Audio Engineering Society

Convention Paper 10236

Presented at the AES 159th Convention
2025 October 23–25, Long Beach, CA, USA

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

See Your Audio 3D: A Novel Optical Approach for Simultaneous Visualization of Sound Fields and 3D Objects

Kenji Ishikawa[1]

[1]*NTT, Inc.*

Correspondence should be addressed to Kenji Ishikawa (ke.ishikawa@ntt.com)

ABSTRACT

While sound transmits linguistic and musical information through its temporal variations, its spatial distribution encodes vital information related to three-dimensional perception and spatial audio representation. However, since sound is invisible to the human eye, grasping its spatial characteristics is challenging. This paper proposes a fully optical visualization method for spatial audio waves alongside the 3D geometry and texture of sound sources. The proposed method involves reconstructing a three-dimensional sound field from a high-speed video of projected sound fields measured using optical interferometry. The continuous rotation of the sound source enables the reconstruction of a volumetric sound field from a single video with a spatial resolution finer than 1 mm. The 3D surfaces and textures of the sound source are constructed from an ordinary video using a neural radiance field-based mesh generation technique. The volumetric sound field and 3D objects are integrated and visualized simultaneously in a 3D space, allowing us to visually ascertain where and how the sound is generated and propagates. The proposed method is demonstrated with three examples: sound radiation from a loudspeaker, a smartphone, and an ultrasonic transducer array.

1 Introduction

Sound is typically captured as a temporal waveform via microphones. However, for a comprehensive understanding of acoustic phenomena, merely recording the distance-traveled sound waveforms is often insufficient. Instead, we need to sense how sound propagates through real space and grasp the interactive relationship between the sound field and the 3D scene. Particularly in the development of acoustic devices and spatial audio playback systems, a precise understanding of the sound sources and objects involved in sound generation and propagation is the key to success in producing high-fidelity experiences.

Various methods have been studied to present the spatial characteristics of a sound field together with visual information. For instance, visualizing the position of a sound source by overlaying acoustic quantities recorded by microphone arrays or scanning probes onto camera images are used in various industrial applications [1, 2, 3, 4]. In addition, Inoue et al. proposed an augmented reality method to present data obtained by microphone probes via a head-mounted display [5, 6]. However, microphone-based methods often face practical limitations in spatial resolution. When using microphone arrays, the resolution is constrained by the number of microphones and the physical dimensions of the array, while scan-based approaches require long measurement times to achieve high spatial resolution.

As an alternative to the microphone-based measure-

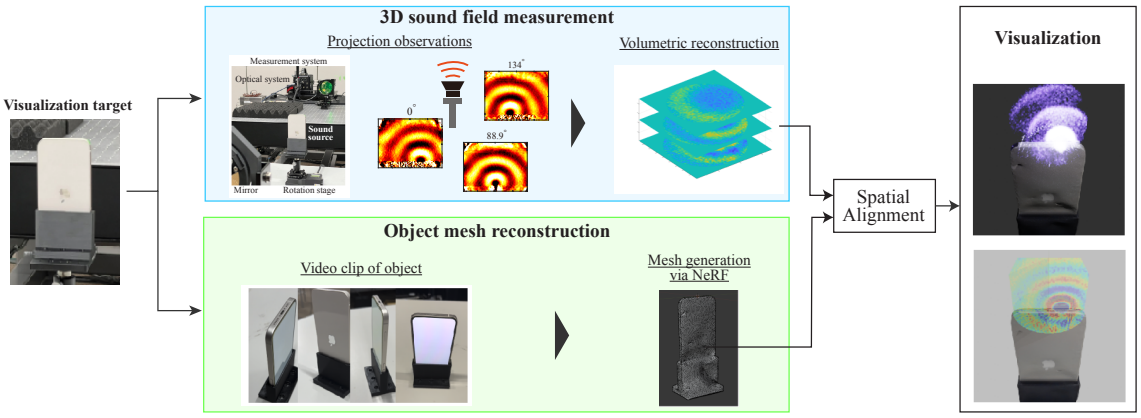


Fig. 1: Processing flow of the proposed method. The sound field and the sound source are measured separately, from which a volumetric sound field and a 3D mesh are generated, respectively. These are spatially aligned and integrated for visualization on a 3D platform.

ment, high-spatial-resolution imaging of the sound field has been achieved using acousto-optic sensing [7]. Extensive research has been conducted on visualizing sound fields using optical imaging systems with high-speed cameras, including parallel phase-shifting interferometry (PPSI) [8] and digital holography [9, 10]. Using high-speed cameras enables the capture of sound fields instantaneously with the spatial resolution defined by the imaging system. In optical imaging, when objects are present within the imaging area, their shapes are observed as projected 2D silhouettes. The silhouettes can be exploited to analyze the acoustic phenomena occurring in the vicinity of the objects [11, 12]. Rajput *et al.* have conducted simultaneous imaging of objects and sound using digital holography, however, both are visible only from a fixed viewpoint, making it inappropriate for 3D representation [13]. To date, no method has successfully achieved both the three-dimensional and high-resolution visualization of the sound field and objects.

In this work, we propose a method for 3D visualization of a sound field along with a textured mesh of sound sources, both of which are generated from short videos. The sound field is reconstructed from a high-speed video of optical interferograms [8]. The continuously rotating measurement system of the sound field was developed, enabling the acquisition of 3D volumetric data from just a 2-second measurement. The sound source mesh and texture are obtained from *nerf2mesh* [14], a neural radiance field-based mesh generation. Combining the volumetric data of the sound field with the 3D object mesh makes it possible to visualize how the sound is generated from the source and propagates through space. The proposed method is applied to visualize loudspeaker, smartphone, and ultrasound transducer array. The diffraction of acoustic waves in single- and double-slit experiments are also analyzed.

2 Methods

Figure 1 illustrates the processing flow of the proposed method. To obtain a 3D representation of both the sound field and the object, two imaging procedures are performed: one captures the sound field using optical interferometry with a high-speed camera, and the other records a 360° video of the sound source using a conventional camera. The sound field and object mesh are then reconstructed and integrated. The following sections describe each step in detail.

2.1 3D sound field measurement

To obtain 3D sound fields, optical high-speed imaging using PPSI [8] and sound-field reconstruction based on filtered back projection (FBP) [15, 16] are employed. By measuring sound with the high-speed camera while continuously rotating the sound source by 180 degrees, 3D sound field reconstruction from only two seconds of measurement is achieved.

The measurement system setup is shown in Fig. 2. The optical system is configured as PPSI [17], observing interference fringe images using a polarized high-speed

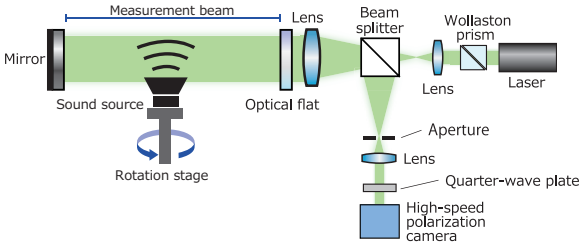


Fig. 2: Optical measurement system using PPSI and a high-speed polarized camera. The optical setup is configured as a Fizeau-type interferometer with a measurement beam diameter of 100 mm, projecting the sound field onto the camera. The sound source is mounted on a rotation stage.

camera [8]. The sound source to be measured is positioned near the measurement beam. The stage rotates 180° in about 1.7 seconds, during which the high-speed camera captures the perturbation of the measurement beam caused by the sound. During the measurements, a sinusoidal wave at a given frequency was generated from the sound source, and the harmonic response of the sound field was analyzed.

The optical measurement of the sound is based on a physical phenomenon known as the acousto-optic effect [15]. Since sound is a variation in air density, the refractive index, which represents the relative propagation speed of light in air, fluctuates due to sound. These fluctuations in the refractive index of the air manifest in the phase of light passing through the field, allowing for the measurement of sound by detecting the phase changes in light. The relationship between the light intensity of the interferometer and sound is given by

$$I = A + B\cos(\phi_s + \phi_0), \tag{1}$$

where A and B are the background illumination intensity and amplitude of the interference fringes, respectively, ϕ_s is the phase variation of light due to sound, and ϕ_0 is a constant component depending on the configuration and geometry of the interferometer. The phase variation ϕ_s is determined by the acousto-optic effect as

$$\phi_s = kC_{\mathrm{AO}} \int_L p\, dl, \tag{2}$$

where k is the wavenumber of light, C_{AO} is a constant representing the acousto-optic effect determined by the state of the air, and p is the sound pressure, with the line integral taken along the optical path L. In the optical system depicted in Fig. 2, the phase term of the cosine function, $\phi_s + \phi_0$, is quantitatively obtained based on the phase extraction algorithm [18]. By performing a Fourier transform along the time direction to the extracted optical phase, a complex-valued amplitude of the projected sound field is obtained.

To obtain a volumetric sound field from the projection observations, it is necessary to measure the projections from multiple angles and perform tomographic reconstruction [15, 19]. In the proposed method, instead of moving the heavy and bulky measuring system, the sound source is placed on a rotating stage, and the projections are measured while the sound source is continuously rotating. The optical phase variations acquired while rotating the sound source can be written as

$$\begin{aligned}\Phi(s,z,t) &= \phi_s(s,z;\theta(t)) \\ &= kC_{\mathrm{AO}} \int_{u_1}^{u_2} p(s\cos\theta - u\sin\theta, \\ &\qquad s\sin\theta + u\cos\theta, z, t)du, \end{aligned} \tag{3}$$

where (s,z) correspond to the coordinates along the horizontal and vertical axes of the captured image, $\theta(t)$ represents the rotation angle of the sound source at time t, u_1 and u_2 are the start and end points of the line integral. The definitions of the coordinates are illustrated in Fig. 3(a). The line integral is performed along the fixed laser path, which appears tilted at angle $\theta(t)$ viewing from the local coordinate of the sound source. In the constructed system, the stage rotates 180 degrees in 1.68 seconds. With the frame rate of the high-speed camera set to 50,000 fps, the total images are 84,000. The sound source rotates 0.0021° per frame.

The 84,000 recorded images are separated by a specific time width and transformed into projection data from multiple angles by Fourier transform, as depicted in Fig. 3(b). The segment width controls the tradeoff between the amount of noise in the Fourier transform bins and the angular resolution of the projections. Here, the time window length was set to 1000, with no segment overlap. The time duration for 1,000 frames is 20 ms, and the rotation angle during this period is 2.14°. Figure 3(c) shows the examples of the projection images.

After obtaining the projections, 3D sound field reconstruction is performed. When reconstructing the volumetric field from multiple projections, each data must be phase synchronized. However, when the time duration of each segment is not an integer multiple of the

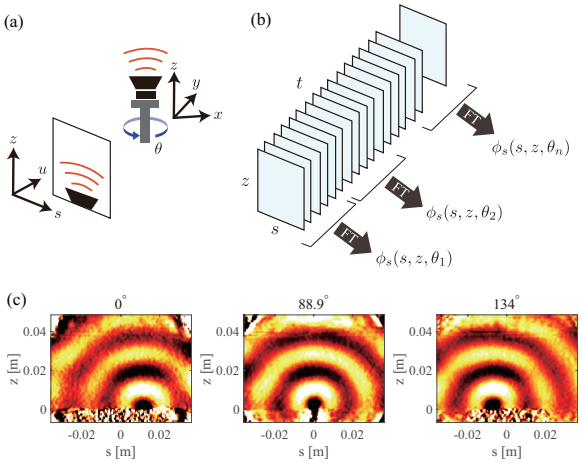


Fig. 3: Optical projection measurement of the sound field. (a) Definitions of the sound source and image plane coordinate. (b) The time-sequential images are segmented, and time-domain Fourier transform (FT) is calculated for each segment, which yields acoustic-field projections at different angles. (c) Examples of the measured projections.

periods of sound, or if there is a slight discrepancy between the clocks of the sound source and the camera, an unknown phase shift occurs. Here, this phase shift was corrected using only the observed data. Since the data at 0° and 180° are mirror images of each other, their sound field should be the same. Therefore, the phase shift between the 0° and mirrored 180° projections was calculated, and the phase shifts were compensated by linear interpolation. For the reconstruction, FBP was used [15, 16] due to its simplicity and efficiency. To perform 3D reconstruction, each z slice was independently reconstructed, as shown in Fig. 4. As the harmonic projection data is complex-valued, the real and imaginary parts were handled separately and then converted into a complex number to restore the volumetric complex sound field.

2.2 Mesh generation from video

In this study, nerf2mesh [14] was used to generate a 3D model of the sound source from multi-angle images, using an implicit neural field to produce textured 3D polygonal meshes. The mesh generation followed the procedure provided by the original authors of nerf2mesh. First, the sound source used in the measurement was recorded using a conventional video camera.

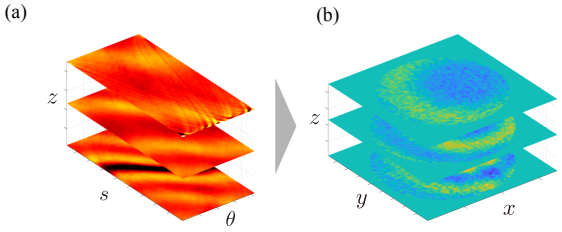


Fig. 4: 3D sound field reconstruction. (a) Projections of the sound field. (b) Volumetric data is reconstructed from the projections. FBP is applied to each z-index plane and combined into a 3D array representing the volumetric sound field.

The camera parameters were then estimated, and the background was removed. Next, in stage 0 optimization, the neural field was trained, followed by stage 1 optimization, in which the surface geometry and texture were reconstructed. Both stage 0 and stage 1 were trained for 1500 iterations each. From a one-minute video, approximately 200 images were extracted for mesh generation. Each optimization stage took less than 15 minutes to train using a single NVIDIA RTX A4000 GPU.

2.3 Integration and visualization

In this study, we visualized both the volumetric sound field obtained through sound field reconstruction and the 3D object obtained through nerf2mesh. To visualize the object mesh with the reconstructed sound field, alignment is conducted according to the definition of the coordinate axes of the measurement space. The volumetric sound field is a 3D grid data with coordinates in meters based on the conditions of the measurement system. The origin of the volumetric data was set to the location on the object surface along the rotation axis. In optical images, the object shape can be captured as a silhouette, allowing the origin to be identified from the boundary of the silhouette. In contrast, the mesh data generated from NeRF does not match the measurement coordinate system in terms of scale, coordinates, and angles. The scale of the mesh data was adjusted in 3D software so that a representative length of the object mesh (such as the width of a smartphone or the diameter of a speaker) matched an actual object length measured by a caliper. The origin was chosen at points on the object mesh that matched the definition. Additionally, the rotation was manually adjusted to align

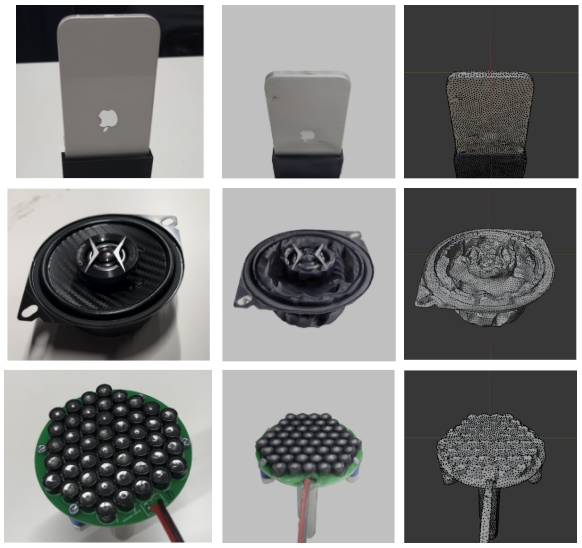


Fig. 5: The sound source and generated texture and mesh from the video. (from left to right) a photograph, 3D rendering with texture, and mesh. (from top to bottom) iPhone 12 stood upside down on a stand, a coaxial loudspeaker, and an ultrasonic transducer array.

with the definition of the measurement coordinate system. These procedures ensured that the alignment of the positions and scales of the sound field and objects in dimensions denoted in meters.

For the visualization, the 3D software Blender was used. The volumetric data of the sound field was visualized both as slice planes and through volumetric mapping using openvdb [20]. For the slice planes, image files with given colormaps and ranges are loaded into Blender, locating the images to the appropriate positions and adjusting their transparency. For the volumetric mapping, volumetric data saved in VDB format was imported into Blender, and the emission color and strength were determined corresponding to the sound pressure values at each point. These procedures were automated by Python.

3 Experiments

The proposed method was demonstrated with three sound sources and at various frequencies. For all experimental conditions, the frame rate of the high-speed camera (Photoron CRYSTA PI-1P) was set to 50,000 fps, and the resolution of the sound field images was

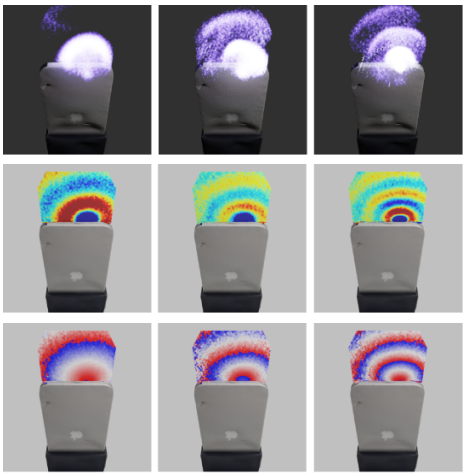


Fig. 6: Visualization of the sound field generated by the iPhone. (from left to right) 8 kHz, 12 kHz, and 16 kHz. (from top to bottom) volumetric representation, the real part of the complex-valued sound field, and the phase of the sound field.

175×134. Reconstruction was performed from data of 84 angles, resulting in volumetric data consisting of $175 \times 175 \times 134$ points. The distance between voxels varied slightly depending on the data but was approximately 0.5 mm.

The sound sources included a smartphone (Apple iPhone 12), a loudspeaker unit (Pioneer TS-F1040), and a circular ultrasonic transducer array consisting of 48 small transducers. Figure 5 shows photographs of each sound source and their corresponding reconstructed mesh data. The overall shape and texture of each sound source were generally well-reproduced. When viewing the mesh data without texture, some complex geometries—such as the concave region of the loudspeaker and the inter-element spaces of the ultrasonic transducer array—were not accurately captured. However, for the purpose of audio-visual visualization, the obtained textured meshes are sufficiently acceptable.

For the iPhone, as sound radiates from acoustic ports at the bottom of the casing, measurements were conducted with the device mounted upside down on the rotating stage. The volume of the device was set to maximum, and sine waves of 8, 12, and 16 kHz were played for the measurement. The imaging results are shown in Fig. 6. The top row represents the real part

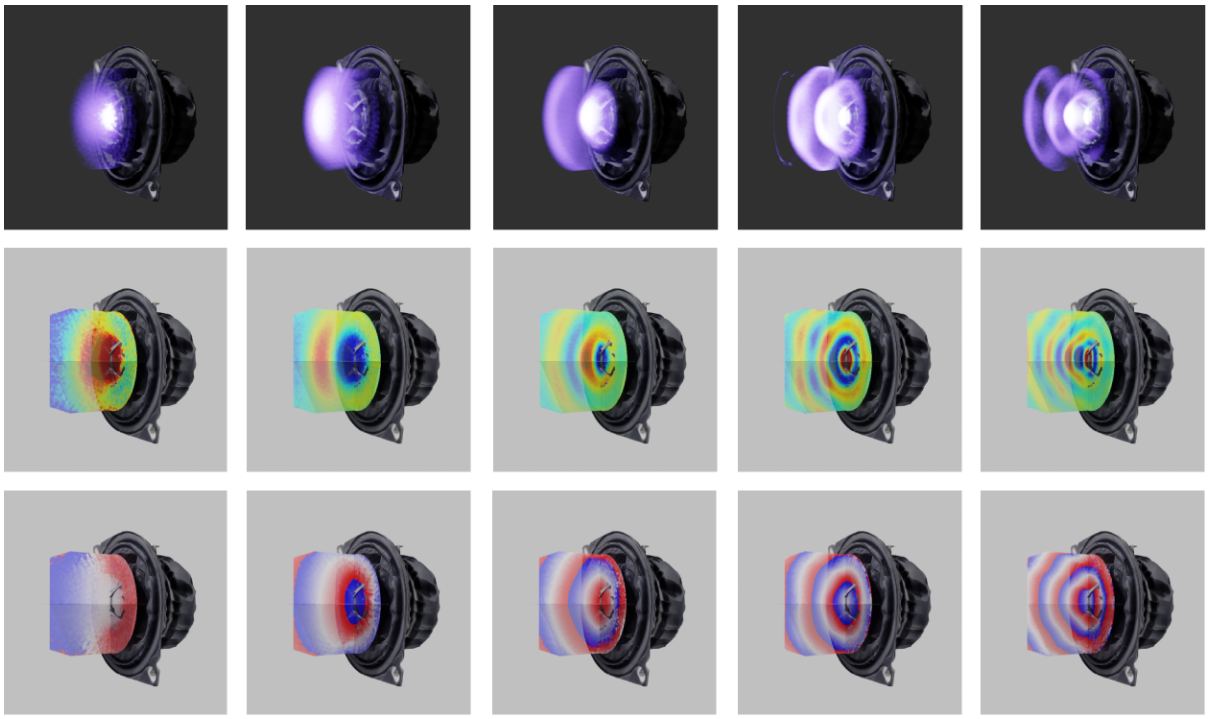


Fig. 7: Visualization of the sound field generated by the loudspeaker. (from left to right) 4 kHz, 8 kHz, 12 kHz, 16 kHz, and 20 kHz. (from top to bottom) volumetric representation, the real part of the complex-valued sound field, and the phase of the sound field.

of the complex sound field by the volumetric visualization. It is observed that sound waves are radiating primarily from the upper right area, where the acoustic ports exist. Due to the different wavelengths at different frequencies, a single hemispherical wavefront is visible at 8 kHz, whereas three layers of wavefronts are visualized at 16 kHz. The single slices of the sound field are displayed; Fig. 6(b) and (c) show the real part and angle of the complex field, respectively. The wavefront of the sound waves and the position where the sound emerges are clearly observed.

The loudspeaker unit was tested as a sound source at frequencies of 4, 8, 12, 16, and 20 kHz. The results are shown in Figure 7. This speaker has a coaxial configuration with a tweeter positioned in front of a full-range speaker unit. From the volumetric display in the top row, higher energies are observed at the central position of the tweeter. From the slice visualizations, the sound wavefronts at 16 and 20 kHz are not simple monotonic curves but show more undulating patterns. This is thought to be due to the effects of split vibrations and diffraction caused by the shorter wavelengths of the sound.

The 40 kHz ultrasonic waves from the transducer array were measured. Since the frame rate of the camera was set to 50 kfps, the sampling theorem was not satisfied for 40 kHz ultrasound. However, since the high-speed camera has no anti-aliasing filter, the 40 kHz signal was observed at 10 kHz due to aliasing at 25 kHz. Therefore, by extracting the aliased 10 kHz component from the Fourier-transformed spectrum of each segment, the recording of the 40 kHz sound field was successfully achieved. The results of the ultrasonic transducer array are shown in Fig. 8. Figure 8 (a) visualizes the multiple zones of locally high sound pressure near the transducers. Figure 8 (b) displays the magnitude of sound pressure on the plane close to the transducer array. The high-magnitude bright zones overlap in front of each transducer. Figures 8 (c) and (d) show slices of the real parts and phase of the sound pressure along the planes perpendicular to the array. In the vicinity of the transducers, sounds emitted from each transducer are discerned individually; however, upon propagating

approximately four wavelengths, these sounds merge to create a unified planar wavefront.

Lastly, to demonstrate the extensibility of the proposed method to investigate object–acoustic interactions other than sound sources, slit experiments were conducted. Thin plates with single and double slits of width of 5 mm and height of 50 mm were fabricated using a 3D printer. The spacing between the double slits was set to 20mm. They were fixed at a distance of 7.7 mm from the array surface, and the diffraction and interference of sound passing through the slits were observed. For the slit plates, the same 3D printing data were also used for visualization, located relative to the mesh of the array. Figures 8 (e) and (f) show the results. The sound waves transmitted through the slits are successfully visualized. For the single slit, a planar wavefront is observed when viewed in a slice that includes the long axis of the slit, whereas a planar wavefront is apparent in the perpendicular plane. In the double slits, more complicated patterns are formed due to the interference of the two waves. Note that since the sound waves incident on both slits were not exactly the same, the interference patterns were not perfectly symmetric.

4 Summary

In this work, we demonstrated the feasibility of integrated visualization method for audio and visual. By utilizing optical sound-field imaging technology, it is possible to reconstruct 3D sound fields with a resolution of less than one millimeter from just a 2-second video, which is quite difficult with conventional microphone-based methods. By employing a NeRF-based 3D mesh generation method, an integrated representation of vision and spatial audio in 3D is achieved. The potential applications may include device development, ultrasonic engineering, and new content creation in spatial sound presentation, education, and video promotion.

References

[1] Bader, R., *Microphone Array*, pp. 1179–1207, Springer New York, New York, NY, 2014, ISBN 978-1-4939-0755-7.

[2] Comesana, D. F., *Scan-based sound visualisation methods using sound pressure and particle velocity*, Ph.D. thesis, University of Southampton, 2014.

[3] Döbler, D. and Heilmann, G., "Perspectives of the Acoustic Camera," in *Internoise 2005*, 2005.

[4] McCormack, L., Delikaris-Manias, S., and Pulkki, V., "Parametric acoustic camera for real-time sound capture, analysis and tracking," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, pp. 412–419, 2017.

[5] Inoue, A., Yatabe, K., Oikawa, Y., and Ikeda, Y., "Visualization of 3D sound field using see-through head mounted display," in *ACM SIGGRAPH 2017 Posters*, SIGGRAPH '17, New York, NY, USA, 2017, ISBN 9781450350150.

[6] Inoue, A., Ikeda, Y., Yatabe, K., and Oikawa, Y., "Visualization system for sound field using see-through head-mounted display," *Acoustical Science and Technology*, 2019.

[7] Verburg, S. A., Ishikawa, K., Fernandez-Grande, E., and Oikawa, Y., "A Century of Acousto-Optics- From Early Discoveries to Modern Sensing of Sound with Light," *Acoustics Today*, 19(3), pp. 54–62, 2023.

[8] Ishikawa, K., Yatabe, K., Chitanont, N., Ikeda, Y., Oikawa, Y., Onuma, T., Niwa, H., and Yoshii, M., "High-speed imaging of sound using parallel phase-shifting interferometry," *Optics Express*, 24, pp. 12922–12932, 2016.

[9] Matoba, O., Inokuchi, H., Nitta, K., and Awatsuji, Y., "Optical voice recorder by off-axis digital holography," *Optics Letters*, 39, pp. 6549–6552, 2014.

[10] Rajput, S. K., Matoba, O., Kumar, M., Quan, X., and Awatsuji, Y., "Sound wave detection by common-path digital holography," *Optics and Lasers in Engineering*, 137, p. 106331, 2021.

[11] Ishikawa, K., Yatabe, K., and Oikawa, Y., "Seeing the sound of castanets: Acoustic resonances between shells captured by high-speed optical visualization with 1-mm resolution," *Journal of the Acoustical Society of America*, 148, pp. 3171–3180, 2020.

[12] Tanigawa, R., Yatabe, K., and Oikawa, Y., "Experimental visualization of aerodynamic sound sources using parallel phase-shifting interferometry," *Experiments in Fluids*, 61(9), 2020.

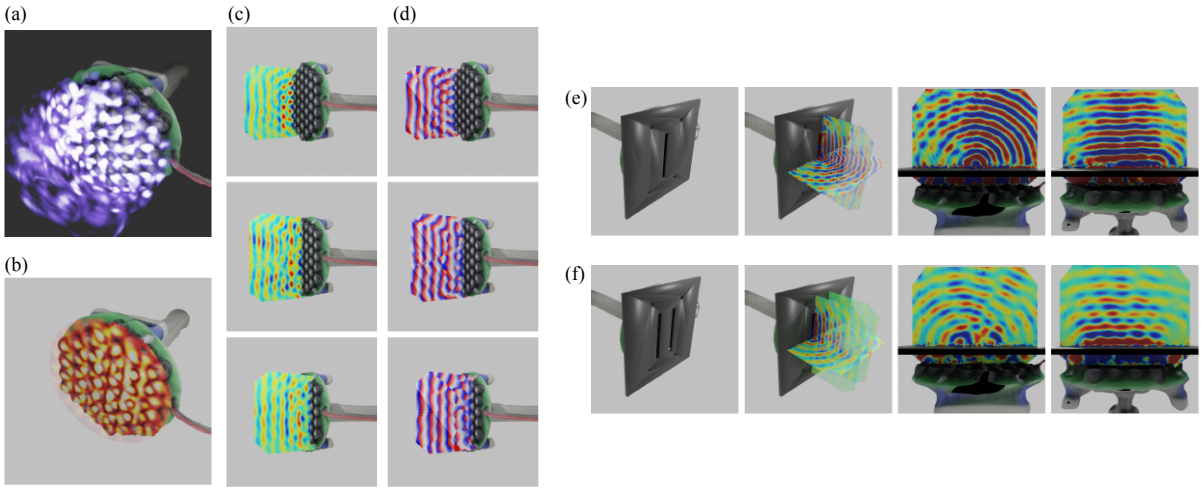


Fig. 8: Visualization of the ultrasound transducer array. (a) volumetric representation, (b) magnitude of the sound field at a plane close to the array surface, (c) and (d) real part and phase of the sound field at different planes, respectively. (e) The sound wave propagates through a single slit. (f) The sound wave propagates through double slits.

[13] Rajput, S. K., Matoba, O., and Awatsuji, Y., "Holographic multi-parameter imaging of dynamic phenomena with visual and audio features," *Opt. Lett.*, 44(4), pp. 995–998, 2019.

[14] Tang, J., Zhou, H., Chen, X., Hu, T., Ding, E., Wang, J., and Zeng, G., "Delicate Textured Mesh Recovery from NeRF via Adaptive Surface Refinement," in *International Conference on Computer Vision (ICCV)*, 2023.

[15] Oikawa, Y., Goto, M., Ikeda, Y., Takizawa, T., and Yamasaki, Y., "Sound field measurements based on reconstruction from laser projections," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 4, pp. iv/661–iv/664 Vol. 4, 2005.

[16] Torras-Rosell, A., Barrera-Figueroa, S., and Jacobsen, F., "Sound field reconstruction using acousto-optic tomography," *Journal of the Acoustical Society of America*, 131, pp. 3786–3793, 2012.

[17] Awatsuji, Y., Sasada, M., and Kubota, T., "Parallel quasi-phase-shifting digital holography," *Applied Physics Letters*, 85(6), pp. 1069–1071, 2004.

[18] Yatabe, K., Ishikawa, K., and Oikawa, Y., "Hyper ellipse fitting in subspace method for phase-shifting interferometry: practical implementation with automatic pixel selection," *Opt. Express*, 25(23), pp. 29401–29416, 2017.

[19] Verburg, S. A. and Fernandez-Grande, E., "Acousto-Optical Volumetric Sensing of Acoustic Fields," *Phys. Rev. Appl.*, 16, p. 044033, 2021.

[20] Academy Software Foundation (ASWF), "OpenVDB," Accessed:21st, Jan. 2024.