Scene-Aware Audio Rendering via Deep Acoustic Analysis

Zhenyu Tang, Nicholas J. Bryan, Dingzeyu Li, Timothy R. Langlois, and Dinesh Manocha

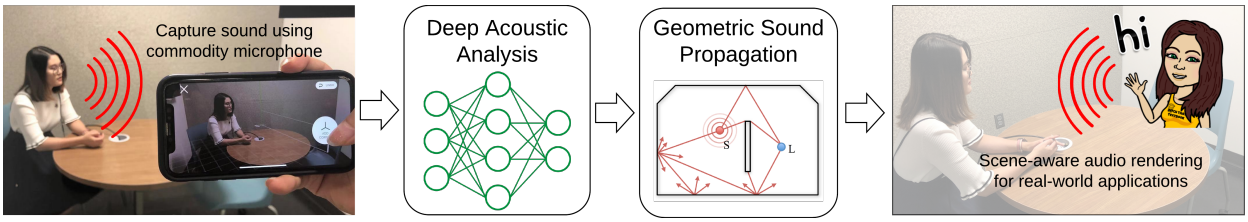


Fig. 1: Given a natural sound in a real-world room that is recorded using a cellphone microphone (left), we estimate the acoustic material properties and the frequency equalization of the room using a novel deep learning approach (middle). We use the estimated acoustic material properties for generating plausible sound effects in the virtual model of the room (right). Our approach is general and robust, and works well with commodity devices.

Abstract— We present a new method to capture the acoustic characteristics of real-world rooms using commodity devices, and use the captured characteristics to generate similar sounding sources with virtual models. Given the captured audio and an approximate geometric model of a real-world room, we present a novel learning-based method to estimate its acoustic material properties. Our approach is based on deep neural networks that estimate the reverberation time and equalization of the room from recorded audio. These estimates are used to compute material properties related to room reverberation using a novel material optimization objective. We use the estimated acoustic material characteristics for audio rendering using interactive geometric sound propagation and highlight the performance on many real-world scenarios. We also perform a user study to evaluate the perceptual similarity between the recorded sounds and our rendered audio.

Index Terms—Audio rendering, audio learning, material optimization.

1 Introduction

Auditory perception of recorded sound is strongly affected by the acoustic environment it is captured in. Concert halls are carefully designed to enhance the sound on stage, even accounting for the effects an audience of human bodies will have on the propagation of sound [2]. Anechoic chambers are designed to remove acoustic reflections and propagation effects as much as possible. Home theaters are designed with acoustic absorption and diffusion panels, as well as with careful speaker and seating arrangements [47].

The same acoustic effects are important when creating immersive effects for virtual reality (VR) and augmented reality (AR) applications. It is well known that realistic sounds can improve a user's sense of presence and immersion [33]. There is considerable work on interactive sound propagation in virtual environments based on geometric and wave-based methods [7, 43, 53, 72]. Furthermore, these techniques are increasingly used to generate plausible sound effects in VR systems and games, including Microsoft Project Acoustics[1], Oculus Spatializer[2], and Steam Audio[3]. However, these methods are limited to synthetic scenes where an exact geometric representation of the scene and acoustic material properties are known apriori.

In this paper, we address the problem of rendering realistic sounds that are similar to recordings of real acoustic scenes. These capabilities are needed for VR as well as AR applications [11], which often use recorded sounds. Foley artists often record source audio in environments similar to the places the visual contents were recorded in. Similarly, creators of vocal content (e.g. podcasts, movie dialogue, or video voice-overs), carefully re-record content made in different environment or with different equipment to match the acoustic conditions. However, these processes are expensive, time-consuming, and cannot adapt to spatial listening location. There is strong interest in developing automatic spatial audio synthesis methods.

For VR or AR content creation, acoustic effects can also be captured with an impulse response (IR) – a compact acoustic description of how sound propagates from one location to another in a given scene. A given IR can be convolved with any virtual sound or dry sound to generate the desired acoustic effects. However, recording the IRs of real-world scenes can be challenging, especially for interactive applications. Many times special recording hardware is needed to record the IRs. Furthermore, the IR is a function of the source and listener positions and it needs to be re-recorded as either position changes.

Our goal is to replace the step of recording an IR with an unobtrusive method that works on in-situ speech recordings and video signals and uses commodity devices. This can be regarded as an acoustic analogy of visual relighting [13]: to light a new visual object in an image, traditional image based lighting methods require the capture of real-world illumination as an omnidirectional, high dynamic range (HDR) image. This light can be applied to the scene, as well as on a newly inserted object, making the object appear as if it was always in the scene. Recently, Gardner et al. [20] and Hold-Geoffroy et al. [25] proposed convolutional neural network (CNN)-based methods to estimate HDR indoor or outdoor illumination from a single low dynamic range (LDR) image. These high-quality visual illumination estimation methods enable novel interactive applications. Concurrent work from LeGendre et al. [34] demonstrates the effectiveness on mobile devices, enabling photorealistic mobile mixed reality experiences.

In terms of audio "relighting" or reproduction, there have been several approaches proposed toward realistic audio in 360° images [29],

- *Zhenyu Tang and Dinesh Manocha are with the University of Maryland. E-mail: {zhy,dm}@cs.umd.edu.*
- *Nicholas J. Bryan, Dingzeyu Li and Timothy R. Langlois are with Adobe Research. E-mail: {nibryan,dinli,tlangloi}@adobe.com.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

[1] https://aka.ms/acoustics
[2] https://developer.oculus.com/downloads/package/oculus-spatializer-unity
[3] https://valvesoftware.github.io/steam-audio
Project website `https://gamma.umd.edu/pro/sound/sceneaware`

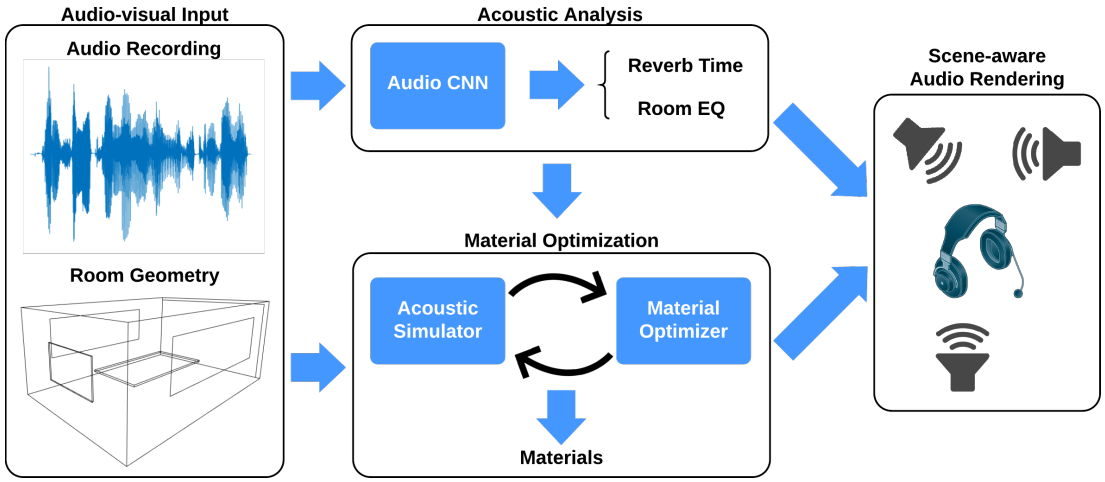


Fig. 2: **Our pipeline:** Starting with an audio-video recording (left), we estimate the 3D geometric representation of the environment using standard computer vision methods. We use the reconstructed 3D model to simulate new audio effects in that scene. To ensure our simulation results perceptually match recorded audio in the scene, we automatically estimate two acoustic properties from the audio recordings: frequency-dependent reverberation time or T_{60} of the environment, and a frequency-dependent equalization curve. The T_{60} is used to optimize the frequency-dependent absorption coefficients of the materials in the scene. The frequency equalization filter is applied to the simulated audio, and accounts for the missing wave effects in geometrical acoustics simulation. We use these parameters for interactive scene-aware audio rendering (right).

multi-modal estimation and optimization [52], and scene-aware audio in 360° videos [35]. However, these approaches either require separate recording of an IR, or produce audio results that are perceptually different from recorded scene audio. Important acoustic properties can be extracted from IRs, including the reverberation time (T_{60}), which is defined as the time it takes for a sound to decay 60 decibels [32], and the frequency-dependent amplitude level or equalization (EQ) [22].

Main Results: We present novel algorithms to estimate two important environmental acoustic properties from recorded sounds (e.g. speech). Our approach uses commodity microphones and does not need to capture any IRs. The first property is the frequency-dependent T_{60}. This is used to optimize absorption coefficients for geometric acoustic (GA) simulators for audio rendering. Next, we estimate a frequency equalization filter to account for wave effects that cannot be modeled accurately using geometric acoustic simulation algorithms. This equalization step is crucial to ensuring that our GA simulator outputs perceptually match existing recorded audio in the scene.

Estimating the equalization filter *without an IR* is challenging since it is not only speaker dependent, but also scene dependent, which poses extra difficulties in terms of dataset collection. For a model to predict the equalization filtering behavior accurately, we need a large amount of diverse speech data and IRs. Our key idea is a novel dataset augmentation process that significantly increases room equalization variation. With robust room acoustic estimation as input, we present a novel inverse material optimization algorithm to estimate the acoustic properties. We propose a new objective function for material optimization and show that it models the IR decay behavior better than the technique by Li et al. [35]. We demonstrate our ability to add new sound sources in regular videos. Similar to visual relighting examples where new objects can be rendered with photorealistic lighting, we enable audio reproduction in any regular video with existing sound with applications for mixed reality experiences. We highlight their performance on many challenging benchmarks.

We show the importance of matched T_{60} and equalization in our perceptual user study §5. In particular, our perceptual evaluation results show that: (1) Our T_{60} estimation method is perceptually comparable to all past baseline approaches, even though we do not require an explicit measured IR; (2) Our EQ estimation method improves the performance of our T_{60}-only approach by a statistically significant amount ($\approx$ 10 rating points on a 100 point scale); and (3) Our combined method (T_{60}+EQ) outperforms the average room IR ($T_{60} = .5$ seconds with uniform EQ) by a statistically significant amount ($+10$ rating points) – the only reasonable comparable baseline we could conceive that does not require an explicit IR estimate. To the best of our knowledge, ours is the first method to predict IR equalization from raw speech data and validate its accuracy. Our main contributions include:

- A CNN-based model to estimate frequency-dependent T_{60} and equalization filter from real-world speech recordings.
- An equalization augmentation scheme for training to improve the prediction robustness.
- A derivation for a new optimization objective that better models the IR decay process for inverse materials optimization.
- A user study to compare and validate our performance with current state-of-the-art audio rendering algorithms. Our study is used to evaluate the perceptual similarity between the recorded sounds and our rendered audio.

2 Related Work

Cohesive audio in mixed reality environments (when there is a mix of real and virtual content), is more difficult than in fully virtual environments. This stems from the difference between "Plausibility" in VR and "Authenticity" in AR [29]. Visual cues dominate acoustic cues, so the perceptual difference between how audio sounds and the environment in which it is seen is smaller than the perceived environment of two sounds. Recently, Li et al. introduced scene-aware audio to optimize simulator parameters to match the room acoustics from existing recordings [35]. By leveraging visual information for acoustic material classification, Schissler et al. demonstrated realistic audio for 3D-reconstructed real-world scenes [52]. However, both of these methods still require explicit measurement of IRs. In contrast, our proposed pipeline works with any input speech signal and commodity microphones.

Sound simulation can be categorized into wave-based methods and geometric acoustics. While wave-based methods generally produce more accurate results, it remains an open challenge to build a real-time universal wave solver. Recent advances such as parallelization via rectangular decomposition [38], pre-computation acceleration structures [36], and coupling with geometric acoustics [48, 73] are used for interactive applications. It is also possible to precompute low-frequency wave-based propagation effects in large scenes [45], and to perceptually compress them to reduce runtime requirements [44]. Even with the massive speedups presented, and a real-time runtime engine, these methods still require tens of minutes to hours of pre-computation depending

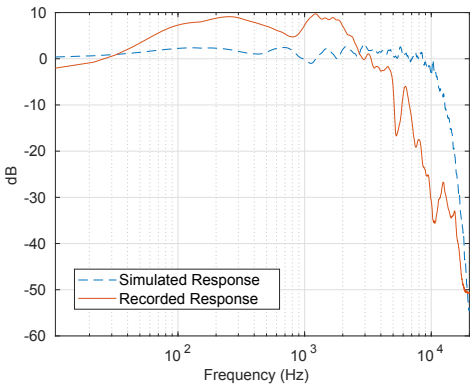


Fig. 3: The simulated and recorded frequency response in the same room at a sample rate of 44.1kHz is shown. Note that the recorded response has noticeable peaks and notches compared with the relatively flat simulated response. This is mainly caused by room equalization. Missing proper room equalization leads to discrepancies in audio quality and overall room acoustics.

on the size of the scene and frequency range chosen, making them impractical for augmented reality scenarios and difficult to include in an optimization loop to estimate material parameters. With interactive applications as our goal, most game engines and VR systems tend to use geometric acoustic simulation methods [7, 53, 54, 72]. These algorithms are based on fast ray tracing and perform specular and diffuse reflections [50]. Some techniques have been proposed to approximate low-frequency diffraction effects using ray-tracing [48, 66, 69]. Our approach can be combined with any interactive audio simulation method, though our current implementation is based on bidirectional ray tracing [7]. The sound propagation algorithms can also be used for acoustic material design optimization for synthetic scenes [37].

The efficiency of deep neural networks has been shown in audio/video-related tasks that are challenging for traditional methods [17, 21, 24, 61, 71]. Hershey et al. showed that it is feasible to use CNNs for large-scale audio classification problems [23]. Many deep neural networks require a large amount of training data. Salamon et al. used data augmentation to improve environmental sound classification [49]. Similarly, Bryan estimates the T_{60} and the direct-to-reverberant ratio (DRR) from a single speech recording via augmented datasets [5]. Tang et al. trained CRNN models purely based on synthetic spatial IRs that generalize to real-world recordings [63–65]. We strategically design an augmentation scheme to address the challenge of equalization's dependence on both IRs and speaker voice profiles, which is fully complimentary to all prior data-driven methods.

Acoustic simulators require a set of well-defined material properties. The material absorption coefficient is one of the most important parameters [4], ranging from 0 (total reflection) to 1 (total absorption). When a reference IR is available, it is straightforward to adjust room materials to match the energy decay of the simulated IR to the reference IR [35]. Similarly, Ren et al. optimized linear modal analysis parameters to match the given recordings [46]. A probabilistic damping model for audio-material reconstruction has been presented for VR applications [60]. Unlike all previous methods which require a clean IR recording for accurate estimation and optimization of boundary materials, we infer typical material parameters including T_{60} values and equalization from raw speech signals using a CNN-based model.

Analytical gradients can significantly accelerate the optimization process. With similar optimization objectives, it was shown that additional gradient information can boost the speed by a factor of over ten times [35, 52]. The speed gain shown by Li et al. [35] is impressive, and we further improve the accuracy and speed of the formulation. More specifically, the original objective function evaluated energy decay relative to the first ray received (the direct sound if there were no obstacles). However, energy estimates can be noisy due to both the oscillatory nature of audio as well as simulator noise. Instead, we optimize the slope of the best fit line of ray energies to the desired energy decay (defined by the T_{60}), which we found to be more robust.

Table 1: Notation and symbols used throughout the paper.

T_{60}	Reverberation time for sound energy to drop by 60dB.
t	Sound arrival time.
ρ	Frequency dependent sound absorption coefficient.
e_j	Energy carried by a sound path j.
β	Air absorption coefficient.
m	Slope of the energy curve envelope.

3 Deep Acoustic Analysis: Our Algorithm

In this section, we overview our proposed method for scene-aware audio rendering. We begin by providing background information, discuss how we capture room geometry, and then proceed with discussing how we estimate the frequency dependent room reverberation and equalization parameters directly from recorded speech. We follow by discussing how we use the estimated acoustic parameters to perform acoustic materials optimization such that we calibrate our virtual acoustic model with real-world recordings.

3.1 Background

To explain the motivation of our approach, we briefly elaborate on the most difficult parts of previous approaches, upon which our method improves. Previous methods require an impulse response of the environment to estimate acoustic properties [35, 52]. Recording an impulse response is a non-trivial task. The most reliable methods involve playing and recording Golay codes [19] or sine sweeps [18], which both play loud and intrusive audio signals. Also required are a fairly high-quality speaker and microphone with constant frequency response, small harmonic distortion and little crosstalk. The speaker and microphone should be acoustically separated from surfaces, i.e., they shouldn't be placed directly on tables (else surface vibrations could contaminate the signal). Clock drift between the source and microphone must be accounted for [6]. Alternatively, balloon pops or hand claps have been proposed for easier IR estimation, but require significant post-processing and still are very obtrusive [1, 56]. In short, correctly recording an IR is not easy, and makes it challenging to add audio in scenarios such as augmented reality, where the environment is not known beforehand and estimation must be done interactively to preserve immersion.

Geometric acoustics is a high-frequency approximation to the wave equation. It is a fast method, but assumes that wavelengths are small compared to objects in the scene, while ignoring pressure effects [50]. It misses several important wave effects such as diffraction and room resonance. Diffraction occurs when sound paths bend around objects that are of similar size to the wavelength. Resonance is a pressure effect that happens when certain wavelengths are either reinforced or diminished by the room geometry: certain wavelengths create peaks or troughs in the frequency spectrum based on the positive or negative interference they create [12].

We model these effects with a linear finite impulse response (FIR) equalization filter [51]. We compute the discrete Fourier transform on the recorded IR over all frequencies, following [35]. Instead of filtering directly in the frequency domain, we design a linear phase EQ filter with 32ms delay to compactly represent this filter at 7 octave bin locations. We then blindly estimate this compact representation of the frequency spectrum of the impulse response as discrete frequency gains, without specific knowledge of the input sound or room geometry. This is a challenging estimation task. Since the convolution of two signals (the IR and the input sound) is equivalent to multiplication in the frequency domain, estimating the frequency response of the IR is equivalent to estimating one multiplicative factor of a number without constraining the other. We are relying on this approach to rec-

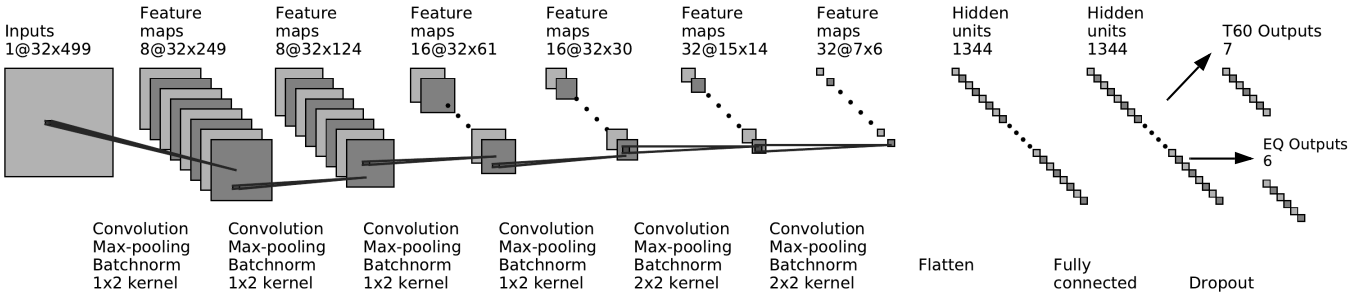


Fig. 4: Network architecture for T_{60} and EQ prediction. Two models are trained for T_{60} and EQ, which have the same components except the output layers have different dimensions customized for the octave bands they use.

ognize a compact representation of the frequency response magnitude in different environments.

3.2 Geometry Reconstruction

Given the background, we begin by first estimating the room geometry. In our experiments, we utilize the ARKit-based iOS app MagicPlan[4] to acquire the basic room geometry. A sample reconstruction is shown in Figure 5. With computer vision research evolving rapidly, we believe constructing geometry proxies from video input will become even more robust and easily accessible [3, 74].

3.3 Learning Reverberation and Equalization

We use a convolutional neural network (Figure 4) to predict room equalization and reverberation time (T_{60}) directly from a speech recording. Training requires a large number of speech recordings with known T_{60} and room equalization. The standard practice is to generate speech recordings from known real-world or synthetic IRs [14, 28]. Unfortunately, large scale IR datasets do not currently exist due to the difficulty of IR measurement; most publicly available IR datasets have fewer than 1000 IR recordings. Synthetic IRs are easy to obtain and can be used, but again lack wave-based effects as well as other simulation deficiencies. Recent work has addressed this issue by combining real-word IR measurements with augmentation to increase the diversity of existing real-world datasets [5]. This work, however, only addresses T_{60} and DRR augmentation, and lacks a method to augment the frequency-equalization of existing IRs. To address this, we propose a method to do this in Section 3.3.2. Beforehand, however, we discuss our neural network estimation method for estimating both T_{60} and equalization.

3.3.1 Octave-Based Prediction

Most prior work takes the full-frequency range as input for prediction. For example, one closely related work [5] only predicts one T_{60} value for the entire frequency range (full-band). However, sound propagates and interacts with materials differently at different frequencies. To this end, we define our learning targets over several octaves. Specifically, we calculate T_{60} at 7 sub-bands centered at $\{125, 250, 500, 1000, 2000, 4000, 8000\}$Hz. We found prediction of T_{60} at the 62.5Hz band to be unreliable due to low signal-to-noise ratio (SNR). During material optimization, we set the 62.5Hz T_{60} value to the 125Hz value. Our frequency equalization estimation is done at 6 octave bands centered at $\{62.5, 125, 250, 500, 2000, 4000\}$Hz. As we describe in §3.3.2, we compute equalization relative to the 1kHz band, so we do not estimate it. When applying our equalization filter, we set bands greater than or equal to 8kHz to -50dB. Given our target sampling rate of 16kHz and the limited content of speech in higher octaves, this did not affect our estimation.

[4] https://www.magicplan.app/

Fig. 5: We use an off-the-shelf app called MagicPlan to generate geometry proxy. Input: a real-world room (left); Output: the captured 3D model of the room (right) without high-level details, which is used by the runtime geometric acoustic simulator.

3.3.2 Data Augmentation

We use the following datasets as the basis for our training and augmentation.

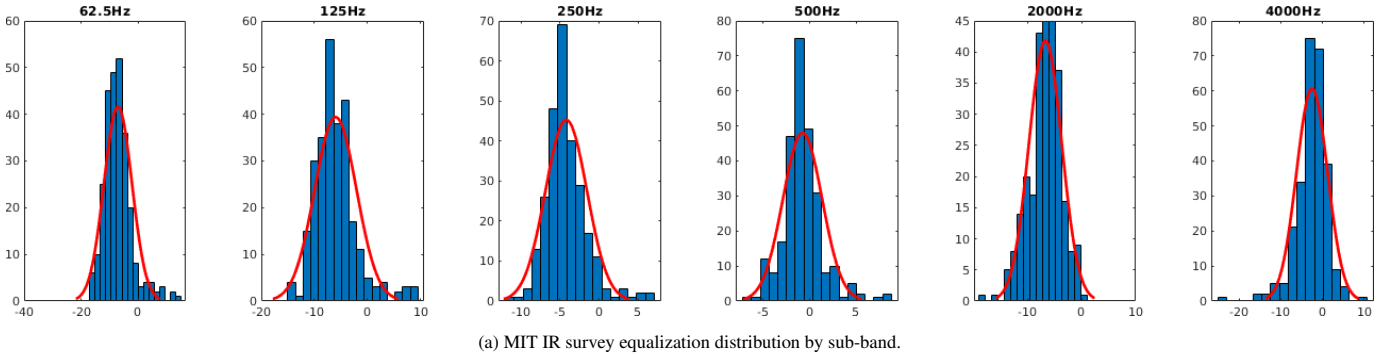
- ACE Challenge: 70 IRs and noise audio [15];
- MIT IR Survey: 271 IRs [68];
- DAPS dataset: 4.5 hours of 20 speakers' speech (10 males and 10 females) [40].

First, we use the method in [5] to expand the T_{60} and direct-to-reverberant ratio (DRR) range of the 70 ACE IRs, resulting in 7000 synthetic IRs with a balanced T_{60} distribution between 0.1–1.5 seconds. The ground truth T_{60} estimates can be computed directly from IRs can be computed is a variety of ways. We follow the methodology of Karjalainen et al. [27] when computing the T_{60} from real IRs with a measurable noise floor. This method was found to be the most robust estimator when computing the T_{60} from real IRs in recent work [15]. The final composition of our dataset is listed in Table 2.

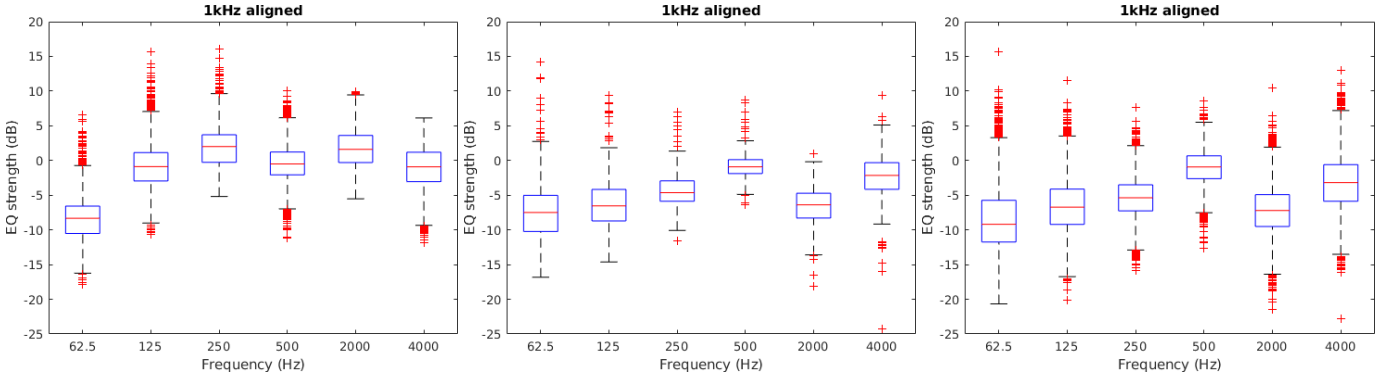
While we know the common range of real-world T_{60} values, there is limited literature giving statistics about room equalization. Therefore, we analyzed the equalization range and distribution of the 271 MIT survey IRs as a guidance for data augmentation. The equalization of frequency bands is computed relative to the 1kHz octave. This is a common practice [70], unless expensive equipment is used to obtain calibrated acoustic pressure readings.

For our equalization augmentation procedure, we first fit a normal distribution (mean and standard deviation) to each sub-band amplitude of the MIT IR dataset as shown in Figure 6. Given this set of parametric model estimates, we iterate through our training and validation IRs. For each IR, we extract its original EQ. We then randomly sample a target EQ according to our fit models (independently per frequency band), calculate the distance between the source and target EQ, and then design an FIR filter to compensate for the difference. For simplicity, we use the window method for FIR filter design [59]. Note, we do not require a perfect filter design method. We simply need a procedure to increase the diversity of our data. Also note, we intentionally sample our augmented IRs to have a larger variance than the recorded IRs to further increase the variety of our training data.

We compute the log Mel-frequency spectrogram for each four second audio clip, which is commonly used for speech-related tasks [9, 16]. We



(a) MIT IR survey equalization distribution by sub-band.



(b) Original synthetic IR equalization.

(c) Target (MIT) IR equalization.

(d) Augmented synthetic IR equalization.

Fig. 6: Equalization augmentation. The 1000Hz sub-band is used as reference and has unit gain. We fit normal distributions (red bell curves shown in (a)) to describe the EQ gains of MIT IRs. We then apply EQs sampled from these distributions to our training set distribution in (b). We observe that the augmented EQ distribution in (d) becomes more similar to the target distribution in (c).

Table 2: Dataset composition. The training set and validation set are based on synthetic IRs and the test set is based on real IRs to guarantee model generalization. Clean speech files are also divided in a way that speakers ("f1" for female speaker 1; "m10" for male speaker 10) in each dataset partition are different, to avoid the model learning the speaker's voice signature. Audio files are generated at a sample rate of 16kHz, which is sufficient to cover the human voice's frequency range.

Partition	Noise	Clean Speech	IR
Training set (size: 56.5k)	ACE ambient	f5∼f10, m5∼m10	Synthetic IR (size: 4.5k)
Validation set (size: 19.5k)	ACE ambient	f3, f4, m3, m4	Synthetic IR (size: 1k)
Test set (size: 18.5k)	ACE ambient	f1, f2, m1, m2	MIT survey IR (size: 271)

use a Hann window of size 256 with 50% overlap during computation of the short-time Fourier transform (STFT) for our 16kHz samples. Then we use 32 Mel-scale bands and area normalization for Mel-frequency warping [62]. The spectrogram power is computed in decibels. This extraction process yields a 32 x 499 (frequency x time domain) matrix feature representation. All feature matrices are normalized by the mean and standard deviation of the training set.

3.3.3 Network Architecture and Training

We propose using a network architecture differing only in the final layer for both T_{60} and room equalization estimation. Six 2D convolutional layers are used sequentially to reduce both the time and frequency resolution of features until they have approximately the same dimension. Each conv layer is immediately followed by a rectified linear unit (ReLU) [41] activation function, 2D max pooling, and batch normalization. The output from conv layers is flattened to a 1D vector and connected to a fully connected layer of 64 units, at a dropout rate of 50% to lower the risk of overfitting. The final output layer has 7 fully connected units to predict a vector of length 7 for T_{60} or 6 fully connected units to predict a vector of length 6 for frequency equalization. This network architecture is inspired by Bryan [5], where it was used to predict full-band T_{60}. We updated the output layer to predict the more challenging sub-band T_{60}, and also discovered that the same architecture predicts equalization well.

For training the network, we use the mean square error (MSE) with the ADAM optimizer [30] in Keras [10]. The maximum number of epochs is 500 with an early stopping mechanism. We choose the model with the lowest validation error for further evaluation on the test set. Our model architecture is shown in Figure 4.

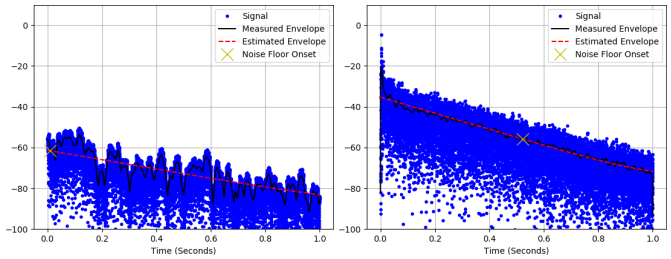
3.4 Acoustic Material Optimization

Our goal is to optimize the material absorption coefficients at the same octave bands as our T_{60} estimator in § 3.3.1 of a set of room materials to match the sub-band T_{60} of the simulated sound with the target predicted in § 3.3.

Ray Energy. We borrow notation from [35]. Briefly, a geometric acoustic simulator generates a set of sound paths, each of which carries an amount of sound energy. Each material m_i in a scene is described by a frequency dependent absorption coefficient, ρ_i. A path leaving the source is reflected by a set of materials before it reaches the listener. The energy fraction that is received by the listener along path j is

$$e_j = \beta_j \prod_{k=1}^{N_j} \rho_{m^k}, \tag{1}$$

where m^k is the material the path intersects on the k^{th} bounce, N_j is the number of surface reflections for path j, and β_j accounts for air absorption (dependent on the total length of the path). Our goal is to optimize the set of absorption coefficients ρ_i to match the energy



(a) 125Hz sub-band. (b) 8000Hz sub-band.

Fig. 7: Evaluating T_{60} from signal envelope on low and high frequency bands of the same IR. Note that the SNR in the low frequency band is lower than the high frequency band. This makes T_{60} evaluation for low frequency bands less reliable, which partly explains the larger test error in low frequency sub-bands.

distribution of the paths e_j to that of the environment's IR. Again similar to [35], we assume the energy decrease of the IR follows an exponential curve, which is a linear decay in dB space. The slope of this decay line in dB space is $m' = -60/T_{60}$.

Objective Function. We propose the following objective function:

$$J(\rho) = (m - m')^2 \tag{2}$$

where m is the best fit line of the ray energies on a decibel scale:

$$m = \frac{n\sum_{i=0}^{n} t_i y_i - \sum_{i=0}^{n} t_i \sum_{i=0}^{n} y_i}{n\sum_{i=0}^{n} t_i^2 - \left(\sum_{i=0}^{n} t_i\right)^2}, \tag{3}$$

with $y_i = 10\log_{10}(e_i)$, which we found to be more robust than previous methods. Specifically, in comparison with Equation (3) in [35], we see that Li et al. tried to match the slope of the energies relative to e_0, forcing e_0 to be at the origin on a dB scale. However, we only care about the energy decrease, and not the absolute scale of the values from the simulator. We found that allowing the absolute scale to move and only optimizing the slope of the best fit line produces a better match to the target T_{60}.

We minimize J using the L-BFGS-B algorithm [75]. The gradient of J is given by

$$\frac{\partial J}{\partial \rho_j} = 2(m - m') \frac{n t_i - \sum_{i=0}^{n} t_i}{n\sum_{i=0}^{n} t_i^2 - \left(\sum_{i=0}^{n} t_i\right)^2} \frac{10}{\ln(10) e_i} \frac{\partial e_i}{\partial \rho_j} \tag{4}$$

4 Analysis and Applications

4.1 Analysis

Speed. We implement our system on an Intel Xeon(R) CPU @3.60GHz and an NVIDIA GTX 1080 Ti GPU. Our neural network inference runs at 222 frames per second (FPS) on 4-second sliding windows of audio due to the compact design (only 18K trainable parameters). Optimization runs twice as fast with our improved objective function. The sound rendering is based on the real-time geometric bi-directional sound path tracing from Cao et al. [7].

Sub-band T_{60} prediction. We first evaluate our T_{60} blind estimation model and achieve a mean absolute error (MAE) of 0.23s on the test set (MIT IRs). While the 271 IRs in the test set have a mean T_{60} of 0.49s with a standard deviation (STD) of 0.85s at the 125Hz sub-band, the highest sub-band 8000Hz only has a mean T_{60} of 0.33s with a STD of 0.24s, which reflects a narrow subset within our T_{60} augmentation range. We also notice that the validation MAE on ACE IRs is 0.12s, which indicates our validation set and the test set still come from different distributions. Another error source is the inaccurate labeling of low-frequency sub-band T_{60} as shown in Figure 7, but we do not filter any outliers in the test set. In addition, our data is intended to cover frequency ranges up to 8000Hz, but human speech has less energy in high-frequency range [67], which results in low signal energy for these sub-bands, making it more difficult for learning.

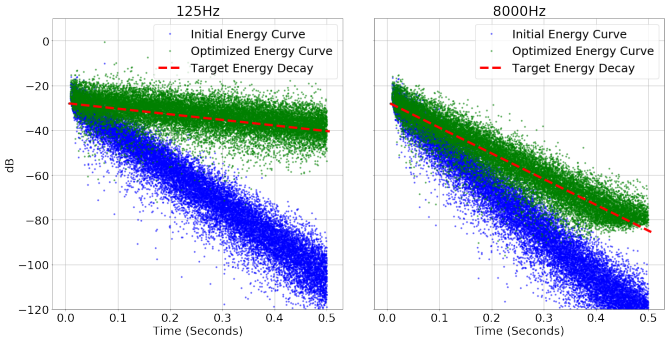


Fig. 8: Simulated energy curves before and after optimization (with target slope shown).

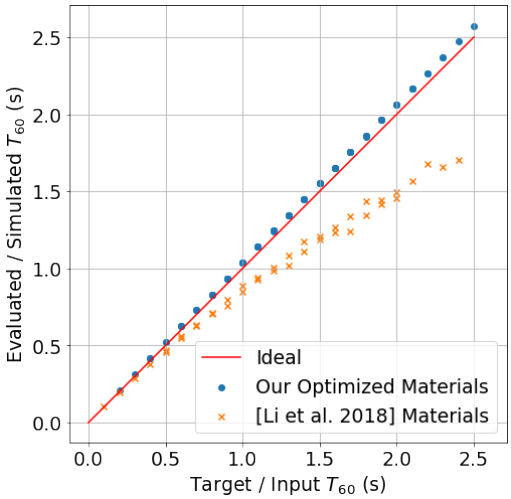


Fig. 9: Stress test of our optimizer. We uniformly sample T_{60} between 0.2s and 2.5s and set it to be the target. The ideal I/O relationship is a straight line passing the origin with slope 1. Our optimization results matches the ideal line much better than prior optimization method.

Material Optimization. When we optimize the room material absorption coefficients according to the predicted T_{60} of a room, our optimizer efficiently modifies the simulated energy curve to a desired energy decay rate (T_{60}) as shown in Figure 8. We also try fixing the room configuration and set the target T_{60} to values uniformly distributed between 0.2s and 2.5s, and evaluate the T_{60} of the simulated IRs. The relationship between the target and output T_{60} is shown in Figure 9, in which our simulation closely matches the target, demonstrating that our optimization is able to match a wide range of T_{60} values.

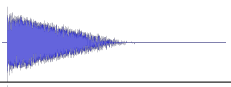
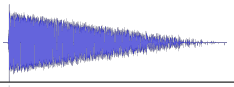
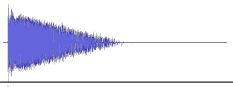
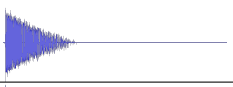
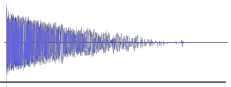
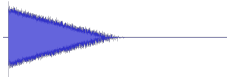
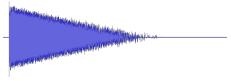
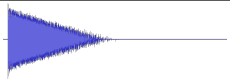
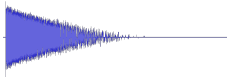
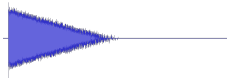
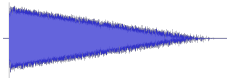
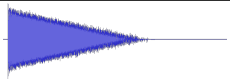
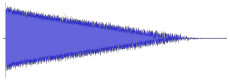
To test the real-world performance of our acoustic matching, we recorded ground truth IRs in 5 benchmark scenes, then use the method in [35], which requires a reference IR, and our method, which does not require an IR, for comparison. Benchmark scenes and results are summarized in Table 3. We apply the EQ filter to the simulated IR as a last step. Overall, we obtain a prediction MAE of 3.42dB on our test set, whereas before augmentation, the MAE was 4.72dB under the same training condition, which confirms the effectiveness of our EQ augmentation. The perceptual impact of the EQ filter step is evaluated in §5.

4.2 Comparisons

We compare our work with two related projects, Schissler et al. [52] and Kim et al. [29], where the high-level goal is similar to ours but the specific approach is different.

Material optimization is a key step in our method and Schissler et al. [52]. One major difference is that we additionally compensate for wave effects explicitly with an equalization filter. Figure 10 shows

Table 3: Benchmark results for acoustic matching. These real-world rooms are of different sizes and shapes, and contain a wide variety of acoustic materials such as brick, carpet, glass, metal, wood, plastic, etc., which make the problem acoustically challenging. We compare our method with [35]. Our method does not require a reference IR and still obtains similar T_{60} and EQ errors in most scenes compared with their method. We also achieve faster optimization speed. Note that the input audio to our method is already noisy and reverberant, whereas [35] requires clean IR recording. All IR plots in the table have the same time and amplitude scale.

Benchmark Scene	 Davis	 301	 501	 620	 750
Size (m^3)	1100 (irregular)	1428 (12×17×7)	990 (11×15×6)	72 (4×6×3)	352 (11×8×4)
# Main planes	6	6	6	11	6
Groundtruth IR (dB scale)					
Li et al. [35] IR (dB scale)					
Opt. time (s)	29	43	25	71	46
T_{60} error (s)	0.11	0.23	0.08	0.02	0.10
EQ error (dB)	1.50	2.97	8.59	3.61	7.55
Ours IR (dB scale)					
Opt. time (s)	13	13	14	31	20
T_{60} error (s)	0.14	0.12	0.10	0.04	0.24
EQ error (dB)	2.26	3.86	3.97	3.46	4.62

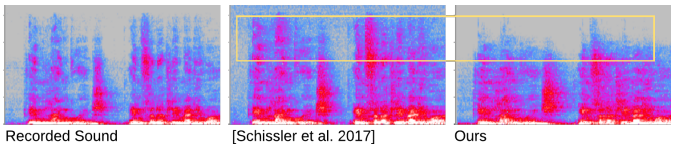


Fig. 10: We show the effect of our equalization filtering on audio spectrograms, compared with Schissler et al. [52]. In the highlighted region, we are able to better reproduce the fast decay in the high-frequency range, closely matching the recorded sound.

the difference in spectrograms, where the high frequency equalization was not properly accounted for. Our method better replicates the rapid decay in the high frequency range. For audio comparison, please refer to our supplemental video.

We also want to highlight the importance of optimizing T_{60}. In [29], a CNN is used for object-based material classification. Default materials are assigned to a limited set of objects. Without optimizing specifically for the audio objective, the resulting sound might not blend in seamlessly with the existing audio. In Figure 11, we show that our method produces audio that matches the decay tail better, whereas [29] produces a longer reverb tail than the recorded ground truth.

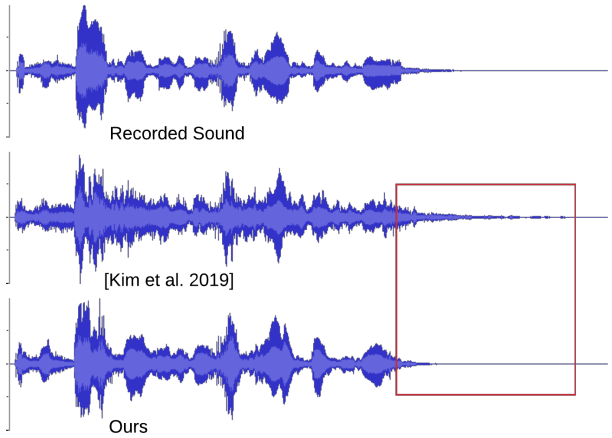


Fig. 11: We demonstrate the importance on T_{60} optimization on the audio amplitude waveform. Our method optimizes the material parameters based on input audio and matches the tail shape and decay amplitude with the recorded sound, whereas the visual-based object materials from Kim et al. [29] failed to compensate for the audio effects.

4.3 Applications

Acoustic Matching in Videos Given a recorded video in an acoustic environment, our method can analyze the room acoustic properties from noisy, reverberant recorded audio in the video. The room geometry can be estimated from video [3], if the user has no access to the room for measurement. During post-processing, we can simulate sound that is similar to the recorded sound in the room. Moreover, virtual characters or speakers, such as the ones shown in Figure 1, can be added to the video, generating sound that is consistent with the real-world environment.

Real-time Immersive Augmented Reality Audio Our method works in a real-time manner and can be integrated into modern AR systems. AR devices are capable of capturing real-world geometry, and can stream audio input to our pipeline. At interactive rates, we can optimize and update the material properties, and update the room EQ filter as well. Our method is not hardware-dependent and can be used with any AR device (which provides geometry and audio) to enable a more immersive listening experience.

Real-world Computer-Aided Acoustic Design Computer-aided design (CAD) software has been used for designing architecture acous-

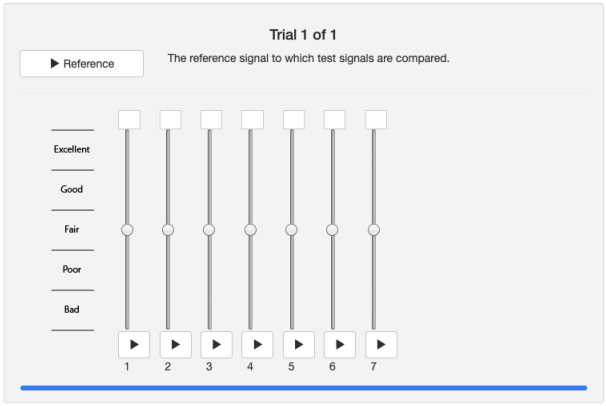


Fig. 12: A screenshot of MUSHRA-like web interface used in our user study. The design is from Cartwright et al. [8].

tics, usually before construction is done, in a predictive manner [31,42]. But when given an existing real-world environment, it becomes challenging for traditional CAD software to adapt to current settings because acoustic measurement can be tedious and error-prone. By using our method, room materials and EQ properties can be estimated from simple input, and can be further fed to other acoustic design applications in order to improve the room acoustics such as material replacement, source and listener placement [39], and soundproofing setup.

5 Perceptual Evaluation

We perceptually evaluated our approach using a critical listening test. For this test, we studied the perceptual similarity of a reference speech recording with speech recordings convolved with simulated impulse responses. We used the same speech content for the reference and all stimuli under testing and evaluated how well we can reconstruct the same identical speech content in a given acoustic scene. This is useful for understanding the absolute performance of our approach compared to the ground truth results.

5.1 Design and Procedure

For our test, we adopted the multiple stimulus with hidden reference and anchor (MUSHRA) methodology from the ITU-R BS.1534-3 recommendation [57]. MUSHRA provides a protocol for the subjective assessment of intermediate quality level of audio systems [57] and has been adopted for a wide variety of audio processing tasks such as audio coding, source separation, and speech synthesis evaluation [8,55].

In a single MUSHRA trial, participants are presented with a high-quality reference signal and asked to compare the quality (or similarity) of three to twelve stimuli on a 0-100 point scale using a set of vertical sliders as shown in Figure 12. The stimuli must contain a hidden reference (identical to the explicit reference), two anchor conditions – low-quality and high-quality, and any additional conditions under study (maximum of nine). The hidden reference and anchors are used to help the participants calibrate their ratings relative to one another, as well as to filter out inaccurate assessors in a post-screening process. MUSHRA tests serve a similar purpose to mean opinion (MOS) score tests [58], but requires fewer participants to obtain results that are statistically significant.

We performed our studies using Amazon Mechanical Turk (AMT), resulting in a MUSHRA-like protocol [8]. In recent years, web-based MUSHRA-like tests have become a standard methodology and have been shown to perform equivalently to full, in-person tests [8,55].

5.2 Participants

We recruited 269 participants on AMT to rate one or more of our five acoustic scenes under testing following the approach proposed by Cartwright et al. [8]. To increase the quality of the evaluation, we pre-screened the participants for our tests. To do this, we first required that all participants have a minimum number of 1000 approved Human Intelligence Task (HITs) assignments and have had at least 97 percent of all assignments approved. Second, all participants must pass a hearing screening test to verify they are listening over devices with an adequate frequency response. This was performed by asking participants to listen to two separate eight second recordings consisting of a 55Hz tone, a 10kHz tone and zero to six tones of random frequency. If any user failed to count the number of tones correctly after two or more attempts, they were not allowed to proceed. Out of the 269 participants who attempted our test, 261 participants passed.

5.3 Training

After having passed our hearing screening test, each user was presented with a one page training test. For this, the participant was provided two sets of recordings. The first set of training recordings consisted of three recordings: a reference, a low-quality anchor, and a high-quality anchor. The second set of training recordings consisted of the full set of recordings used for the given MUSHRA trail, albeit without the vertical sliders present. To proceed to the actual test, participants were required to listen to each recording in full. In total, we estimated the training time to be approximately two minutes.

5.4 Stimuli

For our test conditions, we simulated five different acoustic scenes. For each scene, a separate MUSHRA trial was created. In AMT language, each scene was presented as a separate HIT per user. For each MUSHRA trial or HIT, we tested the following stimuli: hidden reference, low-quality anchor, mid-quality anchor, baseline T_{60}, Baseline T_{60}+EQ, proposed T_{60}, and proposed T_{60}+EQ.

As noted by the ITU-R BS.1534-3 specification [57], both the reference and anchors have a significant effect on the test results, must resemble the artifacts from the systems, and must be designed carefully. For our work, we set the hidden reference as an identical copy of the explicit reference (required), which consisted of speech convolved with the ground truth IR for each acoustic scene. Then, we set the low-quality anchor to be completely anechoic, non-reverberated speech. We set the mid-quality anchor to be speech convolved with an impulse response with a 0.5 second T_{60} (typical conference room) across frequencies, and uniform equalization.

For our baseline comparison, we included two baseline approaches following previous work [35]. More specifically, our Baseline T_{60} leverages the geometric acoustics method proposed by Cao et al. [7] as well as the materials analysis calibration method of Li et al. [35]. Our Baseline T_{60}+EQ extends this and includes the additional frequency equalization analysis [35]. These two baselines directly correspond to the proposed materials optimization (Proposed T_{60}) and equalization prediction subsystems (Proposed T_{60}+EQ) in our work. The key difference is that we estimate the parameters necessary for both steps *blindly from speech*.

5.5 User Study Results

When we analyzed the results of our listening test, we post-filtered the results following the ITU-R BS.1534-3 specification [57]. More specifically, we excluded assessors if they

- rated the hidden reference condition for $> 15\%$ of the test items lower than a score of 90
- or, rated the mid-range (or low-range) anchor for more than 15% of the test items higher than a score of 90.

Using this post-filtering, we reduce our collected data down to 70 unique participants and 108 unique test trials, spread across our five acoustic scene conditions. Among these participants, 24 are females and 46 are males, with an average age of 36.0 and a standard deviation of 10.2 years.

We show the box plots of our results in Figure 13. The median ratings for each stimulus include: Baseline T_{60} (62.0), Baseline T_{60}+EQ (85.0), Low-Anchor (40.5), Mid-Anchor (59.0), Proposed T_{60} (61.5), Proposed T_{60}+EQ (71.0), and Hidden Reference (99.5). As seen, the Low-Anchor and Hidden Reference outline the range of user scores for our test. In

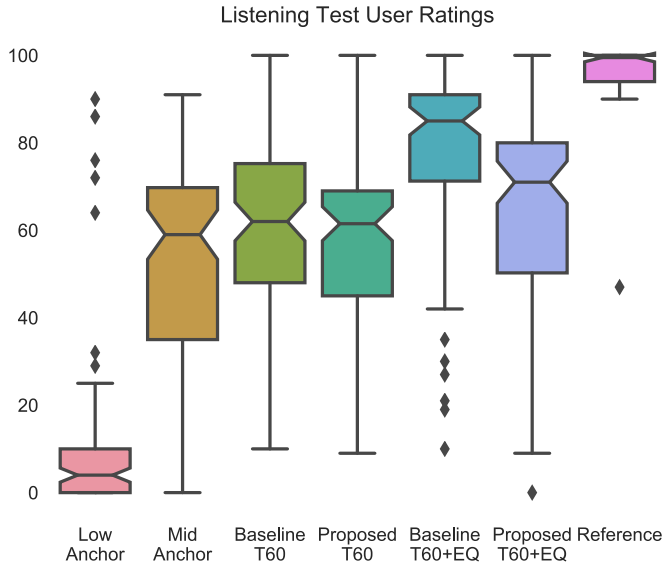


Fig. 13: Box plot results for our listening test. Participants were asked to rate how similar each recording was to the explicit reference. All recordings have the same content, but different acoustic conditions. Note our proposed T_{60} and T_{60}+EQ are both better than the Mid-Anchor by a statistically significant amount ($\approx$10 rating points on a 100 point scale).

terms of baseline approaches, the Proposed T_{60}+EQ method achieves the highest overall listening test performance. We then see that our proposed T_{60} method and T_{60}+EQ method outperform the mid-anchor. Our proposed T_{60} method is comparable to the baseline T_{60} method, and our proposed T_{60}+EQ method outperforms our proposed T_{60}-only method.

To understand the statistical significance, we performed a repeated measures analysis of variance (ANOVA) to compare the effect of our stimuli on user ratings. The Hidden Reference and Low-Anchor are for calibration and filtering purposes and are not included in the following statistical tests, leaving 5 groups for comparison. Bartlett's test did not show a violation of homogeneity of variances ($\chi^2 = 4.68, p = 0.32$). A one-way repeated measures ANOVA shows significant differences ($F(4,372) = 29.24, p < 0.01$) among group mean ratings. To identify the source of differences, we further conduct multiple post-hoc paired t-tests with Bonferroni correction [26]. We are able to observe following results: a) There is no significant difference between Baseline T_{60} and Proposed T_{60} ($t(186) = -1.72, p = 0.35$), suggesting that we cannot reject the null hypothesis of identical average scores between prior work (which uses manually measured IRs) and our work; b) There is a significant difference between Baseline T_{60}+EQ and Proposed T_{60}+EQ ($t(186) = -5.09, p < 0.01$), suggesting our EQ method has a statistically different average (lower); c) There is a significant difference between Proposed T_{60} and Proposed T_{60}+EQ ($t(186) = -2.91, p = 0.02$), suggesting our EQ method significantly improves performance compared to our proposed T_{60}-only subsystem; d) There is a significant difference between Mid-Anchor and Proposed T_{60}+EQ ($t(186) = -3.78, p < 0.01$), suggesting our method is statistically different (higher performing) on average than simply using an average room T_{60} and uniform equalization.

In summary, we see that our proposed T_{60} computation method is comparable to prior work, albeit we perform such estimation directly from a short speech recording rather than relying on intrusive IR measurement schemes. Further, our proposed complete system (Proposed T_{60}+EQ) outperforms both the mid-anchor and proposed T_{60} system alone, demonstrating the value of EQ estimation. Finally, we note our proposed T_{60}+EQ method does not perform as well as prior work, largely due to the EQ estimation subsystem. This result, however, is expected as prior work requires manual IR measurements, which result in perfect EQ estimation. This is in contrast to our work, which directly estimates both T_{60} and EQ parameters from recorded speech, enabling a drastically improved interaction paradigm for matching acoustics in several applications.

6 Conclusion and Future Work

We present a new pipeline to estimate, optimize, and render immersive audio in video and mixed reality applications. We present novel algorithms to estimate two important acoustic environment characteristics – the frequency-dependent reverberation time and equalization filter of a room. Our multi-band octave-based prediction model works in tandem with our equalization augmentation and provides robust input to our improved materials optimization algorithm. Our user study validates the perceptual importance of our method. To the best of our knowledge, our method is the first method to predict IR equalization from raw speech data and validate its accuracy.

Limitations and Future Work. To achieve a perfect acoustic match, one would expect the real-world validation error to be zero. In reality, zero error is only a sufficient but not necessary condition. In our evaluation tests, we observe that small validation errors still allow for plausible acoustic matching. While reducing the prediction error is an important direction, it is also useful to investigate the perceptual error threshold for acoustic matching for different tasks or applications. Moreover, temporal prediction coherence is not in our evaluation process. This implies that given a sliding windows of audio recordings, our model might predict temporally incoherent T_{60} values. One interesting problem is to utilize this coherence to improve the prediction accuracy as a future direction.

Modeling real-world characteristics in simulation is a non-trivial task – as in previous work along this line, our simulator does not fully recreate the real world in terms of precise details. For example, we did not consider the speaker or microphone response curve in our simulation. In addition, sound sources are modeled as omnidirectional sources [7], where real sources exhibit certain directional patterns. It remains an open research challenge to perfectly replicate and simulate our real world in a simulator.

Like all data-driven methods, our learned model performs best on the same kind of data on which it was trained. Augmentation is useful because it generalizes the existing dataset so that the learned model can extrapolate to unseen data. However, defining the range of augmentation is not straightforward. We set the MIT IR dataset as the baseline for our augmentation process. In certain cases, this assumption might not generalize well to estimate the extreme room acoustics. We need to design better and more universal augmentation training algorithms. Our method focused on estimation from speech signals, due to their pervasiveness and importance. It would be useful to explore how well the estimation could work on other audio domains, especially when interested in frequency ranges outside typical human speech. This could further increase the usefulness of our method, e.g., if we could estimate acoustic properties from ambient/HVAC noise instead of requiring a speech signal.

Acknowledgments

The authors would like to thank Chunxiao Cao for sharing the bidirectional sound simulation code, James Traer for sharing the MIT IR dataset, and anonymous reviewers for their constructive feedback. This work was supported in part by ARO grant W911NF-18-1-0313, NSF grant #1910940, Adobe Research, Facebook, and Intel.

References

[1] J. S. Abel, N. J. Bryan, P. P. Huang, M. Kolar, and B. V. Pentcheva. Estimating room impulse responses from recorded balloon pops. In *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[2] M. Barron. *Auditorium acoustics and architectural design*. E & FN Spon, 2010.

[3] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison. Codeslam - learning a compact, optimisable representation for dense

visual slam. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] I. Bork. A comparison of room simulation software-the 2nd round robin on room acoustical computer simulation. *Acta Acustica united with Acustica*, 86(6):943–956, 2000.

[5] N. J. Bryan. Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation. *arXiv preprint arXiv:1909.03642*, 2019.

[6] N. J. Bryan, J. S. Abel, and M. A. Kolar. Impulse response measurements in the presence of clock drift. In *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[7] C. Cao, Z. Ren, C. Schissler, D. Manocha, and K. Zhou. Bidirectional sound transport. *The Journal of the Acoustical Society of America*, 141(5):3454–3454, 2017.

[8] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman. Fast and easy crowdsourced perceptual audio evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 619–623. IEEE, 2016.

[9] L. Chen, Z. Li, R. K Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[10] F. Chollet et al. Keras. `https://keras.io`, 2015.

[11] A. I. Conference. Audio for virtual and augmented reality. *AES Proceedings*, 2018.

[12] T. J. Cox, P. D'Antonio, and M. R. Avis. Room sizing and optimization at low frequencies. *Journal of the Audio Engineering Society*, 52(6):640–651, 2004.

[13] P. Debevec. Image-based lighting. *IEEE Computer Graphics and Applications*, 22(2):26–34, 2002.

[14] M. Doulaty, R. Rose, and O. Siohan. Automatic optimization of data perturbation distributions for multi-style training in speech recognition. In *Spoken Language Technology Workshop*, 2017.

[15] J. Eaton, N. D. Gaubitch, A. H. Moore, P. A. Naylor, J. Eaton, N. D. Gaubitch, A. H. Moore, P. A. Naylor, N. D. Gaubitch, J. Eaton, et al. Estimation of room acoustic parameters: The ace challenge. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(10):1681–1693, 2016.

[16] S. E. Eskimez, P. Soufleris, Z. Duan, and W. Heinzelman. Front-end speech enhancement for commercial speaker verification systems. *Speech Communication*, 99:101–113, 2018.

[17] C. Evers, A. H. Moore, and P. A. Naylor. Acoustic simultaneous localization and mapping (a-slam) of a moving microphone array and its surrounding speakers. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6–10. IEEE, 2016.

[18] A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

[19] S. Foster. Impulse response measurement using golay codes. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11, pp. 929–932. IEEE, 1986.

[20] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017.

[21] S. Gharib, H. Derrar, D. Niizumi, T. Senttula, J. Tommola, T. Heittola, T. Virtanen, and H. Huttunen. Acoustic scene classification: A competition review. In *IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2018.

[22] C. Hak, R. Wenmaekers, and L. Van Luxemburg. Measuring room impulse responses: Impact of the decay range on derived room acoustic parameters. *Acta Acustica united with Acustica*, 98(6):907–915, 2012.

[23] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135. IEEE, 2017.

[24] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

[25] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J.-F. Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7312–7321, 2017.

[26] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

[27] M. Karjalainen, P. Antsalo, A. Makivirta, T. Peltonen, and V. Valimaki. Estimation of modal decay parameters from noisy response measurements. In *Audio Engineering Society Convention 110*. Audio Engineering Society, 2001.

[28] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. Sainath, and M. Bacchiani. Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home. In *Interspeech*, 2017.

[29] H. Kim, L. Remaggi, P. Jackson, and A. Hilton. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. *Proceedings IEEE VR2019*, 2019.

[30] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] M. Kleiner, P. Svensson, and B.-I. Dalenbäck. Auralization: experiments in acoustical cad. In *Audio Engineering Society Convention 89*. Audio Engineering Society, 1990.

[32] H. Kuttruff. *Room Acoustics*. Taylor & Francis Group, London, U. K., 6th ed., 2016.

[33] P. Larsson, D. Vastfjall, and M. Kleiner. Better presence and performance in virtual environments by improved binaural sound rendering. In *Virtual, Synthetic, and Entertainment Audio conference*, Jun 2002.

[34] C. LeGendre, W.-C. Ma, G. Fyffe, J. Flynn, L. Charbonnel, J. Busch, and P. Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5918–5928, 2019.

[35] D. Li, T. R. Langlois, and C. Zheng. Scene-aware audio for 360° videos. *ACM Trans. Graph.*, 37(4), 2018.

[36] R. Mehra, A. Rungta, A. Golas, M. Lin, and D. Manocha. Wave: Interactive wave-based sound propagation for virtual environments. *IEEE transactions on visualization and computer graphics*, 21(4):434–442, 2015.

[37] N. Morales and D. Manocha. Efficient wave-based acoustic material design optimization. *Computer-Aided Design*, 78:83–92, 2016.

[38] N. Morales, R. Mehra, and D. Manocha. A parallel time-domain wave simulator based on rectangular decomposition for distributed memory architectures. *Applied Acoustics*, 97:104–114, 2015.

[39] N. Morales, Z. Tang, and D. Manocha. Receiver placement for speech enhancement using sound propagation optimization. *Applied Acoustics*, 155:53–62, 2019.

[40] G. J. Mysore. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?a dataset, insights, and challenges. *IEEE Signal Processing Letters*, 22(8):1006–1010, 2014.

[41] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[42] S. Pelzer, L. Aspöck, D. Schröder, and M. Vorländer. Integrating real-time room acoustics simulation into a cad modeling software to enhance the architectural design process. *Buildings*, 4(2):113–138, 2014.

[43] N. Raghuvanshi and J. Snyder. Parametric wave field coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 33(4):38, 2014.

[44] N. Raghuvanshi and J. Snyder. Parametric wave field coding for precomputed sound propagation. *ACM Trans. Graph.*, 33(4):38:1–38:11, July 2014.

[45] N. Raghuvanshi, J. Snyder, R. Mehra, M. Lin, and N. Govindaraju. Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes. *ACM Trans. Graph.*, 29(4):68:1–68:11, July 2010.

[46] Z. Ren, H. Yeh, and M. C. Lin. Example-guided physically based modal sound synthesis. *ACM Transactions on Graphics (TOG)*, 32(1):1, 2013.

[47] L. Rizzi, G. Ghelfi, and M. Santini. Small-rooms dedicated to music: From room response analysis to acoustic design. In *Audio Engineering Society Convention 140*. Audio Engineering Society, 2016.

[48] A. Rungta, C. Schissler, N. Rewkowski, R. Mehra, and D. Manocha. Diffraction kernels for interactive sound propagation in dynamic environments. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1613–1622, 2018.

[49] J. Salamon and J. P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[50] L. Savioja and U. P. Svensson. Overview of geometrical room acoustic

modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 2015. doi: 10.1121/1.4926438

[51] R. W. Schafer and A. V. Oppenheim. *Discrete-time signal processing*. Prentice Hall Englewood Cliffs, NJ, 1989.

[52] C. Schissler, C. Loftin, and D. Manocha. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE Transactions on Visualization and Computer Graphics*, 24(3):1246–1259, 2017.

[53] C. Schissler and D. Manocha. Interactive sound propagation and rendering for large multi-source scenes. *ACM Transactions on Graphics (TOG)*, 36(1):2, 2017.

[54] C. Schissler and D. Manocha. Interactive sound rendering on mobile devices using ray-parameterized reverberation filters. *arXiv preprint arXiv:1803.00430*, 2018.

[55] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre. Towards the next generation of web-based experiments: A case study assessing basic audio quality following the itu-r recommendation bs. 1534 (mushra). In *1st Web Audio Conference*, pp. 1–6, 2015.

[56] P. Seetharaman and S. P. Tarzia. The hand clap as an impulse source for measuring room acoustics. In *Audio Engineering Society Convention 132*. Audio Engineering Society, 2012.

[57] B. Series. Recommendation ITU-R BS. 1534-3 method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radio Communication Assembly*, 2014.

[58] P. Series. Methods for objective and subjective assessment of speech and video quality. *International Telecommunication Union Radiocommunication Assembly*, 2016.

[59] J. O. Smith III. *Spectral Audio Signal Processing*. 01 2008.

[60] A. Sterling, N. Rewkowski, R. L. Klatzky, and M. C. Lin. Audio-material reconstruction for virtualized reality using a probabilistic damping model. *IEEE transactions on visualization and computer graphics*, 25(5):1855–1864, 2019.

[61] A. Sterling, J. Wilson, S. Lowe, and M. C. Lin. Isnn: Impact sound neural network for audio-visual object classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 555–572, 2018.

[62] S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[63] Z. Tang, L. Chen, B. Wu, D. Yu, and D. Manocha. Improving reverberant speech training using diffuse acoustic simulation. *arXiv preprint arXiv:1907.03988*, 2019.

[64] Z. Tang, J. Kanu, K. Hogan, and D. Manocha. Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks. In *Interspeech*, 2019.

[65] Z. Tang, H.-Y. Meng, and D. Manocha. Low-frequency compensated synthetic impulse responses for improved far-field speech recognition. *arXiv preprint arXiv:1910.10815*, 2019.

[66] M. Taylor, A. Chandak, Q. Mo, C. Lauterbach, C. Schissler, and D. Manocha. Guided multiview ray tracing for fast auralization. *IEEE Transactions on Visualization and Computer Graphics*, 18:1797–1810, 2012.

[67] I. R. Titze, L. M. Maxfield, and M. C. Walker. A formant range profile for singers. *Journal of Voice*, 31(3):382.e9 – 382.e13, 2017.

[68] J. Traer and J. H. McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016.

[69] N. Tsingos, T. Funkhouser, A. Ngan, and I. Carlbom. Modeling acoustics in virtual environments using the uniform theory of diffraction. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 545–552. ACM, 2001.

[70] V. Välimäki and J. Reiss. All about audio equalization: Solutions and frontiers. *Applied Sciences*, 6(5):129, 2016.

[71] T. Virtanen, M. D. Plumbley, and D. Ellis. *Computational analysis of sound scenes and events*. Springer, 2018.

[72] M. Vorländer. Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm. *The Journal of the Acoustical Society of America*, 86(1):172–178, 1989.

[73] H. Yeh, R. Mehra, Z. Ren, L. Antani, D. Manocha, and M. Lin. Wave-ray coupling for interactive sound propagation in large complex scenes. *ACM Transactions on Graphics (TOG)*, 32(6):165, 2013.

[74] S. Zhi, M. Bloesch, S. Leutenegger, and A. J. Davison. Scenecode: Monocular dense semantic reconstruction using learned encoded scene representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[75] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, Dec. 1997.