Speaker recognition method for short utterance

Miao Guo1, Jiaxiong Yang1, Shu Gao1*
1 Department, College of Computer Science & Technology, Wuhan University of Technology, Wuhan, Hubei, China
*Corresponding author’s e-mail: 237078@whut.edu.cn

Abstract. Speaker recognition is a technology that uses identity information in the human voice for identity recognition, which owns many advantages in convenient information gathering, low gathering cost and high recognition accuracy. However, the difficulty in gathering messages within short utterance declines the voiceprint recognition function rapidly. We propose a recognition model based on SincNet in the aim of obtaining enough feature information in short utterance. The model used a set of learnable Sinc-based filter banks to extract feature directly from primordial voice in featured extraction layer, which enabled neural networks to discover more valuable voiceprint information; In the pooling layer, we designed the pooling method of dual attention mechanism, which combined multiple self-attention mechanism and self-attention mechanism to enrich the feature information and enhance the differentiation degree of key features so as to solve the defect of short speech with less information; choose ArcFace as the loss function, which can maximize the classification limit in the Angle space, thus improving the classification ability of the model. Experimental results demonstrate that the proposed model performs better than the benchmark model.

1. Introduction
Speaker recognition is an important branch of biometrics, which uses the hidden features in the speaker's audio signal to distinguish the speaker's identity. With the increase of voice interaction scenarios, such as smart phone assistants, smart homes, voiceprint locks, a common application scenario is to use short utterance for speaker recognition[1, 2]. The short utterance contains less information and is more difficult to extract. Therefore, the speaker recognition for short utterance poses new challenges.

Compared with knowledge-based acoustic features, directly using less processed features as model input can achieve better recognition accuracy[3]. The reason is that data-driven methods can train better models. Jee-weon Jung et al. proposed an end-to-end neural network model for speaker recognition. This model combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) to obtain feature information directly from the original voice waveform[4]. Ravanelli et al. proposed a new CNN architecture called SincNet, which uses the Sinc-filter bank to directly perform time domain convolution on the original waveform to obtain features for speaker recognition[5].

The loss function also has an important impact on the classification performance of the speaker recognition model. Inspired by Residual Networks (ResNets), Baidu Deep Speaker established a deep residual CNN model[6], and used the Triplet loss[7] to train the network, thereby improving speaker recognition Accuracy. Oxford University imitated Visual Geometry Group (VGG), proposed the VGGVox network structure, and also used the Triplet loss function to train the network, and the recognition effect was significant[8]. Compared with Triplet loss, Generalized End-to-End (GE2E) loss is compared with multiple data during each iteration, which can effectively shorten the training time and improve the recognition accuracy[9]. Zili Huang et al. introduced A-Softmax as the loss function in the
speaker recognition system, and used the end-to-end speaker recognition framework based on CNN and the speaker recognition framework using i-vector for verification. Obvious performance improvements have been achieved under these frameworks\cite{10}.

In order to improve the accuracy of short-speech speaker recognition, this paper proposes a short-speech speaker recognition model based on SincNet. The model uses SincNet as the basic structure and directly uses raw speech as input. The pooling layer of this model is optimized by a dual attention mechanism. First, a multi-headed self-attention mechanism is used to map voice features to multiple feature representation subspaces to achieve the purpose of enriching features. Then, in order to improve the multi-head attention mechanism, it is necessary to assume a unified head correlation defect, and the self-attention mechanism is again used to converge all the feature vectors. In order to further improve the performance of the short speech speaker recognition model, ArcFace\cite{12}, which performs well in the field of face recognition, is selected as the loss function during the model training. The ArcFace loss function is an improved version of the Softmax loss function, which can directly maximize the classification boundary in the angular space and improve the classification accuracy.

2. Materials and Methods

This paper constructs a short-speech speaker recognition model based on SincNet. The model consists of an input layer, a feature extraction layer, a pooling layer, a fully connected layer, and an output layer. Figure 1 shows the overall architecture of the model.

1) Input: use a short speech digital signal with a sampling rate of 16KHz as input.

2) Feature extraction layer: directly input the digital signal representing the short speech into multiple learnable Sinc filter banks, and the filter bank performs time domain convolution operation on the original speech signal (see section 2.1 for details).

3) Pooling layer: adopts Dual Attention Mechanism (DAM) to pool the input feature vector. (see section 2.2 for details).

4) Fully-connected layer: The feature vector output by the pooling layer is used as input. After multiple full connection operations and normalization operations, the posterior probability of the predicted speaker label is finally obtained.

5) Output: Take the posterior probability of the audio attributed to the speaker as input, and obtain the speaker label corresponding to the audio through the ArcFace loss function (see section 2.3 for details).
2.1. Feature extraction method based on Sinc-filter bank

In the feature extraction layer, a learnable filter bank is used for time domain convolution to obtain features from short utterance. The convolution operation is defined as follows:

\[ y[n] = x[n] * h[n] = \sum_{l=0}^{L-1} x[l] \cdot h[n-l] \]  

(1)

with \( x[n] \) describe a short utterance, \( h[n] \) is a filter of length \( L \), and \( y[n] \) is the filtered output. The \( L \) elements of each filter are learned from the data. In this model, a predefined function \( g[\cdot] \) is used to perform the convolution operation, and it contains only a few learnable parameters \( \theta \), which are defined as follows:

\[ y[n] = x[n] \ast g[n, \theta] \]  

(2)

with \( x[n] \) describe a short utterance, \( h[n] \) is a filter of length \( L \), and \( y[n]\) is the output finally filtered. In this case, all the elements of \( h \) are learnable parameters. SincNet proposes to replace \( h \) with a predefined function \( g[\cdot] \) that only depends on much fewer parameters to describe its behavior. In[5], the authors implemented \( g[\cdot] \) as a filterbank composed of rectangular bandpass filters. Such function can be written in the time domain as:

\[ g[n, f_1, f_2] = 2f_1 \sin(2\pi f_1 n) - 2f_2 \sin(2\pi f_2 n) \]  

(3)

with \( f_1 \) and \( f_2 \) the two learnable parameters that describe the low and high cutoff frequencies of the bandpass filters, and \( \sin(x) = \sin(x)/x \). Such parameters are randomly initialized in the interval \([0, f_s/2]\), with \( f_s \) equal to the input signal frequency sampling.

2.2. Pooling method of dual attention mechanism for short utterance

Aiming at the characteristic of short speech that contains less characteristic information, the attention mechanism is used for pooling operation, so that the model can focus on more effective information. The attention mechanism pooling strategy makes the model pay more attention to the various features that contribute to the performance, but the speech features extracted by only one attention mechanism pooling still have limitations, so the self-attention mechanism is again used to optimize the model. The dual attention mechanism pooling method used in this article is shown in Figure 2.

The first multi-head attention mechanism will perform multiple different mappings on the speech features extracted by the Sinc filter bank, and perform self-attention pooling in multiple feature representation subspaces. The hidden state sequence obtained after the original speech is convolved in the time domain can be expressed as \( h = [h_1, h_2, \ldots, h_N] \). Suppose there are \( m \) self-attention heads, and the hidden feature sequence \( h_j = [h_{j_1}, h_{j_2}, \ldots, h_{j_m}] \) calculated by the first self-attention head. The calculation method of each attention head can be expressed as:
with \( w_j \) represents the attention weight of the j-th attention head on the hidden state sequence \( h_j, u \)
represents a set of trainable parameters. Then, calculate each attention head in the same way as self-
attention, and its expression is:

\[
\exp \left( \frac{T_{jj} h_j \cdot w_j}{\sqrt{d_h}} \right)
\]

(4)

\( c_j = h_j^T w_j \) (5)

c_j represents the feature representation of the j-th attention head. After processing by the multi-head
attention mechanism, the feature vectors corresponding to all heads are connected as the final output
feature \( c = [c_1, c_2, \ldots, c_n] \). Using only the multi-head self-attention pooling method requires the
assumption of uniform head correlation. To improve this shortcoming, the self-attention mechanism is
used again. The second self-attention mechanism is used to calculate the weights of different attention
heads, and then merge the results of multiple attention heads. The calculation formula is:

\[
\exp(c_j^T u) \cdot \sum_{i=1}^{k-1} \exp(c_i^T u)
\]

(6)

\[
c = \sum_{j=1}^{k} c_j^T w_j
\]

(7)

c is the weighted average of the output vectors of each attention head, which is the final output result.

2.3. Design of loss function based on ArcFace

Short utterance contains fewer features, so it is more difficult to accurately recognize. Efficient loss
function has better distinguishing ability for features, can make the same type of samples have a smaller
intra-class distance, and can enlarge the distance between samples of different categories. The ideal loss
function has a better decision boundary, which can make a larger angular interval between classes, so as
to obtain better classification results. We use ArcFace as the loss function, which is defined as follows:

\[
L_{\text{arcface}} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s \cos(\theta_j + m))}{\sum_{j=1}^{N} \exp(s \cos(\theta_j))}
\]

with \( N \) denotes the number of training short utterance samples in a single batch, \( m \) is the angle margin,
\( m \) denotes angle margin, \( x_i \) denotes the feature vector corresponding to the i-th short utterance sample,
\( y_i \) denotes the true speaker label of the i-th short utterance sample, and \( \theta \) denotes the scaling
parameter when the feature is normalized.

3. Results & Discussion

In order to verify that our method is effective, we conducted four sets of experiments using the TIMIT\(^5\)
dataset. The frame error rate (FER) and equal error rate (EER) results of the four sets of experiments
are shown in Table 1. At the same time, we also paid attention to the FER change process during the
model training process, as shown in Figure 3.

| Number | Model     | Attention mechanism | Loss function | FER(%) | EER(%) |
|--------|-----------|---------------------|---------------|--------|--------|
| 1      | SincNet   |                     | Softmax       | 41.99  | 8.2    |
| 2      | SincNet   | DAM                 | Softmax       | 32.58  | 6.7    |
| 3      | SincNet   |                     | ArcFace       | 23.87  | 5.2    |
| 4      | SincNet   | DAM                 | ArcFace       | 21.27  | 4.9    |
Experiment 1 and Experiment 2 under the condition that only the attention mechanism is different, the performance of the model using the dual attention mechanism is better. The frame error rate FER of the SincNet model after introducing the dual attention mechanism is reduced by 9.41%. Comparing Experiment 3 and Experiment 4, in the attention mechanism comparison experiment that both use the ArcFace loss function, experiment 4 has improved performance, but the performance improvement is small, and the frame error rate FER is only reduced by 2.6%. Under the condition of using the Softmax and ArcFace, the model using the dual attention mechanism reduces the equal error rate EER by 1.7% and 0.3% respectively.

In order to further verify the performance of our model, compare the speaker recognition results of this model with GMM-UBM model[13], i-vector/PLDA model[14], VGGVox[8] model on the Voxceleb1[13] dataset. The final experimental results are shown in Table 2.

| Model         | EER(%) |
|---------------|--------|
| GMM-UBM       | 15.0   |
| i-vector/PLDA | 8.8    |
| VGGVox        | 7.8    |
| SincNet       | 6.5    |
| Our Model     | 4.3    |

The equal error rate EER of the model proposed in this paper reaches 4.3%, which is the best performance compared with the above model, and has a great improvement. The reasons for the excellent performance of the model in this article may be:

- Using the Sinc-filter bank to capture features directly from the original speech can not only ensure that more effective voiceprint information can be extracted from the short speech, but also can find more valuable filters to effectively reduce the feature dimension;
- The multi-head self-attention mechanism can enrich the extracted feature information when the amount of short speech information is small, and can effectively enhance the discrimination of key features. By using the self-attention mechanism on the features of different heads, the correlation between different heads can be found, and the overall feature vector obtained is more reasonable;
- Compared with the standard cross-entropy loss, ArcFace increases the distance between different categories by setting the angle margin, so that the features are more concentrated in the weight center of the category. Using ArcFace can make the short speech speaker recognition model have better classification performance.

4. Conclusions
Based on the SincNet architecture, this paper constructs a short-speech speaker recognition model that
combines dual attention mechanisms, and uses ArcFace to optimize the model during the training process. Different from other speaker recognition models, the model proposed in this paper aims at the short speech information which is less and difficult to mine. It uses a Sinc-filter bank instead of the standard convolutional layer in the CNN structure to fully mine the hidden speaker features in short speech. The dual attention mechanism is adopted in the pooling layer, which not only uses the multi-head self-attention mechanism to focus on the features that contribute to the model performance, but also weakens the interference of invalid features. At the same time, the self-attention mechanism is used to fuse the features obtained by multiple attention heads. In addition, in the training process of the model, a more excellent loss function is selected, and ArcFace is used to replace the softmax loss function to improve the classification ability of the model. This paper evaluates the model on two voiceprint data sets. Experiments show that the model constructed in this paper has excellent performance on voiceprint tasks.

In future work, we will further consider the adverse effects of complex noise in real application scenarios on short speech recognition, and optimize the proposed model to make it have practical application value.

References
[1] Zheng S, Liu G, Suo H, et al. (2019) Autoencoder-Based Semi-Supervised Curriculum Learning for Out-of-Domain Speaker Verification. In: Interspeech 2019. Graz. pp. 4360-4364.
[2] Hajavi, A. and A. Etemad. (2019) A Deep Neural Network for Short-Segment Speaker Recognition. https://arxiv.org/abs/1907.10420.
[3] Jung J, Kim S, Shim H, et al. (2020) Improved RawNet with Feature Map Scaling for Text-independent Speaker Verification using Raw Waveforms. In: Interspeech 2020. Shanghai. pp. 1496-1500.
[4] Jung J, Heo H S, Kim J, et al. (2019) Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. https://arxiv.org/abs/1904.08104.
[5] Ravanelli M, Bengio Y. (2018) Speaker recognition from raw waveform with sincnet. In: 2018 IEEE Spoken Language Technology Workshop (SLT). Athens. pp. 1021-1028.
[6] Li C, Ma X, Jiang B, et al. (2017) Deep speaker: an end-to-end neural speaker embedding system. https://arxiv.org/abs/1705.02304.
[7] Schroff F, Kalenichenko D, Philbin J. (2015) Facenet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston. pp:815-823.
[8] Chung J S, Nagrani A, Zisserman A.(2018) Voxceleb2: Deep speaker recognition. https://arxiv.org/abs/1806.05622.
[9] Wan L, Wang Q, Papir A, et al.(2018) Generalized end-to-end loss for speaker verification. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary , pp: 4879-4883.
[10] Huang Z, Wang S, Yu K. (2018) Angular Softmax for Short-Duration Text-independent Speaker Verification. In: Proceedings of Interspeech 2018. Hyderabad. pp: 1076-1080.
[11] Nunes J A C, Macêdo D, Zanchettin C. (2019) Additive margin sincnet for speaker recognition. In: 2019 International Joint Conference on Neural Networks (IJCNN). Budapest. pp: 1-5.
[12] Deng J, Guo J, Xue N, et al. (2019) Arcface: Additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach. pp: 4690-4699.
[13] Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. Computer Speech & Language, 60, 101027.
[14] Haniçli C, Çeliktas H. (2018) Turkish text-dependent speaker verification using i-vector/PLDA approach. In: 2018 26th Signal Processing and Communications Applications Conference (SIU). Izmir. pp: 1-4.