은닉 마르코프 모델을 이용한 음성에서의 감정인식

Emotion recognition in speech using hidden Markov model

김성일, *정현열

Sung-Ill Kim, *Hyun-Yeol Chung

중국 청화대학 음성기술센타

Center of Speech Technology,
State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science & Technology,
Tsinghua University, Beijing, 100084, China

*영남대학교 공과대학 정보통신공학과

*Department of Information and Communication Engineering,
Yeungnam University, 214-1, Kyung-San City, Kyungbuk, 712-749, Korea.

요약

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 결과로서, 음성에서의 감정 인식률은 적응 샘플수의 증가에 따라 점차적으로 증가함을 보여준다.

Abstract

This paper presents the new approach of identifying human emotional states such as anger, happiness, normal, sadness, or surprise. This is accomplished by using discrete duration continuous hidden Markov models(DDCHMM). For this, the emotional feature parameters are first defined from input speech signals. In this study, we used prosodic parameters such as pitch signals, energy, and their each derivative, which were then trained by HMM for recognition. Speaker adapted emotional models based on maximum a posteriori(MAP) estimation were also considered for speaker adaptation. As results, the simulation performance showed that the recognition rates of vocal emotion gradually increased with an increase of adaptation sample number.

Key words : Emotion Recognition, HMM, MAP Estimation, Speaker Adaptation, Prosody

I. Introduction

In the human-human interaction, we sometimes feel the emotional states, contained in voices, such as anger, surprise, or sadness in the course of 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. Therefore, understanding those nonverbal communications has been one of the most important subjects for the ultimate goal to a human-like robot. Figure 1 shows the virtual conversation between human being and robot system understanding human feelings or mental states. In this example, we can see that the further information on the present emotional state of user is revealed through emotion recognition.

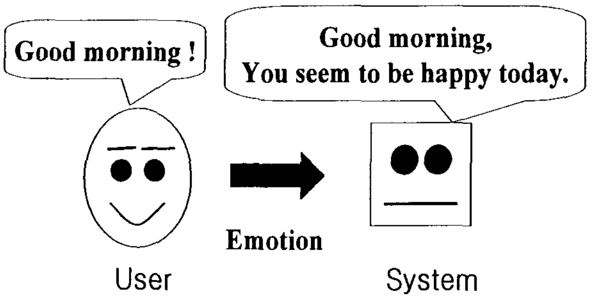


그림 1. 사용자와 로봇간의 감정인식을 통한 가상 대화
Fig. 1. Virtual communication through emotion recognition between user and robot.

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[1,2,3,4] which are based on mathematical classification methods or pattern recognition techniques .

In this paper, new approach of discriminating emotional states was attempted to realize using a statistical model based on hidden Markov model(HMM)[5,6,7,8] 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 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.

II. Extraction of Emotional Feature

The emotional feature parameters are first extracted from voice signals that contain emotional information. The prosodic information [9,10,11,12] is well known as an indicator of the acoustic characteristics of vocal emotions [13,14,15,16]. 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[17,18] in the course of training based on HMM.

Figure 2 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 in HMM.

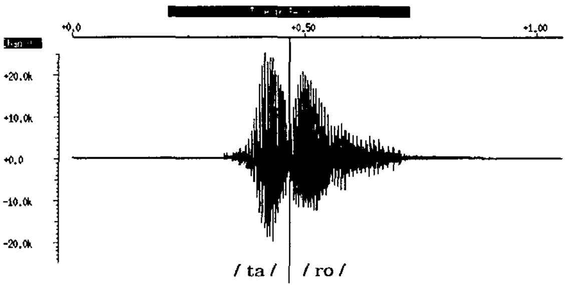


그림 2. /Ta/와 /Ro/로 레이블된 음성 파형의 예
Fig. 2. Example of speech waveform labeled by two parts /Ta/ and /Ro/.

Figure 3 and 4 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. Figure 5 and 6 show the time-differential emotional features as derivative elements[17,18] of both pitch and energy signals, respectively. From each feature signals shown in figure 3,4,5, and 6, 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 a training process using HMM.

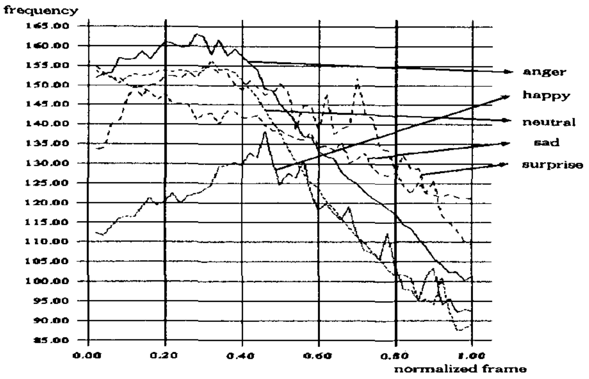


그림 3. 감정특징 파라메타를 위한 피치신호
Fig. 3. Pitch signals for emotional feature parameters

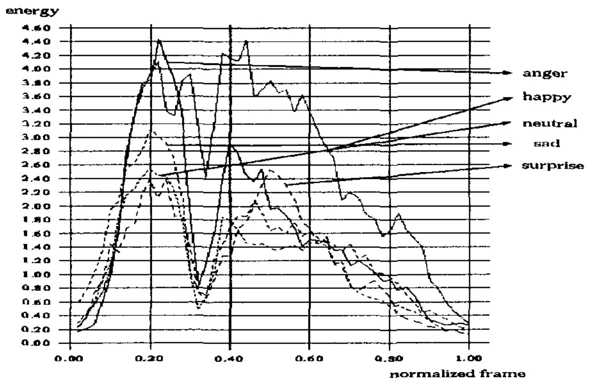


그림 4. 감정특징 파라메타를 위한 에너지신호
Fig. 4. Energy signals for emotional feature parameters.

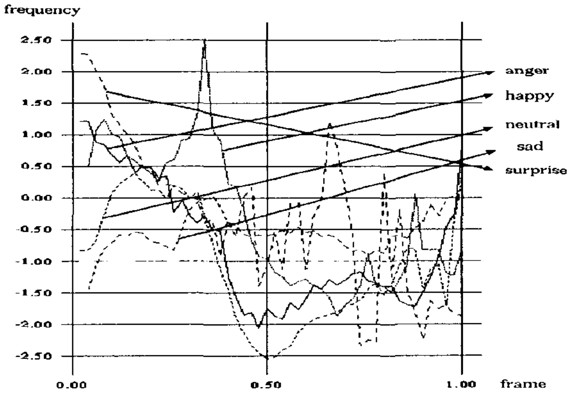


그림 5. 감정특징 파라메타를 위한 시간미분피치신호
Fig. 5. Time-differential pitch signals for emotional feature parameters

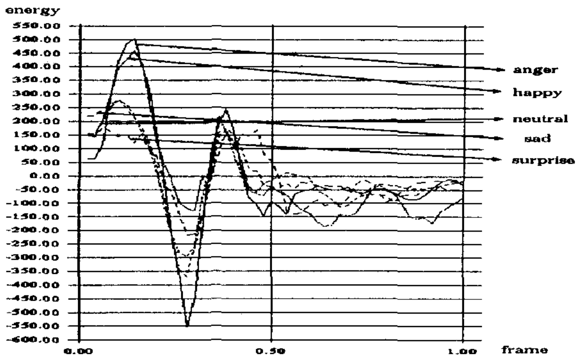


그림 6. 감정특징 파라메타를 위한 시간미분에너지신호
Fig. 6. Time-differential energy signals for emotional feature parameters.

III. Emotional model adaptation

The MAP estimation[18,19] 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,

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

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

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

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 speaker adaptation is performed with successive training of speaker-independent(S-I) models using small amounts of adaptation speech data. There are two adaptation methods. One is well known as the supervised speaker adaptation(SSA) which achieves the adaptation in accordance with correct label sequences. Another model is known as the unsupervised speaker adaptation(USA) where the label sequences are provided automatically by the Viterbi segmentation.

Figure 7 shows a block diagram of unsupervised speaker adaptation for emotion recognition based on

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

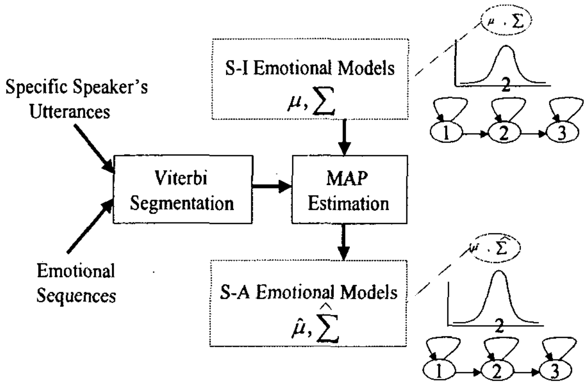


그림 7. 감정인식을 위한 자율 화자적응

Fig. 7. Block diagram of unsupervised speaker adaptation for emotion recognition.

Ⅳ. 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. Figure 8 illustrates HMM network of emotional states in /Taro/ speech as an example where five different syllable emotional models are concatenated into one-word emotional model.

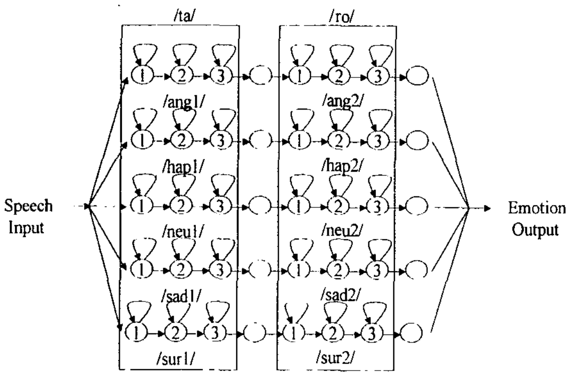


그림 8. 단어에서의 감정상태의 HMM 네트워크

Fig. 8. HMM network of emotional states in word

Figure 9 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, 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. Therefore, the system finally recognizes emotional states using the trained S-A emotional models.

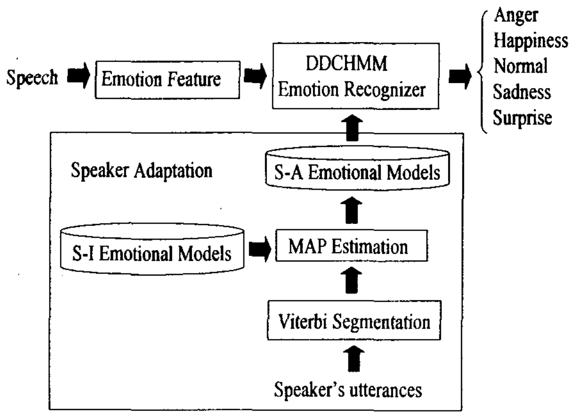


그림 9. 화자 적응된 감정모델을 가진 감정인식 시스템

Fig. 9. Overall emotion recognition system with speaker adapted emotional models.

Ⅴ. 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 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)[18,19], energy RGC as well as discrete duration information.

표 1. 음성신호 분석

Table 1. Analysis of speech signals

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

In simulation experiments, we performed two kinds of recognition tests on five and two different emotional states, respectively. Figure 10 shows the one of the test results in which the recognition rates depend on speaker adaptation in five different emotional states. 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.

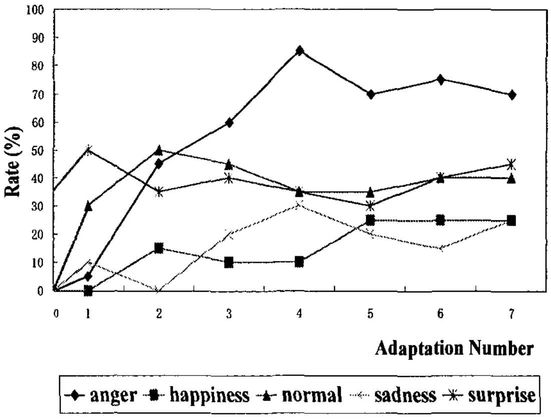


그림 10. 5가지 감정상태에서의 화자적응 감정인식률

Fig. 10. Emotion recognition rates dependent on speaker adaptation in five different emotional states.

The other experiments are shown in figure 11, 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 would 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 correct the unnatural conversation.

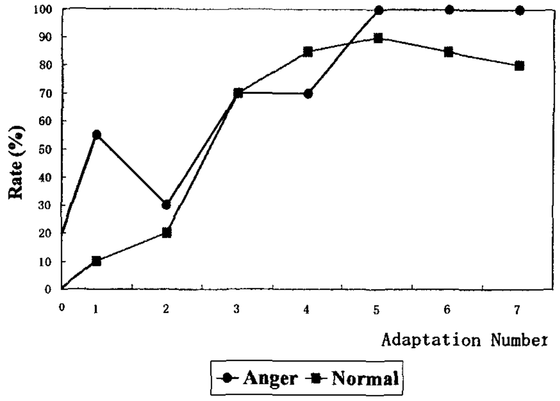


그림 11. 분노와 평정상태에서의 화자적응 감정인식률

Fig. 11. Emotion recognition rates dependent on speaker adaptation in anger and normal states.

Ⅵ. 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 was 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.

접수일자 : 2002. 6. 7 수정완료 : 2002. 7. 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. Y. Yu, E. Chang, C. Li, "Computer Recognition of Emotion in Speech", The 2002 Intel International Science and Engineering Fair, 2002.

5. L. R. Rabiner, R. W. Schafer, "Digital Processing of Speech Signal", Book: Prentice-Hall, 1978.

6. C. Becchetti, L. P. Ricotti, "Speech Recogniton: Theory and C++ Implementation", Book: John Wiley & Sons, 2000.

7. S. Nakagawa, A. Kai, T. Itoh and S. Kogure , "Speech recognition and understanding of spoken dialogue", Proc. Int. Symposium on Spoken Dialogue, pp.5-1-5-6, 2000.

8. R. Fernandez, "Stochastic Modeling of Physiological Signals with Hidden Markov Models: A Step Toward Frustration detection in Human- Computer Interfaces", MIT EECS Thesis for M.Sc. degree in Electrical Engineering and Computer Science, 1997.

9. Waibel, A, "Prosody and Speech Recognition", Doctoral Thesis, Carnegie Mellon Univ. 1986.

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

11. David Talkin. "A robust algorithm for pitch tracking (RAPT)," in Speech Coding and Synthesis, Elsevier Science, Amsterdam, pp. 495-518, 1995.

12. Alice E. Turk, James R. Sawusch, "The processing of duration and intensity cues to prominence", Journal of the Acoustical Society of America, 99(6):3782-3790, June 1996.

13. A. Fernald, "Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages", Developmental Psychology, Vol. 64, pp 657-674, 1993.

14. Rosalind W. Picard, "Affective Computing", MIT Press, Cambridge, MA, 1997.

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

16. E. Vyzas, "Recognition of Emotional and Cognitive States Using Physiological Data", Mechanical Engineer's Degree Thesis, MIT, June 1999.

17. K.F.Lee, "Automatic Speech Recognition; The Development of SPHINX System", Kluwer Academic Publisher, Norwell, Mass., 1989.

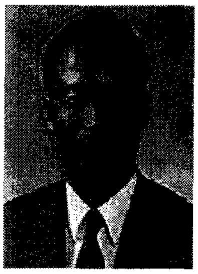
18. L. Rabiner, B-H. Juang, "Fundamentals of Speech Recognition", Prentice Hall Signal Processing Series, 1993.

19. Y. Tsurumi, S. Nakagawa, "An Unsupervised Speaker Adaptation Method for Continuous Parameter HMM by Maximum a Posteriori Probability Estimation", Proc. of ICSLP'94, pp.431-434, 1994.

Profile

- Sung-Ill Kim

Sung-Ill Kim was born in Kyungbuk, Korea, 1968. He received his B.S. and M.S. degrees in the Department of Electronics Engineering from Yeungnam University, Korean, in 1997, and Ph.D. degree in the Department of Computer Science & Systems Engineering from Miyazaki University, Japan, in 2000. During 2000 to 2001, he was a postdoctoral researcher in the National Institute for Longevity Sciences, Japan. Currently, he has been working in the Center of Speech Technology, Tsinghua University, China. His research interests include speech/emotion recognition, neural networks, and multimedia signal processing. E-mail; ksistar02@hanmail.net



- Hyun-Yeol Chung

Hyun-Yeol Chung was born in Kyungnam, Korea, 1951. He received his B.S. and M.S. degrees in the Department of Electronics Engineering from Yeungnam University, in 1975 and 1981, respectively, and the Ph.D. degree in the Information Sciences from Tohoku University, Japan, in 1989. He was a professor from 1989 to 1997 at the School of Electrical and Electronic Engineering, Yeungnam University. Since 1998 he is a professor in the Department of Information and Communication Engineering, Yeungnam University. During 1992 to 1993, he was a visiting scientist in the Department of Computer Science, Carnegie Mellon University, Pittsburgh, USA. He was a visiting scientist in the Department of Information and Computer Sciences, Toyohashi University, Japan, in 1994. He was a principle engineer, Qualcomm Inc., USA, in 2000. His research interests include speech analysis, speech/speaker recognition, multimedia and digital signal processing application. E-mail; hychung@yu.ac.kr