퍼지 및 지능시스템학회 논문지 2005, Vol. 15, No. 2, pp. 173-178

멀티미디어 신호처리에 기초한 스마트홈 가상대화 시스템

Virtual Dialog System Based on Multimedia Signal Processing for Smart Home Environments

김성일*, 오세진**
Sung-Ill Kim and Se-Jin Oh

* Division of Electronic and Electrical Engineering, Kyungnam University
** Radio Astronomy Division, Korea Astronomy and Space Science Institute

요 약

본 논문은 보다 편리한 가정 생활환경 구축을 목적으로 한 가상대화시스템 구현에 관한 연구이다. 이를 실현하기 위하여 본 논문은 음성인식, 음성합성, 비디오 신호 및 센서신호처리 등의 멀티미디어 신호처리에 그 기술적 기반을 두고 있다. 대화시스템의 중요한 모듈로서의 음성합성기, HM-Net(Hidden Markov Network)에 기반한 실시간 음성인식기, 픽셀의 밝기차를 이용한 실시간 움직임 검출 및 터치센서 등을 대화시스템에 통합함으로써 이루어진다. 실제 구동 실험에서, 주위 노이즈 환경의 영향으로 시뮬레이션 결과보다는 성능이 떨어지나, 소파에 앉아있는 동안 작동되는 시스템의 실험 평가에서 가전제품 등의 컨트롤이 비교적 사용하기 쉬웠다는 결과를 얻었다.

Abstract

This paper focuses on the use of the virtual dialog system whose aim is to build more convenient living environments. In order to realize this, the main emphasis of the paper lies on the description of the multimedia signal processing on the basis of the technologies such as speech recognition, speech synthesis, video, or sensor signal processing. For essential modules of the dialog system, we incorporated the real-time speech recognizer based on HM-Net(Hidden Markov Network) as well as speech synthesis into the overall system. In addition, we adopted the real-time motion detector based on the changes of brightness in pixels, as well as the touch sensor that was used to start system. In experimental evaluation, the results showed that the proposed system was relatively easy to use for controlling electric appliances while sitting in a sofa, even though the performance of the system was not better than the simulation results owing to the noisy environments .

Key Words : Smart Home, Dialog System, Speech Recognition, Motion Detection, Multimedia Signals

1. Introduction

Generally speaking, smart home[1,2,3,4,5] or the home of the future refers to a house with networked products that can interact with each other and with house settings(example: heating system), which can be electronically predetermined and controlled by the inhabitants from central and/or mobile input devices. Namely, the infrastructure of smart home consists of a large variety of different networked sensors and systems, which may interplay in a defined manner. When talking about smart home, therefore, the focus often lies on technical grounds regarding home network infrastructures. In addition, many researches put most of their efforts in developing home network related works, but only little efforts in interactive concepts between inhabitants and their living environments at home.

Figure 1 shows the virtual conversation between user and smart home environment which is connected with essential components such as speech recognition, speech synthesis, video signal processing, automatic control system and acquisition system of collecting necessary information on the living room environment. The underlying idea is based on the fact that the place we spend most time at home is our living room, particularly in a sofa. The concept is started on the assumption that virtual dialog can be built when user sits in the sofa that is interconnected with the dialog system. As a consequence, the system enables people to interact with their home, so that our daily lives at home can be more convenient and comfortable. The ideal concept of smart home is possible to lead the living environments to the

접수일자 : 2004년 10월 16일
완료일자 : 2005년 2월 18일
감사의 글 : This work was supported by Kyungnam University Foundation Grant.

most suitable condition for inhabitants by using interactive and intelligent system as shown in the example.

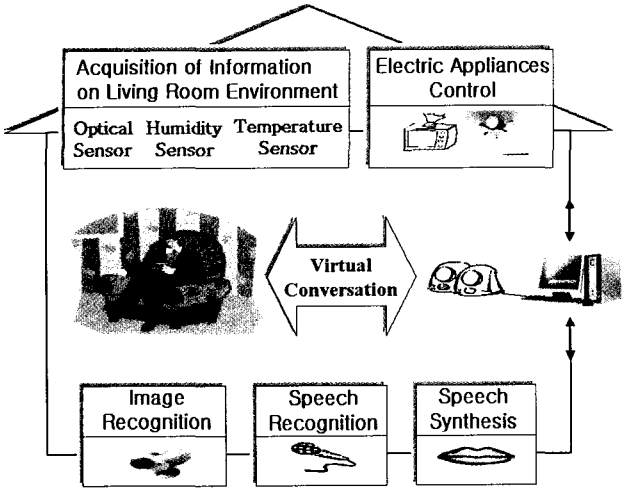


Fig. 1. Concept of virtual dialog conducting while sitting in a sofa of the living room in smart home environments.

In order to enhance a quality of life, this study aims the development of virtual dialog system for interactive home environments where home is possible to converse, just like friends or family members, with inhabitants as user. The proposed system can be realized by making a use of signal processing technologies using multimedia signals including image, speech, and sensor signals.

The system puts emphasis on an easy-to-use and user-friendly man-machine interface for smart home environments. As a result, we adopted, what we call, the human interface technologies including speech and image signal processing. For speech signal processing, in this study, the speech recognition was used to execute user's commands. For image signal processing, on the other hand, the motion detection was used to operate the overall system, which was integrated with the module of speech recognition.

2. HM-Net Speech Recognition System

Recent large vocabulary continuous speech recognition (LVCSR) systems are chiefly based on the state-clustered HMM(Hidden Markov Model). In this paper, we used HM-Net(Hidden Markov Network)[6,7] which is an efficient representation of context-dependent phonemes for LVCSR. The HM-Net, which has various state lengths and share their states one another, is automatically generated by PDT-SSS(Phonetic Decision Tree- based Successive State Splitting)[7,8,9,10].

The PDT-SSS is a powerful technique to design topologies of tied-state models, and is possible to generate highly accurate HM-Net. Each state of HM-Net has the information such as state index, contextual class, lists of preceding and succeeding states, parameters of the output probability density distribution and the state transition probability. If contextual information is given, the model corresponding to the context can be determined by concatenating several associated states within the restriction of the preceding and succeeding state lists. The final result of state splitting gets a network of states that efficiently represents a collection of context-dependent models, as illustrated in Figure 2. In this figure, "p/aa/s" denotes a Triphone for phoneme "aa" when the preceding context of "p" and the succeeding context of "s" are given. In contrast to the training process of the existing HMM, the architecture of the models can be automatically optimized according to the duration of utterances. As a result, the number of states in vowel increases more than that of states in consonant in the architecture.

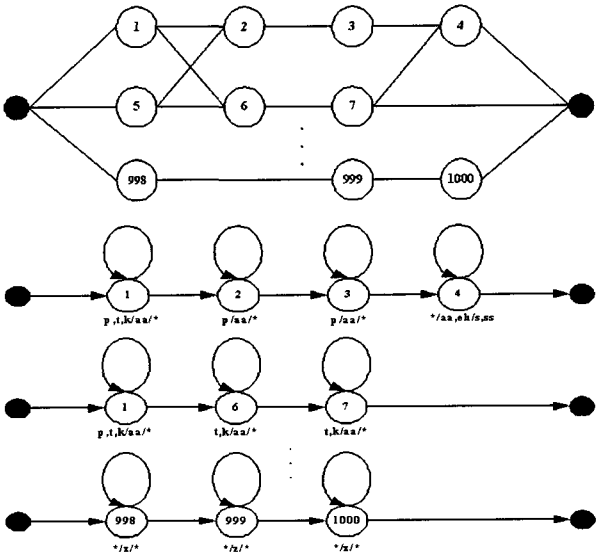


Fig. 2. An example of HM-Net models.

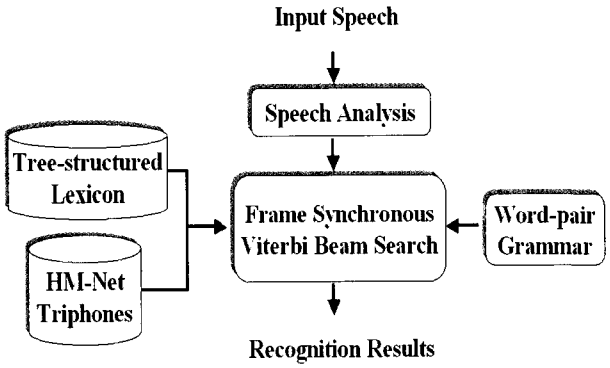


Fig. 3. Overall schematic of HM-Net speech recognition system.

Figure 3 shows an overall schematic of HM-Net speech recognition system[9,10]. In case speech signals are given to the system, the acoustic features are first

picked out for preprocessing, and then given to the search module that uses tree structured lexicon, and HM Net Triphones as well. The final recognition results are then obtained by frame synchronous Viterbi beam search algorithm using word pair grammar.

The speech recognition system used in this paper has been proved that it produced better performance than the conventional HMM speech recognizer in the experiments of phoneme, word, and continuous speech recognition [9,10].

3. Virtual Dialog System

The flow diagram of the processing based on the proposed system, which is operated in real time, is shown in figure 4. It illustrates how to build the dialog between system and user. If user sits in a sofa located in a living room, the touch sensor attached on the sofa gets the input signals.

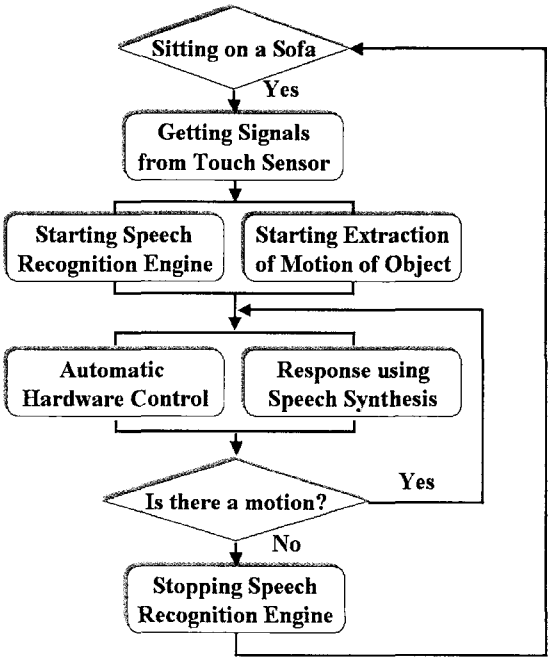


Fig. 4. Algorithm for building virtual dialog between user and system.

Then the system automatically makes both the speech recognition engine and the motion detector to start to get speech and image signals, respectively. If the uttered commands are recognized by speech recognition engine, the system then outputs the hardware control signals as well as the synthesized voices using a speech synthesis. The system with 18 kinds of different commands has been built to be able to control MP3 music player, video player, and several electric appliances such as an electric fan and two kinds of lamps. The mode of recognizing speech is maintained until user leaves the sofa. If the value of difference between the previous and the current brightness in pixel does not exceed the threshold value during the fixed time, a function of speech recognition enters a pause mode since it is regarded that user has left the sofa. In the proposed system, the list of the registered speech recognition candidates can be automatically updated according to the recognition results.

For a motion detection based on video signal processing, first of all, the color images captured from a web camera are converted into a grey scale by a simple average of the colors. The following algorithm is then applied to the converted grey scale images. In this algorithm, I(x,y,t) and I(x,y,t-Δt) represents the functions of brightness in pixels at time t and t-Δt, respectively.

$$if\ (|\ I(x,y,t) - I(x,y,t-\Delta t)\ |> Th1) \quad (1)$$

$$Count\_Motion++ \quad (2)$$

$$Count\_Big\_Motion = \frac{Count\_Motion}{Total\_Num\_Pixels} * 10{,}000 \quad (3)$$

$$if\ (Count\_Big\_Motion > Th2) \quad (4)$$

$$Count\_Big\_Motion++ \quad (5)$$

If the absolute value between two functions exceeds the threshold value Th1, it is estimated that there are motions of objects. In addition, the big motions that exceed the threshold value Th2 are counted to detect relatively big motions. In this algorithm, the threshold values, such as Th1 and Th2, were determined experimentally.

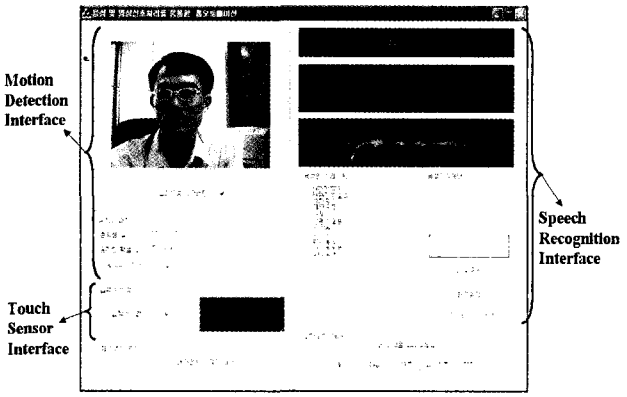
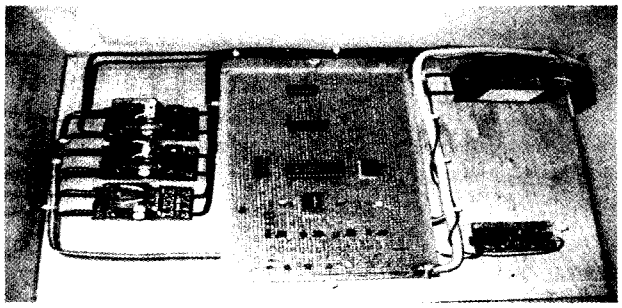


Fig. 5. Main frame of user interface window.

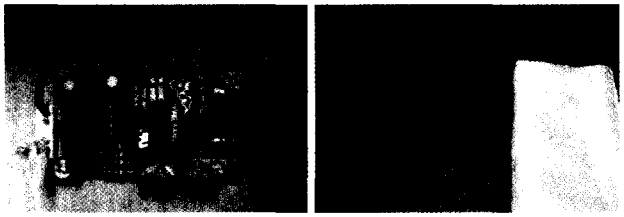
Figure 5 shows the main frame of user interface, which has been made by VC++, with the modules of speech recognizer, motion detector, and sensor signal detector. By utilizing the human interface such as speech recognition and video image processing, the need for a keyboard or mouse can be eliminated in real-world application.

For controlling several electric appliances, we designed

hardware interface for the print-port control shown in figure 6 and 7. The interface receives both the input signals of touch sensor and the result values of speech recognition, which are then given to the corresponding output signals of AC power relay units.



(a) Overall hardware interface



(b) Touch sensor(left) attached on the sofa(right)

Figure 6. Hardware interface for dialog system.

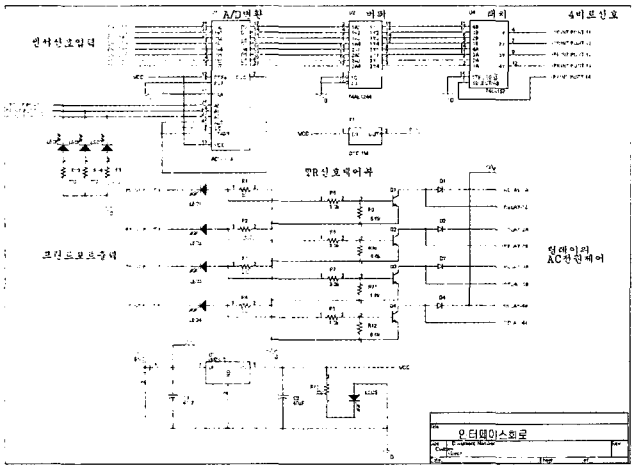


Figure 7. Circuit diagram for both the input of touch sensor and the output of print-port used in the hardware interface for virtual dialog system.

4. Experiments and Discussion

4.1 Speech Database and Signal Preprocessing

Table 1 shows the analysis of speech signals. All speech data were sampled at 16kHz, quantized at 16 bits, pre-emphasized with a transfer function of $(1-0.97z^{-1})$, and processed to extract acoustic features using a 25ms Hamming window with a 10ms shift. The feature parameters consisted of total 39 order LPC MEL Cepstrum coefficients including normalized log-power, 1st and 2nd order delta coefficients.

Table 1. Analysis of speech signals

Sampling rate	16kHz , 16bits
Pre-emphasis	0.97
Window Function	25 ms Hamming window
Frame period	10ms
Feature Parameters	13 order LPC MEL Cepstrum +13 order ΔLPC MEL Cepstrum +13 order ΔΔLPC MEL Cepstrum =Total 39 order LPC MEL Cepstrum

The speech database used in the speaker independent speech recognition consisted of two kinds of database, one of which was made by ETRI(The Electronics and Telecommunications Research Institute), and the other was made by KLE(Center for the Korean Language Engineering). Table 2 shows the database and its contents used in the HMM training and recognition process. When ETRI speech data is used for recognition, it means the speaker-independent and task-dependent condition. When KLE data is used, on the other hand, it means the speaker-independent and task-independent condition.

Table 2. Database used in the module of speech recognition

Process	Database	Content
Training	ETRI	(200 male speakers*280 utterances) + (200 female speakers * 280 utterances) = 112,000 utterances
Recognition	ETRI	(25 male speakers * 100 words) + (25 female speakers * 100 words) = 5,000 words
	KLE	3 male speaker * 452 words = 1,356 word

4.2 Simulation Results of Speech Recognition

For the preliminary experiments, the speech recognition was performed using a frame synchronous Viterbi beam search algorithm with the phonotactic constraint of Korean language using word-pair grammar. Figure 8 and 9 shows the word recognition accuracies, according to the changes of number of HMM mixtures and states, using ETRI and KLE recognition data, respectively. It is noticed in the recognition results that the accuracies grow gradually with an increase of the number of both mixtures and states.

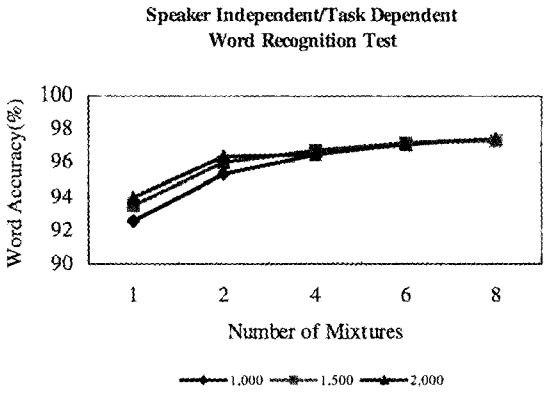


Fig. 8. Speaker-independent/task-dependent word recognition accuracies, according to the changes of number of HMM mixtures and states, using ETRI recognition data.

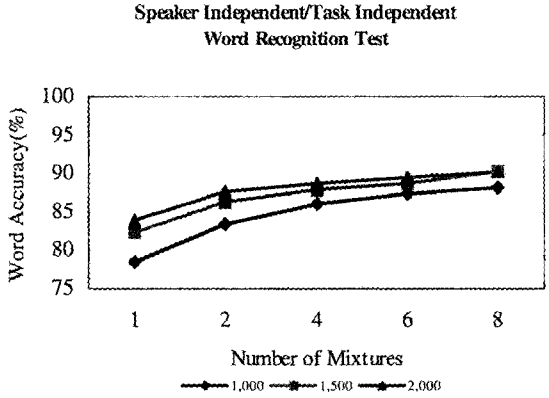


Fig. 9. Speaker-independent/task-independent word recognition accuracies, according to the changes of number of HMM mixtures and states, using KLE recognition data.

4.3 Questionnaire Results of the Proposed System

For speech recognition engine incorporated into the proposed system, we adopted HMM speech recognizer with 2,000 states and 4 mixtures per state, which had nearly equal performance when comparing with the accuracy with 2,000 states and 8 mixtures as illustrated in the preliminary experiments.

For experiments, total 41 male college students were participated in the evaluation of the system. For examining the human performance on the accuracies of the overall system, we first showed them a demonstration of how to use and operate the system, and made them to use it themselves. The evaluation was performed in the laboratory with the noises such as computer cooling fan or buzz of voices. Table 3 shows the average recognition accuracies in each module such as touch sensor, speech recognition, and motion detection. As illustrated in the results, the accuracy of speech recognition was unsatisfactory owing to the noisy environments.

As an evaluation using questionnaire, all participants marked scores from 1 to 5-point about how easy they thought the system was to use. We can see in table 4 and figure 10 that the system was relatively easy to use(average rank is 3.4).

Table 3. Recognition accuracies for touch sensor, speech recognition, and motion detection.

Modules	Accuracies (%)
Touch sensor	100.0
Speech recognition	58.8
Motion detection	100.0
Average	86.3

Table 4. Evaluation of the proposed system, using questionnaire.

Types	Ranks					Sum
	1	2	3	4	5	
Participants	1	6	11	20	3	41
%	2.4	14.6	26.8	48.8	7.3	100
Average Ran	3.4					

Question: Was the system easy to use?
Score: 1(very difficult) 5(very easy)

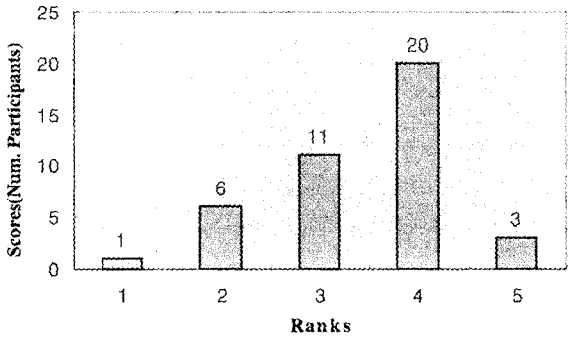


Fig. 10. Scores of the system according to ranks.

As results of the experimental evaluation, we could obtain several ideas on the system as future works. The first thing we should do for future works is that the proposed system is needed to integrate with the interfaces of spontaneous speech recognition, facial or gesture recognition, as natural ways of communication, so that the integrated system would allow users to feel more convenient and natural in virtual conversation for their smart home environments. In addition, it would be greatly significant to attempt to explore a new approach of avoiding misunderstanding in conversation as well as supplementing a function of speech recognition.

5. Conclusion

This paper has described the virtual dialog system based on multimedia signal processing using speech, video and sensor signal processing for smart home environments. The present study aims the interactive home that is more convenient to live a daily life in the living environments. For realizing this, we incorporate the modules of speech recognition and synthesis, video signal, and sensor signal processing into the proposed system. In evaluation, the results presented that the performance of real-time speech recognition in the proposed system was not better than the simulation results owing to the ambient noisy environments. Nevertheless, the results from the questionnaire showed an affirmative possibility for building interactive system that might give us much more convenient and comfortable living environments.

References

[1] J. Machate, *"Being natural - on the use of multimodal interaction concepts in smart homes"*, HCI(2), pp.937-941, 1999.

[2] M. Kohler, "Special Topics of Gesture Recognition Applied in Intelligent Home Environments", *Lecture Notes in Computer Science*, Vol.1371, pp.285-233, 1998.

[3] M. Mozer, "The neural network house: An environment that adapts to its inhabitants", *Proc. AAAI Sym. on Intelligent Environments*, pp.110-114, 1998.

[4] D. J. Cook, M. Youngblood, E. Heierman, K. Gopalratnam, S. Rao, A. Litvin, and F. Khawaja, "MavHome: An Agent-Based Smart Home", *Proc. IEEE Int. Conf. on Pervasive Computing and Communications*, pp.521-524, 2003.

[5] P. M. Corcoran, F. Papai, A.Zoldi, "User Interface Technologies for Home Appliances and Networks", *IEEE Trans. on Consumer Electronics*, August 1998.

[6] M. Suzuki, S. Makino, A. Ito, H. Aso and H. Shimodaira, "A new HMnet construction algorithm requiring no contextual factors," *IEICE Trans. Inf. & Syst.*, Vol. E78-D, No. 6, pp. 662-669, 1995.

[7] M. Ostendoft and H. Singer, "HMM Topology design Using Maximum Likelihood Successive State Splitting", *Computer Speech and Language* Vol. 11, pp. 17-41, 1997.

[8] J. Takami and S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling," *Proc. ICASSP'92*, Vol. 1, pp. 573-576, 1992.

[9] S-J Oh, C-J Hwang, B-K Kim, H-Y Chung, and A. Ito, "New state clustering of hidden Markov network with Korean phonological rules for speech recognition," *IEEE 4th work. on Multimedia Signal Processing*, pp. 39-44, 2001.

[10] S-J Oh, C-J Hwang, B-K Kim, H-Y Chung, "Performance Evaluation of HM-Nets Speech Recognition System using the Large Vocabulary Korean Speech Databases," *Proc. Kyushu-Youngnam Joint Conf. on Acoustics*, pp. 49-52, 2003.

저 자 소 개

Sung-Ill Kim

Feb. 1994 : B.S. from Dept. of Electronics Eng., Yeungnam Univ.
Feb. 1997 : M.S. from Dept. of Electronics Eng., Yeungnam Univ.
Mar. 2000 : Ph.D. from Dept. of Computer Science & Systems Eng., Miyazaki Univ., Japan.
Apr. 2000—Mar. 2001 : researcher at the National Institute for Longevity Sciences, Japan.
Apr. 2001—Feb. 2003 : researcher at the Center of Speech Technology, Tsinghua Univ., China.
Mar. 2003—Current : full-time lecturer at the Div. of Electrical & Electronic Eng., Kyungnam Univ.
research interests : speech/emotion recognition, neural networks, and multimedia signal processing.

Phone : +82-55-249-2632
Fax : +82-55-249-2839
E-mail : kimstar@kyungnam.ac.kr

Se-Jin Oh

Feb. 1996 : B.S. from Dept. of Electronics Eng., Yeungnam Univ.
Feb. 1998 : M.S. from Dept. of Electronics Eng., Yeungnam Univ.
Mar. 2002 : Ph.D. from Dept. of Electronics Eng., Yeungnam Univ.
Sep. 2001—Dec. 2002 : full-time lecturer at the Information & Communication Div., Taegu Science College.
Dec. 2002—Current : senior engineer at the Radio Astronomy Div., Korea Astronomy & Space Science Institute.
research interests : radio signal processing, digital signal processing, speech processing.

Phone : +82-42-865-3280
Fax : +82-42-865-3272
E-mail : sjoh@trao.re.kr