www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

PAN ARCHIVES OF ACOUSTICS
Vol. **37**, No. 1, pp. 63–72 (2012)

DOI: 10.2478/v10168-012-0009-4

Objectivization of Audio-Visual Correlation Analysis

Bartosz KUNKA, Bożena KOSTEK

Multimedia Systems Department, Gdańsk University of Technology
Narutowicza 11/12, 80-233 Gdańsk, Poland; e-mail: {kuneck, bozenka}@sound.eti.pg.gda.pl

(*received December 9, 2011; accepted February 5, 2012*)

Simultaneous perception of audio and visual stimuli often causes concealment or misrepresentation of information actually contained in these stimuli. Such effects are called the "image proximity effect" or the "ventriloquism effect" in the literature. Until recently, most research carried out to understand their nature was based on subjective assessments. The authors of this paper propose a methodology based on both subjective and objectively retrieved data. In this methodology, objective data reflect the screen areas that attract most attention. The data were collected and processed by an eye-gaze tracking system. To support the proposed methodology, two series of experiments were conducted – one with a commercial eye-gaze tracking system Tobii T60, and another with the Cyber-Eye system developed at the Multimedia Systems Department of the Gdańsk University of Technology. In most cases, the visual-auditory stimuli were presented using a 3D video. It was found that the eye-gaze tracking system did objectivize the results of experiments. Moreover, the tests revealed a strong correlation between the localization of a visual stimulus on which a participant's gaze focused and the value of the "image proximity effect". It was also proved that gaze tracking may be useful in experiments which aim at evaluation of the proximity effect when presented visual stimuli are stereoscopic.

Keywords: sound perception, sound source localization, virtual sound source shifting, bi-modal perception, cross-modal perception, image proximity effect, ventriloquism effect, visual attention, perceptual illusion, eye-gaze tracking.

1. Introduction

Multimodal perception (or cross-modal perception) is a well-known phenomenon. Concurrent stimulation of sight and hearing is a comprehensive audio-visual stimulus different from the sensations achieved if the two senses were stirred separately (Bogdanowicz, 2000; McGurk, MacDonald, 1976). While presented simultaneously, audio and visual content influences the perception of information and may lead to its misrepresentation or produce perceptual illusions. This is due to the fact that vision is a dominant sensory modality. One of such illusions is the McGurk effect. Visual information derived from looking at a person that is speaking alters what is being heard by observers (McGurk, MacDonald, 1976) – real utterances may be wrongly taken for other sounds. Another example of a perceptual illusion is a shift of a virtual sound source location towards the direction of the visual stimulus presented. This phenomenon happens when auditory and visual information sources are only slightly separated in space. (Bertelson, 1998; Bertelson, Radeau, 1981; Brook *et al.*, 1984; Gardner, 1968; Komiyama, 1989; Kostek, 2005; Nakayama *et al.*, 2003). There also exists a perceptual illusion known as the *image proximity effect*. Hitherto, most of studies conducted in order to understand simultaneous auditory and visual stimuli perception were based entirely on subjective assessments. It is worth mentioning that subjective tests are a very important issue in audio engineering (Beerends, Stemerdink, 1992; Kostek, 2005; Kostek, Sankiewicz, 2011; Sabin *et al.*, 2011). Subjective tests are utilized in the areas that need perceptual verification (Beerends, De Caluwe, 1999; Klonari *et al.*, 2011; Kin, Plaskota, 2011; Sitek, Kostek, 2011) or in cases that it is an only way to carry out assessment of some perceptual phenomena (Rakowski, Rogowski, 2010; 2011). Examples of such a research are very frequent in the literature. However, there is often a need to make subjective tests more objective. The aim of the presented research is to demonstrate the existence of the image proximity effect using an eye-gaze tracking system, thus, utilizing a methodology which may help to objectivize subjective evaluation results.

1.1. Influence of view direction on sound perception

Audio-visual correlations have been researched in different scientific domains. The authors propose to classify this research into four groups. In the first group, audio-visual correlations are researched in the context of auditory and visual stimuli synchronization (Abel *et al.*, 2009; Liu, Sato, 2008). The second group includes studies on sound source localization in the stereo basis with accompanying visual stimulus (Bertelson, 1998; Bertelson, Radeau, 1981; Brook *et al.*, 1984; Gardner, 1968; Komiyama, 1989; Kostek, 2005; Nakayama *et al.*, 2003). Audio-visual correlations can be also researched in the context of perceived experiences (*Quality of Experience* – QoE domain) (Bech *et al.*, 1995; Davis *et al.*, 1999; Hollier, Voelcker, 1997; Storms, Zyda, 2000). The second and third group focus on how visual stimulation changes auditory perception. Finally, the fourth group is audio-visual correlations employed in the development of video compression methods based on information contained in the soundtrack (Chen, Rao, 1998; Lee *et al.*, 2010; Mujal, Kirlin, 2002; Rao, Chen, 1998).

The research presented in this paper refers to the second group of audio-visual correlations and focuses on how the view direction influences the sound perception. So far, no research of the second group mentioned above has considered the direction in which participants focus their gaze when assessing the impact of the visual stimulus on the virtual sound source localization. It is worth mentioning that some scientists have conducted audio-visual correlation experiments with a simultaneous analysis of the viewer's gaze direction (Kunka, Kostek, 2010a; 2010b; Lewald, 1997; Rorden, Driver, 1999). Lewald (1997) constructed a test stand including nine speakers arranged in a circle with the radius of 3.35 m (one in front of the participant, four to his/her left and four to his/her right side). The speakers were located 2.75° apart from each other. Five LEDs were arranged in a circle of a 1.3 m radius, each one 22.5° away from another. The tested subjects were instructed to gaze at the LED which emitted the light. The results of Lewald's experiment indicated that the subjects did not locate the virtual sound source consistently. Some claimed that the sound came from the direction opposite to gazing. Others stated that the virtual sound source was shifting towards the direction of the visual stimulus (Lewald, 1997; Lewald, Ehrenstein, 1998). Rorden (Rorden, Driver, 1999) was the first to employ an eye-gaze tracking system in his audio-visual correlations experiments. Our approach to apply a gaze tracking technique differs from the Rorden's in two important points. Firstly, Rorden employed a head-mounted eye-gaze tracking system, while we used two contactless gaze trackers: a commercial system Tobii T60 and a Cyber-Eye system developed at the Multimedia Systems Department (MSD). Secondly, we exploited a different visual stimulus. Rorden applied LED-emitted light, similarly as Lewald (1997). We used fragments of a professional stereoscopic video content (3D movies). Moreover, nowadays, technologies based on image processing enable analyzing the subject's visual activity, in this case – in audio-visual correlations experiments with a 3D visual stimulus.

1.2. Research configuration

As mentioned above, the research employed two contactless eye-gaze tracking systems: Tobii T60 and the Cyber-Eye. Both systems are based on infrared illumination. The emission of light in this wavelength range supports image processing and significantly enhances the accuracy of the fixation point determination. Advantages of infrared illumination were described in detail by Kunka (2010a).

The first of the two eye-gaze tracking systems employed is the commercial Tobii T60 system. It determines a gaze point 60 times per second (time resolution is 60 Hz) and estimates it with a 5.2 mm accuracy assuming that the user is about 60 cm away from the monitor screen. Such precision provides a 0.5° spatial resolution. Such a high exactness allows applying the eye-gaze tracker not only in audio-visual correlation experiments but in website usability studies as well. Figure 1 presents the Tobii T60 interface hardware used in the conducted experiments.

Fig. 1. Tobii T60 – eye-gaze tracking system.

Cyber-Eye is a uni-modal interface developed at the Multimedia Systems Department. Its time and spatial resolution are worse in comparison to commercial gaze trackers. Cyber-Eye was developed as an interface supporting education of children with special needs. Such an application does not require a high accuracy of fixation points. Cyber-Eye's resolution is determined by the webcam it uses. The webcam utilized in the experiments can process image frames of 1600×1200 pixels five times per second. It means that Cyber-Eye's time resolution is 5 Hz. The accuracy of the gaze point evaluation is ca. 3.3°, assuming a 60 cm distance from the monitor. More detailed information on the Cyber-Eye software and hardware can be found in earlier

publications of the authors (Kunka, Kostek, 2009; 2010a; 2010b; Kunka *et al.*, 2010). The Cyber-Eye system is shown in Fig. 2.

Fig. 2. Cyber-Eye system.

The eye-gaze tracking systems are used in the experiments of audio-visual correlation to trace points of fixation as the viewer looks at a video content stimulus displayed on the screen. The gaze point coordinates recorded by the systems provide objective information on the image areas which attracted the most attention. Distribution of the visual attention synchronized with the video stream was used to draw conclusions on how the view direction impacts the sound perception.

2. Research on visual attention

Development of the eye-gaze tracking techniques enabled a study of viewers' visual activity. Tracking a viewer's gaze direction is an objective way to acquire information about the areas which attract the most visual attention, so called regions of interests (ROIs). Two eye-gaze tracking systems used in the experiment – Tobii T60 and Cyber-Eye – recorded the x, y coordinates of the fixation points and reflected the viewer's visual activity. Fixation points were determined with the assumption that the origin-point $(0, 0)$ is located in the upper left corner of the monitor screen. Information on the gaze direction is relative to the defined ROIs.

Regions of interest are defined using a vision content indexing based on the XML structure. During the process of indexing, areas of images containing visual stimuli are investigated at certain time intervals. This approach enables tracking the viewers' visual activity for both conventional and stereoscopic movies. In the case of stereoscopic vision, a ROI is described by an additional attribute '*depth* 3*D*' which indicates the location of a virtual object in the perceived spatial scene. The location of the object in a 3D scene was stable in most tested samples. '*Depth* 3*D*' takes three attribute values:

- "+" – behind the screen plane (positive parallax),
- "0" – on the screen plane (zero parallax – 2D image),
- "–" – in front of the screen plane (negative parallax).

Figure 3 explains the object location in a 3D scene when the stereoscopic parallax and attribute values of '*depth* 3*D*' are used.

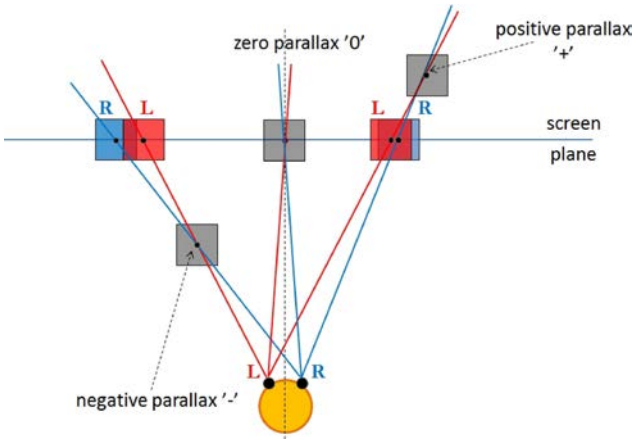


Fig. 3. Stereoscopic parallax and '*depth 3D*' attribute values.

Assigning a relevant attribute to a selected ROI allows tracking the localization of fixation points within tridimensional image elements. Defining many ROIs in different intervals within one audio-visual sample is also possible. Indexing data is provided on the XML-based structure and consists of:

- metadata – general information about an audio-visual sample (sample name, duration, image resolution);
- area – specifies ROI dimensions in pixels [px] and stores a ROI label;
- interval – determines time intervals of a specified ROI [ms].

The aim of indexing the vision content is to compare the coordinates of fixation points obtained from eye-gaze tracking systems with the coordinates of particular ROIs. Relating the information of view directions with defined ROIs enables determining relative times of focusing the visual attention (denoted by a) on a visual stimulus directly associated with a ROI.

3. Audio-video test sample characterization

Four audio-video samples with five characteristic visual stimuli were prepared for the experiments. All the samples were presented twice – first the soundtrack alone, and then the soundtrack with the accompanying video. In both cases subjects used a slider on the interface screen to indicate the perceived sound direction. The range of $[-30°, +30°]$ possible values was assumed. All test samples were fragments of 3D professional movies. We decided to study the image proximity effect using a stereoscopic video content for two important reasons. First, a 3D image provides viewers with more information about the observed scene, therefore, it engages them more than

a conventional 2D image does. It is also worth noting that the subjects assessed the 3D effect focusing on the stereoscopic depth of the projected movies. Thus, it may be assumed that the stereoscopic content-based analysis of viewer's visual activity is more accurate. Soundtracks of all the samples were prepared in a stereophonic system. A detailed description of the samples is presented in Table 1.

Table 1. Test samples characterization.

Sample No./ Stimulus ID	Description	Visual stimulus
1	fragment of *Avatar* movie; five shots made with the use of a camera dolly	character's face located in the central-right part of the frame; subtle change of the stimulus location
2	fragment of *Avatar* movie; five shots; camera tracks the character	character's face located in the central-right part of the frame; quick, dynamical movements of the character
3 (3_*A*, 3_*Q*)	fragment of *Alice in Wonderland* movie; nine static shots; fixed camera	face of the character No. 1 (*Alice*) located in the right part of the frame (3_*A*)
		face of the character No. 2 (*Queen*) located in the central-left part of the frame (3_*Q*)
4	fragment of *Piranha 3D* movie; one static shot; fixed camera	character's face located in the left part of the frame

4. Experiments

The aim of the presented experiments was to investigate, by means of an eye-gaze tracking system, how the view direction influences the sound perception. Tracking the subject's visual attention enables revealing relations between subjective and objective data. Subjective data are based on subjects' evaluations provided during the tests. It indicates the shift of a virtual sound source towards the direction of a visual stimulus after an audio-visual sample projection. The value of this shift is denoted as V [°]. Objective data may be of two types. The first one is associated with the visual stimulus characteristics, and it points to its angular deviation from the screen centre (location of the visual stimulus in the frame). This parameter is denoted as c [°]. The second type of objective data is directly related to the coordinates of fixation points determined by an eye-gaze tracking system. The comparison of a viewer's gaze fixation points with ROIs defined during the indexing process enables evaluating the viewer's visual attention a expressed by Eq. (1). The parameter n_{ROI} represents the number of fixations registered in a ROI, and parameter t_{ROI} denotes the time of visual focusing on a ROI. Quantities in the denominator denote respectively the number of all viewer's fixations (N) and time of the sample duration (T).

$$a = \frac{n_{\mathrm{ROI}}}{N} = \frac{t_{\mathrm{ROI}}}{T} \quad [\%]\,. \tag{1}$$

Two series of experiments with two different eye-gaze tracking systems – Tobii T60 and Cyber-Eye – were conducted. In each series, 15 subjects were tested. Although among them there were people wearing glasses, none of the subjects who took part in the experiments reported problems with the 3D perception. It should be added here that both systems have a robust eye tracking algorithms, thus, they work properly with glasses. The test presentation schedule for the audio and audio-visual samples is represented in the scheme shown in Fig. 4.

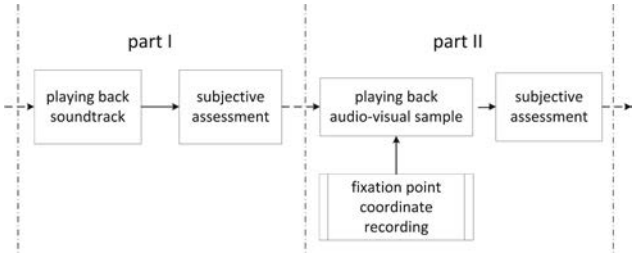


Fig. 4. Test presentation schedule of the conducted experiments.

All subjective assessments obtained in the first and second part of the experiments were analyzed using the ANOVA test. ANOVA (analysis of variance) enables verifying the homogeneity of the compared variables or group means at a specified significance level (Rutkowska, Socha, 2005). We assumed the significance level of 0.05 in the analysis of the experiments results.

4.1. Conditions of the experiments

The research was carried out in an auditory room in which stable conditions were maintained. The auditory room was dimmed, thus, the test participants were not distracted and could concentrate on the displayed visual content. Due to the limitation of the Tobii system which was not provided with a multichannel sound card the experiments were conducted using a two-channel stereo sound system. The auditory room was equipped with the NEXO speaker system, a table on which the eye-gaze tracking system was placed (in the first experiment – Tobii T60, in the second – Cyber-Eye), a special stand where a subject could lean his/her head, and a notebook designed

for filling in the form for subjective evaluations. The special stand was actually a part of a slit lamp used in ophthalmic consulting rooms. Figure 5 presents the equipment and the layout of the auditory room where all the studies were carried out.

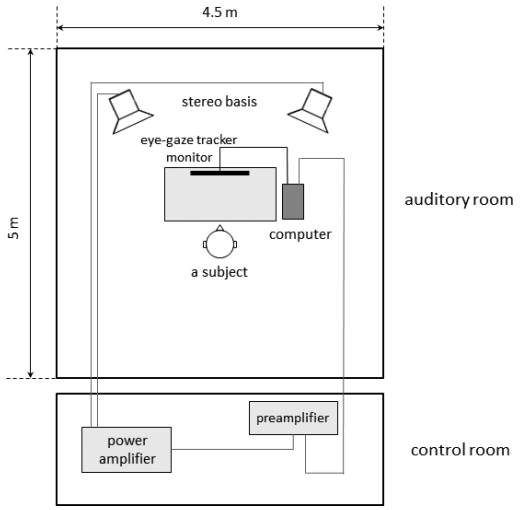
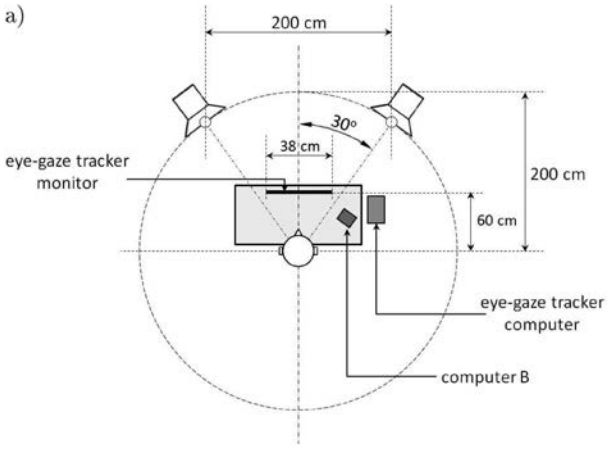


Fig. 5. Setup of the auditory room during the experiments.

More details of the test stand configuration are shown in Fig. 6. The stereo base width was set in compliance with the ITU-R BS.1116-1 recommendation (ITU, 1994–1997), and it equalled to 200 cm. According to this recommendation, the radius of the circle on which the speakers are deployed must be within the range of 200–300 cm. Figure 6 presents the test stand in two perspectives.



computer B – computer designed for subjective evaluating (filling in the electronic form)

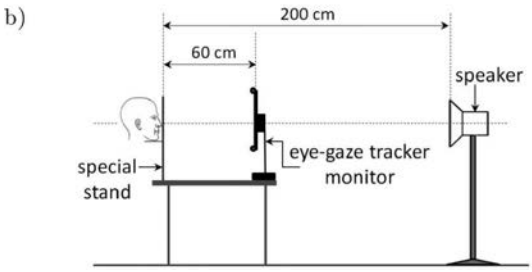


Fig. 6. Listening arrangement with a stereophonic sound system and eye-gaze tracking system.

It is worth mentioning that the distance between the eye-gaze tracking system and the test participant was 60 cm, which is a Cyber-Eye requirement for operating correctly. Generally, it could be regarded as a limitation, but in the context of the conducted research this constraint proved to be advantageous, as it offered the repeatability of listening conditions. Each subject's head was placed exactly in the same position relatively to the centre of the system screen, and in the sweet spot as well (see Fig. 7). Unfortunately, one assumption of the ITU-R BS.1286 recommendation (ITU, 1997) was not met. According to this recommendation, the ratio of the distance between the subject and the screen to the height of the screen should be in between 3 and 6. The Cyber-Eye screen height was 30 cm, therefore, the ratio was 2. Thus, the requirement ensuring the Cyber-Eye correct operation obstructed the recommendation requirement of the optimum distance between the viewer and the display.

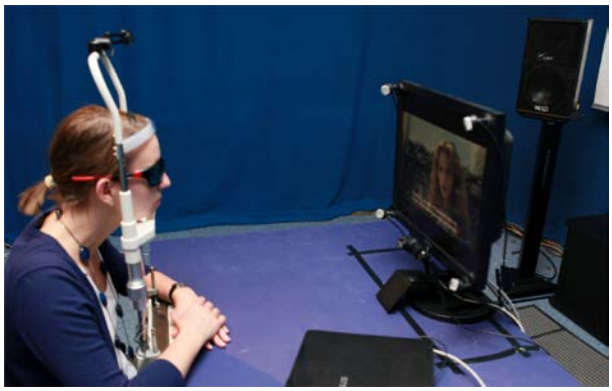


Fig. 7. Subject during the experiment employing the Cyber-Eye system.

4.2. Results of the experiments

It should be noted that the stereo basis width was related to the width of the display. The location and the relative size of the visual stimulus were plotted as a thick black frame in the 'box and whisker' plots.

The authors performed another type of analysis of the data associated with the viewer's visual attention. Although this form of presentation was not directly used in the data analysis, it clearly visualizes the viewers' attention on the displayed visual content. This form of presentation is called a 'dynamic gaze plot'. Figure 8 presents samples of video frames with superimposed dynamic gaze plots generated by two eye-gaze tracking systems. These snapshots of the test samples present characters whose faces are regarded as visual stimuli.

The results of the conducted experiments are presented in the following Sections. Statistical significance

a)

b)

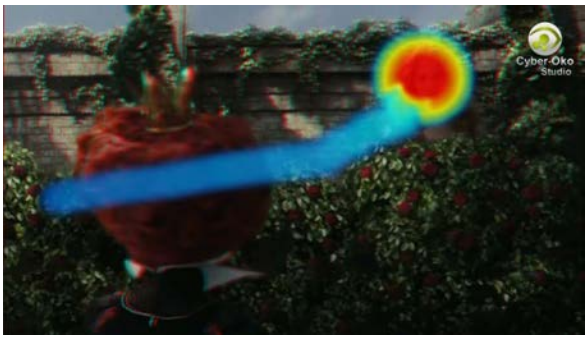


Fig. 8. Examples of dynamic gaze plots: a) generated by the Tobii T60 system, b) generated by the Cyber-Eye system.

of the obtained subjective data was analyzed in Subsubsec. 4.2.1. Then, the relationship between subjective and objective data was investigated.

4.2.1. Statistical significance

Demonstration of the statistical significance of subjective evaluations is an important stage in their analysis. Two conditions that are necessary to perform the ANOVA test were checked for subjective data of each sample. The first condition – normal distribution – was confirmed with Shapiro-Wilk test. The second condition – homogeneity of variance – was checked with Levene's test.

In this subsection, the analysis of the results was divided into two parts depending on the eye-gaze tracking system used. It is worth noting that in the box and whisker plots the value of the viewers' relative attention is inscribed in the area referring to the visual stimulus. The box and whisker plots were presented only for samples No. 1 and 4. During the first experiment, the Tobii T60 system was employed.

The Tobii T60

The ANOVA test showed that the influence of a visual stimulus on the image proximity effect is statistically significant ($F(1,28) = 33.092$, $p < 0.05$) in sample No. 1. Additionally, the shift of the virtual sound source towards the direction of the visual stimulus is presented in the box and whisker plot in Fig. 9. In Subsubsec. 4.2.2, the relation between the objective value of $\boldsymbol{a}$ and observed virtual sound source shifting was investigated.

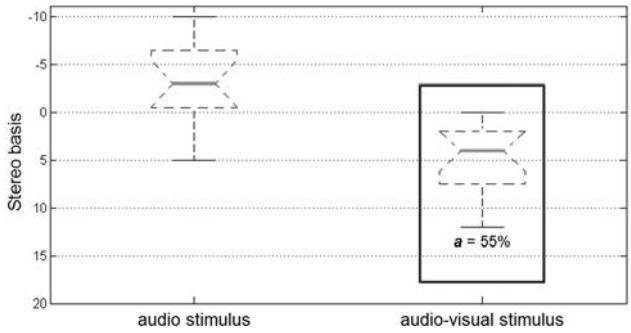


Fig. 9. Box and whisker plot of sample No. 1 (Tobii T60).

Sample No. 2 is characteristic because of its visual stimulus. The main character in this fragment moves dynamically in a substantial part of the frame. The difference between the perceived direction of the sound source in the cases of the audio stimulus and the audio-visual stimulus is significant statistically ($F(2,42) = 11.386$, $p < 0.05$). When assessing sample No. 2, the subjects were asked to answer the question: *Does the virtual sound source change its position during the projection?* Their responses are summarized in Table 2.

Table 2. Summary of the subjects' responses (Tobii T60).

	audio stimulus	audio-visual stimulus
YES	3	10
NO	12	5

When analyzing the subjects' responses, it is noticeable that the visual stimulus influenced the way of the sound perception when it was presented simultaneously with the related visual content.

Sample No. 3 contained two visual stimuli: the ROI associated with the face of character No. 1 (3_A) and the ROI related to the face of character No. 2 (3_Q). The change of the virtual sound source localization was statistically significant in both cases (3_A stimulus: $F(2,42) = 15.12$, $p < 0.05$, 3_Q stimulus: $F(3,56) = 5.335$, $p < 0.05$).

In the case of sample No. 4, the condition of normal distribution for the participants' audio stimulus subjective assessments was not met. Thus, ANOVA test could not be performed, and the value of the alternative Kruskal-Wallis test was determined ($H = 21.98$, $p < 0.05$). The computed p-value indicates that the shift of the virtual sound source localization caused by the visual stimulus is statistically significant. Figure 10 shows distributions of subjective evaluations for the audio alone and audio-visual stimulus. It is also worth paying attention to the parameter a. Its value implies that the subjects were not concentrating on an eye-catching visual stimulus for most sample duration, nonetheless, the observed shift of the perceived sound direction is relatively large and equals to 15°.

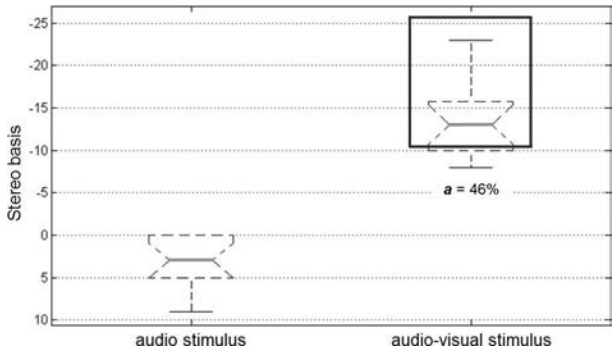


Fig. 10. Box and whisker plot of sample No. 4 (Tobii T60).

One should note that for the Tobii T60 eye-gaze tracking system all the results indicating a shift of the perceived sound direction are statistically significant.

The Cyber-Eye

The second series of experiments in which the Cyber-Eye system was used, was conducted with the same participants in a week's time. In this case, not all visual stimuli influenced the sound perception significantly. Definitive conclusions were specified taking the results of both series of the experiment into account.

In the case of sample No. 1, a significant difference between the subjects' evaluations distributions was observed. This was confirmed by the ANOVA test: $F(1,28) = 12.179$, $p < 0.05$. Figure 11 shows the box and whisker plot for the visual stimulus of sample No. 1.

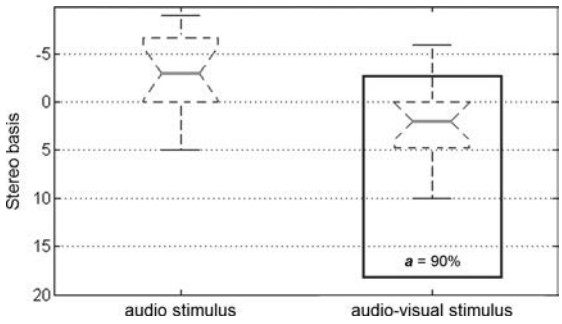


Fig. 11. Box and whisker plot of sample No. 1 (Cyber-Eye).

The first necessary condition of the ANOVA test was not met for sample No. 2. The variable representing responses to the audio-visual stimulus was not specified by the normal distribution. Therefore, the ANOVA test could not be performed, and the value of the alternative Kruskal-Wallis test was determined ($H = 2.89$, $p = 0.236$). The p-value was greater than the accepted level of the statistical significance, thus, the observed shift of the virtual sound source was not significant in the case of sample No. 2.

Moreover, as well as in the previous series of experiments, the subjects answered the following question which referred to sample No. 2: *Does the virtual sound source change its position during the projection?* Responses summarized in Table 3 indicate that 6 out of 15 test participants experienced the perceptual illusion related to the shift of the perceived sound direction.

Table 3. Summary of the subjects' responses (Cyber-Eye).

	audio stimulus	audio-visual stimulus
YES	0	6
NO	15	9

Within sample No. 3, two visual stimuli were tested (3_A and 3_Q). The shift of the virtual sound source location in the case of 3_A stimulus was significant ($F(3,56) = 5.211$, $p < 0.05$), while in the case of 3_Q stimulus it was not ($F(3,56) = 2.246$, $p = 0.093$). For sample No. 3 the 3_A stimulus is much more distant from the centre of the frame in comparison to 3_Q stimulus.

For the last sample (sample No. 4), the shift of the virtual sound source localization reached the level of significance of $F(1,28) = 34.906$, $p < 0.05$. This phenomenon can be observed in Fig. 12 as well.

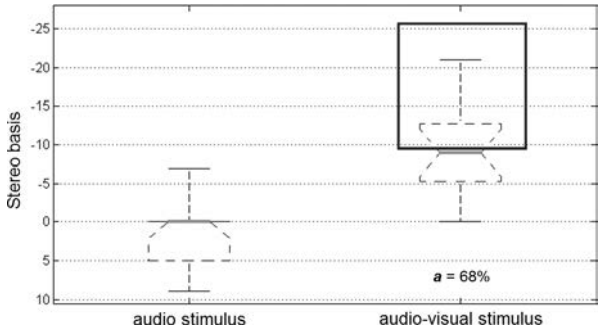


Fig. 12. Box and whisker plot of sample No. 4 (Cyber-Eye).

Concluding the analysis of the results' statistical significance, it is worth noting that in the first series of the experiments (exploiting the Tobii system), all the results of the subjects' assessments were significant. However, in the case of the second series of the experiments (exploiting the Cyber-Eye system), the level of significance was reached for three out of the five visual stimuli. The first stimulus of an insignificant shift of the virtual sound source (sample No. 2) was caused by the fast-moving character – the direction of the sound perception associated with that stimulus was disturbed. In the case of the second stimulus, the location of the eye-catching visual content in the frame was dominant.

4.2.2. Relationship between subjective and objective data

In this subsection the relationship between subjective data represented by the value of the virtual sound source shift V and the objective data represented by the visual stimulus location (c) and the viewer's visual

attention (a) was revealed. This relationship can be expressed by Eq. (2):

$$V = f(c, a). \tag{2}$$

However, the type of the analyzed variables ($a > 0$), allows splitting Eq. (2) into formula (3) and (4):

$$V = f(c), \tag{3}$$
$$|V| = f(a). \tag{4}$$

The analysis of this relationship was carried out employing both eye-gaze tracking systems. It stands to reason that the analysis of the obtained results is not precise enough due to the fact that only five visual stimuli were used. Nevertheless, the outcomes of this analysis reflect the trend in the relation between the subjective and objective data. Therefore, the combined results of the Tobii and Cyber-Eye systems were plotted in Figs. 13 and 14 presented below.

The scatterplots of the analyzed variables clearly indicate their nonlinear dependence, so the Spearman's rank correlation coefficient was used. The correlation coefficient determined for the variables V and c equalled 0.95. This value reflects a strong relationship between the shift of the virtual sound source and the location of the visual stimulus in the frame. Moreover, the correlation coefficient is positive, which means that the increase of one variable causes the increase of the other one too. Generally, the farther from the centre of the screen, the greater the V value is. Figure 13 shows the change in the value V as a function of the visual stimulus location in relation to the screen centre.

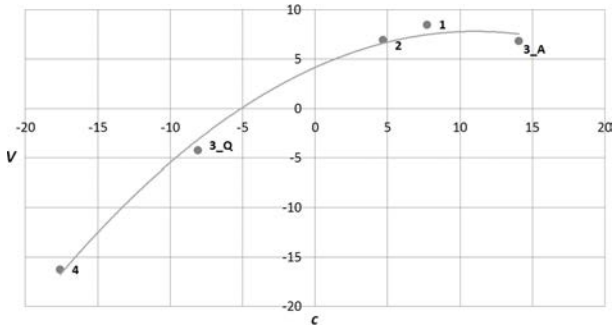


Fig. 13. Scatterplot graph of V and c variables.

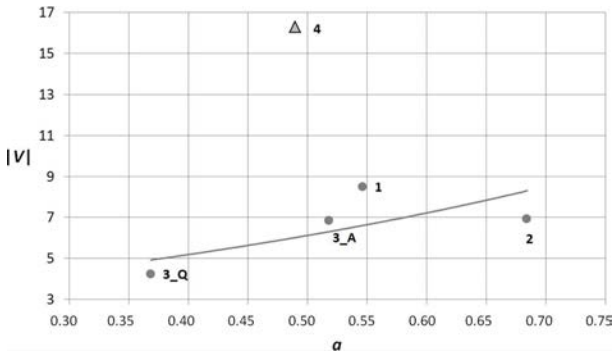


Fig. 14. Scatterplot graph of $|V|$ and a variables.

Formula (4) was investigated in the second stage of the subjective and objective data relationship analysis. The Spearman's rank correlation coefficient determined for the V and a variables equalled 0.3. Nevertheless, it should be noted that the location of the point marked as a triangle (a point apart) in Fig. 14 differs significantly from other measurement points. This point represents sample No. 4 which is characterized by the location of the visual stimulus in the far left part of the frame. Therefore, it can be concluded that the visual stimulus location is a dominant factor influencing the image proximity. Thus, visual stimuli characterized by $|c| \gg 0$ do not represent any appropriate research material for testing the impact of visual attention on the virtual sound source localization.

Given the above assumption it was decided to redetermine the Spearman's rank correlation coefficient excluding sample No. 4. The computed correlation coefficient of $|V|$ and a equalled 0.8. The relationship between visual attention to ROIs and the observed shift of a virtual sound source towards the direction of the visual stimulus has been demonstrated.

4.2.3. Assessment of the 3D effect

It was mentioned that all the research material was displayed in the 3D technology, more specifically, in the anaglyph technique. The test participants assessed the 3D effect quality using a 10-point grading scale. According to the analysis of the perceived stereoscopic depth', the mean value of the 3D effect in the series of experiments with the Tobii equalled 4.98 (standard deviation – S.D. equals 1.62), whereas in the series of experiments with the Cyber-Eye system the mean value equalled 5.21, S.D. = 1.53. Based on these outcomes, it can be concluded that the perceived 3D effect was assessed relatively low. Taking into account the fact that a stereoscopic video was displayed in the anaglyph technique it was decided to conduct an additional experiment in which a polarization technique was used. Seven subjects took part in this experiment. A stereoscopic video was displayed on the Zalman Trimon ZM-M220W polarization monitor. The most characteristic audio-visual sample of the research material (sample No. 3) was used. In this sample, the evaluated sound source was the voice of character No. 1 (*Alice*) and the voice of character No. 2 (*Queen*). The characters were in different locations of the scene depth. Sample No. 3 was projected in the anaglyph and polarization technique. The outcomes of the subjective evaluations obtained for both techniques were analyzed statistically. The distribution of the variable representing the results of assessments for the anaglyph technique did not correspond to a normal distribution, and thus did not meet the first condition of the ANOVA test. Therefore, the value of Kruskal-Wallis test was calculated, and it equalled 5.7, $p < 0.05$. The experiment showed statistically significant differences between the

3D effect' assessments of the anaglyph and polarization techniques. The mean value of the perceived 3D effect in the case of the anaglyph technique equalled 4.57, S.D. = 1.40, whereas the 3D effect in the polarization technique equalled 7.57, S.D. = 1.92.

A research on virtual sound source localizations in the context of the visual stimuli impact on the sound perception indicates that the image proximity effect exists although it is not statistically significant. Thus, despite the statistically significant increase of the perceived 3D effect, the projection of a stereoscopic video in the polarization technique does not ensure any statistically significant differences in the observed shift of a virtual sound source towards the direction of a visual stimulus. At the same time, it should be noted that these results may not reflect precisely the phenomenon of the image proximity effect in the case of a video content' presentation in the polarization technique because the test was conducted employing one audio-visual sample and the perceived 3D effect was not assessed highly enough.

5. Conclusions

In this article, the methodology of studying audio-visual correlation employing an eye-gaze tracking system was proposed and verified. Simultaneous analysis of subjective and objective data is an innovative aspect of the presented research. Exploiting two eye-gaze tracking systems – a commercial Tobii T60 and the Cyber-Eye system) – enabled comparing the functionalities of both interfaces in the context of audio-visual correlations experiments. While completing the form of subjective assessments, the test participants gave answers about the distraction and comfort of each system. The analysis of their answers indicates that they evaluated the comfort of both eye-gaze tracking systems at a comparable level. It is worth mentioning that the experiments did not reveal any significant difference between the systems. Moreover, the Cyber-Eye is characterized by several additional advantages. It enables displaying a stereoscopic video in the polarization technique. The lack of the possibility to display video in any other technique but the anaglyph one is the limitation of the Tobii system. In addition, the Cyber-Eye is compatible with the stereo and surround sound systems, and it enables to conduct audio-visual correlation experiments by displaying the visual content on the projector screen due to a special frame imitating the Cyber-Eye monitor screen.

A strong relationship between the observed shift of a virtual stereo sound source and the visual stimulus location in the frame was proved. The correlation coefficient between V and c variables equals 0.95. The research result analysis showed the correlation between the image proximity effect and the time of the viewer's concentration on visual stimuli (visual attention). Simultaneously, it is worth mentioning that in case of studying the relationship between the observed image proximity effect and the viewer's visual attention using audio-visual samples characterized by $|c| \gg 0$, the shift of the virtual sound source localization in the stereo basis is determined by the visual stimulus location. In such a case, it is the location of the defined ROIs that is a dominant factor in the context of the image proximity effect, and not the visual attention.

In conclusion, it should be emphasized that the proposed methodology of conducting an audio-visual research can be used by researchers involved into experiments of the audio-visual perception, since it may provide objectivization of subjective test results.

Acknowledgment

The research was funded within the project No. SP/I/1/77065/10 entitled: "Creation of universal, open, repository platform for hosting and communication of networked resources of knowledge for science, education and open society of knowledge", being a part of the Strategic Research Program "Interdisciplinary system of interactive scientific and technical information" supported by the National Centre for Research and Development (NCBiR, Poland).

References

1. 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*, Biometric ID Management and Multimodal Communication, Madrid, Spain, 16–18.
2. Bech S., Hansen V., Woszczyk W. (1995), *Interaction Between Audio-Visual Factors in a Home Theater System: Experimental Results*, 99th Audio Eng. Soc. Conv., New York, Preprint No. 4096.
3. Beerends J.G., De Caluwe F.E. (1999), *The Influence of Video Quality on Perceived Audio Quality and Vice Versa*, Journal of the Audio Engineering Society, **47**, 5, 355–362.
4. Beerends J.G., Stemerdink J.A. (1992), *A perceptual audio quality measure based on a psychoacoustic sound representation*, J. Audio Eng. Soc., **40**, 12, 963–978.
5. Bertelson P. (1998), *Starting from the ventriloquist: The perception of multimodal event*, M. Sabourin, F.I.M. Craik, M. Robert [Eds.], Advances in psychological science, **1**. Biological and cognitive aspects, Hove, U.K.: Psychology Press, 419–439.
6. Bertelson P., Radeau M. (1981), *Cross-modal bias and perceptual fusion with auditory-visual spatial discordance*, Perception and Psychophysics, **29**, 6, 578–584.
7. Bogdanowicz M. (2000), *Perceptual-motor integration, theory – diagnosis – therapy* [in Polish], Method-

ological Centre for Psychological and Pedagogical, issue III, Warsaw.

8. Brook M., Danilenko L., Strasser W. (1984), *Wie bewertet der Zuschauer das stereofone Fernsehes* [in German], 13 Tonemeistertagung; Internationaler Kongres, 367–377.
9. Chen T., Rao R.R. (1998), *Audio-visual integration in multimodal communication*, Proceedings of the IEEE, **86**, 5, 837–852.
10. Davis E.T., Scott K., Pair J., Hodges L.F., Oliverio J. (1999), *Can audio enhance visual perception and performance in a virtual environment?*, 43rd Human Factors and Ergonomics Society Annual Meeting, Houston.
11. Gardner M.B. (1968), *Proximity image effect in sound localization*, J. Acoust. Soc. Amer., **43**, 163.
12. Hollier M.P., Voelcker R. (1997), *Objective performance assessment: video quality as an influence on audio perception*, 103rd Eng. Soc. Conv., New York, Preprint No. 4590.
13. ITU-R BS.1116-1 (1994–1997), *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems.*
14. ITU-R BS.1286 (1997), *Methods for the subjective assessment of audio systems with accompanying picture.*
15. Kin M.J., Plaskota P. (2011), *Comparison of sound attributes of multichannel and mixed-down stereo recordings*, Archives of Acoustics, **36**, 2, 333–345.
16. Klonari D., Pastiadis K., Papadelis G., Papanikolaou G. (2011), *Loudness assessment of musical tones equalized in A-weighted level*, Archives of Acoustics, **36**, 2, 239–250.
17. Komiyama S. (1989), *Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems*, J. Audio Eng. Soc., **37**, 4, 210–214.
18. Kostek B. (2005), *Perception-based data processing in acoustics. Applications to music information retrieval and psychophysiology of hearing*, Springer Verlag, Berlin, 389–400.
19. Kostek B., Sankiewicz M. (2011), *Retrospecting Polish Audio Engineering Society membership on 20th anniversary of the Polish Section of the Audio Engineering Society*, Archives of Acoustics, **36**, 2, 187–197.
20. Kunka B., Kostek B. (2009), *A new method of audio-visual correlation analysis*, Proc. of 2nd International Symposium on Multimedia – Applications and Processing MMAP'09, **4**, 497–502, Mrągowo, Poland.
21. Kunka B., Kostek B. (2010a), *Exploiting audio-visual correlation by means of gaze tracking*, International Journal of Computer Science and Applications, **7**, 3, 104–123.
22. Kunka B., Kostek B. (2010b), *Objectivization of audio-video correlation assessment experiments*, 128th Audio Engineering of Society Convention, Paper No. 8148, London.
23. Kunka B., Kostek B., Kulesza M., Szczuko P., Czyżewski A. (2010), *Gaze-tracking based audio-visual correlation analysis employing quality of experience methodology*, Intelligent Decision Technologies Journal, **4**, 3, 217–227.
24. Lee J.S., De Simone F., Ebrahimi T. (2010), *Efficient video coding based on audio-visual focus of attention*, Journal of Visual Communication and Image Representation, Elsevier.
25. Lewald J. (1997), *Eye-position effects in directional hearing*, Behavioural Brain Research, **87**, 35–48.
26. Lewald J., Ehrenstein W.H. (1998), *Auditory-visual spatial integration: a new psychophysical approach using laser pointing to acoustic targets*, J. Acoust. Soc. Am., **104**, 3, 1586–1597.
27. Liu Y., Sato Y. (2008), *Recovering audio-to-video synchronization by audio-visual correlation analysis*, 19th International Conference on Pattern Recognition (ICPR 2008), Tampa, Florida, USA.
28. McGurk H., MacDonald J. (1976), *Hearing lips and seeing voices*, Nature, **264**, 746–748.
29. Mujal M., Kirlin R.L. (2002), *Compression enhancement of video motion of mouth region using joint audio and video coding*, 5th IEEE Southwest Symposium on Image Analysis and Interpretation.
30. 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.
31. Rao R.R., Chen T. (1998), *Exploiting audio-visual correlation in coding of talking head sequences*, IEEE Trans. on Industrial Electronics, **45**, 1, 15–22.
32. Rakowski A., Rogowski P. (2010), *Pitch strength of residual sounds estimated through chroma recognition by absolute-pitch possessors*, Archives of Acoustics, **35**, 3, 331–347.
33. Rakowski A., Rogowski P. (2011), *Absolute pitch and its frequency range*, Archives of Acoustics, **36**, 2, 251–266.
34. Rorden C., Driver J. (1999), *Does auditory attention shift in the direction of an upcoming saccade?*, Pergamon, Neuropsychologia, **37**, 357–377.
35. Rutkowska L., Socha J. (2005), *Statistical data analysis employing STATISTICA program* [in Polish], Lecture Notes, Forestry Faculty, University of Agriculture, Cracow.
36. Sabin A.T., Rafii Z., Pardo B. (2011), *Weighting-Function-Based Rapid Mapping of Descriptors to Audio Processing Parameters*, J. Audio Eng. Soc., **59**, 6, 419–430.
37. Sitek A., Kostek B. (2011), *Study of preference for surround microphone techniques used in the recording of choir and instrumental ensemble*, Archives of Acoustics, **36**, 2, 365–378.
38. Storms R.L., Zyda M.J. (2000), *Interactions in perceived quality of auditory-visual displays*, Presence: Teleoperators and Virtual Environment, **9**, 6, 557–580.