Enhancing Speaker Recognition through Generative Adversarial Networks and Feature Fusion with ResNet18 Networks

Yabo Xu, Xi Li, Runqiang Chen, Weichen Wang

***Abstract*—Speaker recognition, reliant on voiceprint features, is often constrained by insufficient data samples. To mitigate this deficiency, this study proposes a novel approach that leverages generative adversarial networks (GANs) to augment speech features and the ResNet18 network for speaker recognition. Initially, spectrograms are extracted from speech signals, serving as input features. GANs are then employed to generate additional voiceprint features based on the original spectrograms. These generated features are fused with the original spectrograms at varying fusion ratios. The fused features are subsequently fed into a ResNet18 network for speaker recognition. Experimental results prove that the proposed method significantly enhances recognition performance, achieving a peak recognition rate of 97.92% at a 20% fusion ratio. This study underscores the effectiveness of GANs-based feature augmentation in improving speaker recognition accuracy, especially in scenarios with limited training data.**

***Index Terms*—speaker recognition, generative adversarial network, spectrograms, ResNet18 network, feature fusion**

I. Introduction

Voiceprint recognition, also termed speaker recognition, is a biometric identification method that distinguishes individuals based on unique vocal characteristics embedded in their speech signals. This technique determines speaker identity by analyzing voiceprint features [1]. Speaker recognition techniques can be classified into three categories according to the task to be recognized and the application scenario: speaker verification [2], speaker identification, and speaker diarization [3]. According to the content of recognition, it can be divided into two categories: text-dependent and text-independent [4]. The text-dependent implementations, speakers must articulate predefined linguistic content. Tsai et al. [5] advanced this paradigm through evocative word modeling, demonstrating enhanced recognition accuracy when incorporating psychologically salient vocabulary in authentication phrases. Text-independent imposes no constraints on linguistic content and the focus of recognition is to confirm the identity of the speaker. Mohammadi M et al. proposed a method of text-independent speaker verification in noisy environments by fusing different features such as Mel-Frequency Cepstral Coefficients, Linear-Frequency Cepstral Coefficients, and Power Normalized Cepstral Coefficients [6]. In forensic applications, Wang et al. [7-9] pioneered text-independent speaker recognition systems that analyze crime scene voice evidence to establish suspect identification, demonstrating particular efficacy in evidentiary voiceprint matching. Text-dependent speaker recognition is easy to implement, but the obvious limitation is that the same sentence has to be uttered during the recognition process [10]. Text-independent speaker recognition is challenging and usually requires more data to achieve better results, but is more versatile and has large developmental potential [11].

In recent years, the advancement of deep learning has driven extensive adoption of speech spectrograms as acoustic feature inputs to neural networks for enhanced recognition performance. [12, 13]. Within speaker recognition research, in order to improve the recognition rate, Zhang et al. [14] demonstrated that directly feeding speech spectrograms into neural networks with depth-separable convolution operations improves computational efficiency while reducing model complexity. To address feature stability, Jia et al. [15] proposed linear superposition of short-time speech spectrograms to construct pronunciation-stable representations for network training and classification. Zhu et al. [16] further expanded spectrogram datasets through segmentation of single-speaker audio, effectively mitigating data scarcity in deep learning model training. Shafik et al. [17] enhanced noise robustness by integrating Ladon-transformed spectrograms with convolutional neural networks.

The GANs have demonstrated remarkable success across image generation and speech processing tasks, including feature representation learning [18], speech conversion [19], and emotion conversion [20]. Innovative integrations of GANs with other architectures have emerged: Cao et al. [21] combined GANs with variational autoencoders (VAEs) to disentangle and recombine emotion-related and content-related speech features, enabling emotion transfer while preserving speaker identity and linguistic content. Kameoka et al. [22] developed a GAN variant (Saran) that outperformed variational autoencoder-based methods in generating high-fidelity speech with improved speaker

Manuscript received November 26, 2024; revised May 24, 2025.

This work was supported in part by the Graduate Research and Practice Innovation Program of Jiangsu Normal University under Grant 2024XKT06430.

Yabo Xu is a postgraduate student of Jiangsu Normal University, Jiangsu, 221116 China (e-mail: 1114872761@qq.com).

Xi Li is a postgraduate student of Jiangsu Normal University, Jiangsu, 221116 China (e-mail: 443918675@qq.com).

Runqiang Chen is a postgraduate student of Jiangsu Normal University, Jiangsu, 221116 China (e-mail: 1049110759@qq.com).

Weichen Wang is a professor of Jiangsu Normal University. Jiangsu, 221116 China (corresponding author to provide phone: +86-18168772633; e-mail: 18168772633@163.com).

similarity. Complementing these advances, Chen et al. [4] leveraged pre-trained autoencoders to extract discriminative otoacoustic emission features, significantly boosting speaker recognition performance.

Despite these advancements, deep learning-based speaker recognition systems remain constrained by their dependence on large-scale datasets, and the recognition rate is difficult to improve effectively when only insufficient data are available. To address this limitation, we utilize a ResNet18 network containing residual blocks to enhance feature extraction and classification [23]. In this paper, text-independent speaker recognition is taken into account for its versatility and developmental potential. In order to obtain better model training results and accurate recognition, generative adversarial network is used to generate additional vocal features, the generated features are fused with the original vocal features, and the fusion results are taken as the input to the ResNet18 network.

II. Feature Extraction for Spectrograms

A large amount of information related to the utterance properties of speech is contained in the spectrogram, which combines the features of spectrum and time-domain waveforms to visualize the evolution of spectrograms, with any frequency component's energy at a given moment represented by color depth [24]. A spectrogram's lines of varying degrees of color are called voiceprints [25]. Speech spectrogram is one of the main speech signal features used in voiceprint recognition models. The feature extraction process for spectrograms is subdivided into pre-emphasis, framing, windowing, and short-time Fourier transform (STFT), as shown in Fig. 1, where α is the pre-emphasis coefficient, $x(n)$ is the speech signal, $w(n)$ is the Hamming window, $h(\tau\text{-}t)$ is the window function, and N is the window length.

The main purpose of pre-emphasis is to enhance the power spectrum of the high-frequency component of the speech signal [26]. After the pre-emphasis process, the high-frequency portion of the speech signal is increased, and the spectrum becomes flatter, making it easier to study and analyze the spectrogram later. Framing aims to convert an unsteady signal into a short-time steady signal. After framing, there will be breakpoints at the beginning and end of each frame, so the more frames are segmented, the greater the error with the original signal [27]. The purpose of adding a window is to make the signal continuous after framing, and each frame will show the characteristics of the periodic function. To combine the time-domain and frequency-domain characteristics of audio signals, this paper uses the short-time Fourier transform (STFT) to analyze the speech signal segment by segment after the windowing of subframes. Thus, the STFT can be reduced to the Fourier transform with windowing [28].

III. Generation and Discrimination of New Spectrogram Features Based on Generative Adversarial Network

A generative adversarial network consists of a generator and a discriminator network. The input to the generator network is random noise, and the output is fake data similar to real data. The goal of the generator network is to generate high-quality fake data that is as close as possible to real data distribution. The discriminator network compares the output generated by the generator network with the real spectrogram to distinguish between real and fake data. The generator network continuously optimizes the data it generates to increase the difficulty of the discriminator network's judgment; the discriminator network optimizes the loss function to make the judgment more accurate. The relationship between the two forms a confrontation, which coined the "adversarial network" term [29]. The two networks are trained through constant adversarial training, enabling the generator to gradually generate more realistic samples and eventually reach an indistinguishable level from the real samples. At the same time, the discriminator can distinguish the difference between real and generated samples. In this paper, the discriminator network compares the original spectrogram with the new features generated by the generator network so that the latter forms feature similar to the original spectrogram. The overall workflow of the GANs is shown in Fig. 2.

(1) Generating new spectrogram features using generator network

The workflow for the generator network is shown in Fig. 3. Considering the running time of the generator network and the clarity of the generated speech maps, in this paper, the size of the network output is set to 64 × 64× 3, i.e., the size of the generator network used to generate new speech maps. The network input is a random vector of 100 × 1× 1, with the aim of generating a different spectrogram each time. The vectors are converted to 4 ×4 × 512 arrays by Projection and Reshape operations, and then the generated arrays are scaled up to 64 × 64 × 3 arrays by a series of the Transposition Convolutions, the Batch Norm (BN) layers, and the ReLu layers, as shown in Fig. 3.

The Transposed Convolution layers are used in the generator network to transform low-dimensional random noise into a high-resolution image. The four Transposed Convolution layers are used in the generator network, each layer has a decreasing number of filters and padding of the edge data. The purpose is to gradually multiply the length and width of the matrix while making the number of channels gradually decrease. For the final transposed convolutional layer, three filters are specified corresponding to each of the three RGB channels of the generated image. In the generator network except for the output layer using the Tanh activation function, all the other layers use the ReLu activation function, and the goal is to minimize the discriminative rate of the discriminator network. The structure of the generator network is shown in Fig. 3, where n is the number of convolution kernels, k is size of the convolution kernels, s is convolution stride.

(2) Discriminating new spectrogram features using discriminator networks

The workflow for the generator network is shown in Fig.4. The input of the discriminator network is the same as the output of the generator network, with the size of 64 × 64 × 3, and the predicted score is returned through a series of the convolutions, the BN and the Leaky ReLu layers. The output of the network is the prediction score, and after several iterations, the final ideal situation is that the discriminator

network cannot distinguish whether the samples come from the output of the generative network or the real spectrogram, i.e., the final prediction score is 0.5.

There are four Convolution layers in the discriminator network, each layer has an increasing number of filters and padding of the edge data. A Dropout layer is used to add noise to the input spectrogram to improve robustness. The existence of the BN layers in the discriminator and the generator networks helps to deal with training problems caused by poor initialization, accelerates model training, and improve training stability. However, the BN for all layers can lead to sample oscillations and model instability, so the BN layers are not used for output layer of the generative network and input layer of the discriminator network. In the discriminator network, the Leaky ReLu activation function is used for all layers except output layer and the goal is to maximize the discriminative rate of the discriminator network. The structure of the discriminator network is shown in Fig. 4, where p is dropout rate, a is size of slope.

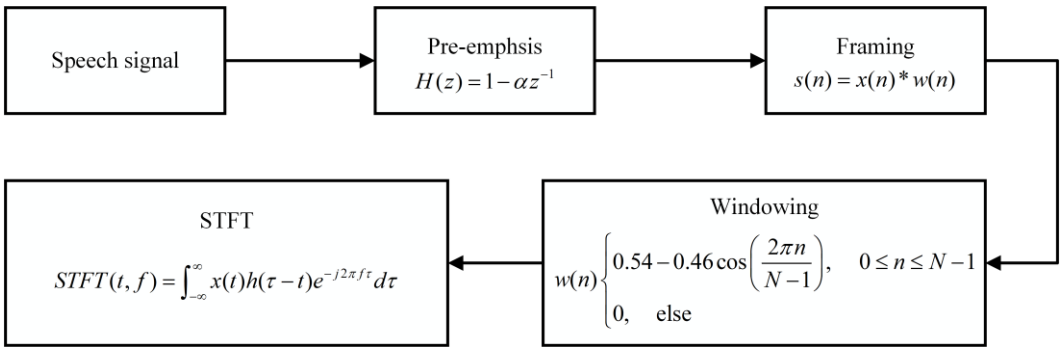


Fig. 1 Feature extraction process for spectrograms

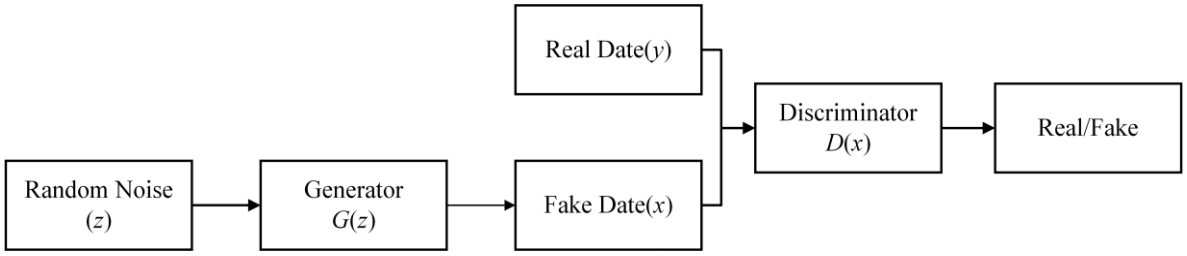


Fig. 2 Overall workflow of the GANs

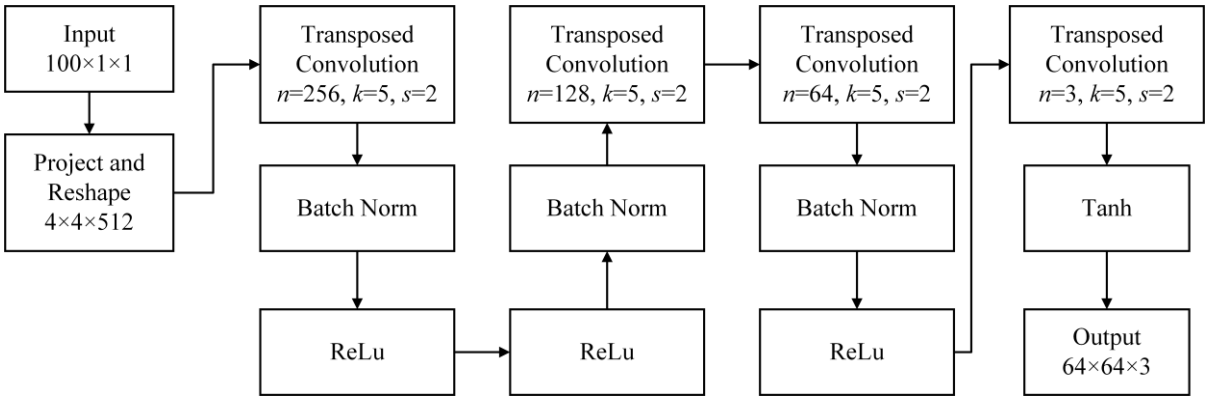


Fig. 3 Generator network structure

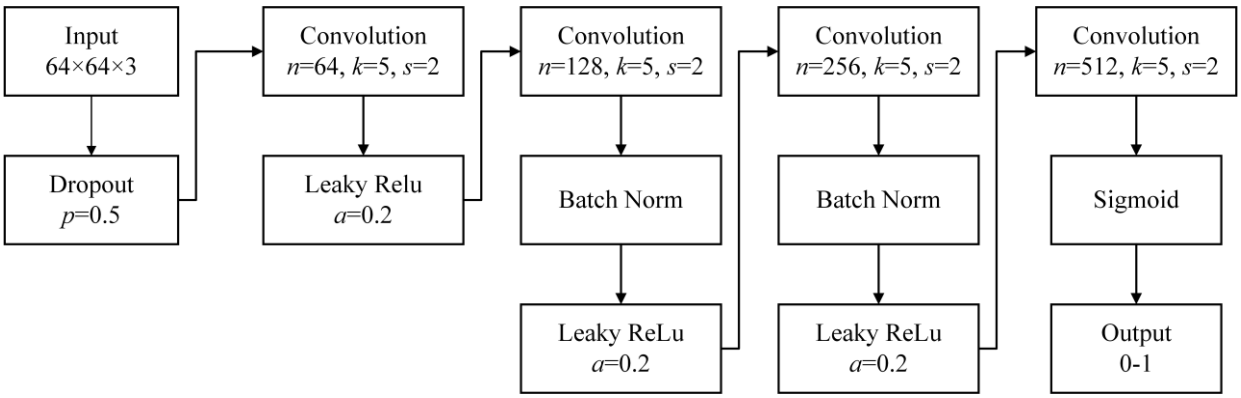


Fig. 4 Discriminator network structure

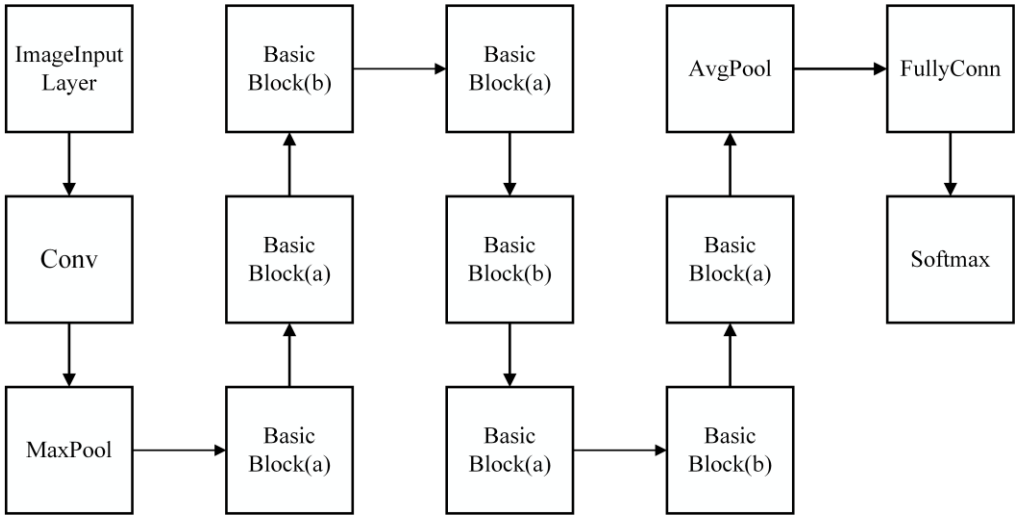


Fig. 5 ResNet18 network structure model

TABLE. I
NETWORK PARAMETERS

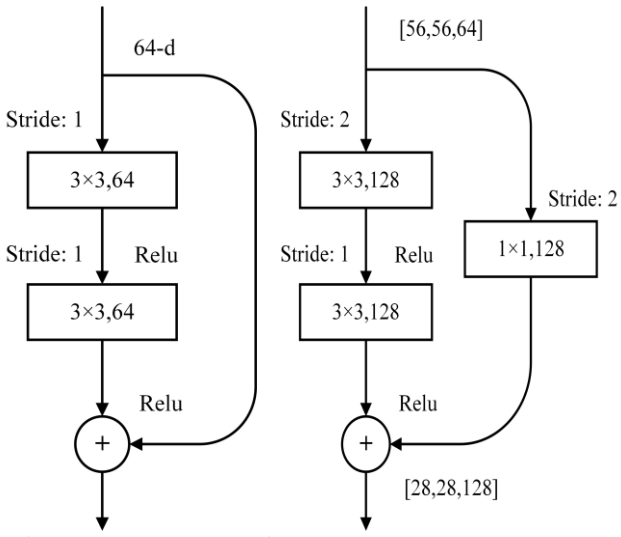
Input	Operator	*k*	*s*	*p*
224^2×3	Conv	7	2	3
112^2×64	MaxPool	3	2	1
56^2×64	Basic Block(a)	3	1	1
56^2×64	Basic Block(a)	3	1	1
56^2×128	Basic Block(b)	1	2	1
28^2×128	Basic Block(a)	3	1	1
28^2×256	Basic Block(b)	1	2	1
14^2×256	Basic Block(a)	3	1	1
14^2×512	Basic Block(b)	1	2	1
7^2×512	Basic Block(a)	3	1	1
1^2×512	Avgpool			
512×n	Fullyconnect			

IV SPEAKER RECOGNITION BASED ON RESNET18 NETWORK

ResNet18 is a conventional deep convolutional neural network model with good feature extraction and classification performances. It has been widely applied for image classification, target detection, and other problems. The residual block was introduced to ResNet18 to mitigate the problems of gradient vanishing and gradient explosion in deep convolutional neural networks [30]. This study selected the ResNet18 network for the speaker recognition task. The network structure is depicted in Fig. 5. The network parameters are listed in Table. I, where *k* is the size of the filter convolution kernel, *s* is the step size, and *p* is the image padding.

The two structures of the basic block are shown in Fig. 6. In the residual block, the shortcut connection allows the layer output to directly skip one or more layers and connect to the input of the subsequent layers, permitting the network to learn the residual information for better feature extraction and processing, as shown in Fig.6(a). When a shortcut connection is performed in the network, the input and output dimensions in the same residual block should be the same, and the information in the shallow layer can be directly transferred to the deep layer through the jump connection, solving the degradation problem and being treated as feature reuse. If the front and back dimensions differ, a dimension upgrading operation is required, as shown in Fig. 6(b).

Figure 6(a) contains two filters with 3 × 3 convolutional kernels and the ReLu layers, with a stride size of 1 and a channel count of 64. The front and back dimensions are not the same in Fig. 6(b), so the dimension upgrading operation is required. The main branch in Fig. 6(b) contains two filters with 3 × 3 convolutional kernels and the ReLu layers. The first convolution has a step size of 2, and the second has a step size of 1, both with 128 channels. In the ResNet structure, the output feature matrix shape of the shortcut and the main branch must be the same; therefore, as shown in the right side of Fig. 6(b), a 1 × 1 convolution kernel upscales the shortcut branch and the resolution is changed by setting the step size to 2, which ultimately matches the number of channels.



（a）Shortcut connection （b）Dimension upgrading operation

Fig. 6 The two structures of Basic Block

V Experimental Setup and Result Analysis

(1) Experimental dataset and parameter settings

The voice data used in this paper are acquired from the voice set of twelve characters in a game, which are recorded in Chinese in a professional recording studio. The voices of the twelve characters are extracted from the voice set, and 150 voice segments are randomly selected from each character's lines to generate spectrograms. The duration of each voice exceeds 2 s. Each spectrogram data set in the experiment is subdivided into training and validation sets at 8:2.

The original spectrograms and the generated ones are preprocessed into images in size of 224 × 224× 3 as input to the classification network. The sampling frequency of the speech signal is 16,000 Hz, the frame length is 25 ms, the frameshift is 10 ms, and the number of Fourier transform points is 1024. The model is executed on a graphics processing unit (GPU), which is usually more efficient than a central processing unit (CPU) for neural network processing, and the computer configurations used in the experiment are listed in Table II.

TABLE. II
Experimental Environment Configurations

Hardware Platform	Parameters
CPU	AMD Ryzen 5 7500F 6-Core Processor
GPU	NVIDA RTX4060 Ti 8 GB
Operating Memory	32GB

The backpropagation algorithm based on stochastic gradient descent is applied to the ResNet18. The model parameters are optimized by minimizing the Cross-Entropy Loss function, which is expressed as the difference between the true probability distribution and the predicted probability distribution, as shown in Equation (1). The smaller the cross-entropy value, the better the model prediction. The training options that need to be pre-set in the ResNet18 network contain the initial learning rate, number of training rounds, number of classifications, validation frequency, and batch size. Considering the dataset size, computer configurations, network structure, and other factors, the values of the training parameters are set in Table III.

$$Loss = -\sum_{i=1}^{n} y_i \log(y_i - z_i)^2 \quad (1)$$

where *yi* is actual value, *zi* predicted value, and *n* sample size.

The proposed framework comprises the GANs model and the ResNet18 mode, as shown in Fig. 7. Each speech is extracted into a spectrogram as a feature input. In the GANs network, spectrograms are used as input to generate new spectrograms at different ratios. In the ResNet18 network, new spectrograms are fused with the original speech spectrograms as input to train the network. In the testing phase, 20% of the original spectrograms are input to the ResNet18 network to predict the result.

(2) Experimental program and result analysis

To verify the effectiveness of spectrograms generated by the GANs network, 150 original spectrograms of each character are input to the GANs network, and additional voiceprint features are generated according to different ratios of the original spectrograms. The additional voiceprint features generated by the GANs are fused with the original speech spectrogram for experimental comparison, and eleven groups are designed according to the fusion ratio and labeled as Groups 1~11. The number of the original spectrograms in each group is 1440, and the new voiceprint features are generated for each group at ratios of 0%, 5%, …, and 50%, respectively. Group 1 includes only the original spectrograms. Group 2 consists of 90 new and 1440 original spectrograms. Group 3 contains 180 new and 1440 original spectrograms. The eleven sets of data are used to train the ResNet18 network. A data set of 30 spectrograms pre-separated from the original spectrograms at 20% is used for testing the recognition rate. The numbers of new and total spectrograms in the eleven groups are listed in Table IV.

The relationship between the recognition rate and fusion ratio is shown in Fig. 8. Experimental simulations are performed for the nine data sets, and the results are listed in Table. V.

TABLE. III
Training Options and Parameters

Training options	Parameters
Initial learning rate	0.01
Number of training rounds	30
Number of classifications	12
Validation frequency	20
Batch size	64

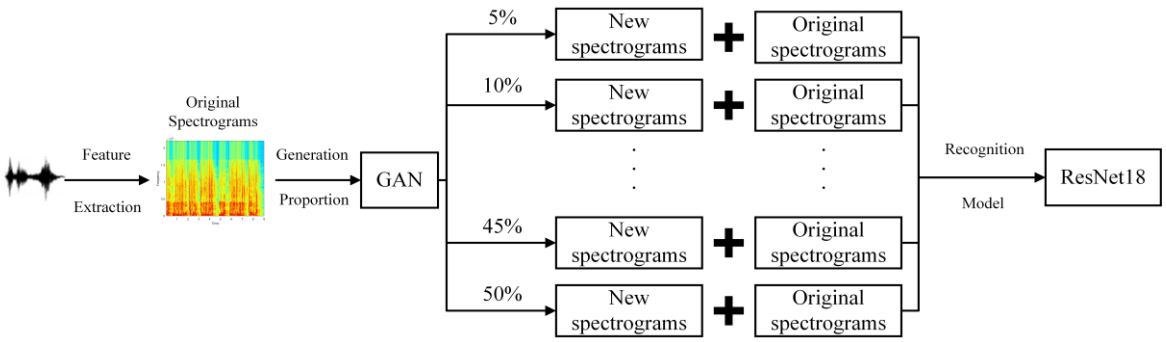


Fig.7 Architecture of the proposed framework

TABLE. IV
THE NUMBER OF NEW SPECTROGRAMS AND TOTAL SPECTROGRAMS

Group	Fusion proportion	Number of new spectrograms	Total spectrograms
1	0%	0	1440
2	5%	90	1530
3	10%	180	1620
4	15%	270	1710
5	20%	360	1800
6	25%	450	1890
7	30%	540	1980
8	35%	630	2070
9	40%	720	2160
10	45%	810	2250
11	50%	900	2340

TABLE. V
RECOGNITION RATE IN DIFFERENT FUSION RATIOS

Group	Fusion proportion	Recognition rate	Increase Rate
1	0%	96.39%	
2	5%	97.22%	0.83%
3	10%	97.47%	1.08%
4	15%	97.78%	1.39%
5	20%	97.92%	1.53%
6	25%	97.5%	1.11%
7	30%	97.22%	0.83%
8	35%	96.95%	0.56%
9	40%	96.63%	0.24%
10	45%	96.52%	0.15%
11	50%	96.29%	-0.10%

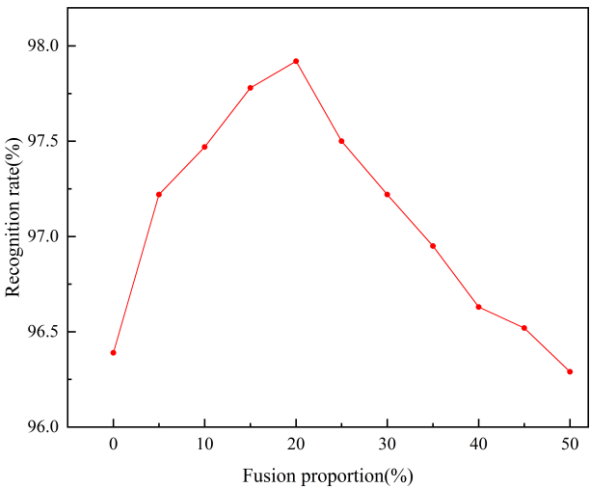


Fig. 8 Fusion proportion vs. recognition rate

(1) At fusion ratios below 40%, the recognition rate of Group 2~10 with fusion features is improved to a certain extent compared with Group 1, including only the original spectrograms. Further calculations show no significant improvement in the recognition rate at fusion ratios exceeding 40%.

(2) At fusion ratios below 20%, the recognition rate increases with the fusion ratio. The reason is that the new spectrograms generated by the GANs networks combine different features of the original spectrograms, thus enriching the diversity of the training spectrograms. And the new spectrogram is trained as an additional feature, which effectively expands the number of samples and solves the problem of insufficient sample size. The recognition rate reaches its peak at a 20% fusion ratio, exceeding that of the original spectrograms by 1.53%.

(3) At fusion ratios exceeding 20%, the recognition rate starts to decrease with the fusion ratio. At a 50% fusion ratio, the recognition rate is already lower than that of the original spectrogram data. This can be attributed to the fact that at fusion ratios exceeding a certain value, an excessive number of new spectrograms with lower resolution and graphical quality are generated. Another reason may be the insufficient number of samples at too high fusion ratios; the generated results will produce overfitting, affecting the recognition accuracy when speaker recognition is performed.

Figure 9 shows the relationship between the recognition rate and the number of iterations of several data sets for clarity. The recognition rate tends to stabilize after 15 iterations, the model converges, and the recognition rate obtained from the five data sets differs. The fusion feature method based on the generative adversarial network proposed in this paper can effectively improve the recognition rate.

To further verify the effectiveness of this method, its recognition rate is compared with those of several other speaker recognition methods for the experimental results, as listed in Table VI. The comparative analysis of the results in Table VI shows that the recognition performance of Group 1 before feature fusion in the experiments of this paper is lower than the two classical models, D-VECTOR and X-VECTOR, while as the fusion ratio grows, the recognition performance of the different groups is gradually better than that of the existing models. Group 3 with 10% fusion ratio is higher than D-VECTOR and lower than X-VECTOR in terms of recognition performance, but Group 5 with 20% fusion ratio is higher than all the compared models. This proves that the proposed method of feature fusion based on a GANs network mitigates the problem of low recognition rates due to insufficient data samples.

TABLE. VI
COMPARISON OF RESULTS

Method	Recognition rate
MFCC+LSTM	95.86%
GROUP1	96.39%
D-VECTOR	96.46%
GROUP3	97.47%
X-VECTOR	97.62%
GROUP5	97.92%

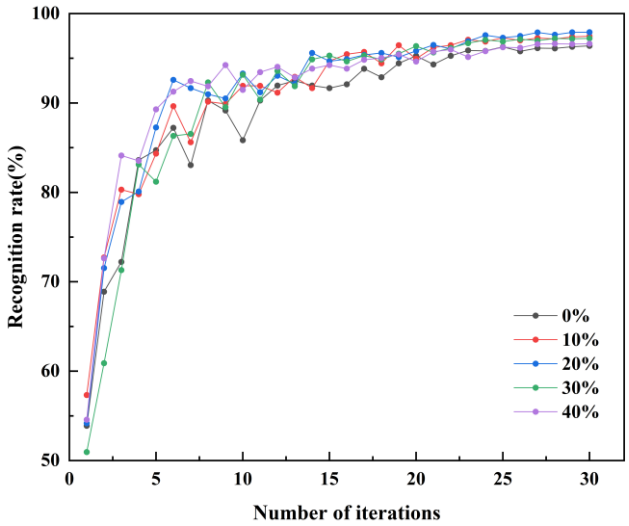


Fig. 9 Number of iterations vs. recognition rate for several dataset

VI Conclusion

Training of deep learning network models usually requires large data samples, implying that training results are less effective if the data samples are insufficient. In this paper, we propose a method of speaker recognition based on the GANs and ResNet18 network, given that the speech spectrogram contains rich voiceprint information, utilizing the speech spectrogram as voiceprint features. The GANs network generates additional voiceprint features according to different ratios of the original spectrogram, and the additional voiceprint features are fused with the original spectrogram for speaker recognition using the ResNet18. The simulation results show that in the example, the recognition rate is improved with the fusion ratio less than 40%; when the fusion ratio is larger than 40%, the recognition rate hardly improves but decreases instead. Therefore, it is appropriate to control the fusion ratio within a suitable range when using this method for voiceprint recognition. The proposed method of speaker recognition based on the GANs and feature fusion can effectively improve the recognition rate in the situation of insufficient training sample data. This study's findings provide a methodological basis and technical guidance for developing speaker recognition systems with high recognition rates.

References

[1] Costantini, G., Cesarini, V., & Brenna, E. (2023). High-Level CNN and Machine Learning methods for speaker recognition. Sensors (Basel, Switzerland), 23(7), 3461.

[2] Costantini, G., Cesarini, V., & Brenna, E. (2023). High-Level CNN and Machine Learning methods for speaker recognition. Sensors (Basel, Switzerland), 23(7), 3461. https://doi.org/10.3390/s23073461

[3] Tiwari, V. (2010). MFCC and its applications in speaker recognition. International journal on emerging technologies, 1(1), 19-22.

[4] Ye, F., & Yang, J. (2021). A deep neural network model for speaker identification. Applied Sciences (Basel, Switzerland), 11(8), 3603. https://doi.org/10.3390/app11083603

[5] Chen, Y.-L., Wang, N.-C., Ciou, J.-F., & Lin, R.-Q. (2023). Combined bidirectional long short-term memory with Mel-frequency cepstral coefficients using autoencoder for speaker recognition. Applied Sciences (Basel, Switzerland), 13(12), 7008. https://doi.org/10.3390/app13127008

[6] Tsai, T.-H., Hao, P.-C., & Wang, C.-L. (2021). Self-defined text-dependent wake-up-words speaker recognition system. IEEE Access: Practical Innovations, Open Solutions, 9, 138668–138676. https://doi.org/10.1109/access.2021.3117602

[7] Mohammadi, M., & Mohammadi, H. R. S. (2017, May). Robust features fusion for text independent speaker verification enhancement in noisy environments. In: 2017 Iranian Conference on Electrical Engineering (ICEE), pp. 1863-1868. https://doi.org/10.1109/IranianCEE.2017.7985357

[8] Wang, Z., & Hansen, J. H. L. (2022). Multi-source domain adaptation for text-independent forensic speaker recognition. ACM Transactions on Audio, Speech, and Language Processing, 30, 60–75. https://doi.org/10.1109/taslp.2021.3130975

[9] Shome, N., Saritha, B., Kashyap, R., & Laskar, R. H. (2023). A robust DNN model for text-independent speaker identification using non-speaker embeddings in diverse data conditions. Neural Computing & Applications, 35(26), 18933–18947. https://doi.org/10.1007/s00521-023-08736-1

[10] Meuwly, D. (2013). Forensic speaker recognition. In Wiley Encyclopedia of Forensic Science pp. 1–5. https://doi.org/10.1002/9780470061589.fsa147.pub2

[11] Velayuthapandian, K., & Subramoniam, S. P. (2023). A focus module-based lightweight end-to-end CNN framework for voiceprint recognition. Signal, Image and Video Processing, 17(6), 2817–2825. https://doi.org/10.1007/s11760-023-02500-7

[12] Hansen, J. H. L., & Hasan, T. (2015). Speaker Recognition by Machines and Humans: A tutorial review. IEEE Signal Processing Magazine, 32(6), 74–99. https://doi.org/10.1109/msp.2015.2462851

[13] Demircan, S., Örnek, H.-K. (2020). Comparison of the effects of mel coefficients and spectrogram images via deep learning in emotion classification. Traitement du Signal. https://hdl.handle.net/20.500.13091/426

[14] Bai, Z., & Zhang, X.-L. (2021). Speaker recognition based on deep learning: An overview. Neural Networks: The Official Journal of the International Neural Network Society, 140, 65–99. https://doi.org/10.1016/j.neunet.2021.03.004

[15] Zhang, Y., Zhang, Z. (2022). Application of DenseNet in voiceprint recognition. Computer Engineering & Science, 44(01), 132.

[16] Jia, Y.-J., Xi Chen, Yu, J.-Q et al. (2019). Fast speaker recognition based on characteristic spectrogram and an adaptive self-organizing feature map. Science Technology and Engineering, 19 (15), 211-218.

[17] Zhu, M.-F., Wang, Z.-C., Dai, S.-B. (2022). Gender and Age Recognition Based on Spectrogram and Dense Convolutional Neural Network. Instrumentation Technology, 66-70.

[18] Shafik, A., Sedik, A., Abd El-Rahiem, B., El-Rabaie, E. S. M., El Banby, G. M., Abd El-Samie, F. E., ... & Iliyasu, A. M. (2021). Speaker identification based on Radon transform and CNNs in the presence of different types of interference for Robotic Applications. Applied Acoustics, 177, 107665. https://doi.org/10.1016/j.apacoust.2020.107665

[19] Chen, M., Shi, Y., & Hain, T. (2021, June). Towards low-resource stargan voice conversion using weight adaptive instance normalization. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5949-5953. https://doi.org/10.1109/icassp39728.2021.9415042

[20] He, X., Chen, J., Rizos, G., & Schuller, B. W. (2021). An improved stargan for emotional voice conversion: Enhancing voice quality and data augmentation. arxiv preprint arxiv, 107, 08361. https://doi.org/10.48550/arXiv.2107.08361

[21] Rizos, G., Baird, A., Elliott, M., & Schuller, B. (2020, May). Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3502-3506. http://doi.org/10.1109/ICASSP40776.2020.9054579

[22] Cao, Y., Liu, Z., Chen, M., Ma, J., Wang, S., & Xiao, J. (2020). Nonparallel emotional speech conversion using VAE-GAN. In: Interspeech, pp. 3406-3410. http://doi.org/10.21437/Interspeech.2020-1647

[23] Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018). Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks. In: 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 266-273. http://doi.org/10.1109/SLT.2018.8639535

[24] Veit, A., Wilber, M., & Belongie, S. (2016). Residual networks behave like ensembles of relatively shallow networks. In: arXiv [cs.CV]. http://arxiv.org/abs/1605.06431

[25] Lukic, Y., Vogt, C., & Dürr, O. (2016). Speaker identification and clustering using convolutional neural networks. In: 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP). https://doi.org/10.1109/MLSP.2016.7738816

[26] Li, B. (2011). On identity authentication technology of distance education system based on voiceprint recognition. In :2011 IEEE 30th Chinese Control Conference, pp. 5718–5721.

[27] Jahangir, R., TEh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., Akhtar, M. Z., & Ali, I. (2020). Text-independent speaker iden

tification through feature fusion and deep neural network. IEEE Access: Practical Innovations, Open Solutions, 8, 32187–32202. https://doi.org/10.1109/access.2020.2973541

[28] Cheng, S., Shen, Y., & Wang, D. (2022). Target speaker extraction by fusing voiceprint features. Applied Sciences (Basel, Switzerland), 12(16), 8152. https://doi.org/10.3390/app12168152

[29] Gabor, D. (1947). Theory of communication. Journal of the Institution of Electrical Engineers - Part I General, 94(73), 58–58. https://doi.org/10.1049/ji-1.1947.0015

[30] Li, D., Yang, Z., Wang, Z., & Yang, H. (2023). Identity Retention and Emotion Converted StarGAN for low-resource emotional speaker recognition. Speech Communication, 151, 39–51. https://doi.org/10.1016/j.specom.2023.05.007