Research Paper

Pursuing Listeners' Perceptual Response in Audio-Visual Interactions – Headphones vs Loudspeakers: A Case Study

Bartłomiej MRÓZ[(1),(2)], Bożena KOSTEK[(2)*]

[(1)] *Multimedia Systems Department*
Gdansk, Poland

[(2)] *Audio Acoustics Laboratory*
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Gdansk, Poland
[*]Corresponding Author e-mail: bokostek@audioacoustics.org

(*received May 26, 2021; accepted December 29, 2021*)

This study investigates listeners' perceptual responses in audio-visual interactions concerning binaural spatial audio. Audio stimuli are coupled with or without visual cues to the listeners. The subjective test participants are tasked to indicate the direction of the incoming sound while listening to the audio stimulus via loudspeakers or headphones with the head-related transfer function (HRTF) plugin. First, the methodology assumptions and the experimental setup are described to the participants. Then, the results are presented and analysed using statistical methods. The results indicate that the headphone trials showed much higher perceptual ambiguity for the listeners than when the sound is delivered via loudspeakers. The influence of the visual modality dominates the audio-visual evaluation when loudspeaker playback is employed. Moreover, when the visual stimulus is present, the headphone playback pattern of behavior is not always in response to the loudspeaker playback.

Keywords: human perception; audio-visual interaction; 3D perception; binaural spatial audio.

1. Introduction

The proposed study explores audio-visual interactions in relation to binaural spatial audio. Therefore, the presented study aimed to compare the perceptual auditory experience between the loudspeaker and headphone playback while interacting with visual stimuli. This is made between loudspeakers and headphones combined with the head-related transfer function (HRTF) plugin playback with/without coupling a visual stimulus. It should be mentioned that various aspects of HRTF are of high interest to researchers, such as applying principal component analysis (PCA) to modelling the magnitude of the HRTFs (Ramos, Tommasini, 2014) and automatic measurement that allows the measurement of HRTF with high spatial resolution within a considerably short time (Pruchnicki, Plaskota, 2008). Storek *et al.* (2016) show analyses of the differential head-related transfer function (DHRTF) performance, an alternative transfer function for headphone-based virtual sound source positioning within a horizontal plane. The authors of this paper used this method to reduce processing and avoid timbre affection while preserving signal features important for sound localisation (Storek *et al.*, 2016). Additionally, it should be remembered that as per the research on audio-visual correlation, individual HRTFs help to provide a better auralisation experience in virtual reality systems (Vorländer, 2014; 2020). With regard to our experiments, the work by Yao *et al.* (2017) is of importance, as it refers to an increasing number of binaural systems embedded with HRTFs, so listeners can experience virtual environments via conventional stereo loudspeakers or headphones. This is a state-of-the-art, as it uses multi-layer feed-forward neural network to estimate the goodness of fit of each

HRTF dataset for a subject. Finally, it is vital to indicate that studies on HRTFs employ subjective tests, checking research assumptions.

Since our study should be regarded from the perception perspective, so a short overview of the research carried out in that area is presented. Human perception, in which audio and visual stimuli are the most dominant and interrelated, is multisensory or cross-modal (Blauert, Braasch, 2020; Bizley *et al.*, 2016; Woodcock *et al.*, 2019). Bizley *et al.* (2016) defined cross-modal integration as a term applied to many phenomena in which one sensory modality influences the task performance or perception in another sensory modality. This means that the human brain integrates incoming signals to form a perceived cohesion of the external world based on spatial and temporal cues (Blauert, Braasch, 2020; Chiou, Rich, 2012). However, as suggested by Regan and Spekreijse (1977), rather than the time-locking of physiological signals, the correspondence between the perceived auditory space and the perceived visual space may be influenced by subjective criteria. Contrarily, Sorati and Dawn (2021) referred to studies in which interactions between auditory and visual perception may be detected by observing early auditory event-related potentials (ERPs). Visual information received prior to the onset of the corresponding acoustic event can provide visual cues and predict the upcoming auditory sound. Such a phenomenon is at the forefront of audio-visual (AV) interaction.

Another notion related to cross-modal perception is auditory-visual integration, which was thoroughly researched by Ecker and Heller (2005). They performed two experiments to measure the combined perceptual effect of visual and auditory information on the perception of the trajectory of a moving object. Additionally, they observed that the sound condition influenced whether observers were more likely to perceive the object as rolling back in-depth on the box floor or jumping in the frontal plane. Moreover, they reported that the object speed was an indirect measure of the perceived path because, as a result of the geometry of the box and the viewing angle, a rolling object (a ball) would travel a greater distance than a jumping ball in the same time interval.

Chiou and Rich (2012) suggested that cross-modal correspondence (e.g. high-pitched sounds associated with bright, small objects located high up) affect multisensory integration. They focused on the association between auditory pitch and spatial location. In particular, they focused on how cross-modal mapping affected the allocation of attention with an attentional cueing paradigm.

This aspect was also pursued by Walker *et al.* (2012); they suggested that cross-modality is an effect of bidirectional cross-activation between dimensions of connotative meaning. This was researched in another context by Ildirar *et al.* (2017): perceptual integration may be more complex and challenging in unfamiliar environments.

Despite the cross-modality and interactions in perception, Woodcock *et al.* (2019) indicated that audio technology is often researched and evaluated in isolation from the visual component. Furthermore, there may be a problem of spatial audio-visual coherence (i.e. whether audio/visual signals arrive from the same direction, especially in stereoscopic 3D movies, when the dialog is reproduced on the central loudspeaker without any regard to the visual position on the screen) (Hendrickx *et al.*, 2015). The outcomes of these experiments indicated an improvement in audio experience when coherence in azimuth was achieved (Hendrickx *et al.*, 2015). However, Pike and Stenzel (2016) concentrated on direct measures (e.g. preference and annoyance) and indirect measures (biosignals and reaction times) to determine how viewers perceive audio and audio-visual attributes concerning spatial coherence testing. This aspect of audio-visual coherence was also investigated by Stenzel *et al.* (2017; 2019) and Stenzel and Jackson (2018); they conducted a simple forced-choice test with subsequent modelling of the psychometric function for evaluating audio-visual spatial coherence. Following this, they employed a psychometric function to semantically vary audio-visual stimuli. The consequent results indicated that the maximum accepted offset angle did not depend on semantic categories, but it was linked to the audio feature classes with harmonic sounds, which led to higher acceptable offsets. They pursued this topic by applying a two-alternative forced-choice (2AFC) using the reaction time (Stenzel, Jackson, 2018). In contrast to prior research, the results obtained by them suggested that, even for speech signals, small audio-visual offsets subconsciously influence spatial integration (Stenzel *et al.*, 2019). This research also identified a problematic issue that subjective tests and standardised scales may not always be adequate for evaluating perceptual response to audio and video objects presented simultaneously (Stenzel *et al.*, 2019).

One of the primary outcomes of the audio-visual relationship is a phenomenon called *ventriloquism*, corresponding to the shift of a virtual (phantom) sound source toward the visual stimulus (Alais, Burr, 2004; Bertelson, 1998; Bertelson, Aschersleben, 1998; Vroomen, De Gelder, 2004). This phenomenon was meticulously investigated by many researchers (Frissen *et al.*, 2004; Kohlrausch, Par, 2005; Morein-Zamir *et al.*, 2003; Radeau, Bertelson, 1977; Vroomen *et al.*, 2001). It is regularly experienced while watching movies or playing video games: the voices seem to propagate from other objects rather than from the actual sources of the sound. Recently, the ventriloquism phenomenon became *digital ventriloquism*, thereby allowing smart

speakers to render sound onto everyday objects. This way, the voice agent is not assigned to the given device when it should be contextually and spatially emanated elsewhere (Iravantchi *et al.*, 2020). This phenomenon is interesting, especially when considering binaural audio and the ongoing research on spatial localisation using the head-related transfer (HRT) functions. The bimodal localisation's precision is usually better than the visual or the unimodal auditory presentation; however, most spatial audio studies exclude visual cues in the experiments (Komiyama, 1989). There is also an "image proximity effect"; it refers to how vision affects sound source localisation on a stereo basis (Gardner, 1968; Komiyama, 1989; Kunka, Kostek, 2012; 2013). Nevertheless, it should be remembered that the presentation of a binaural signal requires a head-related impulse response (HRIR) and/or a binaural room response compensation for the headphone response (Komiyama, 1989).

The paper is organised as follows. Section 2 presents the experimental setup. In Sec. 3, a study procedure underlying the experiment is provided. Results and statistical analysis performed along with the Dunn-Šidák post hoc method are presented in Sec. 4. This is followed by the discussion, concluding remarks, and further research plans contained in Sec. 5.

2. Experimental setup

To answer the question regarding the extent of the way of listening to the audio stimulus with/without accompanying visual cues (i.e. pulsing squares) changes, we followed the experimental procedures that originated in earlier studies. Thus, the proposed method involves examining audio-visual interaction but focuses on how a visual object influences sound perception when listening to the audio stimulus over headphones or coming from the loudspeaker. The overall block diagram of the test scenarios is presented in Fig. 1.

The research procedure consisted of a loudspeaker setup and a headphone setup, with the display screen placed in front of the listener (see Fig. 2). The distance between the listener and the screen was 0.5 m while the loudspeakers were set at the 2 m distance. The loudspeakers were arranged in stereo pairs in the following order: ±30°, ±45°, ±60°, ±75°, ±90°. The first pair, ±30°, was placed following the ITU-R recommendation for stereo speaker placement. The number of speaker pairs and their spread along the listening area allowed for creating a wider stereo image. This is especially important in binaurally rendered Ambisonics, where the localisability of sound sources is much less restrained. The research was carried out in an acoustically treated shoebox-shaped recording room, in which the appropriate acoustic conditions were maintained. There is a window between the control room and the recording room. Hence, a sound-absorbing panel was positioned

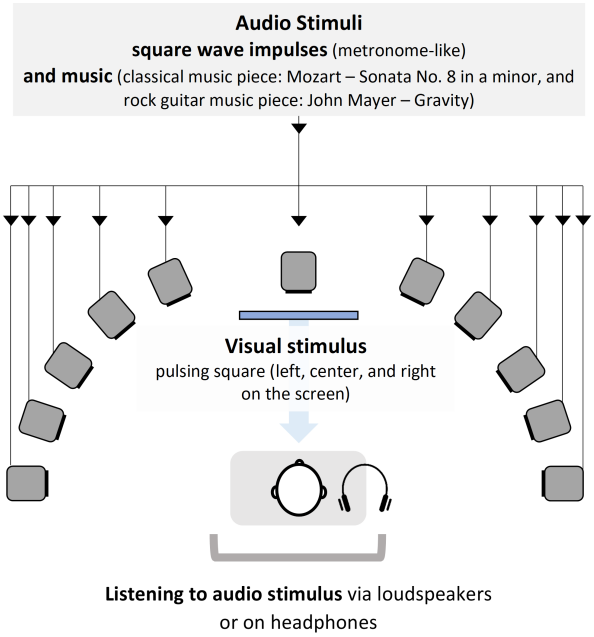

Fig. 1. Block diagram of the methodology.

Fig. 2. Experimental setup.

behind the listener in order to avoid reflections from a glass window. The equipment utilised was as follows: Genelec 8020D loudspeakers, Sennheiser closed type headphones, hi-speed USB interface, head-tracking device (Romanov *et al.*, 2017).

The loudspeaker setup comprised five stereo loudspeaker pairs and an additional central loudspeaker in front of a listener. The image source position was achieved by panning the stereo signal. The virtual sound source positions in degrees of azimuth were: [–56, –34, 0, 22.5, 34, 38, 56]. Not all of the stereo pairs were used – some loudspeakers (e.g. the central loudspeaker) were added to obscure the visual cues eventually, so that the participant's answer would not be suggested. Additionally, there was other equipment present in the room, including loudspeakers. We have

Fig. 3. Example of plugins setup for ±30° virtual loudspeaker setup.

decided not to remove the equipment in order to obfuscate the visual cues even further.

The headphone setup consisted of a pair of closed type headphones with a head-tracking device for dynamic scene rotation. The HRTFs were provided with the KU 100 HRIR set (Bernschütz, 2013), embedded into the IEM Plug-in Suite (n.d.) provided by the Institute of Electronic Music and Acoustics in Graz. This *BinauralDecoder* plugin employs the magnitude least-squares method for calculating the filters (Zaunschirm *et al.*, 2018; Zotter, Frank, 2018). There was no headphone equalisation nor was HRTF equalisation applied. The sound level calibration between loudspeakers and headphones was conducted subjectively by ear.

The auralisation of the sound scene was achieved with the *RoomEncoder* plugin. Thus, the loudspeaker setup was reproduced in the Ambisonic domain. The RoomEncoder plugin also allows reproducing the room acoustics, so the conditions from the loudspeaker trials were maintained. Then, the binaural plugin with head-tracking was used to provide compensation of dynamic head rotation by rotating the Ambisonic sound scene, which was rendered to binaural stereo signals.

An example of the plugin setup for ±30° virtual loudspeaker setup is shown in Fig. 3.

3. Audio-visual experiment

The participants of the experiments included 10 males aged in the range of 20 to 50 years, experienced in auditory experiments. None of the participants have hearing impairments. All participants were informed about the study merits and provided informed verbal consent to take part in the tests.

The participant's task was to fill in a paper questionnaire (see Fig. 4, with an excerpt of this form), where numbered circles were printed. This was to determine the perceived direction and then write down the result on a printed circle along with the trial number. Two different audio signals – square wave (40 Hz) impulses (metronome-like, 60 bpm) and music (classical music piece: Mozart – *Sonata No. 8 in a minor*, and rock-guitar music piece: John Mayer – *Gravity*) were presented to the participants. The pulsing square was displayed on different sides of the display: centre, left, and right. This would give roughly −5°, 0°, and +5° of azimuth, respectively. The video excerpt from a concert used in the experiment was static (i.e. the camera was not moving).

Furthermore, trials were performed with and without visual stimuli; metronome-like impulses were presented with the pulsing square in three configurations:

Participant __no._____

Auxiliary question 1

Do you have a visual impairment?

Auxiliary question 2

Do you have a hearing impairment?

Auxiliary question 3

What music do you like to listen? (generally mention species):

Leading question:

Specify the direction from which you think the sound of the instrument comes from. You can, to improve the perception of the direction, close your eyes or slightly twist your head, etc. Take the direction you want to mark in the picture.

ATTENTION! The print is on both sides!

LOUDSPEAKERS/HEADPHONES

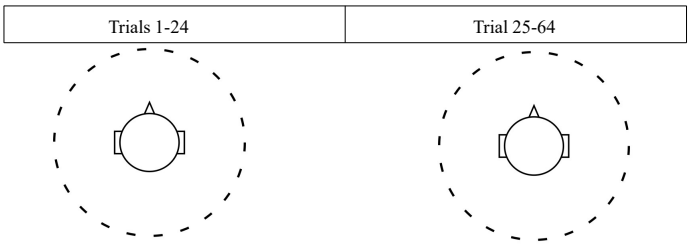

Fig. 4. Questionnaire form (an excerpt) prepared for the tests designed (trials 1 to 24 performed on headphones, trials 25 to 64 employed headphones).

left, centre, and right of the screen. The visual stimuli were displayed at the same tempo (60 bpm) as the audio stimuli. They were either matching or not matching the presented audio stimuli with regard to the direction of the sound source. In the matching case, the sound was coming from the same side as the square was presented on the screen. In the non-matching case, the sound was coming from the opposite side rather than the side of the square shown on the screen.

All audio signals were presented over loudspeakers and then repeated on the headphone playback. The order of stimuli presentation was randomised. However, the visual stimuli were separated from the non-visual; also, the metronome-like pulse stimuli were not mixed with the music stimuli.

In total, the following conditions/variables were changed throughout the experiments performed:

- Presence (or lack) of visual stimuli (present or not);
- Alignment of visual stimuli to the audio stimuli (aligned or not);
- Playback system (loudspeaker of binauralised stereo over headphones);
- Audio-visual stimuli (metronome-like ticks or music pieces).

From these variables/conditions, a total of 62 stimuli/conditions were tested. For loudspeaker playback, there were 10 conditions for audio-only and 14 conditions for audio-visual stimuli. For headphone playback, there were 10 conditions for audio-only and 28 conditions for audio-visual stimuli. In general, fewer conditions were tested for non-visual stimuli since the alignment was not measurable without one of the stimuli present. Additionally, fewer conditions were tested for loudspeaker playback, as the study presented is focused on headphone playback.

Overall, the subjective test participants' task was to fill in the questionnaire form (as presented in Fig. 4) while listening to an audio stimulus (either via headphones or loudspeakers) along with or without the presence of visual stimuli. In addition, they answered several auxiliary questions about having any kind of hearing impairment and music genres they prefer listening to.

4. Results and analyses

After the participants completed their tasks, the results were collected and analysed. The collection of the results was done using angular graduation, which allowed for precise reading of the collected answers. Further on, the differences between the presented stimuli angles and the perceived angles (i.e. incorrect answers given by participants) were calculated. We used a residual error of answers (r_i), which was calculated with the following formula:

$$r_i = \left(X_i - \overline{X}\right)^2, \tag{1}$$

where X_i refers to the observed value, and $\overline{X}$ denotes the predicted value.

The residual errors were used in further statistical analysis. The Kruskal-Wallis method, which is suitable for the non-parametric data, was chosen as the omnibus test. It should be noted that the Kruskal-Wallis test indicates that among several groups of results, at least two are different. To determine which groups differ from each other, a post hoc test was also performed.

The input data were statistically tested in three different comparisons. The first one was comparing stereo pairs of loudspeakers (for headphone cases – virtualised ones). The second test was the visual cues (and the alignment between visual and audio stimuli). The third test case was the comparison of the audio signal type.

In three omnibus tests, the resulting p-values in each case were less than 10^{-3}, stating significant differences between the groups. Therefore, post hoc analyses were performed with the Dunn-Šidák pairwise comparison test. The resulting p-values are presented in Table 1, and the bold font marks p-values ≤0.05. Figure 5 shows the boxplots of residual errors, corresponding to Table 1. The data for the plot were filtered from outliers with the use of the Hampel function (Hampel, 1974) whose goal is to identify and filter outliers in a given series. Thus, it is suitable for the above mentioned task.

Table 1. p-values of the Dunn-Šidák post hoc test for loudspeaker pairs – residual errors of participants' errors (the following abbreviations stand for: LS – loudspeakers; HP – headphones; the bold values indicate p-values ≤0.05).

	Loudspeakers		Headphones			
	±30°	±60°	±30°	±45°	±60°	±75°
LS, ±30°	–	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
LS, ±60°	**<0.01**	–	**<0.01**	**<0.01**	0.45	**<0.01**
HP, ±30°	**<0.01**	**<0.01**	–	1.00	0.55	1.00
HP, ±45°	**<0.01**	**<0.01**	1.00	–	0.20	0.99
HP, ±60°	**<0.01**	0.45	0.55	0.20	–	0.86
HP, ±75°	**<0.01**	**<0.01**	1.00	0.99	0.86	–

Table 2. p-values of the Dunn-Šidák post hoc test for visual cues alignment – residual errors of participants' errors (the full forms of the abbreviations are as follows: VC_{off} – no visual cues; VC_{NA} – visual cues not aligned; VC_A – visual cues aligned. LS and HP refer to loudspeakers and headphones, respectively; the bold values indicate p-values ≤0.05).

	Loudspeakers			Headphones		
	VC_{off}	VC_{NA}	VC_A	VC_{off}	VC_{NA}	VC_A
LS, VC_{off}	–	**0.03**	0.99	**<0.01**	**<0.01**	**<0.01**
LS, VC_{NA}	**0.03**	–	0.35	**<0.01**	0.46	**<0.01**
LS, VC_A	0.99	0.35	–	**<0.01**	**<0.01**	**<0.01**
HP, VC_{off}	**<0.01**	**<0.01**	**<0.01**	–	0.19	1.00
HP, VC_{NA}	**<0.01**	0.46	**<0.01**	0.19	–	0.24
HP, VC_A	**<0.01**	**<0.01**	**<0.01**	1.00	0.24	–

Table 3. p-values of the Dunn-Šidák post hoc test for residual errors (the abbreviations stand for: LS – loudspeakers; HP – headphones; the bold values indicate p-values ≤0.05).

	Loudspeakers		Headphones	
	pulse	music	pulse	music
LS, pulse	–	0.91	**<0.01**	**<0.01**
LS, music	0.91	–	**<0.01**	**<0.01**
HP, pulse	**<0.01**	**<0.01**	–	**0.02**
HP, music	**<0.01**	**<0.01**	**0.02**	–

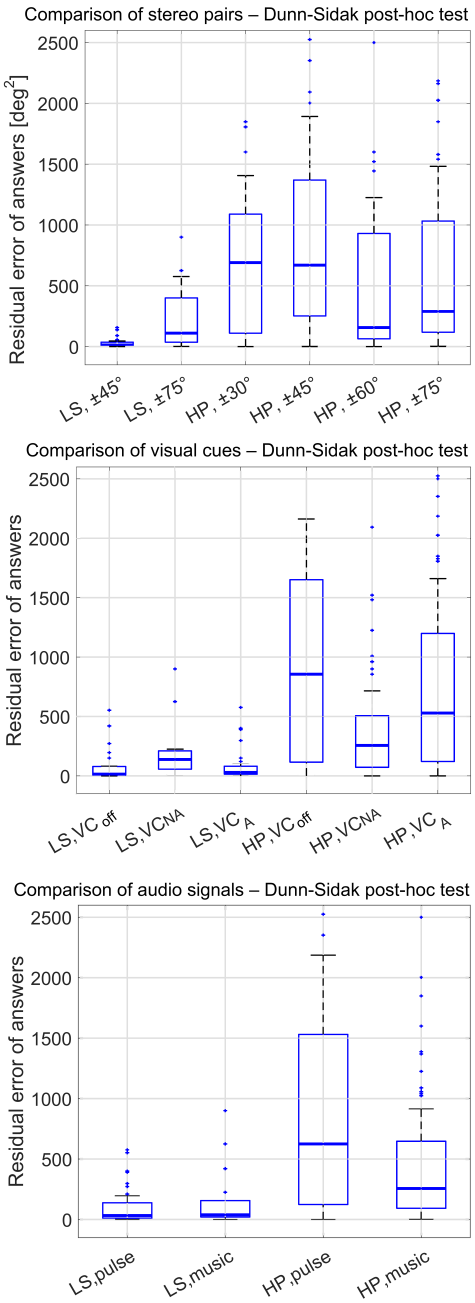

Fig. 5. Residual error of answers. The full forms of the abbreviations are as follows: VC_{off} – no visual cues; VC_{NA} – visual cues not aligned; VC_A – visual cues aligned; LS – loudspeakers; HP – headphones.

As shown in Fig. 5, all headphone trials render much higher residual errors than loudspeaker trials. In the first test, where loudspeaker pairs were compared, the wider spacing of loudspeaker pairs generates lower mean errors; however, they do not yield a statistically significant difference as this occurred in the headphone cases, as shown in Table 1. Additionally, the case of the ±60° stereo pair showed no significant difference regardless of whether loudspeakers or headphones were used. Furthermore, in the comparison of visual cues, it can be observed that the trials with the visual cues present render a lower mean error than in the case when headphones were used. However, only the case with VC_{NA} is not significantly different between the headphones and loudspeakers, suggesting that participants were equally confused in both test cases. It is also important to note that headphone cases are not significantly different among themselves. Only in the last test, where the presented audio signal was compared, the difference for headphone cases is significant, whereas for loudspeaker trials, there is no significant difference.

To present the aforementioned results, the residual error was used. The residual errors can be interpreted as the uncertainty of answers among participants; therefore, a lower residual error means lower uncertainty in evaluation – or higher confidence. In general, for loudspeaker cases, the misalignment of visual stimuli caused the lower confidence on answers, which was expected. Surprisingly, with headphone cases, quite the opposite happens. This is similar to the stereo pair width – for loudspeakers, increasing the width results in lower certainty, whereas for headphones, the higher width yields somewhat higher certainty in answers. In general, headphone trials present much lower confidence among participants' responses. It is also important to note that in the headphone scenarios, the artificial signal, which is a square wave pulse, renders much lower certainty than the music signal. This might be due to the unnatural sounding timbre of the stimulus signal of the headphone setup; for instance, the headphones were of the closed type, whereas open type headphones would be optimal.

Another probable reason behind this result might be the fact that the listeners – regardless of their expertise in acoustic listening tasks – are not too familiar with Ambisonic and binaural tasks, which also could skew their judgment. This may also explain why the loudspeaker examples were more easily recognised and judged more closely to the expected outcome. In the loudspeaker scenarios, visual cues were attracting the participants' focus to a predictable degree. We hypothesise that the headphone scenarios need improvements; for example, the signal heard should match the acoustics of the room where the test is performed, and the placement of the virtual sound sources needs some consideration. Another hypothesis might be that due to the proximity of headphone speakers (i.e. the perception of the listener's envelopment while using headphones), it might be that the visual stimulus affects the listener to a lesser degree, making a person more easily confused in the overall evaluation.

5. Summary

In this paper, a study of researching audio-visual interactions referring to how vision affects the localisation of a sound source on a stereo basis when listened to from loudspeakers and headphones was investigated. More specifically, several test scenarios were proposed, in which various audio signals were presented through loudspeakers or headphones along with the visual stimulus or without it. The hypothesised effect of the presence of visual stimuli was most pronounced in the loudspeaker scenarios.

Future research and method improvement might include using a higher number of loudspeakers to control sound source localisation rather than stereo pairs and panorama. In such a case, the binauralised cases could employ virtual sound sources in the same exact locations rather than stereo pairs replicating the setup. Furthermore, the use of open type headphones, as well as individual (or at least, individualised) HRTFs, might improve the localisation of the headphone scenarios.

Also, the method could include some simple localisation tasks as a form of training for participants before they take part in the evaluation. In such a case, the participants' unfamiliarity with Ambisonic and binaural listening may affect the results to a lesser degree. This may be one of the limitations of our study. Also, in a future study, we may check whether sophisticated visual objects will validate the results obtained. Moreover, a more diversified group of listeners should be engaged in the subjective tests to avoid specific biases such as, e.g. gender, people listening or not to music, working in the music area, etc.

Finally, the use of a gaze tracker might add some additional information and ecological validity to the study. Most gaze tracking techniques enable to determine the respondent's eyes and head movements during the subjective tests. This may be beneficial to check the agreement between the test participants' indications and the direction of their eye-head movement carried out while performing a task prescribed (Munn, Pelz, 2008).

References

1. Alais D., Burr D. (2004), The ventriloquist effect results from near-optimal bimodal integration, *Current Biology*, **14**(3): 257–262, doi: 10.1016/j.cub.2004.01.029.

2. Bernschütz B. (2013), A spherical far field HRIR/HRTF compilation of the Neumann Ku 100, [in:] *Proceedings of the 39th DAGA*, pp. 592–595, Meran, Italy.
3. Bertelson P. (1998), Starting from the Ventriloquist: The perception of multimodal event, [in:] *Advances In Psychological Science, Vol. 2. Biological And Cognitive Aspects*, Sabourin M., Craik F.I.M., Robert M. [Eds], pp. 419–439, Psychology Press/Erlbaum (UK) Taylor & Francis.
4. Bertelson P., Aschersleben G. (1998), Automatic visual bias of perceived auditory location, *Psychonomic Bulletin & Review*, **5**: 482–489, doi: 10.3758/bf03208826.
5. Bizley J.K., Maddox R.K., Lee A.K.C. (2016), Defining auditory-visual objects: behavioral tests and physiological mechanisms, *Trends in Neurosciences*, **39**(2): 74–85, doi: 10.1016/j.tins.2015.12.007.
6. Blauert J., Braasch J. [Eds] (2020), *The Technology of Binaural Understanding*, Springer International Publishing, doi: 10.1007/978-3-030-00386-9.
7. Chiou R., Rich A.N. (2012), Cross-modality correspondence between pitch and spatial location modulates attentional orienting, *Perception*, **41**(3): 339–353, doi: 10.1068/p7161.
8. Ecker A.J., Heller L.M. (2005), Auditory – visual interactions in the perception of a ball's path, *Perception*, **34**(1): 59–75, doi: 10.1068/p5368.
9. Frissen I., Vroomen J., De Gelder B., Bertelson P. (2004), The aftereffects of ventriloquism: generalization across sound-frequencies, *Acta Psychologica*, **118**(1–2): 93–100, doi: 10.1016/j.actpsy.2004.10.004.
10. Gardner M.B. (1968), Proximity image effect in sound localization, *The Journal of the Acoustical Society of America*, **43**(1): 163, doi: 10.1121/1.1910747.
11. Hampel F.R. (1974), The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, **69**(346): 382–393, doi: 10.2307/2285666.
12. Hendrickx E., Paquier M., Koehl V. (2015), Audiovisual spatial coherence for 2D and stereoscopic-3D movies, *Journal of the Audio Engineering Society*, **63**(11):889–899, doi: 10.17743/jaes.2015.77.
13. IEM Plug-In Suite (n.d.), https://plugins.iem.at/.
14. Ildirar S., Levin D.T., Schwan S., Smith T.J. (2017), Audio facilitates the perception of cinematic continuity by first-time viewers, *Perception*, **47**(3): 276–295, doi: 10.1177/0301006617745782.
15. Iravantchi Y., Goel M., Harrison C. (2020), Digital ventriloquism: giving voice to everyday objects, [in:] *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, doi: 10.1145/3313831.3376503.
16. Kohlrausch A., Par S. van de (2005), Audio-visual interaction in the context of multi-media applications, [in:] *Communication Acoustics*, Blauert J. [Ed.], pp. 109–138, Springer, Berlin, Heidelberg, doi: 10.1007/3-540-27437-5_5.
17. Komiyama S. (1989), Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems, *Journal of the Audio Engineering Society*, **37**(4): 210–214, http://www.aes.org/e-lib/browse.cfm?elib=6094.
18. Kunka B., Kostek B. (2012), Objectivization of audio-visual correlation analysis, *Archives of Acoustics*, **37**(1): 63–72, doi: /10.2478/V10168-012-0009-4.
19. Kunka B., Kostek B. (2013), New aspects of virtual sound source localization research–impact of visual angle and 3-D video content on sound perception, *Journal of the Audio Engineering Society*, **61**(5): 280–289, http://www.aes.org/e-lib/browse.cfm?elib=16824.
20. Morein-Zamir S., Soto-Faraco S., Kingstone A. (2003), Auditory capture of vision: examining temporal ventriloquism, *Cognitive Brain Research*, **17**(1): 154–163, doi: 10.1016/s0926-6410(03)00089-2.
21. Munn S.M., Pelz J.B. (2008), 3D point-of-regard, position and head orientation from a portable monocular video-based eye tracker, [in:] *Proceedings of the Eye Tracking Research & Application Symposium*, *ETRA 2008*, pp. 181–188, Savannah, Georgia, USA, doi: 10.1145/1344471.1344517.
22. Pike C., Stenzel H. (2017), Direct and indirect listening test methods – a discussion based on audio-visual spatial coherence experiments, [in:] *143rd Audio Engineering Society Convention*, New York, USA.
23. Pruchnicki P., Plaskota P. (2008), Automatic measuring system for head-related transfer function measurement, *Archives of Acoustics*, **33**(1): 19–25.
24. Radeau M., Bertelson P. (1977), Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations, *Perception & Psychophysics*, **22**(2): 137–146, doi: 10.3758/bf03198746.
25. Ramos O.A., Tommasini F.C. (2014), Magnitude modelling of HRTF using principal component analysis applied to complex values, *Archives of Acoustics*, **39**(4): 477–482, doi: 10.2478/aoa-2014-0051.
26. Regan D., Spekreijse H. (1977), Auditory-visual interactions and the correspondence between perceived auditory space and perceived visual space, *Perception*, **6**(2): 133–138, doi: 10.1068/p060133.
27. Romanov M., Berghold P., Frank M., Rudrich D., Zaunschirm M., Zotter F. (2017), Implementation and evaluation of a low-cost headtracker for binaural synthesis, [in:] *142nd Audio Engineering Society Convention*, Berlin, Germany.
28. Sorati M., Behne D.M. (2021), Considerations in audio-visual interaction models: an ERP study of music perception by musicians and non-musicians, *Frontiers in Psychology*, **11**: 33551911, doi: 10.3389/fpsyg.2020.594434.
29. Stenzel H., Francombe J., Jackson P.J.B. (2019), Limits of perceived audio-visual spatial coherence as defined by reaction time measurements, *Frontiers in Neuroscience*, **13**: 451, doi: 10.3389/fnins.2019.00451.

30. Stenzel H., Jackson P.J.B. (2018), Perceptual thresholds of audio-visual spatial coherence for a variety of audio-visual objects, [in:] *Audio Engineering Society International Conference on Audio for Virtual and Augmented Reality*, Redmond, WA, USA.

31. Stenzel H., Jackson P.J.B., Francombe J. (2017), Modeling horizontal audio-visual coherence with the psychometric function, [in:] *142nd Audio Engineering Society Convention*, Berlin, Germany.

32. Storek D., Rund F., Marsalek P. (2016), Subjective evaluation of three headphone-based virtual sound source positioning methods including differential head-related transfer function, *Archives of Acoustics*, **41**(3): 437–447, doi: 10.1515/aoa-2016-0043.

33. Vorländer M. (2014), Virtual acoustics, *Archives of Acoustics*, **39**(3): 307–318, doi: 10.2478/aoa-2014-0036.

34. Vorländer M. (2020), *Auralization. Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, Springer International Publishing, doi: 10.1007/978-3-030-51202-6.

35. Vroomen J., Bertelson P., De Gelder B. (2001), The ventriloquist effect does not depend on the direction of automatic visual attention, *Perception & Psychophysics*, **63**(4): 651–659, doi: 10.3758/bf03194427.

36. Vroomen J., De Gelder B. (2004), Perceptual effects of cross-modal stimulation: ventriloquism and the freezing phenomenon, [in:] *Handbook of Multisensory Processes*, Calvert G.A., Spence C., Stein B.E. [Eds], pp. 141–150, MIT Press.

37. Walker L., Walker P., Francis B. (2012), A common scheme for cross-sensory correspondences across stimulus domains, *Perception*, **41**(10): 1186–1192, doi: 10.1068/p7149.

38. Woodcock J., Davies W.J., Cox T.J. (2019), Influence of visual stimuli on perceptual attributes of spatial audio, *Journal of the Audio Engineering Society*, **67**(7/8): 557–567, doi: 10.17743/jaes.2019.0019.

39. Yao S.-N., Collins T., Liang C. (2017), Head-related transfer function selection using neural networks, *Archives of Acoustics*, **42**(3): 55–62, doi: 10.2478/aoa-2013-0007.

40. Zaunschirm M., Schoerkhuber C., Hoeldrich R. (2018), Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint, *The Journal of the Acoustical Society of America*, **143**(6): 3616, doi: 10.1121/1.5040489.

41. Zotter F., Frank M. (2019), *Ambisonics. A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, Springer International Publishing, pp. 89–90, doi: 10.1007/978-3-030-17207-7.