ICANDO: INTELLECTUAL COMPUTER ASSISTANT FOR DISABLED OPERATORS

Alexey Karpov and Andrey Ronzhin

Speech Informatics Group, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
SPIIRAS, 39, 14th line, 199178, St. Petersburg, Russia
phone: +7 (812) 328-7081, fax: +7 (812) 328-4450, email: {karpov,ronzhin}@iias.spb.su
web: www.spiiras.nw.ru/speech

ABSTRACT

The paper describes a prospective multimodal system ICanDo (Intellectual Computer AssistaNt for Disabled Operators) developed in SPIIRAS and intended for assistance to 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 multimodal system. The architecture of the system, methods for recognition and tracking, multimodal information fusion and synchronization, experimental conditions and obtained results are described in the paper. The developed system was applied for hands-free work with Graphical User Interface in such tasks as Internet communication and work with documents.

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 multimodal system, which allows controlling a computer without traditional control devices, but using: (1) head (or face) movements to control the mouse cursor on a monitor screen; (2) speech input for giving the control commands. Speech and head-based multimodal control systems have a great potential in improving the life comfort of disabled people as well as independence of their living from other persons.

Disability may affect also the person's neck and head movements along with hands and 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 eye blinking can give the signal for click of mouse button. Among examples of hardware-software eye tacking systems Visual Mouse [1] and Eyegaze System [2] can be mentioned. 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 [3]. Of course, speech input is only one acceptable alternative instead of keyboard for motion-impaired users which cannot move their hands.

In the following sections of the paper the assistive multimodal system ICanDo, which uses the head movements 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 Section 5.

2. AUTOMATIC SPEECH RECOGNITION

ICanDo system can use 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 multimodal applications [4]. 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 [5].

For speech parametrization 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 triphones with mixture Gaussian probability density functions are used. HMM of triphones have 3 meaningful states (and 2 additional states intended for concatenation of triphones in the models of morphemes).

It is necessary to emphasize that for the task of voice command recognition, where the size of vocabulary is less than thousands of words, the vocabulary is composed simply as list of all the word-forms in the task. But for more complex task with medium or large vocabulary the morphemic level of processing should be applied. And in future research it is planned to combine the assistive multimodal 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 voice commands used in the system for hands-free work with a computer. The list of commands for ICanDo contains 41 commands, 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 additional commands, which are often used by a user, allows increasing essentially the velocity of a human-computer interaction.

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 Special Commands class, which contains only the "Calibration" command intended for starting of the tuning process of the head tracking system. However just the mouse manipulator commands have multimodal nature. They use information on coordinates of mouse cursor in a current time moment. All other commands are pure speech commands (unimodal) and the position of cursor is not taken into account at multimodal information fusion.

Table 1 - List of voice commands of ICanDo system

Class of command	Voice Command	Multimodal 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

3. HEAD MOVEMENTS TRACKING

This section describes the head tracking technologies intended for tracking the natural operator's head motions instead of hand-controlling motions.

The head tracking can be performed by two diverse ways: hardware and software-based methods. In the hardware techniques a user should wear some special devices on his head. At present there exist several hardware systems for head tracking in computer market (some examples are presented on Figure 1). For instance, the NaturalPoint company has presented the SmartNAV hands-free mouse. This system consists of the special transmitter-receiver device working in infrared mode and several reflective marks, which should be attached to the face of a user or to special hat. The company InterSence produces professional trackers InterTrax for helmets of virtual reality or computer stereo glasses. Inside of this device there exists a gyroscope, which allows tracking the orientation of a head. Also hardware trackers can be applied based on special device with light emitting diodes and a video-camera. However, all these devices are very expensive and their cost is varied from several hundred till thousand euros. It is one of the reasons why they are not popular in assistive systems for impaired users. The second reason is that users find inconveniently to wear special device on the head during work with a computer.

Figure 1 - Hardware-based head tracking systems

There exist also several effective software methods for head tracking: the methods of optical flow [6], the biological approach based on the retina filters [7], etc.

The first version of the head tracking system, which was applied in SPIIRAS for the assistive multimodal system, used the special hardware (reference device unit) [8]. It was the rigid construction with three light-emitting diodes mounted on the head. A video camera was used in infrared mode to obtain the coordinates of these reference marks. The 3D computer model of the reference device unit was constructed before and having the coordinates of each reference mark on the image the system could calculate the position of the user's head. We rejected this version after the test because real users said about some discomfort wearing any hardware on the head using the system long time.

Therefore new software method for tracking operator's head motions was developed. It is based on the free available software library Intel OpenCV (Open Source Computer Vision Library: http://www.intel.com/technology/computing/opencv). This library realizes many known algorithms for image and video processing.

For video processing USB web-camera Logitech QuickCam for Notebooks Pro with resolution 640x480 and 30 fps is applied. The usage of a professional digital camera (Sony DCR-PC1000E was tested in some experiments) provides better accuracy 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 euros.

Several points on face are tracked by the system to control the mouse cursor on the desktop. These points are: center upper lip, the tip of nose, point between eyebrows, left eye and right eye. It was determined experimentally that these points on face are most stable for tracking points. The average delta over coordinates of 5 points is calculated for two frames of the video and this value is used to set the new position of the mouse cursor like in [9].

The special approach was developed for control of the mouse cursor, which is able to work in real-time mode. It includes two modes of functioning: calibration and tracking. At first short face detection mode the position of face in the video is defined. It is realized by the software module which uses the Haar-based object detector to find rectangular regions in the given image that likely can contain face of a human [10]. This region should not be less than 250 per 250 points that allows accelerating video processing. Then taking into account the standard proportions of a human's face the approximate position of nose is marked by blue point on the image. During several seconds of 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. Four other tracking points are added automatically taking into account proportions of the face for a user. Then the system is started in the tracking mode. Tracking algorithm uses well-known iterative Lucas and Kanade technique for optical flow [6], which is an apparent motion of image brightness. Unfortunately, sometimes the algorithm loses the position of human's nose that is caused by the lack of light or very quick movements of user's head. To solve this problem the special voice command "Calibration" was introduced in the system, which runs the process of calibration (face detection mode) described above.

4. MULTIMODAL SYNCHRONIZATION AND FUSION

In ICanDo system two natural input modalities are used: speech and head movements. As both modalities are active ones [11], then their input must be controlled continuously (non-stop) by the system. Figure 2 shows the common architecture of ICanDo system.

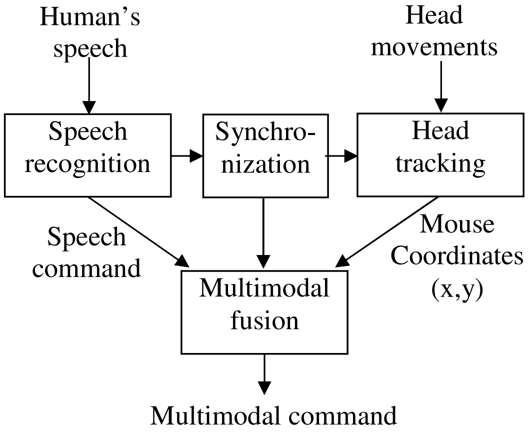

Figure 2 - The common architecture of ICanDo system

The system processes human's speech and head movements in parallel and then combines both informational streams in joint multimodal command, which is used for work with GUI of a computer. Each of the modalities transmits own semantic information: head (nose) position indicates the coordinates of some marker (cursor) in a current time 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 multimodal fusion. Figure 3 illustrates the process of modalities synchronization and data fusion in the system.

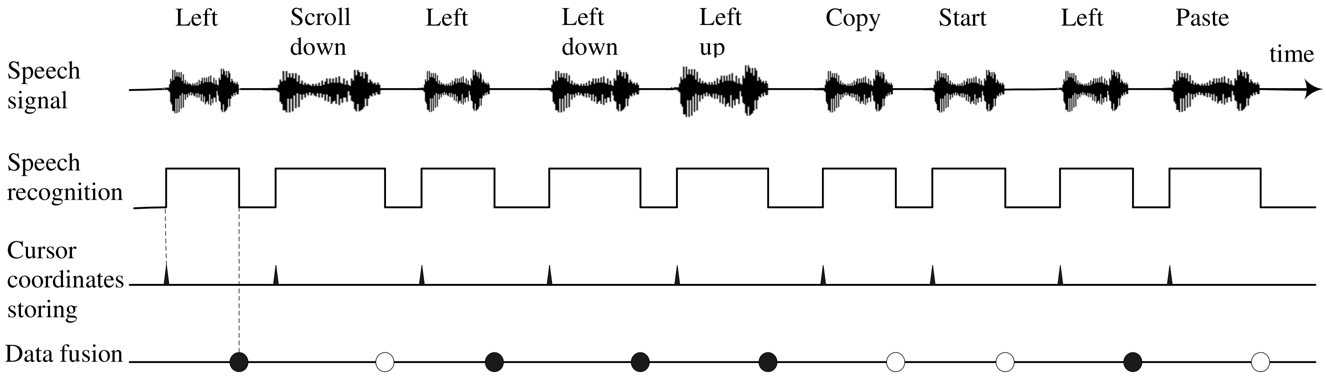

Figure 3 - The example of information synchronization from speech recognition and head tracking modules in the assistive system

This figure shows the process of fulfilment of one scenario for hands-free work with Internet Explorer for obtaining some information at web-portal (sequence of voice commands simultaneously with user's head movements: "Left", "Scroll down" and "Left"), copying this information into the memory buffer (command "Copy"), opening the Notepad 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 speech endpoint detector, which finds the presence of some speech signal different from silence (or permanent background noise). The speech recognition process is finished after finding of best recognition hypothesis of a voice command.

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 Figure 3). 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 command, which must be fulfilled, is appeared in the brain of a human in short time before 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 speech recognition process the corresponding control command is executed. The fields of this structure (frame) are: text of speech command; X coordinate of mouse cursor; Y coordinate of mouse cursor; kind of speech command (multimodal or unimodal). If an input speech command has multimodal 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 unimodal then the coordinates are not taken into account and the message to a virtual keyboard device is sent. The head movements only (without speech modality) can not produce any commands for a computer but they can be used for painting of some pictures in graphical editors.

On Figure 3 a black circle means that the recognized command (for instance, the command "Left down") is multimodal one and white circle means that the command has unimodal nature of human-computer interaction (speech-only, for instance, the command "Copy" or "Paste"). The automatic speech recognition module works in real-time mode, since the voice commands vocabulary is small one, therefore there are minor delays (tens or hundreds of milliseconds depending on the pronunciation of a command) between an utterance of a phrase by a user and fulfillment of the recognized multimodal command and these delays may not be taken into account.

5. EXPERIMENTS

As the hardware for hands-free computer control the miniature web-camera Logitech QuickCam for Notebooks Pro is used. This camera provides a video signal in 640x480x30fps and audio signal, obtained from the microphone built in the camera 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. Figure 4 shows the fragment of hands-free work with a computer by ICanDo system.

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 and Notepad for finding the weather forecast in St. Petersburg at the web-portal www.rbc.ru, selecting, copying and saving this information in a text file and printing this file. The task is divided into several elementary actions, which are accomplished by the multimodal way (head movements + speech input) and by the standard way (mouse + keyboard).

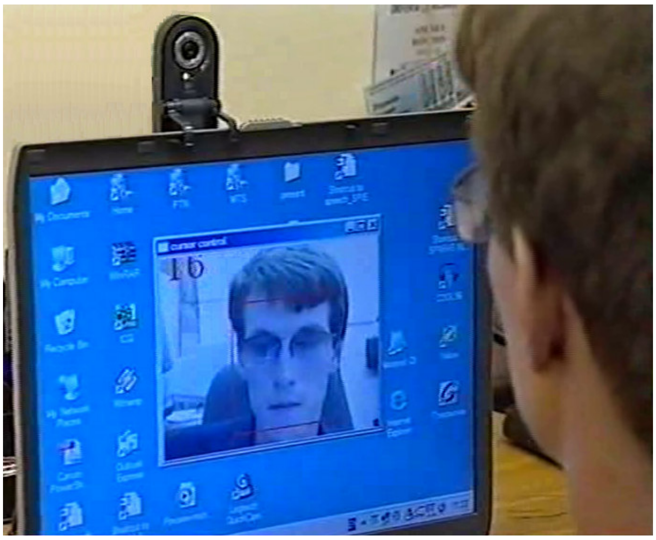

Figure 4 - Impaired person works with a computer by ICanDo assistive multimodal system

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 fulfil the test scenario, and average values for all users are presented in the table. The time for standard way is not exist for the user 6, because he is impaired person without hands and can not use a mouse or a keyboard.

It was determined experimentally that the multimodal way is in 1.9 times slower than the traditional way. However, this decrease of interaction speed is acceptable since the developed system is intended mainly for motor-disabled users. It can be seen from the table 2 that during the experiments accuracy of speech recognition was over 96.5% for each human-operators tested the system.

Table 2 - The comparison of multimodal and standard ways of a computer control

User	Command recognition rate, %	Multimodal 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

Real work of the multimodal system for hands-free computer control based on speech recognition and head tracking was shown in the main Russian TV channel ("First channel") in the news program (“Vremja”) on 6 September 2005. During the demonstration the impaired person 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 of ICanDo assistive multimodal system are available at the web site of Speech Informatics Group of SPIIRAS at the hyperlink: http://www.spiiras.nw.ru/speech/demo/assistive.html.

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

6. CONCLUSIONS

The presented assistive multimodal system ICanDo is aimed mainly for handicapped operators, which have the problems using a computer keyboard and a mouse. The human-computer interaction is performed by voice and head movements. The last version of the system uses a cheap web-camera, which provides video and audio signal with an acceptable quality. 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 Graphical User Interface of operational system 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 speed the multimodal system allows working with a computer without standard mouse and keyboard. Thus the system can be successfully used for hands-free PC control for impaired users. In future research it is planned to combine the developed assistive multimodal system with the dictation system based on SIRIUS speech recognition engine for Russian language.

7. ACKNOWLEDGEMENTS

This research is supported and funded by the EC-FP6-IST Program in framework of SIMILAR Network of Excellence contract № 507609 as well as by the Government of Saint-Petersburg.

REFERENCES

[1] F. Tinto Garcia-Moreno, “Eye Gaze Tracking System Visual Mouse Application Development”. *Technical Report*, Ecole Nationale Supériere de Physique de Strasbourg (ENSPS) and School of Computer Science, Queen’s University Belfast, 77 p., 2001.

[2] LC TECHNOLOGIES, INC. Eyegaze Systems, http://www.eyegaze.com

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

[4] A.L. Ronzhin, A.A. Karpov, A.V.Timofeev, M.V. Litvinov, “Multimodal human-computer interface for assisting neurosurgical system”. In Proc. *11-th International Conference on Human-Computer Interaction HCII’2005*, Las Vegas, USA, 2005.

[5] A. Ronzhin, A. Karpov, I. Li, “Russian Speech Recognition for Telecommunications”. In Proc. *10-th International Conference on Speech and Computer SPECOM’2005*, Patras, Greece, pp. 491-494, 2005.

[6] J.-Y. Bouguet, “Pyramidal implementation of the lucas kanade feature tracker”. *Technical Report*, Intel Corporation, Microprocessor Research Labs, 2000.

[7] A. Benoit, A. Caplier, “Biological approach for head motion detection and analysis”. In Proc. *13-th European Signal Processing Conference EUSIPCO-2005*, Antalya, Turkey, 2005.

[8] A.L. Ronzhin, A.A. Karpov, “Assistive multimodal system based on speech recognition and head tracking”. In Proc. *13-th European Signal Processing Conference EUSIPCO-2005*, Antalya, Turkey, 2005.

[9] D. Gorodnichy, G. Roth, “Nouse 'Use your nose as a mouse' perceptual vision technology for hands-free games and interfaces”. *Image and Vision Computing*, Vol. 22, Issue 12, pp. 931-942, 2004.

[10] R. Lienhart, J. Maydt, “An Extended Set of Haar-like Features for Rapid Object Detection”. In Proc. *IEEE International Conference on Image Processing ICIP’2002*, pp. 900-903, 2002.

[11] S.L. Oviatt, “Multimodal interfaces”. Chapter in *Human-Computer Interaction Handbook*: *Fundamentals, Evolving Technologies and Emerging Applications*. Lawrence Erlbaum Assoc. Mahwah, NJ, chap.14, pp. 286-304, 2003.