# Deep learning based on the application of voice emotion

**Jiang Jinyan**

Southwest University, Chongqing, China

120093765@qq.com

**Abstract.** A language is a valuable tool for human development and progress, and it is also an important medium for human beings to transmit information and express emotions. Language signals are ubiquitous, and it is an indispensable part of human life. This article will take the analysis of language as the starting point, combined with the relevant content of computer deep learning, and summarize various methods of language emotion recognition based on a convolutional neural network. In recent years, with the gradual intelligentization of computers, more in-depth discoveries and research have been made on language emotion research. In the deep neural network sector, most of the models used are CNN, LSTM, MO-LSTM models, and this paper aims to propose a new CLDNN (CONVOLUTIONAL, LONG SHORT-TERM) that integrates CNN, DNN, and LSTM into the same network. MEMORY, FULL CONNECTED DEEP NEURAL NETWORKS) model, compare with it, and summarize the advantages and disadvantages of CLDNN.

**Keywords:** convolutional neural network, CLDNN, emotion recognition, LSTM, voice recognition.

## 1. Introduction

Language is an important symbol of human culture, human ability to create and use symbols, and an important carrier of cultural construction and inheritance. Language and tools are the clear dividing line between humans and animals, and it is these two things that have created such a splendid human civilization. Voice has three major elements: shape, sound, and meaning. These elements constitute the rich and colorful language system of human beings. In addition, gestures and even breathing can also be a way for human beings to express their emotions. From ancient times to the present, with the inheritance and reform from generation to generation, the language and cultural system has been relatively perfect.

But the times are also progressing at the same time. With the rise of the Internet, the Internet, mobile phones, and computers have entered thousands of households. Today’s communication is no longer limited to humans and humans, but also exists between humans and computers. In recent years, the emergence of artificial intelligence has pushed human-computer interaction to a climax. Therefore, how to make humans communicate better with computers, and how to make computers better understand the profound meanings and rich emotions contained in human language, has become a popular research object today. Computers are rigid, and how to understand the deeper meanings and emotions expressed by humans has a long way to go, and this road is also extremely challenging.

## 2. Overview of linguistic emotion recognition

### *2.1. Definition of language*

Language is everywhere, it will be transmitted through sight, hearing, or touch, and it will let you know its existence through your various senses. In layman's terms, language is a set of communication tools established by human beings to express their emotions and thoughts. Strictly speaking, "language is an arbitrary, colloquial system used for human communication [1]." Language and culture are inseparable, language needs culture to enrich it, and culture needs language to continue to pass down, it can even be said that if there is no language, then the culture will be difficult to develop and continue. Each group has its own unique language structure, which reflects the richness and inclusiveness of language. Sapir said, "Language is a unique and non-instinct communication method of human beings. It is a symbolic system to express subjective will such as thoughts, feelings and desires."

### *2.2. Definition of emotion*

"Emotions are a reflection of people's attitudes toward objective things [2]." Attitudes hold emotions, and emotions can be specifically expressed as: pleasure, sadness, happiness, hatred, disgust, and so on. The Dictionary of Psychology believes that "emotion is the attitude and experience of people on whether objective things meet their needs".

Like humans, computers can observe and understand various emotions from various senses, so a set of independent algorithms is needed to analyze them. Human emotions are rich and complex, and can express their emotions through speech intonation, writing style, body movements, facial expressions and so on. "And affective computing allows computers to continuously learn through deep learning methods, and finally through a large number of simulation fittings, they become as able to understand the emotions and emotions contained in human language as human beings [3]."

### *2.3. Ways of expressing emotions in language*

*2.3.1. Emotion recognition system based on body language.* Body language mainly expresses emotions through body parts such as head, eyes, neck, hands, elbows, arms, body, hips, feet, etc. For example, when chatting with close people, they will involuntarily approach each other. Research on body language is relatively mature. For example, document [4], which constructed a human motion emotion data set in daily life scenes, and joint LMA theory with basic emotion model to design and calculate the characteristic information of LMA, from body, force, space, shape, etc. Four aspects are analyzed.

*2.3.2. Emotion recognition system based on facial expression.* Facial expressions are an indispensable part of the human language expression system. They can visually make people quickly understand the real-time emotions of the other party. The emotional expression of facial expressions is difficult to control, and it is also compared to the real ones. When feeling embarrassed, look down involuntarily and so on. In [5], 48 volunteers were recruited to complete three tasks with or without time limit, and data such as expressions and limbs of the volunteers were collected for calculation. Reference [6] recruited 20 volunteers and asked them to pose with different facial emotions, and then extracted multi-scale features by using biorthogonal wavelet entropy, compared various indicators, and accurately calculated and identified different emotions of users.

*2.3.3. Spectrogram-based emotion recognition system.* Sound is produced by the vibration of objects, and the sound frequency and sound loudness corresponding to different emotions are different. For example, when a person is angry, the sound loudness and frequency will be much higher than the level of normal speech. A spectrogram is a picture of a sound, which holds the signal strength at different times and frequencies and displays three-dimensional data in a two-dimensional representation. The spectrogram emotion recognition system works on the principle of signal strength. In the literature [3], the LeNet-5 network is used as the research basis. By stacking convolutional layers and pooling layers, L2 regularization is used to prevent overfitting, and continuous feature learning and simulation training are performed. Finally, on the commonly used public datasets good classification results

## 3. Neural network structure

### *3.1. LSTM*

Long Short-Term Memory (LSTM, Long Short-Term Memory) is a time-cycle neural network, "its key role is to solve the problem of RNN gradient disappearance, because the long-short-term memory network has independent "memory" gate units, also can handle important events with very long delays in forecasting time series. [7]"

In the literature [8], "the unique MO-LSTM structure is used to evaluate the continuous sound signal, and the core structure is still the long short-term memory network LSTM. The input transmits the data into the two-layer LSTM structure, one of which is Forward propagation LSTM, another layer is backward propagation LSTM, and then enter the linear layer, the main function of the linear layer is to perform the learning paranoia function and tile display. Finally, the corresponding four types of output emotion probability values are displayed and given to Label layer, the label layer compares the probabilities, and selects the most likely emotion as the result of this signal recognition. This network structure is not only computationally efficient, but also more difficult to operate and more demanding on equipment than other methods. "

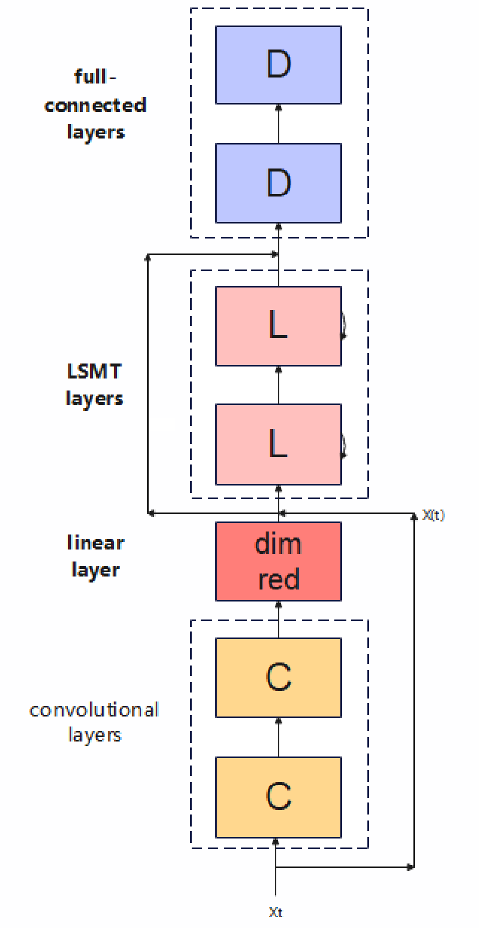

**Figure 1.** CLDNN model diagram.

### *3.2. CLDNN*

CLDNN has a wide range of applications in the field of speech recognition. The general structure of CLDNN (Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks) is several layers of CNN layers, connected to several layers of LSTM layers, and finally divided into full-scale layers according to actual conditions. Connect DNN layers or connect DNN layers using global averaging. Document [9] "Use 300h Chinese noisy voice to use the CLDNN structure for model training. The CLDNN structure, the CNN part is a two-layer CNN, the first layer pooling specification is 3, and the second layer is not connected to the pooling layer, because the second layer is not connected to the pooling layer. The output dimension of the layer CNN is large, and a linear layer is used for dimensionality reduction. The CNN is followed by two layers of LSTM. Finally, after modeling in the frequency and time domains, the output layer of the LSTM is connected to several layers of fully connected DNN layers." As shown in figure (3.2).

## 4. Discussion

### *4.1. The processing of spectrograms*

The essence and purpose of spectrogram processing is to convert the speech signal containing 1D time domain features into 2D signals containing both frequency domain and time domain features, and the most important is to convert 1D feature signals into 2D signals.

First, the speech signal should be weighted, which can highlight the features we need in the frequency domain, while filtering the features we do not need. Then, a complete speech signal is divided into many short speech frames. Then Fourier transform each speech frame. Finally, it is reintegrated into a complete speech signal.

In most cases, a high-pass filter is used to implement the weighted processing operation for winning. To filter out the low-frequency interference we don't need, while enhancing the high-frequency signature. As shown in Equations (3.1) and (3.2).

$$C = \frac{1}{2\pi Rf} \tag{3.1}$$

$$F(x) = 1 - \partial x^{-1} \tag{3.2}$$

"Pre-weighted audio signals are typically split into frames in the range of 20 to 40 milliseconds. [4]" To make the speech continuous, the overlapping segmentation method is usually used to keep the speech information between two adjacent frames unchanged. Then, hamming window function is added to the speech signal after frame segmentation to make the two ends of the signal smooth transition, and the windowed speech signal is obtained. As shown in Equations (3.3)

$$I_w(n) = I(n) \times w(n) \tag{3.3}$$

$I(n)$ represent the original speech signal, $w(n)$ stands for the Hamming window, and $I_w(n)$ stands for the windowed speech signal.

Hamming window:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right], & 0 \le n \le N-1 \\ 0, & otherwise \end{cases} \tag{3.4}$$

By combining equations (3.3) and (3.4), the formula about $I_w(n)$can be obtained.

$$I_w(n) \begin{cases} 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right] \times I(n), & 0 \le n \le N-1 \\ 0, & otherwise \end{cases} \tag{3.5}$$

After adding the Hamming window function, the Fourier transform (SFFT) calculation is performed on each speech signal to realize the conversion from the time domain signal to the frequency domain signal. Then, the data is substituted into the CLDNN neural network, and the feature map Fc is multiplied with the current feature map to obtain a feature map with convolutional attention weight. After obtaining the corresponding output results, the results are repeatedly trained and compared, corresponding to different emotions according to different data results, and finally the results are visualized to achieve the final function.

## 5. Conclusion

Summarizing the current development status of speech recognition, DNN, RNN/ISTM and CNN are several mainstream directions in speech recognition. In 2012, Microsoft Deng Li and Yu Dong introduced the feedforward neural network FFDNN (Feed Forward Deep Neural Network) into the acoustic model modeling and used the output layer probability of FFDNN to replace the output calculated by GMM in the previous GMM-HMM. Probability, leading the trend of DNN-HMM hybrid system. Long Short-Term Memory (LSTM, Long Short-Term Memory) can be said to be the most widely used structure in speech recognition. This network can model the long-term correlation of speech, thereby improving the recognition accuracy. Bidirectional LSTM networks can achieve better performance, but at the same time, there are problems of high training complexity and high decoding delay, which are especially difficult to apply in real-time recognition systems in the industry.

Although the current level of science and technology has reached a good recognition accuracy, the overall average recognition accuracy has not reached more than 90% [4]. Therefore, more data and experimental results are needed to strengthen the reliability and authenticity of the model. At present, the emotions that can be recognized by the system are relatively single, but in fact, human emotions are diverse, so emotion monitoring can not only be done through voice, body, and other ways. In the future, some physiological signals such as heart rate and blood pressure may be added to enrich the algorithm and model. If more kinds of emotions can be identified, it will also provide valuable experience and solutions for the research of emotion recognition methods.

## References

[1] Facial Liu Runqing. The School of Western Linguistics [M]. Revised Edition. Beijing: Foreign Language Teaching Press, 2013

[2] Cao Richang, "Tong Psychology" Volume II, p. 44, People's Education Publishing House, 1979 edition

[3] Yang Lijia Speech Emotion Recognition Algorithm Based on Spectrogram 2021

[4] Jiang Xiuhao Emotion Recognition Based on Human Movement 2021

[5] Huang Kun, Zheng Mingxuan, Luo Shichao, Jin Jian, Research on User Emotions and Influencing Factors Based on Facial Expression Recognition in Exploratory Search 10.13266/j.issn.0252-3116.2022.05.010

[6] Zhang Y D, Yang Z J, Lu H M, et al. Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation[J]. IEEE Access, 2016, 4:8375-8385.

[7] Comprehensive Application of CNN and LSTM in short-term stock price rise and fall prediction of cyclical stocks 2022.6

[8] Yan Yanchao Research on emotion recognitionmethod based on continuous wave sound signal 2021.5

[9] Hou Yixin explains in detail the application of convolutional neural network (CNN) in speech recognition.