MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Spatially-Aware Losses for Enhanced Neural Acoustic Fields

Ick, Christopher; Wichern, Gordon; Masuyama, Yoshiki; Germain, François G; Le Roux, Jonathan

TR2024-169 December 14, 2024

Abstract

For immersive audio experiences, it is essential that sound propagation is accurately modeled from a source to a listener through space. For human listeners, binaural audio characterizes the acoustic environment, as well as the spatial aspects of an acoustic scene. Recent advancements in neural acoustic fields have demonstrated spatially continuous models that are able to accurately reconstruct binaural impulse responses for a given source/listener pair. Despite this, these approaches have not explicitly examined or evaluated the quality of these reconstructions in terms of the inter-aural cues that define spatialization for human listeners. In this work, we propose extending neural acoustic field-based methods with spatially-aware metrics for training and evaluation to better capture spatial acoustic cues. We develop a dataset based on the existing SoundSpaces dataset to better model these features, and we demonstrate performance improvements by utilizing spatially-aware losses.

NeurIPS 2024 Audio Imagination Workshop

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Spatially-Aware Losses for Enhanced Neural Acoustic Fields

Christopher A. Ick[1,2*] **Gordon Wichern**[1] **Yoshiki Masuyama**[1]
François Germain[1] **Jonathan Le Roux**[1]
[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Music and Audio Research Lab (MARL), New York University, Brooklyn, NY, USA

Abstract

For immersive audio experiences, it is essential that sound propagation is accurately modeled from a source to a listener through space. For human listeners, binaural audio characterizes the acoustic environment, as well as the spatial aspects of an acoustic scene. Recent advancements in neural acoustic fields have demonstrated spatially continuous models that are able to accurately reconstruct binaural impulse responses for a given source/listener pair. Despite this, these approaches have not explicitly examined or evaluated the quality of these reconstructions in terms of the inter-aural cues that define spatialization for human listeners. In this work, we propose extending neural acoustic field-based methods with spatially-aware metrics for training and evaluation to better capture spatial acoustic cues. We develop a dataset based on the existing SoundSpaces dataset to better model these features, and we demonstrate performance improvements by utilizing spatially-aware losses.

1 Introduction

What we hear tells us a lot about the acoustic space in which we are listening. The sound of someone speaking in a recording booth will be vastly different to what we perceive if the same speaker is located in a car, a lecture hall, or an auditorium. Furthermore, where the speaker is located relative to the listener will also change what is heard; a speaker speaking face-to-face with a listener will sound significantly different to someone behind, at a distance, or around a solid object. As such, accurately rendering audio in a variety of spatial environments and conditions is a nontrivial problem, and yet one that is essential to obtain accurately-rendered and immersive sound experiences.

Traditional acoustical signal processing describes the relationship between a source, a listener, and their relative locations in an acoustic space via a time-domain transfer function called a room impulse response (RIR) [17]. While room impulse responses can be manually recorded in the field, the process is sensitive to noise and is relatively laborious [5]. Furthermore, it's impractical to record a sufficient number of RIRs at varying source/receiver positions to meet the requirements of a deep neural network. Simulation has provided an attractive solution, but is limited by computational cost, which is often traded for physical accuracy [8].

Recently, literature has emerged demonstrating modeling the RIR as a spatially-continuous neural field [10, 11, 13, 20], allowing us to infer the properties of RIRs at unmeasured locations. However, these approaches lack specific investigation into the spatial properties of the generated RIRs, focusing more on reconstruction ability.

Recent work has demonstrated benefit utilizing this spatial information when available for improving model performance [6]. In this work, we propose a new spatially-aware method for encoding a room's acoustic properties into a neural field. We propose new metrics for training and evaluating

*This work was performed while C. Ick was an intern at MERL.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

neural fields inspired by spatial cues from psychoacoustics, and evaluate the modified models on their ability to capture these spatial cues. We demonstrate the performance of these novel methods, and offer suggestions on how to utilize inter-channel cues to better capture spatial information.

2 Background

Room Acoustics Typically, a room is characterized by its room impulse response (RIR), which is a time-domain signal showing the acoustic response of a room to an instantaneous full-band excitation signal. It is typically dependent on the geometry and materials of the room surfaces, as well as the location of the excitation signal source and receiver. This idea can be extended to any configuration of microphone array, including a human-like binaural configuration, which would produce a binaural room impulse response (BRIR).

Spatial cues in binaural audio In psychoacoustics, acoustic spatial information is characterized by a variety of interaural cues stemming from differences between what a listener hears in their left and right ears [17, 19]. One cue that is thought to be important for spatial information is interaural time delay (ITD), which characterizes the small time delay between the two ears of a direct path, which depends on the azimuth of the direction of arrival of an incoming sound. A second cue of interest is the interaural level differences (ILD), which is the frequency-dependent level differences between what is heard by each ear due to head-shadowing effects.

In practice, this information can be measured in an anechoic room by placing microphones in a subject's ears and measuring impulse responses from various sound source locations, which results in an azimuth and elevation-dependent transfer function known as a head-related transfer function (HRTF). If this measurement is done in a reverberant room of fixed geometry and materials, the recording will also capture characteristics of the room and would instead be measuring a BRIR of a specific head orientation.

Neural Acoustic Fields The prerequisite for use of a neural field is the assumption that the function of interest is defined over a continuous input field. In the setting of a neural acoustic field, a binaural RIR is defined over space, a continuous field in $\mathbb{R}^3$. The typical formulation of a neural acoustic field is a map that estimates a binaural impulse response $h \in \mathbb{R}^{2\times T}$ from a given omni-directional source location $s \in \mathbb{R}^3$, listener location $l \in \mathbb{R}^3$, and listener orientation $\theta \in \mathbb{R}^2$. We assume that there exists some BRIR h for each $(s, l, \theta) \in \mathbb{R}^8$, and our goal is to model some function $f(s, l, \theta) \to h$.

Neural acoustic fields were introduced in [13], in which IR spectrograms were predicted from a 2D set of coordinates. This idea was extended in [20], which disentangles the room geometry from the source and listener, allowing for generalization across multiple scenes. Several newer approaches introduce additional context that can be provided to estimate RIRs, including room materials and geometry [10] and multimodal (visual) input [11].

Most prior works have been trained on SoundSpaces [3], as one of the largest publicly available dataset of spatially distributed BRIRs at a scale sufficient for training deep neural networks.

Loss Functions The most straightforward loss for an audio signal is the time-domain L2 loss $\mathcal{L}_{\text{time}} = \|\hat{h} - h\|_2$. However, it is more common for the loss to be computed in the time-frequency domain, typically via a short-time Fourier transform (STFT) which maps $\text{STFT}(h) \to H \in \mathbb{C}^{2\times F\times T'}$, where F denotes the number of frequency bins and T' the number of time frames. This is often split into magnitude and phase spectrograms, $H_{\text{mag}} = |H|$ and $H_{\text{phase}} = \arctan\left(\frac{\text{Re}(H)}{\text{Im}(H)}\right)$.

From this transform, we can define time-frequency domain magnitude and phase L2 losses:

$$\mathcal{L}_{\text{mag}} = \|\hat{H}_{\text{mag}} - H_{\text{mag}}\|_2, \tag{1}$$

$$\mathcal{L}_{\text{phase}} = \|\hat{H}_{\text{phase}} - H_{\text{phase}}\|_2. \tag{2}$$

In [20], the authors use time-domain L2 loss, magnitude and phase L1 losses, and spectral convergence loss $\mathcal{L}_{\text{sc}}$, which has been proven effective for time-domain signal generation [22]:

$$\mathcal{L}_{\text{sc}} = \frac{\|\hat{H}_{\text{mag}} - H_{\text{mag}}\|_2}{\|H_{\text{mag}}\|_2}. \tag{3}$$

Table 1: Recent neural acoustic field models and their associated loss functions used in training, as well as metrics used in evaluation.*INRAS uses L1-loss for spectral magnitude and phase, rather than L2 as most other models do.

Model	Loss Functions	Evaluation Metrics
NAF [13]	$\mathcal{L}_{\text{mag}}$, $\mathcal{L}_{\text{phase}}$	Spectral Loss, T60
INRAS[20]	$\mathcal{L}_{\text{time}}$, $\mathcal{L}^*_{\text{mag}}$, $\mathcal{L}^*_{\text{phase}}$, $\mathcal{L}_{\text{sc}}$	T60, C50 EDT
NACF[10]	$\mathcal{L}_{\text{mag}}$, $\mathcal{L}_{\text{dcy}}$	T60, C50, EDT
AV-NeRF[11]	$\mathcal{L}_{\text{mag}}$	T60, C50, EDT, MAG, ENV
NAF++[5]	$\mathcal{L}_{\text{mag}}$, $\mathcal{L}_{\text{dcy}}$	Spectral Loss, T60, C50, EDT
INRAS++[5]	$\mathcal{L}_{\text{mag}}$, $\mathcal{L}_{\text{sc}}$, $\mathcal{L}_{\text{dcy}}$	Spectral Loss, T60, C50, EDT

In [10], the authors use a weighted sum of the magnitude L2 loss and the decay loss $\mathcal{L}_{\text{dcy}}$, which is the L1 $\log_{10}$ of the backwards Schroeder integration [18] and characterizes the attenuation of energy of the RIR. The authors in [5] show that adding this loss to NAF and INRAS training improves results.

3 Spatial Losses

While prior approaches have shown success in reconstruction accuracy, the training/evaluation metrics do not explicitly measure inter-channel dependencies that capture spatial properties of the impulse response. For the losses, energy-based metrics such as $\mathcal{L}_{\text{mag}}$ and $\mathcal{L}_{\text{dcy}}$ could implicitly capture early power differences between each channel, but these differences will have proportionally lower weight compared to the long tail of decay where each ear should be similar. Similarly, for loss metrics, RT60, C50, and EDT are concerned with energy contained in the early and/or late reflections of an RIR, which should be similar in a binaural RIR due to the relative proximity of the two ears.

To explore differences between channels in our binaural impulse response h, we investigate the separate channels $h_l, h_r \in \mathbb{R}^T$. Inspired by ITD and ILD, we utilize simplified inter-channel losses that have been shown to be effective in training models with binaural audio [6].

Interchannel delay The ITD of a given two-channel signal in seconds can be estimated by computing the difference in direct-path distance from the ear to each receiver [17]. Because this is typically not known in a real-world setting, it is typically approximated by computing the delay that maximizes the cross-correlation between the left and right channels:

$$\text{ITD} = \frac{1}{f_s} \underset{t_d}{\text{argmax}} \sum_{t=0}^{T} h_l(t) h_r(t - t_d), \tag{4}$$

where f_s is the sampling rate, and $t_d \in \mathbb{Z}$ is a relative sample delay between the left and right channels. Because this computation requires an argmax operation, it is not differentiable, so while it can be used to evaluate a given BRIR, it cannot be used as a loss function in training. To resolve this, one can calculate a range of cross-correlation coefficients over varying delay and use these as a rough estimation for delay confidence. This can be further improved using the normalized coefficients from the generalized cross-correlation phase transform (GCC-PHAT) algorithm [6]:

$$c_{\text{gcc}}(h) = \mathcal{F}^{-1}\Big(\frac{\mathcal{F}(h_l)^* \odot \mathcal{F}(h_r)}{|\mathcal{F}(h_l)^* \odot \mathcal{F}(h_r)|}\Big), \tag{5}$$

where $c_{\text{gcc}} \in \mathbb{R}^{2T+1}$ are the GCC coefficients, $\mathcal{F}$ is the discrete Fourier transform (DFT), and * indicates the complex conjugate of a given DFT. Because GCC-PHAT scales the coefficients with approximate power at a given time-frequency bin, it is typically more robust to noise than standard cross-correlation.

We can define the GCC loss of an estimated BRIR $\hat{h}$ from a reference h by the L2 distance between GCC coefficients:

$$\mathcal{L}_{\text{GCC}} = \|c_{\text{gcc}}(h) - c_{\text{gcc}}(\hat{h})\|_2. \tag{6}$$

Interchannel level differences We can estimate the ILD between h_l and h_r by computing the log of the ratio of their average powers:

$$I(h) = \log_{10}\Big(\frac{||h_l||_2^2}{||h_r||_2^2}\Big) \tag{7}$$

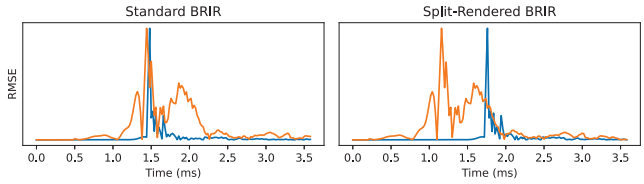

Figure 1: Side-by-side comparison of normalized RMS energy from BRIRs generated on the right side of the listener, comparing the inter-channel time difference by the standard SoundSpaces method, and by the proposed split-rendering method.

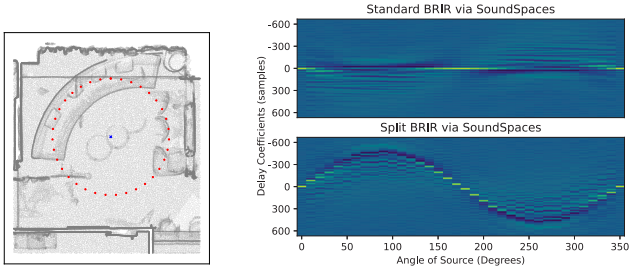

Figure 2: (Left) A circle of emitter locations around a virtual listener. (Right) Side-by-side comparison of the GCC coefficients of BRIRs computed by the standard soundspaces approach, and the split-BRIR approach around the ring.

Because this is differentiable, we can define our loss as:

$$\mathcal{L}_{\text{ILD}} = \|I(h) - I(\hat{h})\|_2. \tag{8}$$

4 Experiments

Data We use the Matterport3D RGB-D dataset to define our acoustic spaces [2], and we use the locations and orientations for the source-receiver pairs from the original SoundSpaces dataset release [3]. Across each environment, points are sampled along a 2D grid at 1.5 m height at a 1 m density, and a BRIR is computed for every source/listener pair at four listener orientations along the cardinal directions in the horizontal plane.

For this task, we utilize the SoundSpaces 2 simulation platform for rendering spatial audio [4], which renders third-order ambisonics RIR for each source/emitter pair, which is then downmixed into a binaural RIR depending on the listener orientation. Because SoundSpaces utilizes a perceptually-driven time-aligned HRTF when rendering binaural RIRs [25], we observe in the time domain that the initial direct path impulses are aligned, which implies an ITD of zero at any listener orientation. estimated using [16] across the dataset. Because the renderer estimates the sound field at a point, the differences in path length and therefore arrival time will not be correctly computed for a binaural listener of finite head size. The difference in rendered BRIRs can be seen in Fig. 1. The effect this has on the approximated delay via GCC coefficients can be see in Fig. 2.

To resolve this, we manually measure a split BRIR by simulating two listeners 20 cm apart. We take the corresponding left and right channel from each measurement to produce a separated BRIR that captures accurate ITD between the left and right channels. The spatially separated BRIR measurements should correctly capture the geometric effects of separate ears, while the HRTF convolution preserves any level differences due to masking from a head.

Model Architecture For these experiments, we use the INRAS [20] model. INRAS uses a combination of a spatial-acoustic feature embedding module, and an impulse response prediction module. The acoustic feature module is composed of 3 units, a scatter module, a bounce module, and a gather module, all of which are based on principles of acoustic radiance transfer. The scatter module uses the scene geometry to embed the relative distance from the emitter to all relevant surfaces, characterizing the emitter. The bounce module embeds scene-level acoustic features, which is characterized by the

Table 2: Performance across several loss configurations. For all metrics, lower is better.

	T60	C50	EDT	ITD	ILD
INRAS[20]	**0.3899**	4.9780	6.559e-3	2.585e-4	0.6471
+ $\mathcal{L}_{\text{ILD}}$	0.4243	**4.0814**	**4.692e-3**	**2.516e-4**	**0.3927**
+ $\mathcal{L}_{\text{GCC}}$	0.3895	4.9103	7.061e-3	2.654e-4	0.6375
+ $\mathcal{L}_{\text{ILD}}, \mathcal{L}_{\text{GCC}}$	0.4348	4.7357	5.970e-3	2.597e-4	0.4066

bounce points calculated in the scatter module. Finally, the gather module associates the listening position with the relevant bounce points, to characterize the receiver. These three modules are then translated into a shared representation as a linear combination of time-dependent basis functions via a fully-connected network, resulting in a set of 3 time-dependent spatial features. These features are concatenated and combined with the head orientation into a learned embedding, which is then used to predict the time-domain impulse responses via an MLP.

Training Losses We trained several variations of the INRAS model with different losses. Following the work of [5], our baseline model made use of magnitude loss, in addition to spectral convergence and decay loss. In our experiments, we added GCC loss as defined in Eq. 6, ILD loss as defined in Eq. 8, and both together. For each room, we use 90% of the data for training and reserve 10% of the data for evaluation.

Evaluation Metrics We evaluate our reconstructions on several quantitative reconstruction metrics. We us clarity (C50), reverberation time (T60), and early decay time (EDT) to quantify the model's ability to reconstruct the RIR [20]. In addition, we examine the ITD and ILD of the estimated RIRs. ITD is calculated using interaural cross-correlation peak delay of the IR envelope [1] and ILD is calculated looking at the time-averaged log-power of the left and right channels.

5 Results

The results of our experiments are shown in Table 2. We can see that the addition of inter-channel level difference loss improves the model's ability to accurately render RIRs with correct interchannel level differences. Furthermore, the addition of ILD error also leads to improvement across reconstruction metrics not directly tied to ILD, with the exception of T60. Because T60 is a measure of the late-decay time of an RIR, the relative power difference between two ears should be relatively low, making it relatively insensitive to errors in ILD. C50 and EDT, in comparison, are measures of direct and early power and the speed in which they decay, in which case masking effects due to head-presence are significant, making the ILD a useful measure for accurate reconstruction. While single-channel loss metrics that measure power can implicitly capture some of this information, directly investigating inter-channel behavior leads to a straightforward improvement.

The addition of GCC error shows limited impact on performance, performing similarly. Because the distance between the ears causes a relatively short time shift (on the order of hundreds of microseconds), the relative effect on T60, C50, and EDT is negligible. However, we see nearly no variation to ITD throughout any model configuration. This is likely due to the dataset defined by SoundSpaces; due to the low sampling density and relatively large multi-room acoustic environment, it is unlikely that the source and listener have a direct path between them. As a result, we do not see the well-behaved ITD behavior we would expect to see from a direct-path source/receiver pair in a shoebox environment. Further work with a dataset constrained to direct-visible source/listener pairs would likely be better suited for this task.

6 Conclusions

In this work, we investigate the role of inter-channel features for neural acoustic fields, which has so far been under-explored in prior approaches. We demonstrate the limitations of current rendering methods for RIRs and propose a simple solution via per-ear simulation. We render a dataset and use it to train and evaluate a neural acoustic field model augmented with spatially-aware interchannel loss functions, and we demonstrate that some of these enhancements improve the quality of RIRs the model is able to generate.

References

[1] A. Andreopoulou and B. F. G. Katz. Identification of perceptually relevant methods of inter-aural time difference estimation. *J. Acoust. Soc. Am.*, 142(2):588–598, 08 2017.

[2] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *Proc. 3DV*, 2017.

[3] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman. SoundSpaces: Audio-visual navigaton in 3D environments. In *Proc. ECCV*, 2020.

[4] C. Chen, C. Schissler, S. Garg, P. Kobernik, A. Clegg, P. Calamia, D. Batra, P. W. Robinson, and K. Grauman. SoundSpaces 2.0: A simulation platform for visual-acoustic learning. In *Proc. NeurIPS Datasets and Benchmarks Track*, 2022.

[5] Z. Chen, I. D. Gebru, C. Richardt, A. Kumar, W. Laney, A. Owens, and A. Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. In *Proc. CVPR*, 2024.

[6] C. Hernandez-Olivan, M. Delcroix, T. Ochiai, N. Tawara, T. Nakatani, and S. Araki. Interaural time difference loss for binaural target sound extraction. In *Proc. IWAENC*, 2024.

[7] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proc. CVPR*, pages 7132–7141, 2018.

[8] C. Ick and B. McFee. Leveraging geometrical acoustic simulations of spatial room impulse responses for improved seld. In *DCASE*, pages 56–60, September 2023.

[9] S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, and J. Brunnström. MeshRIR: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods. In *Proc. WASPAA*, 2021.

[10] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. In *Proc. ICCV*, 2023.

[11] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu. AV-NeRF: Learning neural fields for real-world audio-visual scene synthesis. In *Proc. NeurIPS*, 2023.

[12] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. In *Proc. ICLR*, April 2020.

[13] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan. Learning neural acoustic fields. In *Proc. NeurIPS*, volume 35, pages 3165–3177, 2022.

[14] S. Majumder, C. Chen, Z. Al-Halah, and K. Grauman. Few-shot audio-visual learning of environment acoustics. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Proc. NeurIPS*, 2022.

[15] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020.

[16] K. C. Poole. Spatial audio metrics. 2024.

[17] A. Roginska and P. Geluso. *Immersive sound: The Art and Science of Binaural and Multi-channel Audio*. Focal Press, 1st edition edition, 2018.

[18] M. R. Schroeder. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(3):409–412, 1965.

[19] S. S. Stevens and E. B. Newman. The localization of actual sources of sound. *The American journal of psychology*, 48(2):297–306, 1936.

[20] K. Su, M. Chen, and E. Shlizerman. Inras: Implicit neural representation for audio scenes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Proc. NeurIPS*, volume 35, pages 8144–8158. Curran Associates, Inc., 2022.

[21] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor. Binaural speech enhancement using deep complex convolutional transformer networks. In *Proc. ICASSP*, pages 681–685, 2024.

[22] R. Yamamoto, E. Song, and J.-M. Kim. Parallel Wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Proc. ICASSP*, pages 6199–6203, 2020.

[23] Q. Yang and Y. Zheng. DeepEar: Sound localization with binaural microphones. *IEEE Transactions on Mobile Computing*, 23(1):359–375, 2024.

[24] Y. Zang, Y. Wang, and M. Lee. Ambisonizer: Neural upmixing as spherical harmonics generation. *arXiv preprint arXiv:2405.13428*, 2024.

[25] M. Zaunschirm, C. Schörkhuber, and R. Höldrich. Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint. *J. Acoust. Soc. Am.*, 143(6):3616–3627, 06 2018.