Audio Engineering Society
Convention Paper 10590
Presented at the 152nd Convention
2022 May, In-Person and Online

Moved By Sound: How head-tracked spatial audio affects autonomic emotional state and immersion-driven auditory orienting response in VR Environments

Richard Warp[1], Michael Zhu[1], Ivana Kiprijanovska[2], Jonathan Wiesler[1], Scot Stafford[1], and Ifigeneia Mavridou[2]

[1] *Pollen Music Group, San Francisco CA, USA*

[2] *emteq labs, Sussex Innovation Centre, Brighton, UK*

Correspondence should be addressed to Richard Warp (richard@pollenmusicgroup.com)

ABSTRACT

This paper presents a narrative content-driven virtual reality (VR) experiment using novel biosensing technology to evaluate emotional response to a complex, layered soundscape that includes discrete and ambient sound events, music, and speech. Stimuli were presented in a spatialized vs mono audio format, to determine whether head-tracked spatial audio exerts an effect on physiologically measured emotional response. The extent to which a listener's sense of immersion in a VR environment can be increased based on the spatial characteristics of the audio is also examined, both through the analysis of self-reported immersion scores and physical movement data. Finally, the study explores the relationship between the creators' own intentions for emotion elicitation within the stimulus material, and the recorded emotional responses that matched those intentions in both the spatialized and non-spatialized case. The results of the study provide evidence that spatial audio can significantly affect emotional response in Immersive Virtual Environments (IVEs). In addition, self-reported immersion metrics favour a spatial audio experience as compared to a non-spatial version, while physical movement data shows increased user intention and focused localization in the spatial vs non-spatial audio case. Finally, strong correlations were found between the creators of the sound environment emotional intent for the piece and the significant clusters of valence in the spatialized audio version.

1 Introduction

The creative content industries continue their search for a winning formula for storytelling that stands out from the crowd. Recent advances in extended reality (XR) technology, accelerated by a global shift towards a more virtual ecosystem, increasingly point to immersive media being a primary vehicle for reaching audiences. Spatial audio originally entered the spotlight along with many of the earliest 360 films, virtual reality (VR) games, and music videos, and is becoming increasingly relevant with the introduction of these technologies to contemporary headphones. The narrative around spatial audio has mostly been limited to describing its role in immersion, embodiment, and providing spatial cues directing the audience where to look. But what of its role in engaging emotion? Achieving a strong emotional connection is likely to be a key component for success in this area, but how is such an outcome achieved?

Much of the literature surrounding measures of affective response in Immersive Virtual Environments (IVEs) tends to be focused primarily on the visual modality, for example examining the valence and arousal effects of modification to visual characteristics of IVEs such as colour temperature, intensity, and position [1]. However, there is relatively little research that addresses the effect of

similar auditory modifications in the spatial domain on emotional response. Given the apparent importance of the audio vs visual modality when emotional response is measured physiologically [2], it would be logical to extend such research to the spatial audio domain. Within the literature that currently exists in the area of emotional response to acoustic stimuli, a few important studies have been carried out [3,4] using biosensors similar to those in the current study, but limited to stereo material rather than spatialized, head-tracked audio. Otherwise, much of the available research tends to be based on self-reports rather than continuous physiological measures [5,6]. Else, research on spatial audio effects on physiologically measured arousal and valence has largely taken place outside of VR settings, with fixed head orientation [7,8]. Potential additional insights from head-tracking data in such studies are however excluded.

In terms of immersion, research has shown that there is substantial potential for spatial audio to increase the sense of immersion and presence in AR/VR environments [9,10]. Head movement towards spatial audio-visual cues and longer eye fixation durations were found to correlate with higher sensations of presence and increased focus towards the sound-emitting regions in spatial audio environment experience [11,12]. However, studies that have attempted to investigate this relationship using IVEs have been published scarcely and have historically relied heavily on self-reported measures.

To attain a robust measure of both emotional response and presence levels, a testing environment was built in Unity VR using a novel platform for physiological measurement created by Emteq Labs (Brighton, UK). The emteqPRO system combined with the VIVE Pro Eye VR headset was used for this study [13]. The emteqPRO comprises seven facial electromyographic (EMG), a photoplethysmographic (PPG), and an inertial measurement unit (IMU) sensor integrated within a soft frame. From there, all physiological signals were streamed and processed into selected affect metrics. Affect or 'Core affect' refers to primary processes of valence (pleasure) and arousal (activation), which can be inferred from various physiological and behavioural patterns. Valence describes the levels of pleasure or displeasure induced on a continuous scale, ranging from negative to neutral to positive. Arousal is the level of physiological activation induced, ranging from low to high. Both combined create a 2-dimensional space, wherein all emotional states can be illustrated [14]. In addition to arousal and valence, the system provides other derived metrics including heart-rate variability (HRV), breathing rate, and facial expression detection.

After combining the physiological sensing system with the VR set-up, the team conducted a preliminary study to measure the impact of spatialized audio (referred to here as 'SpA') on the user's experience and emotional intensities induced. For this reason, participants were divided into two groups. The first group (Group A) experienced the VR content in dual mono format (referred to here as 'DmA'), while the second group (Group B) experienced the SpA version. The primary hypothesis was that stronger emotional responses are elicited in the SpA group in terms of valence and arousal than in the DmA group. Secondarily it was hypothesised that the phenomenon of 'immersion' is enhanced for the SpA group compared to the DmA treatment. In addition, the most emotionally intense events within the VR experience were also identified and the mean affective scores for those were compared to the creators' expected response ranges, to assess if the affective impact generated by the scenes within the experience matched their expectations.

2 The VR Experience

The selected stimulus for the study was Google Spotlight Stories' 'Pearl', a VR animated short which takes the form of an 'interactive music video'. Pearl was a highly successful product of the Google Spotlight Stories series, winning multiple Emmy and other awards, and had the distinction of being nominated for an Oscar in 2017 for best animated short. The music and sound design were created by Pollen Music Group.

Pearl was selected as a stimulus partly because it is a relatively unknown piece of content, which allows for more emotional variety in response as opposed to more commonly experienced content that may generate culturally or personally normative emotional

responses [15]. Another motivation for selecting Pearl as a stimulus was the breadth of spatial audio techniques employed to achieve the music and sound design goals of the creators of the experience, which were broadly to support the story and to maximize emotional impact. A number of interwoven audio designs were created in the development of the audio schema of the piece, including multiple Ambisonic audio zones, dynamic object-based audio, both head-locked and spatialized music, and DSP effects such as occlusion, reverb, and EQ to highlight and draw attention to certain aspects of the scene. As the soundscape for the film's scenes moves seamlessly between diegetic music and head-locked score, special attention was needed as regards sound source location, as well as to the interplay between sound design and music.

3 Participants

A total of 23 subjects were included in the current study (46% female, 54% male), ranging between 18 and 64 years old (M: 35.56; SD.: 13.86). Of the 23 participants, only 8 of these had experienced VR prior to this study ('little/basic experience'). All participants had normal to corrected vision and did not require the use of glasses. Additionally, participants reported low or no sensitivity to motion sickness and reported normal hearing ability. Audio volume was self-adjusted to participants' subjective hearing levels. Participants did not suffer from any conditions affecting facial muscle movement, nor any cardiovascular/psychological conditions. From the selected subjects, seven users were excluded from heart rate analyses due to high noise (see section 'Data processing'). The study was conducted in two locations over a period of 2 weeks. Both location environments were controlled for noise and light levels and considered suitable for deploying a VR experience.

4 Materials

4.1 Questionnaires

Pre-VR survey. A basic demographic questionnaire was designed and deployed via SurveyMonkey, including questions on age, sex, and experience with VR. Additional screening questions to address the physiological monitoring aspects of the study included: reporting of facial or skin conditions that might interfere with EMG readings, existence of a heart condition (e.g., arrhythmia), and problems with anxiety or claustrophobia.

Post-VR Presence Questionnaire. In order to attempt to correlate the physiological data, a self--report survey using the ITC-Sense Of Presence Inventory (SOPI) [16] was administered following the task. The SOPI is a state questionnaire measure whose development has been informed by previous research on the determinants of presence and current self-report measures. It focuses on users' experiences of media, with no reference to objective system parameters. The framework identifies four factors for analysis: Sense of Physical Space, Engagement, Ecological Validity (Naturalness), and Negative Effects. The SOPI questionnaire was additionally administered prior to the main experience in a priming task with different content to familiarize the participants with the form and structure of the questions.

4.2 Equipment

EmteqPRO. The Emteq 'Lab in a Box' system records data from multiple biosensors that can be utilized for inferring the underlying emotional states of users. The emteqPRO system incorporates one PPG, seven EMG, and an IMU sensor and allows data to be continuously recorded at fixed rates of 1000 Hz.
Photoplethysmogram (PPG) – PPG is an optical measurement method used for monitoring of cardiac activity. The PPG sensor in the emteqPRO is positioned over the centre of the forehead. The signals obtained from this sensor are used for extraction of useful metrics including the measurement of beats per minute and other heart-rate variability (HRV) measures.
Electromyography (EMG) – EMG is based on measurement of the electrical activity generated by muscle contraction. The EMG sensors incorporated in the emteqPRO system are positioned on the zygomaticus, corrugator, frontalis, and orbicularis muscles. The EMG signals, which are acquired from the surface of the skin, are the result of the facial muscles movement activity and provide an insight on the facial expressivity during the VR experience.

Inertial Measurement Unit (IMU) – The IMU integrated in the emteqPRO system contains an accelerometer, a gyroscope, and a magnetometer. Each sensor outputs data along three axes, the x, y, and z. These sensors provide information about motion and allow for head movement tracking.

VIVE Pro Eye. The Vive Pro Eye is a premium VR headset, comprising built-in eye tracking using the SRanipal framework, a 110-degree field of view, a combined resolution of 2880 x 1600 pixels, and a frame refresh rate of 90hz. The audio system uses built-in on-ear 'Hi-Res Audio Certified' headphones, which were used for the study. While the Vive Pro Eye provides eye-tracking capabilities, this data stream was excluded from the study after analysis as it was determined there were no findings of significance.

Supervision. SuperVision is a 64bit, user-facing Windows 10 compatible application created to enable researchers to control, record and monitor the progress of data collection and the sensors' contact with the skin in real-time. It can be also used for post-experience analysis by providing synchronised affective readings and emotional interpretations corresponding to the recorded data uploaded to the AI ML cloud module.

VR Cinema Environment. The environment for the experiments was created in Unity 3D, using the emteqVR SDK for the biosignal monitoring and automatic event annotation. In this environment, the user was situated in the middle of a dark cinema looking towards the area where the screen would be. Self-report questions were presented in the virtual space as shown in Figure 1.

Figure 1. Left: Participant setup. Right: Unity test environment.

The Unity scene was set up with a landing page for facilitator test coding input, whereby the participant would be alternately assigned to group 'A' or 'B'. A custom sequencing system was created in Unity to collect data and timestamps for easy synchronisation. The system also provided checks for efficient calibration and data saving. A timing schema was explicitly designated between each action in the sequence from calibration to stimulus presentation to recreate the pipeline for the data analysis step.

4.3 Study Procedure

A pre-VR screening survey was conducted via SurveyMonkey prior the day of the study. Participants were selected without significant hearing or visual impairments, as well as conditions such as migraine or motion sickness. Those without significant visual impairments but who habitually wear glasses were requested to remove them or to wear contact lenses should they be selected to participate. Normally, it is possible to wear glasses in VR, however, due to the additional padding required by the emteqPRO insert and the need to achieve tight contact between the facial skin and the EMG sensors this was not possible.

On the day of the study, participants were split into two groups, in which the treatment group (Group A) was presented with dual mono, head-locked (DmA) audio, and the control group (Group B) was presented with spatialized, head-tracked (SpA) audio. The definition of 'SpA' for the purposes of this paper is a fully spatialized and head-tracked treatment in Unity VR using the Oculus Spatializer plugin. The audio format was 44.1khZ, 16-bit WAV. The 'DmA' audio consists of a 44.1khz, 16-bit WAV 'dual mono' collapsed mix from the original default viewing position of the spatialized source in Pro Tools. The visual presentation of the VR piece was identical for both groups, consisting of a non-stereoscopic 'lat-long' .mp4 file attached to a spherical texture object with inverted Normals in Unity. To prepare the participant for the study, the emteqPRO attached to the Vive headset was fitted to the head and face. Upon loading the Unity testing environment, a sensor overlay of the facial EMG sensors was displayed and used as a point of reference for ensuring full contact with the headset's sensors. For increased accuracy of

sensor contact monitoring, a GUI within the Supervision application was available instead of the Unity sensor overlay. Once the survey was completed, participants were assisted in removing the VR headset. Overall, the trial took 20 minutes per participant. The overall testing sequence was executed per Figure 2.

4.4 Calibration

At the beginning of each session, the light level (LUX) was measured in the study location to assure minimal extraneous light while the headset was being worn for the purposes of capturing accurate pupil dilation data. In study location A, a light level of 20 LUX was recorded, and this level matched at location B using dimmable LEDs. The sound level (dB-SPL, A-weighted) of the locations was measured to optimise comparability across all sessions. Average ambient noise levels were 35dB SPL-A. Participants wearing glasses were asked to remove them and use contact lenses if needed, and sanitary wipes were provided to remove any makeup if present. Participants were then seated in the centre of the room, positioning them relative to the Vive Pro Eye Lighthouse sensors so as to be oriented towards the virtual cinema screen. After placing the Velcro straps over the head, the headset was tightened to achieve optimal contact between the EMG sensors (Figure 3) and the skin.

Once good contact was achieved, participants went through two onboarding questions of overall mood that day and hours of sleep the previous night, giving ratings from 1 to 5 (1 = strongly disagree; 5 = strongly agree). Next a baseline was recorded using the emteqPRO device. Additionally, an audio calibration was required by adjusting the volume level of a 1kHz test tone to a subjectively comfortable level using a Unity GUI slider in VR.

Figure 3. EmteqPRO VR headset EMG insert.

5 Data processing

Biosignals Pre-processing. The timestamped, processed, and filtered EMG, PPG, and IMU sensor data (1000Hz) were saved into a combined .csv formatted file. Additionally, the timestamped event annotations were saved in a .json formatted file. The Emteq system also estimated PPG quality to measure contact of the PPG sensors (0 = no contact, 1 = full contact). All subjects with less than 0.5 PPG quality for more than 80% of the time during the VR experience were excluded from the analysis; seven subjects were excluded from heart rate analyses due to this criterion. Heart rate time courses were baseline-corrected using data during the first 5 seconds of the VR experience.

SuperVision Valence Model (cloud). To obtain affective insights for each participant's VR experience, emteqPRO's cloud-based SuperVision application was used. The .csv output files containing sensor data from the calibration session (emteqPRO's mask calibration, where the participant's baseline and expressions were recorded), and a companion file containing sensor data from the VR experience session were uploaded and subsequently processed by emteqPRO's AI Emotion Engine, which recognizes affective states in terms of arousal and valence.

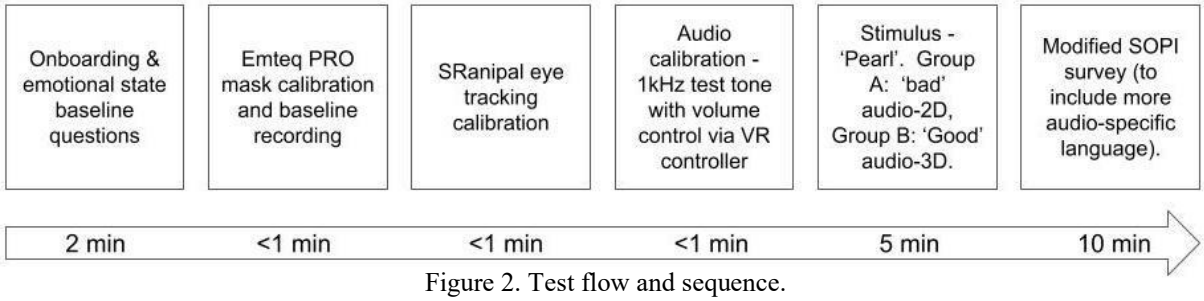

Figure 2. Test flow and sequence.

For the analysis, we were interested only in the output from the valence model, which was provided every 10 seconds. The valence ML model integrated into the AI Emotion Engine recognizes three-level valence – negative, neutral, and positive. Those three classes are derived from the valence rating scale of the Self-Assessment Manikin (SAM) [17]. Namely, the negative class corresponds to valence ratings from 1 to 3, the neutral class ratings from 4 to 6, and the positive class from 7 to 9. Alongside the three-level valence prediction, the model also provided a confidence value for each prediction, representing the model's certainty in the outputted prediction. These values allowed us to produce continuous readings for the valence level in the range from 1 (very negative) to 9 (very positive).

Moving Windows for IMU Data. Axes data (x, y, z) for the gyroscope and accelerometer were analysed separately, i.e., six time series total. For these data, moving averages were applied with 2-second and 1-second bin sizes, respectively. For moving standard deviations, a moving 300ms window for all time series was calculated, followed by a moving average with a 1-second bin size.

Statistical Tests for Time Series Data. For all time series, an unpaired t-test was applied at every time sample to test for differences between treatment and control groups. Significant clusters were defined as 500 or more contiguous time points with significant t-test results ($p < 0.05$, two-tailed).

6 Results

6.1 Experimental Design

To test for the effects of spatial audio on emotional response and sense of immersion during a VR experience, an experiment was conducted with a between-subjects design and two treatment groups: spatialized audio (SpA) and dual mono audio (DmA). Subjects experienced 'Pearl,' a ~5.5 minutes long VR animated short music video, while wearing headphones and the Emteq 'Lab in a Box' headset, which recorded various biosignals, including PPG, EMG, and IMU. After completing the VR experience, subjects also responded to the ITC-SOPI questionnaire [13], which is designed to measure sense of presence while experiencing a piece of media.

6.2 Spatialized Audio Group Reports Higher Sense of Naturalness and Less Negative Effects

The ITC-SOPI questionnaire uses a five-point Likert scale (1 = strongly disagree; 5 = strongly agree) to measure four main factors: (1) 'Spatial Presence', (2) 'Engagement,' (3) 'Naturalness,' and (4) 'Negative Effects.' Figure 4 shows a box plot of group-mean scores for each factor. Because the data are discrete-valued Likert scores and deviated from a normal distribution (Shapiro-Wilk test), we used a nonparametric test (Wilcoxon Rank Sum test) to test for differences between treatment groups per factor. None of these tests found significant differences. To summarize the relative results of these tests: engagement scores were highest among the four factors, with the SpA group reporting slightly higher engagement (U = 0.19, $p = 0.85$, two-tailed); the two largest differences between treatment groups were for Naturalness and Negative effects, with the SpA group reporting a higher feeling of Naturalness (U = 1.22, $p = 0.22$) and less Negative Effects (U = 0.71, $p = 0.48$).

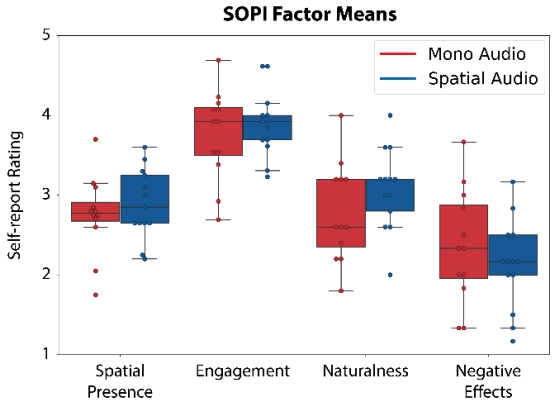

Figure 4. Mean Likert scores for ITC-SOPI factors for DmA (red) and SpA (blue) groups.

6.3 SuperVision Model: Spatialized Audio Group Shows Higher Valence During Key Scenes

To test for the effects of spatialized audio on emotional responses, we used valence score

predictions ranging from 1 to 9 (1 = very negative, 9 = very positive, see 'SuperVision Valence Model (Cloud)' in Methods). Six time clusters were identified with significant differences between the SpA and DmA groups (unpaired t-test, $p < 0.05$, two-tailed,
Table 1). In particular, three of the significant clusters were part of the same time frame spanning 28 seconds from 4:09 to 4:27, during which the SpA group showed higher valence than the DmA group (Figure 5).

Table 1. Time clusters with significant differences in valence predictions between SpA and DmA groups.

Sig. Time Cluster	Mean t-value (p-value)
0:21 – 0:27	t(21) = -2.27, p = .034
0:47 – 0:48	t(21) = -2.13, p = .045
2:08 – 2:09	t(21) = 2.71, p = .013
4:09 – 4:13	t(21) = -2.34, p = .031
4:17 – 4:21	t(21) = -2.38, p = .029
4:23 – 4:27	t(21) = -3.09, p = .009

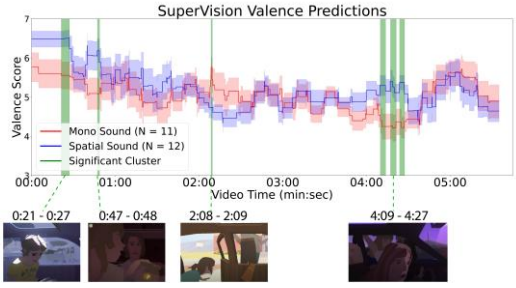

Figure 5. SuperVision Valence predictions across time during the VR experience.

6.4 Relationship between Supervision valence clusters and composers' intended emotion elicitation

Prior to conducting any data analysis and independently of this study, the composers of Pearl attended a retroactive 'spotting session', where they were asked to watch the piece and spontaneously comment on the moments where they intended to 'hit' a specific feeling or emotion. For two significant time clusters found from the Supervision valence model, the composers identified roughly the same time windows as being important moments they had wanted to emphasize musically/sonically. Their comments from the session for these two scenes: 0:47 - "when the daughter catches lightning bugs. Lots of people remember that moment", 4:15 - "Positive moment when she's a teenager vs a kid. In the strobe we see her character change to become the kid again, briefly. But not if you're looking at the tunnel and not her. Blink and you miss it!". These main events in the experience were expected to induce higher valence responses which was the case for the SpA group.

6.5 Spatialized Audio Group Has Higher Heart Rate Throughout the VR Experience

To test for the effects of spatialized audio on arousal and engagement in the VR experience, we plotted the mean heart rate, outputted by the emteqPro system (Figure 6). For the majority of the time during the VR experience — starting from 0:40 through the end of the video — the SpA group showed higher heart rate than the DmA group, with 5 significant clusters of especially large differences between the two groups (unpaired t-test, $p < 0.05$, two-tailed,
Table 2).

Table 2. Time clusters with significant differences in heart rate between SpA and DmA groups.

Sig. Time Cluster	Mean t-value (p-value)
3:00 – 3:13	t(14) = -2.28, p = .040
3:39 – 3:47	t(14) = -2.26, p = .040
4:07 – 4:19	t(14) = 2.38, p = .033
4:40 – 4:48	t(14) = -2.31, p = .036
4:58 – 4:59	t(14) = -2.18, p = .047

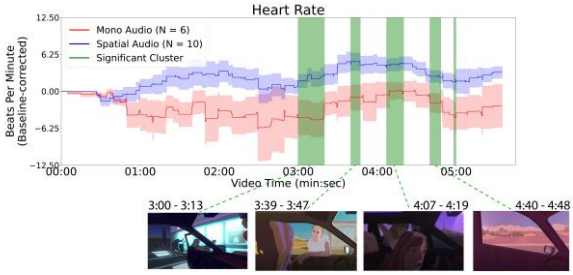

Figure 6. Heart rate (baseline-corrected) across time during the VR experience.

6.6 IMU: Spatialized Audio Group Responds Faster to Sound Cues

IMU data were next analysed to determine if subjects' external responses, i.e., head movements, differed depending on the treatment condition. One effect that appeared during multiple events was faster

acceleration in the SpA group compared to the DmA group. Figure 7 shows three examples of significant time clusters demonstrating this effect (unpaired t-test, $p < 0.05$, two-tailed): (1) 2:21 – 2:23, (2) 3:59 – 4:00, (3) 4:41 – 4:42. Qualitatively, in all three of these scenes in the video, the main sound in each scene is coming from either the far left or right of the scene (e.g., the daughter calling out from the backseat of the car); presumably, the SpA group localised the source of the spatialized sound and thus has a faster response and higher head acceleration in the same direction of the sound.

6.7 IMU: Mono Audio Group Looks Around More During Visually Ambiguous Scenes

To measure short-term variability in head movements, we applied a moving standard deviation window to the IMU data, which resulted in multiple time segments.

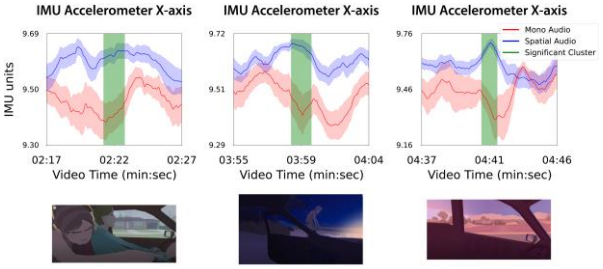

Figure 7. Three examples of scenes when SpA group moved their head faster than DmA.

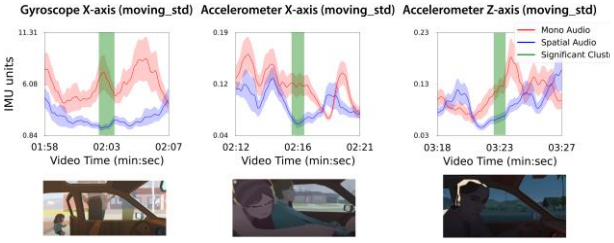

Figure 8. Scenes where DmA moved their head around more than SpA.

Those values in other words indicated how much each group 'looked around' during the VR experience. These time courses revealed differences during events in which the main parts of the scene were ambiguous when relying on visuals alone and when the sound played a pivotal role in directing the viewer's attention (e.g., an argument in the dark with a speaker's voice coming from the left). Figure 8 shows three examples of these scenes, and in all of them, the DmA group has either more head movement (based on the standard deviation of the gyroscope data) or more changes in acceleration of the head (based on the standard deviation of the accelerometer data) (unpaired t-test, $p < 0.05$, two--tailed): (1) 2:02 – 2:03, (2) 2:16 – 2:17, (3) 3:22 – 3:23. Our interpretation of this effect is that the DmA group may be more confused about the events in the scene due to having no spatial source sound information, thus resulting in looking around more during these visually ambiguous scenes.

7 Discussion

This paper presents a novel approach on exploring the effects of spatial audio on continuous physiological, affective, and behavioural responses of users in VR settings, compared to conventional self-reports. Furthermore, it builds on previous important work, particularly in the area of the effectiveness of head-tracked audio spatialization [18,19]. The study's findings validate the extensive research in the free-field domain that shows the extent to which individuals' accuracy and dimensionality of response both on an affective and phenomenological level are enhanced through the ability to interact with the sound field, and it is clear in the context of the current experiment that a similar result can be achieved in VR.

Albeit this was a preliminary investigation study with a relatively small sample size, our findings indicated promising metrics that could be used to inform future studies. Even with this small population, there were a number of results that did crossed the significance threshold during specific event scenes in the VR experience, which suggests that further work in this area with a much larger sample size is likely to yield a diversity of more significant results. There were also some intriguing trends observed in the sub-metrics of HRV and breathing rate that were beyond the scope of this paper. These will be further examined in our next extended study. Our findings are discussed in the following sections.

Spatial Audio affects emotional responses. The SpA group showed stronger valence at key scenes where such responses would be expected (by the creators), and in the direction expected (positive or negative). Normalised heart rate was also higher overall in the SpA group, indicating increased arousal. It was also found that the expected valence magnitude from the point of view of the creators of the piece (i.e., where they considered the emotional high/low points to be, was higher for the SpA group than in the DmA group.

Head movement indicative of spatial audio attention and engagement. The SpA group exhibited less overall head movement as measured by the IMU data, but also a faster orienting response in the correct direction when head movements did occur, indicating an increased capacity for orientation prediction due to heightened localization accuracy. The reduced head movement at other times may indicate longer periods spent focusing on a particular aspect of the scene (fixating), suggesting deeper immersion, focus, or engagement. Increased engagement and presence during those events, were also reflected in the SOPI self-report of the factors of 'naturalness', 'engagement', and 'Spatial Presence', which are higher than in the DmA group. A particular question where SpA scores more highly is 'I felt involved in the scenario', which could be analogous to increased immersion.

Novel VR study approach. In contrast to more traditional stimulus-response studies, once launched into the IVE, participants were free to explore and look around their environment, unconstrained by goal-based tasks or guided interactions. While the latter methods may be more explicitly measurable via a stimulus-response model, the time-series nature of a long-form, emotionally nuanced experience such as Pearl may be well suited to more of a process-based analysis [20].

Limitations. For the current study, 'Pearl' was a monoscopic visual presentation, which means the 360º video did not have the same depth of field as a full VR implementation, and thus possibly contributed to a lack of the same level of perceived immersion as in a stereoscopic presentation. Since the fully interactive 3D environment with object-based audio was converted to a 360 video with Ambisonic rendering (as was done with the study stimulus, for practical reasons as it was not possible to access the source Unity project), it is possible this process would have 'flattened' the mix and created a somewhat less immersive scenario. In addition, the current study involved a 3DoF experience (seated, rather than room-scale where the participant can walk around the experience). It is possible there would be differences in immersion and/or valence & arousal if the participant were able to walk around and physically explore the soundscape.

8 Conclusions

This paper confirms our hypothesis that a sophisticated spatial audio environment with head-tracked music and sound in virtual reality is capable of increasing the magnitude of autonomic arousal and valence when compared to a mono presentation, as well as adhering to the direction of measured emotional response when compared to the creators' original intentions. This finding has important implications for the fields of creative VR production and audio engineering, as well as auditory neuroscience and psychology. From a creative and design standpoint, incorporating such spatial audio schemas adds to the emotional quality of the experience, even if such feelings or perceptions are not explicitly expressed by the audience, and as such warrants further development by creative teams looking to enhance their storytelling and to deepen their emotional impact. It may also be beneficial for creative teams to take advantage of innovations in biosensing technologies to validate creative approaches in development using systems and procedures similar to those described in the study. In terms of the implications for neuro-psychological research into auditory perception, the findings may point to a new direction to be taken in terms of spatial audio's ability to modulate dimensional features of arousal and valence in order to improve patient outcomes for emotional dysregulation in conditions such as PTSD and depression/anxiety disorders.

The complementary finding of increased intentional head movement during auditory attention-led moments and lower overall movement at other times in the spatial audio case is also intriguing and should

be further studied as a potential direct and measurable indicator of engagement and immersion. Although such correlation between levels of immersion and audio-led movement behaviours is limited at this stage, further research may be worth pursuing based on the observations from the current study.

Future work in this area would benefit from creating a custom 3D interactive stimulus that goes beyond 360º video content and 'baked' Ambisonics. Also, additional physiological sensing capabilities such as electrodermal activity (EDA) could be included for the continuous monitoring of arousal responses. In order to increase ecological validity and statistical robustness, a key goal would be to recruit a larger study sample size. Other ways to improve the significance of findings include shortening the length of stimulus, as well as increasing the 'dimensional resolution' of the valence/arousal measures by creating more sonically evocative emotional material, and pre-validating this content for emotional highs/lows via large-scale annotation surveys. By combining these elements in novel ways, it may be possible to attain a more accurate understanding of the underlying processes driving emotional perception of IVEs.

Acknowledgements

The authors would like to thank Stanley Pratt, Rodrigo Marques Almeida de Silva, i2Media Research (Goldsmiths, University of London) and emteq labs team for their guidance, support, and contributions to the development of this work.

References

[1] Marín-Morales, J., Higuera-Trujillo, J.L., Greco, A. et al, "Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors," *Sci Rep*, vol. 8, pp. 13657 (2018).

[2] Richardson, Daniel & Griffin, Nicole & Zaki, Lara & Stephenson, Auburn & Yan, Jiachen & Hogan, John & Skipper, Jeremy & Devlin, Joseph, "Measuring narrative engagement: The heart tells the story," *BioRxiv* [Preprint] (2018).

[3] Bradley MM, Lang PJ, "Affective reactions to acoustic stimuli," *Psychophysiology*, vol. 37, no. 2, pp. 204-15 (2000).

[4] Gatti E, Calzolari E, Maggioni E, Obrist M, "Emotional ratings and skin conductance response to visual, auditory and haptic stimuli," *Sci Data*, vol. 5, pp. 180120 (2018).

[5] Al Alam, R.T. and Dibben, N, "A comparison of presence and emotion between immersive virtual reality and desktop displays for musical multimedia," In: Future Directions of Music Cognition 2021 Virtual Conference Proceedings, Virtual conference, Ohio State University Libraries, Ohio (2021).

[6] Drossos K, Floros A, Giannakoulopoulos A, Kanellopoulos N, "Investigating the impact of sound angular position on the listener affective state," *IEEE Trans Affect Comput*, vol. 6, no. 1, pp. 27–42 (2015).

[7] Y. Hyodo, et al., "Psychophysiological Effect of Immersive Spatial Audio Experience Enhanced Using Sound Field Synthesis," In: 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), Nara, Japan, pp. 1-8 (2021).

[8] Fletcher, M., "The Effect of Spatial Treatment of Music on Listener's Emotional Arousal" in: Journal on the Art of Record Production [Online], Issue 05 (2011).

[9] Poeschl-Guenther, Sandra & Wall, Konstantin & Döring, Nicola, "Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence," *Proceedings - IEEE Virtual Reality*, pp. 129-130 (2013).

[10] Maori Kobayashi, Kanako Ueno, Shiro Ise, "The Effects of Spatialized Sounds on the Sense of Presence in Auditory Virtual Environments: A Psychological and Physiological Study," *Presence: Teleoperators and Virtual Environments*, vol. 24, no. 2, pp. 163–174. (2015).

[11] Wissmath, B., Stricker, D., Weibel, D., Siegenthaler, E., & Mast, F. W., "The illusion of being located in dynamic virtual environments. Can eye movement parameters predict spatial presence?," *Journal of Eye Movement Research*, vol. 3, no. 5. (2010).

[12] Hirway, A., Qiao, Y., & Murray, N., "A QoE and Visual Attention Evaluation on the Influence of Audio in 360 Videos," In *2020 IEEE 21st International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*, pp. 191-193 (2020)

[13] Gnacek, M., Broulidakis, J., Mavridou, I., Fatoorechi, M., Seiss, E., Kostoulas, T., ... & Nduka, C., "EmteqPRO—Fully Integrated Biometric Sensing Array for Non-Invasive Biomedical Research in Virtual Reality," *Frontiers in Virtual Reality*, vol. 3, pp. 781218 (2022).

[14] Posner, Jonathan et al., "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715-34 (2005).

[15] Ali, S. & Peynircioglu, Zehra, "Intensity of Emotions Conveyed and Elicited by Familiar and Unfamiliar Music," *Music Perception*, vol. 27, pp. 177-182 (2010).

[16] Lessiter J, Freeman J, Keogh E, Davidoff J., "A cross-media presence questionnaire: The ITC-sense of presence inventory," *Presence*, vol. 10, pp. 282–297 (2001).

[17] Bradley, M. M., & Lang, P. J., "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49-59 (1994).

[18] Davis, L. S., Duraiswami, R., Grassi, E., Gumerov, N. A., Li, Z., & Zotkin, D. N., "High order spatial audio capture and its binaural head-tracked playback over headphones with HRTF cues," In *Audio Engineering Society Convention 119*. (2005).

[19] Goldenberg, A., Halperin, E., van Zomeren, M., & Gross, J. J., "The process model of group-based emotion: Integrating intergroup emotion and emotion regulation perspectives," *Personality and social psychology review*, vol. 20, no. 2, pp. 118-141 (2016).

[20] Mason, R. D., Kim, C., & Brookes, T., "Taking head movements into account in measurement of spatial attributes," In *Proceedings of the Institute of Acoustics Reproduced Sound Conference*, vol. 30, no. 6, pp. 239-246 (2008).