A Universal Assistive Technology with Multimodal Input and Multimedia Output Interfaces

Alexey Karpov[1,2] and Andrey Ronzhin[2]

[1] University ITMO, St. Petersburg, Russia
[2] St. Petersburg Institute for Informatics and Automation of RAS (SPIIRAS), Russia
{karpov,ronzhin}@iias.spb.su

Abstract. In this paper, we present a universal assistive technology with multimodal input and multimedia output interfaces. The conceptual model and the software-hardware architecture with levels and components of the universal assistive technology are described. The architecture includes five main interconnected levels: computer hardware, system software, application software of digital signal processing, application software of human-computer interfaces, software of assistive information technologies. The universal assistive technology proposes several multimodal systems and interfaces to the people with disabilities: audio-visual Russian speech recognition system (AVSR), "Talking head" synthesis system (text-to-audiovisual speech), "Signing avatar" synthesis system (sign language visual synthesis), ICANDO multimodal system (hands-free PC control system), and the control system of an assistive smart space.

Keywords: Assistive Technology, Multimodal User Interfaces, Multimedia, Universal Access, Audio-Visual Speech, Assistive Applications.

1 Introduction

A lot of people around the world are limited in their possibilities because of hearing, vision, speech, motion dysfunctions and mental impairments. For the help, support and rehabilitation of these people there are various governmental programmes in many countries like e-Accessibility, e-Inclusion, Ambient Assisted Living (AAL) [1].

In the last years, Russian government also pays more attention to the problems of life of people with disabilities. In May 2012, the President of Russia has ratified the Convention on the Rights of Persons with Disabilities [2], which was accepted several years ago by the General Assembly of the United Nations. The key points of this Convention state that countries must create conditions for maximal integration of disabled people into the social life on all its levels (including education and information society). Also, the State Programme "Accessible Environment" [3] intended for 2011-2015 has recently started in Russia. This program supports adaptation of work of governmental, educational, social organizations and information services (including electronic services) for needs of persons with disabilities, as well as provision of information accessibility and computer means for disabled people, creation and embedding of new means of interaction and development of new goods and services that

C. Stephanidis and M. Antona (Eds.): UAHCI/HCII 2014, Part I, LNCS 8513, pp. 369–378, 2014.

apply special interfaces and devices for various groups of people with special needs. Moreover, on December 30th 2012, The President of Russia has confirmed important changes to the law "On social defense of people with disabilities in the Russian Federation" that has essentially increased the status of Russian sign language, which is an official language of Russia now. According to the Ministry of Health of Russia, in the country there are more than 13 million people with disabilities (almost 10% citizens of the Russian Federation), including above 700 thousand children with disabilities. And each year there are up to one million people getting disability that is caused by a lot of reasons (ecological, medical, psychological, etc.).

Nowadays assistive technologies are known as any technical devices, tools or services that increase, maintain or improve functional capabilities of people with disabilities. It is known that the term "Assistive technology" was firstly used in the USA in the state document "Technology-Related Assistance for Individuals with Disabilities Act of 1988 (The Tech Act)" [4], and now it is widely used. At that assistive technology may have electronic, software, mechanical, optical, biological nature, etc.; they are, for example (not limited to), wheelchairs, prostheses, hearing aids, optical glasses, television subtitles, robot-assistants and telepresence robots, wheelchair lifts, sounds of traffic lights, guide dogs with the appropriate equipment and more other.

Among various assistive technologies we may highlight special information technologies that can assist people in human-computer interaction, information access, electronic learning, communication, etc. Therefore, we define the new term "assistive information technology", which is special software and/or hardware that improves information accessibility and communication means for people with disabilities and special needs.

Usually assistive technologies are adapted to disabilities of concrete users, i.e. specific technologies are used for blind people (for instance, speech and haptic interfaces), deaf (e.g. sign language-based interfaces), dumb people (textual interfaces), motor disabled (e.g. hands-free interfaces) and mentally handicapped persons (e.g. touch-screens with simple graphical user interfaces). However, universal assistive technologies are rarely developed. One example of such technology is a computer system that supports natural communication between blind and deaf persons [5, 6]. Some other examples of multimodal human-computer interfaces for universal access framework are presented in [7-11].

The given paper presents a conceptual model (Section 2) and an architecture of software-hardware complex (Section 3) of a universal assistive technology both with multimodal input and with multimedia output user interfaces. At that a modality is considered as a way (process) of producing some information and a media is a process of receiving some information by a human being during human-computer interaction [12].

2 Conceptual Model of the Universal Assistive Technology

The conceptual model of the universal assistive technology is shown in Figure 1. A computer complex is placed in the center of this model, and it is able to hear and see

users by microphones and video-cameras, as well as it can output multimedia information by a display with touch-screen and loudspeakers. A layer of methods and functions of automatic processing (audio, visual and textual information recognition and synthesis) is placed farther from the center and then a layer of user interfaces, which are used for multimodal human-computer interaction. These user interfaces are based on speech recognition, automatic lip-reading, video-based head tracking, text-to-speech synthesis, machine text processing, manual gestures and speech synthesis, which are combined in four assistive multimodal systems for:

1. Automatic audio-visual speech recognition.
2. Text-to-audiovisual speech synthesis ("Talking head").
3. Sign language visual synthesis ("Signing avatar").
4. Multimodal hands-free computer control.

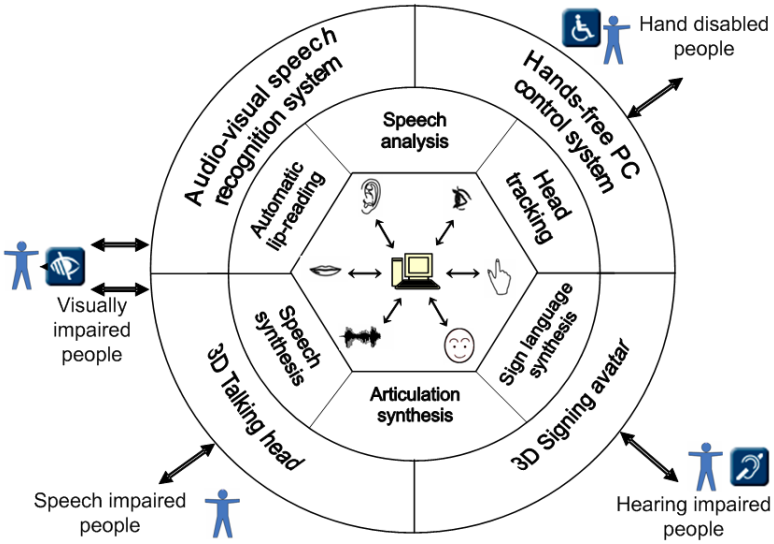

Fig. 1. The conceptual model of the universal assistive technology with multimodal input and multimedia output interfaces

The proposed conceptual model of the universal assistive technology includes only developed or studied by the authors assistive technologies, however, there are some other prospective systems and technologies, not covered in this research, for example, gesture and sign language recognition [13], eye tracking [14], brain-computer interfaces [15], haptic interfaces [16] and so on, which can also be integrated in this framework.

User interfaces for human-computer interaction in assistive technologies must meet some principal requirements of potential users. They should be: universal, multimodal, natural (intuitive), usable (ergonomic), friendly, effective and reliable. Also there

are some basic types of cooperation between modalities: transfer, specialization, equivalence, redundancy, complementarity and concurrency [12].

The proposed assistive information technology is universal one because it is aimed for different categories of users with sensory and physical disabilities: blind people can rely on audio man-machine interface based on speech/sound recognition and synthesis, deaf people focus on text and sign language-based interface, motor disabled people can use multimodal hands-free PC control interface, whereas regular able-bodied users may interact using multimedia (mainly audio-visual) information.

A crucial advantage of multimodal user interfaces is that they provide several alternative ways of human-computer interaction at the same time, and the user may choose how he/she wants (or may) to communicate with information systems. Besides, lacking communicative abilities of a human being can be compensated by some other modalities without any loss of application functionality. At that a set of abilities $P^u = \{p_1^u, ..., p_k^u\}$ of a concrete user $u_i \in U$ (where $U = \{u_1, ..., u_l\}$ is a set of potential users), which are accessible for information input and output, imposes some restrictions on a set of interaction means $S = \{s_1, ..., s_n\}$ of the assistive technology, that determines an optimal interface between the user and the computer system: $I^u = P^u \bigcap S$. At that, interaction means (modalities) can be either input S^I or output S^O for the system (and opposite for the user), i.e. $S = S^I \bigcup S^O$.

3 Software-Hardware Architecture of the Universal Assistive Technology

We have developed and integrated all software modules, components and systems of the model into one software-hardware complex of the universal assistive technology. Figure 2 presents a generalized architecture of this complex; it includes five main interconnected levels (from low to high-level):

1. Level of computer hardware.
2. Level of system software (middleware).
3. Level of application software of digital signal processing.
4. Level of application software of human-computer interfaces.
5. Level of software of assistive information systems.

The low level of computer hardware includes available on the computer market information input sensors and output devices connected to one server: microphones (both stationary ones and portable headset), video cameras (digital camcorders, web-cameras and high speed camera JAI), display (with a touch-screen), and loudspeakers.

The level of system software includes operational system (Microsoft Windows family), drivers of microphones and sound board (provided by manufacturers), drivers of video cameras (provided by manufacturers), computer vision library (OpenCV), computer graphics libraries (OpenGL, DirectX), sound processing libraries (HTK, Julius), and an Internet browser (Microsoft Internet Explorer).

The level of application software of digital signal processing has the modules for voice activity detection, audio and video signals processing/analysis, information fusion, audio and video information synthesis.

The level of application software of human-computer interfaces contains modules for speech and audio events recognition, automatic lip-reading, text-to-speech synthesis, articulation and mimics video synthesis, sign language video synthesis, video-based head tracking, and user fall detection.

The high level of software of assistive information systems includes an audio-visual Russian speech recognition system (AVSR), "Talking head" synthesis system (text-to-audiovisual speech synthesis), "Signing avatar" synthesis system (sign language visual synthesis), ICANDO multimodal system (hands-free PC control system), and a control system of an assistive smart space.

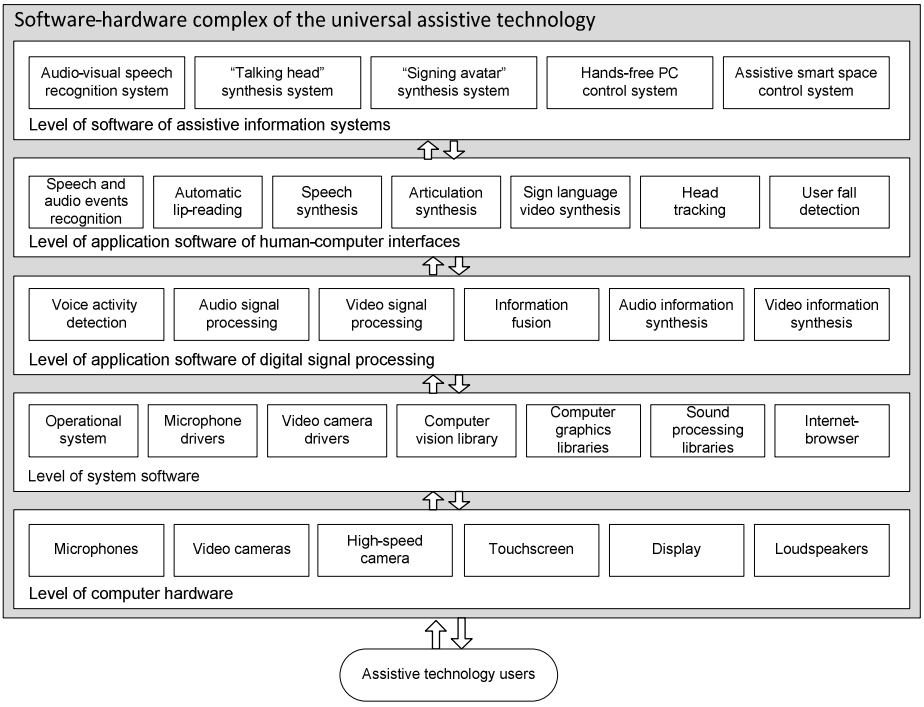

Fig. 2. The architecture of levels of the software-hardware complex of the universal assistive technology

All proposed methods, modules and systems have been implemented as software by C/C++ programming language in Microsoft Visual Studio toolkit using some free available and commercial libraries and software (OpenCV, OpenGL, DirectX, HTK, Julius, MFC, etc.) that works under control of Microsoft Windows operational systems (32/64 bit).

The system for audio-visual Russian speech recognition (AVSR) fuses mathematical models, methods and tools for automatic recognition of auditory speech and reading speech by lips movements [17-19]. The audio-visual system recognition system allows simultaneous processing both audio signal and visual speech (lips articulation) using an information fusion method based on asynchronous Coupled Hidden Markov Models (CHMM) [20] with weights of informativity of these speech modalities depending on audio noises. The recognition approach based on Coupled Hidden Markov Models of the first order allows making information fusion of feature vectors on the level of states of joint probabilistic CHMMs. It provides a possibility to take into account asynchrony (some time lag) between streams of elements of audio speech (phonemes) and visual speech (visemes), which is natural for human's speech production. The bimodal speech recognition system allows increasing accuracy and robustness of automatic speech recognition in noisy environments. It is also aimed for use in speech and multimodal interfaces for human-computer interaction with people having sensory and physical disabilities, including visually impaired and blind people and people with speech disabilities, for example, in the case of whispered speech without vocalization ability, etc.

The computer system for audio-visual Russian speech synthesis (3D "Talking head") [21] integrates virtual 3D models, methods and means for text-to-speech synthesis and video synthesis of lips articulation and facial mimics of the 3D model of human's head. The multimodal system allows processing entered texts and phrases in Russian and generating continuous Russian speech using an original rule-based method for synchronization and fusion of audio and visual modalities of synthesized speech [22]. The proposed method for modalities synchronization allows taking into account natural asynchrony between streams of corresponding visemes of phonemes (visemes always take the lead over phonemes in speech), which is influenced by dynamics of speech production (inertance of human's articulation organs) and co-articulation effects. This method increases both naturalness and intelligibility of generated speech. "Talking head" improves speech perception with respect to audio-only speech synthesizers, and especially in noisy environments. Also it is aimed for creation of human-like embodied conversational agents (ECA) [23] and avatars both for regular users and for persons with disabilities, for example, people with severe speech disabilities may use this system in order to replace own speech, as well as visually impaired people may rely on acoustic modality of synthesized speech for obtaining information from a computer). Multimedia demonstration of this system is available in the Internet [24].

The multimodal system for sign language and speech synthesis (3D "Signing avatar") integrates virtual models, methods and tools for video synthesis of elements of Russian sign language, visual speech (articulation), and audio synthesis of Russian speech [25, 26]. The main components of this system are: a text processor that takes text as an input to generate phoneme and viseme transcriptions, and sequences of HamNoSys (Hamburg Notation System) [27] codes for hand description of manual gestures; text-to-speech module that generates audio speech with time labeling corresponding to the entered text; virtual 3D model of human's head with controlled lips articulation, mimics and facial expressions; control unit for the talking head that

synchronizes and integrates lips movements with synthesized audio signal; virtual 3D model of human's upper body, which is controlled by HamNoSys codes; audio-visual multimodal user interface that synchronizes output audio and visual speech and gesture modalities, integrates all the components for automatic generation of auditory speech, visual speech (articulation and facial expressions) and avatar's gestures of Russian sign language and fingerspelling, as well as outputs multimedia information. "Signing avatar" is aimed for organization of universal human-computer interaction both with regular users, who can perceive multimedia audio-visual-textual information, and with handicapped people, who have severe hearing disabilities or deaf, by generating manual gestures of Russian sign language and fingerspelling, as well as synthesizing visual speech (which is an obligatory component of any sign language), and audio-based verbal communication with visually impaired and blind people. Multimedia demonstration of this system is available in the Internet [28]. There was also some research on automatic recognition of Russian sign language and fingerspelling [29, 30]; however, we have only preliminary results with a prototype of such system.

The multimodal system for hands-free computer control ("ICANDO – Intellectual Computer AssistaNt for Disabled Operators") [31, 32] assembles methods, tools and sub-systems for automatic speech/voice commands recognition in Russian, English and French, video-based user's head tracking in order to interact with the graphical user interface of a PC without use of hands. Instead of use of standard information input devices (such as a keyboard, mouse, touch-screen, touch-pad, etc.) the system proposes to utilize head movements (head gestures) and speech commands. 40 voice commands ("Start", "Escape", "Double click", "Scroll down", etc.) for controlling virtual devices of mouse and keyboard compose system's vocabulary. This system is aimed for organization of multimodal user interface for hands-free human-computer interaction both for regular users (for instance, in edutainment applications, computer games, presentations, in the case when user's hands are busy) and for people with severe hand disabilities (for example, for persons with paralyzed hands or without hands). Demo-version and multimedia demonstration of this system is available on-line [33].

The control system of the assistive smart space [34] combines methods, tools and sub-systems for automatic recognition of speech commands and non-speech audio events (e.g., cry, cough, fall, etc.), which is aimed for analysis and monitoring of audio information in the assistive smart space, and video-based user fall detection, that allows the system to detect involuntary falls of persons inside the assistive smart space, determine extraordinary situations and notify on it. The assistive smart space (assisted living environment) is aimed to help single elderly people and persons with disabilities in independent living. In the case of an extraordinary situation with the user (e.g. at an involuntary fall of the person on the floor, his/her cry or a verbal appeal for help) the control system can detect this and inform a dispatcher service. The scaled-down model of the assisted smart space is equipped with microphone and video-camera arrays, as well as includes developed software modules and tools.

The proposed universal assistive technology is aimed for organization of novel ways of human-computer interaction for support, rehabilitation and education (including electronic learning [35]) of persons with disabilities and special needs, as well as

for improving socio-economical integration of disabled people into the information society and increasing their independence from other persons.

4 Conclusion

We have presented the conceptual model and the architecture of the software-hardware complex of the universal assistive technology. It integrates several multimodal systems: audio-visual Russian speech recognition system, text-to-audiovisual speech synthesis system ("Talking head"), sign language visual synthesis system ("Signing avatar"), multimodal hands-free PC control system (ICANDO), and the control system of the assistive smart space. The proposed universal assistive technology is aimed for organization of novel ways of human-computer interaction for support, rehabilitation and education of individuals with disabilities (visually impaired, deaf people, persons with dysfunctions of hands), and it is useful for regular able-bodied users as well.

Acknowledgements. This research is partially supported by the Russian Foundation for Basic Research (Project № 12-08-01265-a), by the Russian Humanitarian Scientific Foundation (Project № 12-04-12062), and by the Government of Russian Federation (Grant 074-U01).

References

1. Ambient Assisted Living Joint Programme, http://www.aal-europe.eu
2. The Convention on the Rights of Persons with Disabilities of the United Nations, http://www.un.org/disabilities/convention/conventionfull.shtml
3. The Russian State Programme "Accessible Environment", http://zhit-vmeste.ru
4. Tech Act, http://www.ok.gov/abletech/documents/Tech%20Act-Individuals%20with%20Disabilities.pdf
5. Argyropoulos, S., Moustakas, K., Karpov, A., Aran, O., Tzovaras, D., Tsakiris, T., Varni, G., Kwon, B.: A Multimodal Framework for the Communication of the Disabled. Journal on Multimodal User Interfaces 2(2), 105–116 (2008)
6. Hruz, M., Campr, P., Dikici, E., Kindirouglu, A., Krňoul, Z., Ronzhin, A., Sak, H., Schorno, D., Akarun, L., Aran, O., Karpov, A., Saraclar, M., Železný, M.: Automatic Fingersign to Speech Translation System. Journal on Multimodal User Interfaces 4(2), 61–79 (2011)
7. Stephanidis, C., Akoumianakis, D., Sfyrakis, M., Paramythis, A.: Universal accessibility in HCI: Process-oriented design guidelines an tool requirements. In: Proc. 4th ERCIM Workshop on User Interfaces for All, Stockholm, Sweden, pp. 19–21 (1998)
8. Savidis, A., Stephanidis, C.: Unified user interface design: designing universally accessible interfaces. Interacting with Computers 16(2), 243–270 (2004)
9. De Marsico, M., Kimani, S., Mirabella, V., Norman, K.L., Catarci, T.: A Proposal toward the Development of Accessible e-Learning Content by Human Involvement. Universal Access in the Information Society 5(2), 150–169 (2006)
10. Obrenovic, Z., Abascal, J., Starcevic, D.: Universal Accessibility as a Multimodal Design Issue. Communications of the ACM 50(5), 83–88 (2007)

11. Oviatt, S., Cohen, P.: Perceptual user interfaces: multimodal interfaces that process what comes naturally. Communications of the ACM 43(3), 45–53 (2000)
12. Martin, J.-C.: Towards "intelligent" cooperation between modalities. The example of a system enabling multimodal interaction with a map. In: Proc. IJCAI 1997 Workshop on Intelligent Multimodal Systems, Nagoya, Japan (1997)
13. Ong, S., Ranganath, S.: Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(6), 873–891 (2005)
14. Grauman, K.: Communication via Eye Blinks and Eyebrow Raises: Video-Based Human-Computer Interfaces. In: Grauman, K., Betke, M., Lombardi, J., Gips, J., Bradski, G. (eds.) Universal Access in the Information Society, vol. 4, pp. 359–373 (2003)
15. Graimann, B., Allison, B., Pfurtscheller, G.: Brain–Computer Interfaces: A Gentle Introduction. In: Brain-Computer Interfaces. The Frontiers Collection, pp. 1–27. Springer (2010)
16. Colwell, C., Petrie, H., Kornbrot, D., Hardwick, A., Furner, S.: Haptic Virtual Reality for Blind Computer Users. In: Proc. Annual ACM Conference on Assistive Technologies, ASSETS 1998, Marina del Rey, CA, USA, pp. 92–99 (1998)
17. Karpov, A., Ronzhin, A., Markov, K., Zelezny, M.: Viseme-Dependent Weight Optimization for CHMM-Based Audio-Visual Speech Recognition. In: Proc. INTERSPEECH 2010 International Conference, ISCA Association, Makuhari, Japan, pp. 2678–2681 (2010)
18. Karpov, A., Ronzhin, A., Kipyatkova, I., Zelezny, M.: Influence of Phone-viseme Temporal Correlations on Audiovisual STT and TTS Performance. In: Proc. 17th International Congress of Phonetic Sciences, ICPhS 2011, Hong Kong, China, pp. 1030–1033 (2011)
19. Karpov, A., Markov, K., Kipyatkova, I., Vazhenina, D., Ronzhin, A.: Large vocabulary Russian speech recognition using syntactico-statistical language modeling. Speech Communication 56, 213–228 (2014)
20. Nefian, A., Liang, L., Pi, X., Xiaoxiang, X., Mao, C., Murphy, K.: A Coupled HMM for Audio-Visual Speech Recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing, ICASSP 2002, Orlando, USA, pp. 2013–2016 (2002)
21. Karpov, A., Tsirulnik, L., Krňoul, Z., Ronzhin, A., Lobanov, B., Železný, M.: Audio-Visual Speech Asynchrony Modeling in a Talking Head. In: Proc. INTERSPEECH 2009 International Conference, Brighton, UK, pp. 2911–2914 (2009)
22. Karpov, A., Tsirulnik, L., Zelezny, M., Krnoul, Z., Ronzhin, A., Lobanov, B.: Study of Audio-Visual Asynchrony of Russian Speech for Improvement of Talking Head Naturalness. In: Proc. 13th International Conference SPECOM 2009, St. Petersburg, pp. 130–135 (2009)
23. Morales-Rodriguez, M.L., Pavard, B.: Embodied Conversational Agents: A New Kind of Tool for Motor Rehabilitation? In: Proc. 11th Annual International Workshop on Presence, PRESENCE 2008, Padova, Italy, pp. 95–99 (2008)
24. Multimedia demonstration of "Talking head" for audio-visual Russian speech synthesis, `http://www.spiiras.nw.ru/speech/demo/th.avi`
25. Karpov, A., Krnoul, Z., Zelezny, M., Ronzhin, A.: Multimodal Synthesizer for Russian and Czech Sign Languages and Audio-Visual Speech. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2013, Part I. LNCS, vol. 8009, pp. 520–529. Springer, Heidelberg (2013)
26. Karpov, A., Železný, M.: Towards Russian Sign Language Synthesizer: Lexical Level. In: Proc. 5th International Workshop on Representation and Processing of Sign Languages at the LREC 2012, Istanbul, Turkey, pp. 83–86 (2012)

27. Hanke, T.: HamNoSys - Representing sign language data in language resources and language processing contexts. In: Proc. International Conference on Language Resources and Evaluation, LREC 2004, Lisbon, Portugal, pp. 1–6 (2004)
28. Multimedia demonstration of 3D "Signing avatar" for Russian sign language synthesis, http://www.spiiras.nw.ru/speech/demo/signlang.avi
29. Kindiroglu, A., Yalcın, H., Aran, O., Hruz, M., Campr, P., Akarun, L., Karpov, A.: Automatic Recognition of Fingerspelling Gestures in Multiple Languages for a Communication Interface for the Disabled. Pattern Recognition and Image Analysis 22(4), 527–536 (2012)
30. Kindiroglu, A., Yalcın, H., Aran, O., Hruz, M., Campr, P., Akarun, L., Karpov, A.: Multilingual Fingerspelling Recognition in a Handicapped Kiosk. Pattern Recognition and Image Analysis 21(3), 402–406 (2011)
31. Karpov, A., Ronzhin, A., Kipyatkova, I.: An Assistive Bi-modal User Interface Integrating Multi-channel Speech Recognition and Computer Vision. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part II, HCII 2011. LNCS, vol. 6762, pp. 454–463. Springer, Heidelberg (2011)
32. Karpov, A., Ronzhin, A.: ICANDO: Low Cost Multimodal Interface for Hand Disabled People. Journal on Multimodal User Interfaces 1(2), 21–29 (2007)
33. Demonstration of multimodal hands-free PC control system (ICANDO), http://www.spiiras.nw.ru/speech/demo/assistive.html
34. Demiröz, B., Ari, I., Ronzhin, A., Çoban, A., Yalçın, H., Karpov, A., Akarun, L.: Multimodal Assisted Living Environment. Report on research project at eNTERFACE-2011 Summer Workshop on Multimodal Interfaces, Pilsen, Czech Republic (2011), http://www.cmpe.boun.edu.tr/~ari/files/demiroz2011enterface.pdf
35. De Marsico, M., Sterbini, A., Temperini, M.: A Framework to Support Social-Collaborative Personalized e-Learning. In: Kurosu, M. (ed.) HCII/HCI 2013, Part II. LNCS, vol. 8005, pp. 351–360. Springer, Heidelberg (2013)