I.J. Education and Management Engineering **2012, 3, 25-31**
Published Online March 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2012.03.04

Available online at http://www.mecs-press.net/ijeme

Vocal Emotion Recognition Based on HMM and GMM for Mandarin Speech①

Sun Menghan[a,*1], Jiang Baochen[a,*2], Yuan Jing[a]

[a] *School of Mechanical, Electrical& Information Engineering, Shandong University at Weihai, Weihai, China*

Abstract

The recognition of emotions from speech is a challenging issue. In this paper, two Hidden Markov Model-based vocal emotion classifiers are trained and evaluated by an emotional mandarin speech corpus based on Mel-Frequency Cepstral Coefficient features. Up to 6 basic emotion models including angry, fear, happy, sad, neutral and surprise are built under different parameters and the influence of parameter set is investigated. A statistical comparison of the two emotion recognition methods are discussed as well. The overall results reveal that the GMM classifier outperforms HMM classifier taking both computation complexity and recognition rate into consideration with the highest recognition rate of 72.34%.

Index Terms: Speech Emotion Recognition; HMM; GMM

1. Introduction

Recognizing human emotion by computer has gained increasingly interests among researchers and industrial developers since it plays an essential role in efficient human-machine interaction. Recent neurological studies indicate that emotion recognition might enhance computers' ability to make decisions [1]. Compared with the internal physiology signals such as ECG, blood volume pressure, muscle voltage and respiration, vocal emotion expressions could be detected without any physical contact. Hence, emotion recognition from speeches is an active research area which has a wide range of applications such as intelligent robot, e-learning systems and hearing-impaired people assistant equipments [2-4].

Emotion is often portrayed differently in different cultures and language [3]. Mandarin is a tone language and has some significant differences from other languages, many researchers have studied emotional speech in mandarin and published their experiments results on vocal emotion recognition based on various features and different emotion classification approaches, which included some emotional speech samples as well. T.L. Pao

① This work is supported by the science and technology progress program of Shandong Province (2005GGA10111)

* Corresponding author:
E-mail address: *1 emiliesmh@163.com; *2 jbc@sdu.edu.cn

et al. worked with 20 sentences of anger, happiness, sadness, boredom and neutral expressed in mandarin language by 18 males and 16 females [4]. They proposed a weighted discrete K-nearest neighbor classification algorithm for detecting and evaluating emotion. The acoustic features they used are Mel-frequency Ceptral Coefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC). They reported the highest recognition rate is 79.55%. Y. Wang et al. built two emotional speech corpuses for men and women respectively including happiness, sadness, surprise, anger, fear and disgust six emotions [5]. They considered 23 prosodic features including amplitude, short-time energy, pitch, etc. and the derivatives of these parameters. They adopted Genetic algorithm (GA) for feature selection and support vector machine (SVM) combined with GA to recognize emotions. They reported to achieve an overall recognition rate of 88.15%. W.J. Han et al. studied on 1256 mandarin utterances of anger, happiness, sadness and surprise [6]. They took both prosodic features, such as energy and duration and acoustic features as the classifier's input based on artificial neural network (ANN) algorithm. They proposed VQ-based method for MFCC feature generation instead of the statistic method and demonstrated this new method is more effective which revealed a recognition accuracy of 71.1%. Ronald Bock et al. worked on Emo-DB database to compare different feature sets such as MFCCs, LPCs, and Perceptual Linear Predictive cepstral coefficients (PLP) for Hidden Markov Model (HMM) based speech emotion recognizer [7]. Using 3 states HMM model, they found that in case of naive emotions MFCCs provided the test performances.

The present paper investigates classifiers using HMM and GMM method based on MFCC features to detect emotion from our emotional speech corpus in mandarin. A statistical recognition performance comparison of these two methods will be discussed in this paper for 6 basic emotions namely angry, happy, fear, sad, neutral and surprise.

2. Features and Classifiers

The general process of vocal emotion recognition basically involves 2 stages; feature extraction and emotion classifier design. A critical problem of vocal emotion recognition system is the selection of feature sets from the speech utterances. MFCC features are proved to be effective to represent the acoustic characteristics of speech, which is widely used in speech recognition and speaker identification. It models the non-linear auditory response of the human ear that resolves frequencies on a log scale. The MFCC is the discrete cosine transform of the log-spectral energy coefficients of the speech segment. The spectral energy coefficients are calculated over Mel Filter Bank which consists of overlapping triangular filters. Hence, the Mel Filter Banks are logarithmically spaced filters with increasing bandwidths [6].

After getting the features of the emotional utterances, proper classification method should be adopted to detect emotions from speeches. In this work, two emotion recognition algorithms HMM and GMM are applied to classify the extracted features.

Hidden Markov Models (HMM) are statistical Markov models which assume the system to be a Markov process with unobserved states. This model method is widely used in image processing and speech recognition. In a hidden Markov model, the state is directly invisible but the output depending on state is visible. Each state has a probability distribution over the possible output tokens. The structure of the HMM generally adopted for vocal emotion recognition is a left-to-right structure. The probability of observed feature vector is calculated using the forward algorithm [8].

The Gaussian Mixture Model (GMM) is basically a single-state HMM with a Gaussian mixture observation density [11]. It is a popular likelihood ratio detector widely used in various fields of speech signal processing such as speech recognition and speaker verification. For the classifier based on GMM, it assumes that the observed feature vector is generated via a possibility density function (PDF) that is obtained by linearly combing a set of weighted Gaussian PDFs [9].

Both HMM models and GMM models are stochastic models which identify emotion based on comparison of likelihood probabilities exported from each emotion model. A schematic diagram for this process is shown in fig1.

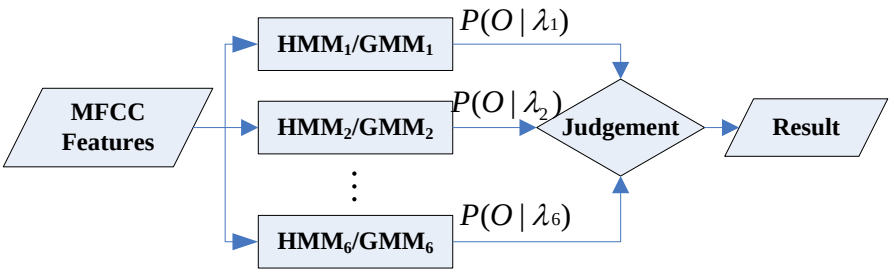

Fig. 1. Likelihood-based classification approach

Usually, the feature vectors are assumed independent, so the likelihood of a model λ for a sequence of feature vectors, $X = \{x_1, \cdots, x_T\}$, is computed as

$$\log p(\mathbf{X}|\lambda) = \sum_{t=1}^{T} \log p(\mathbf{x}_t|\lambda) \tag{1}$$

In particular, HMM/GMM is used to represents exactly one emotional state for emotion classifier and the model with the highest score determines the identified emotion [10].

3. Experiment and Method

The setup of HMM/GMM classifier is based on the specifications of HTK [11].

3.1. Emmotional Speech Corpus

To train the vocal emotion recognizer and evaluate its performance, a speech corpus from Chinese emotional speech database made by Institute of Automation, Chinese Academy of Science (CASIA) was used in the experiment. The experiment corpus covers 300 Chinese sentences with no emotion bias, each of which contains 3 to 13 characters. These sentences are spoken in six emotions (angry, happy, fear, sad, neutral and surprise) by 4 professional actors (2 males and 2 females), which is termed as simulated emotional utterances. The total 7200 utterances are divided into 2 groups, 4800 for training and the rest are objects used to be recognized throughout the evaluation process.

3.2. Emotion Recognition System

The functional components of the speech emotion recognition system are depicted in Fig.2. The system involves training phase and testing phase corresponding to training data and testing data respectively. In this experiment, MFCC feature vectors are derived from the emotional speeches and HMM/GMM algorithms are used to design emotion classifier. The percentage average recognition success score of each emotion as well as the overall recognition rate of all emotions are computed to evaluate the recognition performances then.

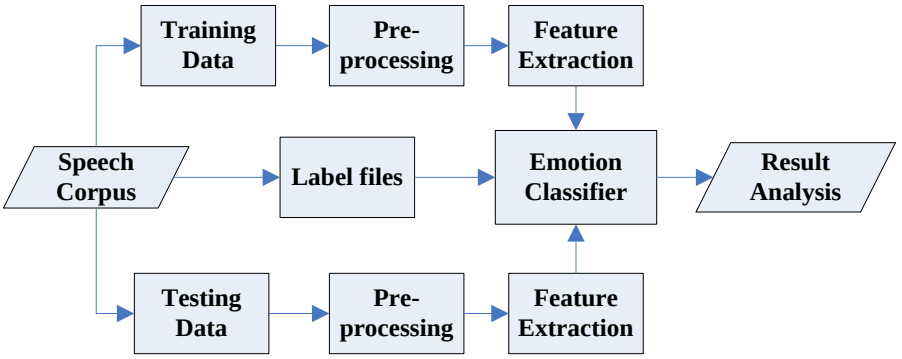

Fig. 2. Block of speech emotion recognition system

To evaluate the recognition performance, the speech utterances should be labeled to compare the detected emotion with the true emotion. The Master Label Files [11] format label files are generated automatically in VC++ 6.0 environment. One sample of the label file versus recognition result file of an emotional utterance is showed in Fig.3. Then, a statistical result analysis for the total 2400 emotional utterances of the test set will be given to compare the recognition performances of the two methods.

Lable file	*"./data/train/feature_H/426.lab"* *HAPPY*
Recogntion file	*"./data/train/feature_H/426.rec"* *0 47200000 HAPPY*

Fig. 3. Label file vs. recognition file for an utterance

3.3. Pre-processing and Feature Extraction

- At first, the speech samples are passed through a high pass filter ($1-0.97z^{-1}$) for pre-emphasis, which gives a spectral tilt to the speech samples. Then the speech signals are segmented into 25ms each frame with a frame shift of 10ms. Each frame is then multiplied by a Hanmming window.
- 12 MFCC, 12 delta MFCC, 12 delta-delta MFCC and a log energy features are computed from each frame using 22 triangular Mel-frequency filter banks. The MFCC features are normalized by cepstral mean subtraction. The total 37 dimensional feature vectors are used as the input of the emotion recognizer.

3.4. HMM/GMM Classifiers and Parameter Sets

In this work, two Hidden Markov Model-based methods (GMM and HMM) are used as the emotion classifiers. Since GMM may be treated as a single-state HMM with a Gaussian mixture observation density, the training and recognizing procedures of the two methods have a similar pattern with isolated words recognition using HMM. The dictionary built for the emotion recognition task is consisted of the six emotions and the grammar is defined as

$EMOTION = ANGRY|NEUTRAL|SAD|HAPPY|FEAR|SURPRISE;

([$EMOTION]) (2)

At first, the initial means and the elements of diagonal covariance instead of the full covariance matrices of the HMMs/GMMs are computed by K-Means algorithm for computation effectiveness. Then the basic Baum-Welch algorithm is used to re-estimate the three main elements of each single model including state transition probability, the output probability and the initial state distribution matrix. The trained classifier is tested with feature vectors extracted from the test utterances based on computation of the total log-likelihood using the probability distribution functions corresponding to each emotion-class. In this paper, up to 6 left-to-right HMM models have been built for the six basic emotions with state number $N \in \{2,4,6,10,14,20,32\}$. While the emotion classifier based on GMM method uses Gaussian mixtures with mixture numbers $M \in \{2,4,8,16,32,64,128\}$ to approximate the emission probability density functions of each emotion.

4. Results and Discussion

The percentage recognition rate for each emotion as well as the average percentage score for all six emotions based on HMM and GMM with different parameter sets are shown in TABLE I and TABLE II respectively. In this experiment, we investigated the influence of the number of emitting states on the classification accuracy based on HMM. For the case of GMM, the influence of Gaussian mixture number on the classifier's performance is studied. The average recognition for the both methods generally tend to increase with the increase of N and M respectively. However, as the computation complexity increases, the recognition accurate of HMM method does not improve so evidently as GMM method. This result may due to the MFCC features derived in utterance level in this experiment. Since GMM method has advantages in dealing with statistical features of acoustic parameters of emotional speeches while HMM is usually used with temporal features in phoneme level [12].

Table 1. Percentage Success Scores by HMM Classifier

N→ Emotion↓	**2**	**4**	**6**	**10**	**14**	**20**	**32**
Angry	48.67	53.00	61.67	71.00	65.33	60.33	70.67
Fear	22.00	47.33	48.67	39.33	38.67	40.33	44.67
Happy	1.33	26.33	28.33	24.00	26.00	34.00	31.67
Neutral	53.33	83.00	90.33	83.33	81.67	85.33	81.67
Sad	36.33	25.00	37.33	36.67	38.33	45.33	59.67
Surprise	43.67	71.67	71.67	57.67	68.00	72.33	67.67
Average	**34.22**	**51.06**	**56.33**	**52.00**	**53.00**	**56.28**	**59.34**

Table 2. Percentage Success Scores by HMM Classifier

M→ Emotion ↓	**2**	**4**	**8**	**16**	**32**	**64**	**128**
Angry	28.67	54.00	63.00	60.33	87.33	87.67	81.67
Fear	19.00	33.00	56.00	28.67	55.33	31.00	43.67
Happy	19.33	23.67	28.67	23.67	45.67	18.33	44.00
Neutral	84.33	86.00	80.00	81.67	75.00	85.67	95.67
Sad	17.00	26.33	46.33	66.67	67.67	56.33	77.33
Surprise	55.33	71.00	73.33	89.00	56.67	80.00	91.67
Average	**40.67**	**49.00**	**57.89**	**58.34**	**64.61**	**59.83**	**72.34**

The GMM classifier achieves the highest average percentage score up to 72.34% at M=128. While for the case of HMM, the highest overall recognition rate reaches 59.34% when the state number N=32. The recognition performances of the vocal emotion recognizer from mandarin speeches of this experiment may be acceptable, because some studies have proved that speech emotion recognition is a difficult task and even human are not perfect emotion recognizer [13][14].

From observations of the results presented in the tables, it is found that under the optimal parameters, surprise and neutral can be detected perfectly. However, fear and happy provide much lower recognition rates than other emotions, because the two emotions are uttered in a less expressive way.

More detailed performance of the confusion matrix are shown by TABLE III and TABLE IV under the parameters set M=8 and N=6.

Table 3. Confusion Matrix Table for HMM (M=6)

True→ Evaluated ↓	**Angry**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
Angry	62.67	3.67	9.33	3.33	4.33	13.00
Fear	3.00	41.33	1.67	1.67	18.67	5.33
Happy	0.00	1.67	30.00	9.33	2.33	2.33
Neutral	14.00	21.00	23.67	82.67	15.33	6.33
Sad	0.00	23.00	2.00	0.67	45.67	0.00
Surprise	20.33	9.33	33.33	2.33	13.67	73.00

Table 4. Confusion Matrix Table for GMM (M=8)

True→ Evaluated ↓	**Angry**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
Angry	63.00	3.00	4.33	0.33	0.00	7.67
Fear	0.67	56.00	4.00	8.67	17.67	6.00
Happy	1.67	8.67	28.67	3.33	7.00	1.00
Neutral	13.33	9.67	21.33	80.00	14.00	11.67
Sad	0.00	14.33	2.00	0.67	46.33	0.33
Surprise	21.33	8.33	39.67	7.00	15.00	73.33

These two classifiers have approximate identical overall recognition rate around 57%. The rows and the columns represent true and evaluated emotion categories respectively. It is observed that the confusion rates of the both emotion recognition methods perform similar tendency: angry and surprise, fear and sad are two emotion pairs that have relatively higher confusion rates than other pairs. It indicates that several emotional states may have the similar acoustic pattern because of similar physiological correlation. However, GMM method has an evident advantage over HMM method in recognizing the emotion fear.

5. Conclusions

In this paper, two HMM-based vocal emotion classifiers are discussed for emotional mandarin speech. After exhaustive experiments, it is found that both the classifiers with proper parameters can achieve much higher recognition rate than random decision. A general increase in performance can be observed using more Gaussian

mixtures for GMM method and more states for HMM method. The GMM emotion classifier outperforms HMM based on MFCC features taking both computation complexity and recognition rate into consideration. The two methods can detect surprise and neutral perfectly but confuse fear and sad, angry with surprise

However, our experiments are limited to the acoustic MFCC features, in further research, other features like prosodic features can be used as well to illustrate the temporal characteristics of each emotion. To achieve better recognition performance, combining classifier can be studied to recognize emotion from speech both on utterance and segment levels.

References

[1] R.W. Picard, "Affective computing," MIT Press, Cambridge, 1997
[2] W. Li, Y.H. Zhang and Y.Z. Fu, "Speech emotion recognition in E-learning system based on affective computing," Third Internetional Conference on Natural Computation, vol.5, pp.809-813, 2007.
[3] Nicholson, K. Takahashi and R.Nakatsu, "Emotion recognition in speech using neural networks," Neural Computing and Applications, vol.9, pp.290-296, December 2000.
[4] T.L. Pao, Y.T Chen, J.H. Yeh and Y.H. Chang, "Emotion recognition and evaluation of mandarin speech using weighted D-KNN classification," Master Thesis, Tatung University, 2005.
[5] Y.Wang, S.F. DU and Y.Z. Zhan, "Adaptive and optimal classification of speech emotion recognition," Fourth International Conference on Natural Computation, vol.5, pp.407-411, 2008.
[6] W. J. Han, H.F. Li and C.Y. Guo, "A hybrid speech emotion perception method of VQ-based feature processing and ANN recognition," WRI Global Congress on Intelligent Systems, vol.2, pp.145-149, 2009.
[7] R. Bock, D. Hubner, A. Wendemuth, "Determining optimal signal features and parameters for HMM-based emotion classification," 15th IEEE Mediterranean Electrotechnical Conference, pp.1586-1590, 2010.
[8] B. Schuller, G. Rigoll and M. Lang, "Hidden markov model-based speech emotion recognition," IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.2, pp.II-1-4, 2003.
[9] A. B. Kandali, A. Routray and T. K. Basu, "Emotion recognition from Assamese speeches using MFCC features and GMM classifier," 2008 IEEE Region 10 Conference, pp.1-5, 2008.
[10] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," Digital Signal Processing, vol.10, pp.19-41, 2000.
[11] S. Ser, C. Ling and L.Y. Zhu, "A hybrid PNN-GMM classification scheme for speech emotion recognition," 19th International Conference on Pattern Recognition, pp.1-4, 2008.
[12] S. Yong, G. Evermann, M. Gales, T. Hain and D. Kershaw, "The HTK Book," Cambridge University Engineering Department, 2006.
[13] D.N. Jiang and L.H. Cai, "Speech emotion reocognition using acoustic features," Tsinghua Univ (Sci & Tech), vol 46 No.1, pp.86-89, 2006.
[14] T.L. Pao, Y.T Chen, J.H. Yeh and P.J. Li, "Mandarin emotional speech recognition based on SVM and NN," 18th International Conference on Pattern Recognion," vol.1, pp.1096-1100, 2006.
[15] V.A. Petrushin, "Emotion recognition in speech signal: experimental study, development, and application," Sixth International Conference on Spoken Language Processing, pp.222-225, 2000.