A Multi-Modal System ICANDO: Intellectual Computer AssistaNt for Disabled Operators

Alexey Karpov[1], *Andrey Ronzhin*[1] *and Alexandre Cadiou*[2]

[1]Speech Informatics Group, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, SPIIRAS, Saint-Petersburg, Russia

{karpov,ronzhin}@iias.spb.su

[2]Advanced Institute of Electronics of Paris, ISEP, Paris, France

alexandre.cadiou@isep.fr

Abstract

The paper describes a multi-modal system ICANDO (an Intellectual Computer AssistaNt for Disabled Operators) developed by Speech Informatics Group of SPIIRAS and intended for assistance to the persons without hands or with disabilities of their hands or arms in human-computer interaction. This system combines the modules for automatic speech recognition and head tracking in one multi-modal system. The developed system was applied for hands-free work with Graphical User Interface of MS Windows in such tasks as Internet communication and work with text documents. The architecture of the system, the methods for recognition and tracking, information fusion and synchronization, experimental conditions and the obtained results are described in the paper.

Index Terms: multi-modal interfaces, assistive technologies

1. Introduction

Many people are unable to operate a personal computer by a standard computer mouse or a keyboard because of disabilities of their hands or arms. One possible alternative for these persons is a multi-modal system, which allows controlling a computer without the traditional control devices, but using: (1) head (or face) motions to control the mouse cursor on the monitor screen; (2) speech input for giving the control commands. Usage of such systems allows supporting the equal participation and socio-economic integration of people with disabilities in the information society and increase their independence from other people.

Disability may affect also the person's neck and head motions along with the hands and the arms. Thus a human can have problems with activity of neck and hence reduced ability to move the head in one or more directions. In many of such cases an eye tracking system can be successfully used instead of head tracking system. Moreover the blinking of the eyes can give the mouse button click signal. However, the usage of eye tracking systems is worse than head tracking systems in such parameters as: task performance, human’s workload and comfort both for untrained and experienced users [1]. Of course, the speech input is only one acceptable alternative to the keyboard for motion-impaired operators which cannot move their hands.

In the following sections of the paper, ICANDO assistive multi-modal system, which uses the head motions tracking for mouse cursor control on a monitor screen and the automatic speech recognition to press the buttons of a keyboard or a mouse, is presented. Section 2 describes the applied automatic speech recognition system, Section 3 presents the head tracking system, Section 4 gives the description of the method for audio and video data synchronization as well as information fusion, and the results of experiments with ICANDO system are presented in the Section 5.

2. Automatic Speech Recognition

ICANDO system can recognize the voice commands of a user in two languages: Russian and English. For automatic speech recognition the SIRIUS system (SPIIRAS Interface for Recognition and Integral Understanding of Speech), developed in Speech Informatics Group, is applied. SIRIUS had already used successfully for automatic speech recognition in several multi-modal applications [2]. This automatic speech recognition system is mainly intended for recognition of Russian speech and contains several original approaches for processing of Russian speech and language, in particular, the morphemic level of the representation of Russian speech and language [3].

For speech parameterization the MFCC features with first and second derivatives are used. The recognition of phonemes, morphemes and words is based on HMM methods. In applied phonetic alphabet for Russian there are 48 phonemes: 12 for vowels (including stressed and unstressed vowels) and 36 for consonants (including hard and soft consonants). As acoustical models the HMMs of the triphones with the mixture Gaussian probability density functions are used. HMM of the triphones have 3 meaningful states (and 2 additional states intended for concatenation of the triphones in the morphemes models) [4].

It is necessary to emphasize that for the task of voice command recognition, where the size of the vocabulary is less than thousands of words, the vocabulary is composed simply as a list of all the word-forms in the task. But for a more complex task with a medium or large vocabulary the morphemic level of processing should be applied. And in further research it is planned to combine the assistive multi-modal system with dictation system based on SIRIUS engine. At present to enter any text in a computer, a user has to use the special program, embedded in MS Windows, the On-Screen Keyboard which is a virtual keyboard on a desktop like in PDA. Table 1 presents the list of the voice commands used for hands-free work with a computer. The list contains 41 commands for ICANDO system, which are similar to the keyboard shortcuts.

Theoretically, two voice commands ("Left" and "Right") could be enough to work with a PC (or a PDA), but introduction of the additional commands, which are often used by a user, allows increasing essentially the velocity of a human-computer interaction.

Table 1. *The list of the voice commands of ICANDO system.*

Class of a command	Voice Command	Multi-modal nature
Mouse manipulator commands	Left	yes
	Right	yes
	Left down	yes
	Left up	yes
	Right down	yes
	Right up	yes
	Double click	yes
	Scroll down	no
	Scroll up	no
Keyboard buttons commands	Shut down	no
	0-9	no
	Escape	no
	Delete	no
	Start	no
Windows Graphical User Interface commands	New	no
	Open	no
	Save	no
	Close	no
	Exit	no
	Cancel	no
	Copy	no
	Cut	no
	Paste	no
	Print	no
	Find	no
	Undo	no
	Redo	no
	Next	no
	Previous	no
	Select all	no
	Say text	no
Special command	Calibration	no

All the voice commands can be divided into four classes according to their functional purpose: mouse manipulator commands, keyboard buttons commands, Windows Graphical User Interface commands, as well as the special commands class, which contains only the "Calibration" command intended to start the tuning process of the head tracking system. However just the commands for the mouse manipulations have multi-modal nature. They use the information on the coordinates of mouse cursor in a current time moment. All other commands are pure speech commands (uni-modal) and the position of cursor is not taken into account at the multi-modal information fusion.

3. Mouse cursor control by head movements

The first version of the head tracking system, developed in SPIIRAS for the assistive multi-modal system, used the special hardware (reference device unit) [6]. It is the rigid construction with three light-emitting diodes mounted on the head. A video camera was used in infrared mode to get the coordinates of these reference marks. The 3D computer model of the reference device unit was constructed and having the coordinates of each reference mark on the image the system could calculate the position of a user's head.

In the last version of ICANDO system we applied a software method for tracking operator's head motions. It is based on the free available software library Intel Open Source Computer Vision Library OpenCV ("http://www.intel.com/technology/computing/opencv"). This library realizes many known algorithms for image and video processing.

It was determined experimentally that the most suitable point on a face for tracking is the tip of nose. It is the center of the face and when we make any gestures by head (turn to the right, left, up or down) the position of the tip of nose moves to that direction and thus it can indicate the position of the mouse cursor on a desktop [7].

For the video processing USB web-camera Logitech QuickCam for Notebooks Pro with the resolution 640x480 and 30 fps is applied. The usage of a professional digital camera (Sony DCR-PC1000E was tested in some experiments) provides better precision of tracking, but taking into account that the system should be available for most users, we apply camera of low-end class with the price under 50 dollars.

The special approach was developed for controlling the mouse cursor able to work in the real-time mode. It includes two stages of the functioning: calibration and tracking. At first short starting stage the position of face is defined in the video. It is realized by the software module which uses the Haar-based object detector to find rectangular regions in the given image that can likely contain a face of a human [8]. This region should not be smaller than 250 per 250 points that allows accelerating the video processing. Then taking into account the standard proportions of a human's face the approximate position of nose is marked by the blue point on the image. During several seconds of the calibration process a user should combine the tip of his nose with the position of this blue point. Then this point is captured by the system and the tracking algorithm is started. This algorithm uses the iterative Lucas-Kanade technique for optical flow [5], which is an apparent motion of image brightness. Sometimes the algorithm loses the position of human's nose that is caused by the lack of light or very quick motions of user's head. To solve this problem the special voice command "Calibration" was introduced in the system, which runs the process of calibration described above. Thus, the tip of nose defines the position of the mouse cursor on the desktop of MS Windows operating system.

4. Modalities synchronization and fusion

In ICANDO system, two natural input modalities are used: speech and head motions. As both modalities are active ones [9], then their input must be controlled continuously (non-stop) by the system. Figure 1 shows the common architecture of ICANDO system.

The system processes human's speech and head motions in parallel and then combines both informational streams in joint multi-modal command, which is used for operating with GUI of a computer. Each of the modalities transmits its own semantic information: head (nose) position indicates the coordinates of some marker (cursor) at a current moment, and speech signal transmits the information about meaning of the action, which must be performed with an object selected by the cursor (or irrespectively to the cursor position). The synchronization of two information streams is made by the speech recognition module, which gives the special signals for storing of the mouse cursor coordinates calculated by the head tracking module, and for multi-modal fusion.

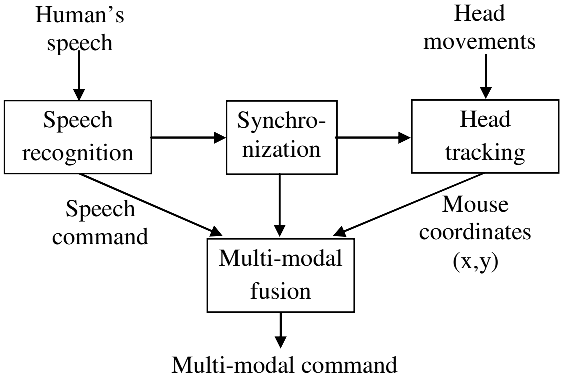

Figure 1. *The common architecture of ICANDO system.*

Figure 2 illustrates the process of modalities synchronization and information fusion in the system. This figure shows the process of fulfillment of one scenario for hands-free work with Internet Explorer for obtaining some information at the web-portal (the sequence of voice commands simultaneously with user's head motions: "Left", "Scroll down" and "Left"), selecting the fragment of page (voice commands "Left down", "Left up" and head motions), copying this information into the memory buffer (command "Copy"), opening the MS Word text editor (commands "Start" and "Left") and paste the information from buffer into the text editor (command "Paste").

The speech signal captured from a microphone is processed continuously by the SIRIUS system. The speech recognition process is started by the speech endpoint detector, which finds the presence of some signal different from silence (or permanent background noise). The speech recognition process is finished after finding the best hypothesis of voice command recognition.

The synchronization of the information streams is activated by the speech recognition module and performed by the following way: concrete mouse cursor position, which is calculated continuously by the head tracking system, is taken at the beginning of a voice command input i.e. at the moment of triggering the algorithm for speech endpoint detection (the markers on "Cursor coordinates storing" line on the Figure 2). It is connected with the problem that during phrase pronouncing a user can move his head and to the end of speech command recognition the cursor can indicate on another graphical object. Moreover a voice command is appeared in the brain of a human in short time before the beginning of phrase input.

For information fusion the frame method is used when the fields of some structure are filled in by required data and on completion of the speech recognition process the corresponding control command is executed. The fields of this structure are: text of speech command, X coordinate of the mouse cursor, Y coordinate of the mouse cursor, kind of speech command (multi-modal or uni-modal). If a speech command has multi-modal nature (see Table 1) then it has to be combined with stored coordinates of the mouse cursor and then the Windows message to a virtual mouse device of operating system is sent automatically. If the voice command is uni-modal then the coordinates are not taken into account and the message to a virtual keyboard device is sent. The head motions only (without speech modality) cannot produce any commands for a computer but they can be used for painting in a graphical editor.

On Figure 2 a black circle means that the recognized command (for instance, the command "Left down") is multi-modal one and a white circle means that the command has uni-modal nature of human-computer interaction (speech-only, for instance, the command "Paste"). The automatic speech recognition module works in real-time mode, since the voice commands vocabulary is small one, therefore there are minor delays between an utterance of a phrase by a user and fulfillment of the recognized multi-modal command and these delays may not be taken into account.

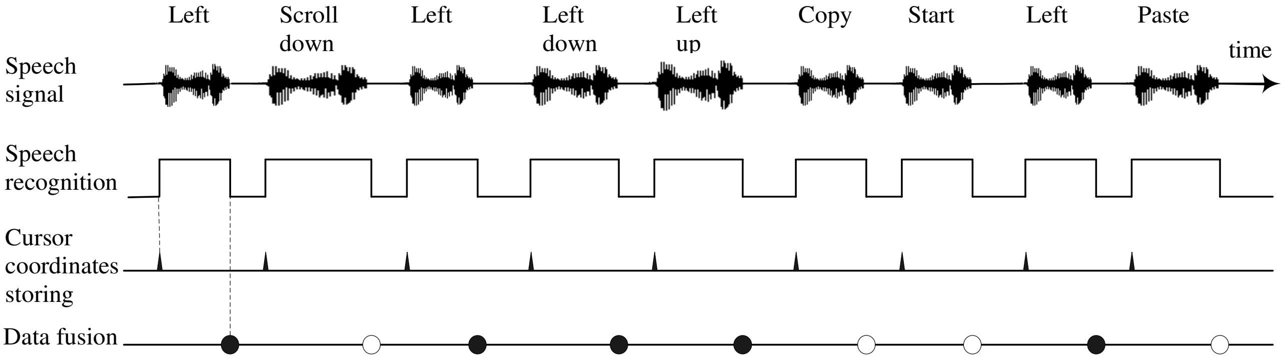

Figure 2. *The mechanism of synchronization and fusion of multi-modal information.*

5. Experimental results

For the head tracking task the miniature web-camera Logitech QuickCam for Notebooks Pro is used. The camera provides a video signal in 640x480x30 fps and an audio signal from camera's microphone with 16 KHz and acceptable SNR level.

The testing of the system was fulfilled by five inexperienced users, which had minor experience of work with a computer as well as by one real handicapped person without hands. The system was tested on the task of GUI control for the operational system MS Windows. The test scenario in the experiments was connected with the work with Internet Explorer for finding the weather forecast at the web-portal "http://www.rbc.ru", selecting, copying and saving this information in a text file in MS Word and printing this file. The task is divided into several elementary actions, which are accomplished by the multi-modal way and by the standard way (mouse + keyboard).

Table 2 presents the results of experiments and comparison of two ways of operation with a computer. The accuracy of speech recognition as well as the time, required to each operator to fulfill the test scenario, and average values for all users are presented. The time for the standard way is not shown for the last user, because he is impaired person without hands and can not use a mouse and a keyboard. The accuracy of speech recognition was over 96.5% for each of the operators.

Table 2. *The comparison of multi-modal and standard ways of computer control.*

User	Command recognition rate, %	Multi-modal way, sec.	Standard way, sec.
1	98.5	84	43
2	97.5	73	36
3	97.5	91	44
4	97.0	88	50
5	96.5	77	42
6	98.0	80	-
Aver.	97.5	82	43

It was determined experimentally that the multi-modal way is in 1.9 times slower than the traditional way. However, this decrease of interaction velocity is acceptable since the developed system is intended mainly for motor-disabled users.

The real work of the multi-modal system for hands-free computer control based on speech recognition and head tracking was shown on the main Russian TV channel ("First channel") in the news program "Vremja" on 5 November 2005. During the demonstration the impaired operator successfully worked with a personal computer by ICANDO system (see "http://www.1tv.ru/owa/win/ort6_main.main?p_news_title_id=82825&p_news_razdel_id=4"). The additional video fragments of testing the ICANDO assistive multi-modal system are available at the web site of Speech Informatics Group (see "http://www.spiiras.nw.ru/speech/demo/assistive.html").

The obtained results allow concluding that the assistive multi-modal system can be successfully used by users with disabilities of their hands for hands-free work with a computer.

6. Conclusions

The presented assistive multi-modal system ICANDO is aimed mainly for challenged users, which have the problems using a computer keyboard and a mouse. The human-computer interaction is performed by voice and head motions. The developed system uses a cheap web-camera, which provides video and audio signals. It simplifies the usage of the system, since no any additional hardware (like a microphone or a helmet) is required. ICANDO system was applied and tested for hands-free operation with GUI of MS Windows in such tasks as Internet communications and work with text documents. The experiments have shown that in spite of some decreasing of operation velocity the multi-modal system allows working with a computer without standard computer control devices.

7. Acknowledgements

This research is supported and funded by the European Community FP6-IST Program in framework of the SIMILAR Network of Excellence, contract № 507609.

8. References

[1] Bates, R., Istance, H. O. "Why are eye mice unpopular? A detailed comparison of head and eye controlled assistive technology pointing devices", Proc. of First Cambridge Workshop on Universal Access and Assistive Technology, USA, 2002.

[2] Ronzhin, A. L., Karpov, A. A., Timofeev, A. V., Litvinov, M. V. "Multimodal human-computer interface for assisting neurosurgical system", Proc. of 11-th International Conference on Human-Computer Interaction HCII'2005, Las Vegas, USA, 2005.

[3] Ronzhin, A. L., Karpov, A. A., Li, I. V. "Russian Speech Recognition for Telecommunications", Proc. of 10-th International Conference on Speech and Computer SPECOM'2005, Patras, Greece, pp. 491-494, 2005.

[4] Karpov, A. A., Ronzhin, A. L. "Speech Interface for Internet Service Yellow Pages", Intelligent Information Processing and Web Mining: Advances in Soft Computing, Springer-Verlag, pp. 219-228, 2005.

[5] Bouguet, J. Y. "Pyramidal implementation of the Lucas-Kanade feature tracker", Technical Report, Intel Corporation, Microprocessor Research Labs, 2000.

[6] Ronzhin, A. L., Karpov A. A. "Assistive multimodal system based on speech recognition and head tracking", Proc. of 13-th European Signal Processing Conference EUSIPCO'2005, Antalya, Turkey, 2005.

[7] Gorodnichy, D., Roth, G. "Nouse 'Use your nose as a mouse' perceptual vision technology for hands-free games and interfaces", Image and Vision Computing, Vol. 22, 2004, pp. 931-942.

[8] Lienhart, R., Maydt, J. "An Extended Set of Haar-like Features for Rapid Object Detection", Proc. of International Conference on Image Processing ICIP'2002, pp. 900-903, 2002.

[9] Oviatt, S. L. "Multimodal interfaces". Chapter in Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications. Lawrence Erlbaum Assoc. Mahwah, NJ, USA, pp. 286-304, 2003.