***This paper has been accepted for publication in Convergence: The International Journal of Research into New Media Technologies***

**Soundtrack Loudness as a Depth Cue in Stereoscopic 3D Media**

**Christos Manolas, Sandra Pauletto and Jon Jang**

**Abstract**

Assisted by the technological advances of the past decades, stereoscopic 3D cues are currently being integrated both in interactive and non-interactive media. Arguably, the main focus of this effort is placed on the creation of an increased sense of visual depth. Considering that human perception relies heavily on the audiovisual integration rather than on visual information alone, it is rather surprising that, in contrast to the evident interest towards the study of realistic 3D audio spatialisation techniques and technologies, relatively little attention has been given so far to the potential effect of the soundtrack on 3D depth perception in terms of the soundtrack. The multisensory nature of human perception suggests that the potential of sound design as a means to influence depth perception in the stereoscopic 3D visual world is worthy of further exploration.

This study reports on our research into the possibilities of using alterations of the volume levels of the soundtrack as a means of affecting the perception of visual depth while viewing stereoscopic 3D animation clips. Based on previous findings indicating that the volume level of the soundtrack may be related to the perception of visual depth, a series of experiments further explored the effectiveness of this auditory cue. Results suggest that, under certain conditions, differences in volume levels of the soundtrack could influence the judgement of visual depth in a way that is opposite to real life expectations. It is suggested that different, more metaphorical perceptual mechanisms are in play when viewing stereoscopic 3D presentations than in real life. In this context, we conclude that stereoscopic 3D media can benefit from further exploration of the effectiveness of certain auditory cues as a means to enhance and support the perception of depth within the 3D environment.

**Introduction**

Although the use of stereoscopic 3D visual cues in moving images presentations is almost as old as cinema itself (Zone, 2007), in the context of mainstream audiovisual media, the use of stereoscopic 3D visual content has been largely regarded as a transient and superficial attraction registering many failed attempts to find its permanent place in the media sector. However, with the rapid technological developments of the past decades, the idea of transitioning from 2D to stereoscopic 3D not only has been revived but starts becoming the preferred and obvious choice for many emerging media forms. Prime examples of this are the recent revival of stereoscopic 3D cinema, the increasing commercial implementation and popularity of numerous bespoke multisensory entertainment platforms (CJ 4DPLEX, 2019; Merlin Entertainments, 2019) and, most notably, the rapid advances in virtual reality and 3D gaming both in terms of research and commercial interest and application (HTC Corporation, 2019).

While this transition takes place, many of the established and tested production ideas that were suitable for traditional media forms may need to either be revisited and revised or even abandoned altogether. At the same time, new forms of production workflows with unique technological and aesthetic characteristics are emerging. It can be, therefore, both intriguing and relevant to question whether the audiovisual production techniques that were developed with 2D visual content in mind are also the optimal choices for stereoscopic 3D productions, as the production process, viewing principles and expectations created in these visual presentation styles can be different (Mendiburu, 2009).

Considering that perception of our surroundings relies heavily on multisensory integration rather than on visual information alone (Shams and Kim, 2010; Stein and Meredith, 1993), the creation of an audiovisual environment with more pronounced spatial characteristics, such as the one created by the use of stereoscopic 3D visuals, could be examined in the context of both the visual and auditory aspects, as well as their combination. In terms of sound design and audio content creation in particular, the adoption of stereoscopic 3D visuals may also introduce new challenges and a spatial environment that is no longer constrained by the strict limitations of the 2D screen (Manolas and Pauletto, 2019). However, at the present time, in the majority of cases the illusion of more pronounced depth created by the stereoscopic 3D visuals does not seem to be reflected in the audio domain in a way that is significantly different to that of traditional 2D productions (Mendiburu, 2009). This is especially true in the context of commercial stereoscopic 3D productions that are also available in 2D, where similar, if not identical, sound design techniques are used for the creation of soundtracks for both the 2D and the stereoscopic 3D versions.

The rather limited interest in exploring the creative implications of stereoscopic 3D visuals on the audio production and creation context is also reflected on the relatively limited volume of research output in this particular direction. Although a significant body of work focuses on the field of auditory depth perception when considering sound without visuals (Bregman, 1994; Blauert, 1997; Coleman, 1963; Moore, 2013) and on the audiovisual depth perception when using relatively simple audiovisual stimuli and strict laboratory conditions (Ecker and Heller, 2004; Kitagawa and Ichihara, 2002; Turner et al., 2011; Vroomen and De Gelder, 2000), research on how auditory depth cues can affect the sense of depth in non-interactive stereoscopic 3D presentations with rich, complex and constantly changing audiovisual content, such as in most stereoscopic 3D cinematic or animation narrative presentations, is rather limited. It is possible that in an environment heavily focused on an increased sense of depth such as this, multisensory cues like sound could assist content producers by supporting and enhancing the impact of the stereoscopic 3D visual information in order to achieve a more pronounced sense of depth and spatial representation. More specifically, an existing stereoscopic 3D object or environment may become more pronounced when accompanied by appropriate auditory cues, therefore increasing the perceived effect of the stereoscopic 3D visuals or the general sense of spatial immersion and envelopment indirectly influencing the sensory interpretation of the stereoscopic 3D environment as a whole. Although it may be true that most of the techniques and approaches used by sound designers in 2D media may be still applicable and effective in stereoscopic 3D, this should not discourage us from exploring alternative ways of creating soundtracks specifically tailored to the unique characteristics of the stereoscopic 3D visual environment.

The potentially large number of ways in which the introduction of stereoscopic 3D visuals may affect the creative decisions of sound designers and the diversity of the parameters that may need to be studied in this context pose a challenge to anyone willing to start exploring this area. In order to identify potential areas of interest, Manolas and Pauletto (2019) proposed a number of stereoscopic 3D production areas that may be of interest to sound designers and content creators. Based on this, a number of subjective tests exploring the effect of basic auditory depth parameters, such as volume attenuation and high frequency content alteration with distance, on the perception of stereoscopic 3D depth of short animations were carried out. One of the observations made during these tests was that, under certain conditions, increased volume levels of the soundtrack may lead to an increased perceived level of visual stereoscopic 3D depth, something that is contrary to real life conditions where increased distance from a sound-emitting object would normally result in a decrease of the volume level. Although the perceptual mechanisms behind this observation are not entirely clear, the authors suggested that in the context of viewings of stereoscopic 3D audiovisual presentations an increased volume level, and therefore a more powerful and imposing sound, may be associated with depth and vastness of space. It may be also possible that a louder soundtrack further disconnects the viewer/listener from the sounds of their surroundings through the mechanism of masking (Bregman, 1994), enabling them to achieve higher levels sensory immersion and envelopment and, potentially, better focus and concentration on the available depth cues during the audiovisual presentations (Witmer et al., 2005).

The current study aimed at exploring the observation that increased volume levels of the soundtrack may lead to an increased perceived level of visual stereoscopic 3D depth further. A series of subjective evaluation tests were designed and carried out in order to test whether the overall volume level of the soundtrack could influence the stereoscopic 3D depth perception of a given stereoscopic 3D animation sequence. More specifically, the tests aimed at investigating whether an increase of the volume level of the soundtrack could also increase the level of the perceived stereoscopic 3D depth as a whole. The experimental hypothesis for the tests was the following:

*(H1) The perceived level of visual depth of a short stereoscopic 3D animation sequence cannot be increased by an amplitude increase of the accompanying soundtrack.*

The independent variable was the overall volume level of the soundtrack and had two states: control (original volume or 0dB) and experimental (increased volume by 6dB). The tests were analysed using the Binomial test method (Harris, 2008; McCluskey and Lalkhen, 2007). The results indicated that increased volume levels of the soundtrack, under certain conditions, can lead to an increased level of perceived depth of stereoscopic 3D animation sequences as a whole.

**Theoretical Background**

Crossmodal Integration

In media utilising conventional 2D visuals, the illusion of visual depth is created by means of monocular depth cues (Mendiburu, 2009). Auditory depth cues, such as volume attenuation, high frequency loss and reverberation, are commonly used to support the visuals and enhance the sense of depth (Beauchamp, 2005; Blauert, 1997; Coleman, 1963). In media utilising stereoscopic 3D visuals, the introduction of stereoscopic visual cues further enhances the illusion of visual depth. However, in real life humans do not rely solely on cues from one sensory stream in order to create representations of our surroundings. Perceptual confirmation of events by comparing inputs from different sensory streams is an important factor in determining the spatial characteristics of our environment. The human brain is extremely good at combining and analysing cues coming from different sensory streams and identifying whether they belong to the same external event (Ecker and Heller, 2004; Shams and Kim, 2010; Stein and Meredith, 1993). This ability of the brain known as crossmodal integration (Spence et al., 2009; Bologniniet al.,2005) and it *‘can enhance perceptual clarity and reduce ambiguity about the sensory environment’* (Lippert et al., 2007; Stein and Meredith, 1993).

In terms of vision and audition, the two senses associated with audiovisual presentations, perceptual organisation of information obtained from visual and auditory sensory streams is a unified, although multifaceted and complex, process. The brain uses cues obtained from each of the two sensory streams to reinforce and verify an audiovisual sensory event (Shams and Kim, 2010; Vroomen and De Gelder, 2000). Changes of the characteristics of one of the sensory streams may also perceptually affect another (Bolognini et al*.*, 2005; Ecker and Heller, 2004). For example, *‘if auditory and visual stimuli are presented synchronously but from different positions, the auditory event is mislocated towards the locus of the visual stimulus’* (Kitagawa and Ichihara, 2002). This means that in addition to comparing and organising auditory and visual cues separately, the brain also monitors and compares cues across the two different modalities. If crossmodal cues confirm each other more accurate and solid conclusions can be drawn about a given event or object (Lippert et al., 2007).

Previous studies have shown that the presence of crossmodal cues can affect the perception of depth in stereoscopic 3D visual environments. Turner et al. (2011) have shown that the perception of distance of stereoscopic 3D visual objects can be affected by changing the distance from the audience of a loudspeaker playing a corresponding sound. Ecker and Heller (2004) have shown that the perceived path of a ball moving in 3D-space can be changed by adding an auditory cue. Kitagawa and Ichihara (2002) have shown that a steady sound can be perceived to be changing in loudness when watching a synchronous 3D object moving in depth. Manolas and Pauletto (2014) have suggested that volume levels can influence the perceived depth of visual elements and environments in stereoscopic 3D animation presentations.

Although a thorough analysis of cross-modal integration concepts is beyond the scope of this study, it is important to note that perception of our surroundings is inherently multimodal. As such, perception of visual cues can be influenced by audition. In the context of stereoscopic 3D media in general and of the current study, this becomes an interesting observation, as auditory cues could be used to influence, enhance or highlight aspects of the stereoscopic 3D environment and, thus, affect the way the stereoscopic 3D world is perceived as a whole.

Realism versus Appropriateness of Auditory Cues for Sound Design Purposes

*'Since Aristotle, many philosophers and psychologists have believed that perception is the process of using the information provided by our senses to form mental representations of the world around us.'* (Bregman, 1994: 3). On a broader perspective, this definition could also be used for describing perception in the context of audiovisual media presentations. However, there are significant differences between the latter and perception in real life conditions. Firstly, perception in the context of audiovisual media presentations is constrained by the fact that limited sensory information is used (i.e. visual and auditory) compared to real life (i.e. information from all five senses). Secondly, in audiovisual media presentations perception is not necessarily used to build representations of a realistic environment that surrounds us, but of a symbolic one that can be largely unrealistic in many ways (i.e. the environment where the visual action takes place is spatially restricted within the screen area). While in real-life we use our perception to understand actual events taking place around us, as audiovisual media audiences we suspend our disbelief in order to make sense of a knowingly unrealistic world (Abrams et al., 2001). Thus, paraphrasing the previously mentioned general definition of perception, it is suggested that, in the typical audiovisual presentation context, perception is the process of using the information provided by some of our senses to form mental representations of knowingly unrealistic worlds created by someone else. An in-depth analysis of the above concept from a physiological, psychological or philosophical perspective exceeds the scope of this study. However, it is useful to highlight the difference between perception in real-life and in narrative audiovisual media presentations, as although the latter is actually a particular case of the former (Langkjær 1997: 93) the two might not necessarily coincide in all respects.

For the purposes of sound design for stereoscopic 3D media, appropriate sound cues may or may not be necessarily the ones that are realistic in a strict sense, rather than ones that are capable of complementing or enhancing the desired sensory effects irrespectively of the level of their apparent realism. As an example, the results of a previous related study indicated that increased volume levels of an ambience sound, under certain conditions, may have a positive effect on the perception of visual depth while viewing short stereoscopic 3D animation sequences (Manolas and Pauletto, 2014). In a realistic context, an increased depth or distance between the viewer/listener and the surrounding sound-emitting objects would be expected to produce lower ambience sound levels. In this case, the perceptual effect is, therefore, defying expectations. Such a relationship between increased volume levels and enhanced visual depth may be due to complex multisensory perceptual processes the thorough analysis of which exceeds the scope of this study. However, if it has a positive effect on the perceived depth of the stereoscopic 3D scene as a whole it can be used as a creative tool when designing soundtracks that accompany stereoscopic 3D visual presentations. Thus, the lack of realism in the way one adjusts the volume levels of the audio cues (e.g. ambience sounds, spot effects), or the soundtrack as a whole, does not necessarily compromise the appropriateness of this auditory cue for the purpose of enhancing visual depth in stereoscopic 3D media presentations.

**Method**

Design

In order to test the main hypothesis:

*(H1) The perceived level of visual depth of a short stereoscopic 3D animation sequence can be increased by an amplitude increase of the accompanying soundtrack,*

a number of subjective tests were designed and carried out. The tests employed a pair comparison design based on the ITU-T Recommendation P.911 (2008) with two versions of a stereoscopic 3D clip (i.e. control and experimental) presented to each participant in succession in a randomised order.

For the purposes of the tests, four short stereoscopic 3D animation clips (i.e. 10-15 seconds) were produced. Two versions of each clip were then created with the only difference being the volume level of the soundtrack (+/-6dB). The soundtrack consisted of a combination of environmental sounds, spot effects, dialogue and/or music as required based on content and narrative. Further details on the sourcing and production of the clips are provided in *Apparatus and Material*.

The independent variable for the tests (i.e. overall volume level of the soundtrack) had two states: control and experimental. In the experimental version, the soundtrack was increased in volume by 6dB compared to the control version. The decision to use a 6dB increase was made based on a number of observations. Firstly, based on the inverse square law (Coleman, 1963), it roughly indicates decrease of the distance between the listener and the sound emitting object(s) by half. This was regarded as a potentially significant change in the terms of mentally calculating depth of a certain sound emitting even or object. Secondly, during pilot testing and refining the test material, it was observed that larger changes (e.g. +/-9-12dB) were too prominent and drew direct attention to the soundtrack itself, something that was not desirable as the purpose of the tests was to evaluate the overall sense of visual depth of the scene without focusing specifically on the auditory stream. Thirdly, smaller changes (e.g. +/-3dB) were found to be too weak to influence the overall sense of depth in the presence of prominent stereoscopic 3D visual cues. Finally, differences of +/-6dB were observed to cause a similar effect in previous related tests (Manolas and Pauletto, 2014). The dependent variable was the overall perceived level of depth of the stereoscopic scene as a whole.

The collected data were analysed using an exact proportion sign test (McCluskey and Lalkhen, 2007; Field, 2005) in SPSS (SPSS, 2019).

Participants

In total, thirty participants were recruited for the tests. The choice of the sample size was based on the minimum requirements for the proportion sign test for a statistical power of .8 or higher, a significance level of .05 and an effect size of .5. The test requirements were calculated using the GPower software (Faul et al., 2007; Heinrich-Heine-Universität, 2013).

Participants were students and staff from SAE Institute in Oxford, UK (SAE, 2019) and they were from various departments and cultural and ethnic backgrounds. Both experts and non-experts in the field of audio and video production participated. The frequency of audio and video production experts in the sample population was 33% and 40% respectively. The mean age of the participants was 25 years (standard deviation = 9.73) with a maximum age of 58 years and a minimum age of 18 years, with 20 % being female and 80 % male.

Participants were asked to rate the overall level of stereoscopic 3D depth of the sequences and, thus, were unaware of the fact that only the audio element of the clips within the sequences was modified prior to taking the tests. This was in order to ensure that their attention would not be purposely directed towards the soundtrack, so that the perceptual mechanism would operate as in a typical audiovisual presentation in normal real-life conditions. The details of the actual purpose of the test were made available to the participants only after all the data was collected. Participants had normal or corrected to normal vision and no significant hearing difficulties.

Apparatus and Material

Audiovisual Content Sourcing and Creation

In total, four stereoscopic 3D animation clips were created based on material sourced from the stereosocpic version of Blender Foundation’s Elephants Dream (Blender Foundation, 2006). The material was under a creative commons license (Creative Commons, 2013). The original soundtrack was removed and four short excerpts (i.e. Clips 1, 2, 3 and 4) were taken from the visual-only source material. Next, original soundtracks were created in order to be added to the four clips. The soundtracks consisted of ambience/environmental sounds, spot effects, dialogue and music appropriate for each clip based on visual content and narrative (**Table 1**).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Ambience** | **Spot FX** | **Dialogue** | **Music** |
| **Clip 1** | x | x |  | x |
| **Clip 2** | x | x | x |  |
| **Clip 3** | x | x | x | x |
| **Clip 4** | x | x | x | x |

Table 1: Soundtrack stems used in each of the four clips

All audio assets were created using Logic Pro X (Apple Inc., 2019). For each of the four clips, two versions of the soundtrack were produced: control and experimental. The volume level of the experimental versions was increased by 6dB in comparison to the control versions. The soundtracks were added to the clips and the resulting audiovisual stereoscopic 3D clips were named in the following manner:

**Clip1\_Ctrl\_0dB:** Original Soundtrack

**Clip1\_Exp\_-6dB:** Original Soundtrack + 6dB

**Clip2\_Ctrl\_0dB:** Original Soundtrack

**Clip2\_Exp\_-6dB:** Original Soundtrack + 6dB

**Clip3\_Ctrl\_0dB:** Original Soundtrack

**Clip3\_Exp\_-6dB:** Original Soundtrack + 6dB

**Clip4\_Ctrl\_0dB:** Original Soundtrack

**Clip4\_Exp\_-6dB:** Original Soundtrack + 6dB

Next, the two versions of each clip were merged into one sequence that played them in succession but separated by a still grey frame displayed for 5 seconds, in line with related ITU-T recommendations (2008). The audiovisual clips and the 5 second grey screen stills were concatenated using the Ffmpeg platform (Ffmpeg.org, 2019). The concatenation resulted in four sequences the structure of which is demonstrated in the following example:

**Sequence1\_Ctrl\_Exp:**

Grey Frame 5sec > **Clip1\_Ctrl\_0dB** > Grey Frame 5sec > **Clip1\_Exp\_+6dB** > Grey Frame 5sec

Next, both possible order combinations for each sequence were produced:

**Sequence1\_Ctrl\_Exp:**

Grey Frame 5sec > **Clip1\_Ctrl\_0dB** > Grey Frame 5sec > **Clip1\_Exp\_+6dB** > Grey Frame 5sec

**Sequence1\_Exp\_Ctrl:**

Grey Frame 5sec > **Clip1\_Exp\_+6dB** > Grey Frame 5sec > **Clip1\_Ctrl\_0dB** > Grey Frame 5sec

Ultimately, this led to the creation of a total of eight sequences, two for each of the four clips:

**Sequence1\_Ctrl\_Exp:**

Grey Frame 5sec > **Clip1\_Ctrl\_0dB** > Grey Frame 5sec > **Clip1\_Exp\_+6dB** > Grey Frame 5sec

**Sequence1\_Exp\_Ctrl:**

Grey Frame 5sec > **Clip1\_Exp\_+6dB** > Grey Frame 5sec > **Clip1\_Ctrl\_0dB** > Grey Frame 5sec

**Sequence2\_Ctrl\_Exp:**

Grey Frame 5sec > **Clip2\_Ctrl\_0dB** > Grey Frame 5sec > **Clip2\_Exp\_+6dB** > Grey Frame 5sec

**Sequence2\_Exp\_Ctrl:**

Grey Frame 5sec > **Clip2\_Exp\_+6dB** > Grey Frame 5sec > **Clip2\_Ctrl\_0dB** > Grey Frame 5sec

**Sequence3\_Ctrl\_Exp:**

Grey Frame 5sec > **Clip3\_Ctrl\_0dB** > Grey Frame 5sec > **Clip3\_Exp\_+6dB** > Grey Frame 5sec

**Sequence3\_Exp\_Ctrl:**

Grey Frame 5sec > **Clip3\_Exp\_+6dB** > Grey Frame 5sec > **Clip3\_Ctrl\_0dB** > Grey Frame 5sec

**Sequence4\_Ctrl\_Exp:**

Grey Frame 5sec > **Clip4\_Ctrl\_0dB** > Grey Frame 5sec > **Clip4\_Exp\_+6dB** > Grey Frame 5sec

**Sequence4\_Exp\_Ctrl:**

Grey Frame 5sec > **Clip4\_Exp\_+6dB** > Grey Frame 5sec > **Clip4\_Ctrl\_0dB** > Grey Frame 5sec

This was in order to be able to later produce a randomised order of sequence viewing for each participant, both in terms of the sequence presentation (i.e. Sequence4 > Sequence2 > Sequence1 > Sequence3) and of the version of the clip for order for each sequence (i.e. Sequence4\_Ctrl\_Exp > Sequence2\_Exp\_Ctrl > Sequence1\_Exp\_Ctrl > Sequence3\_Ctrl\_Exp) (see: *Procedure*). Both the viewing order of the sequence presentation and of the control and experimental version within each sequence was determined using the online randomization tool *Research Randomizer* (Urbaniak and Plous, 2008).

Software and Hardware

The test environment consisted of a square viewing space, a notebook computer, a set of nearfield monitors, a desk and a short throw 3D projector. The dimensions of the space were 3m (Width) x 3m (Length) x 3m (Height). The distance between the projector and the screen was 2.5m and the diagonal size of the viewing area of the screen was 90 inches (2.28 m) with an aspect ratio of 16:9. Two nearfield monitor loudspeakers were placed in a stereo arrangement approximately 1m from the participant. Prior to the tests,the speakers were arranged and calibrated in accordance to the Dolby Labs recommendations (Dolby Laboratories S00/12957, 2000). The stereoscopic 3D sequences were played using the Bino 3D Stereoscopic Player (Bino.org, 2019).

The following hardware and software components were used during the tests:

* Apple Macbook Pro (Apple Inc, 2019).
* Bino 3D Stereoscopic Video Player (Bino.org, 2019; GNU, 2007).
* Optoma GT750 3D projector (Optoma, 2019).
* Optoma DLP 3D Glasses (Optoma, 2019).
* Genelec 8020A Monitor Speakers x 2. (Genelec, 2019).

Procedure

The space arrangements dictated that a single participant should take the test at any given time without compromising the viewing and listening conditions. This also ensured that the order of presentation could be fully randomised and unique for each participant, something that was a requirement of the methodology used for the test. Prior to the test, participants were asked to complete an anonymous personal information form in order to collect the following demographic information: age, gender, ethnic background, as well as whether the participant was considered an expert in audio. Participants were also asked to sign an informed consent form.

After completing the paperwork, participants were asked to take a seat inside the experiment space at a distance of approximately 2.5m from the screen. The test procedure was verbally described by the experimenter and a printed instructions sheet was also handed out. Next, two trial sequences including a control and an experimental version of a stereoscopic 3D clip, similar to the ones created for the actual tests, were played in a random order. After the viewing of each sequence, participants were asked to choose which of the two versions appeared to have the most prominent level of stereoscopic 3D depth. This was in order to test the participants’ visual acuity, stereoscopic vision and hearing ability, as well as the functionality of the equipment. It also allowed them to familiarise with the testing procedure and to ensure that the audiovisual settings used were comfortable.

Once the trial tests were completed, the main tests were carried out with the four main sequences played in a fully randomised order, as previously described (see: *Audiovisual Content Sourcing and Creation*). Prior to the tests, a paper-based answering sheet was provided in which participants could choose either the first or the second clip of each of the four sequences. They also had the option to skip the question if they felt that the difference was too subtle, but they were encouraged to try to select one of the two clips even if the perceived difference between them was minor. Participants were unaware of the actual order of the original clips within each sequence (e.g. control or experimental) and of the order of playback of the four sequences. The answering sheets format is demonstrated below:

**Sequence A**

Clip 1 Clip 2 Skip Question

**Sequence B**

Clip 1 Clip 2 Skip Question

**Sequence C**

Clip 1 Clip 2 Skip Question

**Sequence D**

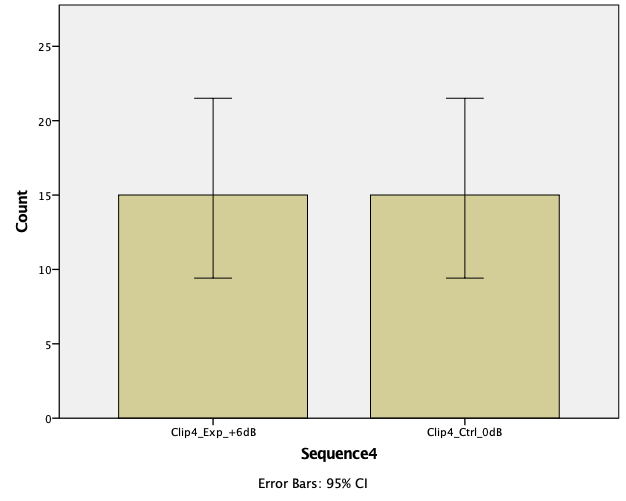
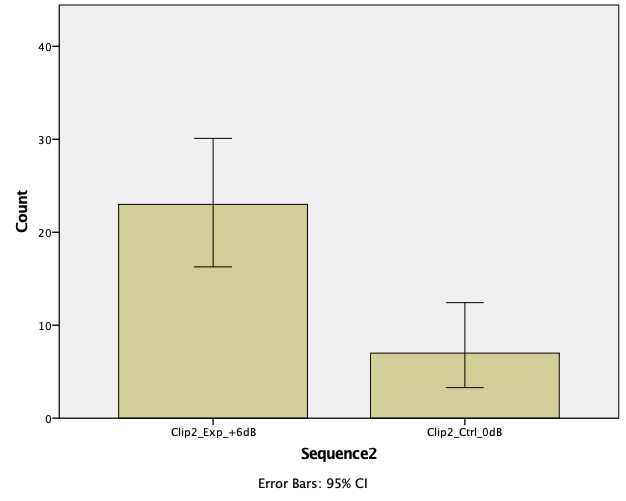
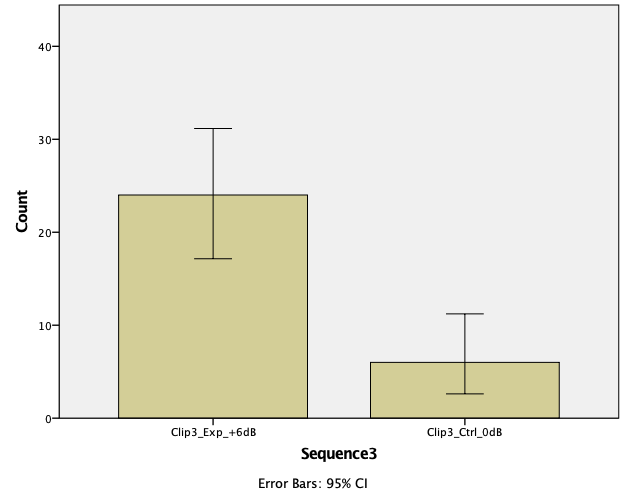
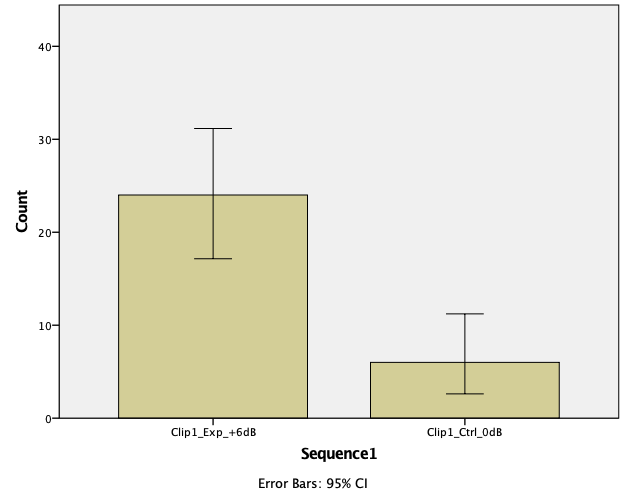
Clip 1 Clip 2 Skip Question

The experimenter had a matrix that was unavailable to the participant indicating the actual order of the original sequences and of the randomised control and experimental version within each the answering sheet information corresponded to. For instance, while a participant may have chosen Clip 1 of Sequence A, the matrix may have indicated to the experimenter that this corresponded to the experimental clip of Sequence 3 (e.g. Clip3\_Exp\_+6dB). Upon completion of the test, the experimenter explained that the purpose of the experiment and the independent variable would be available to the participant after the collection of all the data. This was in order to ensure that the participant would not disclose key information about the test to others, something that could potentially make other prospective participants aware of the details of the test. This, in turn, would have compromised the effectiveness of the tests, as prospective participants could have been aware of the fact that the studied independent variable was audio related and focus their attention on the soundtrack.

Results

The collected data were analysed using the exact proportions (Binomial) sign test (McCluskey and Lalkhen, 2007). The alpha level was 0.05 and the estimated actual power was 0.89 for an effect size (g) of 0.25 based on the total sample size of 30 participants. This was calculated using the GPower software (Faul et al*.*, 2007; Heinrich-Heine-Universität, 2013).

For Sequences 1, 2 and 3 there was statistically significant association between the increase of the volume level of the soundtrack by 6dB (i.e. experimental soundtrack) and an increase in the overall perceived levels of stereoscopic 3D depth of the clip as a whole. No such statistically significant association was observed in Sequence 4 (**Figures 1-4**).



Figures 1-4: Graphs for Sequences 1-4

The 2-tailed exact significance levels for Sequences 1, 2, 3 and 4 were .001, .005, .001 and 1.00 respectively. The Binomial test results are summarised in **Table 2** below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **­­­** | | | | | | |
|  | | Category | N | Observed Prop. | Test Prop. | Exact Sig. (2-tailed) | |
| Sequence1 | Group 1 | Clip1\_Exp\_+6dB | 24 | .80 | .50 | .001 | |
| Group 2 | Clip1\_Ctrl\_0dB | 6 | .20 |  |  | |
| Total | Total | 30 | 1.00 |  |  | |
| Sequence2 | Group 1 | Clip2\_Exp\_+6dB | 23 | .77 | .50 | .005 | |
| Group 2 | Clip2\_Ctrl\_0dB | 7 | .23 |  |  | |
| Total | Total | 30 | 1.00 |  |  | |
| Sequence3 | Group 1 | Clip3\_Exp\_+6dB | 24 | .80 | .50 | .001 | |
| Group 2 | Clip3\_Ctrl\_0dB | 6 | .20 |  |  | |
| Total | Total | 30 | 1.00 |  |  | |
| Sequence4 | Group 1 | Clip4\_Exp\_+6dB | 15 | .50 | .50 | 1.000 | |
| Group 2 | Clip4\_Ctrl\_0dB | 15 | .50 |  |  | |
| Total | Total | 30 | 1.00 |  |  | |

Table 2: Binomial Test Results

**Discussion**

Results showed that, in three out of four occasions (Sequences 1-3), an increase of the overall volume levels of the soundtrack of a stereoscopic 3D animation clip led to a reported increase of the perceived level of the stereoscopic 3D depth of the clip as a whole. Based on this, the null hypothesis:

*H0 The perceived level of visual depth of a short stereoscopic 3D animation sequence cannot be increased by an amplitude increase of the accompanying soundtrack*

can be rejected for Sequences 1, 2 and 3, but not for Sequence 4.

It is possible that the perceptual effect observed in Sequences 1, 2 and 3 may be due to the louder soundtrack allowing for a greater disconnection between the viewer/listener and the sounds of their surroundings through the mechanism of masking (Bregman, 1994). Subtle environmental sounds (e.g. distant environmental noises, natural ambience sound of the space, electrical equipment hiss) that could be possibly audible to the participants when the soundtrack was played at lower volume levels may have been masked when it was played at higher levels. According to Albert (2012), this could potentially *‘destroy the frame that separates the world of the work from that of the viewer’*. This, in turn, could have allowed the participants to achieve higher levels of concentration on the stereoscopic 3D aspects of the audiovisual presentation, something that could be interpreted as a better representation of the overall depth of the stereoscopic 3D environment. Further experimentation in acoustically isolated and controlled environments may be useful in order to explore the validity of this claim.

Altering the volume levels of complex sounds can also affect the perception of *‘the relative loudness of the different frequency components’* and, therefore, the overall *‘tonal balance’* (Moore, 2013: pp. 135-136). It is, thus, possible that the spectral content differences caused by the alteration of the overall volume level of the soundtrack are an influencing factor behind the reported increase of the perceived stereoscopic 3D depth. This could be either in addition to, or instead of any apparent change of the overall volume level. In this context, the effectiveness of volume level increase as a means to increase the perceived stereoscopic 3D depth should be studied across specific frequency bands in order to investigate this possibility.

The fact that volume level increase seems to be affecting perception of stereoscopic depth positively is contradicting real-life experience. In real life, increasing the listener’s distance from sound emitting objects would normally lead to a decrease of the volume levels of the sounds/soundscapes these objects produce, while in our tests the opposite was relationship was observed: increased volume levels led to the perception of increased stereoscopic 3D depth. It is, thus, possible that different interpretative mechanisms are used when viewing stereoscopic 3D animation clips and in real life. We suggest that when viewing stereoscopic 3D clips the viewer/listener tends to interpret a louder sound as the representation of a more imposing, more spacious and therefore deep environment, a perceptual assumption that is not directly based on real life expectation.

Such an interpretation is also supported by another possible explanation for the cause of this effect. While the various different visual and auditory elements of the stereoscopic 3D presentation could be perceptually distinguished and classified as separate objects, on a different level the viewer/listener of a stereoscopic 3D presentation may be also perceiving the whole presentation as a unified *amodal* object (Nudds, 2014). In this case, the soundtrack may be perceived as a unified auditory mixture that represents the whole of the stereoscopic 3D world in the auditory domain rather than merely as a collection of sounds that represent certain events within it. If this is true, a louder soundtrack could also increase the perceived intensity, presence and vividness of the amodal stereoscopic 3D environment as a whole. This, in turn, may cause the viewer/listener to pay more attention to the various audiovisual characteristics of the stereoscopic 3D environment, including stereoscopic depth.

Ultimately, the purpose of the tests was to investigate whether volume level alteration can be used as a creative tool for influencing the perception of visual depth in stereoscopic 3D presentations. In this context, one must also consider the unique spatial characteristics of stereoscopic 3D media that utilise a 2D screen, as opposed to both environments that expand across 360 degrees around the viewer/listener (e.g. virtual reality systems and also in real life) and 2D presentations. In real life there is no screen restricting the visual field, and events are expected to occur anywhere within the 360 degrees sphere surrounding the viewer/listener. In this environment, the viewer/listener is expected to be much more alert to sound sources located outside the field of view, as there is no restrictions to the potential location of the sound-emitting objects. On the other end, in a classic 2D presentation, although the available monoscopic visual depth cues provide a means to calculate the distance of certain visual objects from the viewer/listener, distance estimation is rather symbolic and abstract compared to the more realistic and detailed real-life equivalent. Stereoscopic 3D media could be placed somewhere between these two extremes, utilising more realistic visual spatialisation cues than 2D media but not trying to reach the level of realism found in VR/AR environments or in real life. It is, thus, suggested that audiovisual distance perception in the context of stereoscopic 3D environments could be examined as a unique and separate topic with its own constraints and rules. Considering this, the use of realistic and detailed auditory representations of cinematic objects within the soundtrack (e.g. the attempt to manipulate the volume levels of the sounds of objects that appear on-screen in such a way as to reflect their accurate and realistic supposed distances from the viewer/listener) might simply not be as important in the context of stereoscopic 3D environments as it is in other 3D audiovisual applications that may extend in a 360 degree around the viewer/listener like virtual and augmented reality environments. The results of the tests indicate that other, more metaphorical associations may be at play in stereoscopic 3D presentations than simply realistic and causal relationships. In this context, the effects of altering the volume levels of the soundtrack, or specific elements within it, on the overall perception of depth within the stereoscopic 3D environment could be explored and utilised in more creative and experimental ways, freed from the requirement for realism.

The fact that the effect was not observed in Sequence 4, suggests that soundtrack volume is not always equally effective at influencing the perception of depth in stereoscopic 3D audiovisual presentations and that other factors may be also in play. One such factor may be the visual complexity of the stereoscopic 3D scene. Sequence 4 had arguably the most complex and rich stereoscopic 3D visual content out of the four sequences (**Figures 5-8**).



Figures 5-8: Visual composition and complexity of Sequences 1-4 (top left to bottom right)

This suggests that in the presence of such rich visual content, the effectiveness of the volume level alterations of the soundtrack as a means to influence visual depth may be diminished. This is also in line with findings from previous studies (Manolas and Pauletto, 2014) and with the fact that visual cues tend to dominate over auditory ones when both are available and perceptually assessed for the confirmation of a given audiovisual event (Mastoropoulou, 2006; Shams and Kim, 2010; Woszczyk et al*.*, 1995). It is possible that in absence of strong visual depth cues, participants may associate depth and vastness of space with a louder background sound, while when enough visual depth content becomes available the visual cues take priority over the auditory ones. Further work is required to understand the exact nature of the depth assessment perceptual process in the viewer/listener of a stereoscopic 3D animation clip. However, based on these results we suggest that visual complexity should be taken into consideration when assessing the role of auditory cues in this context.

In addition to the complexity of the stereoscopic 3D visual content, complexity of the soundtrack may be another parameter that affects the perceptual mechanism. As mentioned, Sequence 4 was the richest and more complex of the four sequences in terms of visual content and the soundtrack that was created for it matched this complexity. As such, the soundtrack of Sequence 4 is busier and more complex both in terms of frequency content and variation when compared to the soundtracks of Sequences 1, 2 and 3. In this context, it can be claimed that volume alteration is less effective as a means to influence the perception of depth when used with complex soundtracks. This could be due to the complex soundscape being already immersive and imposing enough to achieve a certain perceptual effect irrespectively of whether its levels were modified within the 6dB range. Further work in this direction is needed to verify the validity of this claim.

**References**

1. Abrams N, Bell I and Udris J (2001) *Studying Film.* London, UK: Arnold.
2. Albert G (2012) Immersion as Category of Audiovisual Experience: From Long Beach to Hollywood. In: Worlds of Audio Vision. Available at: http://www-5.unipv.it/wav/index.php?option=com\_contentandview=articleandid=80andlang=en (accessed 21 July 2019).
3. Apple Inc (2019) Manuals. Available at: https://support.apple.com/en\_GB/manuals (accessed 20 July 2019).
4. Beauchamp R (2005) *Designing Sound for Animation.* Abingdon, UK: Elsevier.
5. Bino.org (2019) Bino 3D Stereoscopic Video Player. Available at: https://bino3d.org (accessed 20 July 2019).
6. Blauert J (1997) *Spatial Hearing: The Psychophysics of Human Sound Localization.* 2nd ed. Cambridge, MA, US: The MIT Press.
7. Blender Foundation (2006) Elephant’s Dream. Available at: [www.elephantsdream.org](http://www.elephantsdream.org) (accessed 20 July 2019).
8. Bolognini N, Frassinetti F, Serino A and Ladavas E (2005) ’Acoustical Vision’ of Below Threshold Stimuli: Interaction Among Spatially Converging Audiovisual Inputs. *Experimental Brain Research* 2005(160): 273–282.
9. Bregman SA (1994) *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA, US: The MIT Press.
10. CJ 4DPLEX (2019) The Innovation of 4DX. Available at: http://www.cj4dx.com/aboutus/aboutus.php (accessed 20 July 2019).
11. Coleman DP (1963) An Analysis of Cues to Auditory Depth Perception in Free Space. *Psychological Bulletin* 60(3): 302–315.
12. Creative Commons (2019) What We Do. https://creativecommons.org/about/ (accessed 20 July 2019).
13. Dolby Laboratories S00/12957 (2000) 5.1 Channel Production Guidelines, Issue 1. San Francisco, CA, US: Dolby Laboratories Inc.
14. Ecker JA and Heller ML (2004) Audio-visual Cue Combination in Depth Perception. *Journal of Vision*, 4(8): 699.
15. Faul F, Erdfelder E, Lang AG and Buchner A (2007) G\*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods* 39(2): 175–191.
16. Ffmpeg.org (2019) About Ffmpeg. Available at: https://ffmpeg.org/about.html (accessed 20 July 2019).
17. Genelec (2019) 8020A Studio Monitor. Available at: https://www.genelec.com/support-technology/previous-models/8020a-studio-monitor (accessed 20 July 2019).
18. GNU (2007) GNU General Public License. Available at: http://www.gnu.org/licenses/gpl-3.0.html (accessed 20 July 2019).
19. Harris P (2008) *Designing and Reporting Experiments in Psychology.* 3rd ed. Maidenhead, UK: Open University Press.
20. Heinrich-Heine-Universität (2019) G\*Power: Statistical Power Analyses for Windows and Mac. Available at: http://www.gpower.hhu.de/en.html (accessed 20 July 2019).
21. HTC Corporation (2019) VIVE. Available at: https://www.vive.com/eu/ (accessed 20 July 2019).
22. ITU-T P.911:12/98(2008)Subjective Audiovisual Quality Assessment Methods for Audiovisual Applications.
23. Kitagawa N and Ichihara S (2002) Hearing Visual Motion in Depth. *Nature* 416(14): 172-174.
24. Langkjær B (1997) Spatial Perception and Technologies of Cinema Sound. *Convergence* 3(4): 92-107.
25. Lippert M, Logothetis N and Kayser C (2007) Improvement of Visual Contrast Detection by a Simultaneous Sound. *Brain Research* 1173: 102–109.
26. Manolas C and Pauletto S (2019) Designing the Stereoscopic 3D Media Soundscape: An Exploration of the Perceptual Effects of Auditory Cues on Stereoscopic 3D Presentations. In: *The European Conference on Media, Communication & Film 2019: Official Conference Proceedings*. UK, Brighton: IAFOR.
27. Manolas C and Pauletto S (2014) Volume Attenuation and High Frequency Loss as Auditory Depth Cues in Stereoscopic 3D Cinema. *3D Research* 5(23): 1-16.
28. Mastoropoulou G (2006) *The Effect of Audio on the Visual Perception of High-fidelity Animated 3D Computer Graphics*. PhD Thesis, University of Bristol, UK.
29. McCluskey A and Lalkhen AG (2007) Statistics III: Probability and Statistical Tests. *Continuing Education in Anaesthesia Critical Care and Pain* 7(5): 167–170.
30. Mendiburu B (2009) *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Oxford, UK: Elsevier/Focal Press.
31. Merlin Entertainments (2019) About Us. Available at: https://www.merlinentertainments.biz (accessed 20 July 2019).
32. Moore CJB (2013) *An Introduction to the Psychology of Hearing.* 6th ed. Leiden, The Netherlands: Brill.
33. Nudds M (2014) Is Audio-Visual Perception ‘Amodal’ or Crossmodal? In: Stokes D., Mohan, M. and Biggs, S., ed., Perception and Its Modalities. Oxford University Press.
34. Optoma (2019) Optoma gt750: Big Screen Full 3d, HD Ready Gaming and Home Entertainment Projector With Integrated Award-winning Optoma 3D-XL Technology. Available at: https://www.optoma.co.uk/press-branding/press-releases/207 (accessed 20 July 2019).
35. SAE Institute (2019) About Us. Available at: https://www.sae.edu/gbr/about-us (accessed 20 July 2019).
36. Shams L and Kim R (2010) Crossmodal Influences on Visual Perception. *Physics of Life Reviews* [7(3](https://www.sciencedirect.com/science/journal/15710645/7/3)): 269-284.
37. Spence C, Senkowski D and Röder B (2009) Crossmodal Processing. *Experimental Brain Research* 198(2-3): 107.
38. Stein BE and Meredith MA (1993) *The Merging of the Senses*. Cambridge, MA, US: The MIT Press.
39. Turner A, Berry J and Holliman N (2011) Can the Perception of Depth in Stereoscopic Images be Influenced by 3D Sound? In: *SPIE Proceedings Volume 7863,* *Stereoscopic Displays and Applications XXII* (ed AJ Woods, NS Holliman and NA Dodgson)*,* San Francisco, USA,15 February 2011, San Francisco: SPIE.
40. Urbaniak CG and Plous S (2008) Research Randomizer. Available at: www.randomizer.org/ (accessed 20 July 2019).
41. Vroomen J and De Gelder B (2000) Sound Enhances Visual Perception: Cross-modal Effects of Auditory Organization on Vision. *Journal of Experimental Psychology: Human Perception and Performance* 26(6): 1583–1590.
42. Woszczyk W, Bech S and Hansen V (1995) Interactions Between Audio-visual Factors in a Home Theater System: Definition of Subjective Attributes. In: *99th Audio Engineering Society Convention*, New York, USA, 6-9 October 1995.
43. Zone R (2007) *Stereoscopic Cinema and the Origins of 3-D Film*. Lexington, US: The University Press of Kentucky.
44. Witmer BG, Jerome CJ, and Singer MJ (2005) The Factor Structure of the Presence Questionnaire. *Presence*, 14(3): 298-312.