Fca-ProRes2Net Speaker Recognition Based on Fusion Feature Dimensionality Reduction

Zhangfang Hu, Yi Wang, Yuan Yuan

Abstract—**With the advancement of the era, speaker recognition technology plays a pivotal role in applications such as remote telephone identity verification and anti-telecom fraud. However, traditional identification methods utilizing single feature parameters often result in inadequate representation and information loss. Additionally, with the introduction of convolutional neural networks, many approaches to adopt deeper and wider network structures to enhance recognition effectiveness, leading to a decline in network performance. In response to this issue, this paper proposes a speaker recognition network model, Fca-ProRes2Net, based on the fusion of feature parameter dimension reduction. Firstly, a novel hybrid parameter is formed by integrating traditional Mel Frequency Cepstral Coefficients (MFCC) with more robust Gammatone frequency cepstral coefficients (GFCC), capturing both mid-to-high frequency and dynamic/static features. Furthermore, 2DPCA is employed to reduce and integrate the feature matrix, addressing the issue of information redundancy caused by high-dimensional feature parameters and thereby enhancing the representation capability of the parameters. Secondly, the Res2Net network is incorporated into the recognition model, utilizing its fully connected form, known as ProRes2Net, to effectively expand the receptive field combination without significantly increasing parameters, thus enlarging the model's acceptance domain. Lastly, Frequency Channel Attention Networks (Fca-Net) are integrated into this model to redistribute weights among feature parameter channels, enhancing the model's recognition ability. Experimental results demonstrate a stable performance improvement of this method on the complex VoxCeleb dataset, particularly evident when data is abundant and tasks are complex, effectively enhancing the accuracy and robustness of the speaker recognition model.**

Index Terms—**Speaker Recognition, 2DPCA, Fca-ProRes2Net, Attention mechanism**

Manuscript received July 11, 2024; revised February 6, 2025.

This work was supported in part by the Youth Fund Program of the National Natural Science Foundation of China (Grant No. 61703067), the Chongqing Basic Science and Frontier Technology Research Program (Grant No. Cstc2017jcyjAX0212), and the Science and Technology Research Program of Chongqing Municipal Education Commission (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).

Yi Wang is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (corresponding author phone: 181-8305-2939; e-mail: 2253356946@qq.com).

Yuan Yuan is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 2250599628@qq.com).

I. INTRODUCTION

SPEAKER recognition, commonly known as voiceprint recognition [1], shares similarities with identity authentication technologies such as fingerprint and facial recognition, as it emphasizes the unique characteristics that distinguish individuals. This technology extracts distinctive features from speech signals to determine the identity of the speaker. The production of human language necessitates the coordinated operation of various physiological components and vocal apparatus [2]. Each individual's speech is marked by variations in short-time spectra, vocal sources, rhythms, and other linguistic characteristics. Voiceprint recognition effectively differentiates individuals by utilizing these feature discrepancies, thereby enabling speaker identification. Compared to other biometric recognition methods, speaker recognition presents several advantages, including ease of data collection, resilience to environmental interference, and a non-contact approach. This technology is particularly beneficial in scenarios where the acquisition of other biometric data, such as facial images or fingerprints, is impractical, such as in remote telephone identity verification and the prevention of telecom fraud. As social interactions and speech contexts become increasingly varied, the demand for speaker recognition technology is growing, alongside heightened concerns regarding privacy and the need for improved system development [3]. Consequently, as the necessity for speaker authentication technology intensifies and relevant advancements occur, enhancing system recognition performance has emerged as a critical area of research in recent years [4].

Speaker recognition can be classified into two primary types based on the ultimate task performed: Speaker Verification (SV) and Speaker Identification (SI) [5]. SV focuses on determining whether a specific speech segment is produced by a designated individual, while SI seeks to identify which individual among a group has spoken a particular segment. Additionally, speaker recognition can be further categorized into three types based on the nature of the content: text-dependent, text-independent, and text-prompted [6]. This paper specifically addresses research related to text-independent speaker verification.

The fundamental framework of speaker recognition consists of two key components: feature extraction and the formulation of recognition models. The primary aim of feature extraction is to identify parameters that capture the unique characteristics of individual speakers, thereby highlighting the differences

among them. These feature parameters are categorized into temporal and spectral features. Temporal features aid in distinguishing between voiced and unvoiced segments, while the perceptual characteristics of various speakers are primarily represented in the spectral domain. As a result, the use of temporal features is somewhat limited due to their inadequate capacity for representing speakers [7]. Recent developments have led to the incorporation of various spectral features, such as Linear Prediction Coefficients (LPC), Filter Bank (F-bank) features derived from filter groups, Mel Frequency Cepstral Coefficients (MFCC), and Gammatone Frequency Cepstral Coefficients (GFCC) [8]. Notably, MFCC features encapsulate the perceptual dimensions of human auditory perception through a series of Mel-frequency filter banks [9], which enhances their discriminative power, making them more widely utilized in feature extraction methodologies. However, MFCC features are somewhat susceptible to noise interference, a limitation that GFCC features effectively mitigate [10].

Traditional speaker recognition models include the Gaussian Mixture Model-Universal Background Model (GMM-UBM), Gaussian Mixture Model-Support Vector Machine (GMM-SVM), Joint Factor Analysis (JFA), and i-vector, among others [11]. These models are generally classified as shallow architectures, which involve limited linear or non-linear processing of raw input signals to achieve signal and information processing goals. Consequently, they exhibit a restricted capacity for modeling complex speech signals and fail to adequately represent structured and high-level information within the signal. The introduction of deep learning has facilitated the development of Deep Neural Networks (DNNs), which have demonstrated significant recognition capabilities across various feature recognition domains [12]. Deep learning has made substantial progress in fields such as data mining, natural language processing, and image processing. In the context of speaker recognition, D-Vector is recognized as one of the initial neural network models that utilizes DNNs for speech recognition, characterized by multiple fully connected hidden layers. Convolutional Neural Networks (CNNs) have been widely adopted as the foundational neural network for speaker recognition systems, enabling the extraction of deeper network parameters. Among these, ResNet is a variant of CNN that addresses issues such as gradient vanishing and explosion. By implementing residual connections, ResNet can construct very deep networks while simultaneously minimizing parameters and computational complexity. However, lightweight CNNs often exhibit limited stability. To address this challenge, Gao et al. [13] proposed Res2Net, which builds upon ResNet and allows for a more detailed hierarchical representation of multiscale features, thereby improving the capture of both deep and shallow features of the speaker. Nonetheless, an indiscriminate increase in the depth and width of network architectures may lead to overfitting. Therefore, the development of an effective network architecture is crucial for enhancing speaker recognition performance.

In consideration of the aforementioned analysis, this research presents an advanced speaker recognition model that incorporates two-dimensional Principal Component Analysis (2D PCA) for the purpose of feature dimension reduction, in conjunction with the Res2Net architecture. This integration is intended to enhance the overall effectiveness of the speaker recognition system. The key contributions of this study are outlined as follows:

1) In the feature extraction phase, a combination of features, including Discrete Cosine Transform (DCT)-operated MFCC, and GFCC, are combined, followed by dimension reduction and integration using 2D PCA, to extract effective combined dynamic and static features.
2) In the speaker recognition model section, a fully connected multi-scale Res2Net network is employed. This network architecture features a larger receptive field than traditional Res2Net networks. Furthermore, a Frequency Channel Attention Network (Fca-Net) is integrated to reweight information from various channels, thereby improving the model's generalization capability and recognition performance.
3) The proposed speaker recognition model is trained and tested on the VoxCeleb dataset, and experiments are conducted comparing it with traditional speaker recognition networks. The results demonstrate the model's effectiveness in extracting highly representative feature parameters and efficiently recognizing speakers.

II. Related works

The speaker recognition model is developed and assessed within a deep learning framework. The foundational architecture for speaker recognition, as presented in this study, is depicted in Figure 1. Enhancements will be introduced in two primary domains: feature extraction and the recognition model, leading to the proposal of a fully connected Fca-ProRes2Net speaker recognition method that utilizes dimensionality reduction via fusion feature 2D PCA [14]. In the feature extraction phase, MFCC, GFCC, ΔMFCC and ΔGFCC, are combined without employing DCT to generate a set of mixed feature parameters. Subsequently, dimensionality reduction techniques are applied to streamline the feature matrix, thereby mitigating the issues of information redundancy associated with high dimensionality and resulting in highly representative parameters. An advanced multi-scale fully connected Res2Net network is then employed for the recognition model, which allows for an expanded receptive field without a substantial increase in the number of parameters. To enhance feature interdependence and bolster the stability and robustness of the system, Fca-Net [15] is incorporated to reweight information across various channels. Following the fully connected layer, the Softmax function is utilized to classify the output results. The optimization of the speaker recognition model is achieved through the application of the cross-entropy loss function, which seeks to identify optimal weight parameters.

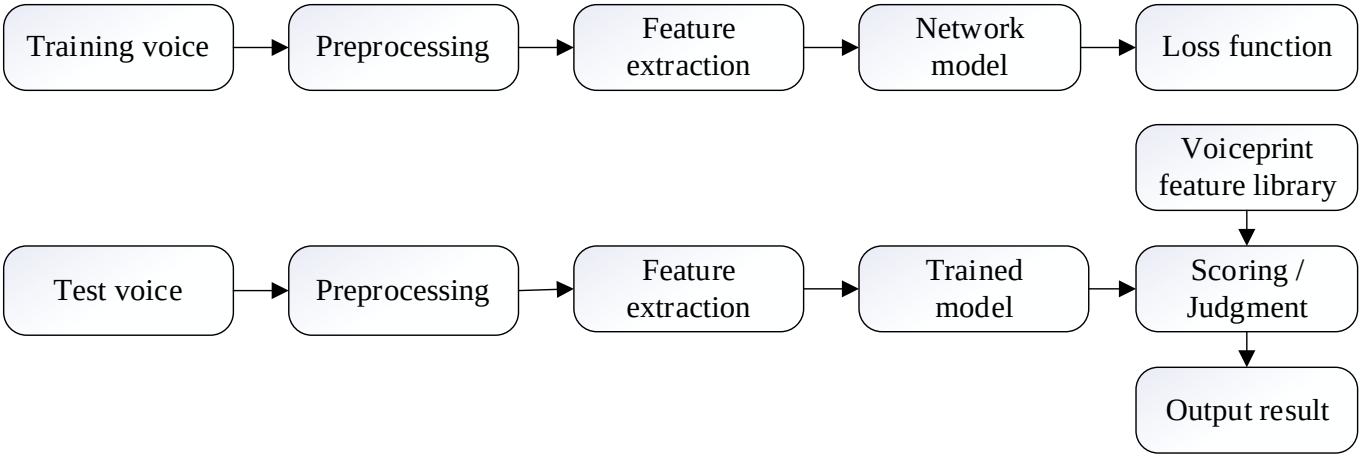

Fig. 1. Depicts the fundamental framework of speaker recognition.

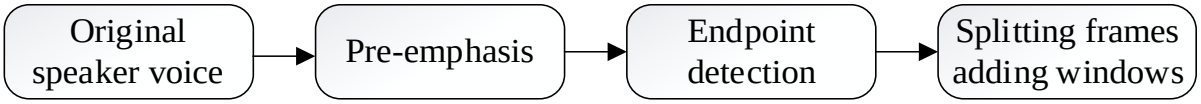

Fig. 2. The process of preprocessing.

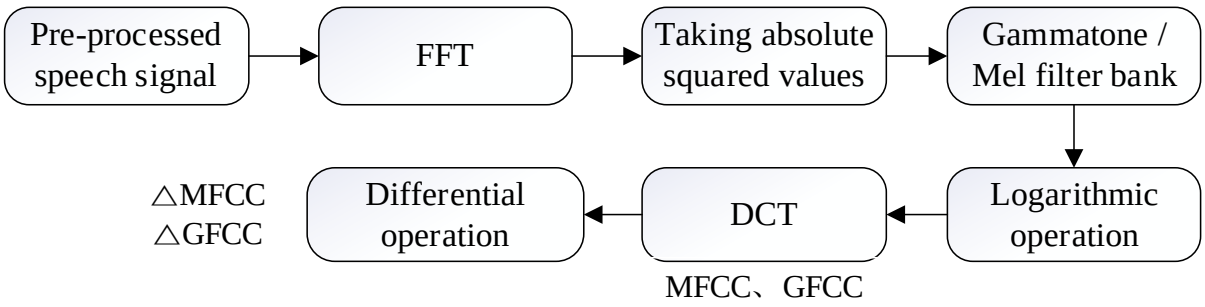

Fig. 3. Illustrates the process of feature extraction.

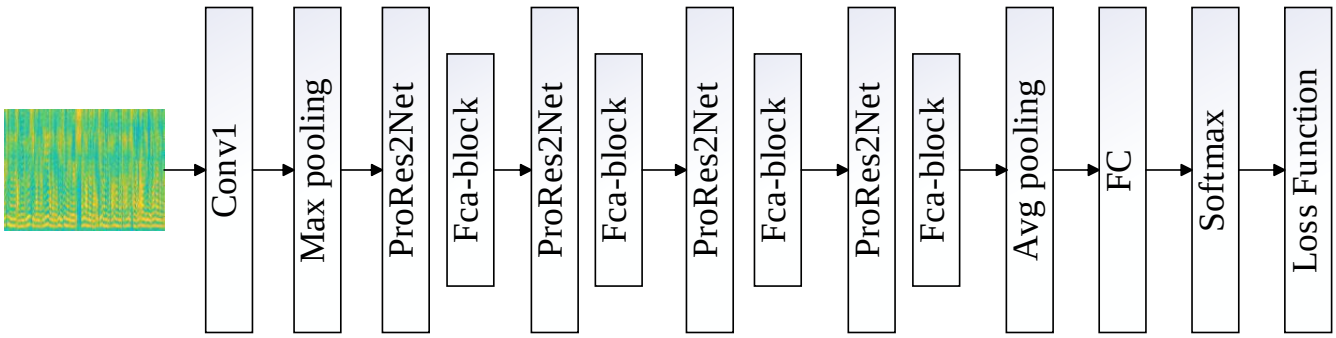

Fig. 4. The framework of the speaker recognition network.

III. Methodology

A. Feature extraction

Due to the unique anatomical and phonetic attributes of each individual's vocal tract, as well as their specific pronunciation patterns, speech manifests distinctive characteristics that set it apart from that of others. Therefore, the extraction of effective and highly representative features is essential for the advancement of a robust speaker recognition system.

1) Preprocessing

The speech signal is acquired through microphone recording, during which the high-frequency components experience significant attenuation. Additionally, the recorded speech may contain segments of silence [16]. Consequently, preprocessing the speech signal is essential to derive appropriate signals for feature extraction pertinent to speaker identification. The preprocessing of speech signals primarily involves pre-emphasis, endpoint detection, and the segmentation of frames with windowing techniques [17].

Pre-emphasis: In the process of generating speech signals, high-frequency components are subject to rapid attenuation. Pre-emphasis serves to amplify these high-frequency elements, thereby counteracting the effects of attenuation and enhancing the high-frequency information present in the signal. This amplification is accomplished by routing the signal through a digital filter specifically designed to enhance high-frequency content.

Endpoint Detection: This procedure entails the identification of both the initiation and conclusion of speech within a signal that contains spoken words. Endpoint detection is crucial for various speech-related tasks that necessitate the analysis or processing of only the speech segments [18]. It is frequently utilized in algorithms for speech encoding and decoding, noise reduction, wake word recognition, and other applications.

Frame Segmentation and Windowing: Speech signals exhibit temporal variations; however, due to their short-time stationary characteristics, they are often divided into brief intervals along the time axis, referred to as frames. To facilitate smooth transitions between frames, adjacent frames typically overlap, with half of their length designated as the overlap length. Windowing addresses issues such as the Gibbs phenomenon and spectral leakage that may occur following Fourier transformation by applying distinct weights to each value within a frame.

2) Feature extraction

Speaker recognition technology commences with the extraction of feature parameters from processed speech data, a pivotal step that significantly influences the final recognition results. The primary process is depicted in Figure 3. While MFCC, which are based on human auditory characteristics, demonstrate limited robustness in noisy environments, GFCC exhibit substantial resistance to noise, thus providing a complementary approach to MFCC. To address the loss of essential information that occurs during the extraction phase as a result of the DCT operation, this study advocates for the elimination of DCT. MFCC and GFCC features primarily capture the static characteristics of speech. Therefore, by implementing differential operations on these two feature sets, ΔMFCC and ΔGFCC are produced, which rectify the inadequacy of dynamic features within the speech signal [19]. Ultimately, these four features are integrated and concatenated to create Mixed Mel Gamma Frequency Cepstrum Coefficients (MMGFCC).

The augmentation of feature parameters can significantly improve the robustness and accuracy of recognition models; however, it simultaneously introduces redundancy and complexity within the subsequent networks. To extract high-quality information while minimizing resource expenditure, dimensionality reduction is applied to the MMGFCC features. A prevalent method for achieving this is Principal Component Analysis (PCA) [20]. Nonetheless, PCA necessitates the transformation of a two-dimensional matrix into a one-dimensional vector, which may not fully leverage the available feature information. As a result, this study employs 2D PCA. In contrast to PCA, 2D PCA does not require the conversion of the feature matrix into a one-dimensional vector; rather, it directly utilizes the original image matrix to construct the covariance matrix. This methodology facilitates the preservation of the local data structure and produces highly representative features while streamlining the computation of feature vectors [21]. Let $\{A_1, A_2, \ldots, A_N\}$ denote the two-dimensional feature image matrices $X \in R^{n\times k}$ that are projected into space $Y \in R^{m\times k}$, resulting in the projected feature vectors $Y \in R^{m\times k}$, as demonstrated in equation (1).

$$Y = AX \tag{1}$$

The ideal projection axis can be identified by analyzing the distribution of feature vectors, particularly through the assessment of the trace of the covariance matrix obtained from the projected features. The criteria employed in this evaluation are as follows:

$$J(X) = tr(S_x) \tag{2}$$

S_x represents the covariance matrix of the feature vectors Y, while $tr(S_x)$ denotes the dispersion of S_x. The definition of S_x is as follows:

$$S_x = E[(A - EA)X][(A - EA)X]^T \tag{3}$$

Therefore, the dispersion is:

$$tr(S_x) = tr(X^T E[(A - EA)^T (A - EA)]X) \tag{4}$$

Assuming a sample set comprises N distinct speech feature images $\{A_i\}_1^N$, with a size of $m \times n$, the covariance matrix of the overall feature images is:

$$G_t = \frac{1}{N}\sum_{i=1}^{N}(A_i - \bar{A})^T(A_i - \bar{A}) \quad (5)$$

Where $\bar{A}$ represents the average of all training samples. Firstly, compute the feature vectors corresponding to G_t, then select d feature vectors $X_1, X_2, \ldots, X_d$ as the principal components of the matrix G_t to form the projection matrix X, obtaining the optimal projection axis. Subsequently, project A_i onto space X to obtain a set of projection feature vectors $Y_1, Y_2, \ldots, Y_d$, which collectively form a matrix $m \times d$ denoted as $M_i = Y_1, Y_2, \ldots, Y_d$.

B. Modeling of Speaker Recognition

CNNs are generally designed by sequentially stacking convolutional layers alongside downsampling layers. However, as the depth of the network increases, several challenges emerge, including vanishing gradients, exploding gradients, parameter redundancy, and heightened computational complexity [22]. To address these challenges, the present study proposes enhancements to the recognition network. The architecture of the proposed recognition network is illustrated in Figure 5.

This research incorporates Res2Net into the speaker recognition framework [23], facilitating the establishment of a fixed receptive field size within the network. In the realm of deep learning, Squeeze-and-Excitation Networks (SE-Net) are frequently utilized to assign weights to significant acoustic feature channels, thereby reflecting their importance in relation to critical features. A higher weight signifies greater relevance. However, SE-Net employs Global Average Pooling (GAP) for dimensionality reduction, which compresses the two-dimensional features of each channel into a single real number. This transformation modifies the feature maps from a three-dimensional configuration of *(H, W, C)* to a two-dimensional format of *(1, 1, C)*, resulting in the loss of certain feature information [15]. To enhance the model's recognition performance, modifications are introduced to the Res2Net architecture by integrating a fully connected structure and the Fca-Net module, which retains a greater amount of feature information. This culminates in the development of a speaker recognition network termed Fca-ProRes2Net. In comparison to the Res2Net architecture, the Fca-ProRes2Net network achieves an expanded receptive field with minimal increases in parameters [23], thereby improving both recognition accuracy and efficiency.

1) Res2Net and ProRes2Net

CNNs, a prominent subset of feedforward neural networks, are distinguished by their incorporation of convolutional operations and deep architectural designs. Their primary purpose is to facilitate the efficient and automated extraction of feature information. This architecture has consistently demonstrated outstanding performance across diverse recognition tasks, establishing itself as a cornerstone of research in the field. To further improve the network's representational power and recognition accuracy, researchers have frequently employed strategies to expand both its width and depth. However, increasing network depth often introduces optimization challenges, such as gradient vanishing and gradient explosion, which can significantly impede effective training. To overcome these challenges, the ResNet was introduced, as depicted in Figure 5, providing a reliable and effective framework for training deep neural networks. Unlike conventional network architectures, ResNet incorporates a direct connection between the input and output of each residual block, thereby allowing the input to be transmitted to the output as an initial result, denoted as $H(x) = F(x) + x$. This innovative design facilitates an increase in network depth without compromising accuracy, while also fostering a more efficient learning process.

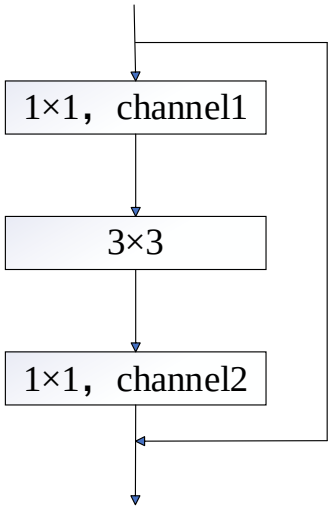

Fig. 5. Structure of the ResNet module.

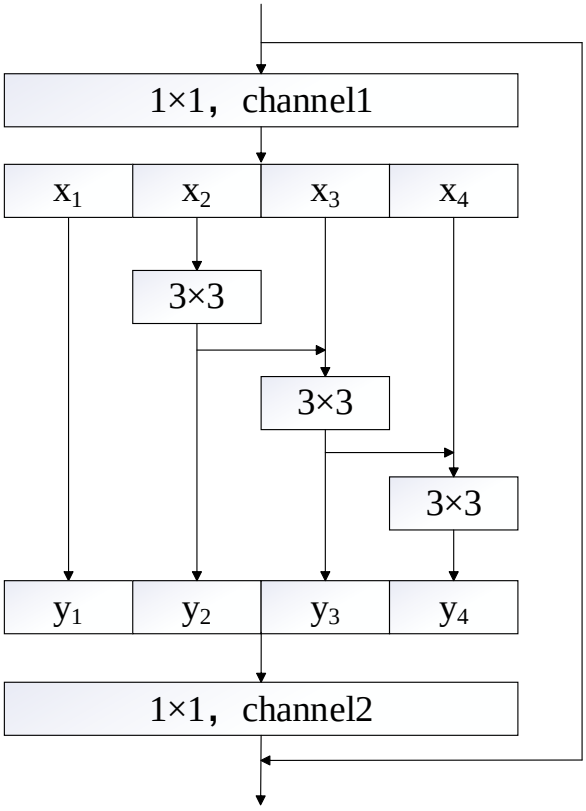

Fig. 6. Structure of the Res2Net module.

Although ResNet demonstrates significant performance advantages in practical applications, its effectiveness still depends largely on increasing network depth. To address this limitation, this study presents an enhanced version of ResNet, called Res2Net[25], which is incorporated into the speaker recognition framework. This variant not only retains the fundamental characteristics of ResNet but also further improves network performance. By introducing hierarchical residual connections within each residual block, Res2Net effectively expands the receptive field of each layer, thereby enhancing the network's feature representation and overall performance. The configuration of the Res2Net module is depicted in Figure 6.

The primary distinction between Res2Net and conventional convolutional neural networks lies in its employment of smaller filters to capture multi-scale feature information. It enhances the single 3×3 convolution structure in ResNet residual blocks by introducing a scale control parameters. For the input feature map $F(H \times W \times C)$, it initially undergoes dimension transformation via a 1×1 convolution, splitting the output into s feature subsets, each denoted as x_i, with shapes represented by $H \times W \times C'$ and $i = 1,2,\ldots,s$, C' denotes the channel number of each subset after averaging. Using $K_i(.)$ to denote a 3×3 convolution, its output is denoted by y_i. The initial feature subset y_i circumvents convolutional computations and transmits directly to the output. In contrast, each subsequent feature subset incorporates the output derived from a 3×3 convolution operation, which is then added to the next feature subset. This iterative process culminates in the following output formula:

$$y_i = \begin{cases} x_i & ,i = 1 \\ K_i(x_i) & ,i = 2 \\ K_i(x_i + y_{i-1}) & ,2 < i \le s \end{cases} \tag{6}$$

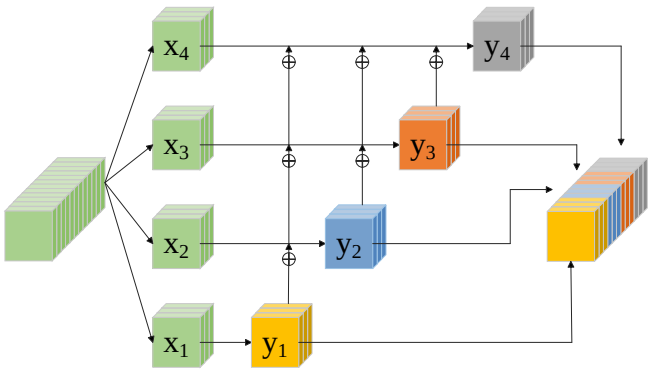

Fig. 7. Structure of the ProRes2Net module.

The generated feature maps are concatenated and integrated according to their channel dimensions. Following this, a 1×1 convolution operation is performed, and the resultant output is merged with the original input features to produce the final feature map, which subsequently serves as the input for the next convolutional layer. In the residual blocks of Res2Net, standard convolutions are substituted with convolutional groups, and hierarchical residual connections are established. This framework of hierarchical residual connections enhances the extraction of a broader spectrum of multi-scale feature information, thereby facilitating the integration of both global and local information. Nonetheless, Res2Net is constrained by its ability to achieve a fixed-size receptive field. To mitigate this limitation and improve the robustness of the recognition model by capturing a wider range of receptive fields, this study employs the fully connected Res2Net module structure, designated as the ProRes2Net module [26], as depicted in Figure 7.

In the Res2Net module, the output from each preceding feature subset is directed solely to the subsequent feature subset. In contrast, as depicted in Figure 7, the ProRes2Net module adopts a different methodology. Within the parallel branches of the ProRes2Net module, each feature subset assimilates all prior outputs before the application of 3×3 convolutional operations. Following this, the outputs from all groups are concatenated along the spatial dimension and processed through a 1×1 convolution operation. This result is then concatenated with the original input features to yield the final feature map.The output y_i is expressed as follows:

$$y_i = \begin{cases} K_i(x_i) & ,i = 1 \\ K_i(x_i + y_{i-1} + \ldots + y_1) & ,1 < i \le s \end{cases} \tag{7}$$

Each subset of features follows a specific processing pathway, integrating diverse feature information and producing outputs via convolutional operations. This approach allows the speaker recognition model to incorporate a broader spectrum of information derived from speech features. As a result, the ProRes2Net module is capable of accessing a wider variety of receptive field sizes, which promotes a more thorough and effective use of speech feature information. Consequently, this enhances the model's performance and demonstrates improved generalization abilities.

2) Fca-Net

The significant efficacy of various attention mechanisms in the domain of computer vision has led to their application in speaker recognition tasks, facilitating models in more effectively capturing the complex interrelationships among input feature parameters. To improve the overall performance of the model, this study integrates the Fca-Net module [15] into the ProRes2Net backbone network. The Fca-Net module is an advancement of the SE-Net, as depicted in Figure 8, which illustrates the structural differences between the SE-Net and Fca-Net modules. The GAP operation employed by SE-Net is limited to retaining low-frequency information [27]; consequently, the literature [15] supports the adoption of Fca-Net, which substitutes the channel attention mechanism's compression with a two-dimensional discrete cosine transformation. This alteration enables the capture of a broader spectrum of information, effectively preserving mid-to-high frequency data, thereby incorporating additional frequency component information that enhances recognition accuracy.

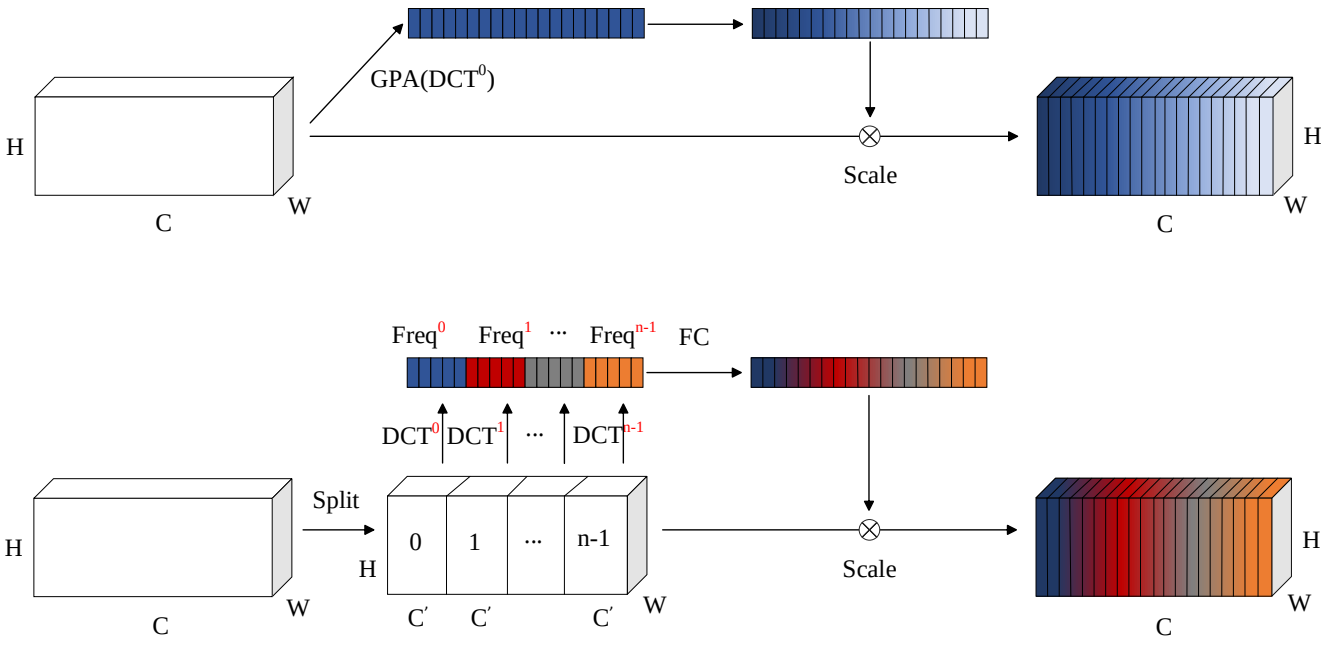

Fig. 8. SE-Net module and Fca-Net module.

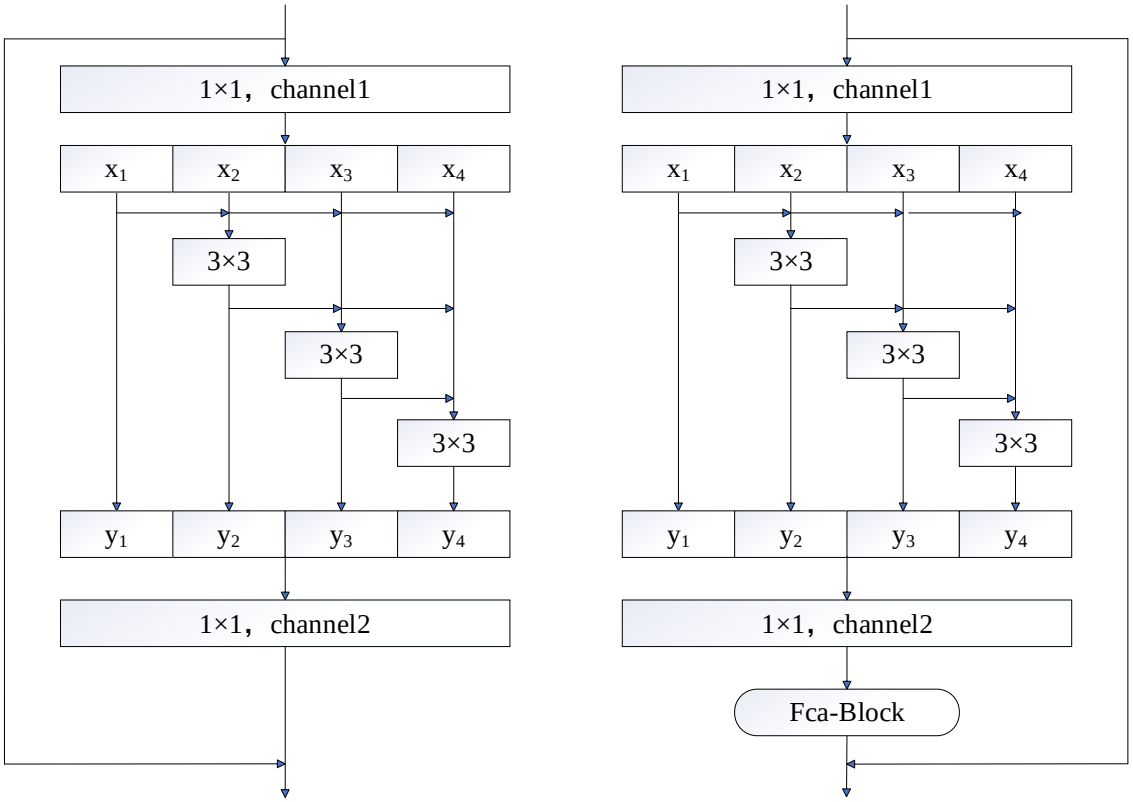

Fig. 9. ProRes2Net module and Fca-ProRes2Net module.

The primary functions of a two-dimensional Discrete Cosine Transform are delineated as follows:

$$B_{h,w}^{i,j} = \cos(\frac{\pi h}{H}(i+\frac{1}{2}))\cos(\frac{\pi w}{W}(j+\frac{1}{2})) \quad (8)$$

Consequently, the two-dimensional Discrete Cosine Transform can be expressed as follows:

$$y_i = \begin{cases} K_i(x_i) & ,i=1 \\ K_i(x_i + y_{i-1} + \ldots + y_1) & ,1 < i \le s \end{cases} \quad (9)$$

Here, $h \in \{0,1,\ldots,H-1\}$, $w \in \{0,1,\ldots,W-1\}$, and f^{2d} represent the two-dimensional Discrete Cosine Transform spectrum, $x^{2d} \in R^{H \times W}$ is the input, and H and W denote the height and width of x^{2d}. The inverse of the two-dimensional Discrete Cosine Transform can be expressed as:

$$x_{i,j}^{2d} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} f_{h,w}^{2d} B_{h,w}^{i,j} \quad (10)$$

When both variables h and w are equal to zero, the above equation becomes:

$$f_{0,0}^{2d} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d} = gap(x^{2d})HW \quad (11)$$

At this juncture, the variable f denotes the lowest frequency component of the 2D DCT, which exhibits a direct proportional relationship with the GAP. Consequently, it can be inferred that GAP serves as a particular representation of the 2D DCT.

Divide along the channel dimension of X into n parts, denoted as $[X^0, X^1, \ldots, X^{n-1}]$, where $X^i \in R^{C' \times H \times W}$, $i = 0,1,\ldots,n-1$, $C' = C/n$, and n can evenly divide C. Each segment is assigned a corresponding 2D DCT frequency component, resulting in the compression of channel attention. Assign a corresponding 2D DCT frequency component to each part, and the resulting output is the compression of channel attention. The compressed C' dimensional vector is:

$$Freq^i = 2DDCT^{u_i,v_i}(X^i) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_{:,h,w}^{i} B_{h,w}^{u_i,v_i} \quad (12)$$

Here, $i = 0,1,\ldots,n-1$ and (u_i, v_i) are the frequency components corresponding to the two-dimensional exponentials of X^i, and finally, by concatenating all the compressed results, we obtain:

$$Freq = compress(X) = cat([Freq^0, Freq^1, \ldots, Freq^{n-1}]) \quad (13)$$

In accordance with the aforementioned rationale, the structural framework of the Fca-Net network is delineated as follows:

$$ms\_att = sigmoid(fc(Freq)) \quad (14)$$

In summary, the integration of the Fca-Net module with the ProRes2Net module serves to enhance the model's recognition capabilities and robustness, as illustrated in Figure 9.

IV. Experiments

A. Dataset

This research involved conducting experiments using the VoxCeleb1 [28] and VoxCeleb2 [29] open-source datasets to support text-independent speaker recognition. The VoxCeleb dataset series consists of an extensive compilation of human speech data sourced from interview videos available on video-sharing platforms. The interview participants represent a wide range of nationalities, ethnicities, accents, ages, and genders, ensuring a balanced gender distribution and the absence of overlap between the development and test sets. Specifically, VoxCeleb1 includes over 100,000 utterances from 1,251 speakers, with 1,211 speakers allocated to the training set and 40 to the test set. The average duration of the utterances is 8.2 seconds, with a maximum length of 145 seconds and a minimum of 4 seconds, predominantly comprising short phrases, all recorded at a sampling rate of 16 kHz, 16-bit, and in mono format. In contrast, VoxCeleb2 offers a more extensive dataset, containing over one million utterances from 6,112 speakers, with 5,994 speakers in the training set and 118 in the test set. For the purposes of this study, the training sets from both VoxCeleb1 and VoxCeleb2 were employed for training, while the VoxCeleb1 test set was utilized for evaluation.

B. Experimental Data

In order to assess the efficacy of the improvements implemented in the speaker recognition system within this research, traditional training methodologies were utilized to evaluate performance. Speech segments, each with a duration of 2.5 seconds, were extracted from individual speakers in the dataset. The selection of an appropriate window function is critical for optimizing the short-term characteristics of the speech signal. The rectangular window, while straightforward, may result in the attenuation of high-frequency components and waveform details, leading to energy leakage. In contrast, the Hamming window provides a broader main lobe and reduced side lobes, thereby facilitating smoother low-pass characteristics. Although both the Hann and Hamming windows are based on cosine functions, the Hamming window is characterized by its smaller side lobes. As illustrated in Figure 10, this study employed a Hamming window with a duration of 25 milliseconds and a sliding step of 10 milliseconds for the preprocessing of each speech segment. A total of 128 filter groups were utilized, yielding feature matrices with dimensions of 250×128 for each feature extraction process. These matrices were subsequently concatenated to create a composite feature matrix measuring 250×512. Following the integration of more representative features through the application of 2DPCA, the dimensions of the resulting matrix were reduced to 250×200. To improve the model's generalization capabilities and to address the limited racial diversity inherent in the VoxCeleb1 dataset, training initially commenced with VoxCeleb1 and was later expanded to incorporate the larger VoxCeleb2 dataset. Performance

evaluation was conducted using the test set from VoxCeleb1, ensuring that the datasets employed were non-overlapping.

In the present study, preliminary experiments were carried out to train and evaluate the performance of the Res2Net, SE-Res2Net, SE-ProRes2Net, and the newly proposed Fca-ProRes2Net networks. Subsequently, a series of ablation studies were performed to assess the efficacy of the proposed methodologies. Ultimately, a comparative analysis and evaluation were conducted among the network models developed in this research, conventional network architectures, and contemporary deep learning models.

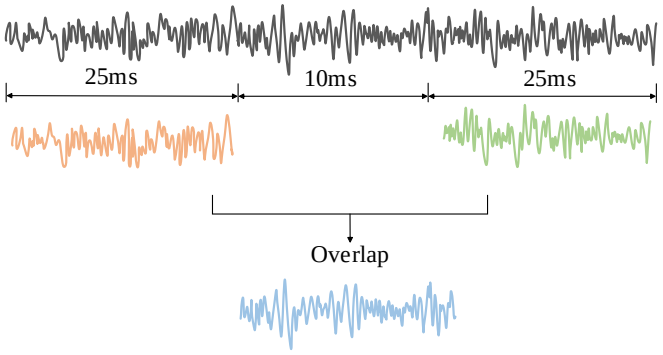

Fig. 10. Schematic of frame length extraction.

C. Experimental evaluation indicators

The study employed a straightforward cosine similarity scoring method to compute the Equal Error Rate (EER) and the minimum Detection Cost Function (minDCF) as evaluative metrics for recognition performance [30].

In the context of speaker recognition systems, False Rejection (FR) and False Acceptance (FA) are two separate types of errors. FR occurs when a genuine speaker is incorrectly classified as an impostor during the assessment, while FA arises when an impostor is erroneously identified as a legitimate speaker. The probabilities related to these errors can be expressed as follows:

$$FRR = \frac{N_{fr}}{N_{target}} \tag{15}$$

$$FAR = \frac{N_{fa}}{N_{non-target}} \tag{16}$$

In this context, N_{fr} and N_{fa} refer to the occurrences of FR and FA during the testing phase, respectively. Additionally, N_{target} and $N_{non\text{-}target}$ signify the total number of genuine trials and impostor trials conducted during testing. Given the existence of both FR and FA errors, it is not feasible to evaluate the performance of a speaker verification system solely on the basis of error rates. Therefore, the EER is utilized as a metric for system evaluation [31], which is defined as follows:

$$FRR = FAR \tag{17}$$

The Detection Cost Function (DCF) [32] takes into account the varying costs linked to the two types of errors, as well as the prior probabilities of genuine speakers and impostors. The DCF is computed using the following formula:

$$DCF = C_{FRR} \times FRR \times PT + C_{FAR} \times FAR \times PI \tag{18}$$

In this context, C_{FRR} refers to the cost incurred when a legitimate speaker is incorrectly rejected, whereas C_{FAR} pertains to the cost associated with the erroneous acceptance of an impostor. PT represents the prior probability assigned to a genuine speaker, while PI indicates the prior probability associated with an impostor. By adjusting parameters such as costs and prior probabilities to align with the particular requirements of diverse recognition tasks, the minDCF threshold can be tailored to suit various recognition contexts.

D. Experimental results and analysis

1) Experiments on Training Parameters

This study investigates the performance disparities among various speaker recognition models by training on datasets from VoxCeleb1 and a subset of VoxCeleb2, utilizing four distinct network architectures: Res2Net, SE-Res2Net, SE-ProRes2Net, and the newly developed Fca-ProRes2Net. The primary objective was to ascertain the optimal number of iterations and training parameters. Figures 11 and 12 illustrate the trends in accuracy and loss function for the different network models applied to the VoxCeleb1 dataset.

The findings indicate that the advanced Fca-ProRes2Net architecture demonstrates superior convergence performance, as evidenced by its lower loss values in comparison to the other methodologies.

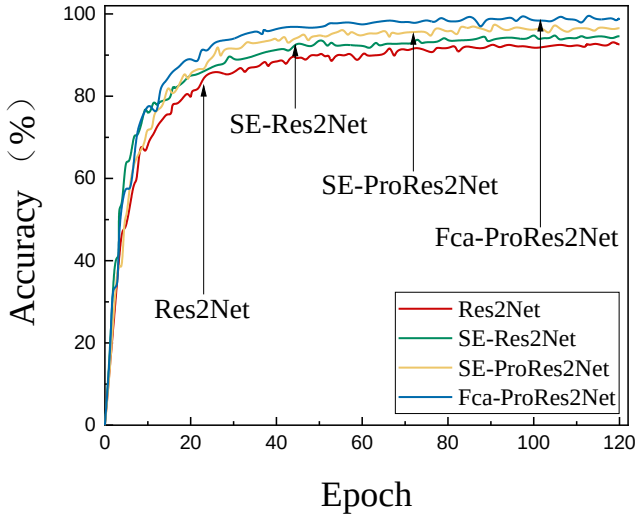

Fig. 11. Transition of Loss functions for different Networks.

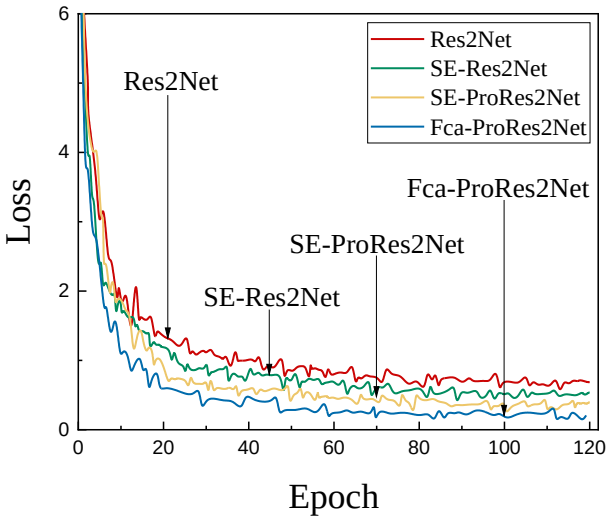

Fig. 12. Transition of accuracy for different Networks.

The SE-Res2Net and SE-ProRes2Net models, which integrate channel attention mechanisms derived from their foundational architectures, exhibit accelerated convergence rates. Specifically, both Res2Net and SE-Res2Net attain convergence in terms of accuracy and loss function within 20 to 40 iterations. The incorporation of the Res2Net module into the fully connected ProRes2Net framework, subsequent to the SE-Res2Net model, significantly enhances the receptive field and improves accuracy, thereby mitigating model error. Furthermore, substituting the SE-Net module with the Fca-Net module in the Fca-ProRes2Net network modifies the feature matrix compression strategy, allowing for the preservation of more representative features and resulting in a 2% increase in accuracy relative to SE-ProRes2Net. The experimental findings regarding loss values and accuracy suggest that the proposed network models exhibit superior feature recognition capabilities, leading to more effective and distinctive outcomes. When subjected to further training on the larger VoxCeleb2 dataset, the recognition rates and loss function values of the network developed in this study demonstrate commendable performance.

2) Identification of network ablation experiments

The experiment was carried out under optimal training conditions and was segmented into two distinct phases. The initial phase focused on training various network models utilizing the VoxCeleb1 training set, which was subsequently evaluated using the VoxCeleb1 test set. The second phase involved training different network models on the more extensive VoxCeleb2 dataset, followed by an assessment of these models using the VoxCeleb1 test set. Given that the training data and parameters remained consistent throughout the experiments, the test results for the VoxCeleb1 and VoxCeleb2 test sets are presented in Tables 1 and 2, respectively.

The data presented in Table 1 demonstrates that the enhanced Fca-ProRes2Net network architecture introduced in this research exhibits superior recognition performance. In comparison to the Res2Net, SE-Res2Net, and SE-ProRes2Net models, the Fca-ProRes2Net model achieves a reduction in EER of 0.9%, 0.4%, and 0.6%, respectively. The integration of the SE-Net module into both Res2Net and its fully connected variant allows SE-ProRes2Net to benefit from an expanded receptive field, thereby enhancing the overall performance of the model. Conversely, the fully connected Res2Net network is negatively impacted by an excess of redundant information, which slightly diminishes its performance. Furthermore, the substitution of the SE-Net module with the Fca-Net module illustrates that the Fca-Net's capacity to assign weights to channel features exceeds that of the SE-Net module, leading to improved performance of the model proposed in this study.

TABLE Ⅰ
PERFORMANCE OF VARIOUS SYSTEMS ON VOXCELEB1 TEST SET

Model	EER(%)	min-DCF
Res2Net	3.437	0.191
SE-Res2Net	2.951	0.156
SE-ProRes2Net	3.102	0.163
Fca-ProRes2Net	2.536	0.143

TABLE Ⅱ
PERFORMANCE OF VARIOUS SYSTEMS ON VOXCELEB1 TEST SET

Model	EER(%)	min-DCF
Res2Net	1.523	0.136
SE-Res2Net	1.431	0.131
SE-ProRes2Net	1.356	0.129
Fca-ProRes2Net	1.285	0.124

The findings in Table 2 further indicate that transitioning to a larger-scale dataset significantly enhances recognition performance across all network models, with the Fca-ProRes2Net model yielding comparatively superior results. As the size of the training dataset increases, the advantages associated with ProRes2Net become increasingly evident. Although SE-ProRes2Net retains some redundant information in relation to SE-Res2Net, the fully connected network demonstrates a greater capacity for information assimilation, particularly when supported by a substantial corpus of speech data. The incorporation of the Fca-Net module alters the approach to information compression, thereby contributing to the enhanced recognition capabilities of the model proposed in this research.

3) Feature extraction ablation experiment

Testing was performed on parameters MFCC, GFCC, MFCC+GFCC+ △ MFCC+ △ GFCC, and the proposed features under identical conditions. A comparative analysis of different network models indicated that the Fca-ProRes2Net network demonstrated enhanced performance compared to the

other models. As a result, the Fca-ProRes2Net network was chosen as the recognition model for the speaker recognition system. The experimental results are detailed in Table 3.

The experimental findings indicate that the proposed feature extraction method demonstrates superior performance, achieving reductions in EER of 3.7%, 6.3%, and 1.8%, respectively. In contrast to MFCC and GFCC, which are limited to singular feature representations and may fail to comprehensively capture the full range of characteristics, the proposed feature parameters exhibit enhanced representational capabilities. These parameters encompass a wider array of information, thereby enabling a more effective modeling of inter-channel correlations and resulting in a notable improvement in network performance relative to the third feature.

TABLE Ⅲ
THE EXPRESSIVENESS OF FEATURE PARAMETERS

Parameters	EER(%)	min-DCF
MFCC	5.023	0.177
GFCC	7.586	0.229
MFCC+GFCC+ △MFCC+△GFCC	3.101	0.148
Proposed	1.285	0.124

4) Model comparison

In this research, we employed established methodologies to assess performance, integrating conventional algorithms such as GMM-UBM [33] and i-vector+PLDA [34-36]. Furthermore, we investigated prominent deep learning algorithms, including TDNN [37], DNN [38], CNN [39-40], as well as fusion techniques based on embedding features [41-43]. A range of residual networks was also incorporated for comparative purposes [23].

TABLE Ⅳ
THE TEST RESULTS OF VARIOUS MODELS UNDER THREE CONDITIONS.

Model	0s-5s	5s-15s	15s-30s
GMM-UBM	61.17%	74.00%	88.00%
i-vector+PLDA	69.50%	86.10%	87.58%
TDNN-UBM	72.51%	80.19%	85.53%
i-vector+DNN	81.55%	88.80%	93.90%
MFCC+CNN	85.16%	83.14%	91.62%
t-vector+LDA	88.32%	84.80%	-
ResNet	90.37%	91. 15%	94.58%
Fca-ProRes2Net	92.18%	92.53%	95.97%

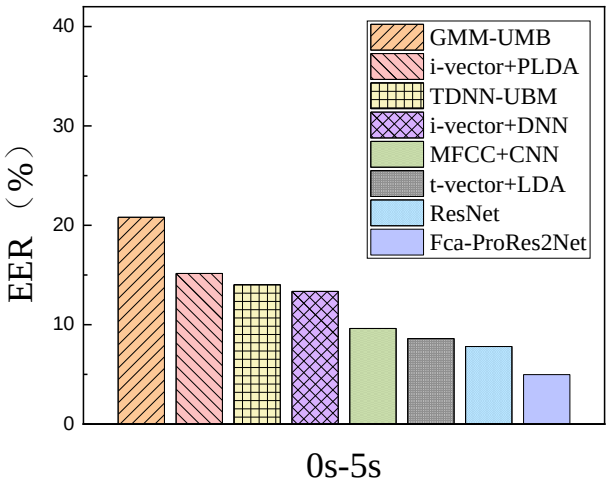

Fig. 13. EER for 0s-5s for each model.

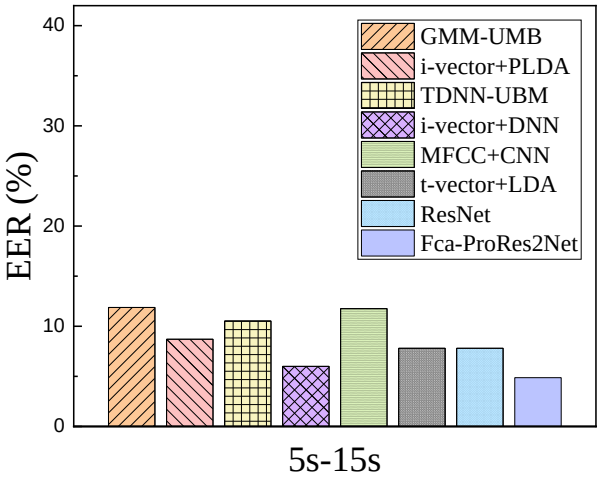

Fig. 14. EER for 5s-10s for each model.

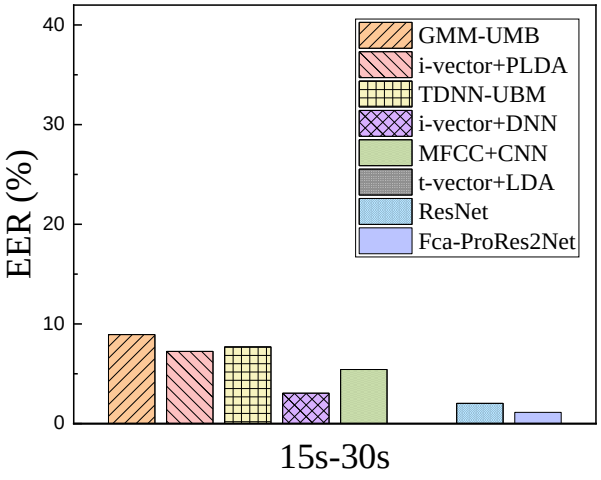

Fig. 15. EER for 15s-30s for each model.

The findings from these experiments are encapsulated in Table 4. To enhance the clarity of performance comparisons across various systems, the EER values for each system have been visually represented in Figures 13, 14, and 15.

The data presented in the charts indicates that the speaker recognition model developed in this research, which employs the Fca-ProRes2Net architecture, demonstrates a substantial enhancement in performance relative to other existing models. As the length of the extracted speech segments increases, the accuracy of this system remains consistently elevated.

Significantly, within the 0 to 5-second range, the Fca-ProRes2Net network significantly surpasses traditional recognition algorithms. In the typical speech duration of 5 to 15 seconds, this system achieves a recognition rate that is approximately 9% higher than that of the MFCC+CNN system, thereby illustrating superior feature representation capabilities compared to single-parameter approaches.

Furthermore, for extended speech segments lasting between 15 to 30 seconds, while the i-vector+DNN and ResNet systems demonstrate satisfactory recognition performance, the recognition rate of the proposed network surpasses that of these systems by approximately 2.07% and 1.39%, respectively. This improvement can be ascribed to the broader receptive field and the efficient processing of mid-to-high-frequency information.

The findings derived from the experiments indicate that the speaker recognition model formulated in this study demonstrates significant effectiveness.

V. Conclusion

This study presents a novel speaker recognition network model that integrates feature fusion and dimensionality reduction techniques based on Fca-ProRes2Net. The model begins by processing a speech segment that has undergone feature fusion, leading to the generation of a speech feature matrix through the implementation of 2DPCA dimensionality reduction. This approach demonstrates superior efficacy in preserving the original information of the features when compared to conventional single speech feature parameters, thereby facilitating the successful extraction of mid-to-high-frequency, dynamic, and static information, which in turn enhances the model's representational capacity. Subsequently, the Res2Net architecture is incorporated into the speaker recognition framework through a fully connected variant known as ProRes2Net. This architecture is proficient in representing multi-scale features with heightened granularity and accommodates a wider array of receptive field combinations. The hierarchical connection structure of Res2Net significantly expands the model's receptive field and promotes the cross-channel fusion of information across different layers, thereby augmenting the model's generalization capabilities. Additionally, the research integrates Fca-Net, a frequency-domain channel attention network, which optimizes the allocation of weights across the voiceprint feature channels, consequently improving the model's recognition performance in text-independent speech contexts. The proposed model is evaluated using the VoxCeleb dataset, demonstrating a 0.4% reduction in EER compared to the SE-Res2Net network. Moreover, it outperforms several existing speaker recognition systems that employ complex architectures, thereby significantly enhancing the model's accuracy and robustness, which has important implications for practical applications.

References

[1] Basyal, Lochan. "Voice recognition robot with real-time surveillance and automation." ArXiv Preprint ArXiv:2312.04072 (2023).

[2] Yang, Qisheng, et al. "Mixed-modality speech recognition and interaction using a wearable artificial throat." Nature Machine Intelligence 5.2 (2023): 169-180.

[3] Kwak, Il-Youp, et al. "Voice spoofing detection through residual network, max feature map, and depthwise separable convolution." IEEE Access (2023).

[4] Bai, Zhongxin, and Xiao-Lei Zhang. "Speaker recognition based on deep learning: An overview." Neural Networks 140 (2021): 65-99.

[5] Kabir, Muhammad Mohsin, et al. "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities." IEEE Access 9 (2021): 79236-79263.

[6] Hébert, Matthieu. "Text-dependent speaker recognition." Springer handbook of speech processing (2008): 743-762.

[7] Vazhenina, Daria, and Konstantin Markov. "End-to-end noisy speech recognition using Fourier and Hilbert spectrum features." Electronics 9.7 (2020): 1157.

[8] Shome, Nirupam, et al. "Speaker Recognition through Deep Learning Techniques: A Comprehensive Review and Research Challenges." Periodica Polytechnica Electrical Engineering and Computer Science 67.3 (2023): 300-336.

[9] Li, Qin, et al. "MSP-MFCC: Energy-efficient MFCC feature extraction method with mixed-signal processing architecture for wearable speech recognition applications." IEEE Access 8 (2020): 48720-48730.

[10] Hu, Wen-long, et al. "Hybrid feature extraction method of MFCC+ GFCC helicopter noise based on wavelet decomposition." Journal of Physics: Conference Series. Vol. 2478. No. 12. IOP Publishing, 2023.

[11] Hansen, John HL, and Taufiq Hasan. "Speaker recognition by machines and humans: A tutorial review." IEEE Signal Processing Magazine 32.6 (2015): 74-99.

[12] Tang, Yong, et al. "Attention based gender and nationality information exploration for speaker identification." Digital Signal Processing 123 (2022): 103449.

[13] Gao, Shang-Hua, et al. "Res2net: A new multi-scale backbone architecture." IEEE Transactions on Pattern Analysis and Machine Intelligence 43.2 (2019): 652-662.

[14] Yuan, Ruixin. "Exploring Principal Component Analysis: A Comprehensive Survey." Science and Technology of Engineering, Chemistry and Environmental Protection 1.6 (2024).

[15] Qin, Zequn, et al. "Fcanet: Frequency channel attention networks." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[16] Reynolds, Douglas A. "An overview of automatic speaker recognition technology." 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 4. IEEE, 2002.

[17] Ma, Hongbin, et al. "The research and design of identity authentication based on speech feature." PROCEEDINGS OF 2013 International Conference on Sensor Network Security Technology and Privacy Communication System. IEEE, 2013.

[18] Ding, Shaojin, et al. "Personal VAD: Speaker-conditioned voice activity detection." ArXiv Preprint ArXiv:1908.04284 (2019).

[19] Zhang, Hongxing, et al. "Feature Extraction of Speech Signal Based on MFCC (Mel cepstrum coefficient)." Journal of Physics: Conference Series. Vol. 2584. No. 1. IOP Publishing, 2023.

[20] Jing, Xinxing, et al. "Speaker recognition based on principal component analysis of LPCC and MFCC." 2014 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). IEEE, 2014.

[21] Jeong, Yongwon, and Hyung Soon Kim. "New speaker adaptation method using 2-D PCA." IEEE Signal Processing Letters 17.2 (2009): 193-196.

[22] Ding, Shaojin, et al. "Autospeech: Neural architecture search for speaker recognition." ArXiv Preprint ArXiv:2005.03215 (2020).

[23] Qiming, Ma, et al. "Intelligent Speaker Recognition Algorithm Based on SE-Res2Net." 2021 International Conference on Intelligent Computing, Automation and Systems (ICICAS). IEEE, 2021.
[24] CHEN, Zhigao, et al. "A Multiscale Feature Extraction Method for Text-independent Speaker Recognition." Journal of Electronics & Information Technology 43.11 (2021): 3266-3271.
[25] Zhou, Tianyan, Yong Zhao, and Jian Wu. "Resnext and res2net structures for speaker verification." 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021.
[26] Wang, Jiji, et al. "Application of Split Residual Multilevel Attention Network in Speaker Recognition." IEEE Access (2023).
[27] Varshaneya, V., S. Balasubramanian, and Darshan Gera. "Res-SE-Net: Boosting Performance of ResNets by Enhancing Bridge Connections." Machine Learning Algorithms and Applications (2021): 61-75.
[28] Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman. "Voxceleb: a large-scale speaker identification dataset." ArXiv Preprint ArXiv: 1706.08612 (2017).
[29] Chung, Joon Son, Arsha Nagrani, and Andrew Zisserman. "Voxceleb2: Deep speaker recognition." ArXiv Preprint ArXiv:1806.05622 (2018).
[30] Zeinali, Hossein, et al. "Non-speaker information reduction from cosine similarity scoring in i-vector based speaker verification." Computers & Electrical Engineering 48 (2015): 226-238.
[31] Doddington, George R. "Speaker recognition—Identifying people by their voices." Proceedings of the IEEE 73.11 (1985): 1651-1664.
[32] Reynolds D, Singer E, Sadjadi S O, et al. "The 2016 nist speaker recognition evaluation." MIT Lincoln Laboratory Lexington United States, 2017
[33] Akula, Aditi, Vijendra Raj Apsingekar, and Phillip L. De Leon. "Speaker identification in room reverberation using GMM-UBM." 2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop. IEEE, 2009.
[34] Li, Rongjin, et al. "Improving the Generalized Performance of Deep Embedding for Text-Independent Speaker Verification." 2018 12th IEEE International Conference on Anti-counterfeiting, Security, and Identification (ASID). IEEE, 2018.
[35] Chakroun, Rania, and Mondher Frikha. "Efficient text-independent speaker recognition with short utterances in both clean and uncontrolled environments." Multimedia Tools and Applications 79.29 (2020): 21279-21298.
[36] Poddar, Arnab, Md Sahidullah, and Goutam Saha. "Performance comparison of speaker recognition systems in presence of duration variability." 2015 Annual IEEE India Conference (INDICON). IEEE, 2015.
[37] Liu, Hui, and Longlian Zhao. "A speaker verification method based on TDNN–LSTMP." Circuits, Systems, and Signal Processing 38 (2019): 4840-4854.
[38] Guo, Jinxi, et al. "Deep neural network based i-vector mapping for speaker verification using short utterances." Speech Communication 105 (2018): 92-102.
[39] Novoselov, Sergey, et al. "Deep cnn based feature extractor for text-prompted speaker recognition." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.
[40] Jagiasi, Rohan, et al. "CNN based speaker recognition in language and text-independent small scale system." 2019 Third International Conference on I-smac (Iot in Social, Mobile, Analytics and Cloud)(I-SMAC). IEEE, 2019.
[41] Toruk, Muhammet Mesut, and Ramazan Gokay. "Short utterance speaker recognition using time-delay neural network." 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD). IEEE, 2019.
[42] Wang, Wenchao, et al. "Multiple temporal scales based speaker embeddings learning for text-dependent speaker recognition." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
[43] Zhang, Chunlei, et al. "UTD-CRSS systems for 2018 NIST speaker recognition evaluation." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.