Audio Engineering Society

Convention Paper 10243

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.

Extending Realism for Digital Piano Players: A Perceptual Comparison of 3DoF and 6DoF Head-Tracked Binaural Audio

Pingkang Chen[1] and Braxton Boren[2]

[1] *Media Arts and Technology Program, University of California, Santa Barbara, CA, USA*

[2] *Audio Technology Program, American University, Washington, DC, USA*

Correspondence should be addressed to Author (pingkang@ucsb.edu)

ABSTRACT

This paper investigates how to achieve and perceptually evaluate the benefits of extending head-tracked binaural audio from three-degrees-of-freedom (3DoF) to six-degrees-of-freedom (6DoF) for digital piano players. Building on previous 3DoF research that tracked only rotational head movements, this study evaluates a 6DoF system that captures both rotational and translational movements during digital piano performance. A real-time HRTF-based binaural rendering system was developed using a custom plugin integrated with face-tracking technology in a Max/MSP–DAW environment to process 6DoF movement data alongside MIDI input from the digital piano. Perceptual evaluation comparing 3DoF and 6DoF tracking modes across five attributes: Dynamic Spatial Realism, Spatial Clarity, Source Stability, Envelopment, and Preference demonstrated significant improvements for 6DoF tracking in four of five attributes. These findings validate the importance of capturing pianists' natural translational head movements and demonstrate that 6DoF head-tracked binaural audio offers enhanced realism and immersion that more closely approximates the acoustic piano experience.

1 Introduction

Traditional playback methods in digital pianos suffer from significant spatial limitations that reduce the acoustic realism of piano performance. Built-in stereo speakers are widely used, but they offer only a fixed stereo image optimized for a narrow sweet spot and fail to adapt to listener movement. Headphone playback, while enabling silent practice, often produces a static and internalized sound image that disconnects the performer from the acoustic experience of playing a real instrument. A previous study [1] addressed these limitations by developing a three-degree-of-freedom (3DoF) head-tracked binaural rendering system that dynamically updated spatial cues based on the listener's yaw, pitch, and roll. Perceptual evaluations showed that this approach significantly enhanced realism and immersion compared to both conventional playback formats. However, while the system successfully accounted for rotational head movement, it did not incorporate translational motion—an important factor in expressive performance and spatial perception. This omission motivates the present study.

Beyond rotational head movement, pianists naturally engage in translational body motions that are deeply embedded in musical performance. From a psychological perspective, expressive playing often involves instinctive forward, backward, and lateral movements as part of emotional communication and interpretive nuance. These gestures are not incidental but reflect a performer's internal connection to phrasing, dynamics, and musical character. Empirical studies confirm that such bodily movements convey expressive intent to observers, even in the absence of sound [2], and are closely tied to internalized structural and affective interpretations of the music [3].

From a biomechanical and physiological perspective, pianists intentionally manipulate their upper body posture to achieve specific musical effects. For instance, leaning forward enhances projection and control in fortissimo passages, while leaning back

facilitates lighter touch and dynamic restraint in pianissimo playing [4]. These positional shifts serve as technical strategies that improve mechanical efficiency, optimize inter-joint coordination, and reduce muscular load during high-speed or forceful keystrokes [5, 6]. Proper use of torso movement, shoulder alignment, and seating posture also supports long-term ergonomic health and performance fluidity [7].

From a compositional standpoint, the architecture of many piano works necessitates full-body movement that extends beyond rotational head motion. Wide-ranging arpeggios, scalar runs, octave jumps, and dramatic register shifts demand the performer's torso and head to move laterally, vertically, and longitudinally along the keyboard. These spatial traversal patterns are not only a technical requirement but also reflect interpretive phrasing structures, often aligned with salient harmonic or melodic events [3, 8]. Gesture analyses show that such translational movements synchronize with phrase boundaries and structural climaxes, suggesting that they are integral to both performance and musical communication [3]. To faithfully capture these embodied performance dynamics, an auditory rendering system must reflect changes not only in head orientation but also in spatial position, as demonstrated in Figure 1.



Figure 1: Pianist upper-body movements during performance: (a) backward lean (−Y axis); (b) forward lean (+Y axis); (c) upright posture with vertical adjustment (Z axis); (d) lateral shift (X axis).

Building on previous findings that 3DoF head-tracked systems enhance spatial realism in piano performance, this study investigates whether incorporating full six-degree-of-freedom (6DoF) listener movement yields additional perceptual benefits. A comparative evaluation is conducted to assess whether extending from rotational to both rotational and translational head tracking further improves the realism and immersion of digital piano performance.

2 Methodology

2.1 Background for 6DoF Dynamic Rendering

Listener head movements in virtual auditory environments are described by six degrees of freedom (6DoF): three translational (X, Y, Z) and three rotational (yaw, pitch, roll), as illustrated in Figure 2. These motions alter the relative spatial relationship between the listener and sound sources, requiring continuous updating of source coordinates.

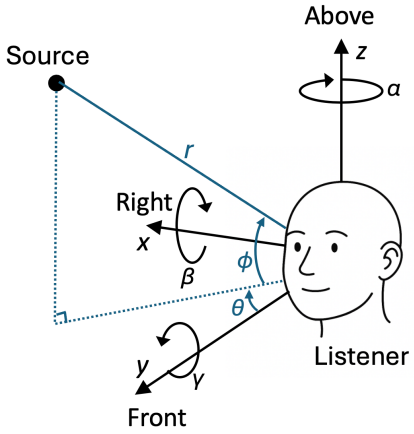


Figure 2. Coordinate system for listener's head.

The system recalculates the source location in spherical coordinates (r, θ, φ) based on current head position (Δx, Δy, Δz) and orientation (α, β, γ). This ensures accurate spatial rendering during both translational and rotational movements. The transformation can be expressed mathematically as:

$$\begin{bmatrix} r'\cos\varphi'\sin\theta' \\ r'\cos\varphi'\cos\theta' \\ r'\sin\varphi' \end{bmatrix} = \begin{bmatrix} \cos\alpha\cos y + \sin\alpha\sin\beta\sin y & -\sin\alpha\cos y + \sin\alpha\sin\beta\cos y & \cos\beta\sin y \\ \sin\alpha\cos\beta & \cos\alpha\cos\beta & -\sin\beta \\ -\cos\alpha\sin y + \sin\alpha\sin\beta\cos y & \sin\alpha\sin y + \cos\alpha\sin\beta\cos y & \cos\beta\cos y \end{bmatrix} \cdot \begin{bmatrix} r\cos\varphi\sin\theta - \Delta x \\ r\cos\varphi\cos\theta - \Delta y \\ r\sin\varphi - \Delta z \end{bmatrix} \quad (1)\ [9]$$

2.2 The Real-Time Rendering Engine

In this study, a real-time rendering system for virtual auditory environments was developed to synthesize binaural signals for digital piano performance. The application has been programmed in C++ using the JUCE audio programming framework [10]. For SOFA file reading and processing, the library Libmysofa [11] was used to extract HRTF data, including linear interpolation functions necessary for continuous spatial updates during listener movement. To ensure real-time performance, JUCE's built-in convolution engine—capable of partitioned convolution [12]—was employed. This method offers higher processing efficiency than the commonly used short-time Fourier transform (STFT) overlap-add convolution, particularly in the context of binaural room impulse response (BRIR) convolution where long reverberation tails are present. Moreover, the JUCE convolution engine inherently enables crossfading between impulse responses, which is essential for eliminating click artifacts during dynamic HRTF or BRIR switching—especially when simulating smooth spatial trajectories of moving sound sources. The audio rendering engine is controlled using standardized Open Sound Control (OSC) messaging, allowing for 6DoF tracking with real-time updates of both listener orientation and position. This modular plugin operates as a VST3 component within a host DAW and supports interaction with external tracking systems. The system was the same as that used in [13].

To model the rendering of virtual sources in reverberant conditions, the plugin incorporates direct sound, early reflections, and late reverberation to simulate binaural sound pressure. As shown in Figure 3, the input signal $e_0(t)$ is processed through three components:

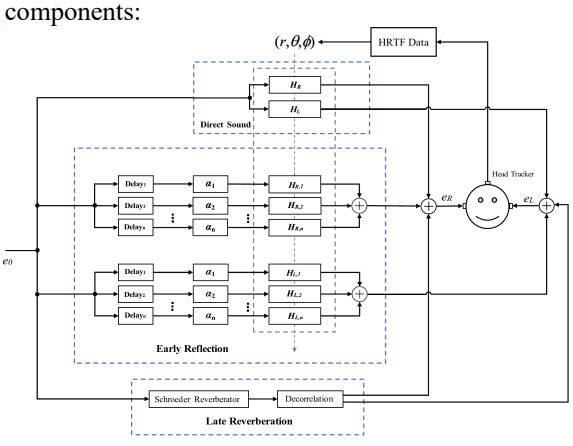


Figure 3. Signal flow diagram for dynamic reverberant rendering.

1: Direct sound is modeled by convolving the signal with the HRTF corresponding to the relative position (r, θ, ϕ) between the virtual source and the listener. In dynamic playback, the listener's head orientation is tracked in real time, and the HRTF lookup and convolution are updated accordingly. Head tracking data provides real-time yaw, pitch, and roll information, allowing the system to select appropriate HRTFs at different angles.

2: Early reflections are simulated using a simplified image-source method [14]. The first- and second-order reflections are treated as independent virtual sources. Given the room geometry and positions of the listener and source, the spatial coordinates of each image source are computed. Each reflection undergoes delay (Delay_i) and attenuation (α_i) to simulate wall absorption and distance-dependent energy decay. These reflections are then spatialized using corresponding HRTFs ($H_{L,i}$, $H_{R,i}$) and the outputs are summed for both ears. During dynamic playback, the image source positions are continuously recalculated based on real-time head-tracking data, ensuring accurate spatialization of reflections.

3: Late reverberation is synthesized using a Schroeder network [15] consisting of four parallel comb filters and two serial all-pass filters. To avoid coloration, the delay lengths in the comb filters are set to irregular, non-integer values. The reverberation tail beyond 80 ms is decorrelated for the left and right ears before being added to the binaural output, ensuring spatial realism. Similar to other perception-based methods for late reflection modeling in virtual auditory environments [16, 17], where late reverberation is treated as a diffuse sound field and directional information is considered perceptually insignificant [18], this plugin models late reverberation as a non-directional component driven solely by room acoustic parameters, independent of source or listener orientation.

The final binaural signal is the sum of all three components. Reverberation behaviour is shaped by user-defined room parameters, including length, width, height, absorption, and RT60.

To further simulate auditory distance cues, the plugin implements a hybrid model for near-field and far-field rendering based on source-listener distance. As illustrated in Figure 4, when the distance between the source and listener is within 1 meter, unnormalized near-field HRTFs are selected to preserve the natural spectral and amplitude variations that occur in close

proximity to the head. For distances beyond 1 meter, 1m far-field HRTFs are used with additional distance simulation effects: inverse-square law attenuation to model amplitude reduction with distance, and air absorption filters that simulate frequency-dependent attenuation of sound traveling through air, as far-field HRTFs are known to be distance-invariant [19, 20].

When BRIR is enabled, this selection process extends conceptually to early reflections, where each reflection point is treated as a new source position requiring appropriate HRTF selection based on its distance from the listener. A final propagation delay is applied to the output based on source-listener distance and speed of sound. This comprehensive approach enhances perceptual accuracy across the plugin's operational range, particularly in reverberant environments where both direct sound and reflections contribute to distance perception.

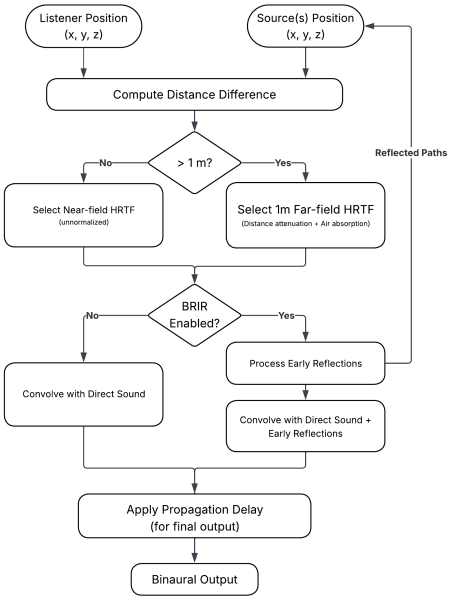


Figure 4. Distance-based HRTF selection and processing flowchart

Based on the rendering architecture described above, a custom graphical user interface (GUI) was developed to provide real-time control over all parameters of the plugin, as shown in Figure 5. The interface allows users to adjust the source position using spherical coordinates (azimuth, elevation, and distance), enable or disable head-tracked playback, and configure the OSC port for communication with external devices. Additionally, users can toggle BRIR simulation and modify room geometry and acoustic parameters, including room dimensions, RT60, and wall absorption. A 3D preview panel displays the relative position of the listener and sound source within a virtual environment, providing spatial feedback during playback.



Figure 5. UI of the developed plugin '6DoF Binauralizer'.

2.3 Technical Implementation of the 6DoF Digital Piano System

In this system, each sound source corresponds to one of the 88 hammer striking points on the piano strings. To establish these source positions, a tape measure was used to obtain the X, Y, and Z coordinates of all 88 hammer striking points on their strings from a well-regulated grand piano. The Cartesian coordinate origin (0,0,0) was established at the head position with the pianist seated in a neutral posture (sit up straight), with the nose aligned with the middle C key and abdomen 20 cm away from the keyboard front edge, placing the middle C hammer striking point at the lateral center (X = 0). These coordinates were then converted to spherical coordinates and rounded to one decimal place for use as the positional parameters of each source sound.

The complete signal flow of the system is illustrated in Figure 6, following a left-to-right path from input capture to binaural audio output. A smartphone running a face tracking application captures the pianist's head movements across six degrees of freedom: three rotational dimensions corresponding to pitch, yaw, and roll, plus three translational dimensions along the X, Y, and Z axes. These tracking data streams are transmitted via Open Sound Control (OSC) protocol over WiFi to the computer.

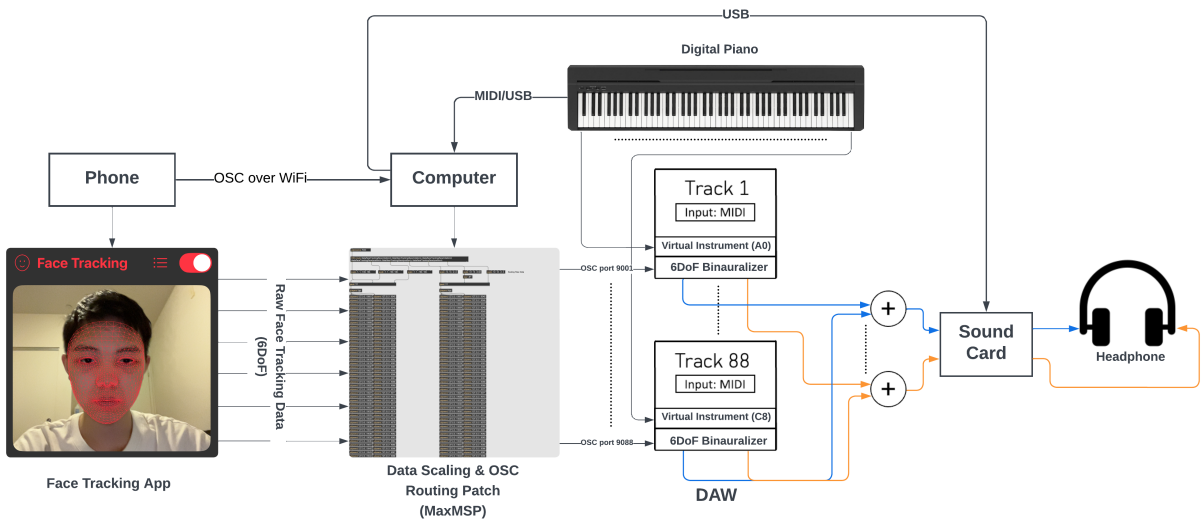


Figure 6. System architecture showing the complete signal flow from face tracking through spatial processing to binaural audio output.

A critical intermediate component is the custom Max/MSP patch that serves dual functions as both a data scaling processor and an OSC routing hub. This scaling process ensures that the raw tracking data properly aligns with the virtual movement parameters in the 6DoF Binauralizer plugin. The patch applies appropriate scaling factors to map the incoming tracking data values to appropriate ranges compatible with the binaural processing parameters.

The digital audio workstation houses 88 independent stereo tracks, each corresponding to a single piano key. Each track contains two plugins: a virtual piano instrument receiving MIDI input, followed by the custom 6DoF Binauralizer plugin. A crucial implementation aspect is the one-to-one mapping between physical keys and tracks. Each virtual instrument instance is configured with a keyboard range limited to a single MIDI note—Track 1 to A0, Track 2 to A#0/Bb0, and so forth. This ensures that when a key is pressed, only its corresponding track produces sound. The spatial coordinates of each hammer striking point are manually configured in each of their corresponding track's 6DoF Binauralizer plugin in DAW.

When the pianist plays the digital piano, MIDI messages are routed to the appropriate virtual instrument tracks. As the pianist moves their head during performance, the face tracking data continuously updates the position parameters in all 88 binauralization plugins simultaneously. This creates a dynamic spatial rendering where the perceived locations of different notes properly respond to both rotational and translational head movements. The final audio outputs from all tracks are mixed and sent to the sound card for headphone playback.

3 Experiment

3.1 Objectives

This study aims to compare two binaural audio tracking modes for playing digital pianos, focusing on determining whether the addition of translational tracking (6DoF) offers perceptual improvements over rotational tracking alone (3DoF) in terms of spatial realism, immersion, and auditory resemblance to playing an acoustic piano. The tracking modes evaluated are:

1: 3DoF tracking mode - rotational movements only: pitch, yaw, roll
2: 6DoF tracking mode - rotational and translational movements: pitch, yaw, roll, X, Y, Z

3.2 Participants

Twelve experienced pianists were recruited from the University of California, Santa Barbara (UCSB). All

participants had substantial piano experience (range: 8-15 years, mean=10.8, SD=2.3) and were currently active performers or students in music programs. The focus on experienced pianists was deliberate, as this population naturally incorporates significant translational movements during piano playing [3, 8], making them suitable candidates for evaluating the perceptual differences between 3DoF and 6DoF tracking modes.

All participants reported normal hearing (self-verified with no history of hearing disorders) and extensive experience with acoustic grand pianos. Each participant played the selected musical excerpts from memory to ensure they could focus entirely on their perceptual evaluation while naturally moving their head and upper body during playing.

3.3 Apparatus and Musical Stimuli

The experimental setup integrated multiple hardware and software components to enable comparative evaluation of 3DoF and 6DoF binaural audio rendering for digital piano performance.

3.3.1 Hardware Components

The primary instrument used was a Yamaha Arius YDP-181 digital piano with MIDI output capability. Head tracking was accomplished using an iPhone 13 running the Data OSC application, which provided both rotational and translational tracking data of the performer's face via the iPhone's front-facing camera. This single tracking solution was used for both experimental conditions, with selective data streams enabled or disabled to achieve 3DoF or 6DoF tracking. Audio was processed through an Audient id22 audio interface and delivered to participants via Beyerdynamic DT 1990 Pro open-back circumaural headphones, as shown in Figure 7.



Figure 7. A participant performing during the perceptual evaluation using the digital piano and 6DoF head-tracked binaural audio system.

The processing and rendering were handled by an HP Victus 16.1 computer running Windows 11.

3.3.2 Software Architecture

The software implementation consisted of four main components. First, the REAPER digital audio workstation was configured with 88 discrete audio tracks, allowing for independent spatial processing of each piano key. Second, the Numa Player 2 virtual instrument plugin (using the "Model D 1983" acoustic piano sound) was employed for high-quality piano sound generation. Third, a custom Max/MSP patch was developed to handle OSC data routing and scaling from the face tracking system. Fourth, the "6DoF Binauralizer" plugin was implemented for the binaural rendering, with one instance inserted on each of the 88 tracks to provide individualized spatial processing for each piano key. The plugin handled both the 3DoF and 6DoF head tracking data to render the piano sounds with appropriate spatial cues according to the current tracking mode, introducing approximately 3.6 ms of processing latency at a buffer size of 512 samples (44.1 kHz sample rate).

3.3.3 Stimuli

Participants performed five musical excerpts, each selected to highlight specific perceptual dimensions:

1: Dynamic Spatial Realism:

• Hanon's "The Virtuoso Pianist" Exercise No. 39 - C major scale (played twice) followed by C major arpeggio (played twice) - Systematic keyboard traversal and varied spatial patterns naturally induce comprehensive lateral body movement for assessing auditory scene responsiveness to translational player movements.

2: Spatial Clarity:

• J.S. Bach's "Prelude and Fugue in C Major" (BWV 846, first 60 seconds) - Multiple independent melodic voices moving between registers create clear opportunities to evaluate source localization accuracy within the sound field.

3: Source Stability:

• Mozart's "Piano Sonata No. 16 in C Major" (K. 545, first movement, first 60 seconds) - Familiar and repetitive thematic material with moderate register changes allows assessment of whether sound sources maintain consistent spatial positions during player movement.

4: Envelopment:

• Chopin's "Nocturne in E-flat Major" (Op. 9, No. 2, first 60 seconds) - Rich harmonies and flowing melodies with dynamic variations provide an ideal test for spatial immersion and sound field surroundness.

5: Preference:

• "The Maiden's Prayer" by Tekla Bądarzewska-Baranowska (first 60 seconds) - Diverse musical textures including melodies, scales, chords, and arpeggios provide a comprehensive test for overall system preference.

3.4 Procedure

The experiment followed a within-subjects design with counterbalanced presentation order to control for potential learning and fatigue effects. Five perceptual attributes were assessed, each precisely defined to participants:

• Dynamic Spatial Realism—How well the auditory scene adjusts to both rotational and translational player movements, maintaining spatial coherence across all axes.

• Spatial Clarity—The degree to which each note or sound source can be accurately localized within the sound field, regardless of head movement.

• Source Stability—The extent to which sound sources remain perceptually stable during player movement, particularly during lateral and vertical translations.

• Envelopment—The degree to which the player feels surrounded and immersed by the sound field, considering both horizontal and vertical spatial cues.

• Preference—The overall perceptual preference rating, allowing participants to indicate which mode provides a more immersive and convincing auditory experience.

Participants rated each criterion on a 5-point Likert scale, where 1 = "Not at all" and 5 = "Extremely." Each participant completed a 20-minute familiarization session to understand the experiment's objectives and evaluation criteria, practicing with both tracking modes and reviewing the evaluation form and rating scale. Before each experimental condition, a standardized calibration procedure was performed where the participant was seated at the digital piano in a neutral playing position, the tracking system was zeroed with the participant facing the keyboard center, the listener position parameters (X, Y, Z) in each 6DoF Binauralizer plugin were reset to 0,0,0 corresponding to the origin head position used during acoustic piano measurement, and the face tracking range was verified to cover participants' normal 6DoF movement while seated at the piano.

For each participant, the experimental sequence proceeded as follows: they performed the five musical excerpts using the first tracking mode and completed evaluation forms immediately after each performance, took a short break (5 minutes), then repeated the process with the second tracking mode. The order of tracking modes (3DoF vs. 6DoF) was counterbalanced across participants, with six participants testing the 3DoF mode first and six testing the 6DoF mode first to control for potential order effects. Participants were informed which tracking mode they were experiencing during each condition.

Switching between conditions was implemented in the Max/MSP patch by enabling or disabling translational tracking data: in the 3DoF mode, only rotational movements influenced the rendering, while in the 6DoF mode, both rotational and translational movements were applied. For both testing conditions, identical acoustic simulation parameters were applied across all 88 instances of the 6DoF Binauralizer plugin (room dimensions: 10 × 10 × 5 meters, RT60: 0.5 seconds) to ensure perceived differences were attributable solely to the tracking mode. The DAW session operated at 44.1 kHz, 24-bit using ASIO drivers. Throughout the experiment, participants were instructed to play naturally and expressively with no constraints on their physical movements.

4 Results and Discussion

4.1 Results

The perceptual evaluation revealed consistent differences between 3DoF and 6DoF tracking modes across all evaluation criteria, with 6DoF showing higher ratings in all categories. Figure 8 presents the mean ratings and standard errors for each perceptual attribute under both tracking conditions.

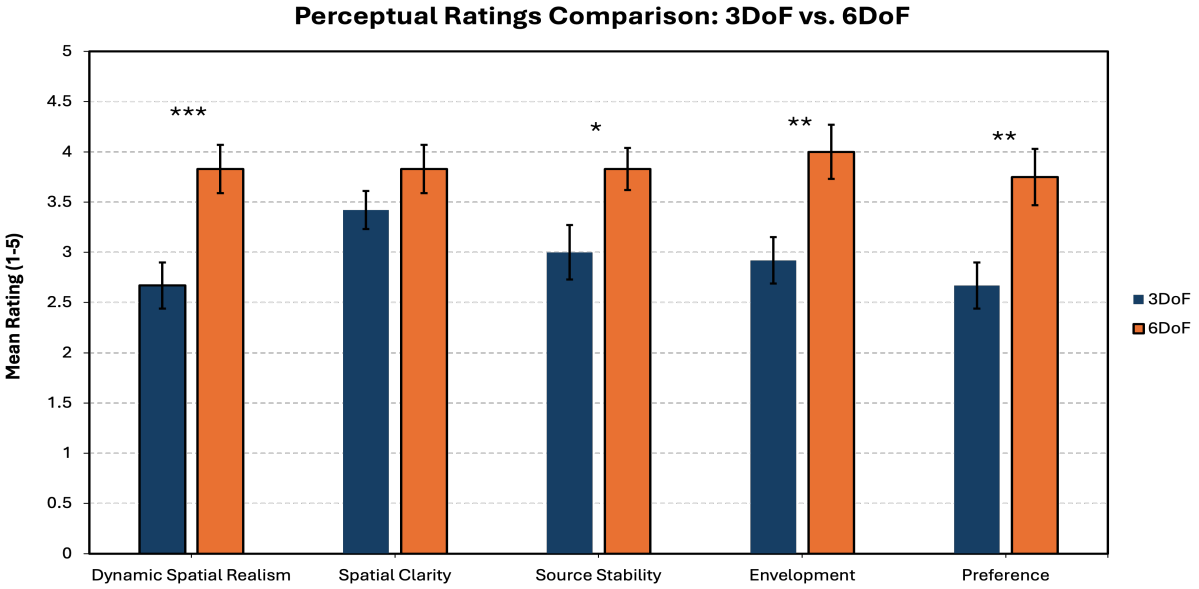


Figure 8. Mean ratings of perceptual attributes for 3DoF and 6DoF head-tracked binaural audio. Error bars represent ±1 standard error of the mean (n=12). Asterisks indicate statistically significant differences between conditions: *p < 0.05, **p < 0.01, ***p < 0.001.

The 6DoF tracking mode consistently outperformed the 3DoF mode across all attributes. The largest improvements were observed in Dynamic Spatial Realism (mean difference = 1.16), Envelopment (mean difference = 1.08), and overall Preference (mean difference = 1.08). Source Stability showed a moderate improvement (mean difference = 0.83), while Spatial Clarity demonstrated the smallest difference between conditions (mean difference = 0.41). These patterns suggest that 6DoF tracking particularly enhances aspects related to spatial responsiveness and immersion rather than basic localization accuracy.

4.2 Statistical Analysis

To assess the perceptual impact of incorporating translational tracking in addition to rotational tracking, both a two-way repeated-measures ANOVA and paired-sample t-tests were conducted. The ANOVA tested for main effects of Tracking Mode (3DoF vs. 6DoF) and Perceptual Attribute (five attributes), as well as their interaction. The results are summarized in Table 1.

Source	Sum of Squares	DF	Mean Square	F Value	P
Corrected Model	30.175[a]	9	3.353	4.826	.000
Tracking Mode (3DoF vs. 6DoF)	25.208	1	25.208	36.287	.000
Perceptual Attribute	2.717	4	.679	.978	.423
Tracking Mode × Perceptual Attribute	2.250	4	.563	.810	.522
Error	76.417	110	.695		
Total	1487.000	120			
Corrected Total	106.592	119			

Table 1. Results of the two-way ANOVA showing the effects of Tracking Mode and Perceptual Attribute on perceptual ratings.

The two-way ANOVA revealed a significant main effect of Tracking Mode, $F(1, 11) = 36.29$, $p < 0.001$, indicating that participants rated the 6DoF condition significantly higher overall than the 3DoF condition. There was no significant main effect of Perceptual Attribute, $F(4, 44) = 0.98$, $p = 0.423$, suggesting that the relative difference between attributes was consistent. Additionally, no significant interaction was found between Tracking Mode and Perceptual Attribute, $F(4, 44) = 0.81$, $p = 0.522$, indicating that the benefit of 6DoF tracking was relatively uniform across attributes.

Paired-sample t-tests revealed significant improvements for the 6DoF condition in four of five attributes:

• **Dynamic Spatial Realism**: 3DoF: M = 2.67 (SE = 0.23), 6DoF: M = 3.83 (SE = 0.24); t(11) = 4.86, p < 0.001 (***), Mdiff = 1.16

• **Source Stability**: 3DoF: M = 3.00 (SE = 0.27), 6DoF: M = 3.83 (SE = 0.21); t(11) = 2.87, p = 0.015 (*), Mdiff = 0.83

• **Envelopment**: 3DoF: M = 2.92 (SE = 0.23), 6DoF: M = 4.00 (SE = 0.27); t(11) = 3.72, p = 0.003 (**), Mdiff = 1.08

• **Preference**: 3DoF: M = 2.67 (SE = 0.23), 6DoF: M = 3.75 (SE = 0.28); t(11) = 3.91, p = 0.002 (**), Mdiff = 1.08

• **Spatial Clarity**: 3DoF: M = 3.42 (SE = 0.19), 6DoF: M = 3.83 (SE = 0.24); t(11) = 1.69, p = 0.118 (ns), Mdiff = 0.41

4.3 Discussion

The results provide compelling evidence for the perceptual advantages of 6DoF head-tracked binaural audio over 3DoF tracking for digital piano performance. The significant main effect of tracking mode ($p < 0.001$) demonstrates that incorporating translational movement tracking substantially enhances the spatial audio experience for pianists.

Enhanced spatial responsiveness: Dynamic Spatial Realism showed the largest and most significant improvement (Mdiff = 1.16, $p < 0.001$), directly validating the core hypothesis that accounting for translational movements enhances perceived spatial responsiveness. This finding aligns with the theoretical framework that pianists' forward/backward leaning and lateral movements are integral to their spatial perception during performance.

Improved immersion and user engagement: Both Envelopment (Mdiff = 1.08, p = 0.003) and overall Preference (Mdiff = 1.08, p = 0.002) demonstrated substantial improvements with identical effect sizes. These results indicate that 6DoF tracking creates a more immersive and satisfying auditory experience that translates directly into increased user engagement with the digital piano system.

Better source coherence: The significant improvement in Source Stability (Mdiff = 0.83, p = 0.015) demonstrates that 6DoF tracking better maintains perceptual coherence of piano keys during performer movements. This is particularly crucial for expressive piano performance, where maintaining stable localization of different registers during dynamic body movements supports musical interpretation and technical precision.

Localization accuracy findings: Spatial Clarity was the only attribute that did not reach statistical significance (p = 0.118), despite showing directional improvement. This suggests that basic sound source localization was already well-established with 3DoF tracking alone. The primary benefits of 6DoF tracking appear to lie in dynamic spatial responsiveness and immersive qualities rather than fundamental localization accuracy, which is consistent with the complementary roles of rotational versus translational spatial cues.

Technical validation: The consistent improvements across multiple perceptual dimensions confirm that the 6DoF system successfully integrated both rotational and translational tracking through camera-based face tracking. Additionally, no participants reported any noticeable latency or artifacts during the experiment, confirming that the specific combination of software and hardware used in this system was effective in maintaining latency below perceptual thresholds. This indicates robust technical implementation without introducing performance issues that could degrade the user experience.

Study limitations: One limitation was the selective deactivation of piano tracks in the DAW based on musical content—for pieces utilizing only middle registers, outer register tracks were disabled to reduce computational load. Future implementations could address this by employing distributed processing across multiple computers, ensuring full 88-key capability regardless of musical content.

The results demonstrate that 6DoF head-tracked binaural audio represents a meaningful advancement in digital piano technology, offering enhanced realism and engagement that more closely approximates the natural spatial experience of acoustic piano performance.

5 Conclusion and Future Work

This study extends prior work on head-tracked binaural rendering by evaluating the perceptual benefits of incorporating six degrees of freedom

(6DoF) tracking into digital piano performance. By capturing both rotational and translational head movements, the custom 6DoF rendering system significantly improved participants' perceptual experiences compared to the 3DoF baseline.

Twelve experienced pianists evaluated both tracking modes across five perceptual attributes. The 6DoF condition produced statistically significant improvements in four of the five dimensions, with the most substantial gains observed in Dynamic Spatial Realism, Envelopment, and overall Preference. These findings highlight the perceptual importance of translational cues—such as forward/backward leaning and lateral shifts—that are naturally embedded in expressive piano playing.

From a technical standpoint, this study demonstrated the feasibility of real-time 6DoF binaural rendering using off-the-shelf components, custom Max/MSP integration, and a low-latency plugin-based architecture. The system consistently outperformed the 3DoF configuration without introducing artifacts or compromising spatial coherence.

Together, these results establish 6DoF head-tracked binaural audio as a meaningful advancement in spatial audio for virtual instruments, offering a heightened sense of realism that more closely approximates the embodied spatial experience of acoustic piano performance.

Building on the success of this single-performer 6DoF implementation, future research will focus on extending the system to multi-user contexts. One promising direction involves integrating the current 6DoF framework with advanced acoustic modeling environments and real-time networking architectures to enable immersive multi-performer scenarios. This would allow multiple musicians, each tracked under 6DoF and wearing HMDs, to rehearse or perform together in a shared virtual space auditory and visually—despite being physically located in different places.

In parallel, future implementations will explore integration with networked music performance (NMP) systems, enabling low-latency spatial interaction between geographically distant performers. Coupled with a synchronized visual representation in VR or MR environments, this would establish a fully immersive, 6DoF-based, multi-user platform for spatialized music performance with headphones. This system would maintain reciprocal spatial relationships, ensuring that each performer hears other participants' sound sources rendered binaurally from their own 6DoF perspective. Such an extension has the potential to support a wide range of novel applications, including immersive remote rehearsals for ensembles and distributed bands.

These directions aim to push the boundaries of immersive audio by scaling 6DoF rendering from solo use cases to collaborative, networked experiences, ultimately contributing to the development of next-generation spatial audio systems for expressive and embodied musical interaction.

References

[1] P. Chen and B. Boren, "Enhancing realism for digital piano players: A perceptual evaluation of head-tracked binaural audio," in *Audio Engineering Society Convention 157*, Audio Engineering Society, 2024.

[2] J. W. Davidson, "Visual perception of performance manner in the movements of solo musicians," Psychology of Music, vol. 21, no. 2, pp. 103–113, 1993.

[3] H. Shoda and M. Adachi, "The role of a pianist's affective and structural interpretations in his expressive body movement: A single case study," Music Perception, vol. 29, no. 3, pp. 237–254, 2012.

[4] B. James, "Pianism: Performance communication and the playing technique," Frontiers in Psychology, vol. 9, p. 2125, 2018.

[5] S. Furuya and H. Kinoshita, "Expertise-dependent modulation of muscular and non-muscular torques in multi-joint arm movements during piano keystroke," Neuroscience, vol. 156, no. 2, pp. 390–402, 2008.

[6] S. Furuya, T. Goda, H. Katayose, H. Miwa, and N. Nagata, "Distinct inter-joint coordination during fast alternate keystrokes in pianists with superior skill," Frontiers in Human Neuroscience, vol. 5, p. 50, 2011.

[7] G. Sandor, On Piano Playing: Motion, Sound, and Expression. New York: Schirmer Books, 1981.

[8] J. MacRitchie, B. Buck, and N. J. Bailey, “Inferring musical structure through bodily gestures,” Musicae Scientiae, vol. 17, no. 1, pp. 86–108, 2013.

[9] Zhang, C., Xie, B. Platform for dynamic virtual auditory environment real-time rendering system. Chin. Sci. Bull. 58, 316–327 (2013). https://doi.org/10.1007/s11434-012-5523-2

[10] *JUCE Audio Programming Library*, Accessed:2025-05-20,https://www.juce.com/.

[11] *Libmysofa Library*, Accessed: 2025-05-20, https://github.com/hoene/libmysofa.

[12] Wefers, Frank. Partitioned convolution algorithms for real-time auralization. Vol. 20. Logos Verlag Berlin GmbH, 2015.

[13] P. Chen, “Auditory distance perception in HRTF-based binaural rendering: Effects of head tracking and reverberation,” in *Audio Engineering Society Convention 159*, Audio Engineering Society, 2025.

[14] Jont B. Allen, David A. Berkley; Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 1 April 1979; 65 (4): 943–950. https://doi.org/10.1121/1.382599

[15] M. R. Schroeder, "Natural-sounding artificial reverberation," Journal of the Audio Engineering Society, vol. 10, no. 3, pp. 219-223, 1962.

[16] Csaba Huszty, Bottyán Németh, Péter Baranyi, Fülöp Augusztinovicz; Measurement-based fuzzy interpolation of room impulse responses. J. Acoust. Soc. Am. 1 May 2008; 123 (5_Supplement): 3771. https://doi.org/10.1121/1.2935382

[17] Kearney, G., Masterson, C., & Boland, F. (2009). Acoustic Impulse Response Interpolation for Multichannel Systems using Dynamic Time Warping. In Audio Engineering Society 35th Conference

[18] Xie, Bosun. Head-related transfer function and virtual auditory display. J. Ross Publishing, 2013.

[19] Li, S.; Peissig, J. Measurement of Head-Related Transfer Functions: A Review. Appl. Sci. 2020, 10, 5014. https://doi.org/10.3390/app10145014

[20] Brungart, D. S., & Rabinowitz, W. M. (1999). Auditory localization of nearby sources. Head-related transfer functions. The Journal of the Acoustical Society of America, 106(3), 1465-1479