IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 27 July 2023, accepted 14 August 2023, date of publication 17 August 2023, date of current version 25 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3306026

RESEARCH ARTICLE

Application of Split Residual Multilevel Attention Network in Speaker Recognition

JIJI WANG[1], FEI DENG[1], LIHONG DENG[1], PING GAO[2], AND YUANXIANG HUANG[2]
[1]College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu 610059, China
[2]Sichuan Tianyi Ecological Garden Group Company Ltd., Chengdu, Sichuan 610093, China

Corresponding authors: Yuanxiang Huang (huangyuanxiang.cn@gmail.com) and Fei Deng (dengfei@cdut.edu.cn)

ABSTRACT Current speaker recognition systems are mainly for the combined application of network architectures and attention mechanisms, however, lightweight networks are not able to extract frame-level features of speaker speech well, and deeper and wider networks also face the problems of slower inference and excessive number of parameters. To this end, we proposes Split-ResNet, a network structure for split residuals, which can obtain a combination of multiple receptive field at a finer-grained level, thus obtaining a variety of feature representations with different scale combinations and producing more informative and comprehensive multi-scale features. In addition we propose a dual time-frequency attention (DTFA) that enhances key features and suppresses unimportant features by focusing on features in the time and frequency domains and learning weights from the time and frequency channels, respectively. We finally tested the speaker recognition system using a combination of Split-ResNet and DTFA against other speaker recognition systems on the Voxceleb1-O test set. The test results show that the speaker recognition system proposed in this paper is 0.98%, 0.39%, 0.69% and 0.47% lower in EER compared with SpeechNAS, RawNet2, Y-vector and CNN+Transformer, respectively, proving that DTFA+Split-ResNet is a speaker recognition system with good speaker audio feature extraction capability and discriminative capability.

INDEX TERMS Attention mechanisms, DTFA, speaker identification, split-ResNet.

I. INTRODUCTION

Speaker identification is the process of determining the identity of a speaker by recognizing the information in the speaker's audio [1]. And since most people have subtle differences in their vocal organs, it is feasible to identify a person by their voice. However, because each person's tone, intonation, speech staccato, and speech content are different, each audio will contain a lot of rich information, and in order to better obtain the speaker's identity information, we need to design a network that extracts audio information from multiple scales and aspects.

Before the rise of deep learning, i-vector systems with probabilistic linear judgment analysis (PLDA) in traditional methods had always been at the forefront in the field of speaker recognition [2], [3], [4]. Subsequently, improved d-vectors [5] and x-vectors [6] based on i-vectors emerged. But with the development of deep learning, deep learning networks (DNNs) have brought breakthroughs in speaker recognition. The DNN architecture system is capable of directly processing the input audio data, then extracting the frame-level features of the input audio through DNN and aggregating the features through an aggregation model to aggregate the extracted frame-level features into discourse-level features, thus making deep learning dominant in the direction of speaker recognition [7], [8].

Extracting effective features in deep learning is the key for us to identify the speaker accurately. Convolutional Neural Network (CNN) is the widely used backbone network for feature extraction, and by adding residual connectivity structure to CNN, Residual Network (ResNet) achieves better frame-level feature extraction inside the same input samples [9]. In general, deeper and more complex network structures can extract more effective features, and many methods are now starting to use deeper, broader and more complex network structures. Zeinali et al. [10] used a 256-layer ResNet

The associate editor coordinating the review of this manuscript and approving it for publication was Yassine Maleh.

network to build a speaker recognition system. In 2020, Jung et al. [11] proposed RawNet using raw audio as input. In 2021, researchers Zhao et al. [12] utilized a 152-layer ResNet network and a multi-headed self-attentive mechanism (MHSA) to develop a speaker recognition system. Similarly, Wang et al. [13] implemented a CNN combined with Transformer for the task of speaker recognition in the same year. Additionally, Zhu et al. [14] proposed a framework known as SpeechNAS, which employed Bayesian optimization-based neural network search to identify the optimal candidate network from trained hypernetworks. Gao et al. [15] proposed a new multi-scale backbone architecture called Res2Net based on ResNet, which replaces a set of 3 × 3 convolution kernels with smaller 1 × 1 filters and connects different filter sets in a hierarchical class residual manner. Also Wang et al. [16] used null convolution in the convolution process to increase the receptive field of the extracted features. However, in speaker recognition, larger features do not necessarily produce better recognition results, and a larger and deeper network structure generates more parametric quantities, making the computational effort increase significantly. Therefore, in order to improve the recognition effect of features without significantly increasing the network parameters and computational complexity, we need to build a lightweight convolutional neural network with strong feature extraction capability.

However, convolutional neural networks also have certain bottlenecks. Convolutional neural networks use fixed-size convolutional kernels to extract features, which also limits the extraction ability of convolutional neural networks to extract more global feature information, making the extracted features not better for speaker recognition. Recently attention mechanism has also shown better performance in computer vision methods, Okabe et al. [17] found that the attention mechanism can be used in the field of speaker recognition to calculate the weighted average of frame-level vectors in a neural network, which enables the speaker's audio features to be focused on the main frames and to obtain feature values with higher discriminative power. Zhou et al. [2] used ResNet in speaker recognition and added Squeeze-and-Excitation (SE) attention mechanism, and concluded that the attention mechanism could indeed improve the speaker audio feature representation at the frame level and improve the feature recognition ability. Desplanques et al. [18] improved on top of TDNN by adding a new jump connection and a global contextual attention mechanism. Yadav and Rai [19] proposed tf-CBAM attention inspired by the Convolutional block attention module (CBAM) attention module in computer vision. Although these attention mechanisms can enhance the feature extraction capability of neural networks globally, only some simple attention mechanisms are used for a particular dimension, ignoring the features of audio in the two-dimensional time, and frequency domain dimensions, and also the pooling operation loses the identity information of the speaker in the audio.

We propose a new Split-ResNet and dual time-frequency attention (DTFA) mechanism to solve the above problems. ResNet solves the problems of gradient explosion and network degradation by introducing jump connections, but to extract more representative features, a deeper and wider network structure is bound to be used. Therefore we propose the improved Split-ResNet, a network that constructs a new multi-scale parallel branching feature fusion network that extracts multi-scale features at a finer-grained level and produces more feature information. Meanwhile, in order to solve the problems arising from existing network attention mechanisms, our proposed DTFA attention mechanism is able to focus on speaker identity information and audio time-frequency information based on the rich and redundant feature information extracted by Split-ResNet. DTFA divides the information in the features into information in the time and frequency domains, enabling attention to be focused on the relationship between these two. It also focuses on the information in the features between channels. We demonstrate the performance of Split-ResNet and DTFA for speaker recognition on the Voxceleb2 dataset through multiple ablation experiments, respectively, which significantly outperforms the performance of existing state-of-the-art speaker recognition systems.

The remainder of this paper is organized as follows. Section II describes the related work on attention and aggregation methods. Section III presents the proposed model. The experimental setup is introduced in Section IV. The results and analysis are shown in Section V. Finally, Section VI concludes the paper.

II. RELATED WORK

A. REVIEW OF SENet

In [20], Hu et al. proposed Squeeze-and-Excitation Networks (SENet). This network is divided into two key operations, Squeeze and Excitation. First is the Squeeze operation, which performs feature compression by following the spatial dimension, turning each two-dimensional feature channel into a real number that somehow has a global receptive field and matching the output dimension to the number of input feature channels. It characterizes the global distribution of the response over the feature channels and makes the global receptive field of available also for layers close to the input, which is very useful in many tasks. The Squeeze operation is shown in Equation (1), where H represents the height of the input data, W represents the width of the input data, and $u_c(i, j)$ represents the average value of the data for this channel.

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j). \tag{1}$$

Next is the Excitation operation, a mechanism similar to the gates in recurrent neural networks. The weights are generated for each feature channel by means of a parameter

w, where the parameter w is learned to explicitly model the correlation between the feature channels, and finally, the importance of each feature channel is automatically obtained by this learning method. The Excitation operation is shown in Equation (2), where z is the result obtained in (1), W_1 represents a weight operator of dimension $C/r \times C$, which is obtained by full concatenation with z to obtain a result with channel number C/r, represents the ReLU activation function, W_2 is a weight operator of dimension $C \times C/r$, which is also recovered to the data with channel number C after computation by full concatenation, σ represents the Sigmoid activation function, and finally s is obtained.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z)). \qquad (2)$$

B. REVIEW OF CBAM

Woo et al. [21] proposed an attention module named Convolutional Block Attention Module (CBAM), which is divided into a channel attention module and a spatial attention module. In the channel attention module, two methods, global maximum pooling and global average pooling, are applied in order to focus attention on the channels that have a greater impact on the results, and the fully connected network is subjected to nonlinear feature changes, and the changed network summation is reactivated to obtain the attention weights.In the spatial attention module, the authors use the spatial relationships between features to generate spatial attention and use maximum pooling and average pooling over channels, respectively, in order to focus the model more on features with more important spatial shapes. Spatial attention is a complement and refinement of channel attention.

Its details are shown in Equation (3), the original CNN feature F, with a shape of $W \times H \times C$, is obtained after a channel attention mechanism M_c to an attention weight $M_c(F)$, with a Shape of $1 \times 1 \times C$. $M_c(F)$ and the input feature F are multiplied to obtain the feature F'; F' is then passed through a spatial attention mechanism M_s to an attention weight $M_s(F')$, with a Shape of $W \times H \times 1$, and $M_s(F')$ and F' are multiplied to obtain the final feature F”, and the Shape of F” is $W \times H \times C$. It can be seen that after CBAM, the Shape of the feature does not change, so CBAM can be inserted after the original CNN feature and the network does not need to be changed.

$$\begin{aligned} F' &= M_c(F) \otimes F, \\ F'' &= M_s(F') \otimes F', \end{aligned} \qquad (3)$$

III. SPEAKER RECOGNITION SYSTEM BASED ON SPLIT-ResNet AND DTFA

Usually, speaker recognition tasks can be divided into open set and closed set. In the closed set, the test set is a subset of the speakers in the training set, and it is relatively easy to identify the speakers. In contrast, in the open set, the speakers in the training set and the speakers in the test set are unrelated [22]. Thus speaker recognition on the open set is more difficult and, at the same time, more realistic. At an earlier time, there was text-related speaker recognition, which required the speaker to say the same thing, a limitation that was difficult to apply in practice. And then, text-independent speaker recognition gradually developed, and in the traditional i-vector speaker recognition system, each step is trained separately on top of the subtask, not jointly optimized [23]. For this reason, an end-to-end speaker recognition system is needed to integrate all subtasks for training [24], and we propose and implement an end-to-end system with unified closed and open sets by integrating all subtasks together.

Our proposed Split Residual Multilevel Attention Network (SPMAN) is shown in Fig. 1. It consists of two parts (1) Split-ResNet, and (2) Dual Time Frequency Attention (DTFA). We use Split-ResNet as the backbone network of the feature extractor, which can obtain multiple combinations of receptive fields at a finer-grained level, thus obtaining multiple feature representations with different scale combinations and producing more informative and comprehensive multi-scale features. In the front-end model, DTFA enhances the feature representation of the network, which can be inserted anywhere in the neural network, but attention-based convolution shows that parallelizing the convolutional layers and the attention module is a more efficient structure for dealing with short and long-term dependencies. Therefore, we embed DTFA into Split-ResNet to produce more effective frame-level features, as shown in the red box in Fig. 1.

A. SPLIT-ResNet

An excellent deep neural network can produce rich or even redundant feature information to ensure a comprehensive understanding of the input features, and redundancy in feature mapping is also an important marker for the success of deep neural networks. In computer vision tasks, multi-scale features have proven to be very effective. By multi-scale features, we mean that features are sampled to different degrees, and different features are observed at different scales. Therefore, multi-scale features also contain more feature information. We introduce its idea to the speaker recognition task, acting on audio features. However, most of the current approaches are multi-scale representations layer by layer, which also leads to increased depth and makes it increasingly complex to design efficient architectures. We have rethought the existing network structure, considering in particular the following three fundamental questions. (i) How to generate more or even redundant information from the input features. (ii) How to facilitate the network to learn stronger feature representations without increasing the computational complexity. (iii) How to achieve better performance and maintain inference time.

To this end, we construct a new multi-scale parallel branching feature fusion network to extract multi-scale features at

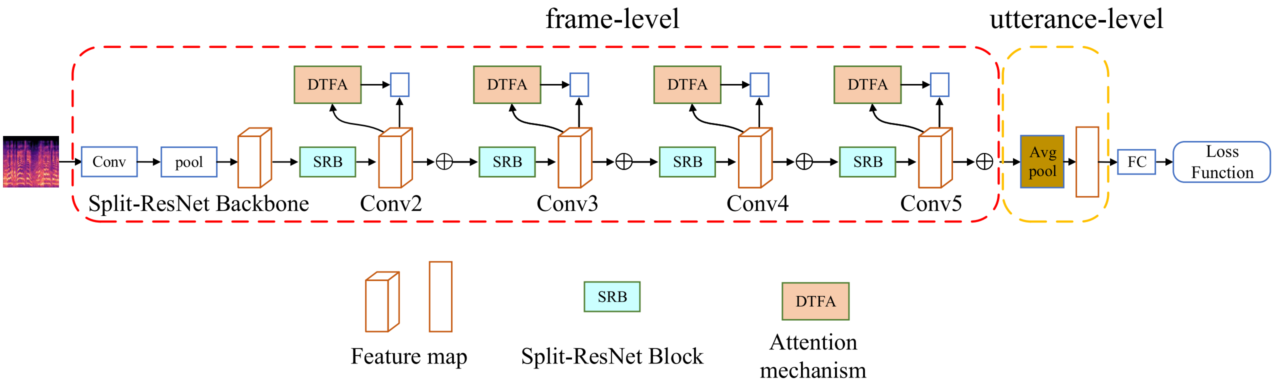


FIGURE 1. Figure/An end-to-end speaker recognition system based on Split-ResNet and DTFA.

a finer-grained level and generate more feature information. The Bottleneck structure in ResNet is shown on the left side of Fig. 2. Instead of using a set of 3 × 3 convolutional operators to extract features as in Bottleneck. We split it into several parallel branches of equal size. This allows the convolution operators on each parallel branch to efficiently learn different time-frequency domain components and also to more efficiently represent multi-scale features in the time-frequency domain while maintaining a similar computational load. Since the proposed neural network module involves residual-like connections within residual blocks, we name it Split-ResNet, and the structure is shown in the red box on the right side of Fig. 2.

After 1×1 convolution, we divide the feature mapping uniformly into s sets of feature images with channel number $w(c = s \times w)$, denoted by x_i, where $i \in \{1, 2, \ldots, \}$ and the dimensionality of x is $x \in R^{T\times F\times C}$. Compared with the input features, each feature subset x_i has the same time-frequency size but the number of channels is $1/s$ of the original number of channels. Except for x_1, each x_i has a corresponding 3 × 3 convolution, denoted as $F_i()$. We denote the output of $F_i()$ by y_i and concatenate the feature subset x_i with the output of $F_{i-1}()$ and feed it to $F_i()$ as shown in Equation (4). We do not use the summation approach because the summation approach (which is used in Res2Net) would change or even destroy the feature representation, whereas the concatenation approach keeps the features intact, and the "⊕" in the formula represents the concatenation operation along the channel dimension. At the same time, the concatenation brings an increase in depth.

$$y_i = \begin{cases} x_i, & i = 1 \\ F_i(y_{i+1} \oplus x_i), & 2 < i \leq s. \end{cases} \tag{4}$$

In this way, each feature subset can receive the output of the previous feature subset, thus increasing the receptive field of the current feature subset, and more information about the identity of the speaker can be obtained, and more and larger combinations of receptive fields are also obtained, as shown in Equation (5).

$$\begin{aligned} y_1 &= x_1 \\ y_2 &= F_2(x_2) \\ y_3 &= F_3(x_3 \oplus y_2) = F_3(x_3 \oplus F_2(x_2)) \\ &\vdots \\ y_n &= F_n\,(x_n \oplus y_{n-1}) = F_n(x_n \oplus F_{n-1}(x_{n-1})) \end{aligned} \tag{5}$$

For example, Equation (6) represents the formula for the receptive field, where l_k represents the receptive field of the k-th layer, $size_k$ represents the convolution kernel size of the k-th layer, and s_i represents the convolution step size of the i-th layer. Then we can conclude that the receptive field is 1 after convolution with a 1 × 1 convolution kernel with a step length of 1. It is also possible to derive the per-layer receptive field of the Bottleneck Block, as shown in Equation (7), and the receptive field of parallel branches in the Split-ResNet Block, as shown in Equation (8). From the comparison results of Equation (7) and Equation (8), it can be concluded that Split-ResNet has more and larger receptive field, which makes Split-ResNet have stronger feature extraction ability than ResNet and can produce better results than the original network.

$$l_k = l_{k-1} + (size_k - 1) * \sum_{i=1}^{k-1} s_i \tag{6}$$

$$l_k(Bottleneck) = \begin{cases} 1, & k = 1 \\ 3, & k = 2, 3 \end{cases} \tag{7}$$

$$l_k(Split - ResNet) = \begin{cases} 1, & k = 1 \\ 5, & 1 < k < s \\ 3, & k = s \end{cases} \tag{8}$$

B. DUAL TIME-FREQUENCY ATTENTION MECHANISM (DTFA)

In the process of extracting frame-level features from Split-ResNet, rich redundant feature information is generated, but not all of the feature information is valid and contains speaker

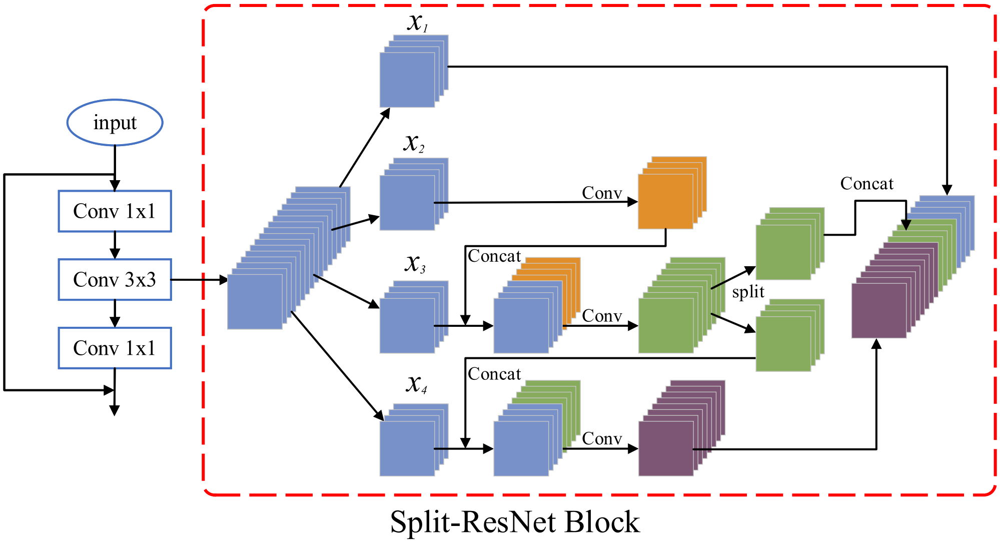


FIGURE 2. Split-ResNet block structure diagram.

identity information. Therefore, in order to focus on the more important regions of these feature information and to compensate for the shortcomings of the convolution operation itself, we construct the dual time-frequency attention (DTFA). Our proposed attention method aims to focus on the important regions in the time channel and frequency channel and obtain the dependencies between features from them to fetch more attention information so that the feature extractor can extract more discriminative frame-level features. The DTFA module is shown in Fig. 3, where we use two-dimensional adaptive averaging pooling to act on the temporal and frequency dimensions of the input audio feature X to obtain X_T and X_F, respectively, with the formulas shown in Equation (9) and Equation (10). After concatenating X_T and X_F, the interdependencies between different channel dimensions are learned using a convolution operator of size $1 \times 1 \times C/r$ with the help of global time and frequency information, as shown in Equation (11), where $r = C/C'$ denotes the dimensionality reduction factor, which is used to reduce the number of parameters.

$$X_T = \frac{1}{T}\sum_{i=1}^{T} X_i \tag{9}$$

$$X_F = \frac{1}{F}\sum_{i=1}^{F} X_i \tag{10}$$

$$X_g = relu(W_1(X_T, X_F)) \tag{11}$$

After encoding the global time and frequency domains, channel attention learning is performed. Split X_g into $X_T{'}$ and X_F', respectively, and then form the time domain channel attention and frequency domain channel attention. The formulas are shown in Eqs. (12) and (13), where W_2 and W_3 are convolution operators with kernel size $(1 \times 1 \times C \times r)$ and δ is the sigmoid activation function for generating the time domain channel attention weights w_T and the frequency domain channel attention weights w_F. The attention module proposed above is called dual time-frequency attention (DTFA) because it forms two different channels of attention with the help of temporal and frequency information. Finally, the resulting time domain channel attention weights and frequency domain channel attention weights w_T and w_F are multiplied by the original input feature X to obtain the final output feature Z. The calculation formula is shown in Equation (14). Also, like the previous attention mechanism, DTFA learns the interdependence between channel features.

$$w_T = \delta(W_2(X_T{'})) \tag{12}$$

$$w_F = \delta(W_3(X_F{'})) \tag{13}$$

$$Z = X \otimes w_T \otimes w_F \tag{14}$$

IV. EXPERIMENTAL SETUP

A. DATASET

The dataset used in our experiments is the Voxceleb dataset, which has been widely used in speaker recognition tasks in recent years [25], [26], [27]. The Voxceleb dataset consists of two parts, the Voxceleb1 dataset and the Voxceleb2 dataset, where Voxceleb1 contains more than 100,000 audio data from 1,251 speakers. Voxceleb2 contains more than 1 million audio data from 5,994 speakers intercepted from Youtube videos. The speakers in the dataset span a wide range of races, accents, and ages, while 61% of the speakers are male and 39% are female. Since the audio data of Voxceleb2 was collected in a real speaking environment, filled with a large number of speaking tones, video sound effects, and murmurs, it greatly enhances the challenge of speaker recognition and also improves the realism of the experiment. So we chose Voxceleb2 as the dataset, and there are three test set files. The first one is VoxCeleb1-O which chose 40 voices from

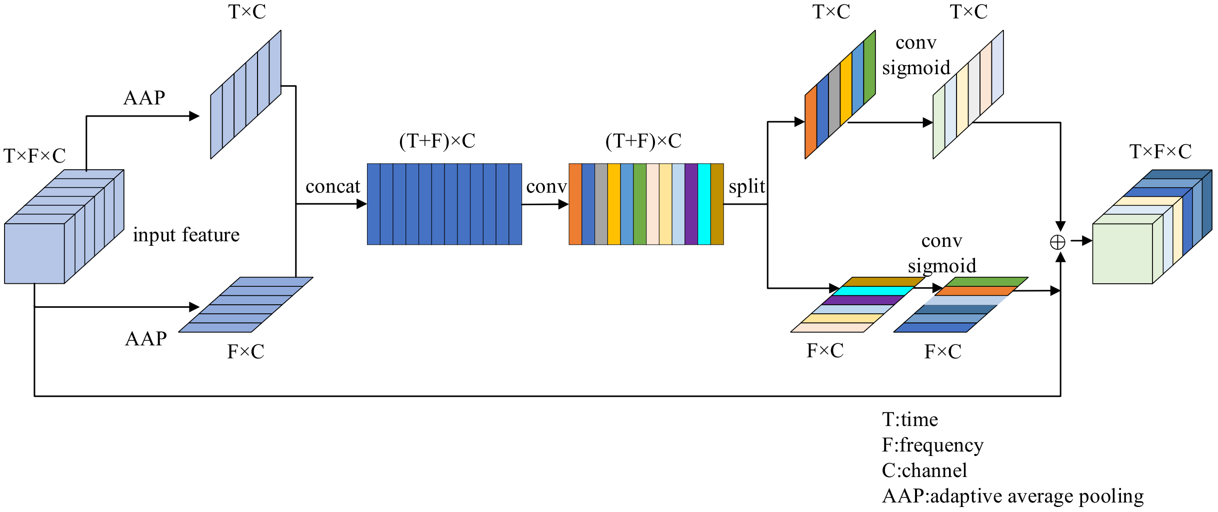


FIGURE 3. Dual time-frequency attention mechanism module.

TABLE 1. Test results of different network structures in Voxceleb1-O.

Voxceleb1-O						
	Model	loss	Dims	Training set	EER(%)	DCF
Chung et al., 2018 [27]	ResNet-50	AM-softmax	512	Voxceleb2	3.95	-
Ours	ThinResNet-50	AM-softmax	512	Voxceleb2	5.04	0.4651
Ours	Res2Net	AM-softmax	512	Voxceleb2	3.32	0.3199
Ours	Split-ResNet	AM-softmax	512	Voxceleb2	2.50	0.2426

Voxceleb1 as the test set; the second one is Voxceleb-E which used the whole Voxceleb1 as the test set; and the third one is Voxceleb-H which took speakers with the same nationality and gender from the whole VoxCeleb1 dataset as the test set.

B. TRAINING PARAMETERS SETTING

To verify the effectiveness of the method proposed in this paper, the same comparison method as in the literature [25], [26], [27] was used for the experiments. The audio data were first converted to a single channel with 16 kHz sampling rate, and then the audio data were filtered (40 groups of Mel filters) with a Hamming window of size 25 ms and step size 10 ms, and then the 40-dimensional Fbank [6] features were obtained by mean and variance normalization in the frequency domain direction as the input to the network. The network parameters were optimized using the Adam optimizer [28], with the initial learning rate set to 0.001 and the learning rate decay set to 0.98, decaying every 5 Epochs.

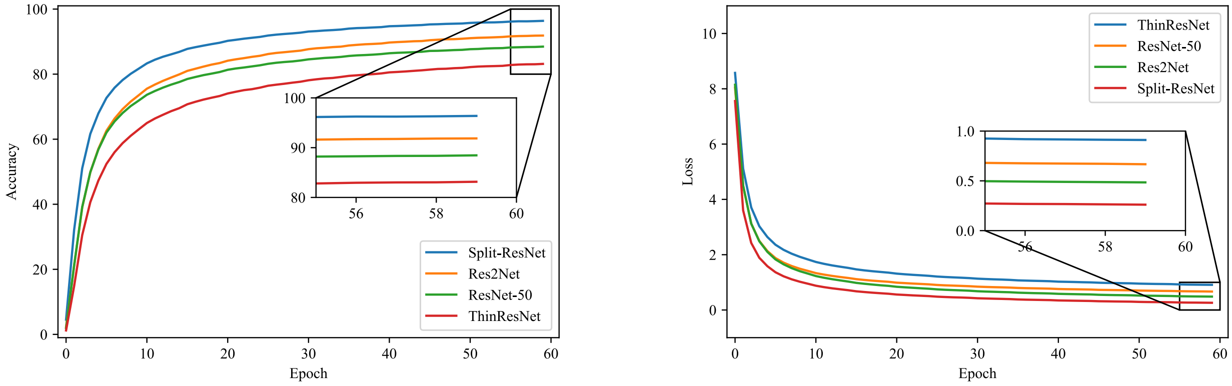
The loss function selected is the AM-softmax function [29], setting margin = 0.1 and scale = 30. Compared to the softmax loss function, AM-softmax introduces a boundary in the corner space to improve the verification accuracy. The formula is shown in Equation (15), where L_{AMS} is the cost of classifying the sample correctly and $\theta_y = arccos(w^T x)$ is the angle from the sample feature to the decision hyperplane w. Both vectors are normalized by L2. The hyperparameter s is the scaling factor, which is fixed to 30 here to improve the difference of *cos* function distribution, increase the inter-class spacing, reduce the intra-class spacing, and improve the convergence speed. m is the inter-class interval, which needs to be set smaller when there is not much difference between classes.

TABLE 2. Params and inference time for different network structures.

	Model	Params(M)	Time(ms)
Ours	ResNet-50	22.4	130
Ours	ThinResNet-50	1.54	60
Ours	Res2Net	2.44	88
Ours	Split-ResNet	2.53	94

$$L_{AMS} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{s\cdot(\cos\theta_{y_i} - m)}}{e^{s\cdot(\cos\theta_{y_i} - m)} + \sum_{j=1,j\neq y_i}^{c} e^{s\cdot\cos\theta_j}} \quad (15)$$

For testing, we use a test set that is completely separated from the training set, read 10 segments of 3s of audio input from each input audio to calculate the features, and then calculate the Euclidean distance from the 10 × 10 segment feature values in both audio segments, after which this distance is averaged to determine the identity of the speaker. To demonstrate the effectiveness of the system proposed in this paper, this paper does comparison experiments with SpeechNAS proposed by Zhu et al. [14], Res2Net proposed by Gao et al. [15], thinResNet proposed by Xie et al. [30], and Y-vector used by Zhu et al. [31], respectively.



(a) Accuracy curves of different networks

(b) Loss rate curves of different networks

FIGURE 4. Loss curves and accuracy curves of different network structures in the training process.

TABLE 3. Test results of different attention methods in Voxceleb1-O.

	Voxceleb1-O						
	Model	Attention	loss	Dims	Training set	EER(%)	DCF
Ours	Split-ResNet	SE	AM-softmax	512	Voxceleb2	2.44	0.2690
Ours	Split-ResNet	CBAM	AM-softmax	512	Voxceleb2	2.38	0.2680
Ours	Split-ResNet	DTFA	AM-softmax	512	Voxceleb2	2.09	0.2426

TABLE 4. Params and inference time for different attention methods.

	Model	Params(M)	Time(ms)
Ours	SE	2.25	77
Ours	CBAM	2.30	80
Ours	DTFA	2.33	85

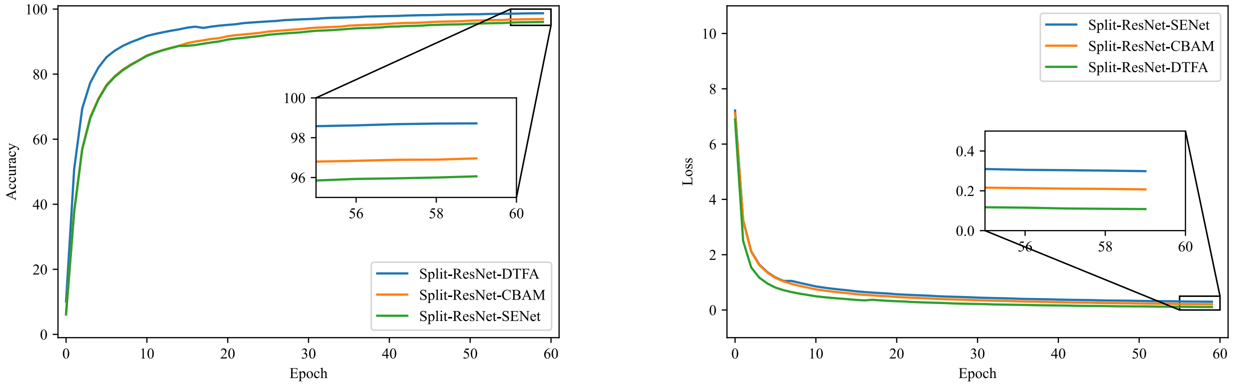
C. EVALUATION INDICATORS

The evaluation metrics selected in this paper are equal error rate (EER) and minimum detection cost function (MinDCF). The smaller the value of both metrics, the better. The formula for MinDCF is shown in Equation (16), where C_{FRR} and C_{FAR} are the weights of false rejections and false acceptances, i.e., penalty coefficients, respectively, and FRR and FAR are the false rejection rate and false acceptance rate. P_{target} is the prior probability of the appearance of the true speaker, and $(1 - P_{target})$ is the prior probability of the appearance of the wrong speaker. We used $C_{FRR} = C_{FAR} = 1$ and $P_{target} = 0.01$, which are the parameters set on NIST SRE2010. MinDCF not only considers the different costs of the two types of errors but also considers the prior probabilities of the two test cases, which is more reasonable than EER.

$$DCF = C_{FRR} \times FRR \times P_{target} + C_{FAR} \times FAR \times (1 - P_{target}) \tag{16}$$

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, two sets of ablation experiments are first conducted to demonstrate the effectiveness of Split-ResNet and DTFA in speaker recognition, respectively. The experimental results of Split-ResNet and other network structures modified in this paper in the Voxceleb1-O test set under the same experimental conditions are shown in Table 1, where no attention method was used in the ablation experiments of the network structures. We can see that the performance of the lighter ThinResNet-50 is relatively poor, with an EER/MinDCF of 5.04/0.4651, while the performance is significantly improved when it is replaced by the larger and deeper ResNet-50, but by comparing the inference time and parameters of different network structures in Table 2 we find that the inference time and parameters of ResNet-50 have also increased a lot. We also tested the effectiveness of Res2Net in our system and concluded that the results are better than ResNet-50, while the parameters and inference time are much reduced. The final result is the test result of our Split-ResNet in the system, and the EER/MinDCF is 2.50/0.2426, which can be seen that the result has been greatly improved compared with the other three network structures, indicating that our Split-ResNet has better feature extraction ability compared with the other three network structures, and can identify the speaker's identity information. Meanwhile, Split-ResNet has only 0.09M more parameters than Res2Net, and the inference time is only 6ms more than Res2Net, but the performance is improved by about 25%. As shown in Fig. 4, the a and b plots show the training accuracy curves and loss rate curves of different networks, respectively. It can be seen that the Split-ResNet network modified in this paper has the highest accuracy and the fastest loss decrease compared with other networks after the same training rounds, always ahead of other networks, which indicates that the Split-ResNet designed in this paper has stronger feature extraction ability.



(a) Accuracy curves for different attention spans

(b) Loss rate curves of different attention spans

FIGURE 5. Loss curves and accuracy curves of different attention mechanisms in the training process.

TABLE 5. Test results of different network structures in Voxceleb1-O.

	Voxceleb1-O						
	Model	Attention	loss	Dims	Training set	EER(%)	DCF
Jung et al.,2020 [11]	RawNet2	tf-SE	softmax	-	Voxceleb2	2.48	-
Wang et al.,2021 [13]	CNN+Transformer	-	-	-	Voxceleb2	2.56	-
Chung et al., 2018 [27]	ResNet-50	-	Softmax+Contrastive	512	Voxceleb2	3.95	-
Chung et al., 2018 [27]	ResNet-34	SE	Softmax+Contrastive	512	Voxceleb2	4.83	-
Zhu et al.,2021 [31]	Y-vector	tf-SE	AM-softmax	-	Voxceleb2	2.78	0.269
Ours	SpeechNAS	-	AM-softmax	512	Voxceleb2	3.07	0.3264
Ours	ThinResNet-50	-	AM-softmax	512	Voxceleb2	5.04	0.4651
Ours	Res2Net	-	AM-softmax	512	Voxceleb2	3.32	0.3199
Ours	Split-ResNet	-	AM-softmax	512	Voxceleb2	2.50	0.2426
Ours	Split-ResNet	DTFA	AM-softmax	512	Voxceleb2	2.09	0.2426

TABLE 6. Test results of different network structures in Voxceleb1-E.

	Voxceleb1-E						
	Model	Attention	loss	Dims	Training set	EER(%)	DCF
Jung et al.,2020 [11]	RawNet2	tf-SE	softmax	-	Voxceleb2	2.87	-
Chung et al., 2018 [27]	ResNet-50	SE	Softmax+Contrastive	512	Voxceleb2	4.42	-
Zhu et al.,2021 [31]	Y-vector	tf-SE	AM-softmax	-	Voxceleb2	2.64	0.270
Ours	SpeechNAS	-	AM-softmax	512	Voxceleb2	3.25	0.3238
Ours	ThinResNet-50	-	AM-softmax	512	Voxceleb2	5.32	0.4804
Ours	Res2Net	-	AM-softmax	512	Voxceleb2	3.12	0.2979
Ours	Split-ResNet	-	AM-softmax	512	Voxceleb2	2.76	0.2931
Ours	Split-ResNet	DTFA	AM-softmax	512	Voxceleb2	2.27	0.2453

In Table 3, we compare the experimental results of the Split-ResNet network structure using different attention mechanisms on the Voxceleb1-O test set. Through the table, we can see that our proposed DTFA mechanism has better results compared to other SE and CBAM attention mechanisms, while the parameters and inference time are similar. We compare the accuracy curves and loss curves during training in Fig. 5 and conclude that DTFA has the highest accuracy and the fastest loss decrease after the same number of training rounds compared to other attention mechanisms and is always ahead of other attention mechanisms. This indicates that the DTFA designed in this paper has a stronger feature extraction ability and can better distinguish the identity of the speaker.

In the end, the combination of Split-ResNet and DTFA is also tested and compared with other advanced, broader, and deeper systems in the Voxceleb-O test set in this paper, and the experimental results are shown in Table 5. The results show that the EER/MinDCF of the Split-ResNet+DTFA system proposed in this paper is 2.09/0.2426, which is significantly better than other advanced systems, proving that the Split-ResNet and DTFA proposed in this paper are better and more excellent than other systems in extracting speaker audio features, and the number of parameters is far less than other complex systems.

Meanwhile, to evaluate the performance of the system more comprehensively and accurately, this paper was tested again in the larger and more difficult test forms Voxceleb1-E and Voxceleb1-H. The test results on Voxceleb1-E, which used the whole Voxceleb1 as the test set, are shown in Table 6, and the results show that the Split-ResNet+DTFA

TABLE 7. Test results of different network structures in Voxceleb1-H.

Voxceleb1-H							
	Model	Attention	loss	Dims	Training set	EER(%)	DCF
Jung et al.,2020 [11]	RawNet2	tf-SE	softmax	-	Voxceleb2	4.69	-
Chung et al., 2018 [27]	ResNet-50	SE	Softmax+Contrastive	512	Voxceleb2	7.33	-
Zhu et al.,2021 [31]	Y-vector	tf-SE	AM-softmax	-	Voxceleb2	4.33	0.377
Ours	SpeechNAS	-	AM-softmax	512	Voxceleb2	5.09	0.4461
Ours	ThinResNet-50	-	AM-softmax	512	Voxceleb2	8.68	0.6504
Ours	Res2Net	-	AM-softmax	512	Voxceleb2	5.29	0.4704
Ours	Split-ResNet	-	AM-softmax	512	Voxceleb2	4.62	0.4193
Ours	Split-ResNet	DTFA	AM-softmax	512	Voxceleb2	3.99	0.3663

TABLE 8. Test results of different network structures in Voxceleb1-O(Extracting 80-dims FBank features).

Voxceleb1-O							
	Model	Attention	loss	Dims	Training set	EER(%)	DCF
Ours	RawNet2	tf-SE	softmax	-	Voxceleb2	2.36	-
Ours	CNN+Transformer	-	-	-	Voxceleb2	2.43	-
Ours	ResNet-50	-	Softmax+Contrastive	512	Voxceleb2	3.88	-
Ours	ResNet-34	SE	Softmax+Contrastive	512	Voxceleb2	4.72	-
Ours	Y-vector	tf-SE	AM-softmax	-	Voxceleb2	2.61	0.2578
Ours	SpeechNAS	-	AM-softmax	512	Voxceleb2	1.52	0.1325
Ours	ThinResNet-50	-	AM-softmax	512	Voxceleb2	4.35	0.4211
Ours	Res2Net	-	AM-softmax	512	Voxceleb2	3.01	0.2752
Ours	Split-ResNet	-	AM-softmax	512	Voxceleb2	1.92	0.1688
Ours	Split-ResNet	DTFA	AM-softmax	512	Voxceleb2	1.81	0.1594

system in this paper still can show better results. While on the Voxceleb-H test set using the same country and gender, the EER/MinDCF of the model both increased due to the reduced differences in accent and intonation, which made the similarity more difficult to distinguish. However, it is still possible to find from the results in Table 7 that the Split-ResNet+DTFA system proposed in this paper is much lower than the other speaker recognition systems with an EER/MinDCF of 3.99/0.3663. In addition, we additionally designed experiments using 80-dimensional FBank features as inputs to each speaker recognition system, and the results are shown in Table 8. From the table, we can see that our Split-ResNet results are worse compared to SpeechNAS, but better compared to all other systems. This is because this paper proposes a lightweight speaker recognition network, so it uses fewer 40-dimensional features to preprocess the speaker's audio features, so it performs worse on 80-dimensional features. But overall from the four test forms it is proved that the Split-ResNet+DTFA system proposed in this paper is a system with more discriminative frame-level features and better able to distinguish speakers with higher similarity.

VI. CONCLUSION

In this paper, we propose a split-residual network Split-ResNet based on the traditional standard 50-layer residual network, which can achieve better results than the original residual network by slicing a piece of data and then superimposing the residuals and convolution of each piece of data. The results show that the improved network improves the performance by about 25% with only 0.09M more parameters and 6ms more inference time than Res2Net. In addition, this paper also proposes an attention mechanism focusing on the time-frequency channel (DTFA), which divides the data into two aspects of features in the time and frequency domains by pooling, and learns information on the time and frequency domains channels, respectively. Using DTFA in combination with Split-ResNet proposed in this paper, an EER/MinDCF of 2.09/0.2426 was achieved in the Voxceleb1-O test dataset, while the results tested in the Voxceleb1-H test dataset and the Voxceleb1-E test dataset were also significantly better than those of other systems. It is proved that the proposed system combining Split-ResNet and DTFA in this paper is a speaker recognition system with good speaker audio feature extraction ability and discriminative ability.

REFERENCES

[1] L. Stoll, *Finding Difficult Speakers in Automatic Speaker Recognition*. Berkeley, CA, USA: Univ. of California at Berkeley, Computer Science, 2011.

[2] J. Zhou, T. Jiang, Z. Li, L. Li, and Q. Hong, "Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function," in *Proc. Interspeech*, Sep. 2019, pp. 2883–2887.

[3] P. Matejka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 4828–4831.

[4] X. Fang, T. Gao, L. Zou, and Z. Ling, "Bidirectional attention for text-dependent speaker verification," *Sensors*, vol. 20, no. 23, p. 6784, Nov. 2020.

[5] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4052–4056.

[6] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333.

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

[8] M. Wang, D. Feng, T. Su, and M. Chen, ''Attention-based temporal-frequency aggregation for speaker verification,'' *Sensors*, vol. 22, no. 6, p. 2147, Mar. 2022.

[9] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[10] H. Zeinali, S. Wang, A. Silnova, P. Matejka, and O. Plchot, ''BUT system description to VoxCeleb speaker recognition challenge 2019,'' 2019, *arXiv:1910.12592*.

[11] J. W. Jung, S. B. Kim, H. J. Shim, J. H. Kim, and H. J. Yu, ''Improved rawnet with filter-wise rescaling for text-independent speaker verification using raw waveforms,'' 2020, *arXiv:2004.00526*.

[12] M. Zhao, Y. Ma, M. Liu, and M. Xu, ''The SpeakIn system for VoxCeleb speaker recognition challange 2021,'' 2021, *arXiv:2109.01989*.

[13] R. Wang, J. Ao, L. Zhou, S. Liu, Z. Wei, T. Ko, Q. Li, and Y. Zhang, ''Multi-view self-attention based transformer for speaker recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6732–6736.

[14] W. Zhu, T. Kong, S. Lu, J. Li, D. Zhang, F. Deng, X. Wang, S. Yang, and J. Liu, ''SpeechNAS: Towards better trade-off between latency and accuracy for large-scale speaker verification,'' in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 1102–1109.

[15] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, ''Res2Net: A new multi-scale backbone architecture,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[16] S. Wang, Y. Qian, and K. Yu, ''Focal KL-divergence based dilated convolutional neural networks for co-channel speaker identification,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5339–5343.

[17] K. Okabe, T. Koshinaka, and K. Shinoda, ''Attentive statistics pooling for deep speaker embedding,'' in *Proc. Interspeech*, Sep. 2018, pp. 1–5.

[18] B. Desplanques, J. Thienpondt, and K. Demuynck, ''ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,'' in *Proc. Interspeech*, Oct. 2020, pp. 1–5.

[19] S. Yadav and A. Rai, ''Frequency and temporal convolutional attention for text-independent speaker recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6794–6798.

[20] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, ''Squeeze-and-excitation networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[21] S. Woo, J. Park, J.-Y. Lee, and I.-S. Kweon, ''CBAM: Convolutional block attention module,'' in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[22] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, ''Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges,'' *Exp. Syst. Appl.*, vol. 171, Jun. 2021, Art. no. 114591.

[23] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, ''Language recognition via ivectors and dimensionality reduction,'' in *Proc. Interspeech*, Aug. 2011, pp. 1–4.

[24] 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,'' 2017, *arXiv:1705.02304*.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, ''VoxCeleb: A large-scale speaker identification dataset,'' in *Proc. Interspeech*, Aug. 2017, pp. 1–6.

[26] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, ''Voxceleb: Large-scale speaker verification in the wild,'' *Comput. Speech Lang.*, vol. 60, Mar. 2020, Art. no. 101027.

[27] J. S. Chung, A. Nagrani, and A. Zisserman, ''VoxCeleb2: Deep speaker recognition,'' in *Proc. Interspeech*, Sep. 2018, pp. 1–6.

[28] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*.

[29] F. Wang, J. Cheng, W. Liu, and H. Liu, ''Additive margin softmax for face verification,'' *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.

[30] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, ''Utterance-level aggregation for speaker recognition in the wild,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5791–5795.

[31] G. Zhu, F. Jiang, and Z. Duan, ''Y-vector: Multiscale waveform encoder for speaker embedding,'' in *Proc. Interspeech*, Aug. 2020, pp. 1–5.

JIJI WANG received the bachelor's degree in computer science from the Chengdu University of Technology, in 2021, where he is currently pursuing the master's degree in computer technology. His research interests include machine learning, deep learning, and image recognition.

FEI DENG received the Ph.D. degree in earth exploration and information technology from the College of Information Engineering, Chengdu University of Technology, China, in 2007. Since 2004, he has been with the College of Computer and Network Security, Chengdu University of Technology, where he is currently a Professor. His research interests include artificial intelligence, deep learning, and computer graphics.

LIHONG DENG received the B.E. degree from the Department of Electronic Information Engineering, Sichuan University of Science and Engineering, China, in 2019. He is currently pursuing the master's degree in software engineering with the Chengdu University of Technology. His research interests include machine learning, deep learning, and speaker recognition.

PING GAO was born in August 1987. She received the B.S. degree in landscape architecture from Beihua University, Jilin, China, in 2010, and the M.S. degree in ornamental plants and horticulture from Beijing Forestry University, Beijing, China, in 2013. From 2016 to 2023, she held positions as a Researcher with the Research Center, Tianyi Group, for three years; as a Manager of the Flower and Shrub Integrated Cultivation Center, for two years; and as a Vice President for two years. Her research interests include the development and utilization of new excellent garden plant resources, the research and application of garden plant breeding technology, fabricated plant landscape, and the landscape construction technology of urban difficult sites. Her awards and honors include Third Prize of Science and Technology, Sichuan, in 2018, and the Third Prize of Science and Technology from the China Society of Landscape Architecture, in 2019.

YUANXIANG HUANG received the master's degree in landscape architecture from Sichuan Agricultural University, in 2008. He is currently a Senior Engineer. He is also the Chairperson and a General Manager of Sichuan Tianyi Ecological Garden Group Company Ltd. His research interests include landscape architecture, green buildings, and landscape architecture information models. He was awarded the leading entrepreneurial talent of the ''Tianfu Ten Thousand Talents Program,'' Sichuan, in 2020. He has won two Science and Technology Progress Awards of Sichuan Province and two Science and Technology Progress Awards from the Chinese Society of Landscape Architecture.