SPECOM'2004: 9th Conference
Speech and Computer
St. Petersburg, Russia
September 20-22, 2004

Assistive multimodal system based on speech recognition and head tracking

Alexey A. Karpov, Andrey L. Ronzhin (1), Alexander I. Nechaev, Svetlana E. Chernakova (2)

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
39, 14th line, St. Petersburg, Russia
(1) Speech Informatics Group
karpov@iias.spb.su, ronzhin@iias.spb.su
(2) Robotics Laboratory
nechaev@iias.spb.su, chernakova@iias.spb.su

Abstract

This paper describes the multimodal system, which was developed by two laboratories of SPIIRAS, for assistance to people with disabilities of hands. It combines automatic speech recognition and head tracking in joint multimodal system. The structure of the system, used methods for recognition and tracking, information fusion and synchronization, obtained results and testing conditions are described in the paper. Developed system was applied for hands-free operation for Graphical User Interface in such common tasks as Internet communication and text editing in MS Word. The experiments have shown that in spite of some decrease of operation speed the multimodal system allows to work 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.

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 systems, which allow to control 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. Here 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 in improving the life comfort of disabled people, their social protectability and independence from other people. Thus a hands-free control devices such as hands-free mouse and keyboard for disabled access to PC is one effective application of these technologies. Users who have difficulties using a standard devices could manipulate an on-screen cursor merely by moving their heads and giving the speech command instead of clicking the buttons.

Of course, there are hardware headgears for tracking the 3D pose of a person's head. In the accessibility community, several companies support software products that perform head tracking and speech control for PC. These products are accurate and reliable, but all they require either expensive dedicated hardware or structuring of the environment (special lighting, markings on the user's face, etc.) to simplify the tracking process [1]. So we propose the multimodal system with minimal cost for hardware and which can be used without any process of user's adaptation.

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 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 the keyboard for motor-disabled users.

Further we describe the developed multimodal system, which uses speech and head movements for input of information into the computer. The speech recognition and head tracking modules will be considered in detail. Then the process of information fusion and synchronization is described. Finally the structure of the complete assistive multimodal system and some examples of its using for hands-free PC control are presented.

2. SIRIUS speech recognition system

The main architecture of automatic system SIRIUS (**S**PIIRAS **I**nterface for **R**ecognition and **I**ntegral **U**nderstanding of **S**peech), developed in Speech Informatics Group of SPIIRAS, is presented in Figure 1. This architecture can be both speaker-independent and speaker-dependent. The difference in these models consists in the acoustical-lexical level.

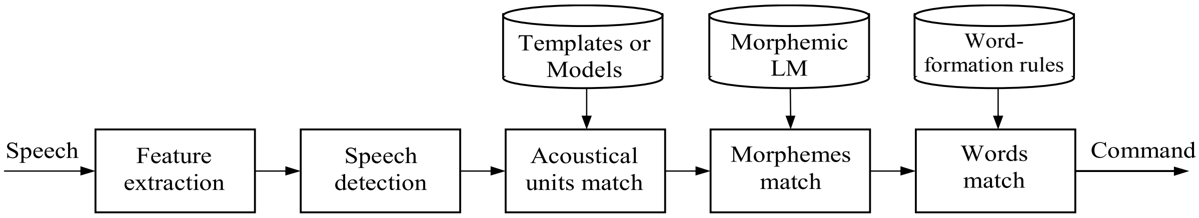

Figure 1. Structure of SIRIUS speech recognition system

For the speaker-dependent speech recognition system the acoustical templates of the whole words are used and for the speaker-independent recognition the stochastic acoustical models based on Hidden Markov Models (HMM) or Artificial Neural Networks can be used. In order to apply this speech recognition system for Russian language several original methods on the all levels of speech processing were developed. Further the used methods for preliminary speech signal processing, continuous speech recognition and language modeling will be described consecutively.

2.1.Initial signal processing

For parametrical representation of the speech signal two methods are used (mel-frequency cepstral features and spectral-difference features). The second method was developed for more robust work in the conditions of variations of the signal amplification level. In this method the subset of pairs of spectral bands is chosen from discrete spectrum and the further processing consists in comparison of the energies of the chosen bands considering some weight coefficients [3].

In principle, this method allows describing any forms of speech spectrum with any required accuracy, but very high accuracy is not necessary since there is redundant variability of spectrum of natural speech.

At present there are many methods for speech endpoint detection based on calculation of short-time signal energy, spectral energy, number of zero-crossing of signal, adaptive threshold values and information about duration of speech fragments. However all these algorithms become less reliable in conditions of non-stationary noise as well as at appearance of diverse sound artifacts (aspiration, lip smacks, etc).

In order to detect the speech signal in conditions of non-stationary noise environments the method based on spectral entropy analysis was developed. The distinction between entropy for speech segments and entropy for background noise is used for speech endpoint detection. Such criterion is less sensitive to variations of signal amplitude. The experiments with the developed method have shown that speech fragments are successfully selected in sound signals, which have diverse kinds of intense noises (including non-stationary) and sound artifacts [4].

2.2.Continuous speech recognition

For continuous speech recognition we have developed speaker-dependent method robust to grammatical deviations in a pronounced phrase and acceptable fast for using in the real time speech understanding model. In the proposed method we rejected completely from the composite templates approach and applied the detection of word hypotheses by the method of sliding analysis of an input signal [5]. The method is based on the following steps: (1) the multi alternative search of word hypotheses in the input signal by “sliding analysis” method with simultaneous estimation of their acoustical likelihood; (2) the recurrent construction of the set of hypotheses of the word chains with any length; (3) the estimation of acoustical lexical probability of these word chains (phrase-hypotheses) based on acoustical probability of words, their mutual time location and total duration of word hypotheses contained in the phrase.

The model was tested by: (1) continuously pronounced digits (obtained accuracy 96%); (2) the database of rephrasing for the demo-version of the model of voice operated flying object (obtained accuracy 86% at vocabulary in 300 words).

Also we compared the complexity of the proposed method (sliding analysis) with the method of complete enumeration of composed templates and also with the method of the isolated speech input. The complexity of the sliding analysis is significantly lower than the method of complete enumeration, and with the increase of vocabulary size or the length of a hypothetical phrase the unacceptable increase of phrase hypotheses is not observed.

The developed module of continuous speech recognition was introduced into the earlier developed base model of integral speech understanding [5]. The developed method of continuous speech recognition and the integral structure of processing provide the robustness to various distorting factors (acoustic-phonetic and grammar deviations in the pronounced phrases, etc.) that allows to make the character of interaction between the speaker and the system more natural and effective.

The speaker-dependent method of speech recognition has shown good results on recognition accuracy, and was successfully applied for many applied tasks on speech control with small vocabulary. But such perspective services as simultaneous voice machine translation, smart room and others which require large vocabulary speech recognition and property of speaker independency is necessary here.

Since pattern recognition methods require the long procedure of system training and do not provide the speaker independency the investigation of stochastic speech recognition methods was conducted.

Table1: Russian phonemes for speech recognition

Symbol	**Keyword**	**English gloss**
a	pa*r**a**	pair
a*	p**a***ra	pair
i	m'e*l'**i**	shoals
i*	m'**i***r	peace
e	d'e*r'**e**va	tree
e*	d'**e***r'eva	tree
ы	dы*r**ы**	holes
ы*	d**ы***rы	holes
u	tul**u***p	crude
u*	t**u**lu*p	crude
o*	g**o***rat	city
э*	tsep'	chain
p	**p**ы*l'	dust
p'	**p'**i*t'	to drink
b	**b**ы*t'	to be
b'	**b'i***t'	to beat
t	**t**o*st	toast
t'	**t'**e*n'	shadow
d	**d**ы*m	smoke
d'	**d'**e*n'	day
k	**k**o*t	cat
k'	**k'**i*t	whale
g	**g**u*s'	goose
g'	**g'**i*pk'ij	flexible
ts	**ts**e*p'	chain
ts'	**ts'**a*j	tea
f	**f**a*rs	farce
f'	**f'**i*z'ika	physics
v	**v**a*za	vase
v'	**v'**i*za	visa
s	**s**ы*n	son
s'	**s'**e*na	hay
z	**z**a*pax	smell
z'	ka**z'** i*na	basket
sh	**sh**ar	ball
sh'	**sh'**uka	pike
z	**z**ы*r	fat
x	**x**l'e*p	bread
x'	**x'**i*trыj	cunning
m	**m**a*j	May
m'	**m'**ata	mint
n	**n**ajt'i*	find
n'	**n'**i*t'	thread
l	**l**u*ts'	ray
l'	**l'**ubo*f'	love
r	k**r**ap	crab
r'	**r'**e*zat'	cut
j	i**j**u*l'	July

Now let's consider the developed speaker-independent speech recognition system. The acoustical models of phonemes in the form of triphones are used. The multi Gaussian HMM are used for phonemes modeling and the models of words are obtained by concatenation of phonemes models. For Russian language we use 48 phonemes (Table 1). This set of phonemes is the modification of SAMPA (Speech Assessment Methods Phonetic Alphabet) for Russian language. In the Table 1 the symbol " * " means that vowel is stressed, and the symbol " ' " means that consonant is soft. Thus we use for speech recognition 12 vowels and 36 consonants.

2.3. Language modeling and words matching

In contrast to English the Russian language has much more variety on word-form level and so the size of recognized vocabulary is sharply increased as well as quality and speed of the processing are decreased. Moreover the usage of syntactic constraints leads to that the errors of declensional endings cause the recognition error of the whole pronounced phrase.

Since during the process of word formation the same morphemes are often used then it will be useful to insert the additional level of speech representation – morphemic level. Owing to division of word-form into morphemes the vocabulary size of recognized lexical units is significantly decreased. At that during recognition the degree of co-ordination between root morphemes will have main significance. As a result of such processing the speed of recognition and robustness to syntactical deviations in the pronounced phrase will be improved [6].

For effective language modeling we use the method based on associative analysis [7]. Associative analysis of morphemes chains (words and phrases) is an alternative to N-gram method and based on the following assumptions:

- Semantic and syntactic relations are realized in human sub-consciousness (and in the brain) by the associations mechanism in the same process.
- The associative connection between two morphemes in words can be evaluated using the bigram statistics or expert estimations.
- Different morphemes chains can be estimated according to their degree of relationship based on inter-morphemic associations.

The associative model contains the compatibility estimations of all ordered pairs of morphemes (or words) of the vocabulary. During the estimation of phrase hypotheses each pair of morphemes contained in the phrase gets the associative estimation by 4-score scale (4 – excellent compatibility; 3 – good compatibility; 2 – satisfactory compatibility; 1 – bad compatibility).

On the output of morphemes matching level we obtain the N-best list of morphemes chains, evaluated by acoustical and semantic-syntactic (associative) criteria. For performing the words matching we use the rules of word-formation for Russian language and the vocabulary of word-forms [6].

The result of speech recognition is the best hypothesis of speech utterance, optimal according to acoustical-lexical and syntactic-semantic estimates. Further the recognized speech command follows to the module of information fusion. At the same time to this module the cursors position data, which is calculated in the head tracking system developed in Robotics Laboratory of SPIIRAS, are entered. In the following section we consider this system in detail.

It is necessary to emphasize that for the task of voice command recognition, where the size of vocabulary does not rich 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.

3. The head tracking system

This section proposes a new intelligent Interface using Head Tracking System (HTS) for tracking natural man-operator's head motion instead of hand-controlling motion. In near future we intend to use the HTS measuring man-operator's gaze direction instead of the mouse or joystick for control of cursor position on the screen.

3.1. HTS hardware design

Hardware of the HTS prototype includes the following units (Figure 2):

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

When HTS realized as active one, CU is equipped with black & white cameras with IR lenses. When a passive variety is studied, CU is equipped with color cameras. Number of cameras in CU and their position are defined by configuration of the HTS hardware [8].

CCU is purposed for power supply to CU and for mutual cameras synchronization. In the case of active HTS, CCU also controls the camera exposition and pulse supply to RDU IR LEDs (Infra Red Light Emitting Diodes).

RDU for active HTS is a rigid construction with LED's or color reference marks for passive HTS, mounted on the head.

VPU is designed as a set of standard card for PC Pentium 3(4). An advantageous feature of the HTS prototype is employment of conventional personal computer Pentium 3(4) for realizing image processing algorithm, HTS hardware control and interfacing with external equipment in real time mode.

3.2. Operational principle of HTS prototype

The HTS hardware is represented by a functional scheme combining two varieties of optical HTS: active and passive. Consider the basic principles of active HTS and differences with passive one (Figure 3).

1) Human operator performs natural head movements in the Head Motion Box (HMB) volume. 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 turn (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 variety of HTS and color for passive one) rigidly mounted on the RDU base and 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 within the HMB.

4) Reference mark images projected on camera's Focal Plane Array (FPA) will have coordinates (*Ximg*, *Yimg*) in the image (camera) coordinate system (*Xc*, *Yc*).

5) The CU control, power supply and synchronize are executed by 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 card at PC Pentium-3 (4). Basic VPU functions are the following: video signal digitization, filtering and selection of reference images on the background and, also, calculating center coordinates with sub-pixel accuracy.

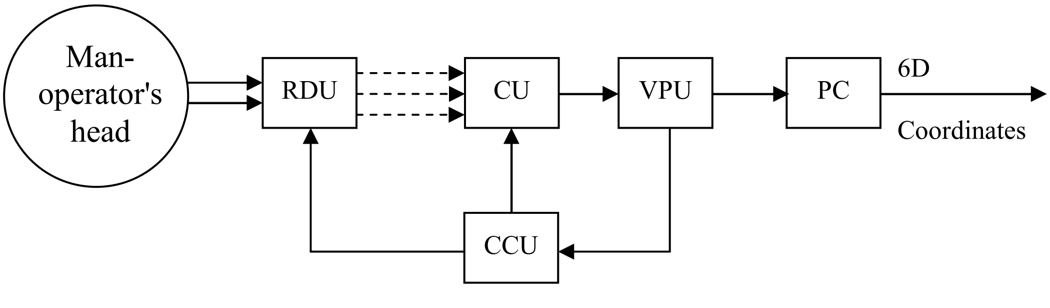

Figure 2. Structure diagram of HTS prototype hardware

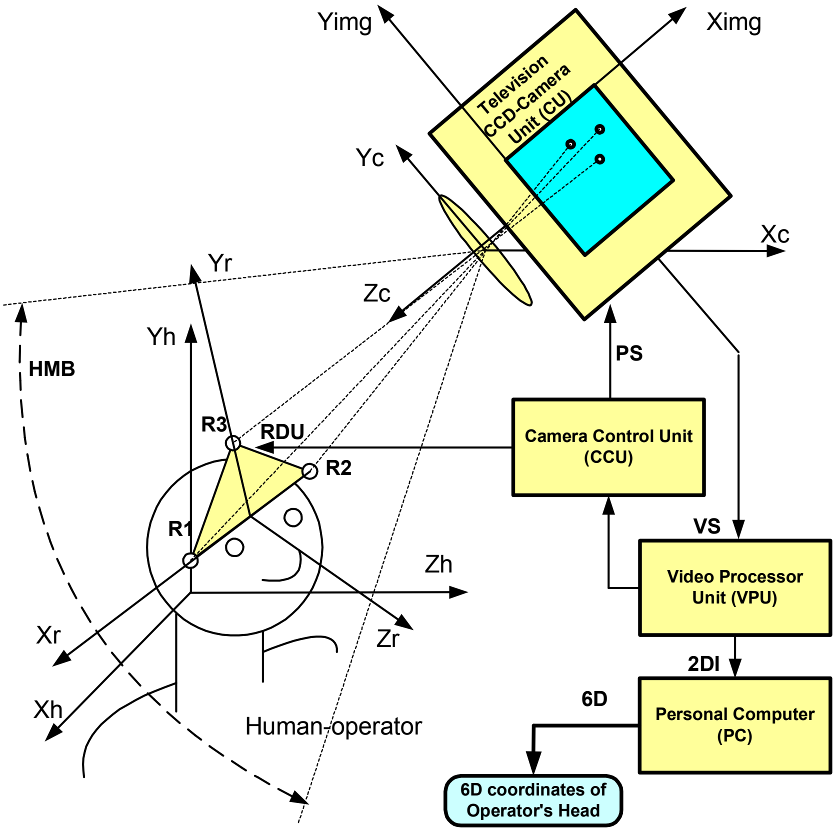
Yimg
Ximg
Television CCD-Camera Unit (CU)
Yc
Xc
Yr
Zc
Yh
HMB
PS
R3
RDU
R2
Camera Control Unit (CCU)
R1
VS
Video Processor Unit (VPU)
Zh
Xr
Zr
2DI
Xh
6D
Personal Computer (PC)
Human-operator
6D coordinates of Operator's Head

Figure 3. Functional diagram of HTS prototype

7) Reference marks' coordinates (*Ximg*, *Yimg*) from VPU are entered the PC memory. For known internal and external parameters of the cameras optical system, which are corrected in the procedure of camera calibration, and for coordinates of reference mark images (*Ximg*, *Yimg*) 3D coordinates of the real position of reference mark in the camera coordinate system (*Xc* , *Yc*, *Zc*) are computed.

8) Using the HTS prototype software, installed at the PC, the reference mark images are processed for selection and identification, and RDU position and orientation in the CU coordinate system are computed too.

3.3. The frame-structural model in Head Tracking System (HTS)

In the HTS image processing and computation of head pose coordinates (position & orientations) are made basing on a priori 3D wire-frame head model.

As the model of head, face, and the 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) [9].

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 defining 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 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, j reference marks in the model (RM_{ij}) are normalized relative 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, 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}).

3.4. 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 types (for the markless HTS – model of operator's face / head) generating it basing on 2D models of images in the camera system of coordinates. Using a 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 places of reference marks on the camera image plane basing on determined speed vectors of their movement.

3) Using color gradient selection of passive reference marks for their localization and identification on the background and, also, for obtaining coordinates of reference mark image centroids with subpixel accuracy.

The HTS' main software for image processing and parameters adjustment with 3D model data in real time mode is presented in Figure 4.

1) Basic information data:

- Input mark images (INPUT IMG);
- Image after spatial-temporal filtering (VI);
- Image after mark feature image selection (PI);
- Image after reference mark identification (2DI);
- Output coordinates of RDU (head) position and orientation (OUTPUT 6D).

2) Auxiliary 3D model data for automatic algorithm's adjustment in real time mode:

- Adjustment parameters for image-temporal filter with image movement prediction (MV);
- Adjustment parameters for variation of mark image properties (size, color, shape, orientation etc.) for RDU (head) image (MP);
- Data for comparison of mark images with 2D model image (M2D);
- Data for correcting 3D RDU model from results of computation (M3D).

3) Auxiliary man-operator's data for manual algorithm's adjustment: (1) Adjustment parameters for spatial-temporal filtering (HV); (2) Adjustment parameters for specific feature selection (HP); (3) Mark image coordinates identified by operator (H2D);

4) Displaying results of spatial coordinates' computation for operator (H3D).

3.5. Head tracking system usage

The design of RDU mounted on the miniature telephone garniture on head of human operator controlling a remote robot-manipulator was demonstrated at International Conference "Novel Information Technologies and Information Assurance", Binghamton, USA, March,4-7, 2002.

This RDU design is universal one operating both in active (with IR diodes) and passive (with colored marks) HTS modes. Some original solutions were used in the design. One of HTS's hardware version is realized on USB-camera with light-weight RDU, (Figure 5).

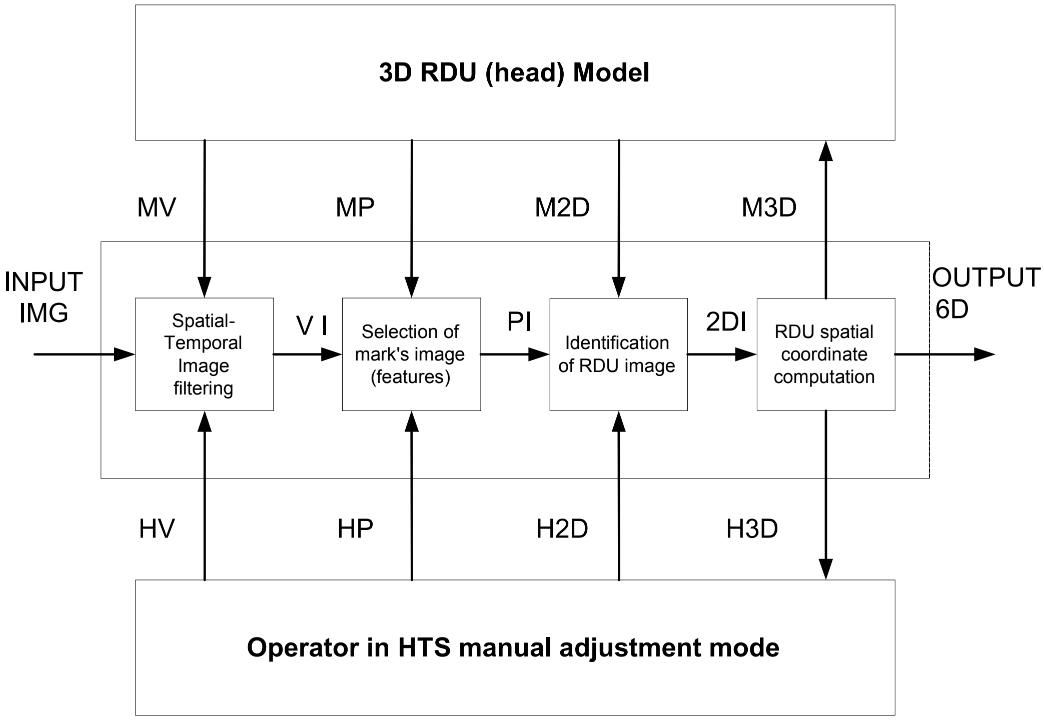
3D RDU (head) Model
MV
MP
M2D
M3D
INPUT IMG
Spatial-Temporal Image filtering
V I
Selection of mark's image (features)
PI
Identification of RDU image
2DI
RDU spatial coordinate computation
OUTPUT 6D
HV
HP
H2D
H3D
Operator in HTS manual adjustment mode

Figure 4. Functional diagram of HTS algorithm

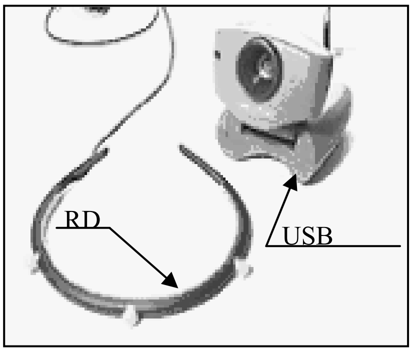

Figure 5. Design of RDU for HTS prototype

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 additional study was carried out for establishing a possibility of creating a camera with the speed four times higher than conventional ones, with the frame rate 100 Hz.

The video processor's (VPU) hardware is implemented on standard card (in PCI slot).

The results of experiments with HTS prototype, RMS error see in Table 2 [10].

Table 2: RMS error

it	1.		2.	
Name of parameter	RMS (σ) error (without filtering)		RMS (σ) error (with median filtering)	
Units	μm	arc. min.	μm	arc. min.
X	23		18	
Y	33		28	
Z	33		40	
φ_x		1		1
φ_y		2,2		1,8
φ_z		2,6		1,8

Head motion box (HMB) measurement results for HTS prototype (see in Table 3).

Table 3: HMB measurements

it	1.		2.			
Name of param.	Maximal zone at distance RDU to CU (Z=400 mm)		Maximal zone at distance RDU to CU (Z=800 mm)		Maximal zone at distance RDU to CU (Z=1600 mm)	
Units	mm	deg.	mm	deg.	mm	deg.
X	250		660		1450	
Y	240		500		1050	
φ_x		±90		±88		±88
φ_y		±40		±33		±30
φ_z		±34		±30		±28

3.6. Main features of HTS usage with SIRIUS for MMI

The sequence of assistive MMI process is a dynamical cooperation (combination) of 2 bi-direction modalities:

(**A**) audio (speech recognition - PC-sound generation),

(**V**) video (image recognition - image displaying - head/hand tracking).

The main features of proposed HTS for MMI:

- low-cost (simple, standard) hardware with special software,
- light-weight microphone garniture with RDU for head tracking,
- optical accuracy measurements with automatic correction,
- natural working conditions (sun-light interference protection),
- 3D control of position and orientation of computer-synthezed objects.
- combine of voice and gesture commands for high reliability,
- text word commands (script, language, rolling of text, text fragment search by natural head movement (up-down, left-right),
- pictures or video, panorama observing and montage, sorting etc.,
- 3D computer design by natural movements of head/hand,
- real (and/or virtual) 3D image viewing and observation.

4. Mechanism of multimodal fusion

The term "information fusion" encompasses any area which deals with utilizing a combination of information acquired from multiple sources (sensor, databases…), either to generate an improved representation, or to reach a more robust decision (for example, in information retrieval systems or in device control systems). Humans utilize multimodal data fusion every day. Some examples are: use of both eyes, seeing and touching, or seeing and hearing which improves intelligibility in noisy situations.

Multimodal fusion of information is very important building block for the various modalities used to exchange information between humans and machines: source coding and channel coding, modality modeling, multimodal and intramodal co-registration, stochastic models of different modalities, perception and degradation of signals (minimizing perception degradation under non-optimal transmission of signals), multimodal fusion of information (like speech prosody and emotions from facial recognition), and machine learning and modalities.

In developed multimodal system two modalities are used: speech and head movements. As both modalities are active [11], 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 irrespective to the cursor). Common architecture of bimodal system is presented in Figure 6.

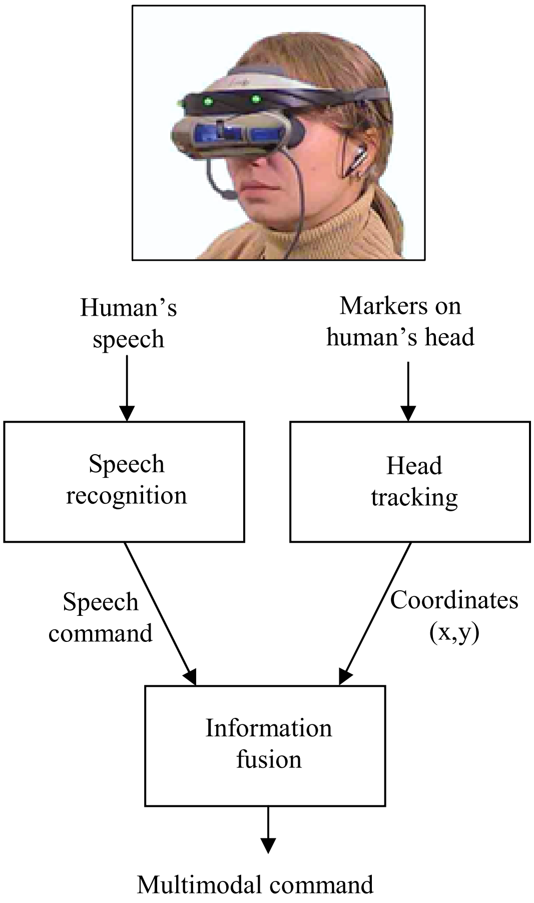

Figure 6. General structure of multimodal system

In contrast to unimodal systems during development of multimodal interfaces it is appeared the new key problems connected with synchronization, joint processing and fusion of multimodal information.

In the developed system the synchronization of modalities is performed by natural 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 phrase utterance the cursor can be moved, but 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 by required data and on completion the signal for command execution is given.

5. The results of MMI system usage

Developed multimodal system is the software-hardware complex for hands-free control for Graphic User Interface (GUI) of PC. AS hardware the following equipment are used: microphone Sony DR-50 with built-in signal amplifier, connected to Sound Blaster Creative Labs Audigy 2 and HTS's hardware version realized on USB-camera with light-weight RDU.

In Table 4 the list of speech commands, which a human can enter into the system, is presented.

Table 4: 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 "Start" button
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
Button down	Mouse left button down
Button 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
End text	End input of text
Write …	Write uttered text
Next	Open next page
Previous	Open previous page
Select all	Select all text in document
Favorites	Open Favorites menu
New	Open blank document
Enter	Press "Enter" button
Escape	Press "Escape" button

Instead of dots (…) any message can be said, which will be recognized in the case when dictation system works. Several commands (for instance, End text) additionally have for sound confirmation of execution.

Such set of commands allows operating by GUI of the operational system Windows. For fulfillment the testing and debugging of the system some scenarios of work with GUI without manual control were selected. All test scenarios were divided into two common groups: work with text editor MS Word and Internet access by means of MS Internet Explorer.

In Table 5 the fragment of operating of a user with Internet Explorer for obtaining information about currencies rate of exchange at web-site www.rambler.ru is presented.

Table 5: Fragment of operation with GUI

N	Task description	Action
1	Open “Start” menu	Start
2	Select “Internet Explorer” icon	(Head)
3	Run Internet Explorer	Left button
4	Open “Favorites” menu	Favorites
5	Select RAMBLER in Favorites	(Head)
6	Run RAMBLER web-site	Left button
7	Scroll the window to bottom	Scroll down
8	Select “Rate of exchange” hyperlink	(Head)
9	Open “Rate of exchange” hyperlink	Left button
10	Close focused window	Close window
…		

For testing 100 simple tasks covering two domains were fulfilled by 5 users (4 untrained users and one experienced user). Approximately half of the tasks comprised the text processing sequence and half the web browsing sequence. In the Table 6 the results on task performance for indicated above tasks sequence are presented.

Table 6: Task performance results

User	Time of select (mouse), sec	Time of select (HTS), sec	Time of click (mouse), sec	Time of command (SIRIUS) sec
1	1,00	2,00	0,10	0,50
2	1,50	3,00	0,20	0,55
3	0,50	1,00	0,08	0,35
4	1,00	1,50	0,15	0,50
5	0,50	1,50	0,10	0,50
Aver.	0,90	1,80	0,13	0,48

From the Table 6 we can obtain the time for one cycle of access to the Internet by means of traditional way (mouse (Δt_I^{T})) and using the system for tracking the head movements jointly with commands giving by voice (Δt_I^{MMI}):

$$\begin{cases} \Delta t_I^{MMI} = 3\ t_{HTS} + 7\ t_{SIRIUS}. \\ \Delta t_I^{T} = 3\ t_{mouse\ select} + 7\ t_{mouse\ click} \end{cases}$$

Time for one cycle of Internet access according to the experiments (Tables 5 and 6) equals:

$$\begin{cases} \Delta t_I^{MMI} = 3 * 1{,}8 + 7 * 0{,}48 = 8{,}76 \text{ (Sec.)} \\ \Delta t_I^{T} = 3 * 0{,}9 + 7 * 0{,}13 = 3{,}61 \text{ (Sec.)} \end{cases}$$

Thus the developed multimodal way of Internet access is in 2,4 times slower than traditional way. However this fall is acceptable since developed system is intended mainly for disabled people.

The experiments according to accuracy of speech recognition have shown that using vocabulary in 110 words the accuracy does not decrease lower than 97% for each of 5 users.

It allows concluding that developed assistive multimodal system can be successfully used for hands-free PC control for users with disabilities of their hands or arms.

Another good application of hands-free cursor control allows users to change the “focus of window” in a window 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. Finally, there are applications of hands-free cursor control for entertainment such as: painting programs, games, designing, etc.

6. Conclusion

Thus the result of joint work of two laboratories of SPIIRAS is developed assistive multimodal system. 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. In future the system will be combined with head gestures (for instance, up-down) recognition system, that will allow to increase the reliability of work due to doubling some speech commands with an involuntary head gestures.

7. References

[1] Toyama K., "`Look, Ma --- No Hands!' hands-free cursor control with real-time 3d face tracking," in

Proc. Workshop on Perceptual User Interfaces (PUI'98), (San Francisco), pp. 49-54, 1998.

[2] Bates R. Istance H.O. “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), Trinity Hall, University of Cambridge, 2002.

[3] Karpov A., Kosarev Yu., Ronzhin A., Lee I., “Development of acoustical features of speech signal robust to variations of signal scale and spectrum deformations”, Proceedings of Conference "Theory and Practice of speech investigations" ARSO-2003, Moscow, Russia, pp. 83-88, 2003.

[4] Karpov A. “The robust method for speech endpoint detection based on spectral entropy”. Scientific journal “Artificial Intelligence” of the Institute of Artificial Intelligence of the National Academy of Sciences of Ukraine, Volume 4, 2004.

[5] Kosarev Yu., Ronzhin A., Lee I., Karpov A.. “Continuous Speech Recognition without Use of High-Level Information”, Proceedings of 15-th International Congress of Phonetic Sciences”, Barcelona, Spain, pp. 1373-1376, 2003.

[6] Ronzhin A.L., Karpov A.A. Implementation of morhpemic analysis to Russian speech recognition. Proceeding of International Conference SPECOM’2004, St. Petersburg: “Evropeiski Dom”, 2004.

[7] Kosarev Yu., Lee I., Ronzhin A., Karpov A., Savage J., Haritatos F. “Robust Speech Understanding for Voice control system”. Proceedings of Workshop SPECOM'2002, St. Petersburg, pp. 13-18, 2002.

[8] Kulakov F.M., Nechaev A.I., Efros A.I., Chernakova S.E., “Hard & software means of MMI for telerobotics using systems tracking human-operator motions”, Proc. of III International conference «Cybernetics and technology of XXI century» Voronezh, Russia, pp. 516-534, October, 2002.

[9] Kulakov F.M., Nechaev A.I., Chernakova S.E., “Modeling of Enviroment for the Teaching by Showing Process”.// SPIIRAS Proceeding, Issue No. 2 , SPIIRAS, Russia, St-Peterburg, pp. 105-113, 2002.

[10] Kulakov F.M., Nechaev A.I., Efros A.I., Chernakova S.E., “Experimental study of man-mashine interface implementing tracking systems of man-operator motions” the Proceedings of Sixth International Seminar on Science and Computing, Moscow, Russia, September 2003.

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