An Assistive Bi-modal User Interface Integrating Multi-channel Speech Recognition and Computer Vision

Alexey Karpov, Andrey Ronzhin, and Irina Kipyatkova

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
SPIIRAS, 39, 14-th line, 199178, St. Petersburg, Russian Federation
{karpov,ronzhin,kipyatkova}iias.spb.su

Abstract. In this paper, we present a bi-modal user interface aimed both for assistance to persons without hands or with physical disabilities of hands/arms, and for contactless HCI with able-bodied users as well. Human being can manipulate a virtual mouse pointer moving his/her head and verbally communicate with a computer, giving speech commands instead of computer input devices. Speech is a very useful modality to reference objects and actions on objects, whereas head pointing gesture/motion is a powerful modality to indicate spatial locations. The bi-modal interface integrates a tri-lingual system for multi-channel audio signal processing and automatic recognition of voice commands in English, French and Russian as well as a vision-based head detection/tracking system. It processes natural speech and head pointing movements in parallel and fuses both informational streams in a united multi-modal command, where each modality transmits own semantic information: head position indicates 2D head/pointer coordinates, while speech signal yields control commands. Testing of the bi-modal user interface and comparison with contact-based pointing interfaces was made by the methodology of ISO 9241-9.

Keywords: Multi-modal user interface; assistive technology; speech recognition; computer vision; cognitive experiments.

1 Introduction

Most of recent research in Human-Computer Interaction (HCI) has focused on equipping machines with means of communication that are used between human beings, such as speech, gestures, tactile interfaces. These interfaces are generally developed for ordinary people without disabilities; however, there is a lack of HCI research towards interfaces that are specifically developed for diverse groups of handicapped people. It is clear that a hearing-impaired person cannot use a speech interface, whereas a hand-disabled person cannot use a manual gesture interface and keyboard/mouse devices. Nowadays the world society pays much attention to the problems of physically and mentally handicapped persons with partial or full dysfunctions of their body parts and organs. There are several kinds of physical disabilities, manifesting themselves in impairments of speech, hearing, vision, and motion impairments such as walking or moving fingers. Many governmental programs have been launched for social and professional rehabilitation and support of disabled people.

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2011, LNCS 6762, pp. 454–463, 2011.

For instance, because of a disaster, or inborn disabilities some people are unable to operate a personal computer and type by a keyboard or a mouse/touchpad due to disabilities of their hands/arms. It much restricts their interaction abilities with diverse information system and results in reducing social status. In our research, it is proposed one solution for these persons; it is a bi-modal system, which allows controlling a computer without the traditional control devices, but using: (1) head (or face) motions to control the mouse cursor through the monitor screen; (2) speech input for giving the control commands. This system relates to the class of multi-modal user interfaces and systems, which are aimed to recognize naturally occurring forms of natural language and behavior, and incorporate several recognition-based technologies (speech recognition, image analysis and computer vision, handwriting recognition etc.) [1]. Multimodal user interfaces can process two or more combined natural user input modes such as speech, touch, manual gestures, or head and body movements, in a coordinated manner with multimedia system output and allow choosing an accessible way of interaction in a concrete application for each concrete user.

The first multi-modal user interface integrating two modalities such as voice information and manual gestures for personal HCI has been proposed by R. Bolt in early 80s [2]. Probably, the first attempt to develop an assistive bi-modal interface employing head pointing and speech control for hands-free HCI oriented to ordinary and impaired people was made in late 90s [3]. There are also more recent similar assistive systems; for instance, in [4] it is proposed a vision-based uni-modal user interface for hands-free HCI based on user's nose detection and tracking by a computer vision system; in [5] it is proposed a voice-based assistive user interface (Vocal Joystick); another research in [6] presents a bi-modal user interface, based on voice input and head tracking, for some home appliances.

In the given paper, it is proposed a novel multi-modal user interface that integrates multi-channel speech recognition and computer vision technologies. Any person, who has troubles with standard computer input devices (e.g. mouse or keyboard) could manipulate mouse cursor by moving head and giving speech command instead of clicking buttons. The bi-modal interface combines modules for automatic speech recognition and head tracking in one system.

2 Architecture of the Assistive Bi-modal User Interface

The proposed bi-modal user interface has been named ICANDO (an Intelligent Computer AssistaNt for Disabled Operators) and it is intended mainly for assistance to persons without hands or with disabilities of their hands/arms. ICANDO integrates one software module for automatic recognition of multi-lingual voice commands in English, French and Russian, as well as another vision-based module for head detection and pointing. It processes natural human's speech and head motions in parallel and then fuses both informational streams in a joint multi-modal command for operating graphical user interface (GUI) of a computer. Each of the modalities transmits own semantic information: head position indicates 2D head/pointer coordinates, while speech signal yields control commands, which must be performed with an object selected by the pointer or irrespectively to its position. The multi-modal user interface has been implemented in two different versions:

1. Low-cost computer interface [7], which is available for most of potential users with any computers. It employs a standard web-camera priced under 50 $. USB web-camera Logitech QuickCam for Notebooks Pro is used in this version. It captures both one-channel video signal in 640x480x25fps mode and one-channel audio signal with 16 KHz sampling rate and mono format with acceptable SNR via the built-in microphone.
2. Advanced multi-modal interface (Figure 1), which simultaneously processes both one-channel video signal obtained by the web-camera Logitech and multi-channel audio signal captured by an array of four professional microphones Oktava connected to the external multi-channel sound board Presonus Firepod.

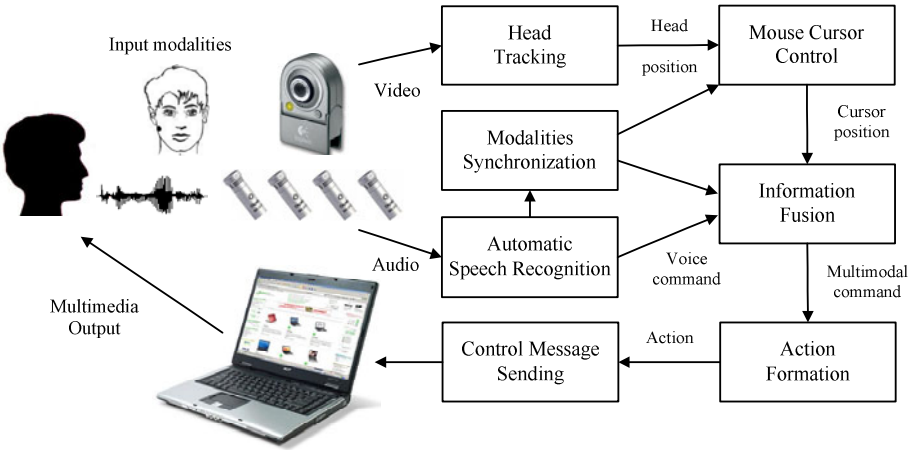

Fig. 1. Software-hardware architecture of the assistive bi-modal user interface

SIRIUS (SPIIRAS Interface for Recognition and Integral Understanding of Speech) speech recognition engine [8] is applied for automatic recognition and interpretation of input voice commands. For the speech parameterization Mel-frequency cepstral coefficients (MFCC) with 1st and 2nd derivatives are used. Modeling and recognition of phonemes and words in the system are based on the first-order continuous Hidden Markov Models (HMM) [9]. The system has been trained to understand 30 voice commands in each of 3 working languages in the speaker-dependent mode. Speaker-dependent automatic speech recognition is more adequate for the present task rather than a speaker-independent one, since it provides a lower word error rate. All the voice commands are divided into four classes according to their functional purpose: pointer manipulation commands ("Left/Click", "Double click", "Scroll up", etc.); keyboard commands ("Enter", "Escape", etc.); GUI commands ("Start", "Next", "Previous", etc.), and special system commands. However, only the pointer manipulation commands are multi-modal ones and require pointer coordinates for the information fusion.

In the advanced assistive interface additionally there is a software module for multi-channel audio signal processing. It helps to improve quality of voice activity

detection (VAD) in noisy environments, for example, when there are some talking people in the room. This module performs spatial speaker localization that is based on the general cross correlation phase transform (GCC-PHAT) method [10] and applies a microphone array in "reversed T" shape (Figure 2, left) [11] consisting of 4 cardioid microphones Oktava MK-012 (Figure 2, right) connected to a laptop via the FireWire interface of the multi-channel sound board Presonus Firepod. The construction of the microphone array is made as a rigid construction in the reversed T-shape, where 3 microphones of 4 are located linearly between the external sound board box, lying on the table, and bottom part of the laptop, whereas the fourth upper microphone locates horizontally above the laptop screen. Microphone cables are plugged into the sound board behind the laptop construction. Thus, all the microphones are maximally far from each other and such microphone's non-linear configuration allows calculating 3D coordinates of sound sources. Estimation of correlation maximum of the mutual spectrum for all the signal pairs allows evaluating phase difference between speech signals in array channels. Further calculation of 3D speech source coordinates is made by the triangulation method. Restriction of a working zone for a user in front of the laptop screen to 0.7 m (radius of the zone) allows the system to eliminate outer acoustic noises and discard useless speech of other people improving quality of voice activity detection. The developed microphone array provides error of speech source localization less than 5º.

Fig. 2. Model of microphone array (left); one microphone of the array (right)

In the proposed bi-modal user interface, the web-camera Logitech is used jointly with the software module of computer vision based on OpenCV library in order to detect and track natural operator's head gestures instead of hand-controlling motions. It captures raw video data in 640x480x25 fps format. At the system start, user's head is automatically searched in video frames, employing an object detection based on Viola-Jones method [12] with Haar-like features of human's head model [13]. It is able to find rectangular regions in a given image that likely contain human's face. Region of interest has to be larger than 220x250 pixels in frames of 640x480 points allowing the system to find only one closest and biggest head in image, accelerating visual processing. Then the computer vision system processes optical flow for continuous tracking of five natural facial markers: tip of nose, center of upper lip,

point between eyebrows, left eye (iris) and right eye. The head tracking method applies a basic iterative Lucas-Kanade algorithm [14] for optical flow with the pyramidal implementation [15]. A mouse pointer controlled by user's nose was also proposed before, for instance in [4]; however, the set of 5 facial points improves robustness of face tracking [7].

Synchronization of two information streams (modalities) is performed by the speech recognition module, which sends messages to store pointer coordinates, calculated by the head tracking module, and for information fusion. Pointer 2D coordinates are taken at a moment of start triggering the voice activity detector, instead of the moment of completion of speech recognition process. It is connected with the problem that, when speaking, a user involuntarily slightly moves his/her head so that, at the end of the speech command recognition, the pointer may indicate another graphical object.

ICANDO uses a late-stage semantic architecture [16], and the recognition results involve partial meaning representations, which are integrated by the information fusion component controlled by the dialogue manager. Fields of a semantic frame (speech command index, X coordinate, Y coordinate, command type) are filled in with the required information, and an action corresponding to the multi-modal command is made on completion of speech recognition. ASR module operates in the real-time mode (< 0.1xRT); since the vocabulary of voice commands is small, there are minor delays between an utterance and fulfillment of corresponding multi-modal command. If a speech command is real multi-modal one (pointer manipulation commands only) then it is combined with stored coordinates of the pointer and a message to the mouse virtual device is sent. If a voice command is uni-modal one, coordinates are not taken into account and a message to the keyboard device is posted.

The developed assistive bi-modal system has been installed on a laptop with a four-cored processor Intel Core2Quad and a wide screen of 15". The multi-modal system takes advantage of the multi-cored architecture of PC in this case.

3 Cognitive Experiments on Human-Computer Interaction

Quantitative evaluation of the developed hand-free interface was carried out using the methodology of ISO 9241-9 [17], which is based on Fitts' law experiments [18] and related works [19]. ICANDO has been quantitatively evaluated by six volunteers, including four beginners in hands-free HCI and two developers/experts. Working with ICANDO the subjects seated at a table about 0.5 meters far from the 15" laptop's screen. Prior to the experience, subjects are shown a short demonstration of the task to be performed. Then, the subjects are allowed a short training period, and instructed to click targets as quickly as possible (in order to comply with Fitts' law hypothesis). The users were instructed to point and to select 16 ordered targets/circles, with a circular layout (Figure 3, left) [20], so pointing movements must be carried out in different directions. When selection occurs, the former target is shadowed and the next target is displayed. Figure 3 (right) shows one sample of real trajectory of mouse cursor movement at hands-free HCI with head pointing.

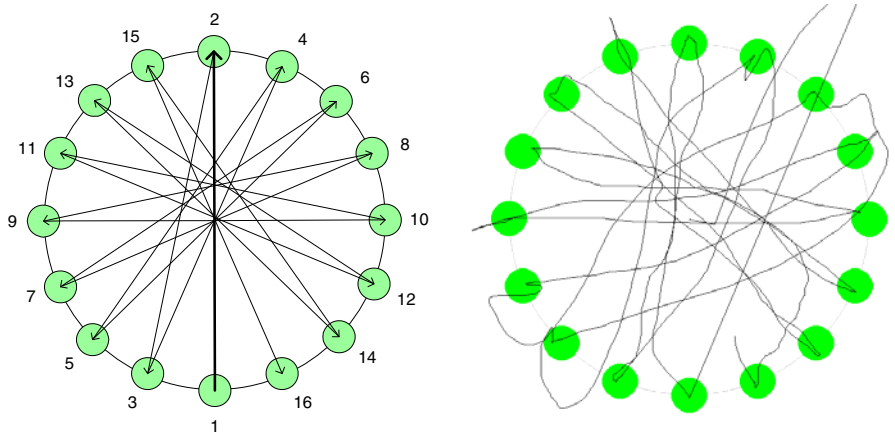

Fig. 3. Layout of round targets on screen for ISO 9241-9 based experiments (left); Example of trajectories of cursor movement at hands-free HCI with head pointing (right)

The experiments with several pairs of targets width/distance, corresponding to different indexes of difficulty were carried out by each subject. The index of difficulty *ID* of the task, measured in bit by $ID = \log_2\left(D/W + 1\right)$, where *D* is the distance between targets and *W* is the target's width. In the experiments *ID* varied in the range from 1.32 till 4.4 (10 different values). However, a location where selection occurs influences both on effective distance and effective width. The effective index of difficulty is $IDe = \log_2\left(De/We + 1\right)$, where *De* is the effective distance between the first and last points of a pointer trajectory; *We* is the effective target width: $We = 4.133\sigma$, where σ is the standard deviation of the coordinates of the point of selection, projected on the axis between the centers of the origin and destination targets [21].

Fitts' law states that the movement time *MT* between two targets is a linear function of *ID* of the task related to the targets' characteristics. Figure 6 shows the movement time *MT* versus the effective index of difficulty *IDe* for ICANDO system. For each trial, the inter-target movement time is defined and measured as the time between two successive selection events, selection occurs counted both inside and outside targets (selection error). Figure 4 plots dependences between *MT* and *IDe* for two groups of users (beginners and experts). These plots show that trained users capable to perform the given task much faster than novices; however, the same is true for any contact-based pointing devices as well.

Throughput *TP* is the linear ratio between *IDe* and *MT* measured in bits per second [22]. Mean *TP* allows comparison between performances of different pointing interfaces. Standard contact-based pointing devices such as a mouse, touchpad, trackball, joystick and 17" touchscreen were also evaluated by the ISO 9241-9 standard in order to compare their performances with those of the proposed contactless interface. Table 1 shows averaged values of movement time *MT* and effective throughput *TPe*, which is a tradeoff between pointing speed and target selection precision.

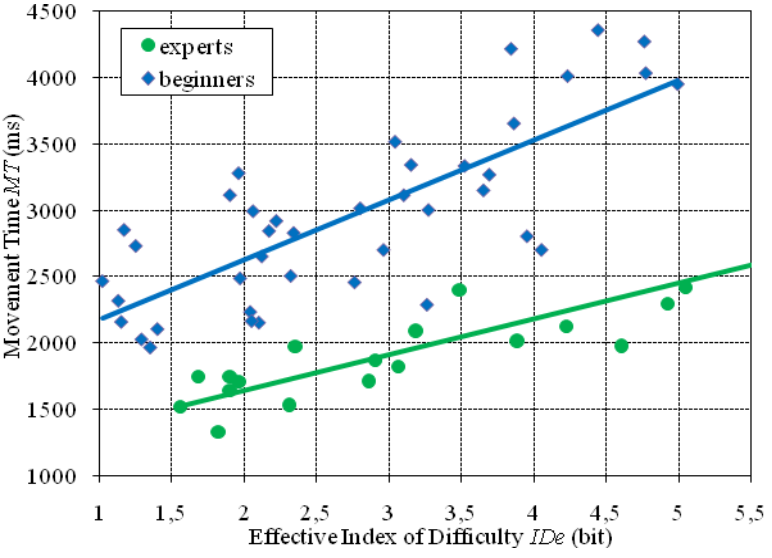

Fig. 4. Dependence of Movement Time *MT* on Effective Index of Difficulty *IDe* for two groups of users

The best *TPe* results have been obtained with the contact-based touch screen and the optical mouse device, taking into account that a touch screen is not so precise for small *W*. Performance results obtained with ICANDO interface, as well as the joystick and trackball devices are rather similar. However, the main advantage of the developed bi-modal interface is that it provides natural hand-free way of human-computer interaction.

Table 1. Quantitative comparison of performances of pointing interfaces using cognitive Fitts' law experiments

Interface of HCI	*MT*, seconds	Selection error, %	*TP*, bit/second
Joystick	2.01	7.00	1.54
Trackball	1.03	3.83	3.51
Touchpad 3"	0.85	4.50	3.72
Optical mouse	0.49	3.17	6.65
Touchscreen 17"	0.50	6.17	7.85
Head pointing + voice	1.98	7.33	1.59

Moreover, one experienced human being applied the basic bi-modal assistive system for the Internet surfing task in several HCI sessions during one day. A statistical analysis of the system's log has shown that about 700 meaningful voice control commands were issued excluding some out-of-vocabulary words and fails of

VAD. However, some speech commands were more frequent than others, and some other commands were not used at all. Figure 5 presents a relative distribution of the issued voice commands in this cognitive experiment.

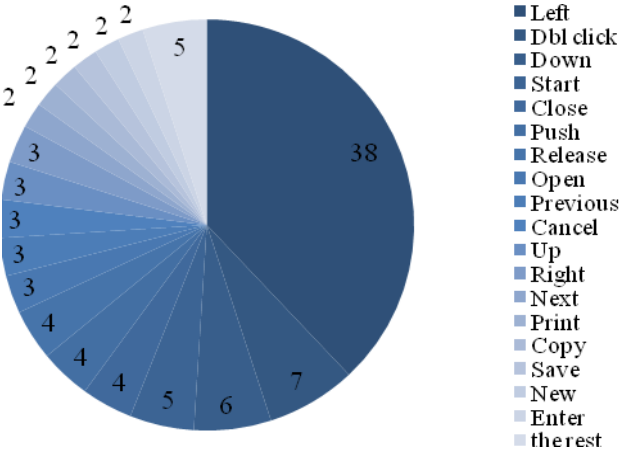

Fig. 5. Relative distribution of voice commands usage at experiments on hands-free HCI

It was predictably that the most important and frequent speech command is "Left" (more than 1/3 of all cases), because it replaces click of the left mouse button that is used very often (for instance, when typing letters in fields of some windows and forms). Internet surfing task sometimes requires to enter text data (for instance, URLs) and user can do it by ICANDO and virtual On-Screen keyboard, confirming pointed letters by the command "Left" (alternatively, Dasher [23] data entry contactless interface can be applied as well). All the other voice commands are distributed among the rest space being 7% at the maximum. Totally above 64% of the speech commands were given in the multi-modal way (simultaneously with the head pointing) and the rest of the commands were uni-modal speech-only commands.

4 Conclusion

The developed bi-modal user interface integrating automatic speech recognition and computer vision technologies was tested by cognitive experiments both by able-bodied human beings and a person with a severe disability (specifically the user has no hands). Video demonstrations of this assistive system are available in WWW: [24] and [25]. In order to quantitatively evaluate the hands-free pointing interface, we used the methodology of ISO 9241-9, which is based on the Fitts' law experiments. Comparisons of the proposed hands-free interface with the contact-based pointing devices (mouse, touchpad, trackball, touch screen and joystick) were made in terms of the effective index of difficulty, movement time and throughput parameters. The best results were shown by the contact-based touch screen and mouse devices, but the bi-modal interface has outperformed the joystick device.

The obtained performance of hand-free HCI is acceptable, since the developed bi-modal interface is intended mainly for human beings with severe motor-disabilities. ICANDO allows supporting equal participation and socio-economic integration of people with disabilities in the information society and improving their independence from other people. However, it can be helpful for ordinary users too for hands-free human-computer interaction in diverse applications, where hands of a human being are busy, for instance when driving or cooking.

Acknowledgements. This research is partially supported by the Federal Targeted Program "Scientific and Scientific-Pedagogical Personnel of the Innovative Russia in 2009-2013" (contracts No. 2360 and No. 2579), by the Grant of the President of Russian Federation (project No. MK-64898.2010.8), and by the Russian Foundation for Basic Research (project No. 09-07-91220-CT-a).

References

1. Oviatt, S.: Multimodal interfaces. In: Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, pp. 286–304. Lawrence Erlbaum Assoc., Mahwah (2003)
2. Bolt, R.A.: Put-that-there: Voice and gesture at the graphics interface. Computer Graphics 14(3), 262–270 (1980)
3. Malkewitz, R.: Head Pointing and Speech Control as a Hands-Free Interface to Desktop Computing. In: 3rd International ACM Conference on Assistive Technologies ASSETS 1998, Marina del Rey, CA, USA, pp. 182–188. ACM Press, New York (1998)
4. Gorodnichy, D., Roth, G.: Nouse 'Use your nose as a mouse' perceptual vision technology for hands-free games and interfaces. Image and Vision Computing 22(12), 931–942 (2004)
5. Harada, S., Landay, J.A., Malkin, J., Li, X., Bilmes, J.A.: The Vocal Joystick: Evaluation of voice-based cursor control techniques. In: 8th International ACM SIGACCESS Conference on Computers & Accessibility ASSETS 2006, Portland, USA, pp. 197–204. ACM Press, New York (2006)
6. Eiichi, I.: Multi-modal interface with voice and head tracking for multiple home appliances. In: 8th IFIP International Conference on Human-Computer Interaction INTERACT 2001, Tokyo, Japan, pp. 727–728 (2001)
7. Karpov, A., Ronzhin, A.: ICANDO: Low Cost Multimodal Interface for Hand Disabled People. Journal on Multimodal User Interfaces 1(2), 21–29 (2007)
8. Ronzhin, A., Karpov, A.: Russian Voice Interface. Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications 17(2), 321–336 (2007)
9. Rabiner, L., Juang, B.: Speech Recognition. In: Benesty, J., et al. (eds.) Springer Handbook of Speech Processing. Springer, New York (2008)
10. Krim, H., Viberg, M.: Two decades of array signal processing research: the parametric approach. Signal Processing Magazine 13(4), 67–94 (1996)
11. Ronzhin, A., Karpov, A., Kipyatkova, I., Železný, M.: Client and Speech Detection System for Intelligent Infokiosk. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 560–567. Springer, Heidelberg (2010)
12. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE International Conference on Computer Vision and Pattern Recognition Conference CVPR 2001, Kauai, HI, USA (2001)

13. Lienhart, R., Maydt, J.: An Extended Set of Haar-like Features for Rapid Object Detection. In: IEEE International Conference on Image Processing ICIP 2002, Rochester, New York, USA, pp. 900–903 (2002)
14. Lucas, B.D., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In: 7th International Joint Conference on Artificial intelligence IJCAI, Vancouver, Canada, pp. 674–679 (1981)
15. Bouguet, J.-Y.: Pyramidal Implementation of the Lucas-Kanade Feature Tracker Description of the algorithm. Intel Corporation Microprocessor Research Labs, Report, New York, USA (2000)
16. Karpov, A., Ronzhin, A., Cadiou, A.: A Multi-Modal System ICANDO: Intellectual Computer AssistaNt for the Disabled Operators. In: INTERSPEECH International Conference, Pittsburgh, PA, USA, pp. 1998–2001. ISCA Association (2006)
17. ISO 9241-9:2000(E) Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), Part 9: Requirements for Non-Keyboard Input Devices, International Standards Organization (2000)
18. Soukoreff, R.W., MacKenzie, I.S.: Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. International Journal of Human Computer Studies 61(6), 751–789 (2004)
19. Zhang, X., MacKenzie, I.S.: Evaluating Eye Tracking with ISO 9241 - Part 9. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4552, pp. 779–788. Springer, Heidelberg (2007)
20. Carbini, S., Viallet, J.E.: Evaluation of contactless multimodal pointing devices. In: 2nd IASTED International Conference on Human-Computer Interaction, Chamonix, France, pp. 226–231 (2006)
21. De Silva, G.C., Lyons, M.J., Kawato, S., Tetsutani, N.: Human Factors Evaluation of a Vision-Based Facial Gesture Interface. In: International Workshop on Computer Vision and Pattern Recognition for Computer Human Interaction, Madison, USA (2003)
22. Wilson, A., Cutrell, E.: FlowMouse: A computer vision-based pointing and gesture input device. In: Costabile, M.F., Paternó, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 565–578. Springer, Heidelberg (2005)
23. Ward, D., Blackwell, A., MacKay, D.: Dasher: A data entry interface using continuous gestures and language models. In: ACM Symposium on User Interface Software and Technology UIST 2000, pp. 129–137. ACM Press, New York (2000)
24. SPIIRAS Speech and Multimodal Interfaces Web-site, TV demonstration, http://www.spiiras.nw.ru/speech/demo/ort.avi
25. SPIIRAS Speech and Multimodal Interfaces Web-site, demonstration 2, http://www.spiiras.nw.ru/speech/demo/demo_new.avi