Proceedings of the 5th WSEAS International Conference on Signal Processing, Istanbul, Turkey, May 27-29, 2006 (pp1-5)

Spectral Features Detection of Speech Emotion and Speaking Styles Recognition Based on HMM Classifier

MONIA KAMMOUN and **NOUREDDINE ELLOUZE**

Laboratoire de Systèmes de Traitement de Signal, Ecole Nationale d'Ingénieurs de Tunis

BP.37 Le Belvédère, 1002, Tunis, Tunisie

Abstract: -This paper deals with the influence of the spectral features in recognizing emotions and speaking styles from speech signals. Through out this study MFCC and Mel band energies are used as the base features. The investigation shows that each feature mentioned has an important impact in recognizing several emotions and speaking styles with a satisfying recognition rate. The best recognition accuracy is reached with Mel band energies attaining 79% for the slow speaking style among 10 different states to be recognized by the system. These results can be significantly enhanced by combining both features. The aim of this work is to identify which of the speaking state will be better recognized with MFCCs or Mel band energies. For this approach, we use the HMM classifier. Results are given on text-independent emotion recognition using SUSAS database (Speech Under Simulated and Actual Stress). We compare emotion recognition performance based on the features mentioned. Experimental results show that stressed and loud styles are better recognized using MFCC features than those of Mel Band energies, by more than 12.5% which is an important improvement. The average recognition accuracy of ten different emotions and speaking styles with HMM models exceeds 60 % using the spectral features. We achieve with the text-independent SUSAS database 62% average accuracy for 10 emotions and speaking styles recognition using the MFCCs features.

Key-words: -MFCC, Mel Band energies, HMM-Based classifier, SUSAS database, Emotion, Speaking Style.

1 Introduction

The motivation for the recognition of emotions and speaking styles in speech comes from the increased role spoken dialog systems are playing in human-machine interaction, especially for deployment of services associated with call centers such as customer care, and for a variety of automatic training and educational applications. For this reason automatic emotion recognition in speech has recently received wide attention. Three different aspects can be easily identified: Speech recognition in the presence of emotional speech, Synthesis of emotional speech, and emotion recognition. In this last case, the objective is to determine the emotional state of the speaker out the speech samples. Possible applications vary from psychiatric diagnosis, to intelligent toys, and are a subject of recent but rapidly growing interest [7]. The performance of an emotion classifier relies heavily on the qualities of the training and testing data. In order to carry out the experimentation, we use the SUSAS database. In this paper, we investigate the classification of emotional information contained in human speech signals. The emotion states explored for this case are neutral, stressed, loud, Lombard, soft, clear, slow, question, angry and fast. The specific focus of this study is to explore the role of spectral features on emotion recognition using Hidden Markov Models. The remainder of this paper is organized as follows. Section 2 illustrates emotion recognition by speech. The description of the Database used is presented in section 3. Section 4 discusses the HMM-based classifiers usually employed for emotion recognition. The framework and the results from HMM approach are presented in section 5 with a recapitulation of the results obtained. Finally, section 6 summarizes the paper and provides conclusions.

2 Emotion Recognition by Speech

Recognizing emotions in speech primarily deals with the search for acoustic components of speech that distinguish a number of emotional states. This topic has received increasing attention during the last ten years. There are several reasons for speech recognition emotion interest such as: technological progress in recording, storing, and processing audio and visual information; the development of non-intrusive sensors; the advent of wearable computers; the urge to enrich human-computer interface from point-and-click to sense-and-feel; and the invasion on our computers of lifelike agents who supposed to be able express, have and understand emotions [2]. Several studies show a high correlation

between some statistical measures of speech and the emotional state of the speaker [3, 4, 5, and 6]. Most of the efforts done so far in emotion recognition are based on determining the sources of information and how can we deal with them. Previous research studies in the field of emotion recognition were focused basically on determining the human state in order to improve the quality of some recognizing systems [1, 11, 7, 8 and 12]. Hidden Markov Models have been used for a long time in speech recognition. The underlying idea is that the voice is not stationary. Instead, voice is modeled as a concatenation of states, which have their own properties. There are two main advantages of using HMM for emotion recognition: first, the structure of HMM may be useful to catch the temporal behavior of speech; second, procedures using HMM for optimizing the recognition framework are well established, since this technology has been applied for a long time for speech recognition purposes. The models can characterize the emotional data and can also reproduce unlimited observation sequences with the same statical properties as the training data, this characteristic is due to the regenerative property of HMM. Since HMM can regenerate a large number of observation sequences.

3 Database Descriptions

The speech data used in this study is a part of the SUSAS database [8]. SUSAS represents a comprehensive speech under stress database which was formulated and collected by the Robust Speech Processing Laboratory at Duke University under the direction of Professor John H. L. Hansen and sponsored by the Air Force Research Laboratory. The database is partitioned into five domains, encompassing a wide variety of stresses and emotions. A total of 32 speakers (13 female, 19 male) with ages ranging from 22 to 76 were employed to generate in excess of 16,000 utterances. The five stress domains include: i) talking styles (slow, fast, soft, loud, angry, clear, question), ii) single tracking task or speech produced in noise (Lombard effect), iii) dual tracking computer response task, iv) actual subject motion-fear tasks (G-force, Lombard effect, noise, fear), v) psychiatric analysis data (speech under depression, fear, anxiety). A common highly confusable vocabulary set of 35 aircraft communication words make up the data base. Simulated speech under stress data consists of data from ten stressed styles (talking styles, single tracking task and Lombard effect domains); while speaking in a noisy environment (i) dual-tracking workload computer tasks, or (ii) subject motion-fear tasks (subjects in roller-coaster rides). Additional speech was also later added from pilots in Apache helicopter flight conditions. Two of the domains employ computer response tasks, and one domain employs entertainment park roller-coaster rides as speaker tasks. The four domains available in the present release of SUSAS consist of a 35 air-craft communication word vocabulary. Subsets include {go, hello, oh, no}, {six, fix}, {white, wide, point}, {degree, three, thirty, freeze}, and {eight, eighty, gain, change}.

4 HMM-Based Classifications

The first problem that arises when trying to build an HMM based recognition framework is the selection of the features to be used. In addition to carrying information about the emotional state, the feature must fit HMM structure. The main consequence of this limitation is that the used features must model the short time behavior of voice. The goal of this study is to explicitly model the spectral information at a local (segmental) level for categorizing emotions. We considered the following possible features: 12 MFCC, 5 MELBAND energies and log energy their velocity and acceleration. They have been widely used in speech recognition because of superior performance than other features. The cepstrum-related spectral features have also been found to be useful in the classification of stress in speech [13]. In this study, the classification of 10 different emotional states and speaking styles - anger, Lombard, loud, slow, soft, clear, fast, question, neutral, and stress - is implemented in the framework of the HMM classifiers, which help model dynamic changes in the emotional state dependent features in an utterance. Previous studies in emotion recognition have used HMM classifiers based on prosodic features to capture the dynamics of expressed emotions [14, 15].

4.1 Base feature set

Log energy, band energies and Mel frequency cepstral coefficients (MFCCs) are selected as the base features; we treat separately the emotion and speaking style recognition with the MFCC feature vector and the band energies feature vector. The frame shift in feature extraction is 10ms. We first segmented only speech parts from an input utterance by using an endpoint detector based on zero crossing rates (ZCR) and frame energy. For each frame of speech signals, we estimate log energy, five Mel band energies, and 12 coefficients MFCC. We also add velocity and acceleration information for MFCC, Mel band energies and log energy respectively, to take the rate of speaking into account and model the dynamics of corresponding temporal change of energy and spectrum. Hence we have 39 streams of features including velocity and acceleration components for MFCC and 18 streams of features for Mel band. These streams are used as input

vectors of HMM based classifiers. For our approach, MFCC coefficients are primarily used separately with Mel band energies.

4.2 Training and testing stage

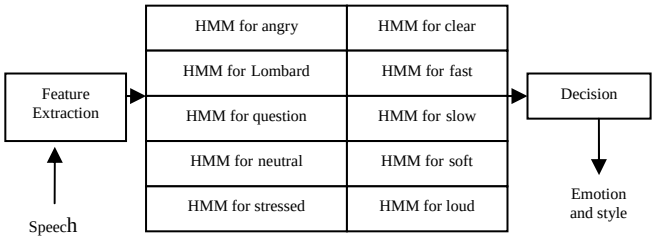

Fig.1: Recognition system

Our proposal consists in modeling short time spectral features with semi-continuous HMM. One HMM is used per class. As spectral features we study the performance of the two features set described in section 3. Each emotion and speaking style is modeled by 5-state HMM with the observation probability distribution of a single full covariance Gaussian distribution. For training, we use 420 words of each emotion and speaking style; the 10 styles are mixed to obtain 4200 words in the training base. For the testing base, 140 words of each style represent the testing data.

5 Experimental Framework and Results

To evaluate performance of the proposed feature extraction method, we chose the SUSAS database designed originally for speech recognition under stress. We used isolated words recorded at 8 kHz sampling rate in various emotions and speaking styles the analysis is performed on samples of 25 ms taken every 10 ms [9]. Hidden Markov Toolkit (HTK) is used [10], to test the performance of the HMM-Based classifier. The input is the 5 Mel band energies with log energy. In order to model the instantaneous values of energy with relying on the absolute value of energy, we also use the first and second derivates of the logarithm of the mean energy in the frame. The acoustical meaning of these measures is related to the sharpness of the energy level, reflecting both the articulation speed and the dynamic range. Then we apply the same procedure for 12 MFCC. Velocity (D) and acceleration (A) are added for all the features. Results are discussed and compared for both of Mel band energies and MFCCs.

5.1 Experimental results with 5 Mel band energies features

%	N	Sd	Ld	Lb	St	C	Sw	Q	A	F
N	70,2	2,8	0,71	3,7	5,89	9,45	1,3	0,92	2,2	2,85
Sd	7,16	60	3,28	3,1	7,86	8,11	0,2	4,41	2,7	3,49
Ld	0,86	2,3	65,2	8,3	0,88	1,25	0,1	3,49	13	4,6
Lb	4,81	5,9	7,52	63	0,54	6,29	0,6	4,68	5,7	1,18
St	12,6	6,1	0,26	0,4	57,6	8,27	9,1	1,37	1,2	3,13
C	6,26	7,7	1,08	5,3	8,61	60,2	1,2	2,54	6,9	0,26
Sw	2,45	0,4	0,04	4,7	10,7	1,82	79	0,59	0,1	0,44
Q	2,17	4,6	2,85	7	1,79	3,64	0,5	50	8,1	19,4
A	1,85	4,6	14,5	5	2,19	9,08	0,3	10,7	46	6,13
F	2,18	3,9	4,12	1,6	5,02	1,16	0,1	17,3	5,7	59

Table 1: confusion matrix of 5 Mel band energies

N: neutral, Sd: stressed, Ld: loud, Lb: Lombard, C: clear, Sw: slow, St: soft, Q: question, A: angry, F: fast.
The accuracies of 10 emotions and speaking styles are mentioned in table1. Results are given for 5Mel band energies, log energy, 6 deltas (D), and 6 accelerations (A). The first interpretation is that the accuracy is higher than 46%; and the best accuracy concerns the slow style and reaches 79%. The neutral style is more confusable with the clear and soft styles; we can also notice that the question and fast styles have a high level error reaching 19.4% and 17.3%, same thing for the loud and angry styles with 14.5% error level. Question speaking style and angry emotion have the worst accuracies. We can also notice that recognizing neutral style produces confusion error of 12.6% with soft style; this error is cut by half when it comes to recognizing soft style from the neutral style 5.89%, and this is due to choices of training data which improve significantly the recognizing system. This is also true for stressed (7.16% to 2.8%), slow (2.45% to 1.3%) and question (2.17% to 0.92%) for which confusing errors decrease when using neutral style as the style to recognize. The smallest confusing error rate concerns the slow and loud styles which reaching 0.04%, however the slow style is confused with soft style with 10.7% error.

5.2 Experimental results with 12 MFCC features

%	N	Sd	Ld	Lb	St	C	Sw	Q	A	F
N	**69,2**	3,81	0,8	2,9	4	12,8	1,12	1,1	2,85	1,52
Sd	2,13	**73,2**	2,2	6,9	6	3,09	0,16	2	3,45	1,23
Ld	1,88	0,68	**73**	6	1	2,28	0,08	3,2	11,7	0,03
Lb	7,93	1,71	4,3	**63**	11	2,23	1,09	2,6	5,87	0,13
St	11,7	5,17	0,2	0,1	**58**	14,7	4,83	2	1,16	2,36
C	10,2	1,79	3,7	5	12	**55,3**	0,87	3,5	5,92	1,51
Sw	2,11	1,93	1	5,7	9	2,73	**77,1**	0,8	0,03	0,02
Q	1,99	4,09	3,1	7	2	2,72	0,13	**44**	11,9	23,6
A	2,03	4,32	17	5	1	9,54	0,37	11	**45,4**	4,16
F	1,53	2,95	12	1,1	3	3,71	0,12	21	5,09	**50,4**

Table 2: confusion matrix of 12MFCC

In this part, MFCC features are selected in order to determinate emotion in speech signal. The input utterance is a 39 variable length feature streams consisting on 12 MFCC, log energy, 13 deltas (D) and 13 accelerations (A). The accuracies of 10 emotions and speaking styles are illustrated in table2. Our results also indicate that slow speaking style is always better recognized in comparison with the other emotions and styles when working with MFCC features. In the case of (12MFCC_E_D_A), all obtained results are normally high than 44%. The best result reached concerns the slow style with recognition rate of 77.1%. The clear style was most confusing with the neutral and soft styles. The question style was most confusing with the fast style and the stressed emotion was most confusing with the Lombard style. We can notice that recognizing neutral style reduces the error in confusing stressed, loud, Lombard, soft and slow speaking styles. However, the confusing error is very important 12.8% for clear style when recognizing neutral style. Recognizing the remaining speaking styles other than the neutral style yields errors which are not variable when changing testing and training data. When comparing these results with the previous one obtained with the Mel band energies features, we can notice that the accuracies of both stressed and loud styles enhance when using the MFCC features. This growth reaches 13.2% for the stressed style. Whereas, the recognition rates of the 6 styles: fast, angry, question, slow, clear, and neutral decrease with more than 1%, this fall attains 8.6% for the fast style. The accuracies of both Lombard and soft styles remain the same.

5.3 Results Recapitulation

Experimental results show that the best average accuracy of the HMM based classifier for 10 class style classification with the 12 MFCC features, log energy their derivate and acceleration is 62%. Table 3 gives the average accuracies for the whole speaking styles.

	MFCC	**Mel Band energies**
Average rate	62%	61%

Table 3: recognition rate average

Table3 shows the average rate for all the combinations of spectral features; the best rate concerns 12 MFCC and reached 62%. However, the Mel band energies features recognition rate attains 61%, which is almost close to the MFCC results. These results are satisfying with regard to the number of emotions and styles recognized.

6 Conclusion

In this paper, an HMM based approach to emotions and speaking styles recognition is presented. For performance evaluation of the proposed feature extraction method, we use the SUSAS database. We evaluate classification accuracy for 10 classes: Neutral, stressed, loud, Lombard, soft, clear, slow, question, angry and fast. Results show an average accuracy higher than 60 % in the text-independent SUSAS database recognition of 10 different emotional and speaking styles. We analyzed the effects of the features in recognizing emotions and speaking styles; MFCC and Mel band energies represent spectral features used in the experimental environment. The study shows that when using the MFCC features the recognition rate improves by more than 12% for the stressed and the loud styles compared with the Mel band features, it was also demonstrated that the proper choice of training data improves significantly the recognizing system. The best accuracy reached concerns the slow style with a 79% recognition rate combining 5 Mel band energies with both log energy and amplitude normalization. This work has proved that in all cases spectral features are useful for recognizing emotions and speaking styles. HMM classifier provided significant classification accuracy improvement for 10 emotions and speaking styles. Previous studies use a maximum of 5 speaking styles [1, 11, and 12]. This work also compares the performance of MFCC features to Mel band energies features. MFCC features are more robust to emotions and speaking styles than Mel Band energies features with 62% average accuracy rate. However, the Mel band energies features tend to be better than MFCC in recognizing fast, angry, question, slow, clear, and neutral styles. We can deduce that the combination of the two different spectral features leads to better performance of the proposed recognition system. The mixture of these features allows the enhancement of the accuracies. The aim of this work is to reach a recognizing system working in real environment with the maximum of emotions and talking styles. Several important issues remain to be addressed, we should investigate better pattern recognition techniques since there is no clear-cut definition of emotion states/categories or their acoustic correlates. Therefore, pattern classification methods that can deal with that uncertainty in the emotion states should be studied and developed. Such studies should also adopt a principled way for incorporating linguistic and dialog information in emotion recognition. Further study is needed to explore new features better representing prosody and timbre such as pitch and formant to reach a better performance.

References:

[1] V. A. Petrushin, "Emotion recognition agents in real world," AIII, Fall Symposium on socially Intelligent Agents: Human in the Loop, 2000.

[2] Elliot, C., Brzezinski, J. "Autonomous Agents as Synthetic Characters," AI Magazine, 19:13-30, 1998.

[3] B. Heuft, T. Portele, and M. Rauth, "Emotions in time domain synthesis," in Proc. of ICSLP, Philadelphia, pp. 1974-1977, Oct. 1996.

[4] R. Cowie and E. Douglas-Cowie, "Automatic statical analysis of the signal and prosodic signs of emotion in speech," in Proc. of ICSLP, Philadelphia, pp. 1989-1998, Dec. 1998.

[5] N. Amir and S. Ron, "Towards an automatic classification of emotion in speech," in Proc. of ICSLP, Sydney, pp. 555-558, Dec. 1998.

[6] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "Acoustic nature and perceptual testing of corpora of emotional speech," in Proc. of ICSLP, Sydney, pp. 225-228, Dec. 1998.

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

[8] J.H.L. Hansen, "Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition," Ph.D. Thesis, Georgia Inst. of Tech., Atlanta, GA, 428 pgs., July 1988.

[9] G. Zhou et al., 'Nonlinear feature based classification of speech under stress," IEEE Trans. Speech, Audio Proc., vol.9, pp.201-216, Mar.2001.

[10] S. Young et al., the HTK Book, 2001.

[11] C. M. Lee and S. Narayanan, "Emotion recognition using a data-driven fuzzy interference system," EUROSPEECH, Geneva, 2003.

[12] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, "Emotion recognition by speech signals," EUROSPEECH, Geneva, 2003.

[13] J. Hansen and B. Womack, “Feature analysis and neural network based classification of speech under stress,” IEEE Trans. On Speech & Audio Processing, vol. 4(4), pp. 307–313, Jul 1996.

[14] C. Lee, “Speaker dependent emotion recognition using speech signals,” in Proc. ICSLP 2000, Beijing, China, Oct 2002.

[15] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Marino, “Speech emotion recognition using hidden markov models,” in Eurospeech’01, 2001.