Speaker Recognition Algorithm Based on Fca-Res2Net

Zhangfang Hu, Caiyun Lv, Changbo Wu

Abstract—**Traditional recognition methods often lead to problems such as speaker information loss and reduced recognition rates. To address these problems, an Fca-Res2Net speaker recognition model incorporating a self-attentive mechanism is proposed in this paper. First, the model uses the modified mel-frequency cepstral coefficients (MFCCs) as the system feature input and combines the inverse mel-frequency cepstral coefficients (IMFCCs) with the MFCCs as the base input features to extract more representative speech spectral features. On this basis, the difference parameters △MFCC and △IMFCC are fused to fully extract the speech dynamic and static features in the high- and low-frequency bands. Second, frequency channel attention networks (FcaNets) are introduced on top of the baseline model (Res2Net: a new multiscale backbone architecture), and the residual module is used to fuse the shallow and deep speaker features to better obtain the different feature channel weights without increasing the number of parameters. In addition, to better introduce temporal information and capture long-span speech features, the self-attention mechanism is integrated to enhance the long-span modelling of speech features. Finally, the classification output results are identified. Experimental results show that the proposed model improves the recognition rate and robustness of speakers in long speech when compared with the current mainstream speaker recognition methods in the VoxCeleb dataset with sufficient data volume.**

Index Terms—**Speaker recognition, self-attention mechanism, Fca-Res2Net, MFCC**

I. Introduction

Speaker recognition, also known as voiceprint recognition, is the process of identifying the person who is speaking by their voice characteristics. This process depends heavily on the speaker who is making the sound. Much of the variation between speakers stems from differences in speaking style, vocal cords and voice form, as well as differences in speaker delivery when conveying a particular meaning [1]. These features are used by the most advanced speaker recognition systems, primarily in real-life security systems. Compared to other biometric features, such as fingerprints, face and iris identification, and DNA, voice capture is more convenient and cost-controllable, and the security of the speaker's privacy is stronger. These advantages are used in fields such as credit card voice protection, telephone banking customer verification, and public security forensics. With the increasing demand for various applications and the maturity of technologies related to speaker recognition, improving system recognition accuracy has become a popular topic for major research this year [2].

The basic speaker recognition system framework consifeature extraction and speaker model building. Feature extraction is the extraction of the speaker's speech signal feature vector as the input to the speaker model so that it can fully reflect the individual differences of the speaker and improve the recognition accuracy. Common time-domain features in speaker features include amplitude, energy, average overzero rate, etc. [3]. However, these features are usually obtained directly from the original speech signal through filters to obtain feature vectors, which are simple to process but have poor stability and poor ability to characterise the speaker's identity and are therefore rarely used in recent years [4]. Common transform domain features include the linear prediction coefficient (LPC), filter bank (Fbank) features based on filter banks, and mel frequency cepstral coefficients (MFCCs). Since the Meier filter is a filter based on the structure of the human ear, it can better fit the characteristics of signals received by the human ear and fully reflect the characteristics of the speaker [5], so the MFCC feature extraction method is mostly used in speaker recognition systems.

With the development of deep learning, speaker recognition methods based on deep learning are becoming increasingly popular due to their excellent modelling capabilities [6]. Deep learning is used to learn more useful features by building models with hidden layers and training large amounts of parametric data to improve the accuracy of classification or recognition [7]. Recent research has shown that using shallow networks to build network models makes the extracted features weak and the model recognition rate insufficient, so efforts have been made to use deeper networks to extract the feature parameters; however, deepening the number of network layers can cause the gradient explosion or gradient disappearance. Based on this, Gao et al. [8] proposed Res2Net using deep residual networks, i.e., a network struc-

Manuscript received March 30, 2023; revised September 3, 2023. This work was supported in part by the National Natural Science Foundation Youth Fund Project (Grant No. 61703067), the Chongqing Basic Science and Frontier Technology Research Project (Grant No. Cstc2017jcyjAX0212), and the Chongqing Municipal Education Commission Science and Technology Research Project (KJ1704072).

Zhangfang Hu is a Professor at the Key Laboratory of Optical Information Sensing and Technology, School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 3565207151@qq.com)

Caiyun Lv is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (corresponding author phone: 177-823-32706; e-mail: 1583519903@qq.com)

Changbo Wu is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 2982954515@qq.com)

ture with hierarchical connections of residual blocks to better fuse shallow and deep layers and speaker feature information from different channels to obtain more powerful feature extraction without increasing the amount of parameter computation.

In previous deep learning networks for speaker recognition, different features and channels were given the same weight, and the resulting speaker recognition models were unable to focus on the most important vocal features; therefore, the introduction of attention mechanisms into speaker recognition is gradually becoming a popular topic [9]. Channel attention is a mechanism that emphasises the regions of interest while suppressing irrelevant background regions through different channels learning to value weights directly. The traditional channel attention mechanism SE-Net (squeeze and excitation net) uses scalars to represent channels and uses global average pooling (GAP) to retain only the lowest frequency information. All components from other frequencies are discarded to better compress the channel and allow more information to be introduced. ZU scholars proposed a new attention mechanism method, a frequency-domain channel attention network called FcaNet [10].

Based on the analysis of the above problem, the attention mechanism network FcaNet using the lightweight network Res2Net as the baseline network is introduced in this paper. An improved speaker recognition model based on FcaNet-Res2Net, which performs better in speaker recognition systems independent of text and long speech aspect recognition performance, is proposed. The main contributions of the article are as follows: Section 2 gives an improved MFCC feature extraction method proposed in this paper. This method combines the inverse Merle cepstral coefficients IMFCC with MFCC and designs a new speaker recognition network combining frequency domain channel attention mechanisms, based on which a self-attention pooling mechanism is used to capture temporal information and assign corresponding weights to different channel information and segment-level features. Section 3 shows the experimental results of training and testing this paper's method on the VoxCeleb dataset with traditional speaker recognition networks and analyses the experimental results, and Section 4 concludes the work of this paper.

II. Proposed Method

Speaker recognition is divided into three stages: training, registration and testing. In the training stage, an optimised recognition model is obtained, the voice fragments of the registrants are extracted by the trained recognition model to save the vocal features in the vocal database, and finally, the voice features of the subjects are extracted and compared with the features of the vocal database. A recognition threshold is set, and those above this threshold are considered the same speaker; otherwise, they are considered different speakers. The specific model building process is shown in Fig. 1.

The Fca-Res2Net-based speaker recognition method presented in this paper starts with both feature extraction and speaker recognition models. To better build a registered vocal database, the MFCC parameters are improved by using feature fusion to obtain a set of feature parameters that can characterise different speaker identities and enhance the expressiveness of speaker features. Then, a FcaNet-Res2Net speaker recognition model incorporating a self-attentive mechanism is proposed, and a new frequency domain channel attention and self-attentive pooling mechanism are used to enhance the recognition capability of the model. Softmax is used to output the output results with different classification probabilities, and the model performance is observed through softmax-loss. The parameters are updated according to the size of the loss function with a reverse gradient, and the optimised recognition model is obtained as shown in Fig. 1.

A. *Speech Network*

Speaker feature extraction is the process of extracting information that characterises the identity of a speaker from the original speech fragment. It is a key part of a speaker recognition system, the framework of which is shown in Fig. 2. Usually, speaker feature extraction requires preprocessing and then different feature extraction methods to obtain different acoustic features. While single feature information is insufficient for fully characterising the speaker identity in-

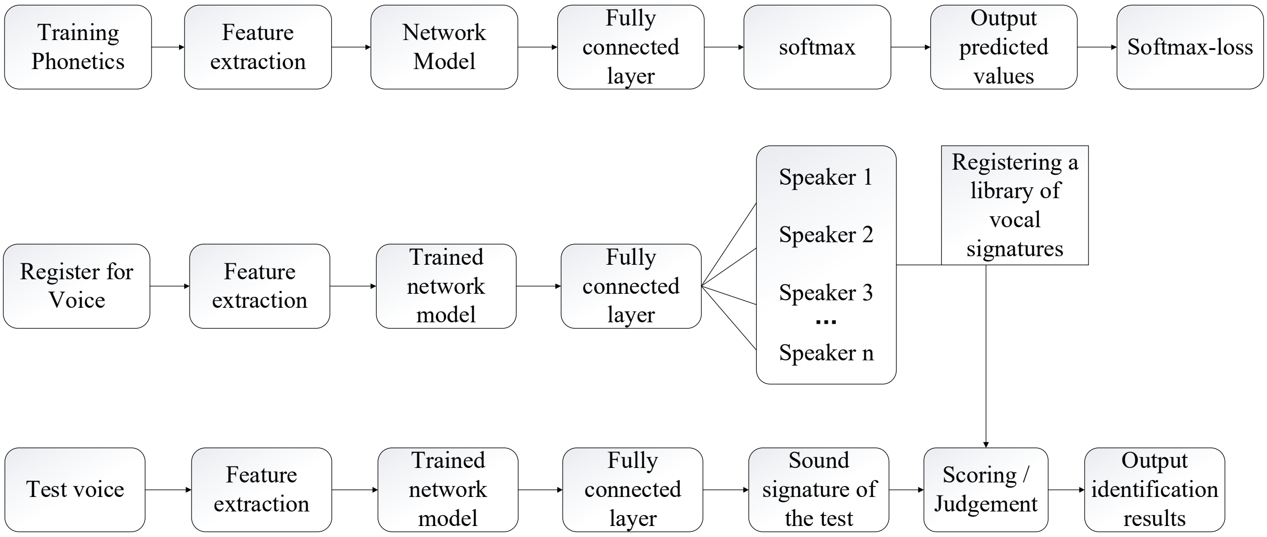

Fig. 1. Architecture of the Basic flow chart for speaker identification

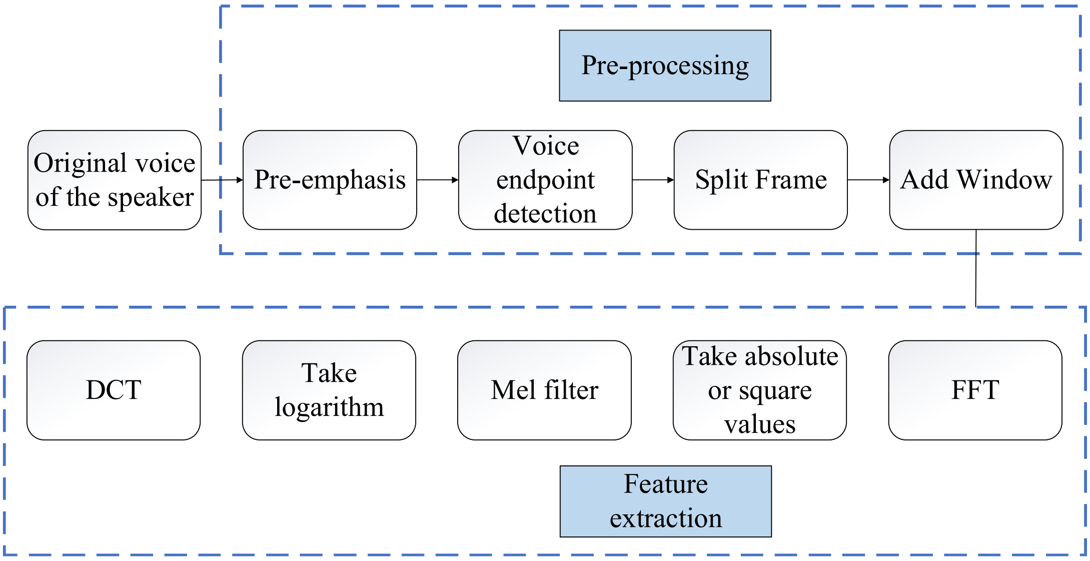

Fig. 2. Architecture of the Speaker feature extraction process

formation, based on the above problems, the fusion feature parameters proposed in this paper exploit the complementary information between different features to improve the performance of speaker recognition [11].

1) Preprocessing

Speech preprocessing includes four steps, namely, pre-emphasis, speech break detection, framing and windowing. It is an important part of feature extraction. In practice, the audio processed is often accompanied by a small amount of noise, pauses and gaps due to the acquisition equipment and the acquisition site, and very clean speech cannot be captured [12]. Therefore, this design requires preprocessing the original audio to eliminate blank frames and background noise to a certain extent before extracting the feature parameters to accurately capture the vocal features in the speech signal.

Preemphasis: By strengthening the high-frequency part of the speech signal, the spectrum of the signal becomes flat and remains in the entire frequency band. This enables the same noise ratio to be used to find the spectrum.

Speech endpoint detection: The raw speech data are screened in the preprocessing stage, which first identifies speech segments and nonspeech segments from the raw audio signal and then screens silent audio segments from them to improve key information extraction for speaker recognition [13].

Framing: Speech signals are short-time smooth signals and therefore cannot be processed for long speech segments, so the speech signal needs to be cut into small segments: frames. To avoid too much variation between two adjacent frames, part of the region between two adjacent frames overlaps to reduce the loss of speech data.

Adding windows: Speech is constantly changing within long ranges of speech and cannot be processed without fixed characteristics, so a windowing operation is required to eliminate the signal discontinuities that may be caused at the ends of individual frames [14].

2) Feature Extraction

Acoustic features such as the MFCC, IMFCC and first-order differential MFCC are obtained after preprocessing by specific algorithms. The basic process is shown in Fig. 3. The preprocessed speech signal is converted into an energy distribution in the frequency domain by fast Fourier transform (FFT) for each frame, and the power spectrum of the speech signal is obtained by taking the square of the frequency spectrum of the speech signal, followed by passing the power spectrum through the Meier filter bank and the inverse Meier filter bank. Finally, the MFCC and IMFCC parameters are obtained by taking the logarithm of the signal and performing the discrete cosine transform (DCT). The difference between the standard parameters is then taken to obtain the dynamic characteristics of the speech signal [15].

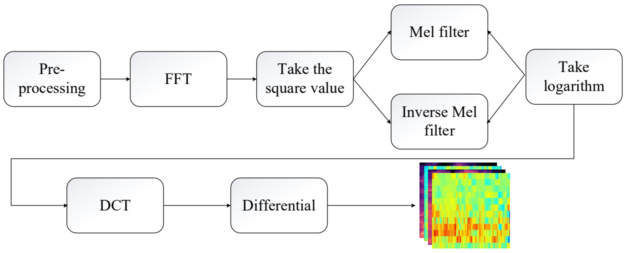

Fig. 3. Architecture of the MFCC, IMFCC fusion feature extraction map

B. Speaker Recognition Model Construction

The shallow convolutional neural network in speaker recognition systems results in poor feature recognition. Deepening the neural network is an effective method to improve feature extraction, but it has been found that many methods use deeper and more complex network structures, increasing the parameters and training volume and even causing problems such as network gradient disappearance and gradient explosion [16]. In this paper, Res2Net, a network applied to target detection tasks, is therefore introduced to the speaker recognition system with Res2Net-50 as the mainline network. Compared with ResNet, Res2Net has a larger perceptual field with almost no increase in the number of parameters [17]. The overall framework of the speaker recognition system is shown in Fig. 5.

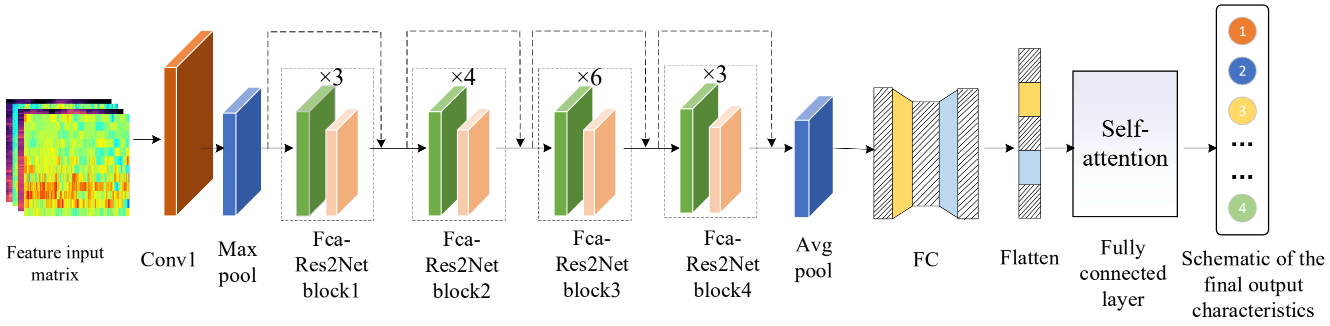

Fig. 4. Architecture of the speaker model framework

In deep learning networks, different features and channels are given the same weight; however, the contribution of each channel feature component to recognition is different. The traditional SE-Net loses all frequency components except the lowest frequency component due to the use of GAP, so to better obtain more feature information, a new attention mechanism method is introduced in this paper: the frequency-domain channel attention network FcaNet is introduced into the recognition mainline network Res2Net-50 to form the new speaker recognition model FcaNet-Res2Net, whose basic framework is shown in Fig. 6. Whereas a single CNN potentially faces the problem of capturing dependencies on long-span expressions within discourse due to its overfocus on local and neighbouring features and cannot fully capture long-span speech features, a self-attention mechanism is integrated in this paper to enhance long-span modelling of speech features [18]. The overall framework of the speaker recognition system is shown in Fig. 4.

TABLE I
NETWORK PARAMETERS OF THE PROPOSED FCA-RES2NET

Layer	Kernel	Stride	Output
Convl	7×7, 16	2	150×128×16
Max Pooling	3×3	2	75×64×16
Res2block1	[3, 3×3, 4] ×3	1	38×32×32
Res2block2	[3, 3×3, 16] ×4	1	19×16×256
Res2block3	[3, 3×3, 32] ×6	1	10×8×512
Res2block4	[3, 3×3, 64] ×3	1	5×4×1024
Avg Pooling	-	-	1×1×1024
FC	-	-	1488

In this paper, the feature size is adjusted to 300×256×4 for input into Fca-Res2Net, and the parameters of each network are shown in Table 1.

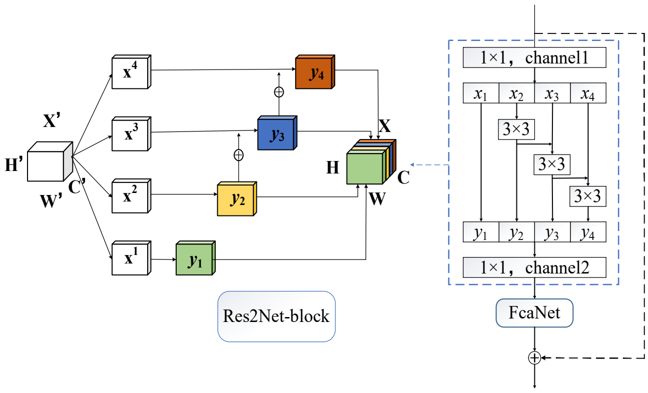

Fig. 5. Architecture of the Fca-Res2Net block

1) Res2Net Network

The Res2Net network was first applied to target detection and has been used in speaker recognition systems in recent years due to its powerful feature extraction capability. The layered parallel network structure also greatly increases the receptive domain of the model by fusing speaker information under different layering across channels [7]. Fig. 7 illustrates a comparison of the Res2Net residual block with the ResNet block.

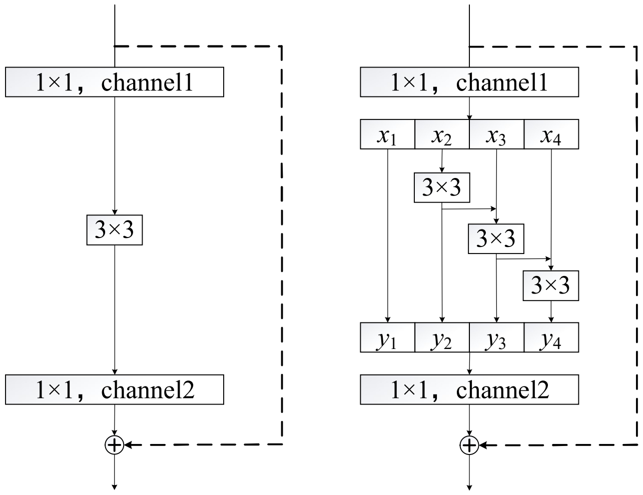

Fig. 6 Architecture of ResNet and Res2Net

Res2Net constructs a channeled residual-like connection based on the residual network block and introduces a new

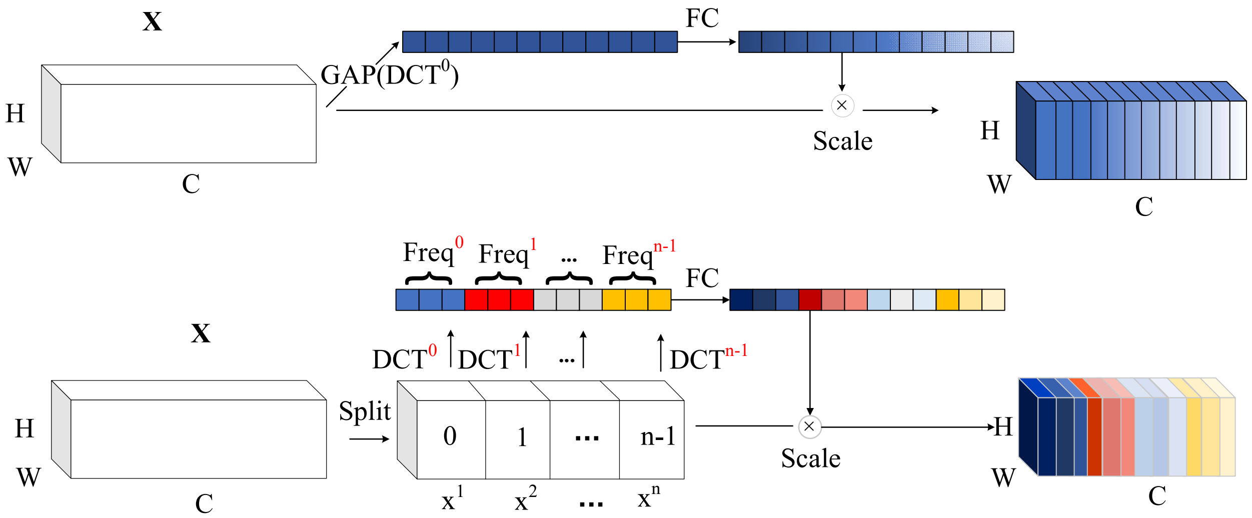

Fig. 7. Architecture of the SENet block and FcaNet block

dimension s (scale) into the convolutional neural network, representing the number of feature partitions mapped in the residual block by uniformly splitting by channel. Unlike the 3×3 convolution common to traditional residual networks, Res2Net, after 1×1 convolution, cuts the input features evenly into s subsets by the number of channels x_i, of which $i \in \{1,2,3...,s\}$, $x \in R^{H\times W\times C}$, $x_i \in R^{H\times W\times C'}$, C is the number of input Res2Net channels, and C' is the average number of channels per subset after segmentation. Except for x_1, each subset is individually convolved with a set of 3×3, starting from $i=3$, x_i is added to the previous subset of convolutional outputs $K_{i-1}(.)$ which is convolved into $K_i(.)$, and the output y_i is expressed as shown in (1).

$$y_i = \begin{cases} x_i, & i=1 \\ K_i(x_i), & i=2 \\ K_i(x_i+y_{i-1}), & 2<i\le s \end{cases} \qquad (1)$$

When $K_i(.)$ receives the output of the previous feature subset y_{i-1}, the previous feature subset corresponds to a succession of different convolution operations of $K_i(.)$ and K_{i-1}. This design of connecting channels in groups within the residual blocks increases the perceptual field of each layer of convolution, allowing the model to increase the receptive field of each layer of the network and extract finer-grained multiscale features. Finally, after a 1×1 convolution operation, all outputs y_i are stitched together as the total output and fed into the next convolution layer.

2) Fca-Net Network

In this paper, we need to obtain high-quality speech feature information, so we add the attention module Fca-block to the backbone network Res2Net-50 and connect it to Res2Net-block, which is used to reduce the weight of low-quality speech information and suppress the information in the input features that is irrelevant to the speaker feature extraction. Fca-block is a new attention mechanism based on the squeeze and excitation block (SE-block) [10] based attention mechanism [9], which was first applied to target detection. Unlike the SE-block, which uses global average pooled GAP to compress the feature map [19], Fca-block uses two-dimensional discrete cosine variation (2D-DCT) to compress the feature map, preserving other frequency components. Using only GAP would result in the feature channels retaining only the lowest frequency component and losing the other frequency components. An overall illustration of the comparison between SE-Net and Fca-Net is shown in Fig. 8.

Suppose the input is X. First, X is divided into n parts

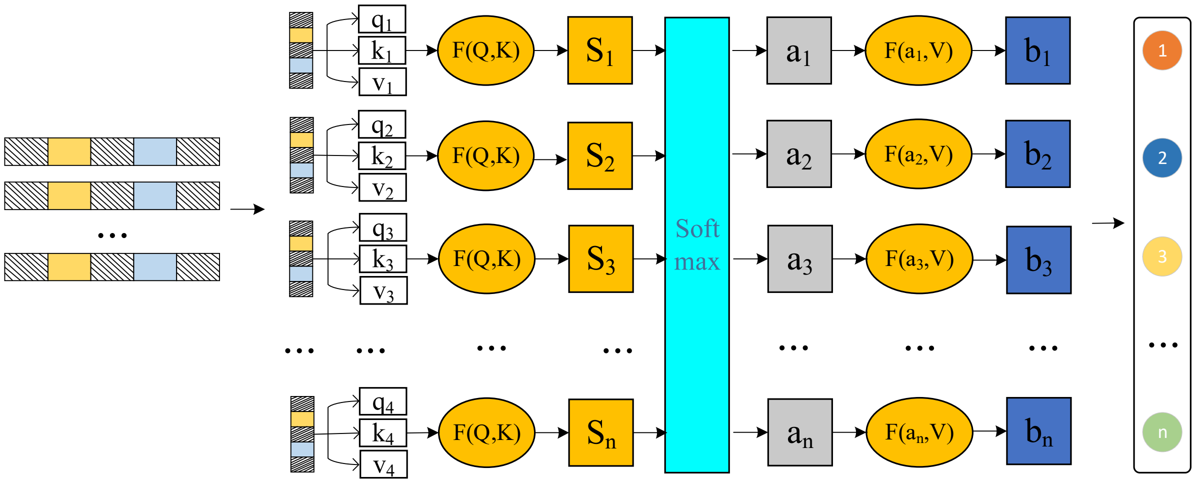

Fig. 8. Architecture of the self-attention framework

along the channel dimension: $\left[X^{0},X^{1},X^{2},...,X^{n-1}\right]$, of which $i\in\left\{0,1,2,...,n-1\right\}$, $X^{i}\in R^{H\times W\times C^{\cdot}}$, $C^{\cdot}=\frac{C}{n}$. For each component, the corresponding 2D-DCT frequency component is calculated and assigned, and the result of the 2D-DCT is used as the result of the compression noted by the channel, 2D-DCT, as shown in (2).

$$Freq^{i}=2DDCT^{u_i,v_i}(X^{i})=\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}X^{i}_{:,h,w}B^{u_i,v_i}_{h,w} \qquad (2)$$

$$B^{u_i,v_i}_{h,w}=\cos(\frac{\pi h}{H}(u_i+\frac{1}{2}))\cos(\frac{\pi w}{W}(v_i+\frac{1}{2})) \qquad (3)$$

As shown in (3), $B^{u_i,v_i}_{h,w}$ is the basis function for 2DDCT, H is the height of X^{i}, W is the width of X^{i}, and (u_i,v_i) is the 2D index of X^{i}. The entire feature information compression vector can be represented by a cascade as follows in (4), where Freq is the obtained multispectral vector.

$$Freq=compress(X)=cat([Freq^{0},Freq^{1},..,Freq^{n-1}]) \qquad (4)$$

The entire Fca-Net framework can be represented as shown in (5).

$$.\ ms\_att=sigmoid(fc(Freq)) \qquad (5)$$

3) Self-attention Network

After the CNN network, the output feature information contains only the spatial information of the speaker and ignores the temporal information. The CNN is difficult to model for longer speech sequences and cannot see all the time sequences at once. The self-attention mechanism can solve the problem arising when the output of the network is multiple identical vectors and there are certain relationships between different vectors. These relationships cannot be fully exploited during training, making the network prediction results poor. CNNs have a better advantage: the input can be segmented and fed into the CNN in parallel, which creates multiple output channels, straightening and coding multiple outputs and feeding them into the self-attention block at the same time. This will result in a long time series of speech features related to the whole time series, decoded as speaker feature information [20] and giving full play to the correlation between different features. The self-attention model framework is shown in Fig. 9.

In this paper, we use the features extracted from Fca-Res2Net as input, obtain a series of input vectors through the encoder, and make the inner product of the input vectors with the matrices w_q, w_k and w_v to obtain q_i, k_i and v_i of each input, respectively, and the inner product of q_i and k_i between two inputs $F\left(Q,K\right)$ to obtain the result: S_n, and the inner product result through the softmax function to obtain the similarity matrix a_n between two inputs, and the inner product of a_n and v_i to add $F\left(a_n,V\right)$, to obtain the output sequence of each input: b_n. Each input of q_i, k_i and v_i is combined to form the input matrix values Q, K and V. The output matrix is then calculated as shown in (6).

$$Attention(Q,K,V)=soft\max(\frac{QK^{T}}{\sqrt{d_k}})V \qquad (6)$$

where d_k is the equal dimension of q_i, k_i and Attention is the output attention score [21]. The output matrix is passed through the decoder to obtain the final output feature values. After self-attention, the output features are the weighted sum of all samples in the time series, which can see all the input time samples, according to the different weights focused on the speaker's own time attention. Finally, the result is sent to the fully connected layer after the softmax function to obtain the speaker recognition results. The cross-entropy error receives the output of softmax and the speaker real label from these data output loss L, and according to the size of the loss function, backpropagation is used to adjust each network weight parameter to train a good speaker recognition model for testing and identify multiple speakers.

III. Experimental

A. Dataset

The datasets used in this paper are the largest independent of text "natural" speaker recognition dataset available, the VoxCeleb dataset. The VoxCeleb1 and VoxCeleb2 datasets, which are large-scale open-source audio and video datasets released by the University of Oxford in 2017 and 2018, respectively [22].

The dataset is a collection of real scenes in a natural environment, with audio and video taken from the YouTube website. It consists of completely realistic English speech, including rich scenes such as celebrity red carpet, reality show interviews, and large stadium commentary. The dataset is text-independent; the speaker range is wide, with a variety of races, accents, occupations and ages; the dataset has a balanced distribution of male and female gender. The speech has additional noise, including ambient sudden noise, background vocals, laughter, speech overlap, echoes, room noise, recording equipment noise, etc.

VoxCeleb1 contains over 100,000 utterances spoken to 1251 celebrities, with a total speech duration of 351 hr, 1,211 speakers in the training set, 40 speakers in the test set, 21,245 videos, 145,265 audio files, 45 to 250 utterances per speaker, and the length of each utterance ranging from 4 s to 145 s. The audio of the VoxCeleb2 dataset contains over 1 million voices from 6,112 different speakers. The two datasets do not overlap, i.e., they do not contain the same speakers. The experiments in this paper use the VoxCeleb1 training set for training, the VoxCeleb2 dataset for validation and finally the VoxCeleb1 test set for model testing scenes.

B. Experimental Data

In this paper, the VoxCeleb dataset was used. A 30 s speech segment of each speaker in the VoxCeleb1 training set was taken, averaged into 10 segments and simultaneously input into the speaker recognition system and they are then input into the training phase.]. To ensure the uniqueness of the prediction data and the accuracy of the recognition model, all the samples used in the training phase were trained with a single speaker. In the training phase, each speaker's speech was extracted from the dataset using a Hamming window with a width of 30 ms, a step size of 10 ms and an interval overlap of 20 ms. The feature extraction schematic is shown in Fig. 9. From the 3 s speech segment sample, 300 speech frames can be obtained. Using a 256-dimensional filter set, each frame can be obtained after feature extraction to obtain 256 MFCC features, 256 IMFCC features, 256 △MFCC features, and 256 △IMFCC features. The combination of the motion and static features of the obtained high- and low-frequency parts can obtain a 4-channel 300×256 speech feature spectrum matrix, which is fed into the speaker recognition network as an input feature map to effectively improve the recognition performance of the system.

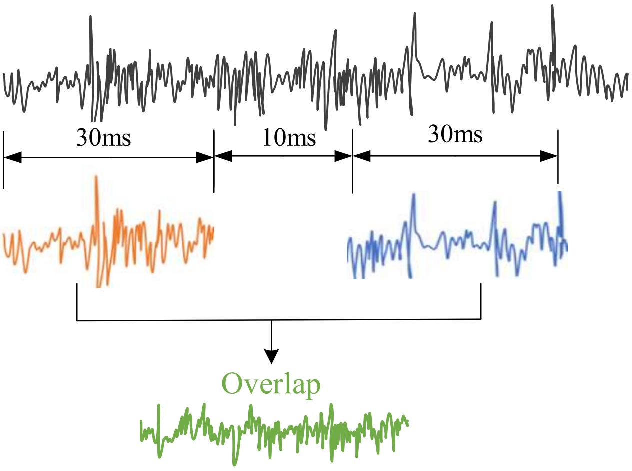

Fig. 9. Architecture of frame feature extraction

The network was trained on the VoxCeleb1 dataset. To improve the generalisability of the network and demonstrate the higher accuracy of the new network, a larger dataset, VoxCeleb2, was used for model training. The network was finally tested with the VoxCeleb1 test set, with no overlap between the datasets used in each part.

C. Experimental Setup

The experiments in this paper are divided into four parts. First, the number of model iterations is trained and updated to find the optimal network parameters. Second, the number of layers of the residual network model is tested, and the optimal network layer model is used as the backbone network for subsequent experiments. The ablation experiment is set up as follows: the model in this paper is compared using a single feature extraction method with the complete feature fusion method, and the experimental effect of a single portion of the model is compared with the experimental effect of the complete model. The experimental results of the single portion are compared with those of the full model. Finally, to verify the advantages of the proposed model, the speaker recognition model is compared with the GMM-UMM and i-vector+PLDA models of traditional speaker recognition algorithms, the network models of popular deep learning algorithms and embedding-based fusion models, and the classical convolutional network ResNet+self-attention. The performance of the self-attention1 module is examined, and experiments are conducted using different speech fragments to verify the effectiveness of the self-attention module for feature extraction of long speech fragments.

D. Experimental Evaluation index

In this paper, Res2Net is used as the backbone network, and the final recognition results are scored and judged by the traditional method, cosine distance. The recognition performance of the Fca-Res2Net speaker recognition model proposed in this paper is experimentally compared to that of other models. Both the equal error rate (EER) and tandem detection cost function (t-DCF) are used as evaluation metrics.

1) Similarity Scoring

After training the vocal recognition model, the vocal feature library registered for the test speech is scored for similarity matching with the target speaker's speech feature vector, with a higher score corresponding to a higher similarity. The final score is then compared with a threshold value set based on a priori knowledge to determine whether the test speech and the target speaker are the same speaker [23]. The speaker system sets up a similarity matching module as a scoring system, which usually uses cosine similarity scoring [24], based on the magnitude of the angle between the two vectors; the larger the angle, the more the cosine value deviates from 1 and the lower the similarity, and vice versa. The two model acoustic feature vectors are $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$. The angle of clamping is θ, and then the cosine similarity between them is calculated as shown in (7).

$$\cos\theta = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i==1}^{n} x_i^2}\sqrt{\sum_{i==1}^{n} y_i^2}} \tag{7}$$

2) Evaluation Metrics

The false rejected ratio (FRR) and the false accepted ratio (FAR) are two key metrics often used in speaker recognition to measure the performance of detection models. Since both vary with the adjustment of the threshold value and cannot accurately evaluate the system performance, most researchers now use EER [25] to make an evaluation metric for the system. EER is introduced when FAR = FRR, which allows the false acceptance rate to be balanced with the false rejection rate. The calculation equations are shown in (8) and (9), respectively.

$$FRR = \frac{Number_{wrong\_reject}}{Number_{all}} \tag{8}$$

$$FAR = \frac{Number_{wrong\_accept}}{Number_{all}} \tag{9}$$

The t-DCF [26] is the case where the FAR and FRR are given different weights according to their different impacts on system performance under a given target prior probability, with the aim of being able to ensure that both error rates are reduced to relatively low levels at the same time. The calculation formula of t-DCF is shown in (10).

$$DCF = C_{fr} \times FRR \times P_{tar} + C_{fa} \times FAR \times (1 - P_{tar}) \tag{10}$$

P_{tar} represents the prior probability of the target speaker, C_{fr} represents the weight of the FRR, C_{fa} represents the weight of the FAR, and the minimum value of the DCF, t-DCF, is found by varying the threshold.

E. Experimental Results and Analysis

1) Experiments on Training Parameters

To explore the performance advantages and disadvantages between different models and identify the best training parameters and the best number of iterations, experiments were conducted on different training times in the training sets VoxCeleb1 and VoxCeleb2. the experimental results are shown below in Fig. 10 and Fig. 11.

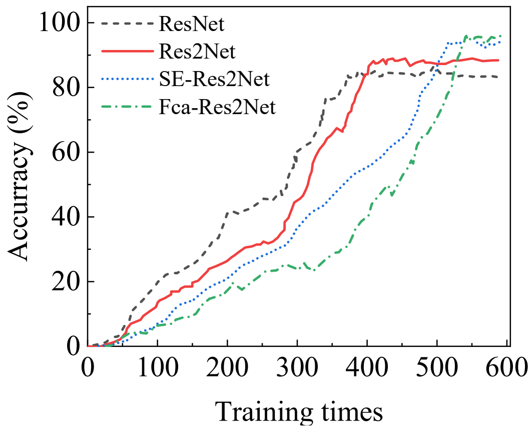

Fig. 10. Iterative parameter training graph: VoxCeleb1

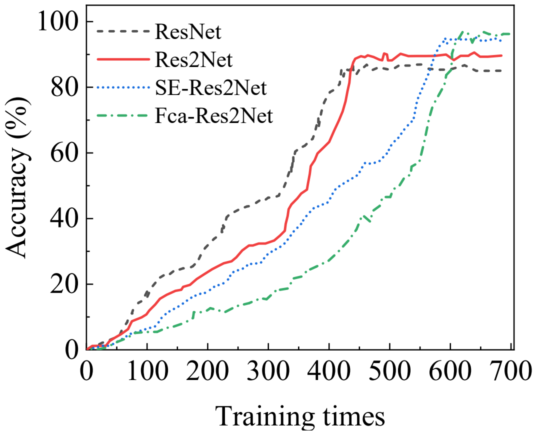

Fig. 11. Iterative parameter training graph: VoxCeleb2

The experimental results show that in the VoxCeleb1 dataset, Model 1 and Model 2 achieve the best recognition accuracy at approximately 380 and 400 times, respectively. Models 3 and 4 stabilise after training at approximately 510 and 580 times, respectively, and obtain the best recognition results at these points. In the dataset VoxCeleb2, Models 1 and 2 achieved the best recognition accuracy by approximately 420 and 450 times, respectively, and Models 3 and 4 stabilised after training by approximately 580 and 640 times, respectively, achieving the best recognition results at these points. This is because the scale of the datasets and the training samples differ. While the network structures of Models 1 and 2 are simple, Model 2 has more parameters than Model 1, and the model is more complex. However, the number of iterations for both models does not change much, which can also be seen from the experimental results. Models 3 and 4 add the attention module to Models 1 and 2, which increases the number of training iterations, slows down the convergence speed and increases the time consumption, thus achieving better recognition results.

2) Network Layer Performance Experiments

To explore the effect of different layers of the model, e.g., the depth of the model, on the performance of the model, experiments using Res2Net-18, Res2Net-34 and Res2Net-50 were conducted in the datasets VoxCeleb1 and VoxCeleb2. The experimental results are shown in Table 2.

The experimental results show that as the number of network layers increases, Res2Net-50 outperforms other residual network layers and improves the speaker recognition performance to a certain extent.

TABLE II
EER FOR RES2NET WITH DIFFERENT LAYERS

Model	VoxCeleb1		VoxCeleb2	
	EER/%	t-DCF	EER/%	t-DCF
Res2Net-18	2.987	0.209	1.743	0.203
Res2Net-34	2.779	0.196	1.675	0.187
Res2Net-50	2.298	0.146	1.298	0.127
Res2Net-101	2.978	0.198	1.729	0.194

3) Network Module Ablation Experiment

The VoxCeleb1 dataset was used to train the speaker recognition models, and the recognition performance of each network model was tested on the VoxCeleb1 test set. The performance of each model was tested when the training data were sufficient, and the results are shown in Table 3. The training data and number of parameters were basically the same for all models.

The experimental results in Table 3 show that the Fca-Res2Net model proposed in this paper has obvious advantages. After cascading the Fca-block module, the model can focus more effectively on the channel frequency information that best reflects the acoustic features during the training process and has 18%, 13% and 10% performance gains, respectively, indicating that the model improves the recognition performance of the model and that the Fca module outperforms the SE module in assigning weights to channel features.

The VoxCeleb2 training set was used to train the model on different network structures. The test results in the table show that the recognition performance of each network model improves significantly as the size of the training dataset increases. The two datasets do not overlap and do not contain the same speakers, so the results in the table also further validate the Fca-Res2Net model proposed in this paper for its generalisability and robustness in the speaker recognition task across datasets.

TABLE III
EER FOR RES2NET WITH DIFFERENT LAYERS

Model	VoxCeleb1		VoxCeleb2	
	EER/%	t-DCF	EER/%	t-DCF
Res2Net-18	2.987	0.209	1.743	0.203
Res2Net-34	2.779	0.196	1.675	0.187
Res2Net-50	2.298	0.146	1.298	0.127
Res2Net-101	2.978	0.198	1.729	0.194

4) Feature Extraction Module Ablation Experiment

In the second part of the experiment, the speaker recognition model based on the Fca-Res2Net network is used to test MFCC, IMFCC, MFCC+IMFCC, and the feature parameters of this paper to verify the effect of feature fusion on the expressiveness of speaker features. The experimental results are shown in Table 4.

TABLE IV
EER FOR FEATURE EXTRACTION ABLATION EXPERIMENTS

Parameters	EER/%	t-DCF
MFCC	5.263	0.181
IMFCC	8.094	0.240
MFCC+IMFCC	3.568	0.153
Proposed	1.298	0.127

To make it easier to observe the impact of the feature parameters on the performance of the network model, the recognition accuracy is visualised in this paper, as shown in Fig. 12 and Fig. 13.

The experimental results show that the fused feature extraction method proposed in this paper works better than single feature extraction methods and can better fit the correlation between channels, bringing some performance gains to the recognition network.

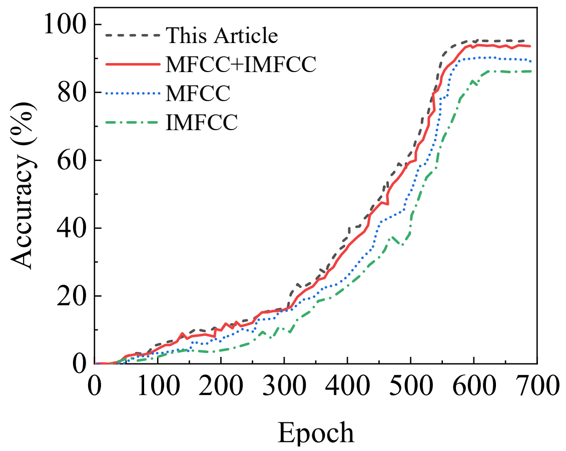

Fig. 12. The feature extraction training map: VoxCeleb1

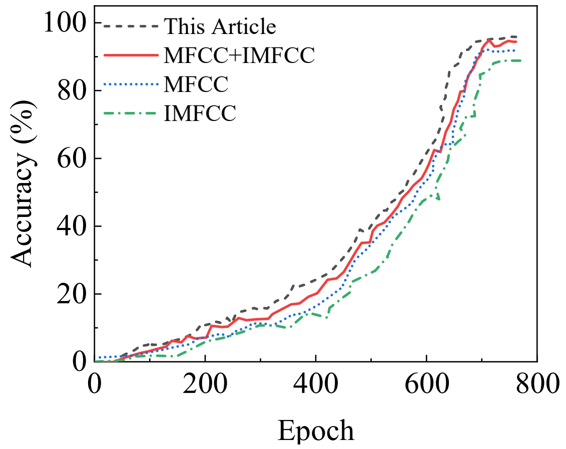

Fig. 13. The feature extraction training map: VoxCeleb2

5) Model Comparison Experiments

The experiments in this section focus on verifying the performance of the proposed speaker recognition model proposed in this paper. The results are compared with those of the traditional model GMM-UBM [27], i-vector+PLDA [28-30], the popular algorithmic model TDNN-UBM [31-32], the network of fusion methods based on embedded features i-vector+DNN [33], MFCC+CNN [34-35]and the popular speaker models SE-ResNet and ResNet+attention [36] for comparison tests. The results of the recognition accuracy experiments are shown in Table 5.

TABLE V
COMPARISON OF RECOGNITION RATES WITH OTHER MODEL METHODS

Model	0s-5s	5s-15s	15s-30s
GMM-UBM[27]	61.17%	74.00%	88.00%
i-vector+PLDA[28-30]	69.50%	86.10%	87.58%
TDNN-UBM[31-32]	72.51%	80.19%	85.53%
i-vector+DNN[33]	81.55%	88.80%	93.90%
MFCC+CNN[34-35]	85.16%	83.14%	91.62%
SE-Res2Net[36]	89.85%	90.59%	92.80%
ResNet+attention[36]	90.13%	90.80%	94.50%
Fca-Res2Net	91.49%	91.80%	95.65%

The comparison results demonstrate that the speaker recognition network architecture based on Fca-Res2Net improves the performance of the speaker recognition system to a certain extent when compared to traditional methods and currently popular algorithms. The recognition rate of the system remains high as the duration of the speaker speech segment increases, and our approach has a significant advantage over traditional algorithms in a 15 s to 30 s speech environment. Compared to the classic recognition model MFCC+CNN, the recognition rate of the Fca-Res2Net-based speaker recognition model improves by almost 3%. It can be seen that the proposed feature extraction algorithm outperforms the algorithm using MFCC alone.

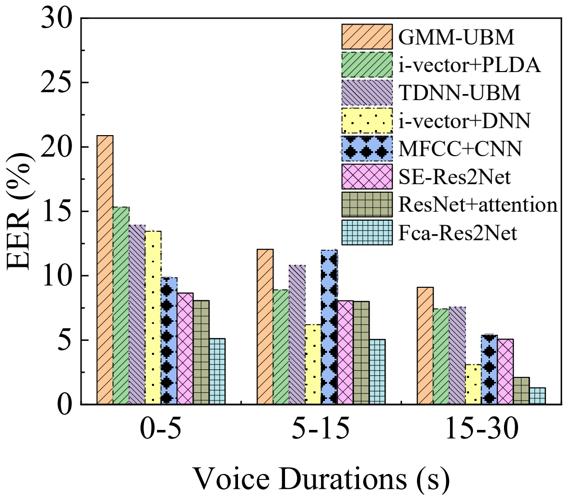

Fig. 14. The EER for different models

In Fig. 14, the traditional i-vector+DNN algorithm model maintains good recognition performance in the long speech range. Our model still outperforms this model, which proves that the recognition model embedded with the self-attention mechanism proposed in this paper has excellent performance in the long speech environment. The above experiments were conducted in a text-independent environment, and therefore, the experimental results show that the proposed model is superior in text-independent recognition.

IV. Conclusion

This paper proposes a speaker network model based on Fca-Res2Net. The input to this model is a speech spectrum matrix obtained after fusion feature extraction. Compared to traditional feature extraction methods based only on MFCC, the improved speaker feature extraction method better preserves the high- and low-frequency information of the speaker and the dynamic information of the speech, ensuring that all key information of the speaker's identity is extracted. Second, we introduce FcaNet into the Res2Net network, forming the Fca-Res2Net module. This module is capable of extracting finer-grained multiscale features, adaptively recalibrating the feature response between channels, expanding the perceptual field and suppressing degradation problems caused by network deepening, thus obtaining high-quality speaker features. In addition, we introduce a self-attention mechanism for a global overview of speech segments, focusing on high-quality speech information. We ultimately obtain highly representative speaker features to represent the speaker's identity. The Fca-Res2Net-based speaker network model was tested on the VoxCeleb dataset. The experimental results show that using the spectrogram of the speech signal can effectively identify the speaker and help improve the recognition performance of the model. Meanwhile, the improved model was tested, and the introduction of Fca-block and self-attention was proven to be able to achieve better recognition results. Compared with other methods, the method proposed in this paper can effectively improve the accuracy and robustness of speaker recognition.

References

[1] Pawar R V, Kajave P P, Mali S N. Speaker Identification using Neural Networks[C]//Iec (prague). 2005: 429-433.

[2] BAI, ZHONGXIN, ZHANG, XIAO-LEI. Speaker recognition based on deep learning: An overview[J]. Neural Networks: The Official Journal of the International Neural Network Society,2021,14065-99. DOI:10.1016/j.neunet.2021.03.004.

[3] Vazhenina D, Markov K. End-to-end noisy speech recognition using fourier and hilbert spectrum features[J]. Electronics, 2020, 9(7): 1157.

[4] Yadava T G, Jayanna H S. Speech enhancement by combining spectral subtraction and minimum mean square error-spectrum power estimator based on zero crossing[J]. International Journal of Speech Technology, 2019, 22: 639-648.

[5] Li Q, Yang Y, Lan T, et al. MSP-MFCC: Energy-efficient MFCC feature extraction method with mixed-signal processing architecture for wearable speech recognition applications[J]. IEEE Access, 2020, 8: 48720-48730.

[6] Mitra V, Franco H, Stern R M, et al. Robust features in deep-learning-based speech recognition[J]. New Era for Robust Speech Recognition: Exploiting Deep Learning, 2017: 187-217.

[7] Tirumala S S, Shahamiri S R. A review on deep learning approaches in speaker identification[C]//Proceedings of the 8th international conference on signal processing systems. 2016: 142-147.

[8] Gao S H, Cheng M M, Zhao K, et al. Res2net: A new multi-scale backbone architecture[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(2): 652-662.

[9] Tang Y, Liu C, Leng Y, et al. Attention based gender and nationality information exploration for speaker identification[J]. Digital Signal Processing, 2022, 123: 103449.

[10] Qin Z, Zhang P, Wu F, et al. Fcanet: Frequency channel attention networks[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 783-792.

[11] Nirmal A, Jayaswal D. Investigating Gammatone Filterbank-Based i-Vectors for Speaker Verification Task[M]//Information and Communication Technology for Competitive Strategies (ICTCS 2021) Intelligent Strategies for ICT. Singapore: Springer Nature Singapore, 2022: 767-775.

[12] Reynolds D A. An overview of automatic speaker recognition technology[C]//2002 IEEE international conference on acoustics, speech, and signal processing. IEEE, 2002, 4: IV-4072-IV-4075.

[13] Ding S, Wang Q, Chang S, et al. Personal VAD: Speaker-conditioned voice activity detection[J]. arXiv preprint arXiv:1908.04284, 2019.

[14] Ayvaz U, Gürüler H, Khan F, et al. Automatic Speaker Recognition Using Mel-Frequency Cepstral Coefficients Through Machine Learning[J]. 2022.

[15] Chandrasekaram B. New Feature Vector based on GFCC for Language Recognition[J]. JOURNAL OF ALGEBRAIC STATISTICS, 2022, 13(2): 481-486.

[16] Ding S, Chen T, Gong X, et al. Auto speech: Neural architecture search for speaker recognition[J]. arXiv preprint arXiv:2005.03215, 2020.

[17] Roy M K, Keshwala U. Res2Net based Text Independent Speaker recognition system[C]//2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2022: 612-616.

[18] Li R, Jiang J Y, Wu C, et al. Speaker identification for household scenarios with self-attention and adversarial training[J]. 2020.

[19] Varshaneya V, Balasubramanian S, Gera D. Res-SE-Net: Boosting Performance of ResNets by Enhancing Bridge Connections[J]. Machine Learning Algorithms and Applications, 2021: 61-75.

[20] An N N, Thanh N Q, Liu Y. Deep CNNs with self-attention for speaker identification[J]. IEEE access, 2019, 7: 85327-85337.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[22] Nagrani A, Chung J S, Zisserman A. Voxceleb: a large-scale speaker identification dataset[J]. arXiv preprint arXiv:1706.08612, 2017.

[23] Ülgen İ R, Arslan L M. Unsupervised Domain Adaptation of Neural PLDA Using Segment Pairs for Speaker Verification[C]//2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2023: 571-576.

[24] SARMAH A, REHMAN R, MAHANTA P, et al. A Novel Approach for automatic speaker identification of Assamese language using cosine similarity and absolute MFCC Feature Matrix[J]. Journal of Theoretical and Applied Information Technology, 2022, 100(21).

[25] Doddington G R. Speaker recognition—Identifying people by their voices[J]. Proceedings of the IEEE, 1985, 73(11): 1651-1664.

[26] Reynolds D, Singer E, Sadjadi S O, et al. The 2016 nist speaker recognition evaluation[R]. MIT Lincoln Laboratory Lexington United States, 2017.

[27] Akula A, Apsingekar V R, De Leon P L. Speaker identification in room reverberation using GMM-UBM[C]//2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop. IEEE, 2009: 37-41.

[28] Li R, Li L, Hong Q, et al. Improving the Generalized Performance of Deep Embedding for Text-Independent Speaker Verification[C]//2018 12th IEEE International Conference on Anti-counterfeiting, Security, and Identification (ASID). IEEE, 2018: 21-25.

[29] Chakroun R, Frikha M. Efficient text-independent speaker recognition with short utterances in both clean and uncontrolled environments[J]. Multimedia Tools and Applications, 2020, 79(29-30): 21279-21298.

[30] Poddar A, Sahidullah M, Saha G. Performance comparison of speaker recognition systems in presence of duration variability[C]//2015 Annual IEEE India Conference (INDICON). IEEE, 2015: 1-6.

[31] Liu H, Zhao L. A speaker verification method based on TDNN–LSTMP[J]. Circuits, Systems, and Signal Processing, 2019, 38: 4840-4854.

[32] Toruk M M, Gokay R. Short utterance speaker recognition using time-delay neural network[C]//2019 16th International Multi-Conference on Systems, Signals & Devices (SSD). IEEE, 2019: 383-386.

[33] Guo J, Xu N, Qian K, et al. Deep neural network based i-vector mapping for speaker verification using short utterances[J]. Speech Communication, 2018, 105: 92-102.

[34] Novoselov S, Kudashev O, Shchemelinin V, et al. Deep cnn based feature extractor for text-prompted speaker recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5334-5338.

[35] Jagiasi R, Ghosalkar S, Kulal P, et al. CNN based speaker recognition in language and text-independent small scale system[C]//2019 third international conference on i-smac (iot in social, mobile, analytics and

cloud)(I-SMAC). IEEE, 2019: 176-179.

[36] An N N, Thanh N Q, Liu Y. Deep CNNs with self-attention for speaker identification[J]. IEEE access, 2019, 7: 85327-85337.