# Feature Fusion Algorithm for Multimodal Emotion Recognition from Speech and Facial Expression Signal

Han Zhiyan[1,a] and Wang Jian[1]

[1]*Bohai University, China*

**Abstract.** In order to overcome the limitation of single mode emotion recognition. This paper describes a novel multimodal emotion recognition algorithm, and takes speech signal and facial expression signal as the research subjects. First, fuse the speech signal feature and facial expression signal feature, get sample sets by putting back sampling, and then get classifiers by BP neural network (BPNN). Second, measure the difference between two classifiers by double error difference selection strategy. Finally, get the final recognition result by the majority voting rule. Experiments show the method improves the accuracy of emotion recognition by giving full play to the advantages of decision level fusion and feature level fusion, and makes the whole fusion process close to human emotion recognition more, with a recognition rate 90.4%.

## 1 Introduction

In recent years, emotion recognition has become a hot spot in human-computer interaction. There are two broad categories: single mode emotion recognition and multimodal emotion recognition. The single mode emotion recognition obtains emotion state from single information channel, as from speech signal, facial expression signal or physiological signal.

For speech emotion recognition, MIT media lab constructed an emotion editor in 1990[1]. Yan [2] used non-uniform subband filter to dig useful information for speech emotion. It increased the identification of all kinds of emotions, and improved the performance of emotion recognition. Mao [3] used parametric filter and fractal dimension to recognize emotion, and got better performance. Zou [4] proposed a emotion recognition method based on improved fuzzy vector quantization, and effectively improved recognition rate. Attabi [5] studied the effectiveness of anchor models applied to the multiclass problem of emotion recognition. Zheng [6] proposed a novel speech emotion recognition method based on least square regression model. This method achived better recognition performance. Mao [7] proposed to learn affect-salient-features using convolutional neural networks. This method led to stable and robust recognition performance in complex scenes.

Ekman [8] developed facial action coding system (FACS) to detect subtle changes in facial expression. Essa [9] proposed dynamic expression description method FACS+ based on video. It solved the problem that there was no time description information in FACS. Rahulamathavan [10] presented a system that addresses the challenge of performing facial expression recognition when the test image is in the encrypted domain. Zheng [11] proposed a novel group sparse reduced-rank regression (GSRRR) model to describe the relationship between the multi-view facial feature vectors and the corresponding expression class label vectors.

For physiological signal emotion recognition, petrantonakis [12] used higher order crossing analysis to extract feature from Electroencephalogram (EEG). Liu [13] used Ant Colony Optimization (ACO) to recognize the emotion from respiration signals. Zacharatos [14] studied the importance of body posture and movement for emotion recognition.

Although single mode emotion recognition has some achievements in scientific research, there are many limitations. Due to human beings express emotion information in many ways. So it has the complexity of expression and the relative property of culture [15]. When the signal is disturbed by noise, the multimodal emotion recognition method can generate complementary effect in some ways, and make up for the shortage to single mode emotion recognition. So the research for multimodal emotion recognition is very necessary. Kim [16] fused electromyogram, electrocardiogram, skin conductivity and respiration changes to recognize the emotion. Zhao [17] fused speech signal and electrocardio signal to recognize the emotion, and obtained a higher fusion recognition rate. But the above methods are all fused physiological signal, the measure of physiological signal must contact body. But the acquisition of signal has a certain difficulty. So the fusion of speech signal and facial expression signal has been widely studied. Busso [18] analysed the complementarity of speech emotion recognition and facial expression recognition. Hoch [19] fused speech and facial expression signal under vehicle environment, and realized the recognition for the three emotions, positive, negative and neural. Sayedelah [20] studied audio-visual feature decision level fusion for spontaneous emotion estimation in speech conversation. In a certain sense, the fusion of different channel information is bottleneck problem for multimodel

[a] Corresponding author: hanzyme@126.com

emotion recognition, and it is directly related to the accuracy of emotion recognition.

So, this paper takes speech signal and facial expression signal as the research subjects, takes the extracting of features, the fusion algorithm of features and recognition algorithm as the research contents, and finally accomplishes recognition for five kinds of human emotion (joy, anger, surprise, sadness, fear).

## 2 System structure frame

The system structure frame is shown in Figure 1. Firstly, emotion signal undergoes a series of preprocessing course. Secondly, we extract speech emotion feature and facial expression feature. Finally, make fusion recognition.

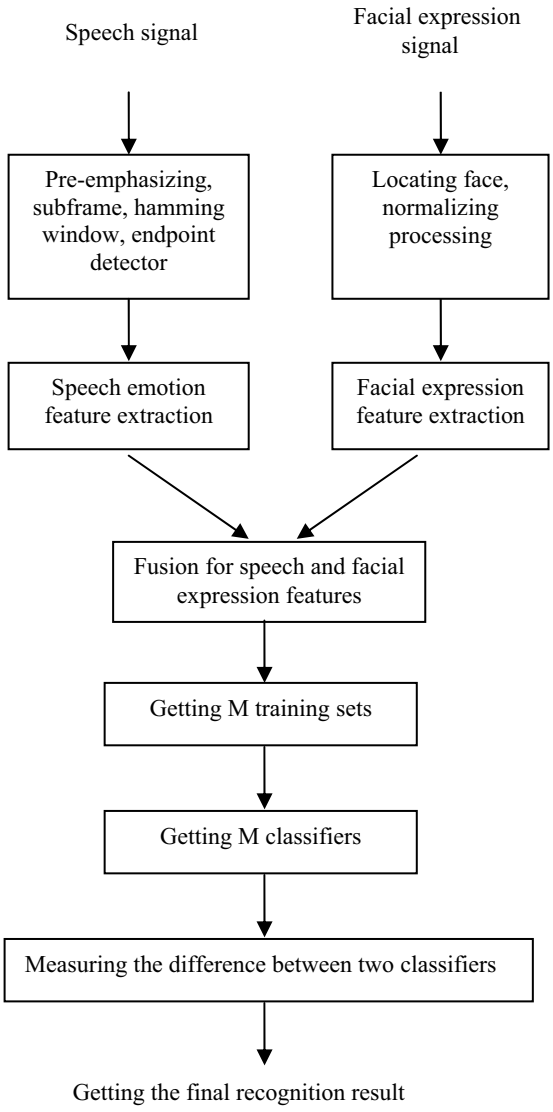


**Figure 1.** The system structure frame.

## 3 Feature extraction

### *3.1 Speech emotion feature extraction*

In the past, the effective extraction of the emotion feature was mainly based on prosodic features. However, in recent years, through deep research, the combination of acoustic features and prosodic features could be more accurate to recognize emotion. Tao [21] found that acoustic feature had a good effect on the distinction between activate dimension relatively close to the emotion, and confirmed the correlation between acoustic feature and valence dimension.

In order to make full use of the emotion information contained in speech signal, we select the ratio of sentence pronunciation duration and corresponding calm statement duration, the average value of pitch frequency, the maximum pitch frequency, the difference of average value of pitch frequency and corresponding calm statement average value of pitch frequency, the difference of maximum pitch frequency and corresponding calm statement maximum pitch frequency, amplitude average energy, dynamic range of amplitude energy, the difference of amplitude average energy and corresponding calm statement amplitude average energy, the difference of dynamic range of amplitude energy and corresponding calm statement value, the average value of the first resonance peak frequency, the average value of the second resonance peak frequency, the average value of the third resonance peak frequency, the mean of harmonic noise ratio, the maximum harmonic noise ratio, the minimum harmonic noise ratio, the variance of harmonic noise ratio, as the emotion recognition features.

### *3.2 Facial expression feature extraction*

Depending on the nature of the image, the extraction of facial expression feature is divided into static image feature extraction and sequence image feature extraction. The deformation features of facial expression are extracted from static images, and the motion features of facial expression are extracted from sequence image. In this paper, we take the static image as the research object, and use gabor wavelet transform to extract facial expression feature. The specific process is as follows:

(1) Put the preprocessed image divide into $25 \times 25$ pixels. Every image has 4 rows 3 columns.

(2) Make convolution with gabor wavelet and grid image. The formula is as follows:

$$r(x,y) = \iint I(\varepsilon,\eta) g(x-\varepsilon, y-\eta) d\varepsilon d\eta \qquad (1)$$

where $I(\varepsilon,\eta)$ is the pixel value of pixel coordinate $(\varepsilon,\eta)$, $g(x,y) = \frac{k^2}{\sigma^2}\exp(-\frac{k^2(x^2+y^2)}{2\sigma^2})(\exp(ik\cdot\begin{pmatrix}x\\y\end{pmatrix}) - \exp(-\frac{\sigma^2}{2}))$, $k = \begin{pmatrix}k_v\cos\varphi_u\\k_v\sin\varphi_u\end{pmatrix}$, $k_v = 2^{-\frac{v+2}{2}}$, $\varphi_u = u\frac{\pi}{k}$, $\sigma$ is a constant related to the frequency bandwidth of the wavelet, and the value is $\sqrt{2}\pi$. The wavelength of Gabor filter is determined by $v$, and the value is 0, 1, 2. $u$ is the direction of the Gabor kernel function, and the value is 1, 2, 3, 4, 5, 6. $k$ indicates the total number of directions, and the value is 6.

(3) Get the mean and variance of $\|r(x,y)\|$ as the facial expression feature.

(4) Make dimensionality reduction using Principal component analysis (PCA).

## 4 Algorithm description

Specific implementation steps are as follows:

Step 1: collect speech signal and facial expression signal through noise stimulation and watching television. For speech signal, undergoes a series of preprocessing course including pre-emphasizing, subframe, hamming window, and endpoint detector. For facial expression signal, firstly, locate face using skin color model, and then normalize processing for image geometry and optical properties.

Step 2: extract speech emotion feature according to section 3.1 and 3.2.

Step 3: combine speech emotion feature and facial expression feature in order, and get multimodal feature vector $u_1,u_2,\cdots,u_r,\cdots,u_W, r=1,2,\cdots,W$ . $W$ is sample number of original training sample set.

Step 4: Training sample set $S_1$ is obtained by putting back sampling N times for multimodal feature vector set, and then continue to obtain training sample set $S_2,\cdots,S_M$ .

Step 5: Train sample set using BP neural network, and then the classifier is obtained on each training sample set.

Step 6: measure the difference between two classifiers by double error difference selection strategy, and then pick out the classifier that more than average difference as recognition classifier. The difference formula $Div(i,j)$ of classifier $H_i$ and $H_j (i \neq j)$ is as follows:

$$Div(i,j)=\frac{num^{00}}{num^{00}+num^{01}+num^{10}+num^{11}} \quad (2)$$

where $num^{ab}$ represents the sample number of correct/error for two classifiers. $a=1$ and $a=0$ respectively represent the correct and error of classifier $H_i$ . $b=1$ and $b=0$ respectively represent the correct and error of classifier $H_j$ .

Step 7: get the final recognition result by the majority voting rule.

## 5 Simulation experiments and results analysis

We did recognition experiment by using the method explained above. In our experiment, five discrete emotional states (joy, anger, surprise, sadness, fear) are classified throughout the work. 100 data per emotion have been used for the training, while a disjunctive set of 100 data per emotion were used testing. The data were recorded over a period of half a year to avoid anticipation effects of the actors.

Table 1 shows the results of the emotion evaluation only using speech emotion feature. Table 2 shows the results only using facial expression feature. Table 3 shows the results of the emotion evaluation by simple combination speech emotion feature and facial expression feature. Table 4 shows the results using proposed method. Columns represent the emotion elected in first choice belonging to the emotion of each row, where "J" stands for joy, "A" stands for anger, "P" stands for surprise, "S" stands for sadness, and "F" stands for fear.

**Table 1.** The results of the emotion evaluation only using speech emotion feature.

| Emotion | J | A | P | S | F |
|---|---|---|---|---|---|
| Joy | 86% | 0 | 11% | 2% | 0 |
| Anger | 4% | 81% | 0 | 7% | 8% |
| Surprise | 20% | 1% | 77% | 2% | 0 |
| Sadness | 5% | 4% | 0 | 88% | 3% |
| Fear | 5% | 10% | 6% | 4% | 75% |

**Table 2.** The results of the emotion evaluation only using facial expression feature.

| Emotion | J | A | P | S | F |
|---|---|---|---|---|---|
| Joy | 85% | 2% | 13% | 0 | 0 |
| Anger | 0 | 79% | 7% | 10% | 4% |
| Surprise | 0 | 0 | 81% | 9% | 10% |
| Sadness | 0 | 20% | 4% | 66% | 10% |
| Fear | 3% | 8% | 2% | 9% | 78% |

**Table 3.** The results of the emotion evaluation by simple combination speech emotion feature and facial expression feature.

| Emotion | J | A | P | S | F |
|---|---|---|---|---|---|
| Joy | 92% | 1% | 6% | 1% | 0 |
| Anger | 1% | 88% | 2% | 7% | 2% |
| Surprise | 4% | 0 | 90% | 4% | 2% |
| Sadness | 5% | 8% | 2% | 85% | 0 |
| Fear | 1% | 3% | 2% | 6% | 88% |

**Table 4.** The results using proposed method.

| Emotion | J | A | P | S | F |
|---|---|---|---|---|---|
| Joy | 96% | 2% | 2% | 0 | 0 |
| Anger | 2% | 88% | 2% | 8% | 2% |
| Surprise | 0 | 3% | 91% | 6% | 0 |
| Sadness | 3% | 7% | 4% | 84% | 2% |
| Fear | 0 | 5% | 1% | 1% | 93% |

From the table 1, we could see that the average recognition correct rate for only using speech signal is 81.4%; From the table 2, we could see that the average recognition correct rate for only using facial expression signal is 77.8%. Due to human beings express emotion

information in many ways. It has the complexity of expression and the relative property of culture. So there are many limitations for single mode emotion recognition. From table 3, we could see that the average recognition correct rate by simple combination speech emotion feature and facial expression feature has some increase, but the improvement is not obvious. So the fusion of different channel information is bottleneck problem for multi model emotion recognition, and it is directly related to the accuracy of emotion recognition. From table 4, we could see that the average recognition correct rate by using proposed method is 90.4%. So the method improves the accuracy of emotion recognition by giving full play to the advantages of decision level fusion and feature level fusion, and makes the whole fusion process close to human emotion recognition more.

## 6 Conclusion

This paper proposed a new multi model emotion recognition method, and improved the accuracy of emotion recognition. But this paper merely aimed at given text to recognize emotion, it has a certain distance with practical level. So the emotion recognition for non specific text will be our next research direction. Certainly, many shortages are lying the selecting of the features, so to find more efficient features and to do further analysis and experiment in a wide field will be our future work.

## Acknowledgment

The authors wish to deeply thank graduate students who collaborated in the experiments and in the development of the system. The work is also supported by grant from the National Natural Science Foundation of China (No. 61503038, No. 61403042).

## References

1. L.L. Yu, Z.X. Cai, M.Y. Chen, Study on emotion feature analysis and recognition in speech signal: an overview, J. Circuits and Systems, **12**, 4 (2007): 76–84.
2. Y.H. Yan, Y. Zhou, Y.Q. Sun, Feature Extraction Method for Speech Emotion Recognition. China: 2010102729713 (2010).
3. X. Mao, L.J. Chen, Speech emotion recognition based on parametric filter and fractal dimension, IEICE T INF SYST, **93**, 8 (2010): 2324–2326.
4. C.R. Zou, L. Zhao, Speech Emotion Recognition Method Based on Improved Fuzzy Vector Quantization. China: 2008101228062 (2008).
5. Y. Attabi, P. Dumouchel, Anchor models for emotion recognition from speech, TAC, **4**, 3 (2013): 280–290.
6. W.M. Zheng, M.H. Xin, X.L. Wang, A novel speech emotion recognition method via incomplete sparse least square regression, IEEE SIGNAL PROC LET, **21**, 5 (2014): 569–572.
7. Q. R. Mao, M. Dong, Z. W. Huang, Learning salient features for speech emotion recognition using convolutional neural networks, IEEE MULTIMEDIA, **16**, 8 (2014): 2203–2213.
8. P. Ekman, W. Friesen, Facial action coding system: a technique for the measurement of facial movement. Palo Alto: Consulting Psychologists Press, (1978).
9. L.H. Liang, H.Z. Ai, G.Y. Xu, A survey of human face detection, Chinese J. Computers, **25**, 5 (2002): 449–458.
10. Y. Rahulamathavan, R.C. W. Phan, J.A. Chambers, Facial expression recognition in the encrypted domain based on local fisherdiscriminant analysis, TAC, **4**, 1 (2013): 83–92.
11. W.M. Zheng, Multi-view facial expression recognition based on group sparse reduced-rank regression, TAC, **5**, 1 (2014): 71–85.
12. P.C. Petrantonakis, L.J. Hadjileontiadis, Emotion recognition from EEG using higher order crossings, IEEE T INF TECHNOL B, **14**, 2 (2010): 186–197.
13. S.L. Lin, G.Y. Liu, H.L. Zhang, Application of ACO algorithm to emotion recognition research based on RSP signal, IJCEA, **47**, 2 (2011): 169–172.
14. H. Zacharatos, H. Gatzoulis, Y.L. Chrysanthou, Automatic emotion recognition based on body movement analysis: a survey, IEEE COMPUT GRAPH, **34**, 6 (2014): 35–45.
15. Z. Zeng, M. Pantic, G.I. Roisman, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE T PATTERN ANAL, **31**, 1 (2009): 39–58.
16. J. Kim, E. Andre, Emotion recognition based on physiological changes in music listening, IEEE T PATTERN ANAL, **30**, 12 (2008): 2067–2083.
17. C.W. Huang, Y. Jin, Q.Y. Wang, Multimodal emotion recognition based on speech and ECG signals, JSEU (NSE), **40**, 5 (2010): 895–900.
18. C. Busso, Z. Deng, S. Yildirim, Analysis of emotion recognition using facial expressions, speech and multimodal information, ICMI 2004, (2004): 205–211.
19. S. Hoch, F. Althoff, G. Mcglaun, Bimodal fusion of emotional data in an automotive environment, ICASSP 2005, (2005): 1085–1088.
20. A. Sayedelahl, R. Araujo, M.S. Kamel, Audio-visual feature-decision level fusion for spontaneous emotion estimation in speech conversations, ICMEW 2013, (2013): 1–6.
21. R. Tato, R. Santos, R. Kompe, Emotion space improves emotion recognition, ICSLP 2002, (2002): 2029–2032.