Emotion Recognition Model Based on Multimodal Decision Fusion

Chunjun Zheng¹, ²*, Chunli Wang¹, Ning Jia²

¹ College of Information Science and Technology, Dalian Maritime University, Dalian, Liaoning, China
² School of Software, Dalian Neusoft University of Information, Dalian, Liaoning, China

*Corresponding author’s e-mail: zhengchunjun@neusoft.edu.cn

Abstract. In the process of human social activities and daily communication, speech, text and facial expressions are considered as the main channels to convey human emotions. In this paper, a fusion method of multi-modal emotion recognition based on speech, text and motion is proposed. In the speech emotion recognition (SER), a depth wave field extrapolation - improved wave physics model (DWE-WPM) is designed. In order to simulate the information mining process of LSTM, a user-defined feature extraction scheme is used to reconstruct the wave and inject it into DWE-WPM. In the text emotion recognition (TER), the transformer model with multi attention mechanism is used to recognize the text emotion combined. In the motion emotion recognition (MER), the sequential features of facial expression and hand action are extracted in groups. Combined with the bidirectional three-layer LSTM model with attention mechanism, a joint model of four channels is designed. Experimental results show that the proposed method has high recognition accuracy in multi-modal, and the accuracy is improved by 9% in the interactive emotional dynamic motion capture (IEMOCAP) corpus.

1. Introduction
Emotion refers to the individual's experience, feeling and behaviour response to the external objective things. It is a kind of psychological activity mediated by the individual's desire and need. Emotion recognition refers to the process that the computer analyses and processes the signal collected from the sensor to obtain the emotion state of the object to be identified. In today's increasingly close human-computer interaction, if the machine can recognize human emotions, and even restore human emotions, it can better work with human division of labour and cooperation.

Human emotion can be expressed through facial expression [1], action posture [2], speech [3], physiological signal [4] and other modalities. Facial expression and action posture are visual, they are called expression modality and action modality. Speech information can be obtained from hearing, also known as speech modality. Emotion judgment from multi-modal is consistent with the emotion experience of daily life, and they can be obtained by non-invasive sensors. This method is simple, convenient and low cost. It is an important subject to judge emotion based on one or more of these modalities. Deep learning method can learn the nonlinear expression of effective information from the information of different modes. At present, this method has been widely used in a variety of emotional modalities.
Speech mode is an important mode that can be used for emotion recognition. There are both explicit linguistic information and nonverbal acoustic information in the signal, which can be used to infer emotional state [5].

Text sentiment analysis is a process of analyzing, processing and extracting subjective text with emotion by using natural language processing and text mining technology [6].

Facial expression is one of the main clues for people to understand each other's emotions in daily communication. The biggest advantage of facial expression recognition is that it is universal and independent of cultural background [7].

We use the popular interactive emotional dynamic motion capture (IEMOCAP) [8] database, which records the emotional data of ten actors, whose behaviours can be expressed through a variety of different modalities.

We propose a multi-modal emotion recognition model with differentiated characteristics, which integrates the information of speech, text and expression. Different feature extraction schemes and matching model structures are designed for speech emotion recognition (SER), text emotion recognition (TER) and motion emotion recognition (MER). Local feature extraction and attention mechanism are designed to enhance the effective information.

2. Proposed method

Building a multimodal emotion prediction model generally includes the extraction of multi-modal emotion features, the design and selection of emotion recognition model, information fusion scheme and so on. The key research direction of our work is to determine the effective mode combination scheme and realize the effective fusion. This paper proposes an emotion recognition method for three modalities of speech, text and motion. An emotion recognition model is designed for each modality, and then the results of all the modalities are fused together with user-defined rules to achieve effective multi-modal emotion information fusion and decision. The overall structure is shown in Fig 1.

\[
\text{result} = \arg \max (\lambda_i \cdot \text{Contact}_i)
\]

\[
= \arg \max \left( \sum \lambda_i \cdot \text{Contact}_i \right)
\]

FIG. 1. Overall design scheme of the model.

2.1. DWE-WPM for Speech Emotion Recognition (SER)

Acoustic signals often have a large number of features, which brings great difficulties for emotion recognition. In this section, a new speech emotion recognition method is proposed, which is based on
depth wave field extrapolation improved wave physics model (DWE-WPM), and can solve the problem of accuracy loss and feature explosion in the process of feature extraction.

Firstly, the fixed step depth recursive acoustic field continuation method is used to obtain the region containing effective emotional information. From the wavefield of a depth, a subset of waves with a fixed step depth is found. The goal of acoustic recursion is to realize fast extension of acoustic field in target area.

In the second step, DWE-WPM is designed to simulate the speech propagation by using the dynamics of wave motion, so as to obtain the expression of local emotional features. The model consists of three stages: forward expansion stage, observation stage and material physical setting stage.

When the waveform is input from the left side of the domain, it is emitted to propagate through the forward expansion region, and the wave velocity distribution and medium parameters can be trained. The detection point at the specified position in the two-dimensional region is used for calculation, and the output information is recorded when the sound wave is transmitted to a certain detection point. The output values of all observation points are spliced into a non-negative vector, which is the feature expression of the current speech in the physical system.

The material and velocity of transmitting wave have a great effect on this model, so the DWE-WPM model is used for back propagation and continuous fine tuning. By correcting the first-order and the second-order matrix, the influence of bias is reduced, and then the random gradient descent is carried out to achieve the final convergence of the model.

The features obtained by DWE-WPM concentrate on the local features of emotion expression, and ignore the global temporal information in speech. We need to integrate the global feature expression to fully reflect the characteristics of speech. Traditional Low-Level Descriptor (LLD) features include prosodic, fundamental frequency, formant and so on. But the recognition accuracy of these LLD is not higher enough. In this paper, two kinds of advanced LLDs are defined, which are sent into SER and TER as global features to achieve higher accuracy of emotion recognition.

2.2. Transformer for Text Emotion Recognition (TER)

In the research of text emotion recognition (TER), transformer is a popular recognition model. Transformer learns the dependencies between words based on self-attention without any recursion or convolution. It has rich forms and fast computing speed, and has been applied to many natural language processing tasks. On the basis of word feature extraction and speech energy feature extraction, we use transformer and attention reduction mechanism to realize text emotion recognition by fusing energy features. The model mainly includes the following aspects.

2.2.1. Embedded Feature Extraction. In the sentence, the vocabulary is constructed according to the training set, the special characters and punctuation are deleted, and each word is embedded into a 300 dimensional vector by using glove. If a word from a validation or test set is not in the default vocabulary, it can be replaced with an unknown tag. Each sentence passes through a single layer LSTM. The size of each sample is \([T, C]\), where \(T\) is the number of words in the sentence.

2.2.2. Projection. The object of projection is the linguistic feature \(x\) and the energy acoustic feature \(y\) with compensation. First, an attention reduction mechanism is used for each modality, and then the weighted sum of elements is used to fuse the two modalities vectors. The attention reduction mechanism contains a soft attention module, which calculates the weighted sum of the two groups of features to the attention weight. It is shown in Eq 1.

\[
\alpha_i = \text{soft max}(v_i^T(W_jx)) \\
\bar{x} = \sum_{i=0}^{T} \alpha_i x_i
\]  \hspace{1cm} (1)

After the reduce mechanism, the input size becomes \([1, C]\). Then perform the operation of Eq 2:

\[
y \sim p = W_\rho(LayerNorm(\bar{x} + \bar{y}))
\]  \hspace{1cm} (2)
Where \( p \) is the probability distribution of each class, and LayerNorm represents standardization for layers.

2.2.3. Modulated Attention Transformer. After giving the language feature \( y \), the acoustic features are coded by module fusion. Two modalities are fused in the acoustic transformer to obtain the output. The scheme can be realized by multi attention and layer normalization. In order to adjust the self-attention of speech through language output, the \( K \) and \( V \) of self-attention are converted from \( y \) to \( \tilde{x} \). The Eq 3 describes the process of creating a new attention layer in Transformer.

\[
y = \text{LayerNorm}(y + \text{Multi\_head\_Attention}(y, x, x))
\] (3)

In the above operation, the characteristic dimensions of \( y \) and \( \tilde{x} \) must be equal. It can be adjusted according to the transformation matrix (eight headers) or the size of LSTM.

2.3. Motion Emotion Recognition (MER) based on Facial Expression and Hand Movement

Motion capture (MOCAP) is a kind of technology which can record the change rule of the wearer's movement in real time through the wearing of sensor equipment, so as to distinguish the wearer's movement track.

In IEMOCAP corpus, we use these two groups of data (facial expression and hand) to form the basis of the multimodal emotion recognition model.

2.3.1. Facial Expression Feature Extraction. When designing the emotion recognition model of facial expression, the second design is carried out on the basis of emotion feature extraction in frame. The muscle deformation around the facial features is the largest, which can be considered as a special expression of emotion. Therefore, the emotional features around the facial features are the most effective features, and other feature points can be ignored. It is necessary to filter out the irrelevant points and combine new features to express more emotional information.

When we use IEMOCAP corpus to detect facial expression, we mainly detect 55 key points of face. These points mainly include the information of eyes, eyebrows, nose, mouth and facial contour. Through the detection of key points, the coordinate information of each point is obtained, which is used to extract emotional features. After extracting these points, 165 facial features are obtained.

2.3.2. Hand Motion Feature Extraction. The hand data consists of a total of 20 dimensional data, which is composed of two-dimensional index and three-dimensional data of six key points. The three-dimensional data correspond to the rotation data of X, Y and Z axes of the key points respectively.

For a single set of points, the subspace with the largest variance in the sliding window is found. We set the size of the sliding window as the window size of the frame. The window moves to the right from the beginning of the sample, and calculates the sum of variance of continuous data in the sliding window. When the sum of variance is the largest, record the starting point of the current window, take out the sum of the absolute difference value of the data inside the window and the variance of the continuous data at this starting point, and use this data as the input of the model.

Based on the realization of SER, TER, MER model, the final emotion recognition result can be obtained by fusing the decision level.

3. Experiments and Results

3.1. Datasets

The IEMOCAP corpus used in this paper is recorded by the University of Southern California, including data and tags of ten emotions (anger, happiness, emotion, sad, friction, fear, surprise, other and neural). The dataset, which uses motion capture and audio/video recording data, is collected from five sessions on ten topics.
3.2. Experimental Setup

This experiment was used to verify the effectiveness of the multi-modal model proposed in Session II. We designed the following models to compare the accuracy. In this experiment, the experimental results of SER in Experiment 1 were used as the baseline, and table 1 showed the accuracy of different recognition models after experimental verification.

| Model                      | WA  | UA  |
|----------------------------|-----|-----|
| Baseline: SER              | 63% | 62% |
| Model 1: Ours (SER+MER)    | 65.4% | 64.8% |
| Model 2: Ours (MER+TER)    | 66.8% | 66.4% |
| Model 3: Ours (SER+TER)    | 72.2% | 72.4% |
| Model 4: Voted (SER+TER+MER) | 73.8 | 73.2 |
| Model 5: Ours (SER+TER+MER) | 74.5% | 75.1% |

Comparing the fusion effect of different modalities in this experiment, we could see that model 1-3 was the fusion of two modes respectively, and their recognition accuracies were significantly improved than that of a single modal. Model 4 and model 5 used voted and custom decision scheme to fuse SER, TER, MER modes respectively. The results showed that the recognition accuracy of the fused model was improved, and model 5 had the best recognition accuracy. It could be seen that different decision-making schemes also had an impact on the recognition results, and the accuracy of our proposed multimodal model was 15% higher than that of a single mode.

Based on IEMOCAP corpus, we compared the accuracies of multi-modal models, as shown in Table 6. It could be found that the multi-modal recognition scheme proposed in this paper had the best recognition accuracy, which was higher than the other models in the Table 2, and the average accuracy was improved to 9%. It proved the effectiveness of the proposed single modal and multi-modal fusion scheme.

| Model          | Accuracy |
|----------------|----------|
| Samarth [9]    | 71.04%   |
| Soujanya [10]  | 71.59%   |
| Gaurav [11]    | 70.1%    |
| Ren [12]       | 60.59%   |
| Ours           | 75.1%    |

4. Conclusions

In this paper, a multi-modal emotion recognition model based on audio, text and action is proposed, which includes three models, namely SER, TER and MER. For each model, the feature extraction scheme and matching model structure are designed respectively. In speech modality, the depth wavefield extrapolation - improved wave physics model (DWE-WPM) and user-defined LLD features are designed to highlight the local and global information of speech mode. The task of text modality is to extract text features and classify emotions from the content expressed by the speaker. The effective sequence features of hand and facial expressions are extracted from motion modality. Four groups of three-layer bidirectional LSTM models are designed by using multi-channel model. Experimental results on the interactive emotional dynamic motion capture (IEMOCAP) emotional corpus show that the proposed method can improve the average accuracy by 9%.

In the future research process, we will further improve the SER model, add the feature extraction scheme and design the video modality, and seek a more general multi-modal network structure. We hope to select some modalities according to the needs to achieve efficient multi-modal emotion recognition.
Acknowledgments
This paper was funded by the Liaoning Province Key Laboratory for the Application Research of Big Data, the Dalian Science and Technology Star Project, grant number 2019RQ120, and the Intercollegiate cooperation projects of Liaoning Provincial Department of Education, grant number 86896244.

References
[1] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency, “Context-dependent sentiment analysis in user-generated videos,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 873–883.
[2] Ding I J , Hsieh M C . A hand gesture action-based emotion recognition system by 3D image sensor information derived from Leap Motion sensors for the specific group with restlessness emotion problems[J]. Microsystem Technologies, 2020(3).
[3] V. Ramanarayanan, R. Pugh, Y. Qian, and D. Suendermann-Oeft: Automatic Turn-Level Language Identification for Code-Switched Spanish-English Dialog. In Proc. of IWSDS 2018, International Workshop on Spoken Dialog Systems, Singapore, Singapore, May 2018.
[4] Sarma P , Barma S . Review on Stimuli Presentation For Affect Analysis Based on EEG[J]. IEEE Access, 2020, PP(99):1-1.
[5] Tiwari U , Soni M , Chakraborty R , et al. Multi-Conditioning and Data Augmentation Using Generative Noise Model for Speech Emotion Recognition in Noisy Conditions[C]/ ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
[6] Bouazizi M , Ohtsuki T . Multi-class sentiment analysis on twitter: Classification performance and challenges[J]. Big Data Mining and Analytics, 2019.
[7] Pranav E , Kamal S , Chandran C S , et al. Facial Emotion Recognition Using Deep Convolutional Neural Network[C]/ 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). 2020.
[8] Busso C , Bulut M , Lee C C , et al. IEMOCP: interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42(4):335-359.
[9] Samarth Tripathi, Sarthak Tripathi, Homayoon Beigi, Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning, 2018.
[10] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain, “Multimodal sentiment analysis: Addressing key issues and setting up the baselines,” IEEE Intelligent Systems, vol. 33, no. 6, pp. 17–25, 2018.
[11] Gaurav Sahu, Multimodal Speech Emotion Recognition and Ambiguity Resolution, 2019.
[12] Ren M , Nie W , Liu A , et al. Multi-modal Correlated Network for emotion recognition in speech[J]. Visual Informatics, 2019, 3(3).