Computer Science and Information Systems 21(1):57–74 https://doi.org/10.2298/CSIS230509066J

Feature Parameters extraction and Affective Computing of Voice Message for Social Media Environment

Peng Jiang[1], Cui Guo[2], and Yonghui Dai[3,*]

[1] Jingan Branch Campus, Shanghai Open University,
Shanghai 200040, China
jzhpmail@163.com

[2] Shanghai Lifelong Education School Credit Bank Management Center,
Shanghai 200092, China
guoc@sou.edu.cn

[3] Management School, Shanghai University of International Business and Economics,
Shanghai 201620, China
daiyonghui@suibe.edu.cn

Abstract. Voice message in social media environment includes a large number of conversation natural languages, which increases the difficulty of emotion tagging and affective computing. In order to solve the above difficulties, this paper analyzes the cognitive differences between the semantic and acoustic features of voice message from the perspective of cognitive neuroscience, and presents a voice feature extraction method based on EEG (Electroencephalogram) experiments, and gets the representation of 25 acoustic feature parameter vectors. Meanwhile, we proposed an affective computing method based on PAD (Pleasure-Arousal-Dominance) dimension emotional space according to the above parameters. Experiments show that the method can effectively solve the affective computing problem of voice message. Overall, there are two main contributions of this paper. Firstly, it comprehensively analyzes the emotional cognitive feature of voice message in social media environment from the perspectives of cognitive neural mechanism, voice acoustic feature and text semantics. Secondly, the segmented affective computing method for voice message based on acoustic feature parameters and PAD emotional state model is proposed.

Keywords: affective feature parameters, cognitive neuroscience, voice acoustic feature, emotion recognition.

1. Introduction

The rise of mobile communication technology and online chat tools has provided the foundation for the development of social media. Wechat, QQ, WhatsApp network communication tools are widely used by people, more and more people exchange information and express opinions through the above tools [32]. The importance of social media in information dissemination and social influence is gradually increasing. From the perspective of communication feature, social media has the feature of fragmented communication content and diversified communication subjects [15]. It is a synthesis of

micro-content, micro-channel and micro-experience. As the most primitive and natural information transmission method for human beings, voice has good convenience and rich emotional feature [30]. Voice message has become a popular way of social communication, and more and more people communicate through the form of voice social communication.

In recent years, the improvement of voice recognition technology has provided support for the promotion of voice social networking. The report released by Baidu company shows that the accuracy of Chinese Mandarin voice recognition is close to 99.8% in a quiet environment, and the technology will be widely used in Baidu voice search product in the future. Because voice not only transmits semantic information, but also contains rich personalized emotional features. The research results and practical cases in recent years have shown that the dissemination of voice message in social media has a huge infectious and synergistic effect, and its emotion has a significant impact on network group cognition, psychology and behavior [3,16]. Great progress has been made in the research of speech emotion recognition in the past few decades. Scholars' research has developed from template matching-based emotion recognition for specific person and simple vocabulary speech to today's statistical model-based emotion recognition for large vocabulary, non-specific person, continuous speech [13, 29]. Many algorithms have been applied to affective computing of speech, such as: K-Means clustering, ant colony, EMD, HMM, SVM algorithms are used in speech recognition [4, 17,].

Previous research provides effective methods and technical tools for solving speech recognition problems in real environments. However, the voice message used in the social media environment such as WeChat, QQ and WhatsApp, which is often completed in the context of the 'small world network' of phrase-based dialogue. The emotion recognition and semantic analysis of voice message are very complicated, and it involves emotional cognitive characteristics. If we only rely on traditional classification methods for affective computing, the effect is not very good. In recent years, EEG technology has made a lot of achievements in cognitive neuroscience research such as the cognitive characteristics of images, voice, text, etc. Therefore, EEG experiments are adapted for our study. In addition, the emotional state of these voice messages is often dynamic, and the previous discrete affective computing methods is not very effective. Therefore, our research has practical significance and good application value.

2. Literature Review

2.1. Affective Feature Parameters of Voice Message and Speech Emotional Model

Some scholars selected ten features such as voice duration, energy, maximum, minimum, median and formant of pitch frequency as emotion recognition features in 2001. Their recognition experimental results of 100 emotion test sentences show that the above features can effectively recognize sadness [40]. Gaussian mixture model was used

to extract the spectral features of speech emotion signals for speech emotion recognition by scholars, their experimental results show that this method has good effect and high recognition efficiency in identifying five emotions: anger, fear, happiness, neutrality and sadness [37]. Some scholars used global control Elman neural network to carry out emotion recognition experiments on speech fused with long-term and short-term features, and achieved a recognition rate of 66.0%. Their experimental research shows that the best recognition segment length of anger, happiness, sadness and surprise is consistent with the prosodic phrase length corresponding to the emotion state [14]. After that, the kernel principal component analysis algorithm, MFCCG-PCA algorithm was proposed. Compared with the general MFCC model, the performance of their model in speech emotion recognition is greatly improved [6].

At present, the existing literature mainly classifies voice emotion based on the traditional six discrete emotions. There are limitations in the analysis of context information and semantics of voice message, and it is difficult to effectively complete emotion recognition in social media environment. Our research is based on the voice emotion recognition, and uses the PAD emotion model of three-dimensional space. The above method can map the emotion types in space, and overcome the discontinuity of emotion classification. It contributes to the research of emotion recognition of voice message in social media environment.

Speech emotion recognition is the ability of a system to recognize human emotions from speech [1]. In order to study speech emotion recognition, phonological affective corpus was proposed, and it was divided into discrete affective corpus and dimensional affective corpus. The former was labeled with language tags, and the latter was labeled with affective spatial coordinates. The representative discrete affective databases with wide influence mainly include: Belfast affective database, Berlin EMO-DB affective phonetic database, FAU AIBO children's German affective phonetic database, CASIA Chinese affective corpus and ACCorpusseries affective database. The above emotional speech database is deduced by the recorder with several kinds of emotions (happiness, anger, neutrality, fear, sadness), which belongs to the performance or guidance emotional database. VAM emotional database selects three emotional dimensions of value, activation and dominion to label [-1, 1]. The database can be used for natural speech understanding and analysis, speech emotion recognition, robust speech recognition research. a fuzzy model of multi-level emotion calculation was constructed according to the emotion dimension to detect various emotions [2]. The recording of Semaine emotion database is carried out in the scene of human-computer interaction, which can be used by researchers free of charge. The audio data is about 7 hours long, sampled at 48Khz, quantized at 24bit, and labeled in five emotional dimensions: value, activation, power, expectation and intensity [26]. Some scholars have studied the PAD emotional model, in which the emotional state is described as a three-dimensional space of Pleasure-Displeasure, Arousal-Nonarousal and Dominance-Submissiveness. So far, PAD emotional model has been widely used in audio-visual speech synthesis, music emotion comparison and speech emotion recognition. In addition, some scholars have studied the speech emotion recognition based on the fusion of HMM and probabilistic neural network. PNN is used to deal with the statistical features in the acoustic feature parameters, HMM is used to deal with the temporal features in the acoustic feature parameters, and then fused through addition and multiplication rules [10].

2.2. Related Research on Brain Cognitive and Neuroscience

Neurocognitive science research shows that the amygdala, prefrontal cortex, hypothalamus and cingulate cortex in human brain structure are related to the processing of emotion, and EEG signals reflect people's emotional state to some extent [22]. In the brain cognitive research of speech emotion, scholars used various emotional speech and music clips as stimulation materials to carry out EEG and ERP experiments and collect data in real time [19, 28]. After denoising, extracting and analyzing the experimental data, they have made a series of important results in the research of speech emotion and brain cognitive function. Some scholars have found that happiness and sadness can cause beta waves changes in the frontal, temporal and parietal regions, and the brain's processing of emotions is related to low-frequency beta waves. In the frontal area, happiness is activated β Wave is more intense than sadness, especially at the frequency of β waves (19.50~25.45 Hz) are the most active [23]. The left frontal lobe related brain area of human brain is easy to be activated by happy and happy speech materials, while the right frontal lobe related brain area is more sensitive to and easier to be activated by fear and sadness speech materials.

The cognitive process of human brain can be reflected by a series of ERP (Event Related Potential) components such as SPN (Stimulus Preceding Negativity), FRN (Feedback Related Negativity) and P300 [31, 35, 41]. Among them, SPN is a continuous and negative slow wave, which responds to attention with expectation, which often appears before the start of task-related stimuli and is closely related to human emotional intelligence [20]. FRN is an important EEG component of brain processing feedback information, and it is mainly distributed in the frontal central region. Its peak appears about 250 milliseconds after the emergence of feedback stimulation, which is closely related to the cognitive evaluation and learning ability of the brain [12]. The latency of P300 component is about 300-600 milliseconds. A large number of experimental studies have found that it is a positive wave of about 300 milliseconds, which is closely related to cognitive processing processes. For example, the allocation of attention resources, the updating process of working memory and inhibition processing [5, 11].

In recent years, the research of cognitive neuroscience has developed rapidly, and its intersection with other disciplines has become a frontier research hotspot. For example: neuro-management, neuromarketing, neuro-entrepreneurship, etc. [9, 33]. The research results of cognitive computing framework and decision-making neural mechanism proposed by scholars on the basis of cognitive neuroscience [34], which provide a reference for our research. One of the innovations and contributions of this paper is to study the cognitive characteristics of the emotional features of speech, which expands the application of cognitive neuroscience.

3. Framework and Method of Voice Affective Computing

3.1. Framework of Voice Affective Computing

In the research of affective computing of voice message, the selection of feature parameters is very important. The cognition of speech information includes not only the cognition of representational information, but also the cognition of semantic information. Among them, the representational information refers to the information that can trigger the rapid initial emotional response of the human brain in a short period of time. Semantic information refers to information that can only produce relatively slow secondary emotional responses after being recognized by the higher cortex of the brain. From the perspective of cognitive neuroscience theory, the physiological structure of human beings receiving speech is the auditory system, which is dominated by the brain, so the cognitive process of speech emotion is the process of the brain's cognition of speech and triggering the corresponding emotional experience. The information conveyed by speech can be divided into two main categories: acoustic information and semantic information [27]. The feature of acoustic information such as pitch, intensity and speed of speech. The semantic information features refer to the characteristics of textual information in speech. Previous studies on brain cognition of speech information have shown differences in cognitive time between acoustic features and word semantic features [21]. With the application of electroencephalography (EEG) in emotion recognition, this paper also uses EEG to perceive emotional features for voice social media. The framework is shown in Fig.1.

In Fig.1, after the voice is preprocessed by phonetic unit interception and voice converted text, affective computing will be performed from both the acoustical features of the voice and the semantic features of the text. Among, speech emotion classifiers mainly include GMM (Gaussian Mixture Model), SVM (Support Vector Machines), HMM (Hidden Markov Model), ANN (Artificial Neural Network) etc. In our study, the combination of SVM and artificial neural network classifier was used for voice emotion classification. There has been a lot of research on the cognition of semantic information in the field of text information communication, but there is still a lack of in-depth research on acoustic feature in the social media environment. This paper studies the response of human's emotional cognition to acoustic features through EEG experiments, so as to provide guidance for the affective computing of voice message and the selection of feature parameters.

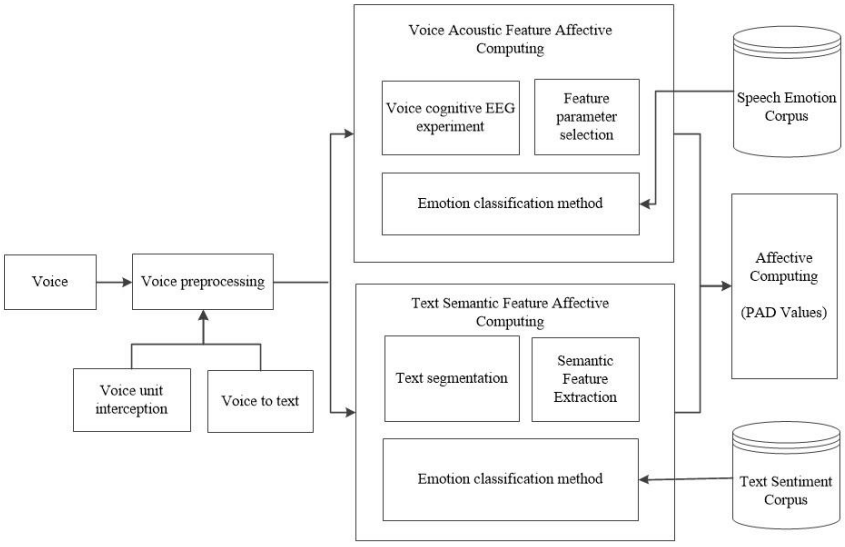

Fig.1. The framework of voice affective computing

3.2. EEG Experiments and Voice Feature Parameters

Considering that the cognition of voice has nothing to do with the identity of subjects, all subjects in EEG experiment come from college students. A total of 15 subjects were recruited in this experiment, aged from 19 to 29 years, with an average of 23.6 years. Among them, 12 (8 males and 4 females) participated in the main experiment and 3 (2 males and 1 female) participated in the control group experiment. All subjects are physically and mentally healthy, right-handed, with normal or corrected-to-normal vision, and no history of mental illness. They had never participated in the EEG experiment before and signed the informed consent before the experiment.

The experimental materials include two parts. Material 1 includes CASIA emotional voice database, which is provided by the human-computer voice interaction research group of the State Key Laboratory of pattern recognition, Institute of automation, Chinese Academy of Sciences. The material is composed of four speakers (two men and two women) who used six emotional types of joy, anger, surprise, sadness, fear and calm, and pronounced 50 sentences once each. It includes 1200 emotional sentence voice, and the sampling rate of each sentence is 16KHz. Material 2 includes 90 units of audio files are collected from social media, film and CD piano music. Each file is divided into audio segments by 8 seconds, and the above segments were labeled with PAD values. The labeled PAD values of these materials are composed of three dimensions: polarity, arousal and dominance, and the range is integer of [-4, 4]. Polarity represents happiness or not, arousal means calm or excitement, and dominance refers to a controlled state.

The processing of EEG experiment is shown in Fig.2.

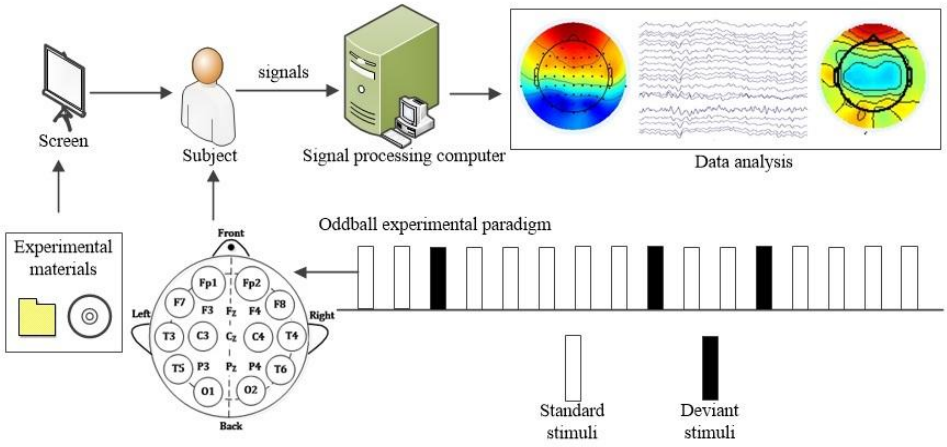

Fig.2. The processing of EEG experiment

In Fig.2, Oddball paradigm is to randomly present two stimuli of the same sensory channel in an experiment, and the probabilities of the two stimuli are very different. The high probability stimuli are called standard stimuli, and the low probability stimuli is called deviant stimuli. In our experimental paradigm, the probability of deviant stimuli is 20%, which meets the criteria of the oddball experimental paradigm.

The experimental process of this study includes four stages as follow: pre-experimental preparation, experimental implementation, post experimental communication, data processing and analysis. Post experimental communication refers to the subjective expression of the experimenter's feelings in the experiment after completing the EEG experiment, such as the emotion at that time and whether it was disturbed. If it was disturbed, the experimental data of the disturbed time segment would be removed in the data processing and analysis stage. According to the feedback form filled by the subjects, combined with their EEG data and voice materials, the corresponding voice characteristics can be gotten by analysis.

In our EEG experiment, the device for collecting data is the product of Shanghai NuoCheng Electric Co., Ltd., and its amplifier model number is NCC Z2N-F-20-C. It adopts wireless WiFi transmission and has strong anti-interference, which has been adopted by many EEG research institutions such as Second Military Medical University. The data sampling rate of the experiment is 128Hz. Because the initial EEG data contains interference components such as EOG (Electro-Oculogram), EMG (Electro-Myogram), it needs to be preprocessed by removing stimulus artifacts, filtering and re-referencing before it can be used for analysis. The sample data of comparison before denoising and after denoising is shown as Fig.3.

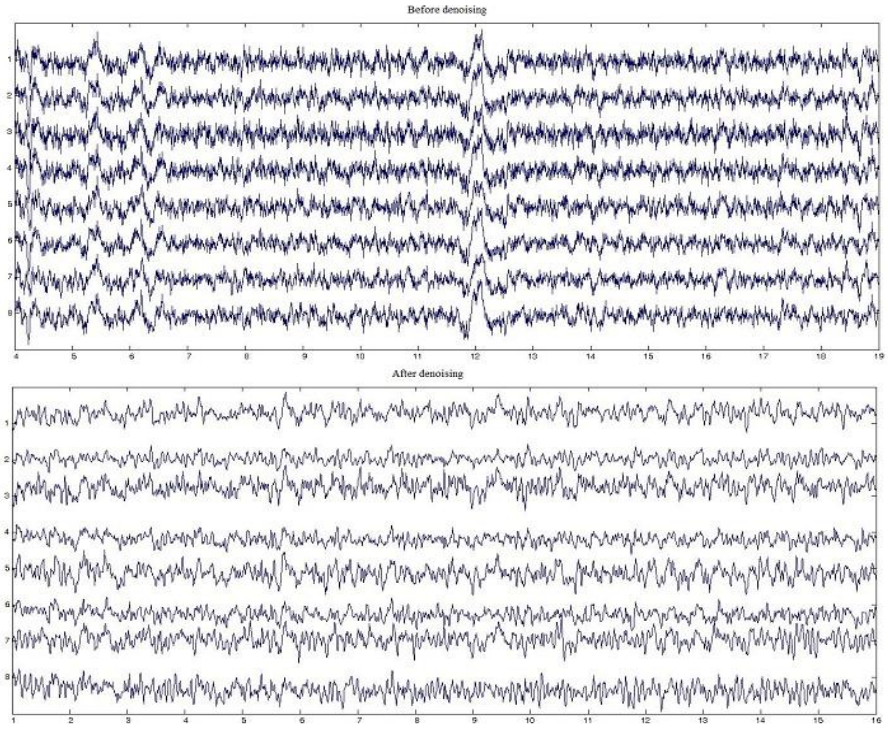

Fig.3. The comparison before and after denoising

After completing the above data preprocessing, we analyzed some electrode point data related to voice cognition by referring to previous research (Li et al. 2018; Yu et al. 2021). The main analysis data came from the signals of eight electrode positions in the frontal, parietal and occipital regions (Fp1, Fp2, T3, T4, C3, C4, O1, O2). Different experimental scenario of brain electrical activity mapping of PSD (Power Spectrum Density) was analyzed, and delta, theta, alpha and beta rhythm is shown as Fig.4.

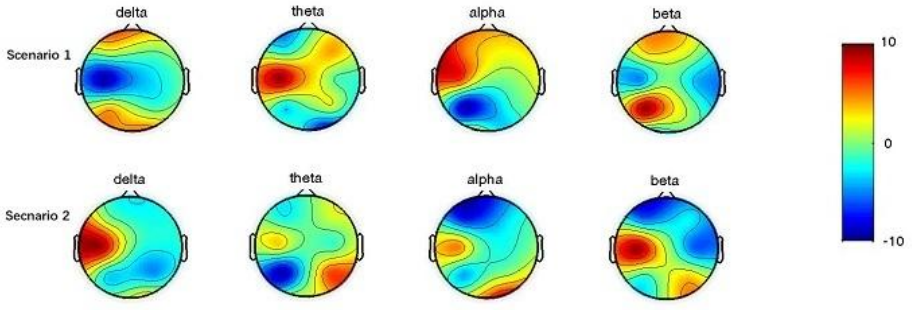

Fig.4. Different experimental scenario of brain electrical activity rhythm

Our experimental results show that beta rhythm waves are more active, especially when they are happy, angry or surprised.

After hearing the fast-speaking or high decibel voice, the power spectral density in the prefrontal, central and occipital regions was significantly more active than in the resting state. Therefore, voice features have an impact on the subjects. After many experiments and analyses, the set of vectors composed of these features can be expressed as follows.

$$I(n) = [P, SE, SZC, MFCC, FF, SF, VS, NVB] \quad (1)$$

In (1), P is the maximum, minimum and average value of the pitch. SE is the maximum, minimum and average values of short-time energy. SZC is the maximum, minimum and average value of the Short Time average Zero-Crossing rate. MFCC is cepstrum coefficient of 12th order Mel frequency. FF is the value of the first formant, SF refers to the value of the second formant. VS is the speed of voice. NVB is the number of breaks in voice. The above value of twenty-five vectors will be used to represent the PAD value.

4. Extraction of Voice Emotion Feature Parameters

In order to get high accuracy in speech feature extraction, it needs to carry out some speech processing, which including voice interception unit, pre-emphasis processing, speech framing and windowing, and end-point detection.

4.1. Voice Signal Pre-emphasis

Voice signal pre-emphasis is mainly to increase the high-frequency part of speech. As the influence of lip pronunciation, the spectrum of high-frequency part of speech will be weakened, resulting in that the main signal left in speech unit information is the spectrum of low-frequency part. In order to avoid this phenomenon, spectrum signal pre-emphasis processing is often adopted. After the speech unit signal is processed by the pre-emphasis filter, it can improve the spectrum of the high-frequency part of the speech signal and filter out the low-frequency interference from 50Hz to 60Hz, which highlights the high-frequency resolution of the speech, so that the random noise can be effectively suppressed in the calculation, which is equivalent to indirectly improving the energy of the voiceless part. The pre-emphasis filter realized by first-order high pass filter is shown as follows.

$$Z(n) = X(n) - a * X(n-1) \quad (2)$$

In (2), the voice sampling value at the nth time is X(n). a is the pre-weighting coefficient, which is between 0.9 and 1.0, usually is 0.98. The code to realize pre-emphasis in MATLAB software is as follows: y = filter ([1 -1], [1 -0.98], X).

4.2. Framing and Windowing

From the whole point of view of speech signal, its feature and parameters characterizing its feature are time-dependent and will change according to time. Previous studies have proved that the gene cycle of voiced voice, the amplitude of voiced voice signal and channel parameters can maintain short-term stability in a short period of time, that is, the spectral feature and some physical characteristic parameters are stable in the above time. Therefore, the speech signal is divided into frames, that is, the voice signal is divided into time periods in frames, and the frame length is generally 10ms ~ 30ms. The non-overlapping part between frames is called frame shift, which is generally set to 50%. After the above processing, the analysis and processing of speech stabilization process can be performed. The framing operation can be completed through the Enframe() function in MATLAB. In this function, the Len() parameter is the frame length (10~30ms) and the Inc parameter is the frame shift (1/3~1/2). In order to keep the emotional signal of speech unit stable in a short time, window function is usually used to process the signal. The commonly used window functions are rectangular window, Hamming window and Hanning window. The mathematical formulas of the three window functions are shown as follows.

$$W(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & Other \end{cases} \tag{3}$$

In (3), it represents a rectangular window, where n is the window length, which is usually represented by frame length.

$$W(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n/(N-1)], & 0 \le n \le N-1 \\ 0, & Other \end{cases} \tag{4}$$

$$W(n) = \begin{cases} 0.5[1 - \cos(2\pi n/(N-1))], & 0 \le n \le N-1 \\ 0, & Other \end{cases} \tag{5}$$

Rectangular window is usually used in the time domain of speech processing. Among, Hamming window or Hanning window is often selected in the frequency domain of speech processing. The code to realize Hamming window in MATLAB software is as follows: X = X'*Hamming(n), where n is the number of data points of each window.

4.3. End-Point Detection

End-point detection is to find the starting point and ending point of a speech. As an important method of pre-processing speech data, Voice Activity Detection (VAD) functions was used to detect presence or absence of human voice in a signal [1]. After the voice is detected by VAD, the invalid voice is removed, and the remaining voice is used for subsequent processing and analysis. In our work, short-term average energy and short-term average zero crossing rate was used to VAD.

Short time average energy. After speech signal is divided into frames, the characteristic parameter used in time domain processing is short-term average energy, which is used to distinguish voiced segment from voiced segment. The short-term average energy is described as follows. If x(n) is the time domain signal of speech waveform and $y_i(n)$ is the i-th frame signal of windowed function w(n), then $y_i(n)$ is calculated as follows.

$$y_i(n) = w(n) * x((i-1) * inc + n), \quad 1 \le n \le L, 1 \le i \le fn \tag{6}$$

In (6), Where $w(n)$ is the window function, $y_i(n)$ is the value of the frame, FN and L refers to the frame length, inc is the frame shift length.

Therefore, the short-time average energy formula of the i-th frame speech signal $y_i(n)$ is calculated as follows.

$$E(i) = \sum_{n=0}^{L-1} y_i^2(n), \quad 1 \le \mathrm{i} \le fn \tag{7}$$

Short time average zero crossing rate. The short-term average zero crossing rate refers to the number of times each frame signal passes through the zero value. It is used to roughly estimate the spectral feature of speech signals. For discrete-time speech signals, the zero-crossing phenomenon means that the algebraic symbols of two adjacent sampled values are different; For continuous speech signal, it should be judged according to how its time-domain waveform crosses the horizontal axis. The formula of short-time average zero crossing rate is calculated as follows.

$$Z(i) = \frac{1}{2} \sum_{n=0}^{L-1} \left| \mathrm{sgn}[y_i(n)] - \mathrm{sgn}[y_i(n-1)] \right|, \ 1 \le i \le fn \tag{8}$$

In (8), sgn[] is a symbolic function, which is is calculated as follows.

$$\mathrm{sgn}\left[x(n)\right] = \begin{cases} 1, \ x(n) \ge 0 \\ -1, \ x(n) < 0 \end{cases} \tag{9}$$

Mel-Frequency Cepstral Coefficients. MFCCs is a set of audio characteristic parameters. This coefficient uses the auditory principle and cepstrum solution to recognize speech. It is a parameter established by imitating the auditory feature of human ears and has high noise resistance. The conversion relationship between MFCC coefficient and linear frequency is calculated as follows.

$$f_{Mel} = 2595 * \log\left[1 + \frac{f}{700}\right] \tag{10}$$

In (10), f is the frequency. MFCC can be used as the frequency sensitivity of human hearing. Human ears show logarithmic changes in frequency perception. The MFCC parameters of order 12 can be obtained by using the function melbankm (m, N, FS) in the voicebox of MATLAB toolbox, where m parameter is the number of filters, n parameter is the speech frame length, and FS parameter is the sampling frequency. The 12th order MFCC parameter feature of the sample voice file is shown in Fig.5.

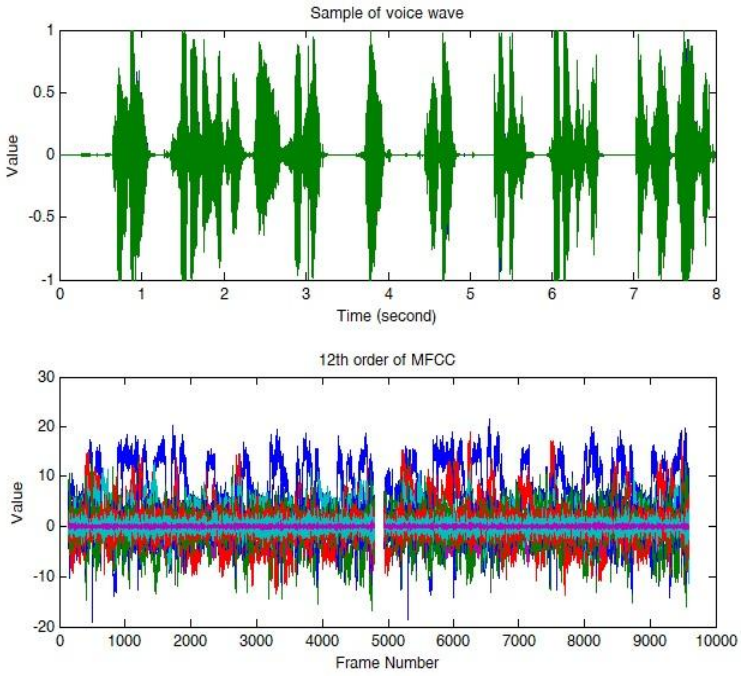

Fig.5. 12th order MFCC feature parameters of sample voice

4.4. PAD Emotional State Model

In the PAD (Pleasure-Arousal-Dominance) emotional state model, the emotional state is described through three-dimensional space. Among them, the positive and negative changes of emotional state are expressed by the score from Pleasure to Displeasure. The emotional physiological activation level and alertness changes are expressed by the scores of Arousal to Nonarousal. The control and influence of emotions on other individuals and the external environment are expressed by Dominance to Submissiveness scores [8, 25]. The calculation of PAD value in this paper is based on the PAD questionnaire scale provided by the Chinese Academy of Sciences. The scale includes three dimensions: pleasure, activation and dominance. Each dimension has four items, which is a 9-point semantic difference scale [39]. Questionnaire scale of PAD emotional state model is shown in Table 1.

Table 1. Questionnaire scale of PAD emotional state model

Question	Emotional state	Value	Emotional state
Q1	Angry	-4, -3, -2, -1, 0, 1, 2, 3, 4	Activated
Q2	Wide awake	-4, -3, -2, -1, 0, 1, 2, 3, 4	Sleepy
Q3	Controlled	-4, -3, -2, -1, 0, 1, 2, 3, 4	Controlling
Q4	Friendly	-4, -3, -2, -1, 0, 1, 2, 3, 4	Scornful
Q5	Calm	-4, -3, -2, -1, 0, 1, 2, 3, 4	Excited
Q6	Dominant	-4, -3, -2, -1, 0, 1, 2, 3, 4	Submissive
Q7	Cruel	-4, -3, -2, -1, 0, 1, 2, 3, 4	Joyful
Q8	Interested	-4, -3, -2, -1, 0, 1, 2, 3, 4	Relaxed
Q9	Guided	-4, -3, -2, -1, 0, 1, 2, 3, 4	Autonomous
Q10	Excited	-4, -3, -2, -1, 0, 1, 2, 3, 4	Enraged
Q11	Relaxed	-4, -3, -2, -1, 0, 1, 2, 3, 4	Hopeful
Q12	Influential	-4, -3, -2, -1, 0, 1, 2, 3, 4	Influenced

The scale and the normalized PAD values are calculated as follow.

$$P = \frac{Q1 - Q4 + Q7 - Q10}{16} \quad (12)$$

$$A = \frac{-Q2 + Q5 - Q8 + Q11}{16} \quad (13)$$

$$D = \frac{Q3 - Q6 + Q9 - Q12}{16} \quad (14)$$

5. Experimental Verification

5.1. Experimental Case

In our research, we take the view of “Shanghai COVID-19” as an experiment in WeChat’s speech transmission. 80 voice chat records about the event were collected from WeChat group. An example of text converted from a sample voice is as follows. “The materials received today include vegetables, meat and eggs. Great! Thank you very much, thank the volunteers!”.

5.2. Measurement of Acoustic Feature Parameters

A total of 25 vectors of acoustic feature parameters are extracted, namely: short-term energy (max, min, mean), pitch (max, min, mean), short-term average zero crossing rate

(max, min, mean), first formant, second formant, speech speed, number of interlanguages pauses and Cepstrum Coefficient of 12th order Mel frequency standard. Fig.6 shows the emotion recognition rate of the test results.

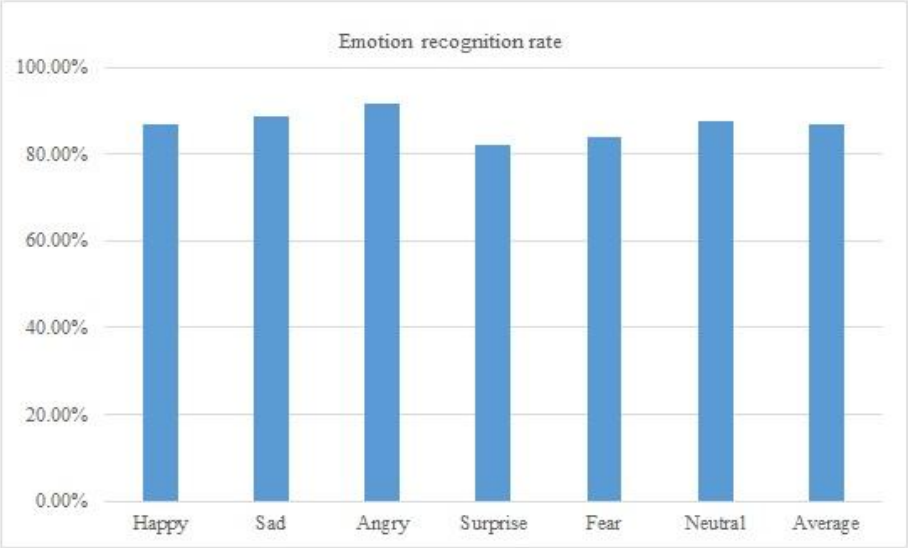

Fig.6. Emotion recognition rate

The results show that the recognition rates of six categories of emotions: happy, sad, anger, surprise, fear and neutral are 86.69%, 88.54%, 91.71%, 81.93%, 83.82% and 87.73%, and the average recognition rate is 86.74%.

In order to extract the voice social parameters of personalized features, we take 80 sentences from the historical voice of WeChat group. The above sentences come from 8 person, with an average of 10 sentences each person. From previous studies, it can be seen that social media voice includes a large number of short sentences, and the changes in emotional states exhibit dynamic characteristics [10]. Therefore, based on previous studies and the analysis of voice social habits, 8 seconds is used as the interval period for processing voice message. Each voice is sampled 8 seconds and trained with BP neural network. After reaching the accuracy requirements ($e<0.001$), the trained BP neural network is retained for the next experiment. The sample voice feature parameters and PAD values are shown in Table 2.

It can be seen from Table 2 that the PAD values are composed of three values (0.42, 0.69, 0.21). Because P value is between 0.4 and 0.53, A value is between 0.67 and 0.73, and D value is between 0.17 and 0.3, its emotion type is happiness according to the distribution range of PAD value of the six emotion types [7].

Table 2. The sample voice feature parameters and PAD values

Acoustic characteristic parameters	Pitch (Max/ Min/ Mean)	Short time energy (Max/ Min/ Mean)	Short time average zero crossing rate (Max/ Min/ Mean)	12th order MFCC	First formant	Second formant	Speed of voice	number of breaks in voice	PAD values
Duration (8S)	190.316	86.971	431	21.7303	799.14	1937.28	0.125/s	18	P=0.42
	76.857	32.528	32	13.8392					A=0.69
	122.534	76.542	224.544	18.9420					D=0.21
				12.6046					
				6.0004					
				1.0265					
				10.6584					
				12.7003					
				13.0288					
				7.7237					
				5.7858					
				1.2557					

6. Discussion and Conclusion

Although speech emotion recognition has made great progress in the past few decades, the emotion calculation of continuous and short-time conversational natural language speech signals in social media still faces many challenges. Especially the calculation of complex mixed emotion and its dynamic change process is still a difficult problem to be explored. In order to solve the above problems, this paper refers to the existing research results and proposes a framework for feature extraction and affective computing of voice message in social media environment. Specifically, it includes voice acoustic feature effective computing and text semantic feature effective computing. This framework is based on EEG experiment and PAD emotion model, which is the main innovation of this paper. In this framework, the emotion of the sample voice is calculated, and the recognition accuracy is 86.74%. The experimental results show that it provides an effective method for computing voice emotion in social media environment.

In general, the contributions of this paper are mainly in two aspects: on the one hand, the emotional cognitive feature parameters of voice message are given. The above parameters have good universality because their extraction is completed on the basis of EEG experiments. On the other hand, a new solution is proposed to transform voice affective computing into PAD emotion model parameter estimation based on acoustic parameters, which provides a reference for the study of feature extraction and affective computing of voice message in social media environment.

In the future, our study will be improved with more sample analysis, cognitive neuroscience experiment and research tools. For example: it combines fMRI (Functional Magnetic Resonance Imaging), FNIRS (Functional near-infrared spectroscopy) instruments for study.

Acknowledgment. This work is supported by the project of Social Science Foundation of Fujian Province of China (No. FJ2022BF033), and Shanghai Open University discipline innovation in 2019 year (No. XK1904).

References

1. Alghifari, M.F., Gunawan, T.S., Nordin, M.A.W., Qadri, A.A.A.: Kartiwi, M., Janin, Z. H.: On the use of voice activity detection in speech emotion recognition. Bulletin of Electrical Engineering and Informatics, Vol. 8, No. (4), 1324-1332. (2019)
2. Bakhtiyari, K., Husain, H.: Fuzzy model of dominance emotions in affective computing. Neural Computing & Applications, Vol. 25, No. 6, 1467-1477. (2014)
3. Bänziger T., Patel, S., Scherer, K.: The Role of Perceived Voice and Speech Feature in Vocal Emotion Communication. Journal of Nonverbal Behavior, Vol. 38, No. 1, 31-52. (2014)
4. Bhagavathsingh, B., Srinivasan, K., Natrajan, M.: Real time speech based integrated development environment for C program. *Circuits and Systems*, 7(3), 69-82. (2016).
5. Challawar, R., Menon. A., Kar, M., Mahapatra, S.C.: Effect of acute bout of moderate exercise on P300 component of event-related potential in young women during different phases of menstrual cycle: A pilot study. Indian Journal of Physiology and Pharmacology, Vol. 64, No. 4, 272-278. （2021）.
6. Chen, W.L., Sun, X.: Mandarin Speech Emotion Recognition Based on MFCCG-PCA. Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 51, No. 2, 269-274. (2015)
7. Chen, Y.L., Cheng, Y.F., Chen, X.Q., Wang, H.X., Li, Chao.: Speech emotion estimation in PAD 3D emotion space. Journal of Harbin Institute of Technology, Vol. 50, No. 11, 160-166. (2018).
8. Choe, Kyung-II.: An Emotion-Space Model of Multimodal Emotion Recognition. Advanced Science Letters, Vol. 24, No. 1, 699-702. (2018)
9. Dai, W.H.: Neuromangement: disciplinary development and research paradigm. Journal of Beijing Technology and Business University (Social Sciences), Vol. 32, No. 4, 1-10. (2017)
10. Dai, W.H., Han, D.M., Dai, Y.H., Xu, D.R.: Emotion recognition and affective computing on vocal social media. Information & Management, Vol. 52, No. 7, 777-788. (2015)
11. Dehaene, S.: The Error-Related Negativity, Self-Monitoring, and Consciousness. Perspectives on Psychological Science, Vol. 13, No. 2, 161-165. (2018)
12. Gerstner, W., Sprekeler, H., Deco, G.: Theory and Simulation in Neuroscience. Science, Vol. 338, No. 6103, 60-65. (2012)
13. Gong, S.P., Dai, Y.H., Ji, J., Wang, J.Z., Sun, H.: Emotion Analysis of Telephone Complaints from Customer Based on Affective Computing. Computational Intelligence and Neuroscience, 2015, 1-9. (2015).
14. Han, W.J., Li, H.F., Han, J.Q.: Speech emotion recognition with combined short- and long-term features. Journal of Tsinghua University (Science and Technology), Vol. S1, 708-714. (2008)
15. Hu, J.: Building a risk prevention system to protect the mainstream ideology in the era of micro communication. Journal of Dalian University of Technology (Social Sciences), Vol. 42, No. 3, 1-6. (2021)
16. Huang, S., Zhou, X., Xue, K., Wan, X.Q., Yang, Z.Y., Xu, D., Ivanović, M., Yu, X.: Neural Cognition and Affective Computing on Cyber Language. Computational Intelligence & Neuroscience, Vol. 2015, 1-10.

17. Jalili, F., Barani, M.J.: Speech Recognition Using Combined Fuzzy and Ant Colony algorithm. International Journal of Electrical & Computer Engineering, Vol. 6, No. 5, 2205-2210. (2016)
18. Jiang, N., Liu, T.: An Improved Speech Segmentation and Clustering Algorithm Based on SOM and K -Means. Mathematical Problems in Engineering, Vol. 1, 1-19. (2016)
19. Kstle, J.L., Anvari, B., Krol, J., Wurdemann, H.A.: Correlation between Situational Awareness and EEG signals. Neurocomputing, Vol. 432, No. 1, 70-79. (2021)
20. Kumar, J., kumar, J.: Affective Modelling of Users in HCI Using EEG. Procedia Computer Science, Vol. 84, No. 5, 107-114. (2016)
21. Kurbalija, V., Ivanovic, M., Radovanovic, M., Geler, Z., Dai, W.H., Zhao, W.D.: Emotion Perception and Recognition: An Exploration of Cultural Differences and Similarities. Cognitive Systems Research, Vol. 52, 103-116. (2018)
22. Lai, Y., Tian, Y., Yao, D.: MMN evidence for asymmetry in detection of IOI shortening and lengthening at behavioral indifference tempo. Brain Research, Vol. 1367, No. 7, 170-180. (2011).
23. Lee, K.J., Park, C.A., Lee, Y.B., Kim, H.K., Kang, C.K.: EEG signals during mouth breathing in a working memory task. International Journal of Neuroscience, Vol. 130, No. 5, 1-10. (2019)
24. Li, H.W., Li, H.F., Ma, L., Bo, H.J., Xu, R.F.: Brain's cognitive law of changes in musical attributes while listening to music-An EEG study. Journal of Fudan University (Natural Science), Vol. 57, No. 3, 385-392, (2018)
25. Li, J., Huang, W., Guo, S.L., Sun, Y.: Research on the Sentiment Intensity Measurement Model of Internet Word-of-Mouth Public Opinion Based on the PAD Model. Journal of the China Society for Scientific and Technical Information, Vol. 38, No. 3, 277-285. (2019)
26. Lin, J.C., Wu, C.H., Wei, W.L.: Error Weighted Semi-Coupled Hidden Markov Model for Audio-Visual Emotion Recognition. IEEE Transactions on Multimedia, Vol. 14, No. 1, 142-156. (2012)
27. Liu, Z.T., Xu, J.P., Wu, M., Cao, W.H., Chen, L.F., Ding, X.W., Hao, M., Xie, Q.: Review of emotional feature extraction and dimension reduction method for speech emotion recognition, Chinese Journal of Computers. Vol. 41, No. 12, 2833-2851. (2018)
28. Ma, Q.G., Feng, Y.D., Xu, Q., Bian, J., Tang, H.X.: Brain potentials associated with the outcome processing in framing effects. Neuroscience letters, Vol. 528, No. 2, 110-113. (2012)
29. Mcaleavy, T., Rhisiart, M.: Harnessing the power of metaphor: uncovering a hidden language of interoperability within the natural speech of emergency managers. International Journal of Emergency Management, Vol. 15, No. 1, 1-25. (2019)
30. Motamed, S., Setayeshi, S., Rabiee, A.: Speech emotion recognition based on brain and mind emotional learning model. Journal of Integrative Neuroscience, Vol. 17, No.12, 1-15. (2018)
31. René, R., Mohr, P.N.C., Kenning, P.H., Davis, F.D., Heekeren, H.R.: Trusting Humans and Avatars: A Brain Imaging Study Based on Evolution Theory. Journal of Management Information Systems, Vol. 30, No. 4, 83-113. (2014)
32. Shankar, S., Tewari, V.: Understanding the Emotional Intelligence Discourse on social media: Insights from the Analysis of Twitter. Journal of Intelligence, Vol. 9, No. 4, 1-17. (2021)
33. Sharma, G.D., Paul, J., Srivastava, M., Yadav, A., Mendy, J., Sarker, T., Bansal, S.: Neuroentrepreneurship: an integrative review and research agenda. Entrepreneurship and Regional Development, Vol. 33, 863-893. (2021)
34. Song, X.Y., Zeng, Y., Tong, L., Shu, J., Li, H.M., Yan, B.: Neural Mechanism for Dynamic Distractor Processing during Video Target Detection: Insights from Time-varying Networks in the Cerebral Cortex. Brain Research, Vol. 1765, 1-9. (2021)

35. Wang, H.L., Feng, T.Y., Suo, T., Liang, J., Meng, X.X., Li, H: The process of counterfactual thinking after decision-making: Evidence from an ERP study. Chinese Science Bulletin, Vol. 55, No. 12, 1113-1121. (2010)
36. Yu, G.M., Wang, W.X., Feng, F., Xiu, L.C. Evaluation of the communication effect of synthetic speech news: The EEG evidence of the effect of speech speed. Chinese Journal of Journalism & Communication, Vol. 43, No. 2, 6-26. (2021)
37. Yun, S., Yoo, C.D.: Loss-Scaled Large-Margin Gaussian Mixture Models for Speech Emotion Classification. IEEE Transactions on Audio. Speech, and Language Processing, Vol. 20, No. 2, 585-598. (2012).
38. Zhang, G.: Quality evaluation of English pronunciation based on artificial emotion recognition and gaussian mixture model. Journal of Intelligent and Fuzzy Systems, Vol. 40, No. 2, 1-11. (2020)
39. Zhang, X.Y., Zhang, T., Sun, Y., Zhang, W., Chang, J.: Emotional Speech Database Optimization and Quantitative Annotation Based on PAD emotional model. Journal of Taiyuan University of Technology, Vol. 48, No. 3, 469-474. (2017)
40. Zhao, L., Qian, X.M, Zhou, C.R., Wu, Z.Y.: A Study on Emotional Recognition in Speech Signal. Journal of Software, Vol. 7, 1050-1054. (2001)
41. Zheng, J.: Cheng, J., Wang, C., Lin, X., Fu, G., Sai, L.: The effect of mental countermeasures on a novel brain-based feedback concealed information test. Human brain mapping, Vol. 43, No. 9, 2771-2781. (2022)

Peng Jiang is an associate professor at Jingan Branch Campus, Shanghai Open University, China. His current research interests include Educational technology and Management Information System. Contact him at jzhpmail@163.com.

Cui Guo is a lecturer at Shanghai Lifelong Education School Credit Bank Management Center, China. Her current research interests include Distance Learning and Computer Application System. Contact her at guoc@sou.edu.cn.

Yonghui Dai is the corresponding author of this paper. He is currently an associate professor at the Management School, Shanghai University of International Business and Economics, China. He received his Ph.D. in Management Science and Engineering from Shanghai University of Finance and Economics, China in 2016. His current research interests include Affective Computing, Intelligence Service and Big Data Analytics. His works have appeared in international journals more than forty papers. Contact him at daiyonghui@suibe.edu.cn.

Received: May 09, 2023; Accepted:September 15, 2023.