PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

ABS-0501

Influence of the presence of congruent visual media on spatial auditory fidelity

Xuan LU;[1] Sungyoung KIM;[1] Miriam A. KOLAR;[2] Doyuen KO[3]

[1] Rochester Institute of Technology, Rochester, NY 14623, USA

[2] Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA 94305, USA

[3] The Mike Curb College of Entertainment and Music Business, Belmont University, Nashville, TN 37212, USA

ABSTRACT

In this research, we investigated the influence of congruent visual media on the recognition of spatial audio attributes in a multimodal virtual reality context. For the immersive auditory data, we recorded solo piano music using three multichannel microphone techniques: a 5-channel conventional surround microphone technique, and a 7-channel and a 9-channel immersive microphone array that both include height-channel microphones. In total, 38 normal-hearing participants were invited to the experiment and separated into two groups for presenting immersive audio with and without congruent visual media. The visual data was collected using a 360-degree camera and projected to a 250-degree wraparound screen for immersive visual rendering. Participants were asked to rate the timbral and spatial attributes for each of the three immersive auditory presentations (solo piano music via the three multichannel recording techniques), and to provide an overall preference rating by using a continuous quality scale (CQS). Experiment results show that the presentation of congruent visual media modulates reported perceptions of auditory attributes from simultaneously presented immersive audio. The experiment demonstrated that the presence of a realistic visual image congruent to the presented auditory material made listeners require higher spatial auditory fidelity in the immersive multimodal context. Consequently, this study's findings can be used to inform immersive and holistic audiovisual and interaction design.

Keywords: audiovisual congruence, immersion, auralization, multimodality, multichannel audio

1. INTRODUCTION

For this study, the authors designed an experiment to study audiovisual congruence in a context of immersive media representation. Specifically, our interest was to investigate the influence of realistic congruent visual media on perceived quality of reproduced soundfields. In the sound-recording literature, many previous sound-quality evaluations have been conducted without the presentation of equivalent visual information. Considering that the music industry has been producing audio-only formats for consumers, those evaluations are not only reasonable but also legitimate. However, the larger media industry has been adapting visual information as content pertinent to musical performances; for example, via YouTube and social media platforms. We wanted to study the dynamic influence of audiovisual congruence on the perception and recognition of specific information about reproduced music.

Previous studies have demonstrated such a dynamic influence for various applications. For example, Bartel and Chon observed a perceptual asymmetry that audio quality had a significant unilateral effect on perceived audio and video quality [1]. Kitagawa et al. investigated audiovisual interactions and found an asymmetric relation: the auditory aftereffect occurs from adaptation to visual motion in depth, but not a visual aftereffect from auditory intensity change [2]. On the other hand, Maempel and Horn [3, 4] suggested methodological criteria for audiovisual experiments, and reported that most unimodal features were rated similarly across environments under both optical

[1] liebelux@gmail.com; sxkiee@rit.edu
[2] kolar@ccrma.stanford.edu
[3] doyuen.ko@belmont.edu

and/or acoustic conditions [4]. A multimodal presentation of audiovisual stimuli also influences a participant's cognitive process. Nittrower and Lowenstein [5] investigated how the lack or presence of visual stimuli influences the memory function for different audio stimuli, and concluded that additional presentations indeed help in the memory process.

One of the limitations of these previous studies lies in their use of 2-channel reproduction of audio information. While a 2-channel audio system provides a successful illusion of a target auditory image, its spatial resolution is limited to the frontal region without allowing listeners to experience immersive soundfields. Chabot and Braasch created a system that puts an equal emphasis on the visual and audio information, utilizing a panoramic display and high-resolution spatial soundfield with multiple loudspeakers [6]. With this kind of system, we can further investigate audiovisual interactions, and particularly the influence of coherent presentations on user experience, especially for a near-immersive presentation of audiovisual information.

2. AIMS

With the advent of interactive media formats including virtual and extended reality (VR and XR), audiovisual congruence has become an important topic for sound recording and reproduction. It is evident that classical music representation — due to its emphasis on fidelity — is a prime target for audiovisual congruence, anticipating that its audiences will seek the most holistically satisfying experience. To provide the appropriate multimedia congruence, can we simply combine immersive auditory and visual information? Or do we need to better understand underlying interactions between auditory and visual information in order to specify the parameters required? This study explored the particular audiovisual interaction scenario of classical music performance for the immersive representation of a concert hall music experience, summarized by the following research question:

Would the presence of realistic and congruent visual information affect perceived fidelity of timbral and spatial attributes of a three-dimensional (3D) reproduced soundfield?

3. EXPERIMENT

For the experiment stimuli, solo classical piano music was recorded in the McAfee Concert Hall of Belmont University in Nashville (TN, USA), using three multichannel microphone techniques: a 5-channel conventional surround microphone technique (5ch) [7], and a 7-channel (7ch) and a 9-channel (9ch) microphone array that both include height-channel microphones [8] (Fig. 1, top). Three microphone techniques were mixed to represent the engineers' assessments of the best tonal and spatial quality when reproduced through 0+5+0, 2+5+0, and 4+5+0 formats respectively. This 'U+M+B' channel format indicates the number of Upper, Middle and Bottom loudspeakers (ITU-R BS.2051-0 [9]).

Thirty-eight self-identified normal-hearing participants from Rochester Institute of Technology (RIT) were invited to the experiment, and randomly separated into two groups for presenting immersive audio with and without visual immersion. The visual data was collected using a 360-degree camera (Ricoh Theta V) and projected to a 250-degree wraparound screen for immersive visual presentation (Fig. 1, bottom).

Three timbral attributes (*Clarity, Brightness, Low Frequency Definition (LFD)*) and three spatial attributes (*Acoustic Energy Spread (AES), Depth, Immersion*) as well as *Overall Preference* were rated by participants using a continuous quality scale (CQS). The participants were also encouraged to report perceived auditory characteristics using their own descriptive adjectives if needed.

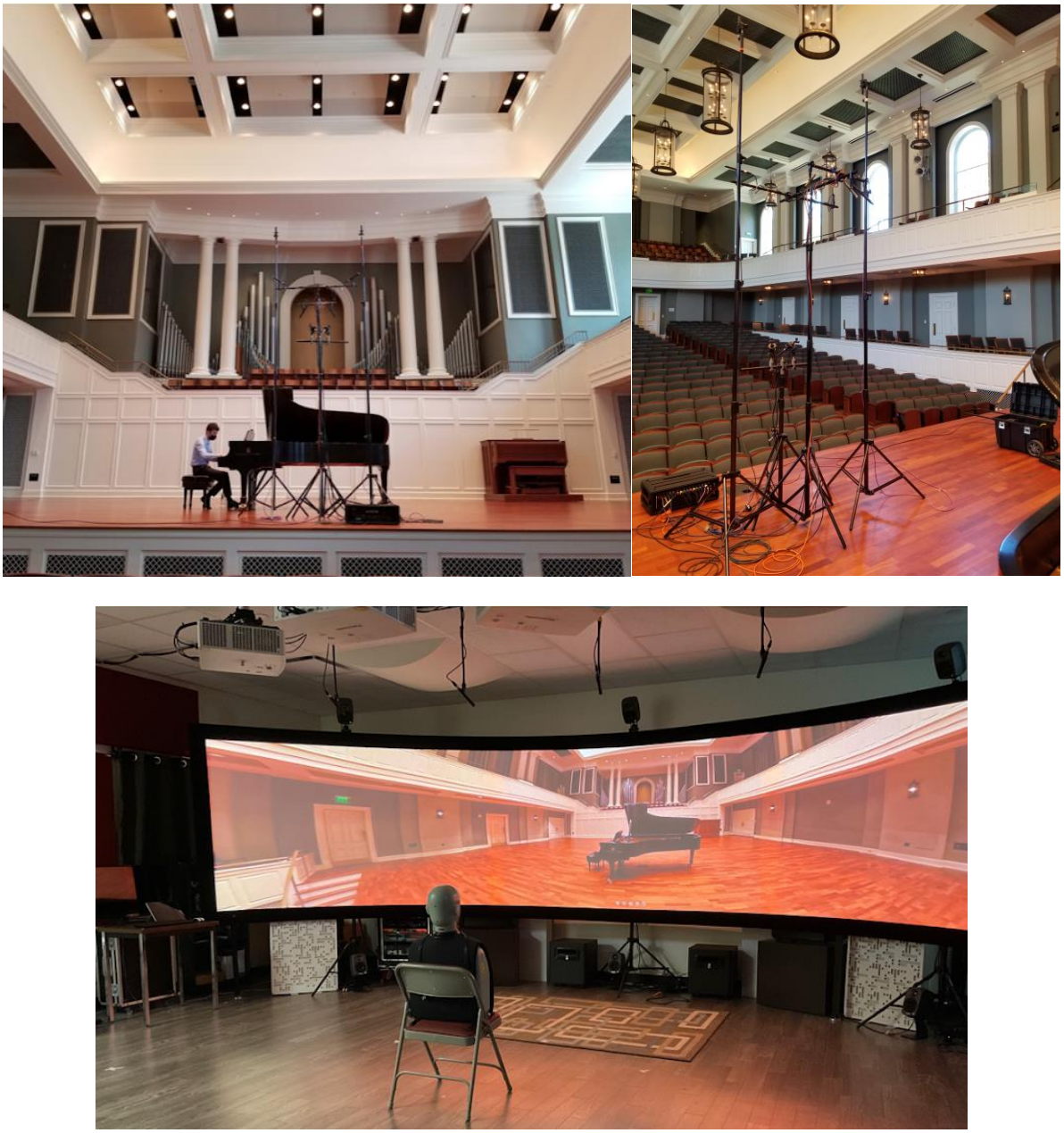

Figure 1. Piano recording in McAfee Concert Hall, Belmont University, using three microphone arrays (top); visual reproduction of the recording via a 250-degree wraparound screen, with 24 loudspeakers for 3D immersive audio reproduction in the RIT immersive experience laboratory (bottom).

4. RESULTS AND DISCUSSION

4.1 Results

A two-way repeated-measures analysis of variance (ANOVA) shows a significant difference in the interaction between the auditory stimuli (3D audio presentation of three different recording techniques) and the condition (the presentation or absence of congruent visual information): $F(2, 72) = 5.4962$, $p < .01$. Subsequent analyses show that the three auditory stimuli were perceived significantly different from each other only with the congruent visual information ($F(2,36) = 5.1988$, $p < .05$) but not without the visual information ($F(2,36) = 2.3004$, $p = .018$). This indicates that audio-visual interaction in this experiment influenced listeners' auditory cognition. In other words, listeners could not differentiate between the three different audio-only stimuli (i.e., the recordings using different microphone arrays) without the presence of congruent visual information.

A significant difference was also found between the two conditions at 5ch ($F(1,36)= 5.4237$, p

< *.05*), whereas no significant difference was observed for either 7ch and 9ch stimuli between conditions. Subsequently, we analyzed five attribute ratings. For the 5ch-reproduced piano sound (audio alone, without visual media), listeners gave higher values in all five attributes. In contrast, for the presentation of piano sound with congruent visual information, the mean score of 5ch attributes (Mean = 3.16, SD = 0.88) was significantly lower than without visual information (Mean = 3.52, SD = 0.83). Specifically, listeners recognized the 5ch-reproduced piano music as more "immersive, spread, and deep," *without* the corresponding visual image. It is important to note that ratings for the 5ch-reproduced piano music alone were comparable with those of the two other formats without visual media; in other words, the presence of congruent visual media significantly changed listeners' auditory spatial recognition.

To better illustrate the previous inferential statistics from the ANOVA test, we generated descriptive spider plots (Fig. 2). These figures visually contrast the difference between with and without the congruent visual presentation. Without the visual information, a significant difference was observed only in *Immersion*: $F(2, 36)= 3.7005$, $p < .05$; whereas with visual information, a significant difference was observed in *Clarity* ($F(2, 36)= 2.6611$, $p < .10$), *AES* (*Acoustic Energy Spread*) ($F(2, 36)= 6.9575$, $p < .01$), *Depth* ($F(2, 36)= 3.3429$, $p < .05$) and *Immersion* ($F(2, 36)= 2.7820$, $p < .10$).

4.2 Discussion

Study results showed that the presence of realistic and congruent, near-immersive visual information modulates perceived magnitudes of auditory attributes for immersive audio reproduction of multichannel-recorded concert piano music. For example, listeners reported that the 5-channel reproduced piano sound brought enhanced impression over its counterparts (7- and 9-channel reproduction) including *Overall Preference* and *Brightness* without corresponding visual information. However, mean ratings of 9-channel reproduction are significantly higher than for those of the 5-channel reproduction with congruent visual presentation. This result may indicate that the visual information adjusted listeners' satisfactory thresholds for auditory quality. Of interest is that these preference changes mostly happened for spatial attributes, but not for timbral attributes (*Brightness* and *LFD*). Thus, the presence of congruent visual media might have set a new quality standard for the listeners to assess required immersion, and the 5-channel reproduced soundfield did not meet this adjusted requirement. One possible alternate explanation is that this result might relate with the idiosyncratic difference of the two specific listener groups in the study. It was possible that participants in the listener group who were not presented with the congruent visual media were, for some reason, able to distinguish spatial attributes between the 5ch and the others (7ch and 9ch) while those in the other group were not. Therefore, to explore how this finding generalizes across more listeners, we plan to conduct a follow-up experiment with listener-groups from diverse backgrounds (e.g., different types of musical training, audio engineering, and 3D sound experience [10], etc.).

5. CONCLUSIONS

In this multimodal virtual reality study, perceived magnitudes of three auditory spatial attributes (*Acoustic Energy Spread, Immersion* and *Depth*) were shown to be influenced by the presence of the realistic and congruent, immersive presentation of the visual image of a solo piano. Study results demonstrate that the presentation of congruent visual media along with the multichannel recorded audio made listeners require higher spatial auditory fidelity in this immersive multimodal context. This finding suggests that for optimal audiovisual presentation of classical music performance, engineers and producers should consider the influence of audio recording techniques in terms of spatial auditory reproduction in order to create holistic immersive audiovisual experiences. Further research could support the proposed extensibility of study findings to other virtual reality contexts.

ACKNOWLEDGEMENTS

This work was supported by the National Endowment for the Humanities (NEH) research and development award, "Digital Preservation and Access to Aural Heritage Via a Scalable, Extensible Method" (Award No. PR-263931-19).

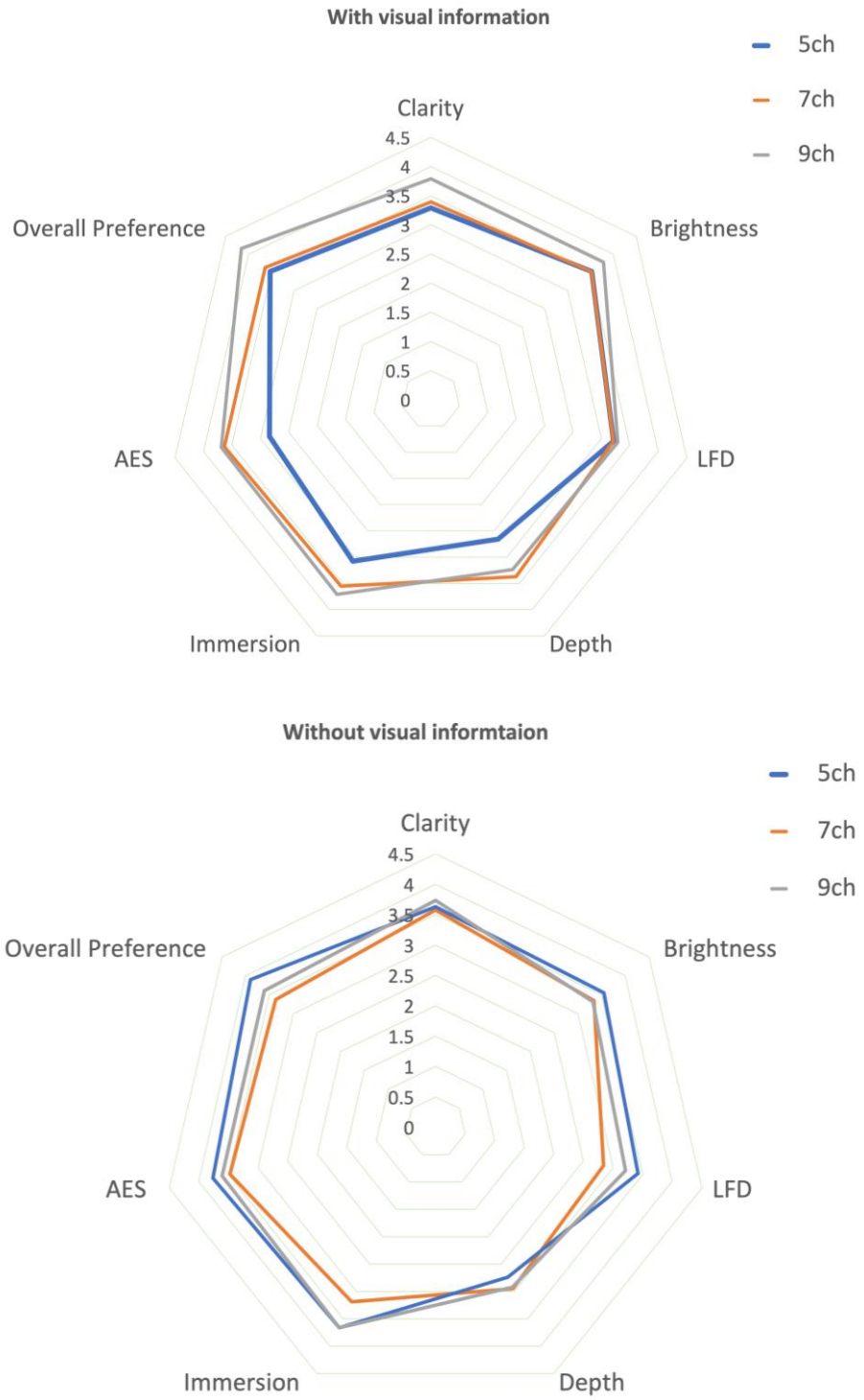

Figure 2. Spider plots illustrating the ratings for the two conditions (with (top) and without (bottom) the visual image). AES stands for Acoustic Energy Spread, and LFD for Low Frequency Definition. The presence of visual information can be seen to have modulated the perceived magnitude of three attributes: Clarity, AES and Depth.

REFERENCES

1. Bartel N, Chon SH. Watching on the Small Screen: The Relationship Between the Perception of Audio and Video Resolutions, Proc 152th AES Conv., May 2022; Paper 10607.
2. Kitagawa N, Ichihara S. Hearing visual motion in depth. Nature. 416; 2022; 172-174.
3. Maempel H. Apples and oranges: a methodological framework for basic research into audiovisual perception. In S. Hohmaier (Ed.), Jahrbuch 2016 des Staatlichen Instituts für Musikforschung Preußischer Kulturbesitz (pp. 361–377). Mainz et al.: Schott. 2019.
4. Maempel HJ, Horn M. Audiovisual perception of real and virtual rooms. Journal of Virtual Reality and Broadcasting, 2017; 14(5).
5. Nittrouer S, Lowenstein J. H. BeyondRecognition: Visual Contributions to VerbalWorking Memory. Journal of Speech, Language, and Hearing Research, 2022; 65(1), pp. 253–273.
6. Chabot S, Braasch J. An Immersive Virtual Environment for Congruent Audio-Visual Spatialized Data Sonifications. ICAD2017; State College, PA, USA; June 20-23, 2017.
7. Fukada A. A challenge in multichannel music recording. Proc AES 19th Conf; Germany, June 2001.
8. Lu X, Kim S, Kolar M, Ko D. Perceptual evaluation of a new, portable three-dimensional recording technique: “W-Ambisonics”. Proc 151st AES Conv. Virtual, Oct 2021; E-Brief 652.
9. ITU-R. 2051-2. Advanced sound system for programme production (ITU, 2018).
10. Howie W, Martin D, Kim S, Kamekawa K, King R. Effect of Skill Level on Listener Performance in 3D Audio Evaluation. J Audio Eng Soc. 2020; 68(9):628-637.