Bimodal Emotion Recognition from Speech and Text

Weilin Ye
Institute of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, China

Xinghua Fan
Institute of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, China

***Abstract*—This paper presents an approach to emotion recognition from speech signals and textual content. In the analysis of speech signals, thirty-seven acoustic features are extracted from the speech input. Two different classifiers Support Vector Machines (SVMs) and BP neural network are adopted to classify the emotional states. In text analysis, we use the two-step classification method to recognize the emotional states. The final emotional state is determined based on the emotion outputs from the acoustic and textual analyses. In this paper we have two parallel classifiers for acoustic information and two serial classifiers for textual information, and a final decision is made by combing these classifiers in decision level fusion. Experimental results show that the emotion recognition accuracy of the integrated system is better than that of either of the two individual approaches.**

Keywords—emotion recognition; acoustic features; textual features; decision level fusion

I. INTRODUCTION

With the advent of information age and the popularity of Internet, more and more kinds of information come to our life. Phoning has become the main means of daily communication and follow-up contacting. We often play some customer service phone to ask for information about some products, after the call we always asked to evaluate the service attitude of the telephone operator, so that the businesses can know the service quality of the staff. However, manual evaluation often has a problem of objectivity and authenticity. Automatic emotion recognition is one of the key techniques of human-computer interaction [1].

In recent years, several research works have focused on emotion recognition. Hoch et al[2] presented a method to recognize three kinds of emotional states in the automotive environment from speech and expression information. Busso et al[3] analyzed the complementarity of speech emotion recognition and facial expression recognition, presented a multi-modal emotion recognition method from feature level fusion and decision level fusion. Wangner et al[4] combined electromyogram, ECG, skin resistance and breathing these four kinds of physiological parameters to recognize emotional state and got a recognition rate of 92%.

However, few approaches have focused on emotion recognition from textual input. Textual information is another important communication medium and can be retrieved from many sources, such as books, newspapers, web pages, e-mail messages, etc. It is not only the most popular communication medium, but also rich in emotion. With the help of natural language processing techniques, emotions can be extracted from textual input. In this paper, a bimodal emotion recognition method is used to extract emotion information from both speech and text input. In this paper, the classifiers recognize emotions according to two simple types: positive and non-positive. This paper designed two parallel classifiers for acoustic information and two serial classifiers for textual information, and a final decision is made by combing these classifiers in decision level fusion. This method can be applied to a telephone service center dialogue system to recognize customers' negative emotions, such as anger, impatience etc. so that to turn the answering service to manual service automatically to avoid losing customers.

II. EMOTIONAL SPEECH CORPUS

At present, there still not have a public database for Chinese speech emotion recognition research. Generally there are two ways to get emotional speech corpus: a) Recording; b) Clipping. Recording method has better customization, and can record emotional speech which meets the speaker, text, emotion categories and other requirements. According to the general rules of building corpus, four college students around the age of 20 with higher emotional expression ability are invited to participate in recording (2 females, 2 males). After five non-recording people's perception experiments, we removed nearly 40% corpus which are not sure which kind of emotion. Finally we picked out a total of 600 available corpuses, including positive and non-positive each 300, where non-positive include anger, sadness, fear and other negative emotions.

III. PREPROCESSING

The purpose of voice and text preprocessing are different. Voice preprocessing is to get pure voice by eliminating the interference of various factors. Text preprocessing is to get relatively clean data sets by filtering noise data.

A. Speech Signal Preprocessing

1) Pre-emphasis

Since speech signal are affected by the glottis excitation and snout radiation, the high frequency part of the speech signal falls down. Pre-emphasis enhance the high frequency part, make the signal spectrum flat over the entire frequency band.

2) Window Function

Commonly used window functions in voice processing are rectangular window and hamming window.

a) Rectangular Window:

$$w(n)=\begin{cases}1 & (0\le n\le N-1)\\ 0 & Other\end{cases}$$

b)

(1)

c) Hamming Window:

$$w(n)=\begin{cases}0.54-0.46\cos\left[\dfrac{2\pi n}{N-1}\right] & (0\le n\le N-1)\\ 0 \quad Other\end{cases} \quad (2)$$

B. Text Preprocessing

Firstly we use the Chinese auto-segmentation system to do the process of word segmentation, and then move the stop words from target text. Finally we can get relatively clean data sets. The process of text pretreatment is showed in figure1.

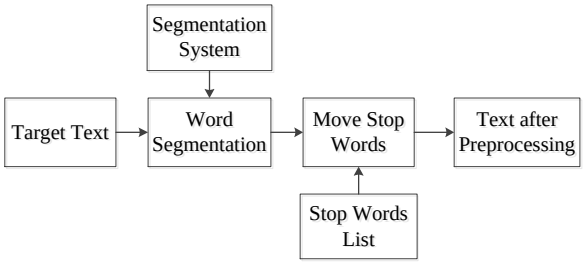

Fig. 1. Flow chart of text preprocessing

IV. Emotional Feature Extraction

A. Acoustic Features

Speech signal is short-time stationary, calculating features based on short time frame are: short-time amplitude energy, short pitch and first-three formant. For the whole speech, every feature is calculated as one-dimensional sequence. However these sequences cannot be directly used as a feature vector for pattern recognition, commonly way to solve this problem is to calculate its statistical value, such as mean and slope.

Speech emotion recognition based on prosodic features has strong robustness and adaptability. Statistical characteristics can better reflect the rhythmic structure of speech. On the basis of previous experiment, we chose 37 identification features, which are shown in Table 1.

B. Textual Features

1) Feature extension

Text orientation classification is different from general classification, words or phrases with semantic orientation or emotional tendencies play a crucial role for classification. In this paper we use three-step feature extension [5] to reconstruct the data sets, which can extent features of the data sets by using list of tendency words, negative words and degree adverbs. This method can enhance expression ability of the textual features by adding words or phrases with semantic orientation to feature sequence.

2) Feature Selection

Commonly used feature selection methods are: document frequency, mutual information, information gain, expects cross-entropy, chi-square statistics etc.

TABLE I ACOUSTIC FEATURES

Feature	Describe	Feature	Describe
1	Maximum energy	20	Mean duration of pitch frequency contour ascent
2	Mean value of energy	21	Maximum of pitch frequency contour decline
3	Median value of energy	22	Mean value of pitch frequency contour decline
4	Rate of change of energy	23	Maximum duration of pitch frequency contour decline
5	Maximum of energy contour ascent	24	Mean duration of pitch frequency contour decline
6	Mean value of energy contour ascent	25	Maximum of the first formant
7	Maximum duration of energy contour ascent	26	Mean value of the first formant
8	Mean duration of energy contour ascent	27	Median value of the first formant
9	Maximum of energy contour decline	28	Rate of change of the first formant
10	Mean value of energy contour decline	29	Maximum of the second formant
11	Maximum duration of energy contour decline	30	Mean value of the second formant
12	Mean duration of energy contour decline	31	Median value of the second formant
13	Maximum of pitch frequency	32	Rate of change of the second formant
14	Mean value of pitch frequency	33	Maximum of the third formant
15	Median value of pitch frequency	34	Mean value of the third formant
16	Rate of change of pitch frequency	35	Median value of the third formant
17	Maximum of pitch frequency contour ascent	36	Rate of change of the third formant
18	Mean value of pitch frequency contour ascent	37	Speed(voice frames / statement of words)
19	Maximum duration of pitch frequency contour aascent		

However these methods are not much suitable for text orientation classification. In this paper we chose the document frequency feature selection formula presented in literature [6] which considered the words tendentiousness.

$$DF\_Sen(t,c)=\frac{\lg(DF_t)}{\lg(N_c)}*\frac{\alpha_t(|\beta_t|+1)}{\alpha_t+\gamma} \quad (3)$$

Among it, DF_t means the number of documents showed in class c of feature t, N_c means the whole numbers of documents in class c, β_t means the intensity values of orientation, α_t means the number of words feature t contains, γ means the weighing coefficient which can be adjusted in experiment. When selecting parameter, we set a threshold $DF\_SEN_{\min}$. If the threshold of a feature is less than a certain value, it will be deleted. In this paper, based on experiments we select 0.04 as the value of threshold. When a feature word appeared at multiple classes, we selected it according to its feature score in

each category. If the absolute value of difference value of feature score in two deferent categories is more than 0.12, the word will be selected as textual feature.

V. BIMODAL FUSION RECOGNITION ALGORISM

Currently, there are two ways to combine different pieces of information: a) feature level fusion, b) decision level fusion. The problem with feature level fusion is the potential of having to face the nurse of dimensionality due to the increase in the input feature dimension. In our case, we have two parallel classifiers for acoustic information and two serial classifiers for textual information, and a final decision is made by combing the outputs results from these classifiers in decision level fusion.

A. Classification of the Acoustic Set

In this paper, we have two parallel classifiers for acoustic information. They are support vector machines (SVM) and BP neural nets.

1) Support Vector Machines

A great interest in Support Vector Machines (SVM) in classification can be observed recently. They tend to show a high generalization capability due to their structural risk minimization oriented training. Non-liner problems are solved by a transformation of the input feature vectors into a generally higher dimensional feature space by mapping function where linear separation is possible. Maximum discrimination is obtained by an optimal placement of the separation plane between the border of two classed.

2) BP Neural Nets

The BP Neural Nets is proposed by a team of scientists led by Rumelhart and McCelland, and it is one of the most widely used neural network model. BP network can learn and store a large amount of mapping relationship of input-output model without pre-revealing the mathematical equations that describe the mapping relationship.

B. Classification of the Textual Set

In this paper, we use the two-step classification proposed in literature [7]. We construct two serial classifiers CF1 and CF2, both of them use equation (4) to select features. Firstly use CF1 for classification. For unreliable part of the classification results, we use CF2 for secondary classification.

1) To construct classifier CF1

CF1 is Naive Bayes classifier. Text d is expressed as $d=(t_1,t_2,K,t_n)$, t_k is feature item of the text. Then Naive Bayes classifier is expressed as follows:

$$P(C_i \mid d) = \max(P(d \mid C_i)P(C_i)), i = 1,2 \qquad (4)$$

In formula (4), $P(d \mid C_i) = \prod_{j=1}^{n} P(t_j \mid c_i)$

2) To construct classifier CF2

Classifier CF2 is expressed as follows:

$$f(d) = \sum_{i=1}^{n} \left(\left| \frac{\lg(P(t_k \mid C_1)/P(t_k \mid C_2))}{\lg(P(t_k \mid C_1)P(t_k \mid C_2))} \right| * \lg\left(\frac{P(t_k \mid C_1)}{P(t_k \mid C_2)}\right) + \lambda Q(t_k) \right) \qquad (5)$$

Where c_1 represents positive emotion, c_2 represents non-positive emotion, $Q(t_k)$ represents the intensity values of feature t_k, $Q(t_k)>0$ means it's a positive word, $Q(t_k)<0$ means it's a negative word. If feature t_k is not in the tendency word list, then $Q(t_k)=0$, λ is adjustment coefficient, if $f(d)>0$, then text d is positive, otherwise text d is non-positive.

C. Fusion Algorism in Decision Level

In this paper we construct two parallel classifiers for acoustic information and two serial classifiers for textual information, and a final decision is made by combing these classifiers in decision level fusion. Flow chat is showed in figure2.

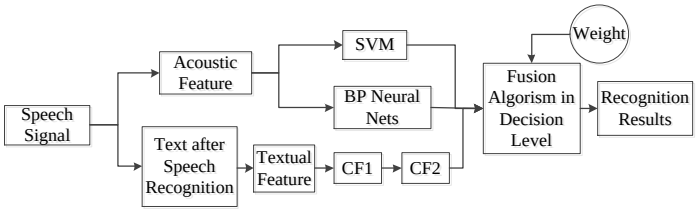

Fig. 2. Fusion algorism in decision level

Assuming that w_1, w_2, w_3 each represents the weight of SVM, BP neural nets and textual classifier. The right value of w_1, w_2, w_3 plays a very important role to the fusion result. As different classifier has a different recognition rate, a number of samples extracted from the training set as a check set. The value of w_1, w_2, w_3 is determined by the recognition rate of the check set. P_1, P_2, P_3 each represents the recognition rate of SVM, BP neural nets and textual classifier. Then the value of w_i is calculated as follows:

$$w_i = \frac{P_i}{\sum_{i=1}^{3} P_i} \qquad (6)$$

The idea of weighted score voting strategy: if the three classifiers have the same recognition result, then the sample will be identify as such class; if two of the three classifiers have the same recognition result, then sum the two weights of the classifiers with the same result, and compare it with the weight of the classifier which has a different result, then the sample will be identify as class recognized by the classifier with bigger weight.

VI. EXPERIMENT AND RESULTS

Both of the training data sets and the testing data sets have two kinds of emotion: positive and non-positive. Each training set contains every emotion 200 speech samples and 200 text samples, each testing sets contains every emotion 100 speech samples and 100 text samples. In experiments, we use the same training sets and testing sets to test every single model classifier.

In this paper, we use 2*2 confusion matrix to evaluate the emotion recognition algorism. The element in row i and column j means the proportion that the real emotion state i is recognized as j. That is to say, the greater values on the diagonal matrix are, the better effect of the emotion recognition algorism is.

Experiment 1: recognition rate of single-mode SVM classifier based on acoustic features is shown in Table 2.

TABLE II ACCURACY OF SVM (%)

Sample Sets	Positive	Non-positive
Positive	82	18
Non-positive	16	84

Experiment2: recognition rate of single-mode BP Neural Nets classifier based on acoustic features is shown in Table 3.

TABLE III ACCURACY OF NEURAL NETS (%)

Sample Sets	Positive	Non-positive
Positive	76	24
Non-positive	22	78

Experiment3: recognition rate of single-mode classifier based on textual features is shown in Table 4.

TABLE IV ACCURACY OF TEXTURL (%)

Sample Sets	Positive	Non-positive
Positive	90	10
Non-positive	12	88

Experiment4: recognition rate of decision level fusion algorism is shown in Table 5.

TABLE V ACCURACY OF FUSION METHOD (%)

Sample Sets	Positive	Non-positive
Positive	94	6
Non-positive	8	92

From table 2 we can see the recognition rate of single-mode SVM classifier based on speech signal is around 83%. Non-positive emotion has shown its importance using value in practical application. Average recognition rate of non-positive emotion is around 81%, which means the acoustic features extracted in this paper have a higher correlation with non-positive emotion, can be used to recognize non-positive emotions. From table 3 we can see the recognition rate of single-mode BP Neural Nets classifier based on speech signal is around 77%. From table 4 we can see the recognition rate of single-mode classifier based on textual features is around 89%. From table 5 we can see the recognition rate of decision level fusion algorism proposed by this paper is around 93%, it's better than all the single-mode classifiers.

From experiment result we can see that the bimodal fusion algorism presented by this paper obtained the expected effect. The advantage of decision level fusion algorithm is that each classifier is independent from each other, when emotional data not available of with low quality in one channel, decision lever can still recognize emotion state, it has good robustness.

VII. CONCLUSION AND PROSPECTS

This paper presents an approach to bimodal emotion recognition from speech signals and textual content. We conduct two parallel classifiers for acoustic information and two serial classifiers for textual information, and a final decision is made by combing these classifiers in decision level fusion. Experimental results show that the emotion recognition accuracy of the integrated system is better than that of either of the two individual.

Emotion recognition cannot only combine speech and text, but also heart rate, blood pressure, skin current etc. physiological characteristics, which can be applied to polygraph, entertainment and many other areas. Affective computing will help artificial intelligence become more and more humanized.

REFERENCES

[1] Z. Zeng, M Pantic and Gl Roisman, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33(1), pp. 39-58, 2009.

[2] S. Hoch, F Althoff, G McGlaun and G. Rigoll, "Bimodal fusion emotional data in an automotive environment," IEEE International Conference on Acounstics, Speech, and Signal Processing, USA, 2005, pp. 1085-1088.

[3] C. Busso, Z. Deng, S. Yildirim, M. Bulut, CM. Lee, A. Kazemzadeh et al. "Analysis of emotion recognition using facial expressions, speech and multimodal information," Prcoceeding of the Sixth International Conference on Multimodal Interfaces, USA, 2004, pp. 205-211.

[4] J. Wagner, J. Kim and E. Andre, "From physiological signals to emotions: implementing and comparing selected method for feature extracton and classification, " IEEE International Conference on Multimedia & Expo. Netherlands, 2005, pp. 940-943.

[5] XH. Fan, P. Wang and P. Zhou. "An two step text tendency analytical method based on extension," Computer engineering and Applications, China, vol. 48(1), pp. 162-165, 169, 2012.

[6] XH. Fan, and H. Wu, "Research of text orientation in the opinion leaders identification", Computer Application Study, China, vol. 30(9), pp. 2613-2615, 2636, 2013.

[7] XH. Fan, and MS. Sun, "A high performance two-class chinese text categorization method," Journal of Computers, China, vol. 29(1), pp. 124-131, 2006.