changmin Kim and Ji-Hwan Kim**

Dept. of Computer Science and Engineering, Sogang University, Republic of Korea

{ssseo,drim,lmkhi,redizard,hosungpark,ohjs,kchangmin,kimjihwan}@sogang.ac.kr

Abstract

The objective of speaker verification is to reject or accept whether or not the input speech is that of a enrolled speaker. Traditionally, i-vector or speaker embeddings system such as d-vector representing the speaker information has been showing high performance with similarity metrics at the backend. Recently it has been proposed an end-to-end system based on previous speaker embeddings approach without additional strategy after extraction. Among the various models, CNN based end-to-end system is showing state-of-the-art performance. CNN based model is trained to classify multiple speakers and speaker embeddings are extracted.

In this paper, we propose shortcut connections based deep speaker embeddings for end-to-end speaker verification system. We construct modified ResNet-18 model so that the activation outputs from bottleneck architecture have shortcut connections to speaker embeddings. Deep speaker embeddings are extracted by jointly training in end-to-end approach. The model was constructed without other sophisticated methods such as length normalization, or additive margin softmax loss. When we tested proposed model on the unconstrained conditions data set called VoxCeleb1, the result showed EER of 3.03% when tested with high dimensional deep speaker embeddings. This is the state-of-the-art performance of end-to-end speaker verification model on VoxCeleb1.

Index Terms: end-to-end speaker verification system, deep speaker embeddings, shortcut connections, ResNet

1. Introduction

Speaker recognition is divided into speaker identification and speaker verification. Speaker identification is a problem of determining which one of the enrolled speakers is the input speech, and verification is a problem of determining whether or not the speech input is the speech of the enrolled speaker. In order to achieve high performance in speaker verification, how to extract speaker representations from speech input has become an important issue [1].

The most well-known method is to extract the i-vector through the gaussian mixture model (GMM) and universal background model (UBM) model [2]. This method involves extracting the i-vector, and then comparing speaker characteristics of enrolled speaker and test speaker through scoring such as probabilistic linear discriminant analysis (PLDA) backend [3, 4, 5, 6].

However, with recent advances in deep learning, method using deep neural networks (DNN) is emerging in speaker verification field [1]. In early system, DNN based model used to classify multiple speakers in Siamese architectures to discriminate same speaker and different speaker pairs with backend similarity metrics [7, 8]. DNN based model is used to extract the speaker embeddings representing the speaker information called d-vector [9]. The d-vector system is trained to classify multiple speakers with softmax loss and the embeddings are extracted by using the activations of the last hidden layer of the trained model. DNN based model can also be retrained by replacing softmax output layer to embeddings layer with contrastive loss [10, 11]. In recent years, an end-to-end speaker verification system based on d-vector approach has been proposed [12, 13, 14, 15, 16, 17], where embeddings will be extracted without further training and other additional strategies after extraction [17]. After extracting speaker embeddings, similarity metric compares pairs of embeddings by scoring.

In end-to-end speaker verification system, among the various models applied to extract speaker embeddings such as time-delay neural network (TDNN) [12, 14] or long short term memory (LSTM) network [16], convolutional neural networks (CNN) is showing promising results [13, 15, 17]. CNN has already established itself as a prominent model for image and face recognition [18, 19], and it has been suggested that deep CNN architectures such as ResNet [18] might be utilized to train speech embeddings where distance would correspond to measure of similarity to the speaker's identity [13, 15, 17, 20, 21]. CNN is suitable for extracting speaker embeddings with robust performance when speech input with unconstrained conditions, which is currently difficult in speaker verification [10, 11].

In this paper, we propose shortcut connections based deep speaker embeddings model for end-to-end speaker verification system. We construct modified ResNet-18 [18] model so that the activation outputs from bottleneck architecture have shortcut connections to speaker embeddings. We extract incorporated deep speaker embeddings by jointly training the proposed model and without the need for further training as in end-to-end speaker verification approach. We trained and tested the unconstrained conditions data set called VoxCeleb1 [10] and compared the performance between the previous models and the proposed model. As a result, proposed model achieved equal error rate (EER) of 3.03%, which was relative error reduction of 21.3% over the previous state-of-the-art model performance.

To summarize, our contributions are as follows: (i) We have constructed shortcut connections based deep speaker embeddings for end-to-end speaker verification system without utilizing more esoteric methods such as retraining, length normalization, and additive margin softmax loss; (ii) Our speaker verification model has produced state-of-the-art EER performance on VoxCeleb1 dataset; (iii) We also analyze

* Corresponding author

http://dx.doi.org/10.21437/Interspeech.2019-2195

the effect of various deep speaker embeddings from different shortcut connections, and we conclude that higher dimensional deep speaker embeddings tend to show better performance in our experiments.

We will introduce previous speaker verification system and ResNet in Section 2. In Section 3, we describe the proposed shortcut connections based deep speaker embeddings method and show the results in Section 4. Finally, conclusions are made in Section 5.

2. Related Works

2.1. i-vector based speaker verification system

Traditionally i-vector with PLDA has been dominant in the field of text-dependent speaker recognition [1]. i-vector [2] is computed by taking the sum of UBM super-vector *m* and total variability matrix *T* times a random vector with standard normal distribution *w* as speaker dependent GMM super-vector *M*. *w* is then trained to be i-vector. With probabilistic PLDA backend on i-vectors compute a similarity score between i-vectors by decomposing the speaker and session variability.

$$M = m + Tw \tag{1}$$

2.2. d-vector based speaker verification system

With recent advances in deep learning, methods using DNN are emerging in speaker verification field [1]. DNN based model is used to extract the speaker embeddings representing the speaker information called d-vector [9]. In this system, DNN is trained to discriminate multiple speaker at the frame-level. After the training is complete, the softmax output layer is discarded and d-vector are produced by activation values of the last hidden layer [14]. Then a trained model can also be retrained to classify d-vector in utterance-level replacing the softmax output layer with embeddings output layer using contrastive cost. Once embeddings for specific speaker are extracted, speech verification is decided by computing the distance between the target d-vector and the test d-vector using similarity metrics.

2.3. End-to-end based speaker verification system

Similar to the d-vector approach, an end-to-end speaker verification system has been proposed. In end-to-end system, there is no further retraining and L2 normalization is omitted [17]. After extracting speaker embeddings, similarity metric compares pairs of embeddings by scoring.

Our baseline end-to-end speaker verification system is depicted in Figure 1. Speaker Embeddings are extracted from trained end-to-end classification networks. Cosine similarity metric is used to compare pairs of speaker embeddings.

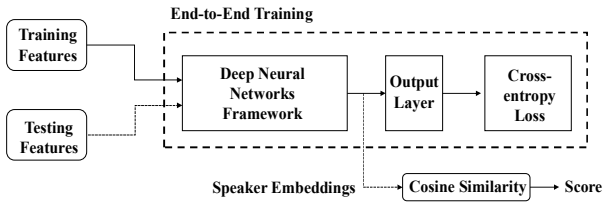

Figure 1: *End-to-end based speaker verification system*

2.4. Shortcut connections in ResNet

In end-to-end speaker verification system, CNN based speaker embedding model is showing high performance [1]. In particular, ResNet architecture, which is similar to standard multi-layer CNN but with shortcut connections [18], has become commonly used in end-to-end speaker verification system [13, 15, 17].

ResNet is based on the idea that one can attempt to approximate the residual functions that were not properly captured by previous CNN architecture depicted in Figure 2. The model attempts to remedy this by adding identity mapping as added layers.

$$y = F(x, \{W\}) [+] x \tag{2}$$

x is the input, y is the output layer, and $F(x, \{W\})$ is the residual mapping to be trained. The central idea is that the identity mapping will remedy the degradation by stacking non-linear layers. One can imagine the extreme case where that if the identity mapping was the optimal case, this model will better capture the reality by simply setting $F(x, \{W\})$ as 0, whereas it would be more difficult to model such case without the identity map [18]. Based on this identity mapping, the shortcut connections method is performed to train the model.

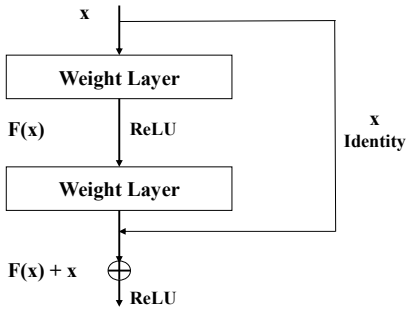

Figure 2: *Shortcut connections process in ResNet (adapted from* [18])

3. Shortcut Connections based Deep Speaker Embeddings

3.1. Overview

In end-to-end speaker verification system, speaker embeddings will be obtained by training DNN system with speaker identities as output layer [1]. The objective of end-to-end system is to extract speaker embeddings while training the network so that it would render additional training in the back end unnecessary [17].

In addition, we have utilized a network of our own that is modified from standard ResNet-18 by incorporating activation outputs from each of bottleneck architecture. Our architecture has been to inspired by shortcut connection nature in ResNet architecture [18]. We assume that the benefit of constructing deep speaker embeddings by incorporating the pooled output from bottleneck architecture would be to capture features that might otherwise be lost from the network. This means that the speaker representations would be richer and would be able to encode more speaker identity information in the training process.

3.2. Model architecture

The baseline network is ResNet-18. This model has total 18 layers including 8 residual blocks. Each residual block has convolution layer, batch normalization, and ReLU [22] activation function. We construct modified ResNet-18 architecture depicted in Table 1. Excluding proposed shortcut connections and fully connected networks at the end, the structure is the equal to the baseline model architecture. The main difference with baseline is how the deep speaker embeddings are extracted. As shown in Table 1, each res# layer (# is index of layer) contains 2 residual blocks and the bottleneck architecture reduced output size between res# layers. Each residual block is constructed identically to the residual block of standard ResNet architecture. The output for speaker embeddings following 3 bottleneck architectures and max pooling layer after conv1 layer are fed into 3 fully connected networks with 3 hidden layers and to the softmax output layer. Output layer represents speaker identities (classes of 1211). The purpose of last fully connected hidden layer size is to decide deep speaker embedding size (1024 dims). Finally, we extract the deep speaker embeddings using activations of last fully connected hidden layer (fc3).

Table 1: *Proposed model architecture*

Layer Name	Modified ResNet-18	Output Size (Training)	Output Size (Embeddings)
conv1 pool1	7×7, 64, stride 2 3×3, max pool, stride 2	$32 \times 150 \times 64$	-
pool2	32×150, avg pool	-	64
res2	$\begin{bmatrix} 3\times3, & 64 \\ 3\times3 & 64 \end{bmatrix} \times 2$	$32 \times 150 \times 64$	-
pool3	32×150, avg pool	-	64
res3	$\begin{bmatrix} 3\times3, & 128 \\ 3\times3 & 128 \end{bmatrix} \times 2$	$16 \times 75 \times 128$	-
pool4	16×75, avg pool	-	128
res4	$\begin{bmatrix} 3\times3, & 256 \\ 3\times3 & 256 \end{bmatrix} \times 2$	$8 \times 38 \times 256$	-
pool5	8×38, avg pool	-	256
res5	$\begin{bmatrix} 3\times3, & 512 \\ 3\times3 & 512 \end{bmatrix} \times 2$	$4 \times 19 \times 512$	-
pool6	4×19, avg pool	-	512
concatenate	-	-	1024
fc1 fc2 fc3 softmax	1024×1024 1024×1024 1024×1024 1024×1211	1024 1024 1024 1211	1024

3.3. Training and extracting deep speaker embeddings

The proposed model is trained to classify training speakers using a categorical cross-entropy objective function. The objective is to extract embeddings to capture speaker characteristics by incorporating outputs from each bottleneck architecture instead of simply using the activations from the penultimate layer as it is done now (depicted in Figure 3). We construct our deep speaker embeddings as:

$$y_{concat} \equiv y_0\,[+]\,y_1\,[+]\,y_2\,[+]\,y_3[+]\,y_4 \quad (3)$$

$[+]$ sign indicates concatenation. First embeddings component y_0 is the average pooling of max pooling output of first convolution layer. Each of the embeddings component y_i is computed as i^{th} output average pooling of bottleneck architecture for $i \geq 1$:

$$y_i = avg\_pool(F(x, \{W_i\})\,[+]\,x) \quad (4)$$

y_i represents average pooling of output of bottleneck architecture of proposed network. y_{concat} is in turn fed into fully connected networks that are connected to the output layer and embeddings are extracted from the output of fully connected networks.

$$output\_layer = softmax(fc(y_{concat})_{embeddings}) \quad (5)$$

Similarity metric such as cosine similarity is then computed for a given speaker with given input speech data. The system will decide whether the speaker verified is 'true' if the computed cosine similarity is above the set threshold and return 'false' otherwise. Model performance evaluation follows computation of EER, which is the valuation when false acceptance rate (FAR) equals false rejection rate (FRR) on detection error tradeoff (DET) curve.

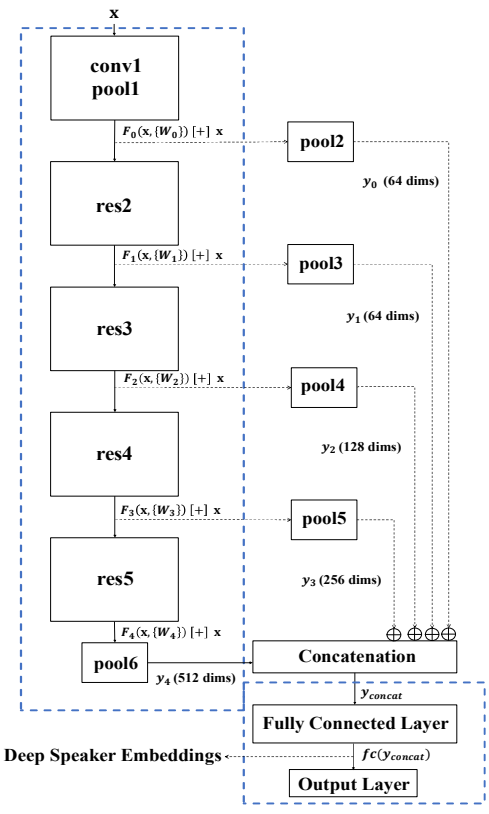

Figure 3: *Proposed shortcut connections based deep speaker embeddings. (dashed boxes: training system, dashed line: process of extracting deep speaker embeddings)*

4. Experiments

4.1. Datasets

We trained our proposed model end-to-end manner on the VoxCeleb1 [10]. It is a large scale text-independent speaker identification and verification dataset in unconstrained condition by collecting various celebrities interview in YouTube videos. It has total of 1211 speakers in the training dataset with 148,642 utterances, with 40 speakers in test dataset with 4,874 utterances. The test protocol for verification contains 37,720 pairs of trials.

4.2. Training details

On VoxCeleb1 dataset, we extracted 64-dimensional log mel-filterbank energies in a sliding window of width 25ms and shift size 10ms without additional processing speech method such as voice activity detection or noise and silence removal.

For model training, we used fixed size log mel-filterbank feature vector of size 64×300 for 3 second segments from each utterance. With batch-size 32 and percentage of train set ratio of 90 and validation set ratio of 10, these temporal feature vector segments were used to train proposed model until validation loss stop decreasing for 100 epochs. We used categorical cross-entropy (softmax loss) as loss function. Stochastic gradient descent (SGD) [23] with momentum 0.9, weight decay 10^{-8}, initial learning rate 10^{-2} reduced by 0.1 decay factor were used. As shown in Figure 4, the validation loss converges when the proposed model was trained.

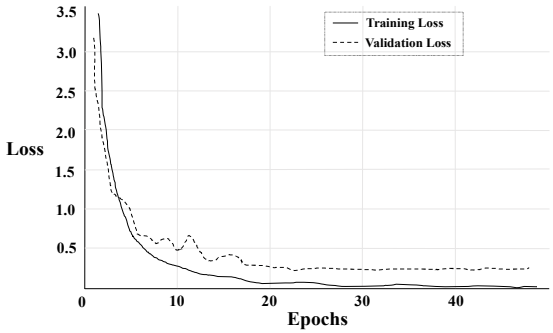

Figure 4: *Training curve of proposed model*

4.3. Evaluation and results

After training, we extracted 1024-dimensional deep speaker embeddings from the trained model based end-to-end manner without any additional method such as retraining on embeddings [10, 11], length normalization [17], and additive margin softmax loss [24]. Extracted deep speaker embeddings was computed by cosine similarity metric and was evaluated using EER performance metric.

We evaluated proposed model on voceleb1 test set. Our model was evaluated with EER performance when comparing with previous models. Previous models for comparison were traditional GMM-UBM [10], i-vector/PLDA system [10] and VGG-M [10], ResNet-34 with angular softmax/PLDA [15], TDNN based model called x-vector [25, 26], ResNet-20 with additive margin softmax loss [27], and L2-normalized deep embeddings with end-to-end manner [17]. In order to study the effect of proposed shortcut connections, we tested ResNet-18 with three fully connected hidden layers (ResNet-18-h3). All of these models were trained and tested on VoxCeleb1 dataset. As Shown in Table 2, proposed model achieved EER of 3.03%, which was relative error reduction of 21.3% over previous state-of-the-art performance on voxceleb1 dataset.

Table 2: *Speaker verification performance in comparing with proposed model and previous models (lower is better)*

Description	Model	Feats	Dims	EER (%)
Nagrani *et al.* [10]	GMM-UBM	-	-	15.0
Nagrani *et al.* [10]	i-vector/PLDA	-	-	8.80
Nagrani *et al.* [10]	VGG-M	spec-512	1024	7.80
Cai *et al.* [15]	ResNet-34	fb-64	128	4.40
Cai *et al.* [17]	ResNet-34	fb-64	128	4.74
Hajibabaei *et al.* [27]	ResNet-20	spec-512	128	4.30
Okabe *et al.* [25]	TDNN	mfcc-40	1500	3.85
Our experiment	ResNet-18-h3	fb-64	512	4.30
Our experiment	Proposed Model	fb-64	1024	**3.03**

Additionally, we tested for effect of deep speaker embeddings dimension in our model. Various size of dimension has been constructed by concatenating the outputs at bottleneck architecture. In Table 3, pool6 represents the output of pool6 which is activations of the last hidden layer depicted in Table 1. In second row, pool6 [+] 64 (576) means that it is a concatenation of first row (pool6, 512 dims) and output with the next bottleneck architecture (pool2, 64 dims) with total dimension of 576, and so on. As result, we conclude that higher dimensional deep speaker embeddings tend to show better performance.

Table 3: *Effect of deep speaker embeddings dimension on speaker verification performance (lower is better)*

Index	Dims	Params (M)	EER (%)
pool6	512	12.58	4.30
pool2	pool6 [+] 64 (576)	12.86	4.20
pool3	pool2 [+] 64 (640)	13.18	3.79
pool4	pool3 [+] 128 (768)	13.87	3.31
pool5	pool4 [+] 256 (1024)	15.56	**3.03**

5. Conclusions

In this paper, we propose shortcut connections based deep speaker embeddings for end-to-end speaker verification system. We construct speaker verification model by modifying ResNet-18 model in order to incorporate the pooled activation outputs from bottleneck architecture by shortcut connections to speaker embeddings. Speaker verification model has produced state-of-the-art EER performance (EER of 3.03%) on VoxCeleb1 data in end-to-end manner. We also conclude that higher dimensional deep speaker embeddings tend to show better performance in our experiments.

6. Acknowledgements

This work was supported by the Technology Innovation Program (10080681, Technical development of Korean speech recognition system in vehicle) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

7. References

[1] A. Irum and A. Salman, "Speaker verification using deep neural networks: a review," *International Journal of Machine Learning and Computing*, vol. 9, no. 1, pp. 20–25, 2019.

[2] N. Dehak, P. 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, pp. 788–798, 2011.

[3] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV 2007 – 11th IEEE International Conference on Computer Vision, October 14-20, Rio de Janeiro, Brazil, Proceedings*, 2007, pp. 1–8.

[4] N. Brummer and E. De Villiers, "The speaker partitioning problem," in *Odyssey 2010 – The Speaker and Language Recognition Workshop, June 28-July 1, Brno, Czech Republic, Proceedings*, 2010, pp. 194–201.

[5] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, August 27-31, Florence, Italy, Proceedings*, 2011, pp. 249–252.

[6] J. Villalba and N. Brummer, "Towards fully bayesian speaker recognition: Integrating out the between-speaker covariance," in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, August 27-31, Florence, Italy, Proceedings*, 2011, pp. 505–508.

[7] K. Chen and A. Salman, "Learning speaker-specific characteristics with a deep neural architecture," *IEEE Transactions on Neural Networks*, vol. 22, no. 11, pp. 1744–1756, 2011.

[8] K, Chen and A. Salman, "Extracting speaker-specific information with a regularized Siamese deep network," in *NeurIPS 2011 – 25th Conference on Neural Information Processing Systems, December 12-17, Granada, Spain, Proceedings*, 2011, pp. 298–306.

[9] E. Variani, X. Lei, E. McDermott, I. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *ICASSP 2014 – 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, April 19-24, Brisbane, Australia, Proceedings*, 2014, pp. 4052–4056.

[10] A. Nagrani, J. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," In *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 20-24, Stockholm, Sweden, Proceedings*, 2017, pp. 2616–2620.

[11] J. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: deep speaker recognition," in *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings*, 2018, pp. 1086–1090.

[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 *SLT 2016 – 2016 IEEE Workshop on Spoken Language Technology, December 13-16, San Diego, U.S.A., Proceedings*, 2016, pp. 165-170.

[13] 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.

[14] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 20-24, Stockholm, Sweden, Proceedings*, 2017, pp. 999–1003.

[15] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Odyssey 2018 – The Speaker and Language Recognition Workshop, June 26-29, Les Sables-d'Olonne, France, Proceedings*, 2018, pp. 74–81.

[16] L. Wan, Q. Wang, A. Papir, and I.L. Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP 2018 – 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, April 15-20, Calgary, Canada, Proceedings*, 2018, pp. 4879–4883.

[17] W. Cai, J. Chen, and M. Li, "Analysis of length normalization in end-to-end speaker verification system," in *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings*, 2018, pp. 3618–3622.

[18] K. He, X. Zhang, S. Ren, and J. Sun "Deep residual learning for image recognition," in *CVPR 2016 – 29th IEEE Conference on Computer Vision and Pattern Recognition, June 26-July 1, Las Vegas, U.S.A., Proceedings*, 2016, pp. 770–778.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR 2015 – 28th IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, Boston, U.S.A., Proceedings*, 2015, pp. 815–823.

[20] Y. Chen, I. Lopez-Moreno, and T. Sainath, "Locally-connected and convolutional neural networks for small footprint speaker recognition," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, Proceedings*, 2015, pp. 1136–1140.

[21] E. Malykh, S. Novoselov, and O. Kudashev, "On residual CNN in text-dependent speaker verification task," in *SPECOM 2017 – 19th International Conference on Speech and Computer, September 12-16, Hatfield, U.K., Proceedings*, 2017, pp. 593-601.

[22] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML 2010 – 17th International Conference on Machine Learning, June 21-24, Haifa, Israel, Proceedings*, 2010, pp. 807–814.

[23] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[24] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive Margin Softmax for Face Verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[25] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings*, 2018, pp. 2252–2256.

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP 2018 – 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, April 15-20, Calgary, Canada, Proceedings*, 2018, pp. 5329–5333.

[27] M. Hajibabaei and D. Dai, "Unified Hypersphere Embedding for Speaker Recognition," *arXiv preprint arXiv*:1807.08312, 2018.