WAVE-FIELD SYNTHESIS AND FORCED PERSPECTIVE FOR CONGRUENT AUDIOVISUAL DISPLAY

Samuel Chabot

Rensselaer Polytechnic Institute
110 8th St
Troy, NY, USA
chabos3@rpi.edu

Jonas Braasch

Rensselaer Polytechnic Institute
110 8th St
Troy, NY, USA
braasj@rpi.edu

ABSTRACT

The spatial audio technique wave field synthesis has the ability to perceptually locate sound sources in front of the loudspeaker array. This can create a challenge for pairing congruent visuals when using a traditional projection screen, such as in many immersive environments. If the sound is placed within the environment beyond the boundary of the screen, how can visuals be displayed such that they appear congruent with their corresponding sound? For this, the technique known as forced perspective, where objects are made to appear of different scales or positions than they are in reality, is utilized. This allows for content to be created to appear as if it is spilling into the environment, and thus remain congruent with its respective sound. Assessments to determine perceptions and the effectiveness of audiovisual sources paired in this way are being planned and will be underway. The technique described here can be utilized for various digital experiences, from artistic performances to 3D advertisements.

1. INTRODUCTION

Spatial audio perceptually positions sound sources around a listener using either headphones or loudspeaker arrays. The area continues to grow as companies like Apple introduce spatial mixes and processing into their streaming services and hardware. One such technique is wave field synthesis (WFS), which utilizes dense arrays of loudspeakers to recreate a continuous wavefront from discrete sources [1]. One strength of the technique is the ability to project "focused" sound sources in front of the array and into a space [2]. In this way, sound sources appear to emanate from a location between the listener and the array, rather than from beyond the array.

However, a challenge is introduced when attempting to pair the focused sound with congruent projected visuals. In immersive environments that place equal emphasis on audio and visual content, such as the CRAIVE-Lab at Rensselaer Polytechnic Institute, how can these visuals be presented to maintain congruence with sound perceived to appear in front of the projection screen?

[3] [4] Previous work has been done to study audiovisual congruence through wave-field synthesis and stereoscopic 3D projection. These thorough perceptual studies investigate perceived congruence with various audiovisual configurations. André et. al present participants with stereoscopic images and sound stimuli presented with various angular errors to determine the perceived spatial accuracy of the each combination. Moulin et. al performs a similar perceptual investigation by analyzing the depth dimension. Combinations of visual and aural stimuli at various distances are presented and perceived congruence is measured. This study finds that there is a range at which stimuli will be perceived as congruent. In neither of these studies do the authors yet investigate the pairing of visuals with focused sound sources. And while stereoscopic 3D projection can create three-dimensional visuals, they do require specialized projectors and additional hardware to be worn by the user (i.e. glasses). Can a similar effect be achieved using more traditional projection hardware without requiring users to wear glasses?

The optical technique of forced perspective provides one solution to this challenge. Objects are visually manipulated so as to appear of different scales or positions than they are in reality. Many examples of this technique exist in history: from flat cathedral ceilings painted to appear to have great depth with frescos [5], such as the Triumph of Saint Ignatius of Loyola, to a corridor which appears about four times larger than it actually is at the Palazzo Spada [6]. This technique can also be used to manipulate projected visual content such that it appears to be emerging from the screen in to the environment, and thus remain congruent with its respective focused sound.

The space in which this work is being developed and evaluated is the CRAIVE-Lab, seen in Figure 1, a human-scale immersive environment with a 10-by-12 meter floorspace. It features a nearly 360-degree, 4.3 m tall projection screen in a rectangular geometry with rounded corners and a permanent horizontal array of 128 loudspeakers. Two additional modular WFS loudspeaker arrays have been developed at the university that will also be used for evaluation. The three arrays feature anti-aliasing frequencies ranging from 700 Hz to about 5.8kHz, one of the highest of any system. The latter is achieved by staggering drivers such that their spacing is only 1.75 cm [7]. For detailed information on the environment, see [8]

2. METHOD

Spatialization of sound sources is achieved using Max and the IRCAM Spat5 library [9]. This library contains the necessary objects to configure a virtual loudspeaker array. Then, virtual sound sources can be placed in the environment relative to the array. These can be manipulated either exactly with defined messages

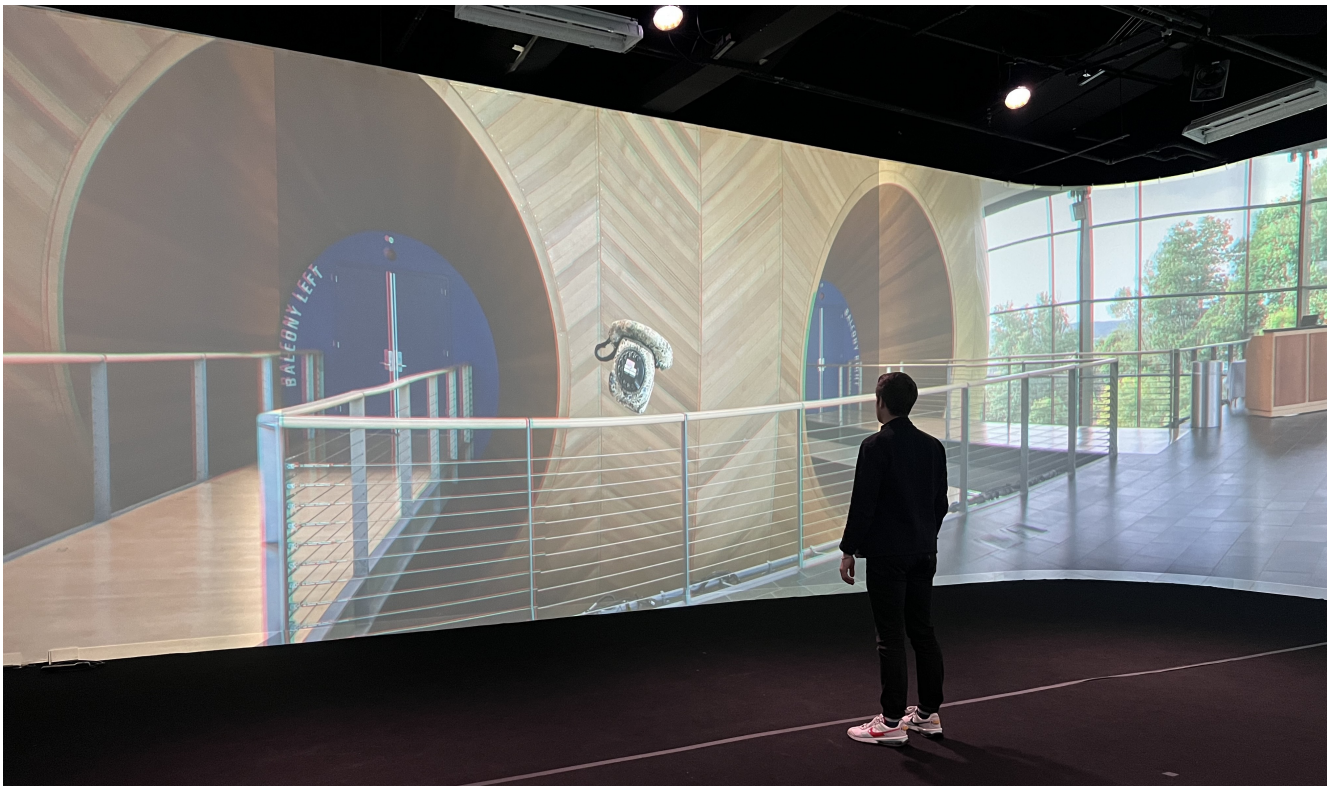

Figure 1: A participant in the CRAIVE-Lab, an immersive environment at Rensselaer used to project the focused sound sources and forced perspective visuals. Behind the acoustically-transparent screen is a dense loudspeaker array. The participant is asked to judge the perceived distance of the objects shown floating on the screen, in this case a telephone.

or approximately with a supported GUI. The contributions to the projected wave-front of the individual loudspeakers are then calculated by the objects and the resulting audio is output along the real-world loudspeaker array.

Perspectival adjustments are already a familiar manipulation in the environment: due to the irregular geometry of the lab, the field of view varies by azimuthal degree. Therefore, a perspectival transform is required to counteract distortion introduced to the projection of equirectangular imagery. This transformation, and the digitally-created forced perspective content, have their roots in computer graphics [10], [11] A 3D rendering software can be employed to create the forced perspective content. The projection screen is modeled in the virtual space. Content that should appear to be emerging from the screen should be modeled doing so in the virtual space. A camera is placed in the virtual space at the desired viewing location of the content. The captured content is then virtually back-projected on to the screen. This back-projection creates what appears to be a distorted image, but when projected on the screen in real-space and viewed from the intended location will match the designed "forced perspective."

3. ASSESSMENT

The lab has previously investigated work in this area. Richie evaluates the perceived distance of various objects projected on the immersive screen [12] These objects are shown in a stereoscopic view using anaglyph imagery and appear to be three-dimensional (a telephone is shown on the screen in Figure 1).

Auditory events are emitted both congruently and incongruently and participants are asked to report the perceived distances of the objects. The work investigated this perception of depth as a function of the interacting qualities of the visual and aural placements, as well as the qualities of the aural, such as the direct-to-reverberant ratio.

The fine-tuning and assessment of these forced perspective pairings is ongoing. Perceptual evaluations in which participants will be asked to report on the effectiveness of the congruently presented sources are planned. Assessments will seek to understand the technique from various perspectives: can participants ascertain congruent from a group of non-congruent presentations?; is the visual or aural effect more convincing given the presence of the other?; how precisely can participants position the focused source at the position of the visual projection?

Other considerations require analysis. The wave field synthesis method creates inhomogeneous sound fields with no sweet-spot, meaning listeners can move around a defined area without loss of fidelity. However, the use of focused sound sources does reduce the size of the listening area. Sources are only perceived spatially accurately when the sound source is between the listener and the array. The forced perspective technique is also performed necessarily with respect to a particular point in space. In what scenarios is the ability worth the trade-off? Can a game engine be employed to enable dynamic, real-time movement of the visual object?

This last solution is being investigated. Game engines are already employed in the environment for the production of walkable auralizations where users can explore the environment in real-time. Using a variety of tracking mechanisms available in the lab (e.g., network of Microsoft Kinects, spherical microphone array), a participant's precise location can be determined. This information could alter the forced projection visuals so as to remain updated with respect to participant location.

4. CONCLUSIONS

Wave field synthesis is an important loudspeaker-based spatial audio technique for creating inhomogeneous sound fields and focused sound sources projected in front of the array. However, in order to maintain audiovisual congruence with the use of a fixed traditional screen requires a technique such as forced perspective to create the illusion that the visual content is emerging from the screen. Planned assessments will show how effective the congruent pairing remains, and in what use cases the ability outweighs the trade-off of a defined viewing point in space. The successful deployment of focused sound sources congruently paired with forced perspective visuals lends itself to a number of future uses, ranging from 3D billboards and advertisements to augmented immersive "field trips" such as to industrial water treatment facilities, for experiential learning opportunities.

5. ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant IIS-1909229.

6. REFERENCES

[1] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of wave field synthesis revisited," in *Proc. 124th Audio Eng. Soc. Conv.*, Amsterdam, The Netherlands, 2008, Art no. 7358.

[2] H. Wierstorf, A. Raake, M. Geier, and S. Spors, "Perception of focused sources in wave field synthesis," *The Journal of the Audio Engineering Society*, vol. 61, 2013.

[3] S. Moulin, R. Nicol, L. Gros, and P. Mamassian, "Audiovisual spatial integration in distance dimension — when wave field synthesis meets stereoscopic-3d," in *2014 Audio Eng. Soc. Int. Conf.: Spatial Audio*, 8 2014, p. 12.

[4] C. R. André, Étienne Corteel, J.-J. Embrechts, J. G. Verly, and B. F. G. Katz, "Subjective evaluation of the audiovisual spatial congruence in the case of stereoscopic-3d video and wave field synthesis," *Int. J. Human-Comp. Studies*, vol. 72, no. 1, pp. 23–32, Jan. 2014.

[5] M. João Paulo Cabeleira, "Inácio Vieira: Optics and perspective as instruments towards a sensitive space," *Nexus Netw. J.*, vol. 13, no. 2, pp. 315–335, Jun. 2011.

[6] J. Wasserman, "Palazzo Spada," *Art Bulletin*, vol. 43, no. 1, pp. 58–63, Mar. 1961.

[7] J. Goebel, "The empac high-resolution modular loudspeaker array for wave field synthesis," in *Proceedings of the 23rd International Congress on Acoustics, integrating 4th EAA Euroregio 2019*, ser. Session: 02 D - Loudspeaker arrays and sound field control, Aachen, Germany, September 2019, pp. 1148–1155, iSSN 2226-7808 and 2415-1599, ISBN 978-3-939296-15-7.

[8] S. Chabot and J. Braasch, "Walkable auralizations for experiential learning in an immersive classroom," *The Journal of the Acoustical Society of America*, vol. 152, pp. 899–910, 8 2022.

[9] *Spatialisateur 5*. (version 5.1.5). IRCAM. Accessed: Aug. 3, 2020. [Online]. Available: https://forum.ircam.fr/projects/detail/spat/

[10] G. Wolberg, "Geometric transformation techniques for digital images: A survey," Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-390-88, Dec. 1988.

[11] C. A. Glasbey and K. V. Mardia, "A review of image-warping methods," *J. Appl. Statist.*, vol. 25, no. 2, pp. 155–171, Apr. 1998.

[12] S. Richie, "Audio and visual distance perception of familiar and unfamiliar objects using wave field synthesis and a stereoscopic display," Master's thesis, Rensselaer Polytechnic Institute, Troy, NY, USA, Aug. 2019.