OPEN ACCESS Freely available online

S. Agrawal S. Bec, K. De Moor and S. Forchhammer, "Influence of Changes in Audio Spatialization on Immersion in Audiovisual Experiences"
J. Audio Eng. Soc., vol. 70, no. 10, pp. 810–823, (2022 October).
DOI: https://doi.org/10.17743/jaes.2022.0034

Influence of Changes in Audio Spatialization on Immersion in Audiovisual Experiences

SARVESH AGRAWAL,[1, 2, *] *AES Student Member*, **SØREN BECH,**[1, 3] *AES Fellow*,
(sraj@bang-olufsen.dk) (sbe@bang-olufsen.dk)
KATRIEN DE MOOR,[4] **AND SØREN FORCHHAMMER**[2]
(katrien.demoor@ntnu.no) (sofo@fotonik.dtu.dk)

[1]*Bang & Olufsen a/s, Struer, Denmark*
[2]*Department of Photonics Engineering, Technical University of Denmark, Lyngby, Denmark*
[3]*Department of Electronic Systems, Aalborg University, Aalborg, Denmark*
[4]*Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Trondheim, Norway*

Understanding the influence of technical system parameters on audiovisual experiences is important for technologists to optimize experiences. The focus in this study was on the influence of changes in audio spatialization (varying the loudspeaker configuration for audio rendering from 2.1 to 5.1 to 7.1.4) on the experience of immersion. First, a magnitude estimation experiment was performed to perceptually evaluate envelopment for verifying the initial condition that there is a perceptual difference between the audio spatialization levels. It was found that envelopment increased from 2.1 to 5.1 reproduction, but there was no significant benefit of extending from 5.1 to 7.1.4. An absolute-rating experimental paradigm was used to assess immersion in four audiovisual experiences by 24 participants. Evident differences between immersion scores could not be established, signaling that a change in audio spatialization and subsequent change in envelopment does not guarantee a psychologically immersive experience.

0 INTRODUCTION

Spatial audio has been an integral part of audiovisual experiences for several decades in the form of channel-based audio. Technological advancements (e.g., object-based audio) have revitalized spatial audio and enabled new, intriguing auditory experiences. Listening tests are crucial for evaluating, understanding, and improving the experiences created using modern spatial audio techniques. Scientific effort has been focused on developing a better understanding of spatial audio reproduction from a perceptual perspective. Nevertheless, the influence of spatial audio reproduction on encompassing concepts such as immersion and quality of experience is poorly understood.

Enhancing the experience for the users is one of the fundamental goals for technologists and creators. However, they have limited control and can only alter the system parameters in the physical domain. Thus, it is vital to assess and establish how changes in the physical domain influence hedonic measurements.

The primary focus of this paper is to determine the influence of technical parameters of the spatial audio system (rendering to different loudspeaker configurations) on the experience of immersion. The authors were interested in audiovisual experiences because the goal of the overall project was to evaluate immersion and lay the foundation for enhancing engagement for entertainment purposes in the home. This study will test the common assumption that spatially superior systems lead to more immersive audiovisual experiences. It will form the foundation for further work aimed at bridging the gap between the technical parameters and immersion. The method of magnitude estimation

*Correspondence should be addressed to Sarvesh Agrawal, e-mail: sraj@bang-olufsen.dk

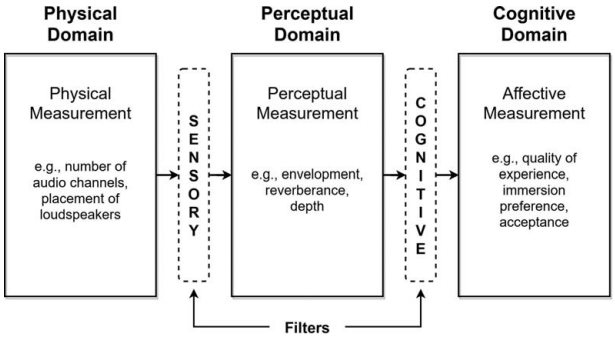


Fig. 1. Filter model as described in [1]. This model is used to study the influence of audio spatialization on envelopment and immersion in the present study.

applied for evaluating perceptual attributes and the subjective assessment methodology for evaluating immersion can be inspirational for audiovisual assessment in a variety of settings.

0.1 Filter Model Framework

The central idea of this paper can be organized with the help of the filter model [1, 2]. The model is applied to this task as depicted in Fig. 1. The model starts with the physical domain where the stimulus can be characterized by objective measurements, such as audio channels, number of loudspeakers, etc. In the context of this study, the primary variable is the loudspeaker configuration for audio rendering. The physical stimulus is perceived upon being filtered by the sensory system. This is called an auditory event and can be broken down into attributes (e.g., envelopment) that can be measured perceptually. Finally, the percept passes through a cognitive filter to form a holistic impression of the auditory event (e.g., immersion). Mood, preference, emotional state, and other non-sensory factors are accounted for by the cognitive filter. Evaluation of quality, immersion, liking, etc. is housed in the cognitive domain and requires an integrative frame of mind.

0.2 Research Question and Organization

The primary research question for this study was "What is the influence of the changes in audio spatialization on the experience of immersion in a domestic audiovisual setting?"

The target relationship to establish is one from the physical domain (changing audio spatialization) to cognitive domain (experience of immersion). However, varying the independent variable and recording immersion responses cannot guarantee that the ratings are obtained because of perceptual differences and not because of cognitive factors. Therefore, the assumption is that the changes in the physical domain lead to a discernible change in the perceptual domain. Evaluating the experiences at each step helps in assessing the effect of the sensory and cognitive filters. The results of the perceptual changes can be interpreted in conjunction to immersion ratings to develop a comprehensive understanding of the changes caused by the independent variable.

The task of choosing the relevant perceptual attribute(s) for spatial audio reproduction can be an experiment in itself. Methods such as check-all-that-apply can be used for attribute selection [3]. In this study, *Recommendation ITU-R BS.2399-0* [4], which presents a sound wheel developed by aggregating perceptual attributes for sound reproduction from over 20 studies, was consulted. From the spatial extent attributes presented in the lexicon (balance, envelopment, width, depth), envelopment was deemed to be the most relevant for this study for the following reasons: 1) envelopment has been a recurring attribute of spatial impression in multichannel audio reproduction [4–8]; 2) it is an attribute that was found to be always important for listener preference when evaluating spatial audio reproduction methods in a study conducted by Francombe et al. [9]; 3) it is applicable to both dimensions: horizontal and vertical; and 4) experience suggested that envelopment will be the attribute most affected by the changes in audio spatialization levels.

In this work, it is assumed that more envelopment should yield more immersion and that more immersive experiences will be preferred. The assumption that more immersive experiences will be preferred has been discussed previously [10] and is at the core of the motivation for studying immersion.

This paper is organized as follows. SEC. 1 describes the program material and experimental setup for the two experiments conducted in this study. SEC. 2 focuses on the influence of changes in audio spatialization on envelopment. Immersion is assessed using the same independent variable as for envelopment in SEC. 3. Finally, the results from both experiments are discussed to answer the research question in SEC. 4.

0.3 Envelopment and Immersion

Spatial audio reproduction over loudspeakers allows for placement of sound sources around the listeners in the horizontal and vertical plane (physical domain). This can lead to an enveloping feeling for the listener (perceptual domain). Envelopment in audio reproduction is referred to as source-related envelopment [11]. It has been defined as the "degree of being surrounded by a source, scene or ensemble" [4] and "are you surrounded by the reproduced sound and does it give a sense of space around you?" [12]. It can be created by dry (direct) and/or reverberant (indirect) sound sources around the listener [13, 11].

Envelopment is commonly explained as being analogous to swimming underwater rather than being sprayed by a water hose [2]. This analogy is similar to the one used for describing immersion and is potentially responsible for causing confusion between the two ideas [14]. Although immersion can be described physically and perceptually [15], it refers to psychological immersion in this study. The definition and description provided by Agrawal et al. [14] is used for this investigation.

> Immersion is a phenomenon experienced by an individual when they are in a state of deep mental involvement in which their cognitive processes (with or without sensory stimulation) cause a shift in their attentional state such that one may experience disassociation from the awareness of the physical world.

Immersion is a holistic concept that is equivalent to mental engagement in an activity. For a detailed discussion on immersion and its assessment, refer to [16, 14, 10].

1 PROGRAM MATERIAL AND EXPERIMENTAL SETUP

The program material and experimental setup were identical for both experiments. These are detailed in the following sections.

1.1 Program Material

Three levels of audio spatialization were chosen for the independent variable—2.1, 5.1, and 7.1.4 reproduction. The rationale for these levels was as follows: The 2.1 reproduction is one of the common reproduction setups in which the speakers are placed in front of the listener. The 5.1 reproduction has been the standard for surround audio reproduction, and it extends audio to loudspeakers placed around the listener in the horizontal plane. Finally, audio can be placed in the median plane with 7.1.4 reproduction to give an impression of height. It is becoming a de facto standard for domestic audio reproduction that includes elevation.

The stimuli were chosen such that they had ultra high-definition (UHD), high dynamic range visuals. The native aspect ratio and chroma sub-sampling were unaltered in the experiments. It should be noted that the stimuli were not mastered in UHD resolution, and the versions of high dynamic range were different among the stimuli. See [10] for additional details.

The material had to satisfy the conditions for assessing immersion as discussed in [14, 10] because it was held constant for both experiments. These conditions include length of stimuli, variety in genre (in an attempt to evoke spatial, emotional, and temporal immersion), and standalone stimuli that do not require prior knowledge of the content.

From the perspective of testing envelopment, efforts were made to pick stimuli in which there was an apparent difference in envelopment when rendered as 2.1, 5.1, and 7.1.4. A clearly perceived difference among the levels was critical because the evaluation of envelopment was to determine the assumption of a perceptual difference for assessing immersion.

The program material was largely inspired by the stimuli mentioned in [10]. They are listed in Table 1. The example excerpt was used only for the part of testing immersion and is explained later in SEC. 3. The abstract excerpt was used for demonstrating the differences in sound envelopment that can be experienced. Please refer to [14, 10] to know more about the rationale for choosing the stimuli and limitations that restrict the stimuli universe.

1.2 Experimental Setup

The experimental setup used in this experiment was identical to the setup described in [10]. The tests were conducted in an IEC 60268-13 standardized listening room [17]. Audio was decoded to 7.1.4 channels and fed to speakers that were at a distance of 2 m around the listening position. The speakers were positioned as per Dolby guidelines [18], time aligned, and level calibrated. Stimuli (at the different audio spatialization levels) were loudness matched by ear at the listening position by two experienced listeners.

Video was displayed on a 65-in OLED screen. It was positioned to yield a zero-degree vertical and horizontal viewing angle. The viewing distance was same as the listening distance, i.e., 2 m, that followed the design viewing distance for UHD resolution stated in *Recommendation ITU-R BT.2022* [19]. Screen settings were tuned based on D65 value and in part based on experience by two viewing experts. Please refer to [10] for further details on the experimental setup of the audiovisual reproduction system.

2 EXPERIMENT 1: INFLUENCE OF AUDIO SPATIALIZATION ON PERCEIVED ENVELOPMENT

Methodologies for the perceptual assessment of audiovisual attributes are well-established (see [1, 3]). Direct scaling methods that employ rating scales are common for evaluating envelopment. Popular methods such as multiple stimulus hidden reference and anchor (MUSHRA) and *Recommendation ITU-R BS.1116-3* employ a reference for comparing the test stimulus against the original reference. Problems arise when the definition of reference stimulus is not established as in the case of frontier technologies and experiences (e.g., spatial audio and virtual reality) [20]. Additionally, end-point effects[1] may be present when the nature and range of the stimulus set is not well-known to the participants, as in the case of new technologies. Furthermore, it should be noted that a number of methods, including MUSHRA and *Recommendation ITU-R BS.1116-3*, are not suitable for conducting tests with inexperienced participants. They are complex and designed for determining fine differences among the stimuli under test.

In this study, the authors were not interested in small differences. Instead, the goal was to uncover the differences discernible to inexperienced participants. The experimental setup described in SEC. 1.2 did not allow for simultaneous comparisons between stimuli. Given the need for absolute (i.e., non-concurrent) evaluations[2] and the lengthy nature

[1] The clustering of scores around the center or toward the end of the scale is because the participants were reluctant to use the end points of the scale. This can restrict the resolution of the scale [1].

[2] The experimental setup restricted simultaneous judgments because the stimuli were presented live from the Blu-ray discs. Additionally, to maintain consistency, envelopment was evaluated without simultaneous comparisons because immersion was assessed in an absolute manner.

Table 1. Audiovisual excerpts used in the experiment.

Excerpt	Content	Genre	UK year of release	Timecode
Example	*Earth: One Amazing Day*	Nature documentary	2018	00:08:50–00:16:49
Abstract	*Kinsetsu: Textures From Planet 9**	Fantasy/Sci-Fi	2018	Around first minute
A	*Dynasties*, "Lion"	Nature documentary	2018	00:16:11–00:20:00
B	*Spider-Man: Into the Spider-Verse*	Animation/Action	2019	00:18:52–00:29:39
C	"The Pines" by Roses & Revolutions*	Alternative/Indie	2018	Complete video (3:10 min)
D	*Jumanji: Welcome to the Jungle*	Action/Adventure	2018	01:15:52–01:21:00

Note. *From DTS demo disc 2018.
The copyright for the content used in this experiment is held by the respective parties. No files were copied or stored during any stage of experimentation.
The genres have been obtained from IMDb (https://www.imdb.com) and selected to reflect the primary genre of the content.
The year of release represents the release on Blu-ray. The actual release year may differ.
The lengths of the excerpts range from approximately 4 to 12 min.

of stimuli, an alternative method for assessing envelopment was needed.

Magnitude estimation, a specific type of ratio scaling approach, avoids the issues with traditional methods for evaluating perceptual attributes. It is a psychophysical scaling technique where the participants are instructed to assign numerical values according to the perceived strength of the stimulus for a specified characteristic. They are tasked with generating numbers that follow the ratio principle, i.e., if stimulus B is perceived to be twice as strong as stimulus A, stimulus B should be rated twice the rating of stimulus A. For example, if stimulus B is thought to be thrice as loud as stimulus A, it should be rated 210 if stimulus A was rated 70. When all data is collected, it can be normalized or equalized (depending on the experimental design choices) [21] and compared among participants or aggregated across participants.

This method is straightforward and relatively simple to understand [20]. Magnitude estimation provides flexibility to the experimenter for designing studies because choices can be made regarding the first sample (fixed vs. random; internal vs. external reference), numerical value assigned to it (fixed vs. free), and order of presentation of stimuli (fixed vs. random) [22]. Historically, magnitude estimation tests have been conducted with inexperienced participants with rarely any training on the method [21]. The method is less susceptible to end-effect because the scale is usually open-ended on the upper side. Please refer to APPENDIX A.1 for information on the background, advantages, and limitations of magnitude estimation.

2.1 Experiment Design

There are many avenues for designing magnitude estimation experiments. The different choices affect the analysis of collected data and can influence the conclusions that can be drawn. Key considerations, reasoning, and implications are discussed in this section.

Selection of the first stimulus is the first question to address for designing magnitude estimation studies. The first stimulus is often referred to as the *standard*, and there are two choices for making a selection. The standard can be fixed in the middle of the stimulus range, or it can be randomly selected. Fixing the standard in the middle of the stimulus range is preferred when the variation in the range of stimuli is so large that it would cause problems if the participants randomly get a high or low standard. For example, if a participant starts with a high standard and rates it 10, rating stimuli that are several magnitudes lower will be difficult because of smaller numbers. Randomly selecting the first sample is the most accepted method [21] and was chosen for this experiment.[3]

Having a different standard for each participant should not affect the relative magnitudes for the stimuli set but affects the actual numbers used by the panelists [21]. A different starting sample among the participants also helps with reducing the round-number tendency bias (tendency where assessors prefer using round numbers). For instance, if there is a fixed standard and all participants start with the same number (e.g., 100), many of the ratings assigned will be the same round numbers.

The number assigned to the first sample (the standard) is called the *modulus*. The modulus can be fixed (fixed modulus), or the participants can be permitted to assign any number of their choice to the standard (free modulus). In case of fixed modulus, the experimenter assigns a predetermined number to the first sample, and the participants are instructed to assign numbers to following stimuli in relation to the fixed number [21]. Fixing the modulus compels participants to use the scale initiated by the experimenter as opposed to the one they wish to use. Importantly, fixing the modulus can exaggerate the round-number tendency.

In this experiment, the participants were permitted to use the modulus of their choice. In an attempt to discourage the participants from picking a very high or low modulus, they were guided to pick a number between 30 and 100 according to the ASTM standard [23]. This increases the workload for the experimenter because they must process the data appropriately to account for the differences in the range of numbers used by the participants. The differences can be accounted by the assessor factor in the analysis of variance (ANOVA) model or by rescaling the scores [22]. The idea and process of rescaling the data are explained in APPENDIX A.2.

[3] Ideally, all samples should occur as the standard an equal number of times.

Following the standard, the order of presentation of stimuli was randomized in this experiment. Randomization is the most common method for presenting stimuli. It is beneficial in reducing the "sequential dependencies" [21] that arise when the order of presentation is fixed. The carryover effect between stimuli can also be reduced by using balanced designs such as Latin square [22].[4] Additionally, it is ideal to balance the order of samples as well because the variances between consecutive samples will be smaller than those between non-consecutive samples, when participants are instructed to assign numbers by comparing the current stimulus to the preceding one.[5]

The nature of the concept under evaluation influenced the instructions provided to the participants. Envelopment was treated as a unipolar attribute, i.e., zero marked the absence of envelopment and the scale stretched to positive infinity (in theory). The usage, analysis, and instructions for bipolar attributes is more complex. Please refer to [21, 24] for a discussion on the topic.

Finally, a warm-up task was included in the design to familiarize the participants with the method of magnitude estimation. A task where the participants had to estimate the length of lines was used as the warm-up task. A practice task is critical in screening participants for their understanding of the task. Screening is especially helpful when inexperienced participants are recruited for the test.

2.2 Assessors

Eleven assessors participated in this experiment.[6] They were seven males and four females between the ages of 22 and 35 years ($\overline{x} = 26.5$, $\sigma = 3.7$). The participants were inexperienced in audiovisual assessment, i.e., they had not received training for perceptual audiovisual evaluations but may have participated in audiovisual tests previously. Usually, evaluation of perceptual attributes, such as envelopment, is performed by experienced assessors because they possess a high degree of sensory sensitivity [1]. Notwithstanding, inexperienced participants were recruited because the results were used to determine the initial condition for testing immersion (an affective concept) experienced by non-expert participants. It was important to ensure that both envelopment and immersion were gauged by participants with a similar level of audiovisual evaluation experience.

All participants were unfamiliar with the method of magnitude estimation. They participated in the familiarization task that doubled as a screening stage. All participants passed the screening. The assessors were permitted to participate after self-reporting auditory and visual acuity.

2.3 Test Procedure and Instructions

The test was completed in one 80-min session with 2-min breaks between stimuli presentations. The participants were asked to judge the *overall sound envelopment* experienced in the 12 excerpts (four stimuli at three spatialization levels). The following description of sound envelopment was synthesized from [4, 12] and provided to the assessors.

> Sound envelopment is the perception of being surrounded by sound(s). The sound scene is said to be enveloping if it wraps around you. Sound envelopment does not need to be restricted to the sound coming from around you; it can also include sounds coming from below or above you.

The participants were asked to report envelopment by assigning numbers to the experiences. They were instructed to assign any number of their choice to the first stimulus. The authors suggested choosing a number between 30 and 100 according to the ASTM standard [23] because the first stimulus was random (could be high, neutral, or low) and the scale was limited by zero on the lower end. The participants were told to assign successive numbers in such a way that it reflected the ratio of envelopment as per their subjective impression. For example, if they found the second presentation to be thrice as enveloping as the first one, they would give it a rating triple that of the first excerpt. This means that if they rated the first excerpt as 70, the second excerpt would be rated 210 if it was thrice as enveloping or 14 if it was one-fifth as enveloping. They were instructed to assign numbers by comparing the current stimulus to the preceding stimulus (except for the standard in which they were free to use any number).

It was emphasized that there was no limit to the range of numbers that the assessors could use. However, they were not permitted to use zero or negative numbers. The use of decimals was stressed because there is a tendency to use round numbers [21]. Participants were encouraged to try their best to match each number to their perception of envelopment because there were no correct answers.

A short abstract excerpt (approximately 1 min) that had exaggerated sound envelopment was used to illustrate the different degrees of sound envelopment before beginning the test. The participants were told that the demonstration was purely to illustrate the idea of sound envelopment and that the envelopment they experienced during the experiment may not be limited to the illustrations. Because the experiment was conducted as an absolute-assessment test without repetitions, the participants were informed that they did not have the ability to directly compare, rewind, or rewatch the excerpts.

[4] Because of uncertainty in recruiting participants during the coronavirus pandemic, a balanced design was not chosen. Instead, the order was randomized, which is recommended when balanced designs cannot be used [22].

[5] Magnitude estimation studies can be run by requiring the participants to memorize the intensity of the standard and rate all stimuli relative to the standard. However, "people probably 'chain' their ratings to the most recent items in the series" [24].

[6] The ISO standard on sensory analysis using magnitude estimation [22] recommends at least 20 assessors when they are newly trained. However, the number of participants depends the closeness of the stimuli under test, required statistical power, importance attached to the results, and so on. Considering that the authors were interested in a perceptual attribute for which there may not be substantial inter-individual differences, it has been recommended that fewer respondents may be sufficient for an accurate measurement of sensory magnitude [21].

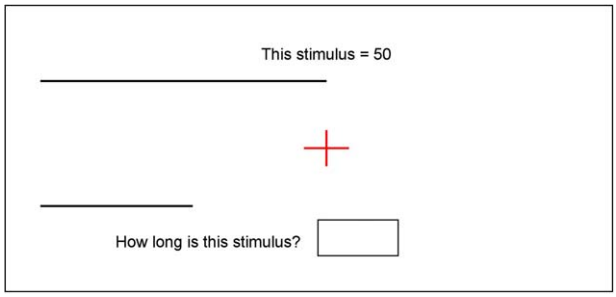


Fig. 2. Example of the length of lines task used for familiarizing and screening the test participants. Here, the top stimulus is a fixed reference, and the assessors are asked to judge the length of the bottom line.

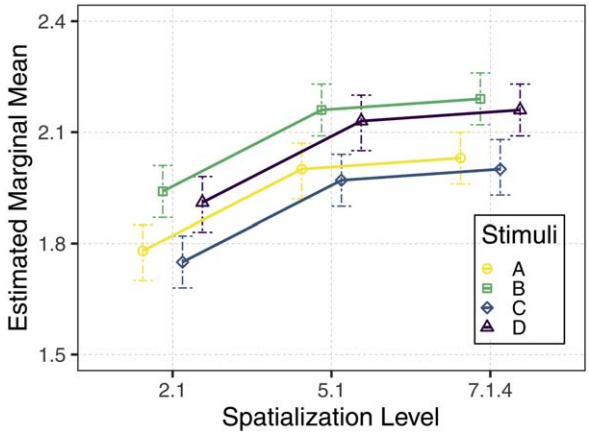


Fig. 3. Estimated marginal means of sound envelopment scores for all stimuli at all spatialization levels. The error bar represents 95% confidence interval. Note that the marginal means are of log-transformed scores. The lines are offset horizontally for better intelligibility.

To further understand the task of assigning numbers to perceptions, a short familiarization and screening exercise was performed using the length of lines task on the computer.[7] Eight levels of line lengths with three repetitions were used. There was a fixed reference with a length of 50 units. The task for the participants was assign numbers appropriately to another line in the presence of the fixed reference. A visual example of the tool is shown in Fig. 2. All participants passed the screening exercise.

The instructions were given provided in writing and discussed verbally by the first author. The entire test was conducted as a pen-and-paper test.[8] Although paradigms in which participants can view their previous ratings do exist [24], the participants did not have access to past ratings in this test, in an effort to limit bias. All recruited participants finished the test.

2.4 Envelopment Data Analysis and Results

The appropriate procedure for analyzing the data obtained from magnitude estimation depends on the experimental choices (complete or incomplete design, presence of replicates, unbalanced number of observations, etc.) and conclusions to be drawn (relative intensities, response relationships, and recalibration of scales.). The ISO standard on the use of magnitude estimation for sensory analysis [22] illustrates the analysis for the different scenarios in detail. Data from this experiment were analyzed using mixed-effect ANOVA after performing total rescaling. The exact method for rescaling and reasoning for logarithmic transformation is explained in APPENDIX A.2.

A mixed-effects ANOVA model described in Eq. (1) was used to analyze the data post–log transformation.

$$y = \mu + spatialization + stimuli + interaction\ between\ stimuli\ and\ spatialization + random\ interaction\ between\ assessor\ and\ stimuli + error. \quad (1)$$

The envelopment score (log-transformed) is modeled in the above equation. Spatialization and stimuli were fixed factors, whereas the interaction between participant and stimuli was a random factor.[9] The interaction between spatialization and stimuli is also taken into consideration in order to ensure that the two could be interpreted independently.

The two interactions were found to be statistically insignificant. Spatialization and stimuli were found to be highly statistically significant. The assumptions of homogeneity of variances and normally distributed residuals were satisfied. Thus, the model was reduced to spatialization and stimuli as fixed factors, and ANOVA was performed again.

There was a significant main effect of audio spatialization on sound envelopment, $F(2, 126) = 27.85$, $p < 0.001$. The post hoc tests (Bonferroni correction) showed that 2.1 audio spatialization was significantly different from 5.1 and 7.1.4 levels ($p < 0.001$). No significant difference was found between 5.1 and 7.1.4 audio spatialization.

The stimulus had a significant effect on sound envelopment as well, $F(3, 126) = 9.70$, $p < 0.001$. Stimuli A and C were statistically different from stimuli B and D as per post hoc comparisons (Bonferroni correction). The estimated marginal means for all four stimuli at every level of audio spatialization is shown in Fig. 3. The figure shows that different stimuli have different perceived envelopment

[7] An online tool at *https://isle.hanover.edu/* was used for the task.

[8] Unlike popular rating tests used in audiovisual assessment, the visual appearance of the ballot is not critical [24]. Instead, the instructions provided to the participants and their comprehension of generating numbers following the ratio principle is important.

[9] Assessor was not included as a factor in the model because the scores were rescaled to equalize the geometric means among the participants.

for the starting spatialization level of 2.1. All stimuli experience gain envelopment when rendered as 5.1 as opposed to 2.1. Envelopment appears to plateau moving from 5.1 to 7.1.4 level.

The reduced ANOVA model did not violate the assumptions of homogeneity of variances as tested by Levene's test. The residuals were not statistically different from the normal distribution as per the Shapiro-Wilk test, $W = 0.98$, $p > 0.1$, with skewness of 0.11 and excess kurtosis of 0.54.

2.5 Discussion of Envelopment Results

The method of magnitude estimation was used in this study to evaluate the differences in envelopment in spatial sound reproduction for four stimuli. Scores were rescaled and analyzed using ANOVA. The observations about difference in spatialization from a visual inspection of Fig. 5 (see APPENDIX A.2) can be confirmed by ANOVA. Audio spatialization was found to be statistically significant, with the 2.1 rendering being different from 5.1 and 7.1.4 rendering. This result was expected because placing loudspeakers around the listener is bound to increase envelopment based on the definition of envelopment in sound reproduction. However, it is interesting to note that an increase in spatialization from 5.1 to 7.1.4 does not substantially increase envelopment.

A possible explanation for this observation can be found in the way commercial content is produced. The 2.1 and 5.1 audio productions have been the norm for a few decades, whereas object-based audio with the ability to add height content for domestic applications is new. It can be argued that there is limited experience for content production using height content, and thus, the capabilities of the reproduction have not been exploited fully [9]. Hence, the authors refrain from interpreting the results to suggest that there is no added benefit of rendering audio to 7.1.4 reproduction systems. Additionally, because a limited set of four stimuli was used in this experiment, the results should be generalized with care.

Audio from the excerpts was decoded to 7.1.4 and recorded to evaluate the content carried by the height channels. It was found that elevation was primarily used for momentary spatialized sound effects and reverberation. Moreover, the content rendered to the elevation channels was found to be substantially lower in level in comparison with the front channels. This can explain the lack of improvement in envelopment between 5.1 and 7.1.4 reproduction. The audio level plots are presented in APPENDIX A.3.

The stimuli factor was found to be significant with two groups of stimuli: A and C were both different from B and D. This demonstrates that not all stimuli provide the same perception of envelopment over the three spatialization levels used in this study. Furthermore, it can be seen in Fig. 3 that different stimuli have varying degrees of perceived envelopment even for 2.1 rendering. These differences are preserved at every spatialization level. All stimuli benefited from an improvement in spatialization from 2.1 to 5.1. The improvement in reported envelopment is marginal from 5.1 to 7.1. There was no interaction between stimuli and spatialization levels as expected, since it is unlikely that envelopment would deteriorate with an increase in spatialization from 2.1 to 5.1 and then increase from 5.1 to 7.1.4 for any of the stimuli under test.

3 EXPERIMENT 2: INFLUENCE OF AUDIO SPATIALIZATION ON IMMERSION

The assessment of immersion is a relatively new and under-explored area. Existing methods can be broadly classified into subjective and objective measures. Objective measures include physiological and behavioral measurements. Because the relationship between objective measures and immersion remains undetermined, this study is restricted to subjective measures that include rating tasks, questionnaires, open-ended interviews, etc. [14, 10]. A complete overview of the methods is presented in [16, 14].

Agrawal et al. [14] pointed out several implications for conducting experiments based on the definition stated in SEC. 0.3. The key implications for developing experimental paradigms are listed below:

- Subjective assessment of immersion must be performed post-experience to avoid disturbing the participant, who may be experiencing immersion.
- Absolute-assessment methods should be picked over comparative methods.
- There should be a sufficient time gap between presentation of consecutive stimuli to allow the assessors to disengage from the previous experience.
- Fatigue should be monitored and all responses should ideally be recorded in the same session.

In this experiment, audio spatialization (three levels) was the independent variable. Typically, it would be possible to use the same group of assessors for evaluating all levels of the independent variable, as is the case in repeated-measures design. However, it has been suggested that repetitions of stimuli can lead to bias, and thus, assessors should evaluate a stimulus only once when assessing immersion [14]. Therefore, an independent-measures design, also known as between-groups or between-subjects design, was used in this experiment.

The participants were divided into three groups, each representing a level of audio spatialization (2.1, 5.1, and 7.1.4). The number of participants in each group was identical. The assignment of participants to the groups was performed randomly. All participants were subjected to the same four audiovisual excerpts, except the audio was rendered according to their group. The order of presentation of stimuli was randomized for all participants.

This experimental design was appropriate for multiple reasons. Foremost, the design accounted for the implications on experimental paradigm outlined in [14]. There was limited transfer of learning for the participants because they evaluated each stimulus once. Limiting the participants from participating in multiple audio spatialization conditions enabled the data collection to be performed in a single session. Furthermore, it controlled the effect of fa-

tigue and considerably reduced the time commitment for the participants.

3.1 Experiment Design

The experimental design choices tested in [10] were used for this investigation. Thus, only the key aspects of the experimental design are discussed here.

The task for the participants in the test was to rate their experience of immersion. The primary benefit of the rating task is that it captures unexpected aspects of the immersive experience that are omitted in questionnaires due to the pre-determined set of questions. A 15-cm long graphic line scale was chosen as the response format. Although it is more difficult to use than other scales, it provides infinite steps (in theory) to report the experience of immersion, and a lack of semantic information reduces bias and clustering of scores [10].

The test was constructed such that the participants could experience each stimulus only once. Switching between excerpts and other user controls, such as rewinding, were prohibited. The participants were provided with a short narrative synopsis before each excerpt to inform them about the narrative background of the excerpt. Interactive distractor tasks were incorporated between presentations with the goal of introducing a pause and helping the participants to shift their attention away from the preceding experience. The three tasks used in this test were a memory game, matchstick puzzle, and Tetris brick puzzle. An iPad was used for all tasks.

3.2 Participants

Twenty-four participants[10] participated in this experiment. Assessors from the previous investigation on envelopment were prohibited from participating in this experiment. The participants were each randomly assigned to one of the three groups: 2.1, 5.1, and 7.1.4 spatialization levels. Thus, each group was comprised of eight participants. The recruited assessors were employees at Bang & Olufsen who were not associated with research and development activities or had expertise in audiovisual assessment. Nevertheless, they may have participated in audiovisual evaluations previously. The mean average age of the participants was 37.5 years ($\sigma = 13.3$). In total, 9 females and 15 males participated in this experiment. All participants self-reported auditory and visual acuity.

[10] The authors were interested in large differences among reported immersion scores as a result of the spatialization levels. The idea was to detect differences that would be of practical significance and justify the resource costs that increase with an increase in audio spatialization. Using the data obtained from previous experiment [10] and pilot tests, the authors determined that eight participants in each group could help detect a difference of ± 1.5 points for an alpha level of 5% and beta level of 20%, provided the variation remained constant between the experiments.

3.3 Experiment Procedure

The experimental procedure for this test was identical to that detailed in [10]. Therefore, only the key points are stated here.

The experiment was conducted within a single session of approximately 50 min. The participants were instructed to rate *overall* immersion in the audiovisual experiences using a graphic line scale as described in [10]. They were provided the following description of immersion:

> Immersion, also known as deep mental involvement, can be described as being mentally lost (absorbed) in the experience. Immersion is encountered when the experience is involving and absorbs you mentally by capturing your attention. For example, immersion may be experienced when reading a book, playing video games, watching a movie, etc.

In addition to rating immersion, assessors were asked to describe their experience and how it influenced their rating in as much detail as possible. They were instructed to use complete sentences and freely describe all aspects of the experience that were relevant to them. The qualitative data was purely to explore the factors considered by the participants for assessing immersion. The discussion of the qualitative analysis results is beyond the scope of this paper and will not be presented here.

An excerpt that may illicit immersion was shown before the experiment. It was explicitly mentioned that the excerpt was only for demonstration, may not illicit immersion, and should not be considered to be a reference. The participants were informed that they will not have a chance to re-watch or rewind the excerpts because it was an absolute-assessment test. They performed a distractor task between the presentations as discussed in [10]. The assessors were assured that there were no correct answers.

3.4 Immersion Data Analysis and Results

Immersion ratings were converted to scores by measuring the distance from the left-most point on the scale. The range of scores was between 0 and 15 because the scale was 15-cm wide (see [10]). A mixed-effects ANOVA model described in Eq. (2) was constructed to analyze the scores.

$$y = \mu + spatialization + stimuli + interaction\ between\ stimuli\ and\ spatialization + random\ effect\ for\ assessors + error. \tag{2}$$

The immersion score is modeled in the above equation. Spatialization (block effect), stimuli, and their interaction were fixed factors. A random effect for each assessor was included in the model because the participant pool was an extremely small subset of the population.

The between-subjects factor, i.e., spatialization level, was found to be statistically insignificant, along with the interaction between stimuli and spatialization. The main effect of stimuli on immersion, $F(3, 63) = 11.92$, $p < 0.001$, was found to be highly statistically significant, showing that

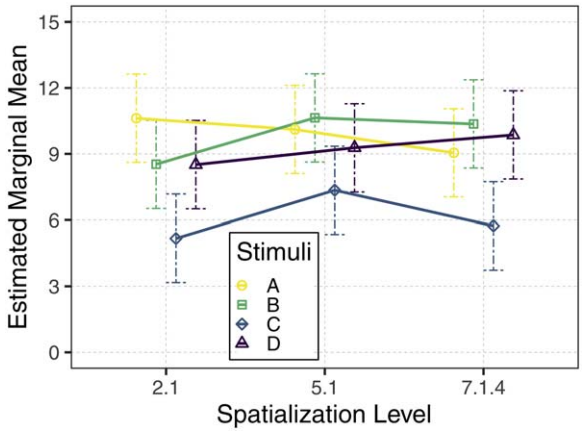


Fig. 4. Estimated marginal means of immersion scores for all stimuli at all spatialization levels. The error bar represents 95% confidence interval. The lines are offset horizontally for better intelligibility.

at least one of the stimuli was different from other. The post hoc test (Bonferroni correction) showed that Stimulus C was significantly different from all other stimuli ($p < 0.001$).

The assessor random effect could explain only about 18% of the total variation in immersion scores. The ANOVA model did not violate the assumptions of homogeneity of variances as tested by Levene's test. The residuals were not statistically different from the normal distribution as per the Shapiro-Wilk test, $W = 0.99$, $p > 0.1$, with skewness of –0.10 and excess kurtosis of –0.29.

The estimated marginal means for all four stimuli at every level of audio spatialization is shown in Fig. 4. There are no obvious patterns that can be deduced from the figure. The confidence intervals are quite large (approximately four units) and overlapping.

3.5 Discussion of Immersion Results

A between-subjects design was used in this experiment to evaluate immersion in four experiences at three audio spatialization levels. A major drawback of between-subjects design is the requirement of a large sample to obtain reliable data. Adding additional conditions considerably increases the experimental effort. For example, if two additional spatialization levels were to be added, 40 participants would need to be recruited.[11] The requirement for resources grows with the addition of variables and their levels, making larger-scale experiments impractical.

In this experiment, the participants were assigned randomly to the groups and evaluated only one condition. The individual variability can obscure the differences between the groups and lead to unreliable results. This can be particularly problematic when the sample size is limited, as in the current study. Participants can be assigned to the groups based on factors such as age, gender, etc. However, this will hamper the generalizability of the results.

[11] This is assuming the same resolution as in the current experiment with three groups.

Audio spatialization, the between-subjects factor, was found to be statistically insignificant, demonstrating that the immersion scores did not improve considerably with an improvement in audio spatialization. Unlike the previous experiment, no difference was observed in immersion between 2.1 and 5.1 reproduction. This is an important finding that suggests that merely improving the spatial audio capabilities of the system is not sufficient for experiencing more intense immersion. It can be argued that insignificant results were obtained because the participants were not subjected to multiple spatialization levels. However, the current experimental design is more ecologically valid than repeated-measures design because the spatialization is seldom changed in domestic applications.

The main effect of stimuli was highly statistically significant, where stimulus C was found to be different from the other three. Interestingly, stimulus C was the only music excerpt in this test. Several participants noted in their comments and told the first author that they did not feel that the music video was helpful for the experience. Instead, they stated that the video was disturbing and in-coherency between the visual and auditory perspectives (e.g., panning of the drums with respect to the visuals) was distracting.

The groups of participants did not rate immersion in the four experiences differently because the interaction between spatialization and stimuli was found to be insignificant. This can be confirmed by visually inspecting Fig. 4.

4 GENERAL DISCUSSION & CONCLUSION

The primary goal of this study was to investigate the influence of changes in audio spatialization on the experience of immersion in audiovisual applications. The authors were interested in changes in immersion ratings that were caused by perceptual differences between the spatialization levels. Thus, the initial condition was that a perceptual difference must exist between the audio spatialization levels.

Envelopment was chosen as the perceptual attribute of interest based on literature and experience. Although the difference in envelopment between 2.1 and 5.1 reproduction is well-documented in the literature, little information is available on the perception of envelopment in 7.1.4 reproduction. Therefore, the experiment was divided into two parts: the first was focused on establishing the difference in envelopment among the three spatialization levels, and the second was targeted at assessing immersion. The program material and experimental setup were identical for both experiments.

The results from the experiment on envelopment showed that there was no statistically significant difference between 5.1 and 7.1.4 reproduction. The initial condition was true only for 2.1–5.1 and 2.1–7.1.4 reproduction pairs. Therefore, the conclusions drawn here are limited to the said pairs. A statistically significant difference between the immersion scores could not be established in the second experiment, despite a large difference in the feeling of envelopment. Hence, the hypothesis that more enveloping experiences are also more immersive could not be supported in this study.

The results suggest that improvement in audio spatialization and a subsequent change in envelopment does not guarantee a psychologically immersive experience. This is an important finding that suggests that spatial audio reproduction may not be the dominant factor for determining immersion in audiovisual experiences. The findings confirm the initial thoughts on the influence of system parameters on immersion expressed in [14].

The results from this study should be interpreted with caution because a limited number of stimuli were reproduced over three audio spatialization levels. The results from the experiment are limited to channel-based and object-based reproduction and do not apply to formats such as wave-field synthesis audio reproduction. The audio production landscape for object-based audio is changing rapidly. The results presented here may no longer be valid as production techniques and usage of spatial audio technologies change.

The visual content was an important factor in this study, and its influence should not be ignored. Presence of visuals can potentially change the perception of envelopment due to multi-modal interactions. Finally, three out of the four stimuli were excerpts from movies. The auditory characteristics of movie clips are considerably different from music. Thus, the results cannot be generalized to music content.

4.1 Directions for Future Research

Magnitude estimation allows for referencing free, quick, and easy-to-comprehend audiovisual assessments. Hence, in applications such as evaluation of spatial audio technologies in which there are no established references, magnitude estimation can be valuable for assessments. Benchmarking of technologies and products can be performed with ease because participants do not need to learn a new scale and because the freedom to choose the appropriate experimental design parameters offers crucial flexibility to the experimenter. Multiple stimulus methods for relative judgments are quite popular for audiovisual assessment. New paradigms can be explored in which the participants have the ability to switch among different stimuli and provide ratings following the ratio principle.

When running tests with inexperienced participants, it is beneficial to have semantic information for conducting useful comparisons between scores. Coupling of ratio data with verbal descriptors has been performed to yield category-ratio scales [24]. Labeled magnitude scale and labeled affective magnitude scale are examples of category-ratio scales. These scales have spaced labels (usually unequal spacing), and the task for the participants is to indicate their rating by determining the category of sensation and then fine-tuning their ratings. These scales should be explored further to ease the task for the participants.

A between-subjects design is not the most efficient design for conducting experiments with a higher number of independent variable levels. It is critical for the community to evaluate the impact of familiarity and repetitions on immersion to potentially reduce the complexity and need for resources to conduct experiments on immersion. Establishing new experimental paradigms and validating existing ones will be useful in setting the foundation for conducting research on immersion.

The work presented here is a first step toward understanding the role of system parameters on the experience of immersion. Future work should aim at determining the different factors that influence the experience of immersion and quantify their impact. Such knowledge will be critical for enabling and enhancing immersive experiences. Although the experiments here were based on a traditional domestic application, the central idea and methods can be applied to technology-mediated experiences in general.

5 ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement number 765911. The authors would like to thank Geoff Martin and Thomas Ottesen for helping with the facilities and experimental setup.

6 REFERENCES

[1] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application* (Wiley, Chichester, UK, 2006).

[2] N. Kaplanis, S. Bech, S. H. Jensen, and T. van Waterschoot, "Perception of Reverberation in Small Rooms: A Literature Study," in *Proceedings of the AES 55th International Conference: Spatial Audio* (2014 Aug.), paper P-3.

[3] N. Zacharov (Ed.), *Sensory Evaluation of Sound* (CRC Press, Boca Raton, FL, 2018).

[4] ITU-R, "Methods for Selecting and Describing Attributes and Terms in the Preparation of Subjective Testsï¿½, *Recommendation ITU-R BT.2399-0* (2017 Mar.).

[5] N. Zacharov and T. H. Pedersen, "Spatial Sound Attributes—Development of a Common Lexicon," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), paper 9436.

[6] C. Colomes, S. Le Bagousse, and M. Paquier, "Families of Sound Attributes for Assessment of Spatial Audio," presented at the *129th Convention of the Audio Engineering Society* (2010 Nov.), paper 8306.

[7] S. Ko, E. Oh, S. H. Park, and H. Shim, "Perceptual Evaluation of Spatial Audio Quality," presented at the *129th Convention of the Audio Engineering Society* (2010 Nov.), paper 8300.

[8] J. Francombe, T. Brookes, and R. Mason, "Evaluation of Spatial Audio Reproduction Methods (Part 1): Elicitation of Perceptual Differences," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 198–211 (2017 Mar.). https://doi.org/10.17743/jaes.2016.0070.

[9] J. Francombe, T. Brookes, R. Mason, and J. Woodcock, "Evaluation of Spatial Audio Reproduction Methods (Part 2): Analysis of Listener Preference," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 212–225 (2017 Mar.). https://doi.org/10.17743/jaes.2016.0071.

[10] S. Agrawal, S. Bech, K. Bærentsen, K. de Moor, and S. Forchhammer, "Method for Subjective Assessment of Immersion in Audiovisual Experiences," *J. Audio Eng. Soc.*, vol. 69, no. 9, pp. 656–671 (2021 Sep.). https://doi.org/10.17743/jaes.2021.0013.

[11] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.*, vol. 50, no. 9, pp. 651–666 (2002 Sep.).

[12] T. H. Pedersen and N. Zacharov, "The Development of a Sound Wheel for Reproduced Sound," presented at the *138th Convention of the Audio Engineering Society* (2015 May), paper 9310.

[13] S. George, S. Zielinski, F. Rumsey, and S. Bech, "Evaluating the Sensation of Envelopment Arising From 5-Channel Surround Sound Recordings," presented at the *124th Convention of the Audio Engineering Society* (2008 May), paper 7382.

[14] S. Agrawal, A. Simon, S. Bech, K. Bærentsen, and S. Forchhammer, "Defining Immersion: Literature Review and Implications for Research on Audiovisual Experiences," *J. Audio Eng. Soc.*, vol. 68, no. 6, pp. 404–417 (2020 Jun.). http://doi.org/10.17743/jaes.2020.0039.

[15] S. Agrawal and S. Bech, "Immersion in Audiovisual Experiences," in M. Geronazzo and S. Serafin (Eds.), *Sonic Interactions in Virtual Environments*, pp. 341–374 (Springer, Cham, Switzerland, 2022).

[16] C. Zhang, "The Why, What, and How of Immersive Experience," *IEEE Access*, vol. 8, pp. 90878–90888 (2020 May). http://doi.org/10.1109/access.2020.2993646.

[17] IEC, "Sound System Equipment - Part 13: Listening Tests on Loudspeakers," *International Standard 60268-13* (1998 Mar.).

[18] Dolby Laboratories, "Dolby Atmos® Home Theater Installation Guidelines," https://www.dolby.com/siteassets/technologies/dolby-atmos/atmos-installation-guidelines-121318_r3.1.pdf (2018 Dec.).

[19] ITU-R, "General Viewing Conditions for Subjective Assessment of Quality of SDTV and HDTV Television Pictures on Flat Panel Displays," *Recommendation ITU-R BT.2022* (2012 Aug.).

[20] A. Brandmeyer, D. Darcy, L. Lu, et al., "Use of the Magnitude Estimation Technique in Reference-Free Assessments of Spatial Audio Technology," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), paper 10273.

[21] H. R. Moskowitz, "Magnitude Estimation: Notes on What, How, When, and Why to Use It," *J. Food Qual.*, vol. 1, no. 3, pp. 195–227 (1977 Oct.). https://doi.org/10.1111/j.1745-4557.1977.tb00942.x.

[22] ISO, "Sensory Analysis — Methodology — Magnitude Estimation Method," *Standard 11056* (2021 May).

[23] ASTM, "Standard Test Method for Unipolar Magnitude Estimation of Sensory Attributes," *Standard E1697-05* (2020 Feb.).

[24] H. T. Lawless and H. Heymann, *Sensory Evaluation of Food: Principles and Practices* (Springer, New York, NY, 2010), 2nd ed. http://doi.org/10.1007/978-1-4419-6488-5.

[25] J. Lim, "Hedonic Scaling: A Review of Methods and Theory," *Food Qual. Prefer.*, vol. 22, no. 8, pp. 733–747 (2011 Dec.). https://doi.org/10.1016/j.foodqual.2011.05.008.

[26] S. S. Stevens, "The Surprising Simplicity of Sensory Metrics," *Am. Psychol.*, vol. 17, no. 1, pp. 29–39 (1962 Jan.). https://doi.org/10.1037/h0045795.

[27] H. T. Lawless, "Logarithmic Transformation of Magnitude Estimation Data and Comparisons of Scaling Methods," *J. Sens. Stud.*, vol. 4, no. 2, pp. 75–86 (1989 Sep.). https://doi.org/10.1111/j.1745-459X.1989.tb00459.x.

A.1 Magnitude Estimation

The background for magnitude estimation, its advantages, and drawbacks of the method are presented in this appendix.

1 BACKGROUND

The technique of magnitude estimation was derived from the idea of cross-modality matching, a general approach to ratio scaling. The task in cross-modality–matching studies is to match the perceived intensity of a stimulus in one domain (e.g., the loudness of an audio clip) to the perceived intensity of another stimulus in another domain (e.g., the brightness of light) by changing physical parameters. One of these is set by the experimenter, and the participants are asked to adjust the physical parameters of the other such that the perception of the two is equivalent. Cross-modality matching allows for representation of the "intensity of a sensation in physical units" [24]. For example, the perceived bitterness of a beverage can be represented in nits (unit for brightness).

The idea of asking participants to assign numbers following the ratio principle appeared long after cross-modal matching [24]. A series of studies conducted on ratio scaling approaches lead to the so-called power law of the form:

$$R = k(I)^n \; ; \quad or \quad log\, R = n(log\, I) + log\, k, \tag{3}$$

where R is the magnitude estimate, I is measured physical intensity, n is an exponent, and k is the constant of proportionality. Here, n determines the perceived intensity as a function of physical intensity. For example, if the value of n is less than one, the magnitude estimates increase slower than the physical intensity [21]. Initially, this equation was used to determine dose-response relationships. Nevertheless, the concept of magnitude estimation can also be applied for determining relative intensities (if A is rated 100 and B is rated 25, A is perceived four times as strong as B) and calibration of scales with verbal descriptors [24].

2 ADVANTAGES

Magnitude estimation provides several important advantages over traditional methods for attribute assessment. Foremost, it allows for drawing inferences among perceptions in ratio terms [25]. Comparisons between perceptions

can be made while considering physical parameters, e.g., increasing the number of audio channels and loudspeakers for audio reproduction improves the quality of experience by 50%. Magnitude estimation can be used for unidimensional and bidimensional scaling. Unidimensional attributes, such as boomy (timbral attribute), for which zero represents the absence of the attribute, and bidimensional concepts, such as liking/disliking, can both be evaluated.

The simplicity of the method makes it effective for the evaluation of perceptual attributes (e.g., envelopment) and for hedonic/affective concepts, such as pleasantness and liking [25]. It is useful for evaluating attributes in which limiting the upper end of the scale restricts the ability to differentiate among extremely strong sensations [24]. For instance, annoyance caused by audio distortions is likely to be rated toward the higher end of the scale. Magnitude estimation is open-ended on the higher side and would allow for better differentiation between annoyance due to stimuli with severe distortions. Magnitude estimation has proved effective in revealing the differences between stimuli just as well as (and sometimes better than) category scaling [25].

3 DRAWBACKS AND LIMITATIONS

The method of magnitude estimation is promising for perceptual and holistic audiovisual evaluations. Nonetheless, the method has several drawbacks that must be considered for designing experiments and interpreting the results. Magnitude estimation lacks efficiency for determining small differences and evaluations around detection thresholds [22]. Contextual factors influence the ratings obtained from magnitude estimation and cannot be avoided by modulus normalization or modulus equalization [21]. For instance, if a set of highly enveloping multi-channel stimuli (e.g., 22.2) are used in a stimulus set with monochannel stimuli that are not enveloping, a contrast effect may be observed in which the multi-channel stimuli are relatively up-rated [21]. Hence, it is advisable to use stimuli of different intensities so that the range can be varied.

In the case of free standard and modulus (as in this experiment), direct comparisons between scores cannot be made without appropriate rescaling [25]. This complicates data analysis in comparison to other popular audiovisual evaluation methods. Readers will notice that the magnitude estimate between spatialization levels or stimuli are not discussed in this paper. This is because of the contextual factors and the question whether "magnitude estimation [is] a ratio scale or simply a scale with ratio instructions?" [24]. Initially, the numbers were taken at face value to conclude that because x is rated twice as y, x is perceived to be twice as strong as y (on the attribute under consideration). However, it is known that assigning numbers is a two-step process comprised of "the psychophysical transformation of energy into conscious sensation and the application of numbers to those sensations" [24]. The step of generating numbers is prone to biases and cannot be used to conclude on ratios and proportions at this point [24].

The round number tendency (explained in SEC. 2) is a common bias observed in magnitude estimation studies. Another bias that is difficult to fix is one in which the participants limit themselves to a certain range of numbers, even when permitted to use an infinite range of numbers [21]. They refrain from using their highest and lowest number in the range. The true boundaries cannot be determined, and there is no known remedy for this issue. A similar issue can be observed when a "categorizing scaler" is encountered [21]. Some participants use a fixed range of numbers and specific numbers within that range. For example, an individual using a 100–500 range might limit their usage to increments of 100. There is not much that can be done because the participant has limited themselves to a convenient set of intervals and treats the chosen numbers as category markers [21].

Inexperienced assessors do not have any known issues with magnitude estimation, but occasionally, they may assign numbers that are completely out of bounds (in comparison to their other ratings) when presented with very high or low intensity stimuli [21]. Such numbers convey extremes, and the numerical values must be interpreted with caution [21].

A.2 Analysis of Magnitude Estimation Data

The data had to be pre-processed for performing ANOVA. The steps for preparing the collected data for analysis are discussed below.

1) Rescaling method: It was important to rescale the data to a common range before performing statistical analysis because the participants were permitted to choose their own number range [24]. Total rescaling was used to rescale the data from this experiment. The reasoning behind total rescaling is that because all participants grade the same stimulus set only once, the total magnitude for the stimuli should be the same [22].

2) Total rescaling: Geometric mean for each participant was calculated (across all four stimuli at the three audio spatialization levels) as the first step of total rescaling. To have a common scale range, it was decided to make the geometric mean of all participants equal to 100.[12] A rescaling factor was calculated for each participant by constructing a ratio of 100 and the geometric mean of that particular participant. The scores for each participant were multiplied by their rescaling factor to force the geometric means for all participants to be equal to 100. The result of total rescaling followed by log transformation is shown in Fig. 5.

It can be seen that most participants found 2.1 audio spatialization to be less enveloping than the others. It is difficult to assess the difference between 5.1 and 7.1.4 audio spatialization levels from the figure. A few participants (E and H in particular) used a much smaller range of numbers to report envelopment. The authors suspect that a flawed understanding of the number generation task or a lack of substantial difference between the perceived magnitudes of

[12] Usually, the grand geometric mean is chosen for rescaling. However, any positive number can be picked because it is used to construct the rescaling factor for each participant [24]. Here, 100 was chosen because log transforming (base 10) the data would yield numbers that are not too small for reading convenience.

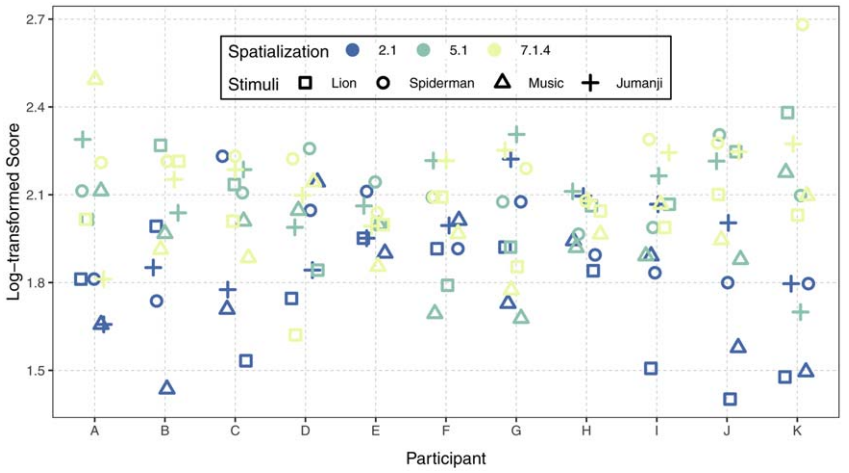


Fig. 5. Rescaled scores for all participants post–logarithmic transformation.

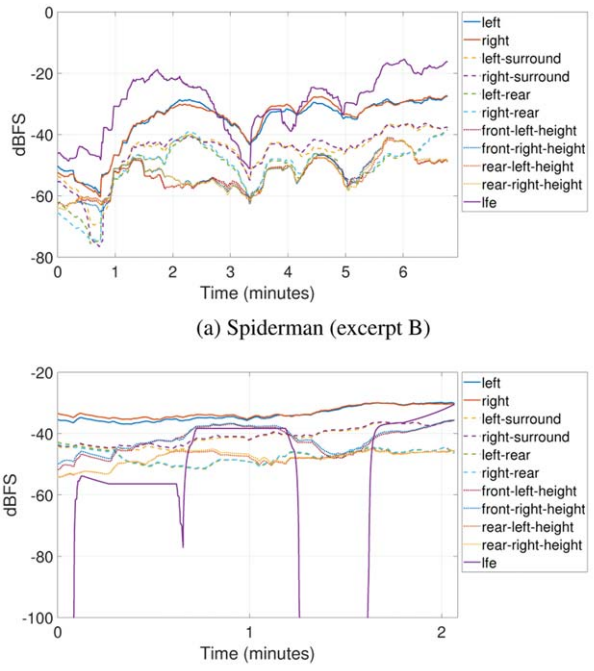
sound envelopment may have caused this. Given a lack of strong reasoning for excluding participants or data points (e.g., highest score by participant K), all collected data was included in further analysis.

3) Logarithmic transformation: The participants were not permitted to use zero or negative numbers in this test. Thus, the scale was bounded by zero and open-ended at the top (goes to infinity in theory). This often causes the data to be positively skewed with an approximate log-normal distribution. The geometric mean is a more appropriate measure of the central tendency as opposed to the arithmetic mean for log normally distributed data [26]. A multiplicative model corresponds to an additive model on the log scale. Therefore, the data can be log-transformed for conducting ANOVA because the geometric mean is the antilog of the arithmetic means of the logs.

Stevens [26] noted that the variability in magnitude estimates increases with the level of the stimulus. Log transformation helps to make the variances homogeneous and satisfy the assumptions for conducting statistical tests (when applicable). There are other reasons for log transforming magnitude estimation data: a) rescaling the data is easier after the transformation as shown in *ISO Standard 11056* [22]; b) power functions are linear in log–log coordinates and easier to interpret for response relationships. Please refer to [27] for a discussion on log transformation of magnitude estimation data.

A.3 Audio Level Plots

Stimuli were recorded, and the level was plotted to assess the level differences between front channels (left and right in this case because of the center being phantom) and the content to be reproduced by the elevation speakers. The excerpt from *Spider-Man: Into the Spider-Verse* (excerpt B) is shown to illustrate the level difference of 20–25 dB in Fig. 6. It was chosen because it was the highest-rated for envelopment and also the most spatialized stimulus, For comparison, the example excerpt that had exaggerated envelopment for demonstrating envelopment is also shown. The example excerpt has sections in which the level difference between the height channels and front channels is only a few decibels. Please note that the plots are unweighted and have been smoothed for intelligibility.



(a) Spiderman (excerpt B)

(b) Kinsetsu (abstract excerpt for envelopment experiment)

Fig. 6. Audio level plots (unweighted) for the (a) *Spider-Man: Into the Spider-Verse* excerpt (excerpt B) and (b) *Kinsetsu: Textures from Planet 9* excerpt (used for demonstrating degrees of envelopment). Please note that the lines have been smoothed for intelligibility.

THE AUTHORS


Sarvesh R. Agrawal


Søren Bech


Katrien De Moor


Søren Forchhammer

Sarvesh Agrawal was born and raised in Mumbai, India. He moved to the United States in 2014 and received a B.S. in audio production with a minor in entertainment technology from Middle Tennessee State University (MTSU) in 2016. Following a brief stay at MSE Audio Group in Kansas, he moved to New York where he attained a graduate degree in architectural acoustics from Rensselaer Polytechnic Institute (RPI) in 2018. Sarvesh joined Bang & Olufsen in 2018 as a research fellow and has been pursuing a Ph.D. from the Department of Photonics Engineering at the Technical University of Denmark (DTU) as an early stage Marie Curie fellow. Psychoacoustics, perceptual evaluation of sound, and subjective assessment of multimodal experiences are his primary research interests. Currently, Sarvesh is the Global Product Manager for Bang & Olufsen's Beolab portfolio.

•

Søren Bech received an M.Sc. and Ph.D. from the Department of Acoustic Technology (AT) of the Technical University of Denmark. From 1982–1992 he was a research fellow at AT studying perception and evaluation of reproduced sound in small rooms. In 1992 he joined Bang & Olufsen where he is Director of Research. In 2011 he was appointed Professor in Audio Perception at Aalborg University, and he is Adjunct Professor at Surrey University (GB) and McGill University (CAN). He is a Fellow of the Acoustical Society of America and Audio Engineering Society (AES). He is past Governor and Vice-President of the AES and now serves as associate technical editor of *JAES*. He has been vice-chair of the International Telecommunication Union working group 10/3. In 2006 he and Dr. Zacharov published the book *Perceptual Audio Evaluation – Theory, Method and Application* (Wiley). His research interests include psychoacoustics and, in particular, human perception of reproduced sound in small-sized and medium-sized rooms. His other interests include experimental procedures and statistical analysis of data from sensory analysis of audio and video quality.

•

Katrien De Moor is an associate professor at the Department of Information Security and Communication Technology at Norwegian University of Science and Technology (NTNU), mainly focusing on socio-technical approaches in information and communications technology (ICT) research. She is co–Editor-in-Chief of the multidisciplinary journal *Quality and User Experience* (Springer) and affiliated researcher at the Research Group for Media, Innovation and Communication Technologies (Ghent University, Belgium). She received her Ph.D. degree in Social Sciences from Ghent University (2012) with a thesis on bridging gaps in Quality of Experience research and its challenges. She has been a visiting researcher at several institutions, including the University of Eindhoven, TU Berlin, and NTNU (as Marie Curie Postdoctoral Fellow). Katrien is passionate about user research and user involvement in user-centric innovation processes. Her main research interests and activities are centered around human/user experiences with technology (Quality of Experience, User Experience, User Engagement, immersive experiences, etc.), related methodological challenges (ecological validity, user diversity, etc.), and ethical implications (e.g., data privacy, human agency, power dynamics in design processes, ecological footprint of ICT). She has published her work in a range of international, peer-reviewed journals, conferences, and books; acts as a reviewer for several international journals; and has served several program committees of international conferences and workshops. She is one of the 20 founding members of the Young Academy of Norway.

•

Søren Forchhammer received M.Sc. and Ph.D. degrees from the Technical University of Denmark, Lyngby. Currently, he is a Professor with DTU Fotonik, Technical University of Denmark, where he has been since 1988. He is Head of the Coding and Visual Communication Group at DTU Fotonik. He is Coordinator of the European Union (EU) Marie Skłodowska-Curie Actions (MSCA) Innovative Training Networks (ITN) RealVision. He is flagship lead in the Danish National Research Foundation (DNRF) Center of Excellence (CoE) Silicon Photonics for Optical Communications (SPOC). His research interests include source coding, image and video coding, processing of image and video, processing for image displays, quality of coded multimedia data, multi-camera and light field images and video, 2D information theory, and visual communications.