Recognition of Emotional states in Speech using Hidden Markov Model

HMM을 이용한 음성에서의 감정인식

Sung-ill Kim, Sang-hoon Lee, Wee-jae Shin, Nam-chun Park
Division of Electrical and Electronic Engineering,
College of Engineering, Kyungnam University
E-mail : kimstar@kyungnam.ac.kr

요 약

본 논문은 분노, 행복, 평정, 슬픔, 놀람 등과 같은 인간의 감정상태를 인식하는 새로운 접근에 대해 설명한다. 이러한 시도는 이산길이를 포함하는 연속 은닉 마르코프 모델(HMM)을 사용함으로써 이루어진다. 이를 위해, 우선 입력음성신호로부터 감정의 특징 파라메타를 정의한다. 본 연구에서는 피치 신호, 에너지, 그리고 각각의 미분계수 등의 운율 파라메타를 사용하고, HMM으로 훈련과정을 거친다. 또한, 화자적응을 위해서 최대 사후확률(MAP) 추정에 기초한 감정 모델이 이용된다. 실험 결과, 음성에서의 감정 인식률은 적응 샘플수의 증가에 따라 점차적으로 증가함을 보여준다.

1. Introduction

In the human-human interaction, we sometimes feel the emotional states, contained in voices, such as anger, surprise, or sadness during communication. It is because the voice is an indicator of the psychological and physiological state of the person, as well as a communicative means. In the human- computer interaction, therefore, it would be quite useful if a computer system can recognize human emotional states that one expresses in conversation. The human-computer interfaces could be made to respond differently if the machine understands the emotional states or feelings of user.

In the recent years, many researches on analysis of human emotional factors have been conducted, particularly, in the fields of emotional voices, facial expressions, and body gestures etc. Therefore, it is the one of essential issues to study real aspects of nonverbal communication of human emotional expressions. In recognizing emotional states or human feelings contained in speech signals, however, there are still few reports[1,2,3], most of which are based on mathematical classification methods or pattern recognition techniques.

In this research, a new approach of discriminating emotional states was attempted to realize using a statistical model based on hidden Markov model(HMM), which has been most widely used in the area of speech recognition. Therefore, this study has advantages that the proposed modules of emotion recognition can be easily integrated with the existing speech recognizer, since both systems are based on the same architecture compatible in the basic algorithms. The emotional features that consist of prosodic information are extracted and then trained to form standard models. In this case, the adapted emotional models using maximum a posteriori(MAP) estimation are also considered for better performance on specific speakers.

2. Extraction of Emotional Features

The emotional feature parameters are first extracted from voice signals that contain emotional information. The prosodic information[4,5] is well known as an indicator of the acoustic characteristics of vocal emotions. In our experiments, we used four kinds of prosodic parameters that consist of pitch, energy, and each derivative element. For incorporating the effect of speaking rate in voices, furthermore, we also used discrete duration information in the course of the training process based on HMM.

Figure 1 shows that the speech samples were labeled at the syllable level (for example, /Ta/ and /Ro/) by a manual segmentation where only voiced regions are considered as data points. The speech signals in the voiced regions were smoothed by a spline interpolation. From the speech waveform, the emotional feature parameters are extracted for the training and recognition using HMM.

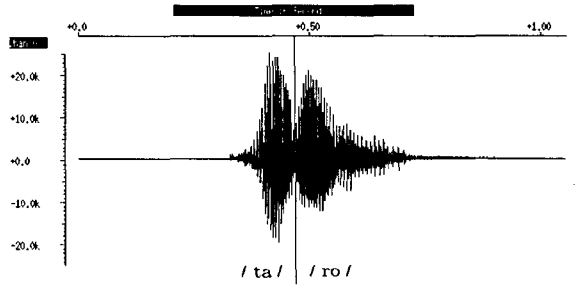

Fig. 1. Example of speech waveform labeled by two parts /Ta/ and /Ro/.

Figure 2 and 3 show the pitch and energy signals, respectively, extracted from emotional speech, "Taro", that was spoken by a female actress. In this figures, it was noticed that the level of feature signals in anger state is, particularly, the highest among five kinds of feature curves.

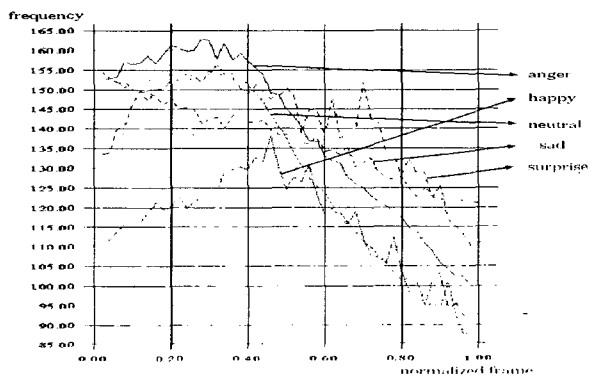

Fig. 2. Pitch signals as an emotional feature.

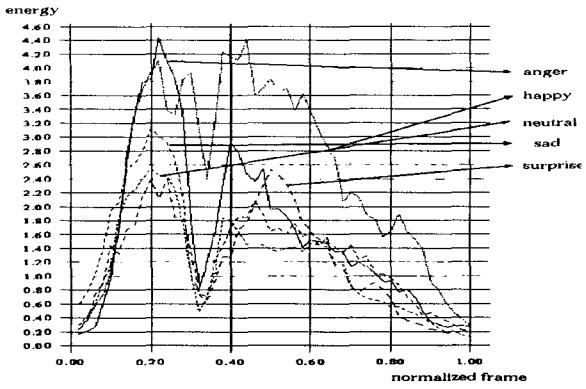

Fig. 3. Energy signals as an emotional feature.

Figure 4 and 5 show the time-differential emotional features as derivative elements of both pitch and energy signals, respectively. From each feature signals shown in figure 2,3,4 and 5, it is found that the feature curves are different in each emotional state. Therefore, we can build five different kinds of characteristic emotional models, respectively, through the training process using HMM.

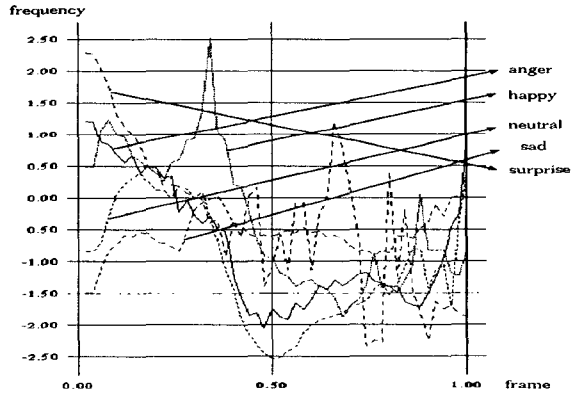

Fig. 4. Time-differential pitch signals as an emotional feature.

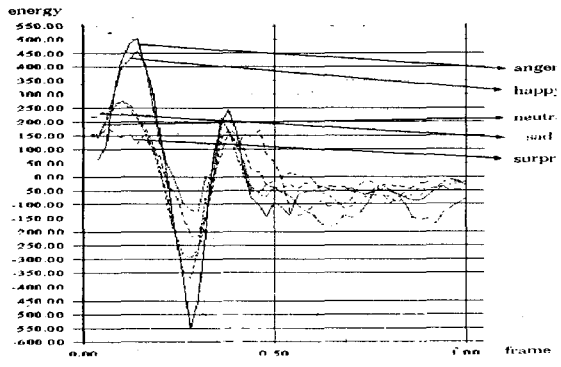

Fig. 5. Time-differential energy signals as an emotional feature.

3. Adaptation of emotional models

The MAP estimation is also called Bayesian successive estimation of HMM parameters for a new speaker in a framework. The estimated mean vector value after given N samples is shown as,

$$\hat{\mu}_N = \frac{\alpha\mu_o + \sum_{i=1}^{N} X_i}{\alpha + N} \quad (1)$$

where α is an adaptation parameter. The estimated covariance matrix using N samples is

$$\sum_N = \frac{1}{\beta + N}\{X_N X_N^T - (\alpha + N)\mu_N \mu_N^T + (\beta + N - 1)\sum_{N-1} + (\alpha + N - 1)\mu_{N-1}\mu_{N-1}^T\} \quad (2)$$

where β is a coefficient. In our experiments, the values of α,β were set at 15 and 50 respectively, which were determined experimentally.

The utterances of specific speaker and the emotional sequences are first given to Viterbi segmentation and then inputted to MAP estimation algorithm, in which speaker-independent(S-I) emotional models are updated to speaker-adapted (S-A) models.

4. Recognition of emotional states

In this study, the syllable label as a basic unit is defined for emotion recognition. Therefore, the basic units can be concatenated to form word or sentence emotional models, so that it would be possible to realize continuous emotion recognition for future works.

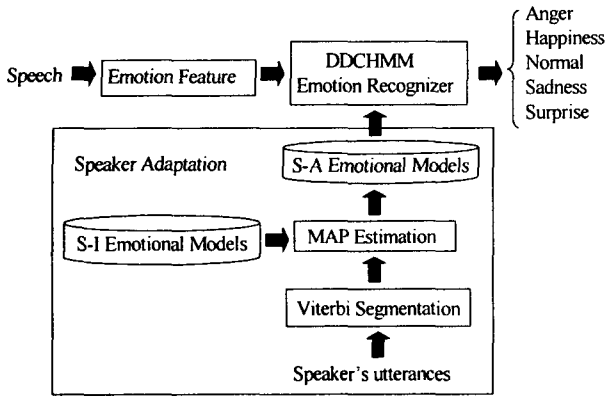

Fig. 6. Overall emotion recognition system with speaker adapted emotional models.

Figure 6 shows an overall emotion recognition system in which speaker adaptation modules based on MAP estimation are incorporated into the main module. In case the speech signals are given to the system, the emotional features are first picked out for pre-processing and then entered HMM emotion recognizer which has an advantage of modeling a duration of each HMM state. In this case, S-A emotional models are trained based on MAP estimation mentioned in the above section. Finally, the system recognizes emotional states using the trained S-A emotional models.

5. Experiments and Discussion

As emotion database, we captured speech samples that were the emotionally induced utterances, simulating five emotional states such as anger, happiness, normal, sadness, and surprise. From the utterances, the semantically neutral word, Japanese name 'Taro' was picked out for evaluation. The 175 samples(7 samples*5 emotions*5 speakers) spoken by 3 actors, 2 actresses were used for training data. On the other hands, the 35 samples(7 samples*5 emotions) spoken by 1 female professional announcer were used for adaptation data. For test data, we used 100 samples(20 samples*5 emotions) spoken by the same speaker in the adaptation procedure.

The speech signals are sampled and analyzed for pre-processing of emotion recognition as shown in table 1. We then extracted four dimensional emotional features that are composed of pitch, energy, pitch regressive coefficient(RGC), energy RGC as well as discrete duration information.

Sampling rate	16Khz , 16 Bit
Pre-emphasis	0.97
Window	16 msec. Hamming window
Frameperiod	5 ms
Feature parameters	pitch signal , energy, pitch RGC, energy RGC, discrete duration information

Table 1. Analysis of speech signals

In simulation experiments, we performed two kinds of recognition tests on five and two different emotional states, respectively. Figure 7 shows the one of the test results in which the recognition rates on five different emotional states depending on speaker adaptation

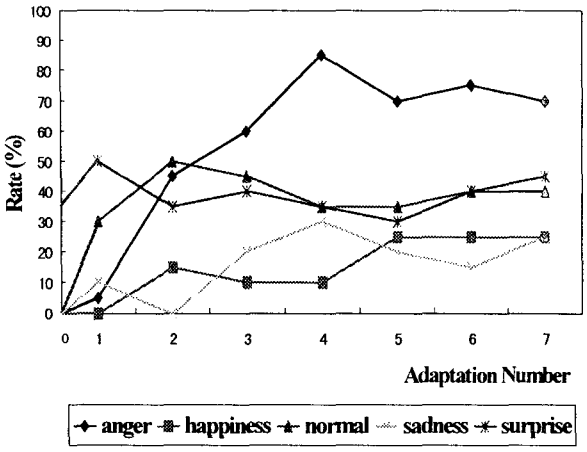

Fig. 7. Emotion recognition rates on five different emotional states depending on speaker adaptation .

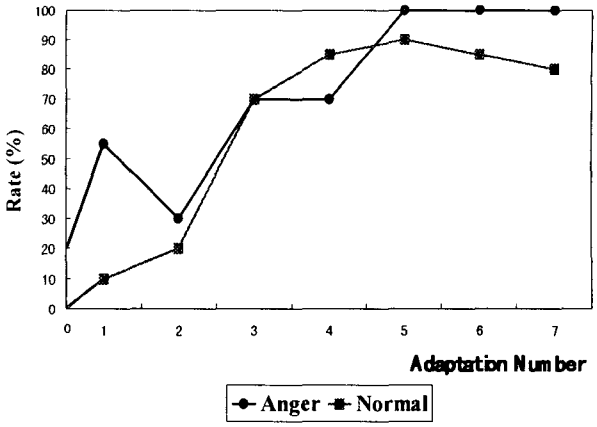

Fig. 8. Emotion recognition rates on anger and normal states depending on speaker adaptation .

It is noticed that the recognition rates in each emotional state grow gradually, in which the anger state has the highest recognition rate. However, the overall rates are unsatisfactory because of an insufficient training of emotional states. This is mainly due to the inadequate amounts of emotion database. Therefore, the performance of emotion recognition would be much better if the training procedure is converged to a relevant level by using enough emotional speech data.

The other experiments are shown in figure 8, in which the recognition rates depend on speaker adaptation in anger and normal states. We can see that the recognition rates in anger and normal states grow increasingly with an increase of adaptation sample number. In practical applications using emotion recognition techniques, it might be quite useful if computer system can recognize only two emotional states such as anger or normal state. For example, the system will be able to advise user to relax when anger state is detected from his or her speech. In addition, the system might perceive the stressful situation that occurs in human-computer interaction, and then correct the unnatural conversation.

6. Conclusion

This paper has described the new approach of recognizing human emotional states contained in voice signals using HMM and MAP adaptation techniques. The present study aims the friendly human-computer interaction by incorporating the nonverbal information such as emotion into general speech information. For realization, the prosodic information were first defined and extracted from speech signals for emotion recognition. The feature parameters were then given to HMM for training and recognition, in which specific speaker's utterances were also used for building adapted emotional models based on MAP estimation. For evaluation, the results presented that the recognition rates in each emotional state grew little by little with an increase of adaptation samples. It was found in the experiments that HMM and MAP estimation algorithms, which have been chiefly used in the area of speech recognition, were also useful in identifying emotional states contained in voice signals.

7. References

[1] F. Dellaert, T. Polzin, A. Waibel, "Recognizing Emotion in Speech", Proc. of the ICSLP'96, October, 1996.

[2] T. Moriyama, S. Ozawa, "Emotion Recognition and Synthesis System on Speech", Proc. of International Conference on Multimedia Computing and Systems(ICMCS'99), Florence, Italy, 1999.

[3] D. Roy, A. Pentland. Automatic, "Spoken Affect Classification and Analysis", Proc. of the 2nd International Conference on Automatic Face and Gesture Recognition, pp 363-367, 1996.

[4] Waibel, A, "Prosody and Speech Recognition", Doctoral Thesis, Carnegie Mellon Univ. 1986.

[5] C Tuerk, "A Text-to-Speech System based on {NET}talk", Master's Thesis, Cambridge University Engineering Dept, 1990.