"When the Elephant Trumps": A Comparative Study on Spatial Audio for Orientation in 360° Videos

Paulo Bala
Madeira Interactive Technologies Institute & Universidade Nova de Lisboa
Funchal, Portugal
paulo.bala@m-iti.org

Raul Masu
Madeira Interactive Technologies Institute & Universidade Nova de Lisboa
Funchal, Portugal
raul.masu@m-iti.org

Valentina Nisi
Madeira Interactive Technologies Institute & Universidade da Madeira
Funchal, Portugal
valentina.nisi@m-iti.org

Nuno Nunes
Madeira Interactive Technologies Institute & IST, University of Lisbon
Funchal, Portugal
njn@m-iti.org

ABSTRACT

Orientation is an emerging issue in cinematic Virtual Reality (VR), as viewers may fail in locating points of interest. Recent strategies to tackle this research problem have investigated the role of cues, specifically diegetic sound effects. In this paper, we examine the use of sound spatialization for orientation purposes, namely by studying different spatialization conditions ("none", "partial", and "full" spatial manipulation) of multitrack soundtracks. We performed a between-subject mixed-methods study with 36 participants, aided by Cue Control, a tool we developed for dynamic spatial sound editing and data collection/analysis. Based on existing literature on orientation cues in 360° and theories on human listening, we discuss situations in which the spatialization was more effective (namely, "full" spatial manipulation both when using only music and when combining music and diegetic effects), and how this can be used by creators of 360° videos.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; **Auditory feedback**; *Empirical studies in HCI*.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300925

KEYWORDS

360° video, Spatial Audio, Cinematic Virtual Reality, Orientation, Virtual Audio Spaces.

ACM Reference Format:
Paulo Bala, Raul Masu, Valentina Nisi, and Nuno Nunes. 2019. "When the Elephant Trumps": A Comparative Study on Spatial Audio for Orientation in 360° Videos. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300925

1 INTRODUCTION

360° video is emerging as a promising media to engage audiences in storytelling, communication, journalism, and marketing among other fields [45]. The immersive nature of the medium provides the viewer with a degree of agency that helps to engage with the content [45]. However, this freedom in orientation introduces unpredictability in the viewer experience, as the audience risks losing essential elements or details [40, 43]. Studies on orientation in cinematic VR towards specific Points of Interest (POIs) are still in their early stages. Nevertheless, they are promising in the potential to improve the quality of VR experiences [28, 29, 31, 38, 44].

Sound often emerges as a helpful element in attracting viewers' attention in a 360° video. For instance, a recent study on storytelling for cinematic VR proposed that "sound plays an important role in peripheral awareness" [37]. Conversely, the uncontrolled use of sound may distract viewers from the primary interest of the plot [1]. Furthermore, studies that focus on the usage of cues demonstrated that sound effects (diegetic cues) could effectively direct attention [17] although the importance of the spatialization of such elements is not conclusive [40]. Therefore, while the spatial rendering of sounds is generally considered as desired for

the immersive and realistic representation of the virtual environment (VE) [22, 41], how to use the spatial properties of sound for notification and orientation purposes is still an open research question.

In this paper, we examine the effect of sound spatial manipulation to direct viewers' attention towards specific POIs in 360° videos. Building on the growing interest in the design of VR applications and 360° video, we ask two high-level research questions (RQs):

- RQ1 - *How can music be spatialized to guide viewers attention in 360° videos?*
- RQ2 - *How can the use of diegetic cues be reinforced by the audio spatialization to guide viewers in 360° videos?*

The questions reflect two typologies of elements in VR, as POIs in a 360° cinematic video can naturally produce sounds (i.e., an animal) or not (i.e., a painting in a museum). Mimicking these two typologies, we chose two tourism 360° videos: a **city** tour (with monuments as silent POIs) and a **safari** (with animals as sounding POIs). Based on the characteristics of the two scenarios, we rely on non-diegetic music for both videos (as a general case, appropriate for cinematic VR). We acknowledge that in scenes with diegetic music, a specific strategy has to be developed, but given the novelty of our approach, we start from a more general case. In the **safari**, we chose to add diegetic sounds as cues over non-diegetic cues, relying on existing works [38, 40].

Mirroring a navigation strategy developed for GPS navigation [21], we designed three spatial audio conditions for each video, according to the level of spatial manipulation (IV with 3 levels "none", "partial", and "full"), as described in Table 1. Finally, we tested both videos with 36 participants (12 participants for each condition). The design of the study relies on *Cue Control*, a tool that facilitates the creation of spatial audio soundtracks for 360° Video, as well as enabling the collection and analysis of captured metrics emerging from the user experience. While adopting a mixed method approach, we combine qualitative and quantitative data, to evaluate and understand participants' behavior.

Our findings led us to frame specific situations in which music manipulation can be effectively used to direct viewers toward POIs in 360° videos. We also contribute to the literature on diegetic sonic elements, by analyzing the importance of spatialization of sound cues and by providing evidence on how audio spatialization can reinforce these elements. We analyze our results against existing theories of musical listening, framing them to suggest implications for design.

2 BACKGROUND AND RELATED WORKS

This research is grounded in attention in VR, and sound perception. Both addressed in the following subsections.

Table 1: Audio Spatialization for the different levels of Spatial Manipulation (C1, C2, and C3)

			Audio Spatialization		
			C1 "none"	**C2** "partial"	**C3** "full"
Audio Cues	City	Track 1: Violin		✓	✓
		Track 2: Piano			✓
	Safari	Track 1: Diegetic Effects		✓	✓
		Track 2: Piano			✓

Attention in VR and the role of sound

Guiding spectators' attention while watching 360° videos is a recent research interest in the field of immersive media, that aims to facilitate the viewers' orientation process and minimize the risk of missing important details. Borji and Itti [5] presented an extensive literature of different models of visual attention, developing a taxonomy of 65 models, and formulating 13 criteria derived from computational and behavioral studies. Specifically for VR, Lin and colleagues [29] developed two methods for focus assistance: Auto Pilot, that brings viewers directly to the target and Visual Guidance, that indicates the direction of the target using arrows. Other techniques take advantage of the hardware setups themselves, such as Xiao et al. [50] and Gugenheimer et al. [19]. In the former, a sparse peripheral display (array of LEDs surrounding the central display) expands the traditional VR field of view and is able to direct attention by rendering only specific POIs in the periphery. In the latter, the SwiVRChair is a swivel chair augmented by a motor and an electromagnetic clutch, that physically directs the viewer toward specific areas of a 360° scene. Another approach was proposed by Nielsen et al. [31] who identified three dimensions of cues that can be used to direct spectators in VR. Firstly, the authors differentiate between explicit cues (that cause top-down shifts in attention) and implicit cues (that instead rely on bottom-up shifts). The second dimension differentiates cues whether or not it limits the user's freedom. Borrowing from film theory, the third dimension discriminates between diegetic cues (elements that belong to the world) and non-diegetic cues. Albeit the cues taxonomy was not specifically targeting audio elements, the idea of diegetic cues has been used to design sonic cues in VR. Following this model, Rothe et al. [40] observed that the presence of new sonic elements induced the viewers to search for the source, but that not all people paid attention to the spatial location of the sound. Similarly, Gödde and colleagues [17] argue that 3D sound is effective for guiding attention and mono sounds can induce the viewer to look for the sound

source. Aligned with this trend, Guhnter et al. [20] presented results supporting that 3D sound reduce the time required to locate objects in a complex virtual environment. Another recent study indicates that binaural audio renderings can introduce advantages to the spatial visual processing and localization [22]. Sheikh et al. [43] tested several unobtrusive techniques for directing a viewer's attention discovering that the combination of audio and visual cues is more potent than visual cues alone. In general, there is a consensus that diegetic sound cues are more effective, when combined with visual cues and movement [38].

Spatial sound in virtual reality

Spatial sound in VR has being studied not only for orientation but from a broader perspective, as in Begault and Trejo's book [3] and Cohen and Barfield's special issue [9], where papers range from spatial sound perception to more applied uses. Grani et al. [18] also present a collection of research focused on two primary areas of sound in VR: the creation of interactive audio experiences, and production of spatial audio for cinema and creative contexts. Concerning manipulation, Jordan et al. [23] proposed a tool for spatially manipulating sound and music in space of a VR environment. Sound in VR has also been studied related to *Presence*. For instance, Serafin and Serafin [42] presented a study that compares the effect on *Presence* of real versus virtual soundscapes, while Poeschl et al. [36] investigated the effect on *Presence* of spatial-sound versus no-sound. In both cases, spatial sound revealed to increase the sense of *Presence*. Walther-Hansen and Grimshaw [47] developed a theoretical framework on the use of sound towards the attainment of *Presence*.

Shifting the function of sound effects: from acousmatic sounds to auditory icons. The importance of diegetic sound effects has been largely acknowledged in traditional cinematics [24]. Among the variety of sound effects, we highlight the role of those sounds defined as acousmatic by Chion [8] in the context of traditional videos: sounds that in the traditional film represent something that is outside the screen (e.g., birds chirping when they are not visible). In the context of 360° videos, these sounds play a different role; they are elements that convey information about events that are happening outside the captured area or the audience's limited point of view. As we have seen above, these elements have been used as diegetic sound cues, specifically directing viewers' attention [38]. In this sense, these sounds became notification elements, denoting the presence of an element of interest outside the view area. This relation between sounds, users, interactive technology, and notification echoes the auditory icon model proposed by Gaver [14], in which sounds were used to notify a user concerning specific states of a computational process. Since the early stages of HCI, sound has been studied for notification purposes. Blatter et al. [4] comprehensively describe sounds with an iconic meaning coining the term Earcons, while Gaver [14] defined the idea of Auditory Icons. In both cases, the representational aspect of sound was studied with the aim of designing sounds that describe the action that is notified. We, therefore, argue that diegetic sonic cues are a subgroup of auditory icons.

Examples of how acousmatic sounds became icons in VR can be found in the literature. For instance, Nordahl [32] discovered that in naturalistic/explorative scenarios, the presence of sound increases the user's amount of movement. Bala et al. [1] observed that the sound of a door in a cinematic VR, was crucial in directing user attention. Similarly, in a study on space in 360° videos, Pope et al. argued that sound plays a fundamental role in peripheral awareness [37].

Sound Perception

Audio as object and events. Studies on VR recognize the importance of acknowledging humans' cognition aspects while dealing with sound and consider different definitions [13]. Sound can be considered "object of the hearing" [35], an event [34], or a combination of the two [33]. The relation between sounds and objects/events appears to be stronger when the sound precisely identifies the object, as in this case, humans tend to describe sounds by the characteristics of their source [27]. Identifying sounds with the objects/events also affects the perception, as our attention tends to interpret everyday sounds by grouping sensory information into clusters based on similarities [13, 30].

Listening modalities. Gaver [15] and Chion [7] developed two different models of the human listening in two fields that contribute to cinematic VR, respectively HCI and Soundtrack. Gaver [15], working on sonic notification in HCI, defined two modalities: Musical Listening and Everyday Listening. In Everyday Listening, people tend to focus on the source rather than on the specific properties of the sound. Chion [7] proposed three main listening modalities: Causal, Semantic, and Reduced Listening. Causal Listening focuses on information about the sound cause; Semantic Listening refers to the process of hearing a language, and Reduced Listening takes sound itself as the primary object considered. We can find similarities between the different modes: Causal Listening mirrors Everyday Listening, and, Reduced Listening echoes Musical Listening. Diegetic cues are a representation of object/event and rely on Everyday Listening.

3 STUDY

Existing literature explored diegetic sonic cues, mainly relying on the representational value of sound and Everyday Listening. In this study, we focus on spatial manipulation of music (RQ1), and music alongside diegetic cues (RQ2). We

were influenced by NavigaTone, a model developed for GPS navigation using spatial manipulation of multitrack music, that demonstrated that the spatialization of multitrack audio could constitute a valid method to orient people [21]. This model could offer a non-intrusive way of directing attention, without sacrificing *presence* or *enjoyment* of the 360° video. In support of our research questions, we developed *Cue Control* a tool for facilitating the creation of spatial soundtracks and the collection and analysis of participants' data. We conducted a mixed-methods study (table 1) using two touristic 360° videos, one for each RQ: **City** video (only with music, RQ1) and **Safari** video (both music and diegetic cues, RQ2). We used three different levels of spatialized auditory elements ("none", "partial", "full"; represented by three conditions, C1, C2 and C3, respectively). C1 ("none") serves as a baseline, as all tracks are head-locked to the user's orientation (meaning that there is no spatial sound); by head-locked we mean that sounds are mixed in the central position with balanced volume. In C2 ("partial"), while a track is head-locked to the user's orientation, the other is responsive. The rationale for the partial type is backed by the NavigaTone model [21] that spacialized the melodic/main track for orientation. In C3 ("full"), both tracks are responsive to the user's orientation (all audio is spatialized).

Cue Control

Cue Control [2] is a prototype design application for spatial soundtracks in 360° video, comprising two components (*Cue Spatializer* and *Cue Playback* plugin). *Cue Control* provides sound object rendering of virtual sound sources in 3D space, using as input mono audio. Built using the Unity (2018.2) game engine, the *Cue Spatializer* (fig. 1) allows for the creation of a cue list representative of the behavior of a soundtrack, over time and space. We used a common interface metaphor in editing or animation software: layers containing keyframes, arranged in a timeline, where layers

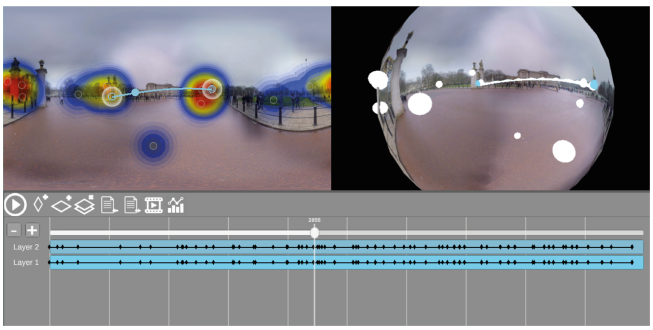

Figure 1: *Cue Control* application showing the timeline interface for soundtrack creation, and heatmaps for analysis.

are representative of an audio source, and keyframes are representative of a change in the audio source (e.g., movement in 3D space). The *Cue Playback* plugin manages the audio playback consistent with the cue list, namely, in the spatialization of tracks as virtual objects to create dynamic and interactive spatial and non spatial soundtracks. Furthermore, *Cue Control* is used for analysis of participant's experience: the *Cue Playback* plugin is responsible for data collection of UX metrics when viewing a video and *Cue Spatializer* provides support for the analysis of that collected data through heatmaps and hotspots.

Soundtrack Creation and Playback. A keyframe is created by direct interaction with the equirectangular texture of a video, being comprised by longitude/latitude (x and y coordinate in the texture), and distance (between the audio source and the origin of the axes, matching the location of the camera). *Cue Control* supports various playback modalities (allowing for dynamic soundtracks that are responsive to the user's viewpoint), but for this study, we only use the static modality (placing sound in the trajectory between the origin and target). At each moment, the current position of a source is calculated based on the start and end keyframe of a segment and the orthodromic distance between those points. Cue lists are exported as XML (Extensible Markup Language) files for use with *Cue Playback* plugin.

Analysis. The *Cue Playback* plugin records head movement (camera rotation quaternion) and exports it in the form of XML files, able to be later retrieved by the experimenter and loaded into *Cue Spatializer*. There, information on the head movement can be over imposed on the 2D visualization of the video and soundtrack in the form of heatmaps for visual analysis. Additionally, head movement data can be transcoded and exported as comma-separated values files for analysis in (spatial) statistics packages.

Media

Using the *Cue Spatializer* prototype, we created soundtracks for two videos. Both videos can be regarded as tourism tools since they depict attractions or activities in a distant location [25]. These videos were chosen because they motivate the exploration and do not have an embedded specific task (e.g., following dialogue between characters). Furthermore, since these videos depict attractions or activities, POIs are easily identifiable, allowing the evaluation of the effectiveness of the spatialized auditory cues in garnering attention to these POIs. Both videos have similar durations (**City** - 4'16"; **Safari** - 3'59"), the same number of scenes, 13, labeled a) to m) all scenes used fixed camera. Music tracks were ad hoc composed by a professional composer (member of the research team), paying attention to keep the most important parameters constant (mode, note ranges, rhythm) [12], aiming to

avoid variations in the viewer's state. While detailing the two videos is out of scope for this paper, we detail the two i) scenes as of interest for the results and discussion.

City video. Depicts a tour of London's famous landmarks from the point of view of a casual tourist. In this case, the POIs are composed of touristic attractions (buildings, monuments, Christmas decorations), elements that do not produce diegetic sounds. Therefore, for this video, we used only music as soundtrack (composed of two tracks, 1. violin and 2. piano). In the **City** video, scene i) depicts the view from the Buckingham Fountains: the camera is placed on a fountain, above the tourists, which is a departure from other scenes; there are two statues, a gold, and a bronze statue, separated by 180° degrees; spatial sound behaves by being stationary in the gold statue, moving to the bronze statue and then being stationary there.

Safari video. Depicts a tour of the African Savanna from the point of view of a tourist on a safari. In this case, the POIs are composed of animals (lions, elephant, birds), elements that produce diegetic sounds. Therefore, for this video, the two tracks were 1. sound effects (diegetic representation of the POIs) and 2. music (piano). For the sound effects of the safari, since these were to be used as icons, we chose natural sound elements (e.g., lion grunts, elephant trumps) stronger in volume as to be recognizable, mirroring post-production methods in standard cinema. In the **Safari** video, scene i) depicts a group of four lions escaping from an elephant: the scene starts with a pack of lions on the right (a lion growl is stationary here, as well as the music track); on the left view, an elephant emerges (the music track moves to the location; after a short time, the accompanying sound of a trump is placed there); the elephant moves closer to the camera, while lions runaway.

Experimental Design

The testing groups were arranged in a between-subjects design: each participant viewed the same one condition for both videos. Independent samples design was chosen due to learnability effects (participants must be naive for each condition). The condition for each participant was randomized, but each condition had an equal number of participants (N=12). For each condition, half of the participants watched the **City** video first to circumvent order effects.

Experimental Procedure. Due to the constraints of the experiment, convenience sampling was used for participant recruitment. Upon arrival, participants were informed about the main goal of the study (without revealing the importance of sound), informed consent form and exclusion criteria. We applied three exclusion criteria to our population: previous conditions refraining them from completing a task in VR, hearing impairment, and professional music education. Only one participant was excluded. Participants were asked to assume a Romberg stance, so that any changes in head movement data reflect postural sway, and given a head-mounted display (HMD) and headphones to view the content. After viewing, they were asked to fill out a questionnaire. Participants then repeated the procedure for the other video. After viewing both videos, participants were then asked to fill out a form with demographic data, and the experimenter proceeded to interview them about their experience viewing the videos. Participants were not financially rewarded for taking part in the study.

Experimental Setup. Participants used a Samsung Galaxy 6 with Gear VR (SM-R322) HMD and V-Moda Crossfade M-100 over-ear headphones with XL cushions. All sessions took place in meeting rooms in our research laboratory.

Measures

After viewing a video, participants were asked to fill out a questionnaire with items on *Presence* (described as the sense of "being there in the environment"[49]), Affect and statements about the video. *Presence* was evaluated through the Igroup Presence Questionnaire (IPQ), a validated *Presence* scale with three subscales and one overall Presence factor (SP, Spatial Presence, the sense of being physically present in the Virtual Environment; INV, Involvement, attention to the real and virtual environment; ER, Experienced Realism, reality judgment of the virtual environment; GP, General Presence). *Positive and Negative Affect* (PA and NA) were measured through the PANAS [48], composed of 20 rating scale items of adjectives with 5 levels, "Very slightly or not at all" to "Extremely". Finally, participants evaluated eight video statements (VS) about the experience of watching the video (VS1 "I enjoyed the video", VS2 "I enjoyed the music", VS3 "The head-mounted display was uncomfortable", VS4 "The headphones were uncomfortable", VS5 "I perceived the audio moving", VS6 "I felt that the video used video filters", VS7 "I expected more audio elements", VS8 "I expected more visual elements"); each statement was a Likert scale item with seven levels (from "Fully disagree" to "Fully agree"). Additionally, user metrics on head direction were collected, imported to *Cue Spatializer* and were converted into *Head movement* (difference in angle, in degrees, between one recording to the next), *Roll* (head tilt), *Yaw* (scan side-to-side), *Pitch* (look up/down), heatmaps and hotspot maps.

After the study, participants were asked to fill out a questionnaire with *demographic* information on gender, date of birth, items on previous experience with VR, 360° VR, 360° video, musical playing and audio production (these last five items were rating scale items with seven levels, "Never" to "Very often"). Furthermore, in order to gather information

on participants' perception and use of audio, we conducted a semi-structured *interview* starting from two specific questions ("Did you use audio to orient yourself?" and "Did you feel audio coming from specific points of the videos?"), then opening the scope by asking participants to provide examples of audio and/or other elements that attract attention and discuss their orientation strategy.

Analysis

Quantitative Data. Data was analyzed using SPSS (v. 25) and used 2-tailed testing at α of 0.05. Testing for Assumption of Normality was done through visual analysis of histograms/boxplots/Q-Q plots, analysis of Kurtosis and Skewness (and their standard errors), and normality tests (Shapiro-Wilk, given the smaller sample size). For **City** video, *Spatial Presence*, *Involvement* and *Positive Affect* data revealed to be normal; for **Safari** video, the same components, plus *Experienced Realism*, were also normal. For data that met parametric assumptions, multivariate analysis of variance (MANOVA) was used to compare conditions and Pearson's product-moment correlation coefficient for correlations; for the remainder of data, Kruskal-Wallis H tests and Median tests were used to compare conditions and Spearman's rank correlation coefficient for correlations. Only significant correlations are reported.

Spatial Data: Heatmaps and Hotspots. Using the *Cue Spatializer* analysis features, two experimenters carried an analysis of the videos, individually reviewing the heatmaps for each condition and proceeding to write notes on user behavior (specifically on the formation of clusters, POIs, and scanning behavior related to the position of spatial audio). The experimenters merged their notes and resolved any conflict that arose by reviewing the heatmaps.

Although the analysis was carried out for all scenes in the videos, we identified specific scenes that were of relevance for further quantitative analysis. The heatmaps revealed a strong cluster formation responsive to the audio spatialization, namely in scene i) for both videos. Moreover, scene i) in both videos is characterized by similar audio spatialization: the new POI is "far away" (>90°) from the previous POI; the audio is stationary before and after the movement, and the change of location is "fast" (<200 frames, ~8 seconds).

For these reasons, we performed a hotspot analysis to quantitatively describe the behavior at the frames identified by the heatmap analysis. This was consistent with Rothe and Hußmann's methodology[39]. Using *Cue Spatializer*, we exported spatial data (longitude and latitude on a sphere) and analyzed it using the geographic information system QGIS 3.2.1, with a Hotspot Analysis plugin (Getis-Ord Gi*), using 2-tailed testing at α of 0.1, 0.05 and 0.01. Each map corresponds to an interval of 5 frames, and the confidence level is represented by a three-color gradient (yellow, orange, red, corresponding to the above α).

Interviews. Interviews were recorded, transcribed, and then analyzed in two phases. First, we quantitatively analyzed the answers to the direct questions, finding how many participants perceive audio as spatialized and how many used audio for orientation. This was repeated for both videos in each condition. The second phase aimed at clarifying more qualitative aspects of the participants' relation with audio and how they oriented in the two videos. We thematically analyzed [6] the statements of the open part of the interview clustered for each condition, by sorting into categories according to the main topic of the statements. Finally, we clustered themes that are common in the different conditions, defining themes common between conditions and themes peculiar to a specific condition.

4 RESULTS

Sample Characteristics

All participants (N=36; 52.8% female) completed their sessions. All conditions had an equal number of participants (N=12). The mean age among participants was 28.59 years (SD=5.96 years; range=19-48). All the items related to previous experience are ordinal, therefore do not respect a normal distribution. In relation to experience with VR, 360° VR and 360° video, most participants (47.2%, 50%, 55.6%, respectively) reported very seldom having experienced them (Mdn=1, IQR=1; Mdn=1, IQR=1; Mdn=1, IQR=0; respectively). Experience with VR was positively and moderately statistically correlated to experience with 360° VR ($rx_s(36)=0.68$, $p<0.0005$) and 360° video ($rx_s(36)=0.61$, $p<0.0005$). Furthermore, a strong and positive correlation was found between experience with 360° VR and 360° video ($rx_s(36)=0.9$, $p<0.0005$). Concerning ability to play a musical instrument or experience with audio production, most participants (52.8% for both) reported never playing/working (Mdn=0, IQR=4; Mdn=0, IQR=3, respectively); a positive and moderate statistically significant correlation was found ($rx_s(36)=0.54$, $p=0.001$) between these items.

Regarding experience with immersive media, our population is consistent with a standard population for whom VR is not yet a popular technology in daily use; we speculate for future studies that a population with experienced VR users might result in better sound localization. Furthermore, the inclusion of population with mostly no musical experience (either playing or producing) is beneficial since results will not be biased towards participants with an experienced hearing ability resulting from musical training. Additionally, we acknowledge that some participants might better at sound localizing [16] but we were not aware of any pre-screening protocols at the time of the study.

City Video

IPQ, PANAS and Head Movement. Table 2 presents the mean scores and standard deviations in IPQ and PANAS components. No statistically significant difference between conditions was found (both for parametric and non-parametric data). SP was strongly positively correlated to INV in two conditions (C1: r(12)=0.82, p=0.001; C3: (r(12)=0.9, p<0.0005) and moderately positively correlated to PA in C2 (r(12)=0.64, p=0.03). Additionally, a moderate and positive correlation between ER and GP in C3 was found (rx_s(12)=0.62, p=0.03). As for mean Head Movement data (Roll, Pitch, Yaw, Angle), no statistically significant differences between conditions were found (using ANOVA) and analyzing a visual distribution of these values over time revealed no visual discrepancies in how participants are expected to behave (higher yaw values representing side to side scanning). Sudden pitch (up/down) changes were noticeable for certain scenes (e.g., looking up at Christmas decorations).

Video Statements. For the statement VS5 "I perceived the audio moving", self-reported values were consistent with the audio spatialization of their conditions; in C1 and C2, most participants (N=5) fully disagreed with the statement (C1: Mdn=1.5, IQR=3; C2: Mdn=1, IQR=3), with only 3 participants in C2 reporting values in the higher end of the scale; in C3, most participants (N=4) fully agreed with the statement (Mdn=4, IQR=5). A Kruskal-Wallis H test showed that there was a statistically significant difference in this statement between conditions, $\chi^2(2) = 6.86$, p=0.03, with a mean rank of 15.79 for C1, 14.88 for C2 and 24.83 for C3.

Table 2: Mean scores and standard deviations for IPQ and PANAS components across conditions for City and Safari video

			C1 "none"	C2 "partial"	C3 "full"
City Video	Presence	GP	3.08 ± 1.78	3.33 ± 1.30	3.75 ± 1.14
		I	3.21 ± 0.69	3.13 ± 0.65	3.19 ± 0.68
		ER	3.00 ± 0.90	2.63 ± 0.91	2.94 ± 1.01
		SP	3.30 ± 1.17	3.55 ± 1.15	3.46 ± 0.98
	PANAS	PA	25.17 ± 8.07	23.75 ± 8.08	26.33 ± 8.55
		NA	11.67 ± 1.61	11.83 ± 2.95	12.92± 6.49
Safari Video	Presence	GP	3.67 ± 1.16	2.75 ± 1.60	3.58 ± 1.31
		I	3.42 ± 1.17	3.12 ± 0.64	3.50 ± 0.71
		ER	2.67 ± 1.06	2.23 ± 0.88	2.92 ± 1.06
		SP	3.63 ± 1.22	3.20 ± 1.10	3.67 ± 0.97
	PANAS	PA	30.00 ± 8.73	21.67 ± 5.98	26.08 ± 7.34
		NA	13.67± 5.79	12.67 ± 5.38	13.25 ± 4.90

Heatmap Analysis. In C1, all participants started by looking down (because of the height placement of the camera) but then corrected the head direction to the horizon; a cluster (N=5) formed on the golden statue, but it dissolved as time progresses. In C2, although participants looked down, no clusters formed and participants did not seem to be reactive to movement of spatial audio. In C3, participants did not react to the change in camera height; a cluster (N=6) formed on the golden statue (when the spatial audio is located there), dispersed and a new cluster (N=5) formed on the bronze statue (when the spatial audio moves to it).

Hotspot Analysis. Based on the heatmap analysis, we identified several timeframe windows in each video for further inspection. Each frame window has a duration of 5 frames (0.2 seconds). Initially, a similar cluster distribution is found in frames 2930-3935 across conditions. In frames 3068-3073, a larger cluster in C3 is formed at the location of the golden statue (where the spatial audio is located and starting to move). In frames 3190-3195, for C3, the largest cluster corresponds to the bronze statue (the new location of the spatial audio), while for C1 and C2, the largest clusters remain on the golden statue. See fig. 2.

Safari Video

IPQ, PANAS and Head Movement. Table 2 presents the mean scores and standard deviations for IPQ and PANAS. No statistically significant differences between conditions were found for IPQ, PANAS, and mean Head Movement data. SP was found to be moderately correlated to INV in all conditions (C1: r(12)=0.61, p=0.036; C2: r(12)=0.62, p=0.021; C3: r(12)=0.59, p=0.043) and strongly correlated to ER in two conditions (C1: r(12)=0.83, p=0.001; C3: r(12)=0.77, p=0.004). Additionally, a moderate and positive correlation between ER and INV in C2 was found (r(12)=0.6, p=0.039). Visual distribution of Head Movement data over time revealed no visual discrepancies in how participants are expected to behave.

Video Statements. For VS5, self-reported values were distributed on the middle to higher end of the scale regardless of their audio spatialization; in C1, most participants (N=5) neither agree or disagree, but a considerable amount of participants (N=4) reported mostly agreeing (Mdn=3, IQR=2); for C2, most participants (N=3) either reported slightly or fully agreeing equally (Mdn=4, IQR=5) while C3, a higher number of participants (N=4) fully agreed with the statement (Mdn=4, IQR=5, for both C2 and C3).

Heatmap Analysis. In C1, participants' attention was initially split between the walking lion and the rest of the pack; a cluster (N=9) forms around the pack, after the elephant trumps (taking around 92 frames ~3.6 seconds). In C2, all participants followed the walking lion; when the elephant trumps,

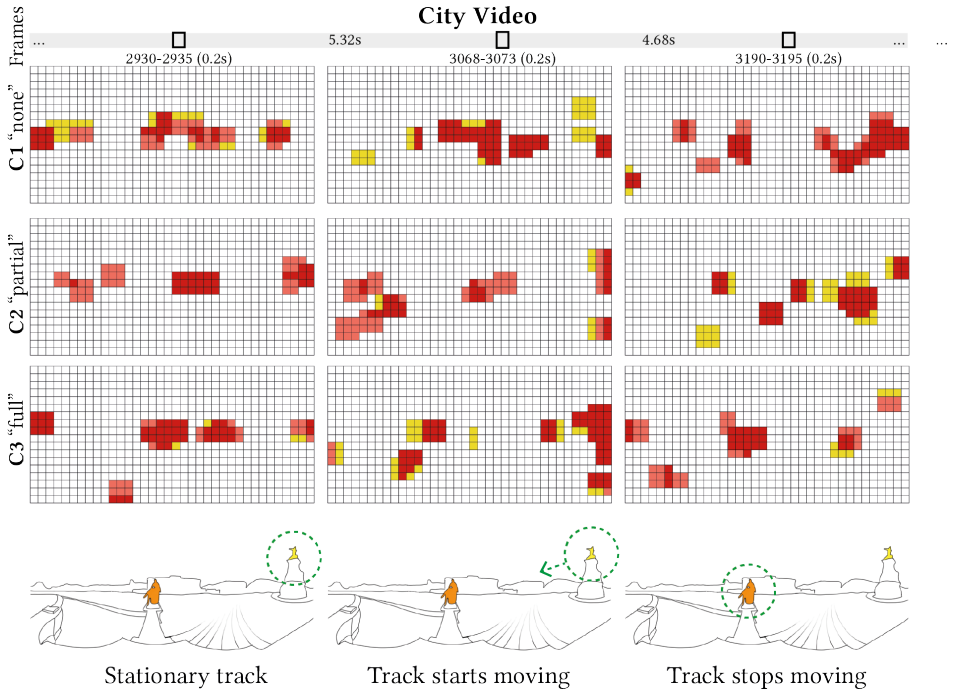

Figure 2: Hotspots for scene i) in City Video. Red, orange and yellow cells represent confidence levels of 99%, 95% and 90%.

only 4 participants remain, while the other 8 clustered on the incoming elephant (taking around 94 frames ~3.9 seconds). In C3, initially all participants started by looking at the lion; as the music moved towards the incoming elephant, a cluster (N=3) formed on it; this cluster grows (N=7) after the trump (taking around 61 frames ~2.4 seconds).

Hotspot Analysis. A similar cluster distribution can be found for the initial timeframe window. During frames 3928-3933, C3 has two similarly sized clusters on the old (lion) and new (elephant) POI, while C1 and C2, have more significant clusters in the old POI. When the elephant stops trumping, C2 and C3 have similar clusters in the old and new POI. See fig. 3.

Post Study Interview

The quantitative analysis of the direct questions showed differences between the two videos. In the **City**, no participants perceived audio as spatialized neither in C1 nor C2, but two participants used audio to orient themselves in C1. In C3, eight participants perceived audio as spatialized, but only four participants declared consistently using audio to orient themselves. In the **Safari**, in C1, most participants (N=9) declared that they used audio to orient themselves. In C2, all the participants (N=12) declared that they used audio to orient themselves, but only eleven perceived it as spatialized. In C3, the majority of participants (N=10) stated that they perceived audio as spatialized and that they used it for orientation. In the thematic analysis, we identified themes shared among the conditions (concerning the diegetic cues, and visual elements) and other themes specific for each condition. Firstly, we report the themes concerning spatialization clustered for each condition (C1 "none", C2 "partial", and C3 "full" spatialization) and then the shared themes.

C1. In C1 we highlight only one theme: *Spatial illusion.* We found 7 instances incorrectly stating that the audio in the **Safari** was spatialized (e.g., P19 "*[I] could tell that the footsteps were coming from a certain direction*").

C2. In C2, we identified two opposite themes. *Correct spatial perception.* 12 instances related to the **Safari**, referred to effects as spatialized (e.g., P8"*the sound of the footsteps [...] was only on one side, and that attracted my attention*", P29"*[...]the elephant [...] and I could feel it in the back* "). *Wrong spatial perception.* 12 instances related to the **City**, referred to music as not spatialized (e.g., P2 "*[I]in the city no [i did not use audio to orient] because everything seemed the same*").

C3. In C3, combining the two scenarios, we highlight three themes.*Complete understanding and usage of music for orientation purposes.* Items from four participants reported that they understood that the music was pointing at the monument, or following the animals and used audio to localize POI (i.e., P3 "*[the music] was kind of pointing at the point of*

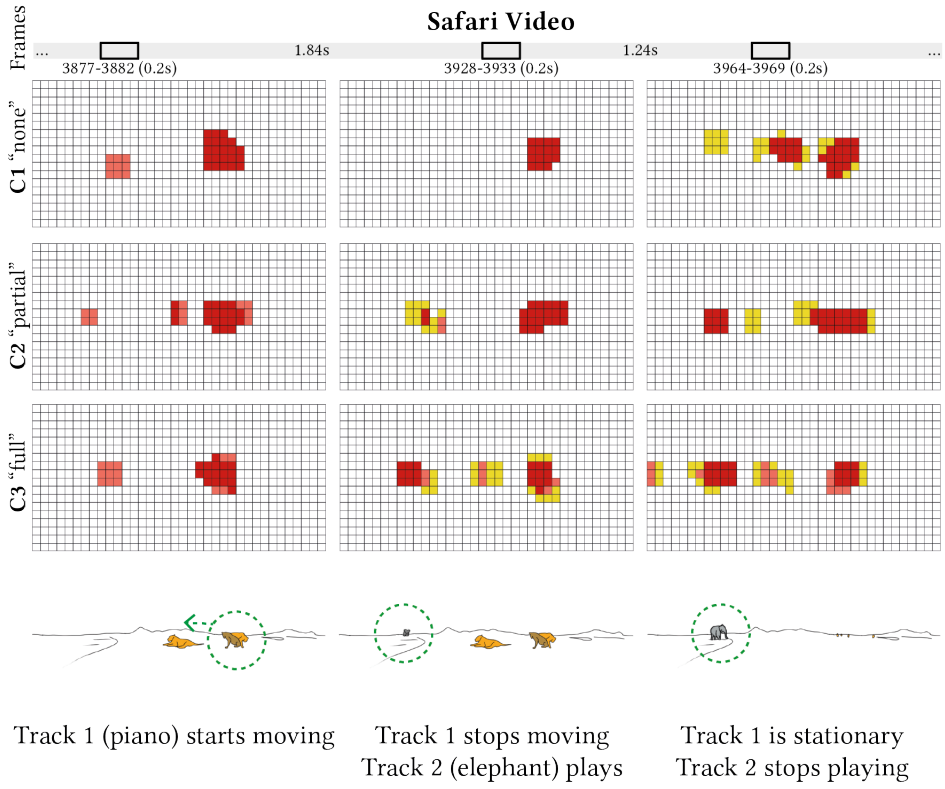

Figure 3: Hotspots for scene i) in Safari video. Red, orange and yellow cells represent confidence levels of 99%, 95% and 90%.

interest [monuments]", P27 "*the music [was] moving with the roars of the lions*"). *Partial understanding of spatialization of music.* We collected items from three participants in both scenarios stating that perceived changes in the location of the music but were not able, or interested in following the music direction (e.g., P33 perceived the music movement only one time and thought that was a way to communicate where to look at; P9 perceived it but was not sure how to use it, because movement was more attracting). *Not understanding spatialization nor usage of audio.* We collected items from five participants in total, who did not perceive nor use the spatial location of the music.

Trans-condition Themes. Diegetic cues. Many statements referred to the semantic content of the audio effect as an important element in the orientation (e.g. P14 "*I heard the sound, and so I tried to see the elephant*"; P7 "*it gave me indicators if there were new animals*"; P8 "*[...] sounds of specific movements*"; P23 "*[i heard] animals that produced sound, and I was using the sound to orient myself*"). Finally, we identified two main themes that refer to visual elements that play a role in the orientation. *Movement.* Regardless of condition, we collected items reporting that participants were attracted by movement both in **City** (e.g., P13 "*I looked at the people going by*") and in **Safari** (e.g., P8 "*the animals moving, I followed those animals*"). *Light.* In the **City**, light also played an important role in orienting people in the night scenes (e.g. P11 "*the lights of the city*", P10 "*a stronger color in the Christmas lights*").

5 DISCUSSION

The main contribution of this paper answers the two proposed RQs. First, we provide evidence of the strengths and weaknesses of the usage of spatial manipulation of music for orientation purposes (RQ1). Secondly, we report findings concerning the spatialization combined with diegetic cues (RQ2). A third contribution is offered by *Cue Control* as a novel support for analysis. We build our discussion, scaffolded on emerged methodological considerations.

Methodological considerations. Studies on 360° videos, such as ours, come against several barriers in the analysis of new techniques. Firstly, although a frequent critique speculates that head direction data (available in low-end HMDs) is not equivalent to eye gaze (requiring expensive HMDs with eye tracking), Sitzmann et al. [44] support that head direction is coupled to eye gaze and can be sufficient to predict saliency with reasonable accuracy. Secondly, the complexity of the

media is a relevant factor since most methodology struggles to deal with the complexity of spatiotemporal data (such as head direction over time). While several studies have faced analysis of head movement for still images [10, 46], for 360° video, since the media is changing over time, there is less methodological support on the analysis and reporting of this dynamic data. Thirdly, existing quantitive and qualitative methods in HCI at times do not meet the idiosyncratic needs of data collected in these studies. The use of quantitative scales after the study, such as *Presence*, have a low level of detail; while they make a judgment on the overall use of a technique, they do not express the behavior during the study. Quantitative scales during the study (e.g., discomfort score while experimenting with a VR sickness reduction technique [11]) could affect the viewing experience in 360° video. Furthermore, both quantitative and qualitative scales may introduce a degree of bias that affects results if they inform the participants about the goal/technique. We address this by introducing other elements/techniques in the video statements and interview, occluding the participants to the real motivation of the study. Finally, existing methods that support spatiotemporal data, such as heatmaps, offer a generalized way to understand user behavior through clustering, but the lack of inferential statistics make it difficult to conclude statistical significance of findings. *Cue Control* played a central role in tackling this issue, allowing us the rapid design and prototype spatial sonic elements, as well as explicitly supporting the quantitative aspects of the mixed-methods approach, through the visualization of heatmaps (with visualizations of spatial audio) and hotspots (treating head direction as geographical data for an analogous QGIS method to Rothe and Hußmann's [39] ArcGIS method).

RQ1 - How can music be spatialized to guide viewers in 360° videos? Our results (from the **City** video) support that partial spatial manipulation (C2) does not, while full manipulation of music (C3) introduces some benefits in directing users attention. We support that partial spatial manipulation (C2) of music revealed to be ineffective with: 1) no participants noticed the spatialization of the violin (from the post-study interview); 2) the perception of audio movement in C2 reported similar and lower values, similarly to C1 (from the post video statements), and 3) in the specific case of scene i) C2, there was no significant cluster formation around the music location (from heatmap and the hotspot analysis).

Complete spatial manipulation of music (C3), instead, is more promising in notifying the location of POIs. We support this with: 1) a statistically significant difference in the self-assessment in the VS5 "I perceived the audio moving" between different conditions, with the C3 having higher values, 2) participants declared that they perceived changes in the location of the music, and used it for locating monuments (thematic analysis), and 3) in scene i), from frame 3190 to 3195, the manipulation of music in C3 created a larger statistically significant cluster around the bronze statue (backed up also by the heatmap observations).

Nevertheless, the techniques did not succeed for the entire number of participants in C3, as evident by the existence of secondary clusters. Moreover, some participants in interviews declared that they did not perceive the music manipulation (supported as well by the video statements). With the objective of finding motivation for this discrepancy among users, we discuss here the main elements that may have interfered with the spatial manipulation. Firstly, the interview analysis provides evidence of the fact that visual elements (e.g., cars and people) may have had a significant impact on the participants. This finding parallels existing studies on visual orientation cues[17].

Concerning Listening modalities, Musical Listening may be a challenging activity, especially for a non-musical population; our sample purposely did not include professional musicians, and more than half of our participants reported having no musical experience at all. From our results, we speculate that the changes in C2 were not strong enough for a non-musical population and also for a sub-group of participants in C3.

RQ2 - How can the use of diegetic cues be reinforced by the audio spatialization to guide viewers in 360° videos? Our results (from the **Safari** video) showed that the spatialization of diegetic cues itself (C2) does not introduce significant benefits in the orientation, while the combination of music and diegetic cues (C3) reveals to be more effective.

The comparison between C1 and C2 highlights that spatial manipulation of the sound effects does not introduce particular benefits in directing attention. This element emerged clearly in scene i) as described in the heatmap and hotspot analysis: when the elephant trumps, participants look for the origin of the sound and found it with a minor difference in reaction time between C1 and C2. Rothe et al. [40] already introduced the hypothesis that spatial location of diegetic cues is not crucial in the orientation as in their study "not all the participants paid attention to the direction of the sound", but did not validate it with a control group (equivalent to C1). Our results confirm their intuition with quantitative evidence and extend it with reaction time comparison.

Moreover, in the interview analysis, we show that diegetic cues were perceived as representations the objects/ events that produce the sounds, as the participants generally described a cue with the animal that produced it, similarly to previous works on sound perception [26]. This underlines the fact that our participants directly connect the sounds with an object/event. The semantic content of the sonic elements appears to be more important than the actual physical

location of the sound, as the sound directly represents the object [35]. For this, we support that participants mainly relied on Everyday Listening, focusing on the object that produced the sound, rather than the physical characteristic (in our case, spatial location). That also may explain why the majority of participants incorrectly declared audio spatialization in C1 (from the interview and video statements).

Our results showcase that the spatialization of music (C3) can highlight the position of elements represented by diegetic cues. Indeed, some participants stated that they perceived the music following the animals (P6 "*audio seems to follow the lions*"). Additionally, participants anticipate the emergence of the elephant, since the cluster formation in C3 (Frame 3928 to 3933) happens before than the clusters in C1 and C2 (Frame 3964 to 3969) in the hotspot analysis. In this case, we argue that participants rely on both Everyday and Musical listening and the two processes support each other, as the semantic content of the cues relies on Everyday listening and the spatialization of the music relies on Musical listening.

To conclude, we highlight the similarities and differences with the study that inspired the usage of multitrack music manipulation for orientation [21]. A first similarity is that spatial manipulation of music is not detrimental to the quality of experience (from *Presence* and PANAS). Similarly to our results, the "stereo condition" (their equivalent to our C3) in the NavigaTone study was more effective in orientation than multitrack manipulation (their equivalent to our C2). However, in our study, evidence showed that C2 was not effective while their multitrack manipulation showed to be effective. We argue that the main reasons for this of this lie behind the context differences: NavigaTone was a task-oriented exercise (participants were primed to the role of the music), and used in the real world, without visual interference.

Implication for Design

Combining results of both videos, we frame suggestions for potential use of spatial manipulation of music for orientation in 360° videos, based on the positive results in scene i). In particular, we present design guidelines for music manipulation: 1) relying on complete manipulation of music rather then multitrack, and 2) creating noticeable changes in the music location: the new POI should be "far away" (>90°) from the previous POI; the audio should be stationary before and after the movement, and the spatial change should be "fast" (<200 frames, ~8 seconds). Concerning diegetic cues, our results confirm that this kind of sonic element constitutes a useful tool for orientation [40] but the spatial location of these sound effects is primary impacting the orientation process. However, spatial manipulation of music can effectively highlight the spatial location of the POIs that produce a given sound.

Future Work

While these guidelines are mainly based on scene i), we do not exclude that others scenes in the videos could have provided other contributions. Given the novelty of the usage of spatial manipulation of audio for orientation purposes in 360° videos, we preferred to analyze in detail one scene in order to provide more solid guidelines, combining quantitative and qualitative methodologies to deal with the limitations of existing methods. One limitation of this study is that we adopted only scenes with fixed cameras, while future research can address other possibilities. Additionally, this work represents a first step in potential research directions besides priming participants to the technique or different combinations of music and diegetic cues. For example, applying the guidelines described to create dynamic soundtracks responsive to the viewer's orientation.

6 CONCLUSION

In this paper, we presented a novel orientation method in 360° video based on spatialization of multitrack audio. We conducted a mixed-methods between-subject study with two videos; for this purpose, we developed *Cue Control*, allowing to easily design spatial cues, while supporting the collection and analysis of user experience metrics. By discussing our results against existing methods and grounding theories on human listening, we frame design implications for 360° video creators (based on our findings of "full" spatial manipulation C3 as a better orientation technique), while outlining future research directions.

ACKNOWLEDGMENTS

The project has been developed as part of MITIExcell (M1420-01-0145-FEDER-000002) and LARSys (UID/EEA/50009/2019). The author Paulo Bala wishes to acknowledge FCT for their support through the Ph.D. Grant PD/BD/128330/2017.

REFERENCES

[1] Paulo Bala, Mara Dionisio, Valentina Nisi, and Nuno Nunes. 2016. IVRUX: A Tool for Analyzing Immersive Narratives in Virtual Reality. In *Interactive Storytelling (Lecture Notes in Computer Science)*, Frank Nack and Andrew S. Gordon (Eds.). Springer International Publishing, 3–11.

[2] Paulo Bala, Raul Masu, Valentina Nisi, and Nuno Nunes. 2018. Cue Control: Interactive Sound Spatialization for 360° Videos. In *International Conference on Interactive Digital Storytelling*. Springer, 333–337.

[3] Durand R. Begault and Leonard J. Trejo. 2000. 3-D sound for virtual reality and multimedia. (2000).

[4] Meera M. Blattner, Denise A. Sumikawa, and Robert M. Greenberg. 1989. Earcons and icons: Their structure and common design principles. *Human-Computer Interaction* 4, 1 (1989), 11–44.

[5] Ali Borji and Laurent Itti. 2013. State-of-the-Art in Visual Attention Modeling. 35, 1 (2013), 185–207. https://doi.org/10.1109/TPAMI.2012.89

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[7] Michel Chion. 2012. The three listening modes. In *The Sound Studies Reader*. Routledge, 48–53.

[8] Michel Chion and Walter Murch. 1994. *Audio-Vision: Sound on Screen* (14th edition ed.). Columbia University Press, New York.

[9] Michael Cohen, Julián Villegas, and Woodrow Barfield. 2015. Special issue on spatial sound in virtual, augmented, and mixed-reality environments. 19, 3 (2015), 147–148. https://doi.org/10.1007/s10055-015-0279-z

[10] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic. 2017. Look around you: Saliency maps for omnidirectional images in VR applications. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, Erfurt, Germany, 1–6. https://doi.org/10.1109/QoMEX.2017.7965634

[11] Ajoy S Fernandes and Steven K. Feiner. 2016. Combating VR sickness through subtle dynamic field-of-view modification. In *2016 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, Greenville, SC, USA, 201–210. https://doi.org/10.1109/3DUI.2016.7460053

[12] Alf Gabrielsson and Erik Lindström. 2010. The role of structure in the musical expression of emotions. *Handbook of music and emotion: Theory, research, applications* 367400 (2010).

[13] Tom A. Garner. 2018. Sound and the Virtual. In *Echoes of Other Worlds: Sound in Virtual Reality*. Springer, 47–82.

[14] William W. Gaver. 1989. The SonicFinder: An interface that uses auditory icons. *Human-Computer Interaction* 4, 1 (1989), 67–94.

[15] William W. Gaver. 1993. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology* 5, 1 (1993), 1–29.

[16] Michele Geronazzo, Erik Sikström, Jari Kleimola, Federico Avanzini, Amalia De Götzen, and Stefania Serafin. 2018. The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios. In *Proc. 17th Ieee International Symposium on Mixed and Augmented Reality (ismar)*. IEEE Computer Society Press, 1–8.

[17] Michael Gödde, Frank Gabler, Dirk Siegmund, and Andreas Braun. 2018. Cinematic Narration in VR–Rethinking Film Conventions for 360 Degrees. In *International Conference on Virtual, Augmented and Mixed Reality*. Springer, 184–201.

[18] Francesco Grani, Dan Overholt, Cumhur Erkut, Steven Gelineck, Georgios Triantafyllidis, Rolf Nordahl, and Stefania Serafin. 2015. Spatial Sound and Multimodal Interaction in Immersive Environments. In *Proceedings of the Audio Mostly 2015 on Interaction With Sound - AM '15* (2015). ACM Press, 1–5. https://doi.org/10.1145/2814895.2814919

[19] Jan Gugenheimer, Dennis Wolf, Gabriel Haas, Sebastian Krebs, and Enrico Rukzio. 2016. SwiVRChair: A Motorized Swivel Chair to Nudge Users' Orientation for 360 Degree Storytelling in Virtual Reality. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (2016). ACM Press, 1996–2000. https://doi.org/10.1145/2858036.2858040

[20] Ryan Gunther, Rick Kazman, and Carolyn MacGregor. 2004. Using 3D sound as a navigational aid in virtual environments. 23, 6 (2004), 435–446. https://doi.org/10.1080/01449290410001723364

[21] Florian Heller and Johannes Schöning. 2018. NavigaTone: Seamlessly Embedding Navigation Cues in Mobile Music Listening. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 637.

[22] Emil R. Hoeg, Lynda J. Gerry, Lui Thomsen, Niels C. Nilsson, and Stefania Serafin. 2017. Binaural sound reduces reaction time in a virtual reality search task. In *Sonic Interactions for Virtual Environments (SIVE), 2017 IEEE 3rd VR Workshop on*. IEEE, 1–4.

[23] Dennis Jordan, Fabian Műller, Colin Drude, Sascha Reinhold, Vanessa Schomakers, and Michael Teistler. 2016. Spatial audio engineering in a virtual reality environment. In *Mensch und Computer 2016 - Tagungsband* (2016), Wolfgang Prinz, Jan Borchers, and Matthias Jarke (Eds.). Gesellschaft fűr Informatik e.V.

[24] Fred Karlin and Rayburn Wright. 2013. *On the track: A guide to contemporary film scoring*. Routledge.

[25] Chelsea Kelling, Heli Väätäjä, and Otto Kauhanen. 2017. Impact of device, context of use, and content on viewing experience of 360-degree tourism video. In *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 211–222.

[26] Guillaume Lemaitre and Laurie M. Heller. 2013. Evidence for a basic level in a taxonomy of everyday action sounds. *Experimental brain research* 226, 2 (2013), 253–264.

[27] Guillaume Lemaitre, Olivier Houix, Nicolas Misdariis, and Patrick Susini. 2010. Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied* 16, 1 (2010), 16.

[28] Yen-Chen Lin, Yung-Ju Chang, Hou-Ning Hu, Hsien-Tzu Cheng, Chi-Wen Huang, and Min Sun. 2017. Tell me where to look: Investigating ways for assisting focus in 360 video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2535–2545.

[29] Yung-Ta Lin, Yi-Chi Liao, Shan-Yuan Teng, Yi-Ju Chung, Liwei Chan, and Bing-Yu Chen. 2017. Outside-In: Visualizing Out-of-Sight Regions-of-Interest in a 360 Video Using Spatial Picture-in-Picture Previews. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. ACM, 255–265.

[30] Brian CJ Moore. 2012. *An introduction to the psychology of hearing*. Brill.

[31] Lasse T. Nielsen, Matias B. Müller, Sune D. Hartmeyer, Troels Ljung, Niels C. Nilsson, Rolf Nordahl, and Stefania Serafin. 2016. Missing the point: an exploration of how to guide users' attention during cinematic virtual reality. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*. ACM, 229–232.

[32] Rolf Nordahl. 2010. Evaluating environmental sounds from a presence perspective for virtual reality applications. *EURASIP Journal on Audio, Speech, and Music Processing* 2010 (2010), 4.

[33] Matthew Nudds and Casey O'Callaghan. 2009. *Sounds and perception: New philosophical essays*. Oxford University Press.

[34] Casey O'Callaghan. 2009. Sounds and Events. In *Sounds and Perception*. Oxford University Press., Oxford, 26–49.

[35] Robert Pasnau. 1999. What is sound? *The Philosophical Quarterly* 49, 196 (1999), 309–324.

[36] Sandra Poeschl, Konstantin Wall, and Nicola Doering. 2013. Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence. In *2013 IEEE Virtual Reality (VR)* (2013-03). IEEE, 129–130. https://doi.org/10.1109/VR.2013.6549396

[37] Vanessa C. Pope, Robert Dawes, Florian Schweiger, and Alia Sheikh. 2017. The Geometry of Storytelling: Theatrical Use of Space for 360-degree Videos and Virtual Reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 4468–4478. https://doi.org/10.1145/3025453.3025581

[38] Sylvia Rothe and Heinrich Hußmann. 2018. Guiding the Viewer in Cinematic Virtual Reality by Diegetic Cues. In *Augmented Reality, Virtual Reality, and Computer Graphics (Lecture Notes in Computer Science)*. Springer, Cham, 101–117. https://doi.org/10.1007/978-3-319-95270-3_7

[39] Sylvia Rothe and Heinrich Hußmann. 2018. Spatial statistics for analyzing data in cinematic virtual reality. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces - AVI '18*. ACM Press, Castiglione della Pescaia, Grosseto, Italy, 1–3. https://doi.org/

10.1145/3206505.3206561

[40] Sylvia Rothe, Heinrich Hußmann, and Mathias Allary. 2017. Diegetic cues for guiding the viewer in cinematic virtual reality. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology - VRST '17*. ACM Press, Gothenburg, Sweden, 1–2. https://doi.org/10.1145/3139131.3143421

[41] Stefania Serafin, Michele Geronazzo, Cumhur Erkut, Niels C. Nilsson, and Rolf Nordahl. 2018. Sonic Interactions in Virtual Reality: State of the Art, Current Challenges, and Future Directions. *IEEE computer graphics and applications* 38, 2 (2018), 31–43.

[42] Stefania Serafin and Giovanni Serafin. 2004. Sound Design to Enhance Presence in Photorealistic Virtual Reality. In *ICAD 2004: The 10th Meeting of the International Conference on Auditory Display, Sydney, Australia, July 6-9 2004, Proceedings* (2004). http://www.icad.org/websiteV2.0/Conferences/ICAD2004/posters/serafin_serafin.pdf

[43] Alia Sheikh, Andy Brown, Zillah Watson, and Michael Evans. 2016. Directing attention in 360-degree video. (2016).

[44] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: How Do People Explore Virtual Environments? *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (April 2018), 1633–1642. https://doi.org/10.1109/TVCG.2018.2793599

[45] Mel Slater and Maria V. Sanchez-Vives. 2016. Enhancing Our Lives with Immersive Virtual Reality. *Frontiers in Robotics and AI* 3 (Dec. 2016). https://doi.org/10.3389/frobt.2016.00074

[46] Evgeniy Upenik and Touradj Ebrahimi. 2017. A simple method to obtain visual attention data in head mounted virtual reality. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, Hong Kong, Hong Kong, 73–78. https://doi.org/10.1109/ICMEW.2017.8026231

[47] Mads Walther-Hansen and Mark Grimshaw. 2016. Being in a Virtual World: Presence, Environment, Salience, Sound. In *Proceedings of the Audio Mostly 2016 - AM '16* (2016). ACM Press, 77–84. https://doi.org/10.1145/2986416.2986425

[48] David Watson, Lee A. Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology* 54, 6 (1988), 1063–1070. https://doi.org/10.1037//0022-3514.54.6.1063

[49] Bob G. Witmer and Michael J. Singer. 1998. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments* 7, 3 (June 1998), 225–240. https://doi.org/10.1162/105474698565686

[50] Robert Xiao and Hrvoje Benko. 2016. Augmenting the Field-of-View of Head-Mounted Displays with Sparse Peripheral Displays. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (2016). ACM Press, 1221–1232. https://doi.org/10.1145/2858036.2858212