Smart Sofa based on Biometric Pattern Recognition

Sung-Ill Kim
Department of Electronic Engineering
Kyungnam University
Email: kimstar@kyungnam.ac.kr

Se-Jin Oh
Radio Astronomy Division
Korea Astronomy and Space Science Institute

Sang-Yong Rhee
Division of Computer Science and Engineering
Kyungnam University

Seung-Gook Hwang
Department of Industrial Engineering
Kyungnam University

***Abstract*— This paper discusses the user-customized interaction for intelligent home environments. The interactive system is based upon the integrated techniques using both speech and face recognition. For essential modules, the speech recognition and synthesis were basically used for a virtual interaction between user and the proposed system. In experiments, the real-time speech recognizer based on the HM-Net(Hidden Markov Network) was incorporated into the proposed system. Besides, the face identification was adopted to customize home environments for a specific user. In evaluation, the results showed that the proposed system was useful and easy to use for intelligent home environments, even though the performance of the speech recognizer was not better than the simulation results owing to the ambient noisy environments.**

I. Introduction

There are many methods of biometrical identification such as eye iris, retina, voice, face etc. Among them, the face recognition has been one of the most widely used biometrics for personal verification. Its advantage is that it does not require physical contact as well as any advanced hardware. It can be used with existing image capture devices such as web cam, security cameras etc. The system of face identification matches the given input face image with the one stored in its database and a degree of similarity is finally computed. If such score is higher than a certain acceptance threshold, then the person is classified as a one of the registered users. In the present paper, the face identification can be used for the interface of user-customized system in the intelligent home environments [1,2,3,4,5].

When talking about the intelligent home, it means different things to different people. Figure 1 shows the basic idea of the interaction between user and intelligent home environment where the face identification is used for user-customization. The interactive system using face identification is integrated with the essential components such as speech recognition, and speech synthesis. Assuming that we sit on the sofa that is interconnected with both touch sensor and face identification subsystem, the user-customized interaction is then automatically formed so that the intelligent home can provide you with more convenient and comfortable living environments.

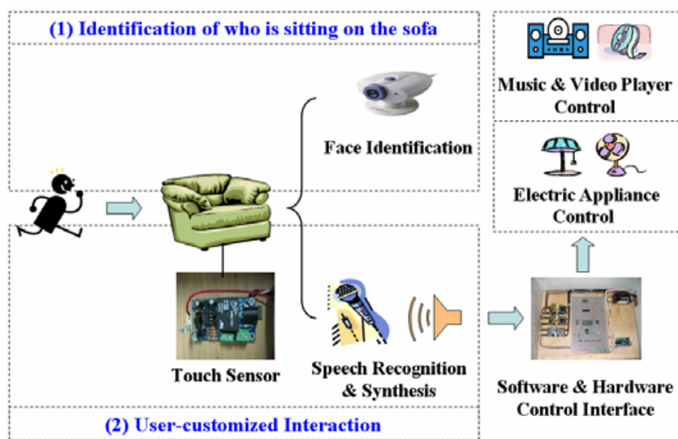

Fig. 1. Concept of the interactive system based on the techniques such as speech recognition, speech synthesis, and fingerprint verification.

The basic idea mentioned above is based on the fact that the place we spend most time at home is our living room, particularly on the sofa. The concept is started on the assumption that the interaction between user and system can be built when user sits on the sofa. The proposed system is designed to allow users to converse with their home based on the user-customized interaction where the system puts emphasis on an easy-to-use and user-friendly man-machine interface. As a consequence, the intelligent home, as an aim of this study, makes it possible to lead the living environments to the most suitable condition for users by integrating speech recognition with face identification. For face identification, in this study, it was used to customize the home environments for a specific user. For speech recognition and synthesis, on the other hand, they were used for interaction between user and intelligent home environment.

II. Interface of Speech Recognition

HMM(Hidden Markov Model) is a mathematical model which has been widely used in speech recognition systems. In this study, we used HM-Net(Hidden Markov Network) [6,7] which is an efficient representation of context-dependent phonemes for speech recognition. The HM-Net, which has various state lengths and shares their states one another, is automatically generated by SSS(Successive State Splitting) [7,8]. The SSS is an iterative algorithm that progressively grows HM-Net, where each state in the network is associated with a 2-component Gaussian mixture.

In the algorithm, a state is selected to be split according to which has the largest divergence between its two mixtures. The state is then split on the contextual and temporal domains, and the one giving greater likelihood is chosen. The affected states are retrained using the Baum-Welch algorithm [9,10]. The above procedure is iterated until getting to a pre-defined number of states.

The PDT-SSS(Phonetic Decision Tree-Successive State Splitting) [11] based on the SSS algorithm 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.

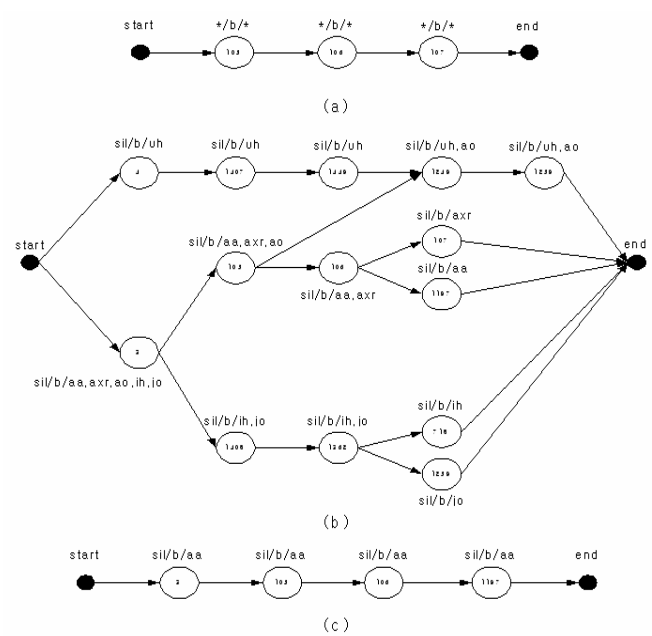

Fig. 2. An example of HM-Net models.

The final result of state splitting is a network of states that efficiently represents a collection of context-dependent models, as illustrated in figure 2. 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 vowels increases more than that of states in consonants in terms of the architecture.

Figure 3 shows an overall schematic of HM-Net speech recognition system. In case speech signals are given to the system, the acoustic features are first extracted for pre-processing, and then given to the first and the second pass search modules that use tree-structured lexicon, HM-Net Triphones, and semantic grammars. The HM-Net speech recognizer has been proved that it produced better performance than the conventional HMM in the experiments of phoneme, word, and continuous speech recognition [12,13].

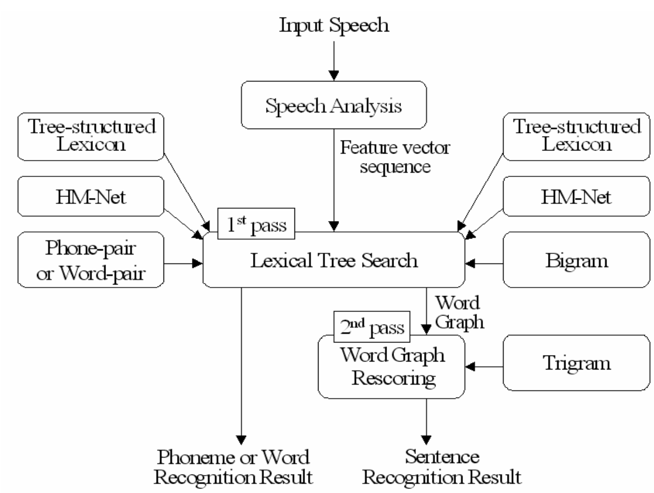

Fig. 3. Overall schematic of HM-Net speech recognition system.

III. Interface of Face Identification

The face identification algorithm implements the advanced face localization, the enrollment and the matching using robust digital image processing algorithms. The interface has two operation modes such as enrollment and matching. It first processes the input face image, extracts features and then writes them to the database. In the mode of face enrollment with features generalization, particularly, it generates the collection of the generalized face features from a number of the face templates of the same person. Each face image is first processed and features are then extracted. In the next step, the collections of features are analyzed and combined into one generalized features collection, which is written to the database. The quality of face recognition increases if faces are enrolled using this mode mentioned above. In the mode of matching, on the other hand, it performs the matching process between the new face image and the face templates stored in the database.

In this study, the interface of face identification was adopted for user-customization of interactive system. In experiments, we used the VeriLook SDK[14] for the interface of face identification. Table **I** shows the technical specification of face identification using a PC with 3GHz Pentium 4 processor.

Figure 4 shows the software interface for face identification module made by VC++. The main application window of the interface has four-pane layout, where two top panes are used for image display and two bottom panes are used for message logging. The face detection pane is used to display the still images, the videos, or the result of face detection algorithm overlaid on image. The matching/enrollment pane is used to display images enrolled to face database or used for matching. The application log pane is used for the system information and the application progress messages. The match results pane is used for listing ID of the subject in the database, most similar to the matched image. Subjects are considered similar if their similarity value exceeds matching threshold set.

TABLE I

TECHNICAL SPECIFICATION OF FACE IDENTIFICATION.

Item	Specification
Recommended minimal image size	640 x 480 pixels
Multiple faces detection time(640 x 480 image)	0.1 seconds
Single face processing time	0.15 seconds
Matching speed	100,000 faces/second
Size of one record in the database	2.3 Kbytes
Maximum database size	unlimited

Fig. 4. Interface of face identification. (num.1: Face detection pane, num.2: Matching/enrollment pane, num.3: Application log pane, num.4: Matching results pane)

IV. THE PROPOSED INTERACTIVE SYSTEM

The proposed system can be built by integrating two main module of both HM-Net speech recognition and face identification, mentioned in the previous chapters. Figure 5 shows the flow diagram of the processing based on the proposed system, which is operated in real time. It shows how to build the interaction between user and system.

If user sits on the sofa, the system catches signals from touch sensor and then activates face identification engine to detect face area. If the system recognizes who is sitting on the sofa, it adapts itself to the new circumstances. The system then activates the speech recognition engine where the virtual interaction between user and system is built using speech recognition and synthesis [15]. In case speech recognition is activated, system provides the user-customized services. It can notify user of the necessary information such as important messages or schedules. In the proposed system, the list of the registered recognition candidates can be automatically updated according to the corresponding recognition results.

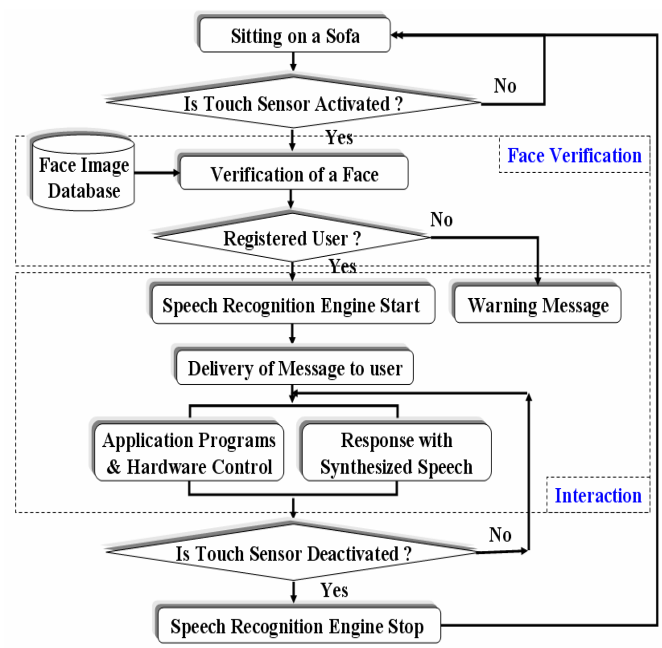

Fig. 5. The flow diagram of building interaction between user and system

Figure 6 shows the main window frame of user interface, which was made by VC++, with the modules of both speech recognition and face identification. The system provides several functions. First, it is possible for user to control multimedia application programs such as video, MP3 player, CD player etc. Besides, several kinds of electrical appliances such as electrical fans, lamps can be controlled by the power relay units of print-port interface.

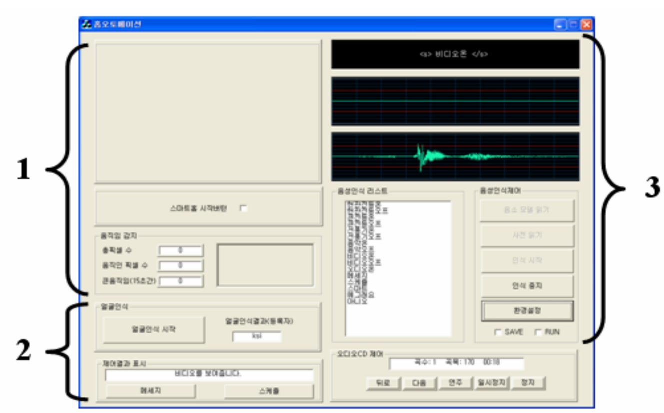

Fig. 6. Main window frame of user interface. (num.1: interface for video processing, num.2: interface of face identification, num 3: interface of speech recognition)

By utilizing the natural human-interfaces such as speech recognition and face identification, the need for a keyboard, mouse, or remote controller can be eliminated in real-world applications. Figure 7 shows the touch sensor, which is one of the components of hardware interface based on print-port control. It was attached on the sofa so that the input signals from the sensor activate or deactivate the system in case someone sits on the sofa.

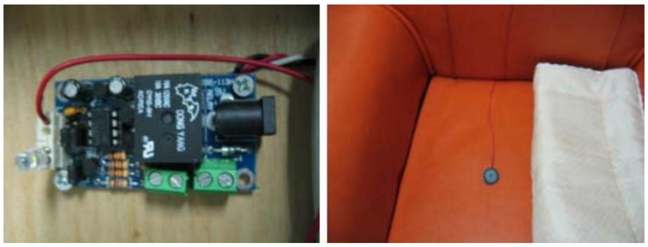

Fig. 7. Touch sensor(left) attached on the sofa(right)

V. Experimental Results

A. Preprocessing for Speech Recognition

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 39th order LPC Mel Cepstrum coefficients including the normalized log-power, the first and the second order delta coefficients. Table Ⅱ shows the analysis of speech signals.

TABLE Ⅱ

Analysis of speech signals.

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

TABLE Ⅲ

Database used in the module of speech recognition.

Process	Database	Contents
Training	ETRI	(200 male speakers*280 utterances) + (200 female speakers * 280 utterances) = 112,000 utterances
Word Recognition	KLE	(3 male speakers * 452 words) = 1,356 words
	YNU	4 male speaker * 200 utterances = 800 utterances

Table Ⅲ shows the speech database and its contents used for both HMM training and recognition process, respectively. For the training process, the database of ETRI(The Electronics and Telecommunications Research Institute) was used. The database used for the recognition, on the other hand, consists of two kinds of database, one of which is made by KLE(Center for the Korean Language Engineering), and the other is made by YNU(YoungNam University).

B. Results based on the Proposed System

For the preliminary experiments, the speech recognition was performed using a frame synchronous Viterbi beam search algorithm with the phonotactic constraint of Korean language. Figure 8 and 9 shows the recognition accuracies for word and continuous speech, respectively. It is noticed in the recognition results that the accuracies, as a whole, grew gradually with the increase of the number of both mixtures and states.

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 proposed system, we first showed them a demonstration of how to use and operate the system, and made them to use it for themselves.

Table Ⅳ shows the average recognition accuracies in each module such as face identification and speech recognition. For the evaluation of speech recognition incorporated into the proposed system, total 738 utterances(41 users * 18 utterances) were used.

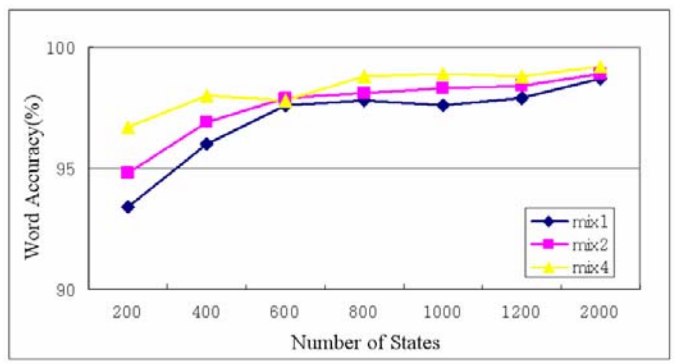

Fig. 8. Speaker and task independent word recognition accuracies according to the number of both mixtures and states using KLE database

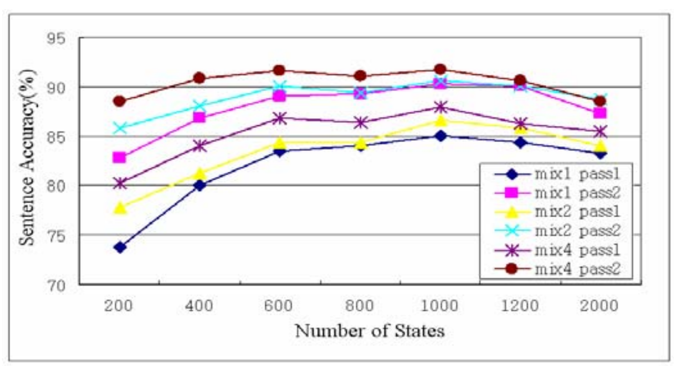

Fig. 9. Speaker and task independent sentence recognition accuracies according to the number of both mixtures and states using YNU database

The evaluation was performed in the laboratory environments with the noises such as computer cooling fan or buzz of voices. In experiments, we adopted speech recognizer with 2,000 states and 4 mixtures per state. For the evaluation of face identification, on the other hand, 41 male college students were first registered in facial image database and the identification test in each user was then conducted.

TABLE Ⅳ

EXPERIMENTAL CONDITIONS FOR SPEECH RECOGNITION AND RECOGNITION ACCURACY.

Module	Accuracy(%)
Face Identification	(40/41)*100=97.6
Speech Recognition	(530/738)*100=71.8

As the evaluation using questionnaire, all participants marked ranks from 1- to 5-point about how easy and how useful they thought the system was to use. We could get the results as shown in figure 10 and 11 that the proposed system was relatively easy to use and would be useful in real applications.

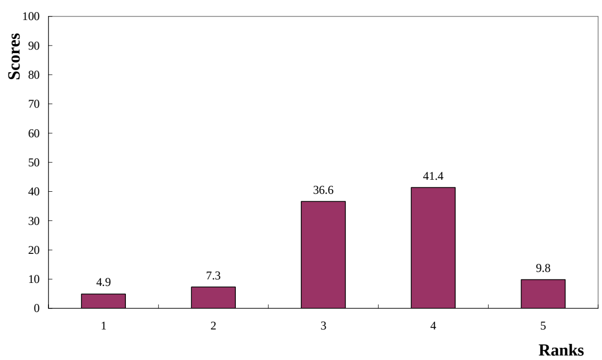

Fig. 10. Evaluation of the system in terms of how easy system was to use.

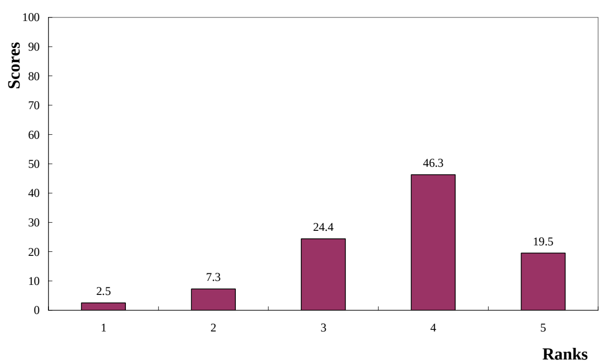

Fig. 11. Evaluation of the system in terms of how useful system would be in real applications.

VI. CONCLUSION

This study has described the user-customized interactive system based on the speech and face recognition for intelligent home environments. The results from the experimental evaluation have shown that the proposed system had relatively good performance. This means a possibility for building a virtual interaction using the system that might give us much more convenient and comfortable living environments. However, the accuracy of speech recognition was unsatisfactory owing to the noisy environments, diverse speaking rates, and speaking styles of users. From the evaluation, we could obtain several ideas on the system as future works. Namely, the future works should be conducted on more natural methods of interaction so that the future systems would allow users to feel more natural in virtual interaction for intelligent home 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. of the AAAI Spring Symposium on Intelligent Environments*, pp. 110-114, 1998.

[4] D.J. Cook, M. Youngblood, E. Heierman, K. Gopalratnam, S. Rao, A. Litvin, F. and Khawaja, “MavHome: An Agent-Based Smart Home”, *Proc. of the IEEE International Conference 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 Transactions on Consumer Electronics* 1998.

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

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

[8] J. Takami, S. Sagayama, “A Successive State Splitting Algorithm for Efficient Allophone Modeling”, *Proc. of ICASSP’92*, vol. 1, pp. 73-576, 1992.

[9] X. Huang, A. Acero, H.W. Hon, *Spoken language processing: a guide to theory, algorithm, and system development*, Prentice Hall, 2001.

[10] L. Rabiner, B.H. Juang, *Fundamentals of speech recognition*, Prentice-Hall International, Inc., 1993.

[11] T. Hori, M. Katoh, A. Ito, M. Kohda, “A Study on HM-Nets using Decision Tree-based Successive State Splitting”, *Proc. of ICSP’97*, pp. 383-387, 1997.

[12] O. Se-Jin, H. Cheol-Jun, K. Bum-Koog, C. Hyun-Yeol, I. Akinori, “New state clustering of hidden Markov network with Korean phonological rules for speech recognition”, *IEEE 4th workshop on Multimedia Signal Processing*, pp. 39-44, 2001.

[13] O. Se-Jin, H. Cheol-Jun, K. Bum-Koog, C. Hyun-Yeol, “Performance Evaluation of HM-Nets Speech Recognition System using the Large Vocabulary Korean Speech Databases”, *Proc. of Kyushu-Youngnam Joint Conference on Acoustics*, pp. 49-52, 2003.

[14] VeriLook SDK(Software Developer’s Kit) version 2.0, Neurotechnologija,. Web Site: http://www.neurotechnologija.com/

[15] i-Talk SDK version 2.2, SL2 Corporation. Web Site: http://www. slworld.co.kr/