THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.21, No.2E 2002. 6. pp. 98~104

Discrimination of Emotional States in Voice and Facial Expression

Sung-Ill Kim[*], Yasunari Yoshitomi[**], Hyun-Yeol Chung[***]

[*]Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems, Dept. of Computer Science & Technology, Tsinghua University, China

[**]Dept. of Environmental Information, Faculty of Human Environment, Kyoto Prefectural University, Japan

[***]Dept. of Information and Communication Engrg., School of Electrical Engrg. and Computer Science Yeungnam University, Korea

(Received 24 May 2002; revised 29 May 2002; accepted 25 June 2002)

Abstract

The present study describes a combination method to recognize the human affective states such as anger, happiness, sadness, or surprise. For this, we extracted emotional features from voice signals and facial expressions, and then trained them to recognize emotional states using hidden Markov model (HMM) and neural network (NN). For voices, we used prosodic parameters such as pitch signals, energy, and their derivatives, which were then trained by HMM for recognition. For facial expressions, on the other hands, we used feature parameters extracted from thermal and visible images, and these feature parameters were then trained by NN for recognition. The recognition rates for the combined parameters obtained from voice and facial expressions showed better performance than any of two isolated sets of parameters. The simulation results were also compared with human questionnaire results.

Keywords: Emotion recognition, HMM, Neural network, Prosody, Facial expression

I. Introduction

In the human-computer interaction, it would be a quite useful if a computer can recognize the affective states of human being in the course of communication. Namely, the human-computer interface could be made to respond differently if computer or robot system perceives the human feelings or mental states. The computer system can interact with persons in a friendly manner. For example, the system can play a comfortable music or advise the person to relax when the user is checked as a tired or unhappy state. Therefore, understanding those nonverbal communications have been one of the most important subjects for the ultimate goal to a humanlike robot, so that our future society would be more convenient.

Recently, many studies have been conducted on analyzing or recognizing the individual nonverbal characteristics such as emotional factors contained in voices, facial expressions, or body gestures etc. Therefore, it is one of the essential issues to study real aspects of human emotional expressions. However, there are still few reports considering and judging nonverbal information, such as emotion, of human being totally.

This study aims more natural human-computer interface with total understanding of human emotional states. In

Corresponding author: Hyun-Yeol Chung (hychung@yu.ac.kr)
Department of Information and Communication Engineering, School of Electrical Engineering and Computer Science Yeungnam University, 214-1, Kyung-San City, Kyungbuk, 712-749, Korea

this paper, the informative integration between voice and facial expression was proposed and practically simulated by extracting emotional features and then recognizing emotional states after training process. This was realized by presenting the integration method among human speech, visible and thermal facial expressions, and by improving the modeling method of emotional states using hidden Markov model (HMM) and neural network (NN).

II. Emotional Feature Extraction

For recognizing emotional states in both voice and facial expression, we need to extract emotional feature parameters from each of them. In a part of voice, we first analyze voice signals that might contain emotional information including four kinds of feature parameters. In a part of facial expressions, we extract useful feature parameters from visible and thermal image.

2.1. Emotional Feature Extraction from Voice

The prosody[1-4] is known as an indicator of acoustic characteristics of vocal emotions[5,6]. In this study, we

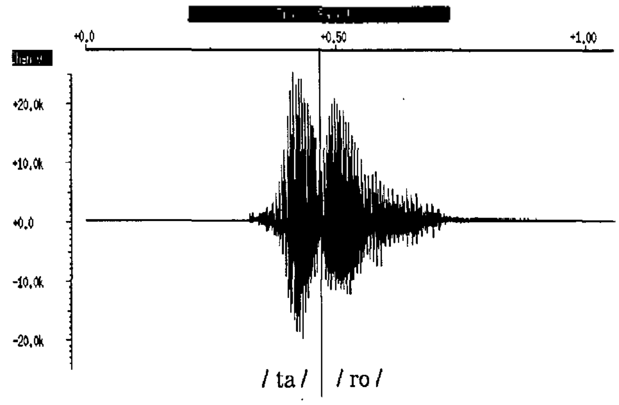

Figure 1. Examples of speech waveform labeled by two parts /Ta/ and /Ro/.

used four kinds of prosodic parameters, which consist of pitch, energy, and each derivative element. The pitch signals in voiced regions were smoothed by a spline interpolation. In order to consider the effect of a speaking rate in voices, furthermore, we incorporated discrete duration information[7,8] in the process of training using HMM.

We analyzed the feature parameters from the speech waveform shown in Figure 1, considering only the voiced regions as data points. All speech samples were labeled

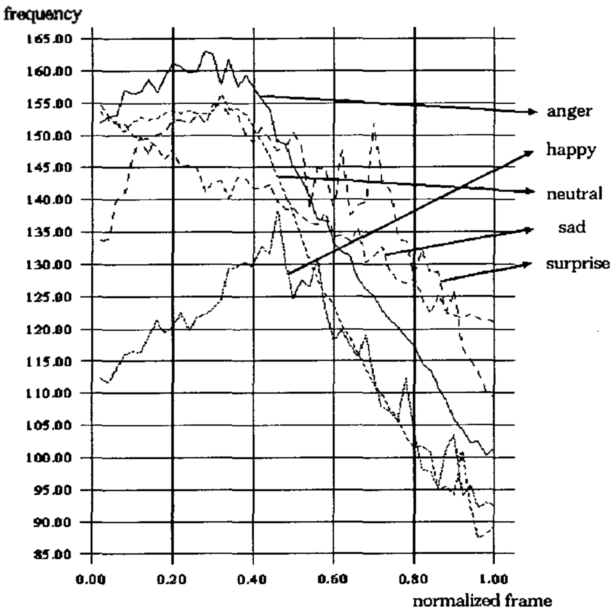

Figure 2. Pitch signals for emotional feature parameters in vocal emotions.

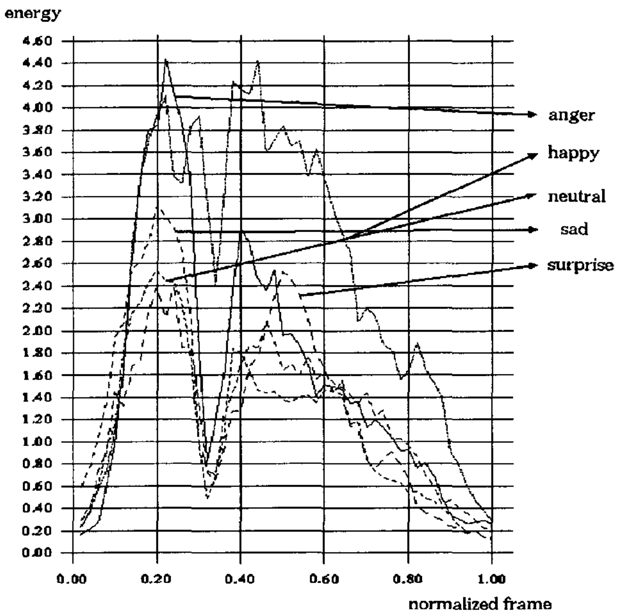

Figure 3. Energy signals for emotional feature parameters in vocal emotions.

at the syllable level (for example, /Ta/ and /Ro/) by manual segmentation in order to train emotional models.

Figure 2 and 3 shows the examples of the pitch and energy signals extracted from each emotional speech that was spoken by a female actress. Particularly, it was noticed in the figures that frequency and energy level in angry state were the highest among five kinds of emotional feature signals such as anger, happiness, normal, sadness and surprise.

2.2. Emotional Feature Extraction from Visible and Thermal Image of Face

Many studies have been performed to tackle the issues of understanding mental states of human being through facial expressions[9,10] using ordinary visible camera. However, those trials still seem to be tough jobs since there is a only slight difference among various facial expressions in terms of characteristic features for the gray level distribution of input image using visible ray (VR). In addition to the existing method using VR, thus, we have attempted to apply thermal distribution image to the recognition of facial expression using infrared ray (IR).

Figure 4 illustrates the examples of female face images by VR and IR. The VR image has the shortcoming that the recognition accuracy of facial expression is greatly influenced by lighting condition including variation of shadow, reflection, or darkness. However, it is perfectly overcome by exploiting IR that is independent of any those conditions.

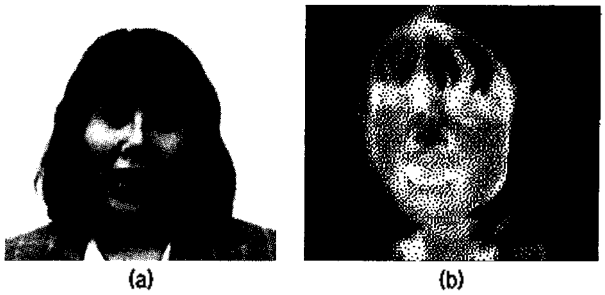

Figure 4. Examples of face images of VR (a) and IR (b).

When the face images are given to recognizer, first of all, it is necessary to correctly extract face-part areas that are important for better performance[11-13] in recognition of facial expressions. Figure 5 and 6 show blocks for extracting face-part areas that consist of three area fractions in VR image and six fractions in IR image, respectively. These blocks were decided based on the psychological and experimental studies. Each normalized area fractions were then used for extracting the values of feature parameters.

In next step, we generate differential images between the averaged neutral face image and the test image, which formed from the extracted face-part areas to perform discrete cosine transformation (DCT). Figure 7 illustrates the procedure of extracting characteristic features of emotional states from VR and IR image, respectively.

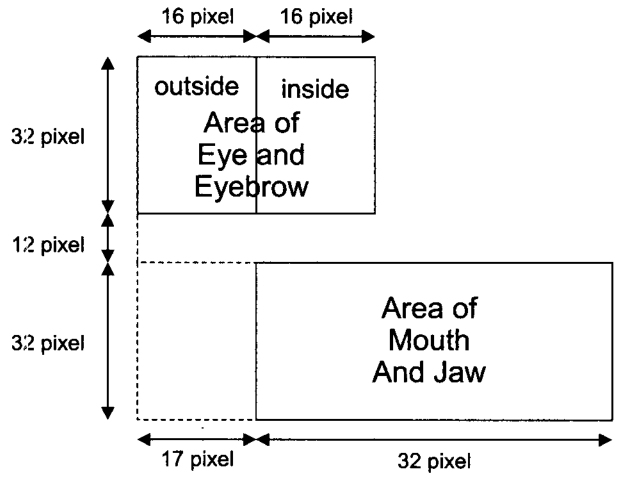

Figure 5. Extraction of blocks for face-part area in VR image.

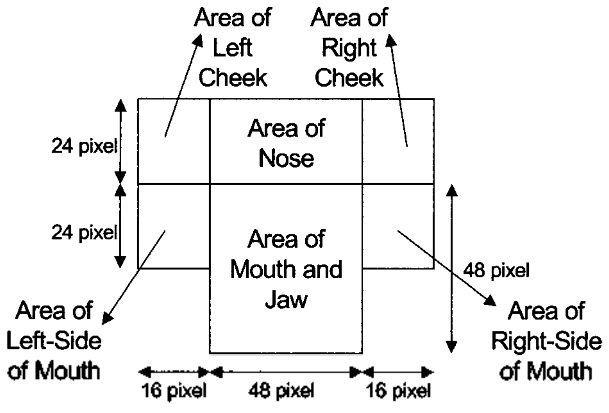

Figure 6. Extraction of blocks for face-part area in IR image.

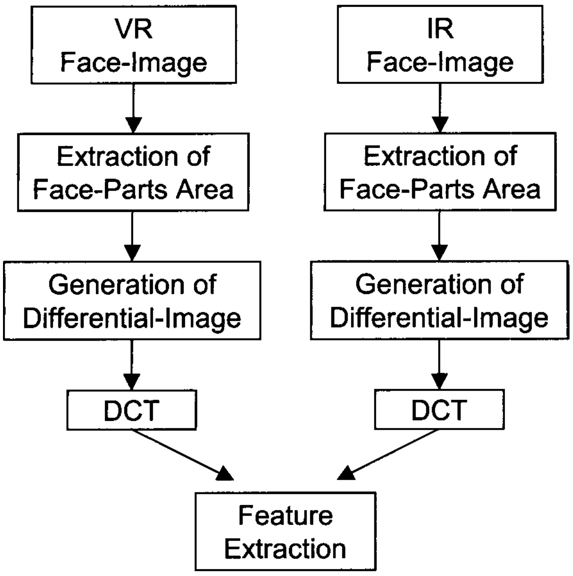
VR Face-Image
IR Face-Image
Extraction of Face-Parts Area
Extraction of Face-Parts Area
Generation of Differential-Image
Generation of Differential-Image
DCT
DCT
Feature Extraction

Figure 7. Procedure of extraction of emotional features from VR and IR facial images.

III. Recognition of Emotion

In case of processing the emotional voice, speech signals are first analyzed for pre-processing of emotion recognition. Table 1 illustrates the experimental conditions in an analysis of speech signals. We extracted emotional feature parameters such as pitch, energy, pitch regressive coefficient (RGC)[7,8], energy RGC, and discrete duration information.

Table 1. Analysis of speech signal.

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

We train emotional models using emotional features and label information based on HMM with three states, and then perform recognition tests. Figure 8 shows a basic concept of the training and recognition using HMM.

In case of processing the facial images, on the other hand, VR and IR images are first analyzed for pre-processing of facial expression recognition. We then train and recognize emotional states in facial expression using NN as shown in Figure 9. The unit number of hidden layer in NN is decided experimentally for improving the recognition accuracy. The unit number of output layer is the number of facial expressions that should be recognized. In this study, 78 and 57 bits of feature parameters are used as an input data for NN with three layers.

For the integration of emotional information from voice as well as VR and IR images, the weighted summation S_i for the emotional state i is calculated by $S_i = \sum_{j=1}^{3} x_{i,j}$ where $x_{i,j}$ is an output value (1 or 0) for an emotional state i using a method j. Therefore, the recognition results are chosen when S_i is maximum. The combination among three different kinds of recognition accuracies forms to

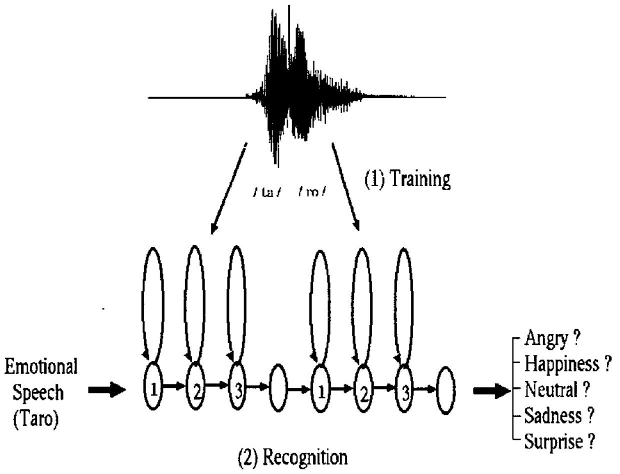

Figure 8. Concept of the training and recognition of emotional states using emotional HMM.

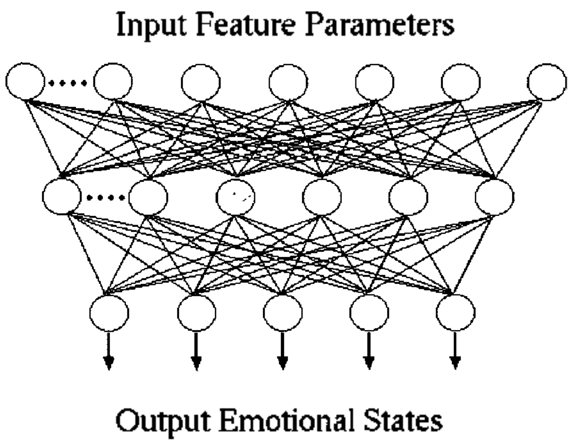

Figure 9. Neural networks for the training and recognition of emotional states.

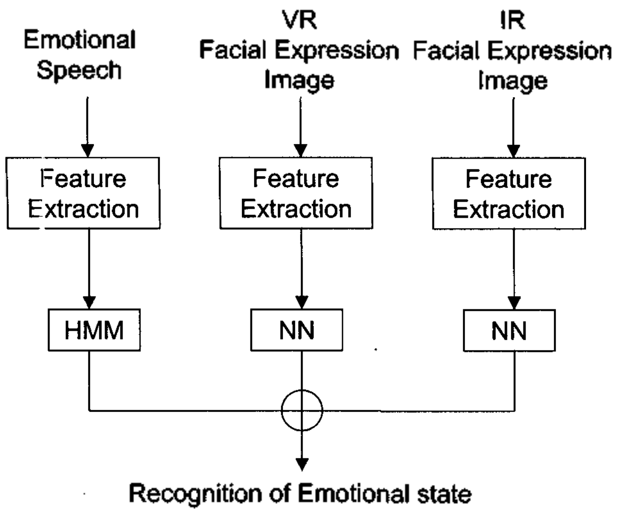
Emotional Speech
VR Facial Expression Image
IR Facial Expression Image
Feature Extraction
Feature Extraction
Feature Extraction
HMM
NN
NN
Recognition of Emotional state

Figure 10. Procedure of recognizing emotional states contained in voice and facial expression.

integrate each source for better recognition of emotional states. Figure 10 shows the overall procedure of recognizing emotional states using integration method among three different sources.

IV. Experiments and Discussion

4.1. Database

The samples consisted of semantically neutral utterance, Japanese name 'Taro', spoken and acted by 3 actors, 2 actresses, and 1 female professional announcer. We recorded voices and image sequences simultaneously, simulating five emotional states such as angry, happiness, normal, sadness, and surprise. We assembled total 100 samples for emotional expressions as training data and 50 samples as test data. In order to confirm whether the test data is objectively valid in representing emotional states, we performed the questionnaire experiments by getting persons to look and listen the data. In the next step, the same data was used in the computer simulation experiments based on the suggested method. Therefore, the validity of emotional state of data would be examined and compared between two experimental results. Moreover, these two kinds of experiments are important for investigating the difference between human cognitive power and computer discrimination ability.

4.2. Questionnaire Results

We first examined human performance on three different types of questionnaire. The emotional states of speech, image samples, and both of them were subjectively estimated by total 21 students who consisted of 14 male and 7 female collage students. Table 2 shows three different results of human performance. It was shown in this table that average recognition rates for five emotional states are 84.0 %, 82.4 %, and 92.5 %, when presenting emotional voices, facial expressions, and both of them, respectively. From these human performances, it was noticed that the questionnaire results integrating both voices and images gave better performance than those separately obtained from voices or images.

Table 2. Human performance for five different emotional states on emotional voices(a), facial expressions(b), and both emotional voices and images(c).

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**86.2**	7.7			5.7
	Hap.	2.8	**75.7**	1.2	0.5	4.3
	Neu.		1.9	**97.1**		2.9
	Sad.	1.4	2.9	1.7	**99.5**	25.7
	Sur.	9.6	11.8			**61.4**

(a) Human performance on emotional voices

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**56.7**	4.0	2.7	2.7	5.3
	Hap.	3.3	**90.0**		3.3	2.7
	Neu.	15.3	2.0	**94.0**	2.0	11.3
	Sad.	21.3	2.0	2.7	**92.0**	1.3
	Sur.	3.3	2.0	0.7		**79.3**

(b) Human performance on facial expressions

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**95.3**	1.3			3.3
	Hap.		**92.0**		0.7	3.3
	Neu.			**100.0**	0.7	4.0
	Sad.		0.7		**98.7**	12.7
	Sur.	4.7	6.0			**76.7**

(c) Human performance on both emotional voices and images

4.3. Simulation Results

We examined how effective our integration method among three different sources was, when simulation results were compared with human performance. We performed the recognition of emotional states over the same test data used in questionnaire test. Table 3 shows the recognition accuracies for each emotional state in the case of emotional voices, VR and IR facial expressions, and total recognition accuracies using integration method among three different recognition results.

Table 3. Recognition accuracies for five different emotional states on emotional voices(a), VR facial expressions(b), IR facial expressions(c), and integration method(d).

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**80**	20		20	10
	Hap.		**20**			
	Neu.		10	**100**	10	20
	Sad.				**50**	10
	Sur.	20	50		20	**60**

(a) Recognition accuracy for emotional voices using HMM

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**40**	40		20	30
	Hap.	50	**60**		10	
	Neu.			**70**		
	Sad.	10			**40**	
	Sur.			30	30	**70**

(b) Recognition accuracy for VR facial expressions using NN

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**40**				
	Hap.		**0**		20	
	Neu.			**70**		
	Sad.		100	30	**80**	50
	Sur.	60				**50**

(c) Recognition accuracy for IR facial expressions using NN

		Input emotion				
		Ang.	Hap.	Neu.	Sad.	Sur.
Output	Ang.	**60**			10	
	Hap.		**20**			
	Neu.			**90**		
	Sad.				**50**	10
	Sur.	10			20	**70**
	No Ans.	30	80	10	20	**20**

(d) Total recognition accuracy using integration method

As shown in this table, the average recognition rates for five emotional states are 62.0 % 56.0 %, and 48.0 %, when using emotional voices, VR and IR facial expressions, respectively.

Overall results are shown in table 3(d) that the total recognition rates amount to 85.0 % (except for no answers) for emotion recognition. Moreover, we could see that the simulation result (85.0 %) using integration method approached human performance (92.5 %) using both emotional voices and images, in spite of slightly low rate compared with questionnaire counterpart.

From the simulation experiments, there are some issues to have to take into account. The one thing is that the failure of emotion recognition in both cases of VR and IR facial expressions was mainly due to the difficulty to correctly extract face-part area because of the change of the face-orientation of subject. Therefore, it would be useful if the face recognition can be preceded by the performance of emotion recognition. The other thing is that the individuality of human feeling or states of mind is one of the important issues. This problem would be solved by producing more reliable standard models made from a large number of emotion database acquired from many subjects.

V. Conclusion

This paper has described the new integration approach of recognizing human emotional states contained in voices and facial expressions. The emotional parameters were trained and recognized by HMM and NN for voices and images, respectively. The recognition results showed that the integration method of recognizing emotional states in both voices and images gave better performance than any isolated methods, which also gave similar results to human questionnaire. Therefore, it was noticed that the discrimination ability of computer simulation had a possibility for realizing a humanlike robot as an ultimate goal.

References

1. A, Waibel, “Prosody and Speech Recognition,” Doctoral Thesis, Carnegie Mellon Univ. 1986.
2. C. Tuerk, “A Text-to-Speech System based on {NET)talk,” Master’s Thesis, Cambridge University Engineering Dept, 1990.
3. D. Talkin, “A robust algorithm for pitch tracking (RAPT),” *in Speech Coding and Synthesis, Elsevier Science*, Amsterdam, 495-518, 1995.
4. A. E. Turk, J. R. Sawusch, “The processing of duration and intensity cues to prominence,” *Journal of the Acoustical Society of America*, **99** (6), 3782-3790, June 1996.
5. A. Fernald, “Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages,” *Developmental Psychology*, **64**, 657-674, 1993.
6. R. W. Picard, *Affective Computing*. MIT Press, Cambridge, MA, 1997.
7. K. F. Lee, “Automatic Speech Recognition: The Development of SPHINX System,” Kluwer Academic Publisher, Norwell, Mass., 1989.
8. L. Rabiner, BH. Juang, “Fundamentals of Speech Recognition,” *Prentice Hall Signal Processing Series*, 1993.
9. Y. Yoshitomi, N. Miyawaki, S. Tomita, and S. Kimura, “Facial Expression Recognition using Thermal Image Recognition and Neural Network,” *Proc. of 6th IEEE Int. Work on Robot and Human Communication*, 380-385, 1997.
10. Y. Sugimoto, Y. Yoshitomi, and S. Tomita, “A method for Detecting Transitions of Emotional States using a Thermal Facial Image based on a Synthesis of Facial Expressions,” *Journal of Robotics and Autonomous Systems*, **31**, 147-160, 2000.
11. Y. Yoshitomi, T. Miyaura, S. Tomita, and S. Kimura, “Face Identification Using Thermal Image Processing,” *Proc. of 6^{th} IEEE International Workshop on Robot and Human Communication*, 374-379, 1997.

12 Y. Yoshitomi, A. Tsuchiya, and S. Tomita, “Face Recognition Using Dynamic Thermal Image Processing,” *Proc. of 7^{th} IEEE International Workshop on Robot and Human Communication*, 443-448, 1998.

13 Y. Yoshitomi, M. Murakawa, and S. Tomita, “Face Identification Using Sensor Fusion of Thermal Image and Visible Ray Image,” *Proc. of 7^{th} IEEE International Workshop on Robot and Human Communication*, 449-455, 1998.

[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, 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 network, and multimedia signal processing.

• Yasunari Yoshitomi

Yasunari Yoshitomi was born in 1956. He received his B.E., M.E. and Dr. Eng. degrees in Applied Mathematics and Physics from Kyoto University in 1980, 1982, and 1991, respectively. He had worked in Nippon Steel Corporation *since 1982 and had been engaged in research on image analysis application* and development of soft magnetic materials. Since 1995 he was in Miyazaki University as an associate professor at the Department of Computer Science & Systems Engineering. Since 2001 he is a professor in the Department of Environmental Information, Faculty of Human Environment, Kyoto Prefectural University. His research interests include image recognition, neural network, and stochastic programming problem. E-mail; yoshitomi@kpu.ac.jp

• 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.