An Experimental Multimodal Command Control Interface for Car Navigation Systems

Kyungnam Kim*, Jong-Gook Ko*, SeungHo Choi**, JinYoung Kim*** and KiJung Kim****

* Department of Information and Communications, Kwangju Institute of Science and Technology(K-JIST)
1 Buk-ku O-Ryong-dong, KwangJu 500-712, Korea
Tel. +82-62-970-2293, Fax +82-62-970-2204
E-mail:knkim@kjist.ac.kr*

** Department of Information and Communications, Dongshin University
*** Department of Electronics, Chonnam University
**** Department of Internet Information Technology, Kwangyang College

Abstract: An experimental multimodal system combining natural input modes such as speech, lip movement, and gaze is proposed in this paper. It benefits from novel human-computer interaction (HCI) modalities and from multimodal integration for tackling the problem of the HCI bottleneck. This system allows the user to select menu items on the screen by employing speech recognition, lip reading, and gaze tracking components in parallel. Face tracking is a supplementary component to gaze tracking and lip movement analysis. These key components are reviewed and preliminary results are shown with multimodal integration and user testing on the prototype system. It is noteworthy that the system equipped with gaze tracking and lip reading is very effective in noisy environment, where the speech recognition rate is low, moreover, not stable. Our long-term interest is to build a user interface embedded in a commercial car navigation system (CNS).

1. Introduction

Humans acquire information from different sources of sensing mode and show their intention through several output modalities. Regarding human-computer interface, so far, conventional input devices such as keyboard, mouse, and touch screen have dominated computer interfaces. These devices have grown to be familiar but tend to restrict the information and command flow between the user and the computer system. So new modalities such as gesture, facial expression, speech, gaze, and lip reading have emerged along with recent technological advances, to resolve the interaction limitation while communicating and commanding. An integrated multimodal human-computer interface may overcome bottleneck of communication between man and machine and make their interaction easy. In this paper, we present an experimental multimodal human-computer interface that incorporates speech recognition, face tracking, gaze tracking, and lip reading. Analysis and integration of individual sensing modalities have been investigated through preliminary experiments.

Our long-term interest is to build a user interface embedded in a commercial car navigation system (CNS). Currently, keypad with touch screen is most popular as an interface for CNS and it is obvious that non-intrusive type input will be useful in implementing instrumentation safe for driving. Further, multimodality may be beneficial to improve the command recognition quality, addition to the speech-based single modality as a non-intrusive type input, under a practical odd environment including engine noises and surrounding interference.

Reported multimodal HCI systems include Bolt's system "Put-That-There", QuickSet, Jeanie, ShopTalk, VisualMan, Finger-Pointer, CUBRICON, Virtual World, ALIVE, Smart Rooms, Neuro Baby, and MDScope. Several promising direction toward multimodal human-computer interaction (HCI) were examined considering some of the emerging novel input modalities for HCI and the fundamental issues in integrating them at various levels by Sharma et al.[1]. They mentioned, in audio sensing, visual sensing modalities such as gaze and lip motion may help in improve speech interpretation. Pastoor[2] and Waibel[3] also reported their multimodal HCI systems.

Up to now, integration of modalities in this field has been pursued as a combination of at least two modes. There has been more research on speech recognition than others such as gaze tracking and lip reading, and its reliable accuracy in the environment without noise has been reported. Our system takes three different human action modalities: speech, lip movement, and gaze, and incorporates them into analysis of user intention with special emphasis on intelligent interpretation of speech recognition with the help of gaze and lip movement information. This circumstance is more imperative inside the car environment, in which speech information may be difficult to recognize due to auditory noise. The analysis of results of three sensing modalities can also be performed in the form of simultaneous integration instead of depending more on speech recognition, when reliability of each performance varies unexpectedly. So far, a PC-based prototype system with a video camera, a frame grabber, a microphone, and a sound card has been developed and tested. The overview of system architecture is illustrated in Fig. 1.

2. System Overview

The overall system, we have implemented, consists of components of speech recognition, face tracking, gaze

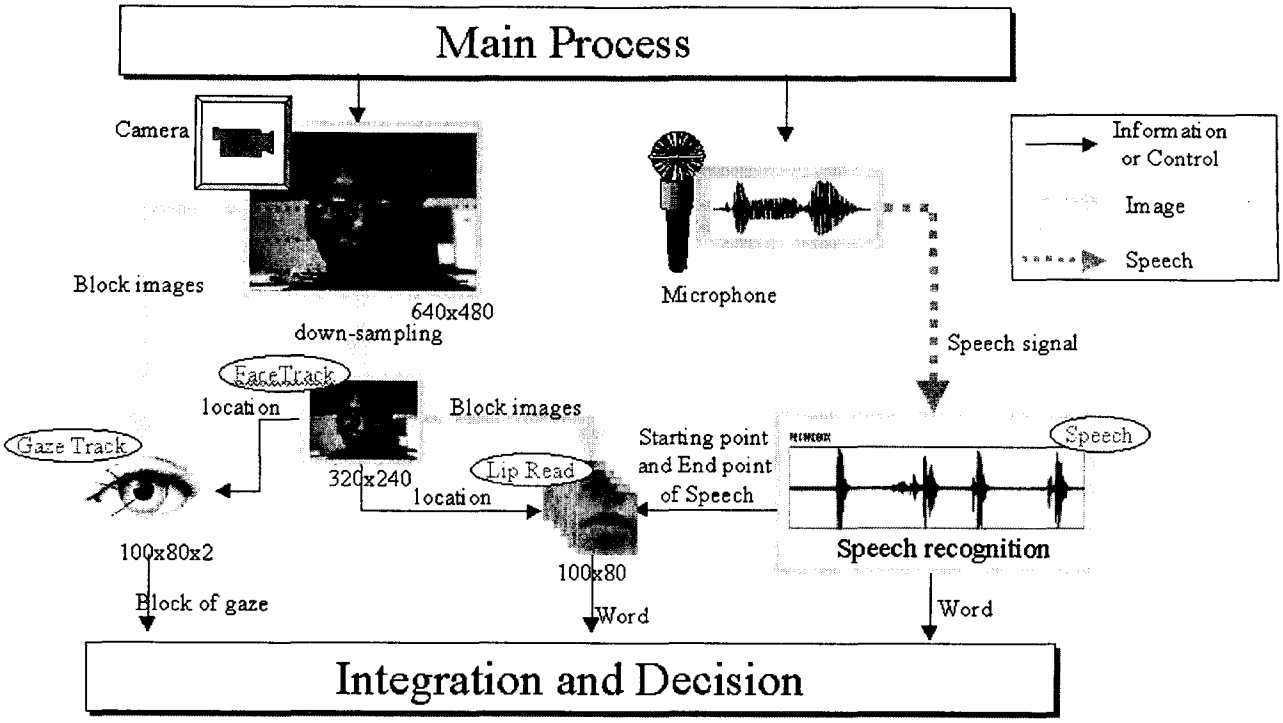

Fig. 1 Overview of system architecture. Four key components are incorporating in parallel in order to interpret a user's intention automatically. This is implemented by multithread programming.

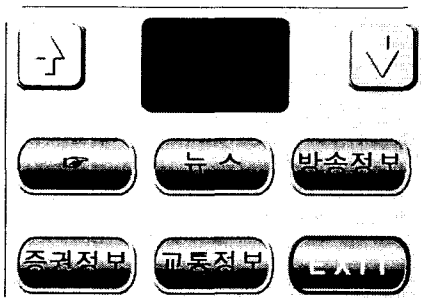

Fig. 2 Display with full screen menu. A user selects one of these menu items.

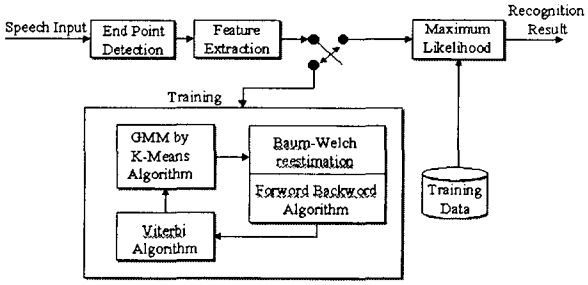

Fig. 3 Block diagram of speech recognition.

tracking, and lip reading. The main function of the system is to select a menu item out of those displayed on the screen when a user speaks the menu name, simultaneously while gazing at the box of that item, by analyzing the user's intention. The inputs are continuous sets of digitized speech and gray-level facial image.

Fig. 2 shows an example of the menu screen, where 6 Korean menu items and 2 up/down items are displayed as icons to be recognized. Basically, the screen is the starting window, and subsequent menu items are hierarchically stored in a tree structure, screen by screen. Further, to jot down the final target item, a few consecutive repetitions of recognition are necessary, as we proceed to the screen including the target item.

For non-intrusive interaction, a microphone is placed properly around a user in order to get clear voice, and a video camera is mounted at the front of display to take images of the user's face including eyes and lip. The overview of system architecture is illustrated in Fig. 1. Face tracking component locates eyes and lip and supports gaze tracking component and lip reading component by giving the location information and sub-images to them. Speech recognition component communicates with lip reading in order to provide the starting and ending point of a speech command. These four components are executed in parallel in the form of threads on a single operating system. Each recognition output weighted by its accuracy is properly combined for constituting the final decision of command selection. These substantial components are described in detail in the following sections.

3. Multimodal Components

3.1 Speech Recognition [4]

Our system employs the speech recognition as a main component due to its higher accuracy of word recognition rate rather than other modes. However, in this paper, we do not focus on its new algorithms, rather, integration of speech recognition with other modes and construction of the system allowing multimodal inputs are addressed. It was implemented using well-known speech recognition algorithms and its roles and

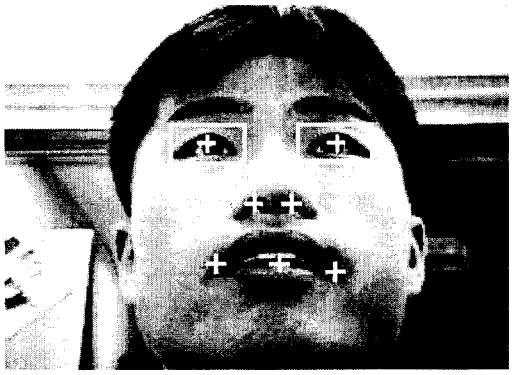

Fig. 4 Result of face tracking - Two rectangles are rough eye regions. The crosses are eyes, nostrils, and mouth center/corners

communication in the system are described hereafter. Building of speech databases including menu items used in our prototype system is also briefly described.

A block diagram of speech recognition is depicted in Fig. 3. The input of speech recognition is a user's speech command of the menu name, and the output is those items sorted by similarity. In preprocessing, the lowpass filtered speech signal is sampled at 8kHz with 16bit quantization. Zero-crossing rate and energy decide the threshold values for the end-point detection in real-time.

The samples are blocked into overlapping frames of 25 msec in duration, where the overlap is set to 10 msec. Each frame is weighted with a Hamming window and then processed by using a 12th-order LPC analyzer. The LPC coefficients are then converted to Mel-cepstrum, where only the first 12 coefficients are retained for further analysis. The feature parameters utilized are 12 Mel-cepstrums, 12 delta-cepstrums, one energy, and one delta energy. The speech recognition component is basically an isolated word recognizer employing continuous Hidden Markov Model(CHMM). For the simplicity of analysis, we presume that the observation probability be governed by the Gaussian mixture.

The speech recognition component in the prototype system covers 22 menu items(command words) and utilizes 8 states, 3 mixtures, and 50~200 frames. These 22 menu names are those in the menu of the experimental system. 50 men and women recorded each word once and 20 of them twice, to build not only speech databases but also lip image files for lip reading. The comparison results of experiment of training data and test data were good and applicable to the practical system. Regarding the processing speed, the required time is less than 0.3 second per word and our algorithm takes 0.1~0.32 second in the PC-based prototype system. Speech recognition component also provides information of starting and end points of speech to the lip reading component, which utilizes it for analysis of lip images.

3.2 Face Tracking [5]

The location information extracted during face tracking is input to both gaze tracking and lip reading: the locations of the eyes and the lip, respectively. It is the first step of video-based components, and subsequently,

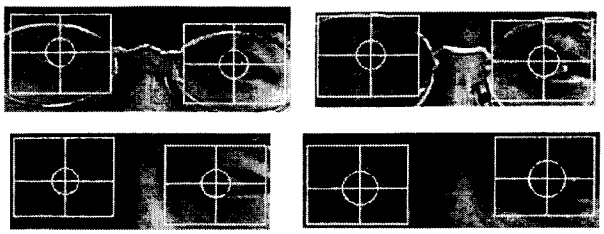

Fig. 5 The precise positions of the eye centers are tracked. The rectangles are the search windows.

it significantly affects the speed of gaze tracking and lip reading.

In our system, we basically employ a feature-based approach to use the information that the iris and the pupil are darker than any other features except hair for locating the facial features, and that there are specific geometrical relationships between facial features. We describe the method of locating facial features and verification in this section. We assumed that the user is in front of computer. At first, the eyes is located using the dark information and geometrical information of the eyes. Also the mouth is located using the information of eyes' position and size. Finally, the nostrils are located using the information of eyes' and mouth's position and size. Fig. 4 depicts the result of face tracking.

3.3 Gaze Tracking

After locating the eyes from the facial images, the eye movement information can be utilized to measure the gazing. By tracking the direction of gaze of the user, the command information from the user to the computer can be interpreted as what the user is looking at, and even designing objects specially intended for the user to look at.

The role of gaze tracking is to estimate the user's current screen block or menu item of fixation by analyzing eye movement information. It starts from rough eye location that is the result of face tracking. The location of face and eyes should be known for tracking eye movement. The investigation of this section concentrates mainly on tracking eye movement. The primary goals are to detect the exact eye position(the iris center) and to decide the screen block or menu item of gaze(intention).

Rough eye position is not sufficient for tracking eye-gaze accurately. Measuring the direction of visual attention of the eyes requires more precise data from eye image. It was decided to track the `iris'. Because the sclera is light and the iris is dark, this boundary can easily be optically detected and tracked. There are some issues, however, which have to be emphasized. They arise, due to the following reasons: coverage of the top and bottom of the limbus by the eyelids, poor quality of the images, and excessive coverage of the eyes by eyelids (in some cases). The techniques proposed in this section effectively deal with the first two, while the last is an inherently hard problem.

After getting the precise eye center, the block/menu item of gaze can be extracted from the current setup and eye movement information. Most hard and important problem on this task is that there is more or less head movement along with eye movement. This

head movement should be compensated by proper manners in order to get real gazing. In previous work, they have dealt with eye movement only or restricted the head mobility. Our approach allowing head movement is quite simple and sufficient for our application, in spite of disadvantage of low resolution.

3.4 Lip Reading [6]

In a noisy environment such as the inside of a car, the performance of speech recognition is not satisfactory. In this regard, lip reading can assist speech processing and improve its accuracy through incorporation, although maximal precision of lip reading is less than 50 percent. One of its roles is to limit a set of menu items, among all to be recognized by speech recognition, in order to enhance speed and performance.

Algorithms for region of interest (ROI) detection, lip parameter detection, and word recognition using HMM are presented in this section. It is commonly observed that visual information provides a precious help to the listener under degraded acoustical conditions. Visual cues are effectively used by human beings to improve speech intelligibility.

First, a lip image is taken from face tracking. lip parameters ,in the mouth region, consist of (1)lip width, (2)outer lip height, and (3)inner lip height. The processing is composed of three stages: ROI detection based on lip width, continuous lip parameter detection, and command word recognition using HMM algorithm.

4. Preliminary Results

We test our system with some users. The performance of each component is like this. That of speech recognition is 98% under no noise. When the noise level is high, the rate of the accuracy drops significantly. As for gaze tracking, the screen resolution is from 3x3 to 4x5. Lip reading has maximum 50% under normal condition.

The whole components are operating in parallel. Advantages of integration are followings: 1) Under noisy environment, the recognition of user's command is benefited from gaze tracking and lip reading components. 2) Integrated modes show better performance of accuracy and robustness than individual one. 3) Lip reading can also give information of the starting point and the end point of the speech.

However, some issues should be considered for real application systems. First, it is needed to enhance the speed of each component. There are two ways to tackle this: one is to upgrade the algorithm of each component. The other is to allocate sub-treads of the components in the program effectively. Second, enhancing recognition rate itself is required for real application. Third, the prototype system will be embedded into CNS. Migration from current system to the other should be carefully prepared in advance (for example, source code conversion).

5. Conclusions

Experimental multimodal system incorporating speech recognition, face tracking, gaze tracking, and lip reading has been developed to enable non-conventional and visual/auditory interaction. The system is not cumbersome and expensive, but require more investigation on real applications. A PC-based prototype has been tested to some users. The results show that visual information assist speech recognition under noisy environment and that the combined result by multimodal integration is better than the individual one in interpreting user's intention.

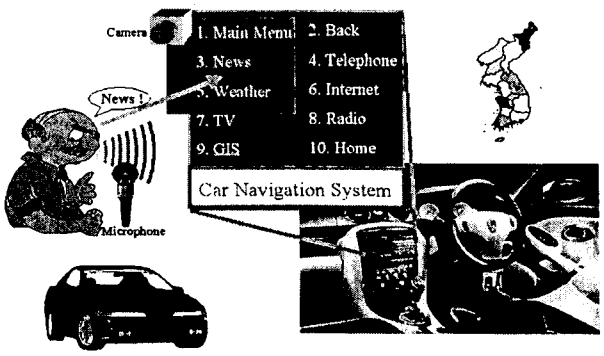

Fig. 6 Final application of the experimental multimodal system in CNS.

Our main goal is to build CNS in which our multimodal interfaces are embedded. Fig. 6 depicts our final product. The multimodal interfaces will be embedded in CNS. We will concentrate future work on enhancing the overall system speed, performance, and robustness. Optimal integration of multimode should be considered through many different approaches. Commercialization will come true if the system could be produced at low-cost and successfully embedded into CNS.

References

[1] Rajeev Sharma, Vladimir I. Pavlovic, and Thomas S. Huang, Toward Multimodal Human-Computer Interface, Proceedings of IEEE., Vol. 86, No. 5, pp. 853-869, May 1998.

[2] Siegmund Pastoor, Jin Liu, and Sylvain Renault, An Experimental Multimedia System Allowing 3-D Visualization and Eye-Controlled Interaction Without User-Worn Devices, IEEE Transactions on Multimedia.}, Vol. 1, No. 1, pp. 41-52, March 1999.

[3] Alex Waibel, Minh Tue Vo, Paul Duchnowski, and Stefan Manke, Multimodal Interfaces, Artificial Intelligence Review., Vol. 10, pp. 299-319, 1996.

[4] Seung-Ho Choi, Kwang-Kook Choi, Kyungnam Kim, Jin-Young Kim, and Ki-Jung Kim, Implementation of Speech Recognition System Using JAVA Applet, Conf. Proc. of ITC-CSCC'2000, July 2000.

[5] Jong-Gook Ko, Kyungnam Kim, SeungHo Choi, JinYoung Kim and KiJung Kim, Facial Feature Tracking and Head Orientation-based Gaze Tracking, Conf. Proc. of ITC-CSCC'2000, July 2000.

[6] DukSoo Min, JinYoung Kim, KyungNam Kim, SeungHo Choi, and KiJung Kim, Stability of Lipreading against the Variations of Rotation, Transition and Scaling, Conf. Proc. of ITC-CSCC'2000, July 2000.