DEEP IMPULSE RESPONSES: ESTIMATING AND PARAMETERIZING FILTERS WITH DEEP NETWORKS

Alexander Richard, Peter Dodds, Vamsi Krishna Ithapu

Reality Labs Research

ABSTRACT

Impulse response estimation in high noise and in-the-wild settings, with minimal control of the underlying data distributions, is a challenging problem. We propose a novel framework for parameterizing and estimating impulse responses based on recent advances in neural representation learning. Our framework is driven by a carefully designed neural network that jointly estimates the impulse response and the (apriori unknown) spectral noise characteristics of an observed signal given the source signal. We demonstrate robustness in estimation, even under low signal-to-noise ratios, and show strong results when learning from spatio-temporal real-world speech data. Our framework provides a natural way to interpolate impulse responses on a spatial grid, while also allowing for efficiently compressing and storing them for real-time rendering applications in augmented and virtual reality.

***Index Terms*—** impulse response estimation, filter interpolation, neural representation learning

1. INTRODUCTION

Robust estimation of impulse responses (IRs) is vital for a variety of audio and speech signal processing applications, including scene and room acoustics modeling, source localization, and audio spatialization. Accurately characterizing room and head-related impulse responses (RIR and HRIR) is required for achieving immersion and realism, while enabling audio and sound pesonalization in augmented and virtual reality applications.

Typically these IRs are spatio-temporal, *i.e.*, each spatial location in 3D corresponds to one short time-domain signal. Measuring IRs for every spatial location in a 3D scene is infeasible in many cases due to the cost and complexity of the setup involved and the setups themselves are typically not portable. Hence, reliable estimation and prediction of such IRs using computational models has received significant attention in the audio and speech signal processing community [1, 2, 3]. IR estimation in linear time invariant (LTI) systems typically corresponds to solving an inverse problem [4] and traditional approaches would require clean data or assume additive noise with known spectral noise characteristics. Neither of these are available for in-the-wild scenarios. While optimal or adaptive filtering techniques have been proposed, they require *apriori* knowledge or some estimate of the underlying signal statistics [5, 6, 7].

Besides robust estimation of IRs in noisy in-the-wild scenarios, interpolation of estimated IRs poses a major challenge. As it is infeasible to measure IRs at every spatial position, bilinear or barycentric interpolation are often utilized to approximate IRs at unmeasured positions [8, 9, 10]. However, such interpolation methods assume that IRs undergo strictly linear transformations as the spatial locations change, which is not an accurate representation of reality. More sophisticated IR interpolation techniques that are domain-specific have been proposed, however they suffer from generalization to arbitrary measurement grids and application domains [11].

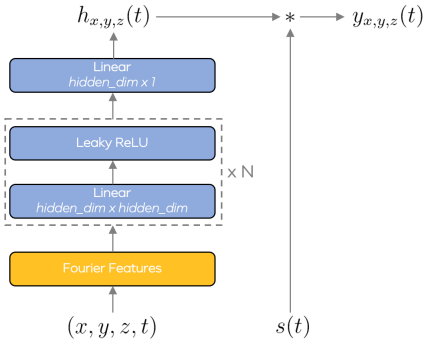


Fig. 1. The proposed IR-MLP. Impulse responses are predicted from spatio-temporal coordinates using an MLP. The final result is obtained by convolution of the IR with the source signal.

In this work, we mitigate some of these drawbacks in IR estimation and interpolation by leveraging recent advances from neural representation learning [12, 13, 14]. We propose a simple model driven by multi-layer perceptrons (MLPs) and a novel loss function for training the MLPs to jointly estimate the IR and the unknown noise characteristics. Deep networks have recently shown to be effective for learning time domain representations [15, 16, 17, 18] and MLPs are one widely studied architectural family of deep networks that are successful in acoustic modeling for speech recognition [19], audio equalization [20], and speech enhancement [21]. MLPs have not only been proven to be able to model high frequency signals [13] but also are excellent interpolation machines, providing a natural mechanism to generate IRs at unmeasured positions. While some recent existing works estimate IRs with neural networks [22, 23, 24, 25], they rely mainly on domain-specific architectures carefully designed to capture some underlying apriori data attributes [26, 27, 28], and do not always yield strong results [24]. Our framework, in contrast, is generic, without domain-specific assumptions, and applicable to any problem where IR estimation is required. In summary, our contributions are:

Impulse response estimation. We propose an effective and robust method to estimate impulse responses in a highly noisy setting where traditional approaches fail or perform worse.

Efficient parameterization of impulse responses. We demonstrate that our approach can store impulse responses with an extremely high compression factor and therefore an extremely low memory footprint, which is a key property for on-device applications.

Native interpolation of impulse responses. We show that the parameterization with a neural network is not only highly memory efficient but also allows to interpolate impulse responses at unseen positions more accurately than traditional interpolation techniques.

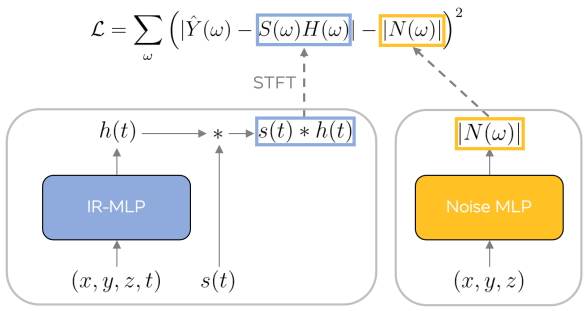


Fig. 2. In addition to the IR-MLP, a Noise MLP predicts the spectral characteristics of additive noise. A noise-robust loss function is then optimized to jointly learn the IRs and noise spectral characteristcs.

2. MODEL FORMULATION

Consider an LTI system characterized by spatio-temporal IRs, where (x, y, z) correspond to the 3D position of source (or receiver). We restrict ourselves to 3D spatial positions in this work for clarity and simplicity. Extension to higher dimensional cases is straightforward. If $s(t)$, $h(t)$ and $y(t)$ are the source, unknown IR, and the observed (target) signal respectively, then we have

$$y_{x,y,z}(t) = s(t) * h_{x,y,z}(t). \tag{1}$$

Estimating such families of filters means to learn a function

$$\mathcal{F} : (x, y, z, t) \mapsto h_{x,y,z}(t) \tag{2}$$

that maps from positional inputs (x, y, z) and the temporal index t of the filter to the t-th sample of the filter. We parameterize $\mathcal{F}$ with an MLP that consumes, as input, the spatial coordinates x, y, z and the temporal index t and predicts, as output, the t-th sample of the finite impulse response $h_{x,y,z}(t)$, see Figure 1. As pointed out in [13], MLPs typically struggle to learn high frequency functions in low dimensional domains. We therefore adopt a popular solution used in neural rendering [14] and compute Fourier features

$$\gamma(p) = \left\{ \sin(2^\ell \pi p), \cos(2^\ell \pi p); 0 \leq \ell < L \right\} \tag{3}$$

of the low dimensional input $p := (x, y, z, t)$ before feeding them into the MLP. The output of the LTI from Equation (1) is then obtained by convolution of the input signal $s(t)$ with the predicted FIR $h_{x,y,z}(t)$. We use an ℓ_2-loss for training the MLP:

$$\mathcal{L} = \sum_t \| \hat{y}_{x,y,z}(t) - y_{x,z,y}(t) \|^2, \tag{4}$$

where $\hat{y}_{x,y,z}(t)$ denotes the (measured) ground truth signal and $y_{x,y,z}(t)$ is the signal obtained by convolution of the MLP output with the input signal.

3. NOISE-ROBUST LEARNING

Practical applications typically include noise, thus a more accurate description of the system than the one in Equation (1) is

$$y(t) = s(t) * h(t) + n(t), \tag{5}$$

where $n(t)$ is an additive noise term. Note that both $h(t)$ and $n(t)$ can potentially be position dependent. To simplify notation, we drop the positional indices. In traditional approaches such as Wiener filtering, noise is typically assumed to be stationary and to have known spectral characteristics. While stationarity is a reasonable assumption, spectral characteristics of noise are generally unknown in practice. Here, we propose a technique to learn the ideal impulse responses in noisy systems with stationary noise but unknown spectral noise characteristics. This is achieved by a learned noise model, see Figure 2.

Given a source signal $s(t)$ and a measured, noisy target signal $\hat{y}(t)$, the ideal IR and noise estimates $h(t)$ and $n(t)$ minimize

$$\mathcal{L} = \sum_t \Big(\hat{y}(t) - \big[s(t) * h(t) + n(t) \big] \Big)^2. \tag{6}$$

It is generally impossible to optimize this function because the exact form of $n(t)$ is uncorrelated to the model's input. With the assumption that noise is stationary, however, we can bypass this problem. We define the residual $r(t) := \hat{y}(t) - s(t) * h(t)$ and apply Parseval's theorem such that we obtain

$$\begin{aligned} \mathcal{L} &= \sum_t \big(r(t) - n(t) \big)^2 = \sum_\omega |R(\omega) - N(\omega)|^2 \\ &= \sum_\omega |R(\omega)|^2 + |N(\omega)|^2 - 2|R(\omega)||N(\omega)| \cos(\phi_R - \phi_N), \end{aligned} \tag{7}$$

where ϕ_R and ϕ_N are the phase of the residual and noise, respectively. We assume that the stationary noise best explains the residual between the measured signal $\hat{y}(t)$ and the result of the convolution $s(t) * h(t)$. This is the case if ϕ_R and ϕ_N are chosen such that Equation (7) is minimal, which is the case for $\phi_R = \phi_N$. Then,

$$\begin{aligned} \mathcal{L} &= \sum_\omega |R(\omega)|^2 + |N(\omega)|^2 - 2|R(\omega)||N(\omega)| \\ &= \sum_\omega \Big(|\hat{Y}(\omega) - S(\omega)H(\omega)| - |N(\omega)| \Big)^2. \end{aligned} \tag{8}$$

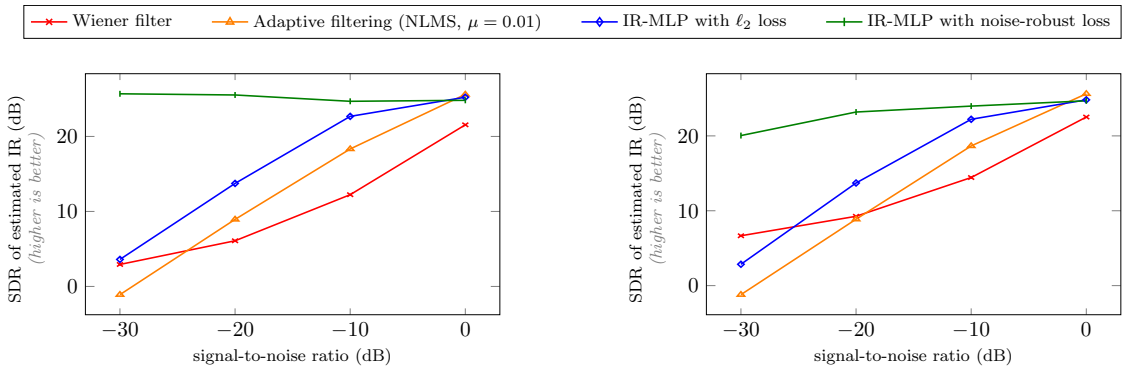
In contrast to Equation (6), this function can be optimized since it no longer depends on $n(t)$ but only on the noise amplitude spectrum $|N(\omega)|$. The latter is either a learnable constant if we assume position independent noise, or a learnable function of (x, y, z) if we assume position dependent noise. Hence, given a model that predicts both $h(t)$ and $|N(\omega)|$ as in Figure 2, the weights of the IR-MLP and the Noise MLP can simultaneously be learnt by optimization of Equation (8). In this formulation, stationarity is the only constraint on noise, knowledge of any spectral characteristics of the noise is not required.

4. EVALUATION

To show the effectiveness of our approach, we evaluate on synthetic data generated from measured HRIRs, such that ground truth IRs are known and we have full control over the noise.

Dataset. We use an in-house dataset of 9,720 HRIRs measured on a sphere around a listener in an anechoic chamber, each of which is a two-channel FIR with 400 taps per channel. We synthetically generate the observed (target) signals $\hat{y}(t)$ by convolving a logarithmic sine sweep (the source signal $s(t)$) with each of the measured IRs and add different kinds of noise. In Section 4.4, we show that our approach works on noisy real-world data as well. All audio data is sampled at 48kHz.

Network Architectures. The IR-MLP, which predicts the IRs, consumes the spatio-temporal coordinates (x, y, z, t) and maps them



(a) IR estimation with position-independent noise

(b) IR estimation with position-dependent noise

Fig. 3. Accuracy of estimated filters for different methods. We compare a Wiener filter, adaptive filtering, the IR-MLP trained with a conventional ℓ_2-loss, and the IR-MLP with the noise robust loss from Section 3. The systems are evaluated for different signal-to-noise ratios, ranging from moderately noisy data (0dB SNR) to highly noisy data (-30dB SNR). We report the SDR between estimated and ground truth filter on the y-axis. Only the IR-MLP trained with the noise robust loss is able to estimate accurate filters under highly noisy conditions.

Architecture	**Efficiency**			**Quality**
	parameters	compression	IR generation time	SDR of predicted IRs
layers=6, hidden size=32	$9.1k$	99.88%	0.646 ms $\pm$0.164 ms	11.993 dB
layers=6, hidden size=64	$30.5k$	99.61%	0.985 ms $\pm$0.247 ms	15.361 dB
layers=6, hidden size=128	$101.2k$	98.58%	1.551 ms $\pm$0.313 ms	19.172 dB
layers=6, hidden size=256	$417.0k$	94.64%	3.890 ms $\pm$0.400 ms	21.472 dB
layers=6, hidden size=512	$1.62M$	79.16%	12.975 ms $\pm$0.822 ms	24.437 dB

Table 1. A comparison of the memory- and computational efficiency versus the quality of predicted IRs for IR-MLPs of different sizes. IR-MLPs allow for a significant compression over naively storing the raw IRs, which would require 7.7M floats, while still being able to recover the parameterized IRs at high quality.

onto a higher dimensional space using Fourier features from Equation (3) with $L = 10$, followed by six fully connected layers with 512 hidden units each. If we assume position independent noise in the signal, the noise model is a simple learnable vector with $1,024$ components representing the static noise amplitude spectrum. In case of position dependent noise, the noise model consumes the spatial coordinates as input and produces a position dependent $1,024$ component noise amplitude spectrum as illustrated in Figure 2. In this case, we use the same architecture as for the IR-MLP with four instead of six layers.

Evaluation Protocol. We report the signal-to-distortion ratio (SDR) between ground truth IRs and estimated IRs,

$$\mathrm{SDR}(h_{\text{true}}, h_{\text{est}}) = 10 \log_{10} \left(\frac{\|h_{\text{true}}\|^2}{\|h_{\text{true}} - h_{\text{est}}\|^2} \right). \tag{9}$$

4.1. Learning from noisy data

We start with an evaluation of the robustness of our approach on noisy data. We investigate two scenarios: learning from data with position independent noise and with position dependent noise.

Position independent noise. To generate target signals $\hat{y}(t)$ with position independent noise, we randomly sample amplitude spectral values and scale the resulting noise spectrogram such that the signal-to-noise ratio (SNR) is at a predefined level and add random phase to the signal. The noisy target data is then obtained by adding the generated noise to the result of the convolution of the sine sweep and the ground truth IR.

Position dependent noise. Position dependent noise, or noise with spectral charachteristics that change based on the spatial position of sound source or listener, is generated similarly to position independent noise. We define two noise frequency bands of width 3kHz which are moved along the spectrum depending on the input positions, *i.e.*, the input position defines the frequency band of the noise. As before, we sample random phase and add the position dependent noise to the result of the convolution of the sine sweep and the ground truth IR.

We compare how accurately our IR-MLP learns the impulse responses when trained with a simple ℓ_2-loss on the raw waveforms and when trained with the noise-robust loss proposed in Section 3.

Baselines. As a baseline, we report the quality of IR estimation using adaptive filtering and a Wiener filter. Since the convergence of adaptive filters requires excitation signals with broadband energy and minimal autocorrelation, we replaced the sine sweep excitation used for the other methods with white noise for the adaptive filtering baseline. The Wiener filter requires knowledge about the spectral characteristics of noise $|N(\omega)|$. In order to estimate the characteristics of noise, first a noisy impulse response

$$\hat{h}_n(t) = \mathfrak{F}^{-1}\left(\frac{\hat{Y}(\omega)}{S(\omega)}\right) = \mathfrak{F}^{-1}\left(H(\omega) + \frac{N(\omega)}{S(\omega)}\right) \tag{10}$$

is estimated from the source signal $s(t)$ and the observed (target) signal $\hat{y}(t)$. Per [29], the final ten percent of the impulse response is assumed to be noise-dominated. While the spectrum of $\hat{h}_n$ does not contain the true spectrum $|N(\omega)|$ of the additive noise in the system, we obtain an estimate for $|N(\omega)|$ by multiplication of the spectrum of the noise-dominated portion of $\hat{h}_n$ with $S(\omega)$. The resulting Wiener filter is then applied to the observed signal $\hat{y}(t)$ to

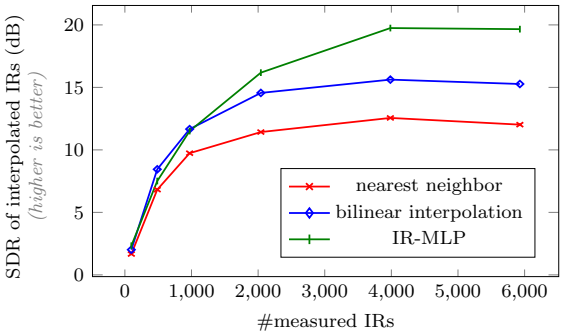


Fig. 4. Quality of interpolated IRs at new (unmeasured) positions from a sparse set of measured IRs. As the number of available measurements increases (x-axis), the quality of interpolated IRs also increases. The IR-MLP achieves much higher quality results compared to a nearest-neighbor baseline and bilinear interpolation.

recover the estimate of the denoised impulse response of the system under observation.

Results. For both, position independent noise (Figure 3a) and position dependent noise (Figure 3b), adaptive filtering, the Wiener filter, and the IR-MLP trained with a conventional ℓ_2-loss fail to estimate the filters accurately in noisy scenarios with a negative SNR. Training the IR-MLP with the noise-robust loss from Section 3, on the contrary, allows for accurate filter estimation even in highly noisy scenarios with an SNR as low as -30dB.

4.2. Interpolation: Predicting Unseen Impulse Responses

In this section, we evaluate the quality of the IR-MLP to predict IRs at positions that are not contained in the set of measured impulse responses. From the entirety of 9,720 measured IRs in the above described HRIR database, we train an IR-MLP using data generated from a limited number of the measured IRs and evaluate how accurately the network can predict the impulse responses at positions that were not part of the training data. The results are shown in Figure 4. We compare to a simple baseline where the nearest measured IR from the training set is returned as approximation to the unseen spatial position (nearest neighbor) and to bilinear interpolation. The number of available impulse responses to generate the training data is gradually increased from 100 to 6,000. Not surprisingly, all methods perform poorly when less than 1,000 measurements are available. However, traditional interpolation methods such as bilinear interpolation saturate at around 15dB SDR, while the IR-MLP produces much higher quality interpolations with an SDR up to 20dB if the spatial density of measured impulse responses is high enough.

4.3. Efficient Parameterization

Besides the previously outlined advantages of estimating and parameterizing IRs with neural networks (robustness to noise, native interpolation to unseen positions), IR-MLPs provide a memory efficient way to store IRs while being able to recover them with low computational effort. This is an important property for on-device applications where memory and compute are scarce resources. We demonstrate the trade-offs between efficiency and quality of IR-MLPs in Table 1 with an example of the HRIR database with 9,720 measured IRs. Storing all measured filters directly requires as much as 7.7M floats. Parameterizing them in a IR-MLP, on the other hand, allows for almost lossless IR recovery ($\approx$ 20dB SDR) with as few as 100k floats, which is a compression of more than 98% compared to naively storing all measured filters. Restoring the IRs requires a forward pass through the network. With the 100k parameter model (6 hidden layers, 128 hidden units per layer), this requires only 1.5ms on a single CPU core on a Macbook Pro. Different neural network sizes equip this approach with great flexibility: in applications where memory efficiency is paramount, the network size can easily be reduced until the requirements are met. In applications where quality is paramount, on the other hand, larger networks allow for higher accuracy of predicted IRs at the cost of memory and compute.

	ℓ_2 loss ($\times 10^3$)	power	phase	latency
DSP baseline [22]	0.485	0.058	1.388	25.0 ms
neural renderer [22]	**0.167**	0.048	**0.807**	†32.8 ms
IR-MLP *(ours)*	0.236	**0.042**	0.933	**13.1** ms

Table 2. Performance of rendering binaural audio where the IRs are learned from real-world speech data. IR-MLPs clearly outperform a traditional DSP baseline and are almost as strong as the neural renderer from [22]. IR-MLPs can be run at much lower latency than the approach from [22], which even requires a GPU (indicated by †).

4.4. Learning from in-the-wild data

We demonstrate that our approach can learn accurate impulse responses from noisy in-the-wild data. Therefore, we use a 2h dataset of binaural speech recordings and tracked source and listener positions [22]. Since the ground truth binaural impulse responses are unknown for such a real-world dataset, we follow the evaluation protocol from [22] and report ℓ_2-loss on the binauralized speech data and on the power spectrum as well as angular error on the phase spectrum. For the IR-MLP, we use six layers with 512 hidden units per layer as in the experiments above. The IR-MLP successfully learns binaural impulse responses from the in-the-wild data, resulting in significantly better performance than a traditional signal processing baseline has, see Table 2. Compared to the neural renderer from [22], our approach performs slightly worse, particularly due to a higher phase angular error. Note, however, that [22] is heavily optimized for binaural rendering and includes physical priors such as time warping, whereas our approach is a generic neural network without explicit assumptions and domain-specific components. Also note that [22] requires a GPU to run at 32ms latency, while our approach runs with only 13ms latency on a single CPU core. The improved latency compared to the DSP baseline (13ms vs. 25ms) can be explained by a lower number of samples that need to be buffered for our approach: the network can produce IRs at dense spatial positions that change smoothly over time, which allows it to produce smooth audio outputs even with small frame sizes and without an overlap-add operation, saving compute and reducing latency.

5. CONCLUSION

We proposed a framework that can effectively estimate and interpolate IRs using deep neural networks which leverage recent advances in neural representation learning. Our approach proves to be robust to low signal-to-noise ratios in the observed signals and allows to handle in-the-wild data. The latter is a particularly important advancement as it allows to estimate IRs directly from user-collected data rather than from idealized lab recordings that require complex and costly equipment. In its current formulation, the framework is generic and directly applicable to a wide variety of IR estimation and interpolation problems.

6. REFERENCES

[1] Yuanqing Lin and Daniel D Lee, "Bayesian regularization and nonnegative deconvolution for room impulse response estimation," *IEEE Transactions on Signal Processing*, 2006.

[2] Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Černockỳ, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[3] BoSun Xie, XiaoLi Zhong, Dan Rao, and ZhiQiang Liang, "Head-related transfer function database and its analyses," *Science in China Series G: Physics, Mechanics and Astronomy*, 2007.

[4] Francesco Dinuzzo, "Kernels for linear time invariant system identification," *SIAM Journal on Control and Optimization*, 2015.

[5] Norbert Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*, The MIT Press, 1949.

[6] Bernard Widrow, John R Glover, John M McCool, John Kaunitz, Charles S Williams, Robert H Hearn, James R Zeidler, Eugene Dong Jr., and Robert C Goodlin, "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE*, 1975.

[7] Weifeng Liu, Jose C Principe, and Simon Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*, Wiley, 2010.

[8] Niccolo Antonello, Enzo De Sena, Marc Moonen, Patrick A Naylor, and Toon Van Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[9] German Ramos and Maximo Cobos, "Parametric head-related transfer function modeling and interpolation for cost-efficient binaural sound applications," *The Journal of the Acoustical Society of America*, 2013.

[10] Hannes Gamper, "Head-related transfer function interpolation in azimuth, elevation, and distance," *The Journal of the Acoustical Society of America*, 2013.

[11] Orchisama Das, Paul Calamia, and Sebastia V Amengual Gari, "Room impulse response interpolation from a sparse set of measurements using a modal architecture," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2021.

[12] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein, "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems*, 2020.

[13] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Information Processing Systems*, 2020.

[14] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European Conf. on Computer Vision*, 2020.

[15] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," in *ISCA Speech Synthesis Workshop*, 2016.

[16] Szu-Wei Fu, Yu Tsao, Xugang Lu, and Hisashi Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2017.

[17] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2018.

[18] Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial audio synthesis," in *Int. Conf. on Learning Representations*, 2019.

[19] Simon Wiesler, Alexander Richard, Ralf Schluter, and Hermann Ney, "Mean-normalized stochastic gradient for large-scale deep learning," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2014.

[20] Giovanni Pepe, Leonardo Gabrielli, Stefano Squartini, and Luca Cattani, "Designing audio equalization filters by deep neural networks," *Applied Sciences*, 2020.

[21] Soha A Nossier, Julie Wall, Mansour Moniri, Cornelius Glackin, and Nigel Cannings, "A comparative study of time and frequency domain approaches to deep learning based speech enhancement," in *Int. Joint Conf. on Neural Networks*, 2020.

[22] Alexander Richard, Dejan Markovic, Israel D Gebru, Steven Krenn, Gladstone Butler, Fernando de la Torre, and Yaser Sheikh, "Neural synthesis of binaural speech from mono audio," in *Int. Conf. on Learning Representations*, 2021.

[23] Israel D Gebru, Dejan Markovic, Alexander Richard, Steven Krenn, Gladstone Butler, Fernando de la Torre, and Yaser Sheikh, "Implicit hrtf modeling using temporal convolutional networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2021.

[24] Christian J Steinmetz, Vamsi Krishna Ithapu, and Paul Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," *arXiv preprint arXiv:2107.07503*, 2021.

[25] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha, "Ir-gan: Room impulse response generator for far-field speech recognition," *Interspeech*, 2021.

[26] Hannes Gamper and Ivan J Tashev, "Blind reverberation time estimation using a convolutional neural network," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[27] Nicholas J Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2020.

[28] Wangyang Yu and W Bastiaan Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[29] Anders Lundeby, H. Bietz, T. Vigran, and Michael Vorlander, "Uncertainties of measurements in room acoustics," *Acta Acustica united with Acustica*, 1995.