

Neural Representation Learning for Impulse Response Generation

Jackie Lin
Aalto University
Master's Programme CCIS / AAT

`jackie.lin@aalto.fi`

Abstract

Impulse response modelling has been a well studied topic in the acoustics and audio signal processing fields. Advancements in deep learning have led to such methods being applied to room impulse response RIR generation, and has yielded some interesting and comparable results to traditional methods. Recently, the development of Neural Radiance Fields NeRF [10] in computer vision has been a breakthrough in neural representation learning, and has been good at interpolating unseen views within a complex 3D scene trained on sparse input views. NeRF has been adapted to acoustic scene representation and has shown to produce RIRs at unseen source-listener pair positions in Neural Acoustic Fields NAF [8]. This seminar reviews the contribution, affordances, and performance of several NeRF-based and non-NeRF-based deep learning-based RIR generation methods. I find that neural methods can be very good IR generators, provided ample ground truth IR training data.

Contents

1	Introduction	2
2	Background	3
2.1	Motivations for Neural Methods	3
2.1.1	Impulse Response Interpolation	3
2.1.2	Generality	4
2.1.3	Cross-modal learning	4
2.2	NeRF	4
2.3	Evaluation metrics	5

3	Analysis	6
3.1	NeRF – NAF	6
3.2	NeRF – IR-MLP	8
3.3	NeRF – AV-NeRF	10
3.4	Notable Non-NeRF IR Generation	12
3.4.1	Few-shotRIR	12
3.4.2	MESH2IR	12
4	Conclusion	13

1 Introduction

An impulse response IR, specifically a room impulse response RIR, between two points in a space is generated by a loudspeaker at one location playing a full frequency signal and recorded by a microphone in a second location. The resulting recorded signal is an acoustic fingerprint of the surrounding physical environment which carry spatial information about the room, such as size, obstacles, surface material, and location of sound source and receiver. Most simply, any dry signal convolved with an RIR will reconstruct how that signal would sound if played and recorded from the source-receiver position within the RIR room.

RIR are important for human localization, navigation, and environmental understanding, so RIR generation is important for realistic audio rendering in gaming and immersive experiences, and acoustic rendering for the architectural building industry. Additionally, there is a need to generate synthetic RIRs to create datasets for speech processing applications (far-field speech recognition, speech enhancement, speech dereverberation, speech separation, etc.).

RIR generation is a well researched area in acoustics and signal processing, but RIR generation continues to be a difficult task. An RIRs typically has over 10,000 time domain values which is difficult to compute. Though, the mapping between the environment information (geometry, surface materials, sound source-receiver positions) and an IR is fully described by the wave equation, it requires large amounts of computation to solve using finite element methods. Geometric methods which produce good physical approximations at high frequencies creates simulation errors for frequencies below 500 Hz, and thus can only be integrated into a hybrid wave-geometric method for high fidelity and faithfulness to the ground truth.

On the other hand, environment and psychoacoustics informed parametric approaches are not physics-based and instead rely on approximating overarching features via

equations and algorithms such as reverberation time T_{60} , direct to noise ratio DNR, and spectral decay envelope. Thus parametric approaches are good at capturing characteristics of an environment but not the best at simulating real sound paths. The RIR generation approaches described above also are largely limited by the availability of accurate geometric and material information.

Advancements in deep learning have led to the development of neural methods for RIR generation which have yielded interesting and comparable results to traditional methods. Specifically, the seminar paper will discuss motivations, strengths, limitations, and future directions of NeRF-based RIR generation methods [8, 16, 7] and non-NeRF-based neural generation methods [15, 9]. These approaches were chosen because they take spatial coordinates of either source or receiver or both as their inputs.

In Section 2 Background, the motivations for neural network IR generation approaches are discussed, and Neural Radiance Fields NeRF is explained. The evaluation metrics used in the approaches are also discussed. Section 3 analyzes NeRF based approaches and comparable non-NeRF neural approaches. Section 4 concludes the paper with a discussion on the benefits of NeRF and non-NeRF approaches, and future research directions.

2 Background

2.1 Motivations for Neural Methods

2.1.1 Impulse Response Interpolation

Impulse response interpolation is a long standing problem in signal processing and acoustics. The IR interpolation task estimates IRs for novel source-receiver positions based on measured or synthesized IRs with known source-receiver positions. Datasets of measured or FEM/geometrically synthesized IRs such as RIRs and head related impulse responses (HRIRs) have been collected for general research purposes, and need to be interpolated for various applications.

Linear interpolation [14, 1], barycentric interpolation, and nearest neighbor interpolation [18], are widely used for spatial interpolation between IRs. For example, one approach [14] in virtual reality acoustic rendering, uses real-time bilinear interpolation of a precomputed bank of spatially distributed IRs to generate IRs for dynamic sound sources. However, the linear and barycentric methods assume IRs undergo strictly linear transformations as spatial locations change, which is not an accurate representation of reality. And nearest neighbor interpolation returns the nearest existing IR, so it is not a continuous interpolation technique.

Neural architectures can be designed to be continuous and differentiable representations which makes them very suitable for interpolation tasks in general.

2.1.2 Generality

Neural methods have been proposed to improve IR generation as they can learn mappings without a priori domain knowledge. Non-neural IR interpolation methods [1, 4] show good results, but can only be applied to specific cases with specific structured data. The hybrid method in [12] uses a fine-tuned and acoustically informed neural network to condition a physical IR approach. Another hybrid IR generation method [6] uses a neural network to estimate RT60 to inform a DSP method. However, many approaches assert that more generality in the method is advantageous as it can adapt to unknown or unknowable acoustic conditions. A deep learning rule-of-thumb is minimize domain-specific manipulations on the data to increase the performance of the model.

2.1.3 Cross-modal learning

Another motivation for neural network IR or audio generation is to learn acoustic scenes from (1) easily obtainable and (2) information-rich data such as non-IR audio, video content, and images. This circumvents the difficulty in obtaining data such as the geometric and material data for wave-based and geometric methods, or real-life IR measurements which require moving loudspeakers and mics many times within a space.

Neural networks can learn acoustic information from non-audio data, which is a type of cross-modal learning. Visuals contain spatial information on geometry and materials which are the physical basis for IRs, and IRs contain spatial information on geometry and materials which are important for view generation. Cross-modal learning uses one or more modalities to guide the learning of related modalities. Research in computer vision tasks that recruit non-visual data have garnered interest, such as audio-visual aided agents [2, 3] and audio generation for video [11, 5, 22], etc. In the acoustics and audio field, cross-modal approaches have been explored for 2d single image to IR or reverberation parameters [19, 6].

2.2 NeRF

Neural Radiance Fields NeRF [20, 10] in computer vision is a breakthrough neural representation technique for interpolating unseen views within a complex 3D scene using an optimized, continuous volumetric representation trained on sparse input views. NeRF represents a complex 3D scene with infinite resolution, parameterized as a basic multilayer perceptron (MLP). NeRF is queried with a camera’s 5D positional vector, and it outputs color and density which is rendered to a 2D image.

The original NeRF architecture proposed by [10] learns a mapping from camera poses (x, y, z, θ, ϕ) to colors $\mathbf{c} = (r, g, b)$ and densities σ :

$$\text{NeRF} : (x, y, z, \theta, \phi) \rightarrow (\mathbf{c}, \sigma) \quad (1)$$

NeRF first maps a 3D position vector with positional encoding to output density σ and an intermediate feature vector with an MLP. The feature vector represents a grid-based spatial latent representation, which is used as spatial conditioning during the image generation phase. Then, the feature vector and viewing direction (θ, ϕ) are mapped to an output color $\mathbf{c}$. Lastly, a classical volume rendering technique accumulates those colors and densities into a 2D image.

The simplicity, human intelligibility, and continuity of NeRF’s spatial query domain makes NeRF an attractive parameterization for a single 3D scene representation. Additionally, NeRF learns a smooth and semantically meaningful, continuous underlying structure. NeRF overcomes the prohibitive storage costs of discretized voxel grids when modelling complex scenes, as it can generate high-resolutions images for novel camera positions.

NeRF has been adapted to acoustic scene representation in Neural Acoustic Fields NAF [8] which has produced binauralized RIRs at unseen source-listener pair positions and listener orientations. IR-MLP [16] is a NeRF inspired impulse response generation method, and AV-NeRF [7] is a NeRF-based binaural audio and video generation method. They are discussed in the Analysis section.

2.3 Evaluation metrics

The papers discussed in this seminar paper used a subset of the following evaluation metrics for generated IRs, lower is better unless noted otherwise:

- **STFT Error a.k.a. Spectral Loss Error (STFTE)** [8, 7] - the average error between predicted and target magnitude spectrograms. STFTE measures the fine-grained time-spectral agreement of a prediction to the target, but it does not measure phase error.
- **RT60 Error (RTE)** [8, 15] - the reverberation time 60 error. This metric measures how close the decay times between generate RIR and target RIR match. It captures global reverberation time but ignores finer properties.
- **Direct to Reverberant Ratio Error (DRRE)** [9, 15] - the error in the estimated energy ratio between direct and reverberant sounds. DRRE and RT60 capture the extent of acoustic mismatch in a prediction.
- **Signal to Distortion Ratio (SDR)** [16] - The ratio between the true time and the estimated domain signal or FIR filter in dB, higher is better,

$$\text{SDR}(h_{\text{true}}, h_{\text{est}}) = 10 \log_{10} \left(\frac{||h_{\text{true}}||^2}{||h_{\text{true}} - h_{\text{est}}||^2} \right)$$

- **Envelope Distance Error (ENV)** [7] - the mean squared error between envelope of the predicted and target signals, which attempt to capture important perceptual characteristics of generated binaural audio.
- **Interaural Cross Correlation Coefficient Error (IACC)** [8] - the IACC is the cross correlation coefficient between the left and right binaural channels. The metric is the mean absolute error of the IACC for a generated binaural RIR and the target binaural RIR. IACC is a commonly accepted metric for spatial localization of sounds sources from binaural audio and believed to be predictive of human localization ability.
- **Mean Option Score Error (MOSE)** [9] - measures the difference in perceptual quality between a prediction and target when convolved with human speech.
- **Early Decay Time Error (EDT)** [15] - size times the time taken for the sound source to decay by 10 dB, and it depends on the type and location of the sound source.

Beyond these metrics, several papers evaluated their approaches using various listening tests. Also, several papers visualized their predictions as wave forms, spectrograms, loudness maps, etc. to demonstrate qualitative aspects.

3 Analysis

This section examines neural methods that map sound source position, and sometimes receiver position, to IRs. This task is well suited for neural representation learning (NeRF based), yet some other non-NeRF neural approaches are discussed because they expand beyond the one-scene-per-model NeRF parameterization.

NeRF-based approaches:

- NAF
- IR-MLP
- AV-NeRF

Other neural approaches:

- Few-shotRIR
- MESH2IR

3.1 NeRF – NAF

Neural Acoustic Fields NAF[8] produces RIRs at unseen source-listener pair positions and listener orientations. NAF is a straightforward adaption of NeRF into the acoustics domain. It is important to note that NAF was motivated to cross-modally condition NeRF’s task of novel 2D image generation, and was demonstrated to improve this computer vision task in this paper.

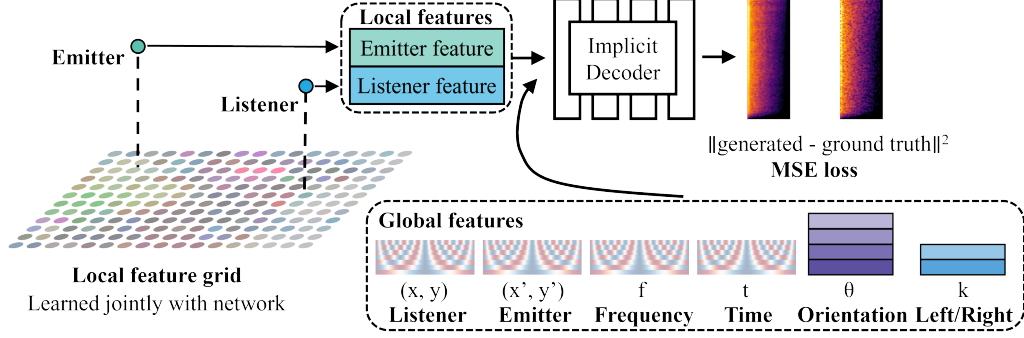


Figure 1: *NAF Architecture. Figure from [8].*

NAF learns a continuous underlying RIR function of a given scene (one or more acoustically connected spaces) from a limited recorded or simulated RIR dataset, and it generates RIRs from novel, variable source and listener positions. NAF’s inputs are source position $\mathbf{q}' = (x', y')$, receiver position $\mathbf{q} = (x, y)$, receiver orientation θ , left/right channel encoding for binaural RIRs $k \in \{0, 1\}$, time bin t , and frequency bin f , with local geometric conditioning using a feature map $grid(\mathbf{q}', \mathbf{q})$ learned during training. The model’s outputs are the log-magnitude and phase angle of the STFT of the predicted IR $\mathbf{v}$ at time-frequency bin (t, f) . Thus NAF generates a spectrogram representation of $\mathbf{v}$.

$$\text{NAF} : (\mathbf{q}', \mathbf{q}, \theta, k, t, f) \rightarrow (\mathbf{v}_{\text{STFT_mag}}(t, f), \mathbf{v}_{\text{STFT_phase}}(t, f)) \quad (2)$$

The strength of NAF is two-fold, and its benefits come from two areas:

1. NAF is highly general: NAF has a very general architecture that does not depend on any physics based parameters or tuning. Its training dataset is simply RIRs with source-listener location metadata.
2. The implicit representation of the acoustic field allows for (a) interpolation that outperforms conventional interpolation techniques and (b) high compression of IRs and fast computation:
 - (a) NAF outperforms bilinear and nearest neighbor interpolation. NAF performance was compared to interpolated IRs coded with Advanced Audio Coding (AAC), and Xiph Opus (Opus) formats, and image source method baselines on reverberation time error T_{60} and spectral loss percentage. A listening test with two-alternative forced-choice task showed that 82.38% NAF RIRs were higher quality than AAC-nearest. Figure 2 qualitatively demonstrates that NAF produces a smooth loudness map that approximates the ground truth better than the other coded-interpolation approaches.

- (b) The implicit representation of the neural field is stored within the parameters of NAF’s neural network. Once a NAF model is trained for a scene, generating an IR is as simple as a forward pass through NAF. NAF is able to better approximate the ground truth at a fraction of the storage of AAC and Opus. The average space consumption across 6 scenes is: NAF - 97.3% compression rate from AAC size (8.41 MB vs. 312 MB). Opus - 47.8% compression rate. The high compression rate suggests that NAF is efficient for IR storage (e.g. HRIR datasets which can contain thousands of IRs for one subject) with in-built interpolation.

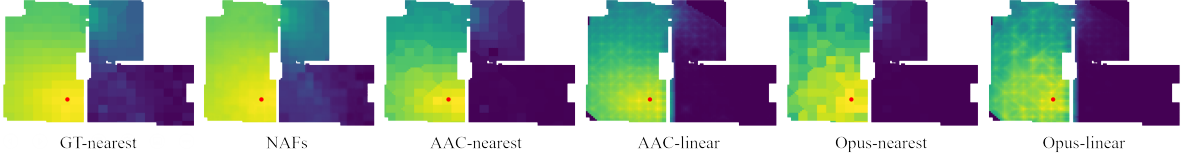


Figure 2: Loudness maps of a multi-room scene. From left to right: Ground-truth nearest neighbor, NAF, AAC nearest, AAC linear interpolation, Opus nearest, Opus linear. Notice how NAF properly captures loudness in the secondary and tertiary room, and coupled room effects at thresholds.

NAF’s reliance on ground truth RIRs hinders its application in real-world scenes where the available data are videos or images or non-IR audio, not carefully measured IR pairs at varying source and receiver positions. Additionally, NAF can only be trained on one scene at a time, so a trained NAF is not able to generalize to novel environments. There was no reporting on the speed of NAF RIR generation.

3.2 NeRF – IR-MLP

IR-MLP [16] is a NeRF inspired approach that estimates impulse responses in a highly noisy setting. The architecture of IR-MLP shown in Figure 3 is a MLP, like NeRF and NAF, and it predicts a time-domain IR as a FIR filter $h_{x,y,z}(t)$ from spatio-temporal coordinates of a sound source.

$$\text{IR-MLP} : (x, y, z, t) \rightarrow h_{x,y,z}(t) \quad (3)$$

In addition to the IR-MLP, a Noise MLP is trained to predict the spectral characteristics of stationary (no change with time) position independent or dependent noise $|N(\omega)|$. The paper proposes a novel noise-robust loss function that combines the outputs of IR-MLP and Noise MLP for joint optimization, Figure 4.

$$\text{Noise MLP} : (x, y, z) \rightarrow |N(\omega)| \quad (4)$$

The dataset used to train IR-MLP is an in-house dataset of 9,720 HRIRs of a single listener recorded in an anechoic chamber. The target signals $\hat{y}_{x,y,z}(t)$ are obtained

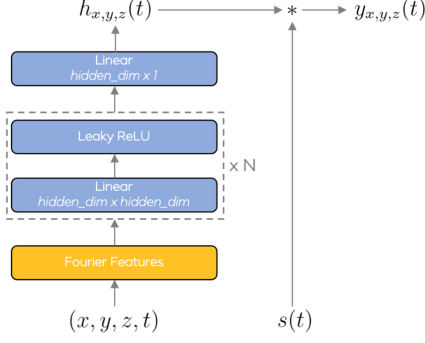


Figure 3: IR-MLP architecture. Figure from [16].

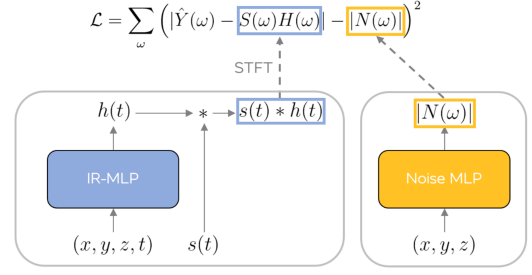


Figure 4: Noise-robust IR-MLP architecture with shared loss function. Figure from [16].

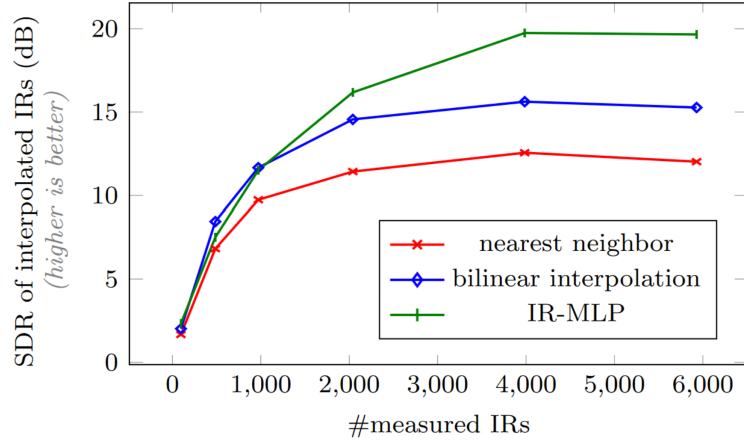


Figure 5: SDR of interpolated IRs at novel positions w.r.t. number of training IRs. IR-MLP achieves much higher quality results compared to traditional interpolation methods. Figure from [16].

by adding generated noise with predefined signal to noise ratio (SNR) $n(t)$ to the convolution of a logarithmic sine sweep $s(t)$ with each of the IRs $h_{\text{true},x,y,z}(t)$.

Signal to distortion ratio (SDR) was the evaluation metric for baseline adaptive filtering [13], baseline Wiener filter [21], IR-MLP with L2-loss, and IR-MLP with noise-robust loss. The results show that noise-robust IR-MLP outperforms the baseline methods and L2-loss IR-MLP, even on highly noisy target audio (SNR = -30 dB). Additionally, IR-MLP interpolates IRs better than nearest neighbor interpolation and bilinear interpolation – bilinear interpolation saturates at 15 dB SDR whereas IR-MLP produces up to 20 dB SDR with the same amount of training data, figure 5.

IR-MLP’s learning was tested on in-the-wild data using binaural speech recordings with tracked source and listener positions. L2-loss on the binauralized speech

data, power spectrum error, and angular error metrics showed that IR-MLP successfully learned binaural impulse responses from the data and resulted in significantly better performance than a traditional signal processing baseline and comparable performance to a neural renderer approach proposed in [17].

As IR-MLP’s parameters implicitly represent a scene, it can store IRs with extremely high compression factor and can generate IRs very quickly via a forward pass through the network. For example, a lightweight IR-MLP (layers = 6, hidden size = 32) achieves a 99.88% compression rate and can generate an IR in $0.646 \text{ ms} \pm 0.164 \text{ ms}$ with an average SDR of 11.99 dB, and a better performing IR-MLP (layers = 6, hidden size = 512) achieves a 79.16% compression rate, generating an IR in $12.975 \text{ ms} \pm 0.822 \text{ ms}$ with an average SDR of 24.437 (better SDR than adaptive filtering and Wiener filtering).

IR-MLP demonstrates again that neural representation learning allows for compact, efficient parameterization of a scene and accurate interpolation of IRs. Similar to NAF, the approach is general without domain-specific priors, and only one scene can be parameterized per model. The paper’s implementation limits the query to a changing sound source position with a static listener position; however, it would be easy to expand IR-MLP’s query to include both positions. The other difference is that IR-MLP predicts the time-domain signal instead of the time-frequency spectrum in NAF, which is a more ”end-to-end” approach. It would be interesting to compare the performance of IR-MLP to NeRF. This paper also proposes a noise-robust loss function that may be helpful when learning from high SNR datasets.

3.3 NeRF – AV-NeRF

AV-NeRF [7] is a NeRF based neural network that learns an implicit visual and binauralized audio representation of a scene. Given the position (x, y, z) and viewing direction (θ, ϕ) of a listener, AV-NeRF renders an image the listener would see and the corresponding binaural audio the listener would hear.

The AV-NeRF architecture consists of two NeRFs, V-NeRF and A-NeRF and the AV-bridge, where V-NeRF generates visual frames, A-NeRF generates acoustic masks, and AV-Bridge is optimized to extract geometric and material information from V-NeRF to inform A-NeRF [6]. The V-NeRF is the same as the original NeRF architecture proposed by [10].

A-NeRF learns a neural acoustic representation that maps camera poses (x, y, z, θ, ϕ) to corresponding acoustic mixture and difference masks $\mathbf{m}_m, \mathbf{m}_d \in R^{2 \times F}$, where F is the number of frequency bins.

$$\text{A-NeRF} : (x, y, z, \theta, \phi) \rightarrow (\mathbf{m}_m, \mathbf{m}_d) \quad (5)$$

The masks are multiplied with the time-frequency representation of the audio signal, and then the inverse STFT is taken, producing a mixture channel $\hat{a}_m$ and difference

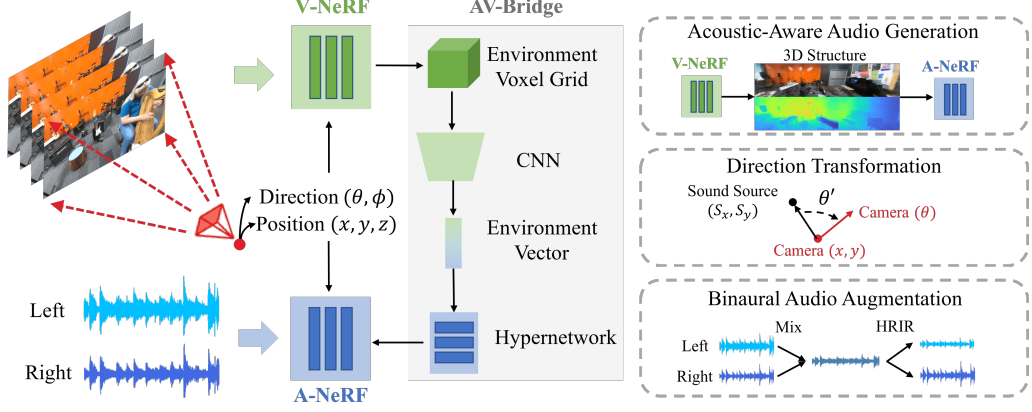


Figure 6: *AV-NeRF Architecture, with V-NeRF, A-NeRF, and AV-Bridge shown. Figure from [7].*

channel $\hat{a}_d$. These are simply recombined to give the left and right binauralized audios: $a_l = (\hat{a}_m + \hat{a}_d)/2$, $a_r = (\hat{a}_m - \hat{a}_d)/2$.

For a scene, two minutes of video and binauralized audio are collected as the training data. V-NeRF is trained separately from A-NeRF and AV-Bridge. However as demonstrated by the NAF paper, audio data could have improved the V-NeRF.

The study compared amplitude envelopes of the baseline and the AV-NeRF for a custom gathered dataset, but did not use STFT error, IACC error, DRR, or other metrics to evaluate the performance of AV-NeRF. The binaural audio generated by AV-NeRF (has AV-Bridge) had amplitude differences between the left and right channels corresponding to the relative position of the sound source to the camera position, which was not present in the standalone A-NeRF (baseline) generated audio. A perceptual user study showed that AV-NeRF achieved the most audio-visual spatial consistency, where the choices were monoaural audio, standalone A-NeRF baseline, and AV-NeRF. No other binaural audio generation methods were compared. The lack of evaluation using commonly-used metrics and comparison to outside methods in the perceptual user study leaves the evaluation of AV-NeRF questionable.

Additionally, I listened to the provided generated samples, and though there was some interaural level differences (ILD) that corresponded to whether the sound source was left or right of the camera orientation, the audio had weak spatial coherence with the generated video orientations and overall was very unconvincing compared to the training data audio. AV-NeRF though with questionable performance, is noteworthy because its training data is just a short video with binaural audio. The other methods are trained on ground-truth IRs.

3.4 Notable Non-NeRF IR Generation

3.4.1 Few-shotRIR

Few-shotRIR [9] is a transformer-based approach that generates RIRs for novel scenes. The trained model consists of an encoder and a decoder. The encoder encodes any environment using a sparse –few-shot– set of egocentric 2D images and binaural RIRs into an implicit representation. The decoder takes the implicit representation and the arbitrary source-receiver location query $\mathbf{Q} = \{s, r\}$, and outputs the log magnitude spectrogram of the predicted RIR.

Few-shotRIR is more adaptable than NAF and IR-MLP because it can produce RIRs with novel source-receiver positions for any environment, given a few observations, without retraining. Few-shot RIR achieves this with cross modal learning from 2d visual images to help IR generation.

The authors showed that *Few-shotRIR performed better than NAF* in STFT error, RTE, DRRE, and MOSE when both models were trained on the same audio dataset. Few-shotRIR has better interpolation capabilities than NAF. Additionally, the training time of Few-shotRIR is comparable to NAF (23 vs 20 hrs). However, NAF only trains on audio data with source-receiver positional tags, but Few-shotRIR trains on images as well as RIRs.

There are a few acoustic cases where Few-shotRIR is inaccurate. The RIR prediction is poor when there is large separation between the query’s source and receiver positions, as modelling how audio gets transformed on a long path becomes challenging due to the few-shot observations which encode the scene. Additionally when the source and receiver positions are close but occluded by geometry, the model incorrectly generates an RIR with a strong direct sound component. This demonstrates that RIR interpolation continues to be a challenging problem.

3.4.2 MESH2IR

[15] MESH2IR is a GAN based RIR generator for complex 3d scenes, where it encodes a complex 3d mesh as a latent vector, concatenates this vector encoding with a source-receiver location pair as input to a generator network, and outputs a fixed-length time-domain IR. The logic of MESH2IR is the same as Few-shotRIR where the model is trained generally to encode any novel environment, combine it with a position query, to output some form of IR. Both are cross-modal learning tasks. MESH2IR is general without too much domain knowledge, however some preprocessing on the ground truth IR targets are needed to reduce the noise in the predicted output.

MESH2IR was evaluated with RTE, early decay time (EDT), and DRR. The approach predicts IRs with less than 10% T_{60} and DRR errors and less than 3% EDT error. which are significantly lower than their just-noticeable-differences.

This approach requires the mesh to be simplified and converted into a graph representation and passed through the graph convolution embedding network which takes on average less than 8 seconds. From there, it can generate an IR in 0.074 ms on a single GPU, which is faster than IR-MLP. And on a CPU, MESH2IR is 200 times faster than a geometric acoustic simulator.

4 Conclusion

The translation of NeRF to the acoustic domain is impressive and valuable for accurate interpolation, efficient storage and IR generation time. Multi-modal learning, discussed in AV-NeRF, Few-shotRIR, and Mesh2IR provide compelling evidence that multi-modal conditioning from visuals or geometry improve IR generation.

Non-NeRF approaches are still competitive as they perform better than bilinear and nearest neighbor and NAF interpolation, as demonstrated with Few-shotRIR. Furthermore, non-NeRF approaches can generalize to novel scenes, whereas NeRFs can only represent one scene. Few-shotRIR, though trained on a large synthetic dataset, is able to generate IRs for a novel scene using a small set of observations, which is realistic on the number of observations that could be gathered in-the-wild or for a dynamic virtual scene. MESH2IR is also trained on a large dataset, but only needs one mesh in IR generation mode to generate any IR within that mesh.

IR generation accuracy, speed, storage, and adaptability to dynamic and novel scenes is crucial in augmented and virtual reality acoustic rendering. Also IR dataset augmentation and generation is useful for general research purposes. Neural IR generation approaches tackle these key areas. It is important to stress that geometric and wave-based acoustic IR computation and real life measurements is still incredibly necessary to build the IR datasets that all of these models train on. An assessment is needed to understand how well these methods perform compared to each other, perceptually and quantitatively. Another important research direction is to understand where, how, and why neural approaches fail in RIR generation.

References

- [1] ANTONELLO, N., DE SENA, E., MOONEN, M., NAYLOR, P. A., AND VAN WATERSCHOOT, T. Room impulse response interpolation using a sparse spatio-temporal representation of the sound field. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 10 (2017), 1929–1941.
- [2] CHEN, C., JAIN, U., SCHISSLER, C., GARI, S. V. A., AL-HALAH, Z., ITHAPU, V. K., ROBINSON, P., AND GRAUMAN, K. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV* (2020).
- [3] CHEN, C., SCHISSLER, C., GARG, S., KOBERNIK, P., CLEGG, A., CALAMIA, P. T., BATRA, D., ROBINSON, P., AND GRAUMAN, K. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *ArXiv abs/2206.08312* (2022).
- [4] DAS, O., CALAMIA, P. T., AND GARÍ, S. V. A. Room impulse response interpolation from a sparse set of measurements using a modal architecture. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), 960–964.
- [5] GAO, R., AND GRAUMAN, K. 2.5d visual sound. *CoRR abs/1812.04204* (2018).
- [6] KON, H., AND KOIKE, H. Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images. In *Audio Engineering Society Convention 144* (2018), Audio Engineering Society.
- [7] LIANG, S., HUANG, C., TIAN, Y., KUMAR, A., AND XU, C. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *arXiv preprint arXiv:2302.02088* (2023).
- [8] LUO, A., DU, Y., TARR, M. J., TENENBAUM, J. B., TORRALBA, A., AND GAN, C. Learning neural acoustic fields. *arXiv preprint arXiv:2204.00628* (2022).
- [9] MAJUMDER, S., CHEN, C., AL-HALAH, Z., AND GRAUMAN, K. Few-shot audio-visual learning of environment acoustics. In *Advances in Neural Information Processing Systems* (2022), A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds.
- [10] MILDENHALL, B., SRINIVASAN, P. P., TANCIK, M., BARRON, J. T., RAMAMOORTHY, R., AND NG, R. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020.
- [11] MORGADO, P., NVASCONCELOS, N., LANGLOIS, T., AND WANG, O. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems* 31 (2018).

- [12] PEZZOLI, M., PERINI, D., BERNARDINI, A., BORRA, F., ANTONACCI, F., AND SARTI, A. Deep prior approach for room impulse response reconstruction. *Sensors* 22, 7 (2022).
- [13] PRÍNCIPE, J. C., LIU, W., AND HAYKIN, S. *Kernel adaptive filtering: a comprehensive introduction*. John Wiley & Sons, 2011.
- [14] RAGHUVANSHI, N., SNYDER, J. M., MEHRA, R., LIN, M. C., AND GOVINDARAJU, N. K. Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes. *ACM SIGGRAPH 2010 papers* (2010).
- [15] RATNARAJAH, A., TANG, Z., ARALIKATTI, R., AND MANOCHA, D. Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes. In *Proceedings of the 30th ACM International Conference on Multimedia* (New York, NY, USA, 2022), MM '22, Association for Computing Machinery, p. 924–933.
- [16] 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)* (2022), pp. 3209–3213.
- [17] RICHARD, A., MARKOVIC, D., GEBRU, I. D., KRENN, S., BUTLER, G. A., TORRE, F., AND SHEIKH, Y. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations* (2021).
- [18] SAVIOJA, L., HUOPANIEMI, J., LOKKI, T., AND VÄÄNÄNEN, R. Creating interactive virtual acoustic environments. *Journal of The Audio Engineering Society* 47 (1999), 675–705.
- [19] SINGH, N., MENTCH, J., NG, J., BEVERIDGE, M., AND DRORI, I. Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2021), pp. 286–295.
- [20] SITZMANN, V., MARTEL, J., BERGMAN, A., LINDELL, D., AND WETZSTEIN, G. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems* 33 (2020), 7462–7473.
- [21] WIENER, N., WIENER, N., MATHEMATICIAN, C., WIENER, N., WIENER, N., AND MATHÉMATICIEN, C. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, vol. 113. MIT press Cambridge, MA, 1949.
- [22] XU, X., ZHOU, H., LIU, Z., DAI, B., WANG, X., AND LIN, D. Visually informed binaural audio generation without binaural audios. *CoRR abs/2104.06162* (2021).