Speech Emotion Recognition Application for Education

Weijia Xian

Sichuan University, Chengdu 610065, China

Abstract. Based on convolutional neural networks, a speech recognition application capable of analyzing human emotions is designed. This speech emotion recognition can better assist teachers to understand students' emotional status in the learning process and enable them to improve their teaching methods with the help of the system, thus achieving the goal of improving students' learning efficiency. The application is based on PAD dimension, convolutional neural network to extract deep speech emotion features, and Least squares support vector machine for emotion recognition, thus improving the recognition accuracy of this application.

Keywords: Speech Emotion Recognition; PAD Dimensions; Convolutional Neural Network (CNN); Least Squares Support Vector Machine (LSSVM).

1. Introduction

Speech contains rich emotional information, and emotion plays a crucial role in speech interaction. Speech emotion recognition refers to computer simulation of human emotion speech perception and understanding process, that is, automatic recognition of the emotional state of speech signals[1]. The research of computer emotion recognition technology can be analyzed from two aspects: one is through facial expression; The second is voice. In the process of communication, voice is the most favorable and direct way to know the motivation and emotion of the other party. The reason why speech can express different emotions is that speech signals contain parameters that can reflect emotional characteristics. It is believed that the phonetic parameters caused by a particular emotional state are roughly the same among different people. Therefore, the computer can use the way of extracting the emotion feature of speech to recognize the human emotion. In recent years, the research on emotion features is constantly enriched. For example, Zbancioc et al. apply improved MFCC and LPCC features to emotion recognition, and the recognition rate reaches 75%[2]. Sun Ying et al. extracted the nonlinear geometric features and optimized the feature parameters to obtain the optimal nonlinear features. The above features are analysis and research of emotion from the perspective of signal processing[3]. The generation of emotion involves people's mental activities.

This application can make teachers to know the emotional status of students in the learning process, so that teachers can adjust their teaching methods through the help of the system, so as to achieve the purpose of improving the learning efficiency of students.

2. Principle of the Speech Emotion Recognition Application

The most important part of the application is the speech emotion recognition system. Similar to speech recognition system, speech emotion recognition system is divided into three parts: speech emotion statement preprocessing, feature parameter extraction and pattern matching. The recognition process is as follows: Firstly, the emotion statement is preprocessed; Secondly, feature parameters are extracted from sentiment statements. Then on this basis to establish the template, the process of establishing the template is called the training process; The process of matching feature parameters with patterns is called recognition process.

2.1 Emotion Model

In order to recognize the emotion of speech, it is necessary to make clear what emotion is. In fact, "what is emotion" has always been a contentious issue. The research of emotion involves the fields of psychology, physiology, medicine, society, science and computer science. Although many scholars

have done a lot of work on "what is emotion", researchers in different fields have made different definitions of emotion, and so far, no unified conclusion has been reached[4].

According to the literature[5], emotion is a dynamic experience of whether objective things meet one's needs. Changes in human physiological information will cause people to have emotions, and the emotional state of people will react on people, resulting in changes in human physiological functions (such as blood pressure, heart rate, body temperature, etc.). Different people tend to show different emotional states at different times and in different situations. To study emotions in depth, researchers need to model them. There are two main models used to describe human emotion: discrete emotion description model and continuous emotion description model.

Human emotions are subtle and complex. For example, emotions such as sadness and joy, tears of joy and mixed feelings do not completely belong to a certain basic emotion category, which poses new challenges to intelligent human-computer interaction. The continuous space theory of emotions provides a solution to this problem[6]. This theory proposes that human emotions are composed of spatial dimensions, which can cover almost all emotion types, and different emotions can change continuously and smoothly. A typical continuous emotion model is PAD 3D emotion model[7], which divides emotion into three dimensions as follows: P represents pleasure-displeasure, representing the positive and negative characteristics of individual emotional states; A means arousal-no arousal, indicating the degree of neurophysiological activation. D indicates dominance-submissiveness, indicating the individual's control over the situation and others. As shown in Fig. 1, any sentiment corresponds to a point in the 3D space.

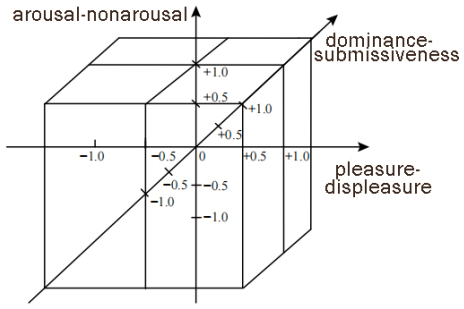

Fig 1. PAD three-dimensional emotion model

2.2 Convolutional Neural Network

CNN is a feedforward neural network proposed by LeCun et al., 1989 [8]. It has translation invariance to input information and is widely used in computer vision, natural language processing, speech recognition and other fields. Convolutional neural networks consist of one or more convolutional layers, pooling layers and fully connected layers.

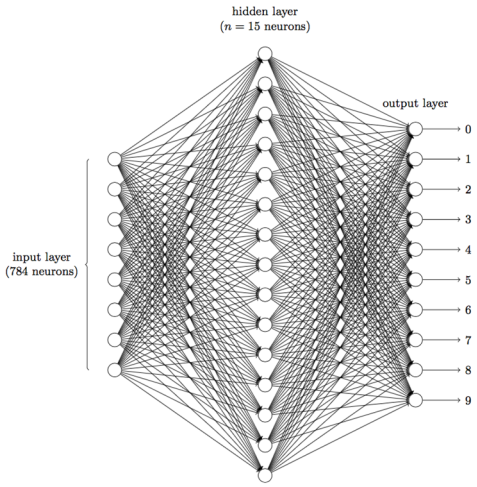

Fig 2. A typical convolutional neural network (CNN)

Convolutional layer is the main component of CNN, and its main function is to extract the features of input data through convolutional kernel. There can be multiple convolution kernels, and different convolution kernels can extract different features. Taking stride =1 as an example, assuming that the input matrix is *I*, the convolution kernel is *K*, *K* is a two-dimensional matrix with the shape of m*n, the bias of the convolution layer is ω_b, f is the activation function, and the output matrix is *O*, then the output obtained through the convolution layer can be expressed as:

$$O_{i,j} = f(\sum_m \sum_n I_{i+m,j+n} K_{m,n} + \omega_b) \tag{1}$$

Where, $I_{i+m,j+n}$ and $K_{m,n}$ denote the multiplication of elements with position (i+m, j+n) on input matrix I and elements with position (m, n) on convolution kernel matrix K. In this process, the input matrix shares the same convolution kernel, which not only greatly reduces the parameters in the neural network, improves the computational efficiency, but also reduces the overfitting and other problems.

If the input of CNN is an image, the size of its convolution kernel can be understood as the range on the image that the human eye can see, which is usually called the receptive field. The larger the size of the convolution kernel in CNN, the larger the receptive field, the larger the range of input data that CNN can see, and the better performance can be obtained theoretically. However, large convolution kernels will lead to computational efficiency reduction, overfitting, gradient disappearance and other problems, thus reducing the performance of CNN. To solve this problem, Yu et al [9]. first proposed void convolution in 2016, which can increase the receptive field without losing resolution.

2.3 Least Squares Support Vector Machine (LSSVM)

LSSVM algorithm introduces the least square linear theory into the support vector machine, and improves the support vector machine. It seeks the nonlinear relationship between the input and output in the mapped high-dimensional space, and then reflects the regression into the original space to get the regression, which reduces the computational complexity[10]. The basic principle of LSSVM is as follows.

Let the sample set $\{x_i, y_i\}, i = 1,2, \dots, n$, where $x_i \in R$ is is the input quantity, $y_i \in R$ the corresponding output quantity,n is the size of the sample set, and the samples are mapped to the high-dimensional space through $\Phi(x)$, and the optimal decision function $y = \omega^T \Phi(x) + b$ is constructed (where ω is the weight vector and b is the deviation). For the input sample x, $|y_i - \omega^T \Phi(x_i) - b_i| \leq e_i$, so the LSSVM optimization problem is

$$\min J(\omega, e) = \frac{1}{2}\omega^T \omega + \frac{1}{2} C \sum_{i=1}^{n} {e_i}^2;$$
$$s.t. \;\; y_i = \omega^T \Phi(x_i) + b_i + e_i, i = 1,2, \dots, n. \tag{2}$$

Where, C is the regularization parameter; e_i is the error variable, $e_i \in R$. Lagrange's method is used to solve the optimization problem, which is transformed into a linear problem: $\begin{bmatrix} 0 & I^T \\ I & K + I/C \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}$;

Where, $I = [1,1, \dots, 1]^T$ is a vector composed of n ones and is an identity matrix of order n.$y = [y_1, y_2, \dots, y_n]^T$; K is the kernel function, $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j), i, j = 1,2, \dots, n$; $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n]^T$ is Lagrange multiplier vector, RBF kernel function $K(x, x_i) = exp(-\|x - x_i\|^2 / (2\sigma^2))$ is adopted, where σ is kernel function width, LSSVM model can be obtained as follows:

$$y(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b \tag{3}$$

The final mapping relationship is shown in the above equation. In this study, *x* represents emotional speech features, and *y(x)* represents the values of P, A and D of emotional dimensions.

3. Experiment and Results

In our app, we collect voice data when children are reading. We analyse the data and get children's three-dimensional continuous emotion state (pleasure-displeasure, arousal-nonarousal and dominance-submissiveness).

We use this kind to data to judge how much work should be assigned to the children. If children are at a more pleasant and arousal state, we could assign more work to them. In a hypothesis scenario, a child has spent his entire afternoon playing football, after that he decided to do some reading work through an AI reading app. However, he felt tired in the process and therefore has a low study efficiency. In this scenario we could detect his emotional state and assign less work to him that day and more work to him later. In this way, we can improve student's efficiency.

To judge whether this method actually works, we could do an A/B test for children. We could randomly assign users into two groups. One group uses this kind of method and the other does not. We make sure two groups have the same learning time and the same amount of work. After a month, we could evaluate their study performance (For example, how many words they have mastered on average and how much their speaking fluency has improved. If the group used this method performs better, we could tell that this method really can improve children's study efficiency.

References are cited in the text just by square brackets [1]. (If square brackets are not available, slashes may be used instead, e.g., /2/.) Two or more references at a time may be put in one set of brackets [3, 4]. The references are to be numbered in the order in which they are cited in the text and are to be listed at the end of the contribution under a heading *References*, see our example below.

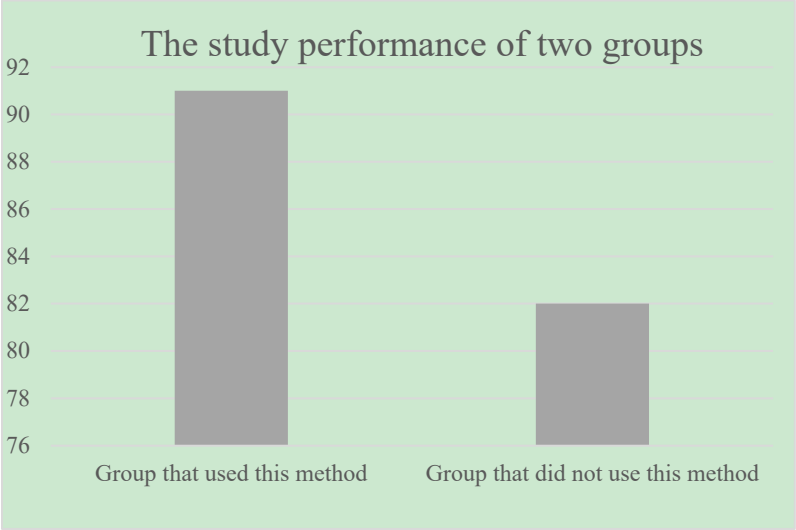

Fig 3. The study performance of two groups

We could do an experiment to judge whether this method can make student more willing to use this study app. We could do another A/B test for children. We could randomly assign users into two groups. One group uses this kind of method and the other does not. After a month, we could evaluate the average time children spend on this app. If the group used this app spend more time on this app, we can tell that we make this app more pleasant to use and have improved the users' experience.

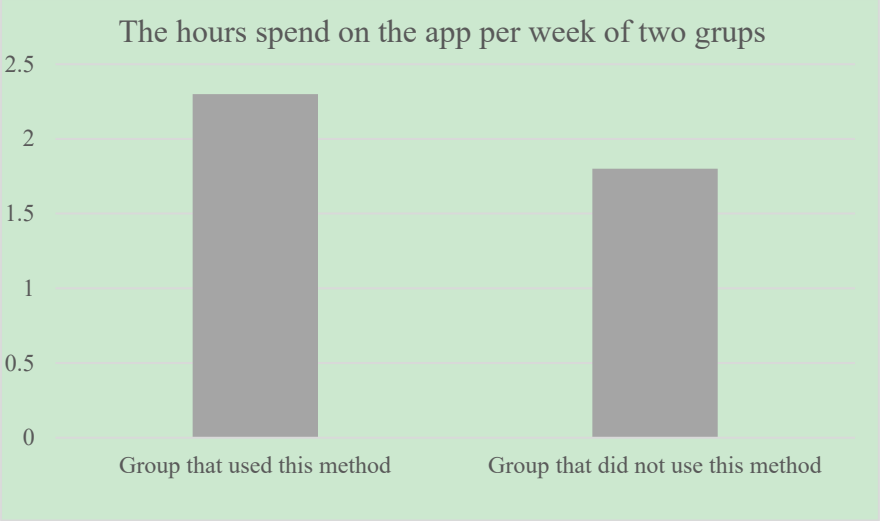

Fig 4. The hours spend on the app per week of two groups

4. Summary

In the process of interaction between artificial intelligence (AI) products and people, if the current emotional state of people can be accurately grasped and responses made according to the emotional state, it can greatly enhance the user's experience of AI products.

Emotion recognition can be applied to many real-world scenarios like educating. We can use this method in the teaching process so that the pace and manner of teaching can be adjusted. In this way we could improve students' study efficiency and increase study time students use on this application.

In the future, emotion recognition algorithms can be improved using some new methods like multi-modal fusion technology. We could combine facial expressions and voice tone data to get better emotion recognition accuracy.

References

[1] Han WJ, Li HF, Ruan HB, et al. A Review of Research Advances in Speech Emotion Recognition[J]. Journal of Software, 2014, 25(1): 37-50.

[2] ZBANCIOC M D, FERARU M. Using the Lyapunov exponent from cepstral coefficients for automatic emotion recognition [C] // International Conference and Exposition on Electrical and Power Engineering. Iasi, Romania: IEEE, 2014: 110–113.

[3] SUN Ying, SONG Chun-xiao. Emotional speech feature extraction and optimization of phase space reconstruction [J]. Journal of Xidian University: Natural Science, 2017, 44(6): 162–168.

[4] Cha Cheng. Research on speech emotion Recognition Algorithm based on feature Learning [D]. Nanjing: Southeast University, 2017:1-2.

[5] Li Danyan. Research on speech emotion Recognition based on deep Learning [D]. Beijing: Beijing University of Posts and Telecommunications,2020:7.

[6] Wang Li. Research on dimensional and continuous emotion prediction in valence-arousal space MEHRABIAN A. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament [J]. Current Psychology, 1996, 14(4): 261–292.

[7] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[8] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[C]. International Conference on Learning Representations, 2016: San Juan, Puerto Rico.

[9] WANG Jian-xin, CHEN Xiao-jie. Application in sintering process modeling using the feature selection algorithm of least squares support vector machine [J]. Machinery Design and Manufacture,2018(3): 75–77.

[10] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech. [C].International Speech Communication Association, Lisbon, Portugal, 2005: 1517-1520.

[11] Wang W, Wu J. Notice of Retraction: Emotion recognition based on CSO&SVM in e-learning[C]//2011 Seventh International Conference on Natural Computation. IEEE, 2011, 1: 566-570.

[12] Poria S, Cambria E, Bajpai R, et al. A review of affective computing: From unimodal analysis to multimodal fusion [J]. Information Fusion, 2017, 37: 98-125.