International Journal of Computer Science and Applications,
©Technomathematics Research Foundation
Vol. 7 No. 3, pp. 104- 123, 2010

EXPLOITING AUDIO-VISUAL CORRELATION BY MEANS OF GAZE TRACKING

BARTOSZ KUNKA

Multimedia Systems Department, Gdansk University of Technology, Narutowicza 11/12, Gdansk, 80-233, Poland
kuneck@sound.eti.pg.gda.pl
http://sound.eti.pg.gda.pl/kuneck/

BOZENA KOSTEK

Multimedia Systems Department, Gdansk University of Technology, Narutowicza 11/12, Gdansk, 80-233, Poland
bozenka@sound.eti.pg.gda.pl
http://sound.eti.pg.gda.pl/kostek

This paper presents a novel means for increasing audio-visual correlation analysis reliability. This is done based on gaze tracking technology engineered at the Multimedia Systems Department of the Gdansk University of Technology, Poland. In the paper, the past history and current research in the area of audio-visual perception analysis are shortly reviewed. Then the methodology employing gaze tracking is presented along with the results of audio-visual experiments performed. It is found that the methodology presented can be used to study audio-video correlation. This is confirmed by a sufficiently high correlation between subjective assessment and objective test results. In this study Pearson's correlation coefficient computed for the subjective responses (MOS) and the objective measure equals 0.63 (at the 95% confidence level).

Keywords: gaze tracking; eye tracking system; audio-visual correlation; interaction between sound and video; image proximity effect; objective tests.

1. Introduction

In this paper a novel method of the analysis of audio-visual correlation is presented. This area has been researched by many scientists in the past. The investigations have been carried out in various context applications, for example a lot of work has been devoted to audio-visual speech recognition enhancement (i.e. automatic lip reading) [Abel *et al.* (2009)], audio-video synchronization [Liu and Sato (2008)] or audio localization dependence on visual cues [Sargin *et al.* (2007)]. The last mentioned subject proves especially interesting in the area of sound source localization. Research conducted during the last decade at the Multimedia Systems Department (MSD) of the Gdansk University of Technology confirmed that visual objects could influence the subjective localization of sound sources significantly [Czyzewski *et al.* (2000a); Czyzewski *et al.* (2000b)] [Czyzewski *et al.* (2002)] [Kostek *et al.* (2001)] [Kostek (2005)] [Odya *et al.* (2001a); (2001b); (2001c)].

Up to now, this type of tests have been carried out as subjective evaluation with the participation of a group of experts. The objective of the method described in this paper is to employ the gaze tracking system developed at the MSD while testing audio-visual correlation on a group of experts. Gaze tracking is an objective technique which displays the fixation points of the user on the computer screen. The received information about the location of the fixation point may enable to obtain credible outcomes of audio-visual correlation analysis in a non-invasive way. Experiments with gaze tracking technique consist in the determination of the part of the screen on which the user is looking and in comparing it with the content of the video image. It is worth mentioning that gaze tracking systems are often used for such tasks as checking the user's attention [Ciger *et al.* (2004)], also in the marketing research. That is why it seems valuable to employ such a system to the domain of subjective testing, where the reliability of the so-called experts is of a great importance.

It should also be stressed that this study was focused towards testing the system as a whole and verifying whether all system functionalities are robust to use in audio-visual subjective tests. Even though the number of subjects taken part in experiments is not sufficient from the statistical point of view, the preliminary tests carried out allowed for providing some valid conclusions on the system design and test procedures. Moreover, it should be remembered that audio-visual correlation is a well-researched subject, especially regarding "image proximity effect", thus we assumed that at this preliminary stage we can start with a smaller number of test participants.

In the paper, a short review of research in the area of audio-visual correlation is given. Then, a description of the preliminary tests carried out with the gaze tracking system applied to the domain of audio-visual correlation is presented. The calibration process of the gaze tracking system is shortly described. Two series of tests are conducted and some conclusions are provided.

2. Review of Audio-Video Correlation Studies

Experiments concerning audio-visual perception have been carried out since the 19th century. Various methods to study the interaction between the perception of sound and image were investigated. Some of these studies are shortly reviewed.

First results published by Stratton, showed that ability of localization of sound sources depend on visual cues. The participants of these tests were watching the video in vertically-flipping glasses and then were localizing the sound sources [Meares (1993)]. Research conducted by Witkin et al. in 1952 enabled to test the influence of the view of the announcer's face on localization of his/her voice [Witkin *et al.* (1953)]. Tests proved that people determine the direction of the heard voice as coming from the center when the announcer's face was seen. In the same case the tested claimed that the voice came from the side when their eyes were closed. The research by Witkin proved that the so-called 'image proximity effect' exists. This effect consists in perceiving the sound source which is shifted towards the existing image in relation to the perceived place when the image is not displayed. It was later confirmed by Sakamoto et al. (1982).

Other researchers conducted the experiments which confirmed observations made by Stratton. Thomas (1941) proved that visual cues do not need to be directly related to sound. He used in his experiments lamplight and bell sound. Development of television caused demand on new research studies – associated with localization of stereo sound under the influence of the television screen.

2.1. *Research dedicated to TV*

Very important experiments demonstrating interaction between audio and video in stereo TV were performed by Brook et al. (1984), and others [Sakamoto *et al.* (1981); (1982)]. Several, studies going out to the late 80th have suggested that ordinary stereophonic systems are not sufficient for HDTV use. These investigations have noted the localization error between picture and sound as a reason, that is why Komiyama performed an experiment designed to investigate the acceptable extent of angular displacement between visual and auditory images for on-axis viewing. The image of metronome with distinct synchronous sound was displayed on the screen. The image was presented in the middle plane of the screen with the angle of $+1^o$, $+7^o$ and -3^o. The metronome sound was played randomly in two loudspeakers, emitting in the range of $0^o – 30^o$ (horizontal plane). It was proved that shifting the image upwards, caused shifting the localized sound in the direction [Komiyama (1989)]. Ohgushi et al. (1987) started their study to examine sound systems for HDTV with defining a few assumptions: (1) HDTV sound systems should offer sounds with a higher level of realism than conventional television, (2) HDTV sound systems should be such that the direction of the sound image is as close as possible to that of the corresponding picture on the screen for various viewing positions, (3) HDTV sound systems should be compatible with movie sound reproduction systems. They performed psychological experiments with seven sound reproduction systems. The semantic differential method was used to measure the impressions of observers. The rating scale for evaluation was a seven-grade scale in which +3 corresponded to complete agreement with a given statement and -3 corresponded to rejection of a given statement. Examples of statements were as follows: "sounds from the audience seem to come from a wide area, the sound stage seems wide", "one feels surrounded by sound reverberation", "the perceived distance of the picture image corresponds with the perceived distance of the sound image in case of a close shot", etc. In general, the authors of this study assigned a clear better sensation of reality between high-resolution TV and associated four-signal sound. Also, Allen in the early 90th reviewed relationship between sound and picture for cinema and television systems. His research led to some parameters of sound matching for television systems defined for future [Allen (1991)].

It is interesting that based on the characteristics described in acoustic psychology as "effect of visual priority" [Komiyama *et al.* (1981)], it was possible to propose the technology-based application, resulted in the US patent [Fukuhara *et al.* (2002)]. The patent description [Fukuhara *et al.* (2002)] shows a technology that provides a loudspeaker capable of matching a sound image and a picture image on a screen of an

image reproducer with each other in simple construction and a simple method of installation. The objective of this patent is to minimize visual interference.

Another research study by Nakayama et al. consisted in checking the impact of voice announcer (male or female) on "image proximity effect". The analysis of this research indicated that the image of announcer female caused more profound proximity effect in the case of the male audience and vice versa. They also studied 3-D sound image reproduction systems associated with 3-D video images with the goal of creating a highly realistic form of 3-D TV broadcasting in the future. In their paper a method of 3-D sound image control using loudspeaker arrays that can control the position of the sound image arbitrarily and continuously was described [Nakayama *et al.* (2003)].

More up-to-date studies of audio-visual correlation are often considered in terms of spatial conditions, i.e. home theater system or games [Bech *et al.* (1995)] [Kostek (2005)] [Meares (1993)] [Rumsey *et al.*(2004)] [Woszczyk *et al.*(1995)] [Zielinski *et al.* (2003). Surround sound systems become a common usage, thus this requires an additional research studies related to the examination of interactions between seeing and hearing. Therefore, there is a need for systematic research in this area to answer the question how the video influences the localization of virtual sound sources in multichannel surround systems, especially as sound and video engineers seek such information in order to optimize the surround sound. This may improve production of movie soundtracks, recording of music events and live transmissions, thus the resulting surround sound may seem more natural to the listener. Up to now the experiments are based on subjective testing of a group of people, so-called experts, listening to the sound with- and without visual stimuli. The obtained results are processed employing statistical analysis in order to find some hidden relations underlying the influence of video on the perception of audio, particularly with regard to the influence of video to the directivity of localization of sound sources in the surrounding acoustical space.

2.2. *Other audio-visual related studies*

In the rich literature on audio-visual perception one may find also study within the context of multi-modal sensory signal integration with a focus on audio-visual integration. Fusing information from audio and video sources has resulted in improved performance in applications such as for example tracking. However, as the researchers from the Sheffield University stated: "crossmodal integration is not trivial and requires some cognitive modeling because at a lower level, there is no obvious way to associate depth and sound sources" [Website, ralyx]. Therefore they addressed the problems of integrating spatial and temporal audio-visual stimuli using a geometrical and probabilistic framework and attack the problem of associating sensorial descriptions with representation of prior knowledge [Website, ralyx]. The system implementation in this case consisted in visual sensors (two cameras) and a pair of microphones mounted on a person's head or on a mannequin head. The cameras were synchronized with the audio. The audio-visual tasks were those of tracking the speaking face, where either the visual or auditory cues add disambiguating information or more varied scenarios with a large

amount of challenging audio and visual stimuli such as multiple speakers, varied amount of background noise, occulting objects, faces turned away and getting obscured.

Another study referred to the results of theoretical and experimental researches of psychophysical and aesthetic aspects of sound and picture interaction. The paper discussed perceptual experiments consisting in examination of the influence of visual factor on threshold sensitivity of hearing, the role of associative links in audio-visual perception, and the correlation between sound and picture images in the perception of spatial localization in multichannel sound systems [Dvorko and Ershov (2002)].

It seems that the outcomes of such studies may be directly applied to audio-visual correlation research shown in our paper.

3. Gaze Tracking System

In general, eye tracking is the process of measuring either the motion of an eye relative to the head or the point of gaze. Additional difference between eye and gaze tracking lies in their layout. If the system is head mounted, then eye-in-head angles are measured. If the measuring system is table/computer mounted camera, then gaze angles are measured.The system engineered at the Multimedia Systems Department (GUT) belongs to the second category. The system allows for tracking eye movements and for estimating a localization of the fixation point, i.e. a point a user is looking at the computer screen [Kunka et al. (2009), Kunka and Kostek (2009)]. To make the system flexible and to offer additional functionality, not available in commercial gaze tracking systems, researchers at the MSD designed and engineered their own gaze tracking system. All functionalities of the system can be changed according to the user's needs and are test-oriented. This make possible to change also operational characteristics of hardware, if required. It is worth mentioning that the system engineered allows for using glasses and makes possible free movement of head. The last mentioned feature is important within the context of subjects' comfort and fatigue during tests. The calibration process that was thought up improves the accuracy of the system working. These issues will be explained later on.

3.1. *System description*

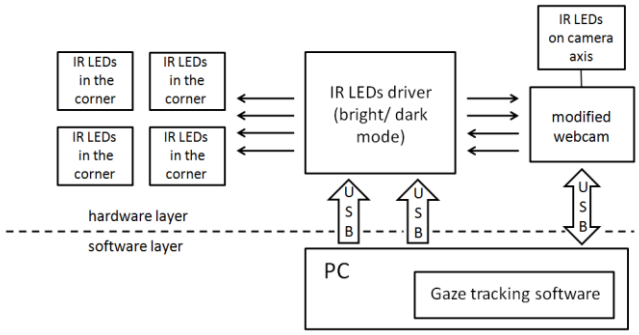
The gaze tracking system presented in Fig. 1a,b comprises of five major components. The USB camera sensitive to the infrared with Pan-Tilt-Zoom (PTZ) mechanism is localized under the monitor. A standard webcam, slightly modified for the purpose of the created interface, was used here. The infrared filter was removed and infrared band-pass filter was mounted on it. The system consists of five sections infrared LEDs. One section is placed around the camera lens. Diodes of this section generate the so-called bright-eye effect which is described in greater detail in Subsection 3.3. The remaining sections of LEDs are placed in the corners of a computer monitor. Another component of the system is the infrared diodes driver which allows for separate activating of IR emitting modules installed on camera axis and display corners. Besides the hardware layer, the interface

created at the MSD consists of software responsible for image processing and control of the IR LEDs driver and camera PTZ mechanism.

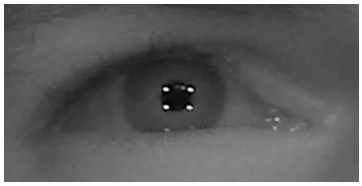
The gaze tracking system developed at the MSD is based on infrared (IR) illumination. Lighting IR LEDs cause arising characteristic corneal reflections, technically known as "glints". Glints are usually the brightest points in the image. They form a shape of quadrangle since they are created by four sections of LEDs located in four display corners. Also, the localization of glints is characteristic. They are localized in the iris and the pupil. The pupil is always very bright because of the IR LED illuminator beaming light along the camera optical axis. It is worth to mention that section diodes along the camera axis cause the fifth glint appearance which is very useful in image processing [Kunka *et al.* (2009)]. The hardware configuration of the developed gaze tracking system, its block diagram and the detected region of the eye are presented below.

a)

b)



c)



d)

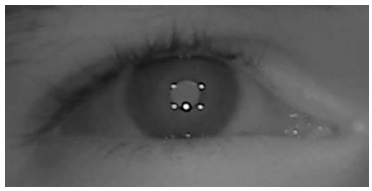


Fig. 1. a) System configuration of the gaze tracker developed by the MSD; b) block diagram of the gaze tracker; c) part of the image with the eye region detected eye illuminated by the IR sources in the dark eye mode and d) in the bright eye mode

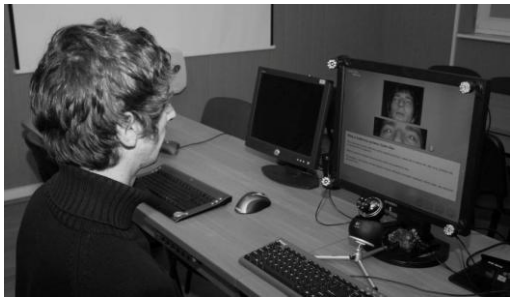


Fig. 2. A user in front of the gaze tracking system

Overall, contrast between the brightness of the iris and the pupil is relatively large, therefore finding area with glints is possible. All data gathered are used in the eye detection procedure and they are usually sufficient to detect eye regions correctly.

3.2. *Spatial resolution and system accuracy*

The main assumption of the gaze tracking system was to have a possibility of distinguishing the nine areas on the computer screen. Such spatial resolution is sufficient in many applications. Nevertheless, the system is able to estimate the fixation point, therefore it is more accurate than determining only one of nine parts of the screen. This fact enables to regard the gaze tracking system as a credible tool for determining the localization of the area on which the user looks.

The second problem may be the accuracy of computing the position of the fixation point in relation to the real fixation point. It is worth mentioning that the calibration

developed for the system procedure improved its accuracy significantly. The calibration procedure is shortly presented in the next two paragraphs.

Due to differences in the iris colors among population and various lighting conditions in the place where the examination is performed the intensity of the IR emission and the reflection pattern (see Fig. 1c, d) which is projected on the user's eye should be appropriately selected. The adaptation of the IR emission intensity is performed during calibration phase which takes place before the system is ready for operation. During calibration the user is required to stare at the nine predefined points sequentially presented on the display. The algorithm compares the estimated fixation point with the known coordinates of the displayed points switching between two possible modes of the system operation:

- bright eye mode – in this case the IR emitting module installed on the camera axis is activated in order to produce so-called bright eye effect;
- dark eye mode – in this case the IR emitting module installed on the camera axis is disabled and the intensity of IR emitted by the modules mounted on the display corners is increased.

Images of the eyes captured by the camera in two abovementioned modes of system operation are presented in Fig. 1c and 1d. Depending on the ambient light conditions, iris color of the user and whether he/she wears (or not) glasses, the algorithm decides which of these two modes of operation gives more reliable results – the effectiveness of the eye detection in the image and the fixation point estimation precision are examined in this way. While the bright eye mode of operation provides generally more accurate fixation point estimates and is usually selected when the ambient light intensity is relatively low, the dark eye mode provides less parasitic reflections on the eye. The dark eye mode is then preferred when the user uses glasses. In this mode there is no reflection coming from the on-axis IR source which when reflected from glasses may mask the eye image and disturb the eye detection process. Furthermore, the errors calculated as a difference between the coordinates of the displayed points and pre-estimated fixation points are monitored in order to extract the fixation point correction terms allowing for more accurate system operation.

4. Experiments

Among many applications of the gaze tracking technology, it could be employed as an objective method supporting the analysis of audio-visual correlations. As mentioned before, until now experiments of audio-visual correlation were based on the subjective opinion of the participants tested. Typically, a viewer was seated before the screen and listened to the accompanying sound. Then the task was to point out the zone in the sound space where he/she seemed to perceive sound.

The authors decided to use the engineered gaze tracking system in audio-visual correlation tests. They conducted two series of experiments. Each of them consisted in filling in a questionnaire form. This phase was related to the so-called "shifting test", the

objective of which was to indicate whether there exists the correlation between the stereo audio basis and the phantom soundscape perceived by a test participant.

4.1. *Preliminary tests*

The system enables to track the way the viewer is looking at the screen. Then, the system analyzes the viewer's concentration. If the concentration is low, this means that the so-called heat (focus) map does not agree with the image and sound related to it [Kunka *et al.* (2010)].

All participants of the preliminary tests sat at the same distance from the screen, in accordance with testing procedure recommendation [Recommendation (2002)]. A group of participants consisted of seven M.Sc. and Ph.D. students of the MSD (males), 25-30 years of age, with a similar education related to multimedia. 12 audio-video samples were presented during the test. Each sample was viewed for approx. 30 s, i.e. time of a single test sample. Table 2 in Subsection 4.1.2 contains precise description of each sample. Two-channel stereo sound was used in these experiments. Each tested participant heard the audio in the headphones and evaluated the shifting of the stereo sound basis in relation to the content of the image.

4.1.1. Sound localization shifting test

The sound localization shifting test is to check whether the observation of video images can change sound localization and if yes then to what degree. The task in the so-called localization shifting test is to evaluate the stereo sound basis in comparison to the phantom soundscape perceived from the video.

Table 1 shows a part of the questionnaire form which enables to analyze stimuli along with the viewer's assessment. Viewers were asked to fill in the evaluation form.

Table 1. Questionnaire form

Stimuli No.	**1**	**2**	**…**	**11**	**12**
Assessment	+2	-1	…	0	-2

Possible levels of assessment are: +2 (too wide), +1, 0 (good), -1, -2 (too narrow).

4.1.2. Description of audio-video samples

The stimuli were prepared at the MSD. To this end, one of the authors prepared an audio-video material, which consists of short scenes presenting musicians playing on their instruments.

It is important to recall first that there are several main types of shots used in filming, i.e. [Website, idrc]:

- ***insert shot*** is used as a shot connecting two shots or emphasizing a detail in a scene. Inserts usually involve an object rather than a person and are often used to make a smooth transition from one scene or shot to another;

- ***close-up*** often used to show the face of the person speaking, and can be a good way to emphasize the importance of what they are saying. Attention can also be directed to part of a person's body by showing it in close-up. Close-ups are more often used on people than objects. They are an effective means of conveying dramatic tension and are widely used in television and movies;
- ***medium shot*** is a camera shot from a medium distance. In such a shot the object or actor and its setting occupy roughly equal areas in the frame. Medium shot shows a person from the waist up. It is used to show people or objects in relation to surroundings. The medium shot is the most common shot used in movies;
- ***medium long shot*** – so-called "American shot", because of its frequent use in westerns, this type of shot shows the subject from the knees up. This shot shows spatial relation between two or more characters in the scene;
- ***long shot*** which are used to show one or several characters from foot to head. Shot which shows all or most of a fairly large object (also character) and usually much of the surroundings.
- ***establishing shot (extreme long shot)*** shows the setting and context where the action takes place. Typically it is a shot at the beginning (or, occasionally, end) of a scene indicating where, and sometimes when, the remainder of the scene takes place. Establishing shots were more common during the classical era of filmmaking than they are at present.

The video prepared does not contain insert or close-up shots. It consists of shots in which the camera moves closer or further from the subject/subjects (medium/long shots), and the same time there are changes done to the accompanying music. In Table 2 examples of prepared stimuli are shown.

Table 2. Stimuli samples

No.	Type of shot	Stereo (two-channel) sound basis
1.	medium	wide
2.	medium	medium
3.	medium	narrow
4.	medium long shot	wide
5.	medium long shot	medium
6.	medium long shot	narrow
7.	long shot	wide
8.	long shot	medium
9.	long shot	narrow
10.	medium	wide
11.	medium	medium
12.	medium	narrow

It is worth mentioning that samples 10-12 do not duplicate samples 1-3. Performers swapped their position in the shot, but the stereo sound image was preserved.

In Fig. 3 an example of the video material is shown (at the left-hand screen), while the right-hand screen presents the so-called heat map, i.e. the effect of the gaze tracking system analysis along with the detected region of the eye.

4.1.3. Preliminary test results

The proposed methodology is based on important assumption: when the generated heat map is focused around the image area related to sound source, then the ratio of audiovisual correlation is high, otherwise sound is not correlated with image. Fig. 4a and 4b present examples of generated heat maps showing two different values of correlation ratio (in colored pictures the scale denotes the frequency of looking at the objects in the image, i.e. generated colors – from blue - the most infrequent to red - the most frequent. To simplify: the larger (more focused) the area in the heat map is, the higher ratio of audio-visual correlation is. Such an observation may be done visually. The methodology of determining audio-visual correlation based on gaze tracking technique requires additional criteria allowing for measuring characteristic features of the generated heat maps. Therefore, two measures were proposed to directly indicate which heat map signifies better focusing. The first measure is based on the amount of pixels which values exceed the determined threshold. For example only red pixels in RGB color space of the heat map layer are being checked. It is necessary to separate the image and the heat map layers from each other because otherwise red pixels of the image content might disturb this procedure (1st measure: "*red pixels*"). The second quantity consists in calculating the sum of lengths of all segments connecting the fixation points revealed in the gaze plot (2nd measure: *length gaze plot (GP) segments*). An example of the gaze plot, which is generated simultaneously with the heat map, is shown in Fig. 5.

Fig. 3. Employing gaze tracking system during the experiment of audio-visual correlation (with stereo sound).

Table 3 contains a summary of test results for both approaches: subjective assessment (based on filling in the form) and objective based on the analysis of the amount of red pixels in heat map layer and the length of gaze plot segments). Each sample in subjective test is represented by two values: arithmetic mean – MOS (*Mean Opinion Score*) and

standard deviation – σ. Another statistical parameter of subjective test is the 95% confidence interval included in Table 3.

Pearson's correlation coefficient (*r*) computed for subjective assessment (MOS) and the "red pixels" objective measure equals 0.62759 (Student's t-test returns the value of -2.54913 at the 95% confidence interval). This confirms the hypothesis that these two measures are correlated. Conversely, Pearson's coefficient for MOS and the length GP is equal to 0.438 (Student's t-test critical value equals 1.54279 in this case). This implicates that the length GP is not useful in the correlation analysis.

Table 3. Preliminary test results – subjective and objective approaches

	Subjective test				**Objective test measures**	
Stimuli No.	**MOS**	**σ**	**95% confidence interval**		**„Red pixels"**	**Length G-P segments [pixels]**
			lower limit	upper limit		
1.	1.35	0.71	0.82	1.88	18058	5312
2.	0.2	0.38	-0.08	0.48	92540	1052
3.	-1.2	0.67	-1.7	-0.7	75289	3541
4.	1.48	0.52	1.09	1.87	35217	4759
5.	0.83	1.22	-0.07	1.73	44223	2548
6.	0.25	0.36	-0.02	0.52	94681	849
7.	1.74	0.91	1.07	2.41	28752	5203
8.	0.8	0.54	0.4	1.2	86514	1267
9.	0.18	0.4	-0.12	0.48	89323	1518
10.	1.1	0.87	0.46	1.74	37813	3891
11.	0.31	0.2	0.16	0.46	98025	820
12.	-0.2	0.53	-0.59	0.19	51976	3827

a)

b)

Fig. 4. Heat maps generated by the gaze tracking system: a) example of high ratio of audio-visual correlation; b) example of low audio-visual correlation.

Fig. 5. An example of the gaze plot

MOS values of the samples assessed in the subjective test should be close to zero. The value of standard deviation should be as small as possible. Samples meeting these conditions are characterized by a higher ratio of audio-visual correlation. Samples No. 2, 6, 9 and 11 could be regarded as samples with a high ratio of audio-visual correlation. Similarly, objective analyzing of the test results may indicate samples with the highest ratio of correlation. Such samples are characterized by the highest amount of red pixels. Samples 2, 6, 8, 9 and 11 meet these conditions. Moreover, by analyzing values of parameters in Table 3 it is possible to indicate samples the soundtracks of which completely do not fit the presented image (the subjective and objective test both). These are samples No. 4 and 7. It appeared that participants preferred watching the performers in the medium shot when the sound sources are located in medium stereo basis (sample No. 2 and 11) and in the long shot with a narrow stereo basis (sample No. 9).

It should be mentioned that seven tested persons do not create a representative group and these preliminary tests could not be regarded as relevant in the statistical sense. Nevertheless, obtained results means a step towards the objectivization of subjective results.

In future we will adopt more sophisticated measures such as proposed in the paper by Nguyen et al. [Nguyen *et al.* (2004)]. They use visual attention (VA) spatial and temporal

characteristics, monitored by a gaze tracking device, to generate a region of interest, the so-called (ROI) 'importance' map. Then, a K-means clustering approach is adopted to group gaze location points into a number of clusters to represent the loci of regions of VA (or ROIs). Several metrics are then derived from the gaze positions and sequences to quantify the relative importance of the K-means clusters. An entropy-weighting strategy is adopted for the combination of these metrics to generate the ROI map. Results obtained by Nguyen and his colleagues show that the ROI map is robust to the number of clusters and different gaze patterns, and can be used in progressive image coding/decoding to enhance the image quality in regions of interest [Nguyen *et al.* (2004)].

4.2. *Second series of experiments*

The second series of audio-visual correlation tests supported by the gaze tracking system referred to research conducted by Witkin in 1952. The image proximity effect was researched for the announcer's voice localized in different parts of the stereo audio basis, with the presenter's face visible in the middle of the frame (see Fig. 7).

Two hypotheses have been assumed and then verified based on the analysis of the visual activity of the viewers: the first one proving the image proximity effect. The second hypothesis enabled to verify the impact fixation point on the presented image on the perception of sound direction.

Seven persons took part in the second series of experiments carried out. Similarly to the first phase, i.e. preliminary tests, a group of participants also consisted of M.Sc. and Ph.D. students of the MSD (males), 25-30 years of age, similar education related to multimedia. Five audio samples and five video samples with the same sound were prepared. Each sample was presented to the test participant for approximately 25 s. Each tested student sat at the same distance from the monitor and listened to audio via headphones. Fig. 6 shows a participant during the test.

Fig. 6. A participant during the second set of experiments

4.2.1. Sound localization shifting test

The test consisted of two phases. The first step consists in indicating the sound direction when only the soundtrack of the sample was presented to the viewer. Each tested person filled in a questionnaire choosing one of the five possible positions of virtual sound source. Fig. 7 shows the potential locations of sound sources in stereo audio basis of the presented samples.

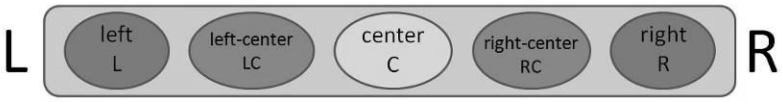


Fig. 7. Possible locations of virtual sound sources

4.2.2. Description of audio-video samples

A set of five samples were prepared. The image was the same in each sample and stereo audio basis was changing. The vision was static. The announcer acting properly sat motionless, only his lips moved. Fig. 8 shows one frame of the samples tested.

Fig. 8. The frame of tested samples

The description of audio and audio-video samples used in the experiment is contained in Tables 4 and 5. The sentence "sound louder in the left channel of 3dB" means that the sound level in the right channel was reduced by 3dB.

Table 4. Audio samples

Sample	Description
audio1.wav	equal level of sound in the left and right channel
audio2.wav	sound louder in the left channel of the 3dB
audio3.wav	sound louder in the right channel of the 3dB
audio4.wav	sound louder in the left channel by 6dB
audio5.wav	sound louder in the right channel by 6dB

Table 5. Audio-video samples

Sample	Description
film1.avi	announcer in the middle frame, the sound from the center (the same level in the left and right channel)
film2.avi	announcer in the middle frame, the sound louder in the left channel by 3dB
film3.avi	announcer in the middle frame, the sound louder in the right channel with 3dB
film4.avi	announcer in the middle frame, the sound louder in the left channel by 6dB
film5.avi	announcer in the middle frame, the sound louder in the right channel by 6dB

Sound samples were played using a standard media player and audio-video samples were played from a dedicated application, created for the purpose of testing interaction between sound and vision. During the playback the gaze tracking system was tracking the viewer's fixation and was saving the point coordinates in the XML file. After the test sample was presented, it was possible to generate the so-called heat map, reflecting the viewer's visual activity. Fig. 9 presents a frame of the video sample with a dynamic heat map associated with it.

Fig. 9. The frame of rendered film sample

4.2.3. Second test results

Table 6 contains the results of the shifting test and results of the evaluated audio-video samples. Results of both tests were obtained from the questionnaire forms filled in by all participants. Values in each row of Table 6 present the differences in perception of sound with and without accompanying the image. Values -2, -1, 0, 1, 2 represent the location of the virtual sound source of the stereo basis: left, left-center, center, right-center, right. Also COG was calculated. The last parameter in the table is the percentage representation of the average time of looking at the announcer's face during projection of the sample. These results were obtained from the analysis of 'film' samples including heat maps.

Table 6. Summary of results

Sample	Location of the virtual sound source of stereo basis					COG	fixation time on the announcer's face [%]
	-2	-1	0	1	2		
audio1.wav	0	4	3	0	0	-0.57	-
film1.avi	0	2	5	0	0	-0.29	92
audio2.wav	4	3	0	0	0	-1.57	-
film2.avi	0	7	0	0	0	-1.00	95
audio3.wav	0	0	0	1	6	1.86	-
film3.avi	0	0	0	4	3	1.43	97
audio4.wav	6	1	0	0	0	-1.86	-
film4.avi	4	3	0	0	0	-1.57	93
audio5.wav	0	0	0	2	5	1.71	-
film5.avi	0	0	0	2	5	1.71	91

The analysis of film samples with marked heat maps (for all participants) confirms the existence of the so-called 'image proximity effect'. Indeed, in the samples No. 1-4 the dynamic heat map is focused around the face of the announcer who is in the middle of the frame. This effect did not occur in all cases to the same extent but the results summarized in Table 6 show that most participants encountered the phenomenon of interaction between sound and vision. This phenomenon is also confirmed by COG (Center-of-Gravity) values. They reflect a change in perception of sound towards the center of the frame where the announcer's face is placed (samples No. 1-4). In addition, the average fixation time (expressed in [%]) confirms that the tested persons were looking at the announcer's face. It may be said that performance on COG was strongly related to subjective perception. This is supported by the graphical presentation of heat maps and by the Pearson's correlation coefficient that is equal to 0.99176 for audio-video samples. Conversely, taking into account the number of participants and the number of audio-video samples, these remarks cannot be supported by the statistical analysis.

Conclusions that may be drawn on the basis of the results of the experiments confirm the validity of using gaze tracking system in audio-visual correlation analysis. Information on the localization of the fixation point during tests is very important because in this way the test participant's answers about the direction of sound perceived may be verified by the gaze tracking system employing heat maps. For most samples, in which respondents experienced image proximity effect, it was noticed that the location of the virtual sound source was in between of the actual sound source and the point of fixation. When the difference between the direction of looking and direction of sound perception was large, the participants felt that there is no match between video and audio and their answers started to be inconsistent.

5. Conclusions

The proposed methodology of researching the influence of the image on perceived sound employing gaze tracking is a novel approach to this domain. The conducted experiments show that it is possible to trace a reaction of the video viewer while listening to sound, and in this way to make the subjective tests more credible.

The first test show that visual objects "attract" the viewers/audience' attention, thus in some cases sound sources may seem to be localized closer to the screen. At the same time it was possible to analyze whether the viewer's attention remains stable through the tests and the audio-visual material keeps his/her interests in order to obtain reliable results. This was supported by a sufficiently high value of Pearson's correlation coefficient (r=0.62759) computed for subjective assessment (MOS) and the "red pixels" objective measure. This confirms the hypothesis that these two measures are correlated. Conversely, Pearson's coefficient for MOS and the length GP is equal to 0.438 (Student's t-test critical value equals 1.54279 in this case). This implicates that the length GP is not useful in the correlation analysis.

The second experiment confirmed the hypothesis introduced by Witkin about the image proximity effect by means of gaze tracking technique. It was observed that the correlation between audio and video samples was strongly correlated (r=0.99176).

Even though a group of tested persons was too small to be regarded statistically relevant, the carried out experiments allowed for formulating some remarks on the system design as well as for building a testing procedure which can be followed step-by-step in the next steps of research.

Acknowledgments

Research funded within the project No. POIG.01.03.01-22-017/08, entitled "Elaboration of a series of multimodal interfaces and their implementation to educational, medical, security and industrial applications". The project is subsidized by the European regional development fund and by the Polish State budget.

References

Abel A., Hussain A., Nguyen Q-D., Ringeval F., Chetouani M., Milgram M. (2009): Maximising Audiovisual Correlation with Automatic Lip Tracking and Vowel Based Segmentation, LNCS 5707 (Fierrez J. et al., eds.), Springer Verlag, Berlin, Heidelberg, **5707**/2009, pp. 65-72.

Allen I. (1991): Matching the Sound to the Picture, 9th Audio Eng. Soc. International Conference: Television Sound Today and Tomorrow.

Bech S., Hansen V., Woszczyk W. (1995): Interactions Between Audio-Visual Factors in a Home Theater System: Experimental Results, 99th Audio Eng. Soc. Conv., Preprint No. 4096, New York.

Brook M., Danilenko L., Strasser W. (1984): Wie bewertet der Zuschauer das stereofone Fernsehes, 13 Tonemeistertagung; Internationaler Kongres, pp. 367-377.

Ciger J., Herbelin B., Thalmann D. (2004): Evaluation of Gaze Tracking Technology for Social Interaction in Virtual Environments, Proc. 2nd Workshop on Modeling and Motion Capture Techniques for Virtual Environments, CAPTECH04.

Czyzewski A., Kornacki A., Kostek B., Odya P., Zielinski S. (2000a): Influence of visual cues on the perception of surround sound, 139th Meeting of the Acoustical Society of America, Atlanta, paper No. 3a, pp. 14.

Czyzewski A., Kostek B., Odya P., Zielinski S. (2000b): Determining Influence of Visual Cues on the Perception of Surround Sound Using Soft Computing, RSCTC'200, Banff, Canada, pp. 507 – 516.

Czyzewski A., Kostek B., Odya P. (2002): Making Surround Audio Considering Image Proximity Effect, Preprint, 112th Audio Eng. Soc. Convention, No. 5583, Munich.

Dvorko N. I., Ershov K. (2002): Audio-visual Perception of Video and Multimedia Programs, 21st International Conference: Architectural Acoustics and Sound Reinforcement, Paper Number: 000122.

Fukuhara S., Kuramitsu I., Omori T., Mizutani T. (2002): Loudspeaker, US Patent 6343132, Patent issued, Application No. 030793 filed on 02/26/1998.

http://ralyx.inria.fr/2007/Raweb/perception/uid37.html (description of the Sheffield University project entitled "Perception and Multimodal Modelling of Space and Motion" concerning audiovisual perception).

http://www.idrc.ca/uploads/user-S/11606750781Sheet13_Video.pdf; (Video, Putting information to work for research projects, Sheet 13: Whether it is film, television or video, the moving image is a very powerful tool of mass communication. The combination of words and images has a greater potential to grab people's attention than any other mass medium).

Komiyama S., Nakabayashi K., Nikaido S. (1981): Experiments on the Interaction Between a Sound Image and a Video Image, Society for Acoustic Research (recalled in the US Patent No. 6343132).

Komiyama S. (1989): Subjective Evaluation of Angular Displacement between Picture and Sound Directions for HDTV Sound Systems, J. Audio Eng. Soc., **37**(4), pp. 210-214.

Kostek B., Krolikowski R., Czyzewski A. (2001): Discovering the Influence of Visual Stimuli on the Perception of Surround Sound Using Genetic Algorithms, 19th International Audio Engineering Society Conference: Surround Sound - Techniques, Technology and Perception.

Kostek B. (2005): Perception-Based Data Processing in Acoustics. Applications to Music Information Retrieval and Psychophysiology of Hearing, Springer Verlag, Series on Cognitive Technologies, Berlin, Heidelberg, New York.

Kunka B., Czyzewski A., Kostek B. (2009): Concentration tests: an application of gaze tracker to concentration exercises, 1st International Conference on Computer Supported Education, Lisbon.

Kunka B., Kostek B. (2009): A New Method of Audio-Visual Correlation Analysis, International Multiconference on Computer Science and Information Technology, **4**, pp. 497 – 502, Mragowo, Poland.

Kunka B., Kostek B., Kulesza M., Szczuko P., Czyzewski A. (2010): Gaze-Tracking-Based Audio-Visual Correlation Analysis Employing Quality of Experience Methodology, Intelligent Decision Technologies, vol. 4, No. 3, pp. 217-227, 2010.

Liu Y., Sato Y. (2008): Recovering Audio-to-Video Synchronization by Audiovisual Correlation Analysis, International Conference on Pattern Recognition (ICPR2008).

Meares D. J. (1993): Perceptual Attributes of Multichannel Sound, The Proceedings of 1st Audio Eng. Soc. International Conference, Copenhagen, Denmark, pp. 171-179.

Nakayama Y., Watanabe K., Komiyama S., Okano F., Izumi Y. (2003): A Method of 3-D Sound Image Localization using Loudspeaker Arrays, 114 Audio Eng. Soc. Convention, Paper No. 5793.

Nguyen A., Chandran V., Sridharan S. (2004): Visual attention based ROI maps from gaze tracking data, 2004 International Conference on Image Processing: (ICIP'2004) International Conference on Image Processing, Singapore, (http://cat.inist.fr/?aModele=afficheN&cpsidt=17612222).

Odya P., Kostek B., Czyzewski A. (2001a): Discovering the Influence of Visual Stimulation the Perception of Surround Sound Using Genetic Algorithms, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York.

Odya P., Czyzewski A., Kostek B. (2001b): Determination of Influence of Visual Cues on Perception of Spatial Sound, 110th Audio Eng. Soc. Convention, Amsterdam, Preprint No. 5311.

Odya P., Czyzewski A., Kostek B., Smolinski T. (2001c): Determining the influence of visual stimuli on the perception of surround sound using data mining algorithms, 142nd Meeting of the Acoustical Society of America, Paper No. 2pPP3.

Ohgushi K., Komiyama S., Kurozumi k., Morita A., Ujihara J., Tsujimoto K. (1987): Subjective evaluation of multi-channel Stereophony for HDTV, IEEE Transactions on Broadcasting, BC-33, No.4.

Recommendation ITU-R BT.500-11, 2002, Methodology for the Subjective Assessment of the Quality of Television Picture.

Rumsey F., Ward P., Zielinski S. K. (2004): Can Playing a Computer Game Affect Perception of Audio-Visual Synchrony?, 117 Audio Eng. Soc. Convention, Preprint No. 6224.

Sakamoto N., Gotoh T., Kogure T., Shimbo M. (1981): Controlling Sound-Image Localization in Stereophonic Reproduction, J. Audio Eng. Soc., **29**(11), pp. 794-798.

Sakamoto N., Gotoh T., Kogure T., Shimbo M. (1982): Controlling Sound-Image Localization in Stereophonic Reproduction: Part II, J. Audio Eng. Soc., **30**(10), pp. 719-721.

Sargin M.E., Yemez Y., Erzin E., Tekalp A.M. (2007): Audiovisual Synchronization and Fusion Using Canonical Correlation Analysis, IEEE Transactions on Multimedia, **9**(7), pp. 1396 – 1403.

Thomas G. J. (1941): Experimental study of the influence of vision of sound localization, J. Exp. Psych., 28, 163-177.

Witkin H. A., Wapner S., Leventhal T. (1952): Sound Localization with Conflicting Visual and Auditory Cues, J. Exp. Psych., **43**, pp. 58-67.

Woszczyk W., Bech S., Hansen V. (1995): Interactions Between Audio-Visual Factors in a Home Theater System: Definition of Subjective Attributes, 99th Audio Eng. Soc. Conv., New York, Preprint No. 4133.

Zielinski S. K., Rumsey F., Bech S., B. de Bruyn, R. Kassier (2003): Computer Games and Multichannel Audio Quality - The Effect of Division of Attention between Auditory and Visual Modalities", 24th International Audio Eng. Soc. Conference: Multichannel Audio, Preprint No. 5856, The New Reality.