Scholar Publishing Group
International Journal of Neural Network
https://doi.org/10.38007/NN.2020.010404
ISSN 2790-2692 Vol. 1, Issue 4: 26-33

Intelligent Speech Recognition and Sentiment Analysis Considering LSTM Neural Network

Deyan Long and Kun Zhan[*]

Anhui University, Hefei, China

[*]*corresponding author*

Keywords: LSTM Neural Network, Intelligent Speech Recognition, Speech Emotion, Attention Mechanism

Abstract: As one of the research hotspots in artificial intelligence application field, speech emotion recognition(ER) is of great value in human-computer interaction system. This paper mainly studies intelligent speech recognition and sentiment analysis considering LSTM neural network. This paper first analyzes the basic flow of speech ER research, analysis of emotion theory. The attention mechanism is used to improve the long term memory network, and the performance of long term memory network for speech ER is optimized. The simulation results show that the optimized model significantly improves the accuracy of recognition.

1. Introduction

Speech emotion recognition is an important direction in the field of Speech research. It is a technology aimed at judging the emotional state of the speaker according to the Speech Speech. It involves signal processing, feature extraction, pattern Recognition, etc. [1]. ER is based on the study of emotion and emotion refers to the person's emotional state, advocating sentiment classification scholars divided the human emotion into a dozen basic emotions, the basic emotion is common in our daily life, such as: happy, sad, angry, etc., all emotions are the basic human emotions and the combination of them. Emotion is purposeful and is a social expression shared by all human beings. All people express basic emotions in a similar way and can be understood by other humans [2]. Due to the universality of emotion across borders, cultures and races and its great significance in social communication, more and more researches have been carried out in the field of ER [3]. With the comprehensive development of information technology in recent years, voice ER plays a wide role in many fields and application scenarios: distance education, emotional video games, telephone customer service and driving assistance systems. In addition to the above, there are numerous speech ER application scenarios, but because of the complexity of the task, the speech ER technology is still cannot achieve the ideal level, the need for more research, innovation and breakthrough, especially in the innovative design on the feature and the algorithm model, the former let model can get voice enough comprehensive information, The latter, as the core part, should make full use of features to provide 4-5 information and improve the performance of ER tasks so that machines can truly understand human emotions [4,5].

In the 21st century, the research of speech ER begins to develop rapidly. The ISCA Conference on Speech and Emotion, held in Ireland, was the first to bring together scholars from around the world who study emotion and phonology [6]. Subsequently, some conferences and journals on affective computing (including speech affective computing) were launched, such as the International Conference on Affective Computing and Intelligent Interaction (ACII), the InterSpeech Emotional Challenge annual competition, the IEEE established the Affective computing theme journal, the first International Audiovisual Emotional Challenge (AVEC) [7]. Speech ER research also joined a growing number of universities and research institutions, such as emotional science center at the university of Geneva Switzerland, cognitive and affective project at the university of Birmingham, UK, the us Massachusetts institute of technology's media lab, Munich university of technology institute of human-computer interaction, palmer at the university of salt complex intelligent systems research institute [8].

Promoting the application development of speech ER will realize the intelligence of human-computer interaction, so as to greatly improve human life, improve people's learning and working efficiency, and promote human beings to enter a highly information society.

2. Speech ER based on LSTM Network

2.1. The Basic Flow of Speech ER Research

The main research process of speech ER is roughly as follows: the first step is the speech emotion information and expression, the collection of data and the establishment and selection of data sets. The data set of ER is relatively important, which is the source of the validity and accuracy of the training model. The second step is to preprocess the speech signal, because the speech signal is analog continuous, so before the model recognition, the speech signal needs to be transformed into a discrete digital speech signal; After the continuous speech signal is converted to digital signal, the speech features are extracted. Finally, the speech recognition algorithm is selected, and the recognition model is established to obtain the speech ER rate [9,10]. The scheme flow of ER by speech signals is shown in Figure 1:

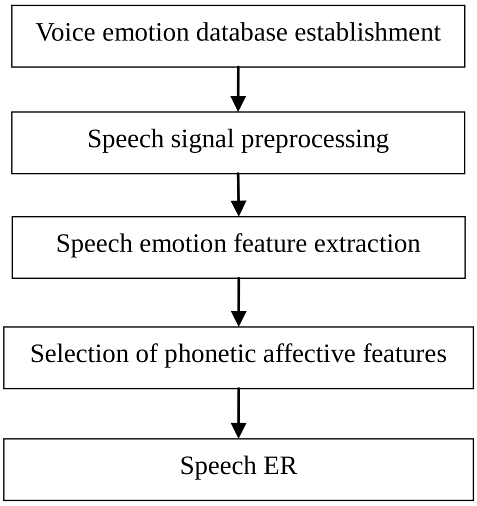
Voice emotion database establishment
Speech signal preprocessing
Speech emotion feature extraction
Selection of phonetic affective features
Speech ER

Figure 1. Flow chart of speech emotion research

2.2. The Emotional Theory

Before the speech ER, the first step is to study and study the emotion theory. After mastering the emotion theory, the speech emotion theory can be further studied [11]. The main research contents of emotion theory include how emotion is formed, the composition of emotion and the composition of emotion expression. These theoretical knowledge provide a basis for the clear analysis of different emotions. This paper expounds the emotion theory from two aspects: on the one hand, dimensional space; on the other hand, finite state set of emotion [12].

(1) Finite state set and dimensional space

So far, by studying and comparing most of the representative affective models, affective theory can be roughly divided into two aspects:

At present, there are two kinds of emotion states in ER: basic emotion state and compound emotion state. Basic emotional states include about 2-11 discrete states [13]. In the aspect of facial expression research, five different emotions of happiness, fear, sadness, anger and disgust have been recognized and selected by more scholars. Some scholars have divided emotions into six categories: anger, boredom, fear, happiness, fun and sadness, and used them in machine emotion modeling [14]. The mixed state of several fundamentally different emotions is called composite emotion expression [15].

Evaluate the spatial model, which is the separation of positive and negative emotions. In practice, arousal degree of emotion to the body belongs to something that arousal degree or activation degree reflects. A scholar proposed a three-dimensional model, whose dimensions are respectively joy - sadness, excitement - quiet, relaxation - tension [16]. People also put forward a four-dimensional model. These four dimensions express respectively the degree of pleasure, the degree of activation, the degree of impulsivity and the degree of certainty, among which the degree of pleasure expresses the main feelings of people. Activation is related to human physiology; Impulsivity expresses the psychological, physical and facial responses to unexpected situations; Certainty mainly expresses the tolerance degree of the subject to emotion [17].

(2) Phonological emotion theory

Sound is one of the main sources of emotional information. Through the experimental analysis of different emotional speech signals, it can be summarized as follows: the emotional characteristic parameters in speech can be learned through the speech ER model, and the emotion in speech signal can be identified at last [18]. The key of speech ER is to enable the model to automatically analyze and recognize the features associated with speech signals and emotion.

2.3. Long and Short Term Memory Network and Optimization

(1) Long and short term memory network

Long Shot Term Memory (LSTM) is a classification and recognition model based on recurrent neural network (RNN) and improved by three gate structures. LSTM can remember the learned speech features for a long time, and solve the problem of gradient disappearance which is not solved by RNN. LSTM has achieved good results in natural language processing (NLP), speech recognition, sequence generation, video analysis and other fields. Its internal structure is similar to RNN. RNN mainly designs a simple unit module, which contains only one activation function TANh or Sigmoid, and uses this function to share parameters. The internal structure of LSTM is the same as that of RNN. The improved part is that this module contains three gate structures, through which input information is screened, and its function is mainly realized by determining the output of

Sigmoid function. Every vector output by Sigmoid layer is a real number between 0 and 1. The lower the probability that the information will pass, the closer to 1, the higher the probability that the information will pass.

LSTM solves the problem of long-term dependence by constantly updating the Cell state through the input gate, output gate, and forgotten gate structure contained in each Cell, so as to control the feature information for a long time.

Forget the door. It is used to determine whether the hidden state in the previous layer Cell should be reserved as input to the next layer Cell. It is mainly represented by probability. Where σ represents the activation function Sigmoid and the probability of activation function through Sigmoid, as shown in Formula (1):

$$f(t) = \sigma(W_f h(t-1) + U_f x(t) + b_f) \tag{1}$$

In the formula, Wf, Uf and BF represent coefficients and bias, the result of output F (t) 0 represents "no information is allowed to pass", and L represents "any information is allowed to pass".

Enter the door. It is mainly used to determine how much information is needed to enter the Cell, which is composed of TANH activation function and Sigmoid activation function, and the output is represented by A (t) and I (t). The output formulas of TANH activation function and Sigmoid activation function are shown in Equations (2) and (3):

$$i(t) = \sigma(W_i h(t-1) + U_i x(t) + b_i) \tag{2}$$

$$a(t) = \tanh(W_a h(t-1) + U_a x(t) + b_a) \tag{3}$$

Where, σ is the activation function Sigmoid, W, U and b represent the coefficient and bias.

Output the door. The main purpose is to update the Cell state and determine the output value of the final output gate through the Cell state. First, the Sigmoid layer determines which part of the Cell state is to be output, and then the TANh layer is to multiply this state with the output state of the previous layer to obtain a value of -L ~+ L, and finally determine the output Oi. Oi calculation method is shown in Equations (4) and (5):

$$O_i = \sigma\left(W_o \cdot [h_{i-1}, x_i]\right) + b_o \tag{4}$$

$$h_i = O_i * \tanh(C_i) \tag{5}$$

Where Oi is the output of the current layer; Hi indicates the status of the hidden layer.

(2) Attention mechanism and improved model

In the experiment of language ER, the speech feature parameters of each frame signal are extracted, but not every frame feature parameters can accurately judge the speech emotion, and different speech feature parameters represent different strength of speech emotion information. If the human brain to accept a voice signal, after listening to quickly identify the voice signal to express emotion category, the process, the brain is focusing on the expression of the most intense emotional speech signal part, such as anger, tone change obviously, the speed will be mutation, then the brain will make judgments, and for other information, will not be so concerned about. In this paper, in the process of building the speech ER model, the attention mechanism can be used to simulate this way to amplify the emotion features of key speech signals and weaken the influence of other signal features.

The Attention Mechanism is mainly inspired by the brain signal processing Mechanism corresponding to the human visual system. For example, when observing an image, humans tend to focus on the important parts and ignore the unimportant parts, so they cannot pay Attention to all the information at once. The idea that the attention mechanism is derived from the formula was put forward by the Google Translation research team, and later received widespread attention from people. It has been introduced into neural machine translation, image description generation, speech ER and other kinds of content processing, and has achieved good results. The attention mechanism is mainly composed of an input mapping Query and a set of key-value pairs. After calculating the weight of each key, the weight coefficient is obtained by the normalization function Softmax. After weighted summation with value, the weight of the corresponding speech features can be obtained.

There are three steps to calculate attention:

The weight. The weight is calculated by the similarity of Query and key, and the calculation method is shown in Equation (6):

$$f(Q,K_i)\begin{cases} Q^T K_i, dot \\ Q^T W_a K_j, general \\ W_a[Q,K_i], concat \end{cases} \tag{6}$$

The normalized. The Softmax function is used to normalize the weight coefficient UI-47, as shown in Equation (7):

$$a_i = soft\max(\, f(Q,K_i)) = \frac{\exp(\, f(Q,K_i))}{\sum_i (\, f(Q,K_i))} \tag{7}$$

Weighted sum. Attention is obtained by weighted summation of weight and key value. The summation formula is shown in (8):

$$Attention\,(Q,K,V)\sum_i a_i V_i \tag{8}$$

3. Simulation Experiment

The experimental platform and related configurations are shown in Table 1:

Table 1. Platform and related configurations

Configuration	Parameter
CPU	I5-9400
Memory	16GB
Operating system	Windows 10
Development platform	Tensorflow

In this paper, three Speech ER models are built: speech-DNN (S-DNN), speech-CNN (S-CNN) and speech-attention-LSTM (S-AL).

4. Comparative Analysis of Experimental Results

4.1. Accuracy of Speech ER

As shown in Table 2, different speech ER methods are compared. It can be seen that the recognition effect of the method based on the long term attention memory network model in this paper is better than that of the other two models, which verifies the effectiveness of the long term attention memory network model.

Table 2. Comparison of speech visual ER results

	Methods	Accuracy
Speech emotion	S-DNN	69.4%
	S-CNN	73.8%
	S-AL	80.2%

4.2. Performance Analysis

Table 3. Recognition rate results of different models

Model	Angry	Excited	Neutral	Sad
S-DNN	65.8%	71.2%	68.4%	72.1%
S-CNN	71.5%	74.9%	75.1%	73.7%
S-AL	81.5%	79.3%	82.1%	77.9%

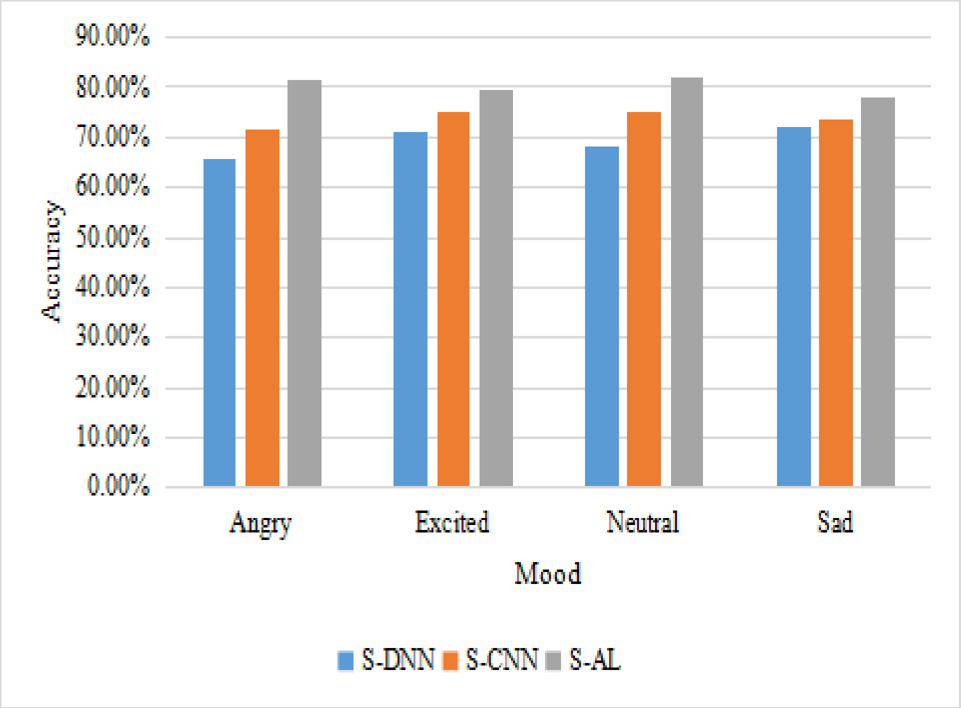

Figure 2. Recognition accuracy of different models under different emotions

As shown in Table 3 and Figure 2, the long term attention memory network model constructed in this paper is superior to the speech deep neural network model and speech convolution network model in different emotions, which verifies that the network model of emotion feature extraction has better performance after optimization.

5. Conclusion

Artificial intelligence has developed rapidly with the progress of social science and technology, and users' demand for human-machine interaction is growing day by day. Humans hope that machines can perceive people's emotional changes and provide them with a variety of high-quality services. Therefore, speech ER technology has practical significance in many fields. In this paper, LSTM+ Attention deep learning model is established as speech emotion classifier. LSTM is responsible for time sequence analysis of feature sequences, and attention machine with P] gives different weight coefficients to speech emotion features. Although the method in this paper improves the performance of ER, and the improved feature parameters and ER classification model have achieved certain effects, there are still great challenges in the field of speech ER, which still needs to be explored continuously.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Dar M N, Akram M U, Khawaja S G, et al. CNN and LSTM-Based Emotion Charting Using Physiological Signals. Sensors. (2020) 20(16):4551. https://doi.org/ 10.3390/s20164551

[2] Shaqra F A, Duwairi R, Al-Ayyoub M. Recognizing Emotion from Speech Based on Age and Gender Using Hierarchical Models-ScienceDirect. Procedia Computer Science. (2019) 151(C):37-44. https://doi.org/10.1016/j.procs.2019.04.009

[3] Henkel A P, Bromuri S, Iren D, et al. Half Human, Half Machine – Augmenting Service Employees with AI for Interpersonal Emotion Regulation. Journal of Service Management. (2020) ahead-of-print (ahead-of-print). https://doi.org/ 0.1108/josm-05-2019-0160

[4] Shinnosuke, Ikeda. Investigation of the Role of Emotion Words on ER from Facial Expression, Affective Voice, and Affective Music by Using Semantic Satiation. Japanese Journal of Research on Emotions. (2018) 26(1):12-18. https://doi.org/10.4092/jsre.26.1_12

[5] Lee, Yoo. Recognition of Negative Emotion Using Long Short-Term Memory with Bio-Signal Feature Compression. Sensors. (2020) 20(2):573. https://doi.org/10.3390/s20020573

[6] Li T, Kuo P H, Tsai T N, et al. CNN and LSTM Based Facial Expression Analysis Model for a Humanoid Robot. IEEE Access. (2019) (99):1-1. https://doi.org/10.1109/ACCESS.2019.2928364

[7] Mersbergen M V, Lyons P, Riegler D. Vocal Responses in Heighted States of Arousal. Journal of Voice. (2017) 31(1):127.e13. https://doi.org/10.1016/j.jvoice.2015.12.011

[8] Green J J, Eigsti I M. Cell-phone vs microphone recordings: Judging emotion in the voice. Journal of the Acoustical Society of America. (2017) 142(3):1261. https://doi.org/10.1121/1.5000482

[9] Massaro A, Savino N, Galiano A M , et al. Voice analysis rehabilitation platform based on LSTM algorithm. International Journal of Telemedicine and Clinical Practices. (2020) 1(1):1. https://doi.org/10.1504/IJTMCP.2020.10034206

[10] Bromuri S, Henkel A P, Iren D, et al. Using AI to predict service agent stress from emotion patterns in service interactions. Journal of Service Management. (2020) ahead-of-print (ahead-of-print). https://doi.org/10.1108/JOSM-06-2019-0163

[11] Bromuri S, Henkel A P, Iren D, et al. Using AI to predict service agent stress from emotion patterns in service interactions. Journal of Service Management. (2020) ahead-of-print (ahead-of-print). https://doi.org/10.1108/JOSM-06-2019-0163

[12] Dimitrova-Grekow T, Klis A, Igras-Cybulska M. Speech ER Based on Voice Fundamental Frequency. Archives of acoustics. (2019) 44(2):277-286. https://doi.org/

[13] Schuller B W. Speech ER Two Decades in a Nutshell, Benchmarks, and Ongoing Trends. Communications of the ACM. (2018) 61(5):90-99. https://doi.org/10.1145/3129340

[14] Vryzas N, Kotsakis R, Liatsou A, et al. Speech ER for Performance Interaction. Journal of the Audio Engineering Society. (2018) 66(6):457-467. https://doi.org/10.17743/jaes.2018.0036

[15] Michalis, Papakostas, Evaggelos, et al. Deep Visual Attributes vs. Hand-Crafted Audio Features on Multidomain Speech ER. Computation. (2017) 5(4):26-26. https://doi.org/10.3390/computation5020026

[16] Alghifari M F, Gunawan T S, Kartiwi M. Speech ER using deep feedforward neural network. Indonesian Journal of Electrical Engineering and Computer Science. Farhoudi Z, Setayeshi S, Rabiee A. Using learning automata in brain emotional learning for speech ER. International Journal of Speech Technology, (2017) 20(3):1-10. https://doi.org/10.1007/s10772-017-9426-0

[17] Kacur J, Puterka B, Pavlovicova J, et al. On the Speech Properties and Feature Extraction Methods in Speech ER. Sensors. (2020) 21(5):1888. https://doi.org/10.3390/s21051888