# Speaker-discriminative Embedding Learning via Affinity Matrix for Short Utterance Speaker Verification

Junyi Peng*, Rongzhi Gu*, Yuexian Zou* and Wenwu Wang†
* Peking University Shenzhen Graduate School, Shenzhen, China
E-mail: zouyx@pku.edu.cn
† Centre for Vision, Speech and Signal Processing, University of Surrey, UK

***Abstract*—Text-independent short utterance speaker verification (TI-SUSV) task remains more challenging compared to the full-length utterance SV task due to inaccurately estimated feature statistics or insufficient distinguishable speaker embeddings. It is noted that recently developed end-to-end SV systems (E2E-SV) achieve the state-of-the-art on several datasets, which directly learn a mapping from speech features to the compact fixed length speaker embeddings. In this study, following the E2E-SV pipeline, we strive to further improve the accuracy of TI-SUSV task. Our research is based on two intuitive ideas: better speech feature representation for SUs and better training loss function to obtain more discriminative embeddings. Specifically, a bi-directional gated recurrent unit network with residual connection (Res-BGRU) is firstly designed to improve feature representation capability. Secondly, a novel affinity loss is proposed where the mini-batch data has been manipulated to obtain more supervision information. In details, a speaker identity affinity matrix formed by one-hot speaker identity vectors is taken as the supervisor of the speaker embedding affinity matrix to obtain better inter-speaker separability and intra-speaker compactness. Experimental results on the Voxceleb1 dataset show that our system outperforms a conventional i-vector and x-vector system on TI-SUSV.**

## I. Introduction

Speaker verification (SV) is the task to verify whether a speech segment belongs to a claimed identity. According to the restriction of the speech content, speaker verification is usually classified into text-dependent speaker verification and text-independent speaker verification.

For decades, i-vector+PLDA approach has been one of the most accomplished approaches for text-independent speaker verification [1], where variable-length utterances are mapped to the fixed-size low-dimensional feature vectors. However, research shows that the performance of the i-vector+PLDA approaches degrades drastically for short utterance SV task (SU-SV). Taking the equal error rate (EER) as a performance measure, one example of EER versus speech duration is shown in Fig. 1. It is clear to see that when the utterance duration becomes shorter than 3 seconds, EER is larger than 12% which is prohibitive for many applications. Further analysis shows that performance degradation may come from linguistic content limitation and nonnegligible statistical variations [2].

To improve the performance of the SU-SV systems, several deep learning methods have been proposed and achieved the state-of-the-art on several public SV datasets. In [3], researchers cast the SU-SV together with the keyword detection as a multi-task learning task, where the additional text phoneme information has been employed to reduce the EER of the SU-SV task effectively. From another perspective, the feature of voice quality is used as prior information to reduce the within-speaker embedding variabilities [4]. Their experimental results validate the performance improvement especially when the speech content was mismatched for SU-SV task.

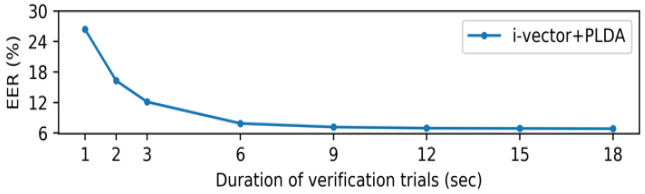


Fig. 1. EER of i-vector+PLDA versus speech durations (verification trials with Voxceleb1). EER increases from 6.8% to 26.4% when speech duration reduces from 18 second to 1 seconds.

Since the SV task is essentially a binary classification task, feature representation of speakers is crucial. In [5, 6, 7], deep neural networks with softmax loss are designed to learn the effective feature representation for speakers and categorical labels for the SU-SV task. At the verification stage, the output from the last hidden layer of the trained model is taken as the speaker embedding. However, the softmax loss does not essentially encourage the speaker embeddings to have inter-speaker separability and intra-speaker compactness. Recently, using different end-to-end loss functions, such as triplet loss [8] and contrastive loss [9], to train speaker discriminative embeddings has drawn more attention[10, 11]. These metric learning based methods achieved the state-of-the-art performance on their self-built short utterance datasets, which demonstrate the effectiveness of the end-to-end deep network for feature learning with the pair-wise loss for SU-SV task. Motivated by these work, we conducted an in-depth analysis and have the following observations: 1) the end-to-end SV model with metric learning approach is effective since the distance information between generated speaker embeddings can be manipulated to enhance the learning ability of the deep

model; 2) the pair-wise loss asks for careful pair selection strategies to avoid suboptimal local minima. One typical pair selection scheme is to increase the weight of hard samples. Hence, the pure data-driven metric learning approach is much more attractive since it does not need to specifically select the training pairs.

Bearing above analysis in mind, in this study, we build a bi-directional gated recurrent unit network with residual connection (Res-BGRU) which directly maps the variable-length hand crafted features to the fixed-size speaker embeddings. Moreover, the speaker identity affinity matrix of the one-hot speaker identity vectors is taken as the supervision information and a novel affinity loss is derived to simultaneously maximize the inter-speaker separability and intra-speaker compactness. Specifically, our proposed affinity loss makes use of all correlation information between the speaker embedding pairs, therefore more discriminative information and better robustness are achieved. Finally, the whole system is trained by optimizing the affinity loss in an end-to-end manner. Experimental results validate the improved performance of our method.

The remainder of the paper is organized as follows. Section II details our proposed SU-SV system. Section III presents the experimental setup and results. Section VI concludes this paper.

## II. Proposed System

As discussed above, in this study, we propose an end-to-end speaker verification system for improving the performance of SU-SV task. The configuration of our designed deep model is shown in Fig. 2. In Fig. 2, BGRU denotes a bi-directional gated recurrent unit layer and the ResBlock consists of one BGRU and one batch normalization layer. The statistic pooling layer is used to convert the frame-level features into a fixed-size representation. Two feedforward fully connection layers are designed to extract speaker-discriminative embeddings. Clearly, our model essentially maps a batch of handcrafted feature vectors (B samples in Fig. 2) into a speaker embedding matrix (S in Fig. 2). To simplify the presentation and distinguish it from other methods, our proposed end-to-end SV system is named as Res-BGRU and the details will be given in part A. Our Res-BGRU is optimized with our proposed affinity loss, which will be introduced in part B. The verification process will be described in part C.

### A. System Design

**Frame-level Feature Extraction**. Motivated by the powerful modeling capability of the recurrent networks for temporal sequence [12], in our design, BGRU is employed to map the variable-length frame-level features to the fixed-length frame-level feature. Compared to the popular Long Short-Term Memory (LSTM), BGRU is easier to train and has a faster convergence [13]. Moreover, the bi-directional structure has the ability to capture past and future information respectively. Besides, motivated by the work in [14], we design to use the residual connections to improve the feature extraction

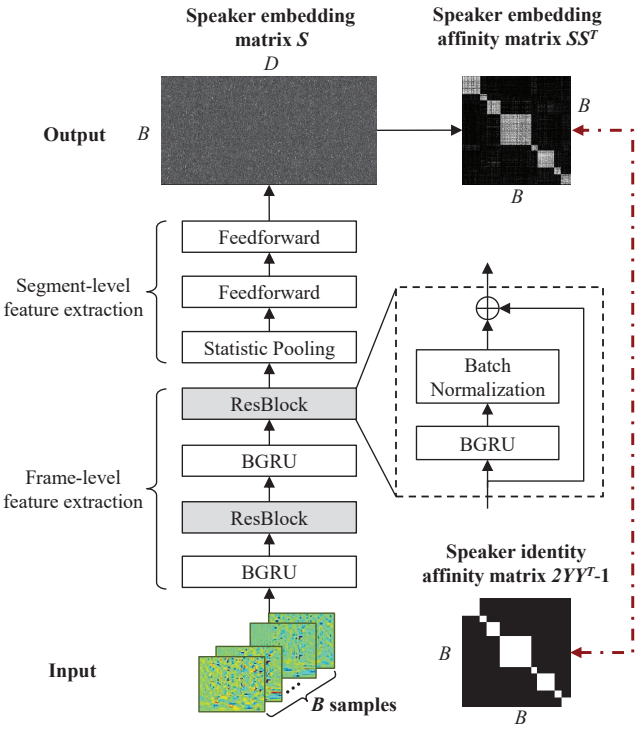


Fig. 2. The architecture of our proposed Res-BGRU system. B denotes the batch size of input data, D is the dimension of output speaker embedding.

capability of the deep networks and speed up its convergence. Combing these two techniques, the ResBlock is delicately designed and the details are shown in Fig. 2.

**Segment-level Feature Extraction**. To aggregate the feature over time steps, following the design in [15], a statistics pooling layer approach is adopted where the frame-level features are taken as its input to generate a single segment-level feature vector. For further extracting the higher level feature vectors, two feedforward fully connection layers are designed to generate D-dim speaker embeddings.

Considering the possible interferences in speech signals, the noise robustness is another issue where the salient speaker information is expected to be preserved and insensitive to noise interferences. Motivated by the work in [16, 17], a Max Feature Map (MFM) is taken as the activation function which is different from the commonly used Rectified Linear Units (ReLU). ReLU discards values smaller than zero while MFM divides output nodes equally and outputs the one with maximum value in element-wise, defined as follows:

$$o_m = max(z_m + z_{m+\frac{M}{2}}) \tag{1}$$

where $1 \leqslant m \leqslant M/2$, $M$ denotes the number of nodes in feedforward layer, $z$ denotes the input tensor, $o \in \mathbb{R}^{\frac{M}{2}}$ represents the output of the MFM operation.

### B. Affinity Loss

Intuitively, the speaker embeddings generated by our deep model are expected to provide good inter-speaker discrimi-

nation and intra-speaker compactness. To achieve this target, we propose a novel affinity loss function (AL) which exploits the correlation information of all speaker embedding pairs of a data batch. It is noted that the discriminability of speaker embeddings extracted by the deep model using contrastive loss and triplet loss depends on the carefully designed pair selection strategy. Our motivation is to remove the pair selection constrain and make use of the information provided by the speaker embedding matrix (in Fig. 2) to enhance the inter-speaker discrimination and intra-speaker compactness. Our method is introduced as follows.

As shown in Fig. 2, assume that our proposed Res-BGRU is parameterized by $\theta$. Res-BGRU maps the feature vector $\mathbf{x}$ to a $D$-dimensional unit-norm speaker embedding $s = f_\theta(\mathbf{x}) \in \mathbb{R}^{1\times D}$, i.e. $\|s\|^2 = 1$. The one-hot label vector $\mathbf{y} \in \mathbb{R}^{1\times D}$ indicates the corresponding speaker identity of $\mathbf{x}$, where $N$ is the total number of speakers involved in training set. While training in a mini-batch, $B$ speech segment features are randomly selected to form a set as $\mathbf{X} = \{\mathbf{x}_b\}_{b=1}^{B}$. Correspondingly, the output of Res-BGRU and label matrix associated with $\mathbf{X}$ can be denoted as $\mathbf{S} = \{\mathbf{s}_b\}$ and $\mathbf{Y} = \{\mathbf{y}_b\}$, respectively. Here $\mathbf{S}$ is named as the speaker embedding matrix and $Y$ is the speaker identity matrix. Following mathematic notations, the matrix $\mathbf{SS}^T \in \mathbb{R}^{B\times B}$ is termed as the speaker embedding affinity matrix and $\mathbf{YY}^T \in \mathbb{R}^{B\times B}$ as speaker identity affinity matrix. In this study, we aim to fully explore the speaker information and the similarity of the speaker embedding pairs in a data batch. A novel affinity loss is proposed as follows:

$$\begin{aligned}\mathcal{L} &= \|\mathbf{1} - \mathbf{SS}^T \odot \mathbf{YY}^T\|_F^2 + \|-\mathbf{1} - \mathbf{SS}^T \odot (\mathbf{1} - \mathbf{YY}^T)\|_F^2 \\ &= \sum_{\substack{i,j \\ \mathbf{y}_i=\mathbf{y}_j}} (1 - cos(\mathbf{s}_i, \mathbf{s}_j))^2 + \sum_{\substack{i,j \\ \mathbf{y}_i\neq\mathbf{y}_j}} (-1 - cos(\mathbf{s}_i, \mathbf{s}_j))^2\end{aligned} \tag{2}$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm and $\odot$ is the Hadamard product. It is noted that $(\mathbf{SS}^T)_{i,j} = \mathbf{s}_i \cdot \mathbf{s}_j^T$ indicates the cosine similarity between $\mathbf{s}_i$ and $\mathbf{s}_j$. If segment $i$ and $j$ belong to the same speaker, then the cosine similarity between $\mathbf{s}_i$ and $\mathbf{s}_j$ should be close to 1. Also, $\mathbf{YY}^T$ is a binary matrix, specifically, if the segment $i$ and $j$ belong to the same speaker (with the same one-hot label vector) then we have $(\mathbf{YY}^T)_{i,j} = 1$. Otherwise, we have $(\mathbf{YY}^T)_{i,j} = 0$. Under a supervised learning framework, $\mathbf{YY}^T$ is known which can be calculated using the training data.

From the second line of 2, the affinity loss is divided into two parts. (a) The first item aims at promoting the similarity between different embeddings of the same speaker; (b) The second item then aims at promoting the discrimination between the speaker embeddings from the different speakers. During the training stage, two parts in 2 are optimized simultaneously.

To ensure that the affinity loss function defined in 2 is mathematically treatable in the forward and back propagation

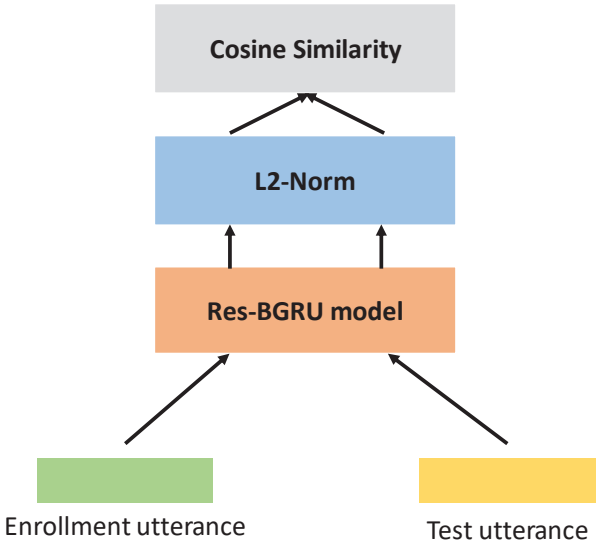


Fig. 3. Schematic diagram of a verification trial.

algorithm, with some mathematical manipulation, 2 has the following equivalent form:

$$\mathcal{L} = \|\mathbf{SS}^T - \mathbf{2YY}^T + \mathbf{1}\|_F^2 \tag{3}$$

From the above analysis, we can see that our proposed affinity loss has following advantages.: 1) Compared to commonly used softmax loss in classification tasks which generates the speaker posterior probability distribution, the affinity loss directly optimizes the similarities between speaker embeddings. Thus an end-to-end trainable system is constructed. 2) Compared with other pair-wise losses (such as the contrastive loss in [9]) which only measure the similarity between a single manually selected pair in training stage, our proposed affinity loss makes use of all speaker embedding pairs which provides more supervision information to achieve better robustness. 3) Compared to the generalized end-to-end loss [18] proposed recently, our proposed affinity loss does not rely on the pair selection strategy and the mini-batch training data can be selected randomly. This property makes the network training more flexible and stable.

### C. Verification Process

After the Res-BGRU is well trained, it works as a feature extractor to generate the utterance-level speaker embeddings. In this study, the cosine similarity is used to measure the similarity score between two speaker embeddings. The final similarity score will be compared against a pre-defined threshold. If the final similarity score exceeds the threshold, the system will accept the claimed speaker and reject otherwise.

## III. Experiments and Analysis

In this section, we first describe the dataset used for the experiments, and then report and analyze the experimental results accordingly.

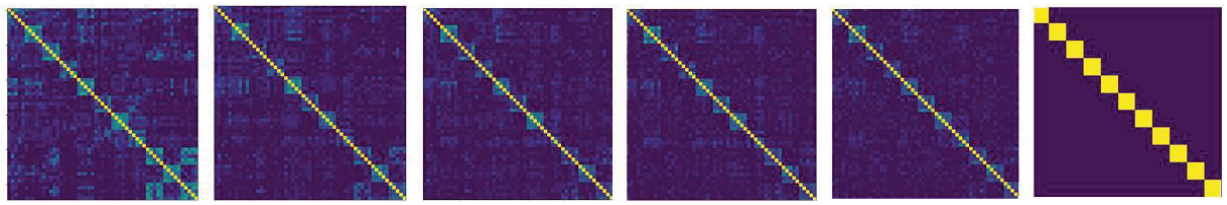


Fig. 4. The visualization of speaker embedding affinity matrix $\mathbf{S}\mathbf{S}^T$ at epochs. The depth of the color is proportional to the level of cosine distance (e.g., brighter color refers to higher similarity). The rightmost is the target speaker identity affinity matrix $2\mathbf{Y}\mathbf{Y}^T - \mathbf{1}$.

TABLE I
ARCHITECTURE OF THE RES-BGRU. T IS THE TOTAL TIME STEP OF THE INPUT SIGNAL, AND MFM IS MAX FEATURE MAP ACTIVATION FUNCTION.

| Layer | Input | Output | Params |
|---|---|---|---|
| BGRU | $T \times 39$ | $T \times 512$ | 456K |
| ResBlock[BGRU] | $T \times 512$ | $T \times 512$ | 1182K |
| BGRU | $T \times 512$ | $T \times 512$ | 1182K |
| ResBlock[BGRU] | $T \times 512$ | $T \times 512$ | 1182K |
| Statistics Pooling | $T \times 512$ | 1024 | - |
| FC1 | 1024 | 1024 | 1049K |
| MFM | 1024 | 512 | - |
| FC2 | 512 | 1024 | 525K |
| MFM | 1024 | 512 | - |
| Total | - | - | 5565K |

### A. Dataset

We evaluate the performance of our method the on Voxceleb1 dataset [19] since it is so far the latest and largest public speaker verification dataset. Specifically, this dataset consists of about 150,000 utterances from 1251 different celebrities. Each celebrity has at least 45 utterances with different lengths. For the speaker verification task, the training set includes 1211 celebrities and 140,664 utterances. The test set comprises 4,715 utterances from 40 celebrities whose names are with an initial E. The performance is reported in terms of EER.

### B. Experiment Setup

The raw features of speech segments consist of 13-dimensional MFCCs and then appended to 39-dimensional speech features by their delta and acceleration. To accelerate the convergence speed, mean and variance normalization are performed on every feature dimension. For the SU-SV task, the utterances in the dataset for the SV task are randomly clipped to 3s during training. The parameter details of our proposed Res-BGRU system are presented in Table I. We train our Res-BGRU model by Keras [20]. The batch size is set as 128 and the weight decay is set to 5e-5 for fully-connected layers. The RMSprop optimizer is used with an initial learning rate of 0.001. Learning rate is decayed if the validation loss has not decreased. We use Xavier as parameter initializer for all layers. In order to get a comprehensive assessment of our proposed system, the following two state-of-the-art speaker verification approaches have been included for performance comparison.

**i-vector**: The i-vector baseline is based on GMM-UBM Kaldi SRE10 V1, as described in [21]. The UBM is composed of 2048 Gaussian components. The dimension of i-vectors is set to 400 and the i-vector extractor is trained by using 60-dimensional MFCC speech features. LDA and PLDA are taken as the scoring functions. All the training dataset is used to train the UBM, T-Matrix and PLDA

**x-vector**: The x-vector baseline is also implemented by the Kaldi toolkit. Five layers of TDNN with the ReLU activation function are used as the frame-level feature extractor and a statistics pooling layer is used to aggregate all the frame-level features. In the end, two fully connected layers are designed to produce the segment-level feature [22].

### C. Performance versus epoch number

To observe the performance of our end-to-end system on SU-SV task, one example of the speaker embedding affinity matrix estimated at different epoch is given in Fig. 4. It can be seen that EER goes down and $\mathbf{S}\mathbf{S}^T$ gets close to $2\mathbf{Y}\mathbf{Y}^T - \mathbf{1}$ with the increase of the epoch. This process indicates that, when the model goes to convergence, the speaker embeddings of the same speaker become similar, while the speaker embeddings of different speakers become dissimilar. As a result, we can conclude that our Res-BGRU model trained with the affinity loss increases the discrimination of learned speaker embeddings.

### D. Results for Short Utterance Speaker Verification

This experiment evaluates the EER performance versus the utterance length. In the verification trial, the duration of both the enrollment and test utterances is set to 3, 6 or 9 seconds. Compared to the large-scale dataset, the training data used is still insufficient. To avoid suboptimal local minima in training, we take a two-stage policy for training Res-BGRU. First, the pre-training is conducted with softmax loss. Second, the fine-tuning is carried with the affinity loss or triplet loss. Both stages are trained with the same dataset.

From Table II, it is encouraging to see that, for both 3s-3s 6s-6s and 9s-9s conditions, our Res-BGRU system with affinity loss (Res-BGRU (AL)) outperforms other compared approaches. Our Res-BGRU system with softmax loss (Res-BGRU (SL)) ranks second. Obviously, i-vector+LDA and

TABLE II
EER (%) PERFORMANCE UNDER 3S-3S, 6S-6S AND 9S-9S CONDITIONS. 3S-3S, 6S-6S AND 9S-9S RESPECTIVELY DENOTE BOTH LENGTHS OF ENROLLMENT, AND THE TEST UTTERANCES OF A VERIFICATION TRIAL ARE IN 3, 6 AND 9 SECONDS.

| System | 3s-3s | 6s-6s | 9s-9s |
|---|---|---|---|
| i-vector+LDA | 19.08 | 13.70 | 12.10 |
| i-vector+PLDA | 12.11 | 7.88 | 7.17 |
| x-vector+PLDA | 7.78 | 6.67 | 5.71 |
| Res-BGRU (SL) | 6.37 | 5.80 | 4.77 |
| Res-BGRU (Triplet Loss) | 6.12 | 5.67 | 4.57 |
| Res-BGRU (AL) | **5.98** | **5.52** | **4.30** |

i-vector+PLDA are not competitive under short utterances conditions. The results also confirm that the deep models have a better capability to learn a more discriminative feature representation than the statistical model-based methods, especially for short utterances. Specifically, when using softmax loss for training, the Res-BGRU (SL) system achieves a relative EER improvement of 18.12%, 13.04% and 16.46% over x-vector+PLDA in 3s-3s, 6s-6s and 9s-9s conditions respectively, which indicates that our end-to-end system has more advantages in feature extraction under short utterance conditions.

Compared with Res-BGRU (SL), the Res-BGRU (AL) system gains 6.12%, 4.83% and 9.85% relative EER improvement in 3s-3s, 6s-6s and 9s-9s conditions respectively, which demonstrate that the affinity loss is capable of reducing the intra-speaker compactness and increasing the inter-speaker separateness under short utterance conditions. Res-BGRU(AL) performs better than Res-BGRU(Triplet Loss). This suggests that using the information of all pairs in a data batch is able to provide more supervision information than using a single manually selected pair. Moreover, from the intrinsic idea in developing our Res-BGRU and affinity loss, we believe that the performance for the SU-SV task could be further improved by training the network with longer utterances.

## IV. CONCLUSIONS

We have presented an effective end-to-end SU-SV system to improve the performance of the speaker verification for short utterances. In our Res-BGRU model, the BGRU is employed to deal with the variable-length speech segments and the MFM is used in the forward layers to filter out speaker irrelevant information. A novel loss, i.e. the affinity loss, has been proposed to leverage the correlations of all speaker embeddings to maximize the intra-speaker compactness and inter-speaker separateness. This leads to an efficient training process. Experiments show that our proposed system achieves significant improvements over the state-of-the-art baselines

## ACKNOWLEDGMENT

This paper was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ20170817160058246 & JCYJ20180507182908274). Special acknowledgements are given to AOTO-PKUSZ Joint Research Center for Artificial Intelligence on Scene Cognition & Technology Innovation for its support.

## REFERENCES

[1] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[2] Arnab Poddar, Md Sahidullah, and Goutam Saha. Speaker verification with short utterances: a review of challenges, trends and opportunities. *IET Biometrics*, 7(2):91–101, 2017.

[3] Shi-Xiong Zhang, Zhuo Chen, Yong Zhao, Jinyu Li, and Yifan Gong. End-to-end attention based text-dependent speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 171–178. IEEE, 2016.

[4] Soo Jin Park, Gary Yeung, Jody Kreiman, Patricia A Keating, and Abeer Alwan. Using voice quality features to improve short-utterance, text-independent speaker verification systems. In *INTERSPEECH*, pages 1522–1526, 2017.

[5] Gautam Bhattacharya, Md Jahangir Alam, and Patrick Kenny. Deep speaker embeddings for short-duration speaker verification. In *Interspeech*, pages 1517–1521, 2017.

[6] Zhifu Gao, Yan Song, Ian McLoughlin, Wu Guo, and Lirong Dai. An improved deep embedding learning method for short duration speaker verification. *Proc. Interspeech 2018*, pages 3578–3582, 2018.

[7] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4052–4056. IEEE, 2014.

[8] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[9] Sumit Chopra, Raia Hadsell, Yann LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, pages 539–546, 2005.

[10] Chunlei Zhang and Kazuhito Koishida. End-to-end text-independent speaker verification with triplet loss on short utterances. In *Interspeech*, pages 1487–1491, 2017.

[11] Sergey Novoselov, Vadim Shchemelinin, Andrey Shulipa, Alexandr Kozlov, and Ivan Kremnev. Triplet loss based cosine similarity metric learning for text-independent speaker recognition. *Proc. Interspeech 2018*, pages 2242–2246, 2018.

[12] Jian Kang, Wei-Qiang Zhang, and Jia Liu. Gated recurrent units based hybrid acoustic models for robust speech recognition. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE, 2016.

[13] Dario Amodei, Sundaram Ananthanarayanan, Rishita

Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, pages 999–1003, 2017.

[16] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[17] Sergey Novoselov, Andrey Shulipa, Ivan Kremnev, Alexandr Kozlov, and Vadim Shchemelinin. On deep speaker embeddings for text-independent speaker recognition. *arXiv preprint arXiv:1804.10080*, 2018.

[18] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE, 2018.

[19] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[20] François Chollet et al. Keras, 2015.

[21] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. Technical report, IEEE Signal Processing Society, 2011.

[22] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.