Journal of Computers Vol. 34 No. 3, June 2023, pp. 75-91
doi: 10.53106/199115992023063403006

End-to-end Speaker Recognition Based on MTFC-FullRes2Net

Li-Hong Deng[1], Fei Deng[1*], Ge-Xiang Chiou[2,3], Qiang Yang[3]

[1] School of Computer and Network Security (College of Computer Science and Cyber Security (Oxford Brookes College)),
Chengdu University of Technology, Chengdu, 610059, China
dengfei@cdut.edu.cn, dengnima233@163.com

[2] Artificial Intelligence Research Center, Chengdu University of Technology, Chengdu 610059, China
zhgxdylan@126.com

[3] School of Control Engineering, Chengdu University of Information Engineering, Chengdu 610059, China
qiangychd@126.com

Received 14 June 2022; Revised 11 September 2022; Accepted 28 October 2022

Abstract. The feature extraction ability of lightweight convolutional neural networks in speaker recognition systems is weak. And recognition accuracy is poor. Many methods use deeper, wider, and more complex network structures to improve the feature extraction ability. But it makes the parameters and inference time increase exponentially. In the paper, we introduce Res2Net in target detection task to speaker recognition task and verify its effectiveness and robustness in the speaker recognition task. And we improved and proposed FullRes2Net. It has better multi-scale feature extraction ability without increasing the number of parameters. Then, we proposed the mixed time-frequency channel attention to solve the problems of existing attention methods to improve the shortcomings of convolution itself and further enhance the feature extraction ability of convolutional neural networks. Experiments were conducted on the Voxceleb dataset. The results show that the MTFC-FullRes2Net end-to-end speaker recognition system proposed in this paper effectively improves the feature extraction and generalization ability of the Res2Net. Compared to Res2Net, MTFC-FullRes2Net performance improves by 31.5%. And Compared to ThinResNet-50, RawNet, CNN+Transformer and Y-vector, MTFC-FullRes2Net performance is improved by 56.5%, 14.1%, 16.7% and 23.4%, respectively. And it is superior to state-of-the-art speaker recognition systems that use complex structures. It is a lightweight and more efficient end-to-end architecture and is also more suitable for practical application.

Keywords: speaker recognition, res2net, attention mechanisms

1 Introduction

The speaker recognition task aims to identify the speaker by obtaining identity information from the audio [1]. Because of the widespread use of voice commands, speaker recognition has become a necessary measure to protect user security and privacy. However, the recording environment may be noisy and contain music, laughter, chatting background sounds, etc. Consequently, some audio may not contain identifying information about the speaker. And the speaker's internal factors (e.g., accent, emotion, intonation, and speaking style) also have an impact [2]. In addition, the duration of the audio may also be as short as 2 to 5 seconds, making it contain very little speaker identity information. Building a lightweight speaker recognition system that can extract discriminative features from variable-length audio is the key to applying speaker recognition in practice.

Before the upswing of deep learning, traditional i-vector systems with probabilistic linear discriminative analysis (PLDA) have been leading the way as the dominant method for speaker recognition [3-4, 7]. With the development of deep learning, deep neural networks (DNNs) have brought substantial improvements to speaker recognition. Compared to traditional i-vector systems, DNN architectures can directly process noisy datasets to extract frame-level features for training through deep neural networks. DNN-based speaker recognition systems have achieved better performance than i-vector systems and have dominated with excellent feature extraction capabilities [5-6]. Building an effective feature extractor (deep neural network) is the key to effective speaker recognition in speaker recognition. CNN-based feature extractors are widely used as the backbone for speaker recognition systems due to their superior feature extraction capability [7-10]. However, the feature extraction

* Corresponding Author

ability of the lightweight convolutional neural network is weak and cannot acquire distinguishing features, resulting in poor recognition. In 2018 Chung et al. used ThinResNet-50 for speaker recognition but achieved far worse recognition results than the i-vector system [7]. Therefore, many current speaker recognition methods use deeper, wider, and more complex network structures to obtain better feature extraction ability and enhance system performance. Such as Wang et al. used a 256-layer residual network for speaker recognition in 2019 [11]. In 2020 He et al. proposed RawNet, a truly end-to-end neural vocoder, which use a coder network to learn the higher representation of the signal, and an autoregressive voder network to generate speech sample by sample [12]. In 2021 Wang et al. proposed a CNN+Transformer speaker recognition system to balance the capabilities of capturing global dependencies and modeling the locality [13]. In 2021, Zhu et al. proposed a Y-vector speaker recognition system. It uses three convolution branches with different time scales to compute speech features from the waveform. These features are then processed by squeeze-and-excitation blocks, a multi-level feature aggregator, and a time delayed neural network (TDNN) to compute speaker features [14]. However, as the depth, width, and complexity increase, the parameters of the model increase exponentially. It makes the application of speaker recognition systems to real life very difficult. In addition, the training and testing conditions are often incompatible in speaker recognition. Simply increasing the depth and width of the network is easy to overfit, which weakens the generalization ability of the whole network. Therefore, building a lightweight network structure with better feature extraction capability is essential. It is also the key to applying speaker recognition to practical applications.

Convolutional neural networks are not perfect either. The convolution uses a fixed-size convolutional kernel to capture the time and frequency information of the audio. And the size of the convolutional kernel limits the receptive field of the features, resulting in the feature extraction capability being also limited. As the attention mechanism has evolved, it has shown excellent performance in computer vision. Therefore, attentional mechanisms are also gradually introduced into speaker recognition [15-19]. The attention mechanism improves the disadvantage that convolutional neural networks can only extract local features. And it effectively enhances the feature extraction and generalization ability of convolutional neural networks. It improves the performance of the system significantly with few parameters. Zhou et al. used Squeeze-and Excitation (SE) attention [15] in ResNet to enhance the feature extraction capability of convolutional neural networks. It is the first time that speaker recognition introduces an attention mechanism. SE attention can capture the dependencies between channels and focus on the more critical channel features. Sarthak et al. were inspired by the Convolutional block attention module (CBAM) attention module in computer vision [18] and proposed the tf-CBAM attention [19]. The tf-CBAM attention emphasizes the time and frequency dimension of the features. Although these attention methods effectively enhance the feature extraction ability of neural networks. However, all only do simple attention learning by using pooling operations, emphasizing a single feature dimension, and ignoring other dimensions and the interaction of time-frequency-channel dimensions. The pooling operation also makes the features lose the speaker identity information.

Based on the above problems, this paper introduces Res2Net in target detection to speaker recognition [20]. And we verify its effectiveness and robustness in speaker recognition. Most current methods enhance the feature extraction capability of the network by increasing the depth, width, and complexity. Res2Net adopts a parallel branch structure without increasing the depth and width but has fewer parameters and better feature extraction capability. However, the network structure of Res2Net limits its feature extraction capability. Therefore, this paper proposes FullRes2Net based on the Res2Net. Compared with Res2Net, the feature extraction capability of FullRes2Net improved significantly. Performance improves by 15.4% with almost no increase in parameters. It can deal with more complex acoustic environments and recognition tasks. This paper proposes an attention method for speaker recognition tasks, the Mixed Time-Frequency Channel attention (MTFC) module, to solve the problems of existing attention methods and improve the feature extraction ability of convolutional neural networks. It can interact with the time, frequency, and channel dimensions, capture the dependencies among features, and get more attention information and global information. Thus, the feature extraction capability of the convolutional neural network is effectively enhanced. To prove the effectiveness of the proposed method, we performed different ablation experiments in this paper. And our designed speaker recognition system based on mixed time-frequency channel attention and FullRes2Net is compared with Res2Net and other current state-of-the-art speaker recognition systems in a variety of experimental settings. It achieved significantly better performance than these state-of-the-art systems. Compared to Res2Net, MTFC-FullRes2Net performance improves by 31.5%. And Compared to ThinResNet-50, RawNet, CNN+Transformer and Y-vector, MTFC-FullRes2Net performance is improved by 56.5%, 14.1%, 16.7% and 23.4%, respectively. Our proposed methods also perform better in more tough environmental situations, such as short-time speaker recognition.

The main contributions of this work are summarized as follows:

(1) We verify its effectiveness and robustness in speaker recognition. It is the first time the network structure from a target detection task is applied to speaker recognition.

(2) We propose FullRes2Net based on Res2Net, which has a better feature extraction ability and can deal with more complex acoustic environments and recognition tasks. The parameters are almost not increased.

(3) We propose mixed temporal-frequency channel attention. It can interact in the temporal, frequency, and channel dimensions and enhance the feature extraction ability of convolutional neural networks more effectively.

(4) We build an end-to-end speaker recognition system with mixed time-frequency channel attention and FullRes2Net, which achieves better performance than current state-of-the-art speaker recognition systems in different settings. It shows the superiority of the methods and has fewer parameters, faster inference time, and higher accuracy. It is also more suitable for the application in reality.

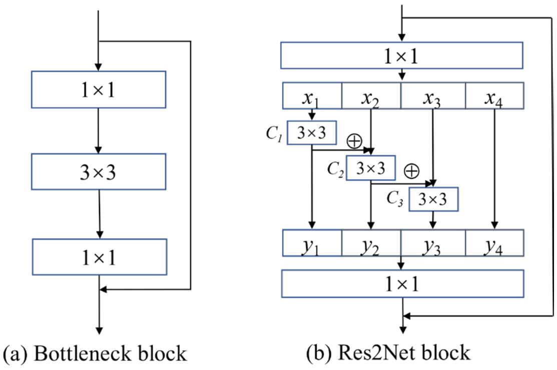
In the rest of this paper, we mainly give a brief review of the related Res2Net, SE attention, and tf-CBAM attention in Section 2. In section 3, we present the details of the proposed method. In the meanwhile, we introduce the dataset used, training details, testing details and testing methods in section 4. We discussed and analyzed the experimental results in section 5. Finally, we summarize the work done and explain the limitations of the current work and future research directions in section 6.

2 Related Works

2.1 Review of Res2Net

Res2Net [20] is a network structure applied in target detection. It aims to improve the feature extraction ability of convolutional neural networks by increasing the size of the receptive field. It takes a parallel branch structure, which is implemented by connecting smaller convolution operators similar to layer residuals. The ResNet residual block and Res2Net block are shown in Fig. 1. Where ⊕ denotes the addition operation. In the ResNet residual block, features proceed through a 1×1 convolution operation and then a 3×3 convolution operation. While in the Res2Net block, the feature $x \in R^{T \times F \times C}$ is divided into s feature subsets $x_i \in R^{T \times F \times w}$ after the 1×1 convolution. Each feature subset has w channels ($C = s \times w$). Therefore, Res2Net also has fewer parameters than ResNet. After the dividing operation, each feature subset x_i ($i \in \{1,2,..., s\}$) has a 3×3 convolution, denoted as C_i. The output y_i is expressed as shown in Equation (1).

$$y_i = \begin{cases} C_i(x_i), & i = 1 \\ C_i(x_i + y_{i-1}), & 1 < i < s. \\ x_i & i = s \end{cases} \tag{1}$$



(a) Bottleneck block

(b) Res2Net block

Fig. 1. ResNet residual block and Res2Net block

In the Res2Net block, when C_i receives a previous feature subset output of $y_{i\text{-}1}$, the prior subset is equivalent to performing the convolutional operations C_i and $C_{i\text{-}1}$ in turn. Thus, it makes the current feature subset have a larger receptive field. Thus, various combinations of different receptive fields are obtained and efficiently express the phonological features, as shown in Equation (2). Finally, all y_i is connected as the output feature y.

$$\begin{aligned} & y_1 = C_1(x_1) \\ & y_2 = C_2(x_2 + y_1) = C_2(x_2 + C_1(x_1)) \\ & \vdots \\ & y_{n-1} = C_{n-1}(x_{n-1} + y_{n-2}) = C_{n-1}(x_{n-1} + C_{n-2}(x_{n-2})). \end{aligned} \tag{2}$$

2.2 Review of SE Attention

SE attention [15] is widely embedded in CNN to learn the inter-dependence of channels in intermediate features. SE attention adopts global average pooling (GAP) to squeeze the 3D temporal-channel-frequency tensor into a 1D channel-wise feature. Then it uses two fully connected (FC) layers and a sigmoid method to weigh different channel importance for boosting speaker representation learning ability. Given the intermediate tensor $x \in R^{T\times F\times C}$ is an input feature in networks, T is the number of frames in the temporal domain, F is the number of frequency bins, and C is the number of channels (convolutional kernels). The average pooling on time and frequency dimension of SE attention is formulated as:

$$x^C = \frac{1}{T \times F}\sum_{i=1}^{T}\sum_{j=1}^{F} x_{ij}. \tag{3}$$

Where $x^C \in R^{1\times 1\times C}$ is the results of the global average pooling of the intermediate feature x on time and frequency dimensions. Next is to learn the channel-wise attention weights of x^C. The channel-wise attention mask learning is formulated as:

$$w = \delta(W_2^{C\times C'} \times (relu((W_1^{C'\times C} \cdot x^C)))). \tag{4}$$

Where · means matrix multiplication. $w \in R^{1\times 1\times C}$ is the channel-wise attention mask. $W_1^{C'\times C}$ and $W_2^{C\times C'}$ are two fully-connected layers which capture the inter-dependence of channels in the intermediate feature x. The dimensional reduction factor $r = C/C'$ indicates the reduction ratio to avoid the parameters overhead. Finally, a sigmoid function δ is used to scale the channel-wise weights. The attention mask is a set of attention factors predicted by supervised speaker classification loss aiming to emphasize essential channels and compress the useless channels. The recalibration operation means that the attention mask dot multiplies with the original tensor x, as shown in Equation (5).

$$y = w \times x. \tag{5}$$

2.3 Review of tf-CBAM Attention

ft-CBAM [19] comprises f-CBAM and t-CBAM applied in parallel on the input features. The output features generated by the two are then averaged. For modeling frequency attention, the f-CBAM module needs to limit the receptive field of the attention module to focus only on the frequency dimensions of the input. f-CBAM module aggregates temporal information by averaging the input feature $x \in R^{T\times F\times C}$ along the time dimensions to generate an efficient feature descriptor $F_{freq} \in R^{1\times F\times C}$ which essentially assigns equal statistical importance to each temporal frame.

$$F_{freq} = AvgPool_{1\times T}(x). \tag{6}$$

Where $AvgPool_{1\times T}$ represents the average pooling operation with a kernel of size 1×T over the input feature x.

It then aggregates channel information by generating two feature maps: F^f_{avg}, $F^f_{max} \in R^{1\times F\times 1}$. F_{avg} and F_{max} denote average and max pooling operations applied across the channel dimension on F_{freq}, and concatenate them. Finally, on this concatenated feature descriptor, it applies a rectangular 7x1 convolution kernel to generate a frequency attention map $M_{freq} \in R^{F\times 1}$.

$$M_{freq} = \sigma(f^{7\times 1}([F^f_{avg}; F^f_{max}]). \tag{7}$$

Here, σ denotes the sigmoid function and $f^{7\times 1}$ represents a convolution operation with a rectangular 7 × 1 kernel. M_{freq} is then broadcasted along the temporal dimension on the original input feature x.

$$x^f = M_{freq} \times x. \tag{8}$$

t-CBAM follows a procedure similar to f-CBAM for modelling temporal attention, albeit limiting the receptive field of the attention module to the temporal dimensions of the input.

$$F_{temp} = AvgPool_{F\times 1}(x). \tag{9}$$

$$M_{temp} = \sigma(f^{1\times 7}([F^t_{avg}; F^t_{max}]). \tag{10}$$

$$x^t = M_{temp} \times x. \tag{11}$$

Where $F_{freq} \in R^{T\times 1\times C}$; and F^t_{avg} , $F^t_{max} \in R^{T\times 1\times 1}$.

3 Proposed MTFC-FullRes2Net Network

Our proposed end-to-end speaker recognition system is based on mixed time-frequency channel attention and FullRes2Net (MTFC-FullRes2Net), as shown in Fig. 1. It consists of two parts: (1) FullRes2Net, and (2) mixed time-frequency channel attention. We use FullRes2Net as the backbone network of the feature extractor. It can generate larger receptive fields and more combinations of different receptive fields, obtaining more informative, comprehensive, and distinctive speaker features. And it has fewer parameters and a faster inference time. To improve the disadvantages of convolution and enhance the feature extraction ability of the convolutional neural network more effectively, we designed the mixed time-frequency channel attention. It can interact with the time, frequency, and channel dimensions to get more attention information and global information, thus effectively enhancing the feature extraction ability of convolutional neural networks. We introduce it into the entire front-end model. Mixed time-frequency attention can be inserted anywhere, but attention-based convolution shows that parallelizing convolution layers and attention mechanisms is a more efficient structure to deal with both short-term and long-term dependencies. Therefore, we embed the mixed time-frequency attention into FullRes2Net, as shown in the red box part in Fig. 2.

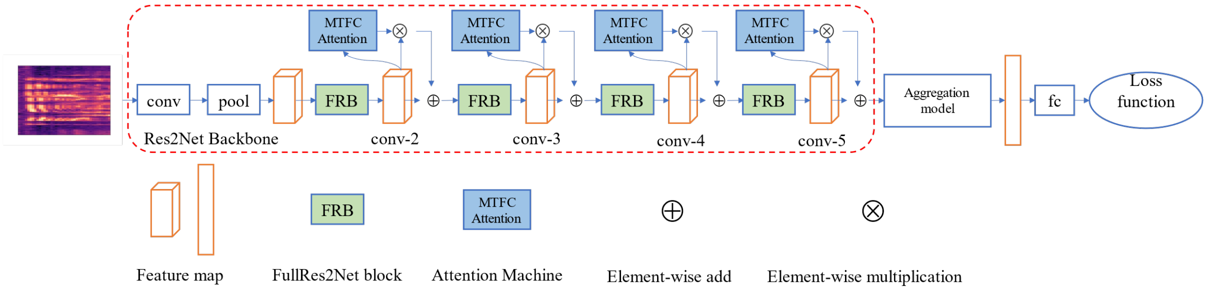


Fig. 2. Overview of the FullRes2Net-Mixed TimeAttention Network

3.1 FullRes2Net

However, the network structure of Res2Net can only generate a fixed combination of receptive fields and cannot obtain larger receptive fields, which also limits its feature extraction capability. Therefore, we propose FullRes2Net, as shown in Fig. 3.

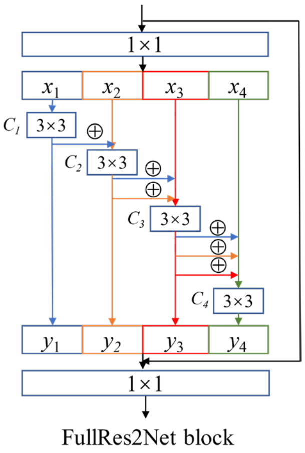


Fig. 3. Overview of the FullRes2Net block

Unlike Res2Net, each feature subset in the parallel branch of FullRes2Net will fuse the outputs of all previous feature subsets, which are then convolved by the convolution operator of the current subset, and passed to the subsequent feature subsets. Finally, all the y_i are connected as the output feature y, as shown in Equation (12). In this way, each feature subset collects the output of all previous feature subsets. Making the receptive field of each previous feature subset increase, thus obtaining more speaker identity information and more combinations of receptive fields, as shown in Equation (13). Therefore, compared with Res2Net, FullRes2Net can gain more comprehensive feature detail information and more effectively extract the speaker identity features in audio. FullRes2Net does not introduce convolutional operations to increase the depth and width of the network, so it only adds some inference time.

$$y_i = \begin{cases} C_i(x_i), & i = 1 \\ C_i(x_i + y_{i-1} + \cdots + C_{i-1}C_{i-2}\cdots C_2(y_1)), & 1 < i \le s \end{cases}. \tag{12}$$

$$\begin{aligned} & y_1 = C_1(x_1) \\ & y_2 = C_2(x_2 + y_1) = C_2(x_2 + C_1(x_1)) \\ & \vdots \\ & y_n = C_n(x_n + \cdots + C_{n-1}\cdots C_2(y_1)) = C_n(x_n + \cdots + C_{n-1}\cdots C_1(x_1)). \end{aligned} \tag{13}$$

As shown in Equation (14), it represents the way of calculating the receptive field. Where RF_i denotes the receptive field of the ith layer, k_i denotes the convolution kernel size of the ith layer, and s_j denotes the convolution step size of the jth layer. Suppose that after the 1×1 convolution operation, the receptive field is 1, and the convolution step size is 1. We can derive the receptive field size on different parallel branches in Res2Net and FullRes2Net according to Equations (2) and (13). The results are shown in Equation (15) and Equation (16). We can find that FullRes2Net has a larger receptive field and more combinations of receptive fields than ResNet, which makes FullRes2Net have a better feature extraction ability than Res2Net.

$$RF_i = RF_{i-1} + (k_i - 1)\prod_{j=1}^{i-1} s_j. \tag{14}$$

$$RF_i(\text{Res2Net}) = \begin{cases} 3, & i = 1 \\ 5, & 1 < i < s \\ 1, & i = s \end{cases}. \quad \textbf{(15)}$$

$$RF(\text{FullRes2Net}) = \begin{cases} 3 & , \quad = 1 \\ 3 + (3-1) \times i, & 1 < i \le s \end{cases}. \quad \textbf{(16)}$$

3.2 Mixed Time-frequency Channel Attention

Although FullRes2Net has better feature extraction capability, it still cannot overcome the disadvantages of the convolutional operation itself. And to solve the problems of current attention methods. We design the Mixed Time-Frequency Channel Attention (MTFC). Mixed time-frequency channel attention can integrate time-frequency and channel information and generate time-frequency attention and channel attention. The MTFC attention mechanism aims to focus on significant regions in time-frequency features and channel features, to obtain the dependencies between different feature dimensions, and to obtain more attention information. It makes the network able to extract more discriminative frame-level features.

As shown in Fig. 4, the MTFC attention mechanism contains two attention methods. It captures the dependencies between local features to map global features, thus obtaining a better feature representation. We use x as the input features. These features are fed into the time-frequency attention module and the channel attention module to produce a two-dimensional time-frequency attention and a global channel feature that is the same size as the input feature. Time-frequency attention and channel attention are weighted in different ways. Time-frequency attention is weighted by multiplication, which has global information to emphasize the more important regions of the time-frequency information. And channel attention is weighted by summation, which simulates long-term dependencies between features and helps improve feature discriminability. Thus the overall process can be summarized as follows:

$$\begin{gathered} x' = M_{TF}(x) \otimes x \\ x'' = M_C(x) \oplus x \\ y = x' + x''. \end{gathered} \quad \textbf{(17)}$$

The M_C denotes the channel attention module. The M_{TF} is the time-frequency attention module. Throughout the process, channel attention spreads along the time-frequency dimension, while time-frequency attention spreads along the channel dimension. y is the final output. The following describes the details of each attention module.

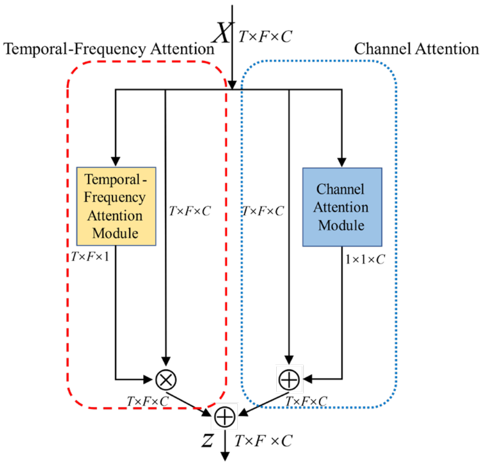


Fig. 4. Overview of the mixed time-frequency channel attention

Time-Frequency Attention. In computer vision, attention-based methods are particularly prospective in modeling spatial relationships. The attention mechanism enables the model to focus on critical features, suppress inessential features, and capture global features. Thus, the feature extraction capability of convolutional neural networks is enhanced. Likewise, the attention method can be applied to modeling the time-frequency relationships. We generate the time-frequency weight matrix using the time-frequency relationship between features, and the weight scale correlates positively with frequency. To make the learned convolutional layers more competitive, we use the softmax function to generate the weight matrix instead of the sigmoid function. The softmax function encourages different convolutional layers to learn distinct features, thus making the model more robust [21]. Finally, the frequency margins of the original features are readjusted according to the generated weight matrix.

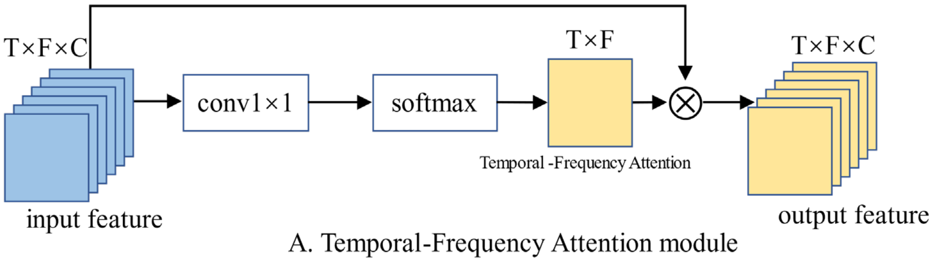

T×F×C
conv1×1
softmax
T×F
T×F×C
Temporal -Frequency Attention
input feature
output feature
A. Temporal-Frequency Attention module

Fig. 5. Details of the time-frequency attention sub-module

As shown in Fig. 5, given a feature input $x \in R^{T\times F\times C}$, the feature x is convolved by a 1×1 convolution operation to obtain a two-dimensional matrix A of size $T\times F$. The 1×1 convolution allows for the interaction of channel information and dimensionality reduction. Also, it increases the nonlinearity of the convolutional neural network. Then, using the softmax function for matrix A, a value is assigned to each position indicating the importance of that position $w(i, j)$,$w \in R^{T\times F}$. The weight matrix A is multiplied with the original input feature x to readjust the activation margins and output the feature x'. The equation is as follows:

$$A(i,j) = W_s(x). \tag{18}$$

$$w(i,j) = \frac{e^{A(i,j)}}{\sum_{i,j} e^{A(i,j)}}. \tag{19}$$

$$x' = x \times w. \tag{20}$$

Where $W_s \in R^{1\times 1\times C}$ denotes a 1×1 convolution.

Channel Attention. The channel attention module can be divided into three parts 1): Background modeling, after 1 × 1 convolution to generate weight values by softmax function, and then weighted mapping to get global channel features. 2) Feature transform, using 1 × 1 convolution for feature transform, captures channel dependence and reduces the model parameters. 3) Feature fusion, we use an additive approach to fuse the global features into the original features. The entire process is shown in Fig. 6.

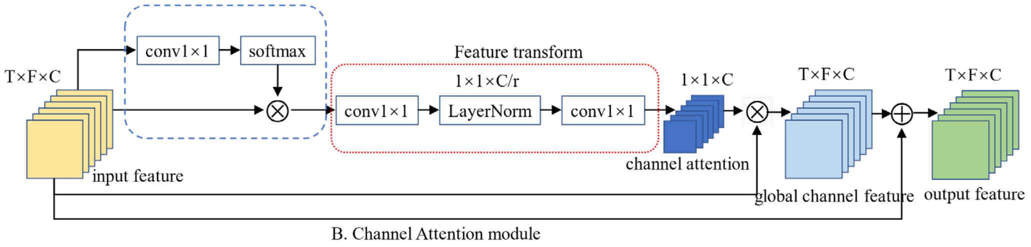


Fig. 6. Details of the channel attention sub-module

Using $x \in R^{T\times F\times C}$ denotes the input features. The 1×1 convolution is used for the interaction of the channel information and dimensionality reduction to make full use of the time-frequency information. And it obtains the feature $h \in R^{T\times F\times 1}$. The softmax function is then used to generate the global weights $w \in R^{N\times 1}$ (N=T×F), and the weights w are matrix multiplied with x to obtain the global channel features. The matrix multiplication aggregates the time-frequency information (by weighting all the time-frequency information with a weight matrix) into the channel information so that the global channel features sufficiently use the time-frequency information. After using feature transforms on global channel features, we use the softmax function to generate channel weights. The channel weights are multiplied with the original input features to readjust the activation margins of the channel information of the features. It then adds to the original input features x, thus simulating the long-term dependency between features and helping to improve feature discriminability. The $\delta(\cdot)$ represents the feature transforms part, and the structure is similar to the SE attention used to reduce the parameters and capture the channel dependence. In the SE attention, the feature transform is implemented by full connection. While experiments in Reference [22] show that full connection capturing channel dependencies is inefficient and unnecessary. Therefore, in Reference [22], the researchers used one-dimensional convolution to capture channel dependence. However, the size of the convolution kernel limits the coverage, and one-dimensional convolution does not operate on all channels. The dependencies between channels captured by 1D convolution are still local. We choose to use a 1×1 two-dimensional convolution, which allows complex learnable interactions of all channel features and can capture dependencies between channels more efficiently:

$$h = W_k x. \tag{21}$$

$$w = \frac{e^{h_i}}{\sum_{m=1}^{N} e^{h_m}}. \tag{22}$$

$$\delta(\cdot) = W_2(LN(W_1(\cdot))). \tag{23}$$

$$w_C = \text{softmax}(\delta(w^T x)). \tag{24}$$

$$x'' = x + w_C \times x. \tag{25}$$

Where r denotes the scale factor, $W_k \in R^{1\times 1\times C}$, $W_1 \in R^{1\times 1\times C/r}$, and $W_2 \in R^{1\times 1\times C}$ are 1×1 convolution operations, and LN is the LayerNorm used to prevent gradient disappearance and accelerate the network convergence. We add the outputs of the two attention modules directly achieving feature fusion, as shown in Equation (17). We do not use cascade operation because more storage space is required. The MTFC attention can be directly inserted into the convolutional network. And it only adds a small number of parameters, but can effectively enhance the feature representation.

4 Experimental Setup

The experimental speaker dataset was adopted from the Voxceleb dataset, which has been commonly used for speaker recognition tasks in recent years. Voxceleb is a large text-independent speaker recognition dataset containing the Voxceleb1 and the Voxceleb2 dataset, and Voxceleb2 contains more than 1 million audio clips of 5,994 speakers extracted from YouTube videos. The average time duration was 7.8S, from different acoustic environments, making speaker identification more challenging. Voxceleb1 contains over 100,000 audios from 1,251 speakers. There are three test sets Voxceleb1-O, Voxceleb1-E, and Voxceleb1-H. Voxceleb1-O is a test set that includes 40 speakers independent of Voxceleb1 and does not overlap with the speakers in Voxceleb1. The Voxceleb1-E test set uses the entire Voxceleb1 dataset, while the Voxceleb1-H test set is more specific. It contains samples from the same country of nationality and the same gender. The Voxceleb2 dataset is an extended version of the Voxceleb1 dataset, but the two datasets are mutually exclusive. As mentioned in reference [23], Voxceleb2 contains some flaws in its annotation. Therefore, it is not recommended to test models. It is widely used for training. As with most existing references, we use Voxceleb2 for training and Voxceleb1 as the test set.

Table 1. Voxceleb training set

Split	Speaker	Utterances
Voxceleb1	1211	148642
Voxceleb2	5994	1092009

Table 2. Voxceleb test set

Dataset	Speakers	Utterances	Pairs
Voxceleb1-O	40	4715	37720
Voxceleb1-H	1251	145375	581480
Voxceleb1-E	1190	138137	552536

4.1 Training Details

The experiment used the same simple training method as the comparative literature to verify the performance improvement of the system. We selected the 40-dimensional FilterBanks (FBank) as the input to the deep convolutional neural network without voice activity detection (VAD) and data augmentation (DA). And We also do not use complex processing at the back-end. Optimization is using the Adam optimizer with an initial learning rate of 0.001. During the training, we cut out a three-second clip from the audio. Choosing the same structure as the ThinResNet-50 network, ThinResNet-50 has the same structure as the standard ResNet-50 network only to reduce the computation cost, and the number of channels per residual block becomes 1/4 of the norm block. The structure of Res2Net and MTFC-FullRes2Net is shown in Table 3. T denotes the time dimension.

Table 3. Structure of different networks

Stage	ThinResNet-50	Res2Net	FullRes2Net	Output
Conv1		Conv2d, 7×7, 16, stride=2 maxpool, 3×3, stride=2		40×T×16
Conv2	[Bottleneck block,16]×2	[Res2Net block,16]×2	[FullRes2Net block,16]×2	40×T×16
Conv3	[Bottleneck block,32]×3	[Res2Net block,32]×3	[FullRes2Net block,32]×3	20×$^{T}/_{2}$×32
Conv4	[Bottleneck block,64]×3	[Res2Net block,64]×3	[FullRes2Net block,64]×3	10×$^{T}/_{4}$×64
Conv5	[Bottleneck block,128]×3	[Res2Net block,128]×3	[FullRes2Net block,128]×3	5×$^{T}/_{8}$×128

4.2 Testing and Testing Standards

During the test phase, we used the same settings as [21], extracted ten 3-second segments from each test audio as samples, and then sent them to the system to extract the utterance-level features of each segment and calculate the distance between all combinations (10×10=100) of pairs of segments. Then, the average of 100 distances denotes the score.

This study adopts the commonly used equal error rate (EER) and minimum detection cost function 2010 (DCF10) as the evaluation indices to objectively evaluate the performance of different aggregation models. They indicate that the smaller the value, the better the performance. The calculation formula of the minimum detection cost function is:

$$DCF = C_{FR} F_{FR} \mathrm{P}_{target} + C_{FA} F_{FA} (1 - P_{target}). \tag{26}$$

Where C_{FR} and C_{FA} are the weights of false rejection rate F_{FR} and false acceptance rate F_{FA}, respectively, and P_{target} and 1-P_{target} are the prior probability of real speaking and impersonation tests. We use the parameters C_{FA}=1, C_{FR}= 1, and P_{target}= 0.01 (DCF10) set by NIST SRE2010. DCF not only considers the different costs of false rejection and false reception but also considers the prior probability of the test. Therefore, MinDCF is more informative than EER in model performance evaluation.

5 Results

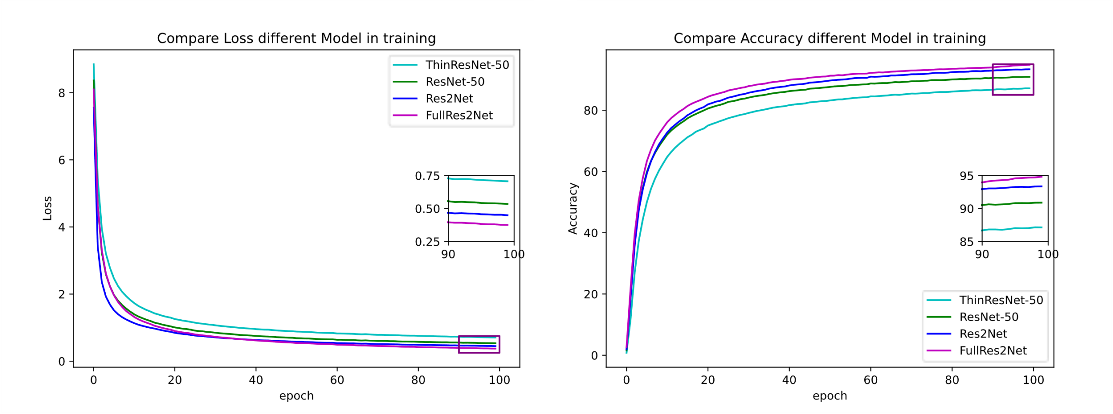
To verify the effectiveness of the proposed FullRes2Net and mixed time-frequency channel attention, we first perform two sets of ablation experiments in this paper. Table 4 shows the results of FullRes2Net with other network structures on the Voxceleb1-O test set under the same experimental conditions (both using the simple training method). The table shows that ThinResNet-50 with a lightweight structure performs the worst result with EER/DCF of 5.04%/0.4551, while ResNet-50 with a deeper and wider network structure shows a significant performance improvement with EER/DCF decreasing to 3.3%/0.33. However, its parameters and inference time also increase exponentially, as shown in Table 5. Introducing Res2Net in the target detection task to the speaker recognition task and testing it, it is observed that Res2Net achieves a lower EER/DCF of 3.32%/0.3199 than ThinResNet-50 and ResNet-50. We can find that compared to ThinResNet-50, the number of Res2Net parameters increased by only 0.89M, and inference time increased by only 18ms, but much less than ResNet-50, as shown in Table 5. It proves that Res2Net is a lightweight network with more excellent feature extraction ability and fewer parameters. And our proposed FullRes2Net further improves the performance compared to Res2Net, achieving the lowest EER/DCF of 2.75%/0.2861. It shows that FullRes2Net has better feature extraction ability and extracts the speaker identity information in the audio more effectively, which results in more discriminative frame-level features. It is also clear from Table 5 that FullRes2Net is a better lightweight feature extractor compared to Res2Net. It increases the number of parameters by 0.06M and the inference time by only 8ms, but the performance improves by 17%. During the training process, we record the loss curves and accuracy curves of networks. The loss value change curve and accuracy curve of the training set are shown in Fig. 7. As shown in the loss curve in Fig. 7(a), the loss value of FullRes2Net proposed in this paper is lower than that of other network structures, and the best convergence effect is achieved. It indicates that FullRes2Net proposed in this paper can extract features more efficiently and converge better. As shown in the accuracy curves in Fig. 7(b), FullRes2Net is consistently the highest compared to all network structures and superior to Res2Net. These experimental results show that our proposed FullRes2Net is a lightweight network structure with better feature extraction capability and obtains frame-level features more effectively and distinctively.

Table 4. Test results of different networks in Voxceleb1-O

	Model	Loss	Dims	Training set	EER (%)	DCF10
Chung et al., 2018	ResNet-50	softmax	512	Voxceleb2	3.95	-
Chung et al., 2018	ThinResNet-50	softmax	512	Voxceleb2	5.04	0.4651
Gao et al., 2019	Res2Net	softmax	512	Voxceleb2	3.32	0.3199
Ours	FullRes2Net	softmax	512	Voxceleb2	2.75	0.2861

Table 5. Number of parameters and inference time for different networks

	Model	Params (M)	Time (ms)
Chung et al., 2018	ResNet-50	21.4	112
Chung et al., 2018	ThinResNet-50	1.34	50
Gao et al., 2019	Res2Net	2.23	68
Our	FullRes2Net	2.29	76



(a) Loss curves for different network structures

(b) Accuracy curves of different network structures

Fig. 7. Loss curves and accuracy curves of different networks in the training

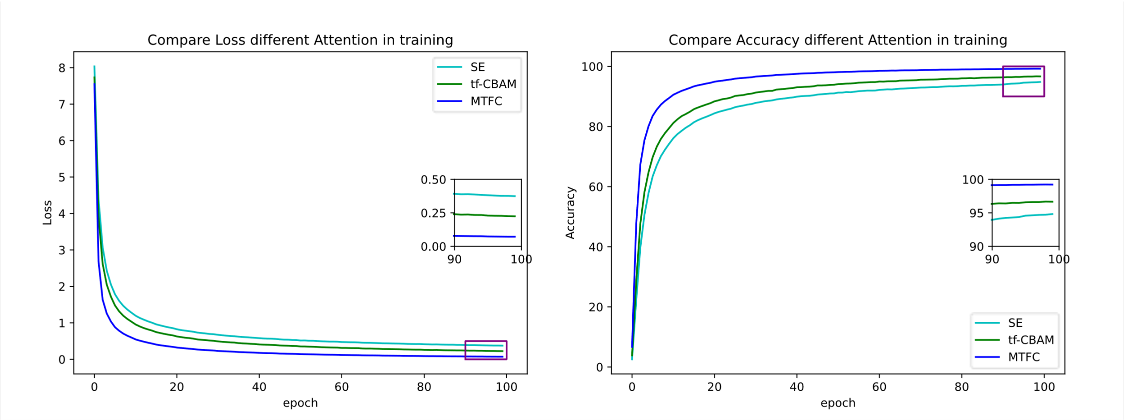
Table 6 shows our proposed mixed time-frequency channel attention with different attention methods on the Voxceleb1-O test set. Again, the test is performed under the same experimental conditions, but the backbone architecture of the feature extractor is directly selected as FullRes2Net. As shown in Table 6, using attention significantly improves the performance of the baseline system and enhances the feature extraction ability of the convolutional neural network. SE attention reduces the EER/DCF of the system to 2.63%/0.2543 by emphasizing the channel dimension of the feature. The tf-CBAM attention further improves the performance by focusing on the temporal and frequency dimensions of the features, achieving a lower EER/DCF of 2.46%/0.2523 than the SE attention. The MTFC attention proposed in this paper obtained the lowest EER/DCF of 2.23%/0.2243 and significantly outperformed the SE and tf-CBAM attention. It proves that MTFC attention is the more excellent attention method, which captures the dependence between different dimensions of the features and obtains more attentional information by interacting on the time-frequency-channel dimensions. Also, using two-dimensional 1×1 convolution captures the dependencies between features more effectively, thus improving the feature extraction ability of convolutional neural networks more efficiently than other attention methods. These experiments support that the using pooling in SE attention and tf-CBAM attention loses the speaker identity information and does not perform accurate attention learning. Table 7 shows the number of parameters and inference time for these attention methods. Compared with the SE attention with the simplest structure and the lowest number of parameters, the proposed MTFC attention increases the number of parameters by 0.03M inference time by 8ms. But the performance improves by 15%. We recorded these attentional loss and accuracy curves during the training process. The loss and accuracy curves of the training set for different attention methods are shown in Fig. 8. The loss curve of the mixed time-frequency channel attention is lower and below the other attention methods, with faster and optimal convergence, as shown in Fig. 8(a). It indicates that the mixed time-frequency channel attention improves the feature extraction ability of the neural network more efficiently compared with other attention methods, which leads to better convergence of the neural network. The MTFC attention shows the best performance, consistently highest compared to other attention methods, and significantly ahead of the tf-CBAM attention, as shown in the accuracy curves in Fig. 8(b). These experimental results show that the MTFC attention, which is more effective in obtaining the dependencies between features, gains more attentional information. Thus, it also enhances the feature extraction ability of the convolutional neural network more effectively and improves the performance of the system.

Table 6. Test results of different attention methods in Voxceleb1-O

	Model	Attention	Loss	Dims	Training set	EER (%)	DCF10
Chung et al., 2018	FullRes2Net	SE	softmax	512	Voxceleb2	2.63	0.2543
Chung et al., 2018	FullRes2Net	tf-CBAM	softmax	512	Voxceleb2	2.46	0.2523
Ours	FullRes2Net	MTFC	softmax	512	Voxceleb2	2.23	0.2243

Table 7. Number of parameters and inference time for different attention methods

	Attention	Params (M)	Time (ms)
Ours	SE	2.29	76
Ours	tf-CBAM	2.31	82
Our	MTFC	2.32	84



(a) Loss curves for different attention methods

(b) Accuracy curves of different attention methods

Fig. 8. Loss curves and accuracy curves of different attention methods in the training

Next, our proposed MTFC-FullRes2Net end-to-end speaker recognition system is compared and evaluated on the Voxceleb1-O test set with these current advanced speaker recognition systems, as shown in Table 8. In previous experiments, RawNet and CNN+Transform speaker recognition systems using complex network structures showed the best performance EER of 2.48% and 2.56%. However, our MTFC-FullRes2Net end-to-end speaker recognition system outperforms the previous best results, achieving the lowest EER/DCF of 2.13%/0.2213. And it also beats most of the current state-of-the-art speaker recognition systems. It proves that the MTFC-FullRes2Net end-to-end speaker recognition system has better feature extraction ability and is a lightweight speaker recognition system suitable for practical applications.

Table 8. Test results of different systems in Voxceleb1-O

	Model (Attention)	Aggregation	Loss	Dims	Training set	EER (%)	DCF10
Rui et al., 2021 [13]	CNN+Transformer	-	-	-	Voxceleb2	2.56	-
Ahilan et al., 2019 [24]	TDNN+PLDA	SP	softmax	-	Voxceleb2	3.10	-
Chung et al., 2018 [10]	ResNet-34 (-)	TAP	Softmax+Contrastive	512	Voxceleb2	4.83	-
Chung et al., 2018 [10]	ResNet-50 (-)	TAP	Softmax+Contrastive	512	Voxceleb2	3.95	-
Etemad et al., 2019 [23]	UtterIdNet (-)	TDV	softmax	512	VoxCeleb2	4.26	-
Jung et al., 2020 [12]	RawNet2 (tf-SE)	GRU	softmax	-	Voxceleb2	2.48	-
Ge et al., 2021 [14]	Y-vector (tf-SE)	SP	AM-softmax	-	Voxceleb2	2.78	0.269
Arsha et al., 2020 [8]	ThinResNet-34 (SE)	GhostVLAD	softmax	512	Voxceleb2	2.87	-
Arsha et al., 2020 [8]	ThinResNet-50 (-)	SAP	softmax	512	Voxceleb2	4.90	0.5049
Gao et al., 2019 [20]	Res2Net (-)	SAP	softmax	512	Voxceleb2	3.11	0.2912
Ours	FullRes2Net (-)	SAP	softmax	512	Voxceleb2	2.63	0.2562
Ours	MTFC-FR (MTFC)	SAP	softmax	512	Voxceleb2	2.13	0.2213

To more comprehensively evaluate the performance of the MTFC-FullRes2Net end-to-end speaker recognition system, we test it again in the more extensive and challenging Voxceleb1-E and Voxceleb1-H test sets in this paper. In Voxceleb1-E, which uses the entire Voxceleb1 as the test set, the proposed MTFC-FullRes2Net end-to-end speaker recognition system still achieves the best results, as shown in Table 9.

Table 9. Test results of different systems in Voxceleb1-E

	Model (Attention)	Aggregation	Loss	Dims	Training set	EER (%)	DCF10
Chung et al., 2018 [10]	ResNet-50 (-)	TAP	Softmax+Contrastive	512	Voxceleb2	4.42	-
Arsha et al., 2020 [8]	ThinResNet-34 (SE)	GhostVLAD	AM-softmax	512	Voxceleb2	2.96	-
Jung et al., 2020 [12]	RawNet2 (tf-SE)	GRU	softmax	512	Voxceleb2	2.87	-
Ge et al., 2021 [24]	Y-vector (tf-SE)	SP	AM-softmax	512	Voxceleb2	2.64	0.270
Ours	ThinResNet-50 (-)	SAP	softmax	512	Voxceleb2	4.92	0.4404
Ours	Res2Net (-)	SAP	softmax	512	Voxceleb2	3.12	0.2979
Ours	FullRes2Net (-)	SAP	softmax	512	Voxceleb2	2.71	0.2732
Ours	MTFC-FR (MTFC)	SAP	softmax	512	Voxceleb2	2.21	0.2249

In the Voxceleb1-H test set using the same country and gender, the differences in accent and intonation decreased, and they were more difficult to distinguish. As a result, the EER/DCF increased for all systems. However, our proposed MTFC-FullRes2Net end-to-end speaker recognition system still holds the lead with an EER/DCF of 3.55%/0.3562 below other methods, as shown in Table 10. It demonstrates that the MTFC-FullRes2Net end-to-end speaker recognition system obtains more distinguishable frame-level features and thus can better distinguish between speakers with higher similarity.

Table 10. Test results of different systems in Voxceleb1-H

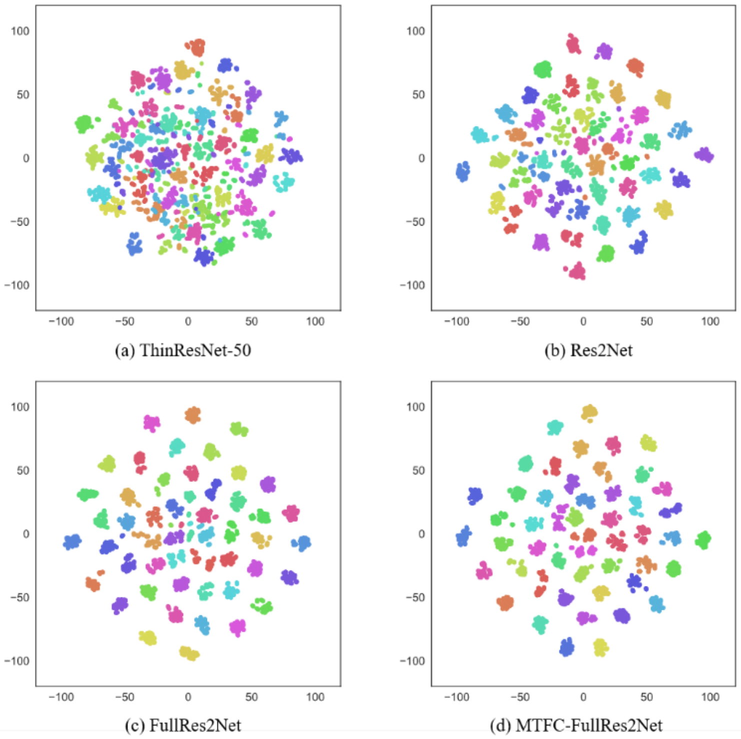
	Model (Attention)	Aggregation	loss	Dims	Training set	EER (%)	DCF10
Chung et al., 2018 [10]	ResNet-50 (-)	TAP	Softmax+Contrastive	512	Voxceleb2	7.33	-
Arsha et al., 2020 [8]	ThinResNet-34 (SE)	GhostVLAD	softmax	512	Voxceleb2	4.93	-
Jung et al., 2020 [11]	RawNet2 (tf-SE)	GRU	softmax	512	Voxceleb2	4.69	-
Ge et al., 2021 [24]	Y-vector (tf-SE)	SP	AM-softmax	512	Voxceleb2	4.33	0.377
Ours	ThinResNet-50 (-)	SAP	softmax	512	Voxceleb2	5.68	0.5504
Ours	Res2Net (-)	SAP	softmax	512	Voxceleb2	4.89	0.4404
Ours	FullRes2Net (-)	SAP	softmax	512	Voxceleb2	4.72	0.4336
Ours	MTFC-FR (MTFC)	SAP	softmax	512	Voxceleb2	3.55	0.3562

In practice, the audio is usually short as 2 to 5 seconds, and the speaker recognition system needs to identify the speaker based on the short audio. Existing speaker recognition systems do not perform well with such short audio and consume much more time to compute the distance between 10 speech combinations. So, we took a random section of 2S, 3S, and 5S of audio from the Voxceleb1-O test set and directly calculated the distance between each pair of audios to get closer to the actual conditions. The results are shown in Table 11. Compared with existing systems, MTFC-FullRes2Net is also better for short audio recognition of different lengths. It indicates that MTFC-FullRes2Net can fully exploit the information in features to obtain more global information and speaker identity information, even if the audio duration is short and contains less information. As the duration of the audio increases, the information contained in the audio becomes richer, and the aggregated utterance-level features are more discriminative. The EER/DCF value of the system is lower, but the inference time also increases.

Table 11. Test results of different systems in short-time audio

Front- model	Aggregation	Dims	Training set	EER (%)	DCF10	EER (%)	DCF10	EER (%)	DCF10
				2S		3S		5S	
ThinResNet-50 (-)	SAP	512	Voxceleb2	10.34	0.5914	6.63	0.4927	4.23	0.3976
Res2Net (-)	SAP	512	Voxceleb2	9.36	0.5841	5.01	0.4822	3.87	0.3841
FullRes2Net (-)	SAP	512	Voxceleb2	8.46	0.5042	4.47	0.4473	3.43	0.3361
MTFC-FR (MTFC)	SAP	512	Voxceleb2	7.25	0.4616	4.01	0.3927	2.63	0.2559

We used the visualization method by Kye S M et al. to visualize the effectiveness of the proposed MTFC-FullRes2Net end-to-end speaker identification system. We formed the visualization map after dimensionality reduction of speaker identity features by the t-SNE [25]. In Voxceleb1-H, fifty speakers were randomly selected to be represented by different colors. Each person randomly selects ten audios, and then ten randomly extracted three-second test segments from each audio. There are a total of 5000 three-second test segments obtained. The visualization maps of ThinResNet-50, Res2Net, FullRes2Net, and FullRes2Net are shown in Fig. 9. (a) shows the speaker feature visualization graph of ThinResNet-50. It is observed that the feature extraction ability of ThinResNet-50 is weak. The result is that the visualization graph is also very poorly classified, with many classification errors and collisions occurring. It suggests that the speaker features obtained by ThinResNet-50 are not discriminative. (b) shows the visualization of Res2Net. It is noticeable that its visualization has been significantly improved compared to (a). It has fewer classification errors and collisions. It verifies the effectiveness and robustness of Res2Net. It also shows that Res2Net has a better feature extraction ability, and the speaker features are more discriminative. However, the intra-class distance of speaker features is bigger, and the inter-class is smaller. And there are still some classification errors and collisions. (c) shows the visualization map of FullRes2Net. Comparing the visualization map of Res2Net, we can find that the collision in the visualization map of FullRes2Net is significantly improved, and almost no speaker features collide. The inter-class distance increases significantly. It also proves that FullRes2Net effectively improves the feature extraction ability by acquiring larger receptive fields and generating more combinations of receptive fields. But there are still a few cases of classification errors and large intra-class distances in FullRes2Net. (d) represents the visualization map of MTFC-FullRes2Net. It not only has no classification errors but also has a larger inter-class distance and closer intra-class distance. It indicates that our proposed mixed time-frequency channel attention is efficient. It effectively improves the feature extraction ability of FullRes2Net to get more distinguishable features. And it makes the speaker features from the same speaker have a higher similarity



(a) ThinResNet-50

(b) Res2Net

(c) FullRes2Net

(d) MTFC-FullRes2Net

Fig. 9. Reduced dimensional view of t-SNE with different network structures

6 Conclusion

The MTFC-FullRes2Net network exploits the information in features more fully through the FullRes2Net network structure. And it introduces the MTFC attention to improve the disadvantages of convolution, interactively integrate the features in the time-frequency channel dimension, and obtain the dependencies between features, to generate more effective global features.

We construct a network model with better feature extraction capability and better adaptation to the speaker recognition task. We achieve the lowest EER and DCF on the noisy Voxceleb1 test set, outperforming more than many existing methods and effectively improving the accuracy of the end-to-end speaker recognition system. MTFC-FullRes2Net end-to-end speaker recognition system also achieved the best performance than all systems tested in the more realistic and challenging short-time audio speaker recognition. Our proposed MTFC-FullRes2Net end-to-end speaker recognition system only increases the number of parameters and inference time very little compared to the Res2Net, but effectively improves the system performance, which is a more efficient and lightweight structure. It is also more suitable for practical application.

The work we have done has mainly improved the feature extraction ability of the speaker recognition system. However, we have not researched the process of feature aggregation. How to aggregate frame-level features more effectively to get more differentiated utterance-level features is also a key factor affecting performance. Therefore, in future work, we will focus on how to aggregate features more effectively.

7 Acknowledgement

This work is supported by National Natural Science Foundation of China, China Grant No. 61972324, Sichuan Science and Technology Program, China Grant No. 2021YFS0313 and 2021YFG0133.

References

[1] J. Hansen, T. Hasan, Speaker Recognition by Machines and Humans: A tutorial review, IEEE Signal Processing Magazine 32(6)(2015) 74-99.

[2] L. L. Stoll, Finding Difficult Speakers in Automatic Speaker Recognition, University of Califo-Rnia, Berkeley, Technical Report No. UCB/EECS-2011-152, December 16, 2011.

[3] P. Matejka, O. Glembek, F. Castaldo, M.J. Alam, O. Plchot, P. Kenny, L. Burget, J. Černocky, Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification, in: Proc. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011.

[4] S. Cumani, O. Plchot, P. Laface, Probabilistic Linear Discriminant Analysis of i-Vector Posterior Distributions, in: Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[5] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[6] M. Mclaren, L. Ferrer, D. Castan, A. Lawson, The Speakers in the Wild (SITW) Speaker Recognition Database, in: Proc. 2016 Interspeech, 17th Annual Conference of the International Speech Communication Association, 2016.

[7] A. Nagrani, J.S. Chung, A. Zisserman, VoxCeleb: a large-scale speaker identification dataset, in: Proc. 2017 Interspeech, 18th Annual Conference of the International Speech Communication Association, 2017.

[8] A. Nagrani, J.S. Chung, W. Xie, A. Zisserman, Voxceleb: Large-scale speaker verification in the wild, Computer Speech & Language 60(2020) 101027.

[9] W. Xie, A. Nagrani, J.S. Chung, A. Zisserman, Utterance-level Aggregation for Speaker Recognition in the Wild, in: Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[10] J.S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep Speaker Recognition, in: Proc. 2018 Interspeech, 19th Annual Conference of the International Speech Communication Association, 2018.

[11] H. Zeinali, S. Wang, A. Silnova, P. Matějka, O. Plchot, BUT system description to VoxCeleb speaker recognition challenge 2019. <https://arXiv.org/abs/1910.12592>, 2019 (accessed 16.10.19).

[12] J.-W. Jung, S.-B. Kim, H.-J. Shim, H.-J. Yu, Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms, in: Proc. 2020 Interspeech, 21st Annual Conference of the International Speech Communication Association, 2020.

[13] R. Wan, J. Ao, L. Zhou, S. Liu, Z. Wei, T. Ko, Q. Li, Y. Zhang, Multi-View Self-Attention Based Transformer for Speaker Recognition. <https://arXiv.org/abs/2110.05036>, 2021 (accessed 11.10.21).

[14] G. Zhu, F. Jiang, Z. Duan, Y-Vector: Multiscale Waveform Encoder for Speaker Embedding, in: Proc. 2021 Interspeech,

22nd Annual Conference of the International Speech Communication Association, 2021.

[15] J. Hu, S. Li, G. Sun, Squeeze-and-excitation networks, in: Proc. 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[16] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in Proc. 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[17] J. Zhou, T. Jiang, Z. Li, L. Li, Q. Hong, Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function, in: Proc. 2019 Interspeech, 20th Annual Conference of the International Speech Communication Association, 2019.

[18] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proc. 2018 European Conference on Computer Vision, 2018.

[19] S. Yadav, A. Rai, Frequency and temporal convolutional attention for text-independent speaker recognition, in Proc. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

[20] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2Net: A New Multi-Scale Backbone Architecture, IEEE Transactions on Pattern Analysis and Machine Intelligence 43(2)(2021) 652-662.

[21] H. Wang, Y. Fan, Z. Wang, L. Jiao, B. Schiele, Parameter-Free Spatial Attention Network for Person Re-Identification. <https://arXiv.org/abs/1811.12150>, 2018 (accessed 29.11.18).

[22] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks, in: Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[23] A. Hajavi, E. Ali, A Deep Neural Network for Short-Segment Speaker Recognition. <https://arXiv.org/abs/1907.10420>, 2019 (accessed 22.07.19).

[24] A. Kanagasundaram, S. Sridharan, G. Sriram, S. Prachi, C. Fookes, A study of x-vector based speaker recognition on short utterances, in: Proc. 2019 Interspeech, 20th Annual Conference of the International Speech Communication Association, 2019.

[25] L. Maaten, G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research 9(2008) 2579-2605.