Speaker Verification System Based on Deformable CNN and Time-Frequency Attention

Yiming Zhang*, Hong Yu†, Zhanyu Ma *
* Beijing University of Posts and Telecommunications, Beijing 100876, China
† Ludong University, Shandong, China

***Abstract*—Speaker verification (SV), especially short utterances SV, needs to be robust under complex noisy and far-field conditions. Majority of recent works apply attention mechanism on aggregation of frame-level speaker embeddings which are extracted by deep neural network. In this paper, a novel speaker verification system based on the deformable convolution module and the time-frequency attention module has been proposed. In the deformable convolution module, the convolutional sampling locations are adaptively adjusted by additional offsets which are learnt from the spectrogram. Meanwhile, in order to extract the more effective speaker discrimination information for short utterances, the time-frequency attention module is used to help the system focus on the important regions of the short utterances along the time and the frequency domain. Experiments on the HI-MIA database show that the proposed modules can improve the equal error rate (EER) of speaker verification system by relatively 24% compared to the baseline model, at a result of 8.51%.**

I. Introduction

Speaker verification (SV) is an important biometric recognition technology. It is developed to verify the identity of speakers by speech signals [1]. SV [2] system is widely used in telephone or network authentication systems such as call center, telephone banking or apartment security. The SV problem can be divided into two subcategories: text-dependent (TD)-SV and text-independent (TI)-SV. In this paper, we will focus on the TD-SV task.

In the past decade, the most popular method for SV was i-vector [3] and probabilistic linear discriminant analysis (PLDA) [4] framework. Recently, with the widely used of deep learning in SV systems, the performance of SV has been significantly improved. The deep neural network (DNN) and convolution neural network (CNN) based methods, such as d-vector [5,6,7] and x-vector [8], have further surpassed statistical-based SV methods, e.g, GMM-UBM and i-Vector+PLDA. However, now available SV systems can be easily influenced by various factors, such as the voice speed, the tone of speech, the background noise and so on. Our previous work [9] has proved that the performance of DNN based SV system will heavily decrease under the complex noisy environment. In order to reduce the influence of far-field and complex noisy environment, researches try to improve the noise robustness of SV systems during the last decade. In the front-end domain, DNN based speech enhancement methods are applied to reduce the adverse impacts of reverberation and noise [10,11]. Some training data argument methods are used to improve the robustness of SV systems [12], by adding some reverberation and different kinds of noise into training utterances to simulate the real-world conditions. In the back-end domain, in order to extract the distinguishable speaker embeddings, many deep neural networks based on attention mechanism have been proposed and achieve remarkable performance. Bhattacharya [7] utilized self-attention for aggregating frame-level embeddings. Geng and Okabe [13,14] combined a weighted statistics pooling layer with statistics pooling to get utterance embeddings.

However, there are still some drawbacks in deep speaker verification methods. First, too much prior knowledge has been used in speech enhancement processing, e.g, noise types and signal to noise ratios (SNRs), which leads to poor performance of SV systems under unknown conditions. Besides, due to the fixed geometric structures of convolution kernel, convolutional neural networks cannot focus on the "interesting" regions effectively. The last but most important problem is that only time domain information, instead of information of the frequency domain, has been considered within current attention based methods.

To address the issues above, in this paper, a novel speaker verification system based on the deformable CNN [15,16] and the time-frequency attention mechanism, namely ResdefNet, is proposed. Different from the traditional SV system, deformable convolutional layer is used to address the problem. Deformable convolutional layer will shift the sampling points of the input feature map, and make them focus on the "interesting" regions. Besides, in order to extract the distinguishable speaker embeddings from the short utterances, the frequency domain is also believed to contain discriminative regions. Therefore, a novel time-frequency attention mechanism for ResdefNet is created and evaluated, so that not only the time domain information will be processed, but also the classification information in frequency domain has also been introduced.

The rest of this paper is organized as follows. Section II presents the new deep speaker verification system in detail. The comparative experiments and results are presented in Section III, and the conclusion is given at Section IV.

II. Deep Speaker Verification System

In order to extract the distinguishable speaker embeddings, we propose a novel deep SV system based on the deformable

Hong Yu is the corresponding author.

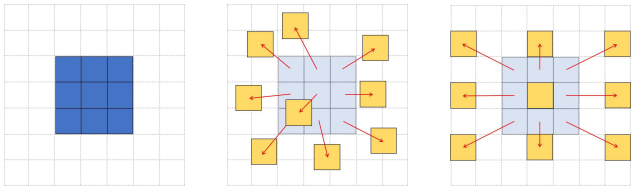

Fig. 1. The sampling points in 3×3 traditional and deformable convolutions. (a) regular sampling grid (blue squares) of traditional convolution. (b)(c) deformed sampling points (yellow squares) with offsets (red arrows) in deformable convolution. (c) is a special case of (b).

convolution module and the time-frequency attention module. This section introduces the module structures of the deformable convolution and time-frequency attention mechanism used in our system.

A. Deformable convolution module

As discussed above, the speech information is not with "regular structure" in feature maps, the traditional convolutional unit which samples on feature maps at fixed locations can not cover the discriminative regions well. In order to improve the flexibility of the convolution unit, we introduce a location deformable sampling method. Different from the regular sampling grid used in traditional convolution layers (Fig.1 (a)), in the deformable convolution layer, points in sampling grid are augmented by some trainable offsets [15]. (Fig1 (b) and (c)). The offsets are learned from the input feature maps by an additional convolution layer, which make the sampling grid pay more attention to the regions that can help the system to distinguish different speakers.

The procedures of the deformable convolution module are shown in Fig. 2:

1) The offset field [15] is obtained by applying a convolutional layer over the input feature map with channel dimension N. The offset field has the same spatial resolution with the input feature map and the channel dimension is 2N, which means two offsets Δx and Δy are leaned for each point of input feature map. In order to prevent the offsets exceed from the range of the input feature map, the clamp function is used to limit the value of the offsets.
   For example, given a pixel P in an input feature map with location (x_p, y_p) and its value is $V_{old}(x_p, y_p)$, the offset pair corresponding to point P in offset fields is $(\Delta x_p, \Delta y_p)$.
2) To facilitate the implementation, instead of arguing sampling grids directly, we change input feature maps into deformable feature maps[16] with learned offsets. The shape of the deformable feature map is same as the input feature map.
   The value of point P in deformable feature maps, $V_{new}(x_p, y_p)$, can be computed by:
$$V_{new}(x_p, y_p) = V_{old}(x_p + \Delta x_p, y_p + \Delta y_p). \quad (1)$$
   Since Δx_p and Δy_p are fractional locations, the value of $V_{old}(x_p + \Delta x_p, y_p + \Delta y_p)$ is replaced by the values of four integer positions around by using the bilinear interpolation method. Then we have :
$$V_{\text{old}}(x_{\text{new}}, y_{\text{new}}) = \sum_{m \in \mathcal{M}} w_{mp} \cdot V_{\text{old}}(x_m, y_m), \quad (2)$$
   where $(x_{new}, y_{new}) = (x_p + \Delta x_p, y_p + \Delta y_p)$ and $\mathcal{M}$ is the set of four integer points closest to P. w_{mp} is the weight for the integer position (x_m, y_m), and can be calculated by
$$w_{mp} = (1 - |x_m - x_p|) \cdot (1 - |y_m - y_p|).$$
3) Traditional convolutions are operated on the deformable feature map and the output can be calculated as:
$$y(x, y) = \sum_{k_w} w_{ij} \cdot V_{new}(x_{ij}, y_{ij}), \quad (3)$$
   where k_w enumerates all the locations of the kernel and w_{ij} represents the corresponding weight of the kernel.

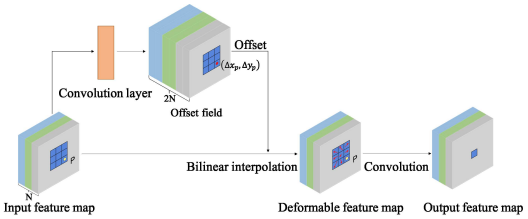

Fig. 2. Illustration of the deformable convolution. The offset field is obtained by applying convolutions on the input feature map to create the deformable feature map.

B. Time-Frequency attention module

The significance of attention has been extensively studied across computer vision, spoken language understanding, and natural language processing. Attention mechanism can make the model focus on the important regions and suppress unnecessary ones. However, the current attention-based speaker recognition methods only consider the information along the time axis. In order to take full advantage of frequency sequence information, we use time-frequency attention module to emphasize speaker discriminative features along both the time and the frequency domains. As shown in Fig. 3, we apply the time and frequency attention modules sequentially, so that each branch can learn time or frequency information, respectively. And the internal structure of frequency attention module is described in Fig. 4.

- *Frequency attention module:*

For modelling the frequency attention, we only focus on the frequency domain of the input feature map. The input feature map of the modules is $F_{input} \in \Re^{C \times F \times T}$, where C denotes the number of input channels, F and T denote the dimensions along the frequency and time domain, respectively.

As shown in Fig. 4. First, we aggregate the channel information of the input feature map by using the average pooling

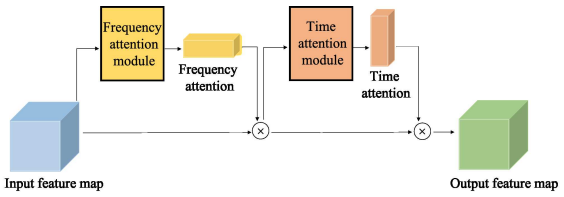

Fig. 3. Illustration of the time-frequency attention module. The module has two sequential sub-modules: time attention module and frequency attention module.

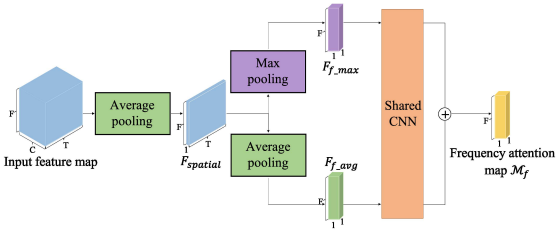

Fig. 4. The frequency attention module.

operations. Then the spatial feature descriptor $F_{spatial} \in \Re^{1\times F\times T}$ is generated. Next, we aggregate time information of the spatial feature map by using the average pooling and the max pooling operations, which generates two different frequency context descriptors, F_{f_avg} and $F_{f\_\max}$ as:

$$F_{f\_avg} = AvgPool\left(F_{spatial}\right), \tag{4}$$

$$F_{f\_max} = MaxPool\left(F_{spatial}\right). \tag{5}$$

Where $F_{f\_avg}, F_{f\_\max} \in \Re^{1\times F\times 1}$, and both F_{f_avg} and $F_{f\_\max}$ are fed into to a shared convolution network with kernel size 7×1. The output features are merged by element-wise summation to produce the frequency attention map $\mathcal{M}_f$:

$$\mathcal{M}_f = \sigma\left(f^{7\times 1}\left(F_{f\_avg}\right) + f^{7\times 1}\left(F_{f\_max}\right)\right). \tag{6}$$

- *Time Attention Module:*

The time attention module follows a similar process as the frequency attention module. However, we use the average pooling and the max pooling operations on the spatial feature map $F_{spatial}$ along the frequency axis. The kernel size of the shared convolution network is 1×7. The time attention map $\mathcal{M}_t$ is:

$$F_{t\_avg} = AvgPool\left(F_{spatial}\right), \tag{7}$$

$$F_{t\_max} = MaxPool\left(F_{spatial}\right), \tag{8}$$

$$\mathcal{M}_t = \sigma\left(f^{1\times 7}\left(F_{t\_avg}\right) + f^{1\times 7}\left(F_{t\_max}\right)\right), \tag{9}$$

where F_{t_avg}, $F_{t\_max} \in \Re^{1\times 1\times T}$.

C. Proposed pipeline

Front-end, we propose a modified 34-layer ResNet, from thin-Resnet [17] as our front-end CNN to encode the spectrogram input. The thin-Resnet has only 3 million parameters compared to the standard ResNet-34 (22 million). The architecture of ResdefNet is illustrated in Table I. ReLU and batch-norm layers are applied to convolutional output, and the utterance embeddings are extracted before the softmax output layer. N denotes the number of speakers.

TABLE I
THE ARCHITECTURE OF RRSDEFNET BASED ON THIN-RESNET

Input Spectrogram $(1\times 257\times T)$	Output Size
conv, 7×7, 64, stride $(1,1)$	$(64\times 257\times T)$
maxpool, 2×2, stride $(2,2)$	$(64\times 128\times T/2)$
$\begin{bmatrix}\text{conv},1\times 1,48\\ \text{def-conv},3\times 3,48\\ \text{conv},1\times 1,96\end{bmatrix}$ $\begin{bmatrix}\text{conv},1\times 1,48\\ \text{conv},3\times 3,48\\ \text{conv},1\times 1,96\\ \text{attention module}\end{bmatrix}$	$(96\times 128\times T/2)$
$\begin{bmatrix}\text{conv},1\times 1,96\\ \text{def-conv},3\times 3,96\\ \text{conv},1\times 1,128\end{bmatrix}$ $\begin{bmatrix}\text{conv},1\times 1,96\\ \text{conv},3\times 3,96\\ \text{conv},1\times 1,128\end{bmatrix}$ $\begin{bmatrix}\text{conv},1\times 1,96\\ \text{conv},3\times 3,96\\ \text{conv},1\times 1,128\\ \text{attention module}\end{bmatrix}$	$(128\times 64\times T/4)$
$\begin{bmatrix}\text{conv},1\times 1,128\\ \text{def-conv},3\times 3,128\\ \text{conv},1\times 1,256\end{bmatrix}$ $\begin{bmatrix}\text{conv},1\times 1,128\\ \text{conv},3\times 3,128\\ \text{conv},1\times 1,256\end{bmatrix}$ $\begin{bmatrix}\text{conv},1\times 1,128\\ \text{conv},3\times 3,128\\ \text{conv},1\times 1,256\\ \text{attention module}\end{bmatrix}$	$(256\times 32\times T/8)$
$\begin{bmatrix}\text{conv},1\times 1,256\\ \text{def-conv},3\times 3,256\\ \text{conv},1\times 1,512\end{bmatrix}$ $\begin{bmatrix}\text{conv},1\times 1,256\\ \text{conv},3\times 3,256\\ \text{conv},1\times 1,512\end{bmatrix}$ $\begin{bmatrix}\text{conv},1\times 1,256\\ \text{conv},3\times 3,256\\ \text{conv},1\times 1,512\\ \text{attention module}\end{bmatrix}$	$(512\times 16\times T/16)$
maxpool, 3×1, stride $(2,2)$	$(512\times 7\times T/32)$
conv, 7×1, 512 stride $(1,1)$	$(512\times 1\times T/32)$
embedding (FC)	512
softmax	N

Deformable convolution module and time-frequency attention module: the deformable convolution module is used in the first residual block of each residual layer and the time-frequency module is integrated in the last residual block of each residual layer, as shown from row 4 to row 7 in Table I.

In the back-end, the PLDA algorithm [4] is used to measure the similarity between the speaker embeddings and the test utterance embeddings.

III. EXPERIMENTS, RESULTS AND DISCUSSIONS

In this section, we present the database and the training details, then report the performance results.

A. Database

Our experiments were conducted on the HI-MIA database [18] which include the AISHELL-wakeup database and AISHELL-2019B-eval database. It is an open source wake-up words database. The content of utterances covers two wake-up words, "ni hao, mi ya" in Chinese and "Hi, Mia" in English. The database contains multi-channel far-field speech data that can be used for far-field speaker verification and wake-up word detection. In the HI-MIA database, it contains 340 speakers and each speaker recorded 160 utterances, with some utterances recorded in a noisy environment and the remaining utterances recorded in the home environment [18].

In this paper, we train our models on the AISHELL-wakeup dataset and use AISHELL-2019B-eval datasets to evaluate the proposed approach. In terms of the enrollment data, five utterances in the far-field condition were randomly selected as the far-field enrollment data. In addition, we also randomly

TABLE II
EER(%) RESULTS OF THE SV SYSTEMS

Method	EER(%)
x-vector	11.25
Thin-Resnet	9.44
Thin-Resnet+def	9.01
Thin-Resnet+attention	9.20
ResdefNet	**8.51**

selected five utterances recorded by close-talking microphone as near-field enrollment data.

B. Training Details

During the training stage, no voice activity detection or automatic silence removal is applied. The short-time Fourier transform was applied to extract the magnitude in frequency domain. Spectrograms are generated by using a 25 ms wide hamming window with a hop length of 10 ms, 512-point FFT corresponding to a random 2-second time crop per utterance, followed by mean and variance normalization. Adam optimizer with an initial learning rate of 0.001 decayed every 10 epochs by a factor of 0.1 is used for training.

C. Performance Evaluation

The equal error rate (EER) is used to evaluate the performance of those models. The method using x-vector [8] was considered as the baseline system. The x-vector embeddings are extracted at the last hidden layer, and the "Thin-Resnet" system extracts the embeddings using the traditional convolution without the deformable convolution module and time-frequency attention module. In the "Thin-Resnet+def" and "Thin-Resnet+attention" systems, the utterance embeddings are extracted by deformable convolution or time-frequency attention module, and the positions of these modules in the residual block are the same as ResdefNet.

The experimental results are listed in Table II. It can be observed that the "Thin-Resnet+def" and "Thin-Resnet+attention" systems significantly outperform the "Thin-Resnet" system and x-vector system, which means that these modules can help the system extract more distinguishable speaker embeddings. In addition, the ResdefNet system achieves the best performance among all the systems. Compared with the baseline systems, the ResdefNet system obtains a relative improvement of 24.0% in terms of EER.

IV. CONCLUSIONS

In this paper, we focus on text-dependent short utterance speaker verification under the far-field and noisy environment. The proposed ResdefNet system outperform the baseline system by a significant margin, achieving an EER of 8.51% on the HI-MIA database. With the help of deformable sampling locations and time-frequency attention mechanism, the proposed ResdefNet system performs well at finding the "interesting" regions that can help the system distinguish different speakers.

ACKNOWLEDGMENT

This work was supported in in part by the National Key R&D Program of China under Grant 2019YFF0303300 and under Subject II No. 2019YFF0303302, in part by the National Natural Science Foundation of China under Grant 61773071, 61922015, and U19B2036, in part by the Beijing Natural Science Foundation Project No. Z200002, in part by the Beijing Academy of Artificial Intelligence (BAAI) under Grant BAAI2020ZJ0204, in part by the Beijing Nova Program Interdisciplinary Cooperation Project under Grant Z191100001119140, in part by the Key Research and Development project of Shandong Province No. 2019GGX101036.

REFERENCES

[1] Wang, Dong, et al. "Deep speaker verification: Do we need end to end?" 2017 *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017.
[2] Reynolds, Douglas A. "An overview of automatic speaker recognition technology." *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 4. IEEE, 2002.
[3] Dehak, Najim, et al. "Front-end factor analysis for speaker verification." *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2010): 788-798.
[4] Garcia-Romero, Daniel, and Carol Y. Espy-Wilson. "Analysis of i-vector length normalization in speaker recognition systems." *Twelfth annual conference of the international speech communication association*. 2011.
[5] Hong, Qian-Bei, et al. "Sequential Speaker Embedding and Transfer Learning for Text-Independent Speaker Identification." *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019.
[6] Snyder, David, et al. "Deep Neural Network Embeddings for Text-Independent Speaker Verification." *Interspeech*. 2017.
[7] Bhattacharya, Gautam, Md Jahangir Alam, and Patrick Kenny. "Deep Speaker Embeddings for Short-Duration Speaker Verification." *Interspeech*. 2017.
[8] Snyder, David, et al. "X-vectors: Robust dnn embeddings for speaker recognition." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
[9] Yu, Hong, et al. "Multi-task adversarial network bottleneck features for noise-robust speaker verification." *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*. IEEE, 2018.
[10] Zhao, Xiaojia, Yuxuan Wang, and DeLiang Wang. "Robust speaker identification in noisy and reverberant conditions." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.4 (2014): 836-845.
[11] Kolbœk, Morten, Zheng-Hua Tan, and Jesper Jensen. "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification." *2016 IEEE spoken language technology workshop (SLT)*. IEEE, 2016.
[12] Qin, Xiaoyi, Danwei Cai, and Ming Li. "Far-Field End-to-End Text-Dependent Speaker Verification Based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation." *Interspeech*. 2019.
[13] Geng, Wang, et al. "End-to-end language identification using attention-based recurrent neural networks." 2016.
[14] Okabe, Koji, Takafumi Koshinaka, and Koichi Shinoda. "Attentive statistics pooling for deep speaker embedding." *arXiv preprint arXiv:*1803.10963, 2018.
[15] Dai, Jifeng, et al. "Deformable convolutional networks." *Proceedings of the IEEE international conference on computer vision*. 2017.
[16] Zhu, Jian, Leyuan Fang, and Pedram Ghamisi. "Deformable convolutional neural networks for hyperspectral image classification." *IEEE Geoscience and Remote Sensing Letters* 15.8 (2018): 1254-1258.
[17] Xie, Weidi, et al. "Utterance-level aggregation for speaker recognition in the wild." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). IEEE, 2019.
[18] Qin, Xiaoyi, Hui Bu, and Ming Li. "Hi-mia: A far-field text-dependent speaker verification database and the baselines." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). IEEE, 2020.