Research on End-to-end Voiceprint Recognition Model Based on Convolutional Neural Network

Hong Zhao[1], Lupeng Yue[1,*], Weijie Wang [1] and Xiangyan Zeng [2]

[1]*School of Computer Science, Lanzhou University of Technology, Gansu, Lanzhou, 730050, China*
[2]*Department of Mathematics and Computer Science, Fort Valley State University, Fort Valley, GA 31030, Georgia*
E-mail: 594286500@qq.com; beifuyangguang@126.com; 1132744259@qq.com; zengx@fvsu.edu
[*]*Corresponding Author*

Received 27 April 2021; Accepted 01 June 2021;
Publication 28 April 2021

Abstract

Speech signal is a time-varying signal, which is greatly affected by individual and environment. In order to improve the end-to-end voice print recognition rate, it is necessary to preprocess the original speech signal to some extent. An end-to-end voiceprint recognition algorithm based on convolutional neural network is proposed. In this algorithm, the convolution and down-sampling of convolutional neural network are used to preprocess the speech signals in end-to-end voiceprint recognition. The one-dimensional and two-dimensional convolution operations were established to extract the characteristic parameters of Meier frequency cepstrum coefficient from the preprocessed signals, and the classical universal background model was used to model the recognition model of voice print. In this study, the principle of end-to-end voiceprint recognition was firstly analyzed, and the process of end-to-end voice print recognition, end-to-end voice print recognition

Journal of Web Engineering, Vol. 20_5, 1573–1586.
doi: 10.13052/jwe1540-9589.20511

features and Res-FD-CNN network structure were studied. Then the convolutional neural network recognition model was constructed, and the data were preprocessed to form the convolutional layer in frequency domain and the algorithm was tested.

Keywords: Convolutional neural network, end-to-end voiceprint recognition, voiceprint recognition model, speech signal, Res-FD-CNN network structure.

1 Introduction

Voiceprint recognition is a kind of technology that uses the speaker's voice to distinguish the speaker, so as to carry out identity identification and verification. As an important biometric identification/identification technology, voiceprint recognition technology is widely used in military security, financial field, judicial identification, voice dialing, telephone banking and many other fields [1]. With the increasing popularity of modern information technology, biometric technology, as a more rapid and natural means of identity authentication, has been more and more widely used in finance, education, social security, medical and health industries [2]. This technology extracts the human's own biometric characteristics for authentication, which can replace traditional passwords and Personal Identification numbers. It is not easy to steal and forget, and is also less vulnerable to counterfeiting than signature authentication. Due to the rapid growth of Internet data, traditional voiceprint recognition methods have been unable to meet the identification accuracy under the condition of large-scale data, and its framework can not meet the current development needs [3]. In the face of multi-channel, background noise, short time and long time problems, the traditional model has poor adaptability, and the steps are complicated, and the calculation is huge. The emergence of deep neural network technology has solved the problem of large-scale data processing to a certain extent. The design of the network framework, the complexity of the structure and the selection of various parameters also determine the final performance of the model [4]. The end-to-end neural network can Increasing the overall fit of the model while reducing network complexity has also become a current research hotspot [5]. Voiceprint recognition technology level unceasing enhancement, for security and identity encryption provides better barrier, people also enjoy more convenient way of life, based on a voiceprint recognition technology broad prospect of application and the present, the difficulties encountered in the

actual environment of research based on convolutional neural network end-to-end voiceprint recognition model has a wide range of social significance and economic significance.

2 End-to-end Voice Print Recognition Principle

2.1 The Process of Voiceprint Recognition

Voiceprint recognition system is mainly divided into three aspects: speech pre-processing, speech feature extraction and model selection. From the perspective of deep learning, voice print recognition can include two aspects [6]. One is training, the other is recognition. The basic tasks of the training stage are data preprocessing, feature extraction and model selection; The recognition part includes that the speech features of the speaker are recognized by the model, and the correct classification results are finally made according to the data features generated by the model. The process of voiceprint recognition is shown in Figure 1:

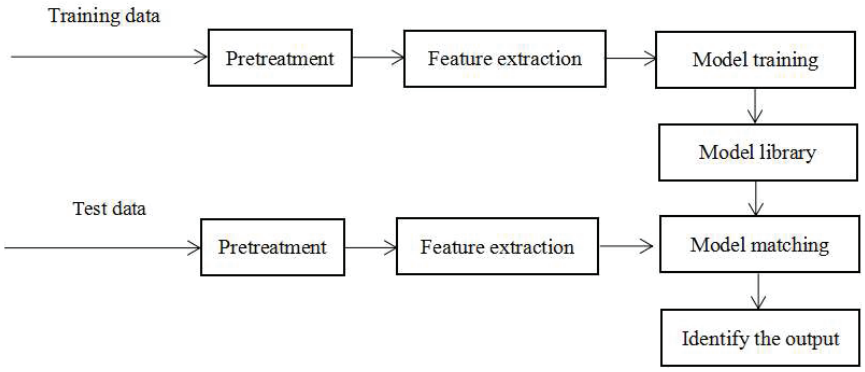


Figure 1 The process of voiceprint recognition.

2.2 End-to-end Voiceprint Recognition Features

Compared with existing voiceprint recognition methods, the end-to-end voiceprint model can capture contextual information, model directly from the discourse level, and combine multiple steps of traditional methods to directly compare the similarities between speakers, and get better results. The end-to-end voice print recognition model is mainly composed of three parts: extraction of acoustic features of speech data, construction of trunk neural network and selection of loss function. Since the acoustic characteristic

parameters are relatively fixed, the effect of the model is usually improved from two other aspects. End-to-end voiceprint recognition model structure is shown in Figure 2:

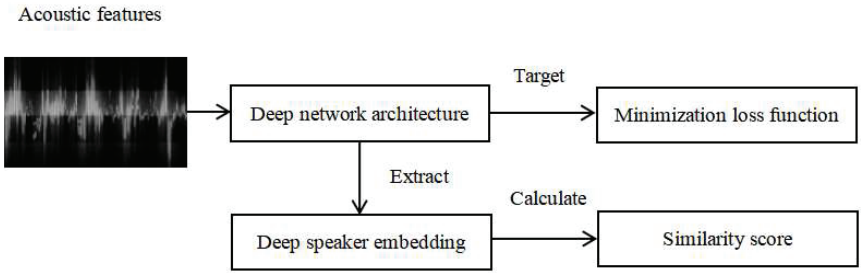


Figure 2 End-to-end voiceprint recognition model structure.

2.3 Res-FD-CNN Network Structure

In this project, multiple residuals blocks of different dimensions are stacked, and the frequency-domain convolution layer is specially added to form the backbone network of the end-to-end model – Res-FD-CNN. Network structure of Res-FD-CNN is shown in Table 1:

Table 1 Network structure of Res-FD-CNN

Network Layer Name	Network Structure	Input Size	Output Size	Step Length	Number of Arguments
Convolution layer-64	$5 \times 5, 64$	$256 \times 64 \times 1$	$128 \times 32 \times 64$	2×2	6K
Convolution layer-128	$5 \times 5, 128$	$128 \times 32 \times 64$	$64 \times 16 \times 128$	2×2	209K
Convolution layer-256	$5 \times 5, 256$	$64 \times 16 \times 128$	$32 \times 8 \times 256$	2×2	823.5K
Residual block-64	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$128 \times 32 \times 64$	$128 \times 32 \times 64$	1×1	246K
Residual block-128	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 3$	$64 \times 16 \times 128$	$64 \times 16 \times 128$	1×1	910K
Residual block-256	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$	$32 \times 8 \times 256$	$32 \times 8 \times 256$	1×1	3.5M
Video convolution layer	1×8	$32 \times 8 \times 256$	$32 \times 1 \times 2048$	1×8	4M
Time pooling layer	8×1	$32 \times 1 \times 2048$	$4 \times 1 \times 2048$	8×1	0
Scale change	–	$4 \times 1 \times 2048$	8192	–	0
Affine layer	8192×512	8192	512	–	4M

Residual combination mode of Res-FD-CNN backbone network refers to the Res-Net-34 structure in the literature. Each basic residual block is composed of a two-layer convolution kernel of 3×3 and step size of 1×1. When the dimensions of the output channel of the residual combination are increased, a specially designed independent convolution layer is adopted in the trunk network of Res-FD-CNN to replace the residual block structure with mapped jumpers in the standard residual network. Res-FD-CNN backbone network only adds one layer of convolution with a convolution kernel of 5×5 and step size of 2×2 before each residual combination. In the trunk network of RES-FD-CNN, the frequency domain dimension of the whole feature after convolution of each layer is constant, so the number of parameters trained in each BN operation is also the same. The activation function is a linear rectifying function with an upper limit of 20, the expression is:

$$\sigma(x) = \min\{\max\{x, 0\}, 20\} \tag{1}$$

The size of the feature graph in the backbone network be $\mathrm{T} \times \mathrm{F} \times \mathrm{C}$, where T and F are time-domain and frequency-domain dimensions respectively, and C is the number of channels. For a mapped jumper residual block whose output dimension is multiplied, the total number of parameters is shown in Equations (1)–(2):

$$\begin{aligned} &num\_params_{projection\ shortcuts} \\ &\quad = (3 \times 3 \times C + 1) \times 2C + (3 \times 3 \times 2C + 1) \times 2C \\ &\qquad + (1 \times 1 \times C + 1) \times 2C + 2FC \times 3 \end{aligned} \tag{2}$$

Which is

$$num\_params_{projections\ shortcuts} = 56C^2 + 6C + 6FC \tag{3}$$

Three 2FCs represent the number of γ and β parameters that need to be trained for batch normalization on each layer sequence, and the residuals of mapped jump-connected blocks have three layers of convolution. The total number of parameters is shown in Equations (1)–(4) after the channel dimensions are doubled directly by using the convolution with a convolution kernel of 5×5 and step size of 2×2.

$$num\_params_{5\times5\ conv} = (5 \times 5C + 1) \times 2C + 2FC \tag{4}$$

Which is

$$num\_params_{5\times5\ conv} = 50C^2 + 2C + 2FC \tag{5}$$

3 Convolutional Neural Network Recognition Model Construction

3.1 Data Preprocessing

Since the end-to-end model constructed by neural network is more suitable for training large-scale data sets with multiple channels, this project adopts the public Vox Celeb2 data set for training, in order to obtain a strong robust model and verify the effect of the model. Vox Celeb is an audio-visual dataset derived from YouTube interview videos. The audio dataset is larger than any publicly available voice-print recognition dataset, and contains nearly 6,000 speakers who have generated millions of voices.

The speakers in the Vox Celeb2 voice data package come from different countries, accents, occupations, and ages. The data are collected without any restrictions [7]. The background noise is in line with the actual real environment. It is more suitable for end-to-end model training to reduce the front-end processing steps of speech signals. The training set of this data contains a total of 5994 different speakers, including more than 1.09 million speeches. Meanwhile, the data set provides the basic scores corresponding to different evaluation indexes of the models constructed by several different methods in speaker recognition and speaker confirmation experiments. The GMM-UBM model takes the 13-dimensional MFCC as the input feature and trains the prior probability of a speaker with 1024 mixed components independent of UBM as the target speaker; In the method of i-vector combined with PLDA, a gender-independent I-vector of 400 dimensions was extracted, which was reduced to 200 dimensions by PLDA [8]; And the use of VGG network as the backbone network to build an end-to-end model, while using comparative loss. For the speaker confirmation system, the dataset further uses Res NET-34 and RES NET-50 as the backbone network to build an end-to-end model, which achieves a better base score.

Convert all data from the Vox Celeb2 training set into a 16K Hz sampling rate, 16bit single-channel wav file, in each round of training, 256 frames are randomly intercepted from the remaining duration of the 1-second mute segment before and after the single selected speech is removed. In this way, even if the same speech is trained in different rounds, the acoustic characteristics of the speaker are not completely the same.

3.2 Frequency Domain Convolution Layer

By comparing the experimental results of speaker identification and speaker confirmation before and after frequency-domain convolution layer was added

to Res-FD-CNN, the cross entropy loss based on Softmax function is also selected as the loss function [9]. The Res-CNN (Deep Speaker Embedding) network only removes the frequency-domain convolutional layer in the Res-FD-CNN backbone network. In order to make the network also invariant in time and position, the high-level frame-level features extracted after multi-layer CNN are passed through the time-average pooling layer with filter size of 8×1. The output size is changed to $4 \times 8 \times 2048$, and the affine layer is used to map the overall feature dimension from 65536 dimension to 512 dimension. Comparison of speaker recognition results before and after adding frequency-domain convolution layer is shown in Table 2:

Table 2 Comparison of speaker recognition results before and after adding frequency-domain convolution layer

Method	Number of Arguments	Top-1 (%)	Top-5 (%)	minDCF	EER (%)
Res-CNN+Softmax	39	79.54	90.51	0.78	9.48
Res-FD-CNN+Softmax	15	82.30	91.62	0.72	8.33

Adding frequency-domain convolution layer to the backbone network can improve the model effect to a large extent. The Res-FD-CNN backbone network adds the frequency domain convolutional layer to multiple independent convolutional layers and residual combinations as the final layer of convolution, which can target the frequency domain based on the correspondence between high-level frame-level features and labels. Information is focused on learning, such as the timbre of the speaker, and the backbone network also has a good balance between the amount of calculation and the effect of the model.

3.3 Algorithm Testing

This study is actually divided into two parts, the self-built database and the TIMIT (TexasInstruments and Massachusetts Institute of Technology) database. The parameters in the experiment are set as follows: the sizes of one-dimensional and two-dimensional convolution kernels are 1×5 and 5×5 respectively, random numbers from 0 to 1 are adopted, and the sizes of down sampling are 1×2 and 2×2 respectively [10]. The convolution operation and down sampling operation are performed only once respectively. MFCC uses 13-dimensional features. The order of GMM-UBM is 16,32,64 and 128 respectively.

Test results of self-built database

The self-built database was made up of a total of 90 people, who were recorded by ordinary microphones with a sampling rate of 8 000 Hz. In the experiment, the number of speakers was set as 60,70,80 and 90, the training speech duration (for each speaker) was 15 s and 30 s, and the test speech duration was 5 s. The recognition rates of the three methods (classical method, 1DConv and 2DConv) were compared respectively. Comparison of different methods to identify the correct number under 15 seconds of training duration(self-built database) is shown in Table 3:

Table 3 Comparison of different methods to identify the correct number under 15 seconds of training duration(self-built database)

		Number of Speakers			
Order	Number of GMM-UBM	60	70	80	90
16	Classic methods	49	50	55	66
	1DConv	38	47	53	50
	1DConv	56	65	76	86
32	Classic methods	43	52	61	67
	1DConv	44	44	52	58
	1DConv	55	66	72	84
64	Classic methods	42	53	56	62
	1DConv	46	49	53	54
	1DConv	56	65	78	90
128	Classic methods	47	56	64	72
	1DConv	46	45	52	56
	1DConv	58	72	81	89

Comparison of different methods to identify the correct number under 30 seconds of training duration(self-built database) is shown in Table 4:

Table 4 Comparison of different methods to identify the correct number under 30 seconds of training duration(self-built database)

		Number of Speakers			
Order	Number of GMM-UBM	60	70	80	90
16	Classic methods	43	50	56	65
	1DConv	35	42	47	57
	1DConv	56	59	67	78
32	Classic methods	48	54	63	62
	1DConv	35	37	42	52
	1DConv	56	64	71	84

(*Continued*)

Table 4 Continued

Order	Number of GMM-UBM	Number of Speakers			
		60	70	80	90
64	Classic methods	55	56	65	65
	1DConv	42	47	48	60
	1DConv	57	66	75	89
128	Classic methods	56	58	67	73
	1DConv	47	50	61	73
	1DConv	57	70	77	91

The following conclusions could be drawn from the experimental results: As the order of GMM-UBM increased, the recognition rates of the three methods all increased; Under different test strips, the overall result of 2DConv was better than that of the classical method and 1DConv, and the recognition rate is improved by about 10%.

Test results of TIMIT database

TIMIT database library was recorded under the condition of high quality microphone by TI(Texas Instruments) and MIT(MassachusettsInstitute of Technology). A total of 630 people (425 men, 205 women) were included, and each speaker recorded 10 sentences, which lasted about 3 s/sentence. The experiment involved 60,70,80, and 90 people randomly selected from 630 people. Comparison of different methods to identify the correct number of training duration(TIMIT database) is shown in Table 5:

Table 5 Comparison of different methods to identify the correct number of training duration(TIMIT database)

Order	Number of GMM-UBM	Number of Speakers			
		60	70	80	90
16	Classic methods	19	25	23	26
	1DConv	32	35	38	44
	1DConv	60	63	72	83
32	Classic methods	24	21	25	26
	1DConv	35	42	45	48
	1DConv	57	69	77	85
64	Classic methods	35	35	42	39
	1DConv	46	55	58	64
	1DConv	56	69	79	86
128	Classic methods	45	44	48	47
	1DConv	55	59	63	73
	1DConv	60	67	82	86

It could be seen from the above that the two-dimensional convolution preprocessing method effectively improved the recognition rate of the system. This should be attributed to the local receptive field and pooling characteristics of convolutional network, which proved that it was also effective for speech signals. In addition, the dimension of data was reduced by convolutional network pretreatment, and the extracted feature dimension was halved after one-dimensional convolution; The feature dimension extracted by two-dimensional convolution was 1/4 of the original one, which was an effective dimension reduction method. In the case of low dimensions, GMM-UBM had a better fit for dimensionality reduction data. When the data was dimensionalized, the model training time was reduced, and the time and space complexity of the algorithm were reduced [11].

4 Conclusion

As a key technology in biometric identification, voice print recognition has the advantages of non-contact and remote verification. However, due to the poor generalization ability of traditional voice print recognition models under the conditions of large background noise, different types and speech time, the main innovation of this study is to conduct in-depth research on the end-to-end voice print recognition model based on convolutional neural network. Use public VOX CeleB2 data collection to develop better performance voice print model. In this study, an end-to-end model backbone network Res-FD-CNN was designed based on the basic unit of frequency domain residual network. Res-FD-CNN adopts stacked direct skip residuals and convolutional layer structure with step size of 2 to increase the network depth, so as to obtain the characteristics of different layers. Res-FD-CNN is the optimal backbone network for the end-to-end voice print model, which can give consideration to both computing load and network performance. In this paper, the end-to-end voice print recognition is preprocessed by using the advantage of convolutional network. Two preprocessing methods are proposed, and the GMM-UBM model is used to model each speaker. The test results of self-built database and TIMIT database show that the proposed method is an effective one.

Acknowledgements

This work was supported in part by the National Science Foundation of China under Grant 51668043, and Grant 61262016, in part by the CERNET

Innovation Project under Grant NGII20160311, and Grant NGII20160112, and in part by the Gansu Science Foundation of China under Grant 18JR3RA156.

References

[1] Q B Nguyen, T T Vu, M L Chi. Improving Acoustic Model for Vietnamese Large Vocabulary Continuous Speech Recognition System Using Deep Bottleneck Features. Advances in Intelligent Systems and Computing, 2015, 326:49–60.

[2] Q Hu, B Y Liu. Speaker recognition algorithm based on convolutional neural network classification. Information Network Security, 2016(04):55–60.

[3] C Zhang, S H Luo, H T Yue, et al. Transformer core voice print pattern recognition method based on MEL time spectrum convolutional neural network. High Voltage Technology, 2020, 327(02):50–60.

[4] Lingfei Yu, Qiang Liu Research and application of voice print recognition method based on deep loop network. Application Research of Computers, 2019, 036(001):153–158.

[5] D Y Du, L J Lu, R Y Fu, et al. Palm vein recognition: An end-to-end convolutional neural network approach. Journal of Southern Medical University, 2019, 039(002):207–214.

[6] C W Sun, C Wen, K Xie, et al. Small sample voice print recognition method based on depth transfer model. Computer Engineering and Design, 2018, 39(12):224–230.

[7] A Nagrani, J S Chung, W Xie, et al. Voxceleb: Large-scale speaker verification in the wild. Computer speech and language, 2020, 60(3):1–15.

[8] J Liu, Y Hu, Huang Heyu. End-to-end deep convolutional neural network speech recognition. Computer Applications and Software, 2020, 037(004):192–196.

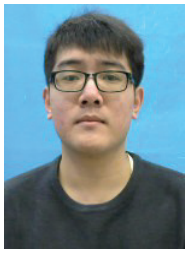
[9] Y Zhao, Y Wang, M G Zhang. Recorded speech detection algorithm based on convolutional neural network. Computer Technology and Development, 2020, 274(02):177–183.

[10] Y C Li, Z F Yan, G P Yan. Edge – Based Double Convolutional Neural Network and Its Visualization. Computer Engineering and Science, 2019, 41(10):1837–1845.

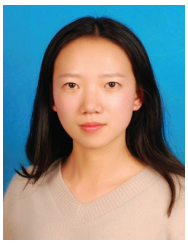
[11] Ji S, Xu W, Yang M, et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(1):221–231.

Biographies

Zhao Hong (1971–), Male, from Xihe, Gansu, Professor, Ph.D., Han nationality, received a bachelor's degree from Northwest Normal University in 1993 and a doctorate degree from Xinjiang University in 2010. Since 1993, he has entered the School of Computer Science, Lanzhou University of Technology, and became a full professor in 2010. He has authored 4 academic works and more than 30 reviewer papers. Current research interests include deep learning, embedded systems and natural language processing.



Yue Lupeng (1995–) Male, from Weihai, Shandong, received a bachelor's degree from Lanzhou University of Technology in 2018. Current research interests include deep learning and speaker recognition.



Wang Weijie (1994–), Female, from Qiqihar, Heilongjiang, received a bachelor's degree from Harbin Finance University in 2016. Current research interests include deep learning and speaker recognition.

Zeng Xiangyan received a bachelor's degree in computer science and information engineering and a master's degree in computer applications from Hefei University of Technology, China in 1987 and 1990, respectively, and a master's degree in electrical and electronic engineering and a doctorate degree in computer science from the University of Ryukyu, Japan in 2001, in 2004. He is currently a professor in the Department of Mathematics and Computer Science at Fort Valley State University in the United States. Wrote more than 40 reference papers. His research interests include computer vision, image processing, pattern recognition and machine learning.