SPEECH EMOTION RECOGNITION BASED ON HMM AND SVM

YI-LIN LIN, GANG WEI

School of Electronic and Information Engineering, South China University of Technology, Guangzhou 5106410, China
E-MAIL: linyilin2000@163.com, ecgwei@scut.edu.cn

Abstract:

Automatic emotion recognition in speech is a current research area with a wide range of applications in human-machine interactions. This paper uses two classification methods, the hidden Markov model (HMM) and the support vector machine (SVM), to classify five emotional states: anger, happiness, sadness, surprise and a neutral state. In the HMM method, 39 candidate instantaneous features were extracted, and the Sequential Forward Selection (SFS) method was used to find the best feature subset. The classification performance of the selected feature subset was then compared with that of the Mel frequency cepstrum coefficients (MFCC). Within the method based on SVM, a new vector measuring the difference between Mel frequency scale sub-bands energies is proposed. The performance of the K-nearest Neighbors (KNN) classifier using the proposed vector was also investigated. Both gender dependent and gender independent experiments were conducted on the Danish Emotional Speech (DES) Database. The recognition rates by the HMM classifier were 98.9% for female subjects, 100% for male subjects, and 99.5% for gender independent cases. When the SVM classifier and the proposed feature vector were employed, correct classification rates of 89.4%, 93.6% and 88.9% were obtained for male, female and gender independent cases respectively.

Keywords:

Emotion recognition; Hidden Markov Model; Support Vector Machine; Sequential Forward Selection; Mel energy spectrum dynamics coefficients

1. Introduction

Research has long been done on emotion in the fields of psychology and physiology. More recently it is the subject of attention by engineers. Its most important application is in intelligent human-machine interaction. In today's human-machine interaction systems, machines can recognize "what is said" and "who said it" using speech recognition and speaker identification techniques. If equipped with emotion recognition techniques, machines can also know "how it is said" to react more appropriately, and make the interaction more natural. Other applications of automatic emotion recognition include psychiatric diagnosis, intelligent toys, and lie detection [1].

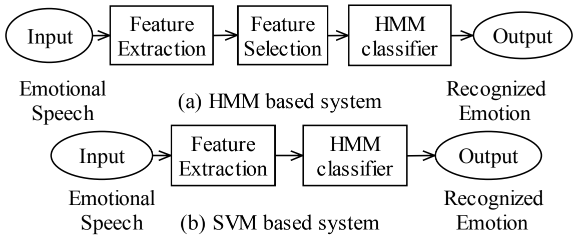
In recent years, a great deal of research has been done to automatically recognize emotions from human speech [1], [2], [3]. In Schuller *et al.* [2], two classification methods, GMM with global statistics and HMM with instantaneous features, were studied, but it was limited to features related to pitch and energy. The discriminating capability of 87 speech features was examined by Ververidis *et al.*, and correct classification rates of 61.1% for male, 57.1% for female and 50.6% for gender independent cases were obtained by employing a Bayes classifier with Gaussian pdfs [3]. The experiments were conducted on the Danish Emotional Speech (DES) Database [4] that expresses five emotional states.

In this paper, two classification methods, the hidden Markov model (HMM) and the support vector machine (SVM), were used to classify five emotional states: anger, happiness, sadness, surprise and a neutral state in which no distinct emotion is observed. The best feature vector with a dimension of five was determined from the 39 candidate instantaneous features before being input into the HMM classifier. That determination was made using the Sequential forward selection (SFS) method. For the SVM classifier, a novel feature vector that measures the difference between Mel scale sub-band energies was used. Classification experiments including gender dependent and gender independent cases were conducted on the DES database.

2. Speech emotion recognition system

The structure of the speech emotion recognition system studied in this paper is depicted in Figure 1. Like the typical pattern recognition system, it contains four main modules: emotional speech input, feature extraction, HMM/SVM based classification, and recognized emotion output. A feature selection module is part of the HMM classifier.

0-7803-9091-1/05/$20.00 ©2005 IEEE



(a) HMM based system

(b) SVM based system

Figure 1. Structure of the Speech Emotion Recognition System.

2.1. Feature extraction

It is believed that prosodic features are the primary indicator of speakers' emotional states. Research to analyze emotional speech indicates that fundamental frequency, energy and formant frequencies are potentially effective parameters to distinguish certain emotional states. In this paper, the five groups of short-term features that were extracted relate to fundamental frequency (F0), energy, the first four formant frequencies (F1 to F4), two Mel Frequency Cepstrum Coefficients (MFCC1, MFCC2) and five Mel frequency sub-band energies (MBE1 to MBE5). A Hamming window was used to segment each speech utterance into frames. The frame shift in feature extraction was 10ms. The fundamental frequency was extracted by the short-term autocorrelation method and then filtered by a 3-step median filter. The formants were estimated using the method in [5], and the two MFCCs were computed as described in [6]. To extract the sub-band energies, a magnitude spectrum of each frame was estimated by a fast Fourier transform (FFT), and then input to a bank of five equi-distant filters located in the Mel frequency scale between 60Hz to 7.6 kHz. The logarithm mean energies of the five filter outputs were calculated to get the five sub-band energies.

Thirty-nine features, including the five groups of features and their first and second derivatives, were extracted as the candidate input to the HMM classifier. Each of the extracted features was linearly scaled to the range of [0, 100] to avoid having values too large or too small.

Spectral energy dynamics provides another possible indicator of the speech emotional states. In this study, for the SVM classifier, a novel parameter vector called the Mel energy spectrum dynamics coefficients (MEDC) is proposed to distinguish the five emotional stats. It was extracted as follows: the magnitude spectrum of each speech utterance was estimated using FFT, then input to a bank of N filters equally spaced on the Mel frequency scale. The logarithm mean energies of the N filter outputs were calculated ($En(i), i=1,...,N$). Then, the first and second differences of $En(i), i=1,...,N$ were computed:

$$\Delta En(i) = En(i+1) - En(i), i=1,...,N-1 \qquad (1)$$

$$\Delta^2 En(j) = \Delta En(j+1) - \Delta En(j), j=1,...,N-2 \qquad (2)$$

The final Mel energy spectrum dynamics coefficients were then obtained by combining the first and second differences:

$$MFDC = [\Delta En(1)...\Delta En(N-1) \quad \Delta^2 En(1)...\Delta^2 En(N-2)] \qquad (3)$$

The value of N was set to 12 in this study, and the coefficients were linearly scaled to the range of [0, 1] before being input to the classifier.

2.2. Feature selection for HMM classifier

Usually, in HMM-based pattern recognizers or classifiers, the dimension of the input vector is pre-defined artificially. Although this study began without a precise idea of how of the 39-candidate features would be needed, using all of the extracted features in the input to the HMM classifier would not guarantee the best system performance and might cause unnecessarily high computational complexity. So, the SFS feature selection method was employed to select the best subset from the 39 features. First, SFS was initialized with the single best feature as determined by its having the maximal correct classification rate using the HMM classifier. When combined with the selected ones, subsequent features that had the maximal correct classification rate were added in turn. The selection of features stopped when adding new ones failed to increase the overall correct classification rate or when the number of the selected features reached a pre-set number.

2.3. Classifier selection

In this study, two different classifiers, the HMM based classifier and the SVM based classifier, were investigated. HMM [7] is a statistical signal model trained by sequences of feature vectors that are representative of the input signal. HMM has a long history in speech recognition. It has the important advantage that the temporal dynamics of speech features can be caught due to the presence of the state transition matrix. Previous study results indicate that some speech features, e.g. F0 and energy, have distinctive short time behavior in different emotional states. In this study, the temporal dynamics modeling capacity of HMM is utilized to discriminate different speech emotions.

SVM [8] is a newer technique for data classification and regression by Vapnik and Chervonenkis derived from statistical learning theory in 1990s. Its main idea is to transform the original input set to a high-dimensional feature space by using a kernel function, and then achieve optimum classification in this new feature space. Many kinds of implementation of SVM can be found in the Internet. The LIBSVM [9] was used in this study.

3. Experimental study

3.1. Emotional speech database

The emotional speech database used in this study is the Danish Emotional Speech (DES) database [4], which includes expressions by four actors familiar with radio theatre, two male and two female. The speech is expressed in five emotional states: anger, happiness, neutral, sadness and surprise. In this study, the speech signal was re-sampled to 16 kHz, and the silence segments at the beginning and the end of the speech were cut out artificially. Then the whole database was divided into four parts for the purpose of cross-validation.

3.2. Experimental results using the HMM based system and discussion

First feature selection was conducted. The selection criterion was the correct classification rates by HMM classifiers, which were estimated by four fold leave-one-out (LOO) cross-validation. Each time one part of the speech database was left for validation and the remainders were used for training the classifier. The five best feature sequences that were selected using this procedure are:

1) the second derivative of F0,
2) the first derivative of F0,
3) the second derivative of MBE5,
4) the second derivative of MBE4, and
5) the first derivative of F1.

After the selection procedure, both gender dependent and gender independent experiments were conducted. Each hidden Markov model employed in these experiments had five states and the observation probability distribution in each state was a mixture of four full covariance Gaussian distributions. A best recognition rate of 100% was obtained when only the emotional speech from male subjects was considered. An accuracy of 98.9% was achieved for female subjects, and in the gender independent case the average correct classification rate of the five emotional states is 99.5%. Recognition accuracy was gained by four-fold leave-one-out cross-validation. Experimental results are shown in Table 1, 2 and 3. In the confusion matrices shown in Table 1 and 2, the columns show the emotions that the speakers tried to induce, and the rows are the output recognized emotions. *N*eu is the abbreviation for neutral, *Ang* for anger, *Hap* for happiness, *Sad* for sadness, and *Sur* for surprise.

Table 1. Confusion matrix of the HMM classifier (Female)

Stimulus	Recognized Emotions (%)				
	Neu	*Ang*	*Hap*	*Sad*	*Sur*
Neu	100	0	0	0	0
Ang	0	100	0	0	0
Hap	0	5.3	94.7	0	0
Sad	0	0	0	100	0
Sur	0	0	0	0	100

The recognition performance of a vector of five MFCC coefficients was also investigated in comparison to that of the selected features. The experimental results are shown in Table 3. The length of the MFCC vector was the same as that of the selected features. Correct classification rates of less than 60% were achieved in these experiments showing that while MFCC coefficients are popular features in speech recognition, they are not suitable for emotion recognition in speech.

Both temporal dynamics and spectral characteristics were taken into consideration by combining HMM with the selected features in analyzing the emotions expressed in the speech. In addition, the speaking rate was also considered by using HMM, producing comparatively good classification results. The experimental results confirm that the temporal dynamics of the fundamental frequency and Mel sub-band energies are important indicators of the emotional content of the speech.

Table 2. Confusion matrix of the HMM classifier (Gender independent)

Stimulus	Recognized Emotions (%)				
	Neu	*Ang*	*Hap*	*Sad*	*Sur*
Neu	100	0	0	0	0
Ang	0	100	0	0	0
Hap	0	2.6	97.4	0	0
Sad	0	0	0	100	0
Sur	0	0	0	0	100

3.3. Experimental results using the SVM based system and discussion

The SVM classifier was used to test the proposed feature vector of MEDC. The radial basis function (RBF) kernel was used in the SVM classifier, and experimental results were obtained by four-fold cross-validation. Both gender dependent and gender independent experiments were performed. The best recognition rate of 93.7% was obtained when only emotional speech from female subjects was considered. For male subjects the correct classification rate was 89.4%. An accuracy of 88.9% was obtained in gender independent case. The experimental results are shown in Table 4. Confusion matrices are shown in Table 5, 6 and 7.

Table 3. HMM recognition results using the selected features and MFCC

	Female	Male	Gender independent
Selected Features	98.9%	100%	99.5%
MFCC	57.9%	41.6%	59.5%

Performance of a K-nearest Neighbors (KNN) based classifier with the K value set to 21 was investigated (see Table 4). This classifier gave recognition rates around 83% and was outperformed by the SVM classifier.

Table 4. Experimental results using MFDC

	Female	Male	Gender independent
SVM	93.7%	89.4%	88.9%
KNN (K=21)	83.2%	85.8%	84.5%

Table 5. Confusion matrix of the SVM classifier (Female)

Stimulus	Recognized Emotions (%)				
	Neu	*Ang*	*Hap*	*Sad*	*Sur*
Neu	95	2.5	0	0	2.5
Ang	0	91.9	0	8.1	0
Hap	5.3	2.5	92.2	0	0
Sad	0	2.5	0	97.5	0
Sur	5.6	0	2.8	0	91.7

Table 6. Confusion matrix of the SVM classifier (Male)

Stimulus	Recognized Emotions (%)				
	Neu	*Ang*	*Hap*	*Sad*	*Sur*
Neu	84.2	0	0	13.3	2.5
Ang	2.8	86.9	0	7.8	2.5
Hap	7.8	0	86.7	5.3	5.3
Sad	0	0	0	1	0
Sur	2.5	2.8	2.8	2.8	89.2

Table 7. Confusion matrix of the SVM classifier (Gender independent)

Stimulus	Recognized Emotions (%)				
	Neu	*Ang*	*Hap*	*Sad*	*Sur*
Neu	89.5	3.9	0	6.6	0
Ang	3.9	90.8	0	5.3	0
Hap	6.6	2.6	82.9	3.9	3.9
Sad	0	3.9	0	96.1	0
Sur	5.3	2.6	5.3	1.3	85.5

By avoiding using the fundamental frequency values, values that affected greatly by the gender and age of the speaker, no significant difference in the recognition performance was found between gender dependant and gender independent cases. However, the effect of the fundamental frequencies was still reflected by the low band dynamics of the Mel sub-band energies.

4. Conclusions

In this paper two speech emotion recognition systems were studied. Features based on the fundamental frequency, energy, formants, and Mel sub-band energies were extracted as the candidate input to the HMM classifier. The best feature set was selected by the SFS method. In the SVM based system, the Mel energy spectrum dynamics coefficients were utilized to classify the five emotional states. Both systems obtained relatively high accuracy in classifying the five emotional states expressed by the DES database.

References

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G., “Emotion recognition in human-computer interaction”, IEEE Signal Processing magazine, Vol. 18, No. 1, pp. 32-80, Jan. 2001.

[2] B. Schuller, G. Rigoll, M. Lang, “Hidden Markov model-based speech emotion recognition”, Proceedings of the IEEE ICASSP Conference, Hong Kong, pp. 1~4, April 2003.

[3] D. Ververidis, and C. Kotropoulos, “Automatic speech classification to five emotional states based on gender information”, Proceedings of the EUSIPCO2004 Conference, Austria, pp. 341-344, Sept. 2004.

[4] Inger. S. Engberg, Anya. V. Hansen, “Documentation of the Danish Emotional Speech Database DES”, Aalborg, Sept 1996.

[5] L. Arslan, Speech toolbox in MATLAB, Bogazici University, http://www.busim.ee.boun.edu.tr/~arslan/

[6] S. B. Davis, and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, IEEE Transaction on ASSP, Vol. 28, No. 4, pp. 357-366, Aug. 1980.

[7] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition”, Proceedings of the IEEE, Vol. 77, No. 2, pp. 257-286, Feb 1989.

[8] Christopher. J. C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, 2(2):955-974, Kluwer Academic Publishers, Boston, 1998.

[9] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm