IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 July 2023, accepted 30 July 2023, date of publication 9 August 2023, date of current version 15 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3303485

RESEARCH ARTICLE

A Method for Speaker Recognition Based on the ResNeXt Network Under Challenging Acoustic Conditions

DONGBO LIU[1], LIMING HUANG[1], YU FANG[1,2], AND WEIBO WANG[1]
[1]School of Electrical and Electronic Information, Xihua University, Chengdu 610000, China
[2]Sichuan Xihua Jiaotong Forensics Center, Xihua University, Chengdu 610000, China

Corresponding author: Yu Fang (yfang_123@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61901393, in part by the Ministry of Education "Chunhui Project" of China under Grant Z2018118, and in part by the Talented Person Projects of School under Grant Z201087 and Grant Z202088.

ABSTRACT Speaker recognition is an indispensable technology for biometrics. It distinguishes individuals based on their vocal patterns. In this paper, a joint confirmation method based on the Akaike Information Criterion (AIC) of reconstruction error (REE) and time complexity (AIC-Time joint confirmation method) is proposed to select the optimal decomposition rank of NMF. Furthermore, non-negative matrix factorization (NMF) is applied to the spectrogram to generate speaker features. The network for speaker recognition is based on Convolutional Neural Networks combining Squeeze Excitation (SE) blocks with ResNeXt, and the best combination is explored experimentally. The SE block conducts a channel-level adaptive adjustment of the feature maps, reducing redundancy and noise interference while improving feature extraction efficiency and accuracy. The ResNeXt convolutional neural network concurrently executes multiple convolutional kernels, acquiring richer feature information. The experimental results demonstrate that compared to speaker recognition based on Gaussian mixture models (GMM), Visual Geometry Group Network (VGGNet), ResNet, and SE-ResNeXt using spectrograms, this method increases the accuracy by an average of 5.8% and 16.24% under the overlaid of babble and factory1 noise with different signal-to-noise ratios, respectively. In the short speech test, the test set is short speech of 1s and 2s, and the noise is superimposed. Compared with other methods, the recognition rate is increased by an average of 8.67% and 11.72%, respectively.

INDEX TERMS Speaker recognition, non-negative matrix factorization, ResNeXt, squeeze-excitation, akaike information criterion.

I. INTRODUCTION

Speaker recognition is using information contained in speech features to achieve speaker classification. These features offer a vast range of data about the speaker, including their gender, emotional state, dialect, etc. The fundamental frequency of speech can distinguish between males and females, with males typically having a lower pitch and females having a higher pitch. The average speech rate, pitch change and resonance peak are closely related to the speaker's emotion and dialect [1]. The sets of features fall into different categories: time, frequency, and combinations [2]. Speaker recognition can be performed accurately through these extensively researched features. The traditional speaker identification method requires less training data and computing resources during the training phase. During the recognition phase, the extracted features are matched against the model library for identification. The speaker with the highest probability is selected and recognized as the target speaker [3]. While conventional speaker recognition methods perform well under ideal circumstances, real-world scenarios have distracting factors such as speech duration, background noise, and the speaker's physical condition. Nonetheless, using deep neural networks (DNN) [4], speaker recognition

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

can overcome these difficulties and present superior performance.

Speaker recognition technologies currently fall into two categories: traditional statistical models and computer-based deep learning models. Statistical modeling approaches such as GMM [5], Hidden Markov models (HMM) [6], i-vectors [7], and joint factor analysis (JFA) [8] are widely utilized for speaker recognition. Statistical model-based recognition requires pre-modeling speaker features and classification based on the existing model during recognition. However, this approach must be improved in effectively handling recognition tasks under complex conditions. Factors such as background noise, shorter speech duration, and lower speech volume impact recognition rates, making it challenging. With the emergence of deep networks and their excellent performance in the ImageNet competition [9], these networks have been successfully applied in speaker sentiment recognition and recognition with equally impressive performance.

Traditional statistical and computer-based deep learning models share similarities in acquiring speech features. The unique acoustic characteristics of the individual, whether in the time domain or frequency domain, form the basis of the features, which are then encoded in a series of continuous speech signals. Feature acquisition for speaker recognition is becoming more advanced, with the combination of time-frequency domains leading to more distinct features.

Some commonly used speaker features, such as Mel-frequency cepstral coefficients (MFCC) [10], linear frequency cepstral coefficients (LPCC) [11], speech spectrograms, and achieve excellent results. In [12], the method uses NMF to decompose spectrograms, using the resulting feature matrix as model input. The experimental results compared the performance of this feature with other features, including in noisy environments and short speech.

Speaker recognition technology research began in the late 1940s. However, due to limitations in technology during that time, feature extraction and recognition accuracy were limited, and the conditions for recognition were also challenging. With the advent of Dynamic Time Warping (DWT) [13] and Vector Quantization (VQ) [14] technologies, speaker recognition greatly improved. Later, GMM and Gaussian Mixture Model-universal Background Model (GMM-UBM) [15], due to their flexibility, simplicity, effectiveness, and robustness, quickly became the mainstream methods and are still utilized today. The artificial neural network approach that emerged afterward has been utilized and improved up to the present day. With the rapid development of computer technology in the 21st century, the computational requirements for neural networks have been satisfied, and an increasing number of DNN models have been proposed and widely utilized in speaker recognition, further enhancing speaker recognition performance [16].

NMF is a widely used matrix factorization technique in machine learning and data analysis designed for feature extraction from non-negative matrices [17]. Its fundamental objective is to factorize a non-negative matrix V into two non-negative matrices W and H, i.e., $V_{M\times N} \approx W_{M\times r}H_{r\times N}$. In this decomposition, the matrix V represents the original data, while matrices W and H represent the latent feature matrices that capture the most important patterns and structures in the data. Crucially, unlike Singular Value Decomposition (SVD) [18] and Principal Component Analysis (PCA) [19], NMF only produces non-negative factors, making it particularly useful for data with naturally occurring non-negative values, such as spectrogram data. Lee et al. [20] utilized NMF to decompose the spectrogram of indoor noise, and the resulting feature matrix was identified utilizing a convolutional neural network (CNN). In [21], NMF is used for source separation of speech and music, also the most widely used field of NMF. The goal of the NMF decomposition process is to find matrices W and H that accurately represent the original data V. This is typically achieved by specifying the target rank of the W and H matrices, which determines the number of features to be extracted. The main objective of the NMF algorithm is to minimize the difference between V and its approximation WH based on distance or dissimilarity measures such as the Frobenius norm, Kullback-Leibler divergence, or Euclidean distance [22]. Nonetheless, the choice of the decomposition rank of NMF is usually selected through empirical values and experimental results, and there is no fixed method to set an optimal decomposition rank.

AIC, as a statistical metric, considers both the goodness of fit and computational complexity. Developed by the Japanese statistician Hirotugu Akaike in 1974 [23], the principle behind AIC is to identify the model that best conforms to the observed data while minimizing the number of parameters. This principle aligns with intending to determine the optimal decomposition rank, which minimizes the computational complexity in the decomposition and is the most appropriate. Based on the maximum likelihood estimation principle, AIC estimates probability distribution parameters using observed data. In [24], Zhiwen Zhao et al. proposes an empirical likelihood-based Akaike information criterion (AIC) to select variables for the generalized random coefficient autoregressive model. In [25], AIC is used to select the best decomposition rank of NMF, and AIC is directly based on the sum of squares of the difference between the original matrix and the reconstructed matrix.

DNN has shown outstanding performance in classification and recognition, and many researchers have applied it to speaker recognition. Variani et al. [7] utilized DNN to extract feature vectors from spectrograms, which they named d-vectors. The final output represents the speaker's d-vector. Snyder et al. [26] proposed an x-vector based on Time Delay Neural Network (TDNN), which can capture long-term speaker features. Nguyen An et al. [27] used Visual Geometry Group Network (VGGNet) and ResNet networks that incorporated a self-attentive structure to handle variable-length speech segments, enabling the learning of

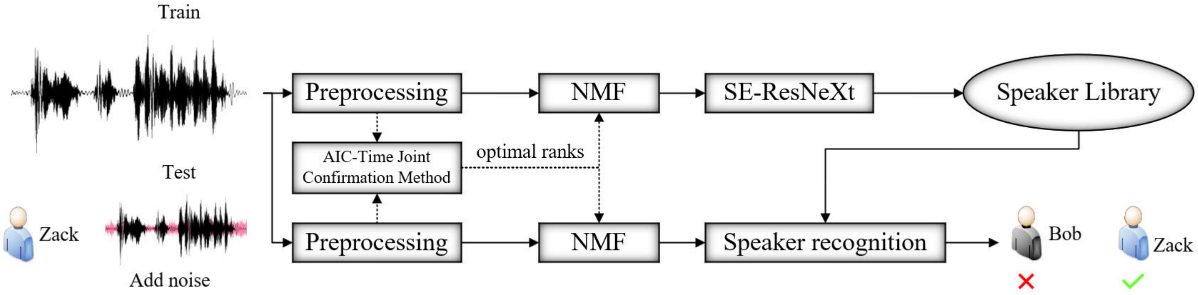

FIGURE 1. **Speaker recognition framework. (The pink waveform in the test is the overlaid noise. The dotted line in the figure indicates the first calculation.)**

speaker characteristics from different aspects of the input sequence. Previous studies have shown that including SE blocks in a model effectively improves classification performance. JAVIER NARANJO-ALCAZAR et al. [28] achieved superior performance in acoustic scene classification by incorporating SE blocks into their model, outperforming the baseline model. Bingzhang Zou et al. [29] achieved impressive classification results in radio signal recognition using a residue fusion network combined with SE blocks.

This paper proposes a method for jointly determining the optimal decomposition rank of NMF based on REE-based AIC and time complexity. The speaker recognition model is based on a network model combining SE blocks and ResNeXt. The article overview is presented below. Section II introduces the overall process and main contributions of the speaker recognition framework proposed in this article. Section III, part A introduces the preprocessing of speech data and the extraction of the spectrogram in detail; part B introduces the feature matrix obtained by the NMF decomposition of the spectrogram and the selection of the optimal rank of NMF by AIC-Time joint confirmation method. Section IV describes the architecture of the neural network model. Section V presents the experimental setup on the AISHELL-1 [30] dataset and the performance comparison of the proposed method with other methods. Finally, Section VI gives the conclusion.

Our main contribution is to select the optimal decomposition rank of NMF using REE-based AIC and the joint time complexity method and to experimentally select the optimal position in the combination of SE block and ResNeXt for speaker recognition.

II. SPEAKER RECOGNITION FRAMEWORK

The proposed speaker recognition method in this article mainly consists of two components. First, use the AIC-Time joint confirmation method to confirm the optimal decomposition rank of NMF in the preprocessing stage, and then use NMF to decompose the spectrogram to obtain the feature matrix. This is a crucial step to eliminate redundant and irrelevant information from the spectrogram and extract only the requisite features required for speaker recognition. Secondly, a deep neural network architecture consisting of ResNeXt and SE blocks trains the speaker recognition model. The optimal form of the combination was also found in follow-up experiments. Combining these two modules lets the network capture deep time-frequency characteristics crucial for speaker recognition. The ResNeXt module enables more efficient and accurate feature extraction, while the SE block adjusts channel-wise feature responses to enhance the quality of extracted spectrogram features.

The proposed speaker recognition method is shown in FIGURE 1. The speech signal is preprocessed in the training phase, and the spectrogram is output. Then use the AIC-Time joint confirmation method to select the optimal decomposition rank of NMF under the current data set. Apply NMF to decompose the spectrogram and generate the feature matrix. The ResNeXt network combined with SE blocks further processes the feature matrix to extract deep time-frequency features and train the speaker model. The proposed method has been evaluated on datasets with different background noise levels, showing good classification performance.

Remarkably, experimental results show that the proposed method surpasses other state-of-the-art techniques in terms of recognition accuracy, demonstrating its effectiveness in terms of the robustness of speaker recognition systems in noisy environments.

The main contribution of this article is the joint determination of the optimal NMF decomposition rank using AIC based on matrix reconstruction error and time complexity and applying NMF to extract speaker features from speech spectrograms. The speaker recognition task is accomplished by training the speaker model using the SE-ResNeXt model. The reasons for using NMF to extract features are as follows:

1) To reduce the dimensionality of input features, decrease the computational load of the network, and improve the model efficiency.

2) To reduce feature redundancy and achieve the goal of feature extraction.

3) Separate signal sources benefit the model recognition of speech and noise signals.

On the other hand, residual networks have more layers and computational complexity than other convolutional neural networks, which can extract more profound speaker features from the input. Group convolution in ResNeXt [31] achieves lower error rates than other residual networks with the exact computational cost. Adding SE blocks enhances

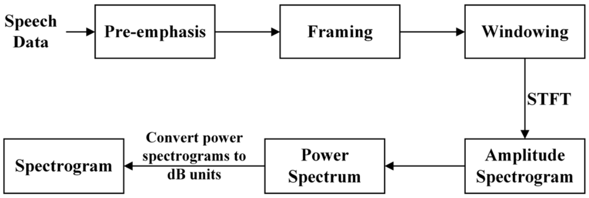

FIGURE 2. Speech data pre-processing. Input data pre-processing output to spectrogram.

valuable features in the network and weakens the weights of less influential features. Therefore, the combination of ResNeXt and SE blocks [32] has been chosen as the neural network model for speaker recognition. The final experimental results confirm the effectiveness of this method in speaker recognition.

III. SPEAKER FEATURE EXTRACTION

A. ACQUIRING SPECTROGRAMS

The preprocessing of speech data is carried out to acquire the spectrogram information of the speaker. Preprocessing involves pre-emphasis, framing, windowing, and short-time Fourier transform (STFT) [33], amongst other steps. Pre-emphasis, similar to a high-pass filter, enhances the speech's high-frequency components, improving the speech signal's intelligibility. The pre-emphasis formula is as follows: the pre-emphasis coefficient α ranges from 0 to 1. In this study, α is set to 0.97.

$$H(z) = 1 - \alpha z^{-1} \tag{1.1}$$

The dataset's sampling frequency is 16kHz, and the experiment's frame length is 25ms. This implies that each frame comprises 400 samples, while the frameshift is at 10ms, which equates to 160 samples. A Hamming window [33] function is used to address signal continuity after framing, while its formula is as shown below,

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \\ 0, n = \text{otherwise} \end{cases} \tag{1.2}$$

where n is the frame signal after framing and N is the size of the frame signal. The process of obtaining a spectrogram for speech data preprocessing is shown in FIGURE 2.

The resulting spectrogram FIGURE 3 contains a wealth of information, including time and spectral data. It assumes a two-dimensional shape, where it possesses three dimensions of information. The x-axis corresponds to the temporal dimension, whereas the y-axis denotes the frequency domain. The bright and dark hues depicted on the chart convey the amplitude dimension's information. The brighter colorations represent stronger intensities, whereas the darker hues portray weaker intensities.

Despite this, there still exists redundant information in the spectrogram. Moreover, that irrelevant information might significantly impact the recognition performance of the model. In the presence of noise, speaker recognition accuracy would be seriously affected, which calls for using NMF to eliminate redundant information in spectrograms and extract relevant features for separation noise and speaker signals. The experimental results show that the feature matrix obtained by NMF decomposing the spectrogram is better than using the spectrogram directly as a speaker feature, especially in the environment of noise and short speech.

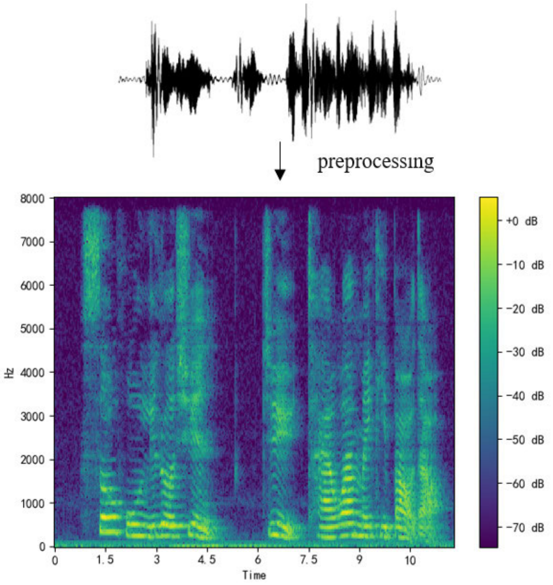

FIGURE 3. Spectrogram of speech. The depth of the color indicates the intensity magnitude of a given frequency component.

B. AIC-TIME JOINT CONFIRMATION METHOD FOR OPTIMAL NMF RANK

In NMF, the feature matrix and coefficient matrix contain different aspects of the original matrix information. The feature matrix represents the original matrix's characteristics and contains the data's original mode and structure. Each column of the feature matrix corresponds to a basic feature of the original matrix, and these basic features can represent all the data points of the original matrix. The coefficient matrix represents the weighted combination of each data point of the original matrix on the feature matrix, including the importance and contribution of the data points in the original matrix [34]. Therefore, the feature matrix and coefficient matrix contains different information about the original matrix, with the feature matrix describing the structural characteristics of the data and the coefficient matrix describing the contribution of the data points to different structures [35].

Usually, W is called a ''basic matrix'' or ''feature matrix'', and H is called a ''coefficient matrix'' or ''activation matrix'', where all elements of $W = \{f_1, f_2 \cdots, f_M\}$ and $H = \{c_1, c_2 \cdots, c_N\}$ are non-negative numbers.

By applying NMF to the feature extraction and data dimensionality reduction of the speaker's spectrogram, the proposed approach in this paper effectively eliminates the redundant information contained in the spectrogram, significantly reducing the data volume of the model. This is a crucial step for speaker recognition, as it reduces computational complexity and enhances the discriminative power of the model. Moreover, NMF can also be used for unsupervised learning tasks such as clustering, which can be helpful in applications such as image segmentation and document clustering [36]. In summary, NMF is a powerful technique that can be applied in various data-driven applications to extract relevant features, reduce dimensionality, and enhance the discriminative power of models.

Table 1 gives the basic decomposition steps of NMF.

TABLE 1. Procedure of NMF.

Input: Spectrogram (Non-negative matrix V^{+}_{M*N}), M and N are the dimensions of the input spectrogram.

Output: The feature matrix W and the coefficient matrix H.

1. Determine the dimension and number of the feature matrix.
2. Initialize feature matrix W and coefficient matrix H.
3. According to the objective function, iterative optimization is performed by alternately updating the feature matrix W and the coefficient matrix H.
4. Stop iterating until the objective function converges or reaches the preset number of iterations.
5. The feature matrix W and the coefficient matrix H are output from the decomposition.

V^{+}_{M*N} in the table is the matrix representation of speech signal Y(n). M represents the frequency resolution of Y(n), and N represents the number of speech frame samples of Y(n).

In the preprocessing section of this paper, the speech signal is subjected to non-negative matrix decomposition on the spectrogram. The feature matrix consists of columns representing the basic spectral composition of local spectral structures in the spectrogram, encompassing critical characteristics specific to the speaker. The decomposition rank R of the feature matrix plays a crucial role in signal analysis by determining the number of basic spectral components into which the spectrogram is decomposed. On the other hand, the coefficient matrix comprises the sound intensity of distinct segments in the spectrogram, with each element representing the weight corresponding to a basic spectral composition. The present paper utilizes the feature matrix as the input of the model network to gain insight into the speaker's essential voice features.

There are four standard methods for initializing non-negative matrix factorization. These methods include random initialization, random NMF initialization, SVD initialization, and random initialization with additive noise [37]. However, since the data in this experiment consists of continuous signal data, it is recommended to utilize the SVD initialization method. This method involves decomposing the spectrogram data matrix using SVD, which allows the initial value to accurately reflect the primary components of the spectrogram [38].

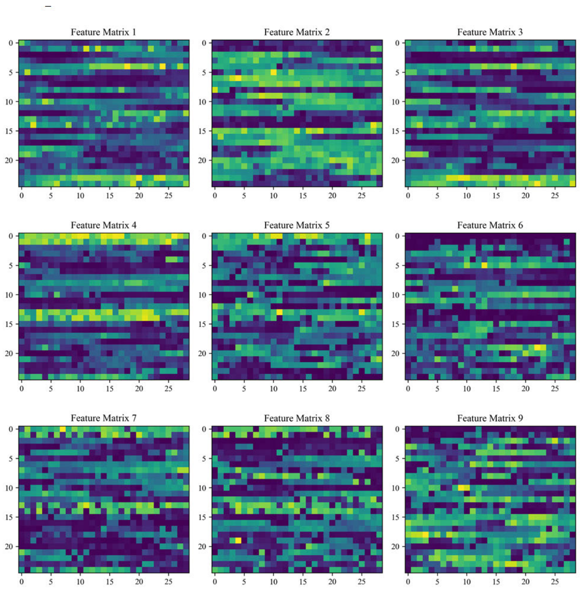

FIGURE 4. The 9 components of the W feature matrix.

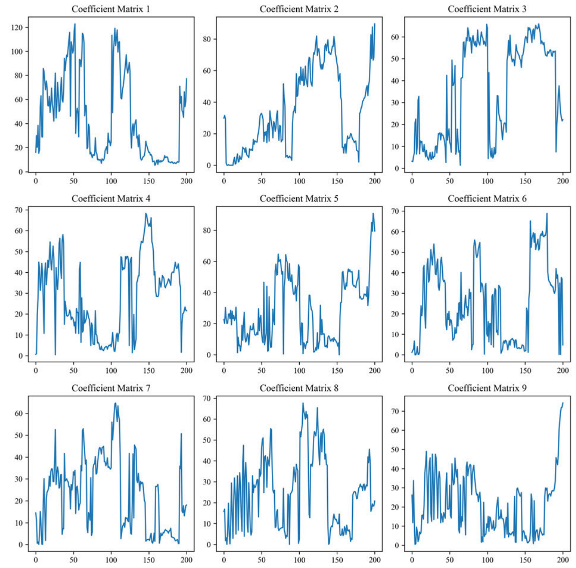

FIGURE 5. The 9 components of the H coefficient matrix.

The W_{M*r} feature matrix and the H_{r*N} coefficient matrix are updated through alternate iterations, and the iterative update rule is shown in formula (1.3), and (1.4). The objective function adopts Euclidean distance, which is more suitable for dealing with continuous signal data. The objective function is shown in the formula (1.5). where V is the original matrix and V' is the reconstructed matrix.

$$W \leftarrow W \frac{[V/(WH)]\,H^{T}}{1_{K\times N}H^{T}} \tag{1.3}$$

$$H \leftarrow H \frac{[V/(WH)]\,H^{T}}{W^{T}1_{K\times N}} \tag{1.4}$$

$$\min \left\|V - V'\right\|^{2} = \sum\nolimits_{ij}\left(V_{ij} - V'_{ij}\right)^{2} \tag{1.5}$$

The feature matrix and coefficient matrix obtained from the speaker's spectrogram by NMF are shown in FIGURE 4, FIGURE 5. The choice of decomposition rank depends on the original matrix's specific structure and data characteristics.

Choosing an appropriate decomposition rank is crucial to the decomposition results, but there is currently no fixed rule to confirm the best decomposition rank.

The REE [39] be a metric for evaluating the decomposition quality. It measures the difference between the original matrix V and the reconstructed matrix V'. Evaluated by the reconstruction error, the decomposition quality is effectively measured with lower values indicating accurate representations of the original data by the reconstruction matrix. The formula for reconstruction error is represented below.

$$REE = ||V - WH||_F = \sqrt{\sum\nolimits_{i,j} (V_{ij} - \sum\nolimits_{k=1}^{r} W_{ik}H_{kj})^2} \tag{1.6}$$

In the formula, the error value is measured by the Frobenius norm, which defines the square root of the sum of the square differences between the corresponding elements of the original matrix V and WH.

$$RSS = \sum\nolimits_{i,j} (V_{ij} - \sum\nolimits_{k=1}^{r} W_{ik}H_{kj})^2 \tag{1.7}$$

It is important to note that the reconstruction error is a crucial indicator to measure the decomposition quality in practice. However, the decomposition rank corresponding to the lowest reconstruction error value sometimes does not translate to the best choice. Other factors need to be considered when choosing the ideal decomposition rank. In this experiment, AIC is used based on the reconstruction error sum of squares (RSS) to find the ideal decomposition level.

The formula for AIC in this experiment is presented below.

$$AIC = 2k - 2\ln(L) \tag{1.8}$$

where k is the number of parameters in the model and L is the likelihood function. In the decomposition experiment, k is the penalty term of matrix decomposition. Likelihood functions are useful for measuring the goodness of fit of a given model and data. In NMF, the observed data matrix V can be fitted by the product of non-negative matrices W and H, i.e., $V \approx WH$. The problem of non-negative matrix factorization can thus be expressed as a maximum likelihood estimation problem. By maximizing the likelihood function, the optimal W and H matrix values of the reconstruction matrix V' can be finally found. The likelihood function is obtained by assuming that the observed data are independent and identically distributed, and the assumed distribution must be consistent with the real distribution.

By assuming that the RSS of the NMF obeys a normal distribution with a mean of 0 and a variance of σ^2, the value of σ^2 can be estimated using the maximum likelihood method. The variance σ^2 represents the noise or variation in the original data that cannot be explained after NMF. In the case where the NMF parameter and the estimated value of σ^2 are known, the calculation of the likelihood function is the joint probability density function of the observed data and adopts the standard distribution form with a mean of 0 and a variance of σ^2, so the joint probability density function of the multivariate normal distribution is used in the calculation of the likelihood function. The calculation deduction of the likelihood function L of the non-negative matrix factorization is as follows.

$$L = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)] \tag{1.9}$$

In the above Gaussian probability density function [39], the d-dimensional column vector of $X = [x_1, x_2, \ldots, x_d]^T$, $\mu = [\mu_1, \mu_2, \ldots, \mu_d]^T$ is the d-dimensional mean vector; $\sum$ is the $d \times d$ dimension Covariance matrix, Σ^{-1} is the inverse matrix of $\sum$, $|\Sigma|$ is the determinant of $\sum$. In NMF, set $V \approx WH$, and add Gaussian noise ε_{ij}, then each data point in V can be represented by

$$V_{ij} = \sum\nolimits_{k=1}^{r} W_{ik}H_{kj} + \varepsilon_{ij} \tag{1.10}$$

where ε_{ij} is Gaussian noise with mean 0 and variance σ^2. Substitute into formula (1.9) to get

$$L = (2\pi)^{\frac{-n}{2}}|\Sigma|^{\frac{-1}{2}} \exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)] \tag{1.11}$$

where $\sum$ is the $n \times n$ -dimensional covariance matrix, and the covariance matrix is a symmetric matrix. On the premise that all variables are independent, the off-diagonal elements of the covariance matrix are 0, so $|\Sigma| = \sigma_1^2\sigma_2^2\sigma_3^2\ldots\sigma_n^2 = \sigma^{2n}$ and the inverse matrix for $\Sigma^{-1} = diag(\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots, \sigma_n^2)^{-1}$. Simplified formula (1.11).

$$L = \left(2\pi\sigma^2\right)^{\frac{-n}{2}} \exp[-\frac{1}{2}(V - WH)^T\Sigma^{-1}(V - WH)] \tag{1.12}$$

Computes the maximum likelihood estimate of the variance σ^2 using the RSS and the number of observations $m \times n$.

$$\sigma^2 = \frac{RSS}{m \times n} \tag{1.13}$$

$$AIC = 2(mr + rn) - 2\ln(L) \tag{1.14}$$

Finally, the calculated likelihood function is substituted into formula (1.8), yielding the AIC values at different decomposition ranks. The penalty term of AIC, which is $2(mr + rn)$, punishes decomposition ranks with more parameters, and the penalty term increases as the decomposition rank increases. The penalty term's purpose is to balance the decomposition rank and computational complexity. The AIC value is a relative measure of the decomposition quality, and a lower AIC value indicates a better fit between the decomposed matrix and the original data [40].

The AIC values of NMF at different decomposition ranks are shown in FIGURE 6. However, it is crucial to select the optimal decomposition rank by considering both the quality of the decomposition and the computational complexity.

It is crucial to note that a single choice cannot determine the optimal decomposition rank of NMF. The choice depends not only on the AIC value (lower is better) but also on the experimental analysis's specific research questions and data requirements. More considerable decomposition ranks

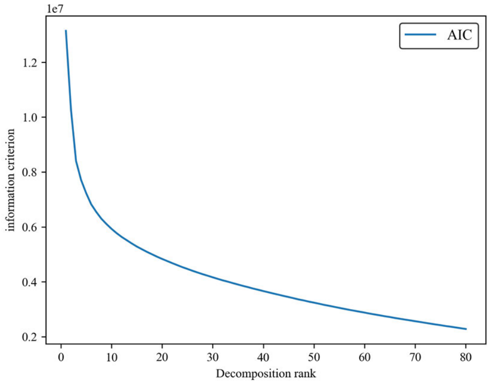

FIGURE 6. AIC values for different decomposition ranks in NMF.

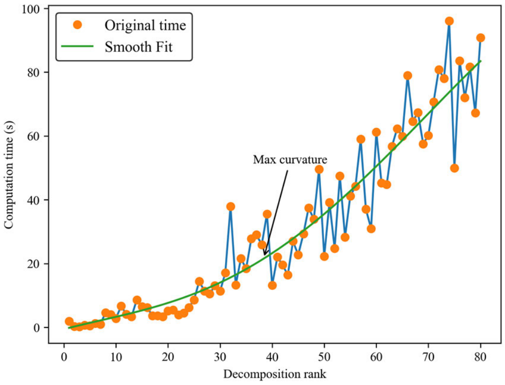

FIGURE 7. Calculation time for different decomposition ranks (consistent computing environment).

provide better outcomes in some cases, but this significantly increases the computational cost and even overfitting. Therefore, it is necessary to select a balance point between the goodness of the decomposition and computational complexity. The current goal of the experiment is to achieve excellent decomposition performance at a low computational time.

Furthermore, the experiment presents the NMF computation time for different decomposition ranks, as displayed in FIGURE 7. The orange dots represent the original computation time, and the green line is the smooth curve of the entire computation time. The smooth curve visually presents the variation of computation time as the decomposition level increases. As seen in the figure, the decomposition time significantly increases as the decomposition rank increases.

Finally, The experiment seeks a balance between computational complexity and decomposition excellence. The minimum-maximum normalization is applied to linearly map the AIC value and the computation time to a value range from 0 to 1, which helps preserve the relative relationship of the original data and simplifies the selection of the optimal decomposition rank. FIGURE 8 illustrates the normalized AIC curve and the smooth computation time curve.

FIGURE 8 exhibits the curves of AIC and computation time normalized to the same scale, and the red circles on the graph represent the intersection points of the curves. The significance of the intersection point is that it indicates the balance point between model complexity (measured by AIC) and computation time. The current optimal decomposition rank is the intersection point corresponding to the decomposition rank (rank=30). In the subsequent NMF calculations, the experiment will apply a factorization rank of 30.

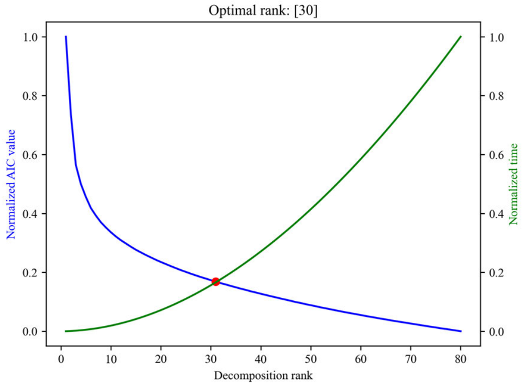

FIGURE 8. Normalized AIC and smoothed time curves.

IV. NEURAL NETWORK MODEL

A. RESNEXT MODEL STRUCTURE

ResNeXt is an aggregated residual network that builds on the residual network (ResNet) initially proposed by Saining Xie et al. [41]. ResNeXt offers higher accuracy than ResNet without increasing the network's computational cost. It can be seen as a combination of ResNet and Inception [42], with a more straightforward design and lower computational requirements than Inception. ResNeXt also adopts the 'splitting-transforming-aggregating' strategy [41], introducing cardinality to control the group size. The 'splitting' step divides the input high-dimensional features into low-dimensional features, the 'transforming' step performs a linear transformation of low-dimensional features, and the 'aggregating' step combines the low-dimensional features from all groups.

Compared to ResNet, the essence of ResNeXt lies in block convolution, which is a method between ordinary convolution and separable convolution. Initially, AlexNet [43] used separate convolution. However, it had to split the convolution operation into two GPUs for training due to GPU limitations, and the training parameters were not shared. The individual branch structure and parameter settings of group convolution are the same, and fewer parameters are required than with ordinary convolution.

The structure of the ResNeXt group convolution is shown in FIGURE 9, where N represents the total number of convolution kernels and C represents the number of group convolutions.

B. SQUEEZE EXCITATION STRUCTURE

The SE block was introduced by Jie Hu et al. [44] in 2017. It is a method for modeling the correlation between feature channels in CNN by selectively amplifying important

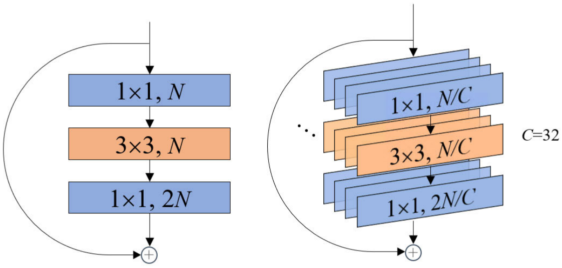

FIGURE 9. Aggregated residual transformations.

features and suppressing less relevant features, thereby enhancing the accuracy of the network [28].

FIGURE 11 shows the combination of the SE block and CNN, with the SE block running in parallel with the output of the network structure. The working principle of the module can be summarized as follows: First, a global average pooling operation is performed on network's output, referred to as the squeeze process, and is shown in the formula (1.15).

The output $1 \times 1 \times C$ data then undergoes two fully connected layers to learn the correlation between channels, referred to as the excitation process. The first fully connected layer performs dimensionality reduction and is activated by ReLU. The second fully connected layer restores the original dimension, and the sigmoid function is applied to limit the data to 0 to 1. The resulting value is multiplied by the network output to generate the input for the next level.

$$z_C = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{1.15}$$

The original feature map is $H \times W \times C$, where H is the height, W is the width, and C is the number of channels. The Squeeze step uses global mean pooling to average the features inside the channel, thus obtaining the global description features.

$$X_c^* = s_c u_c \tag{1.16}$$

X_C^* is the product of the SE block acting on the network, completing the recalibration of the original feature in the channel dimension. Here, u_c is the original output of the network structure, s_c and is the scalar output of the SE block.

C. ESTABLISHING THE SE-RESNEXT NETWORK

Nowadays, CNN has matured in the application of bioinformatics [45]. Common CNN comprises convolutional layers responsible for feature extraction and pooling layers, which reduce feature dimensions.

The speaker recognition method proposed in this article is based on the SE-ResNeXt network, which differs from other speaker recognition models as it has a complex and deep structure. The input data with an initial size of 224∗224 undergoes the residual network structure and is then multiplied with the convolutional output of the SE block. The output is subjected to batch normalization (BN) and Rectified Linear Unit (ReLU) activation function before being fed into the next layer. Finally, the results are produced via global average pooling and a fully connected layer. The network structure is illustrated in FIGURE 10.

Table 2 provides parameter information for each layer of the network. The number 32 in the table denotes the number of groups, wherein the number of groups in conv2 to conv5 is 32. The following 4d indicates the number of convolution kernels for each group in a sequence of group convolutions.

V. EXPERIMENTS

A. EXPERIMENTAL DATA

The AISHELL-ASR0009-OS1, an open-source speech database, is utilized in this experiment, and the total recording duration is 178 hours. The recordings were made in a noiseless environment using three distinct devices: a high-fidelity microphone (with audio down-sampling at 16kHz, and a resolution of 44.1kHz, 16-bit), an Android phone (16kHz, 16-bit), and an iOS phone (16kHz, 16-bit). The recording featured 400 speakers from various regions with different Chinese accents.

The experiment uses speech data from 200 speakers for training and testing. To maintain the uniformity of each speaker's data, the initial 280 data are utilized for training, while the last 30 are used for testing. The training data is divided into a network training set and a validation set at an 8:2 ratio. In the experiment, ambient noise from the NoiseX-92 noise library is deployed, and the noise is mixed with the speech based on SNR levels. The tabulation of the specific use of experimental speech data is shown in Table 3.

B. EXPERIMENTAL EVALUTION INDEX

Speaker recognition is a complex process that involves analyzing a paragraph to identify the speaker. It is important to note that the experiment uses a closed-set data set, meaning that the speaker being tested belongs to a set of known candidates while the recognized speaker is also from a set of known speakers. This experimental design provides a unique challenge for accurately measuring speaker recognition performance.

Throughout the experiment, the standard performance index for speaker recognition is the correct recognition rate. This index is ascertained by determining the proportion of speakers accurately identified within the speech sample, utilizing the formula TNC/TNR. Here, TNC represents the total number of correct recognitions, and TNR represents the total number of recognitions. This performance index results in a more precise and comprehensive evaluation of the speaker recognition model's efficacy in multi-choice settings. This assessment is indispensable in enhancing the model's performance to cater to the diverse requirements of various applications. The formula is as follows.

$$\text{Accuracy} = \frac{TNC}{TNR} \times 100\% \tag{1.17}$$

C. EXPERIMENTAL PLATFORM

The Pytorch deep learning library in the Python library is utilized in the implementation of this experiment, with the

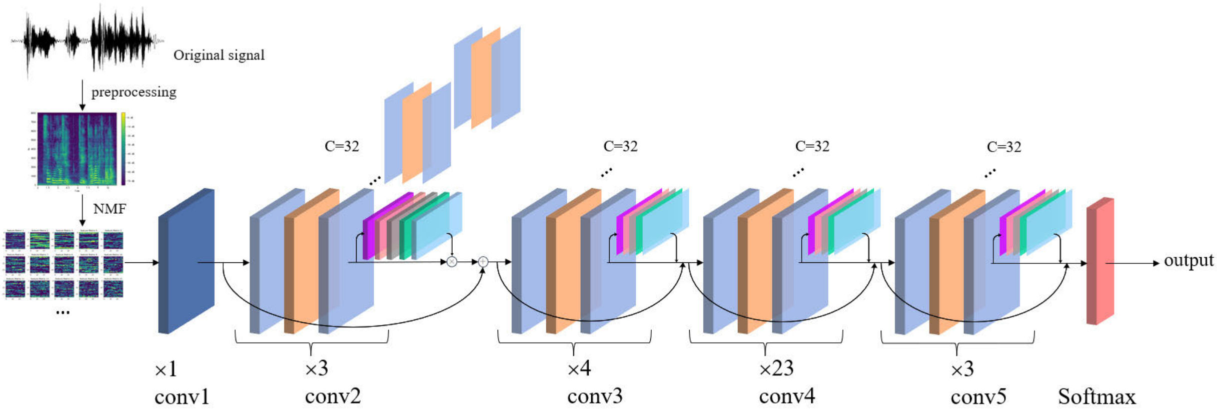

FIGURE 10. Overall structure of SE-RESNEXT network.

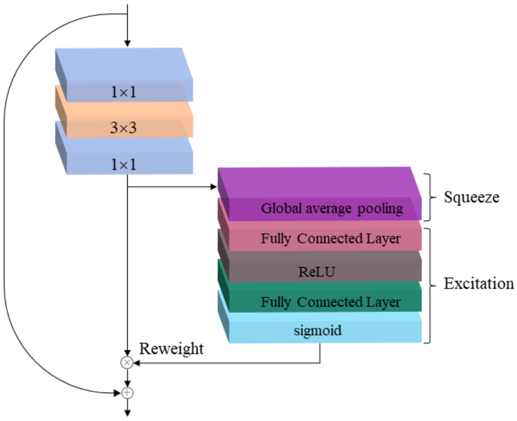

FIGURE 11. Overall structure of SE-RESNEXT network.

computer GPU tapped into for intensive training assistance. The computer specifications utilized in the experiment are as follows: an AMD Ryzen 9 5950 × 16-core processor, 128GB of running memory, and an NVIDIA GeForce RTX3090 24GB graphics card. In model training, Adam is employed as the model optimizer to update network parameters, with the learning rate set to 0.001. The model was trained for 60 batches, with 64 small batches for training and 8 small batches allocated for validation. Training and validation data are shuffled to avoid data input sequence influence on the network, increase randomness, and improve network generalization performance.

Since speaker recognition is a classic classification problem, the network's loss function is cross-entropy, which characterizes the proximity between the actual and anticipated output. In the experiment, the network's parameter and training-related settings are refined to achieve optimal settings.

D. EXPERIMENTAL COMPARISON

During the experiment, strict accuracy evaluation was implemented for every selected model to guarantee model reliability and validity. Input features were meticulously chosen to incorporate critical information and exclude needless noise. Furthermore, the neural network architecture for each model was conscientiously crafted to balance model complexity and efficiency. The experiment was repeated numerous times to confirm the validity of the results, and the data analysis process was meticulously executed to eliminate any potential bias or confounding factors. As a result, the experimental settings were rigorous and comprehensive, reassuring the validity and reliability of the outcomes.

TABLE 2. SE-ResNeXt-101 template.

stage	output	SE-ResNeXt-101(32×4d)
conv1	112×112	7×7,64, stride 2
conv2	56×56	3×3 max pool, stride 2
		$\begin{bmatrix} 1\times1,128 \\ 3\times3,128,C=32 \\ 1\times1,256 \\ fc,[16,256] \end{bmatrix}\times 3$
conv3	28×28	$\begin{bmatrix} 1\times1,256 \\ 3\times3,256,C=32 \\ 1\times1,512 \\ fc,[32,512] \end{bmatrix}\times 4$
conv4	14×14	$\begin{bmatrix} 1\times1,512 \\ 3\times3,512,C=32 \\ 1\times1,1024 \\ fc,[64,1024] \end{bmatrix}\times 23$
conv5	7×7	$\begin{bmatrix} 1\times1,1024 \\ 3\times3,1024,C=32 \\ 1\times1,2048 \\ fc,[128,2048] \end{bmatrix}\times 3$
	1×1	Average pool, 1000-d fc, Softmax

TABLE 3. The distribution of the AISHELL-1 data set in the experiment.

	Speaker	Utterances	Total data	Proportion
Training data	200	280	56000	90%
Test data		30	6000	10%

In the experiment, to measure the performance of different methods and input features on the dataset. In Table 4, the first

TABLE 4. Comparison of each method structure.

Method	Input feature	Model
MFCC-GMM	MFCC	GMM
VGG16	Spectrogram	VGG16
ResNet	Spectrogram	ResNet 50
SE-ResNet	Spectrogram	SE-ResNet 50
SE-ResNeXt (Spec)	Spectrogram	SE-ResNeXt 101
SE-ResNeXt (NMF)	Feature matrix (NMF)	SE-ResNeXt 101

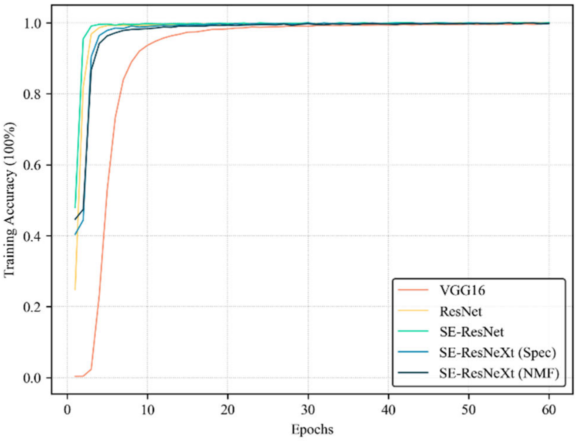

FIGURE 12. The performance of each method on the training set.

method, MFCC-GMM, is a Gaussian mixture model using MFCC as a feature, and the second to fourth methods use a spectrogram as an input feature. The input feature for the last method is the feature matrix of the spectrogram after non-negative matrix factorization.

The FIGURE 12 illustrates the performance of each method on the training set. It can be observed that the residual networks have the fastest increase in iteration accuracy, while VGG16 has lower iteration accuracy than the other methods. Overall, all methods have a consistent accuracy after 30 training iterations. The following experiments were conducted using the aforementioned best training model.

Table 5 presents the speaker recognition accuracy of each method in a clean speech environment. As can be observed from the table above, all compared methods exhibit good recognition performance without the inclusion of noise or other processing methods. SE-ResNeXt (Spec), which employs traditional spectrograms as input, boasts an outstanding recognition rate of 99.45%. SE-ResNeXt (NMF) also exhibits a good recognition rate of 97.67%. The disparity in performance can be attributed to the loss of some feature information on the spectrogram through NMF. However, attaining a clean and informative recognition environment is quite exigent in complex and dynamic recognition environments. Therefore, the ensuing experiments were designed to evaluate the recognition performance of each method in complex settings.

The initial set of experiments aimed to test the performance of each method under different noise levels, thus evaluating the ability of the methods to capture features. To this end, the original speech data from the AISHELL-1 database was utilized for testing, and additional noise data from the NOISE-92 database was introduced to the test set. The experiment aimed to simulate a range of challenging recognition scenarios by incorporating diverse noise samples. Specifically, babble noise was chosen as it closely matches the noise present in real-world recognition environments, thereby increasing the practical applicability of the proposed method. In addition, a commonly-used factory1 noise was also included as overlaid noise. To quantify the effect of noise on the recognition performance, the experiment employed preset SNR levels for superimposing the noise on the speech data. The experimental results are shown in FIGURE 13 and FIGURE 14.

TABLE 5. The number of parameters of the model and the accuracy of the test set.

Method	Params	Data set	Accuracy
MFCC-GMM	—	AISHELL-1	97.97%
VGG16	1.3835e+8		96.17%
ResNet	2.5550e+7		99.18%
SE-ResNet	2.8065e+7		99.32%
SE-ResNeXt (Spec)	9.3528e+7		99.45%
SE-ResNeXt (NMF)	9.3528e+7		97.67%

In the presence of babble noise, it has been observed that each method demonstrates a commendable recognition performance within the range of 15-20 dB of SNR. However, as the SNR decreases further, the accuracy of the other methods drops significantly. Nonetheless, SE-ResNeXt (NMF) maintains a consistent recognition rate of 85.4% even when the SNR is as low as 5 dB. Regarding using SE-ResNeXt (Spec) with spectrogram, it still exhibits impressive performance during testing, and its framework outperforms other methods.

The FIGURE 14 experimental results indicate that all methods' recognition performance is significantly reduced under factory1 noise. However, only the method that utilizes NMF consistently performed better than other methods across all SNR levels, providing further evidence that NMF is conducive to improving recognition rates in challenging noise environments. This is achieved with lower computational complexity while still delivering superior performance. This is due to NMF's characteristic of feature dimensionality reduction and classification signal source, which reduces data redundancy and improves feature extraction efficiency.

VGG16 and ResNet generate feature information from the last layer of the network, often neglecting some shallow texture information and channel-related features, potentially useful for speaker recognition. Traditional MFCC-GMM estimates parameters of acoustic features, while noise in recognition can lead to parameter deviation and variance expansion in Gaussian distribution, significantly affecting recognition

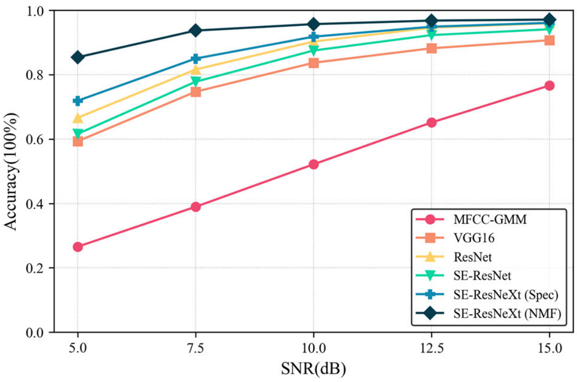

FIGURE 13. The performance of each method overlaid with babble noise.

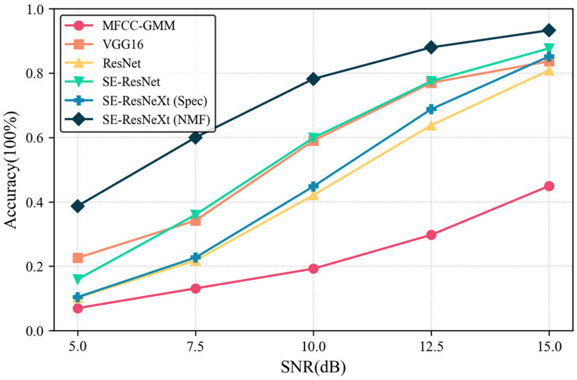

FIGURE 14. The performance of each method overlaid with factory1 noise.

TABLE 6. Short speech recognition performance of each method.

Duration	Method					
	MFCC-GMM	VGG16	ResNet	SE-ResNet	SE-ResNeXt (Spec)	SE-ResNeXt (NMF)
1s	69.4%	74.5%	95.0%	96.0%	96.0%	90.3%
2s	80.1%	94.7%	98.9%	98.9%	99.2%	94.8%

accuracy. As shown in FIGURE 13 and FIGURE 14, MFCC-GMM speaker recognition has the lowest accuracy among all methods.

E. SHORT SPEECH RECOGNITION PERFORMANCE

In order to further investigate the effectiveness of the proposed algorithm in complex recognition environments, a second experiment was conducted. The following table showcases the results of the experiments on the short speech performance of each method, whereby short speech levels of 1s and 2s were utilized to evaluate the recognition performance of each method on short speech segments.

The experiment employed endpoint detection before intercepting the test audio to guarantee the absence of lengthy silent periods in the short speech (1s, 2s) test samples. This procedure ensures that only valid speech clips are fed to the methods.

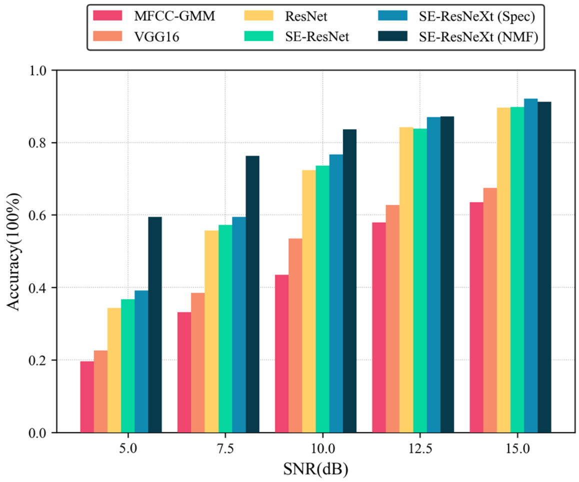

FIGURE 15. The performance of each method on short speech (1s) overlaid babble noise.

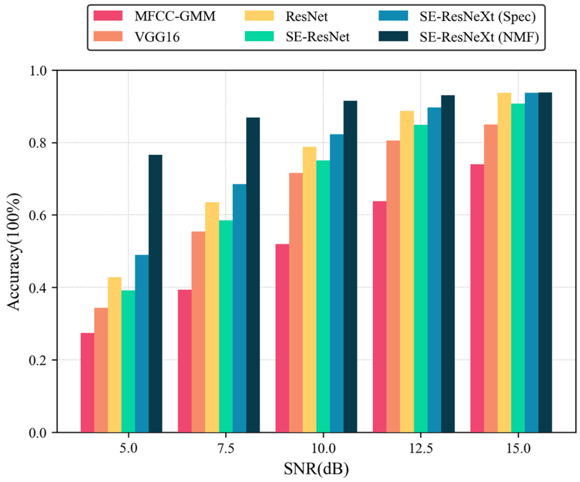

FIGURE 16. The performance of each method on short speech (2s) overlaid babble noise.

Table 6 above shows the performance of six different methods for various durations.

Subsequently, the experiment augmented the test difficulty by introducing noise into the short phonetic samples. The challenging experimental conditions rigorously examined the performance of each method. The experiment comprised two conditions: short speech samples lasting for 1s and 2s, and babble noise. As depicted in FIGURE 15 and FIGURE 16, the recognition rate of each method was evaluated under short speech samples containing babble noise.

From FIGURE 15, it can be intuitively concluded that SE-ResNeXt (NMF) still performs excellently in the noise environment of 1s. SE-ResNeXt (NMF) has maintained a stable recognition level between the SNR of 5-15 and still has a recognition rate of 59.5% even under SNR=5. At the same time, the highest recognition rates of other methods did not exceed SE-ResNeXt (NMF) at the same SNR level. This observation suggests that incorporating SE blocks and NMF into the ResNeXt method enables it to focus on the

TABLE 7. Performance of different decomposition ranks.

NMF	Precision	Recall	Accuracy
R=3	95.6%	95.15%	95.2%
R=5	96.3%	95.9%	95.9%
R=10	96.7%	96.4%	96.5%
R=20	97.1%	96.9%	97.0%
R=30	**97.7%**	**97.5%**	**97.6%**

TABLE 8. Performance of combining SE blocks in different positions.

System	Precision	Recall	Accuracy
Baseline	95.1%	94.4%	94.5%
Conv=2, 3	97.3%	97.1%	97.1%
SE-ResNeXt (NMF)	**97.7%**	**97.5%**	**97.6%**

TABLE 9. Overlaid 10dB of babble noise.

System	Precision	Recall	Accuracy
Baseline	91.8%	89.3%	89.3%
Conv=2, 3	95.7%	95.1%	95.0%
SE-ResNeXt (NMF)	**96.1%**	**95.7%**	**95.7%**

information that significantly contributes to the final classification, thereby improving its performance. Additionally, the integration of NMF has facilitated the removal of the information redundancy in the network input in low SNR conditions, consequently achieving noise reduction and signal source separation. Overall, addition the SE block and NMF improves the method's performance, making it more robust and reliable in the presence of noise.

The results presented in FIGURE 16 indicate that the accuracy rates of all methods have improved when the speech length is increased to 2 seconds. This increase in speech length allows for more speech information in each method, leading to better recognition performance overall. In addition, it is worth noting that when the SNR is between 12.5-15, the recognition rate of the model based on the residual network is higher than that of other models. At low SNR, SE-ResNeXt (NMF) with NMF performs better than other methods with a slower accuracy drop. At an SNR of 5, the recognition rate of SE-ResNeXt (NMF) on 2-second audio snippets was 76.60%, significantly higher than that of other methods at the same level. This can be attributed to the increase in features provided by longer speech samples and the effective separation of noise sources by NMF in noisy environments. Experimental results further emphasize the importance of considering speech length when designing speaker recognition methods.

F. PERFORMANCE OF DIFFERENT DECOMPOSITION RANKS

In prior investigations on NMF [18], the choice of NMF decomposition rank usually relies on commonly used values, such as $R = 3, 5$, or 10. In [20], the experimental results show that when $R < 5$, the number of R has a more significant impact on the system. However, the system performance variation becomes less pronounced as the decomposition rank increases. Therefore, the final choice of decomposition level set by the author is $R = 10$. This paper does not use empirical values, but uses the AIC-time joint confirmation method to determine the most suitable decomposition level for a specific experiment. The experiment compares the recognition accuracy using empirical values. The outcomes of employing various decomposition levels in the system are presented in Table 7, illustrating their respective performance.

Furthermore, this experiment incorporates the assessment metrics of precision rate and recall rate, which are defined as follows.

$$Precision = \frac{n_{TP}}{n_{TP} + n_{FP}} \tag{1.18}$$

$$Recall = \frac{n_{TP}}{n_{TP} + n_{FN}} \tag{1.19}$$

The macro-average method is used for multi-category problems, and the evaluation indicators (Precision/ Recall) of different categories are added to calculate the average. Among them, n_{TP} is True Positive, and the positive class is predicted to be positive; n_{FP} is False Positive, and the negative class is predicted to be positive; n_{FN} is False Negative, and the positive class is predicted to be negative.

In Table 7, In the evaluation of the widely employed empirical value decomposition rank, it is observed that the accuracy rate exhibits an upward trend as the decomposition rank increases. Notably, when the decomposition rank R is below 10, there is a substantial decrease in the model's performance. The optimal performance is achieved using the factorized rank validated through the AIC-time joint confirmation method.

Consequently, the decomposition rank $R = 30$, as determined by this experiment's confirmation method, represents the most advantageous rank.

G. EXPLORATION OF SE BLOCKS AT DIFFERENT STAGES

In [46], it was found that the optimal combination position of SE blocks in speaker verification applications is between conv2 and conv3 in the residual network. Table 8 compares the combination positions proposed in this study with the optimal combination position mentioned in the referenced literature, where the baseline model represents a network without any SE blocks.

The performance of the baseline model on the AISHELL dataset could be better. However, upon integrating the SE block into the model, previous research has demonstrated a remarkable accuracy rate of 97.1%. Remarkably, the employed combination yielded a modest improvement of 0.5% compared to the aforementioned best-performing combination in this experiment.

To further study the model performance of different combinations, in Table 9, 10dB noise is overlaid on the test set, thus revealing the differences in different acoustic environments.

The baseline model exhibits marked inferiority compared to the model incorporating SE blocks in the noise test. The two models with SE blocks have a better recognition rate of noisy speech, and the model used in this paper reaches the highest accuracy rate of 95.7%. The combination of SE blocks in the network model can effectively improve the recognition performance, and the combined position of SE blocks also has different effects on performance. Combining SE blocks after each convolutional layer has the best recognition performance for the current experiments.

VI. CONCLUSION

This paper proposes to use REE-based AIC and a joint computational complexity method to select the optimal decomposition level of NMF and to experimentally select the optimal location of the combination of SE blocks and ResNeXt for speaker recognition. The speaker feature is a feature matrix obtained by applying NMF to the spectrogram. Compared with the traditional spectrogram, it has better robustness and sparsity and can effectively extract feature information. The optimal decomposition rank of NMF is also confirmed by the AIC-Time joint confirmation method. The ResNeXt network utilizes residual connections and group convolution to extract features from speech signals better. The combination with the SE block enables adaptive adjustment of channel weights in feature maps, enhancing the network's representational capacity.

Finally, compared to other commonly used speaker recognition methods, the SE-ResNeXt (NMF) method has significant advantages in noisy environments and short speech segments. However, its performance could improve in a cleaner and perfect recognition environment. Therefore, future research on speaker recognition needs to address and resolve the following issues:

(1) The recognition ability in clean environments needs to be improved, resulting in poorer accuracy in high-SNR scenarios compared to other methods. Improving the overall recognition accuracy of the method is the next research direction.

(2) The generalization of the method in different noise environments could be better, hence the need to improve its generalization ability to different types of noise.

In future research, we will continue to optimize and improve the recognition accuracy of this method, as well as explore its potential application in other relevant fields.

REFERENCES

[1] S. Colby and A. J. Orena, "Recognizing voices through a cochlear implant: A systematic review of voice perception, talker discrimination, and talker identification," *J. Speech, Lang., Hearing Res.*, vol. 65, no. 8, pp. 3165–3194, Aug. 2022, doi: 10.1044/2022_JSLHR-21-00209.

[2] S. Jainar, P. Sale, and B. Nagaraja, "VAD, feature extraction and modelling techniques for speaker recognition: A review," *Int. J. Signal Imag. Syst. Eng.*, vol. 12, nos. 1–2, pp. 1–18, Jan. 2020.

[3] F. Ye and J. Yang, "A deep neural network model for speaker identification," *Appl. Sci.*, vol. 11, no. 8, p. 3603, Apr. 2021, doi: 10.3390/app11083603.

[4] R. Jahangir, Y. W. Teh, N. A. Memon, G. Mujtaba, M. Zareei, U. Ishtiaq, M. Z. Akhtar, and I. Ali, "Text-independent speaker identification through feature fusion and deep neural network," *IEEE Access*, vol. 8, pp. 32187–32202, 2020, doi: 10.1109/ACCESS.2020.2973541.

[5] K. Saakshara, K. Pranathi, R. M. Gomathi, A. Sivasangari, P. Ajitha, and T. Anandhi, "Speaker recognition system using Gaussian mixture model," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Jul. 2020, pp. 1041–1044.

[6] B. Mor, S. Garhwal, and A. Kumar, "A systematic review of hidden Markov models and their applications," *Arch. Comput. Methods Eng.*, vol. 28, no. 3, pp. 1429–1448, 2021.

[7] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 4052–4056, doi: 10.1109/ICASSP.2014.6854363.

[8] N. S. Ibrahim and D. A. Ramli, "I-vector extraction for speaker recognition based on dimensionality reduction," *Proc. Comput. Sci.*, vol. 126, pp. 1534–1540, Jan. 2018.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[10] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 151–167, Jan. 2013, doi: 10.1016/j.csl.2012.01.008.

[11] A. Awais, S. Kun, Y. Yu, S. Hayat, A. Ahmed, and T. Tu, "Speaker recognition using mel frequency cepstral coefficient and locality sensitive hashing," in *Proc. Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2018, pp. 271–276.

[12] B. Karan, S. S. Sahu, J. R. Orozco-Arroyave, and K. Mahto, "Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction," *Comput. Speech Lang.*, vol. 69, Sep. 2021, Art. no. 101216.

[13] F. E. Abualadas, M. Shaban, and A.-E. Messikh, "Speaker identification based on hybrid feature extraction techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 3, pp. 1–12, 2019.

[14] R. Bharti and P. Bansal, "Real time speaker recognition system using MFCC and vector quantization technique," *Int. J. Comput. Appl.*, vol. 117, no. 1, pp. 25–31, May 2015.

[15] S. Peng, T. Lv, X. Han, S. Wu, C. Yan, and H. Zhang, "Remote speaker recognition based on the enhanced LDV-captured speech," *Appl. Acoust.*, vol. 143, pp. 165–170, Jan. 2019, doi: 10.1016/j.apacoust.2018.08.007.

[16] D. Sztahó, G. Szaszák, and A. Beke, "Deep learning methods in speaker recognition: A review," 2019, *arXiv:1911.06615*.

[17] X. Lin and P. C. Boutros, "Optimization and expansion of non-negative matrix factorization," *BMC Bioinf.*, vol. 21, no. 1, pp. 1–10, Dec. 2020.

[18] M. Mulimani and S. G. Koolagudi, "Segmentation and characterization of acoustic event spectrograms using singular value decomposition," *Expert Syst. Appl.*, vol. 120, pp. 413–425, Apr. 2019.

[19] A. Kacha, F. Grenez, J. R. Orozco-Arroyave, and J. Schoentgen, "Principal component analysis of the spectrogram of the speech signal: Interpretation and application to dysarthric speech," *Comput. Speech Lang.*, vol. 59, pp. 114–122, Jan. 2020.

[20] S. Lee and H.-S. Pang, "Feature extraction based on the non-negative matrix factorization of convolutional neural networks for monitoring domestic activity with acoustic signals," *IEEE Access*, vol. 8, pp. 122384–122395, 2020, doi: 10.1109/ACCESS.2020.3007199.

[21] S. Abdali and B. NaserSharif, "Non-negative matrix factorization for speech/music separation using source dependent decomposition rank, temporal continuity term and filtering," *Biomed. Signal Process. Control*, vol. 36, pp. 168–175, Jul. 2017, doi: 10.1016/j.bspc.2017.03.010.

[22] Y. Yi, J. Wang, W. Zhou, C. Zheng, J. Kong, and S. Qiao, "Non-negative matrix factorization with locality constrained adaptive graph," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 427–441, Feb. 2020.

[23] Q. Liu, M. A. Charleston, S. A. Richards, and B. R. Holland, "Performance of Akaike information criterion and Bayesian information criterion in selecting partition models and mixture models," *Systematic Biol.*, vol. 72, no. 1, pp. 92–105, Dec. 2022, doi: 10.1093/sysbio/syac081.

[24] Z. Zhao, Y. Liu, and C. Peng, "Variable selection in generalized random coefficient autoregressive models," *J. Inequal. Appl.*, vol. 2018, no. 1, p. 82, Apr. 2018, doi: 10.1186/s13660-018-1680-4.

[25] V. C. K. Cheung, K. Devarajan, G. Severini, A. Turolla, and P. Bonato, "Decomposing time series data by a non-negative matrix factorization algorithm with temporally constrained coefficients," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 3496–3499, doi: 10.1109/EMBC.2015.7319146.

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.

[27] N. N. An, N. Q. Thanh, and Y. Liu, "Deep CNNs with self-attention for speaker identification," *IEEE Access*, vol. 7, pp. 85327–85337, 2019, doi: 10.1109/ACCESS.2019.2917470.

[28] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, "Acoustic scene classification with squeeze-excitation residual networks," *IEEE Access*, vol. 8, pp. 112287–112296, 2020, doi: 10.1109/ACCESS.2020.3002761.

[29] B. Zou, H. Yan, F. Wang, Y. Zhou, and X. Zeng, "Research on signal modulation classification under low SNR based on ResNext network," *Electronics*, vol. 11, no. 17, p. 2662, Aug. 2022, doi: 10.3390/electronics11172662.

[30] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. 20th Conf. Oriental Chapter Int. Coordinating Committee Speech Databases Speech I/O Syst. Assessment (O-COCOSDA)*, Nov. 2017, pp. 1–5, doi: 10.1109/ICSDA.2017.8384449.

[31] T. Zhou, Y. Zhao, and J. Wu, "ResNeXt and Res2Net structures for speaker verification," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 301–307.

[32] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[33] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, "GMM and CNN hybrid method for short utterance speaker recognition," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3244–3252, Jul. 2018.

[34] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, Mar. 2019, doi: 10.1109/MSP.2018.2877582.

[35] M. Hou, J. Li, and G. Lu, "A supervised non-negative matrix factorization model for speech emotion recognition," *Speech Commun.*, vol. 124, pp. 13–20, Nov. 2020.

[36] G. A. Khan, J. Hu, T. Li, B. Diallo, and H. Wang, "Multi-view data clustering via non-negative matrix factorization with manifold regularization," *Int. J. Mach. Learn. Cybern.*, pp. 1–13, 2022.

[37] G. R. Naik, *Non-Negative Matrix Factorization Techniques*. Cham, Switzerland: Springer, 2016.

[38] H. Qiao, "New SVD based initialization strategy for non-negative matrix factorization," *Pattern Recognit. Lett.*, vol. 63, pp. 71–77, Oct. 2015.

[39] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 66–70, doi: 10.1109/ICASSP.2015.7177933.

[40] Y. Xue, C. S. Tong, Y. Chen, and W.-S. Chen, "Clustering-based initialization for non-negative matrix factorization," *Appl. Math. Comput.*, vol. 205, no. 2, pp. 525–536, Nov. 2008, doi: 10.1016/j.amc.2008.05.106.

[41] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[43] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–13, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

[45] R. Mohd Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Comput. Electr. Eng.*, vol. 90, Mar. 2021, Art. no. 107005.

[46] M. Rouvier and P.-M. Bousquet, "Studying squeeze-and-excitation used in CNN for speaker verification," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 1110–1115, doi: 10.1109/ASRU51503.2021.9687936.

DONGBO LIU is currently an Associate Professor with Xihua University, with a robust academic background. His research interests include intelligent control, pattern recognition, and sensing technology.

LIMING HUANG received the B.S. degree in electrical engineering from Tarim University, Xinjiang, China, in 2021. He is currently pursuing the M.S. degree in electronic information with Xihua University. His research interests include speaker recognition and audio processing.

YU FANG is currently an Assistant Professor with Xihua University. Her research interests include signal processing, system design, and biomedical engineering.

WEIBO WANG received the B.S. and M.S. degrees in information engineering from Xihua University, Chengdu, China, in 2000 and 2003, respectively, and the Ph.D. degree in technology of computer application from Southwest Jiaotong University, Chengdu, in 2011. He was a Visiting Scholar with Dr. Raj Mittra with The Pennsylvania State University, in 2013. He is currently an Associate Professor with Xihua University. His research interests include intelligent signal and information processing and fault detection.