Research on Speaker Recognition of DRNN in Different Noise Environment

chaofeng lan (lanchaofeng@hrbust.edu.cn)

Harbin Industrial University: Harbin University of Science and Technology https://orcid.org/0000-0003-4667-0635

yuanyuan Zhang

Harbin University of Science and Technology

hongyun Zhao

Harbin University of Science and Technology

Research

Keywords: speaker recognition, DRNN, GFCC, MFCC

Posted Date: December 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-124941/v1

RESEARCH ON SPEAKER RECOGNITION OF DRNN IN DIFFERENT NOISE ENVIRONMENT

Lan Chao-feng1*†, Zhang Yuan Yuan 1, Zhao Hong Yun 1

1 School of Electrical and Electronic Engineering,Harbin University of Science and Technology, Harbin 150080, P.R.China

*E-mail: lanchaofeng@hrbust.edu.cn

Abstract: This paper draws on the training method of Recurrent Neural Network (RNN), By increasing the number of hidden layers of RNN and changing the layer activation function from traditional Sigmoid to Leaky ReLU on the input layer , the first group and the last set of data are zero-padded to enhance the effective utilization of data such that the improved reduction model of Denoise Recurrent Neural Network (DRNN) with high calculation speed and good convergence is constructed to solve the problem of low speaker recognition rate in noisy environment. According to this model, the random semantic speech signal with a sampling rate of 16 kHz and a duration of 5 seconds in the speech library is studied. The experimental settings of the signal-to-noise ratios are -10dB, -5dB, 0dB, 5dB, 10dB, 15dB, 20dB, 25dB. In the noisy environment, the improved model is used to denoise the Mel Frequency Cepstral Coefficients (MFCC) and the Gammatone Frequency Cepstral Coefficents (GFCC), impact of the traditional model and the improved model on the speech recognition rate is analyzed. The research shows that the improved model can effectively eliminate the noise of the feature parameters and improve the speech recognition rate. When the signal-to-noise ratio is low, the speaker recognition rate can be more obvious. Furthermore, when the signal-to-noise ratio is 0dB, the speaker recognition rate of people is increased by 40%, which can be 85% improved compared with the traditional speech model. On the other hand, with the increase in the signal-to-noise ratio, the recognition rate is gradually increased. When the signal-to-noise ratio is 15dB, the recognition rate of speakers is 93%.

Keywords: speaker recognition; DRNN; GFCC; MFCC;

1. Introduction

The research on speaker recognition was originated in 1945. L.G.Kesta analyzed the phonetic map and found that all the people's voice signals are clearly different from each other in Bell Labs. These differences can directly reflect the speaker's identity. Combined with the fingerprint recognition technology, L.G.Kesta proposed the concept of voiceprint, and proved the feasibility of using voice signal for the identification, which laid the foundation for the later study of speaker feature extraction.

In 1963, Bogert et al. proposed to apply Cepstrum technology in the scheme design of speaker recognition, and conducted experimental verification[1]. In the same year, Pruzansky raised a method for speaker recognition using template matching, which greatly reduced the computational complexity compared with the traditional recognition methods. In 1965, Tukey and Cooly et al. presented the concept of Fast Fourier Transformation (FFT)[2]. In 1976, Makhoul developed Linear Prediction Cepstral Coefficients (LPCC) as a feature parameter using FFT, and applied LPCC to speaker recognition. In 1980, Davis and Mermelstein et al. studied the Meyer filter and has founded that the Mel Frequency Cepstral Coefficients (MFCC) can reflect the frequency features of the speaker[3]. In 1984, Furui applied MFCC parameter features to speaker recognition technology. In 1987, Burton proposed Vector Quantization (VQ) and applied it to speaker recognition, which achieved good recognition results[4]. In 1988, Carnegie Mellon University developed Sphinx system, which broke the limitation of single vocabulary speech recognition, and made it possible to recognize long-term speech signals of a large number of words. It is a milestone in the research of speech recognition. In 1989, Naik proposed the Hidden Markov Mode (HMM) and applied it to speaker recognition[5]. In 1990, Reynolds proposed the Gauss Mixture Model (GMM) based on the above research results. In 2000, Reynolds proposed the GMM-UBM model, which was enhanced in anti-interference with respect to the GMM model[6].

Speaker recognition has witnessed a rapid development in the 21st century. In 2006, Campbell applied Gaussian Supervector to speaker recognition, demonstrating excellent performance. Corinna Cortes and Vapnik et al. put forward the theory of Support Vector Machine (SVM) and applied it to classifying and recognizing speakers[7]. In the same year, Campbell fused GMM-UBM model with SVM and established Gaussian Mixture Supervector-Support Vector Machine (GSV-SVM), which can greatly improved the effect of speaker classification and recognition[8]. In 2013, Kenny proposed a PLDA model for short speech[9]. In the context of short speech, i-vector was introduced into the PLDA model to improve the recognition rate of short speech. In 2016, Google Variani et al. proposed d-vector and established a speaker recognition system based on deep neural network[10]. In 2017, Alan Papir et al. proposed the loss function model GE2E (Generalized End to End), which can realize end-to-end speaker recognition[11]. This model can extract the speaker's time series information and improve the speech recognition rate.

Compared with foreign countries, the research on speaker recognition in China started relatively later, but due to the broad application prospects of speaker recognition, it has been highly valued by relevant research and development institutions in China, and has achieved rapid development. With the support of research funds such as the National Natural Science Foundation and the Climbing Plan,

China has achieved fruitful results in the research of speaker recognition in recent years[12~16]. At the beginning of this century, machine learning technology has gradually matured, and artificial neural network model has been applied to speaker recognition. In 2011, Fang Xiao used the Error Back Propagation Network (BP) to classify speakers[17]. In 2015, Xue Shaofei improved the training speed of the model by using Deep Neural Network (DNN)[18]. In 2016, Hu Qing and Liu Benyong used Convolutional Neural Network (CNN) to process speech atlas, which has improved the speech recognition rate[19]. In the same year, Pan Hui used RNN to make up for the lack of depth of shallow learning network[20]. In 2018, Li Hao and others introduced softmax classifier into DNN, which solved the problem of gradient dispersion in network training and improved speaker recognition rate[21].

Though the momentum of Speaker recognition technology is very well , there are still bottlenecks to overcome. There are still technical problems to be solved such as how to extract better speech signal characteristic parameters, enhancing the anti-interference of speaker recognition, and improving the speaker recognition rate in a noisy environment. Neural network has a strong feature learning ability. Noisy speech signals can be trained by neural network, which will make them close to the direction of pure speech signals. In the process of training, using function approximation method can reduce the internal parameters and training times, and can quickly improve the accuracy of speaker recognition in noisy environment. In this paper, the Recurrent Neural Networks (RNN) is studied in depth. It is found that the negative feedback of the RNN can be used to identify the characteristic parameters of the noise signal in the speech signal and separate them from the characteristic parameters of the speech signal to achieve the effect of noise reduction. Therefore, on the basis of the existing RNN model, this paper puts the eigenvector zero into the hidden layer, increasing the number of hidden layers, and replacing the activation function, so as to obtain the fast computation speed and noise removal model DRNN (Denoise Recurrent Neural Network); To test and verify the performance of the improved model, the speech recognition effect in the speech database has benn experimented in different noise environments

2. Methods

Speaker recognition is the process of extracting personality features of a certain speaker from a speech signal and matching them with features of a reference model to determine the speaker's identity[22,23]. The pre-treatment method can be used to remove information in the speech signal that is irrelevant for the speaker recognition task, and to enhance the signal-to-noise ratio of the processed speech signal. As a next step, the speech signal is processed using the Fast Fourier Transform (FFT). The transformed signal is filtered using different filter banks and the dimension of the signal is reduced by the discrete Fourier transform. We use two common feature parameters, the Mel Frequency Cepstral Coefficient (MFCC) and the Gammatone Frequency Cepstral Coefficients (GFCC) to determine the speaker on the basis of speech signals: The feature extraction process using MFCC and GFCC is shown in Figure 1.

Once the feature parameters of GFCC or MFCC are obtained, the GMM model is used to train the parameters of multiple speakers, and a reference model library is established [24,25]. The features of the speaker to be identified are compared with each reference feature in the reference model library. The reference feature with the highest matching degree determines the speaker's identity. This is shown in Figure 2.

RNNs have the ability to learn features from time series and have the characteristics of weight sharing. For these reasons, they are well suited for an application in the processing of speech signals. The output of an RNN is correlated with the current input, the network weight and the network before the input. Therefore, this network model is suitable for the processing of voice data with an obvious relationship between front and back text. The forward-propagation within an RNN is a recursive operation on the training data in the time dimension, while the back-propagation is a gradient operation on the training data in the time dimension [26]. Let the training data X_t at a certain time t be transmitted forwards through the Recurrent Neural Network, then the output is Y_t. The structure diagram of the RNN is shown in Figure 3.

The left side of Figure 3 shows the structure of the RNN that has not been expanded in time, the right side shows the structure of the cyclic neural network that has been expanded in time.

Let the transfer function relationship between layers of the RNN model at time t be:.

$$h_t = \sigma(uX_t + wh_{t-1} + b) \tag{1}$$

Where u represents the weight matrix from the input layer to the hidden layer; w represents the weight matrix of the hidden layer; σ(.) represent the nonlinear activation functions within the RNN; b is an offset parameter and h_t denotes the forward transfer function.

The output Y_t can be expressed as:

$$Y_t = vh_t + c \tag{2}$$

Where v represents the weight matrix from the hidden layer to the output layer and c represents the linear offset of the output.

Suppose that the prediction output at time t is Y^t, then the loss function L_t is:

$$L_t = Y^t - Y_t \tag{3}$$

In the RNN, a functional loss exists at each position of the sequence; the final loss can be written as follows:

$$L = \frac{1}{N}\sum_{t=1}^{N} L_t^2 \tag{4}$$

Where N denotes the number of input data.

The process of RNN backpropagation is to evaluate the partial derivatives of L with respect to the five parameters u, v, w, b and c. The partial derivatives of L with respect to v and c can be expressed as follows:

$$\frac{\partial L}{\partial c}=\sum_{t=1}^{\tau}\frac{\partial L_t}{\partial c}=\sum_{t=1}^{\tau}\frac{\partial L_t}{\partial o_t}\frac{\partial o_t}{\partial c}=\sum_{t=1}^{\tau}Y^t-Y_t \quad (5)$$

$$\frac{\partial L}{\partial v}=\sum_{t=1}^{\tau}\frac{\partial L_t}{\partial v}=\sum_{t=1}^{\tau}\frac{\partial L_t}{\partial o_t}\frac{\partial o_t}{\partial v}=\sum_{t=1}^{\tau}(Y^t-Y_t)(h_t)^T \quad (6)$$

Where,*T* means transpose.

In the process of the back propagation, the gradient loss of a sequence position *t-1* is determined by the gradient loss corresponding to the output of the current position and the gradient loss at the sequence index position *t.* The partial derivative of *L* with respect to *u*, *v*, and *b* can only be obtained by calculating the gradient of the hidden position. The gradient is expressed as:

using *δ(t)*, to evlaluate the partial derivatives of *L* with respect to *u*,*v*,*b*:

$$\delta(t)=\frac{\partial L}{\partial h_t} \quad (7)$$

$$\frac{\partial L}{\partial w}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial w}=\sum_{t=1}^{\tau}diag(1-h_t^{\,2})\delta(t)(h_{t-1})^T \quad (8)$$

$$\frac{\partial L}{\partial b}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial b}=\sum_{t=1}^{\tau}diag(1-h_t)^2\delta(t) \quad (9)$$

$$\frac{\partial L}{\partial u}=\sum_{t=1}^{\tau}\frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial u}=\sum_{t=1}^{\tau}diag(1-h_t^{\,2})\delta(t)(x_t)^T \quad (10)$$

All parameters are then updated:

$$u^*=u-\varepsilon\frac{\partial L}{\partial u} \quad (11)$$

$$v^*=v-\varepsilon\frac{\partial L}{\partial v} \quad (12)$$

$$w^*=w-\varepsilon\frac{\partial L}{\partial w} \quad (13)$$

$$b^*=b-\varepsilon\frac{\partial L}{\partial b} \quad (14)$$

$$c^*=c-\varepsilon\frac{\partial L}{\partial c} \quad (15)$$

Where ε is the learning rate of neural networks. The larger the value of ε is, the higher the convergence speed of the neural network is while the oscillation intensity increases. On the contrary, the convergence speed becomes lower and stabler.

When the loss value *L* reaches a certain threshold or changes little, the neural network can be regarded as being convergent, where the updating of the parameters of the whole network can be terminated so as to obtain the optimal parameters of the network.

3. Results

The noise contained in the speech signal is a key issue affecting the recognition rate of the speaker. How to improve the speaker recognition rate in the noisy speech signal is a topic that researchers are keen to do research on. Based on the RNN model, this paper proposes an improved speech denoising model DRNN, which can help achieve high quality speech recognition in noisy environments. The structure of the DRNN model is shown in Figure 4.As shown in Figure 4, compared with the RNN model, the number of hidden layers has been added to the improved DRNN model to improve the approximation speed of the noisy signal to the pure signal during each iteration process, and a connection layer is added to the implicit layer in the middle to reduce the amount of training calculation. Zero-padding is made to the input signal x_i, and the zero-padded x_i, three in one group, is placed in the input layer to improve the information utilization of the input signal and maintain the dimensional consistency.

The training process of DRNN model is shown in Figure 5.

Letting the *i* th packet of the input noisy speech signal be X_i, and the output after passing through the first hidden layer is:

$$P_i^1=\sigma(X_i w_1+b_1) \quad (16)$$

Where w_1 is the weight parameter of the first hidden layer, and b_1 is the paranoid parameter.

The output of the second hidden layer is expressed as:

$$P_i^2=\sigma(P_{i-1}^2 h+P_i^1 W_2+b_2) \quad (17)$$

Where P_{i-1}^2 is the output of the second hidden layer of the *i-1* th group; h is the transfer function between hidden layers, W_2 is the weight parameter, and b_2 is the paranoid parameter.

The output of the third hidden layer is expressed as:

$$P_i^3=\sigma(P_i^2 W_3+b_3) \quad (18)$$

Where W_3 is the weight parameter of the third hidden layer, b_3 is a paranoid parameter.

The output of the fourth hidden layer is expressed as:

$$Y_i = W_4 P_i^3 + c \tag{19}$$

Where W_4 is the weight parameter of the fourth hidden layer, c is the paranoid parameter of the output.

Then the final output of the noisy speech signal after forward transmission can be expressed as:

$$Y = [Y_1, Y_2 ... Y_N] \tag{20}$$

Assuming that Y' represents pure speech signal, the loss function L can be measured by the mean value of vector difference between output signal and pure signal:

$$L = \frac{1}{N}(Y - Y') \tag{21}$$

If L is greater than the preset value, find the partial derivatives of W_1, W_2, W_3, W_4, b_1, b_2, b_3, U and C parameters with respect to L by referring to the RNN principle, calculate new parameters and perform forward transfer until L is less than the preset value and the training ends. Since the trained DRNN model parameters contain the characteristic information of the training noise, the noise-reduced speech signal can be obtained after the speech signal containing the noise is input to the DRNN.

The process of DRNN model speaker recognition is shown in Figure 6.

Using DRNN speech denoising model to train and recognize feature parameters can help achieve certain denoising effects, improving operation speed and saving calculation cost.

Compared with the traditional RNN, the DRNN model has three hidden layer structures. The increase in the number of layers will lead to the increase in the number of neurons and the increase in the network computing complexity. The Sigmoid activation function commonly used in RNN has less derivative results for the neural network with more layers, which may eventually lead to the disappearance of gradient. Therefore, in this part, the Sigmoid activation function is changed to the Relu activation function. Featured by lower computational complexity and faster convergence, it can avoid gradient disappearance, and is suitable for deep neural networks.

The Relu activation function can be expressed as:

$$\sigma_R(z) = \max\{0, z\} \tag{22}$$

Where z represents the input of function.

The characteristics of non-positive zeroing of Relu activation function easily causes the neuron node value to be 0 in the search process, produces dead neurons, affects the training speed and training effect, and is not conducive to the approximation of neural network to complex nonlinear functions, resulting in reduced recognition rate. Therefore, in this paper, Leaky Relu activation function is used to replace Relu activation function in the middle hidden layer to avoid the phenomenon of non-positive zeros, which can be expressed as follows:

$$\sigma_L(z) = \begin{cases} z, z \geq 0 \\ \lambda z, z < 0 \end{cases} \tag{23}$$

Where λ is constant whose value is 0.01. The introduction of Leaky Relu reduces the computational complexity and improves the convergence of DRNN in both forward propagation and backward update.

The experimental software environment selects one PC of Win10 operating system, and the PC processor is Intel (R) Celeron (R) CPU N9300 @ 2.50GHz, with memory of 4G, and the hard disk storage space of 1T. Programming is performed with Matlab R2014b.

The experiment used some voice data from the TIMIT database, and selected 10 voice signals recorded by 200 people each for research. Among them, each person's seven speech segments were used as training set and the rest as test set. The speech duration was 5 seconds and the sampling rate was 16 kHz.The voice signal in the database is not mixed with noise, and the file format was .wav format.

In order to observe the speech recognition rate in different noise environments, the above speech signals are mixed into NOISEX-92 standard noise database and the noise is as follows: White Noise, Babble Noise and Factory Noise.The white noise is generated by an analog noise generator with equal energy per Hertz bandwidth and smoother waveform.Bubbling noise is caused by 100 people speaking in a room with a radius of more than two meters, each of whom can hear all other persons' voices slightly, at a volume of 88 decibels. Factory noise is recorded near the cutting board and welding equipment. The waveforms of the three kinds of noise are shown in Figure 7.

As shown in Figure 7 that the amplitudes of three kinds of noise signals are quite different. When adding noise to the voice database during the experiment, the noise signals are scaled according to the signal-to-noise ratio set by the experiment.

Our papers are double column format; one column width is 8.6cm. All paragraphs must be justified, i.e. both left-justified and right-justified.

In this paper, the DRNN model using Leaky Relu activation function is called the improved DRNN model, and the DRNN model using Relu activation function is called the traditional.

In the experiment, the noise are selected from NOISEX-92 standard noise library. The signal-to-noise ratio is set to -10dB, -5dB, 0dB, 5dB, 10dB, 15dB, 20dB and 25dB, the learning rate is 0.01, the number of neurons is 512, the initial parameters are set to random, the threshold is set to 0.1, the dimension of characteristic parameters is set to 32, and the GMM mixture number is set to 40. In this section, MFCC and GFCC characteristic parameters were subjected to the model. In order to verify the effects of Relu activation function and Leaky Relu activation function on speaker recognition rate in noisy environments, experiments are conducted under different noise environments and different SNR (Signal Noise Ratio).experimental testing of speaker recognition effect, the DRNN model using Relu activation function was defined as R-DRNN-MFCC and R-DRNN-

GFCC, and DRNN model using Leaky Relu activation function was defined as DRNN-MFCC and DRNN-GFCC. The simulation results after 100 times of Monte Carlo simulation training are shown in Figure 8.

In order to study the speech recognition effect of MFCC and GFCC feature parameters in different noise environments, the improved DRNN model and the traditional RNN model proposed in this paper are used for experimental study. The improved DRNN model is defined as DRNN-MFCC and DRNN-GFCC. The traditional RNN model is defined as MFCC and GFCC. The experimental conditions are the same as Section 4.1. The experimental results after 100 Monte Carlo simulation training are shown in Figure 9.

For the intuitive comparison of the size of speech recognition rate between the traditional RNN model and DRNN model, the ratio of speech recognition rate of improved RNN model to that of traditional RNN model is set to R, the variation curve of R with SNR in three noise environments is drawn, as shown in Figure 10. Among them, the marking method of curve drawing is marked as FactoryGFCC based on GFCC in factory noise environment, WhiteGFCC based on GFCC in white noise environment and BabbleGFCC based on GFCC in bubbling noise environment which can be similar with the labeling method for MFCC parameters.

It can be seen from Figure 9 and Figure 10 that, with the increase in SNR, the speech recognition rate in the three noise environments gradually increases. The recognition rate of traditional RNN model is lower than that of improved DRNN model, and the recognition effect of GFCC feature parameters treated by DRNN model is better than that of MFCC. Compared with RNN model, the recognition rate of DRNN model is improved in the range of 1.8-4 when the signal-to-noise ratio is low. Therefore, the improved DRNN model proposed in this paper is beneficial to the improvement of the speaker's speech recognition rate.

In this paper, RNN is applied to the speaker recognition model to make the speaker recognition system acquire the ability of active noise reduction after learning through the neural network. A speech denoising model DRNN based on RNN is proposed to reduce the noise in speech signals.

Leaky Relu is used as the activation function of DRNN model to improve the speaker recognition rate.

Introducing DRNN model into speaker recognition system can significantly improve the performance of speaker recognition system.When the signal-to-noise ratio is 0 dB, compared with the traditional speaker recognition, it improves by 40%, and the speaker recognition rate reaches 85%.With the increase in the signal-to-noise ratio, the recognition rate increases gradually. When the signal-to-noise ratio is 15 dB, the speaker recognition rate reaches 93%.

Equations should be capable of fitting into a two-column print format. If they do not fit into one column they should be floated to the bottom of the page or top of the next and cited in the text by “(see (3))”.

4. List of abbreviations

Not applicable.

Declarations

Availability of data and materials

With the corresponding text in the Availability of data and materials statement:

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by the national natural science youth foundation of china (No.11804068).

Authors' contributions

All authors read and approved the final manuscript.

Acknowledgments

This work was supported by the national natural science youth foundation of china (No.11804068).

References

‘Mingyu Wang, Fang Wang, Shaojun Wei, “A Pipelined Area-efficient and High-speed Reconfigarable Processor for Floating-point FFT/IFFT and DCT/IDCT Computations”[J], Microelectronics Journal, 2016, 47: 19-30.

Robert B, Randall.A,“History of Cepstrum Analysis and its Application to Mechanical Problems”[J], Mechanical Systems and Signal Processing, 2017, 97: 3-19.

Wilson, Richard H, Hamm, Heather M,“Recognition Performance of Interrupted Monosyllabic Words: The Effects of Ten Interruption Locations”[J], Journal of the American Academy of Audiology, 2015, 26(7): 670-677.

Singh, Nilu, R. A. Khan, and Raj Shree, “Applications of Speaker Recognition”, Procedia Engineering 38(2012): 3122-3126.

Junzo Watada, Hanayuki, “Speech Recognition in a Multi-speaker Environment by Using Hidden Markov Model and Mel-frequency Approach”, 2016 IEEE Third International Conference on Computing Measurement Control and Sensor Network (CMCSN), Japan, 2016:80-83.

Mohamad Hasan Bahari, Najim Dehak, Hugo Van hamme, Lukas Burget, Ahmed M. Ali, Jim Glass, “Non-Negative Factor Analysis of Gaussian Mixture Model Weight Adaptation for Language and Dialect Recognition”, IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014 Jul, VOL.22, NO.7, pp. 1117-1129.

M.Jian and L.Yongmei, “An embedded voiceprint recognition system based on GMM”, in Computer Science & Education(ICCSE), 2015 10th International Conference on , 2015, pp. 38-41.

Vincent Christlein, David Bernecker, et al., “Writer Identification Using GMM Supervectors and Exemplar-SVMs”[J], Pattern Recognition, 2017, 63: 258-267.

Qingfang Zhang, Heming Zhao, Xiaojiang Gu, “A Joint Factor Analysis Approach to Whispering Speaker Identification under Mismatched Speaking Manners and Channels”[C], 2012 IEEE International Conference on Signal Processing, HongKong, 2012: 608-612.

Cemal Hanilci, “Data Selection for I-vector based Automatic Speaker Verification Anti Spoofing”[J], Digital Signal Processing, 2018, 72: 171-180.

Alan Papir, Ignacio Lopez Moreno, “Generalized End-to-End Loss for Speaker Verification”, Acoustics, Speech and Signal Processing[C], 2018 IEEE International Conference on IEEE, Qingdao, 2018: 4879-4883.

Krishnamoorthy, P.; Jayanna, H.; Prasanna, S, “Speaker recognition under limited data condition by noise addtion”, Expert Syst. Appl. 2011, 38, 13487-13490.

Shung-Yung Lung, “Improved wavelet feature extraction using kernel analysis for text independent speaker recognition”, Digital Signal Processing, vol.20, no.5, pp.1400-1407, September 2010.

P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint Factor Analysis versus Eigenchannels in Speaker Recognition”, IEEE Transaction on Audio Speech and Language Processing, vol. 15, no.4, pp.14351447, May 2007.

F.Richardson, D. Reynolds, and N. Dehak, “Deep neural network approaches to speaker and language recognition”, IEEE Signal Processing Letters, vol.22, no.10, pp.1671-1675, 2015.

Kenny, P.; Jayanna, H.; Prasanna, S., “Speaker and session variability in GMM-based speaker verification”, IEEE Trans. Audio Speech Lang. Process. 2007, 15, 1448-1460.

Wang, Y.; Lawlor, B, “Speaker recognition based on MFCC and BP neural networks”, In Proceedings of the lrish Signals and Systems Conference(ISSC) , killarney, lreland, 20-21 June 2017 ; pp.1-4.

Zhang, C.; Woodland, P.C. “DNN speaker adaptation using parameterized sigmoid and Relu hidden activation functions”, In Proceedings of the IEEE international Conference on Acoustics Speech and Signal Processing(ICASSP), Shanghai, China, 20-25 March 2016; pp. 5300-5304.

M. yungjong Kim, Beiming Cao, Ted Mau, Jun Wang, “Speaker-Independent Silent Speech Recognition from Flesh-Point Articulatory Movements Using an LSTM Neural Network”, IEEE/ACM Trans Audio Speech Lang Process. 2017 Dec; 25(12): 2323-2336.

M. McLaren, Y. Lei, and L.Ferrer, “Advances in deep neural network approaches to speaker recognition”, in Proc. of ICASSP, 2015, pp.4814-4818.

Sumithra Manimegalai Govindan, Prakash Duraisamy, Xiaohui Yuan, “Adaptive wavelet shrinkage for noise robust speaker recognition”, Digital Signal Processing. vol.33, pp.180-190, 2014.

Waddi Santosh K, Pandey Prem C, Tiwari Nitya, “Speech enhancement using spectral subtraction and cascaded-median based noise estimation for hearing impaired listeners”[C], 2013 National Conference on Communications. Washington, United States: IEEE Computer Society, 2013: 1-5.

Kenny P, Utterances Stafylakis T, “PLDA for Speaker Verification with of Arbitrary Duration”[C], Acoustics, Speech and Signal Processing, International Conference on IEEE, Vancouver, 2013: 7649-7653.

Dionelis Nikolaos, Brookes Mike, “Modulation-Domain Kalman Filtering for Monaural Blind Speech Denoising and Dereverberation”, IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019, Vol.27, no 4, pp.799-814, ISSN:2329-9290.

Yi Zhang, Lei Ni, “Feature Extraction algorithm fusing GFCC and phase information”, 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference(IAEAC 2017), 2017:1163-1167.

Ying Shi, “Optimization and Design of Speech Recognition Scheme Based on Recurrent Neural Network”[D],Beijing. Beijing Jiaotong University, 2017: 27-30.

Figures

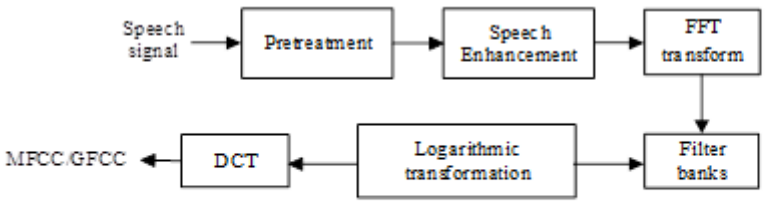

Figure 1

The feature extraction process using MFCC and GFCC

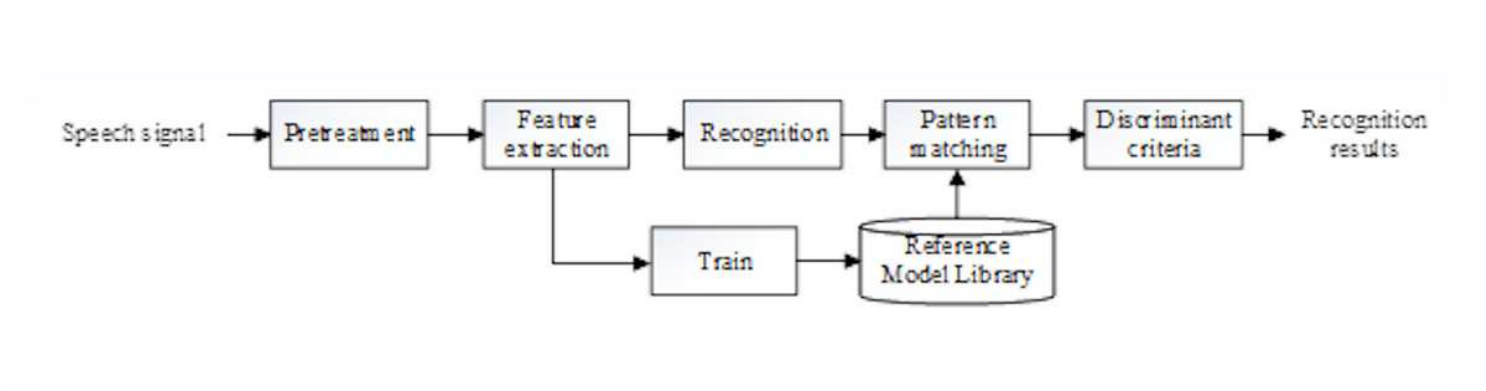

Figure 2

The reference feature with the highest matching degree determines the speaker's identity

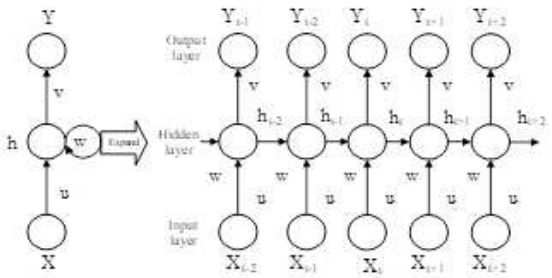

Figure 3

The structure diagram of the RNN

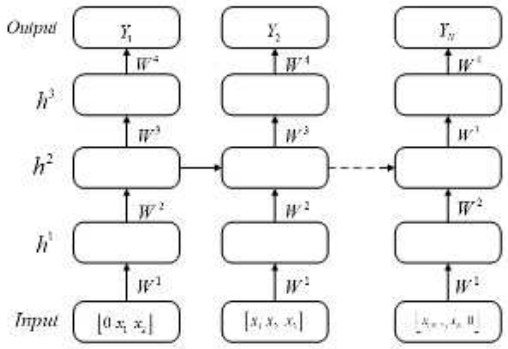

Figure 4

The structure of the DRNN model

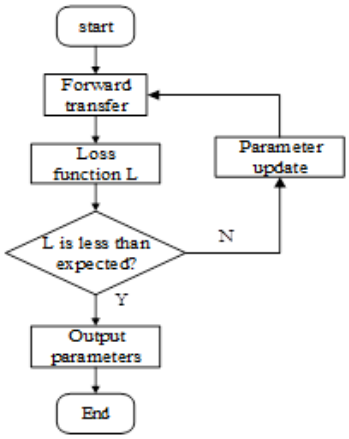

Figure 5

The training process of DRNN model

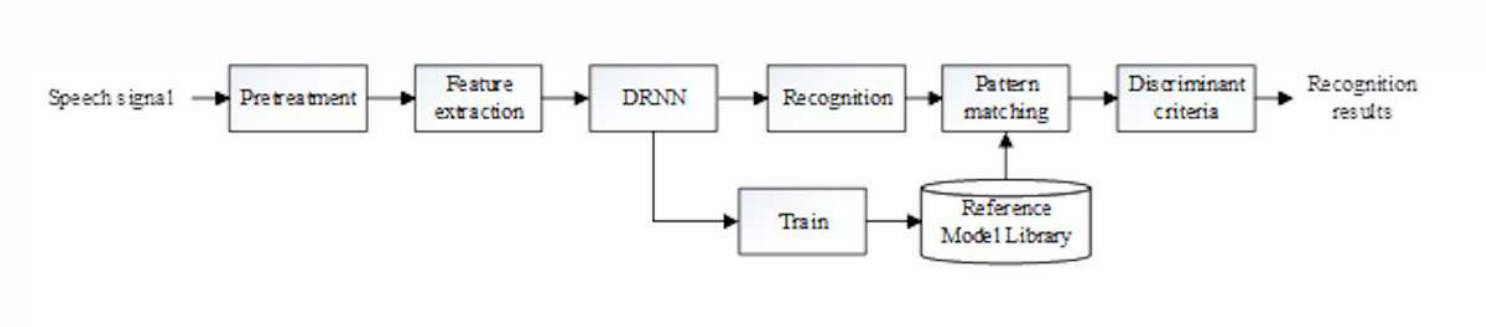

Figure 6

The process of DRNN model speaker recognition

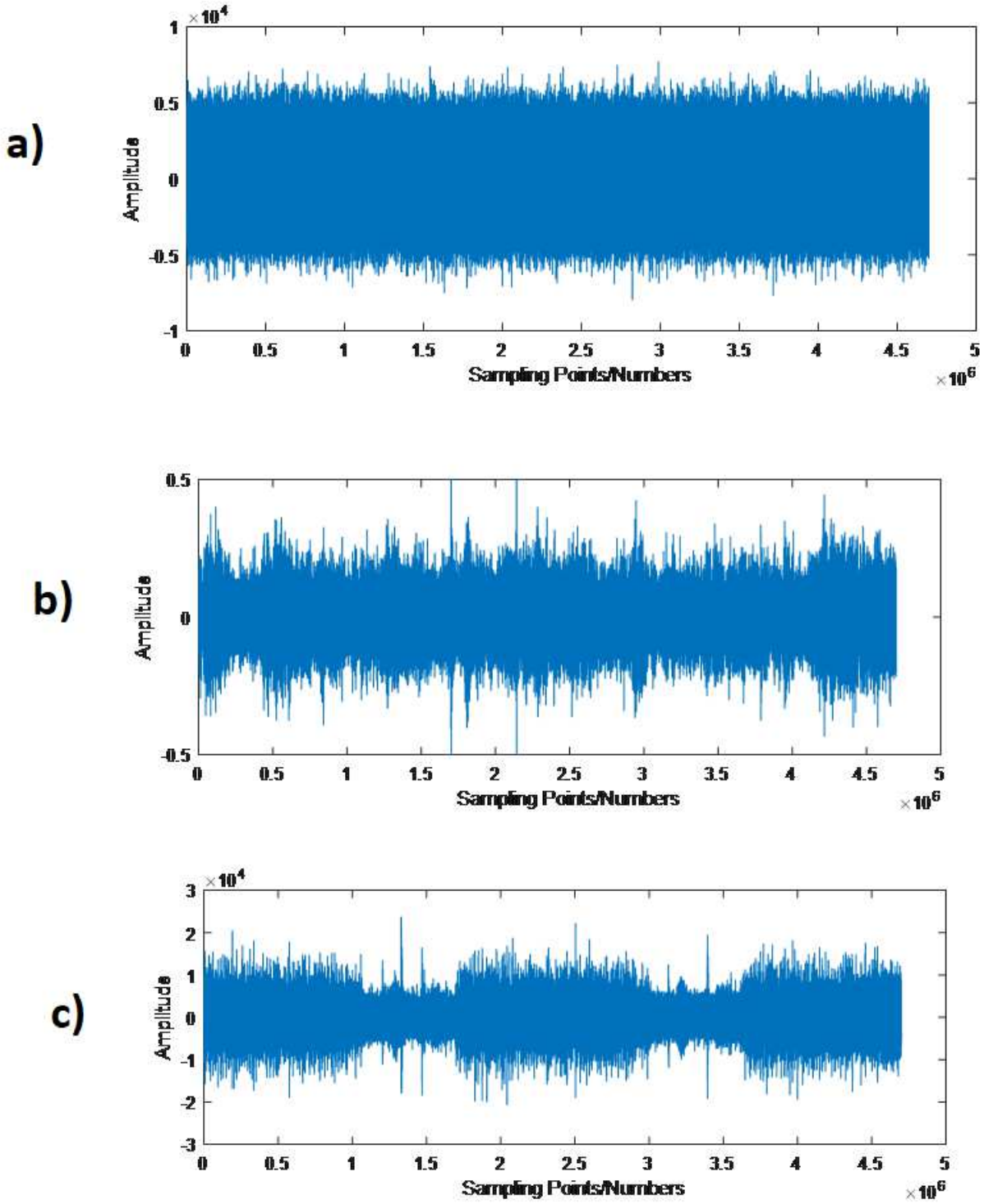

Figure 7

The waveforms of the three kinds of noise

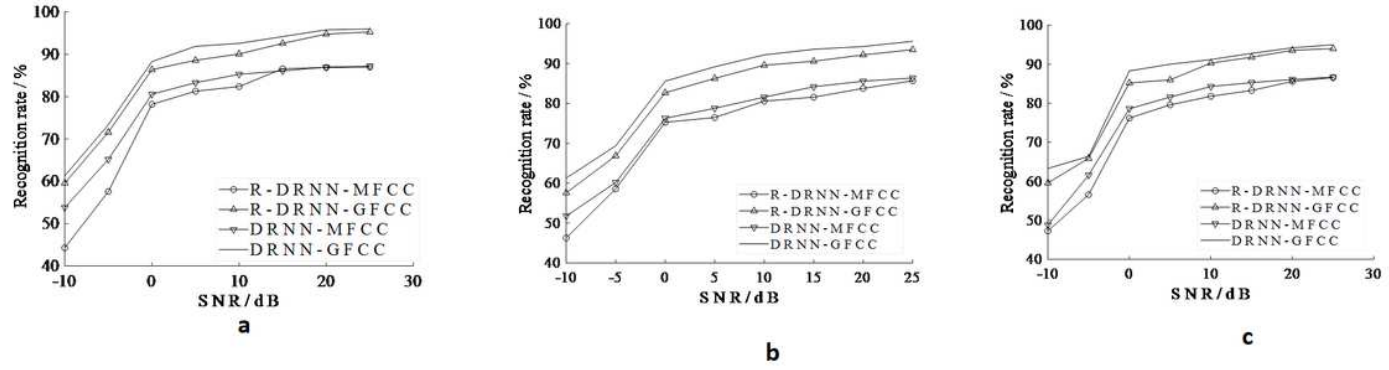

Figure 8

The simulation results after 100 times of Monte Carlo simulation training

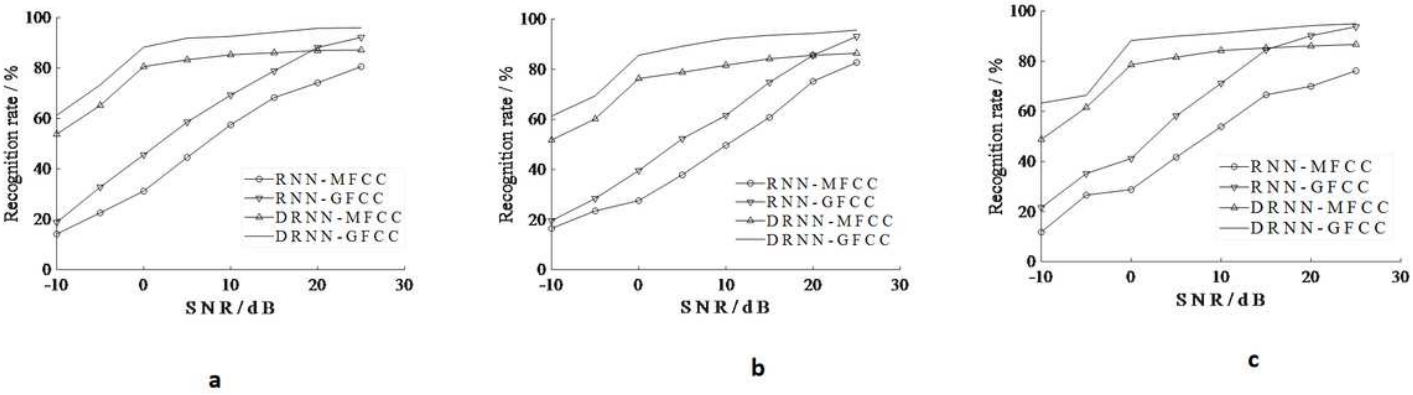

Figure 9

The experimental results after 100 Monte Carlo simulation training

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Highlights.docx