UNIVERSITÀ DI PARMA

UNIVERSITÁ DEGLI STUDI DI PARMA

DOTTORATO DI RICERCA IN
"NEUROSCIENZE"

CICLO XXXV

The perception of audio spatialization during cinematic immersion: an HD-EEG study on the sense of Presence

Coordinatore:
Chiar.mo Prof. Luca Bonini

Tutore:
Chiar.mo Prof. Vittorio Gallese
Chiar.ma Prof.ssa Alessandra Umiltá

Dottorando: Nunzio Langiulli

Anni Accademici 2019/2020 – 2021/2022

INDEX

ABSTRACT

Although many studies have investigated spectators' cinematic experience, only a few of them explored the neurophysiological correlates of the sense of Presence evoked by the spatial characteristics of audio delivery devices. Nevertheless, nowadays both the industrial and the consumer markets have been saturated by some forms of spatial audio format that enrich the audio-visual cinematic experience, reducing the gap between the real and the digitally mediated world. The increase in the immersive capabilities correspond to the instauration of both the sense of Presence, the psychological sense of being in the virtual environment, and embodied simulation mechanisms. While it is well known that these mechanisms can be activated in the real world, they may be elicited even in virtual environments and could be modulated by the acoustic spatialization cues reproduced by sound systems. Hence, the present study aims to investigate the neural basis of the sense of Presence, together with the emotional and physical involvement, evoked by different forms of mediation by testing different sound delivery presentation modes (Monophonic, Stereophonic and Surround). To these aims, a behavioral investigation and a high-density electroencephalographic (HD-EEG) study have been developed. A large set of ecological and heterogeneous stimuli extracted from feature movies were used. Furthermore, 32 participants were selected following the Generalized listener selection procedure. We found a significant event-related desynchronization (ERD) in the Surround condition when compared to the Monophonic condition both in Alpha and Low Beta centro-parietal clusters. We discuss the results as an index of embodied simulation mechanisms that could be considered as a possible neurophysiological correlate of the instauration of the sense of Presence.

CHAPTER 1 – INTRODUCTION

The human experience is a complex and multifaceted phenomenon that encompasses both our cognitive and emotional processes as well as our physical and social interactions. Understanding the intricacies of this experience requires a holistic approach that takes into account the interplay between our brain, body, and environment. In recent years, cognitive neuroscience has made strides in shedding light on the neural mechanisms underlying human experience, particularly through the use of neuroimaging techniques, analyzing the role of the body and environmental context in shaping our experience. Aesthetics, as a fundamental aspect of human experience, is a prime example of this interplay. It encompasses our multimodal perception of the world through the body (Gallese & Guerra, 2013) and allows us to understand how humans interact with the real world and with the cultural and technological artifacts produced by our own intelligence and imagination. The field of neuroaesthetics, which seeks to understand the neural basis of aesthetic experience, is relatively new but has already begun to shed light on the complex mechanisms underlying this experience. One of the most powerful art forms for studying the brain-body system is film. The cinematic experience is unique but directly connected to sensory-motor patterns that connect the viewer with the screen, allowing for a form of immersive simulation that exploits the full potential of our brain-body system (Fingerhut & Heimann, 2022; Freedberg & Gallese, 2007; Gallese & Guerra, 2012). The result is an intersubjective relationship between the viewer and the film that blurs the boundary between the real and imaginary worlds (Gallese & Guerra, 2019).

Cinema is a highly complex art form that combines various elements such as moving images, sound, and music to create a cohesive and immersive experience. While the visual component of cinema has traditionally been the focus of both popular understanding and scientific research, the role of sound in the cinematic experience has been largely overlooked. This bias towards the visual aspect can be attributed to a cultural tendency (Sterne, 2003) to prioritize sight over hearing, as well as the fact that the human brain is wired to process visual information more efficiently than auditory information (Kitagawa & Ichihara, 2002; Sbravatti, 2017).

The integration of sound in cinema began in the 1930s with the advent of synchronized sound technology, which allowed for the integration of voices, sound effects, and music in film. Over the past century, there has been a steady technological development that has contributed to maximize control over the sound field and the ability to create more realistic and believable soundscapes. The use of multichannel sound in cinema has undergone significant developments over the past several decades. In 1992, Dolby Digital Surround was introduced, featuring a 5.1 channel configuration, with five normal channels and a channel dedicated to low frequency effects. This system marked a significant advancement in surround sound technology and became a standard in the film industry. This system allowed for a more immersive and realistic sound experience, with improved spatial localization and sound source area. In 2012, Dolby Atmos was introduced, featuring a 10-channel configuration in a 9.1 setup. This system allowed for even more accurate spatial localization, creating a more realistic and immersive audio experience. It was a big step in the evolution of surround sound technology, and it's widely used today (Sbravatti, 2017).

Empirical research on the relationship between visual images and sound in cinema is limited, however, some authors have suggested that sound has the ability to enhance the immersive

qualities of the two-dimensional cinematic experience (moving images) by creating a sense of three-dimensional reality (Elsaesser & Hagener, 2015). This concept is also supported by the idea that surround sound formats, such as 5.1, have the capability to envelop the viewer in a 360-degree auditory space, as opposed to the traditional 180-degree visual space (DiDonato, 2010).

Despite the acknowledged importance of sound in the cinematic experience, research on the psychological and behavioral effects of multichannel sound in cinema is still relatively limited. Studies on this topic have only recently begun to emerge, and there is still a significant gap in knowledge about how surround sound impacts the viewer's perception and engagement with the film. One study conducted by Lipscomb and colleagues suggests that the use of surround sound can significantly affect the subjective experience of audio-visual content. Participants generally gave more positive verbal evaluations when the stimuli were presented in surround sound. The authors also emphasized the importance of considering the participants' expertise in evaluating the effectiveness of surround sound in enhancing the cinematic experience. It should be noted that the sample size of this study is small, and only six stimuli, several minutes long, were used (Lipscomb & Kerins, 2004).

Some studies have been conducted to investigate the relationship between surround sound, and the "sense of presence" in the cinematic experience. For example, Västfjäll found that six-channel audio reproductions received significantly higher presence and emotional realism scores than stereo (two-channel) and mono (one-channel) reproductions (Västfjäll, 2003). Kobayashi and colleagues (2015) examined the influence of spatialized sounds on the sense of presence in virtual environments by using both physiological and psychological measures (Kobayashi et al., 2015). They utilized a three-dimensional playback system with 96 speakers to reproduce an acoustic stimulus (several people clapping asynchronously around a microphone) under two

experimental conditions: surround condition, monophonic condition (one playback channel). The results showed that the presence ratings for sounds in the spatialized audio condition were higher. Furthermore, physiological measures such as heart rate and skin conductance level indicated that the sympathetic nervous system was activated to a greater extent by sounds in the spatialized audio condition, similar to the responses elicited during intrusions into peri-personal space in real-world scenarios (such as clapping near the participant).

In a 1997 paper, Slater and Wilbur critically examined for the first time the often-confused concepts of immersion and presence by disambiguating their meanings. The two authors defined immersion as an objective property of the technological playback system and presence as the subjective psychological experience of feeling situated in a mediated environment (Slater & Wilbur, 1997).

Cumming and colleagues investigated the relationship between the immersive quality of a mediated environment and the level of presence experienced by the participant. The study employed a meta-analytic approach to examine the overall effect of immersion on presence, specifically exploring how some of the most commonly employed and theoretically interesting immersive features contributed to users' reports of spatial presence. The results of the study found that several immersive features that offer high-fidelity simulations of reality such as surround sound had a significant effect on presence (Cummings & Bailenson, 2015a). Additionally, these results offer some interesting theoretical implications, supporting the formation of presence as outlined by the spatial situational model framework proposed by Wirth et al. (2007).

The spatial situational model framework suggests that the experience of presence in a mediated environment is achieved through a two-step process. The first step is the construction of a spatialized mental model of the mediated environment, in which participants are able to perceive the environment as a space and locate themselves within it. The second step is the embodiment of the mediated environment; participants can potentially interact with this mediated environment. This framework suggests that spatial presence is multidimensional and certain features of the mediated environment, such as surround sound, stereoscopy, and field of view, are particularly important for the formation of a spatialized mental model, and thus for the experience of spatial presence (Wirth et al., 2003).

Furthermore, Gallese proposes that “Film experience and film immersion do not depend just on concepts and propositions, but rely on sensory-motor schemas, which get the viewer literally in touch with the screen, shaping a multimodal form of simulation, which exploits all the potentialities of our brain–body system” (Gallese, 2019), referring to Embodied simulation, a cognitive process described as the ability to simulate the actions, emotions, and sensations of others by activating the same neural circuits that are used to perceive one's own experiences. This mechanism allows individuals to understand the meaning of others' behaviors and experiences by directly relating to them through the activation of sensory-motor representations in bodily format (Gallese, 2009).

The neural substrate of the embodied simulation mechanism for actions corresponds to a particular functional group of neurons called "mirror neurons," first discovered in area F5 of macaques during an intracortical recording of the premotor cortex (DiPellegrino et al., 1992). Keysers and colleagues suggest that these neurons encode actions abstractly regardless of the source of the information (auditory or visual), so multisensory integration can be used to provide

meaningful representations and recognize relevant actions in the environment (Keysers et al., 2003). The mu rhythm, the alpha rhythm commonly recorded over sensorimotor areas, is a common marker of the mirror mechanism in humans (Muthukumaraswamy et al., 2004; Muthukumaraswamy & Johnson, 2004a, 2004b). Voluntary action execution and observation correlate with event-related desynchronization (ERD) in alpha bands, generally between 8 and 13 Hz, and in low beta bands, generally between 14 and 18 Hz, and recorded over sensorimotor areas (Perry et al., 2010; Pfurtscheller et al., 1994; Toro et al., 1994).

The only study that investigated the effect of acoustic spatialization on the sensation of presence using electroencephalography (EEG) was by Tsuchida and colleagues. They used a surround sound reproduction system called BoSC (62 speakers), designed to simulate the presence of other individuals or objects by providing a highly realistic sound field reproduction, to reproduce an acoustic stimulus under two experimental conditions: spatialized condition, monophonic condition (one playback channel). EEG results showed that mu rhythm suppression occurred for action-related sounds, but not for non-action-related sounds. Furthermore, this suppression was significantly greater in the surround (62-channel) condition, which generates a more realistic sound field, than in the 1-channel speaker condition. Additionally, the motor cortical activation for action-related sounds was influenced by the sense of presence perceived by the study participants, as they perceived significantly higher sound realism in the surround condition (Tsuchida et al., 2015). It should be noted that this study had a small sample size and the results should be considered in light of the limitations of the study design. Further research with larger sample sizes and different stimuli is needed to fully understand the effect of acoustic spatialization on the sensation of presence.

1.1 Aims of the study

The aim of this study was to investigate the time course and neural correlates of audio presentation modes on participants' sense of presence during cinematic immersion. To test this hypothesis, we designed a behavioral experiment and a high-density electroencephalographic (HD-EEG) experiment. We initially selected a diverse set of naturalistic stimuli, consisting of validated cinematic excerpts, which were presented to participants in different audio presentation modes (Monophonic, Stereophonic, and Surround) while their neural and behavioral responses were measured. Our hypothesis was that the enhanced spatialization of sound in the surround presentation mode would lead to a greater embodiment as compared to the monophonic and stereophonic presentation modes.

CHAPTER 2 - MATERIALS AND METHODS

2.1 Experimental Stimuli Selection and Validation

The first stage of this research project was the extraction, selection and validation of the cinematic excerpts to be used as stimuli for the next experiments. Bearing in mind that previous studies often used few stimuli, in the present project we employed a large set of naturalistic and, to some degree, heterogeneous stimuli extracted from movies. This approach will, also, allow for more generalizable results (Sonkusare et al., 2019).

All participants provided written informed consent to participate in the studies, which were approved by the local ethical committee "Comitato Etico Area Vasta Emilia Nord" and were conducted in accordance with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards (World Medical Association, 2013).

2.1.1 Participants

The experimental sample consisted of 100 participants (67 males and 33 females, with mean age *M* of 23.6 and standard deviation *SD* of ± 4.9 years, within a *range of* 18 to 38 years).

Recruitment took place *online* through the Prolific platform (Palan & Schitter, 2018). The selection was based on three characteristics: participants had to be between the ages of 18 and 40, had to be fluent in English, and with no prior history of psychiatric disorders. All participants were rewarded through the Prolific platform with an hourly rate of approximately £7.5/hour.

2.1.2 Stimuli

The cinematic excerpts were extracted by an experimenter after watching 31 Academy Award-nominated feature films in the categories of "Best Sound" and "Best Sound Editing" from the editions between 1979 and 2019.

A set of guidelines were established to select specific portions of cinematic sequences for extraction. These were chosen based on three criteria:

- No music should be present in the scene: this criterion aims to ensure that the emotional response elicited by the scene is not influenced by the presence of music, which can often be used to enhance emotional responses in films (Wöllner et al., 2018).
- There should be no dialogue in the scene: dialogue can often give context or emotional information about a scene, which can activate cognitive processes related to human voice perception that are not relevant to the experiment (Latinus & Belin, 2011).
- The scene should be characterized by good acoustic spatialization: this criterion aims to ensure that the scene is acoustically well-produced and that the acoustic aspects of the scene can provide a more immersive and realistic experience for the participants.

We extracted 10 seconds-long cinematic excerpts in the original PCM Surround 5.1 format from the MatrosKa audio-video container format and converted it in MPEG-4 Part 14 (ISO/IEC 14496-14:2003) format. No modifications to the bitrate and sampling frequency were made.

From this initial selection procedure, we obtained 185 **sound-only** cinematic excerpts.

2.1.3 Procedure

The experimental paradigm for stimulus validation was written in PsychoPy (v3.0) and hosted by the Pavlovia service for *online* program execution and management of collected data (Peirce et al., 2019). The experimental sample was randomly divided into 10 groups of 10 participants each. Nineteen different stimuli were randomly assigned to nine groups and 14 different stimuli to one group, for a total of 185 stimuli. A pilot group was utilized in order to assess the feasibility of the experimental paradigm, to ensure that the procedure was being executed as intended and to identify any potential issues with the methodology. We assigned a limited number of stimuli (n=14) to this pilot group.

The participants were instructed to listen to randomly presented stimuli, exclusively on a PC, using headphones or external speakers. After each stimulus, participants were prompted to respond to one of three questions, selected randomly, that appeared on the screen:

1. "*How would you judge the **dynamicity** of the scene?"*
2. "*How would you judge the **Emotional Valence** of the scene?"*
3. "*How would you judge the **Emotional Intensity** of the scene?"*

During the experimental session, each stimulus was repeated three times and was followed by a different question after each presentation for a total of 57 trials (42 only for the first control group) and for a total duration of about 15 minutes.

The questions required the participant to make a personal judgment about the Dynamism, Emotional Valence, and Emotional Intensity of audio-visual stimuli using a Visual Analogue Scale (VAS) ranging from 0 to 100 for the questions on Dynamism and Emotional Intensity, and from -50 to +50 for the question on Emotional Valence. The questions were aimed at identifying

stimuli with high arousal and negative emotional valence, measured according to Russell's dimensional theory of emotion (Russell, 1980). In addition to VAS scores, response times (RTs) were recorded.

2.1.4 Stimuli Analysis and Selection

It is important to note that this experiment was conducted online without the direct supervision of the experimenters. As a result, the participant's execution could not be fully controlled, which may have resulted in a lower quality of data compared to that obtained from laboratory-based studies (Sauter et al., 2020). In order to exclude potential participants' errors such as accidental mouse clicks or distractions during the execution, scores assigned with response times (RTs) less than or equal to 500 milliseconds or greater than 10 seconds were excluded from the analysis. This resulted in the exclusion of 303 data points from the analysis. Outlier detection was also performed on the scores assigned to each stimulus for each question and RTs. This resulted in the exclusion of additional 183 data points from the analysis.

There is evidence to suggest a correlation between immersion and intensity of perceived emotions as found by Visch and colleagues (Visch et al., 2010). Additionally, negative emotions have been shown to have processing priority, triggering defense mechanisms through early activation of the amygdala (Figueiredo et al., 2003; Leppänen & Nelson, 2009). Furthermore, research has shown that audio-visual stimuli with negative valence extracted from films can induce greater arousal than stimuli with positive valence (Fernández-Aguilar et al., 2019)

With these findings in mind, among the 185 stimuli, those with high Dynamicity, high Emotional Intensity, and negative Emotional Valence were selected to be used as stimuli for the next experiments.

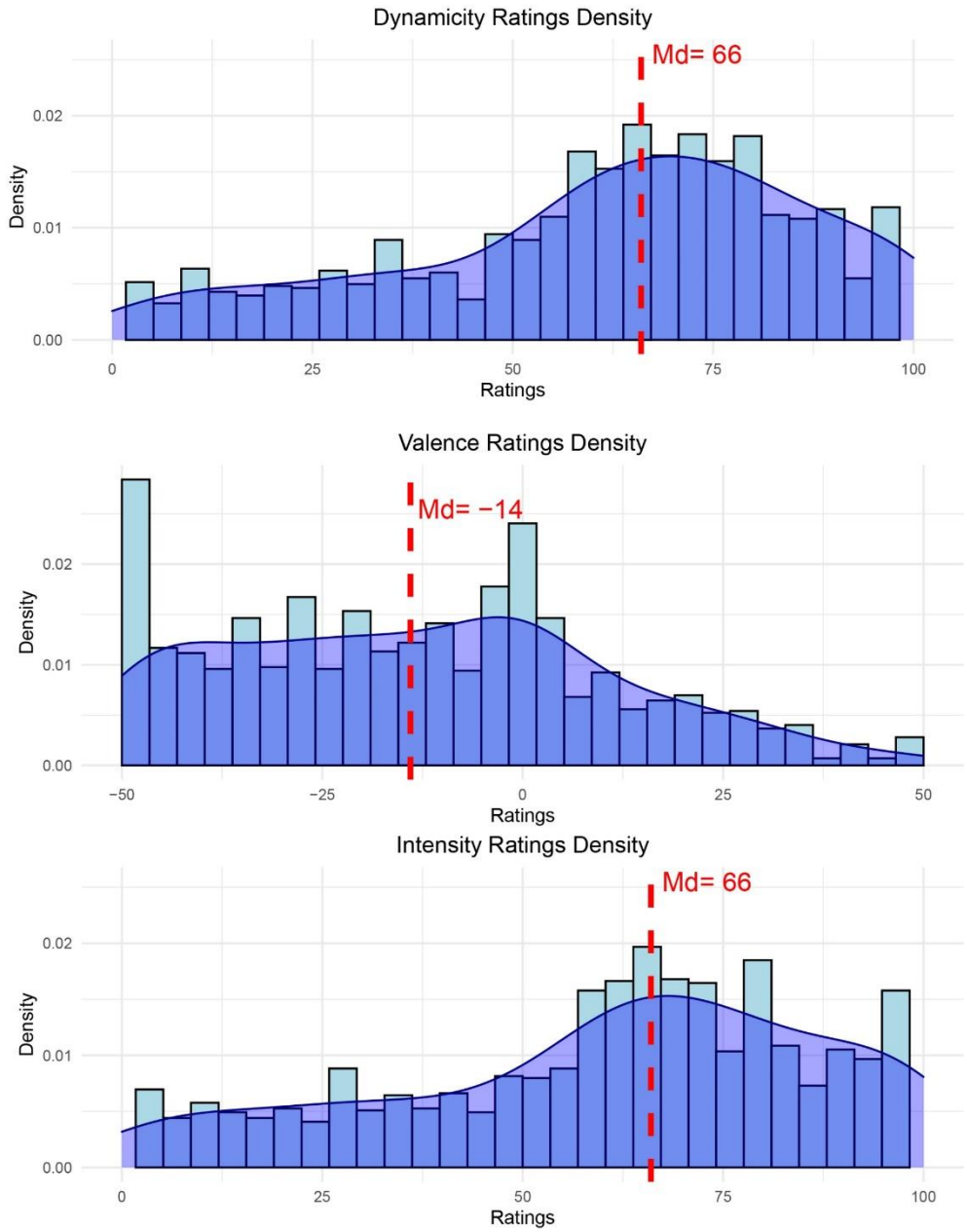

Figure 1. *Scores Density Distributions and Medians (dashed red line).*

We used the median as a cutoff value in order to select the stimuli (see Figure 1):

- The stimulus Dynamism score had to be <u>higher</u> than the sample Dynamism median score (*Md* Dynamism = 66);

- The stimulus Emotional Intensity score had to be higher than the sample Emotional Intensity median score (*Md E*motional Intensity = 66);
- The Emotional Valence score had to be lower than the sample Emotional Valence median score (*Md* Emotional Valence = -14).

We selected **50** stimuli that had the desired characteristics, from the initial sample of 185 stimuli sample, with a mean Dynamicity score of 75, mean Valence score of -23 and mean Intensity score of 78 (Table 1).

	Mean	**SE**	**Min**	**Max**	**IQR**	**Range**
Dynamicity	74.95	0.83	67.2	94.7	9.1	27.6
Valence	-23.21	0.84	-40.4	-14.4	8	26.1
Intensity	77.71	0.77	67.6	90.1	8.14	22.5

Table 1. *Summary of Selected Stimuli Scores (SE= Standard Error, IQR=Interquartile Range).*

2.1.5 Acoustic Features Analysis

In order to perceptually uniform the acoustic features of our stimuli we extracted pitch and loudness, two perceptual properties of sound that can be quantified, measured and normalized.

Pitch is dependent on the frequency of a sound and can be quantified in terms of its fundamental frequency and the amount of overtones (harmonic or not harmonic). The pitch of the acoustic stimuli was extracted using the YIN algorithm, a widely used pitch detection method in speech and music analysis implemented in MATLAB (Figure 2). The extracted pitch values were then analyzed to identify the range of pitches present in the stimuli sample. The minimum and maximum pitch values were determined and outliers, defined as pitch values deviating more than

3 standard deviations from the mean, were identified and excluded from the sample. We excluded 11 stimuli from the initial sample of 50 stimuli. This exclusion was done in order to ensure that the stimuli pitch values were not skewed and to ensure that the stimuli sample had consistent and comparable perceptual features across all stimuli, allowing for a more uniform effect on the participants.

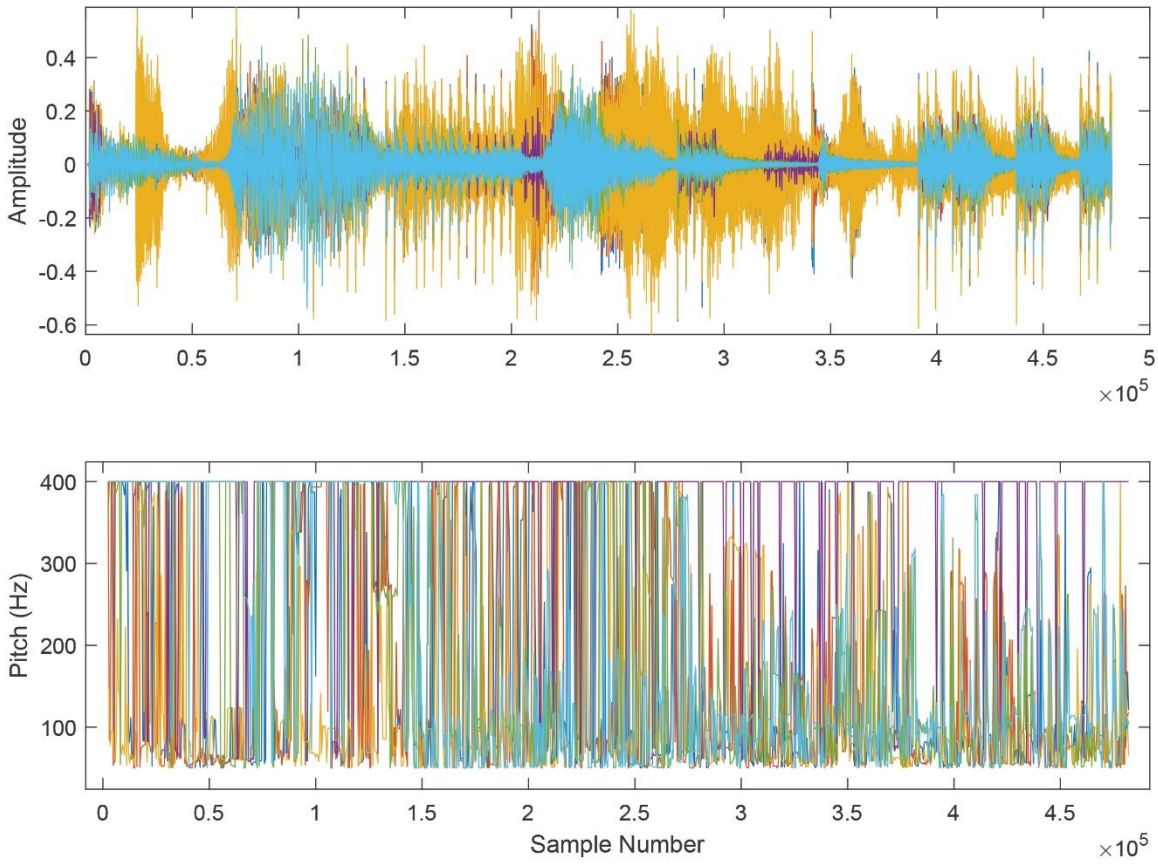

Figure 2. *YIN Pitch Amplitude.*

Loudness, on the other hand, is dependent on the energy or pressure of a sound. The Loudness Unit Full Scale (LUFS) is a commonly used standardized measurement of audio loudness described in the European Broadcasting Union (EBU) standard R 128 (EBU, 2014), where 1 LU

corresponds to a relative measurement of 1 dB on a digital scale and 0 LU = -23 LUFS. There are three distinct methods used to measure loudness (Figure 3):

- Momentary loudness, which uses a sliding time window of 400 ms to describe instantaneous loudness;
- Short-term loudness, which uses a sliding time window of 3 seconds to describe a short time average loudness;
- Integrated loudness, which describes average loudness over the duration of the entire sequence.

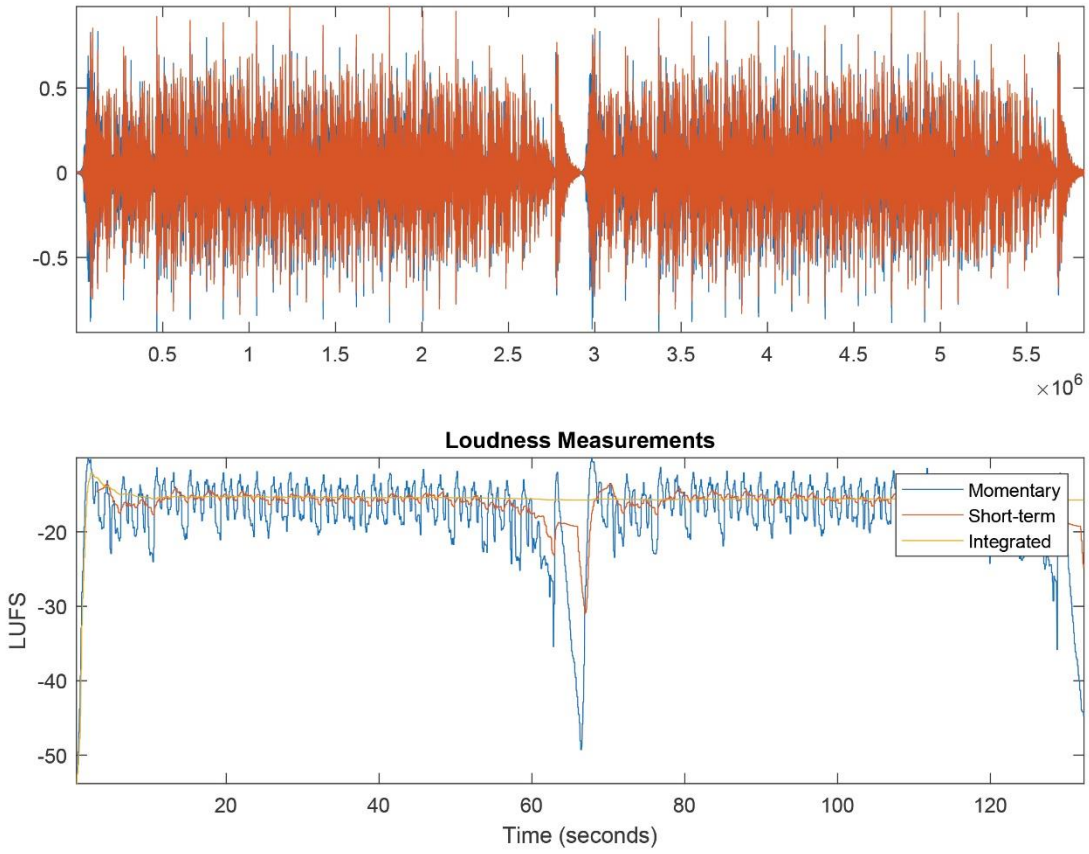

Figure 3. *Momentary loudness, Short-term loudness, Integrated loudness.*

In order to eliminate the potential influence of loudness level differences during the acoustic reproduction of stimuli we followed the European Broadcasting Union EBU R128 recommendation that defines the target loudness level at -23 LUFS. To normalize the stimuli to this level, a simple loudness adjustment operation was applied to the integrated loudness using MATLAB. This operation was performed on all the stimuli in the sample, ensuring that the integrated loudness levels of all stimuli were consistent to -23 LUFS. In addition to normalizing the integrated loudness of the stimuli, the procedure also involved analyzing the peak loudness and the loudness range of all the stimuli. Peak loudness refers to the highest level of loudness within a sound or audio file, while loudness range is the difference between the highest and lowest level of loudness within a sound or audio file. The minimum and maximum peak and range values were determined for all the stimuli. Outliers, defined as values deviating more than 3 standard deviations from the mean, were identified and excluded from the sample. We excluded 12 stimuli from the initial sample of 50 stimuli.

Considering the exclusion of both pitch and loudness related outliers the final stimuli samples was composed by **27** stimuli.

2.1.6 Acoustic Qualitative Descriptions

In order to provide a detailed understanding of the acoustic properties of the stimuli selected we provided an acoustic description from a qualitative perspective, focusing on the different elements present in the acoustic sequence and how they change over time with some emphasis on the different elements present in the surround channels (Table 2). To clearly distinguish between the stimuli, each one was assigned a unique alphanumeric code.

Stimulus	Qualitative Description
AS04	Gunshots, slight clang. Acute impact of bullet on the driver's body. Blood spurts on the glass. Machine gun fire. Wagon motor, explosion, wagon impact on another car and then on ruins. Soldiers' groans in center channel. Wagon arrests. Ricochet. Metal sound of bullets hitting the ground. LFE regularly used. Impacts and gunshots prominent also in surround channels.
AS07	Shots, ricochets, glass breaking. Metal sounds of bullets falling on the floor. Impact of the bullets on the wall. Point of audition clearly shifts consistently with the visual perspective. Surround channels are prominent (the left a bit more so), containing sounds of gunshots. LFE used slightly to enhance the gunshots.
BR02	Moderately regular thuds of the head on the wall, groans. Noises of the wall breaking, fragments banging on the floor and impact of the two bodies falling. Most sounds are spread in all the five main channels. Point of audition shifts consistently with the point of view: full-frequency sounds in the room where the two men are, same sounds muffled heard through the wall in the adjacent room. LFE use subtly and then prominently, when the wall collapses and the two man fall.
BR09	Mechanical noise of the missile approaching to the building. Explosions, glass breaking, debris impacts. The sound stream is slightly fragmented. Afterwards, the soundscape rarifies, there is far reverberation, while the point of view moves outside the building (long shot). Some subtle groans in center channel, spread in the surrounds which are regularly used – as is the LFE, albeit being prominent especially during the first explosions.
CA04	Silent sound field, low background of the aircraft. Slight sound of the patch removed from the finger. Sudden, intense acoustic blast, reinforced by LFE and surround channels as well (which continue to be used though less prominently). Noisy sound field: air flux,

	impact of objects also in surrounds, metal and electromechanical sounds.
CA07	Noises of impacts, electromechanical sounds, clatter. Hiss and air also from surround channels. LFE only at beginning and end.
CA11	Spacious underwater point of audition, muffled. Water movements, hiss, string rubbing on metal, clink, creak. Ominous deep and cracking noises suggesting the movement of the aircraft. Sounds subtly spread in the surround channels.
CA13	Rain, which spreads also in the surround. Crescendo of the turbine whir, explosion (includes LFE) accompanied by a couple of whistles. Right surround contains a very subtle whoosh (water or wind).
CA14	Much muffled underwater sounds. Point of audition moves repeatedly under and above the water level. Sounds of water and rain (mixed also in the surround channels), wind, low animal-like sound of the airplane, distant electronic tones.
CA15	Wind, rain, water moving and gurgling. Left surround contains mostly wind, right surround rain only.
HR02	Shouts, moans, spread throughout the front channels (and for an instant in the surround ones as well). Gunshots, hisses, explosions, bloody impacts of bullets on bodies, footsteps running on the ground. Surrounds and LFE regularly used.
HR03	Shouts, gunshots, bloody impacts of bullets on bodies, body impacts on the ground, hisses and explosions, hand grenade unlocked and thrown. Surrounds and LFE regularly used.
HR05	Groans, mostly from the center channel but slightly spread throughout the other main channels. Gunshots, hisses and explosions, machine gun fire, bullet impact on the ground and on a soldier's metal helmet. Surrounds and LFE regularly used. Two whooshes before the first bomb explosion are prominent in the right surround, consistently with the visual perspective.

HR09	Hisses and explosions, shouts and groans (mostly from the center channel), gunshots, Wilhelm scream, clang, impact on ground. Surrounds and LFE regularly used.
HR10	Gunshots, prominent in left surround. Shouts and groans, mostly from the center channel. Hisses and explosions, gunshots, various impacts (bullets on bodies, bodies on other bodies and on the ground). Surrounds and LFE regularly used,especially some gunshots from the left surround.
HR11	Shouts and groans, mostly from the front channels. Gunshots, body impacts, hisses and explosions, mold falling over the soldiers, blood spurts. Surrounds and LFE regularly used.
HR15	Cannon shots (much reinforced by LFE), explosions, distant shouts (spread throughout the five main channels), nearby breath (center channel only). Point of audition changes in terms of distance. Surrounds and LFE regularly used.
ID01	Weapon release, explosion, glass shattering, screams (front channels, slightly spread in the surrounds), fire. Subtle growls. Surrounds and LFE regularly used.
ID05	Roar of flames expanding towards the camera. Squeals, metal noises, explosions, glass shattering, screams, siren, animal-like voices and roars subtly mixed with the noises. Body impacts on car's hood, breaking the windshield. Surrounds and LFE regularly used.
JU01	Whir, impacts of the airplane on the branches and the trees, alarm, glass shattering, a woman's scream, mechanical sounds, debris. Surrounds are prominent until the airplane stops, then they mute. LFE enhances most of the impacts. Various acoustic accents.
LS11	Gunshots, explosions and debris in the center channel, bullet impacts. Bazooka launches missile which impacts on the ground. Human movements in the woods. Two musical chords (electronics and maybe low strings) in the background. Surrounds and LFE regularly used, the latter especially for the two explosions.

LS16	Gunshots and ricochets, nearby groan in the center channel, distant shouts in the center channel but slightly spread throughout left and right. Much reverberation, also thanks to the surrounds which are regularly used. LFE enhances gunshots.
LS19	Gunshots to the helicopter, shouts in the center channel, descending mechanical noises, blades whirring, thump of the helicopter breaking in two parts. Strong impacts of the falling helicopter on the mountain, some explosions. Surround (especially the left one) only at the beginning, then slightly for the first impact on the mountain, and in the end. Point of audition gets further consistently with the point of view.
MC04	Noise of the cannon shot, series of impacts, wood clatter, roar of the ship, whooshes, sounds from the sea (objects sinking, ship sailing). Surrounds and LFE regularly used, the former being particularly prominent.
MC09	Wind howling (almost choir-like), sea waves, squeak, ropes whooshes, roar of the ship, mast sinking in the sea. LFE used for an instant only, almost inaudible. Surrounds used constantly and prominently.
SK01	Wheels on the rails, electricity buzzing, screeches, impacts, debris, glass. LFE used regularly and intensely. Surround channels are always used, especially during the first impact of the train and the one on the camera. The right one is slightly more prominent (along with the right front one), consistently with the visual perspective.

Table 2. *Stimuli Acoustic Qualitative Descriptions.*

Participants were also asked to listen to each stimulus and indicate if they recognized any broadly defined action-related sound. If they stated that there was an action-related sound present, they were asked to write down what action sound they recognized. This information was then used to verify that participants were able to recognize the correct action-related sounds in each stimulus.

2.1.7 Acoustic Formats

The original multi-channel Surround audio stimuli, encoded in 5.1 channels format were used to derive a Stereophonic sound version (2.0 channels) and a Monophonic sound version (1 channel). The conversion was performed using MATLAB by downmixing the Surround sound to the Stereophonic sound, and then further downmixing to the Monophonic sound averaging the amplitude values of the channels. Care was taken to ensure that the downmixing process did not alter the temporal and spectral characteristics of the stimuli, such as the pitch and loudness, to maintain the integrity of the original stimuli. The original PCM format was converted in Waveform Audio File Format (sample rate= 48000 Hz; sample size= 16 bits).

2.2 Acoustic Experimental Setup

A silent audiometric cabin (IAC-Acoustics) 2 meters high, 2.5 meters wide, and 2.1 meters deep was set up with a *surround* sound reproduction system consisting of five APART MASK4C speakers (impedance 8 Ohms), one APART SUBA165 *sub-woofer* (impedance 4 Ohms), and a TV set (LG UHD, 42 inches, 16/9). The participant was positioned at the center of the silent audiometric cabin, while the six speakers were positioned and oriented as shown in Figure 4 following the ITU-R BS.1116-1 recommendation, so as to direct the sound to a central point that identified the correct listening position (ITU-R., 1997). Specifically, the television screen and three speakers were attached on VESA mounts, at 95 cm height, arranged as follows: two at the sides of the television, (left channel “L”, and right channel “R”) and one above it (center channel “C”). At the rear corners, two floor stands, 95 cm high, were placed with the last two speakers (left surround channel “Ls”, and right surround channel “Rs”); the *sub-woofer* (“LFE” or *Low Frequency Effects*) was placed on the floor, approximately below the “L” channel.

Surround sound stimuli were reproduced by all six channels in Surround Reproduction Mode, Stereophonic sound stimuli were reproduced by “L” and “R” channels in Stereophonic Reproduction Mode, and Monophonic sound stimuli were reproduced only by “C” channel in Monophonic Reproduction Mode.

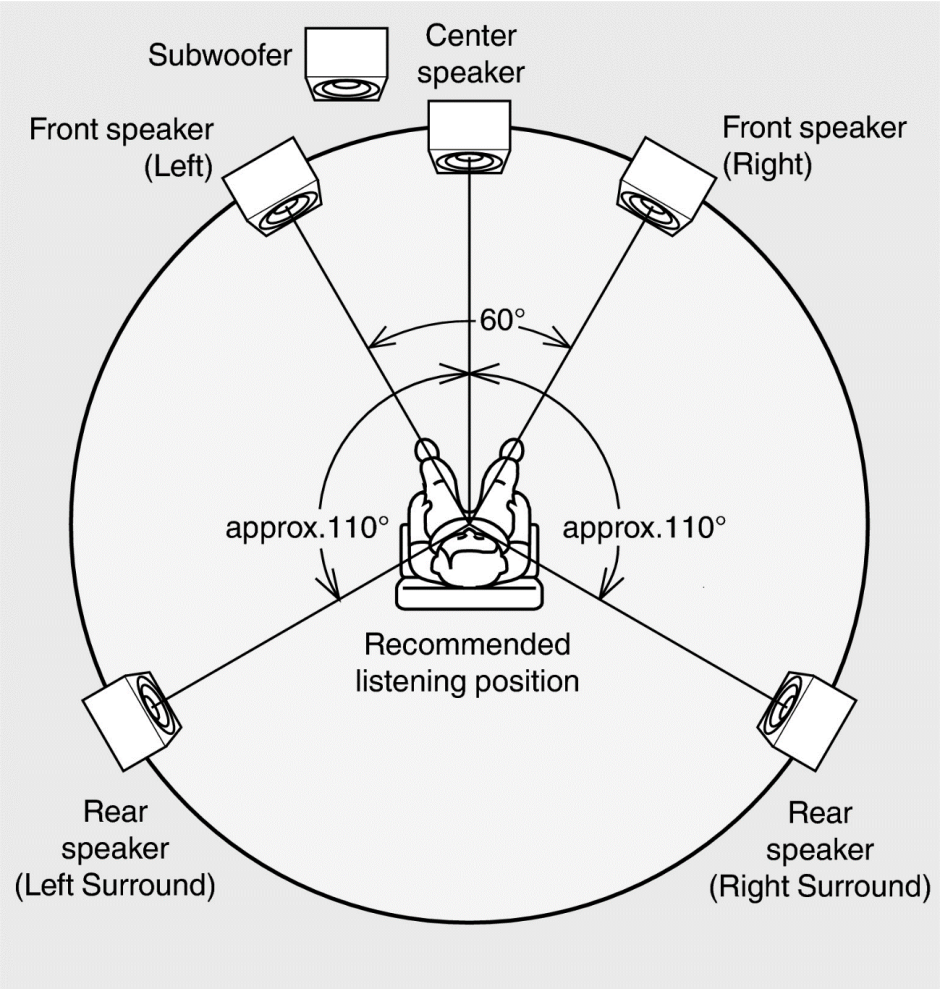

Figure 4. *Surround System Layout - ITU-R Recommendation BS.1116-1.*

The surround sound system was driven by a DENON AVR-X1600H amplifier, positioned outside the silent audiometric cabin, which amplified the audio signal. In order to ensure an accurate and precise auditory experience, the acoustic reverberation of the cabin and the phase of acoustic reproduction were measured. This was done using the microphone provided with the

amplifier and positioning it at the central listening position. The phase response of the different audio channels was then corrected using a room correction technique implemented in the Audyssey software (Paul & A., 2009).

At this point, the sound pressure level of all stimuli reproduced by the surround sound system was recorded with a sound level meter (Gain Express, applied standard IEC651 type 2, type ANSI 2 SI 0.4) placed at the listening position. Analysis of the recordings made it possible to verify that the set level of acoustic reproduction was below the hazardous hearing threshold (85 dB, A-weighted, for 8 consecutive hours) defined and standardized by the National Institute for Occupational Health and Safety (NIOSH) in the ONE (Occupational Noise Exposure) recommendation (Murphy & Franks, 2002).

2.3 Generalized Listener Selection Procedure

In order to conduct a thorough and rigorous study on the perception of audio spatialization during cinematic immersion and to ensure that any differences in perception of audio spatialization are not due to extraneous factors, it is crucial to select participants who meet certain criteria:

- Normal hearing: participants should have normal hearing abilities, as determined by an audiometric test, in order to ensure that any differences in perception of audio spatialization are not due to hearing impairments.
- Age range: it is important to consider the age range of the participants, as age can affect auditory perception and cognitive processing. It is recommended to limit the age range of the participants to a specific range, such as 18-45 years old, to minimize the effects of aging on the results (Howarth & Shone, 2006).

- Participants had to be able to perceive differences between different acoustic stimuli and between different modes of sound reproduction: this criterion ensures that the participants have the ability to perceive and distinguish between different acoustic features and different sound reproduction modes.
- Participants had to be "un-trained/naive subjects" as described in ITU-T Recommendation P.800 (ITU-R., 1996): This criterion ensures that the participants do not have specific technical skills in the evaluation of acoustic reproduction systems, which would affect the results. This criterion should be met in order to generalize the results to the population and not only to a specific group of "experienced/expert/trained subjects" described in ITU-T Recommendation P.832 (ITU-R., 2000).

To check for all the listed requirements, an adaptation of the Generalized Listener Selection (GLS) procedure described by Zacharov and colleagues (Bech & Zakharov, 2007; Mattila & Zakharov, 2001) was used. Our adaptation of this procedure included:

- Six Questionnaires: questionnaires were used to gather information about the participants' demographics, hearing history, knowledge of spatialization techniques, and movie-watching experience as well as to control for other factors such as age, gender, education level, and cultural background.
- Audiometric test: this test was used to measure the participants' hearing abilities and ensure that they met the criterion for normal hearing. This test typically involves playing a series of tones at different frequencies and intensities and asking the participants to indicate when they can hear them.

- Two screening tests to assess the discriminative abilities of both loudness and sound source localization.

The audiometric and screening test administrations were performed using MATLAB (ver. R2021a).

2.3.1 Questionnaires

Each participant had to fill out via Google Forms, before the experiment, a battery of questionnaires composed as follows:

1. GLS-1 (Generalized Listener Selection): a questionnaire consisting of 13 items constructed ad hoc by the experimenter investigating general aspects of interest: participants should not have any specific skills or training in the evaluation of surround systems and should not be professionally involved in acoustics. Additionally, the participants should not have previously participated in formal listening tests, as this would skew their ability to accurately perceive and report on the audio spatialization. (Bech & Zacharov, 2007).
2. QEAV (Questionnaire on Audio-Visual Experience): a questionnaire consisting of 11 items constructed ad hoc by the experimenter that measures the degree of knowledge in filmmaking and audio-visual editing.
3. B-MEQ (Brief Music Experience Questionnaire): a questionnaire consisting of 53 items that measures the degree of knowledge in music as both a listener and a musician, divided into six subscales (Werner et al., 2006):

a. *Commitment to Music*: the centrality of pursuit of musical experiences in the participant's life;
b. *Innovative Musical Aptitude*: self-reports of musical performance ability and of the ability to generate musical themes and work;
c. *Social Uplift*: the experience of being stirred and uplifted in a group-oriented manner by music;
d. *Affective Reactions*: affective and spiritual reactions to music;
e. *Positive Psychotropic Effects*: calming, energizing, integrating reactions,
f. *Reactive Musical Behavior*: behavioral responses including humming, swaying, etc. along with music.

2. IRI (Interpersonal Reactivity Index): a questionnaire consisting of 28 items that quantify empathic response, defined as an individual's ability to vicariously affect the emotional states of observed others (Albiero et al., 2006). The questionnaire consists of four subscales:
    a. *Perspective Taking*: refers to the ability to adopt the psychological point of view of others, as a cognitive and social-cognitive process. It is considered a key aspect of empathy and it allows individuals to understand and predict the behavior of others by simulating their mental states.
    b. *Fantasy*: is considered a cognitive and affective process, it refers to the ability of an individual to imaginatively project themselves into the feelings and actions of fictitious characters in books, movies, and plays. This construct is often used as a measure of fantasy proneness, as it taps into individuals' tendencies to engage in mental simulations of fictional scenarios.

c. *Empathic Concern*: this is a construct that assesses "other-oriented" feelings of sympathy and concern for unfortunate others. It is considered a key aspect of empathy, as it reflects the ability to feel and respond to the emotional states of others.

d. *Personal Distress*: this is a construct that measures "self-oriented" feelings of personal anxiety and unease in tense interpersonal settings. It is considered a negative aspect of empathy, as it reflects the individual's level of discomfort and stress when exposed to the negative emotions of others.

3. ITQ (Immersive Tendencies Questionnaire): a questionnaire consisting of 34 items that measures participants' ability or tendency to immerse themselves in different mediated environments (Witmer & Singer, 1998).
4. VMIQ-2 (Vividness of Movement Imagery Questionnaire): a questionnaire consisting of 36 items on bodily self motor imagery (Roberts et al., 2008). The questionnaire is divided into three subscales:
    a. *External self Visual Imagery* (EVI): is a construct that measures third-person visual imagination;
    b. *Internal first-person Visual Imagery* (IVI): is a construct that measures first-person visual imagination;
    c. *Kinesthetic Imagery* (KIN): is a construct that measures imagination of kinesthetic sensations related to imagined movement.

At the end of the experimental session, the participant had to fill out the following two questionnaires:

1. F-IEQ (Film Immersive Experience Questionnaire): a questionnaire consisting of 31 items for measuring the viewer's immersive ability while watching films (adapted to the acoustic experience only), divided into four subscales (Rigby et al., 2019):
    a. *Captivation*: this is a construct that measures the viewer's enjoyment, level of interest, and motivation.
    b. *Real-world dissociation*: this is a construct that measures the viewer's awareness of their real-world surroundings while watching the movie.
    c. *Comprehension*: this is a construct that measures the viewer's understanding of the concepts and themes presented in the movie.
    d. *Transportation*: is a construct that measures the viewer's sense of experiencing the events portrayed in the movie as if they were happening to themselves, and how much they felt they were located in the world portrayed in the movie.
2. Post-Q (Post-Experiment Questionnaire): questionnaire consisting of nine items constructed ad hoc by the experimenter to obtain more information about the participant's familiarity with the stimuli and playback system used during the experimental session.

2.3.2 Audiometric Test

Each participant was subjected, in the silent audiometric cabin, to a brief tonal audiometric test to measure hearing threshold using the "Frequency Response of the Ear, Hearing Test" application developed in MATLAB (Rawashdeh, 2021). The participant wore *over-ear* headphones (Audio-technica ATH-MSR7). A pure tone to both ears was simultaneously played, starting from a frequency of 1 KHz up to a frequency of 16 KHz with intervals of 1 KHz, at decreasing loudness levels. The participants' task was to indicate whether he heard the played

tone pressing a button. The next frequency in the series was played only when the participant failed to press the button, indicating that they were unable to hear the previous tone. The resulting audiometric curves were then compared with audiometric curves for otologically normal participants differentiated by age according to the ISO Standard 7029 (ISO, 2017) and participants whose hearing performance did not appear to be in the normal range were excluded from the experimental sample.

2.3.3 Screening Test 1: Loudness

The first screening test was conducted to verify the participant's ability to discriminate sounds played at different loudness levels. The participant was positioned in the central listening position of the silent audiometric cabin. Five different acoustic stimuli, 500 milliseconds long, were created and played simultaneously on five channels of the surround sound system (LFE channel excluded). The stimuli consisted of pink noise sampled at 48000 Hz with different loudness levels, indicated in Loudness Units relative to Full Scale (LUFS).

The highest loudness level was set at -23 LUFS, while the loudness level of other stimuli was set at 3 LUFS lower than the previous one, based on the minimum audible difference described by Larsen and colleagues (Larsen et al., 2008). The stimuli had the following loudness levels:

- -23 LUFS;
- -26 LUFS;
- -29 LUFS;
- -32 LUFS;
- -35 LUFS.

The participant was given a forced-choice task, in which they had to listen to two stimuli played in succession and indicate which stimulus was played with the higher loudness level or whether both stimuli were played with the same loudness level. Stimuli were presented in all different possible pairings (including pairings between equal stimuli) for a total of 25 random trials (Table 3). Each stimulus was accompanied by a visual cue, the first stimulus in the pair was accompanied by the label "First" while the second stimulus played was accompanied by the label "Second".

Trial	**"First"**	**"Second"**	**Trial**	**"First"**	**"Second"**
1	-23 LUFS	-23 LUFS	14	-29 LUFS	-32 LUFS
2	-23 LUFS	-26 LUFS	15	-29 LUFS	-35 LUFS
3	-23 LUFS	-29 LUFS	16	-32 LUFS	-23 LUFS
4	-23 LUFS	-32 LUFS	17	-32 LUFS	-26 LUFS
5	-23 LUFS	-35 LUFS	18	-32 LUFS	-29 LUFS
6	-26 LUFS	-23 LUFS	19	-32 LUFS	-32 LUFS
7	-26 LUFS	-26 LUFS	20	-32 LUFS	-35 LUFS
8	-26 LUFS	-29 LUFS	21	-35 LUFS	-23 LUFS
9	-26 LUFS	-32 LUFS	22	-35 LUFS	-26 LUFS
10	-26 LUFS	-35 LUFS	23	-35 LUFS	-29 LUFS
11	-29 LUFS	-23 LUFS	24	-35 LUFS	-32 LUFS
12	-29 LUFS	-26 LUFS	25	-35 LUFS	-35 LUFS
13	-29 LUFS	-29 LUFS	-	-	-

Table 3. *Loudness Screening Test Trials.*

At the end of each playback, the question "Which sound had the loudest volume?" was displayed on the screen and the participant had to answer, within 5 seconds, with the mouse by choosing from the options "First", "Equal" and "Second".

The cut-off criterion for participants' performance was set at 80% correct responses (20 out of 25) as indicated in the description of the matching tests on auditory abilities performed by Bech and Zacharov (Bech & Zacharov, 2006).

2.3.4 Screening Test 2: Localization of the Sound Source

The second screening test was conducted to verify the participant's ability to detect the sound source (sound localization). The participant was positioned in the central listening position of the silent audiometric cabin. Five different acoustic stimuli (Table 4), 500 milliseconds long, were created and played simultaneously on five channels of the surround sound system (LFE channel excluded). The stimuli consisted of a background pink noise sampled at 48000 Hz reproduced in four channels and a pure 1000 Hz cue tone reproduced by only one channel in each stimulus (Table 4). Loudness difference between background noise (-48 LUFS) and cue tone (-23 LUFS) was set at a -25 LUFS. Each stimulus was repeated four times, for a total of 20 trials.

Stimulus	***"Ls"***	***"L"***	***"C"***	***"R"***	***"Rs"***
1	Cue	Background	Background	Background	Background
2	Background	Cue	Background	Background	Background
3	Background	Background	Cue	Background	Background
4	Background	Background	Background	Cue	Background
5	Background	Background	Background	Background	Cue

Table 4. *Localization of the Sound Source Screening Test Stimuli.*

The participants were given a forced-choice task, in which they had to listen to the stimulus and indicate the origin of the cue tone. At the end of each playback, the question "Which speaker reproduced the *beep*?" was displayed on the screen and the participant had to answer, within 5 seconds, by choosing from the options shown in Figure 6.

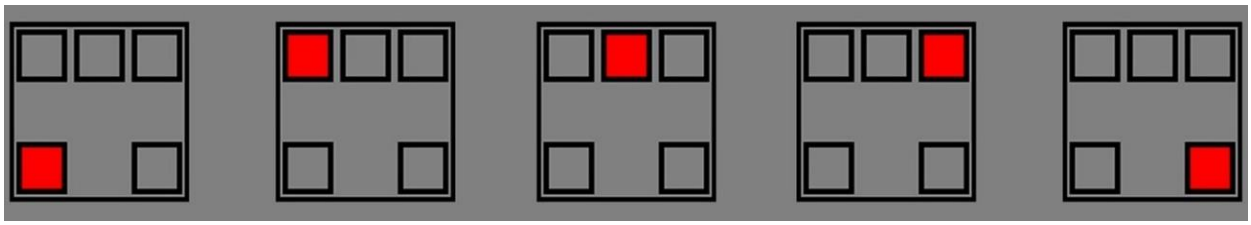

Figure 6. *Location Maps (from left to right) for "Ls", "L", "C", "R", "Rs" channels.*

The cut-off criterion for participants' performance was set at 80% correct responses (16 out of 20) based on the results of sound source location discrimination reported by Lopatka and colleagues (Lopatka et al., 2016).

2.4 Study 1: The Experience Of "Presence" Modulated by Audio Setting While Listening to Film Sound Sequences: A Behavioral Study

In the second stage of this research project, the difference in emotional and bodily involvement and in spatial audio perception among acoustic presentation modes were investigated using the stimuli validated in the first stage. Specifically, our interest was addressed to the perceived immersion, presence, emotional involvement, motor resonance, depth perception, and realism, that the surround presentation mode and a smaller-closer or larger-farther device could elicit.

2.4.1 Participants

The Generalized Listener Selection (GLS) procedure was implemented to select participants for the study, resulting in the selection of 32 participants out of the 41 screened. The sample consisted of 14 males and 18 females, with a mean age M of 28.7 years (standard deviation SD of ±6.3) within a range of 22 to 42 years. The selected sample was balanced in terms of gender

and had a high education level (*M* =15.5 years, *SD*= ± 2.3 years). Questionnaires descriptive statistics are shown in Table 5.

Questionnaire	Min.	Max.	M	SD
BMEQ C	13	28	19.3	4.8
BMEQ I	7	33	20.0	7.1
BMEQ S	8	18	12.2	2.8
BMEQ A	36	50	43.3	3.5
BMEQ P	34	71	58.5	11.1
BMEQ R	19	45	35.2	6.3
IRI CE	3	13	7.4	3.6
IRI DP	11	26	17.7	4.8
IRI PT	0	25	8.6	5.6
IRI FS	0	18	8.1	4.5
ITQ	46	94	73.8	13.4
VMIQ-2 EVI	12	48	26.8	10.3
VMIQ-2 IVI	12	55	27.8	13.2
VMIQ-2 KIN	12	55	26.8	13.0
F-IEQ CAP	44	77	62.8	10.6
F-IEQ DIS	10	20	15.4	2.6
F-IEQ COM	9	27	15.5	4.5
F-IEQ TRA	16	27	23.3	3.3

Table 5. *Questionnaire Descriptive Statistics (Min.= minimum score, Max.=maximum score, M=mean, SD=tandard deviations*).

B-MEQ: Brief Music Experience Questionnaire (C: Commitment to music, I: Innovative Musical Aptitude, S: Social uplift, A: Affective reactions, P: Positive psychotropic effects, R: Reactive musical behavior); IRI: Interpersonal Reactivity Index (PT: Perspective Taking, FS; Fantasy,

CE: Empathic Concern, DP: Personal Distress); ITQ: Immersive Tendencies Questionnaire; VMIQ-2: Vividness of Movement Imagery Questionnaire-2 (EVI: External self Visual Imagery, IVI: Internal first-person Visual Imagery, KIN: Kinesthetic Imagery); F-IEQ: Film Immersive Experience Questionnaire (CAP: Captivation, DIS: Real-world dissociation, COM: Comprehension, TRA: Transportation).

2.4.2 Experimental Stimuli

The 27 selected and validated stimuli (see Paragraph 2.1) were used in Surround sound, Stereophonic sound, and Monophonic sound conditions (three levels of the experimental condition) for a total of 81 experimental stimuli.

2.4.3 Experimental Paradigm

A total of 81 auditory stimuli were presented to the participants in a random sequence for two repetitions, resulting in a total of 162 trials. These trials were divided into three blocks, with each block consisting of 54 stimuli and lasting approximately 15 minutes.

The experimental trial was composed as follows: a fixation cross was presented on the screen for 1.5 seconds, followed by the playback of a 10-second experimental stimulus, followed by two random questions that the participant had to answer as quickly and accurately as possible on a Visual Analog Scale (VAS) within a predetermined time limit of 5 seconds, and followed by a 3.5-second inter-trial interval (ITI) period.

The primary task of the participants during the experiment was to simply listen to the auditory stimuli and answer the two questions as quickly and accurately as possible. The questions were designed to assess different aspects of cinematic immersion and sense of presence such as enjoyment, emotional involvement, physical immersion, and realism:

- "How much did you like the scene?" (VAS 0-100) measuring Enjoyment (EN);
- "How much did you feel emotionally involved?" (VAS 0-100) measuring Emotional Involvement (EI);
- "How much did you feel physically immersed?" (VAS 0-100) measuring Physical Immersion (PI);
- "How much realistic did you judge the scene?" (VAS 0-100) measuring Realism (RE).

In order to familiarize with the experiment, a training test was performed before the experimental phase. Participants were also instructed to interpret the questions correctly, particularly regarding realism, defined as the level of fidelity of reproduction of a sound scene, and physical immersion, defined as the impression of being physically present within the reproduced sound scene.

Stimuli were presented with MATLAB extension Psychtoolbox-3 (Brainard, 1997).

2.5 Study 2: The Perception of Acoustic Spatialization While Listening to Film Sequences: An EEG Study

In the third stage of this research project, we investigated the time course and neural correlates of the audio presentation. We designed a behavioral experiment and a high-density electroencephalographic (HD-EEG) experiment using the stimuli validated in the first stage. We hypothesized that the enhanced spatialization of sound in the surround presentation mode would lead to greater activation of embodied mechanisms as compared to the monophonic presentation mode.

2.5.1 Participants

The Generalized Listener Selection (GLS) procedure was implemented to select participants for the study, resulting in the selection of 24 right-handed participants out of the 32 screened. The sample consisted of 11 males and 13 females, with a mean age M of 24.3 years (standard deviation SD of ±2.4) within a range of 21 to 30 years. The selected sample was balanced in terms of gender and had a high education level (*M* =15.2 years, *SD*= ± 1.5 years). Questionnaires descriptive statistics are shown in Table 6.

Questionnaire	**Min.**	**Max.**	**M**	**SD**
BMEQ C	17	23	18.3	2.3
BMEQ I	15	32	21	6.4
BMEQ S	9	15	13.3	2.5
BMEQ A	34	49	41.4	3.5
BMEQ P	33	65	52.4	10,1
BMEQ R	17	43	33.2	5.3
IRI CE	5	14	5.6	3.8
IRI DP	13	23	15.2	3.7
IRI PT	2	20	7.5	4.4
IRI FS	3	17	6.9	4.1
ITQ	50	91	69.5	11.3
VMIQ-2 EVI	14	51	24.3	9.9
VMIQ-2 IVI	13	55	25.1	11.3
VMIQ-2 KIN	15	54	26.4	12.1
F-IEQ CAP	45	70	63.8	9.8
F-IEQ DIS	11	21	12.1	2.1
F-IEQ COM	13	22	13.4	4.1

F-IEQ TRA	11	24	22.3	2.3

Table 6. *Questionnaire Descriptive Statistics (Min.= minimum score, Max.=maximum score, M=mean, SD=tandard deviations*).

B-MEQ: Brief Music Experience Questionnaire (C: Commitment to music, I: Innovative Musical Aptitude, S: Social uplift, A: Affective reactions, P: Positive psychotropic effects, R: Reactive musical behavior); IRI: Interpersonal Reactivity Index (PT: Perspective Taking, FS; Fantasy, CE: Empathic Concern, DP: Personal Distress); ITQ: Immersive Tendencies Questionnaire; VMIQ-2: Vividness of Movement Imagery Questionnaire-2 (EVI: External self Visual Imagery, IVI: Internal first-person Visual Imagery, KIN: Kinesthetic Imagery); F-IEQ: Film Immersive Experience Questionnaire (CAP: Captivation, DIS: Real-world dissociation, COM: Comprehension, TRA: Transportation).

2.5.2 Experimental Stimuli

The 27 selected and validated stimuli (see Paragraph 2.1) were used in Surround sound and Monophonic sound conditions. In order to control for potential biases in the study, a set of control stimuli were also created by manipulating the original audio stimuli. Specifically, 27 control stimuli were generated, in both the Surround sound and Monophonic sound, respectively, by scrambling the original tracks in a way that retained all the acoustic characteristics on the frequency level, but in a random temporal sequence that made them unintelligible to the participant. We excluded the Stereophonic condition to reduce the length of the experimental session and to increase the perceptual difference between the Experimental conditions. Therefore, the experimental condition consisted of four levels: Surround, Monophonic, Monophonic Control, and Surround Control for a total of 108 experimental stimuli.

2.5.3 Setup EEG

EEG data were acquired by a Geodesic Sensor System which includes the Net Amps 300 amplifier and a 128-channel HydroCel Geodesic Sensor Net (Figure 7) and recorded using Net Station 5.4 EGI software (Electrical Geodesic Inc., Eugene, OR).

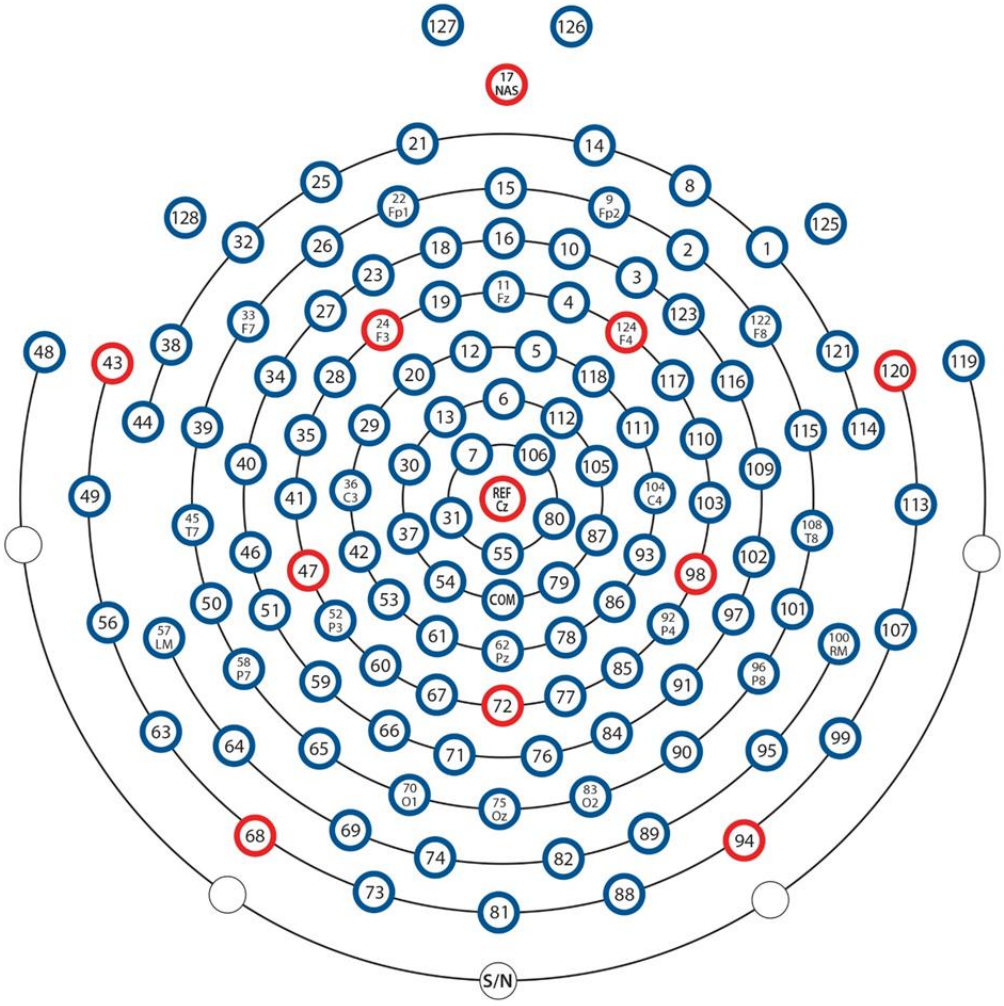

Figure 7. *EGI 128-channel Sensor Net Layout.*

The EEG net size used on each participant was dependent on the actual head circumference wrapped at Glabella (brow ridge) and Occipital Protuberance (OP), the most prominent bump on the back of the skull. The EEG net was immersed in an electrolyte solution for 5 minutes to

decrease electrode impedances were kept below 50 kΩ. Raw EEG was sampled at 500 Hz and was recorded with the vertex (Cz) as the online reference. The vertex was at the intersection of the Preauricular Midpoint and the Nasion-Inion midpoint.

Also an Electromyography (EMG) signal was acquired with an AD Instruments PowerLab 35 (ADInstruments, U.K.), and LabChart 8 Pro software was used for recording. EMG activity was bipolarly recorded on the left Extensor Digitorum Communis and left Tibialis Anterior with 4 mm standard Ag/Ag-Cl electrodes. Before being attached over the muscle regions the participants' skin was cleaned with an alcohol solution and the electrodes were filled with gel electrode paste (Fridlund & Cacioppo, 1986). EMG was sampled at 2 kHz and recorded with an online Mains Filter (adaptive 50 Hz filter).

2.5.4 Experimental Paradigm

A total of 108 auditory stimuli were presented to the participants in a random sequence for two repetitions, resulting in a total of 216 trials. These trials were divided into four blocks, with each block consisting of 54 stimuli and lasting approximately 15 minutes.

The experimental trial was composed as follows: a fixation cross was presented on the screen for 1.5 seconds, followed by the playback of a 10-second experimental stimulus, followed by the questions that the participant had to answer as quickly and accurately as possible on a Visual Analog Scale (VAS) within a predetermined time limit of 5 seconds, and followed by a 3.5-second inter-trial interval (ITI) period.

The primary task of the participants during the experiment was to simply listen to the auditory stimuli and answer to one question about Physical Immersion (PI) as quickly and accurately as

possible ("How much did you feel physically immersed?"). In order to familiarize with the experiment, a training test was performed before the experimental phase. Participants were also instructed to maintain a fixed position and to not blink during the EEG recording.

Stimuli were presented with MATLAB extension Psychtoolbox-3 (Brainard, 1997).

CHAPTER 3 - ANALYSIS AND RESULTS

3.1 Study 1: The Experience Of "Presence" Modulated by Audio Setting While Listening to Film Sound Sequences: A Behavioral Study

3.1.1 Behavioral Analysis and Results

In this experiment, participants were asked to assign a score between 0 and 100 to the level of Enjoyment (EN), Emotional Involvement (EI), Physical Immersion (PI), and Realism (RE) they perceived after each auditory stimulus. To test for significant differences between the scores given by participants in the Presentation modes condition (Monophonic Condition, Stereophonic Condition, Surround Condition), a linear mixed model was employed.

A hierarchical approach was followed in the model selection process. A null model was first created, and other parameters were added incrementally to assess the impact of each on the fit of the model. Model selection criteria such as the likelihood ratio test, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) were used to rigorously choose which parameters improved the fit of the model to the data.

The final model included the participants' responses as the dependent variable, with the Presentation modes condition (Monophonic Condition, Stereophonic Condition, Surround Condition) and Question (EN, EI, PI, and RE) as the fixed independent variables. The participants were included as a random intercept and the experimental condition as a random slope. This approach accounted for the within-subject and between-subject variability in the data. Outliers were identified and excluded from the analysis based on the standardized model residuals and a threshold value of Cook's distance (threshold=1).

The means, standard error, and limits of the confidence intervals relative to Questions scores in the different levels of the experimental condition are shown in Table 7.

Condition	Question	Mean	SE	Lower IC	Upper IC
Mono	Emotional Involvement	38.25	2.82	32.57	43.94
Mono	Enjoyment	40.47	3.41	33.58	47.36
Mono	Physical Immersion	38.97	2.48	33.94	44.00
Mono	Realism	46.92	2.24	42.36	51.48
Stereo	Emotional Involvement	46.66	2.32	41.93	51.39
Stereo	Enjoyment	47.12	3.02	40.96	53.28
Stereo	Physical Immersion	53.89	1.90	50.02	57.76
Stereo	Realism	57.91	1.57	54.71	61.11
Surround	Emotional Involvement	54.09	2.53	48.96	59.22
Surround	Enjoyment	54.99	3.18	48.53	61.44
Surround	Physical Immersion	63.78	2.15	59.41	68.15
Surround	Realism	64.90	1.87	61.10	68.70

Table 7. *Questions Scores Descriptive Statistics.*

The linear mixed model hierarchically selected demonstrated an acceptable fit to the data with a marginal R^2_m value of 0.22 and a complex R^2_c value of 0.85. These goodness-of-fit parameters indicate that the variance of the dependent variables explained by the model is 22% when considering only the fixed effect, and 85% when taking into account the effect of the random variables (intercept and slope). The model revealed a significant main effect of Presentation modes ($\chi^2_{(2)} = 65.16,\ p < .001$), Question ($\chi^2_{(3)} = 71.57,\ p < .001$) and a significant interaction Presentation modes*Condition ($\chi^2_{(6)} = 269.36, p < .001$).

To investigate the specific differences between the levels of the experimental condition, post-hoc tests were conducted, using Tukey's correction for multiple comparisons and adjusting for the Kenward-Roger degrees of freedom.

Mean differences, standard deviations, confidence intervals, T-test values, and associated p-values for the pairwise comparisons between the levels of the experimental condition are reported in Table 8, Table 9, Table 10, Table 11.

Contrasts	Difference	SE	D.o.F.	Lower IC	Upper IC	T test	p
Monophonic - Stereophonic	-10.24	1.64	30.99	-14.27	-6.22	-6.26	<.001
Monophonic - Surround	-18.29	2.36	31.00	-24.08	-12.49	-7.76	<.001
Stereophonic - Surround	-8.04	1.04	30.94	-10.59	-5.49	-7.76	<.001

Table 8. *Main effect of Presentation modes Post Hoc Pairwise comparisons.*

Presentation modes post hoc comparisons (Table 8) showed that participants attributed significantly higher absolute scores when stimuli were presented in the Surround condition than when they were presented in the Stereophonic condition or in the Monophonic conditions. At the same time, participants attributed significantly higher scores when stimuli were presented in the Stereophonic condition than when they were presented in the Monophonic conditions.

Contrasts	Difference	SE	D.o.F.	Lower IC	Upper IC	T test	p
EI - EN	-1.19	1.30	30.95	-4.72	2.33	-0.92	n.s.
EI - PI	-5.88	1.06	30.93	-8.75	-3.00	-5.55	<.001
EI - RE	-10.24	1.80	30.99	-15.12	-5.36	-5.70	<.001
EN - PI	-4.69	2.15	31.00	-10.51	1.14	-2.18	n.s.
EN - RE	-9.05	2.49	31.00	-15.81	-2.29	-3.63	<.01
PI - RE	-4.36	1.35	30.98	-8.02	-0.70	-3.23	<.01

Table 9. *Main effect of Question Post Hoc Pairwise comparisons*

Question post hoc comparisons (Table 9) showed that participants attributed higher scores on Realism (RE) than on Enjoyment (EN), Emotional Involvement (EI), and Physical Immersion

(PI). In addition, participants attributed higher scores to Physical Immersion (PI) than to Emotional Involvement (EI).

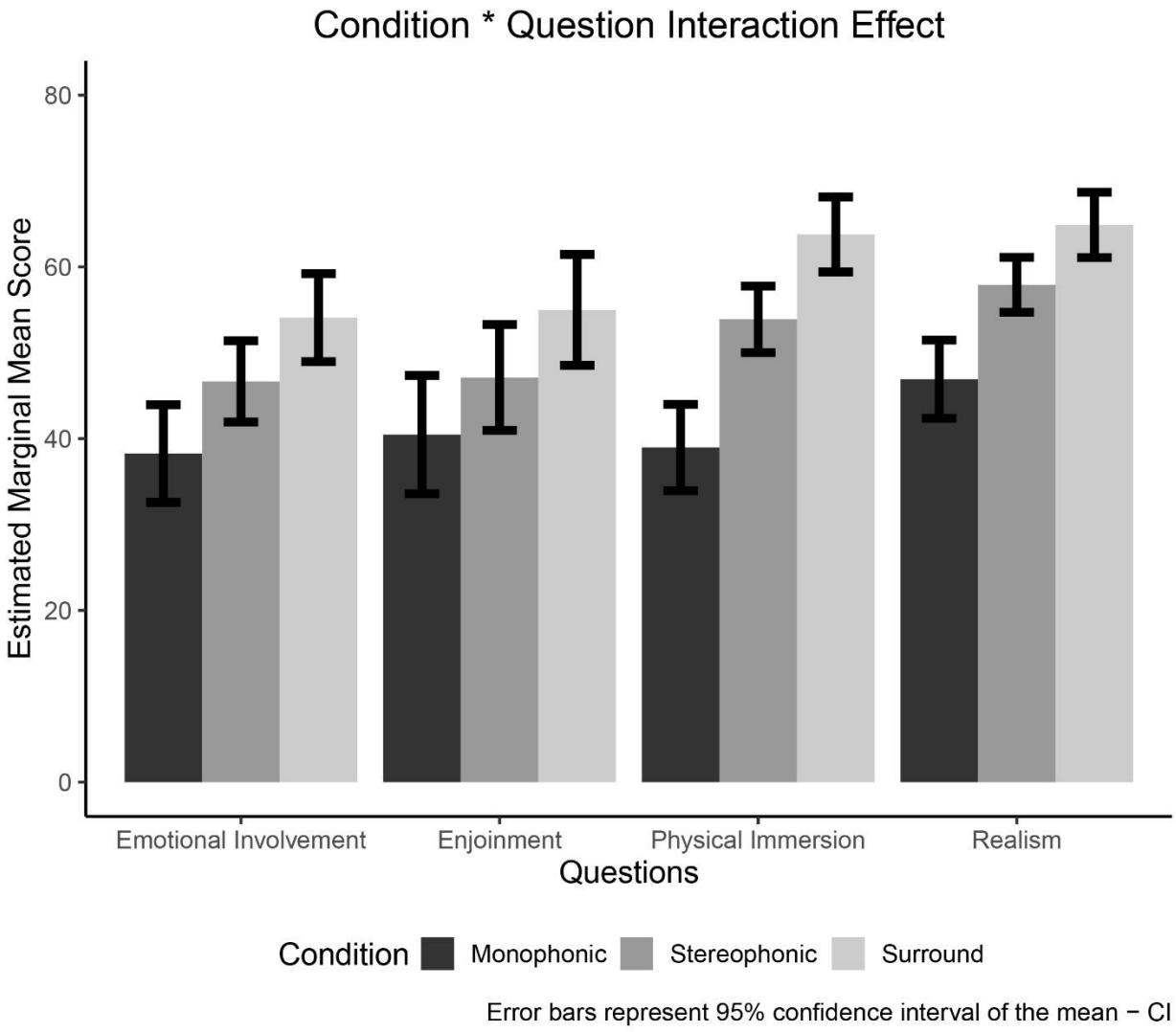

Figure 8. *Interaction effect Post Hoc Pairwise comparisons by Question.*

Interaction post hoc comparisons showed that independently from the Question (Table 10) participants attributed always significantly higher absolute scores when stimuli were presented in the Surround condition than when they were presented in the Stereophonic condition or in the Monophonic conditions (Figure 8). Also independently from the Question, participants attributed

significantly higher scores when stimuli were presented in the Stereophonic condition than when they were presented in the Monophonic conditions.

Contrasts	Question	Difference	SE	D.o.F.	Lower IC	Upper IC	T test	p
Monophonic - Stereophonic	EI	-8.41	1.69	35.23	-12.54	-4.27	-4.97	<.001
Monophonic - Surround	EI	-15.84	2.40	33.18	-21.71	-9.96	-6.61	<.001
Stereophonic - Surround	EI	-7.43	1.12	42.10	-10.15	-4.71	-6.64	<.001
Monophonic - Stereophonic	EN	-6.65	1.69	35.01	-10.78	-2.52	-3.94	<.001
Monophonic - Surround	EN	-14.52	2.39	32.86	-20.38	-8.65	-6.07	<.001
Stereophonic - Surround	EN	-7.86	1.12	41.87	-10.58	-5.15	-7.03	<.001
Monophonic - Stereophonic	PI	-14.92	1.70	35.78	-19.07	-10.78	-8.80	<.001
Monophonic - Surround	PI	-24.81	2.40	33.46	-30.70	-18.92	-10.33	<.001
Stereophonic - Surround	PI	-9.89	1.12	42.73	-12.61	-7.16	-8.81	<.001
Monophonic - Stereophonic	RE	-10.99	1.69	35.63	-15.13	-6.85	-6.49	<.001
Monophonic - Surround	RE	-17.98	2.40	33.30	-23.86	-12.10	-7.50	<.001
Stereophonic - Surround	RE	-6.99	1.12	42.55	-9.71	-4.26	-6.23	<.001

Table 10. *Interaction effect Post Hoc Pairwise comparisons by Question.*

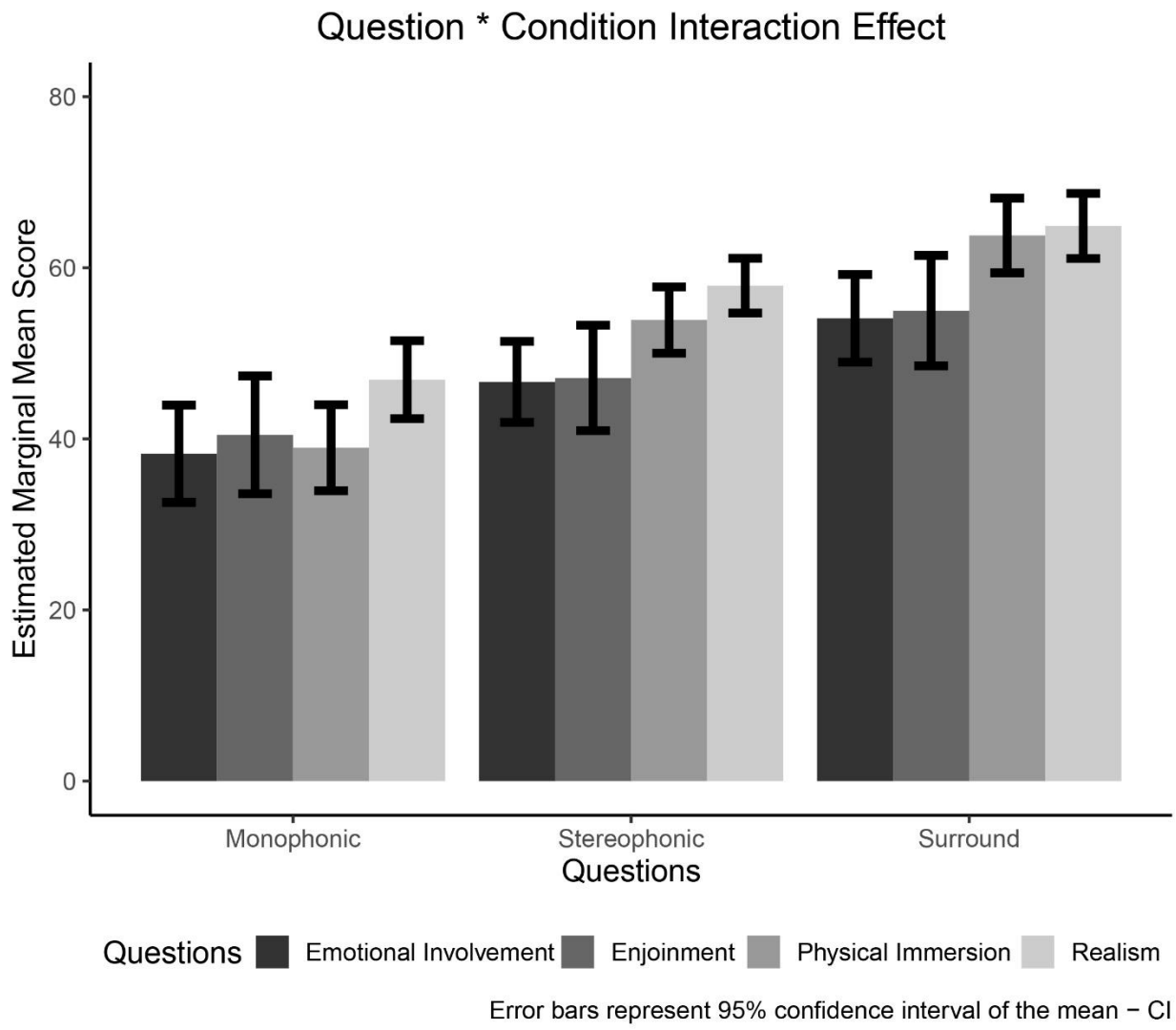

Figure 9. *Interaction effect Post Hoc Pairwise comparisons by Presentation Modes.*

If we consider only the Monophonic Presentation mode (Table 11; Figure 9), participants attributed significantly higher scores on Realism (RE) than on Emotional Involvement (EI) and Physical Immersion (PI).

If we consider only the Stereophonic Presentation mode (Table 11; Figure 9), participants attributed significantly higher scores on Realism (RE) than on Emotional Involvement (EI) and Physical Immersion (PI). In addition, participants attributed significantly higher scores to Physical Immersion (PI) than to Emotional Involvement (EI).

If we consider only the Surround Presentation mode (Table 11; Figure 9), participants attributed significantly higher scores on Realism (RE) than on Emotional Involvement (EI) and Physical Immersion (PI). In addition, participants attributed significantly higher scores to Physical Immersion (PI) than to Emotional Involvement (EI) and Enjoyment (EN).

Contrasts	**Condition**	**Difference**	**SE**	**D.o.F.**	**Lower IC**	**Upper IC**	**T test**	**p**
EI - EN	Monophonic	-2.22	1.36	37.07	-5.87	1.44	-1.63	n.s.
EI - PI	Monophonic	-0.71	1.14	41.94	-3.77	2.35	-0.62	n.s.
EI - RE	Monophonic	-8.66	1.85	34.57	-13.65	-3.68	-4.69	<.001
EN - PI	Monophonic	1.50	2.19	33.39	-4.41	7.41	0.69	n.s.
EN - RE	Monophonic	-6.45	2.53	32.72	-13.28	0.38	-2.55	n.s.
PI - RE	Monophonic	-7.95	1.42	38.09	-11.77	-4.13	-5.59	<.001
EI - EN	Stereophonic	-0.46	1.36	36.74	-4.11	3.18	-0.34	n.s.
EI - PI	Stereophonic	-7.23	1.12	39.37	-10.25	-4.22	-6.43	<.001
EI - RE	Stereophonic	-11.25	1.84	33.84	-16.21	-6.29	-6.13	<.001
EN - PI	Stereophonic	-6.77	2.18	33.10	-12.67	-0.87	-3.10	n.s.
EN - RE	Stereophonic	-10.79	2.52	32.58	-17.61	-3.96	-4.28	<.001
PI - RE	Stereophonic	-4.02	1.40	36.19	-7.79	-0.24	-2.86	n.s.
EI - EN	Surround	-0.89	1.36	37.26	-4.55	2.76	-0.66	n.s.
EI - PI	Surround	-9.69	1.14	41.39	-12.74	-6.64	-8.50	<.001
EI - RE	Surround	-10.81	1.84	34.40	-15.78	-5.83	-5.86	<.001
EN - PI	Surround	-8.79	2.18	33.30	-14.70	-2.89	-4.02	<.001
EN - RE	Surround	-9.91	2.52	32.66	-16.74	-3.08	-3.93	<.001
PI - RE	Surround	-1.12	1.41	37.25	-4.92	2.68	-0.79	n.s.

Table 11. *Interaction effect Post Hoc Pairwise comparisons by Presentation Modes.*

3.2 Study 2: The Perception of Acoustic Spatialization While Listening to Film Sequences: An EEG Study

3.2.1 Behavioral Analysis and Results

In this experiment, participants were asked to assign a score between 0 and 100 to the level of Physical Immersion (PI) they perceived after each auditory stimulus. To test for significant differences between the scores given by participants in the Presentation modes condition (Monophonic Condition, Surround, Monophonic Control, Surround Control), a linear mixed model was employed.

A hierarchical approach was followed in the model selection process. A null model was first created, and other parameters were added incrementally to assess the impact of each on the fit of the model. Model selection criteria such as the likelihood ratio test, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) were used to rigorously choose which parameters improved the fit of the model to the data.

The final model included the participants' responses as the dependent variable, with the Presentation modes condition (Monophonic, Surround, Monophonic Control, Surround Control) as the fixed independent variable. The participants were included as a random intercept and the experimental condition as a random slope. This approach accounted for the within-subject and between-subject variability in the data. Outliers were identified and excluded from the analysis based on the standardized model residuals and a threshold value of Cook's distance (threshold=1).

The means, standard error, and limits of the confidence intervals relative to Question scores in the different levels of the experimental condition are shown in Table 12.

Condition	Mean	SE	Lower IC	Upper IC
Monophonic	69.95	3.40	62.66	77.25
Monophonic Control	17.72	2.61	12.11	23.33
Surround	76.38	3.36	69.17	83.6
Surround control	30.67	2.92	24.41	36.93

Table 12. *FI Scores Descriptive Statistics.*

The linear mixed model hierarchically selected demonstrated an acceptable fit to the data with a marginal R^2_m value of 0.82 and a complex R^2_c value of 0.98. These goodness-of-fit parameters indicate that the variance of the dependent variable (Physical Immersion) explained by the model is 82% when considering only the fixed effect (Presentation modes), and 98% when taking into account the effect of the random variables (intercept and slope). The model revealed a significant main effect of the experimental condition ($F_{(3, 13.88)}$=61.36, p <.001).

To investigate the specific differences between the levels of the experimental condition, post-hoc tests were conducted, using Tukey's correction for multiple comparisons and adjusting for the Kenward-Roger degrees of freedom. Mean differences, standard deviations, confidence intervals, T-test values, and associated p-values for the pairwise comparisons between the levels of the experimental condition are reported in Table 13.

Contrasts	Difference	SE	D.o.F.	Lower IC	Upper IC	T test	p
Monophonic - Monophonic Control	52.13	4.12	13.99	40.13	64.13	12.62	<.001
Monophonic-Surround	-6.78	1.0	13.98	-9.70	-3.85	-6.73	<.001
Monophonic - Surround Control	40.03	4.46	13.99	27.06	53.0	8.97	<.001
Monophonic Control - Surround	-58.91	4.39	13.99	-71.68	-46.13	-13.40	<.001

Monophonic Control -Surround Control	-12.1	2.32	13.99	-18.84	-5.35	-5.21	<.001
Surround -Surround Control	46.81	4.41	13.99	33.97	59.64	10.59	<.001

Table 13. *Post Hoc Pairwise comparisons.*

These post hoc comparisons showed that participants attributed significantly higher absolute scores when stimuli were presented in the Surround condition than when they were presented in the Monophonic condition or in the respective Control conditions. In addition, participants attributed significantly higher scores when stimuli were presented in the Monophonic condition than when they were presented in the Control conditions (Monophonic Control - Surround Control). Interestingly, even if we consider only the Control conditions, participants attributed significantly higher scores when the control stimulus was presented in the Surround condition than when it was presented in the Monophonic condition.

Thus, it is clear from the results of the behavioral data analysis that the participants felt more physically immersed when the acoustic stimuli were presented in the Surround condition.

3.3.2 EEG Data Pre-processing

The EEG recordings were pre-processed using MATLAB toolbox EEGLAB ver. 2022.1 (Delorme & Makeig, 2004) in order to remove acquisition artifacts. These artifacts can be either related to participant or related to recording issues. Those related to the participant are physiological and are generally attributed to the participant's movement, the electrocardiogram (ECG) signal picked up by the electrodes, blinks and eye movements, sweat, and should be excluded. Artifacts related to recording technical issues, on the other hand, are due, for example, to interference from radio frequencies, transient fluctuations in electrode impedance, and cable

movement, and must be also removed. To do so we developed a custom script in MATLAB, using EEGLAB ver. 2022.0 toolbox, to execute a pre-processing pipeline:

a. **High-pass FIR filter and Line Noise Removal**

   We employed a high-pass filter to increase the signal-to-noise ratio and reduce data distortion and low-frequency noise. A high-pass filter removes low frequency noise and drift, which can obscure neural signals of interest. We chose to use a cutoff frequency of 0.5 Hz with a transition window of 0.25 Hz for the high-pass filter. The cutoff frequency is the point at which the filter begins to remove low frequency noise, and the transition window defines the range in which the filter gradually reduces the amplitude of frequencies near the cutoff frequency. In addition to the high-pass filter, we also employed a method called ZapLine to remove the power line noise at 50 Hz and its harmonics. The ZapLine method is a technique that uses a digital filter to remove this noise from the data, which is a common source of telecommunication frequencies and power grid related contamination in EEG data.

b. **Bad channels Interpolation detection: flatline, low-frequency, noise**

   We chose to interpolate channels that met one or more of the following criteria: having a variance greater than 25 times the mean variance of all channels, having interruptions in recording, or exceeding a threshold value of 100 μV of detected activity. By interpolating these channels, we were able to replace the missing or noisy data with estimates based on the activity of the surrounding electrodes. We used the spherical interpolation method for this step. The spherical interpolation method is based on the assumption that the brain generates electrical activity on a smooth and continuous surface, and it uses the activity of surrounding electrodes to estimate the activity at the interpolated electrode. By

applying this channel interpolation step, we were able to improve the quality of the data by replacing noisy or missing data with estimates based on the activity of surrounding electrodes.

c. Channels Montage Reduction

We discarded the outermost belt of electrodes of the sensor net, which included 13 peripheral channels (Ch48, Ch49, Ch56, Ch63, Ch68, Ch73, Ch81, Ch88, Ch94, Ch99, Ch107, Ch113, Ch119). These electrodes are often associated with residual muscle activity, which can introduce noise and artifacts into the EEG data that are not related to brain activity. Additionally, to remove activity related to blinks and eye movements, we also removed 11 frontal channels (Ch1, Ch8, Ch14, Ch17, Ch21, Ch25, Ch32, Ch125, Ch126, Ch127, Ch128) which are known to be highly susceptible to these types of artifacts. In total, this channel removal reduced the number of electrodes from 128 to 104 as in Pedroni and colleagues (Pedroni et al., 2019).

d. **Epochs**

Continuous EEG data were divided into 12-second epochs, which included 2 seconds of baseline (1.5 seconds of fixation cross period and 0.5 seconds of ITI), or pre-stimulus activity, and 10 seconds activity during the presentation of the stimulus in the various presentation modes.

e. EMG Bad Epochs Rejection

We identified and removed epochs with muscle activity using EMG data. The RMS amplitude of the EMG epoch was compared to a threshold value to determine if it contained muscle activity. The threshold value was set at 3 times the mean RMS amplitude of the baseline period. The epochs that exceeded this threshold were

considered muscle activity and were removed from the EEG analysis. Also, peak amplitude spikes of 0.5 volts or more were considered movement artifacts.

f. **Independent component analysis (ICA)**

We then applied Independent Component Analysis (ICA) to the remaining 104 channels of EEG data. ICA is a technique that allows for the decomposition of a multivariate signal into its independent components, each of which may represent a different source of activity or noise. ICA is a powerful tool for removing artifacts and noise from EEG data because it allows for the separation of independent sources of activity, rather than relying solely on spatial information. By identifying and isolating the independent components that are most likely to reflect noise or artifacts, we can improve the signal-to-noise ratio and increase the accuracy of our subsequent analysis.

g. **ICs Artifactual Components Rejection**

Once the Independent Components (ICs) were estimated, we applied an automated recognition algorithm, MARA, developed by Winkler and colleagues to identify and exclude components that represent noise or artifacts (Winkler et al., 2011, 2014). MARA stands for Multiple Artifact Rejection Algorithm, and it is an automated algorithm that uses the Minimum Description Length (MDL) principle to identify and remove artifactual independent components (ICs) while preserving the neural signal of interest. The algorithm uses predefined features, such as spatial distribution, power spectrum, and temporal dynamics of the ICs, to calculate the MDL values for each IC and assigns a probability of being artifactual to each IC. Artifactual ICs were removed from EEG data.

h. **Common average Re-referencing**

EEG data, recorded with the vertex (Cz) as the online reference, were re-referenced offline to the common average. The goal is to eliminate potential bias and increase the signal-to-noise ratio by removing the average potential across all electrodes. This is achieved by subtracting the average of the signal at each electrode from the EEG signal at that electrode for each time point. The idea behind this method is that the head can be approximated to a sphere and it is assumed that the sum of all potentials recorded on the sphere, due to current sources inside, is zero (by Ohm's law), resulting in a neutral reference.

i. **Baseline Normalization**

The mean activation values in the 2 seconds baseline were used to normalize the data for each epoch.

3.4.3 EEG Data Analysis and Results

The time-frequency analysis was performed using the Hanning taper method, which involves applying a Hanning window to the data prior to the computation of the fast Fourier transform. The window length was fixed at 0.5 seconds, with frequency intervals of 1 Hz, spanning from 3 to 32 Hz. This allowed for the examination of event-related spectral perturbation (ERSP) in Alpha (8 – 13 Hz, Figure 10) and Beta (14 – 32 Hz, Figure 11) frequency bands. Additionally, this analysis allowed for the investigation of dynamic changes in spectral power over time, revealing temporal patterns of neural activity that may be important for understanding the effects of the experimental conditions on the perception of audio spatialization during cinematic immersion. We averaged the Monophonic Control condition and the Surround Control condition and considered them as one Control condition.

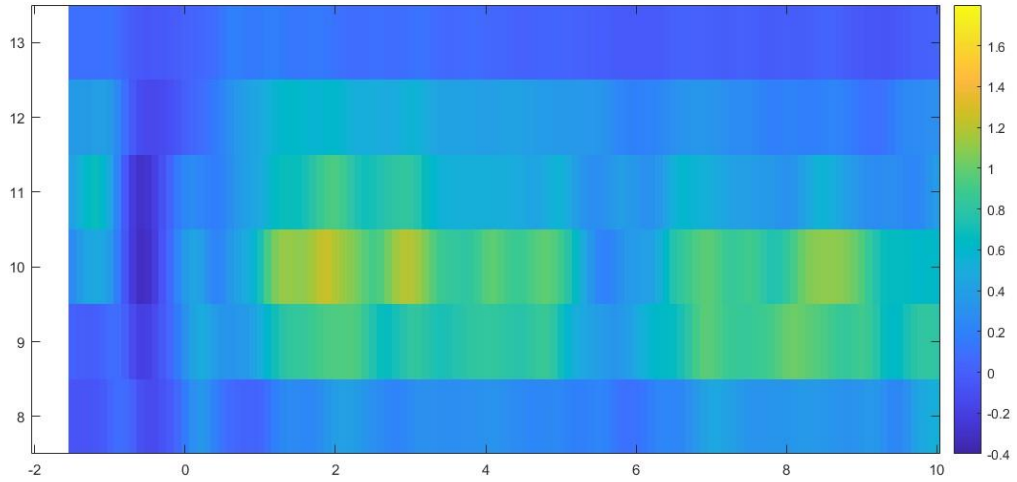

Figure 10. ERSP *Alpha Frequency Band Surround Presentation Mode.*

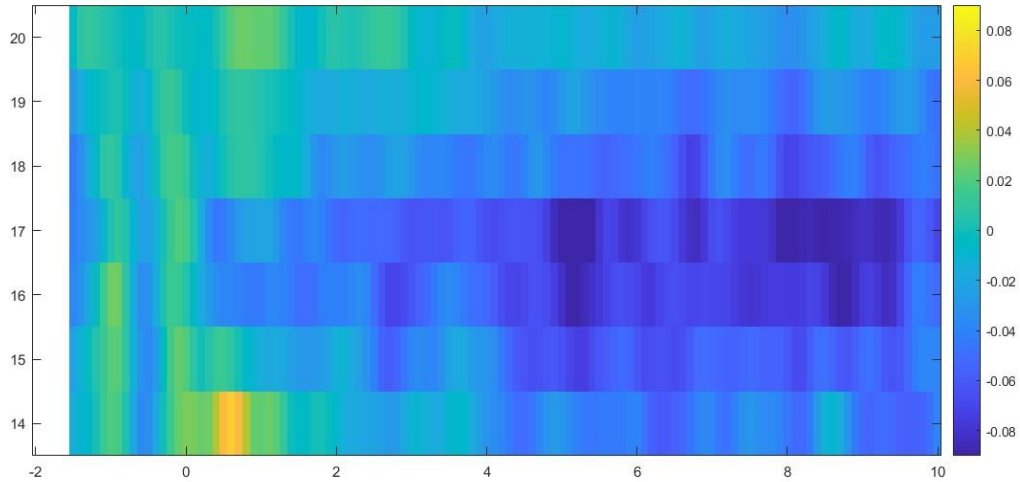

Figure 11. ERSP *Low Beta Frequency Band Surround Presentation Mode.*

In the statistical analysis of EEG data, it is necessary to address the multiple comparisons problem (MCP). This problem arises from the fact that EEG data has a spatiotemporal structure, meaning that the signal is recorded at multiple channels and multiple time points. The MCP arises due to the large number of statistical comparisons that need to be made, which is often in the thousands. This makes it difficult to control the Family-wise error rate (FWER), which is the probability of falsely concluding that there is a difference between experimental conditions at

one or more (channel, time) pairs. A solution for the MCP is to use non-parametric statistical testing and control the family-wise error rate (FWER) using a cluster-based test statistics for within-subjects experiments. The cluster-based test statistics is calculated by comparing experimental conditions at the sample level, selecting samples with t-values above a certain threshold, clustering them based on temporal, spatial, and spectral adjacency, and taking the sum of t-values within each cluster. The significance probability is then calculated using the Monte Carlo permutation method with 500 random draws. A p-value is calculated by comparing the observed test statistic to the distribution of test statistics obtained through random partitions of the data. A cluster is considered significant if its p-value is less than the critical alpha level of 0.05.

This data-driven approach allows us to identify specific time windows and electrodes clusters where there is a significant difference in neural activity between experimental conditions without any spatial cluster and frequency band assumption and highlight regions of interest for further analysis. Once specific electrodes clusters and time windows were identified, we extracted the power of the event related spectral perturbation (ERSP) and performed parametric statistics.

Two significant clusters were identified, one in the Alpha frequency band and the other in the Low Beta frequency band. The first significant cluster was identified in the time window from 3 to 7 seconds in the Alpha frequency band (8 to 10 Hz). This suggests that there is a significant difference in neural activity in this frequency band and time window between the experimental conditions. Specifically, this cluster is characterized by an event-related desynchronization (ERD) during the Surround condition compared to both the Monophonic (Figure 12) and Control conditions (Figure 13), with a peak difference around 5 seconds from the stimulus onset.

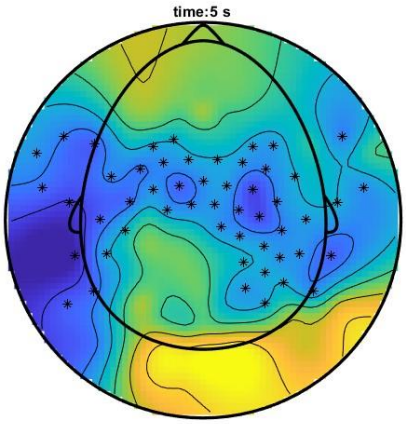

Figure 12. *Surround - Monophonic Alpha band cluster.*

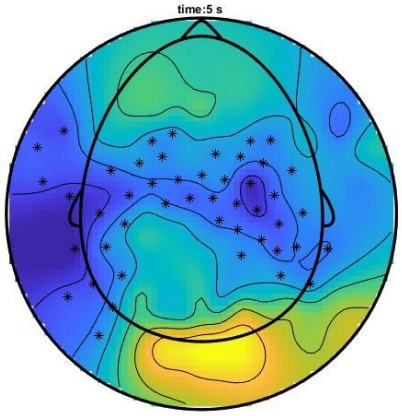

Figure 13. *Surround - Control Alpha band cluster.*

The second significant cluster was identified in the time window from 2 to 7 seconds in the Low Beta frequency band (16 to 18 Hz). This suggests that there is a significant difference in neural activity in this frequency band and time window between the experimental conditions. Similarly to the first cluster, this cluster also is characterized by an ERD during the Surround condition compared to both the Monophonic (Figure 14) and Control (Figure 15) conditions, with a peak difference around 4.5 seconds from the stimulus onset.

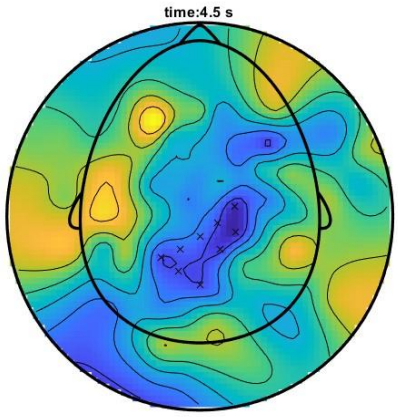

Figure 14. *Surround - Monophonic Low Beta band cluster.*

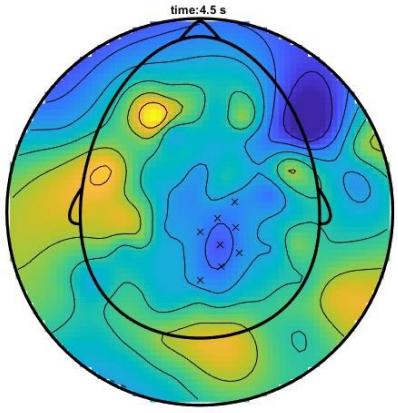

Figure 15. *Surround - Control Low Beta band cluster.*

Since the Rolandic alpha frequency band of interest (8–13 Hz) overlaps with the posterior alpha band, recordings in central areas might be affected by this posterior activity. However, given that significant clusters were detected only in central and parietal areas, we can exclude that our results were related to attentional/vigilance factors originating from parieto-occipital cortex.

From significant electrodes in the two clusters of interest, we then extracted the log-ratio frequency power in the significant time window/frequencies and analyzed them in a linear mixed

model. The log-ratio is used to represent the change in power in a more interpretable way, as it is a unitless value that allows for direct comparisons between different frequency bands.

The Alpha cluster model included the log-ratio frequency power as the dependent variable, with the Presentation modes condition (Monophonic, Surround, Control) as the fixed independent variable. The participants were included as a random intercept. Outliers were identified and excluded from the analysis based on the standardized model residuals and a threshold value of Cook's distance (threshold=1).

The model explained 57% of the variance in power, taking into account the random effects ($R2m = 0.18;\ R2c = 0.57$). The model revealed a significant main effect of the experimental condition ($\chi^2_{(2)} = 74.05$, $p < .001$).

To investigate the specific differences between the levels of the experimental condition, post-hoc tests were conducted, using Tukey's correction for multiple comparisons and adjusting for the Kenward-Roger degrees of freedom. Mean differences, standard deviations, confidence intervals, T-test values, and associated p-values for the pairwise comparisons between the levels of the experimental condition are reported in Table 14.

Contrasts	Difference	SE	D.o.F.	Lower IC	Upper IC	T test	p
Control - Monophonic	0.09	0.02	Inf	0.03	0.14	3.77	<.001
Control - Surround	0.20	0.02	Inf	0.15	0.25	8.55	<.001
Monophonic - Surround	0.11	0.03	Inf	0.05	0.18	4.14	<.001
Control - Monophonic	0.09	0.02	Inf	0.03	0.14	3.77	<.001

Table 14. *Alpha Cluster Post Hoc Pairwise comparisons*

The post-hoc comparisons performed on the Alpha cluster model revealed that there was a significant event-related desynchronization (ERD) in the Surround condition when compared to both the Monophonic and Control conditions. This means that there was an increase in neural activity in the centro-parietal areas during the Surround condition when compared to the Monophonic and Control conditions, with a peak difference around 5 seconds after stimulus onset. Furthermore, the post-hoc comparisons also showed that there was a significant ERD in the Monophonic condition when compared to the Control condition. This suggests that there is a distinct neural response in the centro-parietal areas to the different auditory stimuli presented in the Monophonic and Control conditions, and that this response is further modulated by the introduction of the surround sound in the Surround condition.

In Figure 16, we plotted the log-ratio of the frequency power extracted from the Surround condition and the Monophonic condition relative to the Control condition in the Alpha cluster. The plot provides a visual representation of the changes in frequency power over time, and it can help visualize the time windows where there are significant differences in power between the conditions. It seems that we detect a partially different EDR pattern between the Surround and Monophonic Condition relative to control. Even if the significant difference is detected 3 seconds after stimulus onset, the ERD in Surround relative to control starts after stimulus onset followed by a power rebound 6 seconds after stimulus onset, while in Monophonic relative to control we distinguish an ERS in the first 2 seconds after stimulus onset and an ERD 2 seconds after stimulus onset followed by a power rebound 6 seconds after stimulus onset.

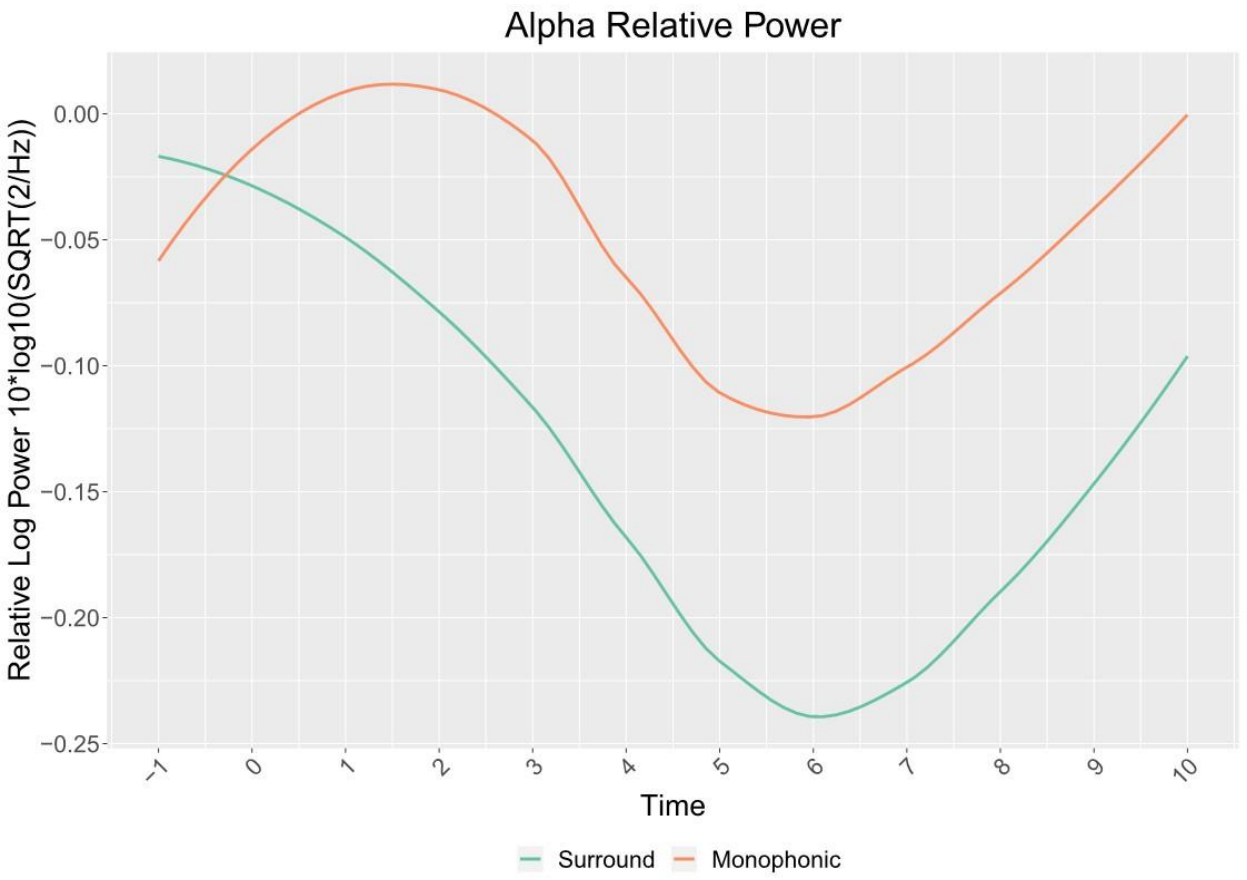

Figure 16. *Alpha Power in Surround and Monophonic Condition Relative to Control.*

The Low Beta cluster model included the log-ratio frequency power as the dependent variable, with the Presentation modes condition (Monophonic, Surround, Control) as the fixed independent variable. The participants were included as a random intercept. Outliers were identified and excluded from the analysis based on the standardized model residuals and a threshold value of Cook's distance (threshold=1). The model explained 89% of the variance in power, taking into account the random effects ($R2m = 0.2$; $R2c = 0.89$). The model revealed a significant main effect of the experimental condition ($\chi^2_{(2)} = 9.79$, $p < .001$).

To investigate the specific differences between the levels of the experimental condition, post-hoc tests were conducted, using Tukey's correction for multiple comparisons and adjusting for the Kenward-Roger degrees of freedom. Mean differences, standard deviations, confidence intervals,

T-test values, and associated p-values for the pairwise comparisons between the levels of the experimental condition are reported in Table 15.

Contrasts	Difference	SE	D.o.F.	Lower IC	Upper IC	T test	p
Control - Monophonic	0.00	0.00	Inf	-0.01	0.01	-0.04	n.s.
Control - Surround	0.01	0.00	Inf	0.00	0.02	2.94	<.01
Monophonic - Surround	0.01	0.00	Inf	0.00	0.02	2.58	<.05
Control - Monophonic	0.00	0.00	Inf	-0.01	0.01	-0.04	n.s.

Table 15. Low Beta *Post Hoc Pairwise comparisons.*

The post-hoc comparisons performed on the Low Beta cluster model revealed that there was a significant event-related desynchronization (ERD) in the Surround condition when compared to both the Monophonic and Control conditions. This means that there was an increase in neural activity in the centro-parietal areas during the Surround condition when compared to the Monophonic and Control conditions with a peak difference around 4.5 seconds after stimulus onset.

In Figure 17, we plotted the log-ratio of the frequency power extracted from the Surround condition and the Monophonic condition relative to the Control condition in the Low Beta cluster. The plot provides a visual representation of the changes in frequency power over time, and it can help visualize the time windows where there are significant differences in power between the conditions. It seems that we detect a partially different EDR pattern between the in Surround and Monophonic Condition relative to control. Even if the significant difference is detected 2 seconds after stimulus onset, the ERD in Surround relative to control starts after stimulus onset followed by a power rebound 6 seconds after stimulus onset, while in

Monophonic relative to control we distinguish a ERS in the first 2 seconds after stimulus onset and an ERD 2 seconds after stimulus onset followed by a power rebound 5 seconds after stimulus onset.

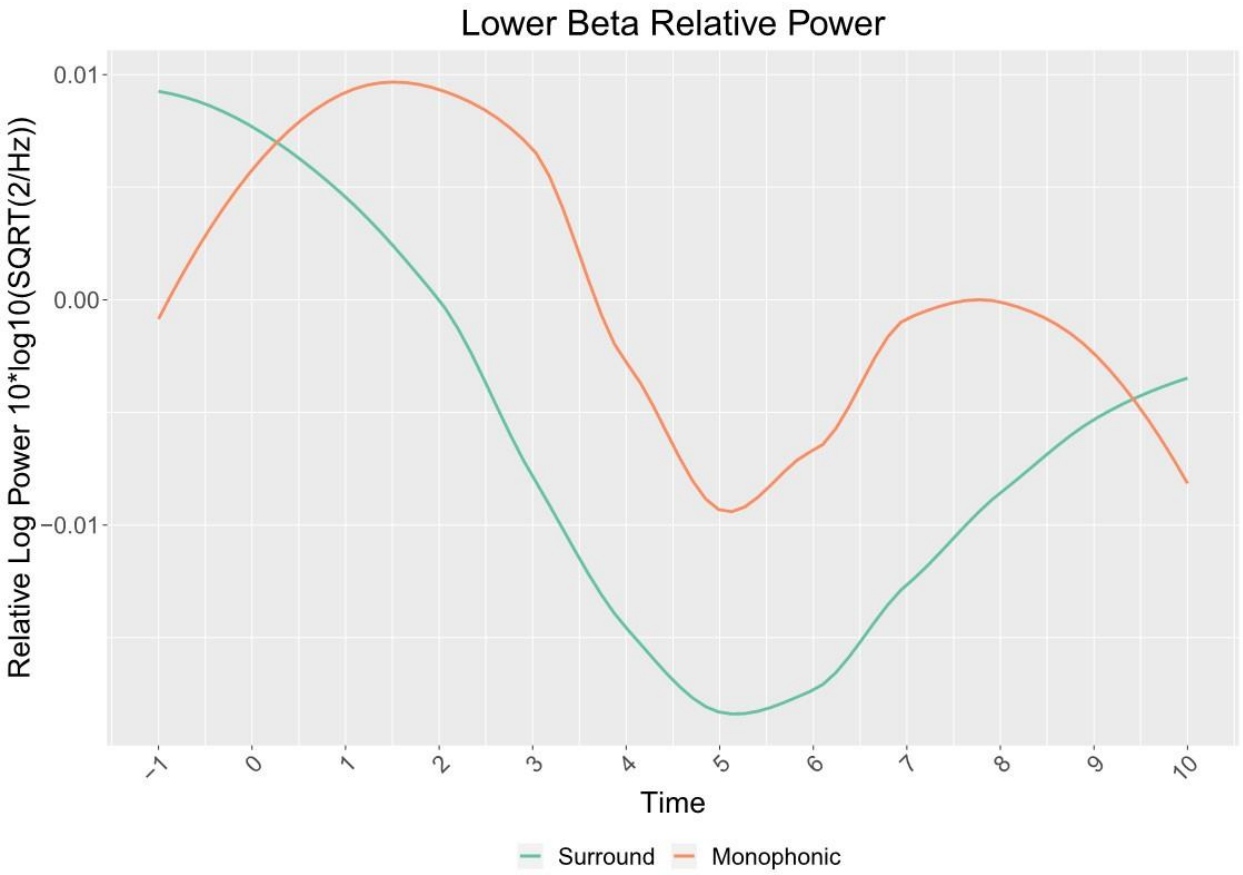

Figure 17. *Low Beta Power in Surround and Monophonic Condition Relative to Control.*

These results provide evidence for the role of auditory spatial information in modulating neural activity in the centro-parietal areas and can contribute to the understanding of the neural mechanisms underlying spatialized auditory perception.

CHAPTER 4 - DISCUSSION AND CONCLUSIONS

The aim of this study was to investigate the time course and neural correlates of audio presentation modes on participants' sense of presence during cinematic immersion. To test this hypothesis, the research project was divided into three main stages. In the first stage, we extracted, selected and validated a diverse set of naturalistic stimuli consisting of validated cinematic excerpts. This approach allowed for a more diverse range of stimuli and more generalizable results, compared to previous studies (Sonkusare et al., 2019).

The participants in the study were chosen based on their ability to distinguish between various acoustic features and different modes of sound reproduction, even though, in order to ensure that the results could be generalized to a larger population, they were "un-trained/naive" participants. In the second stage, we conducted a behavioral experiment to investigate the differences in emotional and bodily involvement and in audio perception among different audio presentation modes (Monophonic, Stereophonic and Surround). The validated stimuli from the first stage were used, and participants assessed Enjoyment (EN), Emotional Involvement (EI), Physical Immersion (PI), and Realism (RE). Results of the behavioral data analysis showed a significant main effect of the experimental condition (Presentation mode) and the type of Question on the participants' scores. Post hoc comparisons revealed that participants consistently rated, in all Questions, stimuli higher when they were presented in the Surround condition compared to the Monophonic or Stereophonic conditions. Specifically, we found that Surround presentation mode was particularly effective in eliciting a sense of Realism (RE), Emotional Involvement (EI) and Physical Immersion (PI) among participants. These data are in line with the meta-analysis by

Cummings and colleagues, who report that the spatial presence experience, evoked by the Surround presentation mode, correlates positively with the level of immersion of the system (Cummings & Bailenson, 2015b). We also corroborate, with more robust results and heterogeneous and ecological stimuli, previous studies results confirming that the sense of presence can be modulated by the spatial sound reproduction (Kobayashi et al., 2015; Lessiter & Freeman, 2001; Pettey et al., 2010; Västfjäll, 2003).

In the third stage, we designed a high-density electroencephalographic (HD-EEG) experiment to investigate the time course and neural correlates of audio presentation modes perception. The validated stimuli from the first stage were again used, and the HD-EEG recordings were used to measure neural activity as participants listened to the stimuli in the different audio presentation modes (Control, Monophonic, and Surround). The main focus was to compare the neural activity in the surround presentation mode to that in the monophonic and control presentation modes, with the hypothesis that the enhanced spatialization of sound in the surround mode would lead to greater activation of embodied simulation mechanisms viewed as physiological index of sense of presence. Using a data-driven approach that allows us to identify specific time windows and electrodes clusters where there is a significant difference in neural activity between experimental conditions without any spatial cluster and frequency band assumption, we identified two significant centro-parietal clusters: the first in Alpha frequency band (8 to 10 Hz) and in the time window from 3 to 7 seconds, the second in the Low Beta frequency band (16 to 18 Hz) and in the time window from 2 to 7 seconds. Further analysis revealed a significant event-related desynchronization (ERD) in the Surround condition when compared to both the Monophonic and Control conditions both in Alpha and Low Beta centro-parietal clusters, confirming previous results (Tsuchida et al., 2015). Comparing our results with the course of typical mu rhythm

desynchronization we can affirm that we observed a significant late ERD peak (around 4.5/5 seconds), even if we observed, in Surround relative to control, a non-significant ERD after stimulus onset. Avanzini and colleagues found that during the observation of specific motor acts, there was a desynchronization of central cortical rhythms across different frequency bands, including alpha (8-13 Hz), and low beta (13-18 Hz). This desynchronization occurred almost as soon as the observed movement started with a peak after around 700 ms and continued for about 400-600 ms after the movement ended. This was then followed by a prolonged power rebound, with the rebound onset delay possibly corresponding to the time necessary for active inhibition to take place following previous cortical excitation. The study also found that there were differences in timing and amplitudes of the rebound in the different frequency bands, with the largest rebound observed in the low beta band and the earliest rebound in the upper beta band. Additionally, there was a greater modulation in parietal regions relative to central regions (Avanzini et al., 2012).

One possible explanation for this timing difference is the sensory modality and the nature of the stimuli used in our research project. We used naturalistic and, in some degree, heterogeneous stimuli extracted from movies, which did not have a time-locked action sound onset. This may have influenced the timing of the neural response observed in our study, as the participants were not presented with a clear and consistent action sound. Furthermore, it is also possible that the use of naturalistic stimuli, as opposed to stimuli created ad hoc, may have led to a more complex and nuanced neural response delayed peak of desynchronization. Furthermore, the heterogeneity of the stimuli may have prevented us from making a definitive conclusion about the influence of vocal actions such as groans, shouts, and other vocalizations on the evoked bodily simulation.

Previous research revealed different source locations and reactivity for the alpha and beta subcomponents of the mu rhythm desynchronization active during action execution and action observation, supporting the idea that they serve distinct functions (Hari, 2006; Hari & Salmelin, 1997; Hobson & Bishop, 2016; Pfurtscheller et al., 1997; Press et al., 2011). The alpha subcomponent is thought to reflect a sensorimotor function, while the beta component is more closely linked to motor cortical control. Indeed, we need to further investigate the different functions of the alpha and beta subcomponents and how they relate to different audio presentation modes.

The results of Heimann and colleagues suggest that there may be a relationship between the perception of approaching stimuli and the feeling of involvement in the scene (Heimann et al., 2014). This may be due to the presence of more depth cues, which more closely resemble real-life vision. The level of similarity between the perceptual experience elicited by video clips and the visual experience during real-life movements is believed to depend on the filming technique. A similar mechanism can be hypothesized for the audio-cinematic component. A similar hypothesis can be made for the audio component in cinematic immersion, where the surround sound presentation can more closely resemble real-life hearing and activate embodied simulation processes.

Further research is needed to fully understand the underlying mechanisms and factors that contribute to this neural response and the functional significance of the activation of embodied simulation mechanisms. Regardless, we can state that immersion, as an objective property of the technological playback system, was a defining characteristic of our stimuli delivery setup, and this was reflected by the instauration of the sense of Presence in participants revealed by stronger engagement of spectators.

This study provides new data on how the level of spatial detail of a scene presented in the acoustic mode alone can influence the participant's perceptions and sensations giving rise to the sense of presence. By further understanding the relationship between sound and visual in the cinematic experience, we can gain insight into how the brain processes information and how it can be used to enhance the immersive experience for the viewer. Furthermore, this understanding can also be applied to other areas such as virtual reality and augmented reality, which also rely on the integration of sound and visual information to create immersive experiences. Overall, the study research project can provide a deeper understanding of the human experience and how technology can be used to enhance it.

REFERENCES

Albiero, P., Ingoglia, S., & Cocco, A. L. (2006). Contributo all'adattamento italiano dell'Interpersonal Reactivity Index. *TESTING PSICOMETRIA METODOLOGIA*, *13*(2), 107–125.

Avanzini, P., Fabbri-Destro, M., Volta, R. D., Daprati, E., Rizzolatti, G., & Cantalupo, G. (2012). The Dynamics of Sensorimotor Cortical Oscillations during the Observation of Hand Movements: An EEG Study. *PLoS ONE*, *7*(5), e37534. https://doi.org/10.1371/journal.pone.0037534

Bech, S., & Zacharov, N. (2006). *Perceptual Audio Evaluation–Theory, Method and Application*. https://doi.org/10.1002/9780470869253

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Cummings, J. J., & Bailenson, J. N. (2015a). How Immersive Is Enough? A Meta-Analysis of the Effect of Immersive Technology on User Presence. *Media Psychology*, *19*(2), 272–309. https://doi.org/10.1080/15213269.2015.1015740

Cummings, J. J., & Bailenson, J. N. (2015b). How Immersive Is Enough? A Meta-Analysis of the Effect of Immersive Technology on User Presence. *Media Psychology*, *19*(2), 272–309. https://doi.org/10.1080/15213269.2015.1015740

DiDonato, M. (2010). *La spazializzazione acustica nel cinema contemporaneo. Tecnica, linguaggio, modelli di analisi* (Onyx, Ed.).

DiPellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research*, *91*(1), 176–180. https://doi.org/10.1007/bf00230027

EBU. (2014). Recommendation 128, Loudness normalisation and permitted maximum level of audio signals. *European Broadcasting Union*.

Elsaesser, T., & Hagener, M. (2015). *Film Theory An Introduction through the Senses*. Routledge.

Fernández-Aguilar, L., Navarro-Bravo, B., Ricarte, J., Ros, L., & Latorre, J. M. (2019). How effective are films in inducing positive and negative emotional states? A meta-analysis. *PLOS ONE*, *14*(11), e0225040. https://doi.org/10.1371/journal.pone.0225040

Figueiredo, H. F., Bodie, B. L., Tauchi, M., Dolgas, C. M., & Herman, J. P. (2003). Stress Integration after Acute and Chronic Predator Stress: Differential Activation of Central Stress

Circuitry and Sensitization of the Hypothalamo-Pituitary-Adrenocortical Axis. *Endocrinology*, *144*(12), 5249–5258. https://doi.org/10.1210/en.2003-0713

Fingerhut, J., & Heimann, K. (2022). Enacting Moving Images: Film Theory and Experimental Science within a New Cognitive Media Theory. *Projections*, *16*(1), 105–123. https://doi.org/10.3167/proj.2022.160107

Freedberg, D., & Gallese, V. (2007). Motion, emotion and empathy in esthetic experience. *Trends in Cognitive Sciences*, *11*(5), 197–203. https://doi.org/10.1016/j.tics.2007.02.003

Fridlund, A. J., & Cacioppo, J. T. (1986). Guidelines for Human Electromyographic Research. *Psychophysiology*, *23*(5), 567–589. https://doi.org/10.1111/j.1469-8986.1986.tb00676.x

Gallese, V. (2009). Mirror Neurons, Embodied Simulation, and the Neural Basis of Social Identification. *Psychoanalytic Dialogues*, *19*(5), 519–536. https://doi.org/10.1080/10481880903231910

Gallese, V. (2019). Embodied Simulation. Its Bearing on Aesthetic Experience and the Dialogue Between Neuroscience and the Humanities. *Gestalt Theory*, *41*(2), 113–127. https://doi.org/10.2478/gth-2019-0013

Gallese, V., & Guerra, M. (2012). Embodying Movies: Embodied Simulation and Film Studies. *Cinema: Journal of Philosophy and the Moving Image*, *3*, 183–210.

Gallese, V., & Guerra, M. (2013). Film, Corpo, Cervello: Prospettive Naturali Per La Teoria Del Film. *Fata Morgana*.

Gallese, V., & Guerra, M. (2019). *The empathic screen: Cinema and neuroscience.* (O. U. Press., Ed.).

Howarth, A., & Shone, G. R. (2006). Ageing and the auditory system. *Postgraduate Medical Journal*, *82*(965), 166. https://doi.org/10.1136/pgmj.2005.039388

ISO. (2017). Standard 7029:2017, Statistical distribution of hearing thresholds related to age and gender. *International Organization for Standardization*.

ITU-R. (1996). Recommendation ITU-R P. 800, Methods for subjective determination of transmission quality. *International Telecommunication Union*.

ITU-R. (1997). Recommendation BS.1116-1, Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems. International Telecommunications Union. *International Telecommunication Union*.

ITU-R. (2000). Recommendation ITU-R P. 832, Subjective performance evaluation of hands-free terminals. *International Telecommunication Union*.

Keysers, C., Kohler, E., Umiltà, M. A., Nanetti, L., Fogassi, L., & Gallese, V. (2003). Audiovisual mirror neurons and action recognition. *Experimental Brain Research*, *153*(4), 628–636. https://doi.org/10.1007/s00221-003-1603-5

Kitagawa, N., & Ichihara, S. (2002). Hearing visual motion in depth. *Nature*, *416*(6877), 172–174. https://doi.org/10.1038/416172a

Kobayashi, M., Ueno, K., & Ise, S. (2015). The Effects of Spatialized Sounds on the Sense of Presence in Auditory Virtual Environments: A Psychological and Physiological Study. *PRESENCE: Teleoperators and Virtual Environments*, *24*(2), 163–174. https://doi.org/10.1162/pres_a_00226

Larsen, E., Iyer, N., Lansing, C. R., & Feng, A. S. (2008). On the minimum audible difference in direct-to-reverberant energy ratioa). *The Journal of the Acoustical Society of America*, *124*(1), 450–461. https://doi.org/10.1121/1.2936368

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology*, *21*(4), R143–R145. https://doi.org/10.1016/j.cub.2010.12.033

Leppänen, J. M., & Nelson, C. A. (2009). Tuning the developing brain to social signals of emotions. *Nature Reviews Neuroscience*, *10*(1), 37–47. https://doi.org/10.1038/nrn2554

Lessiter, J., & Freeman, J. (2001). Really hear? The effects of audio quality on presence. *4th International Workshop on Presence*, 288–324.

Lipscomb, S. D., & Kerins, M. (2004). An empirical investigation into the effect of presentation mode in the cinematic and music listening experience. *8th International Conference on Music Perception & Cognition*.

Lopatka, K., Kotus, J., & Czyzewski, A. (2016). Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations. *Multimedia Tools and Applications*, *75*(17), 10407–10439. https://doi.org/10.1007/s11042-015-3105-4

Murphy, W. J., & Franks, J. R. (2002). Revisiting the NIOSH criteria for a recommended standard: Occupational noise exposure. *International Congress and Exposition on Noise Control Engineering*, 19–21.

Muthukumaraswamy, S. D., & Johnson, B. (2004a). *Changes in rolandic mu rhythm during observation of a precision grip*. *41*(1), 152–156.

Muthukumaraswamy, S. D., & Johnson, B. W. (2004b). *Primary motor cortex activation during action observation revealed by wavelet analysis of the EEG*. *115*(8), 1760–1766.

Muthukumaraswamy, S. D., Johnson, B. W., & McNair, N. A. (2004). *Mu rhythm modulation during observation of an object-directed grasp*. *19*(2), 195–201.

Paul, & A. (2009). *Audyssey DSX 10.2 Surround Sound Overview*.

Pedroni, A., Bahreini, A., & Langer, N. (2019). Automagic: Standardized preprocessing of big EEG data. *NeuroImage*, *200*, 460–473. https://doi.org/10.1016/j.neuroimage.2019.06.046

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Perry, A., Troje, N. F., & Bentin, S. (2010). *Exploring motor system contributions to the perception of social information: Evidence from EEG activity in the mu/alpha frequency range*. *5*(3), 272–284.

Pettey, G., Bracken, C. C., Rubenking, B., Buncher, M., & Gress, E. (2010). Telepresence, soundscapes and technological expectation: putting the observer into the equation. *Virtual Reality*, *14*(1), 15–25. https://doi.org/10.1007/s10055-009-0148-8

Pfurtscheller, G., Pregenzer, M., & Neuper, C. (1994). *Visualization of sensorimotor areas involved in preparation for hand movement based on classification of μ and central β rhythms in single EEG trials in man*. *181*(1–2), 43–46.

Rawashdeh, S. (2021). *Frequency Response of the Ear , Hearing Test. MATLAB Central File Exchange.* https://www.mathworks.com/matlabcentral/fileexchange/16101-frequency-response-of-the-ear-hearing-test

Rigby, J. M., Brumby, D. P., Gould, S. J. J., & Cox, A. L. (2019). Development of a Questionnaire to Measure Immersion in Video Media: The Film IEQ. *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*, 35–46. https://doi.org/10.1145/3317697.3323361

Roberts, R., Callow, N., Hardy, L., Markland, D., & Bringer, J. (2008). Movement Imagery Ability: Development and Assessment of a Revised Version of the Vividness of Movement Imagery Questionnaire. *Journal of Sport and Exercise Psychology*, *30*(2), 200–221. https://doi.org/10.1123/jsep.30.2.200

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, Hosting and Recruiting: A Brief Introduction to Running Behavioral Experiments Online. *Brain Sciences*, *10*(4), 251. https://doi.org/10.3390/brainsci10040251

Sbravatti, V. (2017). *La cognizione dello spazio sonoro filmico: un approccio neurofilmologico*. Università di Roma La Sapienza.

Slater, M., & Wilbur, S. (1997). A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, *6*(6), 603–616. https://doi.org/10.1162/pres.1997.6.6.603

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Sterne, J. (2003). *The Audible Past Cultural Origins of Sound Reproduction*. Duke University Press.

Toro, C., Deuschl, G., Thatcher, R., Sato, S., Kufta, C., & Hallett, M. (1994). *Event-related desynchronization and movement-related cortical potentials on the ECoG and EEG*. *93*(5), 380–389.

Tsuchida, K., Ueno, K., & Shimada, S. (2015). Motor area activity for action-related and nonaction-related sounds in a three-dimensional sound field reproduction system. *NeuroReport*, *26*(5), 291–295. https://doi.org/10.1097/wnr.0000000000000347

Västfjäll, D. (2003). The Subjective Sense of Presence, Emotion Recognition, and Experienced Emotions in Auditory Virtual Environments. *CyberPsychology & Behavior*, *6*(2), 181–188. https://doi.org/10.1089/109493103321640374

Visch, V. T., Tan, E. S., & Molenaar, D. (2010). The emotional and cognitive effect of immersion in film viewing. *Cognition & Emotion*, *24*(8), 1439–1445. https://doi.org/10.1080/02699930903498186

Werner, P. D., Swope, A. J., & Heide, F. J. (2006). The Music Experience Questionnaire: Development and Correlates. *The Journal of Psychology*, *140*(4), 329–345. https://doi.org/10.3200/jrlp.140.4.329-345

Winkler, I., Brandl, S., Horn, F., Waldburger, E., Allefeld, C., & Tangermann, M. (2014). Robust artifactual independent component classification for BCI practitioners. *Journal of Neural Engineering*, *11*(3), 035013. https://doi.org/10.1088/1741-2560/11/3/035013

Winkler, I., Haufe, S., & Tangermann, M. (2011). Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals. *Behavioral and Brain Functions : BBF*, *7*(1), 30–30. https://doi.org/10.1186/1744-9081-7-30

Wirth, W., Hartmann, T., Bocking, S., Vorderer, P., Klimmt, C., Schramm, H., Saari, T., Laarni, J., Ravaja, N., Gouveia, F. R., Biocca, F., Sacau, A., Jäncke, L., Baumgartner, T., & Jäncke, P. (2003). *Constructing Presence: A Two-Level Model of the Formation of Spatial Presence Experiences*.

Witmer, B. G., & Singer, M. J. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, *7*(3), 225–240. https://doi.org/10.1162/105474698565686

Wöllner, C., Hammerschmidt, D., & Albrecht, H. (2018). Slow motion in films and video clips: Music influences perceived duration and emotion, autonomic physiological activation and pupillary responses. *PLoS ONE*, *13*(6), e0199161. https://doi.org/10.1371/journal.pone.0199161

Hari, R. (2006). Action–perception connection and the cortical mu rhythm. *Progress in Brain*

Research, *159*, 253–260. https://doi.org/10.1016/s0079-6123(06)59017-x

Hari, R., & Salmelin, R. (1997). Human cortical oscillations: a neuromagnetic view through the skull. *Trends in Neurosciences*, *20*(1), 44–49. https://doi.org/10.1016/s0166-2236(96)10065-5

Heimann, K., Umiltà, M. A., Guerra, M., & Gallese, V. (2014). Moving Mirrors: A High-density EEG Study Investigating the Effect of Camera Movements on Motor Cortex Activation during Action Observation. *Journal of Cognitive Neuroscience*, *26*(9), 2087–2101. https://doi.org/10.1162/jocn_a_00602

Hobson, H. M., & Bishop, D. V. M. (2016). Mu suppression – A good measure of the human mirror neuron system? *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *82*, 290–310. https://doi.org/10.1016/j.cortex.2016.03.019

Pfurtscheller, G., Stancák, A., & Edlinger, G. (1997). On the existence of different types of central beta rhythms below 30 Hz. *Electroencephalography and Clinical Neurophysiology*, *102*(4), 316–325. https://doi.org/10.1016/s0013-4694(96)96612-2

Press, C., Cook, J., Blakemore, S.-J., & Kilner, J. (2011). Dynamic Modulation of Human Motor Activity When Observing Actions. *Journal of Neuroscience*, *31*(8), 2792–2800. https://doi.org/10.1523/jneurosci.1595-10.2011

World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA*, *310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053