OPEN ACCESS Freely available online

New Aspects of Virtual Sound Source Localization Research–Impact of Visual Angle and 3-D Video Content on Sound Perception

BARTOSZ KUNKA,[1] *AES Member,* **AND BOZENA KOSTEK,**[2] *AES Fellow*
(kuneck@multimed.org) (bokostek@audioacoustics.org)

[1]*Multimedia Systems Department*
[2]*Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications & Informatics, Gdansk University of Technology, Gdansk, Poland.*

The influence of image on virtual sound source localization, called the "image proximity effect" or the "ventriloquism effect," is a well-known phenomenon. This paper focuses on other aspects related to this effect, namely the impact of the visual angle of the presented object and 3-D video content on sound perception. The research conducted confirmed that the visual angle of the presented object determines the image proximity effect regardless of the screen size. An interesting observation was made when studying the impact of 3-D video on virtual sound source localization. When two objects are displayed in a 3-D scene, the viewer's attention is more attracted by the object that is closer to the viewer (negative parallax). Two eye-gaze tracking systems were exploited in the presented experiments to objectivize the obtained results.

0 INTRODUCTION

One sensory stimulus can make various human senses respond simultaneously. We are accustomed to perceiving the world based on combined inputs from our senses resulting in interactions between two or more different sensory modalities. This phenomenon is called multimodal (or cross-modal) perception. The most well known example of multimodal perception is the McGurk effect. The McGurk and McDonald experiment consisted in presenting audio and visual stimuli with inconsistent contents to participants. They asked the participants to watch the face of a person saying the syllables "ga-ga" while the audio presented the syllables "ba-ba" synchronously to the image. The majority of subjects claimed they heard syllables "da-da" in the presented audio-visual sample [1].

Much research dedicated to multimodal perception phenomena has been carried out over the years [2–14]. There are interactions between different modalities. Harrar and Harris [14] noticed that the perceived location of a visual stimulus, an auditory stimulus, and a tactile stimulus are shifted when the head is not aligned with the body. There are many other aspects of multimodal perception apart from the orientation of the head relative to the body. It is worth noting that multimodal perception is currently a very popular research topic. For example it is widely understood in sound engineering that stimulating a tactile sense while "consuming" multimedia content enhances the perceptual experience and makes it more profound [15], [16]. The employment of three communication channels (vision, hearing, and touch) in virtual environments (VEs) is a common procedure. Interesting experiments focusing on the parameters that influence the quality of audio-tactile VEs were conducted by Altinsoy [17]. In our work we investigated a phenomenon of audio-visual correlations. As mentioned before, the McGurk effect is the most well known example of audio-visual illusion related to multimodal perception. Nevertheless, there also exist other interesting perceptual phenomena resulting from simultaneous auditory and visual stimulation. We focused specifically on the aspect of how vision affects the localization of a sound source in the stereo basis (two-channel stereophony). It should be emphasized that most 3-D video content is accompanied by surround sound. Although a lot of research has been so far done on different aspects of 3-D video and surround sound [18], [19] technologies, the impact of a stereoscopic object on sound perception is still seen as a relatively complex phenomenon. Therefore, we concentrated on more basic features of this phenomenon, i.e., on a two-channel stereophony accompanying a 3-D video. The novelty of our research lies within the advanced technology of eye-gaze tracking that we used in the explorations.

The shift of virtual sound source toward the visual stimulus is often described by the term *ventriloquism effect* [10],

[20–25] or *image proximity effect* [11], [26]. The former refers to the effect that may appear when the measured sound directions are "pulled" toward the visible anchors (*ventriloquism effect*) [27], [28]. Generally, both phenomena based on bimodal perception may result in the virtual sound source shift toward the visual stimulus. This auditory sound source shift is usually determined in a subjective way. The subjects are asked to indicate the sound direction by pointing to the apparent location of the virtual sound source in the stereo basis. A shift in the direction of the visual stimulus is observed when subjects' responses are compared to the controlled condition (localization of sound direction in the case of audio stimulation only) [23].

An appropriate illustration of this phenomenon is the experiment carried out by Witkin et al. in 1952 [13]. They researched the influence of the announcer's face on localization of his voice. In the first step of the experiments, subjects watched the announcer and listened to his voice. In the next step, they listened to his voice only. Subjects' responses referring to the first case indicated that the announcer's voice was located in the center of the stereo basis. In the second case (when the subjects had closed their eyes), they perceived the direction of the voice as being on the left or right side. Toole observed that the announcer's voice is often localized in the part of the screen where the moving lips show, and he related this phenomenon to the precedence effect [29]. It is worth noting that the ventriloquism effect also occurs while watching audio-visual content in the form of movies. For this reason, all test samples used in our experiments were fragments of professional movies realized by a stereoscopic technique.

Sound source localization could also be regarded from a different perspective. There are many studies dedicated to improving the localization of sound sources during the process of music recording. Sound source localization could find application in movie production, in virtual reality environments or in teleconferences and distance learning applications using 3-D audio [30], [31].

Moreover, the localization of sound sources impacts the auditory experience of listeners. Roginska [32] proved that stimuli perceived inside a listener's head cause a more accurate and faster response as to their localization than stimuli perceived as externalized.

Spatial audio and virtual sound source localization are important in different contemporary applications. For example it seems desirable to create a feeling of a real-life meeting with each speaker's position properly recognized [33] during teleconferences of several participants.

Furthermore, advanced multichannel audio systems support impression of a viewer's tele-presence. Hamasaki et al. [34], [35] achieved this with the 22.2 multichannel audio system for ultrahigh-definition video with 4000 scanning lines and an advanced multichannel sound system with frontal loudspeakers placed in several rows for reproducing the live sound field.

The main purpose of the research presented in this paper is to investigate two important factors. The first is associated with the visual angle of an object displayed on screens of various sizes. We examined the influence of image on virtual sound source localization depending on screen sizes. The auditory shift in direction of the visual stimulus in the case of each screen size may indicate that the researched phenomenon is scalable. This aspect of the image proximity effect was researched by Bech et al. [36] but in a different context. Also, Emoto et al. [37] carried out a study to establish a clear and quantitative relationship between the viewing angle of the displayed images and the viewer's sensation of presence while watching them.

The second phenomenon investigated was how the observed image proximity effect depends on the position of the presented object in stereoscopic depth. Within this research topic other aspects of the impact of 3-D video on sound perception were studied as well.

Two eye-gaze tracking systems were employed in the conducted audio-visual correlation experiments. The role of these systems was to record the subject's fixation points referring to his or her visual attention. The exploitation of an eye-gaze tracking technique in the investigation of the impact of visual stimuli on virtual sound source localization has been published by the authors previously [38–41]. The information about the direction of the viewer's gaze allows attractive elements of the presented visual content to be tracked. These data are useful in the objectivization of the test procedure results obtained during the subjective evaluation. Moreover, apart from the above mentioned context some additional conclusions related to 3-D video content in audio-visual correlations were drawn. They were based on other experiments carried out and supervised by the authors [41], [42].

It should be emphasized that the study of the interaction of sound and visual stimuli on human perception may contribute to the introduction of some changes to the preparation of audio-visual content. As was indicated, it is possible to enhance the experience of a viewer who is watching a movie; at the same time this enhanced experience based on multimodal perception does not depend on the screen size.

In Section 1 all aspects of experiments conducted within this research are addressed. This Section covers description of experimental procedures, test conditions, characteristics of stimuli, and subjects. In Section 2 we present analysis of results of studied aspects, especially concerning impact of visual angle and 3-D video content on sound perception. Discussion to the obtained results is included in Section 3. Final remarks of the carried out research are provided in Section 4.

1 EXPERIMENTS

1.1 Experimental Procedure

As mentioned before, one of the aims of this paper was to examine the influence of image on virtual sound source localization depending on screen sizes. The visual angle of the presented object does not depend on screen size, as shown in Fig. 1. Fig. 1 presents three different screen sizes: large, medium, and small. The width of each screen is denoted by w. It is worth noting that the visual angle of

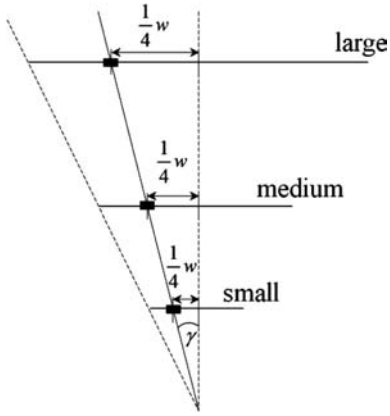


Fig. 1. A constant visual angle of the presented object for different screen sizes.

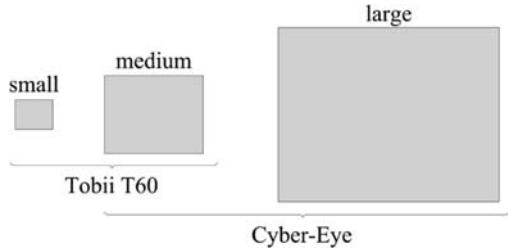
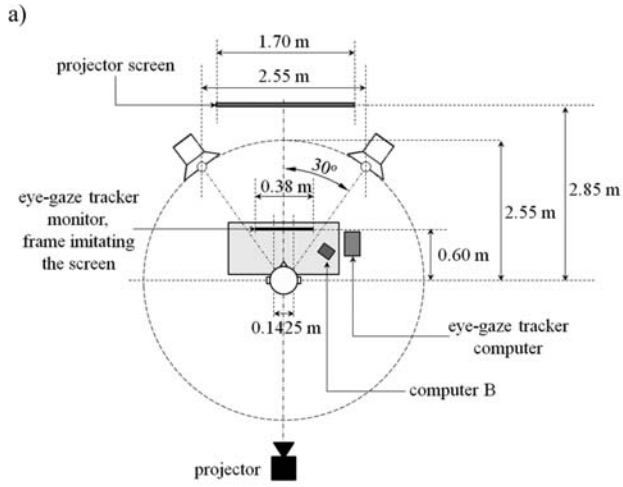


Fig. 2. Two stages of the experiment depending on the employed eye-gaze tracking system.

the displayed object (γ) is constant regardless of the screen size.

In order to study the image proximity effect for three screen sizes with the registration of participants' visual attention it was necessary to split the experiment into two stages. In the first stage we researched the image proximity effect when the research material was presented on small and medium screen sizes. The small screen size referred to in-flight entertainment displays used onboard aircraft, for example. We assumed that its width was equal to 0.1425 m. The medium screen size reflected a standard 19" computer screen (4:3 format) with a width of 0.38 m. In the second stage of the experiment the video content was viewed on medium and large screen sizes. The large screen size referred to the projector screen. In our experiment the width of the display area was 1.70 m. Samples of small and medium display areas were presented on the computer screen. It was dependent on the Cyber-Eye system, which requires a constant distance between the user and the system monitor.

The experimental setup is illustrated in Fig. 2, which shows a two-stage sample presentation. This division was necessary because of the limited functionality of eye-gaze tracking systems. Among the most important characteristics of the eye-gaze tracking systems are their spatial and time resolutions. The commercial gaze-tracking system employed in our experiments is characterized by an angular (spatial) resolution equal to 0.5° and high speed (time resolution) equal to 60 Hz. The second eye-gaze tracking system, developed at the Multimedia Systems Department of the Gdansk University of Technology is called Cyber-Eye.



computer B – computer designed for subjective evaluation

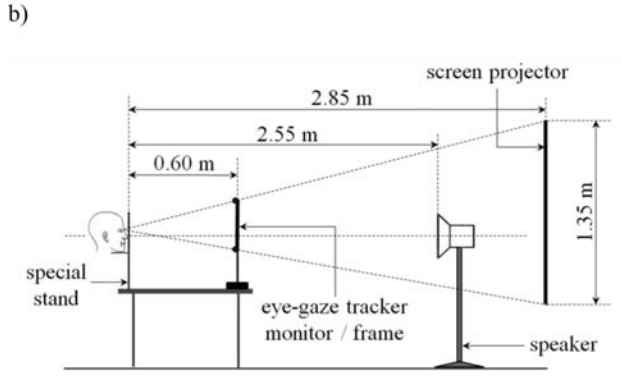


Fig. 3. Listening arrangement with three screen sizes, a stereophonic sound system, and eye-gaze tracking system: a) top view, b) side view.

It was presented thoroughly in our previous publications [38–41], and thus here we recall only the temporal and spatial features of the system, i.e., an angular resolution of ca. 3.3° and time resolution of ca. 5 Hz [38], [39], [41]. Nevertheless, the Cyber-Eye system makes it possible to track the viewer's visual activity on large screens using a special frame imitating the standard computer screen (see Figs. 3b and 4a). Moreover Cyber-Eye works with multichannel sound, but since the first system can use only a two-channel sound configuration, the experiments were limited to two-channel sound presentation.

1.2 Test Conditions

The experiments were conducted in an auditory room maintaining stable control conditions. The subjects were not distracted and could concentrate on the displayed visual content because lights in the auditory room were dimmed.

The presentation of the audio-video samples on small- and medium-sized screens was associated with the commercial system (the first configuration), and presentation on large- and medium-sized screens was associated with the Cyber-Eye system (the second configuration). The sound stereo basis of both arrangements was set in compliance with the ITU-R BS.1116-1 recommendation [43]. The

Fig. 4. The second configuration of the listening arrangement during the experiment: a) a subject during the experiment, b) a view of the video content displayed on the projector screen (subject's perspective).

test stand configuration differed in the stereo basis width (2.00 m in the first configuration, 2.55 m in the second) and the distance between the subject and the screen plane (0.60 m in the first configuration, 2.85 m in the second). The change of the stereo basis width in the second setup is not proportional to the change in the distance between the subject and screen. The listening arrangement of the second configuration is shown in Fig. 3.

Keeping the appropriate ratio of the stereo base width and the subject–screen distance would require an increase in the stereo basis to 9.50 m, while the maximum width of the basis equals 3.00 m, according to the ITU-R BS.1116-1 [43] recommendation. The optimum arrangement of the loudspeaker sound system can be achieved if one of the conditions is met. The first condition is a shift of the projector screen toward the viewer. The second condition is the displacement of the loudspeakers behind the projector screen while maintaining the width of the stereo basis according to the ITU-R BS.1116-1 recommendation.

Nevertheless, neither a shift of the projector screen nor a change of the loudspeaker setup was possible in our auditory room. We plan to carry out our future experiments in surroundings that would meet the requirements stated in the mentioned earlier recommendation and provide the possibility of changing the distance between the projector screen and the viewer.

Table 1. Test samples used in the *basic aspect* (BA) of the investigation

Sample ID	Description	Visual stimulus
1_{BA}	Fragment of *Avatar* movie	Character's face located in the center-right part of the frame; subtle changes of the stimulus location
2_{BA}	Fragment of audio-video recording of a violin and piano concert	Violin (and violinist) located in the left part of the frame

It is worth mentioning that for the second arrangement of the listening test, a special frame imitating the screen of the eye-gaze tracking system was constructed. To provide identical conditions of viewing and listening, each of the subjects leaned his or her head on a special stand in a fixed position with regard to both the center of the system screen and the sweet spot. The stand was an element of a slit lamp used in ophthalmic consulting rooms. The second configuration of the listening test showing the special frame prepared for these tests is presented in Figs. 4a and b.

1.3 Stimuli

Three groups of audio-video samples prepared using the anaglyph technique were employed in the experiments conducted according to the studied phenomenon. The first group of samples was used in studying the so-called *basic aspect* of virtual sound source localization in the stereo basis. First, they were exploited to investigate the influence of screen size on the image proximity effect (impact of the visual angle of the presented object on virtual sound source localization).

Moreover, in the case of the first group of samples we investigated the relation between the observed shift of virtual sound source in the direction of the visual stimulus and length of time during which the viewer's attention was focused on an object/character associated with this sound source, since two eye-gaze tracking systems supported the experiments. A detailed sample description of the first group is presented in Table 1.

In the case of the *basic aspect* of the investigation, subjects evaluated the location of the sound source in the stereo basis. Such an approach is commonly studied due to the dominant role of vision in human perception of the surrounding environment. Nevertheless, the influence of the sound effect on the direction of the viewer's gaze also seems to be an interesting problem. Therefore, we decided to study this aspect by exploiting test samples of the second group (we called this the *additional aspect*).

Test samples of the second group were characterized by the sound effect that occurred at the beginning of the sample. Both samples of the second group were fragments of 3-D movies. The viewer's visual attention was registered before the attractive object that was the sound source appeared in the frame. A detailed description of the samples is included in Table 2.

Table 2. Test samples used in the *additional aspect* (AA) of the investigation

Sample ID	Description	Audio-visual stimulus
1_{AA}	Fragment of *Avatar* movie	Steps of the robot moving from right to left; sound of footsteps becomes louder; the robot appears in the frame when the virtual sound source location refers to the right edge of the screen
2_{AA}	Fragment of *Resident Evil: Afterlife* movie	Engine sound of an approaching light aircraft; sound is heard in the center of the stereo basis and becomes louder; the light plane appears in the bottom of the frame when the engine sound is loudest

Table 3. Test samples used in the investigation of the impact of the 3-D video on virtual sound source localization

Sample ID	Description	Audio-visual stimulus
$3D_P$	An object behind the screen – in the positive parallax area	The object is slowly moving along the axis of symmetry of the screen (0°); the sound source is a voice related to the moving object and is localized in the left part of the stereo basis (–25°)
$3D_N$	An object in front of the screen – in the negative parallax area	The object is slowly moving along the axis of symmetry of the screen (0°); the sound source is a voice related to the moving object and is localized in the left part of the stereo basis (–25°)

As mentioned before, the aim of the experiments conducted was to investigate the influence of 3-D video content on the observed image proximity effect. Specifically, we focused on the relation between the observed shift of virtual sound source localization and the position of the 3-D object attracting the viewer's attention in the stereoscopic depth ("eye-catching" 3-D object). It was noted that professional 3-D movies are mostly characterized by stereoscopic depth related to positive parallax (a 3-D scene is perceived behind the screen plane). This is reasonable because the area of visual comfort for 3-D perception (the so-called 3-D comfort zone) contains an area of positive parallax for the most part [44]. However, the impact of a 3-D image on virtual sound source localization can be studied in a different context. We focused on observation of the image proximity effect when subjects concentrate their gaze on an object in the positive parallax area (behind the screen) and in the negative parallax area (in front of the screen). The research material associated with the investigated context consisted of two prepared samples consisting of short animations (Table 3). The main element of each animation was an object moving from background to foreground. Other fragments of the presented 3-D scene were static and did not distract the subjects. It should be noted that soundtracks of all groups of samples were reproduced in the two-channel sound system due to the limitation of the commercial tracking system, which was not provided with a multichannel sound card.

It should be mentioned that stimuli contained in Tables 1 and 3 were presented twice during the experiments. The standard procedure of virtual sound source localization research was utilized. The soundtrack of each test sample was reproduced first (unimodal stimulus). Then, the audio-visual (bimodal) stimulus was presented.

1.4 Subjects and Their Tasks

Fifteen students of Gdansk University of Technology (5 females, 10 males) participated in our experiments (average age: 24.27; standard deviation (S.D.) = 1.8). They did not know about the issues related to the research topic. Each subject sat with his or her head leaning on the special stand in front of the monitor or projector screen as shown in Figs. 3 and 4a. The vision and hearing acuity of all subjects was examined during regular tests that the students of the Sound and Vision Engineering specialization take in the course of laboratory sessions concerning sound and vision perception. No problems with hearing and stereoscopic vision were reported. A comfortable sound level was set the same for all participants. Three of them wore glasses. It should be added that both our eye-gaze tracking systems work properly with glasses.

In the case of the *basic aspect* researched using the first group of samples, the subjects' task was to indicate the perceived sound direction using a slider on the interface screen. The range of possible values was referred to the width of the stereo basis according to the ITU-R BS.1116-1 recommendation that covers all aspects of subjective quality assessment of sound signals [43].

As mentioned before, the presentation of test samples from Tables 1 and 3 consisted of two stages. In the first stage the sound was presented as a unimodal stimulus, and the virtual sound source localization determined by the listener created a reference point. The second stage consisted in the presentation of audio-visual (bimodal) stimuli. In both cases the subjects' task was to localize a virtual sound source in the stereo basis. Subjects' responses referring to sound source localization were compared to the reference location, which allowed an auditory shift of a few degrees in the direction of the visual stimulus to be identified [24].

2 ANALYSIS OF RESULTS

2.1 Analysis Assumptions

Statistical data analysis of test samples used in the basic aspect was carried out by the ANOVA test. Since each sample was presented in two stages according to the experimental setup presented in Fig. 2, the obtained results were analyzed in two stages as well. We assumed that scalability of the image proximity effect is observed when the null

Table 4. The ANOVA test results for the first experiment (*basic aspect*)

	Tobii T60		Cyber-Eye	
Sample ID	F	p	F	p
1_{BA}	33.09	0.000004	12.18	0.0016
2_{BA}	3.32	0.079	Kruskal-Wallis: $H = 11.262$	0.0008

hypothesis of the ANOVA test is not rejected (homogeneity of the compared mean values of subjects' assessments). Therefore, the expected p-value should be greater than 0.05.

The analysis of results obtained during the presentation of the second group of samples (*additional aspect* of the investigation) was based on descriptive statistics. We determined the average relative time during which the subjects' gaze was focused on the so-called expected region of interest (ROI). The expected ROI is the area where the object that is the sound source appears.

The statistical analysis of the impact of the 3-D video content on the shift of virtual sound source localization in the stereo basis was based on the ANOVA test and mean value analysis. First, we compared the localization of the virtual sound source indicated after the unimodal (sound only) stimulation and then bimodal (audio-visual) stimulation. There was an estimated shift of virtual sound source localization (in degrees) for both $3D_P$ and $3D_N$ samples. Then, the mean value of sound source localization was determined for each sample. In the last step, the statistical significance of the difference between the image proximity effects observed for samples of positive and negative parallax areas was evaluated.

The three groups of prepared test samples refer to three experiments. The results of the first experiment were related to subjective data. The research on the virtual sound source shift was based on the evaluation of the statistical significance of differences between indications of sound source directions for unimodal and bimodal stimuli. The conclusions drawn in the second experiment were based on objective data acquired by the eye-gaze tracking system. The results of the third experiment were analyzed including subjective data (similar to the first experiment). The first two experiments were carried out twice, employing the commercial gaze-tracking and Cyber-Eye systems with a time interval of a week. It was not necessary to utilize the eye-gaze tracking systems in the case of the first experiment. However, we decided to exploit the system to maintain the same conditions in the investigations. The third experiment was conducted only once, with another group of participants and at a different time to the two previous experiments.

As mentioned above, the analysis of the results obtained for the first and the third experiments was based on the ANOVA test. To perform the ANOVA test it is necessary to check that two conditions are met: a normal distribution (examined with the Shapiro-Wilk test) and homogeneity of variance (checked with Levene's test) [45]. When one of the above conditions was not met then the alternative Kruskal-Wallis test [45] was performed.

2.2 Results for Basic Aspect Stimuli

2.2.1 Shift of Virtual Sound Source in the Stereo Basis

Table 4 presents the results for the statistical significance of virtual sound source shift when the video content was displayed on the medium-sized screen. We assumed a significance level (α) of 0.05 in the analysis of the experiment results. According to this assumption the image proximity effect unquestionably occurred in the case of sample 1_{BA} (p-value close to 0) in both sessions of this experiment. In the case of sample 2_{BA}, a virtual sound source shift toward the visual stimulus was observed, but in the first session utilizing the commercial gaze-tracking system it was not statistically significant ($p > 0.05$). According to the results of the second session the image proximity effect was confirmed for this sample.

2.2.2 Relation between the Observed Shift and Viewers Visual Attention

It is worth mentioning that viewers' visual attention (based on data acquired from both eye-gaze tracking systems) has an impact on the observed shift of the virtual sound source localization. We studied this relation for each subject independently. Then, we determined the Spearman's rank correlation coefficient for two vectors, the first representing the shift of virtual sound source toward the visual stimulus and the second, the relative length of time for which visual attention was focused on the visual stimulus. The obtained values of the correlation coefficient for all configurations are shown in Table 5.

According to the values of the correlation coefficients presented in Table 5 the relationship between the viewer's visual attention and the observed shift of the virtual sound source localization has been demonstrated for both samples of the first group. Nevertheless, this relationship was presented in the first session of the conducted experiments. The results of the second test session (employing the Cyber-Eye system) are not representative because the same group of subjects participated in both experiments. Therefore, we can conclude that knowing the content of the research material has a negative influence on the reliability of the obtained results.

2.2.3 Scalability of the Image Proximity Effect

Samples of the first group were employed to investigate the impact of screen size on the observed image proximity effect. This aspect is associated with scalability of the image proximity effect and corresponds to checking whether the visual angle of the presented object determines the observed

Table 5. The Spearman's rank correlation coefficient determined for both sessions employing two eye-gaze tracking systems

Sample ID	Tobii T60	Cyber-Eye
1_{BA}	0.87	0.42
2_{BA}	0.75	0.31

Table 6. Results of the ANOVA test for subjective assessments of virtual sound source localization depending on screen size

	small + medium		medium + large	
Sample ID	F	p	F	p
1_{BA}	1.16	0.29	0.63	0.43
2_{BA}	1.06	0.31	Kruskal-Wallis: $H = 0.34$	0.56

Table 7. Results for the viewers' visual attention before the appearance of a visual stimulus in the frame for both eye-gaze tracking systems

	Tobii T60		Cyber-Eye	
Sample ID	N	m_{VA} [%]	N	m_{VA} [%]
1_{AA}	5	13	10	67
2_{AA}	14	29	5	34

where:
N – the number of subjects who focused their gaze in the ROI before the appearance of an object associated with the sound effect heard in the frame (the studied group consisted of 15 participants)
m_{VA} – the mean value of the relative duration of concentration of gaze in the ROI for all subjects who focused their visual attention on the ROI

virtual sound source shift. Table 6 includes the results of this investigation.

The p-values obtained for both stages of this examination (dependent on screen size: small + medium, medium + large) for each sample are greater than 0.05. In this case we could not reject the null hypothesis of the ANOVA test. Consequently, the mean values of subjective evaluations of virtual sound source location when displaying the video content on different screen sizes are homogeneous for each analyzed sample. In this context, we confirmed the scalability of the image proximity effect. This means that the virtual sound source shift is observed for the same visual angle of the presented object regardless of the screen size.

2.3 Results for Additional Aspect Stimuli

Within the second experiment we investigated the additional aspect related to the influence of the sound effect on the viewer's direction of gaze. According to the assumption of this experiment we observed the direction of gaze of each subject at the beginning of the sample. This means that we analyzed the subject's visual attention before the appearance of an object referring to the eye-catching visual stimulus (the so-called ROI) in the frame. Table 7 includes the results obtained from the examination of the viewer's visual attention with both eye-gaze tracking systems.

It is worth noting that reliable results were obtained only for the first session of this experiment. The nature of the test required the subjects to be unacquainted with the sample content. Therefore, although the results of the second session of the experiment conducted using the Cyber-Eye system are quite promising, they should be omitted in the formulation of conclusions for this study. According to the values included in the third column of Table 7, the subjects' visual attention to the ROI was relatively low. Moreover, in the case of sample 2_{AA} only five of 15 respondents looked toward the expected ROI. These observations allowed us to conclude that the sound effect heard before the appearance of the object that was the source of this sound in the frame has an insignificant impact on the direction of the viewer's gaze.

2.4 Results for 3-D Content Stimuli

The third experiment consisted in studying the relation between the observed shift of the virtual sound source and the position of the 3-D visual stimulus in the stereoscopic depth. First, we determined the virtual sound source shift based on its location indicated after the unimodal stimulation to the location designated after presentation of the audio-visual sample. According to the statistical analysis the observed shift of the virtual sound source in the direction of the 3-D object is statistically significant for both samples. For sample $3D_P$ the value of the F test was equal to 25.62 with $p = 0.000023$. This means that the image proximity effect occurs for 3-D content included in the positive parallax area. Also, the results for sample $3D_N$ indicate that the observed shift of the virtual sound source is statistically significant ($F = 12.54$, $p = 0.0014$). Subsequently, we decided to check whether the observed shift of the virtual sound source is correlated with the type of stereoscopic parallax. We determined the mean value (m) and S.D. for all subjects' evaluations in the case of each tested configuration. It should be noted that the sound source was localized in the left part of the stereo basis (–25°). According to the data analysis performed, the mean of assessments indicating the location of the virtual sound source was equal to –18.53°, with S.D. = 5.84. Means determined for samples with the accompanying 3-D video demonstrated some differences ($3D_P$: $m = -8.47$, S.D. = 7.19; $3D_N$: $m = -5.6$, S.D. = 5.47). In order to verify the statistical significance of this discrepancy we performed the ANOVA test. The obtained results did not confirm the expected correlation ($F = 1.41$, $p = 0.245$). The observed shift of the virtual sound source is greater when the 3-D object is closer to the viewer (negative parallax) than when the object is perceived behind the screen (positive parallax). However, this relationship is statistically insignificant.

Investigation of the impact of 3-D video content on the image proximity effect remains an open research area. We plan to continue our research in this context.

Based on the observation of the multimodal perception presented in this paper, as well as in our previous publication [41], we may propose a simple expression for modeling

the localization of virtual sound source (L) as a function of audiovisual stimulus characteristics according to (1).

$$L = f\begin{pmatrix} position\_of\_visual\_stimulus, \\ character\_of\_visual\_stimulus \\ viewer's\_visual\_attention, \\ content\_of\_audio\_stimulus \end{pmatrix} \quad (1)$$

3 DISCUSSION

Within this article we presented a research study on two interesting aspects of the image proximity effect. The first aspect was related to the influence of the screen size on the observed shift of the virtual sound source. With regard to the second aspect we studied the relationship between the observed image proximity effect and the location of the 3-D object in the stereoscopic depth. In addition, we carried out an experiment in which we investigated the impact of the sound effect on the direction of the viewer's gaze. It is worth mentioning that eye-gaze tracking systems were employed in all experiments conducted to make the obtained results more objective.

The image proximity effect was observed in the first experiment conducted. We studied this effect by exploiting two different visual stimuli: the face of a talking character (1_{BA}) and a musical instrument (violin, 2_{BA}). According to the obtained results, the virtual sound source shift is unquestionable for sample 1_{BA}. In the case of the second sample the image proximity effect was noted, but it achieved a statistical significance for only one session of the experiment. These observations enable us to consider that the observed shift of the virtual sound source depends on the video content and the type of "eye-catching" (ROI) visual stimulus. When the viewer is focused on the character's speech, then he or she localizes the voice closer to the character's face. This means that the musical instrument attracted the subjects' attention to a lesser extent.

An important aspect of the research conducted was to investigate the image proximity effect for the presentation of video content on small, medium, and large size screens. The results obtained provide clear evidence that the visual angle of the presented object determines the image proximity effect regardless of screen size. The observed scalability of the image proximity effect was demonstrated for two types of audio-visual stimulus: the character's face and the musical instrument (the violin).

As mentioned above, an additional aspect of our studies was to check the relation between the sound effect and the viewer's visual attention. We can conclude that for the tested samples the sound effect heard before the appearance of the sound source in the frame does not influence the direction of the viewer's gaze. Nevertheless, research on this aspect should be continued in the future. We believe that spatial sound (reproduced by the multichannel sound system) may affect the viewer's visual attention to a greater extent.

In the third experiment we found a difference in the observed image proximity effect for 3-D video content presented in positive and negative parallax areas. The observed shift of the virtual sound source is greater when the 3-D object is closer to the viewer than when the object is perceived behind the screen. Nonetheless, this relationship is not statistically significant ($F = 1.41$, $p = 0.245$). It is worth mentioning that our observations were supported by the findings of our student, who performed similar tests [42]. He noticed that 3-D objects of positive or zero parallax are dominated by the 3-D object of negative parallax in attracting the viewer's attention. When two objects are included in the 3-D scene, the viewer's attention is more attracted by the object of negative parallax. However, this domination is observed during the first few seconds. Therefore, this context should be researched further in audio-visual correlation studies to verify our observations.

4 FINAL REMARKS

We employed two eye-gaze tracking systems in the first two experiments: the commercial system and the Cyber-Eye system developed at the Multimedia Systems Department of the Gdansk University of Technology. It should be stressed that we did not notice any significant difference between the systems in the context of the experiments carried out. However, Cyber-Eye is provided with better functionalities than the commercial eye-gaze tracker in the context of audio-visual correlations research. It allows a stereoscopic video to be displayed using any 3-D technique and is compatible with both the stereo and surround-sound systems. In addition, Cyber-Eye allows the viewer to be observed when the video content is displayed on a large projector screen.

It is worth noting that the interpretation of results obtained within different researches related to audio-visual correlation is relatively complex. Moreover, evaluation of the sound quality may be problematic because it concerns many aspects of assessment [46–48]. However, the use of eye-gaze tracking technology in such experiments supports the analysis and makes this research more objective.

5 ACKNOWLEDGMENTS

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 Center for Research and Development (NCBiR, Poland).

We would like to thank Krzysztof Głośnicki for the input he gave to our study by sharing the results of experiments conducted within his M.Sc. thesis, which supported our observations and conclusions.

6 REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, vol. 264, no. 5588, pp. 746–748 (1976).

[2] J. G. Beerends and F. E. De Caluwe, "The Influence of Video Quality on Perceived Audio Quality and Vice Versa," *J. Audio Eng. Soc.*, vol. 47, pp. 355–362 (1999 May).

[3] P. Bertelson and M. Radeau, "Cross-Modal Bias and Perceptual Fusion with Auditory-Visual Spatial Discordance," *Percept. Psychophys.*, vol. 29, no. 6, pp. 578–584 (1981).

[4] P. Bertelson, J. Vroomen, B. de Gelder, and J. Driver, "The Ventriloquist Effect Does Not Depend on the Direction of Deliberate Visual Attention," *Percept. Psychophys.*, vol. 62, no. 2, pp. 321–332 (2000).

[5] E. T. Davis, K. Scott, J. Pair, L. F. Hodges, and J. Oliverio, "Can Audio Enhance Visual Perception and Performance in a Virtual Environment?" *Proc. of 43rd Human Factors and Ergonomics Society Annual Meeting* (1999).

[6] I. Frissen, J. Vroomen, B. de Gelder, and P. Bertelson, "The Aftereffects of Ventriloquism: Generalization across Sound-Frequencies," *Acta Psychologica*, vol. 118, no. 1–2, pp. 93–100 (2004).

[7] M. P. Hollier, A. N. Rimell, D. S. Hands, and R. M. Voelcker, "Multi-Modal Perception," *BT Technology J.*, vol. 17, no. 1, pp. 35–46 (1999).

[8] S. Morein-Zamir, S. Soto-Faraco, and A. Kingstone, "Auditory Capture of Vision: Examining Temporal Ventriloquism," *Cognitive Brain Research*, vol. 17, pp. 154–163 (2003).

[9] M. Radeau and P. Bertelson, "Adaptation to Auditory-Visual Discordance and Ventriloquism in Semirealistic Situations," *Percept. Psychophys.*, vol. 22, no. 2, pp. 137–146 (1977).

[10] P. Bertelson, "Starting from the Ventriloquist: The Perception of Multimodal Event," *Advances in Psychological Science*, M. R. M. Sabourin, F. I. M. Craik, Ed. (Psychology Press, 1998), pp. 419–439.

[11] M. B. Gardner, "Proximity Image Effect in Sound Localization," *J. Acous. Soc. Am.*, vol. 43, no. 1, p. 163 (1968).

[12] S. Komiyama, "Subjective Evaluation of Angular Displacement between Picture and Sound Directions for HDTV Sound Systems," *J. Audio Eng. Soc.*, vol. 37, pp. 210–214 (1989 Apr.).

[13] H. A. Witkin, S. Wapner, and T. Leventhal, "Sound Localization with Conflicting Visual and Auditory Cues," *J. Experimental Psych.*, vol. 43, no. 1, pp. 58–67 (1952).

[14] V. Harrar and L. R. Harris, "Eye Position Affects the Perceived Location of Touch," *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation cérébrale*, vol. 198, no. 2–3, pp. 403–10 (Sep. 2009).

[15] S. Merchel, M. E. Altinsoy, and M. Stamm, "Touch the Sound: Audio-Driven Tactile Feedback for Audio Mixing Applications," *J. Audio Eng. Soc.*, vol. 60, pp. 47–53 (2012 Jan./Feb.).

[16] F. Rumsey, "Audio in Multimodal Applications," *J. Audio Eng. Soc.*, vol. 58, pp. 191–195 (2010 Mar.).

[17] M. E. Altinsoy, "The Quality of Auditory-Tactile Virtual Environments," *J. Audio Eng. Soc.*, vol. 60, pp. 38–46 (2012 Jan./Feb.).

[18] Y. Nakayama, K. Watanabe, S. Komiyama, F. Okano, and Y. Izumi, "A Method of 3-D Sound Image Localization Using Loudspeaker Arrays," presented at the *114th Convention of the Audio Engineering Society* (2003 Mar.), convention paper 5793.

[19] C. E. Vegiris, K. A. Avdelidis, C. A. Dimoulas, and G. V. Papanikolaou, "Live Broadcasting of High Definition Audiovisual Content Using HDTV over Broadband IP Networks," *Intl. J. Digital Multimedia Broadcasting*, vol. 2008, pp. 1–18 (2008).

[20] P. Bertelson, "Automatic Visual bias of Perceived Auditory Location," *Psychonomic Bulletin and Review*, vol. 5, no. 3, pp. 482–489 (1998).

[21] D. Alais and D. Burr, "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration," *Current Biology?: CB*, vol. 14, no. 3, pp. 257–62 (Feb. 2004).

[22] P. Bertelson, "Ventriloquism, Sensory Interaction, and Response Bias: Remarks on the Paper by Choe, Welch, Gilford, and Juola," *Attention, Perception & Psychophysics*, vol. 19, no. 6, pp. 531–535 (1976).

[23] J. Vroomen, P. Bertelson, and B. de Gelder, "The Ventriloquist Effect Does Not Depend on the Direction of Automatic Visual Attention," *Percept. Psychophys.*, vol. 63, no. 4, pp. 651–9 (May 2001).

[24] J. Vroomen, "Perceptual Effects of Cross-Modal Stimulation: Ventriloquism and the Freezing Phenomenon," pp. 1–23, in *Handbook of multisensory processes*, Cambridge, 2004.

[25] P. Bertelson and M. Radeau, "Cross-Modal Bias and Perceptual Fusion with Auditory-Visual Spatial Discordance," *Percept. Psychophys.*, vol. 29, no. 6, pp. 578–84 (Jun. 1981).

[26] B. Kostek, "Perception-Based Data Processing in Acoustics," in *Studies in Computational Intelligence*, J. Kacprzyk, Ed. (Berlin : Springer, 2005), pp. 389–400.

[27] H. Wittek, F. Rumsey, and G. Theile, "Perceptual Enhancement of Wavefield Synthesis by Stereophonic Means," *J. Audio Eng. Soc.*, vol. 55, pp. 723–751 (2007 Sep.).

[28] M. Cobos and J. J. Lopez, "Resynthesis of Sound Scenes on Wave-Field Synthesis from Stereo Mixtures Using Sound Source Separation Algorithms," *J. Audio Eng. Soc.*, vol. 57, pp. 91–110 (2009 Mar.).

[29] F. E. Toole, "Loudspeakers and Rooms for Sound Reproduction—A Scientific Review," *J. Audio Eng. Soc.*, vol. 54, pp. 451–476 (2006 Jun.).

[30] C. Dimoulas, K. Avdelidis, G. Kalliris, and G. Papanikolaou, "Sound Source Localization and B-Format Enhancement Using Sound Field Microphone Sets," presented at the *122*nd Convention of the Audio Engineering Society (2007 May), convention paper 7091.

[31] C. Dimoulas, G. Kalliris, K. Avdelidis, and G. Papanikolaou, "Improved Localization of Sound Sources Using Multi-Band Processing of Ambisonic Components," presented at the *126th Convention of the Audio Engineering Society* (2009 May), convention paper 7691.

[32] A. Roginska, "Effect of Spatial Location and Presentation Rate on the Reaction to Auditory Displays," *J. Audio Eng. Soc.*, vol. 60, pp. 497–504 (2012 Jul./Aug.).

[33] J. Herre, C. Falch, D. Mahane, G. Del Galdo, M. Kallinger, and O. Thiergart, "Interactive Teleconferencing Combining Spatial Audio Object Coding and DirAC Technology," *J. Audio Eng. Soc*., vol. 59, pp. 924–935 (2011 Dec.).

[34] K. Hamasaki, T. Nishiguchi, K. Hiyama, and K. Ono, "Advanced Multichannel Audio Systems with Superior Impression of Presence and Reality," presented at the *116th Convention of the Audio Engineering Society* (2004 May), convention paper 6053.

[35] K. Hamasaki, K. Hiyama, T. Nishiguchi, and R. Okumura, "Effectiveness of Height Information for Reproducing the Presence and Reality in Multichannel Audio System," presented at the *120th Convention of the Audio Engineering Society* (2006 May), convention paper 6679.

[36] S. Bech, V. Hansen, and W. Woszczyk, "Interaction between Audio-Visual Factors in a Home Theater System: Experimental Results," in *Perception and Subjective Evaluation* (1995), Paper No. 4096.

[37] M. Emoto, K. Masaoka, M. Sugawara, and F. Okano, "Viewing Angle Effects from Wide Field Video Projection Images on the Human Equilibrium," *Displays*, vol. 26, no. 1, pp. 9–14 (2005).

[38] B. Kunka and B. Kostek, "A New Method of Audio-Visual Correlation Analysis," in *Computer Science and Information*, 2009, pp. 497–502.

[39] B. Kunka and B. Kostek, "Exploiting Audio-Visual Correlation by Means of Gaze Tracking," *International J. Computer Science and Applications*, vol. 7, no. 3, pp. 104–123 (2010).

[40] B. Kunka and B. Kostek, "Objectivization of Audio-Video Correlation Assessment Experiments," presented at the *128th Convention of the Audio Engineering Society* (2010May), convention paper 8148.

[41] B. Kunka and B. Kostek, "Objectivization of Audio-Visual Correlation Analysis," *Archives of Acoustics*, vol. 37, no. 1, pp. 63–72 (2012).

[42] K. Głośnicki, "Realization of 3-D Audio-Video Recordings' Database Dedicated to Audio-Visual Correlations Research (in Polish)," Gdansk University of Technology, 2011.

[43] "Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems, " ITU-R BS.1116-1.

[44] B. Mendiburu, *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen* (Focal Press, 2009).

[45] W. J. Conover, *Practical Nonparametric Statistics* (John Wiley & Sons, 1999).

[46] J. Blauert and U. Jekosch, "A Layer Model of Sound Quality," *J. Audio Eng. Soc*., vol. 60, pp. 4–12 (2012 Jan./Feb.).

[47] F. Rumsey, "Sound Quality Evaluation," *J. Audio Eng. Soc*., vol. 58, pp. 853–858 (2010 Oct.).

[48] F. Rumsey, "Hear, Hear! Psychoacoustics and Subjective Evaluation," *J. Audio Eng. Soc*., vol. 59, pp. 758–763 (2011 Oct.).

THE AUTHORS

Bartosz Kunka

Bozena Kostek

Bartosz Kunka received his M.Sc. degree in 2007 from the Faculty of Electronics, Telecommunications and Informatics, Technical University of Gdansk. His thesis was related to stereoscopic imaging, particularly concerning anaglyph movie. In 2012 he completed his Ph.D. dissertation entitled "*The Eye-Gaze Tracking System Supporting Audio-Visual Correlations Research*."

His scientific interests are associated with image processing, video recording and editing, and telemedicine. His research focuses on of eye-gaze tracking systems to various applications. His current work concentrates on a method for consciousness level evaluation of post-comatose patients based on gaze direction determination.

●

Bozena Kostek holds professorship at the Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology (GUT), Poland. She is now Head of the Audio Acoustics Laboratory. She is a Fellow of the Audio Engineering Society. She received her M.Sc. degree in sound engineering from the Technical University of Gdansk (1983) and her second M.Sc. in organization and management in 1986. In 1992, she supported her Ph.D. degree with honors from the Technical University of Gdansk. In March 2000 she supported her D.Sc. degree at the Institute of Research Systems of the Polish Academy of Sciences in Warsaw. In 2005 she was granted the title of professor from the President of Poland.

Her research activities are interdisciplinary, however the main research interests focus on cognitive bases of hearing and vision, music information retrieval, musical acoustics, studio technology, quality-of-experience, human-computer-interaction (HCI) as well as applications of soft computing and computational intelligence to the mentioned domains. She has published more than 450 scientific papers, three books, and dozens of book chapters.

In 1991 she helped to form the Polish Section of the Audio Engineering Society and since then has served as a member of its Committee. In 2003, 2005, and 2009 she was elected Vice-President of the Audio Engineering Society for Central Europe, and in 2007 and 2011 she was elected as Governor of the AES. She is now Editor-in-Chief of *JAES*.