Gaze-tracking-based audio-visual correlation analysis employing quality of experience methodology

Bartosz Kunka, Bozena Kostek, Maciej Kulesza, Piotr Szczuko and Andrzej Czyzewski
Multimedia Systems Department (MSD), Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology (GUT), Gdansk, Poland

Abstract. This paper investigates a new approach to audio-visual correlation assessment based on the gaze-tracking system developed at the Multimedia Systems Department (MSD) of Gdansk University of Technology (GUT). The gaze-tracking methodology, having roots in Human-Computer Interaction borrows the relevance feedback through gaze-tracking and applies it to the new area of interests, which is Quality of Experience (QoE). Results of subjective tests carried out at the MSD showed a strong dependency between video presented in the screen and the perceived audio. It has also been shown that the application of gaze-tracking to the audio-visual correlation analysis allows for the objectivization of results obtained in subjective tests. Therefore this research study concentrates on the possibility to apply this methodology to the area of Quality of Experience.

Keywords: Gaze-tracking, audio-visual correlation, Human Computer Interaction (HCI), Quality of Experience (QoE), image proximity effect

1. Introduction

The motivation behind this research study is related to the need of systematic research related to the examination of interactions between seeing and hearing to answer the question how the video influences the localization of virtual sound sources, especially as sound and video engineers seek often such information in order to optimize audio signals accompanying video. This may improve production of movie soundtracks, recording of music events and live transmissions, thus the resulting 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 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 directivity of localization of sound sources in the sound field.

Scientists have researched the interactions between seeing and hearing for many years, examples of such research may be found in various domains, e.g. TV, HDTV, QoE, multimedia. Research studies carried out at the Multimedia Systems Department (MSD) – conducted during the last decade – enabled to confirm that subjective localization of sound image depends on visual objects [4–6,10,11,17–19,24].

This paper presents a new methodology applied to the analysis of interaction between stereo sound and video, employing the gaze-tracking system. As already mentioned the investigation of correlation between the perceived sound and video is based on subjective tests, which are the basic methodology of Quality of Experience (QoE) [9]. At the same time, gaze-tracking is often described as benchmarking support from the Human-Computer Interaction (HCI) area [4]. That is why the authors thought about the idea to explore these two areas, and create a methodology that presents an opportunity to make subjective test results more reliable. Despite significant advances in objective measuring methods, the only way to evaluate the subjective quality of audio/video signals is to prepare a set of tests in which a tester’s opinion is given [16]. However, by
borrowing the gaze-tracking from HCI, it is possible to observe and measure the performance of the tester at the same time.

The gaze-tracker developed at the MSD may be regarded as a safe device which could support the research of audio-visual correlations. Experiments with gaze-tracking technique consist in determination of the part of the screen the user is looking on 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 attention, for example in concentration tests dedicated for children [3, 12]. That is why it seems valuable to employ such a system to the domain of subjective testing, where the reliability of so-called experts is of great importance.

In the paper, a short review of the audio-visual interaction studies was first presented. Then, the notion of Quality of Experience was described and explained in relation to audio-visual correlation analysis. Also, subjective tests carried out were presented and their results analyzed. Then, the hardware setup of the gaze-tracking system developed at the MSD was presented. Finally, the proposed methodology based on gaze-tracking technique was described and results of the audio-visual interaction were shown.

2. Review of audio-visual correlation experiments

Stratton was perhaps the first researcher who published outcomes of his experiments related to interaction between two senses: seeing and hearing. This research was conducted in the 19th century. Stratton had proved that human ability of sound sources localization depends on content of the image being watched by the viewer. His research has been recalled by Thomas [25]. In his own experiments, Thomas used two lights that were presented at spatially separated positions, and a buzzer that was presented from various positions between the lights. Subjects made judgments of left, right, or middle depending on the region from which they heard the sound. One of the lights flickered in the same rhythm as the buzzer sounded. The results showed a tendency for the subjects to skew the judgments of the sound localization toward the flickering light or the in-rhythm light. The findings have been interpreted by Thomas as being a phenomenon of inter-sensory relations [25]. Gardner demonstrated how an image can affect the perceived distance between a sound source and the listener [10]. Then, a research study of this type was conducted by Witkin et al. in 1952, testing the influence of the announcer’s face presence in the image on localization of his/her voice [27]. These tests proved that observers determined the direction of the heard voice as coming from the center when the announcer’s face was seen, despite the speakers being located at substantial angle from the screen. With eyes closed the participant perceived the voice as coming from the side. Thus, the so-called ‘image proximity effect’ had been proved by this experiment [27]. Other researchers conducted experiments which confirmed observations made earlier by Stratton. Thomas [25] proved that visual cues do not need to be directly related to sound.

Brook and others dedicated a lot of research to experiments demonstrating interaction between seeing and hearing senses in stereo television [2, 21, 22]. Then, Bech, Hansen and Woszczyk tried to assess impact of these interactions for Home Theater Systems [28]. Also, Zielinski and his colleagues carried out some studies on audio-visual correlation. His experiments were concentrated on the audio set-up and on finding the optimum position in surround system (5.1 standard). The outcomes of these tests indicated that listeners prefer the limitation of channels over the limitation of bandwidth [32]. In general, current researches of interactions between seeing and hearing are often considered in terms of spatial conditions, i.e. home theater system or computer games [1, 10, 14, 20, 28, 32].

3. Quality of experience methodology

The Quality of Experience (QoE) notion emerged due to a renewed interest in audio-visual signals of degraded quality, and in rich literature on this subject is often referred to as a part of Quality of Service (QoS) perceived subjectively by the end-user. There exist, however, several definitions of this notion, in which differences between QoE and QoS are pointed out, i.e. the aim of the network and services should be to achieve the maximum user rating (QoE), while network quality (QoS) is the main building block for reaching that goal effectively. Nevertheless, it always refers to human side of the service, i.e. the overall acceptability of an application or service, as perceived by the end-user [9]. It should also be mentioned that QoE concerns many domains, such as for example multimedia, telecommunications, networks, business, medicine, etc.

However, it seems that bases of this domain have roots in earlier research associated with the evaluation of high quality audio-visual signals. For this purpose subjective tests were thought up and a formal statistical
analysis was proposed. A wide range of recommendations how to carry out such tests were standardized by international organizations such as EBU, ANSI or ITU. Also, listening/viewing conditions, primary application, non-applicability, scales, subject requirements, number of subjects, technical requirements, stimuli as well as reproducibility of tests were considered and included as guidelines in these recommendations. The ITU-T standard contains listening test methods related to perceptual audio evaluation, subjective video quality assessment, subjective audiovisual quality assessment as well as interaction, audiovisual communications, all of the above within the context of multimedia applications, and in addition, methodology for subjective assessment of the quality of TV pictures, multichannel stereophonic sound system with and without accompanying picture, and methods for subjective assessment of audio systems with accompanying picture [1]. Apart from the subjective evaluation there exist a wide range of objective measurement techniques, the outcomes of which are parameters, measures, numbers and quantities. The problem that remains still open is mapping quantitative measurement results on human perception evaluation quality. Many studies can be found trying to resolve this problem [10], but the techniques employed are not always transferable from one to another domain. It may be said that quality assessment of perceived stimuli is a challenging task, since there many factors affecting perception, quality and the assessment process. That is why the step towards objectivization of subjective methodology and reliability of results obtained is very important.

Therefore, the concept of this new approach is to use the subjective testing methodology from QoE and at the same time applying gaze-tracking (HCI) to this area.

4. Audio-visual correlation tests – experiment analysis

The first stage of research conducted at the MSD encompasses subjective tests devoted to audio-visual interaction, while not using the gaze-tracker. Test participants had been listening to the especially prepared sound samples while watching the accompanying video. Participants’ task was then to fill in the questionnaire form. Preparing audio signals dedicated to tests was performed in several stages. First, audio signals were recorded in the MSD studio, characterized by the desired acoustical properties, very low reverberation and low level of noise. A song was performed by a soloist and a guitarist. The recordings produced in the studio were treated as semi-anechoic (very low reverberation and a good sound distribution). To simulate the acoustics of the auditorium, shots which were used in the accompanying video, required rendering appropriate virtual acoustic field, i.e. auralization process. Therefore recorded audio signals were convolved with the auditorium impulse responses. Moreover, the guitar sound recorded in studio was panned in some proportion to central and left channels, and the vocal – to central and right channels. Resulted from this process, the produced audio signals contained samples with different configurations of localization of the guitar and vocal in the stereo sound basis. Different localizations of virtual sound sources enabled to conduct the so-called ‘proximity shifting test’.

During filming, special attention must be paid to continuity – visual, auditory, temporal, and spatial. All these aspects should be considered individually and also as interrelated constraints in order to achieve film of good quality. The next stage of preparing test samples encompassed making films using different shot techniques.

In short, one may discern several main types of shots used in filming, i.e. [26]:

- insert shot, which is used to bridge two shots or emphasize 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 video;
- medium shot requires the camera slightly farther back, showing a person from the waist up.
- 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;
- long shot which are used to show one or several characters from head to foot;
- establishing shot show the setting and context where the action takes place.

While shooting, it is also important not to overuse the zoom shot. On the other hand, such techniques may used to emphasize an expression or a dramatic situation.
Table 1

<table>
<thead>
<tr>
<th>No.</th>
<th>Type of shot</th>
<th>Stereo sound basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>medium</td>
<td>wide</td>
</tr>
<tr>
<td>2.</td>
<td>medium</td>
<td>medium</td>
</tr>
<tr>
<td>3.</td>
<td>medium</td>
<td>narrow</td>
</tr>
<tr>
<td>4.</td>
<td>medium long shot</td>
<td>wide</td>
</tr>
<tr>
<td>5.</td>
<td>medium long shot</td>
<td>medium</td>
</tr>
<tr>
<td>6.</td>
<td>medium long shot</td>
<td>narrow</td>
</tr>
<tr>
<td>7.</td>
<td>long shot</td>
<td>wide</td>
</tr>
<tr>
<td>8.</td>
<td>long shot</td>
<td>medium</td>
</tr>
<tr>
<td>9.</td>
<td>long shot</td>
<td>narrow</td>
</tr>
<tr>
<td>10.</td>
<td>medium</td>
<td>wide</td>
</tr>
<tr>
<td>11.</td>
<td>medium</td>
<td>medium</td>
</tr>
<tr>
<td>12.</td>
<td>medium</td>
<td>narrow</td>
</tr>
</tbody>
</table>

In Fig. 1 examples of video shots are presented that were produced according to the above remarks, i.e.: medium, medium long and long shot [24].

The ‘proximity shifting test’ was conducted in order to indicate the relation between width of the stereo sound basis and the type of shot. The meaning of this test was to check whether video image observation can change sound localization. Tests consisted of 12 video samples presenting the soloist and the guitarist in an auditorium with shifted (in stereo sound basis, panning closer to center, or farther apart) sound sources (guitar and vocal). The sequence of samples is presented in Table 1. It is worth noticing that samples 10–12 do not duplicate samples 1–3. Performers swapped their position in the shot, but the stereo image was preserved.

Experiments were conducted always in the same conditions (a quiet room functioning as a listening studio) with the participation of 55 persons, students of the Multimedia Systems Department, GUT. Test participants were watching samples presented and evaluating the audio-visual correlation. The assessment was performed within a five-level rating scale, i.e.: +2 (too wide), +1, 0 (good), −1, −2 (too narrow) which enabled to evaluate width of stereo sound image in comparison to shift of sound sources in the video image. Results of 55 students’ answers were gathered and they are presented in Fig. 2. The horizontal axis denotes the consecutive samples (according to Table 1), and the vertical axis – levels of assessment. Each sample is represented by two values: arithmetic mean and standard deviation.

Obtained results of the carried out experiment indicate the best correlated samples. 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 in the long shot with narrow stereo basis (sample No. 9). The results analyzed proved that the image proximity effect exists. Indeed, the content of video images influences the perceived localization of sound sources. Results of samples 10–12 reflect this relation well. According to participants the best correlated sample is No. 12. It represents a medium shot with accompanying sound with a narrow stereo basis spread. The narrow stereo sound basis is the closest to the presented content of image, therefore a brain may easily adapt the sound source localization to this content. On the basis of this analysis it was also confirmed that viewers prioritized video image over perceived audio.

5. Gaze-tracking technique

Gaze tracking is a technique which allows for tracking eye movements and for estimating a point a user is looking at the computer display [15,30]. In order to estimate the fixation point the eye is illuminated by infrared (IR) light sources which are invisible to the user and do not disturb his/her interaction with computer. If IR sources are properly installed on the camera and computer display, they produce unique reflections on the user’s eye. The dedicated algorithm analyzes the
image captured by the camera in order to detect the IR reflections present on the eyeball and performs mathematical calculation resulting in coordinates of the point where the user is looking at. Although there are many commercial gaze-tracking system of various properties available in the market, they are expensive and usually very sensitive to the user head motions (head motions significantly degrades the accuracy of the fixation point estimation) [31]. Therefore, instead of employing off-the-shelf gaze tracking system, the system engineered in the MSD was employed during the audio-visual correlation experiments. The block diagram of the developed system is presented in Fig. 3.

The presented gaze tracking system comprises of five major components:

- USB camera sensitive to the infrared with Pan-Tilt-Zoom (PTZ) mechanism, customized lenses and infrared band-pass filter mounted on it;
- IR emitting module installed on camera axis;
- four IR emitting modules installed on computer display corners;
- IR LEDs driver – the USB controlled device allowing for separate activating of IR emitting modules installed on camera axis and display corners;
- software responsible for image processing and control of the IR LEDs driver and camera PTZ mechanism.

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. 4) 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 pre-defined 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.

The images of the eyes captured by the camera in two abovementioned modes of system operation are presented in Fig. 4. 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.

After accomplishing the calibration phase, the system starts continuously monitoring the point the user is looking at. In the Fig. 5 the block diagram of the algorithm for fixation point estimation is presented.

The image captured by the camera is segmented into the set of overlapped regions of dimensions related to the expected eye size. Further, each of the sub-images is verified whether it contains candidates for glints (reflections from IR sources). For all sub-images detected as the candidates for eye images the group of descriptors is calculated. The descriptors are related to the distinctive properties of the eye image illuminated by the IR sources either in bright or in dark eye mode of the system operation. After the regions containing eyes are detected, the center of the pupil is estimated using ellipse fitting procedure. The position of glints is estimated using the center of gravity method, which was found to provide more reliable results comparing to the simple method of detecting the pixel of highest brightness. Finally, the fixation point is estimated basing on the detected shape and coordinates of the IR reflection pattern as well as correction terms determined during calibration phase. The estimated fixation point is then displayed along with so called “heat maps” indicating the regions of user most interest on the second monitor.

Although there exist gaze tracking systems which operate with less number of IR sources mounted on the computer display, they are usually significantly more sensitive to the head movement of the user [7,29]. Consequently, when the user head changes its position comparing to the initial position during calibration, the accuracy of the fixation point estimation is significantly degraded – the calibration process must be repeated. The system presented in the paper provides reliable results even if the user head changes its position within approximately ± 30 cm and ± 20 cm ranges horizontally and vertically, respectively. However, when the user feels freely to change the position when sitting in front of the computer display, the eyes may be no longer present in the image captured by the camera. In order to prevent such a situation the position of the camera is mechanically adjusted (pan and tilt is triggered) basing on the positions of the eyes found in image. If only one eye is found in the image, it is crucial for the proper PTZ control to determine whether it is right or left eye.

The gaze tracking system set-up employed during audio-visual correlation experiments is presented in Fig. 6. According to the requirements outlined in the IEC 60825-12:2004 standard the intensity of the IR
light emitted by the system was found to be low enough to be safe for the users eyes [8].

6. Audio-visual correlation analysis based on gaze-tracking

Among many applications of the gaze-tracking technology, such a system may be employed as an objective method supporting the analysis of audio-visual correlation. The function of a gaze tracking system can be either passive or active. A system can identify user's message target by monitoring the user’s gaze, or the user could use his gaze to directly control an application or launch actions [23]. In the case of the experiments presented in the paper the usage of the gaze-tracker is passive. To date audio-visual correlation was subjectively evaluated and analyzed after filling in a questionnaire form. The method described in this Section enables to carry out the same experiment in more objective way. Preliminary tests of audio-visual correlations, conducted at the MSD, engaging gaze-tracking system were based on the viewer’s concentration [13]. This means that the system generated the gaze plot or/and
heat map which showed parts of the screen the viewer was looking at. If the concentration was low, this means that the so-called heat (focus) map did not agree with the image and sound related to it.

Figure 7 presents the hardware setup and the main concept of the gaze-tracking system employed to the analysis of interaction between seeing and hearing senses. The monitor in front of the user (master display) is equipped with infrared LEDs, and serves for presentation of the testing material. The second monitor (slave display) enables to supervise the application and to control the user’s behavior. The heat map is generated in real time and it is presented on this monitor.

The methodology proposed is based on an important assumption: when the heat map generated is focused around the image area related to sound source, then ratio of audio-visual correlation is high, otherwise sound is not correlated with the image. Figure 8 presents 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).

Obviously, this new methodology of determining audio-visual correlation requires additional criteria allowing for measuring characteristic features of the generated heat maps. Therefore, two additional criteria were proposed to directly indicate which heat map signifies better focusing. The first criterion 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 layer and the heat map layer because otherwise red pixels of the image content might disturb this procedure. The second criterion relies on calculating the sum of lengths of all segments connecting the fixation points revealed in the gaze plot. An example of the gaze plot which is generated simultaneously with the heat map is showed in Fig. 9.

6.1. Experiment result analysis

Fifteen persons took part in the audio-visual correlation test. This group is not an ideal example of the statistical population (males, one age group, 25–30 years; similar education related to multimedia), nevertheless some important conclusions could be specified, especially as the test participants may be called ‘experts’ since they are students of the Multimedia Systems De-
Fig. 10. Audio-visual correlation analysis results.

First, the proposed method of reading this type of graph should be explained. Gray, wider bars show how much red pixels exceed the determined threshold. The vertical axis for this parameter is assigned to the left side of the chart. The black bars are generated from the gaze plot (G-P). Values of pixels describing this parameter should be as small as possible, indicating small changes of the fixation point location. The vertical axis of this parameter is assigned to the right side of the graph. To find samples which were evaluated as ones with good interaction between seeing and hearing, one should choose those characterized by a high value of red pixels in the heat map layer and small value of the total length of segments in the gaze plot. According to this assumption, samples No. 2, 4, 5, 8, 9, 11 and 12 can be discerned. The remaining samples can be evaluated as weakly correlated. Comparing these results to subjective test outcomes (samples with the mean score close to 0 in Fig. 2), one can claim that samples No. 2, 5, 9 and 12 characterize a strong correlation. Outcomes of the objective method supported by the gaze-tracking system may serve as the confirmation of this statement.

It is worth mentioning that the gaze-tracking results, such as gaze plots or heat maps should be analyzed with the context of the content sample, because the audio-video content may influence tests outcomes and their interpretation. Therefore, the proposed audio-visual correlation analysis should be regarded as a good supporting tool accompanying subjective tests.

7. Conclusion

The paper recalled the problem of audio-video correlation in the context of Quality of Experience field. It was also shown that other areas, namely Human-Computer Interaction and gaze interaction may serve as a tool to analyze perception of sound accompanying video. The proposed method and the discussed system for relevance feedback through gaze-tracking seem a very promising approach for researching the influence of video image on the perceived sound.

The gaze-tracking system determines the fixation point and its accuracy is sufficient to distinguish at least nine areas on the screen. Therefore, the gaze-tracking methodology proposed by authors could be regarded as a new objective method supporting the audio-visual correlation examination. This can also ensure the earlier assessment results of the researched audio-visual interaction.

The experiment results showed that visual objects strongly “attract” the viewers’ attention, therefore in some cases sound sources were perceived closer in relation to the screen center comparing to their actual placement in virtual spatial field. At the same time it was possible to analyze whether the viewer’s attention...
remained stable through the tests and the audio-visual material kept his attention in order to obtain credible results.

In the future, the gaze tracking system will be applied to audio-visual correlation tests employing spatial sound, however, this requires new hardware configuration of the gaze-tracking system. Also, additional tests will be performed to evaluate the individual user’s attention given to a particular aspect of performance while changing the video content.

Acknowledgements

Research funded within the project No. POIG.01.03.01-22-017/08, entitled “laboration 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 and fund by the Polish State budget.

References


