# ASSISTIVE MULTIMODAL SYSTEM BASED ON SPEECH RECOGNITION AND HEAD TRACKING

*Andrey Ronzhin, Alexey Karpov*

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

## ABSTRACT

In this paper the assistive multimodal system is presented, which is aimed for the disabled people, which need other kinds of interfaces than ordinary people. The group of users of this system is persons with hands disabilities. The interaction between a user and a machine is performed by voice and head movements. It gives the opportunity for disabled people to carry out a work with PC. The work of the multimodal systems is presented during EUSIPCO-2005 Conference in framework of Similar Demonstration Session "Multimedia Tools for Disabled".

## 1. INTRODUCTION

Many people are unable to operate a standard computer mouse or keyboard because of disabilities of their hands or arms. One possible alternative for these persons is multimodal system, which allows controlling a computer without using standard mouse and keyboard, for example: (1) using head movements to control the cursor across the computer screen; (2) using the speech for giving the control commands.

Automatic speech recognition and head tracking in joint multimodal system are combined in the system. The multimodal system processes the input information and controls the PC devices. For output of information the standard means of PC are used: PC monitor for graphical output and sound card + speakers for audio output.

In the multimodal system we combine two modalities only: speech and head movements. It is concerned with specific application area for hand-disabled people, so such modalities as gestures, haptics, handwriting can not be used. On the other side using emotion recognition, facial moves, eye detection, etc. the system can be enhanced in future.

Speech and head-based control systems have a great potential for improving the life comfort of disabled people, their social protectability and independence from other people. Hands-free control devices such as hands-free mouse and keyboard for access to PC are effective applications of these technologies. Users who have difficulties using the standard PC control devices could manipulate cursor merely by moving their heads and giving the speech command instead of clicking the buttons.

Unfortunately, a person's disability may affect his neck and head movements along with hands and arms. For instance, a person may have reduced active neck range of motion and hence reduced ability to move the head in one or more directions. In many of such cases the gaze tracking system can be successfully used instead of head tracking system. Though, usage of the gaze tracking system is worse in such parameters as task performance, human's workload and comfort both for untrained user and for experienced user, than the head tracking system. Of course, speech input is only one acceptable alternative to the keyboard for motor-disabled users.

Another good application of hands-free cursor control allows users to change the "focus of windowÓ in GUI without mouse movement. It is helpful because ordinary human at typing uses both hands and during this typing he cannot move the mouse. The usage of hands-free mouse cursor control is effective way to increase the speed of information input. Let's imagine that there are two GUI opened side-by-side on the desktop. Instead of having to laboriously switch the active window by moving and clicking on the mouse, the user could simply turn head towards desired window and say the speech command after that keyboard input will flow into the appropriate document. Also, there are applications of hands-free cursor control for entertainment such as: painting programs, games, designing systems, etc.

## 2. TECHNICAL ISSUES

The system is intended for usage in operating system MS Windows 9x and above. Hardware of the Head Tracking System (HTS) includes the following units:

- Reference Device Unit - RDU;
- Camera Unit - CU;
- Video Processor Unit - VPU;
- Personal Computer, Pentium 3(4) – PC;
- Camera Control Unit – CCU.

When HTS is used in active mode, CU is equipped with black & white cameras with IR lenses. In passive mode CU is equipped by color cameras.

RDU for active HTS is a rigid construction with LED's (Infra Red Light Emitting Diodes) or color reference marks for passive HTS, mounted on the head. Both for active and passive varieties of the HTS prototype we use miniature commercial video cameras: black & white and color one (PAL) with resolution no worse than 400 TV lines. The video processor's (VPU) is designed as a set of standard card for PC Pentium 4.

For automatic speech recognition the SIRIUS system is used [1]. This system is aimed for recognition of Russian speech and was developed in Speech Informatics Group of SPIIRAS. The main feature of this system is usage of morphemic level of speech and language representation. Owing to division of word-form into morphemes the vocabulary size of recognized lexical units is significantly decreased as well as accuracy and speed of processing are increased.

For speech recognition the microphone and PCI sound card are required. In our experiments we used microphone Sony DR-50 with built-in signal amplifier, connected to Sound Blaster Creative Labs Audigy 2.

Common architecture of developed assistive bimodal system [2] is presented in Figure 1. In developed multimodal system two modalities are used: speech and head movements. As both modalities are active, then their input into the system must be controlled continuously (non-stop) by the computer. Each of the modalities transmits own semantic information: head position indicates the coordinates of some marker (cursor) in current time moment, and speech transmits the information about meaning of the action, which must be performed with an object selected by cursor (or irrespectively to the cursor).

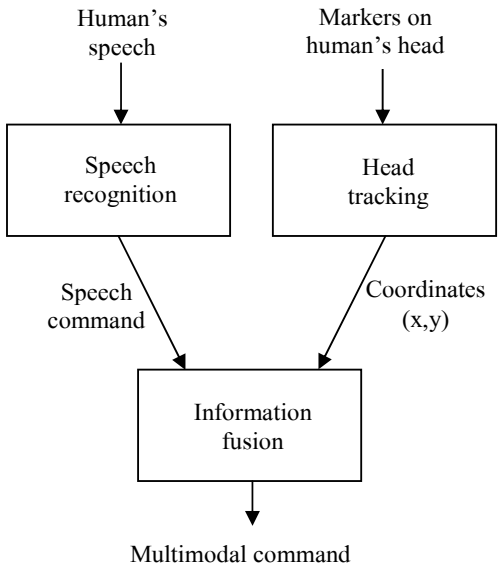


Figure 1. General structure of multimodal system

In the developed system the synchronization of modalities is performed by following way: concrete marker position is calculated at beginning of the phrase input (i.e. at the moment of triggering the algorithm for speech endpoint detection). It is connected with the problem that during pronouncing the phrase the cursor can be moved. For information fusion the frame method is used when the fields of some structure are filled by required data and on completion the signal for command execution is given.

## 3. FUNCTIONALITY

In Table 1 the list of the basic speech commands, which a human can enter into the system, is presented.

Table 1: The list of speech commands

| Speech command | Action |
|---|---|
| Left | Click mouse left button |
| Right | Click mouse right button |
| Open | Open file or program |
| Close | Close window or file |
| Exit | Exit from program |
| Save | Save current file |
| Scroll down | Scroll text down |
| Scroll up | Scroll text up |
| Cancel | Cancel the action |
| Start | Click "StartÓbutton |
| Shut down | Shut down computer |
| Copy | Copy selected object |
| Cut | Cut selected object |
| Paste | Paste buffered object |
| 0-9 | Write digits 0-9 |
| Print | Print current file |
| Find | Open find window |
| Left down | Mouse left button down |
| Left up | Mouse left button up |
| Double click | Left button double click |
| Say text | Say selected text (TTS) |
| Undo | Undo last action |
| Redo | Redo last action |
| Delete | Delete selected object |
| Next | Open next page |
| Previous | Open previous page |
| Select all | Select all text in document |
| New | Open blank document |
| Enter | Press "EnterÓbutton |
| Escape | Press "EscapeÓbutton |

The control of cursor is provided by head movement. The cursor coordinates are tied with the position of head and small change of head position produces the cursor movement. At that the recognized speech command are fulfilled concerning the cursor position. It allows operating GUI of the operational system Windows and peripheral devices.

## 4. INTERFACE AND USABILITY

Figure 2 shows the dialogue windows of speech recognition system (top window) and head tracking system (bottom window). In work mode these windows are minimized. In the top window the result of speech recognition is showed in Russian and English.

The bottom window shows the coordinates of markers on Reference Device Unit obtained through the camera-recorder. This coordinates are transformed into the mouse cursor position.

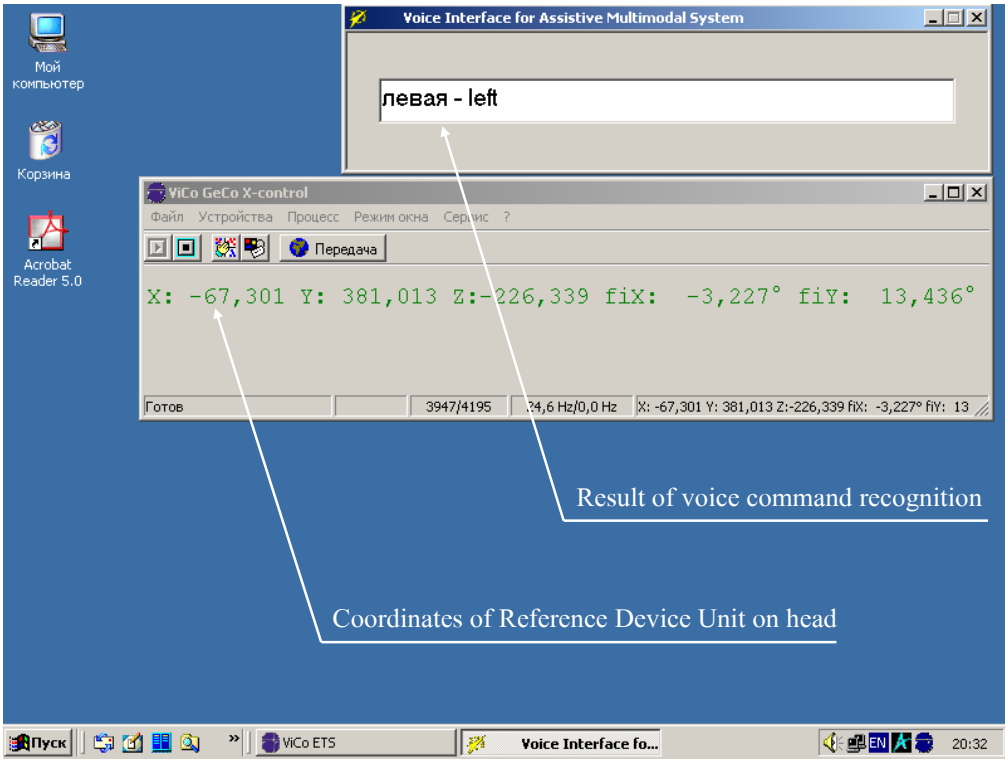


Figure 2. Screenshot of the dialogue windows

The speech command is fulfilled at once after the recognition of speech but the concrete marker position is calculated at beginning of the phrase input (i.e. at the moment of triggering the algorithm for speech endpoint detection). It is connected with the problem that during phrase pronouncing the cursor can be moved and to the end of speech command recognition the cursor can indicate of other graphical object, moreover the command which must be fulfilled is appeared in the brain of a human in short time before beginning of phrase input.

To provide required accuracy of speech recognition the system should be used in quiet environments and no other conversations should present in the room. For usage of head tracking system the camera-recorder and reference device unit should be adjusted for best usage comfort and optimal performance.

## 5. EVALUATION

The testing was fulfilled by 5 beginning users, which had no essential experience of work with personal computer. To estimate the performance time of the cursor movement by mouse and by head we conducted the following experiment. Two shortcuts were located in the desktop with 15 centimeters distance between them. During the experiment we calculated how many times the user can move the cursor from one to another shortcut during one minute. As a result it was found that users operated by mouse in 2.1 times faster than by head.

Then we added the “clickÓaction in the experiment. The task included clicking the shortcuts one after another by mouse and by head movements + voice command. A time for mouse click is insignificant and the time practically was not changed in comparison with mouse movement without click. At that at operating by head and voice the time was increased in 1.4 times in average. Thus the operation by mouse is fulfilled in 2.9 times faster than by developed multimodal system.

Above experiment showed the comparison of performance of cursor operating without attaching to the concrete applied task. Now we consider the testing of the developed system at the task of control GUI of the operational system Windows. The task included work with text editor MS Word and Internet access by means of MS Internet Explorer. The set of spoken commands is adjusted to perform over 100 actions with GUI.

The examples of usage of developed assistive multimodal system are available in the Internet at Web site of Speech Informatics Group of SPIIRAS http://www.spiiras.nw.ru/speech/demo/assistive.html.

Three video fragments show different tasks fulfilled by the system:

1. Opening the MS Internet Explorer, opening the web portal www.rambler.ru and finding the table of currency exchange within this portal.
2. Opening the Calculator and converting 1350 Swiss francs into Russian roubles.
3. Obtaining information about TV program (MTV) for today evening at web-site www.rambler.ru, copying

this information into new .doc file, saving at desktop, printing this file and closing the opened programs.

Table 2 describes the fragment of operating with Internet Explorer and Word for obtaining information about TV program (third scenario). This task is divided on some elementary actions, which can be accomplished by multimodal interface (head movement + speech input) or standard way (mouse + keyboard). The total time spent for this scenario is presented in the end of the table.

Table 2: Fragment of operation with GUI

| N | Description of actions | System | |
|---|---|---|---|
| | | Head + speech | Mouse + keyboard |
| 1 | Select hyperlink “TV ProgramÓ | (Head) | Mouse |
| 2 | Open hyperlink “TV ProgramÓ | Left | Left click |
| 3 | Scroll down screen | Scroll down | Wheel down |
| 4 | Scroll down screen | Scroll down | Wheel down |
| 5 | Select hyperlink “MTVÓ | (Head) | Mouse |
| 6 | Open hyperlink “MTVÓ | Left | Left click |
| 7 | Set cursor on beginning | (Head) | Mouse |
| 8 | Left button down | Left down | Left button down |
| 9 | Set cursor on ending | (Head) | Mouse |
| 10 | Left button up | Left up | Left button up |
| 11 | Copy selected text | Copy | Ctrl+C |
| 12 | “StartÓmenu opening | Start | Mouse, left click |
| 13 | MS Word icon selection | (Head) | Mouse |
| 14 | MS Word opening | Left | Left click |
| 15 | Paste the text | Paste | Ctrl+V |
| 16 | Save the file | Save | Ctrl+S |
| 17 | Set cursor on “FolderÓ | (Head) | Mouse |
| 18 | Open tree of folders | Left | Left click |
| 19 | Selection of “Desktop Ófolder | (Head) | Mouse |
| 20 | Set current folder | Left | Left click |
| 21 | Set cursor position on “SaveÓbutton | (Head) | Mouse |
| 22 | Click “SaveÓbutton | Left | Left click |
| 23 | Print the file | Print | Ctrl+P |
| 24 | Set cursor position on “PrintÓbutton | (Head) | Mouse |
| 25 | Click “PrintÓbutton | Left | Left click |
| 26 | Close MS Word | Close | Alt+F4 |
| 27 | Close MS IE | Close | Alt+F4 |
| Total time | | 80 sec. | 28 sec. |

Thus the developed multimodal way of Internet and MS Word access works in 2.85 times slower than traditional way. However this fall is acceptable since developed system is intended mainly for disabled people. During the experiments the accuracy of speech recognition was over 97% for each of 5 users.

## 6. CONCLUSION

The presented assistive multimodal system for disabled people was developed in SPIIRAS. The interaction between a user and a machine is performed by voice and head movements. In order to process these data streams the modules of speech recognition and head tracking were developed. The fusion of information, synchronization and performing the command are realized in the main module. The developed system was applied for hands-free operations with Graphical User Interface in such common tasks as Internet communications and text editing in MS Word. The experiments have shown that in spite of some decreasing of operation speed the multimodal system allows working with computer without using standard mouse and keyboard. Thus the developed assistive multimodal system can be successfully used for hands-free PC control for users with disabilities of their hands or arms.

## 7. ACKNOWLEDGEMENTS

This research is supported and funded by Sixth Research Framework Programme FP6 of the European Union in framework of the SIMILAR European Network of Excellence FP6-507609.

## REFERENCES

[1] A.A. Karpov, A.L. Ronzhin. Speech Interface for Internet Service Yellow Pages. Intelligent Information Processing and Web Mining: Advances in Soft Computing, Proc. of the International IIS:IIPWM’05 Conference, Poland, Springer Verlag, 2005, pp. 219-228.

[2] A. Karpov, A. Ronzhin, A. Nechaev, S. Chernakova. Multimodal system for hands-free PC control. 13th European Signal Processing Conference EUSIPCO-2005, Turkey, September 2005.