TELKOMNIKA, Vol.11, No.1, January 2013, pp. 83~90
ISSN: 2302-4046

Research of Emotion Recognition Based on Speech and Facial Expression

Yutai Wang*[1,2], Xinghai Yang[1], Jing Zou[1]
[1]School of Information Science and Engineering, University of Jinan, Jinan, 250022, China
[2]Shandong Provincial Key Laboratory of Network based Intelligent Computing, Jinan, 250022, China
*corresponding author, e-mail: ise_wangyt@ujn.edu.cn

Abstract

The paper introduced the present status of speech emotion recognition. In order to improve the single-mode emotion recognition rate, the bimodal fusion method based on speech and facial expression was proposed. The emotional databases of Chinese speech and facial expressions were established with the noise stimulus and movies evoking subjects' emtion. On the foundation, we analyzed the acoustic features of Chinese speech signals under different emotional states, and obtained the general laws of prosodic feature parameters. We discussed the single-mode speech emotion recognitions based on the prosodic features and the geometric features of facial expression. Then, the bimodal emotion recognition was obtained by the use of Gaussian Mixture Model. The experimental results showed that, the bimodal emotion recognition rate combined with facial expression was about 6% higher than the single-model recognition rate merely using prosodic features.

Keywords*: Gaussian mixture model; Prosodic features; Facial expression; Bimodal recognition.*

1. Introduction

With the wide application of computers in various fields, speech recognition as the key technology of human-computer interaction has attracted more and more attention. However, current research on emotion recognition do not go far enough. The studies on many aspects have not led to a systematic theory such as the establishment of the emotional speech databases, the selection and parameter extraction of emotional features, emotion recognition methods [1]. The related studies on English, Japanese, and etc are comparatively more than those on Chinese.

Speech emotion recognition methods are mainly concentrated on speech signals. Emotion feature selection is mainly on prosodic parameters. Emotional analysis methods mainly include Principal Component Analysis, Gaussian Mixture Models, Hidden Markov Models, Support Vector Machines, and etc [2]. There have got been certain results in those areas. But there has been less research on Multi-modal speech emotion recognition which integrates facial expression and human physiological signals.

Because of the inherent defects of voice in the emotion detection, using voice signals to identify the emotional state, the recognition rate can only reach about 80%, and the robustness of the recognition results can not be guaranteed. Emotion detection from a single channel has become increasingly unable to meet the actual needs of the project [3-5]. Therefore, complementary features extracting from the dual-mode has become new ways to improve speech emotion recognition rates.

This paper presents a dual-mode recognition method based on prosodic features and facial expression to increase the rate of speech emotion recognition and robustness.

2. Research Method

2.1. The System Model

As shown in Figure 1, the entire process is roughly divided into two major parts: signal processing, and emotional training recognition [6]. Firstly, we take some proper preprocessing for the audio signals to obtain an effective voice signal. The preprocessing includes pre-

Received September 27, 2012; Revised November 22, 2012; Accepted November 29, 2012

emphasis, endpoint detection, framing, window adding, and etc. After that we change the prosodic features extracted from the processed voice signals and facial geometric features extracted from processed pictures into feature vectors respectively.

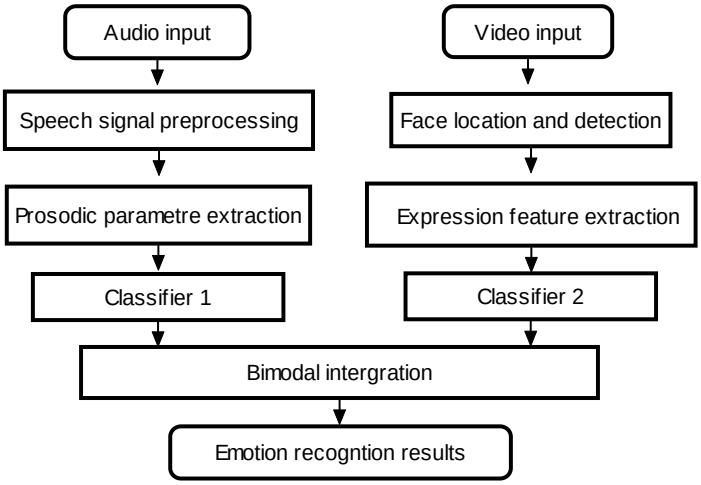

Figure 1. Bimodal emotion detection system

In the recognition stage, After extracting prosodic feature parameters and reducing dimensions of the audio test samples and the video test samples, it is the turn to input the Gaussian mixture model trained for recognition. Finally, we integrate the recognition results of two classifiers for the final judgment.

2.2. Emotion and Features

Psychological studies have shown that changes in human emotions reflect through prosodic parameters of speech. Generally, acoustic features associated with the emotions includeing pitch, duration time, energy, formant, and average, maximum, minimum, intermediate values, ranges, the first derivative, the second derivative and change rates derived from them [7]. After repeated experiments, this paper eventually selected the following prosodic features: phonation time, speech rates, basic frequency averages, basic frequency ranges, basic frequency change rates , Amplitude averages, Amplitude change ranges, formant change averages, formant change ranges, and formant change rates.

Face features generally include three kinds of: Geometric features, physical features, mixing features. The physical features refer to the features using the whole face image pixels, reflect the underlying information of face images, and focus on extracting the subtle changes of local features [8]. However, the number of feature point? extracted is too many that resuling to the higher dimension and the complex calculations. Mixing features combine the geometric features with physical features. The calculation of it is also complex, and the initial point is difficult to obtain [9]. The recognition effect of the geometric features require a higher accuracy of the Datum point extracted. The recognition effect of requiring a higher accuracy of the Datum point extracted. Meanwhile extracting the geometric features ignores the other information of faces(such as skin texture changes etc.) But it can describe the macro structural changes of the face, and the easy way to extract and the lower dimension making it quite comply with the requirements of our emotional system.

We preprocess on the expression samples of detection and location the light compensation, Normalizing, graying, Gaussian smoothing and preprocessing to obtain valid information of expression pictures. Then we extract facial geometric features from the preprocessed expression pictures to form feature vectors, and compare them with the samples of expression template library established after training, thus distinguish the images from different emotianal categories.

2.3. Multi-modal Fusion Recognition Algorithm

To take full advantage of speech prosodic features and facial expression features for bimodal speech emotion recognition, we need to analyze the feature fusion algorithm related.

Firstly,we design two expression classifiers respectively based on the speech prosodic features and facial expressions, and then we judge and integrate the two classifers according to certain rules [10]. Emotional categories to be identified include five kinds of happiness, anger, surprise, sadness and calm. For the two kinds of classifiers we both used Gaussian mixture model (GMM) to train the probability models of each emotion category [11]. GMM is a weighted sum of the M members' density and can be represented as the following:

$$P(X_t|\lambda) = \sum_{i=1}^{M} a_i b_i(X_t) \quad (1)$$

In this formula, x_t is the tth random vector of D-dimensions; a_i is the Mixed weights; $b_i(x_t)$ is the members'density. Each member's density is a Gaussian function of D-dimensional variables on the average vector Ui and covariance matrix Σ_i, and can be represented as the following form:

$$b_i(X_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(X_t - U_i)^T \Sigma_i^{-1}(X_t - U_i)\right\} \quad (2)$$

We use iterative calculation of EM algorithm to obtain the parameter estimation of GMM model.

When there is noise, the performance of speech classifier will fall; therefore it is necessary to consider classifier's confidence level at some point when we select fusion algorithm of judgment layer [12]. Here we choose a sample adaptive method to measure whether it is a reliable judgment of the current sample by the classifiers. We assign the classifer of high confidence with high fusion weights, while that of low confidence with low fusion weights. We consider the sample is in the non-overlapped region of the probability distribution model, and the judgment confidence of the sub-classifier is relatively high. Therefore, the mixed weights of each sub-classifier can be expressed as:

$$w_j = \frac{\left|\ln(P(X|\lambda_1)) - \ln(P(X|\lambda_2))\right|}{\left|\sum_{k=1}^{2} \ln(P(X|\lambda_k))\right|} \quad (3)$$

Which, X is feature vector; GMM likelihood of two kinds of emotional categories given by the sub-classifiers(voice classifer and expression classifer) is expressed as $P(X|\lambda_k)$, k=1,2. If the judgement of classifer is more reliable, the difference is bigger;and vice versa. we get the final output of classifier fusion judgment.

$$Y = \arg\max\left\{\sum_{j=1}^{2} w_j p^j(X|\lambda_k)\right\} \quad (4)$$

3. Experiments and Results Analysis

3.1. Establishment of Database

In the experiment, we select happiness, anger, surprise, sadness, and calm as typical emotion to be studied. During the collection process of voice sample, we choose Chinese

sentences that are frequently used in the daily life and will not be ambiguous under a variety of emotional states as the recording. The texts are shown in Table 1.

Table 1. Recording texts

Number	Voice texts	Number	Voice texts
1	What's your name?	6	I'm glad today.
2	You are back.	7	He is very practical.
3	Today is the weekend.	8	The weather is more and more hotter.
4	I had a dream.	9	This thing is so regrettable.
5	I didn't expect he would be like this.	10	This is my book.

We select four men and women respectively with strong capacity of emotional expression. Firstly we conduct the speakers into the required emotional states by words, music and movies. Then we record videos and voice using microphone and camera. each text of the same emotional state is recorded 10 times, and then we get 800 videos and 800 voice samples. We test the samples of the same sentences under each emotional state subjectively by hearing, and retain effective samples. And then we convert video data into continuous pictures of BMP format, selecting expression samples with relatively good quality. Finally we get the 400 voice samples of good quality and 400 expression samples for training and testing to establish the emotion database of this research project.

Figure 2 shows a woman speaks the joint sample of "Today is the weekend" with happy emotion.

(1) Image series (2) Audio signal

Figure 2. Joint sample of "happy-today is the weekend"

3.2. Prosodic Features Extraction

In this study, we select ten prosodic feature parameters of voice signals such as the pronunciation time, speech rates, basic frequency averages, basic frequency ranges, basic frequency change rates , Amplitude averages, Amplitude change ranges, formant change averages, formant change ranges,formant change rates to analyze and find the structural characteristics and distribution patterns of different emotional signal features. The experimental results are shown in Table 2.(a),(b).

Table 2.(a) Five prosodic feature parameters under different feature states

Emotion category	Time	Speed	Basic frequency averages	Basic frequency ranges	Basic frequency change rates
calm	0.72	5.23	285.31	158.41	21.09
happy	0.86	4.21	290.22	212.13	22.41
surprise	0.81	4.69	293.45	187.82	36.08
sad	0.91	4.13	280.02	231.23	26.81
angry	0.60	6.31	292.86	172.76	27.89

(b) The remaining prosodic feature parameters under different feature states

Emotion category	Amplitude averages	Amplitudechange ranges	Formant change averages	Formant change ranges	Formant change rates
calm	1.06	1.64	725.25	2048.21	120.13
happy	1.36	3.16	721.07	2056.33	126.11
surprise	1.63	7.03	868.51	2051.52	137.26
sad	0.68	2.81	801.13	2057.47	125.92
angry	1.52	2.13	695.09	2059.56	128.74

Through the experiment we can know that the ten emotion feature parametres selected can reflected the emotion features of different voice signals well. The experimental data in Table 2 also verified the distribution rules of prosodic features of voice signals under different emotional states.

When a person is under the emotional state of anger, her speaking speeed will become faster, the volume will become larger, and the tone will become higher; In the pronunciation duration of the signal,the expression of anger and surprise last shorter, while that of happiness and sadness last longer. The amplitude of voice signals under the three emotional states of happy, anger, surprise is large, while that under the states of sad and calm is little. The differences of time, speed, energy parametres in different emotional sentences are shown in Table 3.

Table 3. Time speed energy ratios under different sppech emotion

Emotion category	calme	happy	surprise	sad	angry
Time ratio	1.000	1.194	1.125	1.264	0.833
Speed ratio	1.000	0.805	0.897	0.790	1.207
Average energy ratio	1.000	1.646	2.362	0.416	2.056

Through the analysis of prosodic features of emotional speech, we extract ten parametres of emotional features from the voice samples in expression databases and eatablish the foundation for emotion recognition through the feature extraction.

However, the number of feature parameters selected is not the more the better. The effective feature parametres which have been choosed will form a lot of feature vectors through statistical calculations, and it will bring some difficulties to the identification process. However part of these feature vectors can contribute little to the emotion recognition. Related researches indicate that the rate of pattern recognition is not proportional to the number of dimensions of feature space. Therefore, after the extraction of feature parametres, we should consider the problems of feature selection and dimensionality reduction of the feature parameters in necessary conditions.

We change the ten feature parameters extracted into feature vectors and reduct the feature dimensions with Principal Component Analysis (PCA) method to obtain the optimal feature parameter group and ensure the better realization of expression recognition.

3.3. Recognition Based on the Voice

In the speech emotion recognition, the training samples are compose of a data matrix and use the method of Gaussian Mixture Model for emotion recognition experiments. We extract the parameters of prosodic features for each training sample and convert them into feature vectors for normalization. For each emotion category, each feature parameter of the speech samples corresponds to a probability density function of Gaussian distribution. The weighted Gaussian distribution corresponding to each feature parameter is composed of a Gaussian mixture model. In the emotion recognition, we extract the emotion feature parameters from the test speech samples and substitute them into the probability density function of Gaussian Mixture Models corresponding to each emotion category, and the one of largest probability is the recognition result. The results are shown as Table 4.

Table 4. The emotion recognition results of Chinese speech samples

Emotion catogory	Test set size	Correctly recognized sentences	Recognition rate(%)
Calm	400	355	88.8
Happy	400	330	82.5
Surprise	400	322	80.5
Sad	400	347	86.8
Angry	400	326	81.5

The resulting data shows that the best rate of recognition can reach 88.8% under the emotional state of calm. The recognition results under surprise states are relatively poor with the rate of 80.5%. The emotion recognition rates under different emotional states are slightly

different, mainly related to the sample quality and the individual's habit with expression of emotions. In the experiments, the average recognition rate of five kinds of emotional states is about 84%. In some single-language speech recognition systems with less requirements on the emotion recognition results and more restrictions on the amount of computation, the method of recognizing single-language speech samples based on the basic rhythm features has a certain degree of reliability and availability.

3.4. Facial Expression Recognition

In the experiments, images of facial expressions in the emotion database, due to interference from various factors, often have the problem of relatively low quality. It's very difficult to process the data like this by computer, so there is a need for preprocessing to enhance image quality, such as removal of noise in the image. Meanwhile, there is also a need for the light compensation, suppression from the influence of light and elimination of the deviation of colors.

After light compensation for the facial image, its size, camera angle and the position of facial organs will be slightly different [6]. In order to process and analyze the images, there is a need for the normalization of image sizes to transform the face in the picture into the same size and the same position.

On this basis, we process the expression samples of detection, location and garying to obtain effective information of expression pictures. And then we compared feature pictures, which composed of facial geometric features we extract from the expression pictures, with the samples in the expression library. Finally, we can distinguish the images belonging to which emotion category.

We read the facial expression samples from the test set, and transform them into standard data types. And then we select the first 30,000 pixels, to avoid exceeding the storage range, and use the Gaussian Mixture Model to find the training picture closest to the identified picture. The largest probability is the recognition result. Through the program simulation, the recognition result of 400 facial expressions samples in the emotional face database is shown in Table 5.

Table 5. Emotion recognition results of expression samples in face database

Emotion catogory	Test set size	Correctly recognized samples	Recognition rate(%)
Calm	400	176	44
Happy	400	164	41
Surprise	400	144	36
Sad	400	168	42
Angry	400	148	37

It can be seen from Table 3 that the recognition rate under the calm state is the best, up to 44%. The average recognition rate based on this face database can reach 40%.

3.5. Bimodal Integration Results

For the bimodal emotion recognition of integrated facial expressions, we choose the fusion method on decision layer, and select the single-mode voice emotion classifier and expression classifier respectively. And then we evaluate the confidence of the two sub-classifiers, and get the final results of the emotion recognition through the judgment. Substituting prosodic features and expression features into the probability density function of the Gaussian Mixture Model for each emotion category, the probability of the largest is the recognition result. The experimental result is shown in Table 6.

Table 6. Bimodal emotion recognition results

Emotional category	Test set size	Correct recognition	Recognition rate（%）
Calm	400	377	94.3
Happy	400	354	88.5
Surprise	400	349	87.3
Sad	400	370	92.5
Angry	400	359	89.8

It can be seen from Table 4 that, for the 400 test set samples, the highest recognition rate can reach 94.3%. The average recognition rate is about 90.5%.

Compared single-mode recognition results based on the prosodic features with bimodal recognition results integrated with expression information, the result is shown as Figure 3.

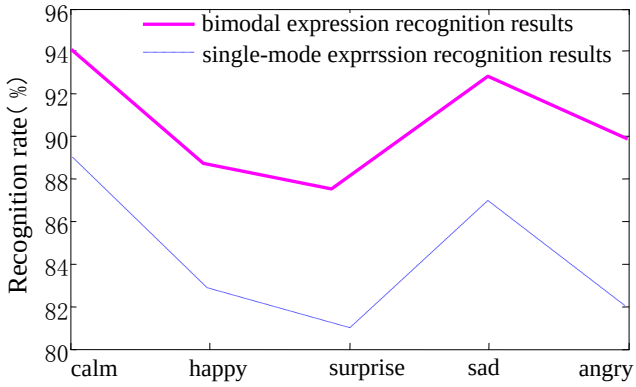

Figure 3. Recognition results contrast between single-mode and bimodal fusion

From the above chart we can see that the bimodal emotional recognition rate increase about 6% more than the single-mode recognition rate. It can be concluded that on the basis of emotion recognition based on speech prosodic features, integrating with facial expression information, the overall emotion recognition rate can be improved. This has reached the purpose of the experiment, and the results are satisfactory.

4. Conclusion

In the experiment, we analyze and compare time, amplitude energy, basic frequency and formant feature parameters under different emotional states, and find out the distribution laws of different emotional signal features. On this basis, we classify five emotional states of calm, sadness, happiness, surprise, and anger. The recognition results show that on these basic Prosodic information we can initially recognize basic emotional catagories, and apply it into the emotion recognition system, which limits the amount of storage and computation and doesn't have strict recognition accuracy. Meanwhile, the prosodic features, integrating with facial expression information, recognizes emotional categories with Bimodal, reaching a higher recognition rate.

Although the emotional recognition performance combinating with facial expressions has improved, The recognition rate doesn't improve significantly. This is mainly because in the terms of obtaining the emotional information, there is a similar correlation between the adjacent video frames, due to the continuity of the facial expression changes. But we didn't take this correlation into consideration when catching the instant face image to analyze separately. On the other hand, when the facial expression changes, the shape and the location of the organs on the face, will change accordingly. In this paper,although the image analysis method based on Gaussian mixture algorithm has a higher recognition rate for the face contour, it lacks of detailed characterization of changes in the eyes, nose, mouth and other facial organs. Based on the above two reasons, in order to truly improve the system performance, we need to build a correction model associated with the expressions containing a variety of rules, and modified the image recognition results using the model. In addition, in the term of real-time applications, besides enhancing the robustness of the system and improving the accuracy, the efficiency of the recognition algorithm is also a key factor. The strategies such as codebook pruning, data compression can also improve the recognition rate effectively.

Multi-modal recognition systems intergrating with images, voice and other emotional information is the inevitable trend of future human-computer interaction development. Although

there are still many insurmountable technical problems, with the continuous progress of science and unremitting efforts of the researchers, the real-time systems of multi-modal speech recognition will have more potential development.

References

[1] LS Zhao, Q Zhang, XP Wei. A Research Progress in Speech Emotion Recognition. *Computer Application and Research.* 2009; 26(2): 428-432.
[2] CW Huang, Y Zhao, Y Jin. A Study of Practical Speech Emotion Features Analysis and Recognition. *Electronics and Information Journal.* 2011; 33(1): 112-116.
[3] LL Xu, ZX Cai, MY Chen. A Study Review of Emotion Feature Analysis and Recognition of Speech Signals. *Circuits and Systems Journal.* 2007; 12(4): 77-84.
[4] YM Huang, GB Zhang, HB Liu. Emotion Detection Based on New Bimodal Fusion Algorithm. *Journal of Tianjin University.* 2010; 43(12):1067-1072.
[5] CW Huang, Y Jin, QY Huang. Mutil-Modal Emotion Recognition Based on Speech Signals and ECG. 2009; 40(5): 895-900.
[6] B Xie. A Study of Key Technologies on Mandarin Speech Emotion Recognition. *Zhejiang: Zhejiang University Computer Science and Technology Major*. 2006.
[7] YM Huang, GB Zhang, X Li. Speech Emotion Recognition Base on Small Samples of Global Features and Weak-Scale Integration Strategy. *Acoustics Journal.* 2012; 37(3): 330-338.
[8] LQ Fu, YB Wang, CJ Wang. Speech Emotion Recognition Based on Mutil-Feature Vectors. *Computer Science*. 2009; 36(6): 231-234.
[9] XQ Jiang, SY Cui, YH Yin. Speech Emotion Processing in the Man-Machine Speech Interaction. *Jounal University of Jinan : Version of nature science.* 2008*;* 22(4): 354-357.
[10] YL Xue, X Mao, Y Guo. A Research Pogress in Facial Expression Recognition. *Chinese Image and Graphic Journal.* 2009; 14(5): 764-772.
[11] M Fan. Emotional Speech Recognition Based on Facial Expression Analysis. *Shandong: Shandong University Circuits and Systems Major.* 2009.
[12] TF Zhang, R Min, BY Wang. Facial Expression Recognition Based on Automatic Segmentation of the Characteristic Regions. *Computer Engineering.* 2011*;* 37(10): 146-151.