다중신호처리를 이용한 인터렉티브 시스템

김성일, 양효식, 신위재, 박남천, *오세진
경남대학교 전자전기공학부, *한국천문연구원 전파천문연구부 KVN사업본부

Interactive System using Multiple Signal Processing

Sung-Ill Kim, Hyo-Sik Yang, Wee-jae Shin, Nam-chun Park, *Se-Jin Oh
Division of Electronic and Electrical Engineering, Kyungnam University,
*Radio Astronomy Division, Korea Astronomy and Space Science Institute

Abstract

This paper discusses the interactive system for smart home environments. In order to realize this, the main emphasis of the paper lies on the description of the multiple signal processing on the basis of the technologies such as fingerprint recognition, video signal processing, speech recognition and synthesis. For essential modules of the interactive system, we adopted the motion detector based on the changes of brightness in pixels as well as the fingerprint identification for adapting home environments to the inhabitants. In addition, the real-time speech recognizer based on the HM-Net(Hidden Markov Network) and the speech synthesis were incorporated into the overall system for interaction between user and system. In experimental evaluation, the results showed that the proposed system was easy to use because the system was able to give special services for specific users in smart home environments, even though the performance of the speech recognizer was not better than the simulation results owing to the noisy environments.

I. Introduction

The term 'smart home'[1,2,3] or the home of the future means different things to different people. Generally speaking, smart home refers to a house with networked products that can interact with each other and with house settings(e.g. 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 users and their living environments at home.

In order to enhance a quality of life, this study aims the development of interactive system for smart home environments where home is possible to converse with inhabitants as users, just like friends or family members. The proposed system can be realized by making a use of interactive technology, between user and smart home, mainly based on the multiple signal processing including vision, speech, and fingerprints.

Since the proposed system puts emphasis on an easy-to-use and user-friendly man-machine interface, in this paper, we made a use of the integrated system based on multiple signal processing. Why should we be interested in multimedia interfaces of multiple signal processing technologies for building smart home environments? In our daily life, we use voice, gesture, vision and tactile channels in combination. Although human beings have a natural facility for managing and exploiting multiple input and output media, computers do not. Consequently, providing computers with multimedia interfaces, which have the ability to interpret multiple signal inputs and generate coordinated multiple signal output, would be a valuable facility for key application such as interactive system.

II. HM-Net Speech Recognizer

Recently, the 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)[4,5] 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).

As the chief idea of HM-Net, 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 is a network of states that efficiently represents a collection of context-dependent models. 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.

Figure 1 shows an overall schematic of real-time HM-Net speech recognition system. In case speech signals are given to the system, acoustic features are first picked out for pre-processing, 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.

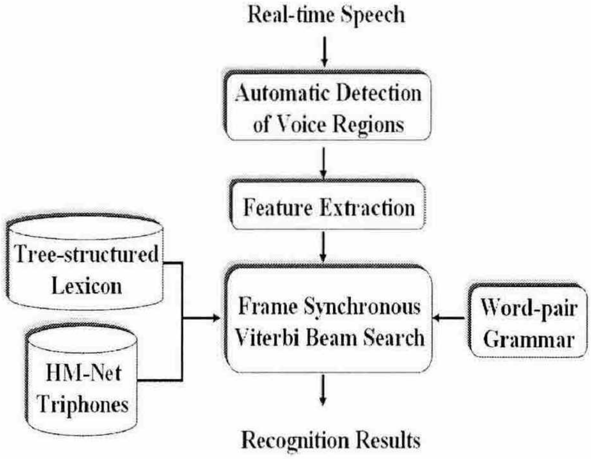


Fig.1. Overall schematic of HM-Net speech recognition system with a real-time operation

III. The Proposed Interactive System

The figure 2 shows the flow diagram of the processing based on the proposed system operated in real time. It illustrates how to build the interaction between user and system. When user gets into the home through an identification of his or her fingerprint, the system presents a greeting with a synthesized speech and simultaneously operates the motion detector. In case user sits in a sofa located in a living room, relatively big motions are detected to start speech recognition engine. The interaction between user and system is then built using speech recognition and synthesis. The mode of speech recognition is maintained during motions are detected. 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.

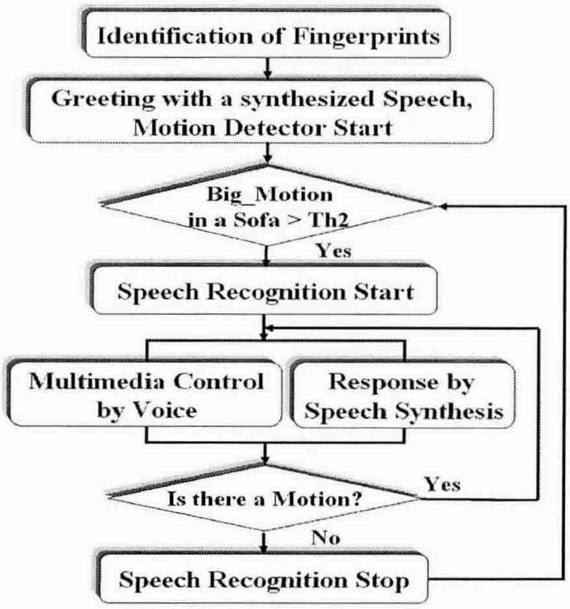
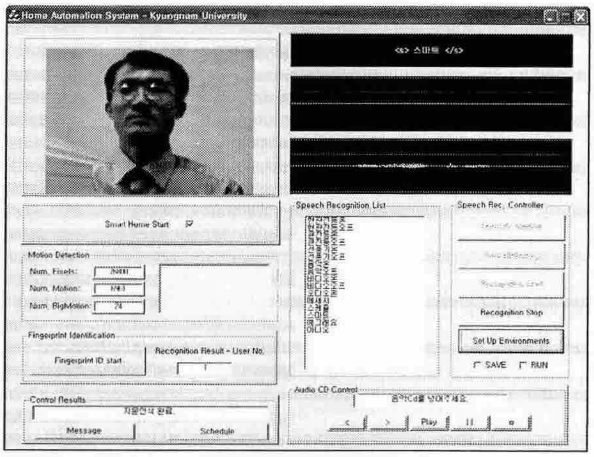

Identification of Fingerprints
Greeting with a synthesized Speech, Motion Detector Start
Big_Motion in a Sofa > Th2
Yes
Speech Recognition Start
Multimedia Control by Voice
Response by Speech Synthesis
Is there a Motion?
Yes
No
Speech Recognition Stop

Fig.2. Algorithm for building an interaction between user and system.

In the motion detection based on video signal processing, the color images captured from a web camera are converted into a grey scale by a simple average of the colors. The big motions that exceed the threshold value, $Th2$, are then counted to detect whether there are relatively big motions.

Figure 3 shows the main frame of user interface, which has been made by VC++, with the modules of speech recognizer, motion detector, and fingerprint identification. By utilizing the multiple signal processing such as speech recognition, video signal processing, and fingerprint identification, the need for a keyboard or mouse can be eliminated in real-world applications. In experiments, we used the Fingkey Hamster of NITGEN for fingerprint input device as well as biometric service provider SDK[6] for software interface.


Home Automation System - Kyungnam University

<s> 스마트 </s>

지문인식 완료.

음악Cd를 넣어주세요.

Fig.3. Main frame of user interface window.

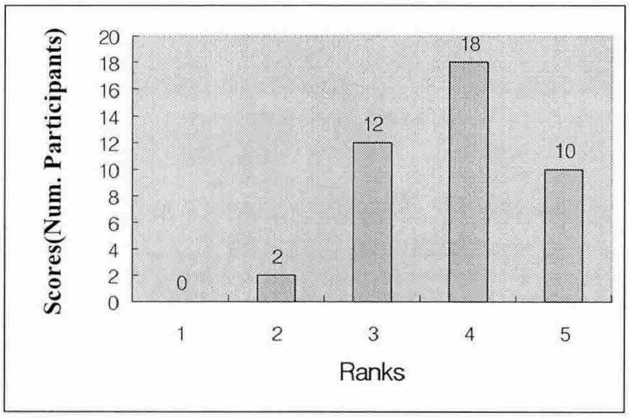
Ⅳ. Experiments and Discussion

All speech data for speech recognition 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.

For the training process of acoustic models, we used ETRI database of 112,000 utterances spoken by 200 male and 200 female speakers. The acoustic models used in the HM-Net speech recognizer consisted of 2,000 states and 4 mixtures per state.

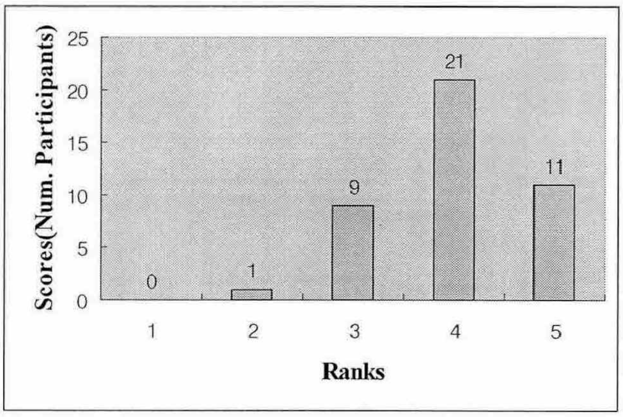
For evaluation, total 42 male college students were participated in the evaluation of the system. For examining the human performance on the accuracies of the overall system, we first showed them a demonstration of how to use and operate the system, and made them to use it themselves. Table 1 shows the procedure of experiments and checking points in each case. The evaluation was performed in the laboratory environment with the noises such as computer cooling fan or buzz of voices

As an evaluation based on a questionnaire, all participants marked scores from 1- to 5-point about how easy and how useful they thought the system was to use, respectively. As indicated in figure 4, we can see that the system is relatively easy to use(average rank is 3.9) as well as useful in real applications(average rank is 4.0).



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

(a)



Question: Do you think the system will be useful?
Score: 1(No) 5(Very useful)

(b)

Fig.4. Evaluation of the system

V. Conclusion

This paper has described the interactive system based on the multiple signal processing technologies including speech, vision, and fingerprints, for smart home environments. The present study aims the interactive home that will be more convenient for the future living environments. For realizing this, we integrated the modules of speech recognition, speech synthesis, fingerprint identification and video signal processing into the proposed system. In evaluation, the results presented that the performance of real-time speech recognition in the proposed system was unsatisfactory than we have expected. It is mainly due to the ambient noisy environments, diverse speaking rates, and speaking styles of users. Nevertheless, the results from questionnaire showed a positive possibility for building interactive system that might give us much more convenient and comfortable living environments in the near future.

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. ofthe AAAI Spring Symposium on Intelligent Environments, pp.110-114, 1998.

[4] 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

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

[6] NITGEN Biometric Service Provider SDK(Software Developer's Kit) version 4.01, NITGEN Corporation. Web Site: http://www.nitgen.co.kr/.