Audio Engineering Society

Convention Paper 10199

Presented at the AES 157th Convention

2024 October 8-10, New York, NY, 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.

Enhancing Realism for Digital Piano Players: A Perceptual Evaluation of Head-Tracked Binaural Audio

Pingkang Chen[1], Braxton Boren[1]

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

Correspondence should be addressed to Pingkang Chen (pc1026a@student.american.edu)

ABSTRACT

This paper outlines a process for achieving and perceptually evaluating a head-tracked binaural audio system designed to enhance realism for players of digital pianos. Using an Ambisonic microphone to sample an acoustic piano, followed by leveraging off-the-shelf equipment, the system allows players to experience changes in the sound field in real-time as they rotate their heads while wearing headphones under three degrees of freedom (3DoF). The evaluation criteria included spatial clarity, spectral clarity, envelopment, and preference. These criteria were assessed across three different listening systems: stereo speakers, stereo headphones, and head-tracked binaural audio. Results showed a strong preference for the head-tracked binaural audio system, with players noting significantly greater realism and immersion.

1 Introduction

Digital pianos, designed to simulate their traditional acoustic counterparts, have continually evolved in recent years. This development primarily stems from advancements in sampling technology and key bed mechanics. Consequently, the market demand for digital pianos has gradually increased, inversely correlating with the demand for acoustic pianos, as evidenced by sales statistics in the United States from 2005 to 2021 [1,2]. One of the major advantages of digital pianos, beyond their portability, lower maintenance costs, and more affordable prices, is the ability to play through headphones when it is necessary to avoid using its built-in loudspeakers. This feature effectively prevents disturbances to others and is particularly beneficial for residents in multi-unit apartment buildings where sound insulation may be inadequate. However, both the built-in speaker systems and the current headphones system have limitations.

To better understand these limitations, consider an acoustic piano, which has 88 independent strings, each functioning like a directional 'speaker'. When a note is played, only one ‘speaker' produces sound, and any head movement by the player results in changes in the sound field due to variations in Interaural Time Difference (ITD) and Interaural Level Difference (ILD) associated with that specific 'speaker'.

In contrast, most digital pianos use two built-in speakers to playback sound, one on each side (left and right), to pan all the notes through stereo imaging. This means that when the player plays a note and moves their head, the volume ratio between the left and right speakers remains constant, as it was preset from the sweet spot when the head is fixed. Consequently, this setup only provides the correct sound field when heard from the sweet spot, rather than allowing for a changeable volume ratio between the two speakers in real-time according to the player's head direction.

Additionally, when players use headphones with the built-in speakers turned off, they hear the stereo format. They hear only the stereo version of the music being played with a fixed frontal image, and the entire soundscape rotates with them when they rotate their head, thus failing to replicate a true sound field. Consequently, sounds heard through headphones are generally perceived as internalized.

Accordingly, this paper presents a design for a more immersive and realistic sound environment that dynamically adjusts to the head movements of the player under three degrees of freedom (3DoF), closely mimicking the authentic sound experience of playing a grand piano when playing a digital piano with headphones on.

2 Background

2.1 Binaural Sampling

Binaural recording, the method of capturing three-dimensional audio experiences by simulating the perceptual cues of human hearing, involves the use of two microphones placed in a way that mimics the human ear's natural listening experience [3]. This technique captures the spatial characteristics of sound, including direction, distance, and the natural filtering effects of the head, torso, and pinna, characterized by the head-related transfer function (HRTF) [4]. This technique has been recently employed to sample acoustic pianos, allowing players to play and hear the recorded binaural signals through headphones. This is achieved by employing a binaural microphone positioned from the player's perspective to record the sound of each key on the piano, as shown in Figure 1 where a dummy head microphone was used. The dummy head captures the true path of sound propagation as it travels to the player's ears, providing a natural and realistic auditory experience. By utilizing binaural recording technology, the sound heard through headphones encompasses a full 360-degree sound field, offering significant spatial enhancement over traditional stereo playback [4].



Figure 1. Using a dummy head to sample an upright piano [5].

Despite its advantages, binaural sampling still has limitations. One notable limitation is that when a listener rotates their head, the entire soundscape rotates with them [6]. This occurs because using a dummy head for sampling captures sound from a fixed direction, resulting in a static sound field that replicates auditory experiences from only one specific perspective. In contrast, pianists often need to move their heads to look at the keys, especially when playing pieces with extensive notes in either the low or high registers, prompting them to turn their heads to better locate the keys. Additionally, they may tilt or lift their heads during expressive musical passages.

Another limitation is that the dummy head used for recording may not match the anatomical features of all listeners. Because head-related transfer functions (HRTFs) are unique to everyone, the HRTF of the dummy head may not accurately represent the auditory characteristics of all listeners, meaning a recording may only sound optimal to a limited number of people [6]. Consequently, the static binaural sound field does not fully replicate the dynamic auditory experience of playing an acoustic piano [7].

2.2 Ambisonics Recording to Head-Tracked Binaural

To address these limitations, Ambisonic-based binaural rendering was later developed to deliver a dynamic and personalized auditory experience. This approach was well documented by McKeag and McGrath in 1996, using 1st order Ambisonic recordings to feed head-tracked binaural audio over headphones [6].

One notable method within Ambisonic-based binaural rendering is 'Virtual Ambisonics Approximation' [8], which organically integrates Ambisonics and head-tracking. This method is based on the idea of decoding Ambisonics to virtual loudspeakers and has been integrated to combine convolutions with computationally efficient spherically decomposed HRTF sets for binauralizing Ambisonic sound fields, effectively reducing computational costs.

2.2.1 Ambisonic Recording and Encoding

For Ambisonic recording, the original method developed by Gerzon [9] involves using a tetrahedral

microphone. This setup contains four cardioid or subcardioid capsules mounted very close together at the vertices of a tetrahedron and pointing outwards, as seen in Figure 2. Since a cardioid microphone can be broken down into an omnidirectional component and a figure-eight component (cardioid mic = 1/2 (omni mic + figure-of-8 mic)), the four cardioid microphones in Gerzon's tetrahedral array, being coincident, can collectively be considered as one omnidirectional microphone and three figure-eight microphones oriented along the x-axis, y-axis, and z-axis [10].

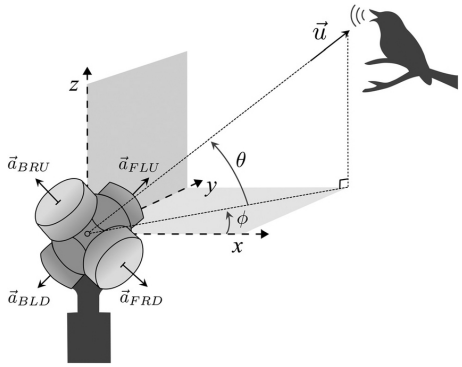


Figure 2. A tetrahedral mic [11].

When the sound field is being recorded, all four microphone capsules are active simultaneously, each capturing sound from their respective positions onto separate tracks referred to as FLU (Front-Left-Up), FRD (Front-Right-Down), BLD (Back-Left-Down), and BRU (Back-Right-Up) [10]. This recorded audio (the group of four raw signals from the microphones) is termed "A-format" and does not easily translate to intuitive spatial information. In order to recreate the sound field and to hear and work with the spatial information, signals from each capsule have to be encoded into the Ambisonics channels, which are often referred to as B-format signals, through:

$$W = \text{FLU} + \text{FRD} + \text{BLD} + \text{BRU} \quad (1)$$
$$X = \text{FLU} + \text{FRD} - \text{BLD} - \text{BRU}$$
$$Y = \text{FLU} - \text{FRD} + \text{BLD} - \text{BRU}$$
$$Z = \text{FLU} - \text{FRD} - \text{BLD} + \text{BRU}$$

These four Ambisonics channels comprise the first-order directional information: X (front/back), Y (left/right), Z (up/down), and W, an omnidirectional 0th-order [10]. Additionally, it is necessary to add some filtering to correct for the fact that not all four capsules are coincident, which is relevant at high frequencies.

2.2.2 Ambisonics Scene Rotation

When the user rotates their head, the tracking system captures the movement in three directions: pitch, yaw, and roll, which allows the Ambisonics B-format to be easily rotated. Under an arbitrary rotation, characterized by a rotation matrix R, the B-format is transformed as follows:

$$W' = W \quad (2)$$
$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \mathbf{R} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

The W channel (omnidirectional component) remains unchanged, while the X, Y, and Z channels (figure-of-eight components) are adjusted with coefficients based on the head-tracking data, effectively rotating the Ambisonic sound field in the opposite direction of the head movements. This process ensures that the positions of the virtual loudspeakers, simulated by HRTFs, remain fixed, while the audio signals sent to these virtual loudspeakers are updated in real time to reflect the listener's current orientation. This eliminates the need for continuous computational convolution and intensive interpolation across a large dataset of HRTFs or BRIRs, greatly increasing computational efficiency [12].

2.2.3 Ambisonics Decoding to Headphones

For the next step in the process, the Ambisonic signals, once rotated to compensate for head movements, are fed into an Ambisonic decoder. The role of the Ambisonic decoding is to transform these encoded Ambisonic channels into a set of virtual loudspeakers (SPK1 to SPK(N)) simulated by HRTFs, each corresponding to a specific direction in the Ambisonic sound field. These decoded signals are then ready for convolution with Head-Related Transfer Functions (HRTFs). The left and right ear signals, L(t) and R(t), are calculated as follows:

$$L(t) = \sum_i HRIR_L\,(t;\, \phi_{i,}\, \delta_n) * s_i(t), \quad (3)$$
$$R(t) = \sum_i HRIR_R\,(t;\, \phi_{i,}\, \delta_n) * s_i(t),$$

where $HRIR_L$ and $HRIR_R$ are the Head-Related Impulse Responses for the left and right ears, respectively, and $s_i(t)$ are the signals from the virtual loudspeakers.

As mentioned in section 2.1, one of the limitations of binaural recording is that the HRTF of the dummy head may not accurately represent the auditory characteristics of all listeners. Therefore, the use of individualized HRTFs ensures that the spatial and frequency-dependent content in binaural reproduction is accurately delivered, as perceived by each listener's ears. This point also applies to the use of HpEQs (Headphone Equalization), as correct HpEQ ensures that the spectral features of the HRTF are precisely reproduced at the eardrum during playback. The best choice is to use individualized HpEQs, which improve the consistency in distance perception and externalization compared to generic HpEQs [13, 14]. These filters ensure a flat target spectrum by canceling out the headphone transfer function (HpTF) through regularized inversion [15].

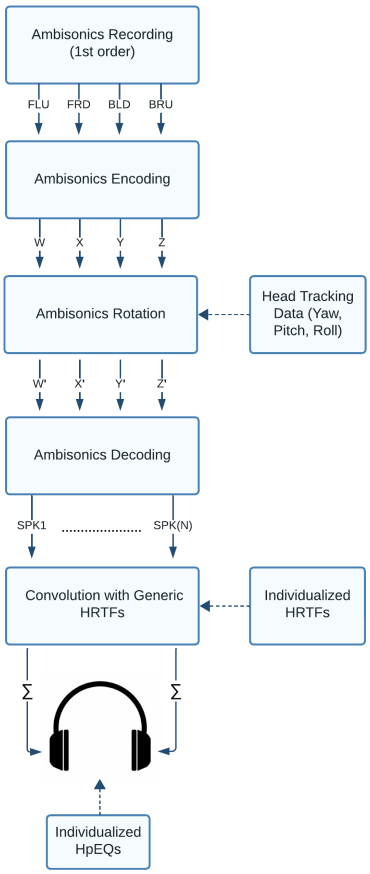


Figure 3. Block diagram illustrating audio signal chain for 1st-order Ambisonics recording to head-tracked binaural audio over headphones.

3 Experiment Design

3.1 Objectives

This study aims to compare three playback systems for playing digital pianos, focusing on determining whether head-tracked binaural audio can offer a superior experience for the players compared to the other two in terms of greater auditory realism. The playback systems evaluated are:"

1: Built-in speakers (stereo) on digital pianos.
2: Headphone (stereo)
3: Headphone (head-tracked binaural audio)

3.2 Implementation of Head-Tracked Binaural Audio System

In preparation for the official perceptual evaluation on the digital piano, the workflow detailed in section 2.2 was followed for achieving the head-tracked binaural playback system. This involved first using the Ambisonic microphone 'Zoom H3-VR' positioned in front of the piano at a height of approximately 1.2 meters, similar to the average height of human ears when seated on the piano bench for playing the piano, to sample the Yamaha GB1 baby grand piano from the pianist's perspective at a sample rate of 24bit/48kHz, as seen in Figure 4, in the Live Room A at Kreeger Studio, American University, Washington, D.C. Besides the directionality of each key being sampled, 8 velocity layers for each key (sustain pedal on/off) were captured at different dynamics, including *ppp, pp, p, mp, mf, f, ff, and fff.* Additionally, the noises associated with key trigger releases and pedal movements (both pressing down and lifting) were also sampled.



Figure 4. Sampling the Yamaha grand piano using the Ambisonics microphone 'Zoom H3-VR'

Since the mic preamplifier of 'Zoom H3-VR' already does the A to B format conversion in its hardware, all samples were recorded directly as B format audio. Finally, these recorded samples were loaded into the sampler plugin 'KONTAKT,' which serves as a bridge in the Digital Audio Workstation (DAW) Reaper for further spatial processing. This allowed the sampled sounds to be evaluated directly on the digital piano, which was connected to the computer via a MIDI cable.

For the employment of head-tracking, the plugin 'SceneRotater' from the 'IEM Plug-in Suite'[16] was used to rotate the scene. This rotation is controlled by an external head tracker from 'Supperware Ltd,' which can be attached to the headphone and sends OSC data to the plugin in real-time. Notably, the 'SceneRotater' plugin has been measured to have 0 samples of processing latency, ensuring smooth and responsive head-tracking. For the final step of decoding to headphones, the plugin 'BinauralDecoder' from IEM was used, which has a processing latency of 126 samples, equivalent to 2.63 ms [17].

Considering the time cost of obtaining individualized HRTFs and HpEQs for each of the 12 participants, the present study utilized the generic HRTF built into the 'BinauralDecoder' plugin and the built-in generic HpEQ corresponding to the headphone used in the experiment from the 'SoundID Reference' application.

3.3 Participants

A total of 12 participants were selected from the piano performance program at the Xi'an Conservatory of Music in China. All participants were experienced pianists who regularly practiced and performed on grand pianos. The participants were divided into three groups based on their academic positions: undergraduate students, master's students, and teaching faculty members. This division allowed for the assessment of potential effects related to their academic level and experience on the perceptual evaluation.

The overall mean age of the participants was 26.25 years (SD = 9.423). The mean ages and standard deviations (SD) for each group were as follows:

- Undergraduate students: Mean age = 19 years (SD = 0.816)
- Master's students: Mean age = 23.25 years (SD = 0.433)
- Teaching faculty members (including professors): Mean age = 38 years (SD = 2.828)

All participants self-reported no hearing disorders that could impact the study. Their years of piano experience ranged from a minimum of 8 years to an average of 13 years across the full sample. Importantly, the participants had primarily practiced and performed on grand pianos regularly, ensuring their familiarity and expectations of directivity aligned with the grand piano I sampled.

3.4 Apparatus and Musical Stimuli

The digital piano utilized for the evaluation was the Yamaha P-48B 88-Key Weighted Action Digital Piano. This model is equipped with two built-in stereo speakers (12 cm x 2), positioned under the piano and pointing downwards towards the floor. Each speaker is powered by a separate amplifier (6 W + 6 W). The piano also has the capability to connect to headphones. It is because most digital pianos on the market do not support playing external samples from a computer through their built-in speakers, so the built-in stereo speakers and the headphone (stereo) output share the same piano sample that comes inherently with this digital piano, named 'Grand Piano.' The head-tracked binaural audio uses the sample from the Yamaha Baby Grand. Additionally, these digital piano features a MIDI input port, enabling it to be interfaced with a laptop (M1 Max MacBook Pro) via a USB cable. The laptop was connected to the audio interface (Audient id22). The headphone (Beyerdynamic DT 1990 Pro) was connected to this audio interface to deliver the rendered binaural audio. Figure 5 shows the experiment in progress.



Figure 5. The experiment in progress.

Stimuli:

Participants will play the following musical excerpts, each consisting of the initial 60-90 seconds of the respective pieces, on the digital piano using each of the three playback systems. After playing each excerpt on a given system, participants will immediately rate the playback based on the corresponding criterion for that excerpt before moving on to the next one.

1: Spatial Clarity:

• J.S. Bach's Invention No. 1 in C Major (BWV 772): The opening 60-90 seconds of this piece will be used to rate spatial clarity. This excerpt features the clear separation and independence of two melodic lines that continuously interweave and cross over each other, creating an intricate contrapuntal texture from the very beginning.

2: Spectral Clarity:

• Rachmaninoff's Prelude in C-sharp Minor, Opus 3 No. 2: The first 60-90 seconds of this piece will be used to assess spectral clarity, with its rich harmonic textures and dense chordal structures.

3: Envelopment:

• Chopin's Nocturne in E-flat Major, Op. 9, No. 2: The initial 60-90 seconds of this piece will be used to assess envelopment, with its flowing melodies and dynamic variations.

4: Preference:

• "The Maiden's Prayer" by Tekla Bądarzewska-Baranowska: The first 60-90 seconds of this piece will be used to provide the preference rating. This opening segment encompasses a variety of musical textures, including melodies, scales, chords, arpeggios, and dynamics, which collectively engage the three previously evaluated criteria (spatial clarity, spectral clarity, envelopment).

3.5 Procedure

The listening experiment was conducted individually for each participant at the Digital Piano Lab at Xi'an Conservatory of Music and consisted of a training phase and an experiment phase. During the training phase, the experiment procedures and the principles of the three different playback systems were explained to the participants. Additionally, the definitions of each criterion to be evaluated were provided.

In the experiment phase, participants were asked to play the same evaluated piece using each of the three playback systems, one at a time. After playing the piece on each system, they were required to rate the perceived levels of the following four criteria for each excerpt:

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

Spectral Clarity—The degree to which different frequencies/components are distinct and clearly audible.

Envelopment—The degree to which the sound surrounds and immerses the listener.

Preference—The choice of which system offered a better overall experience and to what degree.

The definitions of these criteria were included in the information sheet provided to the participants. These definitions were inspired by previous studies, as referenced in [18] and [19].

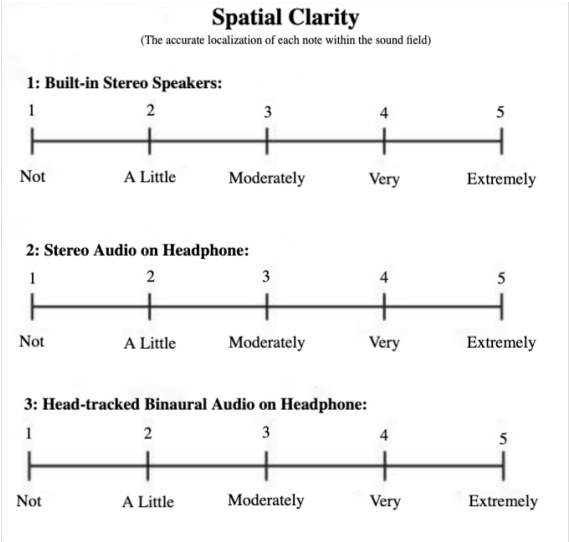


Figure 6. A 5-point Likert scale was used for rating each criterion, this rating sheet was reproduced from [19].

Training Phase:

During the training phase, participants were first introduced to the purpose and objectives of the study. They were then provided with detailed explanations of the three playback systems (built-in stereo speakers, static stereo headphones, and head-tracked binaural headphones) and how they differ in terms of sound presentation and responsiveness to head movements.

Next, the evaluation criteria (spatial clarity, spectral clarity, envelopment, and preference) were thoroughly explained to the participants. The definitions of each criterion were presented, along with illustrative audio examples showcasing high and low levels of each attribute. Participants were encouraged to ask questions and seek clarification to ensure a clear understanding of the concepts.

To familiarize participants with the rating process, they were presented with practice trials featuring audio samples of varying quality. For these samples, I first improvised and recorded a piano excerpt myself in advance, incorporating diverse musical elements such as melodies, scales, chords, arpeggios, and dynamics. This was done to ensure the sample engaged all three previously introcuded criteria: spatial clarity, spectral clarity, and envelopment. The recording was made from the player's perspective using the same Ambisonics microphone 'Zoom H3-VR' on the sampled piano in the Live Room A at Kreeger Studio, at a sample rate of 24bit/48000Hz.

I then processed this Ambisonics signal to create three different versions: binaural, stereo, and mono, by downmixing the Ambisonics signal accordingly. In this case, the binaural excerpt was expected to have the highest degree of all criteria, corresponding to ratings of 4 or 5 on the 5-point Likert scale. The stereo version represented a mid-range quality, corresponding to a rating of 3. Finally, the mono version's quality was intentionally degraded by reducing the sample rate from 24-bit/48kHz to 16-bit/41kHz and boosting certain low-frequency bands to make the sound muddier, corresponding to ratings of 1 or 2.

Participants rated these samples using the provided 5-point Likert rating scales and discussed their assessments with the experimenter. Any discrepancies or misunderstandings were addressed to establish a consistent interpretation of the rating scales among all participants.

Experimental Phase:

In the experimental phase, participants were seated at the digital piano and provided with the necessary instructions and materials, including the musical scores and rating sheets. Each participant was given ample time to familiarize themselves with the selected pieces and practice playing them on the digital piano.

The order of the playback systems was counterbalanced across participants to minimize potential order effects. Participants performed the evaluation process for each playback system separately. They began by playing the specific excerpts chosen for spatial clarity, spectral clarity, and envelopment assessments, followed by the excerpt of the common piece for preference evaluation.

After completing the playing on a given playback system, participants immediately filled out the corresponding section of the rating sheet. They assigned ratings for spatial clarity, spectral clarity, envelopment, and preference based on their listening experience. Participants were encouraged to provide qualitative comments and feedback to supplement their numerical ratings.

This process was repeated for each of the three playback systems, with short breaks between evaluations to minimize fatigue. Once all systems had been assessed, participants completed an exit survey, ranking the playback systems in order of overall preference and providing detailed insights into their experiences with each system.

Throughout the experimental phase, the experimenter was present to guide participants, answer any questions, and ensure adherence to the evaluation protocol. The experimenter maintained a neutral demeanor to avoid influencing participants' judgments.

4 Results

4.1 Overview

The perceptual evaluation involved participants rating three different playback systems (Built-in Speakers, Headphone, and Headphone with Head-tracked Binaural) across four rating criteria: Spatial Clarity, Spectral Clarity, Envelopment, and Overall Preference. The mean ratings for each playback system were calculated and analyzed using a two-way ANOVA to determine the effects of playback systems and rating criteria on participants' ratings.

4.1.1 Spatial Clarity

The mean ratings for spatial clarity are shown in Figure 7. The Head-tracked Binaural system received the highest mean rating (4.17), followed by the Built-in Speakers SPKs (2.83) and the Stereo Headphones (2.67).

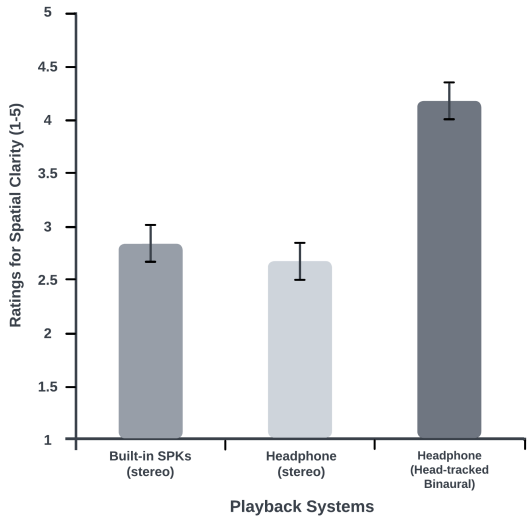


Figure 7. Spatial Clarity scales and standard errors for these three different playback systems.

4.1.2 Spectral Clarity

The mean ratings for spectral clarity are presented in Figure 8. The Head-tracked Binaural system again received the highest mean rating (4.5), followed by the Stereo Headphones (3) and the Built-in Speakers (1.83).

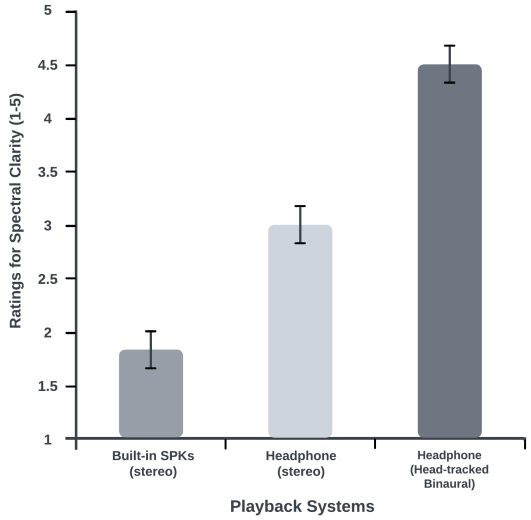


Figure 8. Spatial Clarity scales and standard errors for these three different playback systems.

4.1.3 Envelopment

The mean ratings for envelopment are shown in Figure 10. The Head-tracked Binaural system had the highest mean rating (4.17), while the Built-in Speakers and Stereo Headphones had mean ratings of 3 and 2.33, respectively.

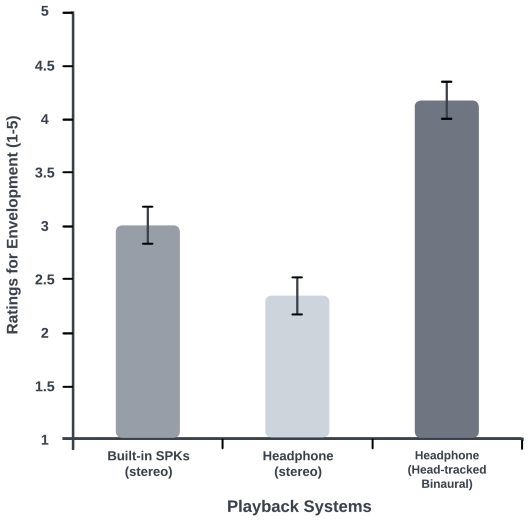


Figure 9. Envelopment scales and standard errors for these three different playback systems.

4.1.4 Preference

The mean ratings for preference are presented in Figure 10. The Head-tracked Binaural system was rated highest (4.67), followed by the Built-in Speakers (2.5) and the Stereo Headphones (1.58).

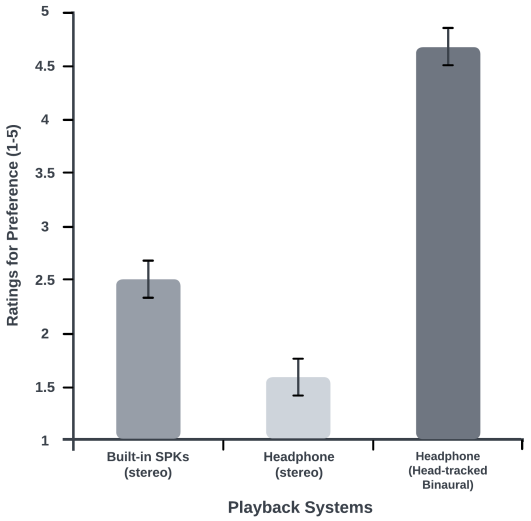


Figure 10. Preference scales and standard errors for these three different playback systems.

4.2 Statistical Analysis

A two-way ANOVA was conducted to examine the main effects of playback systems and rating criteria on the participants' ratings, as well as their interaction effects. The results are summarized in Table 1.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	120.028[a]	5	24.006	43.241	.000
Intercept	1381.361	1	1381.361	2488.253	.000
Playback Systems	117.722	2	58.861	106.027	.000
Rating Criterion	2.306	3	.769	1.384	.250
Error	76.611	138	.555		
Total	1578.000	144			
Corrected Total	196.639	143			

Table 1. Results of the two-way ANOVA showing the effects of Playback Systems and Rating Criteria on the perceptual ratings.

The two-way ANOVA revealed a significant main effect of playback systems on the ratings ($F = 106.027$, $p < 0.001$), indicating that the type of playback system significantly influenced the ratings across all criteria. The intercept, which represents the grand mean of the ratings, was also significant ($F = 2488.253$, $p < 0.001$).

The main effect of the rating criteria was not significant ($F = 1.384$, $p = 0.250$), suggesting that the different criteria (Spatial Clarity, Spectral Clarity, Envelopment, and Preference) did not significantly affect the ratings independently of the playback system used.

The interaction effect between playback systems and rating criteria was also not significant ($F = 0.769$, $p = 0.555$). This indicates that the influence of the playback system on the ratings did not vary significantly across the different rating criteria.

The corrected model, which includes both main effects and the interaction effect, was significant ($F = 43.241$, $p < 0.001$), explaining a significant portion of the variance in the ratings. The error term, representing the unexplained variance, had an F value of 0.555.

The total sum of squares was 1578.000, and the corrected total sum of squares, which accounts for the mean differences, was 196.639. The model explained 120.028 of the corrected total sum of squares, indicating a substantial proportion of the variance was accounted for by the playback systems and rating criteria.

Overall, the statistical analysis demonstrates that the type of playback system is a significant factor in the participants' perceptual ratings, while the specific rating criteria and their interaction with the playback systems did not have a significant impact.

4.3 Discussion

The results of the perceptual evaluation clearly demonstrate the superiority of the head-tracked binaural audio system across all rating criteria. In addition to the high ratings received from participants, the head-tracked binaural audio system's real-time performance was validated by the absence of any noticeable latency or artifacts reported by participants throughout the evaluation. This confirms that the specific combination of software and hardware used in this system was effective in maintaining latency below perceptual thresholds.

It is worth discussing the differences observed between the built-in stereo speakers and the headphone (stereo). For the spectral clarity, as shown in Figure 8, the built-in stereo speakers received a lower mean rating compared to the stereo headphones. This difference can be attributed to several factors. First, the limited number of speakers on the digital piano may affect the perceived clarity and realism of the audio due to reduced spatial resolution and sound fidelity. Second, the room acoustics of the experimental space likely influenced the perceived spectral clarity. The presence of standing waves in the digital piano lab, which lacked proper reverberation absorption, could have resulted in the boosting of specific frequency bands, leading to an uneven frequency response. Moreover, the samples used for the digital piano's built-in speaker playback already contained some inherent reverberation. When combined with the reverberation created by the room, the resulting sound heard may have become muddied and less clear.

To further analyze the spectral clarity differences, I compared the overall spectra of the same note, E2, from both the built-in sample of the digital piano and the Ambisonic sample. The built-in sample of E2 was recorded to a computer in stereo first with the connection of two TRS cables to the audio interface connected to the computer, followed by the export to a stereo audio file. This sample was then downmixed from stereo to mono, and the Ambisonic sample of E2 was also downmixed from four channels to mono to

ensure a fair comparison. The normalized amplitude spectra of both samples are shown in Figure 11.

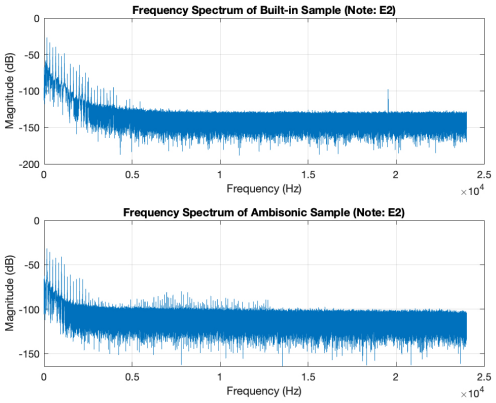


Figure 11. Frequency Spectrum of Built-in Sample (Note: E2) and Ambisonic Sample (Note: E2).

From the spectral plots, we observe that the Ambisonic sample maintains a more consistent spectral energy distribution across frequencies, while the built-in sample shows a more uneven distribution with certain frequency bands being more pronounced. This difference in spectral energy distribution likely contributes to the higher ratings for spectral clarity in the Ambisonic sample.

Additionally, the reverberant characteristics of both samples were examined using spectrograms (Figure 12). The built-in sample's spectrogram shows a longer decay time, indicating more reverberation compared to the Ambisonic sample. This visual difference further explains the disparity in spectral clarity ratings, as the built-in sample's prolonged reverberation may have led to a muddier sound, affecting clarity.

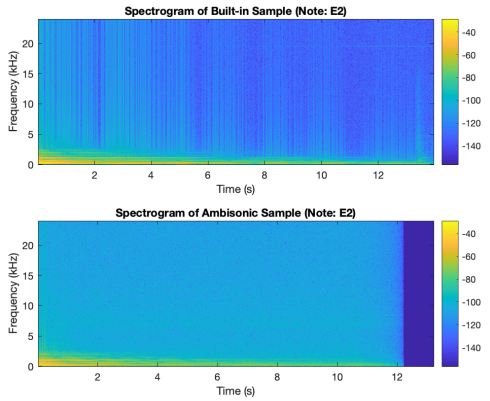


Figure 12. Spectrogram of Built-in Sample (Note: E2) and Ambisonic Sample (Note: E2).

In contrast, for envelopment, as shown in Figure 9, the room acoustics contributed to a higher rating for the built-in stereo speakers compared to the stereo headphones. The room reverberation, created by sound reflections within the space, enhanced the sense of surroundness for the built-in speakers. This is because the reverberant sound field envelops the listener, providing a more spacious and immersive auditory experience. On the other hand, the headphones used in this study, the Beyerdynamic DT 1990 Pro with analytical earpads, primarily provide a frontal sound image, lacking the spatial cues and reflections that contribute to a sense of envelopment. The analytical earpads provide a more neutral frequency response, similar to free-field equalized headphones, which aim to simulate the frequency response of a loudspeaker in an anechoic chamber. However, regardless of the type of equalization (free-field or diffuse-field), headphone playback in stereo generally has limited ability to recreate the spatial cues and reflections that are present when listening to speakers in a room. The equalization techniques mainly address the frequency response of the headphones and do not directly influence the spatial characteristics of the sound field. As a result, even with an appropriate frequency response, the headphone playback in stereo was unable to fully replicate the spatial characteristics of the room acoustics, resulting in a lower envelopment rating compared to the built-in stereo speakers.

Despite the differences between the built-in stereo speakers and stereo headphones, the head-tracked binaural audio system consistently outperformed both playback methods across all rating criteria (Figures 7-10). The superior performance of the head-tracked binaural audio system can be attributed to its ability to capture and reproduce spatial cues and directional information, creating a more natural and immersive auditory experience that closely mimics playing an acoustic piano. These findings highlight the potential of head-tracked binaural audio in enhancing the listening experience for digital piano players, offering a more realistic alternative to traditional playback systems. Moreover, this setup can serve as a useful training environment for players who are eventually going to be performing on a grand piano, as they will be experiencing a sound field that is more similar to their final performance environment.

5 Conclusion and Future Work

The results of the perceptual evaluation clearly indicate that the head-tracked binaural audio system significantly outperformed the other two playback systems across all criteria. The head-tracked binaural system received the highest mean ratings for spatial clarity, spectral clarity, envelopment, and preference, highlighting its effectiveness in providing a more immersive and realistic auditory experience. These findings are further supported by the two-way ANOVA results, which showed a significant main effect of playback systems on the ratings ($F = 106.027$, $p < 0.001$). The rating criteria did not independently affect the ratings significantly, nor did the interaction between playback systems and rating criteria, suggesting that the type of playback system was the primary factor influencing participants' ratings.

The higher order Ambisonics could be employed to improve the spatial resolution and sound fidelity of rendered sound. This may involve using a higher-order Ambisonics microphone for recording or an Ambisonics order upmixer to convert lower-order Ambisonics signals into higher-order formats, such as 7th-order Ambisonics. By increasing the spatial resolution, the system could provide an even more accurate and immersive auditory experience.

To further develop and spread this system for immersive music activities, take the digital piano as an example: the mono signal for each of these 88 keys needs to be dry (without reverberation), which can be convolved with different angle's BRIRs corresponding to the current directivity information of each key being pressed. The BRIRs can be obtained from the Ambisonics Impulse Response signals recorded using an Ambisonics microphone in various acoustic environments. This relative directivity information for each key can be changeable in real time when the player's head is rotating. By capturing the Ambisonics IRs from the player's perspective in different venues, the piano sound can be adaptively rendered with the desired reverberation characteristics. This approach eliminates the need to sample all 88 keys of a grand piano in multiple locations, providing greater flexibility in simulating different acoustic spaces used for this head-tracked binaural audio system. Furthermore, integrating this system with Virtual Reality (VR) technology could offer an even more immersive experience for digital piano players, players wearing VR headsets could feel as if they are playing in different venues both visually and auditorily. This multisensory experience could greatly enhance the realism and engagement of playing digital pianos.

Exploring the implementation of a 6 degrees of freedom (6DoF) system using Ambisonic translation from a higher-order sample could further enhance the immersive experience for digital piano players. Soundfield translation techniques such as soundfield re-expansion allow for accurate reproduction of spatial cues and directional information for off-center listening positions. By employing these techniques with higher-order Ambisonic recordings, the system could enable listeners to move freely within the virtual acoustic space, experiencing realistic changes in the soundfield based on their position and orientation.

References

[1] Statista Research Department (2023, June 22). Number of digital pianos sold in the United States from 2005 to 2021. Statista. https://www.statista.com/statistics/452777/number-of-digital- pianos-sold-in-the-us/

[2] Statista Research Department (2023, June 22). Number of acoustic pianos sold in the United States from 2005 to 2021.Statista.https://www.statista.com/statistics/452772/number-of-pianos-sold-in-the-us/

[3] Møller, H. (1992). Fundamentals of Binaural Technology. Applied Acoustics, 36(3-4), 171-218.

[4] Blauert, J. (1997). Spatial Hearing: The Psychophysics of Human Sound Localization. MIT Press.

[5] Papetti, Stefano & Avanzini, Federico & Fontana, Federico. (2019). Design and Application of the BiVib Audio-Tactile Piano Sample Library. Applied Sciences. 9. 914. 10.3390/app9050914.

[6] McKeag A, McGrath DS. Sound field format to binaural decoder with head tracking. Proceedings of the Audio Engineering Society Convention 6r. 1996.

[7] Werner, S., Klein, F., & Götz, G. (2017). Influence of Head Tracking on the Externalization of Auditory Events at Divergence between Synthesized and

Listening Room Using a Binaural Headphone System. AES Convention.

[8] Noisternig M, Musil T, Sontacchi A, Holdrich R. 3d binaural sound reproduction using a virtual ambisonic approach. In: IEEE International Symposium on Virtual Environments, Human- Computer Interfaces and Measurement Systems. IEEE; 2003. p. 174–178.

[9] Michael Gerzon. 1975. The design of precisely coincident microphone arrays for stereo and surround sound. 50th Audio Engineering So- ciety Conference, (AES preprint 8083L20).

[10] R. Nicol, “Sound Field,” in Immersive Sound: The Art and Science of Binaural and Multichannel Audio, A. Roginska and P. Geluso, Eds., New York; London, Routledge, 2018, pp. 276-310.

[11] Tolkova, Irina & Klinck, Holger. (2022). Source separation with an acoustic vector sensor for terrestrial bioacoustics. The Journal of the Acoustical Society of America. 152. 1123-1134. 10.1121/10.0013505.

[12] P. Giddings. Multi-channel sound reproduction moving into the nineties. Engineering Harmonics, February 15, 1991.

[13] Sunder, Kaushik & Tan, Ee-Leng & Gan, Woon-Seng. (2014). Effect of headphone equalization on auditory distance perception. 137th Audio Engineering Society Convention 2014.

[14] Kim SM, Choi W. On the externalization of virtual sound images in headphone reproduction: a Wiener filter approach. J Acoust Soc Am. 2005 Jun; 117(6):3657-65. doi: 10.1121/1.1921548. PMID: 16018469.

[15] Braxton Boren, Michele Geronazzo, “Comparison of Distortion Products in Headphone Equalization Algorithms for Binaural Synthesis”, in Proceedings of the 150th Audio Engineering Society Convention, Online, 2021.

[16] M. Frank, F. Zotter, and A. Sontacchi, "Producing 3D Audio in Am-bisonics," in Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology-Cinema, Television and the Internet. Audio Engineering Society, 2015.

[17] M. Tomasetti, A. Farina and L. Turchet, "Latency of spatial audio plugins: a comparative study", Proceedings of the International Conference on Immersive and 3D Audio, 2023.

[18] G. Marentakis, F. Zotter and M Frank, Vector-base and ambisonic amplitude panning: A comparison using pop, classical, and contemporary spatial music. Acta Acustica united with Acustica, 100(5), 945- 955 (2014).

[19] H. Lynch and R. Sazdov, "A Perceptual Investigation into Spatialization Techniques Used in Multichannel Electroacoustic Music for Envelopment and Engulfment," in Computer Music Journal, vol. 41, no. 1, pp. 13-33, March 2017, doi: 10.1162/COMJ_a_00401.