# EXPLORING AUDIO IMMERSION USING USER-GENERATED RECORDINGS

*Daniel Gomes, João Magalhães, Sofia Cavaco* *

NOVA LINCS, Departamento de Informática
Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
2829-516 Caparica, Portugal
ddl.gomes@campus.fct.unl.pt, {jmag,scavaco}@fct.unl.pt

## ABSTRACT

The abundance and ever growing expansion of user-generated content defines a paradigm in multimedia consumption. While user immersion through audio has gained relevance in the later years due to the growing interest in virtual and augmented reality immersion technologies, the existent user-generated content visualization techniques are still not making use of immersion technologies.

Here we propose a new technique to visualize multimedia content that provides immersion through the audio. While our technique focus on audio immersion, we also propose to combine it with a video interface that aims at providing an enveloping visual experience to end-users. The technique combines professional audio recordings with user-generated audio recordings of the same event. Immersion is granted through the spatialization of the user generated audio content with head related transfer functions.

## 1. INTRODUCTION

Considering a society in transformation and transition to immersive content, such as video games and virtual reality, it is natural to extend the immersion to other content such as user-generated content (UGC). To answer this tendency we propose a technique that uses the audio from UGC to achieve immersion through the audio. By immersion we mean spatial presence, as defined by Madigan [1].

Audio immersion can serve multiple purposes ranging from different areas of interest. As an example, it can be used for education, training and entertainment of blind and visually impaired (BVI) people. *Navmol* is an application that uses audio immersion to help BVI chemistry students to interpret and edit the representation of molecular structures [2]. Once a reference atom is selected, the application uses a speech synthesizer to inform the user about the neighboring atoms. The speech signal is spatialized using head related transfer functions (HRTFs). In this way, users wearing headphones will hear the atoms' descriptions coming from different angles in space. Similarly, immersive audio can be used to train orientation and mobility skills for BVI people. Cavaco, Simões and Silva propose a virtual environment for training spatial perception in azimuth, elevation and distance [3]. The audio is spatialized with HRTFs. Immersive audio has also been used for entertainment of BVI people. *Demor* is a shooting game based in 3D spatialized audio that aims at providing entertainment to both BVI and sighted players [4]. Despite the entertainment component, Cohen *et. al* also attempt to improve BVI people emancipation in the sighted world by training mobility and spatial orientation. The game requires players to localize sounds in space that represent targets to shoot before they reach the player, who is equipped with a laptop, a GPS receiver, a head tracker, headphones and a modified joystick, all attached to a backpack. The kit continuously tracks player location and orientation and updates the sound accordingly.

Immersive audio can also be applied to information delivery. Montan introduced a low cost audio augmented reality prototype for information retrieval [5]. In the study, the author created a headset with a head movement tracker for a use case of museum interactive audio-guides. As the users rotated their heads, the tracker registered head orientation and the system rendered the audio properly. The rendering is performed in real time using HRTFs according to the relative position and orientation of the listener and the emitters. In another study, Guerreiro proposed to take advantage of the cocktail party effect to convey information about digital documents to BVI people using only audio [6]. Instead of using a common text-to-speech system that converts textual information into a speech signal that contains a single voice, the author proposed to use various voices simultaneously at different angles. HRTFs were used to change the speech synthesizer signal.

Here we propose an audio immersion technique for UGC. The technique combines several UGC recordings of the same event, modified with HRTFs, in order to immerse audibly the user. Such recordings are distributed in space and are reproduced from different angles. While the technique focus on audio immersion, we also propose to combine it with a video interface that aims at providing an enveloping visual experience (more details in section 2).

In order to demonstrate and validate the proposed technique, we built a prototype for mobile devices that includes a video player (section 3). The proposed tool is designed to play concert videos, although it it not limited to this single use case scenario. We used this prototype in a user test that validates the proposed technique. The test focuses on three attributes: immersion, sense of space and directional quality. Section 4 describes the tasks performed in more detail, section 4.1 describes the data used in the user test, and section 5 discusses the user test's results.

Since the scope of the presented work required diverse and abundant UGC, musical concerts were chosen as a good use case scenario to draw useful conclusions. The database chosen is composed of multiple events (*i.e.*, audio recordings of several concerts), which in their turn have several recordings. User-generated

* This work was supported by the H2020 ICT project COGNITUS with the grant agreement No 687605 and the Portuguese Foundation for Science and Technology under project NOVA-LINCS (PEest/UID/CEC/04516/2019).

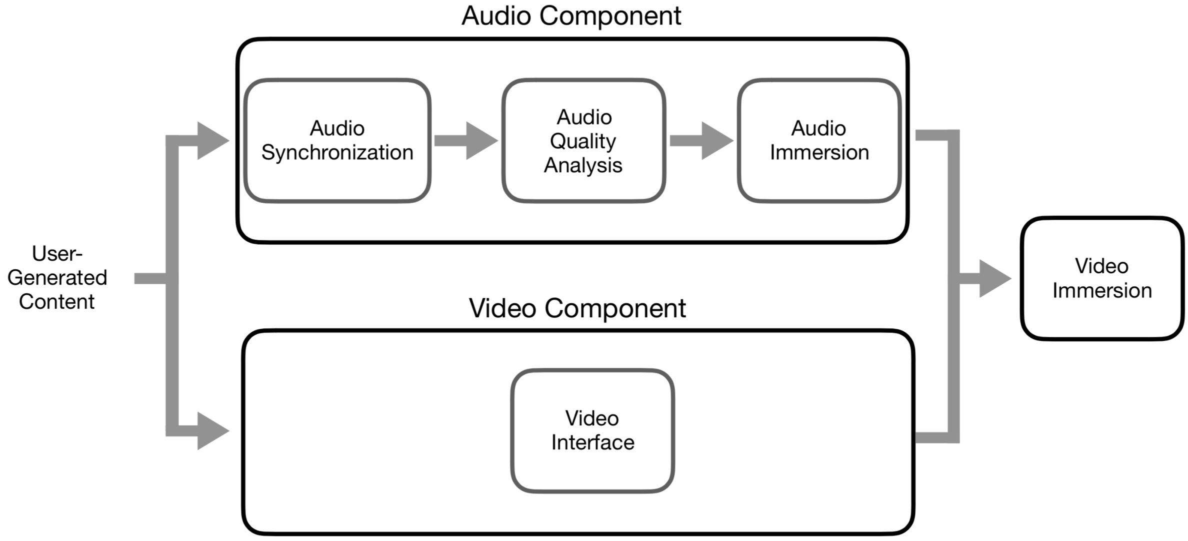


Figure 1: The proposed UGC audio immersion technique scheme.

recordings raise some challenges, in particular noise and a time sparse nature. Different devices have distinct recording qualities which impact in the level of noise captured. Additionally, recordings capture different parts of the event, considering possible overlaps. Thus it is required to deal with these UGC challenges. We propose to categorize recordings of the chosen data set by subjective level of noise and select those with best quality to be played. As explained in section 2, we propose to use a quality analysis and data synchronization technique based on audio fingerprints.

## 2. THE UGC AUDIO IMMERSION TECHNIQUE

The proposed immersion technique can be seen as the interaction between two main components: the audio and the video components (figure 1). The audio component immerses the user audibly by presenting multiple audio sources distributed in a multi-dimensional space. The video component consists of a user interface that aims at providing an enveloping visual experience to the end user. Note that we do not aim to achieve video immersion in this prototype.

Since our main focus in this paper is the audio component, in this section we focus only on this component. Nonetheless, we developed two different simple approaches for the video component for validation purposes. These approaches are discussed in section 3.1.

To achieve audio immersion using UGC, we propose to combine several recordings from the same event. Our technique consists of changing the original recordings such that they are reproduced from different angles, and when played together they provide a combined audio signal that can give the perception of immersion.

It is important to highlight that UGC have diverse audio qualities inherent to the different characteristic of the devices used to capture the audio. Also different recordings of the same event (for instance, the same music in a concert), can capture different portions of the event with possible overlapping sections. Thus, even before we process the audio signals to provide a sense of immersion, there are other steps we must perform, namely audio synchronization and analysis of the signals' quality.

### 2.1. Audio synchronization

Given a data set of recordings from the same event, it is important to understand the chronological order of the events and identify the recordings' overlapping sections. Following our previous work on organization of user generated audio content (UGAC), we propose to use audio fingerprints to create a timeline with the event's recordings [7, 8].

The resistance to noise of fingerprinting techniques is particularly relevant when dealing with low quality music recordings. This characteristic is suitable for our proposed immersion technique because it enables synchronizing samples with quite different quality and noise levels, which is a characteristic of UGAC.

Mordido *et al.* use audio fingerprints to identify common sections between the audio recordings [8]. This technique identifies the overall offsets of all recordings of the same event, as well as the duration of each recording. This gives us information on which recordings cover different portions of the timeline. Thus, we can organize an event with *timeline segments*, such that segments coincide with the time interval of overlapping recordings. The final result is a timeline with all the recordings aligned, such that overlapping sections of different signals are played simultaneously. Figure 2 shows the timeline for a set of five recordings from the same event. The timeline is organized into timeline segments $T_1$ to $T_7$.

### 2.2. Audio quality analysis

Once the signals are chronologically organized, we need to choose which signals to use. Given a set of recordings from the same event, we will choose only a few. More specifically, for each timeline segment, $T_i$, we choose $n_i$ recordings (we choose a low number, such as three or four at most). In order to choose the $n_i$ recordings for each timeline segment, we start by measuring the quality

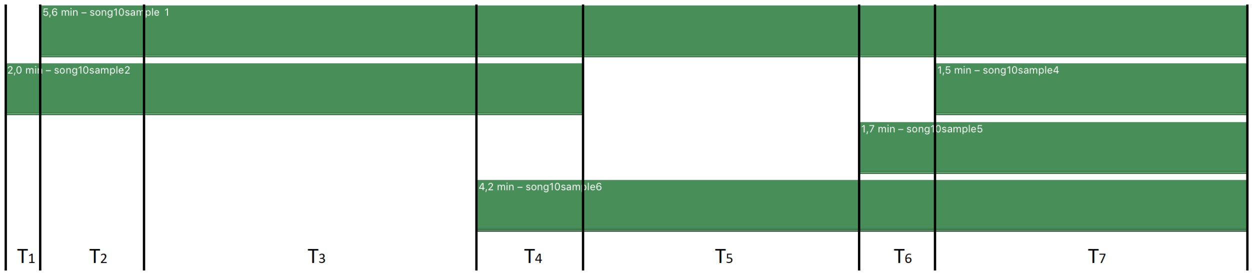


Figure 2: Timeline and segments, $T_1$ to $T_7$, for a set of recordings (in green) from the same event.

of all the data in the set, and choose the recordings with the higher quality.

For the recordings audio quality analysis, we propose to use our previous work with audio fingerprints for quality inference of UGAC [7, 8]. Like with the audio synchronization technique explained above, this method uses audio fingerprints to detect overlapping segments between different audio recordings. In addition it assumes that the recordings with more matching landmarks have higher quality. (A landmark is a pair of two frequency peaks and contains information about the frequency for the peaks, the time stamp of the first peak, and the time offset between the two peaks.)

Alternatively, some interesting recordings can also be chosen manually. For instance, let us assume there is a recording with voices or clapping in the audience that we want to use, but has lower quality. While this recording may be ranked as having low quality, it can still be chosen. In fact, we manually chose the recordings used in our prototype because we wanted to have recordings with quite different characteristics.

### 2.3. Audio immersion

As shown in figure 1, the following step is audio immersion. In this step, we change the original $n_i$ audio signals selected for each timeline segment $T_i$, such that when heard individually through headphones, they can be perceived as if coming from different directions, and when heard together, they give a sense of space.

Our proposal is to change each original signal $s_j$ with HRTFs. That is, we apply HRTFs to the left and right channel of each signal $s_j$, such that the modified signal, $s'_j$ is perceived as if coming from angle $\theta_j$. Angle variations are performed in azimuth and elevation. Finally, the timeline built in the audio synchronization phase is used when playing the modified signals. Thus, for each timeline segment $T_i$, we will play a final signal, $S_i$, that consists of adding together all selected modified signals from that timeline segment. For instance, let us assume that timeline segment $T_i$ has the overlapping signals $s_1$, $s_2$ and $s_3$. We modify these original signals with HRTFs such that when heard individually, the modified signals $s'_1$, $s'_2$ and $s'_3$ are perceived from directions $\theta_1$, $\theta_2$ and $\theta_3$. Hearing the three signals played simultaneously can give a sense of immersion in which we hear the common music (present in all three recordings) in the surrounding space and we hear the specific individual noises or sounds (like clapping or voices) from each recording as if coming from different directions. Figure 3 illustrates this example.

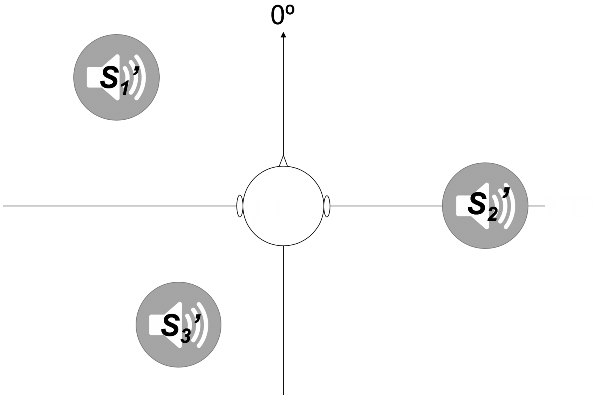


Figure 3: Signals $s'_1$, $s'_2$ and $s'_3$ distributed spatially at directions $\theta_1$, $\theta_2$ and $\theta_3$, respectively. (The users' initial orientation is used to define 0°, which is the direction ahead of the user).

## 3. THE VIDEO PLAYER PROTOTYPE

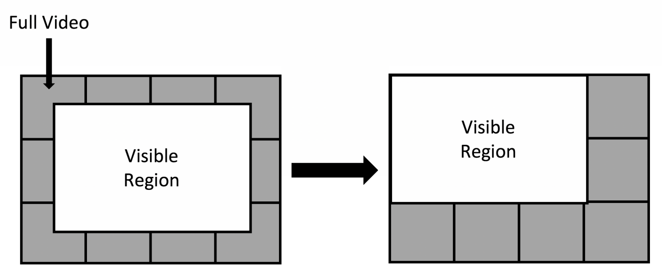
In order to validate our UGC audio immersion technique, we developed a prototype that was used in the user tests. This prototype includes an audio component and a video interface.

The prototype's video component was developed with Unity Game Engine. Audio spatialization was granted by Google's Resonance Audio Software Development Kit which uses Sadie HRTF library (University of York SADIE KU100 data set). In the current context, the application was built for iOS devices and requires the use of headphones for audio immersion.

### 3.1. The video component

The design of our graphical user interface was inspired on the work proposed by Chen, who presented an image-based approach to virtual environment navigation [9]. Chen presented two types of video player: a panoramic and an object player. The first was designed for looking around a space from the inside, while the second was designed to view an object from the outside. Among other features, the panoramic player allows the user to perform continuous panning in the vertical and horizontal directions.

Since, the current state-of-the-art in multimedia content creation by users is from planar smartphone cameras, we developed a graphical user interface that has similarities to the one proposed by Chen. Our user interface does not show the video completely (figure 4). Instead, as shown in the figure, there is a visible region that



(a) (b)

Figure 4: The image display process. The white region is visible, while the dark region is not visible. The user can slide the white rectangle to the dark regions so that the visible region changes. (a) The visible region is in the center of the original video. (b) The visible region changes when the user slides it.

the users can pan continuously through the entire video region (*i.e.*, navigation in horizontal, vertical and diagonal). To implement this visible region we used Unity's orthographic camera projection.

We developed two different approaches for user interaction with the application:

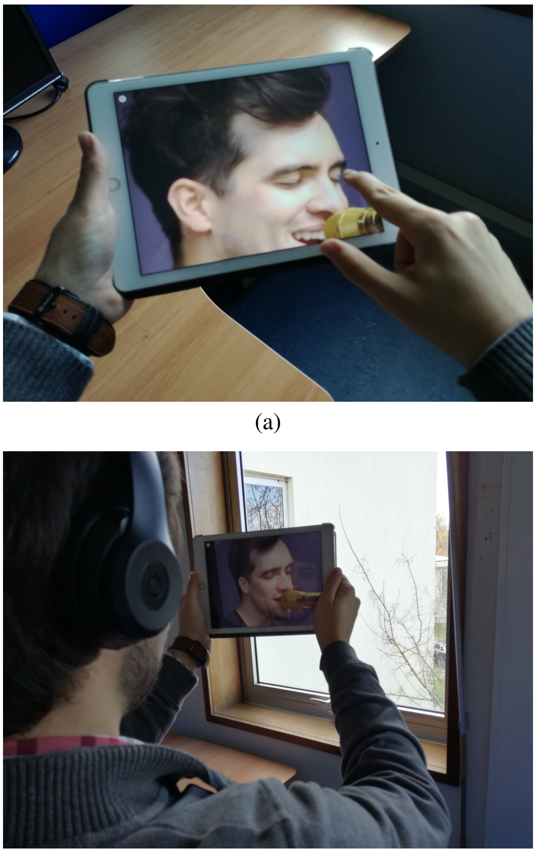
- In the *touch screen approach*, the interaction is processed using the device's touch screen. Users can navigate through the video making use of the device's touch screen to move the visible region around (figure 5.a). For instance, a sliding movement towards the left makes the visible region move to the right.
- The *gyroscope approach* uses the device's gyroscope. Here, users can interact and move the visible region using the gyroscope (figure 5.b). Moving the screen towards the right makes the visible region move to the right. Moving the screen upwards, makes the visible region move upwards.

### 3.2. The audio component

While developing our prototype, we focused our attention especially in the audio immersion box from figure 1 and the interaction between the video and audio components. The audio recordings used in the prototype were manually selected and synchronized.

As explained above, we use HRTFs to spatialize the original signals $s_j$, such that each modified signal $s'_j$ is located at a specific direction in space ($\theta_j$) and the final sound $S_i$ for each timeline segment is the combination of the modified signals $s'_j$. Sliding the visible region into a certain direction, produces changes in each signal $s_j$, and, as a consequence, the combined signal $S_i$ also changes.

There are two parameters that change for each signal: the intensity and the relative angle to the user. Sliding the visible region into a certain direction, is mapped into head rotations. In other words, when the user slides the visible region (figure 4), the direction of each signal $s_j$ relative to the user changes. This way, when hearing the sounds, the users will perceive changes in the sounds that cause the sensation of having performed head rotations. The changes can be in azimuth and elevation. For instance, when the user moves the visible region to the right, the samples $s_j$ suffer a rotation to the left, as if the user had rotated his/her head clockwise (a change in the azimuth). When the user moves the visible region up, the samples $s_j$ suffer a downward rotation in elevation.



(a)

(b)

Figure 5: The graphical user interface. Navigation in the video using (a) the touch screen approach, and (b) the gyroscope approach.

Let us define $\vec{d}$ as the vector that represents the sliding movement in a 2D space whose $x$ and $y$-axes are parallel to the screen edges. $\vec{d}$ defines the movement direction and displacement. Let $\vec{d_x}$ be the projection of $\vec{d}$ into the $x$-axis, and $\vec{d_y}$ be the projection of $\vec{d}$ into the $y$-axis. For each signal $s_j$, the rotation in the azimuth is given as a function of $\vec{d_x}$ and the change in elevation is given as a function of $\vec{d_y}$.

The intensity of the signals may also change. We increase the intensity of sounds whose directions $\theta_j$ are approximate to the users final orientation, and decrease the intensity of other sounds. Sound intensity changes are described by a linear function of the relative angle to the users' orientation.

## 4. USER TESTS

In order to evaluate the proposed technique, we run a user test to evaluate spatial sound quality. Pulkki *et al.* propose that spatial sound quality evaluation should consider the evaluation of envelopment, naturalness, sense of space, directional quality and timbre [10]. Among the presented group of attributes, the ones of interest to our study are **directional quality** and **sense of space**. In addition, we introduced the **immersion** factor to be tested.

There were 15 volunteers participating in the study (10 men on

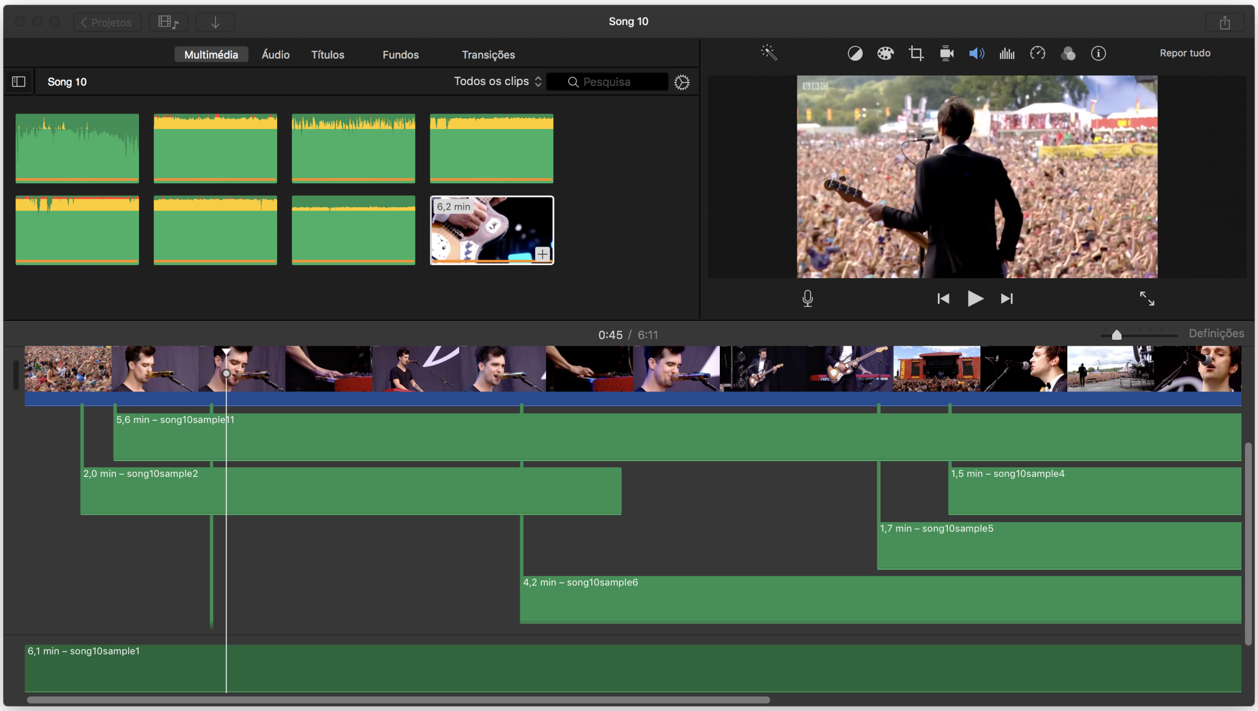


Figure 6: Global timeline of audio and video (developer's view). The green bar for the professional recording (*i.e.*, sample 1) is at the bottom of the image. The UGC recordings are displayed with their respective sample name and length, and starting at their start time $t_i$. The user's view is presented at the top right corner.

and 5 women) with ages ranging from 19 to 26 years old and all university students. Four of these participants had musical training (either by actively playing an instrument or attending music classes). Only one participant declared having hearing problems and one had a temporary hearing condition. The remaining participants asserted having no hearing problems that could affect their participation.

The user test consisted of five related tasks. For the first four tasks users used a computer while for the last one they used a tablet. The volunteers wore headphones for all tasks and received written instructions and a demonstration for every task. Additionally, at the end of each task, the volunteers were provided with a form in which they were queried about each task. All volunteers were attributed with a numeric reference in order to guarantee data protection.

### 4.1. Data

The three first tasks tested directional audio quality and used musical instrument sounds generated with iPad's GarageBand application and spatialized according to the technique described in section 2.3. These consisted of:

- A sequence of three sustained piano notes (C, E, G, in the presented order).
- A sequence of three guitar notes (C, G, F, in the presented order).
- A sequence of drum notes from three cymbals (snare, tom high and tom low, at no specific order).

The remaining tasks used recordings from music concerts that were extracted from Mordido's data set [7]. These recordings provide different components under different conditions (*e.g.*, users recording part of a concert in distinct places at different angles to the stage). Our data includes recordings from two events: two musics, each from a different concert. The first music chosen was a cover performed by Panic at the Disco! band of the popular Queen music Bohemian Rhapsody performed at the 2015 Reading Festival. For the second music we chose a live performance of Sing, by Ed Sheeran at the 2014 Glastonbury Festival. For each event, the data set includes a professional recording of the music and two to four user recordings of that music in the same concert.

The professional recordings have higher sound quality and less noise than the remaining samples in the data set. In this group of samples, it is possible to hear the audience singing along, cheering and clapping. Task 4 used the Queen's concert samples, and task 5 used both concerts.

The defined data set is used to produce immersive sound. Professional recordings are combined with UGC using the techniques and timeline explained above. For each event (that is, for each concert), we spatialize the original sound signals such that each modified signal ($s'_j$) is assigned a different direction ($\theta_j$): (1) The professional recordings are always assigned the same direction: 0°. This direction is determined by the user's initial orientation. (2) The remaining recordings are placed in lateral or rear-user an-

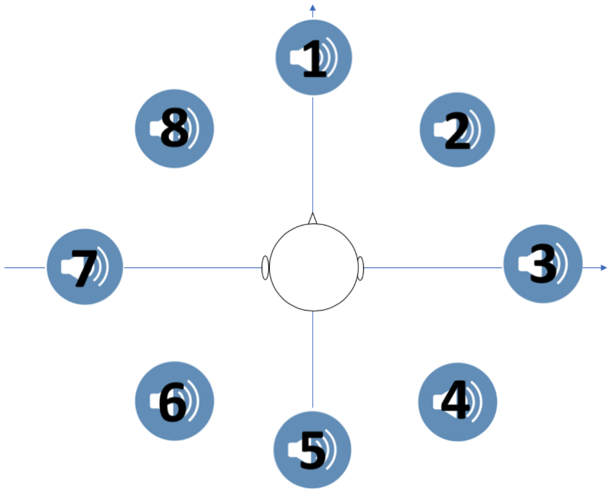


Figure 7: The audio sources' directions used in the user test's tasks 1 to 3.

gles.

For each event's timeline, the professional recording, $s_1$, starts at $t_1 = 0$ seconds while the start time for each remaining sample $s_i$ is $t_i \geqslant 0$ seconds. Hence, every selected user recording $s_i$, for $i > 1$, is present in the timeline with $t_i \geqslant 0$ seconds. Figure 6 illustrates an example. This shows the protocol's developer's view, which shows the timeline (bottom blue and green bars).

### 4.2. Tasks

Task 1 to 3 were used as training tasks but we also used them to assess the proposed technique's audio directional quality. In these tasks, participants train their sound localization ability for the next tasks (task 4 and 5).

Tasks 1, 2 and 3 consist of the reproduction of one, two and three audio sources simultaneously. The first task uses the piano sequence, task 2 uses the piano and guitar sequences and task 3 uses the three instruments sequences (section 4.1). The notes sequences are played several times (16, 9 and 6 times in task 1, 2 and 3, respectively). The directions of the instruments changed randomly. The possible directions are indicated in figure 7.

For every sequence reproduction, the volunteers were asked to register the perceived direction using the numbers provided in figure 7. The volunteers were asked to picture themselves in the center of the referential, with the circle numbered as 1 exactly in front of them, the circle numbered as 3 exactly at their right, etc. In order to better determine the direction of the sequences, the volunteers can simulate head rotations using the mouse and are provided with a button that allows them to return to the original orientation.

These three tasks focused mainly on identifying the audio source locations, in order to test directional audio quality. Therefore, the visual component of those tasks was ignored to keep them simple and have the user focusing on the audio. On the contrary, the fourth and fifth tasks consider both the visual and audio components in the context of the real application's goal.

The main goal of task 4 was to test all parameters simultaneously (*i.e.*, directional quality, sense of space and immersion). In this task, the video player (user's view in figure 6) displayed a concert video with a professional and two UGC recordings with some overlapping and some non-overlapping regions. A clock system was used for sound source location as presented in figure 8. 12 o'clock represents 0° as in figure 3.

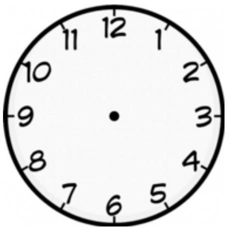


Figure 8: Clock system for audio source location.

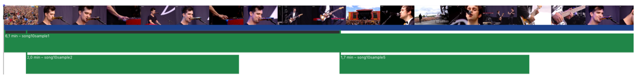


Figure 9: Timeline of the audio and video used for the task 4. The blue bar aggregated to video screen shots represent the video. The green bar immediately below the blue bar is the professional recording. Both smaller green bars at the bottom are the UGC recordings.

The professional recording was located at 12 o'clock, while the UGC recordings were placed at 8 o'clock and at 6 o'clock, by order of appearance respectively. Figure 9 displays the timeline for this task. The recordings in task 4 were played from static different directions. That is, the directions of the three recordings did not change during this task.

In task 5 the directions of the sounds were not static. This task tested if the users perceive a sense of space and directional audio when the directions of the sounds change dynamically.

In this task, participants used the two approaches developed for the video component (section 3.1) in an iPad.

## 5. RESULTS

In the first three tasks,we used the following classification scheme, where from *error type 1* to *4* the test subject failed to identify the audio source location:
– *Error 0* – the test subject identified successfully the audio source location;
– *Error type 1* – the answer provided was the location at 45 °from the correct audio source location;
– *Error type 2* – the answer provided was the location at 90 °from the correct audio source location;
– *Error type 3* – the answer provided was the location at 135 °from the correct audio source location;
– *Error type 4* – the test subject answered the location in the opposite location (*i.e.*, at 180 °).

Figures 10, 11 and 12 present the results of tasks 1, 2 and 3, respectively. All audio sources for all tasks exceeded more than 70% of right answers, which shows the directional quality of the spacialized sounds obtained with our technique.

The results of task 4 show that the sounds combined and spatialized by our technique give a sense of space and of directional audio. Users perceive that recordings played simultaneously (*i.e.* overlapping recording in the same timeline segment) have different directions. As mentioned above (section 4.2) this task used a

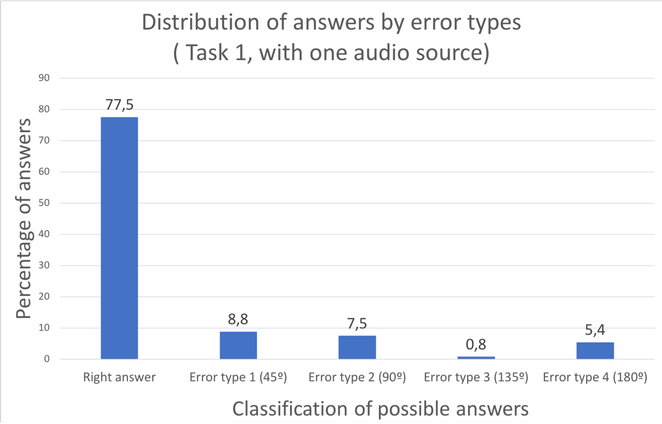


Figure 10: One audio source error distribution.

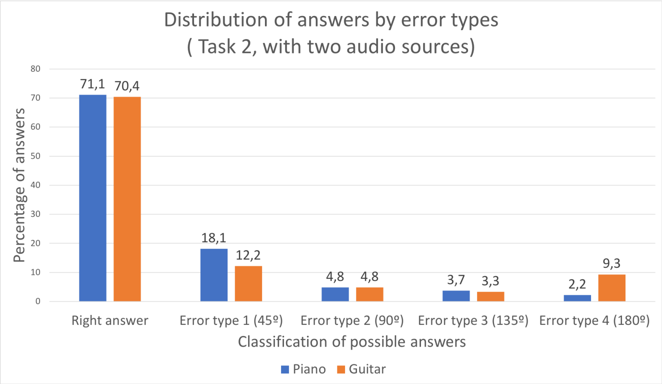


Figure 11: Two audio source error distribution.

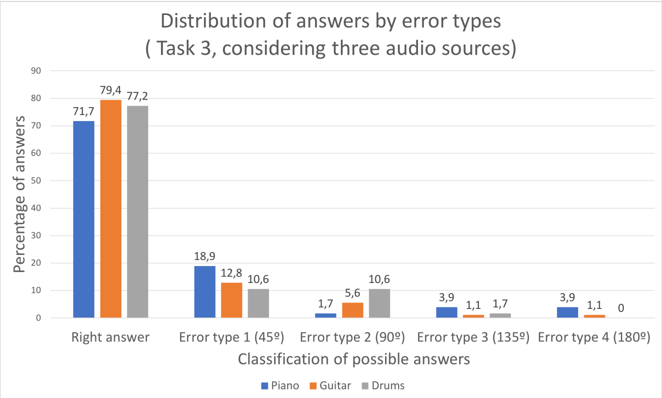


Figure 12: Three audio source error distribution.

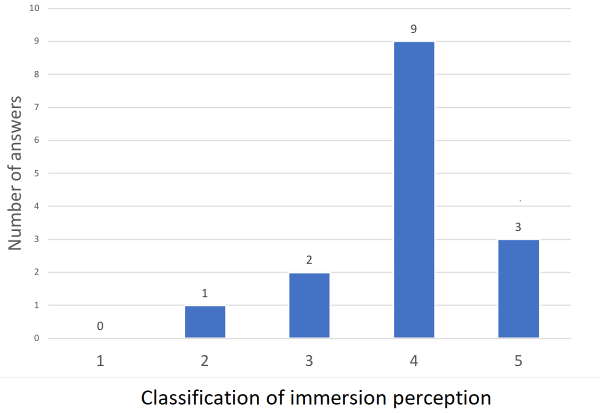
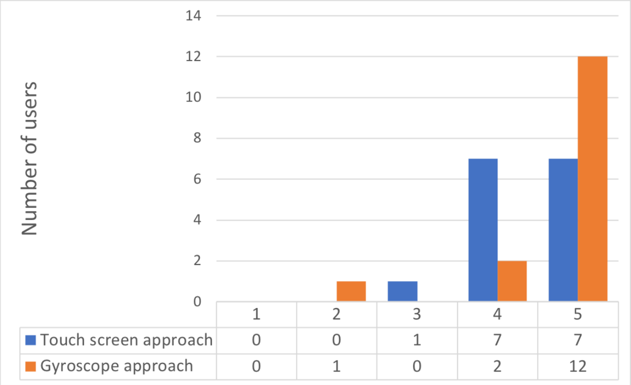
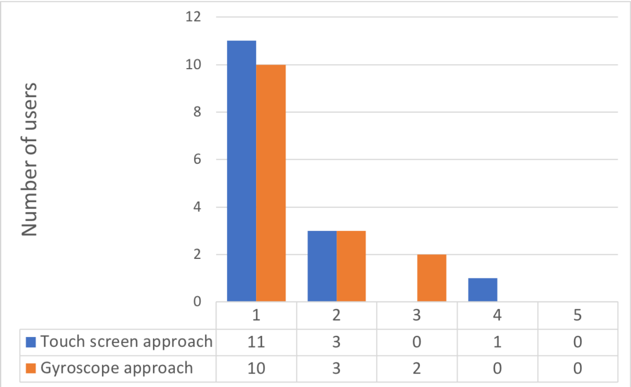


Figure 13: Immersion perception level in task 4.



| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 🟦 Touch screen approach | 0 | 0 | 1 | 7 | 7 |
| 🟧 Gyroscope approach | 0 | 1 | 0 | 2 | 12 |

Figure 14: Azimuth perception level with the touch screen approach (approach 1, in blue) and with the gyroscope approach (approach 2, in orange).



| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 🟦 Touch screen approach | 11 | 3 | 0 | 1 | 0 |
| 🟧 Gyroscope approach | 10 | 3 | 2 | 0 | 0 |

Figure 15: Elevation perception level with the touch screen approach (approach 1, in blue) and with the gyroscope approach (approach 2, in orange).

professional recording, $s_1$, located at 12 o'clock, one UGC recording, $s_2$, at 8 o'clock and another, $s_3$, at 6 o'clock. $93.3\%$ of the subjects localized $s_1$ correctly, and $73.3\%$ localized $s_3$ correctly. Only one subject identified the audio source at 8 o'clock correctly but $66.7\%$ of the participants chose the 9 o'clock direction, showing that the perception of the direction of $s_2$ was close to the real one (and within the $30^{\circ}$ localization precision of humans [11]).

The results from task 4 also show that the technique provides the sense of immersion. Subjects were asked to classify the level of audio immersion they felt while performing this task. A 5-point Likert scale was used, where 1 was *the experience was not immersive* and 5 was *the experience was strongly immersive*. The results are shown in figure 13. While only $20\%$ of the subjects chose answer 5, $60\%$ of them chose answer 4 (*the experience was very immersive*), which results in $80\%$ of the subjects judging the experience as very or strongly immersive.

For task 5, users were asked if they perceived direction variations when moving the visible region (figure 4) horizontally and vertically. A 5-point Likert scale was used, where 1 was *no variation perceived* and 5 was *strong direction variation perceived*. The results are shown in figures 14 and 15.

Figure 14 shows that most users perceive variations in azimuth very easily. On the other hand, figure 15 shows that most users did not perceive variations in elevation. This task shows that users can identify variations in azimuth when the directions of the sounds change dynamically, which indicates that the sense of space and directional audio is not lost with dynamic direction changes. On the other hand, variations in elevation remained unnoticed. This was an expected result as humans do not perceive elevation easily. Since the results for the azimuth depend on the interaction approach (touch screen *vs* gyroscope), this task also show that the sense of directional audio is dependent on the type of user interaction with the application.

## 6. CONCLUSIONS

While the industry has been developing domestic solutions on user immersion that have a particular focus on the visual component, here we focus on audio immersion. We propose a technique that combines user-generated content (and possibly professional recordings) of the same event, to create a final spatialized immersive audio signal that can be combined with video in an interactive tool. The proposed technique spatializes the individual user recordings usings HRTFs, and organizes and synchronizes them with an audio fingerprinting based technique.

We run a user test that showed that the combination of the different recordings from the same event with the proposed technique, where each recording has its own individual characteristics and quality, provides a sense of immersion that the user can experience when listening to the recordings through headphones. The results also show that the proposed technique gives a sense of space and directional audio for azimuth direction changes. Nonetheless, the sense of directional audio is dependent on the type of user interaction with the application.

The current version of the prototype lacks HRTFs individualization. Since different people have different pinnae, the HRTFs set used in our prototype does not adapt equally to all people. As future work, we can extend the prototype to use more HRTFs sets such that it will be possible to choose the HRTF functions that best adapted to each listener, in order to produce more individual and personalized results that best fit the listener.

## 7. REFERENCES

[1] J. Madigan, *Getting Gamers: The Psychology of Video Games and Their Impact in the Peope who Play Them*, Ebook. Maryland: Rowman Littlefield Publishers, 2015.

[2] I. Rodrigues, R. Teixeira, S. Cavaco, V. Bonifácio, D. Peixoto, Y. Binev, F. Pereira, A. Lobo, and J. Aires-de Sousa, “2D spatial audio in a molecular navigator/editor for blind and visually impaired users,” *20th International Conference on Digital Audio Effects*, 2017.

[3] S. Cavaco, D. Simões, and T. Silva, “Spatialized audio in a vision rehabilitation game for training orientation and mobility skills,” *18th International Conference on Digital Audio Effects*, 2017.

[4] Y. Cohen, J. Dekker, A. Hulskamp, D. Kousemaker, T. Olden, C. Taal, and W. Verspage, “Demor, location based 3d audiogame,” 2004.

[5] N. Montan, “AAR - an audio augmented reality system,” M.S. thesis, Department of Microeletronics and Information Technology, KHT, Royal Institute of Technology, Stockholm, 2012.

[6] J. Guerreiro, *Enhancing Blind People's Information Scanning with Concurrent Speech*, Ph.D. thesis, University of Lisbon, Lisboa, Portugal, 2016.

[7] G. Mordido, J. Magalhães, and S. Cavaco, “Automatic organisation, segmentation, and filtering of user-generated audio content,” *25th European Signal Processing Conference (EUSIPCO)*, 2017.

[8] G. Mordido, J. Magalhães, and S. Cavaco, “Automatic organisation, segmentation, and filtering of user-generated audio content,” *IEEE 19th International Workshop on Multimedia Signal Processing*, 2017.

[9] S. E. Chen, “Quicktime VR: An image-based approach to virtual environment navigation,” in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. ACM, 1995, pp. 29–38.

[10] V. Pulkki, *Spatial sound generation and perception by amplitude panning techniques*, Ph.D. thesis, 2001.

[11] D. Wang and G. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE press, 2006.

[12] J. Rees-Jones and D. Murphy, “A comparison of player performance in a gamified localisation task between spatial loudspeaker systems,” .

[13] S. Handel, *Listening: An Introduction to the Perception of Auditory Events*, A Bradford book. MIT Press, 1989.

[14] W.A. Yost and D.W. Nielsen, *Fundamentals of Hearing: An Introduction*, Holt, Rinehart and Winston, 1977.

[15] A. Farina and E. Ugolotti, “Subjective comparison of different car audio systems by the auralization technique,” in *Audio Engineering Society Convention 103*. Audio Engineering Society, 1997.

[16] G. Mordido, “Automated organization and quality analysis of user-generated audio content,” M.S. thesis, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, 2017.