INTERNATIONAL JOURNAL OF COMPUTERS
Issue 4, Volume 1, 2007

Mandarin speech emotion recognition based on a hybrid of HMM/ANN

Xia Mao, Lijiang Chen, Bing Zhang

***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. Optimal state sequences, exported from HMMs, are normalized to be one of the inputs of ANN; hence different methods of state normalization are compared. Adopting Beihang University Database of Emotional Speech (BHUDES), comparison between isolated HMMs and hybrid of HMMs/ANN proves that the classifier introduced in this paper is more effective, and the average recognition rate of five emotion states has reached 83.9%.

***Keywords*—**Speech emotion recognition; HMM; ANN; State normalization.

I. INTRODUCTION

Human computer interaction has been the focus of artificial intelligence research for several years now, and the research has moved ahead from the simple information exchange towards the affective communication. Introducing emotional intelligence to computers is an interesting yet difficult challenge that will bring about a positive revolution in the existing relationship between human and automated systems. Affective human computer interaction technology could be widely applied in virtual reality, especially in the field of entertainment and games. Moreover, the virtual human and psychiatric aid are the further application prospects for affective human computer interaction.

Making computer recognize the emotion of human being is the foundation of affective human computer interaction. The main carriers of human emotion, including facial expressions, posture and speech, are the primary channels for computer to recognize human's emotion. Speech is one of the most effective methods for people to communicate with each other. Emotion recognition of speech as a significant part of affective human computer interaction technology has become a challenge to speech processing. The accustomed way for speech emotion recognition is to distinguish the utterance between a defined set of discrete emotions. Manifold classifiers have been employed in this field. 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). The recognition rates of the most researches on a speaker-independent mode ranged from 55% to 95%, since even the recognition rates of human beings could hardly reach 60% when communicating with strange speakers [4].

HMM, with advantage on dynamic time warping capability has been long time studied for speech recognition. Moreover, it has been widely used 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.

The organization of this paper is as follows. Section II is dedicated to explain our system to recognize speech emotion, including five segments such as feature extraction, modeling emotion speech by HMMs, state normalization, distortion and emotion recognition by ANN. In section III, we design an experiment to check up the effect of our new approach compared with the results of isolated HMMs. Finally, we conclude with future directions of this research.

II. 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, which is used to classify emotions finally. Figure 1 illustrates the structure of the speech emotion recognition system developed in this paper.

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

Also this work is supported by the Specialized Research Fund for the Doctoral Program of Higher Education (20070006057).

 Manuscript Received June 23, 2007; Revised Received Dec. 22, 2007

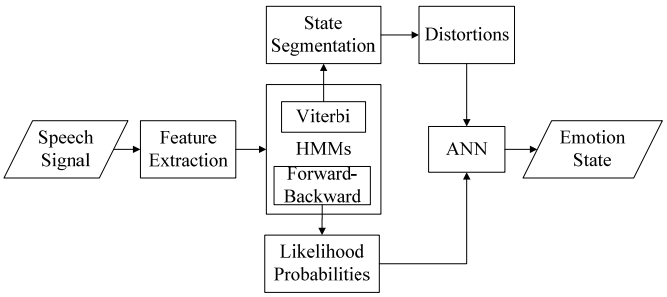

Fig.1. Speech Emotion Recognition System based on the hybrid of HMMs/ANN

A. Feature Extraction

To select suited features carrying information about emotion is necessary for emotion recognition. Studies on emotion of speech indicate 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. Each frame is 23ms and frame shift is 11ms. For each frame, six common features, including pitch, amplitude energy, logenergy, 10-order LPCC, 12-order MFCC and formant, are extracted. All of them form the candidate input feature sequences with their first and second derivatives.

B. Modeling Emotion Speech by HMMs

As shown in figure 1, 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 of 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.

C. 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 different methods to normalize the states.

1) State Normalization by Polynomials Expansion

Method of orthogonal polynomials expansion is employed 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:

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

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:

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

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

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

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

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

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 Lengendre 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) State Normalization by Statistical Method

Statistical features are computed to obtain fixed-length vectors. The statistics include mean, standard deviation, maximum, and so on. Isolated and combined statistics compose different normalized-states. Compared with the state normalization by polynomials expansion, the statistical method is more effortless to compute.

D. 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.

E. Emotion Recognition by ANN

Since ANN possesses excellent discriminate power and learning capabilities and represents implicit knowledge, the hybrid classification in this paper takes advantage of a one hidden layer and 11 hidden nodes net to classify emotions. The input of the ANN consists of distortions and likelihood probabilities, while the output is the assumed emotion.

III. EXPERIMENT

A. Speech Corpus

To evaluate the performance of the proposed classifier in this paper, Beihang University Database of Emotional Speech (BHUDES) was set up to provide speech samples. Besides a set of criterion of BHUDES, an emotional speech evaluation system was established to ensure the reliability of the speech samples. Emotional speech which was accurately recognized by at least 70 percent of the listeners was collected into the experiment corpus. This corpus contains Mandarin utterances of five emotions, twenty texts and five actors, two males and three females. 323 utterances covering all texts and speakers are used for training, while the utterances left are objects to be recognized throughout the evaluation process.

B. Results of Isolated HMMs

In this paper, a speech emotion recognition system based on isolated HMMs is used for comparison. Figure 2 illustrates this system, and likelihood probabilities exported from HMMs, the models of different emotional speeches, are compared to classify emotion states.

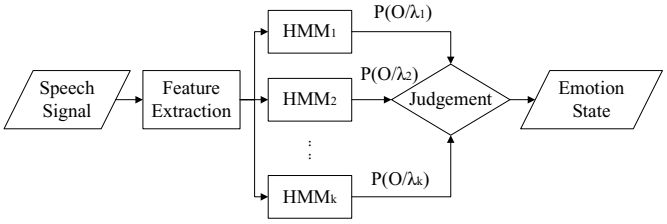

Fig.2. Speech Emotion Recognition System based on isolated HMMs

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. The experiments are performed by using BHUDES, and three sets of features are proved to be relative preferable. Table1 lists the experimental results using 5-states HMMs, all three sets containing subset of first and second derivative of pitch, first and second derivative of amplitude energy. Experiments shows that the optimal performance of isolated HMMs is 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
subset + LPCC	**73.0**
subset + MFCC	**75.1**
subset + LPCC + MFCC	**79.6**

The optimal experiment result deriving from isolated HMMs using suited feature set is shown in Tab.2

Table 2 Experimental results using Mandarin database

emotion	**Recognized emotion(%): 79.6 on average**				
	anger	happiness	sadness	disgust	**surpris**
anger	71.4	19.1	0	9.5	**0**
happiness	6.2	68.8	6.3	12.5	**6.2**
sadness	0	0	93.8	6.2	**0**
disgust	7.2	0	7.1	85.7	**0**
surprise	**8.7**	**8.7**	**0**	**4.3**	**78.3**

C. Results of Hybrid Based on HMMs/ANN

In this paper, we set up a hybrid of HMMs/ANN to classify emotion states of speech. Methods of state normalization are compared, and the experiment results are listed in Tab.3, Tab.4 and Tab.5. The results show that state normalization using statistical method is more effective.

Table 3 State normalization using polynomials expansion

emotion	**Recognized emotion(%): 81.6 on average**				
	anger	happiness	sadness	disgust	**surpris**
anger	76.2	14.3	0	9.5	**0**
happiness	12.5	68.8	6.2	6.2	**6.3**
sadness	0	0	87.5	12.5	**0**
disgust	0	0	7.1	92.9	**0**
surprise	**8.7**	**8.7**	**0**	**0**	**82.6**

Table 4 State normalization by statistics (maximum+mean)

emotion	**Recognized emotion(%):83.9 on average**				
	anger	happiness	sadness	disgust	**surpris**
anger	61.9	28.6	0	9.5	**0**
happiness	6.3	75.0	0	12.5	**6.2**
sadness	0	0	93.8	6.2	**0**
disgust	0	0	7.1	92.9	**0**
surprise	**4.3**	**0**	**0**	**0**	**95.7**

Table 5 State normalization by statistics (mean)

emotion	**Recognition result(%)**
mean	**81.9**
maximum	**82.5**
standard deviation	**80.5**
mean + maximum	**83.9**

IV. 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 BHUDES. State normalization by statistical method outperformed it by polynomials expansion. Still recognitions of the hybrid classification based on both state normalization methods have been proved more effective than isolated HMMs. Our future work is to explore the possibilities to integrate other channels such as facial expression to increase the recognition rate.

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..*

Xia Mao was born in Yiwu Zhejiang province China in 1952. She received her M.S. degree and Ph.D. degree from Saga University, Japan in 1993 and 1996 respectively.

She is currently a professor at School of Electronic and Information Engineering, Beihang University, Beijing, China. Her current research interests include affective computing, artificial intelligence, pattern recognition and Human-Computer Interaction. So far, she has published over 80 pieces of papers both domestically and overseas, many of them have been cited by the SCI, EI, ISTP etc.

Dr. Mao is leading several projects supported by the National High-tech Research and Development Program (863 Program), National Natural Science Foundation and Beijing Natural Science Foundation.

Lijiang Chen received the B.Sc. degree in Electronic and Information Engineering, Beihang University, Beijing, China, in 2007. He is currently pursuing the Ph.D. degree in Electronic and Information Engineering, Beihang University, Beijing, China

His main research interests include speech signal processing, pattern recognition and speech emotion recognition.