Selective 3D Audio Presentation System for a Moving Individual Tracking Using a Pair of Parametric Speakers

Hiroyuki Minematsu[1], Hyuma Auchi[1], Ayuto Togashi[1], Rina Masuda[1], Yohei Shida[2,3] and Keiichi Zempo[2,*]

[1]Graduate School of Science and Technology, University of Tsukuba, Tsukuba, Japan
[2]Institute of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan
[3]School of Computing, Institute of Science Tokyo, Yokohama, Japan

Abstract

With the rise of multi-user AR/MR environments, there is an increasing demand for auditory interfaces that can provide individualized spatial audio without adding to environmental noise or cognitive load, not only in public spaces but also in interactive digital contexts. Conventional loudspeakers disperse sound broadly, disturbing non-target listeners, while headphones isolate users from their surroundings and conflict with the open and multimodal nature of AR/MR. Parametric array loudspeakers (PALs) offer extremely high directivity; however, previous research has primarily focused on static users, leaving unresolved the technical challenge of achieving both selective acoustic intervention and stable sound localization for moving individuals in multi-user scenarios. Here, we present a system that employs a pair of tracking PALs, guided by depth-camera-based motion capture, to deliver spatialized 3D audio exclusively to a walking target. Two experiments evaluated (i) selective acoustic intervention and (ii) localization accuracy while walking. Results showed that only the tracked target consistently received stable sound pressure, while non-target individuals experienced minimal exposure, and that localization accuracy during walking was more stable compared with fixed PALs. These findings demonstrate that tracking PALs can simultaneously achieve selectivity and stability in dynamic multi-user environments, paving the way for immersive and noise-conscious auditory interfaces in public guidance and AR/MR applications.

Keywords

spatial audio, auditory perception, human motion tracking

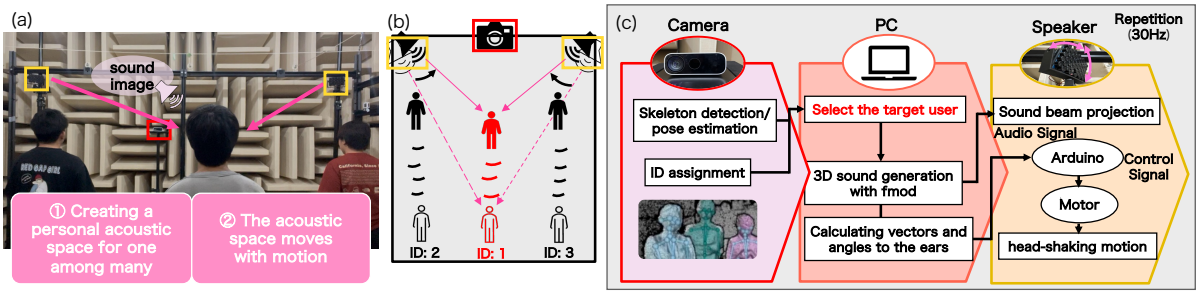

Figure 1: (a) An example of how this system can be used to deliver sound to only one specific person among multiple people walking around. The system automatically tracks the user and creates a personal acoustic space without the need for headphones. (b) A conceptual diagram of the core of this system. The camera detects each person's skeletal structure and assigns a "Body ID" to each individual, allowing it to selectively track one target. This enables the novelty of this research: presenting sound to one person among multiple people walking around. (c) The system processing flow. The camera calculates the ear position of the target with a Body ID in real time, and a motor-driven speaker physically follows the target and delivers an acoustic beam. This high-speed tracking loop ensures a stable audio experience that is uninterrupted even when moving.

APMAR'25: The 17th Asia-Pacific Workshop on Mixed and Augmented Reality, Sep. 26-27, 2025, Busan, South Korea
*Corresponding author.
s2520818@u.tsukuba.ac.jp (H. Minematsu); auchi.hyuma.24@aclab.esys.tsukuba.ac.jp (H. Auchi); s2420510@u.tsukuba.ac.jp (A. Togashi); s2210014@u.tsukuba.ac.jp (R. Masuda); shida@sk.tsukuba.ac.jp (Y. Shida); zempo@iit.tsukuba.ac.jp (K. Zempo)
0009-0006-3771-1748 (H. Minematsu); 0009-0007-5440-8625 (H. Auchi); 0009-0004-1217-9017 (A. Togashi); 0009-0002-3382-0315 (R. Masuda); 0009-0003-8207-8417 (Y. Shida); 0000-0003-2339-5298 (K. Zempo)

1. Introduction

With the advancement of Mixed Reality (MR) and Augmented Reality (AR) technologies, interactive experiences that merge real-world and virtual information are expanding across diverse domains [1, 2]. Among these, auditory information, particularly spatial audio in combination with visual information, has been shown to greatly enhance realism and immersion [3, 4, 5, 6, 7]. In the context of MR and AR, there is a growing demand for real-time and high-precision audio presentation that adapts to user movements and gaze shifts [8, 9, 10]. Furthermore, MR/AR systems designed for multiple users have recently emerged [11, 12]. Consequently, in addition to highly accurate spatial audio rendering, it is becoming increasingly important to deliver audio individually tailored to each user within the same physical space. Applications of spatial audio extend beyond MR/AR to public environments such as train stations, commercial facilities, exhibitions, and digital signage in urban spaces [13, 14, 15]. However, conventional loudspeaker-based methods indiscriminately diffuse sound to large audiences, where unintended listeners may perceive it as unwanted noise, thereby increasing cognitive load and stress [16, 17, 18]. Thus, there is a growing necessity for selective and high-precision spatial audio presentation tailored to individual users, not only in MR/AR but also in public spaces.

Conventional approaches to spatial audio reproduction with external loudspeakers have been designed for fixed configurations, targeting sound localization within a so-called "sweet spot."However, in dynamic scenarios where users are walking or changing orientation, maintaining consistent localization is difficult, often degrading the accuracy of audio presentation [19, 20]. In addition, MR/AR use cases frequently involve multiple simultaneous users, necessitating techniques to provide individualized audio information. To address this issue, the parametric array loudspeaker (PAL), which exhibits extremely high directivity, has drawn attention. For example, Kuratomo et al. controlled an ultrasonic directional loudspeaker toward both ears of a static user, presenting spatial audio exclusively to that individual [21, 22]. However, these studies were limited to single, stationary users, and the effectiveness of selective presentation in multi-user environments or the stability of localization during walking has not been sufficiently verified.

To address these challenges, this study develops a system that recognizes the positions and postures of multiple users in real time and dynamically steers directional loudspeakers to follow a target user. Even in scenarios where multiple users are walking within the same space, the system delivers audio exclusively to the designated individual, while suppressing sound leakage to non-target users and maintaining stable spatial audio presentation.

The objectives of this study are to investigate the following research questions:

RQ1: In walking scenarios with multiple users, can the proposed system achieve selective audio delivery to a specific individual?

RQ2: To what extent can the proposed system maintain sound localization accuracy for users while walking?

By addressing these research questions, we verify the effectiveness of the proposed system in dynamic and multi-user environments. This work demonstrates the potential for a new form of spatial audio presentation applicable to MR/AR contexts involving multiple users and mobile conditions.

2. Related Work

2.1. Selective Acoustic Intervention

With the rise of AR and MR technologies, scenarios involving multiple users working within the same space have become increasingly common, thereby requiring methods for individualized audio presentation [11, 12]. In public spaces, conventional audio presentation has primarily relied on loudspeakers, which can result in increased stress and cognitive load due to noise [16, 17, 18].

To address this issue, extensive research has been conducted on selective acoustic presentation meth-

ods that deliver sound only to a specific area or individual. A promising technology for achieving such selectivity is the parametric array loudspeaker (PAL), which utilizes ultrasonic waves to generate audible sound in midair, thereby forming highly directional acoustic beams [23]. Many studies have focused on enhancing PAL directivity. For example, Fan et al. proposed a method that employs phase-randomized arrays to suppress grating lobes, improving beam-steering accuracy toward the desired direction [24]. Kinjo et al. developed a spot-delivery system capable of controlling the irradiation point with an error margin of within ±1° based on 3D position estimation using stereo cameras, demonstrating that users could clearly perceive audio beams targeted at themselves [25]. Furthermore, Zhuang et al. proposed a sound-zone control method using a minimal setup of a single PAL to generate multiple audible zones, achieving performance comparable to conventional multi-loudspeaker systems [26]. In addition, simulation studies on sound-zone control using PAL arrays have shown superior performance and robustness compared to electrodynamic loudspeakers under high-frequency and low-SNR conditions [27].

However, most of these studies have focused on static single users. Systematic verification of whether it is possible to continuously track and selectively intervene with a specific user in a multi-user environment, while suppressing sound leakage to surrounding individuals, has not been sufficiently conducted. In this study, we quantitatively evaluate the feasibility of selective intervention in environments where multiple users are in motion using the proposed system.

2.2. Spatial Audio

Spatial audio is an indispensable technology for creating high levels of presence and immersion, and its importance is widely recognized in MR/AR research [7, 10]. For example, Kern et al. demonstrated that incorporating natural environmental sounds and footsteps synchronized with user actions into VR environments significantly enhances presence and realism, proving that spatial audio complements visual information and deepens immersion [6]. Similarly, Rumiński et al. reported that in an AR hidden-object search task, spatialized sound presentation significantly improved task completion speed and efficiency compared to non-spatial conditions, demonstrating the effectiveness of spatial audio for navigation support in AR [10].

The two primary approaches to spatial audio presentation in AR/MR environments are headphones and loudspeakers. While headphones provide highly accurate localization, they block real-world sounds and hinder the fusion with reality, which is central to AR/MR. Loudspeakers, on the other hand, offer a more open auditory experience but suffer from the limitation that accurate sound perception is confined to a narrow sweet spot [19]. Furthermore, cross-talk—where sound from one loudspeaker reaches the opposite ear—is known to degrade localization accuracy [20].

One approach to addressing these issues is to leverage the high directivity of PALs. Kuratomo et al. demonstrated that, by steering directional loudspeakers toward both ears of a static user based on depth-camera position estimation, it is possible to present spatialized sound exclusively to a specific individual while maintaining stable localization even during head rotation [21, 22]. Nakayama et al. proposed a method that combines PALs with conventional loudspeakers, controlling the ratio of direct sound to reverberation in order to manipulate perceived source distance and reduce cross-talk [28].

However, these prior studies primarily focused on static users. The extent to which sound localization accuracy and tracking performance are maintained for users while walking remains insufficiently explored. Therefore, in this study, we use the proposed system to continuously present spatial audio to users in motion and quantitatively evaluate localization accuracy under dynamic conditions.

3. Proposed Method

3.1. Overview

In this study, we developed a spatial audio system designed to track a specific individual among multiple users in dynamic environments involving movement and rotation, and to present a clear sound

image exclusively to that person. The system estimates the positions of the user's ears and head orientation, and based on this information, it controls the direction of parametric array loudspeakers (PALs) in real time, thereby realizing selective acoustic presentation that delivers sound precisely to any arbitrary point in space.

The system consists of two PALs, a depth camera (Azure Kinect), and FMOD Studio for audio playback, all integrated under unified control in C++. The depth camera captures skeletal information of users, enabling target user selection, ear coordinate computation, angle calculation, loudspeaker orientation control, and physical sound playback to operate in real time. Furthermore, playback status is controlled depending on the presence of a tracked user: when the target exits the field of view, audio is immediately muted. This design enables precise sound image presentation to a single user even within interactive spatial environments.

3.2. Tracking and Target Switching Logic

Through skeletal tracking by the depth camera, multiple joint positions such as the user's left and right ears, head, and neck are obtained frame by frame. Target user selection is managed using the Body ID assigned by Kinect; when tracking begins, the first detected person within the frame is registered as the target.

If the current Body ID is no longer detected in subsequent frames (e.g., when the user leaves the camera's field of view), the system reassigns the target to the first newly detected person. This sequential switching ensures continuous audio presentation to one user even in dynamic public spaces where people are frequently entering and exiting.

If no individuals are detected at all, ongoing audio playback is paused. This prevents unintended acoustic presentation to non-targets and minimizes sound leakage. When a user is detected again, playback automatically resumes, providing autonomous responsiveness to user appearance and disappearance.

3.3. Speaker Angle Calculation and Control

In this system, two PALs are fixed on the left and right sides of the Kinect camera. The left speaker is controlled to direct sound toward the left ear, and the right speaker toward the right ear, respectively. Each speaker is connected to an Arduino via an independent serial port, through which real-time horizontal and vertical angles are transmitted.

For control, the difference vector between each speaker position and the target ear coordinates is computed. After applying rotational correction to transform into the local coordinate system considering the speaker's mounting angle, the angles are calculated as follows:

$$\theta_{\text{pan}} = \text{clamp}\left(90 - \frac{180}{\pi} \cdot \arctan\left(\frac{x'}{z'}\right), 0, 180\right) \tag{1}$$

$$\theta_{\text{tilt}} = \text{clamp}\left(90 - \frac{180}{\pi} \cdot \arctan\left(\frac{y}{\sqrt{x'^2 + z'^2}}\right), 0, 180\right) \tag{2}$$

Here, (x', z') are the local coordinates after compensating for the physical tilt of the speaker, and $\text{clamp}(v, a, b)$ is a function restricting a value v within a–b. A rotation of $+45^\circ$ is applied for the left speaker and -45° for the right speaker, so that the ear-directed vectors are recalculated in each respective local coordinate system. In our implementation, the servo motors allowed the PALs to rotate within a range of $\pm 90^\circ$ horizontally and $\pm 45^\circ$ vertically, which was sufficient to cover typical head and body movements during walking.

The computed results are converted to integer angles, serialized as strings, and transmitted through the corresponding serial port to each speaker. The Arduino receives these values, generates PWM signals, and drives the motors to control the physical speaker angles in real time.

This computation is continuously performed in synchronization with skeletal frame updates from Kinect (approximately 30 Hz), enabling the speakers to follow the ear positions even while the user is moving.

3.4. FMOD-based Sound Playback

FMOD Studio is used for sound playback, where pre-prepared audio files are looped in spatial audio mode. The virtual sound source is fixed at the user's frontal position at $z = 1.0$ [m]. The user's head position detected by Kinect is converted into meters and set as the FMOD listener position. Additionally, the head orientation (yaw angle) is estimated from the quaternion (w, x, y, z) of the neck joint using the following equation:

$$\theta_{\text{yaw}} = \arctan 2\left(2(yz + wx), w^2 - x^2 - y^2 + z^2\right) \tag{3}$$

This yaw angle is used to update the forward vector of the FMOD listener, ensuring that sound images are perceived from the correct direction relative to the user's orientation. Thus, even when the user rotates, the perception of a frontal sound image is maintained.

3.5. Temporal Update Loop and Synchronization

The entire control algorithm is executed in synchronization with body frame updates from Kinect, operating at approximately 30 Hz. The following processes are repeated for each frame:

1. Determine the presence of a tracked target
2. Acquire ear coordinates and compute horizontal/vertical angles
3. Send loudspeaker control angles via serial communication
4. Toggle audio playback ON/OFF
5. Update FMOD listener position and orientation

Through this processing loop, smooth and accurate sound image presentation is achieved, enabling continuous and precise auditory tracking in environments where users are constantly in motion.

4. Experiments

4.1. Evaluation of Selective Acoustic Interventions

4.1.1. Method

To evaluate the feasibility of selective acoustic intervention, we conducted a sound pressure measurement experiment. The experimental setup is shown in Fig. 2(a). The experiment was conducted in an anechoic chamber with dimensions of approximately 3.0 m × 3.0 m. Three participants were positioned in the space, labeled as person 1, person 2, and person 3 from front to back. They were spaced 1.5 m apart, and each walked a distance of 3.0 m at a speed of approximately 0.5 m/s.

The experimental conditions included three setups: two-channel loudspeakers, fixed PAL, and tracking PAL. In the fixed PAL condition, sound was directed toward the center of the space, approximately 1.5 m along the walking line of person 2. In the tracking PAL condition, the sound was continuously directed in real time toward the ear position of person 2.

As the test sound, we used white noise, which is frequently employed in prior studies on noise evaluation, and the sound pressure level was adjusted to approximately 55 dB SPL at the point of maximum sound pressure, corresponding to typical voice-guidance levels [29, 30, 13]. An omnidirectional microphone (Behringer ECM8000) was used for measurement, which each participant held in front of their face. The recorded sound pressure levels were used to evaluate sound leakage to person 1 and person 3 when the sound image was directed at person 2.

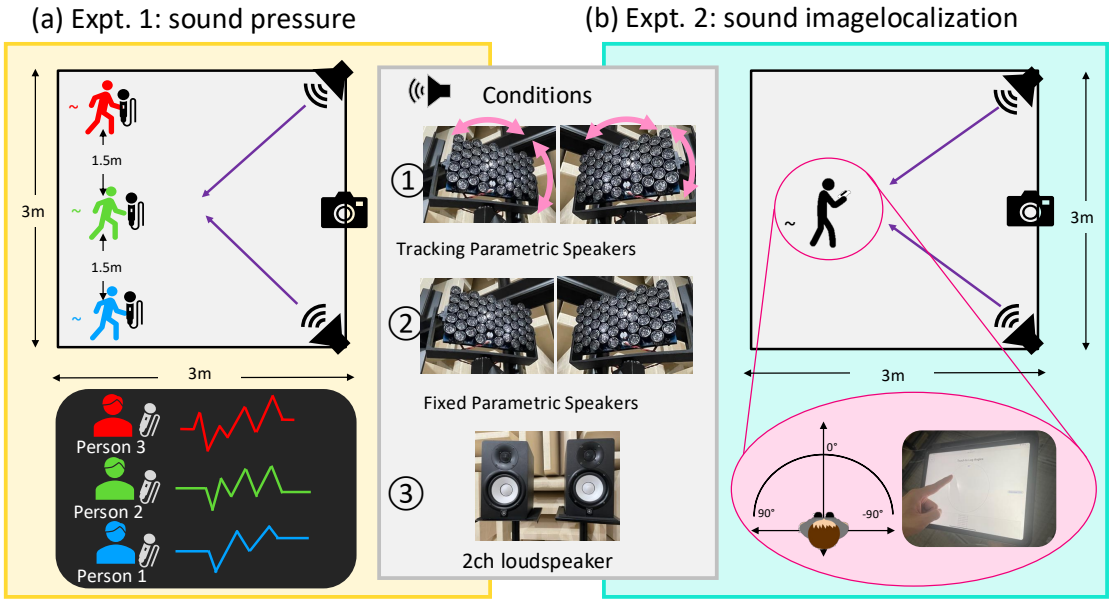

Figure 2: Experimental environment and conditions for Experiments 1 and 2

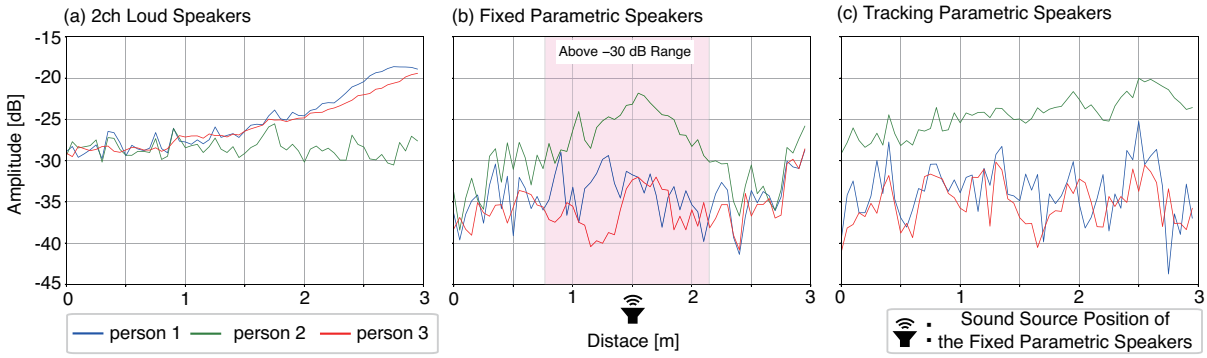

Figure 3: Sound pressure according to distance under each condition

4.1.2. Results

The results of this experiment are shown in Fig. 3. The graph illustrates the sound pressure distribution while participants walked from the starting point (0 m) to the end point (3.0 m).

In the two-channel loudspeaker condition (Fig. 3(a)), person 2 consistently exhibited a sound pressure level of approximately −30 dB, while person 1 and person 3 started near −30 dB but gradually experienced increasing sound pressure as they walked. In the fixed PAL condition (Fig. 3(b)), person 1 and person 3 consistently experienced sound pressure levels below −30 dB, while person 2 showed sound pressure above −30 dB only around the region 1–2 m, where the PAL beam was directed. Finally, in the tracking PAL condition (Fig. 3(c)), person 1 and person 3 were almost always below −30 dB, while person 2, the designated target, consistently experienced sound pressure levels above −30 dB.

4.2. Evaluation of Sound Localization while Walking

4.2.1. Method

To evaluate sound localization accuracy, we conducted an experiment in the same anechoic chamber as Experiment 1, where sound sources were presented from different directions, and participants indicated

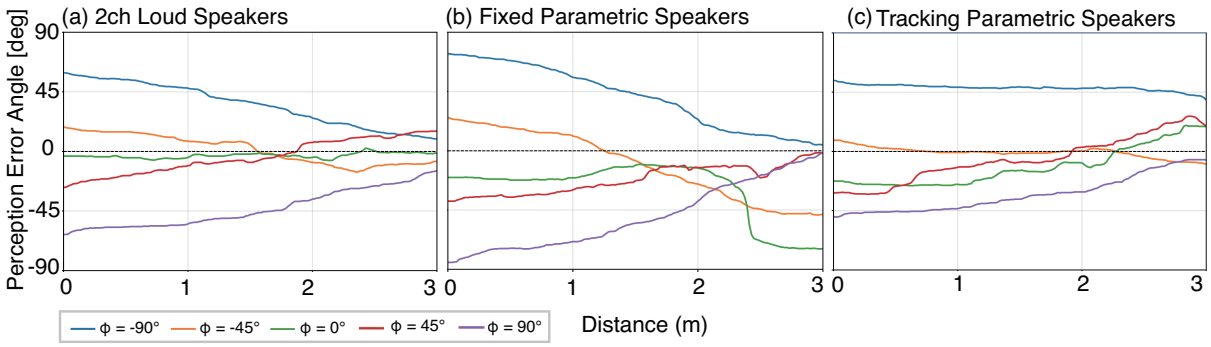

Figure 4: Sound image localization error under each condition

how accurately they perceived the direction. The experimental setup is shown in Fig. 2(b). Taking the camera's frontal direction as $\varphi = 0^\circ$, pink noise—commonly used in localization experiments [31]—was presented from five angles: 0°, $\pm 45^\circ$, and $\pm 90^\circ$. The sound pressure level was set to approximately 55 dB SPL, consistent with Experiment 1.

While walking, participants indicated the perceived direction of the sound image in real time using an evaluation application. The participants included three males and one female (N = 4). The same three conditions as in Experiment 1 were compared: tracking PAL (proposed method), fixed PAL, and two-channel loudspeakers.

4.2.2. Results

The results of the sound localization accuracy experiment are shown in Fig. 4 and Tab. 1. The graph illustrates the localization error angles perceived by participants while walking from 0 m to 3.0 m for each of the five presentation angles ($\varphi = 0^\circ, \pm 45^\circ, \pm 90^\circ$). Table 1 presents the root mean square error (RMSE) for each angle under each condition, as well as the overall average RMSE across all directions.

In the two-channel loudspeaker condition (Fig. 4(a)), the overall average RMSE was the smallest among the three conditions, at 26.97. In particular, at 0°, the RMSE was the lowest and most stable compared to the other two conditions. For $\pm 90^\circ$ and $\pm 45^\circ$, the error decreased gradually as participants approached the sound source from the initial position.

In the fixed PAL condition (Fig. 4(b)), the overall average RMSE was the largest among the three conditions, at 39.08. At 0°, errors were relatively small as the user passed through the beam's focal region (1.5–2.0 m), but beyond that range, the error increased sharply, producing a distinctive pattern in the graph.

In the proposed tracking PAL condition (Fig. 4(c)), the overall average RMSE was 30.54. Although this was larger than that of the two-channel loudspeakers, the errors at 90° and -45° were smaller. Furthermore, the error variation remained relatively stable across all presentation angles.

The violin plots in Fig. 5 further illustrate these results. In the fixed PAL condition (Fig. 5(b)), the distribution exhibited large variability for all presentation angles. Additionally, for 0° in the tracking PAL condition (Fig. 5(c)), the distribution was wider than that of the fixed PAL condition.

5. Discussion

5.1. Selective Acoustic Intervention (Answer to RQ1)

RQ1: Can the proposed system achieve selective acoustic intervention for a specific user while multiple individuals are walking?

In this experiment, the feasibility of selective acoustic intervention was evaluated by analyzing the sound pressure recorded using omnidirectional microphones held in front of participants'faces as they

Table 1
Numerical values of RMS localization error $\varepsilon(\Phi)$ [deg] corresponding to Fig. 4

$\varepsilon(\Phi)$	(a) 2ch Loud Speakers	(b) Fixed Parametric Speakers	(c) Tracking Parametric Speakers
$\varepsilon(\Phi = 90^\circ)$	46.95	51.04	38.68
$\varepsilon(\Phi = 45^\circ)$	24.09	34.15	24.44
$\varepsilon(\Phi = 0^\circ)$	8.67	36.64	24.93
$\varepsilon(\Phi = -45^\circ)$	16.56	27.03	13.03
$\varepsilon(\Phi = -90^\circ)$	38.55	46.55	51.61
RMSE Average	26.97	39.08	30.54

Note: $\varepsilon(\Phi)$ represents the RMS error [deg].

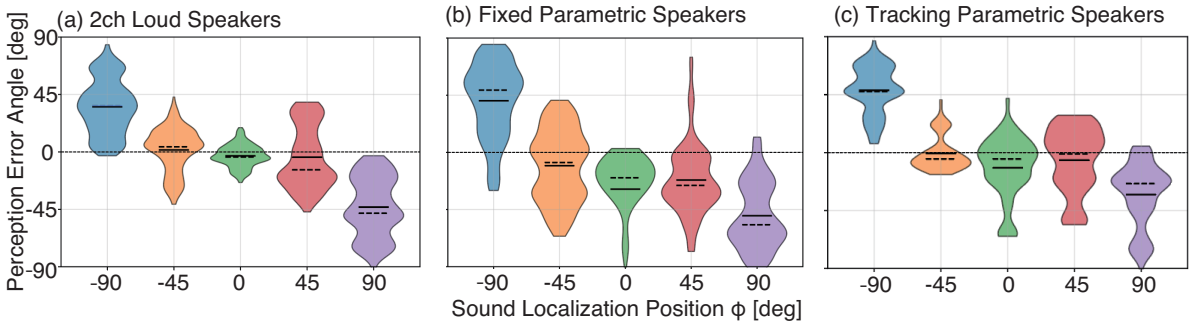

Figure 5: Violin plot of sound localization error for each sound presentation method

walked at equal intervals in the anechoic chamber. Three speaker conditions were compared: two-channel loudspeakers, fixed PAL, and tracking PAL.

As shown in Fig. 3(a), in the two-channel loudspeaker condition, the sound pressure for person 2 remained nearly constant at approximately −30 dB. This result can be attributed to the broad directivity of loudspeakers, which distribute sound evenly throughout the space, leading to uniform sound pressure regardless of distance. In contrast, for person 1 and person 3, the sound pressure increased as they approached the speakers during walking, reaching nearly −30 dB at around 2.0–3.0 m. This was likely due to their proximity to the speakers installed at both ends. These findings suggest that two-channel loudspeakers disperse sound across the entire space, making selective acoustic intervention for a single individual difficult.

As shown in Fig. 3(b), in the fixed PAL condition where sound was directed at person 2 around 1.5 m, the sound pressure for person 1 and person 3 consistently remained below −30 dB. In contrast, only person 2 exhibited sound pressure levels above −30 dB within the 1.0–2.0 m range, where the PAL beam was directed. This demonstrates that the high directivity of PALs enables selective acoustic intervention for person 2 within this range. However, for person 2, sound pressure fell below −30 dB in the 0–1.0 m and 2.0–3.0 m ranges, indicating reduced audibility. Thus, selective presentation by fixed PALs is limited to approximately ±0.5 m around the beam's focal point.

As shown in Fig. 3(c), in the tracking PAL condition, the sound pressure for person 1 and person 3 consistently remained below −30 dB, while the designated target, person 2, always exhibited sound pressure above −30 dB. These results demonstrate that tracking PALs can provide continuous and selective acoustic intervention during walking, regardless of the target's distance.

In summary, the proposed tracking PAL method enables selective acoustic intervention more effectively than two-channel loudspeakers or fixed PALs. Importantly, even under multi-user walking conditions, it allows continuous selective presentation while reducing noise for surrounding individuals.

5.2. Sound Localization while Walking (Answer to RQ2)

RQ2: To what extent can the proposed system maintain sound localization accuracy when the target user is walking?

In this experiment, the maintenance of sound localization accuracy during walking was evaluated by comparing three conditions: two-channel loudspeakers, fixed PAL, and tracking PAL. While the overall average RMSE indicated that conventional two-channel loudspeakers achieved the best results, differences in characteristics beyond simple accuracy rankings were revealed.

As shown in Fig. 4 and Tab. 1(a), the two-channel loudspeaker condition achieved the lowest average RMSE of 26.97 among the three conditions. This was primarily due to the exceptionally high localization accuracy at 0°. Additionally, because the experimental setup placed the loudspeakers at the 3.0 m endpoints, participants experienced increased sound pressure as they approached the speakers, thereby enhancing acoustic cues and contributing to improved accuracy. These results indicate that conventional loudspeakers excel at frontal localization and can provide stable localization within close proximity to the speakers (approximately 1.0 m in this experiment). However, the localization error varied greatly with distance, making it difficult to consistently present sound from a fixed direction to moving users.

As shown in Fig. 4 and Tab. 1(b), the fixed PAL condition exhibited the largest average RMSE of 39.08. This large error was particularly evident for the 0° direction. In this condition, when users passed through the sweet spot of the PAL beam (1.5–2.0 m), localization accuracy was high, but once they moved beyond this region, the source physically shifted behind them, leading to a sharp increase in error. For angles other than 0°, the error decreased as users approached the physical loudspeaker positions, similar to the two-channel condition. Thus, due to the highly restricted effective localization range, fixed PALs are also unsuitable for presenting sound images to moving users.

As shown in Fig. 4 and Tab. 1(c), the tracking PAL condition resulted in an average RMSE of 30.54, which was higher than that of two-channel loudspeakers. However, the key feature of this method was that the error variation remained relatively stable across the entire walking path, independent of the user's position. Unlike the other two conditions, where localization error fluctuated greatly with distance, the proposed method continuously tracked the user and maintained consistent sound pressure, thereby avoiding abrupt error changes. This explains why the overall RMSE was larger than that of the two-channel loudspeakers, as the error remained constant rather than being reduced near the speakers.

The violin plots provide further insights. In Fig. 5(a), the 45° and −45° conditions showed relatively small mean and median errors, yet the distributions were spread approximately ±45°, indicating that some participants perceived the sound as frontal or lateral at different times. Moreover, Tab. 1(b)(c) shows that at 0°, the tracking PAL achieved lower RMSE than the fixed PAL; however, Fig. 5(c) indicates that the tracking PAL distribution was wider. While most errors clustered around 0°, a few outliers degraded accuracy. These distribution issues are likely attributable to individual differences in HRTFs. Given the small sample size of four participants, outliers had a greater impact on the results.

Additionally, both Fig. 4 and Fig. 5 show that errors were particularly large for ±90° under all three conditions. Although some participants reported errors closer to 0°, suggesting the influence of outliers due to the small sample size, the overall trend remained consistent. A major factor contributing to the increased error is the small number of loudspeakers (two) used in this experiment. In contrast, Brungart et al. [32] evaluated walking sound localization using 64 loudspeakers and reported average errors below 9°. This comparison suggests that the particularly low lateral localization accuracy observed here was due to the limited number of loudspeakers, which made lateral localization more difficult than in multi-speaker systems.

From these findings, we conclude that the proposed tracking PAL system successfully overcomes the sweet-spot limitation of fixed PALs and achieves significantly improved localization accuracy. Compared to conventional two-channel loudspeakers, it yielded slightly larger average error but maintained nearly comparable accuracy while preventing sound diffusion to non-target users—an advan-

tage unique to parametric loudspeakers. Therefore, the proposed method represents a highly promising approach for public spaces and multi-user AR/MR environments where both noise reduction and accurate localization are desired. On the other hand, limitations such as reduced accuracy at ±90° and inter-individual variability in HRTFs highlight areas for future improvement.

6. Limitations and Future Work

This study demonstrated that the tracking PAL system is effective for selective acoustic intervention toward a moving target user and for maintaining stable sound localization. However, several limitations remain in both the evaluation and the system itself. Future work should address the following points.

First, there are challenges related to sound localization accuracy. In our experiments, the average RMSE of the proposed method was larger than that of two-channel loudspeakers. One reason is that this study represents a proof-of-concept stage, where the number of speakers and participants was limited. Moreover, the stability of the tracking PAL maintained a consistent baseline error, in contrast to conventional methods that showed extremely small errors near the speakers, which resulted in higher average error for our method. Beyond these factors, however, the fundamental causes of this baseline error remain unidentified. Possible contributing factors include system-wide latency from skeletal estimation by Kinect to loudspeaker motor actuation, as well as hardware limitations such as servo motor precision. As future work, these delays and hardware constraints should be quantitatively measured to identify the primary sources of error. Based on these findings, improvements such as faster and more precise tracking systems, or software-based compensation that predicts user motion to reduce latency, could be implemented. Additionally, experiments with larger numbers of speakers and participants will be necessary for more robust evaluation.

Second, the simplicity of the experimental environment poses limitations. The experiments were conducted in an anechoic chamber, free of acoustic reflections and external noise, under the simplified condition of linear walking. However, MR/AR environments and public spaces, which are the intended applications of this system, are acoustically complex, filled with noise and reverberation. Moreover, user movements in these contexts may include turning, stopping, and changing directions, beyond simple linear motion. In particular, when users move freely, the distance from the speakers can vary greatly, and if they move too far away, the perceived loudness may decrease. A potential solution is to install multiple PAL units at elevated positions such as the ceiling and dynamically switch or hand over the active speaker based on the tracked user's position. This multi-speaker handover approach would enable the system to maintain audibility and scalability in larger spaces without relying solely on distance compensation. Future research should therefore include evaluations in real-world environments such as offices and commercial facilities, as well as assessments of the system's tracking performance and localization accuracy under more complex user behaviors.

Finally, a limitation lies in target selection and switching in multi-user environments. In this study, target identification relied solely on Kinect's skeletal tracking. This approach was chosen for its robustness in detecting users even when faces were not visible, its efficiency with low computational load and real-time performance, and its anonymity in avoiding personal identification, thereby respecting privacy. Based on this policy, the system reassigns the target to the oldest detected ID when the current target leaves the detection area. However, this mechanism does not allow for intentional dynamic target selection. In crowded environments where users frequently enter and exit, maintaining selectivity becomes difficult. To address this issue, future work may explore intuitive interfaces such as gesture-based or gaze-based target switching. Furthermore, attention should also be given to the act of delivering sound itself. Potential directions include methods to reduce discomfort when sound unintentionally reaches non-target users, and approaches to deliver notifications perceivable only by the intended recipient. These aspects highlight opportunities for further exploration in the design of acoustic presentation.

7. Conclusion

This study addressed the challenge of delivering selective and stable spatial audio to specific moving users in dynamic environments with multiple people, such as public spaces and commercial facilities. To tackle this problem, we developed and evaluated a system that dynamically steers a pair of PALs toward a user's ears, based on real-time tracking with Kinect.

The evaluation of the proposed system yielded two key findings. First, regarding selective acoustic intervention (RQ1), the system successfully delivered sound exclusively to a specific walking user while minimizing sound leakage to surrounding individuals. Second, with respect to maintaining sound localization accuracy (RQ2), although the proposed method did not outperform conventional two-channel loudspeakers in terms of average error, it provided stable and consistent localization performance that was independent of user position, unlike existing methods.

In summary, the contribution of this study lies in demonstrating the effectiveness of a method that simultaneously fulfills two essential values in acoustic presentation for moving users: "selectivity"and "stability."Ensuring stability such that auditory information is not disrupted by user motion is critically important for all forms of dynamic acoustic interaction. The proposed tracking PAL system is expected to serve as a foundational technology for next-generation acoustic interfaces in dynamic and multi-user environments, including voice guidance in public spaces such as train stations to provide individualized navigation instructions while reducing noise, personalized advertisements in commercial facilities, exhibition and museum spaces offering visitor-specific audio explanations, and AR/MR experiences such as collaborative design sessions, remote maintenance support, or educational field trips in shared mixed reality environments where selective auditory presentation enhances immersion without disturbing others.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Gemini in order to: grammar and spelling checks, Paraphrase and translation. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. Wang, X. Bai, M. Billinghurst, S. Zhang, X. Zhang, S. Wang, W. He, Y. Yan, H. Ji, Ar/mr remote collaboration on physical tasks: a review, Robotics and Computer-Integrated Manufacturing 72 (2021) 102071.

[2] C. E. Mendoza-Ramírez, J. C. Tudon-Martinez, L. C. Félix-Herrán, J. d. J. Lozoya-Santos, A. Vargas-Martínez, Augmented reality: survey, Applied Sciences 13 (2023) 10491.

[3] M. Kobayashi, K. Ueno, S. Ise, The effects of spatialized sounds on the sense of presence in auditory virtual environments: a psychological and physiological study, Presence: Teleoperators and Virtual Environments 24 (2015) 163–174.

[4] N. Langiulli, M. Calbi, V. Sbravatti, M. A. Umiltà, V. Gallese, The effect of surround sound on embodiment and sense of presence in cinematic experience: a behavioral and hd-eeg study, Frontiers in Neuroscience 17 (2023) 1222472.

[5] C. Hendrix, W. Barfield, The sense of presence within auditory virtual environments, Presence: Teleoperators & Virtual Environments 5 (1996) 290–301.

[6] A. C. Kern, W. Ellermeier, Audio in vr: Effects of a soundscape and movement-triggered step sounds on presence, Frontiers in Robotics and AI 7 (2020) 20.

[7] I. Mavridou, E. Seiss, G. Ugazio, M. Harpster, P. Brown, S. Cox, F. Panchevski, C. Erie, D. Lopez Jr, R. Copt, et al., " did you hear that?": Software-based spatial audio enhancements increase self-reported and physiological indices on auditory presence and affect in virtual reality first author 1*, second author 2, third author 3, forth author 4, fifth author 1, sixth author 5, seventh author 5,

eighth author 4, nineth author 4, tenth author 5, eleventh author 4, twelfth author 4, thirteenth author 4*, fourteenth author 4, Frontiers in Virtual Reality 6 (2025) 1629908.

[8] N. Kuratomo, H. Uchida, T. Ebihara, N. Wakatsuki, K. Mizutani, K. Zempo, Spatialphonic360: Accuracy of the arbitrary sound image presentation using surrounding parametric speakers, in: Companion Proceedings of the 2022 Conference on Interactive Surfaces and Spaces, ISS Companion '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 32–36. URL: https://doi.org/10.1145/3532104.3571462. doi:10.1145/3532104.3571462.

[9] X. Su, J. E. Froehlich, E. Koh, C. Xiao, Sonifyar: Context-aware sound generation in augmented reality, in: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, 2024, pp. 1–13.

[10] D. Rumiński, An experimental study of spatial sound usefulness in searching and navigating through ar environments, Virtual Real. 19 (2015) 223–233. URL: https://doi.org/10.1007/s10055-015-0274-4. doi:10.1007/s10055-015-0274-4.

[11] S. Feng, W. He, X. Zhang, M. Billinghurst, S. Wang, A comprehensive survey on ar-enabled local collaboration, Virtual Reality 27 (2023) 2941–2966.

[12] B. Sonkoly, B. G. Nagy, J. Dóka, Z. Kecskés-Solymosi, J. Czentye, B. Formanek, D. Jocha, B. P. Gerő, An edge cloud based coordination platform for multi-user ar applications, Journal of Network and Systems Management 32 (2024) 40.

[13] N. Kuratomo, H. Miyakawa, T. Ebihara, N. Wakatsuki, K. Mizutani, K. Zempo, Attracting effect of pinpoint auditory glimpse on digital signage, IEEE Access 11 (2023) 42779–42794.

[14] M. Glaser, L. Hug, S. Werner, S. Schwan, Spatial versus normal audio guides in exhibitions: Cognitive mechanisms and effects on learning, Educational technology research and development 73 (2025) 169–198.

[15] N. Kuratomo, H. Miyakawa, S. Masuko, T. Yamanaka, K. Zempo, Effects of acoustic comfort and advertisement recallability on digital signage with on-demand pinpoint audio system, Applied Acoustics 184 (2021) 108359.

[16] N. Kuratomo, B. Karic, C. Kray, Explicit vs. implicit auditory displays for managing people flow in a pandemic: An exploratory study, Interacting with Computers (2025) iwaf008.

[17] M. Basner, W. Babisch, A. Davis, M. Brink, C. Clark, S. Janssen, S. Stansfeld, Auditory and non-auditory effects of noise on health, The lancet 383 (2014) 1325–1332.

[18] R. Thompson, R. B. Smith, Y. B. Karim, C. Shen, K. Drummond, C. Teng, M. B. Toledano, Noise pollution and human cognition: An updated systematic review and meta-analysis of recent evidence, Environment international 158 (2022) 106905.

[19] G. Theile, On the naturalness of two-channel stereo sound, Journal of the Audio Engineering Society 39 (1991) 761–767.

[20] M. Morimoto, Y. Ando, On the simulation of sound localization, Journal of the Acoustical Society of Japan (e) 1 (1980) 167–174.

[21] N. Kuratomo, H. Uchida, T. Ebihara, N. Wakatsuki, K. Mizutani, K. Zempo, Spatialphonic360: Accuracy of the arbitrary sound image presentation using surrounding parametric speakers, in: Companion Proceedings of the 2022 Conference on Interactive Surfaces and Spaces, ISS Companion '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 32–36. URL: https://doi.org/10.1145/3532104.3571462. doi:10.1145/3532104.3571462.

[22] H. Uchida, N. Kuratomo, T. Ebihara, N. Wakatsuki, K. Zempo, Spatialphonic360: Acoustic space for arbitrary sound image presentation based on both ears tracking, in: Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers, UbiComp/ISWC '22 Adjunct, Association for Computing Machinery, New York, NY, USA, 2023, p. 123–125. URL: https://doi.org/10.1145/3544793.3560319. doi:10.1145/3544793.3560319.

[23] T. Nishiura, High-realistic acoustic sound field reproduction: Research trend with parametric array loudspeaker, IEICE Fundamentals Review 10 (2016) 57–64.

[24] F. Fan, Y. Zhu, J. Yang, A grating lobe suppression method for a steerable parametric array loudspeaker, in: Proceedings of Meetings on Acoustics, volume 52, Acoustical Society of America,

2023, p. 055002.
[25] S. Kinjo, S. Fujiwara, T. Fujioka, Y. Nagata, Parametric loudspeaker steering system using output pointing interface, IEICE Technical Report; IEICE Tech. Rep. 120 (2020) 45–49.
[26] T. Zhuang, S. Li, F. Niu, J.-X. Zhong, J. Lu, Generating localized audible zones using a single-channel parametric loudspeaker, arXiv preprint arXiv:2504.17440 (2025).
[27] T. Zhuang, J. Zhong, J. Lu, The feasibility of sound zone control using an array of parametric array loudspeakers, in: 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, 2024, pp. 66–70.
[28] M. Nakayama, T. Ekawa, T. Takahashi, T. Nishiura, Virtual sound source construction based on direct-to-reverberant ratio control using multiple pairs of parametric-array loudspeakers and conventional loudspeakers, Applied Sciences 15 (2025) 3744.
[29] Y. Deng, K. Chen, H. Li, J. Zhang, Matched standard samples method in laboratory listening tests for annoyance perception, Applied Acoustics 224 (2024) 110103.
[30] D. Yunyun, L. Hao, D. Bo, L. Jianben, et al., The white noise standard sample method and application for subjective noise evaluation, Xibei Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University 40 (2022) 746–754.
[31] S. Aoki, M. Toba, N. Tsujita, Sound localization of stereo reproduction with parametric loudspeakers, Applied Acoustics 73 (2012) 1289–1295.
[32] D. S. Brungart, S. E. Kruger, T. Kwiatkowski, T. Heil, J. Cohen, The effect of walking on auditory localization, visual discrimination, and aurally aided visual search, Human factors 61 (2019) 976–991.