Audio Engineering Society

Conference Paper

Presented at the AES 2026 International Conference on
Audio for Virtual and Augmented Reality and Immersive Games
2026 June 30 – July 3, Sorbonne University/d'Alembert & IRCAM/STMS, Paris, France

Sub-band Neural Interpolation of Binaural Room Impulse Responses for Personal Sound Zone Control

Yazhou Li, Lin Wang, and Joshua Reiss

Center for Digital Music, Queen Mary University of London

Correspondence should be addressed to Yazhou Li (yazhou.li@qmul.ac.uk)

ABSTRACT

To enable dynamic control in transaural personal sound zone (PSZ) systems, accurate binaural room impulse responses (BRIRs) at various listener positions are needed. Since it is impractical to measure BRIRs at all possible positions, interpolation from a sparse set of measured positions can be used. Although numerous BRIR interpolation methods exist, their effectiveness in sound field control applications remains unclear. In this paper, we propose a sub-band interpolation method that combines linear interpolation for frequencies below 2000 Hz with sinusoidal representation networks for frequencies above 2000 Hz. The interpolated BRIRs are then applied in a PSZ control system. Simulation results demonstrate that this hybrid approach significantly improves system performance over a wider frequency range. [1]

1 Introduction

Personal sound zone (PSZ) reproduction from loudspeakers aims to provide different audio content with minimum interference to multiple listeners in the same physical space without the usage of headphones. Design methods include pressure matching (PM) [1], acoustic contrast control (ACC) [2], amplitude matching [3], and variable span trade-off filters (VAST) [4].

There are two approaches for PSZ reproduction. One is to create large sound zones, where a large number of room impulse responses (RIRs) covering the listening area are needed to control the sound zones. The other is to create small sound zones only at listeners' ears, where dynamic binaural room impulse responses (BRIRs) are required to track listeners' movement [5], which is termed as transaural PSZ. A similar principle lies in crosstalk cancellation (CTC), which aims to provide binaural audio to the listener through loudspeakers. Despite their similarities, PSZ presents additional challenges compared to CTC, as personal audio perception requires a higher energy ratio between the bright zone and the dark zone than spatial audio perception [6]. Therefore, accurate BRIRs at all positions are needed for filter design.

However, conducting exhaustive measurements across different rooms is often impractical. To address this challenge, various strategies have been proposed to improve the robustness to room acoustics or listening positions in PSZ control. One approach improves robustness to transfer function errors through worst-case optimization and probability-model optimization [7].

[1] The code is available at https://github.com/li630925405/BRIRs-interpolation-for-PSZ.

Another approach uses the equivalent source method, where measurements are taken only along the boundary of the sound zone [8]. Alternatively, room geometry can be incorporated via the image source method to model early reflections, thereby improving robustness to room acoustic variations [9]. Another possible solution is to interpolate the IRs at unmeasured positions based on sparse measurements. For example, [10] applies singular value decomposition to interpolate RIRs, and the interpolated RIRs are then used for PSZ control.

Traditional DSP methods for BRIR interpolation include using dynamic time warping to find corresponding reflection peaks of two adjacent BRIRs, then doing linear time interpolation of the onset times of two peaks, and linear magnitude interpolation of the shapes of two peaks [11, 12]. However, the onset time and magnitude of the reflection peaks may not change linearly, and modeling the non-linear composition is necessary [12]. Other traditional methods utilize the discrete prior of data distribution and use kernel interpolation [13] or compressed sensing [14] for RIR interpolation.

Currently, many works have started to use deep learning for head-related impulse responses (HRIR) interpolation, including currently popular generation models, HRTF field [15, 16] and GAN HRTF [17]. There are also emerging works using deep learning for RIR interpolation, including super-resolution methods [18] and implicit neural representation (INR). INR models learn the continuous data representation from discrete observations, which is first applied in computer vision for 3D scene reconstruction [19]. Another benefit of using INR to represent the sound field is that it consumes less memory to store the data, compared to traditional audio encoding methods, like advanced audio coding (AAC) [20] and Xiph Opus [21].

For audio applications, INR has been used to represent audio signals in different domains. Time domain representation helps model the details in the time signal, while time-frequency domain and frequency domain usually learn smoother global features. [22] uses INR to estimate RIRs in the time-frequency domain. [23] does it in the time domain to learn the sudden peaks of reflections. [24] models the early reflections using the time domain representation and the late reverberation using the time-frequency domain representation separately. [25] estimates HRTFs using INR in the frequency domain. [26] extends volume rendering in computer vision to acoustic volume rendering, so spatial room impulse responses (SRIRs) can be estimated from monaural measurements. [27] applies INR to BRIR interpolation.

Existing physics-informed neural networks (PINNs) for acoustic modeling are also built upon INR architectures. These methods incorporate physical constraints into the loss function to enforce consistency with physical laws. Such constraints typically involve derivatives with respect to continuous spatial or temporal coordinates, which are naturally supported by INR-based formulations. PINNs have been applied to RIRs [28, 29], HRIRs [30] and BRIRs [31]. However, these works mainly report time-domain metrics and do not include downstream PSZ evaluation, and frequency-domain transfer-function accuracy is not the primary focus in those evaluations.

A well-known problem in INR is the spectral bias problem [32], which means the network tends to learn the low frequencies of the data and struggle to capture fine details. This behavior can be explained using neural tangent kernel [33]. To mitigate this issue, Fourier feature encoding [34] and sinusoidal representation networks (SIREN) [35] have been proposed to map the low-dimensional input into trigonometric functions, enabling the modeling of high frequency signal components. However, both methods require empirical parameter tuning for signals with different frequency ranges. Also, SIREN is prone to overfitting high-frequency content and is sensitive to noise, which can degrade its low-frequency performance. Many methods have been proposed to enable INR to control the frequency range in the signal. Some works [36] control the bandlimit by filtering the spectrum. [37] adds different initialization and bounding schemes so that the network focuses more on the low frequencies, preventing the overfitting problem.

In this study, we propose a time-domain frequency-partitioned interpolation method: linear interpolation is applied to low-frequency components, while SIREN models the high-frequency components. This is based on the observation that BRIRs vary smoothly with the change of spatial positions at low frequencies, so linear interpolation should perform well. Other methods that could address the spectral bias of SIREN include modifying the network architecture to enable SIREN to learn low frequencies better. However, in this work, we just use linear interpolation to sidestep the spectral bias problem because linear interpolation performs well enough in low frequencies for our problem. We

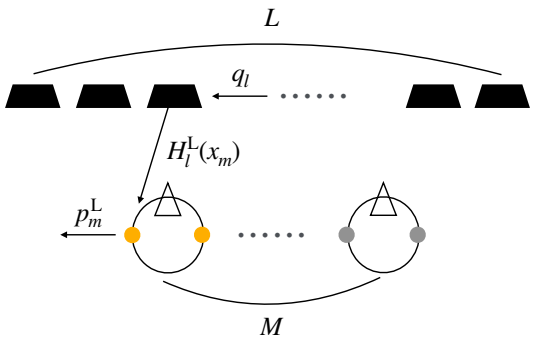


Fig. 1: There are L loudspeakers and M listeners, and one microphone at each ear, $2M$ microphones in total. The yellow circles are the bright zones and the grey circles are the dark zones. q_l is the filter of the l-th loudspeaker, $H_l^{\mathrm{L}}(x_m)$ is the ground truth transfer function from the l-th loudspeaker to the m-th listener's left ear, and p_m^{L} is the measured SPL at the m-th listener's left ear. These quantities are implicitly dependent on frequency ω_k.

further investigate the impact of the turning frequency that separates low and high frequencies in different experiment setups.

2 Problem Definition

We consider a transaural personal sound zone (PSZ) reproduction system with L loudspeakers and M listeners. Each listener is modeled by two microphones located at the left and right ears, resulting in a total of $2M$ microphones.

Let $x \in \mathbb{R}$ denote the one-dimensional spatial position of the center of the listener's head along the measurement line. The BRIRs from the l-th loudspeaker to the left and right ears at head position x are denoted by $h_l^{\mathrm{L}}(n,x)$ and $h_l^{\mathrm{R}}(n,x)$, respectively, where $n = 0,\ldots,N-1$ is the discrete time index. Note that the left- and right-ear BRIRs share the same head center position x. For notational convenience, we represent the time-domain BRIRs in matrix form as $\mathbf{h}_l^{\mathrm{L}} \in \mathbb{R}^{N\times Z}$, and similarly for the right ear, $\mathbf{h}_l^{\mathrm{R}}$.

BRIRs are measured at a set of Z discrete head positions and are estimated at M evaluation head positions through interpolation. The estimated BRIRs from the l-th loudspeaker to the left and right ears at evaluation position x_m are denoted by $\tilde{h}_l^{\mathrm{L}}(n,x_m)$ and $\tilde{h}_l^{\mathrm{R}}(n,x_m)$, or equivalently in matrix form as $\tilde{\mathbf{h}}_l^{\mathrm{L}}$ and $\tilde{\mathbf{h}}_l^{\mathrm{R}} \in \mathbb{R}^{N\times M}$.

For each loudspeaker and each ear, the estimated BRIRs are transformed into frequency-domain transfer functions via the discrete Fourier transform (DFT), yielding $\tilde{H}_l^{\mathrm{L}}(\omega_k,x_m)$ and $\tilde{H}_l^{\mathrm{R}}(\omega_k,x_m)$, where ω_k, $k = 1,\ldots,K$, denotes the frequency bins. At each frequency bin ω_k, these transfer functions are assembled into the left- and right-ear transfer matrices $\tilde{\mathbf{H}}^{\mathrm{L}}(\omega_k) \in \mathbb{C}^{M\times L}$ and $\tilde{\mathbf{H}}^{\mathrm{R}}(\omega_k) \in \mathbb{C}^{M\times L}$. The estimated transfer matrix used for PSZ control is then

$$\tilde{\mathbf{H}}(\omega_k) = \begin{bmatrix} \tilde{\mathbf{H}}^{\mathrm{L}}(\omega_k) \\ \tilde{\mathbf{H}}^{\mathrm{R}}(\omega_k) \end{bmatrix} \in \mathbb{C}^{2M\times L}. \tag{1}$$

As shown in Figure 1, L loudspeakers are used for reproduction and $2M$ microphones are used for measuring the sound pressure levels (SPLs) when there are M listeners. B listeners who can hear the audio are referred to as the bright listener, and $D = M - B$ listeners who cannot hear the audio are referred to as the dark listener. Each ear is seen as a sound zone, so there are $2B$ bright zones and $2D$ dark zones. The SPLs in the bright zones are maximized and the SPLs in the dark zones are minimized.

3 Background

3.1 SIREN and linear interpolation for BRIRs

INR networks are neural networks that represent a discrete signal as a continuous function. By inputting the coordinate in the signal, the network outputs the value at that coordinate. It can intrinsically perform interpolation of a spatially discrete signal. In this work, we adopt the time-domain SIREN architecture [35], which is a multilayer perceptron (MLP) with sinusoidal activation functions. The network output is given by

$$\begin{aligned} \tilde{h}_l^{\mathrm{L}}(n,x) &= \boldsymbol{W}_n(\phi_{n-1}\circ\phi_{n-2}\circ\ldots\circ\phi_0)(n,x) + \boldsymbol{b}_n, \\ \phi_i(\boldsymbol{x}_i) &= \sin(\omega_0(\boldsymbol{W}_i\boldsymbol{x}_i + \boldsymbol{b}_i)), \end{aligned} \tag{2}$$

where ϕ_i denotes the i-th network layer and $\circ$ represents function composition. The input to the i-th layer is $\boldsymbol{x}_i \in \mathbb{R}^{N_{i-1}}$, while $\boldsymbol{W}_i \in \mathbb{R}^{N_i\times N_{i-1}}$ and $\boldsymbol{b}_i \in \mathbb{R}^{N_i}$ are the corresponding weight matrix and bias vector. Here, N_i denotes the number of neurons in layer i.

ω_0 is a hyperparameter that controls the frequency scaling of the sinusoidal activation functions. Specifically,

it determines the range of frequencies the network can represent, influencing its ability to model signals with varying levels of detail. Larger values of ω_0 allow the network to capture higher-frequency variations in the signal. The SIREN implementation used in this work follows our previous study [31].

For each loudspeaker–ear pair, the loss function is the time-domain mean squared error (MSE):

$$\mathscr{L} = \frac{1}{NZ} \sum_{(n,x)} \left| \tilde{h}_l^L(n,x) - h_l^L(n,x) \right|^2. \quad (3)$$

Another interpolation method is linear interpolation, which can be written as:

$$\tilde{h}_l^{\mathrm{L}}(n,x_m) = \frac{x_2 - x_m}{x_2 - x_1} h(n,x_1) + \frac{x_m - x_1}{x_2 - x_1} h(n,x_2), \quad (4)$$

where x_m is the evaluation listener's position, x_1, x_2 are the two measurement positions that are the nearest to the evaluation listener.

3.2 Pressure matching for PSZ

We use frequency-domain PM [38] as the optimization algorithm to design filters to control the driving signals of loudspeakers [9]. The loudspeaker filters $\mathbf{q} \in \mathbb{C}^{L\times 1}$ are designed using the estimated transfer function matrix $\tilde{\mathbf{H}} \in \mathbb{C}^{2M\times L}$ from L loudspeakers to $2M$ microphones. In the PM formulation, all quantities in matrix form are implicitly dependent on frequency ω_k. The optimization problem with Tikhonov regularization is formulated as

$$\min_{\mathbf{q}} \|\tilde{\mathbf{H}}\mathbf{q} - \mathbf{p}_T\|^2 + \beta \|\mathbf{q}\|^2, \quad (5)$$

where $\mathbf{p}_T = [..., 1, ..., 0, ...]^T \in \mathbb{C}^{2M\times 1}$ is the target pressure at $2M$ microphones, which is 1 in the bright zones and 0 in the dark zones, and β is the weight of Tikhonov regularization. The closed-form solution is

$$\mathbf{q} = (\tilde{\mathbf{H}}^H \tilde{\mathbf{H}} + \beta \mathbf{I})^{-1} \tilde{\mathbf{H}}^H \mathbf{p}_T. \quad (6)$$

During evaluation, the achieved SPL $\mathbf{p} \in \mathbb{C}^{2M\times 1}$ is computed using the ground-truth transfer functions $\mathbf{H} \in \mathbb{C}^{2M\times L}$

$$\mathbf{p} = \mathbf{H}\mathbf{q}. \quad (7)$$

4 Method

4.1 System overview

In contrast to existing work that treats BRIR interpolation as a signal reconstruction problem, this paper treats it as an integral part of the PSZ control pipeline. Even without the proposed hybrid scheme, using SIREN-interpolated BRIRs for PSZ control already constitutes a new system formulation. The hybrid method is then introduced to address the frequency-dependent limitations of SIREN. Specifically, BRIRs at unmeasured positions are first obtained using 4 different interpolation methods as shown in the next section, and subsequently used for loudspeaker filter design via PM.

4.2 Sub-band hybrid interpolation

We propose a hybrid method that uses linear interpolation for low frequencies and SIREN for high frequencies. Specifically, for each loudspeaker and each ear, the estimated full-band transfer functions from linear interpolation are low-pass filtered (LPF) with the cut-off frequency f_t, and the estimated full-band transfer functions from SIREN are high-pass filtered (HPF) with the cut-off frequency f_t. The cut-off frequency is also referred to as the turning frequency. The two filtered transfer functions are then added to form the broadband transfer function $\tilde{\mathbf{H}}_l^{\mathrm{L}}$, as shown in Figure 2.

We also have two schemes for choosing the turning frequency f_t. One is the fixed method, where f_t is set to 2000 Hz. The other is the adaptive method, where f_t is calculated as the first frequency from which the performance of SIREN exceeds the performance of linear interpolation for a consecutive 5 frequency bins in order to mitigate the influence of noise, and this frequency is referred to as f_c. The performance metric used to calculate f_c is acoustic contrast or NMSE, which will be defined later in the experiment section. The turning frequency is then set as $f_t = f_c$.

The above interpolation process only produces $\tilde{\mathbf{H}}_l^{\mathrm{L}}$ of a single sound field, in other words, from a single loudspeaker to the left ear. To get $\tilde{\mathbf{H}} \in \mathbb{C}^{2M\times L}$ from L loudspeakers to both the left and right ears, the above process needs to be repeated for $2L$ times.

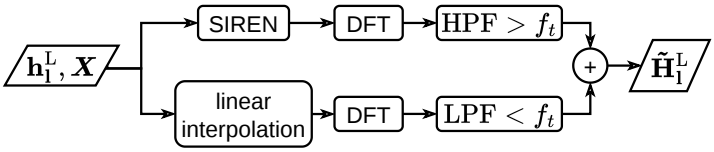


Fig. 2: Block diagram of the signal processing procedure of the hybrid method. For the l-th loudspeaker, the left-ear transfer functions $\tilde{\mathbf{H}}_l^{\mathrm{L}}$ of the evaluation listeners are estimated from the measured BRIRs $\mathbf{h}_l^{\mathrm{L}}$ of the measurement listeners. Here $\tilde{\mathbf{H}}_l^{\mathrm{L}} \in \mathbb{C}^{K\times M}$ denotes the full-band transfer functions at K frequency bins, and $\boldsymbol{X}$ is the spatial positions of the evaluation listeners.

5 Experiment

5.1 Dataset

Figure 3 shows the experiment setup when $Z = 11$. The experimental configuration consists of 6 loudspeakers arranged vertically from 2 m to 3 m along the $x = 3$ m line, and 11 heads positioned from 0.5 m to 4.5 m at a 0.4 m interval along the $x = 2$ m line. The BRIRs for the left and right ears are trained separately, and each loudspeaker's BRIRs are trained separately, resulting in a total of $6 \times 2 = 12$ networks trained for a single experiment setup.

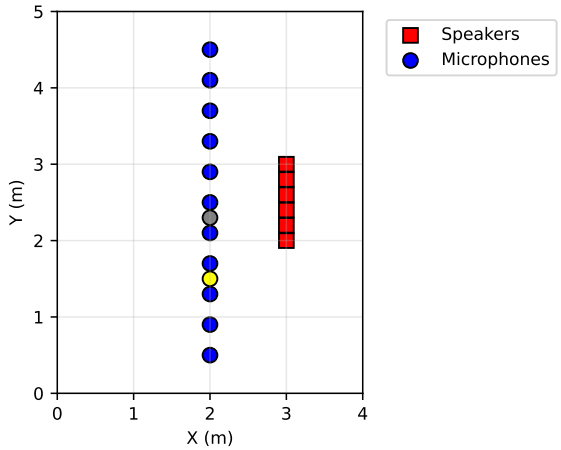


Fig. 3: Experimental setup showing loudspeaker array (red squares), microphone array (blue circles), bright listener (yellow circle), and dark listener (gray circle).

We consider the two listeners ($M = 2$, $B = D = 1$) scenario, where the right listener (facing the loudspeaker array) is the bright listener, and the left listener is the dark listener. This can be easily extended to more listeners. The distance between the bright listener and the dark listener is 0.8 m, and the listener's position is in the center of two adjacent measurement microphones. Although the listener can move along the microphone array on $x = 2$ m at any position, the performance when the listener is near the measurement microphone is better for both linear interpolation and SIREN. Therefore, we only consider the worst case when the listener is the farthest away from the measurement microphone, which is in the middle of the two microphones.

For a comprehensive assessment, we traverse the listener's position from the start to the end of the microphone array to calculate the average metrics at different positions. We only consider the positions when both the bright listener and the dark listener are within the range y = 1.5 m to y = 3.5 m, because it is difficult for the loudspeakers to control the microphone that is too far away from the loudspeakers. The bright listener starts at the position 1.5 m and ends at the position 2.7 m. The movement size of the listeners matches the spacing between two adjacent measurement microphones. For instance, in the first evaluation step, the bright listener is positioned at 1.5 m, as illustrated in Figure 3, and then moved to 1.9 m in the next step. To investigate the effect of f_t under varying spatial resolutions and measurement microphone numbers, Z is varied from 11 to 111 with an interval of 10, resulting in a total of 11 settings.

The BRIR dataset is created using RAZR simulator [39]. RAZR is a MATLAB toolbox that calculates SRIRs and then convolves HRIRs from different directions with SRIRs to synthesize BRIRs. We choose HRIRs measured with the small KEMAR dummy head from the CIPIC database [40] for BRIRs synthesis. Head rotation is not modeled, and only translational changes of the head position are considered. The head

orientation is therefore fixed and faces the loudspeaker array throughout the simulation.

The room is a shoebox with the size of $4\times5\times3$ m^3. The wall materials are plaster sprayed, whose absorption coefficient is about 0.7, and the room reverberation time is about 0.1 s. A short reverberation time is adopted because modeling long BRIR sequences with SIREN increases estimation noise and leads to degraded performance. Extending SIREN to longer time sequences is therefore left for future work. The sampling rate is 44100 Hz and the length of the BRIR is 1000 samples, which is 22.7 ms. The BRIRs are band-limited to the frequency range 0–7000 Hz. As discussed in the Introduction, SIREN struggles to model the full frequency range, and restricting the learning task to the lower-frequency sub-band improves modeling accuracy. Moreover, frequencies up to 7000 Hz are sufficient to achieve perceptual separation between bright and dark zones for a wide range of audio content.

5.2 Baseline methods

We compare the proposed hybrid method with three baseline methods: linear interpolation, SIREN, and dynamic time warping (DTW). Linear interpolation and SIREN have been introduced in Section 3.1. The implementation of DTW is based on [11].

5.3 Evaluation metrics

The evaluation metric of the PSZ system is acoustic contrast (AC), which describes the energy difference between the bright zones and the dark zones. AC at frequency ω_k is presented by

$$\mathrm{AC}(\omega_k) = 20\log_{10}\frac{|p_b(\omega_k)|}{|p_d(\omega_k)|}, \tag{8}$$

where $p_b(\omega)$ is the average SPL in the bright zones and $p_d(\omega)$ is the average SPL in the dark zones.

Normalized mean squared error (NMSE) is used to evaluate the frequency-dependent accuracy of the interpolated BRIRs. At each frequency bin ω_k, NMSE is defined as

$$\mathrm{NMSE}(\omega_k) = 10\log_{10}\frac{\|\tilde{\boldsymbol{H}}(\omega_k,:,:) - \boldsymbol{H}(\omega_k,:,:)\|_F^2}{\|\boldsymbol{H}(\omega_k,:,:)\|_F^2}, \tag{9}$$

where $\boldsymbol{H}(\omega_k,:,:) \in \mathbb{C}^{2M\times L}$ and $\tilde{\boldsymbol{H}}(\omega_k,:,:) \in \mathbb{C}^{2M\times L}$ denote the ground-truth and interpolated acoustic transfer matrices from L loudspeakers to $2M$ ear microphones, and $\|\cdot\|_F$ denotes the Frobenius norm.

5.4 Training

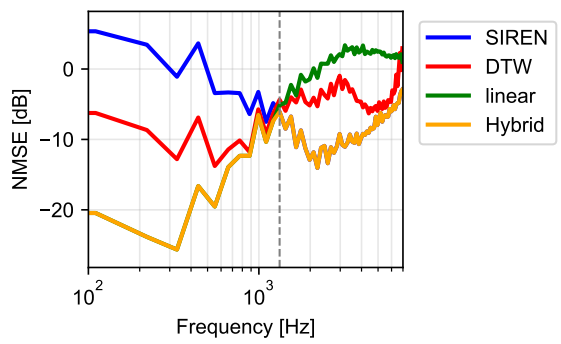
The network architecture is a 5-layer MLP with 128 neurons in each layer. The activation function initialization parameter ω_0 is 10 for all layers. The network weights initialization scheme is the same as [35]. The inputs are normalized to [-1, 1] before being input to the network. The model is trained for 20,000 iterations with a batch size of 400. We use the AdamW optimizer with a learning rate of 1×10^{-4} and an exponential learning rate scheduler with $\gamma = 0.9$ applied every 2000 steps. The random seed is set to 3407 for reproducibility.

6 Results

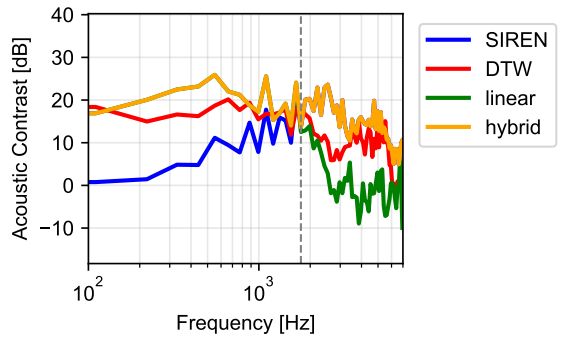
6.1 Performance comparison of the four methods

To investigate the relationship between BRIR modeling accuracy and PSZ performance, we examine both AC and NMSE across the frequency range. As shown in Figure 4, the hybrid method aligns with linear interpolation at low frequencies before the turning frequency, and with SIREN at high frequencies after the turning frequency. This is consistent with the observation that BRIRs vary smoothly at low frequencies, where linear interpolation can perform well, while SIREN performs better in high frequencies because it uses trigonometric functions to model the fluctuations in the signals and is good at modeling high frequencies. DTW's performance lies between linear interpolation and SIREN in both low and high frequencies.

Figure 4 (b) presents the average AC across all microphone positions. The average AC of linear interpolation, SIREN and the hybrid method are 4.96 dB, 12.85 dB and 15.40 dB, and the corresponding NMSEs are -1.96 dB, -7.12 dB and -9.89 dB. The hybrid method improves AC by 2 dB compared to SIREN and 9 dB compared to linear interpolation. The average metrics of SIREN are much better than linear interpolation because more frequency bins in high frequencies are used to calculate the average. Although the average AC is lower than 20 dB, the performance could be easily improved by using a larger number of loudspeakers; nevertheless, even with just six loudspeakers, the achieved contrast is already acceptable for many practical scenarios.



(a) NMSE (the lower the better)



(b) Acoustic Contrast (the higher the better)

Fig. 4: NMSE and acoustic contrasts when the number of measurement microphones is 21. Four lines are the average metrics of the four methods, and the shaded areas are the standard deviation around the average metric. The grey vertical dotted line is the calculated turning frequency f_c.

By comparing the shaded area in the left and right figures in Figure 4, we can see that the shaded area in AC is larger than that in NMSE, where the standard deviation of AC of the hybrid method is 5.64 dB and the standard deviation of NMSE of the hybrid method is 1.16 dB. This is due to the variation of AC across the microphone positions. When the listeners are near the center of the loudspeaker array, AC is higher, while AC is lower when the listeners are far from the center of the loudspeaker array.

6.2 Analysis of the turning frequency

Figure 5 shows how f_c and acoustic contrast improvement of the hybrid method over linear interpolation vary with the number of measurement microphones Z. With the increase of Z, the AC improvement first increases and then decreases, and f_c first decreases and then increases. f_c of AC is empty at $Z = 11$ because AC of SIREN is lower than AC of linear interpolation across all frequencies. This indicates that f_c can be used to compare the performance of SIREN and linear interpolation. When f_c is small, SIREN performs better than linear interpolation in a wider frequency band.

The result suggests that when Z is small, SIREN struggles to learn the data representation from a small dataset and cannot interpolate BRIRs accurately. When Z becomes 21, SIREN starts to perform better than linear interpolation, and the performance improvement is the most obvious at this point. However, as Z becomes larger, AC improvement starts to decrease and f_c starts to increase. This is because linear interpolation also becomes accurate at higher frequencies with increased measurement resolution, reducing the performance gap between the two methods.

From Figure 5 (left), we can see that f_c of NMSE is always lower than f_c of AC, which means even if SIREN performs better than linear interpolation at a certain frequency, AC of SIREN may still be lower than AC of linear interpolation at that frequency, indicating that although SIREN becomes more accurate for BRIR interpolation, residual noise can hinder filter design and AC performance. However, f_c of NMSE and f_c of AC follow a similar trend, confirming that a low NMSE can result in a high AC in general cases.

Although f_c varies with Z, as shown in Figure 5 (left), the difference in AC improvements of the hybrid approach over linear interpolation between the fixed method and the adaptive method is negligible, as shown in Figure 5 (right). Here f_c of AC is used for the adaptive method. This is because the difference between SIREN and linear interpolation is small around f_c, and the performance does not change significantly when f_t varies around f_c. Therefore, it should be sufficient to use a fixed f_t of 2000 Hz without the need to adaptively change f_t for different setups in practical implementation.

7 Conclusion

In this work, we propose a sub-band BRIR interpolation method that combines linear interpolation for low

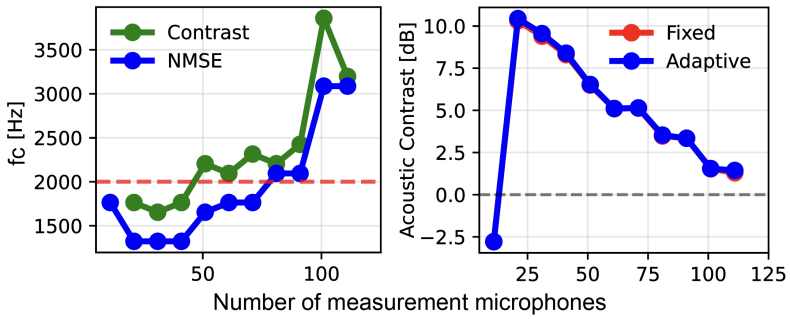


Fig. 5: Left: The calculated turning frequency f_c of NMSE and acoustic contrast. The red horizontal dotted line represents a fixed 2000 Hz. Right: Average acoustic contrast improvement of the hybrid approach over linear interpolation across all frequencies for both the adaptive and fixed methods.

frequencies with SIREN for high frequencies. This hybrid approach is applied to personal sound zone control, and the results demonstrate substantial improvements in acoustic contrast over using either method alone. We also explore the choice of the turning frequency that separates the low and high-frequency bands in different experiment setups. While adaptive selection can offer slight performance gains, we find that a fixed turning frequency of 2000 Hz generally performs well across different measurement densities, simplifying the system design for practical implementation.

Future work will focus on extending the proposed framework to handle longer BRIR time sequences and on exploring alternative neural architectures to further alleviate spectral bias. In addition, subjective listening tests and validation through physical experiments will be essential to assess the perceptual impact and practical feasibility of the proposed approach.

References

[1] Vindrola, L. et al., "Pressure matching with forced filters for personal sound zones application," *Journal of the Audio Engineering Society*, 68(11), pp. 832–842, 2020.

[2] Choi, J.-W. and Kim, Y.-H., "Generation of an acoustically bright zone with an illuminated region using multiple sources," *The Journal of the Acoustical Society of America*, 111(4), pp. 1695–1700, 2002.

[3] Abe, T. et al., "Amplitude Matching for Multizone Sound Field Control," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, pp. 656–669, 2023, doi:10.1109/TASLP.2022.3231715.

[4] Lee, T., Nielsen, J. K., and Christensen, M. G., "Signal-adaptive and perceptually optimized sound zones with variable span trade-off filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, pp. 2412–2426, 2020.

[5] Jackson, P., Fazi, F., and Coleman, P., "Personalising sound over loudspeakers," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, 2019.

[6] Canter, N. and Coleman, P., "Delivering personalised 3D audio to multiple listeners: Determining the perceptual trade-off between acoustic contrast and cross-talk," in *Audio Engineering Society Convention 150*, 2021.

[7] Zhu, Q. et al., "Robust personal audio reproduction based on acoustic transfer function modelling," in *AES International Conference on Sound Field Control*, 2016.

[8] Du, B., Zeng, X., and Vorländer, M., "Multizone sound field reproduction based on equivalent source method," *Acoustics Australia*, 49(2), pp. 317–329, 2021.

[9] Li, Y., Wang, L., and Reiss, J., "Impact of mismatched room acoustic modeling on transaural reproduction with loudspeaker arrays," in *Audio Engineering Society Convention 155*, 2023.

[10] Zhu, Q. et al., "An experimental study on transfer function estimation using acoustic modelling and singular value decomposition," *The Journal of the Acoustical Society of America*, 150(5), pp. 3557–

3568, 2021.

[11] Bruschi, V. et al., "An innovative method for binaural room impulse responses interpolation," in *Audio Engineering Society Convention 148*, 2020.

[12] Garcia-Gomez, V. and Lopez, J. J., "Binaural room impulse responses interpolation for multimedia real-time applications," in *Audio Engineering Society Convention 144*, Milan, 2018.

[13] Koyama, S., Kimura, K., and Ueno, N., "Weighted pressure and mode matching for sound field reproduction: Theoretical and experimental comparisons," *arXiv preprint arXiv:2303.13027*, 2023.

[14] Zea, E., "Compressed sensing of impulse responses in rooms of unknown properties and contents," *Journal of Sound and Vibration*, 459, p. 114871, 2019.

[15] Zhang, Y., Wang, Y., and Duan, Z., "HRTF field: Unifying measured HRTF magnitude representation with neural fields," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, 2023.

[16] Masuyama, Y. et al., "NIIRF: Neural IIR Filter Field for HRTF Upsampling and Personalization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[17] Hogg, A. O. et al., "HRTF upsampling with a generative adversarial network using a gnomonic equiangular projection," *TASLP*, 32, pp. 2085–2099, 2024, doi:10.1109/TASLP.2024.3375635.

[18] Lluis, F. et al., "Sound field reconstruction in rooms: Inpainting meets super-resolution," *The Journal of the Acoustical Society of America*, 148(2), pp. 649–659, 2020.

[19] Mildenhall, B. et al., "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, 65(1), pp. 99–106, 2021.

[20] Bosi, M. et al., "ISO/IEC MPEG-2 advanced audio coding," *Journal of the Audio engineering society*, 45(10), pp. 789–814, 1997.

[21] Valin, J.-M., Vos, K., and Terriberry, T. B., "Definition of the Opus Audio Codec," RFC 6716, Internet Engineering Task Force, 2012.

[22] Luo, A. et al., "Learning neural acoustic fields," *Advances in Neural Information Processing Systems*, 35, pp. 3165–3177, 2022.

[23] Su, K., Chen, M., and Shlizerman, E., "INRAS: implicit neural representation for audio scenes," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, New Orleans, 2022.

[24] Ge, Z., Li, L., and Qu, T., "A Hybrid Time and Time-frequency Domain Implicit Neural Representation for Acoustic Fields," in *Audio Engineering Society Convention 156*, Madrid, 2024.

[25] Di Carlo, D. et al., "Neural steerer: novel steering vector synthesis with a causal neural field over frequency and direction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, 2024.

[26] Lan, Z. et al., "Acoustic Volume Rendering for Neural Impulse Response Fields," *Advances in Neural Information Processing Systems*, 2024.

[27] Qiao, Y. and Choueiri, E., "Neural modeling and interpolation of binaural room impulse responses with head tracking," in *Audio Engineering Society Convention 155*, New York, 2023.

[28] Pezzoli, M., Antonacci, F., and A., S., "Implicit neural representation with physics-informed neural networks for the reconstruction of the early part of room impulse responses," in *Forum Acusticum*, Turin, 2023.

[29] Karakonstantis, X. et al., "Room impulse response reconstruction with physics-informed deep learning," *The Journal of the Acoustical Society of America*, 155(2), pp. 1048–1059, 2024.

[30] Ma, F. et al., "Physics informed neural network for head-related transfer function upsampling," *arXiv preprint arXiv:2307.14650*, 2023.

[31] Li, Y., Wang, L., and Reiss, J., "Binaural room impulse responses interpolation using physics-informed neural networks in three dimensions," 51st German Annual Conference on Acoustics, 2025.

[32] Rahaman, N. et al., "On the Spectral Bias of Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[33] Jacot, A., Gabriel, F., and Hongler, C., "Neural tangent kernel: convergence and generalization in neural networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

[34] Tancik, M. et al., "Fourier features let networks learn high frequency functions in low dimensional domains," in *Proceedings of the 34th Interna-*

tional Conference on Neural Information Processing Systems, 2020, ISBN 9781713829546.

[35] Sitzmann, V. et al., "Implicit neural representations with periodic activation functions," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, 2020.

[36] Dou, Y. et al., "Multiplicative Fourier Level of Detail," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[37] Novello, T. et al., "Tuning the Frequencies: Robust Training for Sinusoidal Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[38] Kirkeby, O. and Nelson, P. A., "Reproduction of plane wave sound fields," *The Journal of the Acoustical Society of America*, 94(5), pp. 2992–3000, 1993.

[39] Wendt, T., Van De Par, S., and Ewert, S., "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," *Journal of the Audio Engineering Society*, 62(11), pp. 748–766, 2014.

[40] Algazi, V. et al., "The CIPIC HRTF Database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, New Paltz, 2001.