SPEECH EMOTION RECOGNITION BASED ON A HYBRID OF HMM/ANN

XIA MAO, BING ZHANG, YI LUO
School of Electronic and Information Engineering
Beihang University
37th Xueyuan Road, Haidian District, Beijing
CHINA

Abstract: - Speech emotion recognition, as a vital part of affective human computer interaction, has become a new challenge to speech processing. In this paper, a hybrid of hidden Markov models (HMMs) and artificial neural network (ANN) has been proposed to classify emotions, combining advantage on capability to dynamic time warping of HMM and pattern recognition of ANN. HMMs, which export likelihood probabilities and optimal state sequences, have been used to model speech feature sequences, while ANN has been employed to make a decision. The recognition result of the hybrid classification has been compared with the isolated HMMs by two speech corpora, Germany database and Mandarin database, and the average recognition rates have reached 83.8% and 81.6% respectively.

Key-Words: - Speech emotion recognition; Hidden Markov model; ANN;

1 Introduction

Affective human computer interaction has been the focus of artificial intelligence research for several years now, and the research has moving ahead from the simple information exchange between human and computer towards the affective communication. Affective human computer interacion technology could be widely applied in virtual reality, especally in the field of entertainment and games. Besides, the virtual human and psychiatric aid are the further application prospects for affective human computer interacion. Making computer recognize the emotion of human being is the foundation of affective human computer interaction. The main carriers of human emotion, including facial expression, posture and speech, are the primary channels for computer to recognize human's emotion. Emotion recognition of speech as a significant part has become a challenge to speech processing. The accustomed way for speech emotion recognition is to distinguish between a defined set of discrete emotions. Manifold classifiers have been employed in this research. The recent approaches relate to K-nearest Neighbors (KNN)[1], hidden Markov model (HMM)[2][3], Gaussian mixtures Model (GMM), support vector machine (SVM)[3] and artificial neural net (ANN). Most advanced researches on a speaker-independent mode achieve recognition rates from 55% to 95%, since even humans could hardly reach emotion recognition rates of about 60% from unknown speakers[4].

HMM, with advantage on dynamic time warping capability has been long time studied for speech recognition. Moreover, it has been proved useful in dealing with the statistical and sequential aspects of the speech signal for emotion recognition[2][3]. However, the classify property of HMM is not satisfying. Meanwhile, ANN is a new approach to pattern recognition, but generally used for classification of static inputs with no sequential processing. In this paper, we design a hybrid classifier for speech emotion recognition based on modeling sequences by HMMs, and making decision by ANN.

2 Speech Emotion Recognition System

It has been proved that both statistical and temporal features of the acoustic parameters affect the emotion recognition of speech[5]. In this paper, HMMs are used to deal with the temporal features, getting likelihood probabilities and state segmentations. Many practical applications proved there is often some physical significance attached to the states of HMM[6]. Therefore distortions based on state-segments are introduced in this paper. Finally, the distortions and likelihood probabilities derived from HMMs are combined to be the input of ANN, and ANN is used to classify emotions. Figure 1 illustrates the structure of the speech emotion recognition system developed in this paper.

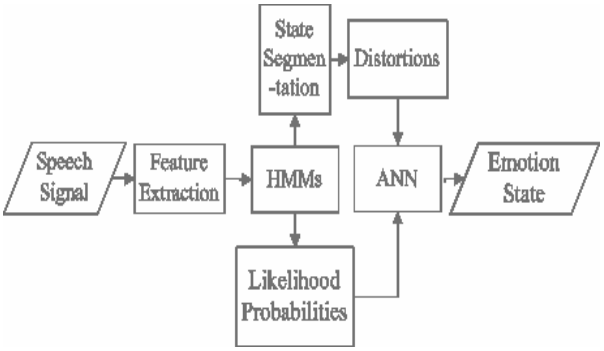

Figure 1. Speech Emotion Recognition System

2.1 Feature Extraction

To select suited features carrying information about emotion is necessary for emotion recognition. Study on emotion of speech indicates that pitch, energy, formant, Mel prediction cepstrum coefficient (MPCC) and linear prediction cepstrum coefficient (LPCC) are effective features to distinguish certain emotions[5, 7-8]. Feature extraction is based on partitioning speech into frames. For each frame, six common features, including pitch, amplitude energy, logenergy, 10-order LPCC, 12-order MFCC and formant, are extracted. These features form the candidate input feature sequences with their first and second derivatives.

2.2 Modeling Emotion Speech by HMMs

HMMs, dealing with the input speech observation sequences containing temporal features, are used to model the emotion utterances, exporting likelihood probabilities and 'optimal' state sequences. In this paper, the HMMs are left-right discrete models, whose input vectors need to be vector quantization (VQ), and the capacity of VQ codebook is key impact to the performance of modeling. The most pervasive methods, Forward-Backward Procedure, Viterbi Algorithm and Baum Welch re-estimation are employed in this paper. Baum Welch re-estimation based on likelihood training criterion is used to train the HMMs, each HMM modeling one emotion; Forward-Backward Procedure exports the likelihood probability; Viterbi Algorithm, focusing on the best path through the model, evaluates the likelihood of the best match between the given speech observations and the given HMMs, then achieving the 'optimal' state sequence.

2.3 State Normalization

Viterbi algorithm could not make time alignment to the observation sequence in accordance with a fixed time scale. Therefore, the state-segments have different lengths. In order to obtain isometric state segments, this paper adopts the method of orthogonal polynomials expansion to normalize the states. Orthogonal polynomials possess the property that makes it possible to expand an arbitrary function f(x) as a sum of the polynomials. We choose Legendre polynomials to be the orthogonal bases. Assuming m is the number of feature vectors in state i, the set of feature vectors can be represented by the following expression:

$$\left\{\overrightarrow{x_1^i}, \overrightarrow{x_2^i}, ..., \overrightarrow{x_j^i}, ..., \overrightarrow{x_m^i}\right\} \quad (1)$$

where $\overrightarrow{x_j^i} = \left[x_{j1}^i, x_{j2}^i, ...x_{jL}^i\right]$, and L indicates the length of feature vector. We list the feature vectors to get a matrix as follows:

$$C = \begin{bmatrix} x_{11}^i & x_{12}^i & ..... & x_{1L-1}^i & x_{1L}^i \\ x_{21}^i & x_{22}^i & ..... & x_{2L-1}^i & x_{2L}^i \\ & & & & \\ & & & & \\ x_{m1}^i & x_{m2}^i & ..... & x_{mL-1}^i & x_{mL}^i \end{bmatrix} \quad (2)$$

In this paper, each column of the matrix C as m-order polynomial coefficients structure a polynomial as follows:

$$f(x) = x_{1n}^i + x_{2n}^i + ... + x_{mn}^i \qquad n=1,2\ldots L \quad (3)$$

The polynomial is expanded on [-1,1] via the orthogonal Legendre polynomials as follows:

$$C_n = \frac{2n+1}{2}\int_{-1}^{1} f(x)P_n(x)dx \quad (4)$$

Where $P_n(x)$ denotes the Legendre polynomials and C_n denotes the expansion coefficient. All Legendre polynomials constitute a complete set of orthogonal function. To simplify Calculation, only six Legendre polynomials have been chosen to be the orthogonal bases. Although m is a variable, each m-order polynomials can be expanded to six coefficients. As a result, each matrix composed of m feature vectors for one state can be normalized to 6L coefficients, and L, which stands for the length of feature vector, is a constant. Then each state can be represented by the 6L coefficients as the normalization features.

2.4 Distortion

As one part of the input to ANN, distortion which derives from normalized state-segments represents the ratio of distance of the intra-class to the differentia of the extra-class. The distance of the intra-class is the distance between the normalized state-segment vectors of given speech and the model

of one emotion. The model is obtained from the set of normalized state-segment vectors in training speech. By execution of Linde-Buzo-Gray (LBG) VQ design algorithm, state-segment vectors of one emotion generates the codebook. The distortion is achieved by adding weight to distance, and the weight is the sum of distance from models of other emotions.

2.5 Emotion Recognition by ANN

Since ANN possesses excellent discriminate power and learning capabilities, the hybrid classification in this paper takes advantage of a one hidden layer and 9 hidden nodes net to classify emotions. The input of the ANN consists of distortions and likelihood probabilities, while the output is the assumed emotion.

3 Experiment

3.1 Speech Corpus

Because there is no standard in selecting the samples of emotion speech, two speech corpora from Berlin database of emotional speech and Beihang University Mandarin Chinese emotion database were used in the experiment. The experiment corpus from Berlin database of emotional speech covers Germany utterances of four emotions, ten texts and six actors, four males and two females. However, corpus from Beihang University Mandarin Chinese database contains Mandarin utterances of five emotions, fourteen texts and four actors, two males and two females. Twenty two utterances per emotion, including all texts and speakers are used for training, and the left utterances are samples to be classified in the experiments using Germany database. Meanwhile, for the Mandarin corpus, twenty six utterances per emotion, covering all texts and speakers are used for training, while the utterances left are objects to be recognized throughout the evaluation process.

3.2 Feature Set

Suited feature sequences used to recognize emotion not only should carry information of emotion, but also need to fit classification. The classifications in this paper are based on isolated HMMs and hybrid of HMMs and ANN. Thus, performances of different feature sets are compared in this paper. First, the experiments are performed by using Germany corpus, and three sets of features are proved to be relative preferable. Table1 lists the experimental results using isolated HMMs and the hybrid classification, all three sets containing subset of first and second derivative of pitch, first and second derivative of amplitude energy. Then, experiments implemented using Mandarin database shows that the optimal performance of isolated HMMs are also obtained from feature sequences set which contains 10-orders LPCC, first and second derivative of pitch, first and second derivative of amplitude energy and 12-orders MFCC.

Table1 recognition results by using different feature sets

	HMM	hybrid
Subset + LPCC	73.0	77.0
Subset + MFCC	75.1	79.2
subset + LPCC + MFCC	81.9	83.8

3.3 Results of Isolated HMMs

As mentioned in the paragraph above, distinct capacity of VQ codebook could result in different performance of modeling. The experimental results which are listed in figure 2 and figure 3 show that the relative optimal codebook capacity is 128 for Germany and 2048 for Mandarin.

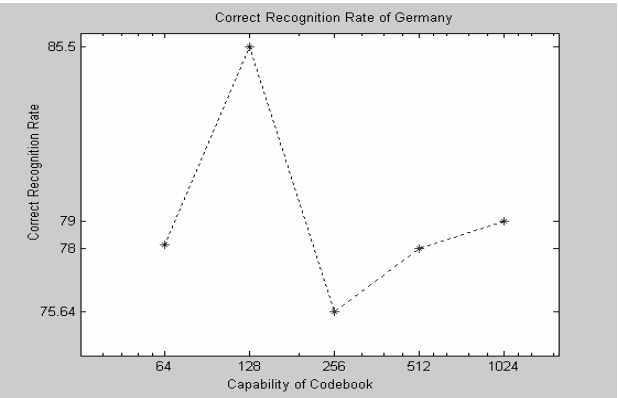

Figure2. Recognition rate based on 5-states HMMs, different capacity of codebook using Germany database.

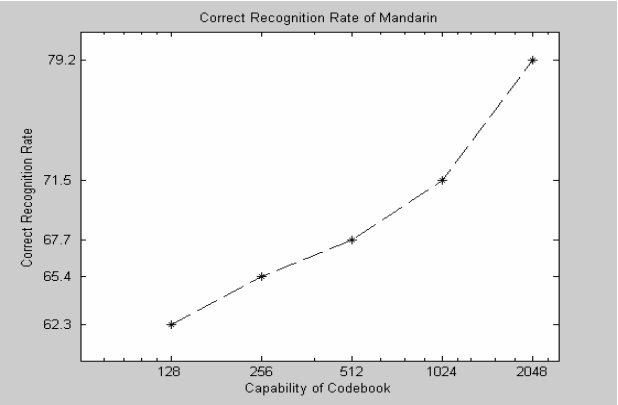

Figure3. Recognition rate based on 5-states HMMs, different capacity of codebook using Mandarin database.

According to the analysis above, the optimal recognition rates 81.9% obtained by using Germany database and 79.23% by Mandarin database are listed in detail in table 2 and table 3.

Table 2. Experimental Results using Germany database

emotion	Recognized emotion(%): 81.9 on average			
	anger	happiness	sadness	Disgust
anger	61.5	19.2	0	19.2
happiness	16.7	77.7	0	5.6
sadness	0	0	94.1	5.9
disgust	0	0	5.6	94.4

Table 3. Experimental results using Mandarin database

emotion	Recognized emotion(%): 79.2 on average				
	anger	happiness	sadness	disgust	surprise
anger	69.2	15.4	15.4	0	0
happiness	7.7	84.6	0	0	7.7
sadness	0	0	88.5	11.5	0
disgust	0	7.7	7.7	84.6	0
surprise	3.8	19.2	0	7.7	69.2

3.4 Results of Hybrid Based on HMMs/ANN

The hybrid classifier proposed in this paper proved to be more effective than the isolated HMMs. Recognition rates by means of the same speech corpora in the experiments of isolated HMMs are 83.8% and 81.6%. The experimental results are listed in detail in table 4 and table 5.

Table 4. Experimental results using Germany database

emotion	Recognized emotion(%): 83.8% on average			
	anger	happiness	sadness	disgust
anger	69.2	27.0	0	3.8
happiness	22.2	77.8	0	0
sadness	0	11.8	88.2	0
disgust	0	0	0	100

Table 5. Experimental Results using Mandarin database

emotion	Recognized emotion(%): 81.6 on average				
	anger	happiness	sadness	disgust	surprise
anger	61.6	23.1	3.8	7.7	3.8
happiness	3.8	92.3	0	3.8	0
sadness	0	0	100	0	0
disgust	0	7.7	11.5	73.1	7.7
surprise	7.7	11.5	0	0	80.8

4 Conclusion

In this paper, we have studied on emotion speech recognition by means of HMMs, and we believe that HMM makes significant impact on speech emotion recognition. Furthermore, a speech emotion recognizer that combines of HMMs and ANN has been proposed. Performances of the hybrid classification and isolated HMMs were collected by experiments using two speech corpora. Recognition by the hybrid classification has been proved more effective than isolated HMMs. Our future work will further explore the possibility to integrate other channels such as facial expression to increase the recognition rate.

5 Acknowledgement

This work is supported by High Technology Research and Development Program of China(863 Program, NO. 2006AA01Z135).

References:

[1] Kand B.S., Han C.H., Lee S.T.. Speaker dependent emotion recognition using speech signals. In Proc. ICSLP, 2000, pp.383-386.

[2] Schuller R., Rigoll G., Lang M.. Hidden Markov model-based speech emotion recognition, Porceeding of IEEE ICASSP Conference, Vol.2, 6-10, 2003, pp.1-4.

[3] YI-LIN LI, Gang Wei. Speech emotion recognition based on HMM and SVM, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Vol.8, 18-21 Aug. 2005, pp.4898 – 4901.

[4] Natascha Esau, Lisa Kleinjohann, Bernd Kleinjohann. Fuzzy emotion recognition in natural speech diaglogue. Robot and human interactive communication, 2005, pp.317-322.

[5] Dan-Ning Jiang, Lian-Hong Cai. Speech emotion classification with the combination of statistic features and temporal features. IEEE international conference on Mutinedia and Expo, 2004, pp.1968-1970.

[6] Rabiner, L.R., A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, vol.77, 1989, pp. 257-286.

[7] Tao J.H., Kang Y.G.. Features importance analysis for emotional speech classification, In Proceedings of lecture notes in computer science 3784 Springer, 2005, pp.449-457.

[8] Cowie R., Douglas-Cowie E.. Automatic statistical analysis of the signal and prosodic signs of emotion in speech, In Proc. 4th Int. Conf. Spoken Language Processing. Philadelphia, PA, 1996, pp.1989-1992.