Influence of directional sound cues on users' exploration across 360° movie cuts

Belen Masia
Universidad de Zaragoza, I3A

Javier Camon
Universidad de Zaragoza, I3A

Diego Gutierrez
Universidad de Zaragoza, I3A

Ana Serrano
Universidad de Zaragoza, I3A, Max Planck Institute for Informatics

***Abstract*—Virtual reality (VR) is a powerful medium for 360° storytelling, yet content creators are still in the process of developing cinematographic rules for effectively communicating stories in VR. Traditional cinematography has relied for over a century in well-established techniques for editing, and one of the most recurrent resources for this are cinematic cuts that allow content creators to seamlessly transition between scenes. One fundamental assumption of these techniques is that the content creator can control the camera, however, this assumption breaks in VR: users are free to explore the 360° around them. Recent works have studied the effectiveness of different cuts in 360° content, but the effect of directional sound cues while experiencing these cuts has been less explored. In this work, we provide the first systematic analysis of the influence of directional sound cues in users' behavior across 360° movie cuts, providing insights that can have an impact on deriving conventions for VR storytelling.**

CINEMATIC VIRTUAL REALITY experiences are richer and intrinsically different than those provided by traditional displays. Since the user has control of the camera and can choose in which direction to look, it becomes crucial to understand how users explore their environment in these new experiences. Therefore, analyzing user behavior and attention in 360° environments has become a growing topic of interest during the last years [1], [2], [3]. Among the different types of content shown in VR, narrative experiences are of crucial relevance. However, they present a big challenge for the medium and have led to great controversy among VR practitioners about good practices and

guidelines for content creation. Analyzing user behavior and attention in this content can thus help derive conventions for VR cinematography and content creation from first principles.

In traditional cinematography, *cuts* are a fundamental and well-established technique to switch between different scenes and sequences, and to elicit different feelings in the spectator [4]. This technique has been developed over many decades, and it strongly depends on the director's placement of the camera before and after the cut. However, cuts in VR have a number of implications compared to traditional media, mainly due to three reasons. First, users are not only passively watching the scene on a screen, but are immersed in it; therefore, when a cut takes place they get suddenly teleported to a different location without any forewarning. Second, users in VR have active control over the orientation of the camera and can create their own experiences; due to this, it becomes necessary to make assumptions about which regions of the scene they will be paying attention in order to effectively convey a story, resulting in a much more complex endeavor. Third, the multimodal nature of the experience takes on a new dimension in VR: Sound has played a key role in traditional cinematography for decades [5]; in VR, auditory information can be spatialized in the 360° environment together with the visual input, leading to a more complex interaction between auditory and visual stimuli, and their influence on user attention.

In the context of VR cinematography, previous works have studied the impact of different types of cuts in users' behavior [6], [7], [8]; however, none of these works has analyzed the influence of directional sound cues in such cuts. Conversely, works have studied the influence of sound in users' attention during 360° viewing [9], [10], but never in the context of analyzing 360° movie cuts. In this work we aim to bridge this gap, exploring to what extent directional sound cues play a key role in user behavior during cuts in 360° movies in VR.

We focus on diegetic directional sound cues as described in film-theory and previous work [9]. Diegetic cues are those that come from elements present in the scene (e.g., a character playing an instrument), while non-diegetic cues come from outside of the scene (e.g., a narrator). Previous research has shown that diegetic methods typically perform well in cinematic VR since they provide useful cues without breaking the immersive experience [3], [11], [12]. Visual diegetic cues have been extensively used in traditional cinema for drawing attention [4]. In VR, direct use of these visual techniques is only possible if they are present within the field of view that the user chooses to explore. Instead, it is always possible to hear diegetic directional sound cues regardless of the field of view; therefore, they may be more effective [3].

We perform an experiment to evaluate whether the presence of directional sound cues has a significant influence on users' viewing behavior across 360° movie cuts. Additionally, we also explore the influence of the configuration and alignment of the regions of interest (ROIs) in the scene. These ROIs correspond to focal points of the scene, areas that are designed to draw the attention of the viewer (e.g., the main character, or a part of the scene relevant for the action). Using quantitative metrics of users' viewing behavior, our analysis leads to a number of findings with potential implications for VR content creation. Our insights indicate that, in the context of cinematic cuts, directional sound decreases significantly the time that users take to converge to the main region of interest (ROI) after a cut: this insight can be directly used as guidance for establishing the editing pace. We have found, in accordance with previous work [6], that multiple ROIs before the cut elicit a more exploratory behavior after the cut. In the presence of directional sound cues this behavior holds, but users' gaze paths are more predictable than without directional sound. This is likely due to them being guided by the directionality of the sound. Our results also suggest that in the presence of directional sound cues, when multiple ROIs are in different fields of view, users pursue a more exploratory behavior rather than locking their attention to a single ROI. We hypothesize that the directional sound cues are key for identifying the presence of these multiple ROIs when they are not visible in the current field of view, which is in agreement with previous research in different fields showing that directional sound can be effectively used for attention guidance [13],

[9]. Finally, we interview practitioners in the field and provide a brief discussion of their insights.

To our knowledge, our work is the first to attempt a systematic analysis on the influence of directional sound cues in users' exploration patterns during 360° movie cuts. We believe that our findings are a step forward towards understanding how to effectively communicate stories in VR, and we will make our data available to foster new research in this direction.

Related work

Multimodality in VR environments.
In traditional displays, directional sound can help boost the feeling of presence by making viewers feel there are other elements in the scene surrounding them; however, the only available visual information is presented on the screen in front of the viewer. In contrast, in 360° VR environments users are completely surrounded by visual stimuli, and directional sound has the potential of acting as a powerful cue for attention guidance. Multimodality, including auditory and visual information, plays a key role in VR scenarios. It has the potential to improve the experiences, but at the same time brings in new challenges, such as the need for adequate coherence between sensory inputs [14]. In the particular case of audio-visual interaction, Steuer et al. [15] defined sound as one of the key elements for presence in VR environments. In order to enhance presence, directional audio and sound quality are particularly important [16]. Therefore, in the last years, there has been a number of works analyzing several aspects of VR content generation featuring directional sound. For example, Bala et al. [17] proposed an audio editor that facilitates the creation of soundtracks for 360° videos by letting the user control the location of the sonic elements. Further, several researchers have reported that directional sound can improve efficiency in different tasks. In particular, Grohn et al. [13] have shown that auditory cues alone can be sufficient for successfully navigating virtual environments. Given the importance of reproducing the spatial properties of sound, a large body of research has been devoted to the sonification of spatial data, and different methods have been proposed for estimating and reproducing spatial audio for real-world rooms [18], and captured 360° content [19], [20].

However, while directional sound is indeed a powerful cue that has an impact in users' experience in virtual environments [21], [22], how to use the spatial properties of sound for supporting cuts in 360° cinematographic content, where narrative plays an important role, is still an important open research question.

Cinematic VR.
Several works have made efforts towards analyzing users' attention in VR cinematography, typically focusing on *visual* attention. Ferghail et al. [8] analyze different cuts and how they affect storytelling by identifying which scene elements attract users' attention. Following a similar line of work, Knorr et al. [23] include in their analyses the director's intended viewing orientation in order to compare it with the actual exploration patterns followed by users. Serrano et al. [6] study user behavior under different types of cinematic cuts and different dispositions of the regions of interest in the scene before and after the cuts, introducing a set of metrics for quantifying exploration patterns and user behavior. Later, Marañes et al. [7], built upon this work and proposed a similar analysis in the context of professionally edited narrative VR. Despite all the valuable insights provided by these works, all have focused on understanding users' behavior by parameterizing the space of visual stimuli only, leaving out the potential influence of directional sound cues. Recently, Rothe et al. [9] explored three different cinematic methods for guiding users' attention: light, movement, and sound, showing that the attention of the viewer can be directed through the last two, and that reproducing new sounds may induce the viewer to search for the source of the sound. Sheikh et al. [10] analyzed the effectiveness of different unobtrusive attention directing techniques embedded in the narrative (such as motion across main characters, or following gestural and audio cues), concluding that audio cues alert the viewer that there is something important to see, while having the advantage that no assumption is made about the viewer's focus of attention at the time of the cue.

In contrast with these works, we focus on studying the influence of directional sound cues in VR movie editing techniques. In particular, we seek to analyze and quantify the extent to which

directional sound cues influence users' patterns of exploration while experiencing movie cuts in cinematographic content in VR.

Driving attention in cinematic VR.
Previous research has shown that diegetic cues typically perform well in cinematic VR since they provide useful cues for driving attention without breaking the immersive experience [3], [10], [11]. Visual diegetic cues, such as movements, lights, and characters, are well known for drawing attention in traditional movies [4]. However, these cues can only be used if they are in the field of view. Sound coming from the direction of a region of interest can be more effective in 360° visualization, since even though the source is not visible, it is possible to hear it. However, there might be the case of story parts in which maybe no suitable cues (neither visual nor acoustic) are naturally present in the narrative. In these cases, if it is still necessary to guide the attention to some regions of the scene, non-diegetic methods such as halos or arrows can be considered [24], at the risk of breaking immersion. Other alternatives, such as introducing *interactive shot orientation* so viewers can quickly re-orient to the important content by pressing a button have been presented [25]. This, however, has been shown to diminish users' exploratory behavior, since it minimizes the portion of shot traversed. As the authors discuss in their paper, in some cases it is important to allow users to traverse the scene themselves to find the regions of interest (for instance, video creators can build suspense in a horror scene by eliciting exploratory behaviors).

Although there are still many open questions in this field, it has been argued that storytellers should not try to artificially force viewers to look where they want them to look, but rather guide them through the story using visual and auditory cues integrated with the narrative (diegetic cues), if the narrative allows it. Therefore, in this work we focus on diegetic methods, in particular, diegetic directional sound cues.

Measuring the influence of sound

Our goal is to analyze user behavior across movie cut boundaries in the presence of diegetic directional sound cues. In this section we describe the procedure we have followed to assess the effects of introducing directional sound cues in viewers' behavior during cuts when viewing 360° cinematic content. Given the high dimensionality of the space composed by all potential cuts that could take place in a cinematic video, we chose to follow previous work [6] and characterize this space as a function of the regions of interest (ROIs) present in the video. ROIs are described following common terminology as the particular regions of the 360° in which the main action takes place (see Figure 1 for examples). In particular, we focus on the following main parameters (variables of influence): the presence or absence of directional sound cues; the misalignment between the region of interest (ROI) before and after the cut; and the number and location of such ROI before, and after the cut. The goal of this parameterization is to cover a comprehensive set of simple but commonly used scene setups, since covering in a single work all the space of potential cuts would become intractable. In addition to these parameters, we include as main condition to test the presence of directional sound cues. To compare our results against previous findings, we use as baseline for our analysis the data made publicly available by Serrano et al. [6], being the key difference of our work with respect to this previous work that their videos were not presented with directional sound. Please note that we carefully design our experiment in order to strictly replicate the conditions of this previous work, so we can compare our results to theirs. In the following we describe in more detail our stimuli, variables of influence, and procedure.

Stimuli and variables of influence

We use as stimuli a subset of the stimuli presented in the work of Serrano et al. [6][1]. Each stimulus (clip) is created from 360° monocular high-quality videos, with a resolution of 3840 × 1920 pixels, and consists of two shots of six seconds each, separated by a cut, i.e., each stimulus spans a total of 12 seconds. The videos depict four different scenes (Stairs, Kitchen, Living Room, Study) and were recorded using two different rigs: a *GoPro Omni* (a 360° video rig consisting of six GoPro Hero4 cameras), and a

[1]Their videos and data are publicly available at http://webdiis.unizar.es/~aserrano/projects/VR-cinematography

Figure 1. ROI configurations covered in our experiment. We analyze three possible ROI configurations for the scene before and after the cut in which the scene can contain a single ROI, two ROIs whithin the same field of view, or two ROIs in different fields of view. ROIs are marked with an orange square, while the blue speaker icon indicates that each of these ROIs has an spatially-aligned associated sound source.

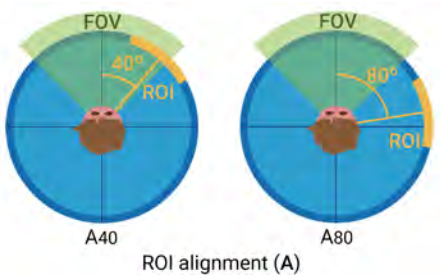

Figure 2. ROI alignments covered in our experiment. We analyze two different ROI alignments in which the ROI after the cut is misaligned by $40°$ or $80°$ with respect to the ROI before the cut. The depicted field of view (FOV) encompasses $90°$.

Freedom360 rig (with three GoPro Hero4). The sound cues corresponding to each scene were recorded using a *Zoom F8* recorder with wireless microphones. All the performers appearing in the scene (sound sources) wore a microphone, therefore each of the sources was recorded as separate sound tracks. The recorded sounds were those naturally produced by the performers while carrying out simple tasks such as washing the dishes or stirring coffee. Note that, since we record the sounds with local microphones, some effects related to sound transport throughout the scene that would be experienced from the camera's point of view are not included in our stimuli. We are therefore assuming that global effects such as reverberation do not play a relevant role in our scenario. The experiment was designed in Unity, and each of the sound sources (playing each of the performers' tracks) was played by a virtual speaker manually aligned with the 360° video feed so that the direction of each sound cue was spatially accurate. There are several spatializer plugins for creating directional sound in VR (including Oculus Spatializer, Google Resonance, or Steam Audio). For sound spatialization in this paper, the Head-Related Transfer Function (HRTF) was simulated through the default Unity Audio Spatializer, using the head orientation from the default Oculus tracking.

Our variables of influence, borrowing terminology from Serrano et al. [6], are the following:

*Alignment of ROIs (*A*).*

We test two different alignment conditions for the regions of interest (Figure 2) : (i) a misalignment between the ROIs before and after the cut that is just within the field of view of the HMD, such that the viewer can have a hint of where the new action is taking place; as in previous work, we chose 40° of misalignment. And (ii) a misalignment that is outside the field of view; again following previous work, we chose 80 °. These conditions are named $A = \{A_{40}, A_{80}\}$.

ROI configuration ($R_{\{b|a\}}$).

To analyze different configurations of the ROIs (number and location) before and after the cut, two variables are introduced, R_b, and R_a respectively (Figure 1). This variable is extremely hard to parameterize due to the large amount of possible configurations it may take, so we restrict our configurations to those of Serrano et al.: a single ROI ($R_{\{b|a\},0}$), two ROIs within the same field of view ($R_{\{b|a\},1}$), and two ROIs not sharing the same field of view ($R_{\{b|a\},2}$). All the possible configurations of R_b and R_a yield a total of nine different conditions.

Figure 3. Example of users' gaze for a frame sequence after the cut with (ours, top), and without (Serrano et al., bottom) directional sound cues. In the first frame we mark the ROIs (orange) and their associated sound source (blue icon); color points represent gaze samples from different users. This frame sequence shows gaze patterns after a cut with a ROI misalignment of $80°$ (A_{80}): notice how users converge faster to the main action in the presence of directional sound cues.

*Directional sound cues (*D*).*
This is the main variable in our experiments, as we seek to analyze the influence of directional sound cues in users' exploration patterns during cuts. This condition can take values such that *D* = $\{Dir, nDir\}$ depending on whether the stimulus presents directional sound cues or not. Note that *D* = $\{nDir\}$ (no directional sound cues) is equivalent to the condition analyzed in Serrano's work [6], so we use their publicly available data for our analysis and comparisons.

Summary.
In order to sample the described parameters of influence we include three different clips per condition, resulting on a set of 54 clips (2 (alignments) x 9 (ROI configurations) x 3 (clips) = 54 stimuli). The subset of clips has been chosen so that sound sources (performers) are placed at similar distances to the observer, thus having comparable intensities, in order to avoid attention biases.

Procedure

We displayed our stimuli on an Oculus DK2[2] equipped with a Pupil Labs[3] binocular eye tracker, recording data at 120 Hz with an accuracy of $1°$ of visual angle. We use a pair of stereo isolation headphones to reproduce the sound (*Vic Firth SIH1*). Users were standing during the experiment to facilitate the visualization of the $360°$ scenes. A total of 35 subjects voluntarily participated in the experiment (8 female, age 26.4 ± 5.2 years old), and all of them reported normal or corrected-to-normal vision. The procedure included an eye-tracking calibration step before starting the experiment which was repeated in case the calibration was not successful. To avoid fatigue, each user was presented only a subset with 27 randomly selected trials of the total 54 different clips. Following Sitzmann et al. [1], before starting the visualization of each of the trials, a gray background with a red square was displayed; users had to align their head direction (displayed as a black cross) in order to launch a new trial, which would start 500 ms after successful alignment. This allows to ensure the same starting alignment for all the participants, and it also allows for resting during the experiment if needed. Users were asked to freely watch the scene while trying to follow the sequence of actions being performed. The Unity game engine was used to carry out the experiments; manual alignment of the sound sources ensured a perfect alignment to their corresponding positions

[2]We use this device to perfectly reproduce the viewing conditions of Serrano et al. [6], since we use their data as baseline for our analysis.
[3]https://pupil-labs.com/

in the video. The experiment had a duration of approximately 20 minutes, including a prior questionnaire about the subject's visual health and previous experience with VR. We recorded raw head and gaze samples, performed outlier rejection, and computed fixations and scanpaths (sequences of gaze samples over time) as described in Appendix A. From this processed data, we analyze users' behavior across cuts with the metrics described in the next section.

Metrics

To analyze the influence of stereo directional sound cues we compare our data to the baseline provided in previous work (without directional sound) [6], therefore we use the metrics presented in their work. We include here a brief description of such metrics.

Frames to reach a ROI (framesToROI).
This metric indicates the number of frames that users spend until they fixate on a ROI after the occurrence of the cut. It serves as an indicator of the time of convergence to the main action after the cut.

Percentage of total fixations inside the ROI (percFixInside).
This metric computes the percentage of the total fixations that fall inside the region of interest after fixating on the ROI after the cut. It is thus independent of the time that takes users to find the ROI, and it only accounts for the behavioral effects once the ROI is found again. Intuitively, this metric can be seen as an indicator of the interest of the viewer in the ROI(s).

Number of fixations (nFix).
This metric is computed as the ratio between the number of fixations and the total number of gaze samples (after fixating on the ROI after the cut). This metric accounts for the number of saccades and fixations performed by users (a lower value corresponding to a higher amount of saccades), and can be seen as an indicative of the exploratory behavior of the user (more saccades corresponding to more exploration).

Scanpath error (scanpathError).
This metric is computed as the RMSE error between a *baseline scanpath* and the scanpaths calculated for each cut. It shows how gaze behavior is altered by the cut itself. Again, this metric is computed after fixating on the ROI after the cut, in order to make it independent of the time in which the user searches for the ROI (framesToROI). For computing the *baseline scanpath*, eye-tracking data from ten different subjects watching the unedited video sequences is used (i.e., users visualize the full scene without cuts, therefore avoiding their influence). Baseline scanpaths are computed as the mean scanpath of all users, for each of the videos. Serrano et al. [6] demonstrated that a high consistency between users was achieved (by analyzing the *Inter-Observer Congruency*), ensuring that this data can be effectively used as a behavioral baseline of the unedited clips.

Analysis and results

Given that each user sees a different subset of the conditions, we cannot assume that the observations are independent in our collected data. We thus employ multilevel modeling in our analysis, which is preferred for related data. Multilevel modeling accounts for random effects among the predictors, so it allows us to include each particular subject as a random effect (modeled as a random intercept). For all our metrics the effect of the subject was found significant ($p < 0.05$ in Wald's test), indicating that we cannot treat the samples as independent, therefore confirming the need of multilevel modeling. From now on, we report significance values as given by the multilevel regression.

We perform two separate analysis with our data. First, we analyze our experiment in isolation. We include in the regression the three factors described in Section (alignment of ROIs *A*, and ROI configuration before R_b, and after the cut R_a), as well as their first-order interactions. We code our categorical variables as dummy binary variables to include them in the regression. The goal of this first analysis is to assess whether our observed effects are in accordance with those of Serrano et al. [6]. We describe the main findings of this analysis below, under *Influence of alignment* and *Influence of ROI configurations*. Then, we perform a second analysis in which we additionally account for the effect of directional sound cues by introducing a new variable *D*. In

Figure 4. Example of users' gaze for a frame sequence after the cut with (ours, top), and without (Serrano et al., bottom) directional sound cues. In the first frame we mark the ROIs (orange) and their associated sound source (blue icon); color points represent gaze samples from different users. For scenes with two ROIs in different fields of view after the cut, users are incited by the directional sound to switch between the two ROIs rather than locking their attention to a single ROI.

addition to our data ($D = Dir$), we include in this analysis the data provided by Serrano et al. corresponding to our conditions but without directional sound ($D = nDir$). The goal of this second analysis is to investigate the influence of directional sound cues in users' behavior. We report its main findings in the subsection entitled *Influence of directional sound cues*.

Influence of alignment A
We have observed an effect of the alignment factor in the *framesToROI* metric ($p < 0.001$). As expected, the more misaligned the ROI is with respect to the previous clip, the longer it takes user to reach it and fixate on it. The *nFix* metric also reveals a significant effect: the number of fixations is significantly lower ($p = 0.001$) for A_{80} than for A_{40}: this could suggest, as noted by Serrano et al., that viewers could be more inclined to explore when they are presented with a larger misalignment at the cut between clips.

Interestingly, different from Serrano et al., we did not find a significant influence of the alignment in the *scanpathError* metric (how much the path followed during the exploration differs with respect to the baseline data) nor in the *percFixInside* metric (how much viewers fixate on the ROI(s)). This could indicate that even if a more exploratory behavior is encouraged (as hinted by the *nFix* metric), users tend to follow a more similar path, and their attention to the ROI is preserved. We hypothesize that directional sound acts as an incentive to return to the ROI, and that the direction of the sound helps guiding the exploration resulting in more similar scanpaths, even during these more exploratory

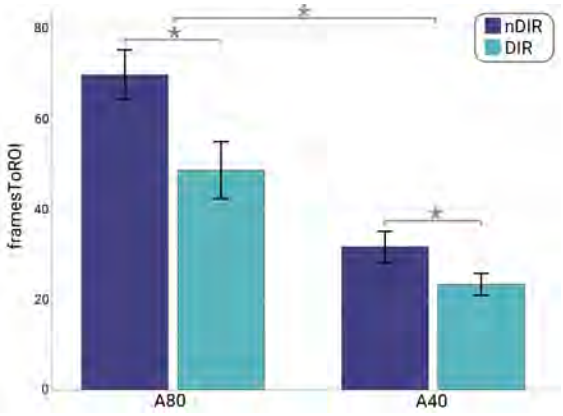

Figure 5. Mean number of frames to reach the ROI (*framesToROI*) for different misalignments (A40 and A80). The general trend prevails with respect to Serrano et al., as expected: larger misalignmets require more time for converging to the main action. Additinally, the time needed for convergence is greatly reduced in the presence of directional sound cues. Asterisks mark significant differences. Error bars represent a 95% confidence interval.

patterns. These findings seem to indicate that even when misalignment fosters exploration, users tend to follow a more predictable path in the presence of directional sound cues.

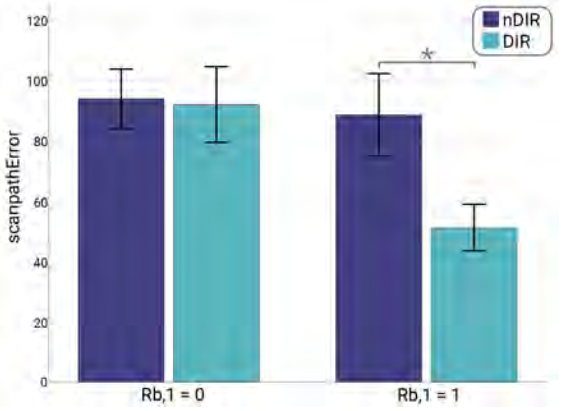

Figure 6. Mean scanpath error (*scanpathError*) for different configurations of the ROI before the cut. The deviation of the scanpath with respect to the baseline decreases for clips with two ROIs in the same field of view before the cut ($R_{b,1} = 1$). This trend is only visible under the presence of directional sound cues, indicating that users use sound as a cue to guide their exploration in search of the ROIs they have previously seen before the cut. Asterisks mark significant differences. Error bars represent a 95% confidence interval.

Influence of ROI configurations R_a and R_b

We have observed a significant effect of ROI configuration before the cut (R_b) on the *percFixInside* metric ($p = 0.006$). In particular, two ROIs in the same field of view before the cut ($R_{b,1}$) lead to less fixations on the ROI(s) after the cut. This finding is in accordance to Serrano et al., and as suggested in their work, this could indicate that multiple ROIs before the cut elicit a more exploratory behavior after the cut. We have also observed a significant influence of the ROI configuration after the cut (R_a) on the *scanpathError* metric ($p < 0.001$). The deviation of the scanpath with respect to the baseline is significantly higher when two ROIs do not fall within the same field of view after the cut ($R_{a,2}$). This is to be expected, since, regardless of the presence of directional sound cues, two ROIs can not be attended simultaneously, and therefore the scanpaths followed by users after the cut will tend to differ.

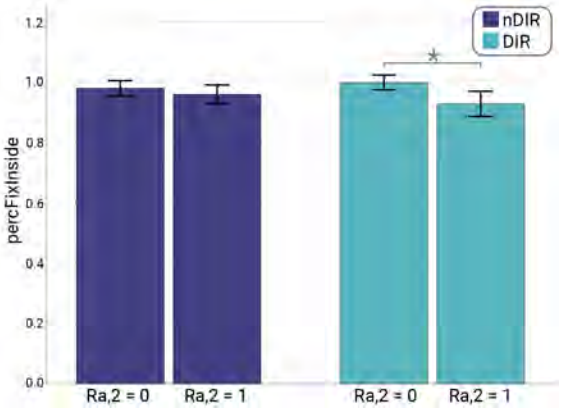

Figure 7. Mean percentage of fixations inside the ROI (*percFixInside*) for different configuration of the ROI before the cut. The metric *percFixInside* decreases in the presence of two ROIs in different fields of view after the cut ($R_{a,2} = 1$) when directional sound cues are used. This trend, although slightly observed without directional sound, is stronger under the presence of directional sound, indicating that the directional sound is a strong cue of the presence of different ROIs even if they are not visible in the current field of view, inciting users to search for them rather than locking into a single one. Asterisks mark significant differences. Error bars represent a 95% confidence interval.

Influence of directional sound cues D

We have discovered that the presence of directional sound cues has a strong effect in the time that users take to converge to the main action. There is a significant influence of D in the *framesToROI* metric ($p = 0.004$): using directional sound cues heavily decreases the time needed to find the new ROI after the cut. This effect can be clearly seen in Figure 5. We also show an example of users' gaze patterns for a video sequence (after the cut) in Figure 3 where it can be seen that users converge faster to the main action, and following more predictable patterns, even if this new ROI is misaligned with respect to the previous ROI.

Interestingly, there is also a significant influence of the interaction between directional sound, and the presence of two ROIs in the same field of view before the cut (D*$R_{b,1}$) in the *scanpathError* metric ($p = 0.01$). Even though multiple

ROIs before the cut elicit a more exploratory behavior after the cut, in the presence of directional sound cues the deviation of the scanpath with respect to the baseline decreases for clips with two ROIs in the same field of view before the cut. This could indicate that users use sound as a cue to guide their exploration in search of the ROIs they have previously seen before the cut, making their trajectories more similar. This effect can be appreciated in Figure 6.

Another interesting behavior can be observed in the interaction between directional sound, and the presence of two ROIs in different fields of view after the cut ($D^*R_{a,2}$). The metric *percFixInside* decreases significantly ($p = 0.012$) in the presence of two ROIs in different fields of view after the cut when directional sound cues are used. This can be an indicative of users pursuing a more exploratory behavior instead of locking their attention to a single ROI: the directional sound is a strong cue of the presence of different ROIs even if they are not visible in the current field of view, and users may be more inclined to switch between these two different ROIs rather than focusing their attention to the only visible one. This trend is observed in Figure 7, and also in the frame sequence in Figure 4, where it can be seen that users only fixate in one of the ROIs in the absence of directional sound, while they effectively explore the scene and find the two ROIs when directional sound cues are present.

Discussion and Conclusions

In this work we have studied and quantified the influence of directional sound cues in users' gaze patterns across cuts while watching 360° videos. We have demonstrated that directional sound cues are a key element that can strongly influence users' attention during cuts while watching cinematic content in VR. Our main finding shows that users converge much faster to the main action after a cut when directional sound cues are present, even in the presence of strong misalignments between the ROIs before and after the cut. While in the absence of directional sound cues users had to fully explore the scene in order to find the new region of interest, directional sound is a strong cue that serves as guidance to find the orientations of the ROIs. We have also confirmed that several trends found in previous works hold: the larger the misalignment introduced between the ROIs before and after the cut, the longer it takes users to converge.

Further, we find additional interesting insights: In the presence of directional sound cues, this time to convergence is not only shorter, but also the difference between different misalignments is less stressed. This reduction in time to convergence is expected, but, to our knowledge, it is the first time that it has been actually shown through careful, controlled experimentation. We have explicitly quantified this effect, so it could serve as guidance for establishing a desired editing pace: faster cuts or larger misalignments may be possible under the presence of directional sound cues since users reorient to the main action quickly after the cut.

We have also observed that the percentage of fixations inside the ROI decreases in the presence of two ROIs in different fields of view after the cut when directional sound cues are present. This can be an indicative of users pursuing a more exploratory behavior instead of locking their attention to a single ROI: Directional sound is a cue of the presence of different ROIs even if they are not visible in the current field of view, and users are more inclined to switch between these. This insight confirms that directional sound cues can be effectively used to alert the viewer about actions taking place at different orientations outside the field of view, and is in accordance with previous findings [17], [9] regarding orientation in 360° content.

Finally, we have shown that even though multiple ROIs before the cut elicit a more exploratory behavior after the cut, in the presence of directional sound the deviation of the scanpaths with respect to the baseline decreases for clips with two ROIs in the same field of view before the cut. This could indicate that users rely on sound as a cue to guide their exploration in search of the ROIs they had previously seen before the cut, making their trajectories more similar. This insight could have direct implications not only for cinematic VR and content creation, but also for scanpath prediction, which is an open challenge in VR: Our results suggest that predicting scanpaths may become an easier task when directional sound cues are used, and that users' exploration patterns can be anticipated in such cases.

Interviews with practitioners

Spatial audio has been proven to be of great importance in virtual reality production [26]. We have interviewed practitioners with over 20 years of experience in the field, specialized in virtual reality productions and immersive experiences[4] to discuss our insights and their applicability to cinematic VR production. As expressed by the interviewed practitioners, the main consideration they must take into account when employing directional audio in production is that this increases the recording and post-processing complexity, and therefore increments the budget of the production. Given this limitation, it is indeed very useful to know in which conditions directional sound may have a stronger impact on the users' viewing behavior, in order to assess whether this increased complexity is actually justified. In many cases, knowing in which conditions directional audio may be more beneficial is learned by trial an error: Our insights may serve as initial steps to support this learning process, and as baseline to start exploring more complex scenarios.

Limitations and future work.

As in similar studies, our insights may not extrapolate to conditions outside of our study; more analysis under a wider variety of scenes and conditions (such as more complex scenes, or the presence of more sound sources and distractors) could be needed in order to generalize our insights. More parameters and metrics could be analyzed in the future, such as different intensities and/or distances of the sound sources to the observer; these variables could also influence which regions of interest users choose to explore.

In this work we focus on diegetic methods, in particular, diegetic directional sound cues. In cases diegetic sounds are not naturally present in the narrative, artificially introducing a sound could also be considered. As previously discussed, this sound would ideally be diegetic (i.e., somehow related to the narrative, such as some local audio source). Non-diegetic sounds need to be introduced with caution: in Peck et al.'s work [27], they propose a distraction method they call *distractor audio*, in which the user is asked to turn their head to follow the sound of a hummingbird's wings. It was successful at inducing rotation, but overall considered unnatural by users. All these effects should be taken into consideration, and the interaction between diegetic and non-diegetic cues, as well as the potential influence of sound distractors, are open and challenging lines of research that must be carefully explored. We have experimented with different ROI configurations before and after the cut (a single ROI, two ROIs within the same field of view, and two ROIs not sharing the same field of view), resulting in nine different conditions for this factor. We believe some of our insights may still hold when more ROIs are present in the scene since directional sound will still play a crucial role in alerting the viewer about actions taking place at different orientations. However, further research is needed to understand the implications of introducing multiple ROIs, for example, previous work has suggested that if there are too many sound sources in a VR movie (clutter), it can be difficult to follow a spatial audio signal which should guide to a ROI [3]. Some of the fundamental behaviors we have observed have been shown to be consistent for different applications involving orientation in 360° [9], [17]: Our principled methodology can serve as a baseline to progressively analyze how well our insights will carry over to increasingly complex scenes, and to explore these more complex interactions.

We have shown that an accurate representation of direct sound paths is sufficient to effectively use the spatial properties of sound to support cuts in 360° cinematographic content. Nevertheless, our stimuli's audio is recorded from local microphones, and therefore the influence of some aspects of sound transport throughout the scene (e.g., reverberation levels) are not included. Our assumption is that these effects do not play a key role in the scenarios here tested. While we take here the first step towards analyzing the influence of directional sound cues in cuts, investigating more thoroughly the effects of ambisonic recording, modeling the complete sound transport of the scene, is definitely an interesting avenue of future work. We believe our work could serve as a baseline for comparisons, in addition to providing a valid methodology for future experiments. While a few works have looked into the influence of environmental settings on sound perception,

[4] Ábaco digital: https://www.abaco-digital.es/web/en

it is still unclear to what extent they have an influence: For instance, Engel et al. [28] suggest that there may not be strong perceivable benefits in using high order ambisonics encoding (beyond first order) for room auralisation as long as the direct sound is rendered with enough accuracy. On the other hand, Bormann et al. [29] compares directional but non-attenuated sound (similar to our setup) and fully spatialized sound for the task of finding the direction of a sound source, and concludes that fully spatialized sound increases performance. Therefore, we expect that introducing a more accurate spatialization would be beneficial for orientation during cinematic cuts. Future work could thus investigate the combination of ambisonic recording and direct sound cues, and analyze the perceptual implications and trade-offs in terms of accuracy of the representation and usefulness for directing attention.

In summary, we believe that our work provides valuable insights for assisting the creation of 360° cinematic experiences. It provides the first systematic study of the influence of directional sound cues in guiding the attention across cuts in VR narrative. We hope that our findings and data will foster new research in this direction in order to further our understanding on how to effectively communicate stories in VR.

Acknowledgements

We would like to thank Jaime Ruiz-Borau for his help setting up the experiments. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (CHAMELEON project, grant agreement No 682080), from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreements No 765121 and 956585, and from the Spanish Ministry of Economy and Competitiveness (projects TIN2016-78753-P and PID2019-105004GB-I00).

REFERENCES

1. V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in vr: How do people explore virtual environments?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 36, no. 4, 2018.
2. D. Martin, A. Serrano, and B. Masia, "Panoramic convolutions for 360° single-image saliency prediction," in *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, 2020.
3. S. Rothe, D. Buschek, and H. Hußmann, "Guidance in cinematic virtual reality-taxonomy, research status and challenges," *Multimodal Technologies and Interaction*, vol. 3, no. 1, p. 19, 2019.
4. D. Bordwell, K. Thompson, and J. Smith, *Film art: An introduction*. McGraw-Hill New York, 1993, vol. 7.
5. R. Viers, *Sound Effects Bible*. Michael Wiese Productions, 2011.
6. A. Serrano, V. Sitzmann, J. Ruiz-Borau, G. Wetzstein, D. Gutierrez, and B. Masia, "Movie editing and cognitive event segmentation in virtual reality video," *ACM Transactions on Graphics (SIGGRAPH 2017)*, vol. 36, no. 4, 2017.
7. C. Marañes, D. Gutierrez, and A. Serrano, "Exploring the impact of 360° movie cuts in users' attention," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2020.
8. C. O. Fearghail, C. Ozcinar, S. Knorr, and A. Smolic, "Director's cut-analysis of aspects of interactive storytelling for vr films," in *International Conference on Interactive Digital Storytelling*. Springer, 2018, pp. 308–322.
9. S. Rothe, H. Hußmann, and M. Allary, "Diegetic cues for guiding the viewer in cinematic virtual reality," in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. ACM, 2017, p. 54.
10. A. Sheikh, A. Brown, Z. Watson, and M. Evans, "Directing attention in 360-degree video," 2016.
11. L. T. Nielsen, M. B. Møller, S. D. Hartmeyer, T. C. Ljung, N. C. Nilsson, R. Nordahl, and S. Serafin, "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*, 2016, pp. 229–232.
12. S. Grogorick and M. Magnor, "Subtle visual attention guidance in vr," in *Real VR–Immersive Digital Reality*. Springer, 2020, pp. 272–284.
13. M. Groehn, T. Lokki, L. Savioja, and T. Takala, "Some aspects of role of audio in immersive visualization," in *Visual Data Exploration and Analysis VIII*, vol. 4302. International Society for Optics and Photonics, 2001, pp. 13–22.
14. D. Martin, S. Malpica, D. Gutierrez, B. Masia, and A. Serrano, "Multimodality in vr: A survey," 2021.
15. J. Steuer, "Defining virtual reality: Dimensions determining telepresence," *Journal of communication*, vol. 42, no. 4, pp. 73–93, 1992.

16. J. J. Cummings and J. N. Bailenson, "How immersive is enough? a meta-analysis of the effect of immersive technology on user presence," *Media Psychology*, vol. 19, no. 2, pp. 272–309, 2016.
17. P. Bala, R. Masu, V. Nisi, and N. Nunes, "Cue control: Interactive sound spatialization for 360 videos," in *International Conference on Interactive Digital Storytelling*. Springer, 2018, pp. 333–337.
18. Z. Tang, N. J. Bryan, D. Li, T. R. Langlois, and D. Manocha, "Scene-aware audio rendering via deep acoustic analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 5, pp. 1991–2001, 2020.
19. H. Kim, L. Hernaggi, P. J. Jackson, and A. Hilton, "Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 120–126.
20. D. Li, T. R. Langlois, and C. Zheng, "Scene-aware audio for 360 videos," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–12, 2018.
21. S. Serafin, M. Geronazzo, C. Erkut, N. C. Nilsson, and R. Nordahl, "Sonic interactions in virtual reality: state of the art, current challenges, and future directions," *IEEE computer graphics and applications*, vol. 38, no. 2, pp. 31–43, 2018.
22. M. Naef, O. Staadt, and M. Gross, "Spatialized audio rendering for immersive virtual environments," in *Proceedings of the ACM symposium on Virtual reality software and technology*. ACM, 2002, pp. 65–72.
23. S. Knorr, C. Ozcinar, C. O. Fearghail, and A. Smolic, "Director's cut - a combined dataset for visual attention analysis in cinematic vr content," in *The 15th ACM SIGGRAPH European Conference on Visual Media Production*, 2018.
24. S. Burigat, L. Chittaro, and S. Gabrielli, "Visualizing locations of off-screen objects on mobile devices: a comparative evaluation of three approaches," in *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, 2006, pp. 239–246.
25. A. Pavel, B. Hartmann, and M. Agrawala, "Shot orientation controls for interactive cinematography with 360 video," in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, 2017, pp. 289–297.
26. "Oculus connect - bring your 360 videos to life with spatial audio." https://www.youtube.com/watch?v=MgI85YQch10, accessed: 2020-12-01.
27. T. C. Peck, H. Fuchs, and M. C. Whitton, "Evaluation of reorientation techniques and distractors for walking in large virtual environments," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 3, pp. 383–394, 2009.
28. I. Engel, C. Henry, S. V. A. Garí, P. W. Robinson, D. Poirier-Quinot, and L. Picinali, "Perceptual comparison of ambisonics-based reverberation methods in binaural listening," 2019.
29. K. Bormann, "Presence and the utility of audio spatialization," *Presence: Teleoperators & Virtual Environments*, vol. 14, no. 3, pp. 278–297, 2005.
30. T. C. Kübler, K. Sippel, W. Fuhl, G. Schievelbein, J. Aufreiter, R. Rosenberg, W. Rosenstiel, and E. Kasneci, "Analysis of eye movements with eyetrace," in *International Joint Conference on Biomedical Engineering Systems and Technologies*. Springer, 2015, pp. 458–471.

Appendix

In this appendix we describe the processing procedure of the collected gaze points (Pupil-labs eye-tracker) and head positions (Head Mounted Display). In order to compare our data with that of Serrano et al. [6] we strictly follow the same procedure described in their work:

Gaze processing:
The head orientation tracking from the Head Mounted Display has a lower sampling rate than the eye-tracker. Therefore, to match the sampling rates, we first assign to each gaze sample the head position with the closest timestamp. Eye tracker measurements with a confidence below 0.9 are linearly interpolated (confidence ranges from 0 to 1). Then, we compute the final gaze point by taking into account head orientation and relative gaze from the eye-tracker, and we register these final gaze points to each frame in the videos. Since the videos run at 60fps and the eye tracker records at 120Hz, two gaze positions are recorded for each frame: we merge them into a single gaze point per frame by computing their mean position.

Prior to gaze processing, trials in which the mean confidence for both eyes is below 0.6 are discarded.

Fixation detection:
Fixations are detected by means of a velocity-based fixation detector [30], in which a gaze point is considered a fixation if its velocity is below a

certain threshold. This threshold is computed as 20% of the maximum velocity for a given gaze scanpath, after discarding the top 2% velocities for increased robustness.

Outlier rejection:
Following standard practice, we discard observations that differ significantly from other users' behavior based on the interquartile difference, with a factor of 1.5. Additionally, we discard a trial when less than 40% of the total number of fixations before the cut occurred inside the ROI. We consider that in such cases users were not paying attention, or did not understand the task.

Belen Masia is an Assistant Professor at the Universidad de Zaragoza. Her research in virtual reality, computational imaging, displays, and applied perception has received a number of awards, including the Eurographics Young Researcher Award in 2017, a Eurographics PhD Award in 2015, a Leonardo BBVA Foundation Award in 2020, or a NVIDIA Graduate Fellowship in 2012. She is a Eurographics Young Fellow. Contact her at bmasia@unizar.es.

Javier Camon is product design engineer. He received his Bs.C. in Product Design Engineering from Universidad de Zaragoza in 2017. His areas of interest include 3D design, innovation, virtual reality and cinema. You can contact him at jcamon@outlook.es.

Diego Gutierrez is a Professor at the Universidad de Zaragoza, where he leads the Graphics and Imaging Lab. His areas of research include physically based global illumination, virtual reality and computational imaging. He has published over 100 papers in top journals, including Nature. He has received many awards (including Google or the BBVAFoundation). He received in 2016 an ERC Consolidator Grant. Contact him at diegog@unizar.es.

Ana Serrano is a postdoctoral researcher at the Max Planck Institute for Informatics. She received an Adobe Research Fellowship in 2017, and a NVIDIA Graduate Fellowship in 2018. Her thesis has been awarded with one of the Eurographics 2020 PhD awards. Her work on virtual reality has been published in top venues, including ACM Transactions on Graphics, Scientific Reports, and IEEE TVCG. Contact her at anserran@mpi-inf.mpg.de.