GWA: A Large High-Quality Acoustic Dataset for Audio Processing

Zhenyu Tang
Rohith Aralikatti
zhy@umd.edu
rohithca@umd.edu
University of Maryland
Department of Computer Science
College Park, MD, USA

Anton Ratnarajah
jeran@umd.edu
Department of Electrical and
Computer Engineering
University of Maryland
College Park, MD, USA

Dinesh Manocha
dmanocha@umd.edu
Department of Computer Science,
Department of Electrical and
Computer Engineering
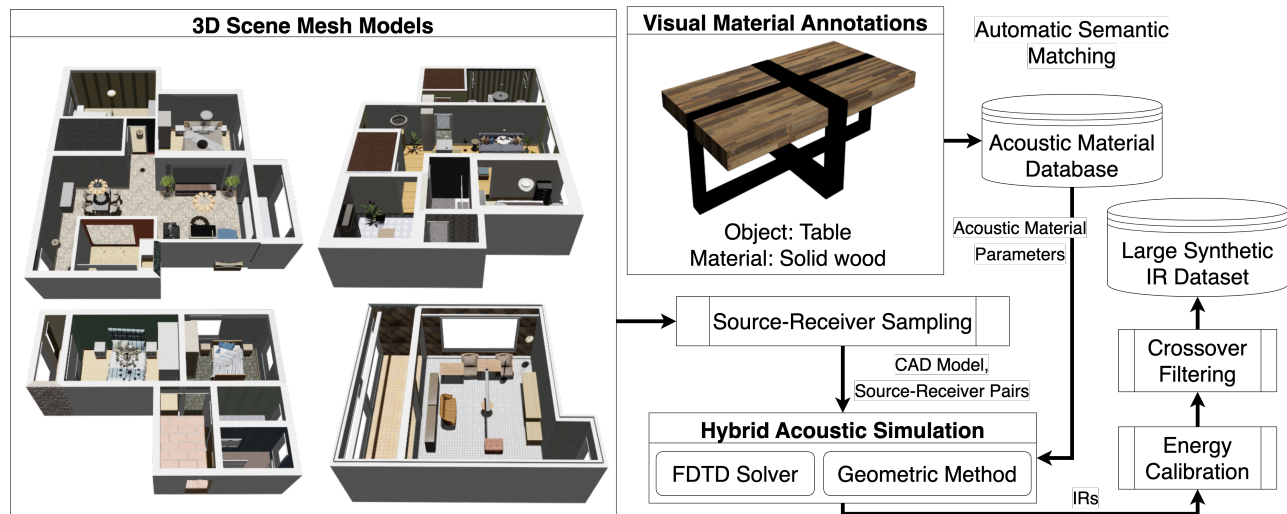University of Maryland
College Park, MD, USA

Figure 1: Our IR data generation pipeline starts from a 3D model of a complex scene and its visual material annotations (unstructured texts). We sample multiple collision-free source and receiver locations in the scene. We use a novel scheme to automatically assign acoustic material parameters by semantic matching from a large acoustic database. Our hybrid acoustic simulator generates accurate impulse responses (IRs), which become part of the large synthetic IR dataset after post-processing.

ABSTRACT

We present the Geometric-Wave Acoustic (GWA) dataset, a large-scale audio dataset of about 2 million synthetic room impulse responses (IRs) and their corresponding detailed geometric and simulation configurations. Our dataset samples acoustic environments from over 6.8K high-quality diverse and professionally designed houses represented as semantically labeled 3D meshes. We also present a novel real-world acoustic materials assignment scheme based on semantic matching that uses a sentence transformer model. We compute high-quality impulse responses corresponding to accurate low-frequency and high-frequency wave effects by automatically calibrating geometric acoustic ray-tracing with a finite-difference time-domain wave solver. We demonstrate the higher accuracy of our IRs by comparing with recorded IRs from complex real-world environments. Moreover, we highlight the benefits of GWA on audio deep learning tasks such as automated speech recognition, speech enhancement, and speech separation. This dataset is the first data with accurate wave acoustic simulations in complex scenes. Codes and data are available at https://gamma.umd.edu/pro/sound/gwa.

CCS CONCEPTS

• **Computing methodologies** → *Modeling methodologies*; *Simulation evaluation*.

KEYWORDS

Acoustic simulation, geometric sound propagation, audio dataset

1 INTRODUCTION

Audio signals corresponding to music, speech, and non-verbal
sounds in the real world encode rich information regarding the sur-
rounding environment. Many digital signal processing algorithms
and audio deep learning techniques have been proposed to extract
information from audio signals. These methods are widely used
for different applications such as music information retrieval, auto-
mated speech recognition, sound separation and localization, sound
synthesis and rendering, etc.

Many audio processing tasks have seen rapid progress in recent
years due to advances in deep learning and the accumulation of
large-scale audio or speech datasets. Not only are these techniques
widely used for speech processing, but also acoustic scene under-
standing and reconstruction, generating plausible sound effects
for interactive applications, audio synthesis for videos, etc. A key
factor in the advancement of these methods is the development
of audio datasets. There are many datasets for speech processing,
including datasets with different settings and languages [Park and
Mulc 2019], emotional speech [Tits et al. 2019], speech source sepa-
ration [Drude et al. 2019], sound source localization [Wu et al. 2018],
noise suppression [Reddy et al. 2020], background noise [Reddy
et al. 2019], music generation [Briot et al. 2017], etc.

In this paper, we present a large, novel dataset corresponding
to synthetic room impulse responses (IRs). In the broader field of
acoustical analysis, the IR is defined as the time domain (i.e. time vs
amplitude) response of a system to an impulsive stimulus [Kuttruff
2016]. It is regarded as the *acoustical signature* of a system and
contains information related to reverberant, signal-to-noise ratio,
arrival time, energy of direct and indirect sound, or other data
related to acoustic scene analysis. These IRs can be convolved with
anechoic sound to generate artificial reverberation, which is widely
used in music, gaming and VR applications. The computation of
accurate IRs and resulting datasets are used for the following areas:

Sound Propagation and Rendering. Sounds in nature are produced
by vibrating objects and then propagate through a medium (e.g., air)
before finally being heard by a listener. Humans can perceive these
sound waves in the frequency range of 20Hz to 20KHz (human
aural range). There is a large body of literature on modeling sound
propagation in indoor scenes using geometric and wave-based
methods [Allen and Raghuvanshi 2015; Funkhouser et al. 1998;
Krokstad et al. 1968; Liu and Manocha 2020; Mehra et al. 2013;
Raghuvanshi et al. 2009; Schissler et al. 2014; Vorländer 1989]. Wave-
based solvers are expensive for high-frequency sound propagation.
As approximations, Wang et al. [2018] use full wave simulation to
model local sound synthesis followed by estimated far-field sound
radiation, while Chaitanya et al. [2020] pre-compute and encode
static sound fields using psycho-acoustic parameters. Geometric
methods, widely used in interactive applications, are accurate for

higher frequencies. We need automatic software systems that can
accurately compute IRs corresponding to human aural range and
handle arbitrary 3D models.

Deep Audio Synthesis for Videos. Video acquisition has become
very common and easy. However, it is hard to add realistic audio
that can be synchronized with animation in a video. Many deep
learning methods have been proposed for such audio synthesis that
utilize acoustic impulse responses for such applications [Li et al.
2018; Owens et al. 2016; Zhou et al. 2018].

Speech Processing using Deep Learning. IRs consist of many clues
related to reproducing or understanding intelligible human speech.
Synthetic datasets of IRs have been used in machine learning meth-
ods for automatic speech recognition [Ko et al. 2017; Malik et al.
2021; Ratnarajah et al. 2021; Tang et al. 2020b], sound source sep-
aration [Aralikatti et al. 2021; Jenrungrot et al. 2020] and sound
source localization [Grumiaux et al. 2021].

Sound Simulation using Machine Learning. Many recent deep
learning methods have been proposed for sound synthesis [Hawley
et al. 2020; Ji et al. 2020; Jin et al. 2020], scattering effect computa-
tion, and sound propagation [Fan et al. 2020; Pulkki and Svensson
2019; Tang et al. 2021]. Deep learning methods have also been used
to compute material properties of a room and acoustic characteris-
tics [Schissler et al. 2017; Tang et al. 2020a].

Other applications that have used acoustic datasets include navi-
gation [Chen et al. 2020], floorplan reconstruction [Purushwalkam
et al. 2021] and depth estimation algorithms [Gao et al. 2020].

There are some known datasets of recorded IRs from real-world
scenes and synthetic IRs (see Table 1). The real-world datasets are
limited in terms of number of IRs or the size and characteristics of
the captured scenes. All prior synthetic IR datasets are generated
using geometric simulators and do not accurately capture low-
frequency wave effects. This limits their applications.

Main Results. We present the first large acoustic dataset (GWA)
of synthetically generated IRs that uses accurate wave acoustic
simulations in complex scenes. Our approach is based on using
a hybrid simulator that combines a wave-solver based on finite
differences time domain (FDTD) method with geometric sound
propagation based on path tracing. The resulting IRs are accurate
over the human aural range. Moreover, we use a large database
of more than 6.8K professionally designed scenes with more than
18K rooms with furniture that provide a diverse set of geometric
models. We present a novel and automatic scheme for semantic
acoustic material assignment based on natural language processing
techniques. We use a database of absorption coefficients of 2,042
unique real-world materials and use a transformer network for
semantic material selection. Currently, GWA consists of more than 2
million IRs. We can easily use our approach to generate more IRs by
either changing the source and receiver positions or using different
set of geometric models or materials. The novel components of our
work include:

- Our dataset has more acoustic environments than real-world
  IR datasets by two orders of magnitude.
- Our dataset has more diverse IRs with higher accuracy, as
  compared to prior synthetic IR datasets.

Table 1: Overview of some existing large IR datasets and their characteristics. In the "Type" column, "Rec." means *recorded* and "Syn." means *synthetic*. The real-world datasets capture the low-frequency (LF) and high-frequency (HF) wave effects in the recorded IRs. Note that all prior synthetic datasets use geometric simulation methods and are accurate for higher frequencies only. In contrast, we use an accurate hybrid geometric-wave simulator on more diverse input data, corresponding to professionally designed 3D interior models with furniture, and generate accurate IRs corresponding to the entire human aural range (LF and HF). We highlight the benefits of our high-quality dataset for different audio and speech applications.

Dataset	Type	#IRs	#Scenes	Scene Descriptions	Scene Types	Acoustic Material	Quality
BIU [Hadad et al. 2014]	Rec.	234	3	Photos	Acoustic lab	Real-world	LF, HF
MeshRIR [Koyama et al. 2021]	Rec.	4.4K	2	Room dimensions	Acoustic lab	Real-world	LF, HF
BUT Reverb [Szöke et al. 2019]	Rec.	1.3K	8	Photos	Various sized rooms	Real-world	LF, HF
S3A [Coleman et al. 2020]	Rec.	1.6K	5	Room dimensions	Various sized rooms	Real-world	LF, HF
dEchorate [Carlo et al. 2021]	Rec.	2K	11	Room dimensions	Acoustic lab	Real-world	LF, HF
Ko et al. [2017]	Syn.	60K	600	Room dimensions	Empty shoebox rooms	Uniform sampling	HF
BIRD [Grondin et al. 2020]	Syn.	100K	100K	Room dimensions	Empty shoebox rooms	Uniform sampling	HF
SoundSpaces [Chen et al. 2020]	Syn.	16M	101	Annotated 3D model	Scanned indoor scenes	Material database	HF
GWA (ours)	Syn.	2M	18.9K	Annotated 3D model	Professionally designed	Material database	LF, HF

- The accuracy improvement of our hybrid method over prior methods is evaluated by comparing our IRs with recorded IRs of multiple real-world scenes.
- We use our dataset to improve the performance of deep-learning speech processing algorithms, including automatic speech recognition, speech enhancement, and source separation, and observe significant performance improvement.

2 DATASET CREATION

A key issue in terms of the design and release of an acoustic dataset is the choice of underlying 3D geometric models. Given the availability of interactive geometric acoustic simulation software packages, it is relatively simple to randomly sample a set of simple virtual shoebox-shaped rooms for source and listener positions and generate unlimited simulated IR data. However, the underlying issue is such IR data will not have the acoustic variety (e.g., room equalization, material diversity, wave effects, reverberation patterns, etc.) frequently observed in real-world datasets. We identify several criteria that are important in terms of creating a useful synthetic acoustic dataset: (1) a wide range of room configurations: the room space should include regular and irregular shapes as well as furniture placed in reasonable ways. Many prior datasets are limited to rectangular, shoebox-shaped or empty rooms (see Table 1); (2) meaningful acoustic materials: object surfaces should use physically plausible acoustic materials with varying absorption and scattering coefficients, rather than randomly assigned frequency-dependent coefficients; (3) an accurate simulation method that accounts for various acoustic effects, including specular and diffuse reflections, occlusion, and low-frequency wave effects like diffraction. It is important to generate IRs corresponding to the human aural range for many speech processing and related applications. In this section, we present our pipeline for developing a dataset that satisfies all these criteria. An overview of our pipeline is illustrated in Figure 1.

2.1 Acoustic Environment Acquisition

Acoustic simulation for 3D models requires that environment boundaries and object shapes be well defined. This requirement can often

be fulfilled by 3D meshes. Simple image-method simulations may only require a few room dimensions (i.e., length, width, and height) and have been used for speech applications, but these methods cannot handle complex 3D indoor scenes. Many techniques have been proposed in computer vision to reconstruct large-scale 3D environments using RGB-D input [Choi et al. 2015]. Moreover, they can be combined with 3D semantic segmentation [Dai et al. 2018] to recover category labels of objects in the scene. This facilitates the collection of indoor scene datasets. However, real-world 3D scans tend to suffer from measurement noise, resulting in incomplete/discontinuous surfaces in the reconstructed model that can be problematic for acoustic simulation algorithms. One alternative is to use professionally designed scenes of indoor scenes in the form of CAD models. These models are desirable for acoustic simulation because they have well-defined geometries and the most accurate semantic labels. Therefore, we use CAD models from the 3D-FRONT dataset [Fu et al. 2021], which contains 18,968 diversely furnished rooms in 6,813 irregularly shaped houses/scenes. These different types of rooms (e.g., bedrooms, living rooms, dining rooms, and study rooms) are diversely furnished with varying numbers of furniture objects in meaningful locations. This differs from prior methods that use empty shoebox-shaped rooms [Grondin et al. 2020; Ko et al. 2017], because room shapes and the existence of furniture will significantly modify the acoustic signature of the room, including shifting the room modes in the low frequency. 3D-FRONT dataset is designed to have realistic scene layouts, and has received higher human ratings in subjective studies [Fu et al. 2021]. Generating audio data from these models allows us to better approximate real-world acoustics.

2.2 Semantic Acoustic Material Assignment

Because the 3D-FRONT dataset also provides object semantics (i.e., object material labels), it is possible to assign more meaningful acoustic materials to individual surfaces or objects in the scene. For example, an object with "window" description is likely to be matched with several types of window glass material from the acoustic material database. *SoundSpaces* dataset [Chen et al. 2020]

also utilizes scene labels by using empirical manual material assignment (e.g., acoustic materials of carpet, gypsum board, and acoustic tile are assumed for floor, wall, and ceiling classes), creating a one-to-one visual-acoustic material mapping for the entire dataset. This approach works for a small set of known material types. Instead, we present a general and fully automatic method that works for unknown materials with unstructured text descriptions.

To start with, we retrieve measured frequency-dependent acoustic absorption coefficients for 2,042 unique materials from a room acoustic database [Kling 2018]. The descriptions of these materials do not directly match the semantic labels of objects in the 3D-FRONT dataset. Therefore, we present a method to calculate the semantic similarity between each label and material description using natural language processing (NLP) techniques. In NLP research, sentences can be encoded into fixed length numeric vectors known as sentence embedding [Mishra and Viradiya 2019]. One goal of sentence embedding is to find semantic similarities to identify text with similar meanings. Transformer networks have been very successful in generating good sentence embeddings [Liu et al. 2020] such that sentences with similar meanings will be relatively close in the embedding vector space. We leverage a state-of-the-art sentence transformer model [1] [Reimers and Gurevych 2019] that calculates an embedding of dimension 512 for each sentence. Our assignment process is described in Algorithm 1.

ALGORITHM 1: Semantic material assignment

Input: Object name string s_0
Data: Embedding model $f : s \longmapsto \mathbb{R}^{512}$;
Acoustic material names $\{m_1, m_2, ..., m_N\}$
Output: Acoustic material assignment
$e_0 \leftarrow f(s_0)$;   // embedding for object name
for $i \leftarrow 1$ **to** N **do**
$\quad e_i \leftarrow f(m_i)$; // embedding for acoustic material
$\quad w_i \leftarrow max\{0, cosineSimilarity(e_0, e_i)\}$;
end
Assign m_i with probability $P(X = m_i) = w_i / \Sigma_{i=1}^{N} w_i$;

The key idea of our assignment is to use the cosine similarity (truncated to be non-negative) between the embedding vectors of the input and target material names as sampling weights to assign materials. Note that we do not directly pick the material with the highest score because for the same type of material, there are still different versions with different absorption coefficients (e.g., in terms of thickness, brand, painting). These slightly different descriptions of the same material are likely to have similar semantic distance to the 3D-FRONT material label being considered. Therefore, we use a probabilistic assignment to increase the diversity of our materials.

2.3 Geometric-Wave Hybrid Simulation

It is well known that geometric acoustic (GA) methods do not model low-frequency acoustic effects well due to the linear ray assumption [Funkhouser et al. 1998; Schissler et al. 2014]. Therefore, we use a hybrid propagation algorithm that combines wave-based methods

with GA. These wave-based methods can accurately model low-frequency wave effects, but their running time increases as the third or fourth power of the highest simulation frequency [Raghuvanshi et al. 2009]. Given the high time complexity of wave-based methods, we also want to use methods that are: (1) highly parallelizable so that dataset creation takes acceptable time on high-performance computing clusters; (2) compatible with arbitrary geometric mesh representations and acoustic material inputs; and (3) open-source so that the simulation pipeline can be reused by the research community. In this paper, we develop our hybrid simulation pipeline from a CPU-based GA software *pygsound* [2] and a GPU-based wave FDTD software *PFFDTD* [Hamilton 2021].

2.3.1 Inputs. The scene CAD models from the 3D-FRONT dataset, each corresponding to several rooms with open doors, are represented in a triangle mesh format. Most GA methods have native support for 3D mesh input. The meshes are converted to voxels to be used as geometry input to the wave-based solver. We sample source and receiver locations in each scene by grid sampling in all three dimensions with $1m$ spacing. This sampling yields from tens to thousands source and receiver pairs in each scene depending on its size. We perform collision checking to ensure all sampled locations have at least $0.2m$ clearance to any object in the scene.

We assign acoustic absorption coefficients according to the scheme presented in § 2.2. These coefficients can be directly used by the GA method and integrated with the passive boundary impedance model used by the wave FDTD method [Bilbao et al. 2015]. The GA method also requires scattering coefficients, which account for the energy ratio between specular and diffuse reflections. Such data is less conventionally measured and is not available from the material database in § 2.2. It is known that scattering coefficients tend to be negligible (e.g., ≤ 0.05) for low-frequency bands [Cox et al. 2006] handled by the wave method. Therefore, we sample scattering coefficients by fitting a normal distribution to 37 sets of frequency-dependent scattering coefficients obtained from the benchmark data in § 3.1, which are only used by the GA method.

2.3.2 Setup. For the GA method, we set 20,000 rays and 200 maximum depth for specular and diffuse reflections. The GA simulation is intended for human aural range, while most absorption coefficient data is only valid for octave bands from 63Hz to 8,000Hz. The ray-tracing stops when the maximum depth is reached or the energy is below the hearing threshold.

For the wave-based FDTD method, we set the maximum simulation frequency to 1,400Hz. The grid spacing is set according to 10.5 points per wavelength. Our simulation duration is $1.5s$, which is sufficient to capture most indoor reverberation times and can be extended to generate additional data.

2.3.3 Automatic Calibration. Before combining simulated IRs from two methods, one important step is to properly calibrate their relative energies. Southern et al. [2011] describe two objective calibration methods: (1) pre-defining a crossover frequency range near the highest frequency of the wave method and aligning the sound level of the two methods in that range; (2) calibrating the peak level from time-windowed, bandwidth-matched regions in the wave and the GA methods. Both calibration methods are used

---

case-by-case for each pair of IRs. However, the first method is not physically correct, and the second method can be vulnerable when the direct sound is not known, as with occluded direct rays in the GA method. Southern et al. [2013] improved the second method by calculating calibration parameters once in free-field condition using a band-limited source signal.

We use a similar calibration procedure. The calibration source and receivers have a fixed distance $r = 1m$ in a large volume with absorbing boundary conditions, and the 90 calibration receivers span a 90° arc to account for the influence of propagation direction along FDTD grids. The source impulse signal is low-pass filtered at a cut-off frequency of 255Hz. Since the source signal is a unit impulse, the filtered signal is essentially the coefficients of the low-pass filter. The simulated band-limited IRs are truncated at twice the theoretical direct response time to further prevent any unwanted reflected wave. The calibration parameter for wave-based FDTD is computed as $\eta_w = \sqrt{E_s/E_r}$, where E_s is the total energy of the band-limited point source, and E_r is the total energy at the receiver point. For multiple receiver points, η_w takes the average value. During wave-based FDTD calibration, each received signal is multiplied by η_w, and we can calculate the difference between the calibrated signal and the band-limited source signal. As a result, we obtain a very low mean error of $0.50dB$ and a max error of $0.85dB$ among all calibration receivers.

For the GA method, we follow the same procedure though the process is simpler since the direct sound energy is explicit in most GA algorithms (i.e., $\frac{1}{r}$ scaled by some constant). Another calibration parameter η_g is similarly obtained for the GA method. This calibration process ensures that the full-band transmitted energy from both methods will be $E = 1$ at a distance of $1m$ from a sound source, although the absolute energy does not matter and the two parameters can be combined into one (i.e., only use $\eta'_w = \eta_w/\eta_g$ for wave calibration). Figure 2 shows an example of simulation results with and without calibration. Without properly calibrating the energies, there will be abrupt sound level changes in the frequency domain, which can create unnatural sound.
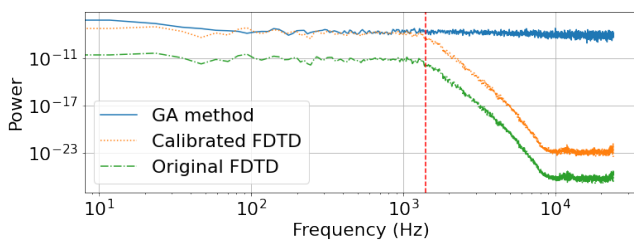

Figure 2: Power spectrum comparison between the original wave FDTD simulated IR and the calibrated IR. The vertical dashed line indicates the highest valid frequency of the FDTD method. Our automatic calibration method ensures that the GA and wave-based methods have consistent energy levels so that they can generate high quality IRs and plausible/smooth sound effects.

2.3.4 Hybrid Combination. Ideally we would want to use the wave-based method for the highest possible simulation frequency. Besides

(a) Occurrence of top visual material names.

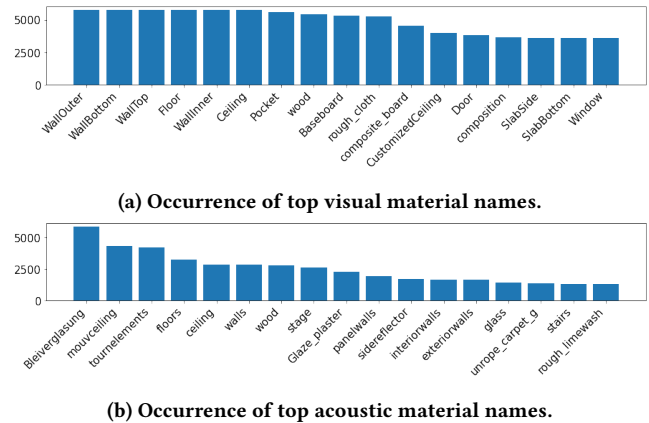**(b) Occurrence of top acoustic material names.**

Figure 3: We highlight the most frequently used materials in our approach for generating the IR dataset. The acoustic database also contains non-English words, which are handled by a pre-trained multi-lingual language model.

the running time, one issue with FDTD scheme is the rising dispersion error with the frequency [Lehtinen 2003]. As a remedy, the FDTD results are first high-pass filtered at a very low frequency (e.g., 10Hz) to remove some DC offset and then low-pass filtered at the crossover frequency to be combined with GA results. We use a Linkwitz-Riley crossover filter [Linkwitz 1976] to avoid ringing artifacts near the crossover frequency, harnessing its use of cascading Butterworth filters. The crossover frequency in this work is chosen to be $1,400$Hz to fully utilize the accuracy of wave simulation results. Higher simulation crossover frequencies could be used at the cost of increased FDTD simulation time.

2.4 Analysis and Statistics

Runtime. The runtime of our hybrid simulator depends on specific computational hardware. We utilize a high-performance computing cluster with 20 Intel Ivy Bridge E5-2680v2 CPUs and 2 Nvidia Tesla K20m GPUs on each node. On a single node, our simulator requires about $4,000$ computing hours for the wave-based FDTD method and about $2,000$ computing hours for the GA method to generate all data.

Distributions. More than 5,000 scene/house models are used. On average, each scene uses 22.5 different acoustic materials. We assign $1,955$ unique acoustic materials (out of $2,042$) from the material database, and the most frequently used materials are several versions of brick, concrete, glass, wood, and plaster. The occurrence of most frequently used materials are visualized in Figure 3.

The distribution of distances between all source and receiver pairs are visualized in Figure 4. We also show the relationship between the volume of each 3D house model and the reverberation time for that model in Figure 5 to highlight the wide distribution of our dataset. Overall, we have a balanced distribution of the reverberation times in the normal range.
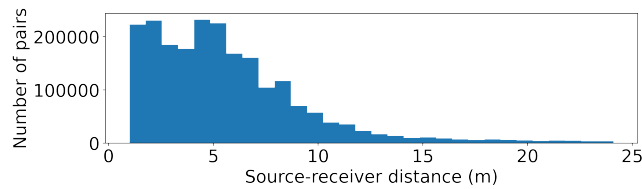
Figure 4: Distance distribution between source and receiver pairs in our scene database. 3D grid sampling is used in each scene followed by collision detection between sampled locations and objects in the scene. The IRs vary based on relative positions of the source and the receiver.
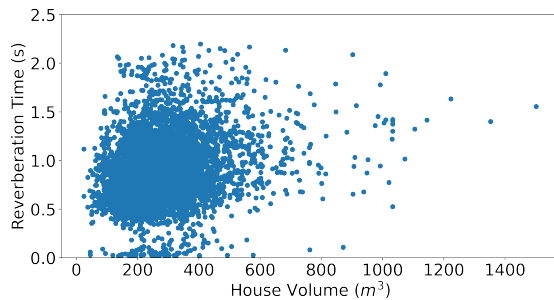

Figure 5: Statistics of house/scene volumes and reverberation times. We see a large variation in reverberation times, which is important for speech processing and other applications.

3 ACOUSTIC EVALUATION

In this section, we evaluate the accuracy of our IR generation hybrid algorithm. We use a set of real-world acoustic scenes that have measured IR data to evaluate the effectiveness and accuracy of our hybrid simulation method.

3.1 Benchmarks

Several real-world benchmarks have been proposed to investigate the accuracy of acoustic simulation techniques. A series of three round-robin studies [Bork 2000, 2005a,b; Vorlander 1995] have been conducted on several acoustic simulation software systems by providing the same input and then comparing the different simulation results with the measured data. In general, these studies provide the room and material descriptions as well as microphone and loudspeaker specifications including locations and directivity. However, the level of detailed characteristics, in terms of complete 3D models and consistent measured acoustic material properties tend to vary. Previous round-robin studies have identified many issues (e.g., uncertainty in boundary condition definitions) in terms of simulation input definitions for many simulation packages, which can result in poor agreement between simulation results and real-world measurements. A more recent benchmark, the BRAS benchmark [Aspöck et al. 2020], contains the most complete scene description and has a wide range of recording scenarios. We use the BRAS benchmark to evaluate our simulation method. Three reference scenes (RS5-7)

are designed as diffraction benchmarks and we use them to evaluate the performance of our hybrid simulator, especially at lower frequencies.

The 3D models of the reference scenes along with frequency-dependent acoustic absorption and scattering coefficients are directly used for our hybrid simulator. We use these three scenes because they are considered difficult for the geometric method alone [Brinkmann et al. 2019].

3.2 Results

We use the room geometry, source-listener locations, and material definitions as an input to our simulation pipeline. Note that the benchmark only provide absorption and scattering coefficients, and no impedance data is directly available for wave solvers. Thus, we only use fitted values rather than exact values. The IRs generated by the GA method and our hybrid method and the measured IRs from the benchmark are compared in the frequency domain in Figure 6. In these scenes, the source and receiver are placed on different sides of the obstacle and the semi-anechoic room only has floor reflections. In the high frequency range, there are fewer variations in the measured response, and both methods capture the general trend of energy decay despite response levels not being perfectly matched. This demonstrates that our hybrid sound simulation pipeline is able to generate more accurate results than the GA method for complex real-world scenes.

4 APPLICATIONS

We use our dataset on three speech processing applications that use deep learning methods. Synthetic IRs have been widely used for training neural networks for automatic speech recognition, speech enhancement, and source separation (as described in § 1). We evaluate the benefits of generating a diverse and high-quality IRs dataset over prior methods used to generate synthetic IRs.

In general, these speech tasks use IR datasets to augment anechoice speech data to create synthetic distant training data, whereas the test data is reverberant data recorded in the real world. When high-quality IR datasets are used, the training set is expected to generalize better on the test data. We simulate distant speech data $x_d[t]$ by convolving anechoic speech $x_c[t]$ with different IRs $r[t]$ and adding environmental noise $n[t]$ from BUT ReverbDB [Szöke et al. 2019] dataset using

$$x_d[t] = x_c[t] \circledast r[t] + n[t+l]. \qquad (1)$$

Then the data is used by different training procedure and neural network architectures on respective benchmarks. In the following tests, we also examine variations of our datasets: **GA** (geometric method only), **FDTD** (only up to 1,400Hz), and **GWA** (hybrid method).

4.1 Automated Speech Recognition

Automatic speech recognition (ASR) aims to convert speech data to text transcriptions. The performance of ASR models is measured by the *word error rate (WER)*, which is the percentage of incorrectly transcribed words in the test data. The AMI speech corpus [Carletta et al. 2005] consisting of 100 hours of meeting recording is used as our benchmark. And we use the *Kaldi* toolkit [Povey et al. 2011] to run experiments on this benchmark. We randomly select 30, 000

(a) RS5: simple diffraction with infinite edge.

(b) RS6: diffraction with infinite body.

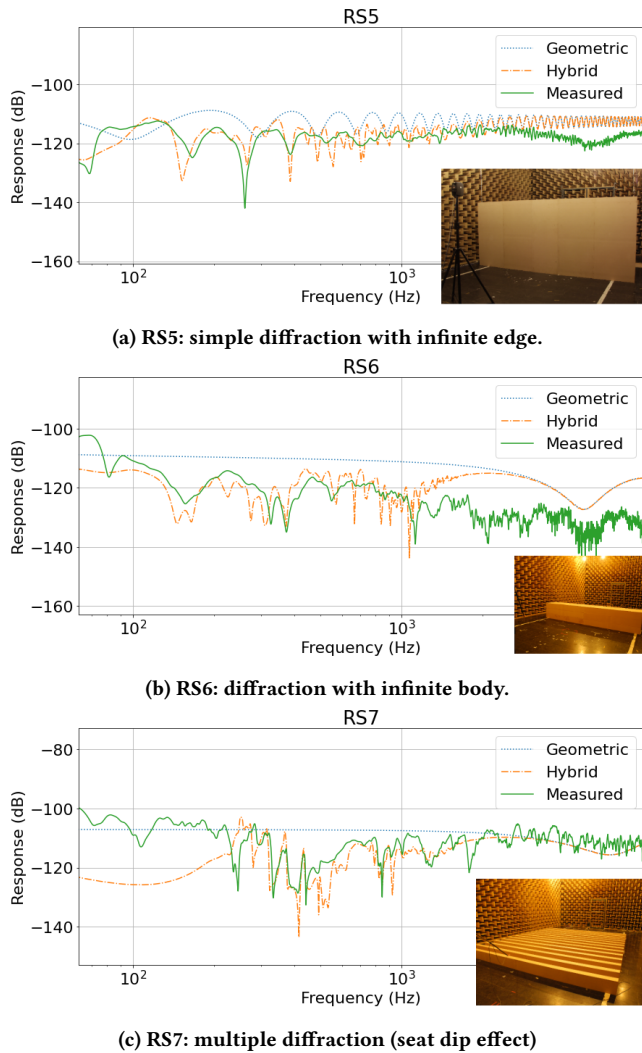(c) RS7: multiple diffraction (seat dip effect)

Figure 6: Frequency responses of geometric and hybrid simulations compared with measured IRs in BRAS benchmarks RS5-7 [Aspöck et al. 2020]. Images of each setup are attached in the corners of the graph. We notice that the IRs generated using our hybrid method closely match with the measure IRs, as compared to those generated using GA methods. This demonstrates the higher quality and accuracy of our IRs as compared to the ones generated by prior GA methods highlighted in Table 1.

IRs out of 2M synthetic IRs in GWA to augment the anechoice training set in AMI, and report the WER on the real-world test set. A lower WER indicates that the synthetic distant speech data used for training is closer to real-world distant speech data. We highlight the improved accuracy obtained using GWA over prior synthetic IR generators in Table 2.

In this benchmark, we include more comparisons with prior datasets from Table 1 and we also incorporate SoundSpaces [Chen et al. 2018] data into our pipeline. We have replaced the original

Table 2: Far-field ASR results obtained for the AMI corpus.

IR used	WER[%]↓
None (anechoic)	64.2
BIRD [Grondin et al. 2020]+Ko et al. [2017]	48.4
BUT Reverb [Szöke et al. 2019]	**46.6**
SoundSpaces (manual) [Chen et al. 2020]	51.2
SoundSpaces-semantic	**50.3**
FDTD	58.1
GA	48.1
GWA subset	48.5
GWA (ours)	**47.7**

manual material assignment in SoundSpaces with our semantic assignment (**SoundSpaces-semantic**), and observe a 0.9% WER improvement. To study the effect of model diversity, we use a subset of house models from 3D-FRONT (**GWA subset**) that has the same number (103) of scenes as SoundSpaces, this results in a 0.8% degradation in WER compared with our full dataset. Overall, we see that our **GWA** dataset outperforms prior synthetic datasets, where **BIRD+Ko et al. [2017]** uses an image-method simulator in shoebox-shaped rooms and **SoundSpaces** uses a ray-tracing based simulator, and has the closest WER to that of the real-world **BUT Reverb** [Szöke et al. 2019] dataset, due to both the inclusion of more diverse 3D scenes and our semantic material assignment.

4.2 Speech Dereverberation

Speech dereverberation aims at converting a reverberant speech signal back to its anechoic version to enhance its intelligibility. For our tests, we use SkipConvNet [Kothapally et al. 2020], a U-Net based speech dereverberation model. The model is trained on the 100-hour subset of Librispeech dataset [Panayotov et al. 2015]. The reverberant input to the model is generated by convolving the clean Librispeech data with our synthetically generated IRs. We

Table 3: Speech enhancement results trained using different synthetic IR datasets.

IR used	SRMR↑
None (anechoic)	4.96
SoundSpaces	7.44
GA	6.01
FDTD	4.78
GWA (ours)	**8.14**

test the performance of the model on real-world recordings from the VOiCES dataset [Richey et al. 2018] which contain reverberant Librispeech data. We report the *speech-to-reverberation modulation energy ratio (SRMR)* [Falk et al. 2010] over the test set. A higher value of SRMR indicates lower reverberation and higher speech quality. As seen from Table 3, our proposed dataset obtains better dereverberation performance as compared to all other datasets.

4.3 Speech Separation

We train a model to separate reverberant mixtures of two speech signals into its constituent reverberant sources. We use the Asteroid [Pariente et al. 2020] implementation of the DPRNN-TasNet model [Luo et al. 2020] for our benchmarks. The 100-hour split of the Libri2Mix [Cosentino et al. 2020] dataset is used for training. We test the model on reverberant mixtures generated from the VOiCES dataset. We report the improvement in *scale-invariant signal-to-distortion ratio (SI-SDRi)* [Le Roux et al. 2019] to measure separation performance. Higher SI-SDRi implies better separation. As seen from Table 4, our GWA outperforms alternative approaches.

Table 4: SI-SDRi values reported for different IR generation methods. We report results separately for the four rooms used to capture the test set (higher is better).

IR used	SI-SDRi↑			
	Room 1	Room 2	Room 3	Room 4
SoundSpaces	1.85	1.74	1.02	1.80
GA	2.25	2.55	1.44	2.55
FDTD	2.36	2.43	1.33	2.46
GWA (ours)	**2.94**	**2.76**	**1.86**	**2.91**

5 CONCLUSION AND FUTURE WORK

We introduced a large new audio dataset of synthetic room impulse responses and the simulation pipeline, which can take different scene configurations and generate higher quality IRs. We demonstrated the improved accuracy of our hybrid geometric-wave simulator on three difficult scenes from the BRAS benchmark. As compared to prior datasets, GWA has more scene diversity than recorded datasets, and has more physically accurate IRs than other synthetic datasets. We also use our dataset with audio deep learning to improve the performance of speech processing applications.

Our dataset only consists of synthetic scenes, and may not be as accurate as real-world captured IRs. In many applications, it is also important to model ambient noise. In the future, we will continue growing the dataset by including more 3D scenes to further expand the acoustic diversity of the dataset. We plan to evaluate the performance of other audio deep learning applications.

ACKNOWLEDGMENTS

REFERENCES

Andrew Allen and Nikunj Raghuvanshi. 2015. Aerophones in flatland: Interactive wave simulation of wind instruments. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 1–11.

Rohith Aralikatti, Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. 2021. Improving Reverberant Speech Separation with Synthetic Room Impulse Responses. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 900–906.

Lukas Aspöck, Michael Vorländer, Fabian Brinkmann, David Ackermann, and Stefan Weinzierl. 2020. Benchmark for Room Acoustical Simulation (BRAS). *DOI* 10 (2020), 14279.

Stefan Bilbao, Brian Hamilton, Jonathan Botts, and Lauri Savioja. 2015. Finite volume time domain room acoustics simulation under general impedance boundary conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 1 (2015), 161–173.

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

Ingolf Bork. 2005a. Report on the 3rd round robin on room acoustical computer simulation–Part I: Measurements. *Acta Acustica united with Acustica* 91, 4 (2005), 740–752.

Ingolf Bork. 2005b. Report on the 3rd round robin on room acoustical computer simulation–Part II: Calculations. *Acta Acustica united with Acustica* 91, 4 (2005), 753–763.

Fabian Brinkmann, Lukas Aspöck, David Ackermann, Steffen Lepa, Michael Vorländer, and Stefan Weinzierl. 2019. A round robin on room acoustical simulation and auralization. *The Journal of the Acoustical Society of America* 145, 4 (2019), 2746–2760.

Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet. 2017. Deep Learning Techniques for Music Generation - A Survey. *CoRR* abs/1709.01620 (2017).

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The AMI Meeting Corpus: A Pre-Announcement. In *International workshop on machine learning for multimodal interaction*. Springer, 28–39.

Diego Di Carlo, Pinchas Tandeitnik, Cedric Foy, Nancy Bertin, Antoine Deleforge, and Sharon Gannot. 2021. dEchorate: a calibrated room impulse response dataset for echo-aware signal processing. *EURASIP Journal on Audio, Speech, and Music Processing* 2021, 1 (2021), 1–15.

Chakravarty R Alla Chaitanya, Nikunj Raghuvanshi, Keith W Godin, Zechen Zhang, Derek Nowrouzezahrai, and John M Snyder. 2020. Directional sources and listeners in interactive sound propagation using reciprocal wave field coding. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 44–1.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. 2020. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 17–36.

Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. 2018. Lip Movements Generation at a Glance. In *The European Conference on Computer Vision*.

Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. 2015. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5556–5565.

P Coleman, L Remaggi, and PJB Jackson. 2020. S3A Room Impulse Responses.

Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. 2020. LibriMix: An Open-Source Dataset for Generalizable Speech Separation. arXiv:2005.11262 [eess.AS]

Trevor J Cox, B-IL Dalenback, Peter D'Antonio, Jean-Jacques Embrechts, Jin Y Jeon, Edgar Mommertz, and Michael Vorländer. 2006. A tutorial on scattering and diffusion coefficients for room acoustic surfaces. *Acta Acustica united with ACUSTICA* 92, 1 (2006), 1–15.

Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. 2018. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4578–4587.

Lukas Drude, Jens Heitkaemper, Christoph Boeddeker, and Reinhold Haeb-Umbach. 2019. SMS-WSJ: Database, performance measures, and baseline recipe for multichannel source separation and recognition. *arXiv preprint arXiv:1910.13934* (2019).

Tiago H Falk, Chenxi Zheng, and Wai-Yip Chan. 2010. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 7 (2010), 1766–1774.

Ziqi Fan, Vibhav Vineet, Hannes Gamper, and Nikunj Raghuvanshi. 2020. Fast acoustic scattering using convolutional neural networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 171–175.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 2021. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10933–10942.

Thomas Funkhouser, Ingrid Carlbom, Gary Elko, Gopal Pingali, Mohan Sondhi, and Jim West. 1998. A beam tracing approach to acoustic modeling for interactive virtual environments. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. ACM, 21–32.

Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. 2020. Visualechoes: Spatial image representation learning through echolocation. In *European Conference on Computer Vision*. Springer, 658–676.

François Grondin, Jean-Samuel Lauzon, Simon Michaud, Mirco Ravanelli, and François Michaud. 2020. BIRD: Big Impulse Response Dataset. *arXiv preprint arXiv:2010.09930*

(2020).

Pierre-Amaury Grumiaux, Srđan Kitić, Laurent Girin, and Alexandre Guérin. 2021. A survey of sound source localization with deep learning methods. *arXiv preprint arXiv:2109.03465* (2021).

Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot. 2014. Multichannel audio database in various acoustic environments. In *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 313–317.

Brian Hamilton. 2021. PFFDTD Software. https://github.com/bsxfun/pffdtd.

Scott H Hawley, Vasileios Chatziioannou, and Andrew Morrison. 2020. Synthesis of musical instrument sounds: Physics-based modeling or machine learning. *Phys. Today* 16 (2020), 20–28.

Teerapat Jenrungrot, Vivek Jayaram, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2020. The cone of silence: Speech separation by localization. *Advances in Neural Information Processing Systems* 33 (2020), 20925–20938.

Shulei Ji, Jing Luo, and Xinyu Yang. 2020. A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions. *arXiv preprint arXiv:2011.06801* (2020).

Xutong Jin, Sheng Li, Tianshu Qu, Dinesh Manocha, and Guoping Wang. 2020. Deep-modal: real-time impact sound synthesis for arbitrary shapes. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1171–1179.

Christoph Kling. 2018. Absorption coefficient database. https://www.ptb.de/cms/de/ptb/fachabteilungen/abt1/fb-16/ag-163/absorption-coefficient-database.html

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5220–5224.

Vinay Kothapally, Wei Xia, Shahram Ghorbani, John HL Hansen, Wei Xue, and Jing Huang. 2020. SkipConvNet: Skip Convolutional Neural Network for Speech Dereverberation using Optimally Smoothed Spectral Mapping. *arXiv preprint arXiv:2007.09131* (2020).

Shoichi Koyama, Tomoya Nishida, Keisuke Kimura, Takumi Abe, Natsuki Ueno, and Jesper Brunnström. 2021. MeshRIR: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 1–5.

Asbjørn Krokstad, S Strom, and Svein Sørsdal. 1968. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration* 8, 1 (1968), 118–125.

Heinrich Kuttruff. 2016. *Room Acoustics* (6th ed.). Taylor & Francis Group, London, U. K.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. 2019. SDR–half-baked or well done?. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 626–630.

Jaakko Lehtinen. 2003. Time-domain numerical solution of the wave equation. *Feb* 6 (2003), 1–17.

Dingzeyu Li, Timothy R. Langlois, and Changxi Zheng. 2018. Scene-Aware Audio for 360° Videos. *ACM Trans. Graph.* 37, 4 (2018).

Siegfried H Linkwitz. 1976. Active crossover networks for noncoincident drivers. *Journal of the Audio Engineering Society* 24, 1 (1976), 2–8.

Qi Liu, Matt J Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278* (2020).

Shiguang Liu and Dinesh Manocha. 2020. Sound Synthesis, Propagation, and Rendering: A Survey. *arXiv preprint arXiv:2011.05538* (2020).

Yi Luo, Zhuo Chen, and Takuya Yoshioka. 2020. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 46–50.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications* 80, 6 (2021), 9411–9457.

Ravish Mehra, Nikunj Raghuvanshi, Lakulish Antani, Anish Chandak, Sean Curtis, and Dinesh Manocha. 2013. Wave-based sound propagation in large open scenes using an equivalent source formulation. *ACM Transactions on Graphics (TOG)* 32, 2 (2013), 19.

Mridul K Mishra and Jaydeep Viradiya. 2019. Survey of Sentence Embedding Methods. *International Journal of Applied Science and Computations* 6, 3 (2019), 592–592.

Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. 2016. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2405–2413.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.

Manuel Pariente, Samuele Cornell, Joris Cosentino, Sunit Sivasankaran, Efthymios Tzinis, Jens Heitkaemper, Michel Olvera, Fabian-Robert Stöter, Mathieu Hu, Juan M. Martín-Doñas, David Ditter, Ariel Frank, Antoine Deleforge, and Emmanuel Vincent. 2020. Asteroid: the PyTorch-based audio source separation toolkit for researchers. In *Proc. Interspeech*.

Kyubyong Park and Thomas Mulc. 2019. Css10: A collection of single speaker speech datasets for 10 languages. *arXiv preprint arXiv:1903.11269* (2019).

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Ville Pulkki and U Peter Svensson. 2019. Machine-learning-based estimation and rendering of scattering in virtual reality. *The Journal of the Acoustical Society of America* 145, 4 (2019), 2664–2676.

Senthil Purushwalkam, Sebastia Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. 2021. Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1183–1192.

Nikunj Raghuvanshi, Rahul Narain, and Ming C Lin. 2009. Efficient and accurate sound propagation using adaptive rectangular decomposition. *IEEE Transactions on Visualization and Computer Graphics* 15, 5 (2009), 789–801.

Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. 2021. IR-GAN: Room Impulse Response Generator for Far-Field Speech Recognition. In *Proc. Interspeech 2021*. 286–290. https://doi.org/10.21437/Interspeech.2021-230

Chandan K. A. Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. 2019. A scalable noisy speech dataset and online subjective test framework. arXiv:1909.08050 [cs.SD]

Chandan K. A. Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke. 2020. The INTER-SPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results. arXiv:2005.13981 [eess.AS]

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

Colleen Richey, Maria A. Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, Paul Gamble, Jeff Hetherly, Cory Stephenson, and Karl Ni. 2018. Voices Obscured in Complex Environmental Settings (VOICES) corpus. arXiv:1804.05053 [cs.SD]

Carl Schissler, Christian Loftin, and Dinesh Manocha. 2017. Acoustic Classification and Optimization for Multi-modal Rendering of Real-world Scenes. *IEEE Transactions on Visualization and Computer Graphics* 24, 3 (2017), 1246–1259.

Carl Schissler, Ravish Mehra, and Dinesh Manocha. 2014. High-order diffraction and diffuse reflections for interactive sound propagation in large environments. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 39.

Alex Southern, Samuel Siltanen, Damian T Murphy, and Lauri Savioja. 2013. Room impulse response synthesis and validation using a hybrid acoustic model. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 9 (2013), 1940–1952.

Alexander Southern, Samuel Siltanen, and Lauri Savioja. 2011. Spatial room impulse responses with a hybrid modeling method. In *Audio Engineering Society Convention 130*. Audio Engineering Society.

Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Černocký. 2019. Building and evaluation of a real room impulse response dataset. *IEEE Journal of Selected Topics in Signal Processing* 13, 4 (2019), 863–876.

Zhenyu Tang, Nicholas J Bryan, Dingzeyu Li, Timothy R Langlois, and Dinesh Manocha. 2020a. Scene-Aware Audio Rendering via Deep Acoustic Analysis. *IEEE Transactions on Visualization and Computer Graphics* (2020).

Zhenyu Tang, Lianwu Chen, Bo Wu, Dong Yu, and Dinesh Manocha. 2020b. Improving reverberant speech training using diffuse acoustic simulation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6969–6973.

Zhenyu Tang, Hsien-Yu Meng, and Dinesh Manocha. 2021. Learning acoustic scattering fields for dynamic interactive sound propagation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 835–844.

Noé Tits, Kevin El Haddad, and Thierry Dutoit. 2019. Emotional Speech Datasets for English Speech Synthesis Purpose: A Review. In *Proceedings of SAI Intelligent Systems Conference*. Springer, 61–66.

Michael Vorländer. 1989. 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 (1989), 172–178.

Michael Vorliander. 1995. International round robin on room acoustical computer simulations. In *15th Intl. Congress on Acoustics, Trondheim, Norway*. 689–692.

Jui-Hsien Wang, Ante Qu, Timothy R Langlois, and Doug L James. 2018. Toward wave-based sound synthesis for computer animation. *ACM Trans. Graph.* 37, 4 (2018), 109–1.

Tao Wu, Yong Jiang, Nan Li, and Tao Feng. 2018. An indoor sound source localization dataset for machine learning. In *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence*. 28–32.

Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. 2018. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3550–3558.