System of Audio-Visual Streams Recording and Synchronization for the Smart Meeting Room

Ronzhin Al.L., Karpov A.A.
Speech and Multimodal Interfaces Laboratory
Institution of the Russian Academy of Sciences St.Petersburg Institute for Informatics and Automation of RAS (SPIIRAS),
St. Petersburg, Russia
{ronzhinal, karpov}@iias.spb.su

Abstract

The problem of automatic detection and recording of active speaker talks among more than thirty participants located in the medium meeting room is not solved completely. In the developed smart meeting room techniques of video tracking and audio source localization are implemented for recording AVI files of speaker messages. Video processing of streams from five cameras serves for registration participants in fixed chair positions, tracking main speaker and recording view to the audience. The experiments showed that the developed audiovisual system captured messages of all the speakers. The detection error of beginning and ending of speech message had acceptable rate.

Keywords: *Video surveillance, Computer vision, Speaker detection, Smart meeting room.*

1. INTRODUCTION

Choosing current active speaker and recording his/her activity during an event are the main tasks for meeting recording and supporting teleconference systems [1, 2]. Panoramic and personal cameras could be employed for simultaneous recording of all participants. Such approach is suitable for small events, where all the participants are located at one table. Increase in participant number leads to space extension, which should to be processed, as well as cost of recording technical equipment.

Several approaches of information presenting such as oral statement, presentation, whiteboard notes, demonstration of audiovisual data may be used for support educational events such as teleconference, lecture, workshop, meeting, which are carried out in rooms with state of art multimedia equipment. General lecture scenario implies that students have to write most fully information of lecture talk. However, students usually may write only short notes and main words. So in order to provide participants with meetings materials the audiovisual system for meetings recording and processing was developed [3]. The first prototype of such system was developed in the Cornell University [4], which consists of two cameras for lecture talk and presentation slides recording. Another system is FLYSPEC [5], which was developed at 2002 year by FX Palo Alto Labs and its intended for supporting teleconference. Two video sensors were implemented in this system: high resolution camera and Pan/Tilt/Zoom (PTZ) camera. The system may automatically control second camera or by analysis of participants requests.

Automatic analysis of audio and video data recorded during a meeting is not trivial task, since it is necessary to track a lot of participants, which randomly change position of their body, head and gaze. In order to detect participant activity several approaches based on using panoramic cameras, intelligent PTZ cameras, distributed camera systems were employed [5, 6]. Besides video monitoring, motion sensors and microphone arrays could be implemented for detecting participant's location and selection of the current speaker [7]. The sound source localization technique is effective for small lections or conference rooms. Personal microphones for all the participants or system of microphone arrays, which set on several walls of smart room, are employed for audio re-cording in medium rooms [8, 9].

Description of the technological framework of the smart meeting room is presented in Section 2. An algorithm describing the interaction of sound source localization and video tracking modules during speaker detection, recording audio video files and synchronization all the data streams is presented in Section 3. The results of the experiment are discussed in Section 4.

2. TECHNOLOGICAL FRAMEWORK OF THE SMART MEETING ROOM

A premises of 72 square meters located inside the institute building was supplied for intelligent meeting room in 2008 at the financial support of the Russian Foundation for Basic Research. Monitoring of the room is performed by 15 video cameras mounted on the walls, ceiling and tables and provides tracking of moving objects, face detection and other functions. Three T-shape 4 channel microphone arrays mounted on the different walls serve for localization of sound sources, far-field recording and following speech processing. Besides video recording the personal web cameras mounted on the tables have internal microphones and are used for recording speech of each meeting participant.

A wide touchscreen plasma panel and multimedia projector (projection screen) are located one under another in the left side of the room and provide output of multimodal information. Operated electro gears are connected to the projection screen and curtain rail. The curtains are made from special light-tight cloth in order to suppress the outside influence on the illumination changing in the room. The processing of recorded audio-visual data, control the multimedia and electro mechanic equipment are performed by six four-cored computers, two multichannel audio boards Presonus FirePod, as well as some devices providing cable and wireless network. The referred service equipment is mounted in a rack and located on the adjacent room from the meeting one. Thus, users inside the meeting room could see only appliances for input/output information, but other devices and computational resources are invisible. To provide service and the same time be hidden for a user is one of the main features of ambient intelligence.

The developed smart room is intended for holding small and medium events with up to forty-two participants. Also there is the

ability to support of distributed events with connection of remote participants. Two complexes of devices are used for tracking participants and recording speakers: (1) personal web-cameras serve for observation of participants, which are located at the conference table; (2) three microphone arrays with T-shape configuration and five video cameras of three types are used for audio localization and video capturing of other participants, which sit in rows of chairs in other part of the room. Description of the first complex could be found in [10]. Status of multimedia devices and participant activity are analyzed for whole mapping current situation in the room.

Location and description of places of seats, multimedia equipment (TV set, Projector), five AXIS Internet-cameras (two PTZ-cameras, two wireless cameras, camera with wide angle lens, installed on the ceiling in the center of the room),10 personal webcams Logitech AF Sphere (on the conference table) could be found in [11]. Three microphone arrays located in the center of the left and right walls and over the sensor plasma (touch screen) serve for sound sources localization and recording phrases of participants. Each array of microphones has T-shaped configuration [12]. The applied software for processing multi-channel audio streams was firstly used at development of multimodal information kiosk [13]. The conference table with installed personal web-cameras for placement of participants of small meetings (round tables) up to 10 people is located on the left side of the hall. The right side of the hall contains rows of seats, which can accommodate up to 32 participants, tracking of which is implemented by the distributed system of cameras and microphone arrays.

3. ALGORITHM OF SPEAKER RECORDING

The algorithm of camera pointing to the current active speaker in the zone of chairs and following recording of his/her speech should be considered in detail. Sound source localization and object tracking by ceiling camera are implemented here. Both modules work together during the all time of the event in the smart room.

The object detection module carries out the search and following tracking of the participants inside in the room. Also the module marks the occupied chairs [14, 15], which will be used as hypotheses for speaker position. The scheme of the algorithm of the speaker detection and recording is shown in Figure 1.

The appearance of a sound signal in the chair zone launches the voice activity detection process and makes query (the event E_1) to the object detection module in order to check the presence of a participant in the chair, which is closest to the determined coordinates of the sound source. If the chair is marked as occupied then corresponding response is transmitted to the module of speech recording as well as the camera serving this zone is being pointed to the selected chair with the current active participant.

To avoid missing of the speech, the decision about useful sound segment is made every 20 ms. Short remarks and noise with duration less half of second are discarded in order to exclude false speaker detection. The frame rate of the camera, which captures the speaker in the chair zone, achieves thirty frames per second. However, the camera pointing takes up to couple seconds owing to a stabilisation period after mechanical setting of direction angles and zooming of the objective lens. So the recording of images to the *bmp* files is started after the camera pointing is accomplished.

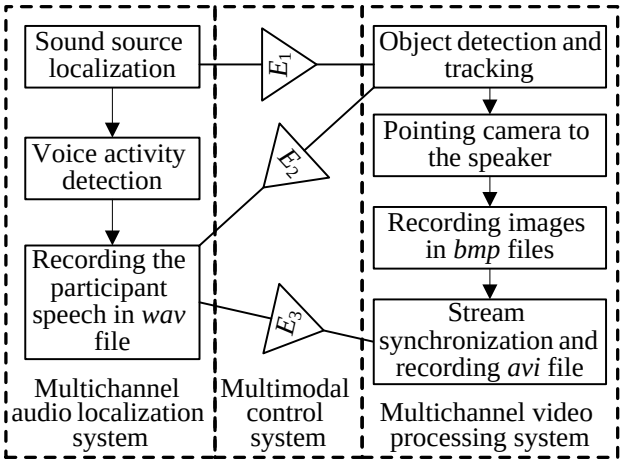

Figure 1: Algorithm of speaker detection and recording.

At the same time, in the multichannel audio localization system the *wav* file with participant speech is recorded in case of speech boundaries detection and positive response from the object detection module (the event E_2) about presence of a participant in the selected chair. After recording *wav* file the corresponding notification (the event E_3) with path to the file transmitted to the video processing system. Then the system goes to the sound source detection and localization stage again.

Participant could make some pauses during the talk that leads to the detection ending boundary of the phrase and recording the separate wav file. As a result during the talks the system could write several audio files, which belongs to the same participant (more precisely put, belongs to chair coordinates assigned to this speaker). Name of audio file includes information about chair number, from which speech signal was recorded.

The creation of *avi* file is started after silence of the current speaker during five seconds or, that more frequent case, detection of an active speaker on other chair, conference table or in the presentation area. The main difficulty of recording *avi* file consists in synchronization of the sets of audio and image files. Frame rate of the camera is not constant owing to various download of the computer, constraints of network and camera hardware. So, the synchronization process is based on analysis of duration and creation time of the *wav* files. Figure 2 shows scheme of synchronization algorithm. All audio files are processed in consecutive order. At first, system detects time interval, in which audio and video files were recorded.

Then to get normal FPS (25 frame per second) starts image duplication in determinate time intervals. Edition of *avi* file is carried out during processing of packets of *bmp* files recorded in interval time approximately equal one second. A data packet structure consists of its duration, first frame number and frames total amount in packet. An analysis of current structure need for elimination of asynchrony appears at recording of audio and video streams, because it allows calculating additional frames total amount. After processing of all *bmp* and *wav* files selected and duplicated images add to an *avi* file, then *wav* files add to it.

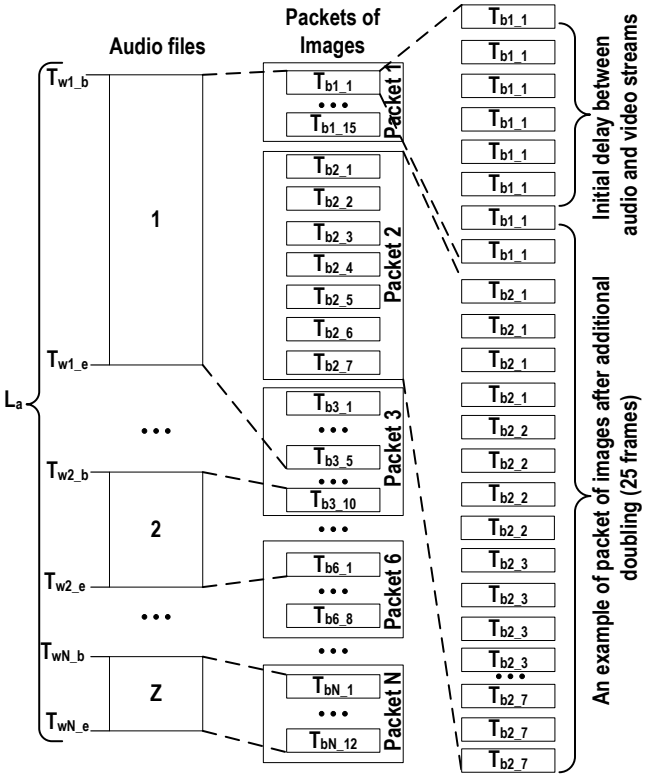

Figure 2: An example of audio and video streams synchronization for recording *avi* file

The described algorithms serve for recording remarks of participants sitting thirty-two chairs of the right side of the smart room. At the end of the meeting the set of *avi* files with all the remarks are recorded. Analogical algorithm is used for tracking main speaker in the presentation area. The description of the approach, which is used to capture activities of participants sitting at the conference table, as well as the logical-temporal model for compilation multimedia content for remote participants of the meeting and support teleconference, is presented in [10].

4. EXPERIMENTS

For an estimation algorithm of detecting and recording active participant speech four criteria were used.

(1) Initial delay between audio and video streams L_{b_d} calculates as difference between first *wav* file creation time T_{w1_b} and *bmp* file T_{b1_1} creation time, which corresponded with a T_{w1_b} time: $L_{b\_d} = | T_{w1\_b} - T_{b1\_1} |$;

(2) A length of recorded *avi* file L_a calculates as summing up of *wav* files length for current speech: $L_a = \sum_{i=1}^{N} T_{wi\_e} - T_{wi\_b}$

(3) Duplicate frames total amount calculates as summing up of L_{b_d} and all duplicated frames in all image packets $P_i : N_{f\_d} = L_{b\_d} + \sum_{i=1}^{N} P_i$; $P_i = P_{AF\_i} + P_{RF\_i}$;

$$P_{AF\_i} = \frac{(P_{FN\_i} - P_{F\_i})}{P_{F\_i}}; \qquad P_{RF\_i} = (P_{FN\_i} - P_{F\_i}) \% P_{F\_i};$$

$$P_{FN\_i} = F_D * \frac{T_{bN\_i} - T_{b1\_i}}{1000};$$

(4) A mean FPS F_a in a video buffer calculates as summing up of image packets size divide on the packets total amount: $F_a = \sum_{i=1}^{i<N} F_i$.

The estimation of the algorithm of detecting and recording active participant speech was carry out in the SPIIRAS smart room. Main attention was paid on detecting active participants in the zone of chairs. Each tester performed the following scenario: (1) take a sit in the room; (2) wait visual confirmation on a smart board about registration of participant in the chair; (3) pronounce the digit sequence from one to ten; (4) move to another chair.

During the experiments were recorded 36 avi files in a discussion work mode. After manual checking was detected that 89% are files with speaker's speech and 11% false files with noises. Such noises are carrying out in process of tester standing up from a chair, because in such moment chair's mechanical details carry out high noise. Also mistakes in detecting sitting participants influence on appearance of false files. Table 1 shows results of estimation files with speaker's speech.

L_{b_d}, ms			L_a, ms			N_{f_d}, frames		
Min	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**
80	2440	724	5312	6432	5608	32	104	59

Table 1: The estimation of algorithm of detecting and recording active participant speech work

A result of experiments shows, that *avi* file in mean consists of 137 frames, 59 of it are duplicated frames, as well as has length 5 seconds. Calculated mean FPS in video buffer is 24 frames per second, this is due to the fact that rounding of values at calculating a required total amount of additional frames in image packets. The total amount of duplicated frames includes initial delay between audio and video streams. Also such total amount of duplicated frames is carry out with changing camera FPS as a result of noises in a network devices as well as limited writing speed of storage devices. Analyses of received data shows that *avi* files form by system include all speeches and a small percent of false records.

5. CONCLUSION

The audiovisual monitoring system of participants was developed for automation of recording events in the smart room. It consists of the four main modules, which realize multichannel audio and video signal processing for participants localization, detection of speakers and recording them. The proposed system allows us to automate control of audio and video hardware as well as other devices installed in the smart room by distant speech recognition of participant command. The verification of the system was accomplished on the functional level and also the estimations of

detection quality of participants, and camera pointing on speaker and speaker detection error were calculated.

6. ACKNOWLEDGMENT

This work is supported by Russian Foundation for Basic Research (projects 10-08-00199-a 11-08-01016-a).

7. REFERENCES

[1] Busso C., Hernanz S., Chi-Wei Chu, Soon-il Kwon, Sung Lee, Georgiou P.G., Cohen I., and Narayanan S. *Proc. IEEE International Conference on Multimedia and Expo: Smart room: Participant and speaker localization and identification. Philadelphia, USA, March 18-23. 2005, pp. 1117–1120.*

[2] Zhang C., Yin P., Rui Y., Cutler R., Viola P., Sun X., Pinto N., and Zhang Z. *IEEE Transactions on Multimedia: Boosting-Based Multimodal Speaker Detection for Distributed Meeting Videos. Vol.10, No.8, 2008, pp. 1541-1552.*

[3] Lampi F. *Automatic Lecture Recording. Dissertation. The University of Mannheim, Germany. 2010.*

[4] Mukhopadhyay, S., Smith, B. *Passive capture and Structuring of Lectures, Proceedings of ACM Multimedia 1999, Orlando, FL, USA, Vol.: 1, pp. 477-487.*

[5] Liu, Q., Kimber, D., Foote, J., Wilcox, L., Boreczky, J. *FLYSPEC: a multi-user video camera system with hybrid human and automatic control, Proceedings of ACM Multimedia 2002, Juan-les-Pins, France, pp. 484-492.*

[6] Erol B. and Li Y. *in Proc. ICASSP: An overview of technologies for e-meeting and e-lecture. 2005, pp. 6-12.*

[7] Kellermann W. *SPECOM 2009: Towards Natural Acoustic Interfaces for Automatic Speech Recognition. St. Petersburg, June 25-29, 2009, pp. 8-17.*

[8] Brutti A., Omologo M. and Svaizer P. *Hands-Free Speech Communication and Microphone Arrays (HSCMA): Comparison be-tween different sound source localization techniques based on a real data collection. Trento, Italy, May 2008.*

[9] Waibel A., Stiefelhagen R. *Computers in the human interaction loop. Berlin: Springer, 2009, 374 p.*

[10] Ronzhin, A., Budkov, V., and Karpov, A., *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] Al.L. Ronzhin, M.V. Prischepa, Budkov V. Yu., A.A. Karpov, A.L. Ronzhin. *Distributed System of Video Monitoring for the Smart Space. In. Proc. GraphiCon'2010. Saint-Petersburg, Russia, 2010 pp. 207-214. (in Rus.).*

[12] Maurizio O., Piergiorgio S., Alessio B., Luca C. *Machine Learning for Multimodal Interaction: Speaker Localization in CHIL Lectures: Evaluation Criteria and Results. Berlin: Springer, 2006, pp. 476–487.*

[13] A. Ronzhin, A. Karpov, I. Kipyatkova, M. Zelezny. *TSD 2010: Client and Speech Detection System for Intelligent Infokiosk. Springer-Verlag Berlin Heidelberg, Petr Sojka et al. (Eds.): TSD 2010, LNAI 6231, 2010, pp. 560–567.*

[14] Schiele B., Crowley J. L. *European Conference on Computer Vision: Object recognition using multidimensional receptive field histograms. Vol. I, pp. 610–619, April 1996.*

[15] Viola P., Jones M., Snow D. *In Proc.of IEEE ICCV: Detecting pedestrians using patterns of motion and appearance. Pages II: 734–741, 2003.*

About the author

Ronzhin Alexander Leonidovich – PhD student of Speech and Multimodal Interfaces Laboratory of the Institution of the Russian Academy of Sciences St.Petersburg Institute for Informatics and Automation of RAS (SPIIRAS), ronzhinal@iias.spb.su

Karpov Alexey Anatol'evich – PhD, senior researcher of Speech and Multimodal Interfaces Laboratory of the Institution of the Russian Academy of Sciences St.Petersburg Institute for Informatics and Automation of RAS (SPIIRAS), karpov@iias.spb.su