ISSN 2616-5775 Vol. 5, Issue 7: 85-89, DOI: 10.25236/AJCIS.2022.050714

Principle Analysis of Voiceprint Identification and Authentication System Based on LSTM

Xirui Wang, Jiachun Wang, Yunqi Cao, Ziyi Ye, Wenlian Zhang*

School of Electrical and Electronic Engineering, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, China
**Corresponding author*

Abstract: *Recurrent neural networks, as an effective method to study the analysis and prediction of large-scale time series, have existed in various applications on time series data, but the traditional RNN neural network has the problem of gradient disappearance and the problem of poor prediction accuracy has not been solved. This paper proposes a vocal recognition system based on LSTM neural network, whose core structure can be divided into four parts: forgetting gate, input gate, cell The core structure can be divided into four parts: forgetting gate, input gate, state and output gate, which are more suitable to analyze the characteristics of different human voice patterns, and can obtain higher recognition accuracy after the process of feature extraction, data enhancement, model training and voice pattern recognition. Finally, the performance of the LSTM neural network is tested on 300 speech data, and the test results prove that the LSTM neural network can obtain high recognition rate with less iterations.*

Keywords: *LSTM neural network model; RNN neural network; MFCC; Voice recognition*

1. Introduction

With the rapid development of digital signal processing technology, the frequency of use of voice signals has increased dramatically, especially in the fields of attendance recording, e-commerce and automatic control with increasing participation. The vocal recognition system uses the data extracted from the pick-up system to establish a verdict model, which can quickly and efficiently achieve sign-in person identification with high foresight. Kangji Du et al. proposed an improved recurrent neural network model by adding the weight matrix of the connected input values with the random perturbation matrix to obtain a new weight matrix, so that the convergence effect of the improved recurrent neural network algorithm could be took advantage of Convolutional Neural Networks (CNN) in spatial feature extraction and LongShort-Term Memory (LSTM) in spatial feature extraction and Long-Short-Term Memory (LSTM) in temporal feature extraction, a hybrid neural network model of CNN-LSTM fusing CNN and LSTM was constructed, Liu Xiaoxuan et al. proposed a Long Short-Term Memory (LSTM) neural network based on A text-independent vocal pattern recognition method was proposed by Liu et al. The traditional RNN neural network has the problem of gradient disappearance and the problem of poor prediction accuracy has not been solved. In response to the above problems, this paper summarizes the research progress of scholars at home and abroad based on the neural network model, on this basis, we propose a voice recognition model based on Long-Short-Term Memory (LSTM) neural network, and complete the identity authentication system design. The system takes different students' voices as the research object, and through training data, builds and optimizes the model, and finally obtains a practically useful voice judgement system to effectively make recognition of the identity of the sign-in person.

2. Establishment of Model

2.1. Overview of Neural Networks

Neural networks are an important machine learning technology, and many scholars at home and abroad have made considerable research in the direction of neural networks. Its learning process is similar to that of a human being.

ISSN 2616-5775 Vol. 5, Issue 7: 85-89, DOI: 10.25236/AJCIS.2022.050714

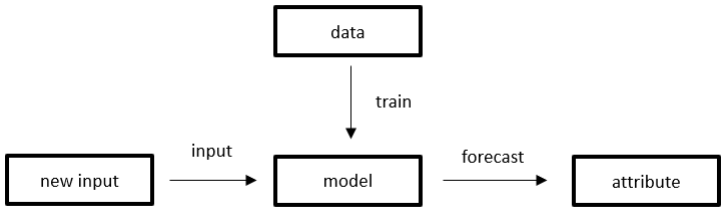

Figure 1: The learning process of a neural network

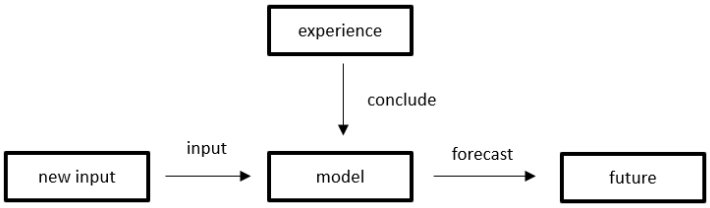

Figure 2: The human learning process

2.2. LSTM Neural Network

Recurrent Neural Network (RNN) is able to obtain good predictions when dealing with shorter sequences. However, because it can only remember the behavior of the previous moment and react based on that behavior and the features of that moment, it is difficult to influence the present based on the distant past when dealing with longer sequences - the present result is only changed by a limited number of "yesterdays". LSTM is a special RNN architecture in the field of deep learning, suitable for predicting important events with relatively long intervals and large delays in a time series. The method circumvents the gradient explosion and gradient disappearance problems caused by chained derivations. LSTM controls the flow of information by adding forgetting gates, input gates and output gates, where x_t is the input layer and h_t shows the current hidden state of the output layer, and the LSTM indicates the long-term and short-term memory mechanisms. Where σ and$tan\ h$ denote the Sigmoid neural network layer and the hyperbolic tangent activation function, respectively. The flow chart shows the forward propagation process for processing input variables and generating hidden states in a sequence, which can be divided into six steps as follows.

(1) The forgetting gate controls the flow of information from the previous moment cell c_t−1 how it accumulates to the current storage cell c_t. By reading h_t and x_t of the information, the forgetting gate, via the σ layer will output a value between 0 and 1, where 0 and 1 indicate that the information is to be discarded completely or retained completely, respectively. Using W and b denote the corresponding weights and deviations (below), respectively, and this step equation is shown below.

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{1}$$

(2) Input gateit Controlling the input to the memory cell and accumulating new information, the σ layer determines which variable will be updated.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{2}$$

(3) Update input variables g_t. By means of the$tan\ h$ activation function, transferring the previous hidden layer h_t−1 and the current x_t state, thus creating a new candidate vector

$$g_t = tan\,h\,(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{3}$$

(4) Update output gate o_t. σ Layer determines how much information can flow from the

storage unit c_t flow into the hidden layer h_t.

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{4}$$

(5) Update the memory unit c_t. The result is that the previous moment storage unit c_t−1 and the forgetting gate f_t and the product of the corresponding term of the input gateit and the current input value g_t The former removes unnecessary information from the previous moment, while the latter adds useful information, where $\odot$ denotes the Hadamard product (below).

ISSN 2616-5775 Vol. 5, Issue 7: 85-89, DOI: 10.25236/AJCIS.2022.050714

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

(6) Replace the transient hidden layer h_t that is, the storage unit c_t and the output gate o_t are multiplied by each other.

$$h_t = o_t \odot c_t \tag{6}$$

$h_t = o_t \odot c_t$ and the last minute h_t is the output of the LSTM forward propagation.

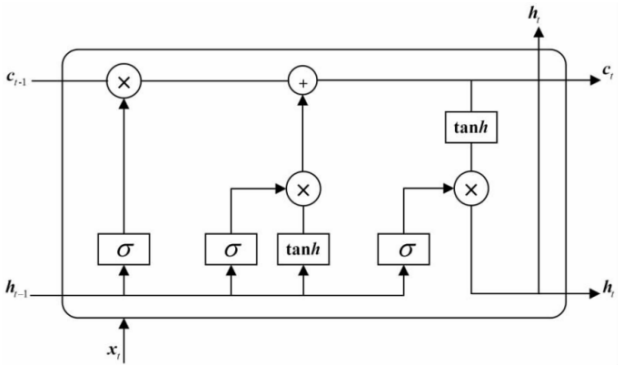

Figure 3: Structure of LSTM neurons

2.3. LSTM based voice recognition authentication system

The speech dataset used in this paper is a Chinese dataset published from the Centre for Speech and Language Technology, Tsinghua University. The speech sampling frequency was 16 000 Hz and the sampling size was 16 bits. first, the THCHS-30 dataset was ranked and filtered, in which 20.

People were selected from the dataset with correct label classification, with a total of 300 speech items, and the training and test sets were divided according to the ratio of 8:2. The human auditory system is a special non-linear system with different sensitivities in response to signals of different frequencies. MFCC is based on experiments in human auditory perception. The equation for converting the ordinary spectrum to the Mel spectrum is as follows.

$$mel(f) = 2595 \times lg(1 + \frac{1}{700}) \tag{7}$$

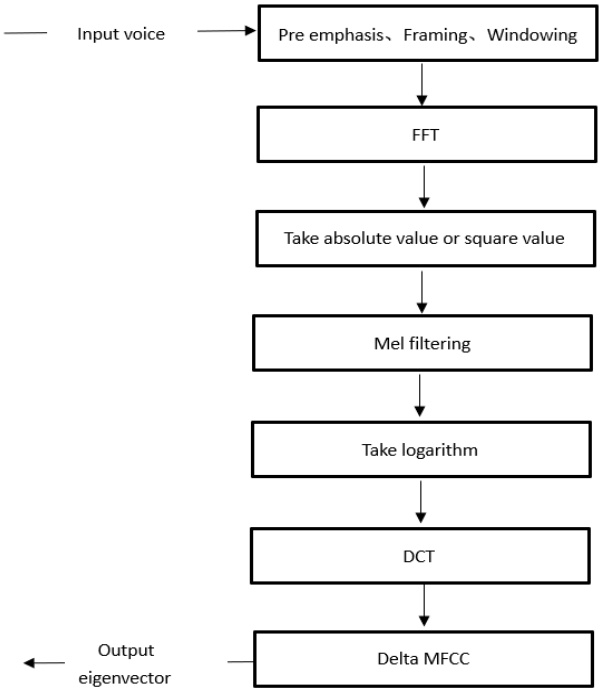

Figure 4: Process of feature vector extraction

ISSN 2616-5775 Vol. 5, Issue 7: 85-89, DOI: 10.25236/AJCIS.2022.050714

A randomly selected 80% of each person's speech fragment was used as the training set, with 12 training pieces, and 3 of the remaining 20% were used as the test set, i.e. to verify and evaluate the performance of the model in terms of accuracy, and to compare the change of the loss function with different iterations under the same training set. In this paper, the LSTM model was constructed by extracting spatial and temporal features, and the number of iterations set in the experiment was 100.

Life is full of distracting factors such as noise, echoes, background sounds, etc. If the system is only learning from "perfect" data it is often unable to adapt to complex real-world situations, so after obtaining the characteristics of the data it still needs to be augmented. During training we randomly overlay the input signal with Gaussian white noise, add echoes, randomly flip and splice, and randomly crop. After processing, the final result is a short-time Fourier transformed amplitude spectrum.

Y is the correct classification of the category, P (Y|X) is the probability of correct classification, this paper uses the loss function (Loss) as the evaluation criterion, then the formula is calculated as follows.

$$Loss = L\big(Y, P(Y|X)\big) \tag{8}$$

3. Simulation and Evaluation of the model

The experimental platform uses Google's open source deep learning framework Tensorflow, and the samples are trained on the Tensorflow platform. In this paper, 300 speech data were selected, and the duration of the selected speech fragments was generally in the range of 3 ~ 4 s. In order to unify the size of the speech spectrogram, the speech fragments with less than 4 s were processed with a complementary 0, so that their duration was unified to 4 s. 80% of the speech fragments of each person were randomly selected as the training set, and the remaining 20% were used as the test set, and echoes and Gaussian white noise were added to the source data for data enhancement. Then the speech data energy was plotted by Fourier variation of the speech temporal information, frequency information and 106 ×80 The speech spectrograms are used as the input data for the network model. In this paper, the LSTM model is constructed by extracting temporal features, and the model is combined with a soft max classifier to form a network model, and the number of iterations set in the experiment is 100. The experimental results are shown below, indicating the variation of the loss function of the LSTM in the test set with the number of iterations. Robustness is the property of a control system to maintain certain performance under certain parameter uptake. It can be seen that the value of the loss function of the network changes slowly and smoothly after dropping to about 0.1, with a relatively stable robust performance and a final accuracy of 95.67%.

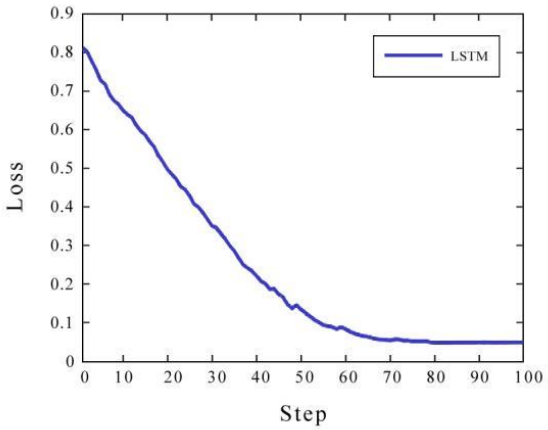

Figure 5: Simulation result

The LSTM network designed in this paper was able to achieve an accuracy of 95.67% in fewer iterations, verifying the effectiveness and efficiency of the MFCC-based LSTM network.

4. Conclusion

This paper uses MFCC features of voiceprint recognition as network model input and feature learning through LSTM neural networks to implement the design of a voiceprint identification system based on LSTM networks, achieving an accuracy of 95.67% for speaker authentication. Based on today's

ISSN 2616-5775 Vol. 5, Issue 7: 85-89, DOI: 10.25236/AJCIS.2022.050714

development trend of neural networks and the widespread of voiceprint authentication, there is an extremely far-reaching development potential to improve the accuracy and stability of voiceprint identification authentication.

References

[1] Du Kangji. Improved recurrent neural network method and its application research [D]. Northeastern Electric Power University, 2021.
[2] DU Xiaoming, GE Shilun, WANG Nianxin. Academic prediction based on CNN___LSTM hybrid neural network model [J]. Modern Educational Technology, 2021, 31(12): 69-76.
[3] Liu Xiaoxuan, Ji Yi, Liu Chunping. LSTM neural network-based vocal pattern recognition [J]. Computer Science, 2021, 48(S2): 270-274.
[4] Yu Lingfei, Liu Qiang. Research and application of deep recurrent network-based vocal recognition method [J]. Computer Application Research, 2019, 36(01): 153-158.
[5] Liu Y. Research on key technology of vocal pattern recognition based on deep learning [D]. Qingdao University of Science and Technology, 2021