Audio Engineering Society

Convention Express Paper 158

Presented at the 155th Convention
2023 October 25–27, New York, USA

Neural modeling and interpolation of binaural room impulse responses with head tracking

Yue Qiao[1] and Edgar Choueiri[1]

[1] *3D Audio and Applied Acoustics Laboratory, Princeton University, Princeton, New Jersey, 08544, USA*

Correspondence should be addressed to Yue Qiao (yqiao@princeton.edu)

ABSTRACT

The use of neural networks for modeling and interpolating binaural room impulse responses (BRIRs) is investigated for facilitating spatial audio applications that require head tracking in multiple degrees of freedom. A deep neural network model is adopted from an architecture originally proposed for neural representation problems to predict unknown BRIRs that contain salient early reflection peaks, given head coordinates. Instead of its original time-domain formulation, a frequency-domain formulation is proposed to enhance the model efficiency and flexibility for band-limited BRIRs. Both model formulations are evaluated with measured and simulated BRIRs in terms of modeling accuracy and interpolation performance, respectively. It is shown that the frequency-domain formulation is more effecient at modeling band-limited BRIRs than its time-domain counterpart as the former only learns the partial frequency spectrum, and that models with both formulations significantly outperform conventional methods for interpolating sparse BRIRs.

1 Introduction

The use of binaural room impulse responses (BRIRs) has been ubiquitous in many spatial audio applications. For headphone-based reproduction, BRIRs mainly serve as audio filters that simulate or reproduce an immersive and perceptually plausible sounding environment; for loudspeaker-based applications such as binaural reproduction with crosstalk cancellation [1, 2] and personal sound zone reproduction [3, 4], they are utilized to approximate the "plant" acoustic transfer functions of the system, based on which audio filters are designed to and achieve the desired system response at listeners' ears. It is often necessary in both cases to update the BRIR with head tracking in multiple degrees-of-freedom (DoFs, referring to head translation and rotation) in order to either enhance the immersion or improve the filter robustness against possible head misalignments [5].

When compensating BRIRs for head movements in multiple DoFs, one can synthesize the BRIR by combining the corresponding anechoic head-related impulse response (HRIR) and the room impulse response (RIR) captured either with a multichannel microphone array or a single microphone [6, 7, 8]. However, such a synthesized BRIR is usually only suitable for au-

ralization applications whose main goal is to achieve perceptual plausibility [9]. In comparison, a BRIR that better represents the actual system response (or more physically accurate) is required for loudspeaker-based applications [2, 4], and therefore interpolating between pre-measured BRIRs is more appropriate. We will mainly focus on the interpolation approach in this paper.

Typically, thousands of acoustic measurements are required for covering a wide range of head movements for a single listener. As an effort to reduce the number of measurements and the space for data storage, the modeling and interpolation of HRIRs have firstly been investigated and various methods have been proposed, including linear and spline interpolation [10], pole/zero modeling [11, 12], parametric filter modeling [13, 14], spherical harmonics decomposition [15], spatial principle component analysis [16, 17], and machine learning-based methods [18, 19]. Compared to HRIRs, BRIRs contain additional salient peaks due to room reflections, which vary greatly with head movements in multiple DoFs. Instead of traditional methods, other DSP-based methods [20, 21] that identify and align early reflection peaks based on dynamic time warping [22] were proposed, but often at a cost of additional computation for peak detection.

In this paper, we utilize deep learning to model and interpolate BRIRs, rather than relying on data-specific audio processing techniques. More specifically, we train a deep neural network (DNN) that takes head coordinates as inputs and predicts the early part of the corresponding BRIR. The DNN architecture is inspired by the work of Richard et al. [19] and the original NeRF paper [23], which introduced the model for solving the view synthesis problem in the field of computer vision. In [19], the DNN was evaluated for the task of estimating and interpolating anechoic HRIRs that only vary with head rotations. Here, we modify the DNN architecture to allow for better parallelization and extend the use of the proposed DNN to BRIRs that contain strong early reflections and vary with not only head rotations but also head translations. Furthermore, we propose a new frequency-domain formulation of the DNN by changing the output and the loss function, which proves to be more efficient than the traditional time-domain counterpart in the case of modeling band-limited BRIRs. In addition, we introduce principles for model optimization based on Fourier analysis, which provide physically meaningful guidelines for tuning

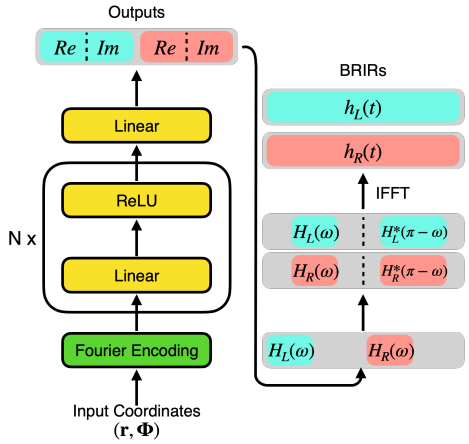

Fig. 1: Illustration of the DNN structure and further processing to generate BRIRs based on the frequency-domain formulation.

the DNN. The DNN models with both formulations are further compared against conventional methods to evaluate their capability of interpolating sparse BRIRs.

2 Model Formulation

Given a fixed sound source in a room, a typical BRIR that varies in multiple DoFs can be expressed as a function of head center position, head orientation, and time, with respect to the left or right ear of the listener:

$$h_{L,R} = h_{L,R}(\mathbf{r}, \Phi, t), \tag{1}$$

where $\mathbf{r} = (x, y, z)$ denotes the 3D coordinate of the head center and $\Phi = (\phi, \theta)$ denotes the head orientation (in this case the azimuth and elevation angles). In [19], the proposed DNN takes both position coordinates and a time index as inputs and predicts the t-th sample of the BRIR. Given that 1) the spatial and temporal variables are often separable in the description of a wave pressure field, and 2) unlike spatial variables which are continuous and require interpolation, the temporal variable in a digital audio signal is discrete by nature and can be directly used as the index of a BRIR vector, we propose a DNN architecture that learns the temporal dependency implicitly and only takes the spatial coordinates as inputs. This allows for the generation of the entire BRIR vector over a single model inference, making the model more suitable for real-time applications.

The general structure of the DNN (shown in Fig. 1) is similar to that in [19], also known as multi-layer perceptrons (MLPs). First, an input layer takes the coordinate

vector $p = (\mathbf{r}, \Phi)$ and pass it to a positional encoding layer that maps the inputs to higher dimensional Fourier features

$$\gamma(p) = \{sin(2^n \pi p), cos(2^n \pi p)\}, n = 0, 1, \cdots, N. \quad (2)$$

It has been shown [24] that such a mapping enables the DNN to better fit high-frequency variation in the data. These features are passed to a series of fully-connected layers with the ReLU activation function, and finally to an output layer that yields the entire BRIR vector. As the BRIR is expressed in the time domain, a common choice for the loss function is the l_2 loss in the time domain between the estimated and the ground truth BRIRs:

$$\mathscr{L} = \sum_t |\hat{h}_{L,R}(t) - h_{L,R}(t)|^2, \quad (3)$$

where the hat symbol denotes the estimated BRIR. However, as the early reflection part in the BRIR significantly increases the output vector length, such a formulation can be computationally expensive. Moreover, only a partial spectrum of the BRIR is of concern in certain audio applications, and therefore learning the entire time-domain vector can lead to modeling redundancy (unnecessary storage and computation cost). An alternative to learning the time-domain representation would be to learn the partial (or full) frequency spectrum of the BRIRs, which can be obtained by taking their Fourier transform. The corresponding loss function is then given by

$$\mathscr{L} = \sum_\omega (|Re\{\hat{H}_{L,R}(\omega) - H_{L,R}(\omega)\}|^2 + |Im\{\hat{H}_{L,R}(\omega) - H_{L,R}(\omega)\}|^2), \quad (4)$$

where $H_{L,R}(\omega)$ denotes the corresponding complex transfer function of the time-domain BRIR $h_{L,R}(t)$, ω denotes the angular frequency, and $Re\{\cdot\}$ and $Im\{\cdot\}$ denote taking the real and the imaginary part, respectively. The magnitude and phase representation of the complex transfer function is not adopted due to the issue of phase wrapping. The final BRIRs are generated by assembling the complex transfer function from the model output and then taking the inverse Fourier transform, as shown in Fig. 1. In this formulation, we keep only the relevant frequency band in the learning process, therefore improving the modeling efficiency.

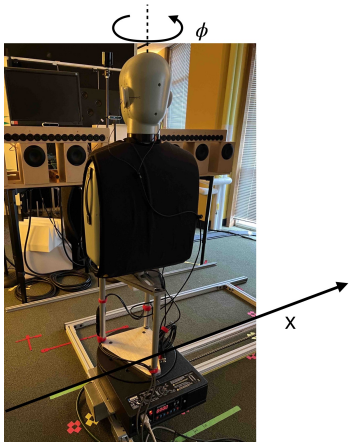

Fig. 2: Measurement setup for BRIR data collection. Note that only one loudspeaker in the photo was used for the measurement.

3 Evaluation

3.1 BRIR Datasets

We used both measured and simulated BRIR datasets to evaluate the performance of proposed DNN models. The datasets share the same scale of head movements and spatial resolution, but differ in room acoustics such as reverberation time, in order to diversify testing conditions.

Measured BRIRs. We conducted in-house BRIR measurements for a fixed sound source and with both head translations and rotations of a dummy listener. For simplicity, the head translation occurred only in one axis, with a grid of (0:1:90) cm; the head was horizontally rotated for a full circle at each new position, with a resolution of 1 degree. In total, there are $91 \times 360 = 32760$ measured BRIRs with each containing two channels. The measurements were conducted with one B&K Head and Torso Simulator (HATS, Type 4100) with in-ear binaural microphones (Theoretica Applied Physics BACCH-BM Pro) in a typical listening room ($RT_{60} = 0.24$ s in the range 1300-6300 Hz). A custom made mechanical translation stage and a turntable (Outline ET250-3D) were used to translate and rotate the dummy head, respectively, for automated measurements (see Fig. 2 for the setup). The BRIRs were measured using synchronized exponential sine sweep (ESS) [25] signals at 48 kHz sampling rate, with each sweep lasting 1 sec.

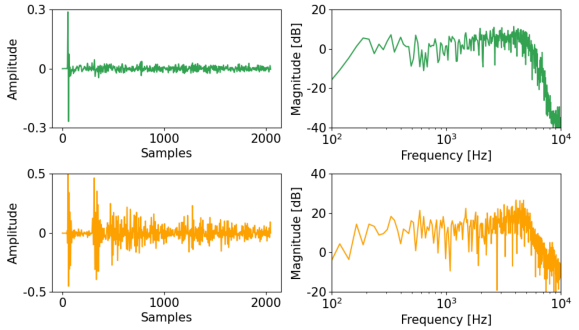

Fig. 3: Examples of measured (top) and simulated (bottom) BRIRs. Left: time-domain representation. Right: corresponding magnitude spectrum.

Simulated BRIRs. We simulated BRIRs using the RAZR room acoustics simulator [6], under a shoebox room model with dimensions of 5×4×3 m and a RT_{60} of 1 s at all frequency bands. An omnidirectional source was placed 1.5 m in front of the listener, and the early part of the BRIR was generated by convolving the spatial room impulse responses synthesized from the image source model (see [6] for details) and the corresponding HRIRs measured in-house [26]. The BRIRs were computed at the same head movement grid and resolution as for the measured BRIRs, and also at 48 kHz sampling rate.

In this study, we truncated all BRIRs to the first 2048 samples (approximately 40 ms at 48 kHz sampling rate) to include only early reflections, as the late reverb tails can be assumed to be stochastic and therefore efficiently synthesized [6]. For measured BRIRs, we performed band-pass filtering between 150 Hz and 7 kHz to exclude low-frequency noise in the room and prevent spatial aliasing as well as low signal-to-noise ratio at high frequencies due to possible head occlusion. Note that such processing steps are also of practical value as they are also used in loudspeaker-based applications (e.g., personal sound zone reproduction [4]) in order to improve the robustness of filters against transfer function mismatches. The same filtering was applied to the simulated BRIRs to allow for comparison between the two datasets. The resulting BRIRs were then used for model evaluation. Fig. 3 gives examples of one measured and one simulated BRIRs.

3.2 Model optimization

We introduce two aspects of DNN model optimization based on the particular setup of the BRIR datasets:

Input normalization. As proposed in [23], the inputs are normalized to an interval of $[-1,1]$ before the Fourier encoding layer. Although this applies to the case of modeling HRIRs with only head rotations, for BRIRs that involve head translations, such a normalization would lead to modeling errors near the two boundaries due to the aperiodic nature of translation (the Gibbs phenomenon). Therefore, we mapped the translational coordinate x to a slightly narrower range, $[-1+\Delta, 1-\Delta]$, where Δ corresponds to the spacing between two adjacent positions. The same $[-1,1]$ interval was used for normalizing the rotation angle ϕ.

Fourier encoding order. It is intuitive to understand that the different components in the Fourier encoding layer represent information at different spatial frequencies. When fitting a dataset with limited spatial resolution, it is important to find an optimal encoding order to avoid possible spatial aliasing [27]; when the data is band-limited, it is also helpful to optimize the encoding order to minimize model redundancy. Here we only consider the case of head translation as an approximation since it has lower spatial resolution compared to head rotation. To determine the maximum encoding order N_{max}, we first re-express the phase term in Eq. 2 with equivalent wavenumber $\tilde{k}$ and translational coordinate x,

$$2^n \pi p = \tilde{k} x, \tag{5}$$

where p is the normalized position coordinate, and for translations we have

$$p = \frac{x}{L/2}, \tag{6}$$

as the normalization maps x from $[-L/2, L/2]$ to $[-1,1]$. Combining the two equations above, the equivalent wavenumber can be expressed as

$$\tilde{k} = \frac{2^{n+1}\pi}{L}. \tag{7}$$

First, according to Nyquist sampling theorem, for a spatial sampling resolution Δx, the maximum wavenumber is given by

$$k_{max,1} = \frac{\pi}{\Delta x}. \tag{8}$$

In addition, considering the highest frequency of the effective bandwidth f_{max}, we also have

$$k_{max,2} = \frac{2\pi f_{max}}{c}, \tag{9}$$

where c denotes the sound speed. In practice, we choose the lower value of the two, which is then combined with Eq. 7 and yields the maximum encoding order N_{max}

$$N_{max} = \lfloor \min\{\log_2(\frac{L}{2\Delta x}), \log_2(\frac{Lf_{max}}{c})\} \rfloor, \tag{10}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. For our dataset, we have $L = 0.9$ m, $\Delta x = 1$ cm, $f_{max} = 7$ kHz, and $c = 340$ m/s, and therefore $N_{max} = 4$.

3.3 Metrics

We adopt two metrics to evaluate the DNN performance: 1) the signal-to-distortion ratio (SDR), which quantifies the modeling accuracy in the *time* domain:

$$SDR = 10\log_{10}(\frac{\|h_{L,R}(t)\|^2}{\|h_{L,R}(t) - \hat{h}_{L,R}(t)\|^2}), \tag{11}$$

and 2) the spectral distortion (SD), which quantifies the magnitude difference between the ground truth and the estimated BRIR in the *frequency* domain, first as a function of the frequency ω:

$$SD_\omega = 10\log_{10}(\frac{|H_{L,R}(\omega)|^2}{|\hat{H}_{L,R}(\omega)|^2}), \tag{12}$$

and then logarithmically averaged:

$$SD = \frac{\sum_\omega |SD_\omega|/\omega}{\sum_\omega 1/\omega}. \tag{13}$$

4 Results

4.1 Modeling efficiency

We first compare the time-domain and frequency-domain formulations in terms of modeling efficiency for band-limited BRIRs. All DNNs evaluated have an encoding order of N=4 and 4 fully-connected layers, but vary in the hidden layer size (and therefore the number of trainable parameters). The output of the time-domain model is the concatenated BRIR vectors of left and right channels, with a size of 4096; the output of the frequency-domain model is the concatenated real and imaginary parts of the two complex transfer functions *only* in between 150 Hz and 7 kHz, yielding a total size of 1168. All DNNs were trained on 70% of the entire dataset (randomly selected from either measured or simulated dataset) and evaluated on the remaining 30%.

Fig. 4 shows the SDR and SD as a function of the number of parameters that corresponds to different hidden layer sizes. Both metrics are averaged across all the evaluation samples and across the two channels of BRIRs. First, comparing the time-domain and frequency-domain models with the same hidden layer size, we note that the frequency-domain model achieves similar or better SDR and SD performance as the time-domain model, but with a much smaller number of parameters. The frequency-domain model leads to better SD performance than the time-domain one for small models (e.g., with a layer size of 32 or 64) in both datasets as the latter needs to learn a much larger output vector; the difference between the two models becomes smaller as the model size increases. Comparing results from two datasets, we note that although the dependency of model performance on the layer size is similar in both datasets, the model performance is generally worse (lower SDR and higher SD) in the simulated dataset than in the measured dataset. It is likely due to the longer reverberation time (1 s compared to 0.24 s in the measured dataset) and more early reflection peaks, as can be seen in Fig. 3. In general, both models efficiently compress the BRIR information while achieving a high level of modeling accuracy. For example, the frequency-domain model with a layer size of 512 only requires 1.4 M parameters, which equivalently compresses the original dataset (38.3 M floats) by 96.3%.

4.2 Interpolation performance

Next, we evaluate the capability of the DNN models to interpolate from known (trained) BRIRs. Both time-domain and frequency-domain DNN models used for evaluation have a fixed size of 4 layers, each with 512 hidden units. We downsample the BRIRs to a spatial resolution of 5 cm and 5 degrees instead of using the original datasets, yielding a total of 1296 BRIRs, to evaluate the performance on sparse data. We then vary the proportion of the training set in the downsampled dataset and use the remainder for evaluation. We

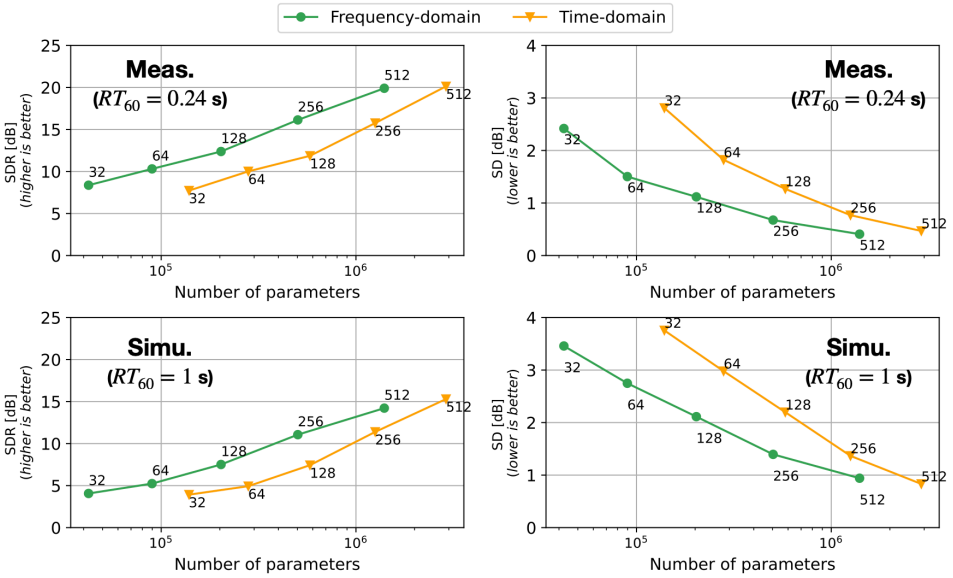

Fig. 4: Model performance in terms of SDR and SD, evaluated for both the frequency-domain and time-domain models on the measured (top) and simulated (bottom) datasets. The data points in both curves represent evaluated DNNs with different hidden layer sizes, which are marked next to each data point in the plots.

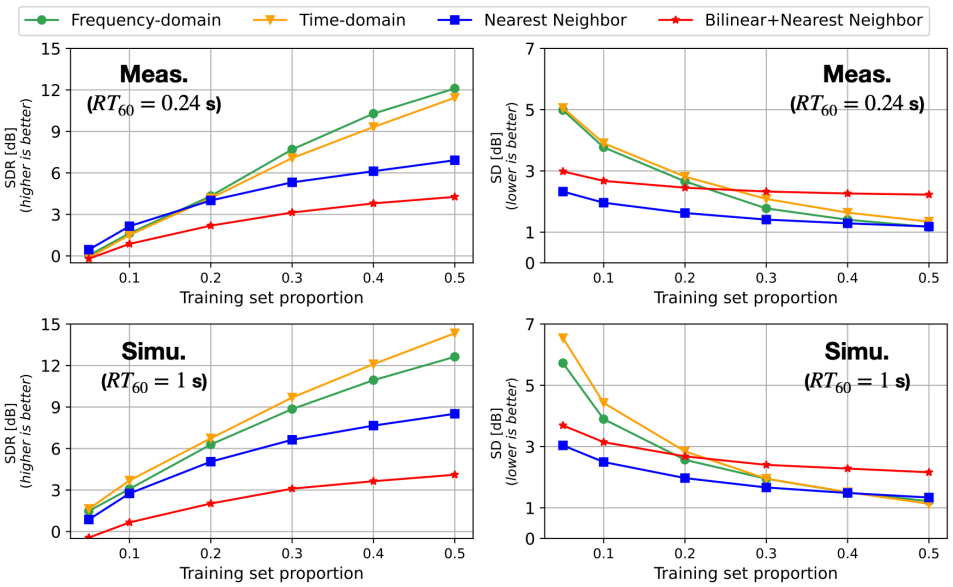

Fig. 5: Interpolation performance in terms of SDR and SD, evaluated for the frequency-domain and time-domain DNN models, the nearest neighbor method, and the bilinear interpolation method. The evaluation was performed on the downsampled version of the measured (top) and simulated (bottom) datasets.

choose the nearest neighbor and bilinear interpolation methods as baselines for comparison. The bilinear interpolation method is generalized for an arbitrary grid by polynomial fitting as the training set is randomly selected and does not follow a uniform spatial grid; in addition, the bilinear interpolation automatically falls back to the nearest neighbor whenever the predicted points are located on the boundaries. Fig. 5 shows the SDR and SD as a function of the training set proportion, for four interpolation methods evaluated on both measured and simulated datasets. Here, the SDR metric is associated with interpolation accuracy in phase, and the SD metric represents the magnitude accuracy. We note that for SDR, both DNN models outperform the two baseline methods, and the difference between the DNN models and the baselines increases as more training data is available for interpolation. The bilinear interpolation method yields worst SDR performance because direct summation of BRIRs without peak alignment would introduce errors to the location of onsets and reflection peaks, resulting in poor phase accuracy. The SD performance of the both DNN models is worse than the baselines when there are few training samples, and gradually improves as more training samples are available, until eventually it reaches the level of nearest neighbor at 50% of the training proportion. This is because the evaluated DNN models are likely underfit given few known BRIRs; we expect better performance of such models when more training data becomes available.

5 Conclusion

In this paper, we examined the use of DNNs for modeling known BRIRs and interpolating unknown sparse BRIRs that vary with head movements in multiple degrees of freedom, based on a DNN architecture originally proposed for neural representation problems and its adaptation for modeling HRIRs. More specifically, we implemented DNN models in both time-domain and frequency-domain formulations and evaluated their performance with measured and simulated BRIRs. In general, both DNN models were able to model BRIRs that contain dense reflections with relatively low distortion and a high data compression rate; the frequency-domain formulation of the DNN was shown to be twice as efficient as its time-domain counterpart for band-limited BRIRs. Such advantages can facilitate many spatial audio applications that require head-tracked rendering but have limitations in storage and/or computation. In terms of interpolating among sparse BRIRs, we also found the DNN models also to have better performance than traditional interpolation methods such as nearest neighbor and bilinear interpolation when there are sufficient training samples. However, as pointed out previously, we note that the performance of DNN models can be largely influenced by the choice of hyperparameters, such as the normalization of input coordinates, the Fourier encoding order, and the model size. Therefore, it is important to optimize the DNN model for the the specific BRIR dataset in order to balance the trade-off between the modeling accuracy and the efficiency.

References

[1] Choueiri, E., "Binaural audio through loudspeakers," in A. Roginska and P. Geluso, editors, *Immersive sound: the art and science of binaural and multi-channel audio*, chapter 6, pp. 124–179, Taylor & Francis, 2018.

[2] Majdak, P., Masiero, B., and Fels, J., "Sound localization in individualized and non-individualized crosstalk cancellation systems," *The Journal of the Acoustical Society of America*, 133(4), pp. 2055–2068, 2013.

[3] Betlehem, T., Zhang, W., Poletti, M. A., and Abhayapala, T. D., "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Processing Magazine*, 32(2), pp. 81–91, 2015.

[4] Qiao, Y. and Choueiri, E., "The Performance of A Personal Sound Zone System with Generic and Individualized Binaural Room Transfer Functions," in *Audio Engineering Society Convention 152*, Audio Engineering Society, 2022.

[5] Qiao, Y. and Choueiri, E., "Optimal Spatial Sampling of Plant Transfer Functions for Head-Tracked Personal Sound Zones," in *Audio Engineering Society Convention 154*, Audio Engineering Society, 2023.

[6] Wendt, T., Van De Par, S., and Ewert, S. D., "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.

[7] McCormack, L., Pulkki, V., Politis, A., Scheuregger, O., and Marschall, M., "Higher-order spatial impulse response rendering: Investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution," *Journal of the Audio Engineering Society*, 68(5), pp. 338–354, 2020.

[8] Arend, J. M., Garí, S. V. A., Schissler, C., Klein, F., and Robinson, P. W., "Six-degrees-of-freedom parametric spatial audio based on one monaural room impulse response," *Journal of the Audio Engineering Society*, 69(7/8), pp. 557–575, 2021.

[9] Lübeck, T., Arend, J. M., and Pörschmann, C., "Binaural reproduction of dummy head and spherical microphone array data—A perceptual study on the minimum required spatial resolution," *The Journal of the Acoustical Society of America*, 151(1), pp. 467–483, 2022.

[10] Nishino, T., Kajita, S., Takeda, K., and Itakura, F., "Interpolating head related transfer functions in the median plane," in *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No. 99TH8452)*, pp. 167–170, IEEE, 1999.

[11] Runkle, P., Blommer, M., and Wakefield, G., "A comparison of head related transfer function interpolation methods," in *Proceedings of 1995 workshop on applications of signal processing to audio and accoustics*, pp. 88–91, IEEE, 1995.

[12] Watanabe, K., Takane, S., and Suzuki, Y., "Interpolation of head-related transfer functions based on the common-acoustical-pole and residue model," *Acoustical science and technology*, 24(5), pp. 335–337, 2003.

[13] Ramos, G. and Cobos, M., "Parametric head-related transfer function modeling and interpolation for cost-efficient binaural sound applications," *The Journal of the Acoustical Society of America*, 134(3), pp. 1735–1738, 2013.

[14] Breebaart, J., Nater, F., and Kohlrausch, A., "Spectral and spatial parameter resolution requirements for parametric, filter-bank-based HRTF processing," *Journal of the Audio Engineering Society*, 58(3), pp. 126–140, 2010.

[15] Zhang, W., Abhayapala, T. D., Kennedy, R. A., and Duraiswami, R., "Insights into head-related transfer function: Spatial dimensionality and continuous representation," *The Journal of the Acoustical Society of America*, 127(4), pp. 2347–2357, 2010.

[16] Xie, B.-S., "Recovery of individual head-related transfer functions from a small set of measurements," *The Journal of the Acoustical Society of America*, 132(1), pp. 282–294, 2012.

[17] Zhang, M., Ge, Z., Liu, T., Wu, X., and Qu, T., "Modeling of individual HRTFs based on spatial principal component analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, pp. 785–797, 2020.

[18] Gebru, I. D., Marković, D., Richard, A., Krenn, S., Butler, G. A., De la Torre, F., and Sheikh, Y., "Implicit hrtf modeling using temporal convolutional networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3385–3389, IEEE, 2021.

[19] Richard, A., Dodds, P., and Ithapu, V. K., "Deep impulse responses: Estimating and parameterizing filters with deep networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3209–3213, IEEE, 2022.

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

[21] Bruschi, V., Nobili, S., Terenzi, A., and Cecchi, S., "An Improved Approach for Binaural Room Impulse Responses Interpolation in Real Environments," in *Audio Engineering Society Convention 152*, Audio Engineering Society, 2022.

[22] Masterson, C., Kearney, G., and Boland, F., "Acoustic impulse response interpolation for multichannel systems using dynamic time warping," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, Audio Engineering Society, 2009.

[23] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R., "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, 65(1), pp. 99–106, 2021.

[24] Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., and Ng, R., "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Information Processing Systems*, 33, pp. 7537–7547, 2020.

[25] Novak, A., Lotton, P., and Simon, L., "Synchronized swept-sine: Theory, application, and implementation," *Journal of the Audio Engineering Society*, 63(10), pp. 786–798, 2015.

[26] Sridhar, R., Tylka, J. G., and Choueiri, E., "A database of head-related transfer functions and morphological measurements," in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.

[27] Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., and Srinivasan, P. P., "Mip-nerf: A multiscale representation for antialiasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5855–5864, 2021.