E-meeting Web-Interface Adaptive to Changing Context and Mobile Devices

Andrey Ronzhin, Viktor Budkov
Speech and Multimodal Interfaces Laboratory
SPIIRAS
St. Petersburg, Russia
e-mail: {ronzhin, budkov}@iias.spb.su

Abstract— **Web-based collaboration using the wireless devices that have multimedia playback capabilities is a viable alternative to traditional face-to-face meetings. To provide quick and effective engagement to the meeting activity, the remote user should be able to perceive whole events in the meeting room and have the same possibilities like participants inside. The proposed logical-time model for compilation of multimedia content of web-based E-meeting system takes into account current situation inside the meeting room, states of audio-, video- and presentation equipment as well as constraints of user mobile devices and provides distant control by equipment and support of distributed meeting participants. The developed web-based application for remote user interaction with equipment of the intelligent meeting room and organization of E-meetings were tested with Nokia mobile phones.**

Keywords- E-meeting; smart space; remote collaboration; mobile communication; multimodal interfaces

I. Introduction

Distributed events organized via specialized web-applications are real alternative to traditional meetings and lectures, where participants interact "face to face." Development of multimedia technologies allows video conference systems not only recording and output of audio and video data, but they also use more sophisticated techniques, such as analysis, pattern recognition and structuring of multimedia data, that certainly enhances services for participants and leads to new ways of access to events in real-time and archive processing [1]. Internet applications for teleconference and distant learning (E-meeting and E-lecture) become more popular in business, research, education and other areas. Such systems allow us to save cost and provide self-paced education and convenient content access and retrieval.

However the main part of the secretary work on documentation and connection of remote participants is performed by a human-operator. Another constraint of E-meeting systems is capacities of communication network and performance of audio and video user devices, which influence on user interface features and sufficiency of remote participant possibilities.

One of the key issues of remote communication is high uncertainty, caused by the spatial and temporal distance between co-participants [2]. The physical and psychological barriers that exist in hybrid meetings make difficult for remote partners to attend a selected conversational flow, which the participants share the same room, and to initiate a new topic of discussion. Thus the main task of the research is to make remote meetings more engaging by giving remote participants more freedom of control in discussions and decision making processes.

Research projects AMI, CHIL, AMIGO, CALO [3,4] were targeted to study various aspects of arrangement of meetings or teleconferencing in smart environments and development of meeting support technology, multi-modal meeting browsers, as well as automatic audio-, and video-based summarization systems. Meeting support includes (semi-) automatic retrieval of information needed for arrangement of remote participation in hybrid meetings, in which one or more participants are remote and others stay in a shared room [5]. Development of technologies for automatic selection of the best camera view, switching on the projector or whiteboard output and selection of microphone of the current speaker is capable to improve audio-visual support for a remote mobile participant. The automatic analysis of audio-visual data of meetings or lectures is complicated by the fact that such events are usually held in large halls with lots of participants, who arbitrarily change positions of their body, head and gaze. Microphone arrays, panoramic cameras, intelligent cameras (PTZ - Pan / Tilt / Zoom) and distributed camera systems are used to improve the capturing and tracking of a group of participants.

In the system Cornell Lecture Browser [6], two video cameras with subsequent synchronization and integration of video streams are used. In the project eClass, videos of lectures were combined with the handwritten notes on a whiteboard. The system AutoAuditorium detects participants by the PTZ-camera, as well as carries out an automatic selection of video from one of three cameras, installed in the meeting room. The spatial sound source localization is used to control of the PTZ-cameras and to track speakers and listeners in the system [7]. Also a list of rules and recommendations for professional operators, assisting to select optimal positions of cameras in the hall, is defined in the paper. The system FlySpec implements the PTZ and omni-directional cameras, management of which is remotely carried out by several participants with the automatic control system adjusting the direction of PTZ-cameras. Application of the panoramic camera allows us to capture the images of all the events taking place in the room and to determinate locations of each participant.

Motion sensors and microphone arrays are used additionally to video monitoring, in order to detect participant positions and the active speaker. Sound source localization by microphone arrays is effective in small lecture or conference rooms only. To record audio signals in a large room, participants and speaker often use personal microphones or apply a system of microphone arrays, distributed in the room [8].

The various approaches have been proposed to record presentation slides projected on the screen during the event [9]. Some systems require loading presentation slides in advance or installation of special software on user's computer, which transmits the current slide to the server.

Parameters of the equipment located inside the intelligent meeting room, which are analyzed at changing the graphical content of the web-page, are considered in Section 2. A model of media stream fusion used for designing actual content of the web-based application is described in Section 3. Experimental results are presented in Section 4. The novelty of the proposed web-based application for E-meeting consists in automatic selection of web-camera of the current active speaker by multichannel speech activity algorithm [10] and designing a user interface adaptive to mobile device features of remote participant.

II. Analysis of the Current Situation Inside the Intelligent Meeting Room

The developed intelligent meeting room is equipped by the projector, the touchscreen TV for smart desk application, several cameras for video monitoring of audience and presentation area, and the personal web-cameras for analysis of behavior of participants sitting at the conference table [11]. Three T-shape microphone arrays mounted on the different walls serve for localization of sound sources, far-field recording and further speech processing. The personal web cameras mounted on the conference table have internal microphones and are used for record of video and speech of each meeting participant. A combination of the desktop web-cameras and microphone arrays provides both spatial localization of sound sources and record of participants' speech with high quality. The multimedia content compiled from audio and video signals captured by the referred devices is used in the web-based application for organization of hybrid E-meetings.

Table 2 contains parameters of objects (equipment, software, participants) located inside the intelligent meeting room, which is used for description of the current situation in the meeting room and taken into account during compilation of the graphical content of the web-page. Behavior of participants at the conference table, as well as the main speaker located in the presentation area is analyzed by developed technologies of sound source localization, video tracking of moving objects, detection and tracking of human face.

TABLE I. The Parameters Representing the Current Situation in the Meeting Room

Object in the meeting room	Parameters		
	Notation	*Values*	*Description*
Projector	p_1	0	Projector turned-off
		1	Projector turned-on
	p_2	0	Presentation is not started
		1	Presentation is started
	p_3	0	Current slide of a presentation is shown longer, than τ_{slide} , ($t_{cur} - t_{slide} > \tau_{slide}$, where t_{cur} is current time).
		1	Slide of a presentation was changed (time of the changing t_{slide} is saved)
Smart desk	d_1	0	Wide touchscreen TV is turned-off
		1	Wide touchscreen TV is turned-on
	d_2	0	Smart desk application is not loaded
		1	Smart desk application is loaded
	d_3	0	Touch input was not used longer than τ_{desk} , ($t_{cur} - t_{desk} > \tau_{desk}$).
		1	Touch input was used (beginning time of touchscreen using t_{desk} is saved)
Main speaker (presenter)	s_1	0	A speaker in the presentation area is not observed by the video monitoring system
		1	Speaker is found in the presentation area
	s_2	0	Speaker face is not found
		1	Speaker head is directed to (the face tracking system founded presenter face)
	s_3	0	Speech activity in the presentation area is not detected
		1	A presenter gives a speech (the sound source localization system detected an activity in the presentation area)
Personal web-cameras assigned with participants sitting at the conference table	c_1	0	Currently there are no speakers at the conference table
		1	Currently a participant at the conference table gives a speech comment
	c_{2i}	0	Personal web-camera i is turned-off
		1	Personal web-camera i is turned-on
	c_{3i}	0	No participant in front of the web-camera i
		1	There is a participant in front of the web-camera i (the video monitoring system estimates degree of changing image background recorded before the meeting)
	c_{4i}	0	No speech activity of the participant sitting in front of the web-camera i
		1	A participant sitting in front of the web-camera i gives a speech comment (the multichannel speech activity detection system determines useful signal in the audio channel of the web-camera i)
	c_{5i}	0	Face of the participant sitting in front of the web-camera i is not found
		1	Face position of the participant sitting in front of the web-camera i is detected (the face detection system found a participant face)

Values of the parameters of hardware and software are determined by a query of its states via TCP/IP protocol or by means of Object Linking and Embedding Automation.

III. E-MEETING WEB-INTERFACE DEVELOPMENT

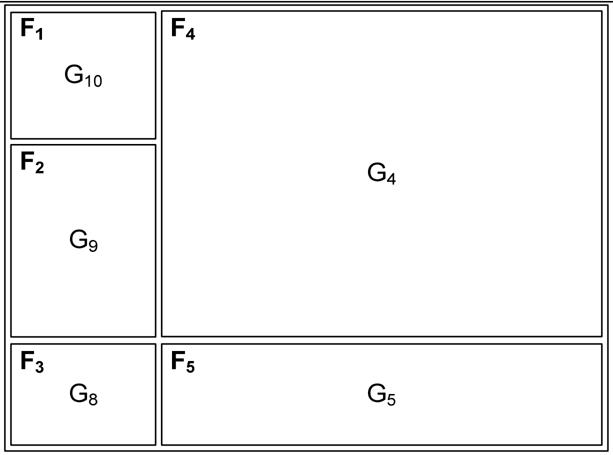
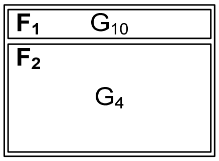
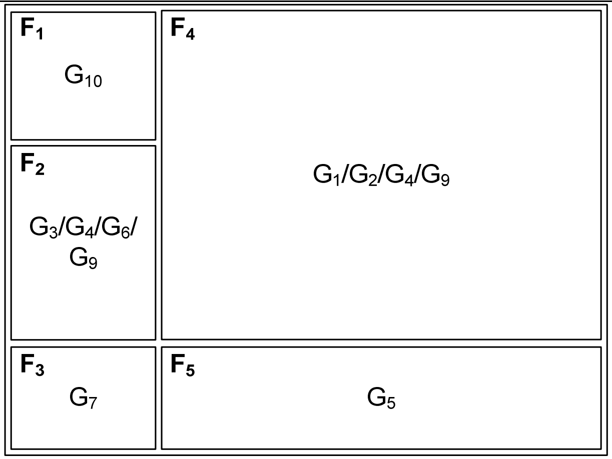
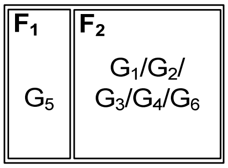
Graphical interface of the web-page, on which remote participants could observe a meeting organized inside the intelligent meeting room, contains several basic forms:

$$F = \{F_1, F_2, ...F_{N_F}\},$$

where N_F is a number of the forms depending on current meeting state and features of browser used in a client device. Two main states in meeting process were selected and corresponding notations of forms were used for the current version of web-page software: (1) registration (preparations before meeting), forms F^{reg}; (2) presentations (main part of meeting), forms $F^{meeting}$. Further the number of the meeting states will be increased taking into account peculiarities of participant behavior and necessity of use of specific technical equipment during the discussion, voting and other formal stages.

Another important factor effecting on the web-page content is a display resolution and correspondingly maximal size of browser window used for remote view of the meeting. So two classes of devices, which have especially different sizes of screen, and corresponding notations of the forms were selected: (1) personal computer, forms $F(PC)$, (2) mobile device, forms $F(MD)$. Table 2 shows the basic variants of the web-page layout depending on the current state of meeting and type of user device.

TABLE II. THE LAYOUT VARIANTS OF THE WEB-PAGE FOR E-MEETING

Meeting state	**Screen of client device**	
	Personal computer ($F(PC)$)	***Mobile device($F(MD)$)***
Registration (F^{reg})		
Presentations ($F^{meeting}$)	 F1 G10 F2 G3/G4/G6/ G9 F3 G7 F4 G1/G2/G4/G9 F5 G5	

Symbol “/” designates that several variants of graphical content are possible in the form. For instance, the current image on the projector or the current image on the smart desk will be represented in fourth form during presentations on the web-page browsed by a personal computer. Content of the forms could be changed during meeting, but it always includes a graphical component from a set:

$$G = \{G_1, G_2, ...G_{N_G}\},$$

where N_G is a number of used components (in the current version $N_G = 10$): G_1 is a component representing the current image on the projector; G_2 is a component representing the current image on the smart desk; G_3 is a component representing the current image captured by the video camera directed to main speaker; G_4 is a component representing the current image captured by the video camera directed to audience; G_5 is a component representing a assemble of the current images captured by personal web-cameras directed on participants sitting at the conference table in the meeting room; G_6 is a component representing the current image captured by a web-camera assigned with a participant, which currently gives a speech comment; G_7 is a component representing an indicator of speech duration; G_8 is a component representing a clock with time labels of the current meeting; G_9 is a component representing a logo of the current meeting; G_{10} is a component representing main data about the current meeting.

The enumerated components are connected with corresponding source, which transmits own graphical data (the projector – a presentation slide; the smart desk – window with handwriting sketches; the video and web-cameras – frames with an image; the software services – time indicators, logo and other data about meeting). Receiving new data on a source leads to updating content of corresponding form in the web-page.

Let us consider a process of graphical content compilation in the forms. Each graphical form F_j on the web-page is described by the following tuple:

$$F_j = \left\langle l_j, u_j, w_j, h_j, g_j \right\rangle,$$

where l_j is upper left corner position of the form at the abscissas axis, u_j is upper left corner position of the form at the ordinates axis, w_j is a form width, h_j is a form height, g_j is a graphical content of the form, which is actual and was chosen from the set G . Sizes of the forms could be changed depending on the current features of browser used in client device.

In the forms $F_2^{meeting}(PC)$, $F_4^{meeting}(PC)$, $F_2^{meeting}(MD)$ the number of the graphical components is changed depending on the parameter values mentioned in Table 1. In other forms the graphical component numbers are kept during the whole meeting. Selection of the current graphical component $g \in G$ for the referred forms is realized by a logical-temporal model of compiling the graphical interface web-page. The following set of logical rules is an essence of the model:

$$g_2^{meeting}(PC) = \begin{cases} G_3, s_1 \wedge s_2 \wedge s_3 \wedge \neg c_1, \\ G_4, \neg s_2 \wedge \neg c_1 \wedge ((p_1 \wedge p_2) \vee (d_1 \wedge d_2)),, \\ G_6, \neg s_3 \wedge c_1, \\ G_9, \;\; overwise. \end{cases}$$

$$g_4^{meeting}(PC) = \begin{cases} G_1, (p_1 \wedge p_2 \wedge (\neg d_1 \vee \neg d_2)) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge \neg d_3) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge d_3 \wedge (t_{slide} > t_{desk})), \\ G_2, ((\neg p_1 \vee \neg p_2) \wedge d_1 \wedge d_2) \vee \\ \quad (p_1 \wedge p_2 \wedge \neg p_3 \wedge d_1 \wedge d_2 \wedge d_3) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge d_3 \wedge (t_{slide} < t_{desk})), \\ G_9, (\neg p_1 \vee \neg p_2) \wedge (\neg d_1 \vee \neg d_2), \\ G_4, \;\; overwise. \end{cases},$$

$$g_2^{meeting}(MD) = \begin{cases} G_1, ((p_1 \wedge p_2 \wedge (\neg d_1 \vee \neg d_2)) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge \neg d_3) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge d_3 \wedge \\ \quad (t_{slide} > t_{desk}))) \wedge \neg s_2 \wedge \neg c_1, \\ G_2, (((\neg p_1 \vee \neg p_2) \wedge d_1 \wedge d_2) \vee \\ \quad (p_1 \wedge p_2 \wedge \neg p_3 \wedge d_1 \wedge d_2 \wedge d_3) \vee \\ \quad (p_1 \wedge p_2 \wedge p_3 \wedge d_1 \wedge d_2 \wedge d_3 \wedge \\ \quad (t_{slide} < t_{desk}))) \wedge \neg s_2 \wedge \neg c_1, \\ G_3, \neg p_3 \vee \neg d_3 \wedge s_1 \wedge s_2 \wedge s_3 \wedge \neg c_1, \\ G_4, \neg p_3 \vee \neg d_3 \wedge \neg s_2 \wedge \neg c_1, \\ G_6, \neg p_3 \vee \neg d_3 \wedge \neg s_3 \wedge c_1, \\ current\ component\ is\ saved, overwise. \end{cases}.$$

Verification of the model was completed manually and during the experiments. The next version of the model will take into account behavior of participants sitting at the left side of the meeting room. Increase of participant number and zones of the meeting room, which should be analyzed, will required definition of some priorities in order to select image of the current speaker.

The component G_5, which consists of actual images of participants sitting at the conference table, is compiled by an analysis of states of personal web-cameras and presence of participants and faces on frame. Let us identify a set of images from the web-cameras as:

$$W = \{W_1, W_2, \ldots W_{N_W}\},$$

where N_W is a number of the web-cameras mounted on the conference table (in the developed meeting room $N_W = 10$). Then the component G_5 consists of the images captured by turned-on web-cameras, in which a participant is detected:

$$G_5 = \bigcup_{i=1}^{N_W} (W_i | c_{2i} \wedge c_{3i} = 1) \cdot$$

ISBN: 978-1-61208-129-8

Taking into account the limited sizes of the forms used for representing the component G_5, the number of displayed participants is reduced by an analysis of his speech activity c_{4i} and/or presence of his face in the frame c_{5i}. Particularly, the form $F_1^{meeting}(MD)$ for mobile device contains up to three participant images, so both parameters are used for selection of more active participants:

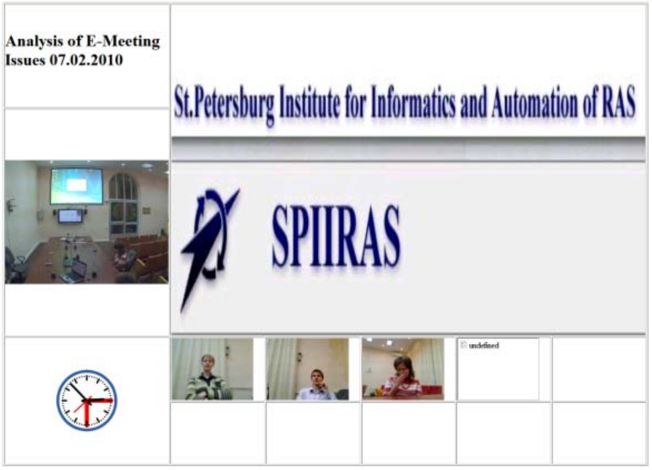
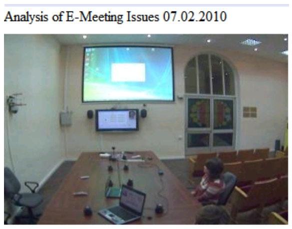
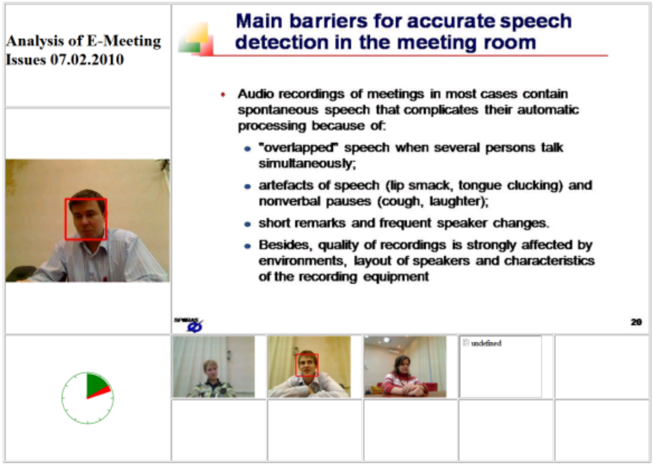
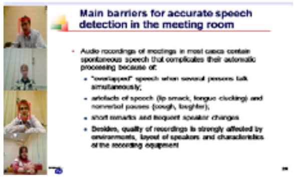
$$F_1^{meeting}(MD)=\bigcup_{i=1}^{N_W}(W_i|c_{2i}\wedge c_{3i}\wedge c_{4i}\wedge c_{5i}=1)\ .$$

The proposed logical-temporal model of compilation of web-page graphical interface was tested on a personal computer and several models of Nokia smartphones. The fourteen different browser resolutions were applied for the tested devices, since the window size, which could be used for web page view, is significantly varied owing to different screen sizes and browser options [10].

IV. EXPERIMENTS

Experimental results were received in a natural scenario where several people discuss a problem in the meeting room about forty minutes. One of the participants stayed in the presentation area and could use the smart desk and the multimedia projector. Other participants were located at the conference table. The main speaker started a talk, when all the participants bring together in the meeting room. Every participant could ask a question after finish of the presentation. Table 3 presents some examples of web-page content generated during the meeting for view on smartphone Nokia N95 with browser resolution 314x200 pixels and a monitor of personal computer with browser resolution 1280x768 pixels. Placement of elements in the web-page is specified using the CSS style tables. The resolution and orientation of screen are checked every 500 ms using Java Script. At the changing screen parameters the corresponding layout of the page is automatically selected and generated for a remote client.

TABLE III. CONTENT EXAMPLES OF THE WEB-BASED APPLICATION FOR MEETING

Meeting state	**Screen of client device**	
	Personal computer	***Mobile device***
Registration	 Analysis of E-Meeting Issues 07.02.2010 St.Petersburg Institute for Informatics and Automation of RAS SPIIRAS undefined	 Analysis of E-Meeting Issues 07.02.2010
Presentations	 Analysis of E-Meeting Issues 07.02.2010 Main barriers for accurate speech detection in the meeting room • Audio recordings of meetings in most cases contain spontaneous speech that complicates their automatic processing because of: • "overlapped" speech when several persons talk simultaneously; • artefacts of speech (lip smack, tongue clucking) and nonverbal pauses (cough, laughter); • short remarks and frequent speaker changes. • Besides, quality of recordings is strongly affected by environments, layout of speakers and characteristics of the recording equipment 20 undefined	 Main barriers for accurate speech detection in the meeting room • Audio recordings of meetings in most cases contain spontaneous speech that complicates their automatic processing because of: • "overlapped" speech when several persons talk simultaneously; • artefacts of speech (lip smack, tongue clucking) and nonverbal pauses (cough, laughter); • short remarks and frequent speaker changes • Besides, quality of recordings is strongly affected by environments, layout of speakers and characteristics of the recording equipment

The sound source localization system and the multichannel speech activity detection system were used for selection of source of audio stream transmitted to remote participant [10]. Speech of a presenter was recorded by the microphone, which is located over the presentation area. The built-in microphone of the web-camera assigned with the presently active participant sitting at the conference table was used for recording his/her speech.

The statistics of base events, which effect to changing situation in the meeting room and selection graphical components is presented in Table 4. During registration stage the all graphical components have own layout, so incorrect web-page content could be compiled during presentation only. Changing states of the smart desk and the projector was correctly detected. Most part of the errors arose at detection of speech activity of the participants sitting at the conference table. This type of errors leads to selection of the camera, which captures another participant, as result the image of the active participant is not displayed on the web-page. However, his speech captured by currently selected camera is transmitted with some attenuation of the sound signal.

TABLE IV. LIST OF EVENTS OCCURRED DURING THE MEETING

Event description	**Number of the event**	
	Determined manually	**Determined automatically**
Number of participants	**7**	**7**
Number of participants sitting at the conference table	6	6
Number of slide changes	14	14
Number of smart desk usage	5	5
Number of speech activity of main speaker	34	37
Number of changes of speakers sitting at the conference table	13	16
Number of temporal inactivity in audience	1	4

Also the miss of speech activity of main speaker led to selection of camera with view to the audience or other participant, whose speech was incorrectly detected. At whole about eighty percent of the graphical components were correctly selected during the analysis of the current situation in the meeting room. At this moment, the developed web-page layout model was tested at the task of support of passive remote participants. To enhance his potentials a toolbar allowing a participant located outside the meeting room to give a question to the presenter and to share the current discussion will be developed.

V. CONCLUSION

The developed intelligent meeting room is a distributed system with the network of software modules, actuator devices, multimedia equipment and audio-visual sensors. Awareness of the room about spatial position of the participants, their activities, role in the current event, their preferences helps to predict the intentions and needs of participants. Context modeling, context reasoning, knowledge sharing are stayed the most important challenges of the ambient intelligent design.

Assignment of easy-to-use and well-timed services, at that stay invisible for user, is one of another important feature of ambient intelligent. In the developed intelligent room all the computational resources are located in the adjacent premises, so the participants could observe only microphones, video cameras, as well as equipment for output of visual and audio information. Implementation of multimodal user interface capable to perceive speech, movements, poses and gestures of participants in order to determinate their needs provides the natural and intuitively understandable way of interaction with the intelligent room.

The proposed logical-temporal model of compilation of web-page graphical interface allows remote participants to perceive whole events in the meeting room via a personal computer or smartphones. The developed web-based application for remote user interaction with equipment of the intelligent meeting room and organization of E-meetings were successfully tested with Nokia mobile phones. Further efforts will be focused on enhancement of capabilities of remote participants during events conducted inside the intelligent meeting room.

ACKNOWLEDGMENT

This work is supported by The Russian Program “Research and Research-Human Resources for Innovating Russia in 2009-2013” (contract П2360), RFBR (projects 10-08-00199-a) and Nokia.

REFERENCES

[1] B. Erol and Y. Li. An overview of technologies for e-meeting and e-lecture. In: IEEE International Conference on Multimedia and Expo, 2005, pp. 6–12.

[2] N. Yankelovich, J. Kaplan, N. Simpson, and J. Provino. Porta-person: telepresence for the connected meeting room. In: Proceedings of CHI 2007, 2007, pp. 2789–2794.

[3] Computers in the human interaction loop. Ed. A. Waibel and R. Stiefelhagen, Springer, Berlin 2009. 374 p.

[4] R. Rienks, A. Nijholt, and P. Barthelmess. Pro-active meeting assistants: attention please! // AI & Society Vol. 23(2), Springer, 2009. pp. 213–231.

[5] R. Op den Akker, D. Hofs, H. Hondorp, H. Akker, J. Zwiers, and A. Nijholt. Supporting Engagement and Floor Control in Hybrid Meetings. Springer, LNAI 5641, 2009, pp. 276-290.

[6] S. Mukhopadhyay and B. Smith. Passive capture and structuring of lectures. ACM Multimedia, 1999, pp. 477–487.

[7] Y. Rui, A. Gupta, J. Grudin, and L. He. Automating lecture capture and broadcast: Technology and videography // Multimedia Systems, Vol. 10, 2004. pp. 3–15.

[8] M. Wienecke, G. Fink, and G. Sagerer. Towards automatic video-based whiteboard reading // Proc. ICDAR’2003, 2003. p. 87–91.

[9] A. Ronzhin and V. Budkov. Multimodal Interaction with Intelligent Meeting Room Facilities from Inside and Outside // Springer-Verlag Berlin Heidelberg, S. Balandin et al. (Eds.): NEW2AN/ruSMART 2009, LNCS 5764, 2009. pp. 77–88.

[10] A. Ronzhin, V. Budkov, and A. Karpov. Multichannel System of Audio-Visual Support of Remote Mobile Participant at E-Meeting // Springer-Verlag Berlin Heidelberg, S. Balandin et al. (Eds.): NEW2AN/ruSMART 2010, LNCS 6294, 2010, pp. 62–71.

[11] R. Yusupov and A. Ronzhin. From smart devices to smart space // Herald of the Russian Academy of Sciences, MAIK Nauka, Volume 80, Number 1, pp. 63-68.