Recognition of Emotion and Emotional Speech Based on Prosodic Processing

Sung-Ill Kim*
*Division of Electronic and Electrical Engineering, College of Engineering, Kyungnam University

(Received July 14 2004; revised August 16 2004; accepted August 23 2004)

Abstract

This paper presents two kinds of new approaches, one of which is concerned with recognition of emotional speech such as anger, happiness, normal, sadness, or surprise. The other is concerned with emotion recognition in speech. For the proposed speech recognition system handling human speech with emotional states, total nine kinds of prosodic features were first extracted and then given to prosodic identifier. In evaluation, the recognition results on emotional speech showed that the rates using proposed method increased more greatly than the existing speech recognizer. For recognition of emotion, on the other hands, four kinds of prosodic parameters such as pitch, energy, and their derivatives were proposed, that were then trained by discrete duration continuous hidden Markov models(DDCHMM) for recognition. In this approach, the emotional models were adapted by specific speaker's speech, using maximum a posteriori(MAP) estimation. In evaluation, the recognition results on emotional states showed that the rates on the vocal emotions gradually increased with an increase of adaptation sample number.

Keywords: *Prosody, Emotional speech, Speech recognition, Emotion recognition, HMM*

I. Introduction

Generally, human voice is an indicator of the psychological and physiological state of the person, as well as a communicative means. Therefore, we can feel the emotional states such as anger, surprise, or sadness in the course of communication. In human-computer interaction, thus, it would be quite useful if a computer system can recognize human emotional states as well as human speech in conversation. Namely, the human-computer interfaces could be made to respond differently if the machine understands the emotional states or feelings of user. As a consequence, understanding those nonverbal communications has been one of the most important subjects for the ultimate goal to a humanlike robot.

In the recent years, many studies in the fields of speech recognition have been conducted. However, the present speech recognizers have often disregarded human speech signals with emotional states. Hitherto, the existing speech recognition systems have dealt with only normal speech that excluded various human emotional states. In addition, so far, there are still few reports[1-4] on classification methods or pattern recognition techniques in recognizing emotional states in vocal signals. Therefore, these are essential issues to study real aspects of nonverbal communication of human emotion. Figure 1 shows the

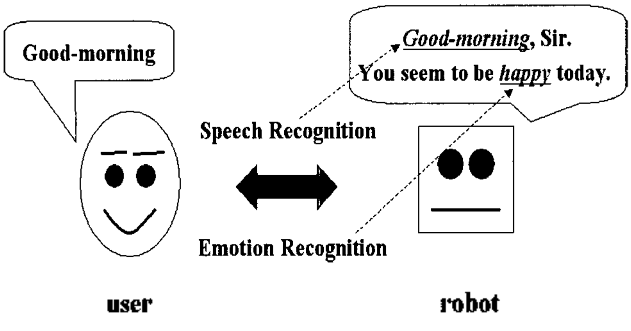


Fig. 1. Dialog between user and robot through speech and emotion recognition.

Corresponding author: Sung-Ill Kim (Kimstar@kyungnam.ac.kr)
Div. of Electronic & Electrical Engineering, Kyungnam University, 449, Wolyoung-dong, Masan, Kyungnam, 631-701, Korea

virtual conversation between user and robot system understanding both speech and human feelings. In this example, we can see that the techniques of handling emotional speech can make the human-computer interaction more natural, and give further information on user's present emotional states that have been usually omitted from the results of ordinarily speech recognition systems. Thus, the technologies dealing with emotional speech is currently being developed with the aim of improving the quality of human-computer interaction.

This study aims the friendly human-computer interaction by incorporating the nonverbal information such as emotion into speech information. For realizing it, the prosodic information was first defined and extracted from speech signals. In this study, two different approaches of discriminating both emotional speech and emotional states in speech are introduced. Each of the approaches was based on the prosodic processing for hidden Markov models (HMM) in common. The first approach presents prosodic identifier using prosodic information incorporated into the traditional acoustic features. The other approach is concerned with the emotion recognition using prosodic features, which can be easily integrated with the existing HMM speech recognizer because both systems are based on the same HMM architecture compatible in basic algorithms. For better performance on specific speakers, the adapted emotional models are also taken into account by using maximum a posteriori(MAP) estimation[5,6].

II. Prosodic Processing in Emotional Speech Recognition

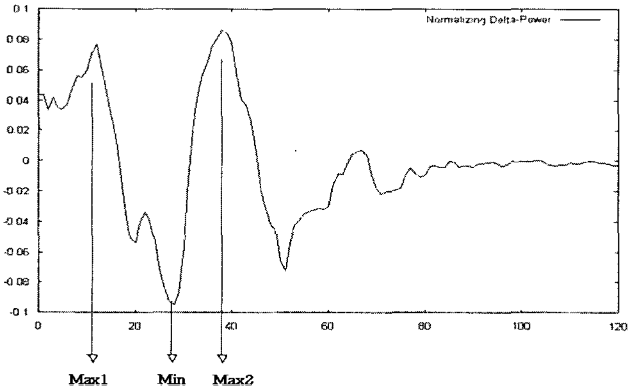


Fig. 2. Extraction of prosodic parameters from the normalized power RGC.

TABLE 1. Proposed prosodic features

Meaning Group	Prosodic Features
Temporal Information	Min Frame Max1 Frame Max2 Frame
Contour Information	Min Value Max1 Value Max2 Value
Num. of Voiced Region	Freq. of occurrence of 0Hz
Speaking Rate	Min Frame/(End Fr. - Min Fr.) Pitch1/(End Fr. - Pitch2)

The prosodic information[7-10] is well-known as an indicator of the acoustic characteristics of vocal emotions [11-15]. In order to improve classification performance, we adopted total nine kinds of features from both the normalized power regressive coefficients (RGC)[5,18] and the pitch signals. In this study, the power RGC means temporal changes in power curves in which rising and falling slopes are an important cue in identifying human emotions. The coefficients are then normalized so that individual variations are absorbed.

Figure 2 shows the signal of the normalized power RGC, from which seven kinds of prosodic feature parameters are extracted. The features are Min, Max1, Max2 frames, and Min, Max1, Max2 values, and Min frame/(End Frame-Min Frame). In this figure, the locations of Max1 and Max2 represent the biggest values in respective phonemes.

Figure 3 shows the pitch signal, from which two kinds of prosodic feature parameters, such as the occurred frequency of 0Hz and Pitch1/(End frame-Pitch2), are extracted. From the figure, we can see that the speaking rates and the number of voiced regions can be simply calculated from the locations of separate pitch signals. In this study,

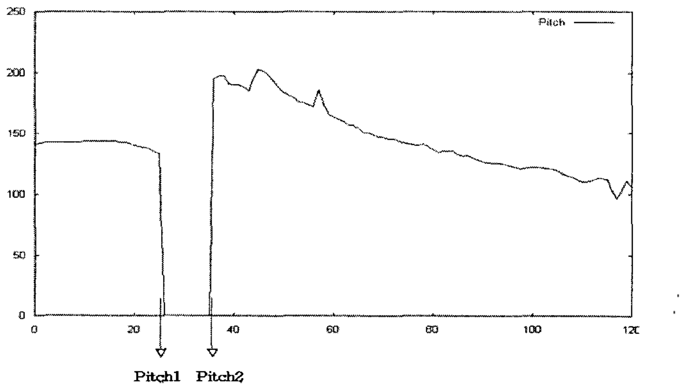


Fig. 3. Extraction of prosodic parameters from the pitch signals.

therefore, total nine kinds of prosodic features are incorporated into MFCC as shown in table 1.

The result values for prosodic identification are determined as following,

$$\Pr osodicId = \frac{1}{9}\sum_{n=1}^{9}\frac{\left|\Pr osodicAv_n\right| - \left|\Pr osodicIn_n\right|}{\left|\Pr osodicAv_n\right|} \qquad (1)$$

$$\begin{aligned} &If(0.0 < \Pr osodicId < 0.6) : Candidate \\ &else : Noncandida\ te \end{aligned} \qquad (2)$$

As shown in this equation, total nine kinds of prosodic averages are calculated based on the prosodic features mentioned above, respectively, using the training speech database. The deviation between average and input value is estimated by the simple equation (1). As a result, the result value is accepted as a candidate if it does not exceed the threshold value determined experimentally as illustrated in equation (2). In this case, the deviation of prosodic identification means the changing rate of emotional fluctuation of the same speech signals.

Figure 4 illustrates the speech recognition system, with the proposed prosodic identifier that is formed as an isolated module from the speech recognizer. In this study, the hypothetic recognition candidates are estimated by a probability score of HMM recognizer. We then decide the final recognition result, among recognition candidates, with the minimum value estimated by prosodic identifier.

III. Prosodic Processing in Emotion Recognition

For emotion recognition, we used four kinds of prosodic parameters that consist of pitch, energy, and their derivative elements. For incorporating the effect of speaking rate in voices, furthermore, we also used a discrete duration information[16, 17] in the course of training process based on HMM. Figure 5 shows that the speech samples are labeled at the syllable level by manual segmentation where only voiced regions are considered as data points. The emotional feature parameters are then extracted from the speech signals for both the training and recognition process of HMM.

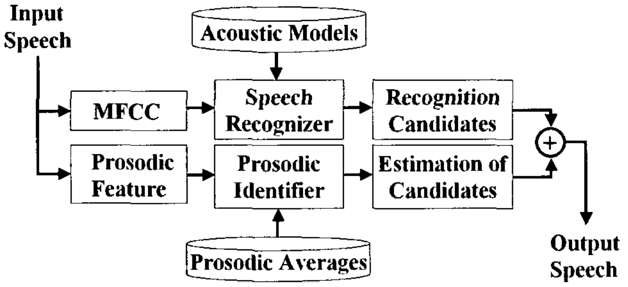


Fig. 4. Speech recognition system with the proposed prosodic identifier.

Figure 6 and 7 show the pitch and energy signals, extracted from emotional speech, /Taro/, that was spoken by a female actress. In these figures, it was noticed that the level of feature signals in angry state is, particularly, the highest among five kinds of feature curves.

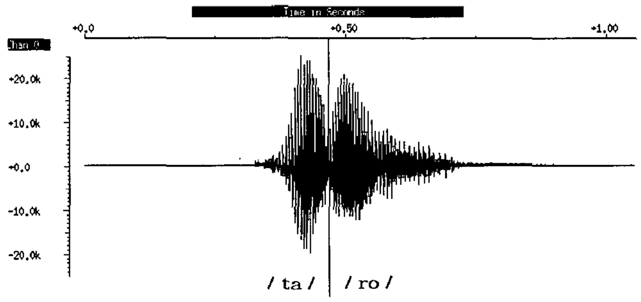


Fig. 5. Example of speech waveform labeled at the syllable level.

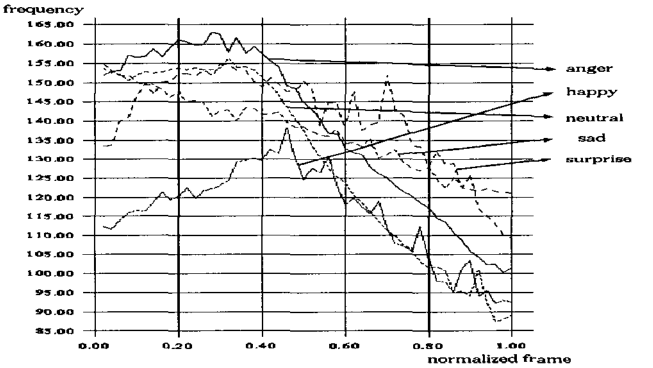


Fig. 6. Pitch signals for emotional feature parameters

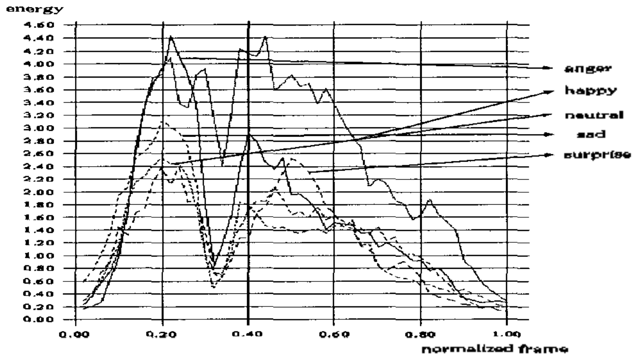


Fig. 7. Energy signals for emotional feature parameters.

The additional emotional features are the time-differential ones such as RGC in both pitch and energy signals, respectively. From the feature signals, it is found that the shapes of feature curves are different in each emotional state. Therefore, we can build each characteristic emotional model through training process using HMM.

Figure 8 shows the recognition procedure of the proposed 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 given to discrete duration continuous HMM (DDCHMM) emotion recognizer which has an advantage of modeling a duration of each HMM state. In this system, the recognition is performed using both speaker independent (S-I) and speaker adapted (S-A) emotional models in which S-A models are trained using 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 S-A models.

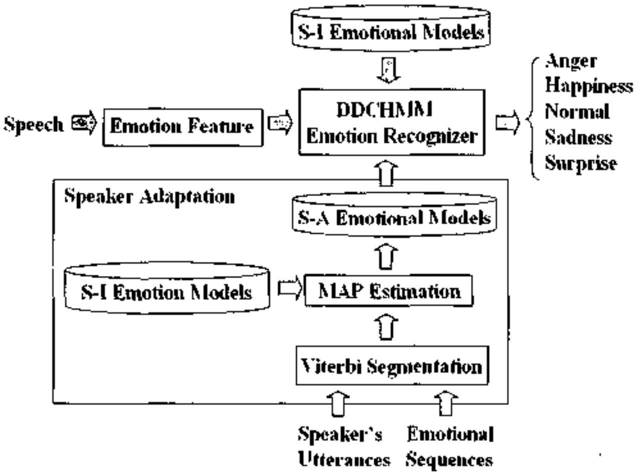


Fig. 8. Procedure of emotion recognition using HMM and MAP estimation.

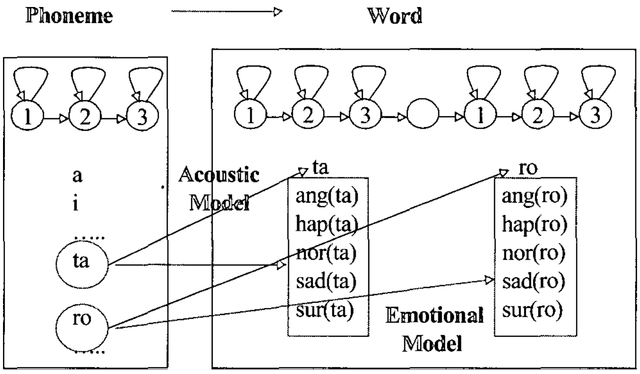


Fig. 9. Phoneme-level emotional HMM models incorporated into acoustic models.

TABLE 2. Analysis of speech signals

Sampling rate	16Khz , 16 Bit
Pre-emphasis	0.97
Window	16 msec. Hamming window
Frame period	5 ms

In this study, the phoneme label is defined as a basic unit for emotion recognition. Consequently, the basic units can be easily concatenated to form word or sentence emotional models, so that it is possible to realize continuous emotion recognition for future works. Figure 9 shows the phoneme-level emotional model where phoneme models are connected to form emotional word model. The techniques used in this study can also facilitate to build an integrated system between speech and emotion recognition. In addition, this idea can be applied to form prosodic averages, mentioned in the previous section, for robust speech recognition system absorbing the emotional variation that might distort original speech signals.

IV. Experiments and Discussion

For emotional speech database, we collected speech samples that were 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 and 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 table 2 shows the analysis of speech signals for pre-processing of recognition.

As preliminary study, we performed the recognition of speech with five kinds of different emotional states, so that we investigated if the current speech recognition systems recognized emotional speech correctly. Figure 10 illustrates the recognition rates on emotional speech using 'Julius' Japanese large vocabulary continuous speech recognition

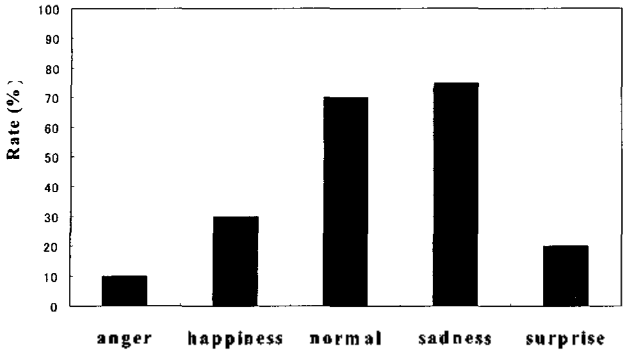


Fig. 10. Recognition rates on emotional speech using LVCSR system.

(LVCSR) system[19]. In this experiment, we used the test data of 100 samples (20 samples*5 emotions) spoken by a female professional announcer.

It was revealed from the graphs that the recognition rates on emotional speech, particularly in anger, happiness, and surprise states, were seriously degraded in performance. This is a limitation of the current speech recognition systems that depend on Mel-frequency Cepstral coefficient (MFCC) as feature parameters. It is because the existing speech recognition systems have dealt with only normal speech that excluded various human emotional states. In the present study, therefore, we explored to incorporate prosodic parameters as mentioned in the previous section, which was well-known as vocal emotional information, into MFCC.

In the simulation experiment of speech recognition on five emotional voices, the proposed speech recognition system with prosodic identification module was compared with the traditional system based on MFCC. Figure 11 shows the comparison of recognition results using MFCC and MFCC + prosodic information, respectively, on five kinds of emotional speech.

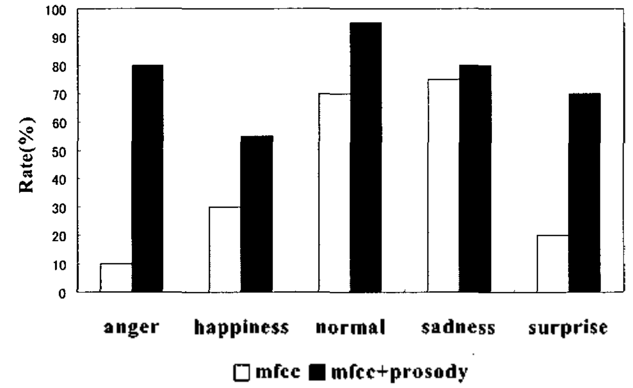


Fig. 11. Comparison of speech recognition system based on MFCC with the proposed system based on integration method between MFCC and prosodic information on five kinds of emotional speech.

In this graph, the recognition rates greatly increased by incorporating prosodic information into MFCC. From the results, it was noticed that the proposed method using prosodic identification was useful for identifying speech with emotional states. Moreover, it was found that the prosodic information could supplement the traditional acoustic features, such as MFCC, in speech recognition with emotional states.

In the experiment of emotion recognition using prosodic features and MAP estimation, on the other hands, we performed recognition test on five different emotional states. Figure 12 shows the results in which the recognition rates depend on speaker adaptation on five different emotional states. It is noticed that the recognition rates in each emotional state gradually grow, in which the anger state has the highest recognition rate.

However, the overall recognition rates were unsatisfactory because of an insufficient training for each emotional state. This is mainly due to inadequate amounts of emotional speech database. Therefore, we can see that the performance of emotion recognition would be much better if training procedure is converged to a relevant level by using enough emotional speech data.

V. Conclusion

The present study has focused on two different recognition tests handling emotional speech with five

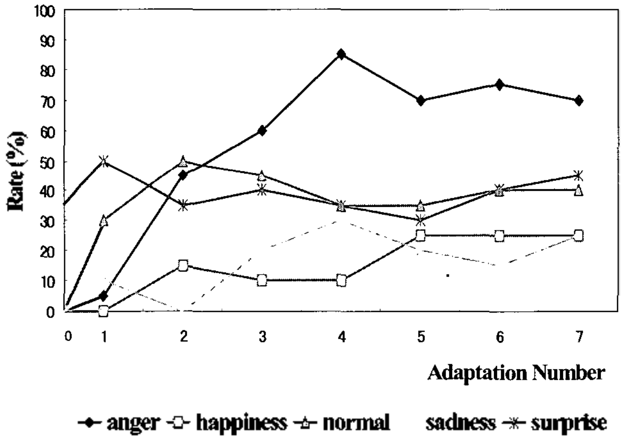


Fig. 12. Recognition rates on five different emotional states dependent on speaker adaptation.

different kinds of human emotional states. One of recognition test was concerned with the recognition of speech with emotional states using prosodic information. In evaluation, the recognition results revealed that the degradation of recognition of emotional speech greatly decreased by incorporating prosodic information into the existing feature parameters such as MFCC.

The other test was concerned with the recognition of emotional states using prosodic features and MAP estimation. In evaluation, the results presented that the recognition rates in every emotional states grew little by little with an increase of adaptation samples. In addition, it was revealed 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.

For the future works, it is greatly significant to attempt to explore the integrated system between emotion recognition and speech recognition robust on emotional variation.

Acknowledgement

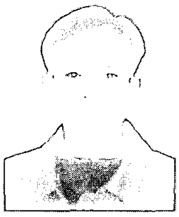
This work is supported by the Kyungnam University Research Fund, 2004.

References

1. F. Dellaert, T. Polzin, A. Waibel, "Recognizing Emotion in Speech", Proc. of the ICSLP'96, October, 1996.
2. 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.
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 and C. Li, "Computer Recognition of Emotion in Speech", The 2002 Intel International Science and Engineering Fair, 2002.
5. L. Rabiner and B-H. Juang, "Fundamentals of Speech Recognition", Prentice Hall Signal Processing Series, 1993.
6. Y. Tsurumi and 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.
7. Waibel, A, "Prosody and Speech Recognition", Doctoral Thesis, Carnegie Mellon Univ. 1986.
8. C Tuerk, "A Text-to-Speech System based on NETtalk", Master's Thesis, Cambridge University Engineering Dept, 1990.
9. David Talkin, "A robust algorithm for pitch tracking (RAPT)," in Speech Coding and Synthesis, Elsevier Science, Amsterdam, pp. 495-518, 1995.
10. 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.
11. A. Fernald, "Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages", Developmental Psychology, 64, pp 657-674, 1993.
12. Rosalind W. Picard, "Affective Computing", MIT Press, Cambridge, MA, 1997.
13. Deb Roy, Alex Pentland, "Automatic spoken affect classification and analysis", IEEE Face and Gesture Conference, Killington, VT, pp. 363-367, 1996.
14. E. Vyzas, "Recognition of Emotional and Cognitive States Using Physiological Data", Mechanical Engineer's Degree Thesis, MIT, June 1999.
15. J. L. Armony, D. Servan-Schreiber, J. D. Cohen, and J. E. LeDoux, "Computational modeling of emotion: Explorations through the anatomy and physiology of fear conditioning", Trends in Cognitive Sciences, 1(1), 28-34, April 1997.
16. L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signal*, (Prentice-Hall), 1978.
17. C. Becchetti and L. P. Ricotti, *Speech Recogniton: Theory and C++ Implementation*, (John Wiley & Sons, 2000).
18. K.F.Lee, *Automatic Speech Recognition; The Development of SPHINX System*, Kluwer Academic Publisher, Norwell, Mass., 1989.
19. 'Julius' Japanese large vocabulary continuous speech recognition system Available: http://winnie.kuis.kyoto-u.ac.jp/pub/julius/index.html

[Profile]

○ **Sung-Ill Kim**



Sung-Ill Kim was born in Kyungbuk, Korea, 1968. He received his B.S. and M.S. degrees in electronic engineering from Yeungnam University, in 1994 and 1997, respectively, and Ph.D. degree in computer science & systems engineering from Miyazaki University, Japan, in 2000. He worked at National Institute for Longevity Sciences, Japan, as a post-doctoral researcher from 2000 to 2001. He worked at Center of Speech Technology, Tsinghua University, China, from 2001 to 2003. Currently, he has been working at Division of Electronic & Electrical Engineering, Kyungnam University, as a full-time lecturer since 2003. His research interests include multimedia signal processing and its applications.