Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 6, No. 2, 2023

Research on Voiceprint Recognition based on Densenet

Yongkang Tang, Xiangguo Sun

School of Mechanical Engineering, Sichuan University of Science & Engineering, Yibin Sichuan, 644002, China

Abstract: To date, most methods for voiceprint recognition are implemented based on convolutional neural networks. In order to combine the spatial features and temporal features of sound and improve computational efficiency, a combined model of DenseNet169 and Bidirectional Gated Recurrent Unit (BiGRU), named DenseNet169-BiGRU, is proposed. Firstly, deep separable convolution is employed to reduce the number of network parameters and enhance computational efficiency. The speech FBank feature map is extracted after preprocessing, and experimental results show that the DenseNet169-BiGRU model outperforms both DenseNet169 and DenseNet169-GRU models, with an improvement in accuracy rate of 2.7% compared to DenseNet169, and an improvement of 1.1% compared to Densenet-GRU. This validates the effectiveness of the proposed method in both improving computational efficiency and achieving good performance.

Keywords: Voiceprint Recognition; Depth Separable Convolutional; FBank; DenseNet; BiGRU.

1. Introduction

Voiceprint recognition, also known as speaker identification, is able to extract sound features from a piece of speech and identify the speaker's identity[1]. Speech is a unique physiological behavior of humans, and for each individual, their speech has its own biological characteristics, such as physiological structural differences caused by genetic variations, voice differences formed after birth, and gender differences[2], among others. Similar to speech recognition, both require processing of speech signals to extract their features. The feature map is used as input to train deep learning models, enabling the models to learn information from the feature map to be used for prediction and classification[3,4]. These features include the vibrational cycle of the vocal cords and the frequency range of the sound, among others. Due to the high security and the ability to collect speech features without user contact, the market share of voiceprint recognition has increased and is being applied to a wider range of fields.

Speaker recognition can be divided into two tasks: speaker verification and speaker identification. Speaker verification is a key technology for intelligent interaction and can be widely applied in fields such as financial payments, criminal investigations, and national defense[5]. Speaker identification is the process of determining a specific reference speaker based on the speech content. Voiceprint recognition with predefined speech content is called text-dependent speaker recognition, while voiceprint recognition that does not require predetermined speech content is called text-independent speaker recognition. The focus of this study is on text-independent speaker identification.

In the past, the main technique used for voiceprint recognition was the Gaussian Mixture Model-Universal Background Model (GMM-UBM)[6,7]. In recent years, with the emergence of spectrograms, researchers have proposed combining spectrograms with Convolutional Neural Networks (CNN) for speaker identification[8–10]. Spectrograms[11] contain rich information about the voice characteristics of speakers. By extracting features from spectrograms using CNN, the accuracy of voiceprint recognition has been significantly improved. In literature[12], an LFBANK feature was designed specifically for female voiceprint recognition, and the results show that this feature has advantages in recognizing female voices. Literature[13] proposed a CNN-LSTM model based on global attention mechanism, which simultaneously considers spatial and temporal features of speech, and uses different weights in LSTM hidden layers through attention mechanism, resulting in noticeable improvement in accuracy compared to CNN and CNN-LSTM models. Literature[14] introduced a voiceprint recognition method combining knowledge distillation and ResNet, which achieved significant error rate reduction. Literature[15] presented an end-to-end voiceprint recognition algorithm based on LSTM, which shortened the model training time and achieved good results. Literature[16] proposed an end-to-end voiceprint recognition method based on stacked bidirectional LSTM with non-linear layers, which extracts deeper abstract features of speech signals using stacked multiple layers of bidirectional LSTM and non-linear layers, resulting in better recognition performance than unidirectional GRU and unidirectional LSTM.

2. Speech Feature Extraction

Currently, two commonly used features in voiceprint recognition are MFCC and FBANK. The human ear's response to sound spectra is nonlinear. The FBANK algorithm processes audio in a similar way to the human ear, fitting the characteristics received by the ear. MFCC is actually obtained from FBANK through Discrete Cosine Transform (DCT), which is a linear transformation that leads to the loss of many nonlinear components in speech signals. In the past, MFCC features combined with GMMs-HMMs methods were frequently used in speech recognition. With the emergence of deep learning algorithms, neural networks are not sensitive to highly correlated information. As a result, FBANK has gained broader application. In this paper, FBANK is used as the speech feature. The process of extracting FBANK feature maps is shown in the following figure.

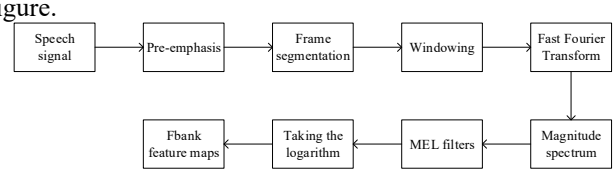

Figure 1. Speech feature extraction process

Taking a speech segment from the dataset used in this paper as an example, the speech content is "Green is the background color of the great Spring scenery". The duration of the speech is 3.5 seconds. First, the speech signal is processed to obtain the temporal information, and the corresponding temporal graph is shown in Figure 2. The horizontal axis represents time, and the vertical axis represents amplitude. It can be observed from the graph that the amplitude varies over time. The temporal signal is then subjected to the following processing steps as input.

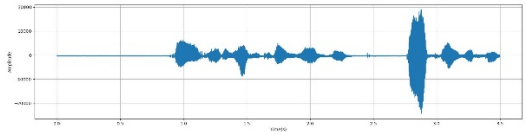

Figure 2. Time-domain graph

(1) Pre-emphasis is a process where the input speech signal is passed through a high-pass filter to compensate for high-frequency signals in the speech signal and flatten the spectrum. The formula is as follows:

$$g(t) = f(t) - \alpha f(t-1) \tag{1}$$

f(t) represents the original speech signal, g(t) represents the pre-emphasis result, α represents the pre-emphasis coefficient, with typical values of 0.95 or 0.97. In this paper, the value of α is taken as 0.97. After pre-emphasis processing, the temporal graph is shown in Figure 3, which is flatter compared to Figure 2.

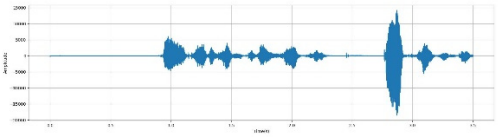

Figure 3. Time-domain graph after pre-emphasis processing

(2) Frame segmentation is performed to facilitate subsequent short-time Fourier transform. The speech signal is divided into small segments of equal duration, with each segment referred to as a frame. To ensure smooth transitions between adjacent frames, there is an overlap region between them. In this paper, the duration of each frame is 25ms, and the overlap region is 10ms. The waveform of the first frame after frame segmentation is shown in Figure 4.

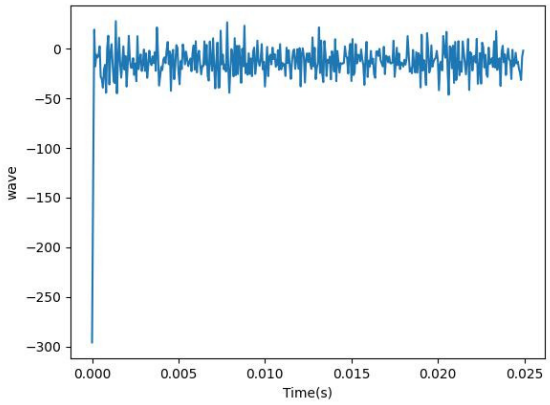

Figure 4. Waveform graph of the first frame

(3) Windowing is performed in this paper using a Hamming window. Each frame of the speech signal is multiplied by the window function to enhance continuity at the edges of each frame. The formula is shown in equation (2), where $w(n)$ represents the Hamming window and is set to 0.46. Figure 5 shows the result of the first frame of the speech signal after windowing. It can be observed that the peaks in the spectrum become narrower, which helps reduce spectral leakage. However, the amplitude of the signal at the edges of each frame is attenuated. To compensate for this, overlapping regions are introduced between adjacent frames during the frame segmentation process.

$$\mathrm{W}(n,\alpha) = (1-\alpha) - \alpha \times \cos(\frac{2\pi n}{N-1}) \tag{2}$$

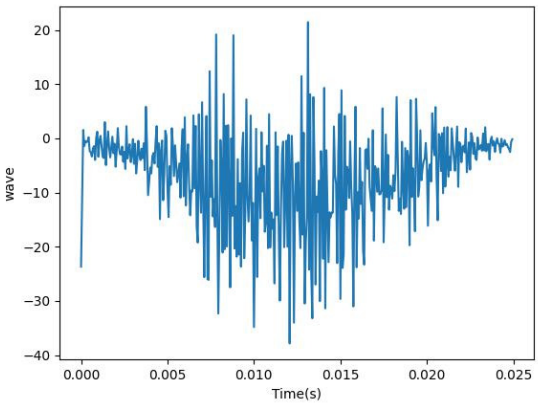

Figure 5. Waveform graph of the first frame after windowing processing

(4)Fourier Transform is used in this case to convert the signal in the time domain into the energy distribution in the frequency domain. Since the speech waveform exhibits periodicity only in the short-time domain, Short-Time Fourier Transform (STFT) is employed. The formula is as follows:

$$S_i(k) = \sum_{n=1}^{N} s_i(n) \exp\{-2j\pi \frac{kn}{N}\}, 1 \le k \le K \tag{3}$$

(5)The magnitude spectrum is calculated using the following formula, where FFT(x) represents the result obtained from the Fourier Transform.

$$P = \frac{\left|FFT(x_i)\right|^2}{N} \tag{4}$$

(6)MEL filter bank is used to simulate the characteristics of human hearing and extract frequency bands. The formula for converting MEL frequency to speech frequency is as follows:

$$f_{mel} = 2595 * \log_{10}(1 + \frac{f}{700}) \tag{5}$$

$$f = 700(10^{\frac{f_{mel}}{2595}} - 1) \tag{6}$$

After the above steps, the FBANK feature map can be obtained. The following image shows the FBANK feature maps obtained when two people pronounce the same speech. The feature map contains a lot of important information. The color variation in the image represents the change in speech energy, where darker colors indicate stronger energy. The vertical axis corresponding to the lowest frequency range of the horizontal stripes represents the fundamental frequency, which can be used to determine the fundamental period, i.e., the vibration period of the vocal cords. Additionally, the darker stripes at a particular moment represent the formants, which refer to the resonant frequencies of the vocal tract and reflect the physical characteristics of the vocal tract.

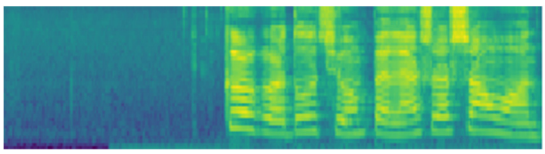

Figure 6. FBANK feature map of Speaker 1

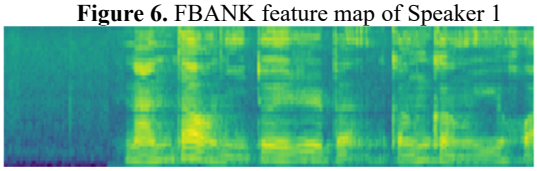

Figure 7. FBANK feature map of Speaker 2

3. Voiceprint Recognition based on Densenet169-BiGRU

3.1. DenseNet Network

DenseNet, which belongs to a type of convolutional neural networks, can be considered as an extension of residual networks. Similar to residual networks, the fundamental building block of DenseNet is a residual block. The basic structure is shown in Figure 8, where it can be observed that the residual block allows for faster forward propagation of the input.

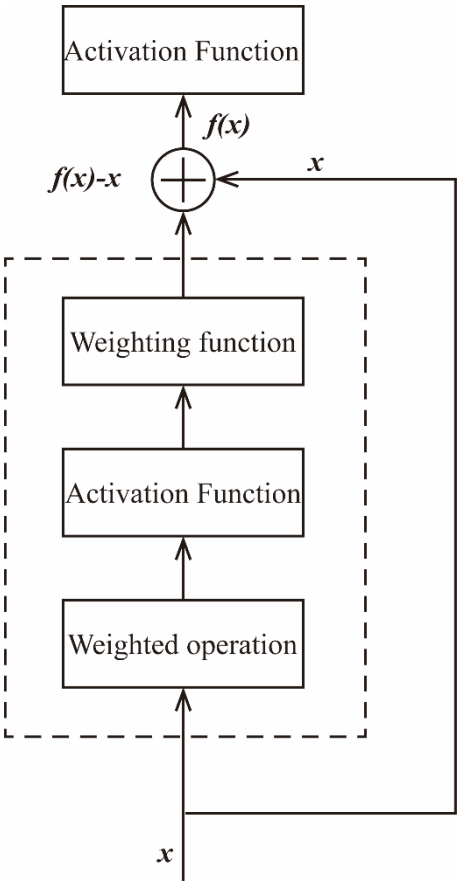

Figure 8. Residual block

The difference between DenseNet and ResNet lies in the way their outputs are combined. While ResNet uses summation, DenseNet uses concatenation. All layers are connected together, meaning that if a DenseNet has K layers, there will be K(K+1)/2 connections in the network. It is because all layers are connected that features are fully utilized, the number of parameters is reduced, and the issue of gradient vanishing is alleviated. The DenseNet connectivity pattern is shown in Figure 9.

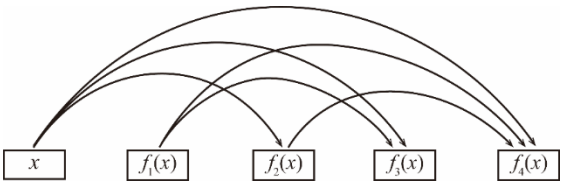

Figure 9. Dense connection

In this study, the Densenet169 network is adopted, and its simplified structure diagram is shown in Figure 10, where 169 represents the sum of convolutional layers and fully connected layers.

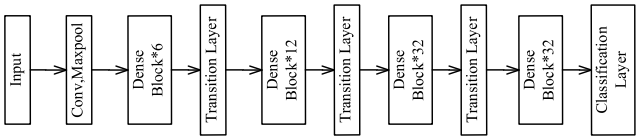

Figure 10. Schematic diagram of DenseNet169 structure

3.2. Optimization of DenseNet Network Structure

As the number of network layers increases, the number of parameters also increases. To address this issue, this study optimizes the DenseNet169 network using the Depth Separable Convolution (DSC) method. DSC is an effective approach to reduce the number of network parameters[17]. The method of Depth Separable Convolution involves performing a depthwise convolution on the input, where each channel of the input is convolved with a separate convolutional kernel. However, due to the individual convolutions across channels, the feature information of different channels at the same location is not fully utilized. Therefore, after the depthwise convolution, a pointwise convolution is applied, where the kernel size is 1*1, aiming to reduce the number of channels and combine the results of the depthwise convolution. The improved DenseNet169 network, compared to the original DenseNet169 structure before optimization, is shown in Figure 11. The improvement involves replacing each 3*3 convolution in each DenseBlock with a 3*3 depthwise convolution and a 1*1 pointwise convolution.

Layers	Output size	DenseNet169	DenseNet169+DSC
Convolution	112*112	7*7conv，stride2	
Pooling	56*56	3*3maxpool，stride2	
DenseBlock（1）	56*56	[1×1 conv; 3×3 conv] ×6	[1×1 conv; 3×3 DWconv; 1×1 PWconv] ×6
Tansition Layer（1）	56*56	1*1conv	
	28*28	2*2average，stride2	
DenseBlock（2）	28*28	[1×1 conv; 3×3 conv] ×12	[1×1 conv; 3×3 DWconv; 1×1 PWconv] ×12
Tansition Layer（2）	28*28	1*1conv	
	14*14	2*2average，stride2	
DenseBlock（3）	14*14	[1×1 conv; 3×3 conv] ×32	[1×1 conv; 3×3 DWconv; 1×1 PWconv] ×32
Tansition Layer（3）	14*14	1*1conv	
	7*7	2*2average，stride2	
DenseBlock（4）	7*7	[1×1 conv; 3×3 conv] ×32	[1×1 conv; 3×3 DWconv; 1×1 PWconv] ×32
Classfication Layer	1*1	7*7 globle average pool	
		1000D fully connected，softmax	

Figure 11. Network structure comparison graph

Assuming the input size is H* W*M and the number of output channels is N, with a convolution kernel size of K*K and padding size of P. For regular convolution, the formula for calculating the number of parameters Yc is given by

equation 7, and the formula for calculating the computational cost Fc is given by equation 8.

$$M \times K \times K \times N \tag{7}$$

$$(H - K + 2P + 1) \times (W - K + 2P + 1) \times M \times K \times K \times N \tag{8}$$

The parameter quantity YDSC of depth separable convolution consists of the parameter quantities of depthwise convolution and pointwise convolution, and the formula is as follows (Equation 9):

$$M \times K \times K + N \times 1 \times 1 \times M \tag{9}$$

Similarly, the computational cost FDSC is calculated as Equation 10:

$$(H-K+2P+1) \times (W-K+2P+1) \times K \times K \times M + 1 \times 1 \times H \times W \times M \times N \tag{10}$$

In this model, K is 3 and P is 1. Substituting the values into the formula, we can obtain the ratio of computational cost:

$$\frac{F_{DSC}}{F_c} = \frac{Y_{DSC}}{Y_c} = \frac{1}{N} + \frac{1}{9} \tag{11}$$

From the above equation, it can be seen that using depth separable convolution can significantly reduce the computational cost. The parameter comparison obtained from experimental results is shown in Table 1.

Table 1. Parameter comparison

NET	Batch Size	Input Size (MB)	Total Params	Params Size (MB)
Densenet169	64	39.35	12617680	50.47
Densenet169+DSC	64	29.35	10368560	41.23

3.3. GRU Network

The gated recurrent unit (GRU) belongs to the family of recurrent neural networks (RNN) [18]. It is an improved version based on the long short-term memory (LSTM) network. Both GRU and LSTM introduce a "gate" mechanism on top of the basic RNN[18], which allows for determining which data should be retained or discarded. The simplified structure of GRU is shown in Figure 12. GRU simplifies the three gate functions in LSTM to two gate functions. The function zt represents the update gate, where a higher value indicates more information from the previous state being incorporated into the current state. The function rt represents the reset gate, which determines how much information from the previous state needs to be forgotten.

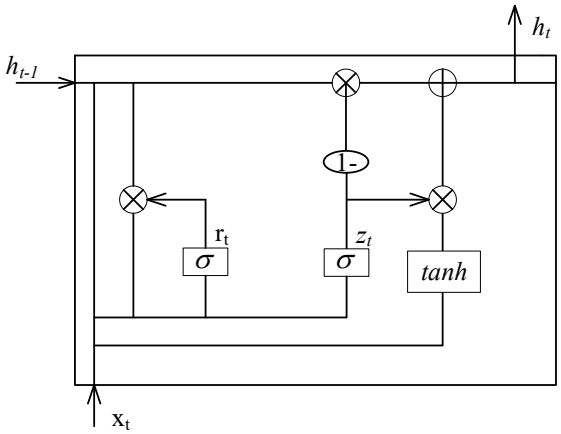

Figure 12. Schematic diagram of the GRU structure

Compared to LSTM, GRU has fewer parameters, higher computational efficiency, and faster training time. However, since the data in GRU is unidirectional, predictions can only be based on previous data. Bidirectional Gated Recurrent Unit (BiGRU) consists of two GRUs with opposite directions, allowing the utilization of information from both previous and future contexts. This leads to more accurate predictions. The structure of BiGRU is shown in Figure 13, where the input information is simultaneously fed into two GRU networks with opposite directions, and the output is determined by both GRU networks.

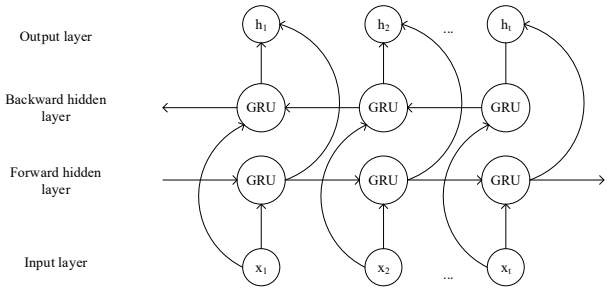

Figure 13. Schematic diagram of the BiGRU structure

3.4. DenseNet169-BiGRU network

The working principle of DenseNet169-BiGRU network is shown in Figure 14. The two networks are concatenated, where the feature information is input into DenseNet169 for feature extraction and dimensionality reduction through convolutional calculations. The output information serves as the input for BiGRU network, which learns the temporal dependencies among the data. The Conca layer integrates the output information from the two GRU networks with opposite directions, and the results are produced through a classifier.

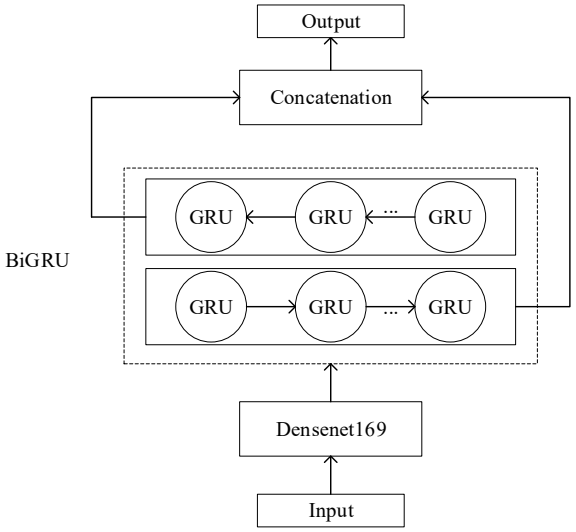

Figure 14. Schematic diagram of the DenseNet169-BiGRU structure

4. Voiceprint Recognition Experiment

4.1. Data Preparation

The speech dataset used in the experiments includes

THCHS-30 from Open Speech and Language Resources and Free ST Chinese Mandarin Corpus. THCHS-30, released by the Center for Speech and Language Technology (CSLT) at Tsinghua University, is an open Chinese speech database. The participants in the Thchs30 Chinese dataset are university students who are fluent in Mandarin. The recorded texts are derived from a large corpus of news, with a total duration of 30 hours[19]. The latter dataset, Free ST Chinese Mandarin Corpus, is provided free-of-charge by Surfingtech (www.surfing.ai). It consists of speech data from 855 speakers, with 120 speech segments per person, totaling 102,600 speech segments. The audio is sampled at a frequency of 16KHZ. In this study, 80 participants (equal number of male and female) were involved. Each participant had 110 speech segments, with the same content. The distribution of speech segments for each individual is shown in Table 2. The training, validation, and testing set proportions are set to 9:1:1.

Table 2. Distribution of speech data

Speaker ID	Quantity	Average duration(seconds)	sampling frequency (KHz)
1	110	3.5	16
2	110	3.5	16
...	...	...	...
80	110	3.5	16

4.2. Evaluation Criteria

This article selects accuracy as the evaluation criteria, and the calculation formula is as follows: TP represents the samples that are positive and are predicted as positive by the model. FN represents the data samples that are positive but predicted as negative by the model. FP represents the data samples that are negative but identified as positive by the model. TN represents the data samples that are negative and predicted as negative by the model.

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

4.3. Experimental Results and Analysis

In order to verify the performance of the model, the network model constructed in this paper was compared with Densenet169 and Densenet169-gru in terms of validation set accuracy. The results are shown in Figure 15. The validation accuracy of DneseNet169-BiGru with a dual-layer Gru reached 98.9%, and the test accuracy was 97.3%. The validation accuracy of DneseNet169-Gru with a single-layer Gru was 97.3%, and the test accuracy was 96.3%. The validation accuracy of DenseNet169 was 95.9%. The test accuracy was 94.6%. This also indicates that the recognition of voice features not only involves the temporal aspect but also the importance of data before and after a certain moment. Moreover, using this information can further improve the accuracy. Figure 16 shows the confusion matrix of the test results of DneseNet169-BiGru, and the accuracy is the ratio of the sum of the diagonal numbers to the total number of the test samples. Next, the accuracy of the test set is compared with the proposed methods in voiceprint recognition research, and the results are shown in Table 3.

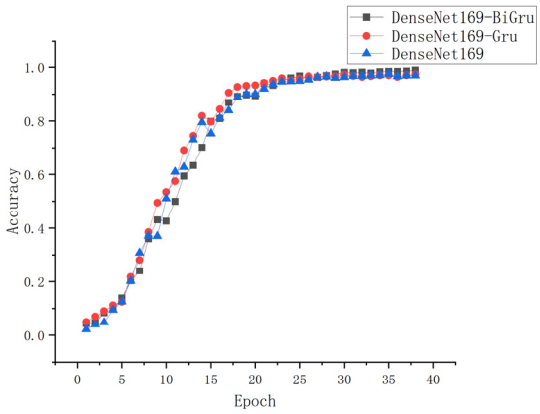

Figure 15. Validation accuracy curves of three network models

Table 3. Distribution of speech data

NET	Accuracy (%)
DneseNet169-BiGru	97.3%
DenseNet169-Gru	96.2%
DenseNet121-BiGru CNN-LSTM Resnet34-se	96.3% 95.42% 90.3

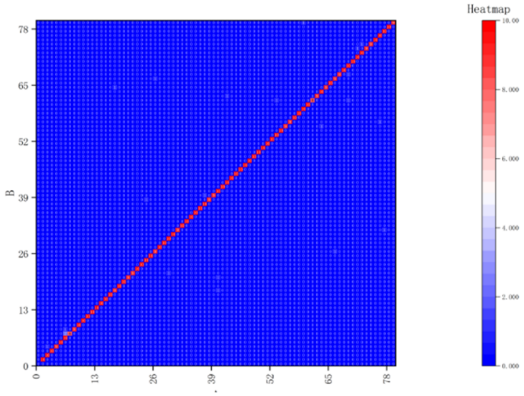

Figure 16. Confusion matrix of DenseNet169-BiGRU

5. Conclusion

This paper proposes a joint model combining DenseNet169 and a two-layer GRU network, and adopts depth-wise separable convolution with fewer model parameters to improve computational efficiency. The experimental results show that the model performs well with a test accuracy of 97.3%. In order to verify the performance, experiments were conducted comparing the model with DenseNet169 and DenseNet169-BiGRU, as well as models proposed by other authors. The advantages of the model in this paper can also be seen from the table.

References

[1] Chen Guo-guo,Chen Jia-yu,Na Xing-yu'et al. Kaldi speech recognition combat[M]. 1st ed. Beijing:Publishing House of Electronics Industry 2020. (in Chinese)

[2] HANSEN J H L, HASAN T. Speaker Recognition by Machines and Humans: A tutorial review[J/OL]. IEEE Signal Processing Magazine, 2015, 32(6): 74-99. DOI:10.1109/ MSP. 2015. 246 2851.

[3] LIU D, XU J, ZHANG P, et al. A unified system for multilingual speech recognition and language identification

[J/OL]. Speech Communication, 2021, 127: 17-28. DOI: 10. 1016/ j.specom.2020.12.008.

[4] ABLIMIT M, XUELI M, HAMDULLA A. Language Identification Research Based on Dual Attention Mechanism[C/OL]//2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML). 2021: 241-246[2023-12-10]. https:// ieeexplore. ieee.org/ document/ 9520699. DOI:10.1109/PRML52754.2021.9520699.

[5] ZHANG C, KOISHIDA K, HANSEN J. Text-Independent Speaker Verification Based on Triplet Convolutional Neural Network Embeddings[J/OL]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26: 1-1. DOI:10.1109/TASLP.2018.2831456.

[6] SARKAR A, TAN Z H. Text Dependent Speaker Verification Using un-supervised HMM-UBM and Temporal GMM-UBM[C/OL]. 2016. DOI:10.21437/Interspeech.2016-362.

[7] HAO M, LIU H, LI Y, et al. Speech Tracking Based on Cluster Analysis and Speaker Recognition[J/OL]. Computer and Modernization, 2020, 4: 11-17. DOI:10.3969/j.issn.1006-2475. 2020.04.002.

[8] LIU Z, WU Z, LI T, et al. GMM and CNN Hybrid Method for Short Utterance Speaker Recognition[J/OL]. IEEE Transactions on Industrial Informatics, 2018, PP: 1-1. DOI: 10. 1109/ TII.2018.2799928.

[9] LUO K, FU L. Research and application of voiceprint recognition based on a deep recurrent neural network[M/OL]. 2019: 309-316. DOI:10.1201/9780429468605-43.

[10] HE Y, YINGYAN D, PENG W, et al. VOICEPRINT RECOGNITION BASED ON CNN-LSTM NETWORK [J/OL]. Computer Applications and Software, 2019 [2023-12-15]. http://en.cnki. com.cn/Article_ en/CJFD Total-JYRJ 2019 04027.htm.

[11] MA Y, YUAN M, QI C, et al. Reacher of Feature Extratcion from spectrogram Based on Pulse Coupled Neural Network in Speaker Recognition[J/OL]. 2005(20)[2023-12-11]. https:// ir. lzu. edu.cn//handle/262010/127329. DOI:10/127329.

[12] CUI L, WANG Z. Study on Voiceprint Recognition Based on Mixed Features of LFBank and FBank[J]. Computer Science, 2022, 49(S2): 621-625.

[13] CHU X, YANG H, YAN D, et al. Research on CNN-LSTM Speaker Recognition Based on Global Attention Mechanism [J/OL]. Audio Engineering, 2022, 46(2): 38-45. DOI:10. 16311/ j.audioe.2022.02.009.

[14] RONG Y, FANG Y, TIAN P, et al. Voiceprint recognition based on knowledge distillation and ResNet[J]. Journal of Chongqing University, 2023, 46(1): 113-124.

[15] WANG F, XU Y. LSTM-Based End-to-End Voiceprint Recognition Algorithm Implementation[J/OL]. Software Engineering and Applications, 2021, 10: 467. DOI:10. 12677/ SEA. 2021.104052.

[16] WANG ZHI-YUE C L. End to End Voiceprint Recgonition Based on Nonlinear Stacked Bidirectional Network[J]. Computer and Modernization, 2022, 0(03): 13.

[17] ZHANG Y, ZHANG Z. Application of DenseNet in voiceprint recognition[J]. Computer Engineering & Science, 2022, 44(1): 132-137.

[18] JIANG G, FU Y. Simulation of User Login Speech Recognition Model Based on Multi-Task Training[J]. Computer Simulation, 2022, 39(9): 190-194.

[19] LIU X, SONG W, CHEN X, et al. BLSTM-CTC SPEECH RECOGNITION BASED ON MULTI-CORE CONVOLUTIONAL FUSION NETWORK[J]. Computer Applications and Software, 2021, 38(11): 167-173.