음성인식과 얼굴인식을 사용한 사용자 환경의 상호작용

User-customized Interaction using both Speech and Face Recognition

Sung-Ill Kim[1], Se-Jin Oh[2], Sang-Yong Lee[3], Seung-Gook Hwang[4]

[1] Department of Electronic Engineerin, Kyungnam University
E-mail: kimstar@kyungnam.ac.kr
[2] Radio Astronomy Division, Korea Astronomy and Space Science Institute
[3] Division of Computer Science and Engineering, Kyungnam University
[4]Department of Industrial Engineering, Kyungnam University

요 약

In this paper, we discuss the user-customized interaction for intelligent home environments. The interactive systemis 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 proposed system. In experiments, particularly, the real-time speech recognizer based on the HM-Net(Hidden Markov Network) was incorporated into the integrated 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 easy to use for intelligent home environments, even though the performance of the speech recognizer did not show a satisfactory results owing to the noisy environments.

Key Words : Smart Home, Speech Recognition, HMM, Face Recognition

1. Introduction

Currently, 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.

When talking about the intelligent home, it means different things to different people. 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 subsystem of face identification, the user-customized interaction is then automatically formed so that the intelligent home can provide you with more convenient and comfortable living environments.

The basic idea 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.

2. Interface for Speech Recognition Using Hm-net

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)[1,2] which is an efficient representation of context-dependent phonemes for speech recognition. The HM-Net, which has various state lengths and sharestheir states one another, is automatically generated by SSS(Successive State Splitting)[2,3]. 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[4]. The above procedure is iterated until getting to a pre-defined number of states.

The Phonetic Decision Tree-Successive State Splitting[5] 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.

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 vowels increases more than that of states in consonants in terms of the architecture.

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[6,7].

3. Interface for Face Identification

The face identification algorithm implements advanced face localization, enrollment and 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 writes them to the database. In the mode of face enrollment with features generalization, particularly, it generates the generalized face features collection from a number of the face templates of the same person. Each face image is processed and features are extracted. Then 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 new face image and face templates stored in the database.

In this study, the interface for face identification was adopted for a user customization of interactive system. In experiments, we used the VeriLook SDK[8] for the interface of face identification.

4. 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 1 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.

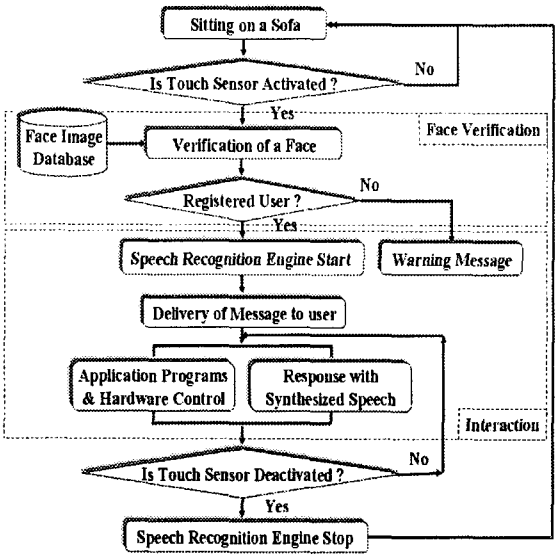

Fig. 1. The flow diagram of building interaction between user and system.

If user sits on 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[9]. In case speech recognition is activated, system provides the user-customized services. It can notify the user of 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.

Figure 2 shows the main window frame of a user interface, which has been made by VC++, with the modules of 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 unitsof print-port interface.

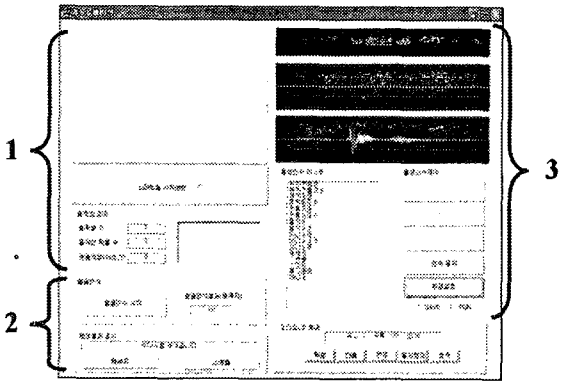

Fig. 2. Main window frame of a user interface. (num.1: interface for video processing, num.2: interface for face identification, num 3: interface for speech recognition)

5. Experimental Result

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 the normalized log-power, the first and the second order delta coefficients. For the training process, the database of ETRI(400 speakers*280 utterance = 112,000 utterance) was used.

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 1 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. 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 1. 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 table 2 that the proposed system was relatively easy to use in real applications.

Table 2. Evaluation of The Proposed System Using Questionnaire (Question: Was The System Easy to Use?, Score: 1(Very Difficult) 5(Very Easy)).

Types	Ranks					Sum
	1	2	3	4	5	
Scores	2	3	15	17	4	41
%	4.9	7.3	36.6	41.4	9.8	100

6. 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.

Reference

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

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

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

[4] L. Rabiner, and B.H. Juang, "Fundamentals of speech recognition," Prentice-Hall International, Inc. 1993.

[5] T. Hori, M. Katoh, A. Ito and M. Kohda, "A Study on HM-Nets using Decision Tree-based Successive State Splitting," Proc. of ICSP'97, Vol.1, pp. 383-387, 1997.

[6] Se-Jin Oh, Cheol-Jun Hwang, Bum-Koog Kim, Hyun-Yeol Chung, and Akinori Ito, "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.

[7] Se-Jin Oh, Cheol-Jun Hwang, Bum-Koog Kim, Hyun-Yeol Chung, "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, Japan, 1. 2003.

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

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