# AUDIOVISUAL RECORDING SYSTEM FOR E-LEARNING APPLICATIONS

Al. L. Ronzhin
*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*

Keywords: Video Surveillance, Computer Vision, Speaker Detection, Smart Meeting Room, Sound Source Localization, Microphone Array.

Abstract: Paper presents the audiovisual recording system, which has been implemented in the SPIIRAS smart room and provides conduct of e-learning meetings. The techniques of video tracking and sound source localization are implemented for recording AVI files of speaker messages in the room. Video processing of streams from five cameras serves for registration participants in fixed chair positions, tracking main speaker and recording view to the audience. During the experiments AVI files with speaker's messages were recorded in a discussion work mode. After manual checking it was detected that 90% of files are files with speaker's speech and 10% of files are false files with noises. 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 message had acceptable rate.

## 1 INTRODUCTION

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. There are several types of recording system, which depend on their goals (Lampi, 2010): 1) video surveillance; 2) meeting recording; 3) presentation recording; 4) documentary recording. Let us consider peculiarity of soft-hardware and methods for audiovisual streams analysis in each type of recording systems.

First type systems record a general view on the auditorium or on the monitoring zone. Such systems use a set of cameras, which are pointed on a direct region. Analysis of image changing in video stream or sound-level allows making general estimation about presence of activity in a room. The image from video stream is displayed on the operator monitor and stored on the hard disk for providing following viewing. The first prototype of such system was developed in the Cornell University (Smith, 1999). The system consists of two cameras for lecture talk and presentation slides recording. In this system a panoramic camera captures the whole area of presenter's movements and detects moment of presentation slides switching. The second camera uses a hardware tracking algorithm for tracking the presenter. At the end of a lecture, processing and combination of recorded video files starts. Thus a single video file has been formed, that allows listening of a lecture talk and seeing presentation slides at e-learning meetings. Such file has been accepted after one hour, which is needed for combination of audio and video streams.

Second type systems are employed for meeting recording and usually equipped by the set of cameras or a camera with 360 degree angle of view (Rui, 2001). An image from white board is also stored if in a room there is such device. The unique feature of systems of this type is that the using of panoramic camera frame processing for presenter tracking and method sound source localization by using microphone array for correction of his/her location are employed. The System cuts an area with

L. Ronzhin A..
AUDIOVISUAL RECORDING SYSTEM FOR E-LEARNING APPLICATIONS.
DOI: 10.5220/0003859905150518
In *Proceedings of the International Conference on Computer Graphics Theory and Applications* (GRAPP-2012), pages 515-518
ISBN: 978-989-8565-02-0

speaker from the image with general view on the auditorium after detection of speaker location (Cutler, 2002). From received fragments video file is created in each frame of it where presence only presenter, that allows to concentrate listener attention on his/her talk without distract attention on behavior of other participants.

Third type system began actively developing in connection with the employing of multimedia presentations during scientifically events. Listeners have to focus their attention all the time for perception of animation objects as well as text information which are displayed on slides. Therefore, at meeting recording main attention is paid on presentation slides, which are used as a background to audio record of lectures talk. The image of a presenter commonly occupies at the most 20 percents of a frame or its size is controlled by user. First such systems may record presentation slides and speaker's talk, so a listener may judge about behavior of participants in the auditorium only by audio signal (Cruz, 1994). Another system is FLYSPEC (Liu, 2002), which was developed at 2002 year by FX Palo Alto Labs and it was intended for supporting teleconference. Two video sensors were implemented in this system: the high resolution camera and the Pan/Tilt/Zoom (PTZ) camera. The system may control second camera automatically or by analysis of participant's requests. At the beginning of a meeting a general view on the auditorium is displayed to assisted and remote participants. During a meeting participants can send commands to the system for PTZ camera pointing to region of interest. The system chooses the optimal camera direction by analysis of received participant requests, which will satisfy the most of remote participants.

Fours type systems are usually employed for films recording, where all parts of it are edited at the end of recording and montage scene by scene to a final version of an audiovisual stream. This systems use set of cameras, which is installed and to be directed on all meeting participants, presenter and projector screen. Furthermore, one of these cameras should record general view on the auditorium. This type of recording is most informative and adequate to artistic style. Meanwhile, self-cost of such recording is rather high, because it uses many devices for recording and processing of audiovisual data. Other disadvantage of this system is presence of human-operator in recording process, which encloses and distracts listener's attention from action in a presentation zone and projector screen, by himself and his devices.

Certain conclusions can be made on the base of analysis of above-listed system classes. These conclusions are useful when developing systems for automation of event writing. At first, recording should be maximum unobtrusive to speakers and listeners. Second, recording and processing should work in real time mode to provide information about current situation in a room to remote participants. Third, recording should consist of presentation slides and talks of all speakers at least. Fourth, because total amount and membership of meeting participants persistently changes and influences on behavior of presenters and listeners, the view on the auditorium can help to remote participants to orientate in a meeting process. Fifth, during technical pauses in speaker's talk, system should add information about meeting, participants or general view on the auditorium in a multimedia report. Also in the process of developing such system cinematograph rules may be used (Rui, 2004).

The developed SPIIRAS smart meeting 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 the other part of the room. Description of the first complex could be found in (Ronzhin Al.L, 2010). Status of multimedia devices and participant activity are analyzed for whole mapping current situation in the room. More information about the SPIIRAS smart meeting room could be found in work (Yusupov, 2011).

## 2 ALGORITHM OF AUDIOVISUAL RECORD FILE CREATION

The creation of the *avi* file is started after silence of the current speaker during five seconds or, that is more frequent case, detection of an active speaker on other chair, conference table or in the presentation area. The main difficulty of recording the *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 1 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.

Participant can 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 can write several audio files, which belong to the same participant (more precisely put, belongs to chair coordinates assigned to this speaker). Name of the audio file includes information about chair number, from which speech signal was recorded.

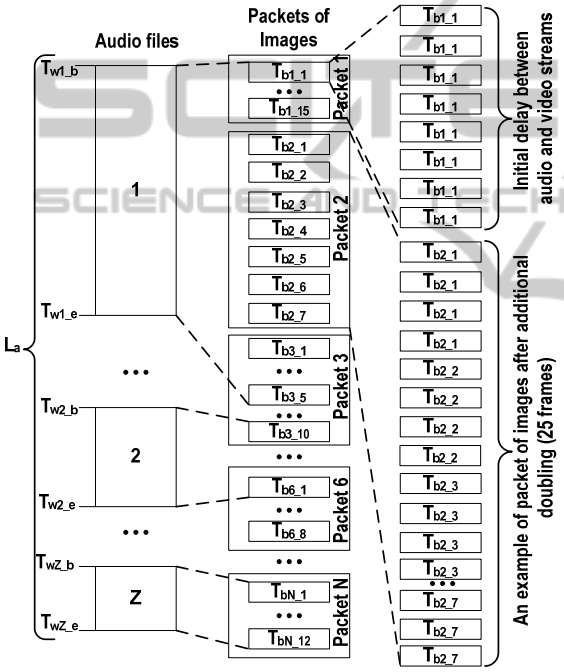


Figure 1: An example of audio and video streams synchronization for recording *avi* file.

Then, duplication of images of determinate time intervals is started to get normal FPS (25 frame per second). Edition of the *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 is needed for elimination of asynchrony appearance at recording of audio and video streams, because it allows calculating total amount of additional frames. After processing of all *bmp* and *wav* files selected and duplicated images are added to an *avi* file, then *wav* files are added to them.

The described algorithm serve for recording remarks of participants sitting on thirty-two chairs of the right side of the smart room (Ronzhin, 2011). 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. After end of *avi* files with speakers messages preparation all these file may be used at e-learning meetings. 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 (Ronzhin, 2010).

# 3 EXPERIMENTS

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

(1) Initial delay between audio and video streams $L_{b\_d}$ is calculated as difference between first *wav* file creation time $T_{w1\_b}$ and *bmp* file $T_{b1\_1}$ creation time, corresponding with a $T_{w1\_b}$ time:

$$L_{b\_d} = \left|T_{w1\_b} - T_{b1\_1}\right|;$$

(2) A length of recorded *avi* file $L_a$ is calculated 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 is calculated as summing up of $L_{b_d}$ and all duplicated frames in all image packets $P_i$:

$$N_{f\_d} = L_{b\_d} + \sum_{i=1}^{N} P_i;\ P_i = P_{AF\_i} + P_{RF\_i};$$

$$P_{AF\_i} = \frac{(P_{FN\_i} - P_{F\_i})}{P_{F\_i}};\ P_{NF\_i} = F_D * \frac{(T_{bN\_i} - T_{b1\_i})}{1000};$$

$P_{RF\_i} = (P_{FN\_i} - P_{F\_i})\%P_{F\_i}$; $F_D$ - is the defined number of FPS, which is 25 frames; $P_{F\_i}$ - is the amount of frames in current packet.

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

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 37 *avi* files were recorded in a discussion work mode. After manual checking it was detected that 90% of files are files with speaker's speech and 10% of files are false files with noises. Such noises are carried 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.

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

| $L_{b\_d}$, ms | | | $L_a$, ms | | | $N_{f\_d}$, frames | | |
|---|---|---|---|---|---|---|---|---|
| **Min** | **Max** | **Mean** | **Min** | **Max** | **Mean** | **Min** | **Max** | **Mean** |
| 80 | 2440 | 730 | 3104 | 6432 | 5530 | 23 | 104 | 58 |

A result of experiments shows, that *avi* file in mean consists of 137 frames, 59 of them are duplicated frames and has length of 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 the 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. An analysis of received data shows that *avi* files formed by the system include all speeches and a small percent of false records.

# 4 CONCLUSIONS

The audiovisual system for e-learning applications 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.

# ACKNOWLEDGEMENTS

This work is supported by the Federal Target Program "Research and Research-Human Resources for Innovating Russia in 2009-2013" (contract 14.7 40.11.0357).

# REFERENCES

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

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

Rui, Y., Gupta, A., Cadiz, J. J., 2001 Viewing meetings captured by an omni-directional camera, *Proceedings of ACM CHI, Seattle, WA, USA, pp. 450-457.*

Cutler, R., Rui, Y., Gupta, A., Cadiz, J. J., Tashev, I., He, L., Colburn, A., Zhang, Z., Liu, Z., Silverberg, St., 2002 Distributed Meetings: A Meeting Capture and Broadcasting System, *Proceedings of ACM Multimedia, Juan-les-Pins, France, pp. 503-512.*

Cruz, G., Hill, R., 1994 Capturing and playing multimedia events with STREAMS, *Proceedings of the second ACM international conference on Multimedia, San Francisco, California, USA, pp. 193-200.*

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

Rui Y., Gupta A., Grudin J., and He L., 2004 Automating lecture capture and broadcast: Technology and videography, *ACM Multimedia Systems Journal. pp. 3–15*

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/ru SMART 2010, LNCS 6294, 2010, pp. 62–71.*

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

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

Yusupov R. M., Ronzhin An. L., Prischepa M. V., Ronzhin Al. L. Models and Hardware-Software Solutions for Automatic Control of Intelligent Hall. *Automation and Remote Control, Vol. 72, No. 7, 2011 pp. 1389–1397.*

Ronzhin An. L., Ronzhin Al. L., Budkov V. Yu. Audiovisual Speaker Localization in Medium Smart Meeting Room. *In Proc. 8th International Conference on Information, Communications and Signal Processing ICICS-2011, Singapore, 2011.*