MULTIMODAL SYSTEM FOR HANDS-FREE PC CONTROL

Alexey Karpov, Andrey Ronzhin, Alexander Nechaev, and Svetlana Chernakova

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@iias.spb.su
web: www.spiiras.nw.ru/speech

ABSTRACT

This paper describes the developed multimodal system intended for assistance to people with disabilities of hands. It combines automatic speech recognition and head tracking in one multimodal system. The structure of the system, the methods for recognition and tracking, information fusion and synchronization, the obtained results and testing conditions are described in the paper. This system was applied for hands-free control of Graphical User Interface for such tasks as Internet communication and text editing in MS Word.

1. INTRODUCTION

Many people are unable to operate PC by means of standard computer mouse or keyboard because of disabilities of their hands or arms. One possible alternative for these persons is a multimodal system, which allows controlling PC without mouse and keyboard but using: (1) head movements to control the mouse cursor position on screen [1]; (2) speech for giving the control commands. Speech and head-based control systems have a great potential in improving the life comfort of disabled people, their social protectability and independence of living from other persons.

Unfortunately, disability may affect person's neck and head movements along with hands and arms. For instance, 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 the eye tracking system can be successfully used instead of head tracking system. Though, usage of the eye 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 [2]. Of course, speech input is only one acceptable alternative to keyboard for motor-disabled users.

2. "SIRIUS" SPEECH RECOGNITION SYSTEM

For ASR in multimodal system the SIRIUS speech recognition system is used. The architecture of SIRIUS (SPIIRAS Interface for Recognition and Integral Understanding of Speech) [3], developed in Speech Informatics Group of SPIIRAS, is presented in Figure 1.

In contrast to English the Russian language has much more variety at word-formation and thus the size of recognized vocabulary sharply increases as well as quality and speed of processing decrease. Moreover the usage of syntactic constraints leads to that the errors of declensional endings cause the errors of recognition of whole pronounced phrase.

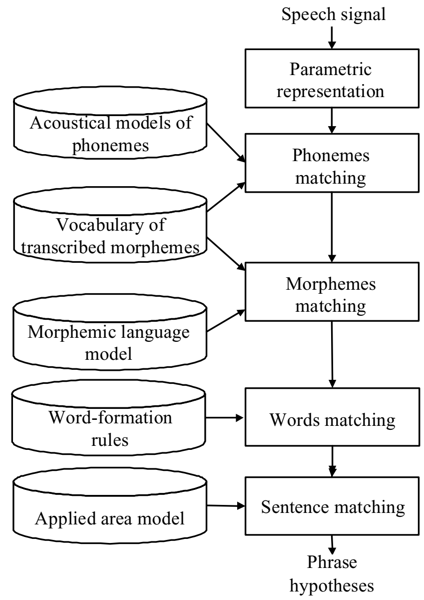

Figure 1. Architecture of SIRIUS system

To avoid these problems the additional level of speech representation (morphemic level) was introduced into speech recognition model [4]. Owing to division of word-forms into morphemes the vocabulary size of recognized lexical units is significantly decreased, since during the process of word formation the same non-root morphemes are used. The databases of various types of morphemes and automatic methods for text processing were elaborated on the base of the rules of Russian word-formation. It has allowed reducing the size of vocabulary of base lexical units in several orders. The coordination of morphemes is calculated using statistics in text corpuses of the concrete applied area. As a result of such processing the speed of recognition and robustness to syntactical deviations in the pronounced phrase are significantly improved.

The speech signal captured from the microphone firstly passes the stage of parametric representation, where starting and ending pauses in the signal are eliminated but the residual part is encoded into the sequence of feature vectors.

The parameterized speech signal follows to the module of phoneme recognition. The recognition of phonemes and formation of morphemes are based on methods of Hidden Markov Models (HMM). As acoustical models we used HMM with mixture Gaussian probability density functions. For feature extraction the mel-cepstral coefficients with first and second derivatives are used. The phonemes are trans-

formed into the triphones (the phoneme in some phonetic context) and Viterbi algorithm is used for the training HMM of triphones by speech databases. HMM of triphone have 3 meaningful states (and 2 "hollow" states intended for concatenation of triphones in the models of morphemes).
For phonetic alphabet we modified the International Phonetic Alphabet SAMPA. In our phonetic alphabet there are 48 phonemes: 12 for vowels (taking into account stressed vowels) and 36 for consonants (taking into account hard consonants and soft consonants).
In contrast to existent analogues our system uses the morpheme level instead of the word one and the language model is created based on statistics of sequences of morphemes in word-forms of Russian language. It is necessary to emphasize that for the task of voice command recognition, where the size of vocabulary does not reach thousands of words; the vocabulary can be composed as list of all word-forms. But for more complex task the additional level of processing (morphemic level) can be successfully applied.
Due to this change the recognition of lexical units is significantly accelerated. After the phoneme recognition and morphemes matching the most probable sequences of morphemes are used for formation of words. The process of word formation from different types of morphemes is realized by means of the oriented graph presented in Figure 2.

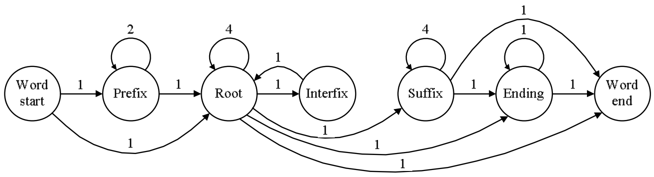

Figure 2. Formation of word from morphemes

In this model there are starting and ending nodes as well as nodes, which present different types of morphemes. The arcs denote possible transitions with maximal number of transitions from one node to another.
The result of speech recognition is best hypothesis of speech utterance, optimal according to acoustical-lexical and semantic-syntactic estimates. Further the recognized speech command follows to the module of information fusion. At the same time to this module the cursor coordinates, which are calculated in the head tracking system developed in Robotics Laboratory of SPIIRAS, are entered. In the following section this system is considered in detail.

3. HEAD TRACKING SYSTEM

This section describes Head Tracking System (HTS) intended for tracking natural man-operator's head motions instead of hand-controlling motions.

3.1 HTS hardware design

HTS prototype includes the following units (Figure 3): (1) Reference Device Unit - RDU; (2) Camera Unit - CU; (3) Video Processor Unit - VPU; (4) Personal Computer - PC; (5) Camera Control Unit – CCU.

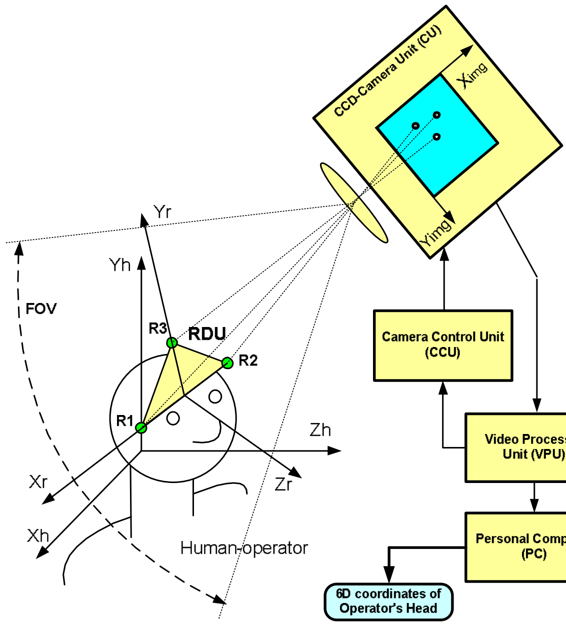

Figure 3. Functional diagram of HTS

The work of this system is described by the following way:
1) Human operator performs natural head movements in the Head Motion Box (HMB) area. In the same time, a helmet-mounted reference device unit (RDU) moves in the HMB for 6 coordinates: three linear translations (x_h, y_h, z_h) and three rotation turns (φ_{xh}, φ_{yh}, φ_{zh}) in the head coordinate system (Xh, Yh, Zh).
2) RDU module has 3 reference marks R1-R3 (IR LEDs for active mode of HTS and color reference marks for passive one) rigidly mounted on the RDU base and the coordinates of each reference mark (xr, yr, zr) are exactly known in the RDU system of coordinates (Xr, Yr, Zr).
3) CCD-Camera Unit (CU), rigidly mounted on the control console base or PC monitor, is aligned so that the reference marks of RDU always remain in the camera FOV while head of operator moves.
4) Reference mark images projected on camera's Focal Plane Array (FPA) will have coordinates in the image (camera) coordinate system (Ximg, Yimg).
5) The CU is intended for control, power supply and synchronization of the camera control unit (CCU). For the active HTS the most important CCU function is synchronization of camera exposition with pulsed emission of IR LEDs. That makes possible a considerable shortening of exposition time (to 5 µs and less) resulting in rejection factor about 1000 against background interference.
6) Camera video signals come for digital processing to the Video Processor Unit (VPU), implemented as standard PCI plat. Basic VPU functions are the following: video signal digitization, filtering and selection of reference images on the background, calculating of the center of coordinates with sub-pixel accuracy.
7) Reference marks' coordinates (Ximg, Yimg) from VPU are entered into the PC memory. For known internal and external parameters of the camera optical system and coor-

dinates of reference mark images 3D coordinates of real position of reference marks in the camera coordinate system (Ximg, Yimg) are computed.
8) Using the HTS prototype software the reference mark images are processed for selection and identification that allows computing RDU position and orientation in the coordinate system (Xh, Yh, Zh).

3.2 The frame-structural model in HTS

In the HTS image processing and computation of head coordinates (position & orientation) are made based on a priori 3D wire-frame head model.
As the model of head, face, and Reference Device Unit (RDU) on head we propose 3D graph-like structure which vertices are tables of parameters (or frames) describing properties of each artificial or actual reference mark (specific feature) [5].
This Frame-Structural Model (FSM) stores simultaneously two kinds of information:
1) Data on characteristic properties of mark images needed for automatic selection and identification of images;
2) Parameters, which define the configuration of marks' mutual positions in a real object (head) specific features.
Therefore, the basic properties of FSM are analogous to both types of known descriptions: visual graphs and frame descriptions. Some analogies are models of crystalline structures and models of molecules wherein the configuration of links and type of atoms in the nodes define properties of substance.
For example, FSM model configuration is described by a set of relative spacings. Spacing between i and j reference marks in the model (RM_{ij}) is normalized relatively to the basic spacing (RM_b) between the marks:

$$RM^{n}_{ij} = \frac{RM_{ij}}{RM_b}$$

where: RM_b – basic spacing length equal, e.g. to the maximal spacing (RM_{ij}) or spacing between specific marks in object. Besides, the configuration is described by a set of spatial angles formed by wire ribs connecting the nearest (neighbor) marks, between radii from the k^{th} mark to i, j marks (αM^{k}_{ij}).

3.3 Head tracking mechanism

The significant features of HTS prototype algorithm are the following:
1) A 3D frame-structured model (FSM) of reference device for active and passive HTS modes (for the markless HTS – model of operator's face / head) is based on 2D models of images in the camera system of coordinates. The usage of 3D model increases reliability of identification of reference marks (characteristic features of face) on the real background.
2) A prediction algorithm for obtaining the most probable points of reference marks on the camera image plane is based on determined speed vectors of their movement.
3) Using color gradient selection of passive reference marks for their localization and identification on the background as well as for obtaining coordinates of reference mark image centroids with subpixel accuracy.

4. MECHANISM OF MULTIMODAL FUSION

In the developed system two modalities are used: speech and head movements. As both modalities are active [6], 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 the current time moment, and speech signal transmits the information about meaning of the action, which must be performed with an object selected by cursor (or irrespectively to the cursor position).
The synchronization of modalities is performed by following way: concrete marker position is calculated at beginning of the phrase input (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 completion of speech command recognition the cursor can indicate on another 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.
For information fusion the frame method is used when the fields of some structure are filled in by required data and on completion the signal for command execution is given.

5. EXPERIMENTAL RESULTS

As hardware the following equipment was used: microphone Sony DR-50 with built-in signal amplifier, connected to Sound Blaster Creative Labs Audigy 2 and HTS's hardware (USB-camera with light-weight RDU). The testing was fulfilled by 5 inexperienced users, which had not 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 a user can move the cursor from one shortcut to another 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 "click" action in the experiment. The task included clicking the shortcuts one after another by mouse and by head movements + voice command. Time for mouse click is insignificant and total 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.
Then we tested the system for 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 in the task contained 110 commands.

Table 1 describes the fragment of operating with Internet Explorer and Word for obtaining information about TV program (MTV) for today evening at web-site www.rambler.ru, copying this information into new .doc file, saving and printing this file. This task is divided into 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 1: Fragment of operation with GUI

N	Description of actions	Performance	
		MMI	Standard
1	select link *TV Program*	(Head)	Mouse
2	open link *TV Program*	Left	Left click
3	scroll down screen	Scroll down	Wheel down
4	scroll down screen	Scroll down	Wheel down
5	select hyperlink *MTV*	(Head)	Mouse
6	open hyperlink *MTV*	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	open *Start* menu	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 *Folder* item	(Head)	Mouse
18	open tree of folders	Left	Left click
19	select *Desktop* folder	(Head)	Mouse
20	set current folder	Left	Left click
21	set cursor on *Save* button	(Head)	Mouse
22	click *Save* button	Left	Left click
23	print the file	Print	Ctrl+P
24	set cursor on *Print* button	(Head)	Mouse
25	click *Print* button	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 was in 2.85 times slower than traditional way. However this fall is acceptable since the developed system is intended mainly for disabled people. During the experiments the accuracy of speech recognition was over 97% for each of 5 users.

The obtained results allow concluding that the developed assistive multimodal system can be successfully used for hands-free PC control for users with disabilities of their hands or arms.

6. CONCLUSION

The multimodal system is aimed for the disabled people, which need other kinds of interfaces than ordinary people [7]. In the developed system the interaction between a user and a computer is performed by voice and head movements. To process these data streams the modules of speech recognition and head tracking were developed. This system was applied for hands-free operations with Graphical User Interface in such 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 funding under the Sixth Research Framework Programme of the European Union in framework of the SIMILAR European Network of Excellence FP6-507609.

REFERENCES

[1] K. Toyama, "`Look, Ma --- No Hands!' hands-free cursor control with real-time 3d face tracking", In *Proc. of Workshop on Perceptual User Interfaces PUI'98*, San Francisco, USA, 1998, pp. 49-54.

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

[3] A.L. Ronzhin, A.A. Karpov, "Large Vocabulary Automatic Speech Recognition for Russian Language", In *Proc. of Second Baltic Conference on Human Language Technologies*, Tallinn, Estonia, 2005, pp. 329-334.

[4] A.L. Ronzhin, A.A. Karpov, "Implementation of morphemic analysis to Russian speech recognition", In *Proc. of 9th International Conference SPECOM'2004*, St. Petersburg: Anatoliya, 2004, pp. 291-296.

[5] F.M. Kulakov, A.I. Nechaev, S.E. Chernakova, "Modeling of Environment for the Teaching by Showing Process", SPIIRAS Proceeding, Issue No. 2, SPIIRAS, Russia, St. Petersburg, 2002, pp. 105-113.

[6] S.L. Oviatt, *Multimodal interfaces: In Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, J. Jacko and A. Sears, Eds. Lawrence Erlbaum Assoc. Mahwah, NJ, chap.14, 2003, pp. 286-304.

[7] D. Tzovaras, G. Nikolakis, G .Fergadis, S. Malasiotis and M. Stavrakis. "Design and Implementation of Haptic Virtual Environments for the Training of Visually Impaired", *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, Vol. 12, No. 2, pp.266-278, June 2004.