Taylor & Francis
Taylor & Francis Group

Journal of Information Display

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjid20

Reconstructing room scales with a single sound for augmented reality displays

Benjamin S. Liang, Andrew S. Liang, Iran Roman, Tomer Weiss, Budmonde Duinkharjav, Juan Pablo Bello & Qi Sun

To cite this article: Benjamin S. Liang, Andrew S. Liang, Iran Roman, Tomer Weiss, Budmonde Duinkharjav, Juan Pablo Bello & Qi Sun (2023) Reconstructing room scales with a single sound for augmented reality displays, Journal of Information Display, 24:1, 1-12, DOI: 10.1080/15980316.2022.2145377

To link to this article: https://doi.org/10.1080/15980316.2022.2145377

Published online: 15 Nov 2022.

JOURNAL OF INFORMATION DISPLAY
2023, VOL. 24, NO. 1, 1–12
https://doi.org/10.1080/15980316.2022.2145377

Taylor & Francis
Taylor & Francis Group

INVITED PAPER

OPEN ACCESS Check for updates

Reconstructing room scales with a single sound for augmented reality displays

Benjamin S. Liang[a], Andrew S. Liang[a], Iran Roman[a], Tomer Weiss[b], Budmonde Duinkharjav[a], Juan Pablo Bello[a] and Qi Sun[a]

[a]Computer Science and Engineering, New York University, New York, NY, USA; [b]Informatics, New Jersey Institute of Technology, Newark, NJ, USA

ABSTRACT

Perception and reconstruction of our 3D physical environment is an essential task with broad applications for Augmented Reality (AR) displays. For example, reconstructed geometries are commonly leveraged for displaying 3D objects at accurate positions. While camera-captured images are a frequently used data source for realistically reconstructing 3D physical surroundings, they are limited to line-of-sight environments, requiring time-consuming and repetitive data-capture techniques to capture a full 3D picture. For instance, current AR devices require users to scan through a whole room to obtain its geometric sizes. This optical process is tedious and inapplicable when the space is occluded or inaccessible. Audio waves propagate through space by bouncing from different surfaces, but are not 'occluded' by a single object such as a wall, unlike light. In this research, we aim to ask the question *'can one hear the size of a room?'*. To answer that, we propose an approach for inferring room geometries only from a single sound, which we define as an audio wave sequence played from a single loud speaker, leveraging deep learning for decoding implicitly-carried spatial information from a single speaker-and-microphone system. Through a series of experiments and studies, our work demonstrates our method's effectiveness at inferring a 3D environment's spatial layout. Our work introduces a robust building block in multi-modal layout reconstruction.

ARTICLE HISTORY
Received 1 June 2022
Accepted 1 November 2022

KEYWORDS
Scene perception; multi-modal; audio listening; acoustic propagation; augmented reality

1. Introduction

Understanding the dimensions and spatial configuration of a 3D physical environment is a fundamental task for augmented reality (AR) displays, and a requirement for basic AR applications such as virtual object placing [1], physics-based interaction [2], and visualization [3,4]. To measure the physical environment, both humans and computer vision systems (e.g. RGBD cameras equipped on Hololens) primarily rely on sensors that record light reflections from the environment's objects and boundaries. However, light-based acquisition and reconstruction systems are commonly limited by spatial occlusions and field-of-view. Consequently, in an indoor scenario, users may have to repetitively scan the whole room to obtain its size and geometries. The process, besides being tedious and inefficient, does not apply to highly occluded, inaccessible, or low-visibility scenes (Figure 1).

To address this problem using light sensors, solutions have been proposed, such as non-line-of-sight (NLOS) imaging [5,6] or prior- and image-based approximations [7]. However, NLOS solutions typically require customized or high-cost devices, which limits its broad deployment in regular consumer ends. While image-based approximations require full visual coverage of the environment for successful reconstructions [8], other modalities for spatial scene processing models are often biased to work in a particular scene, in specific spatial conditions, or with specific hardware [9,10]. Alternatives to light sensing include radar (radio-waves) [11] and sonar (supersonic sound-waves) [12] sensors, but these require specialized hardware and signal processing, which also limits their broad deployment.

Acoustic waves, on the other hand, propagate similarly to light waves, but implicitly carry spatial content while they propagate, reflect, and pass through indoor spaces [13]. Additionally, audio propagation is less restricted by occlusions and can be broadly acquired by the multi-channel microphone arrays with today's AR devices [14]. Therefore, to avoid current tedious processes or customized hardware used toward scene perception for AR displays, we are inspired to present a learning-based pipeline able to predict the size and layout of an environment using the stereo recording of only an audio source. This data-driven method learns the

CONTACT Qi Sun qisun@nyu.edu

ISSN (print): 1598-0316; ISSN (online): 2158-1606

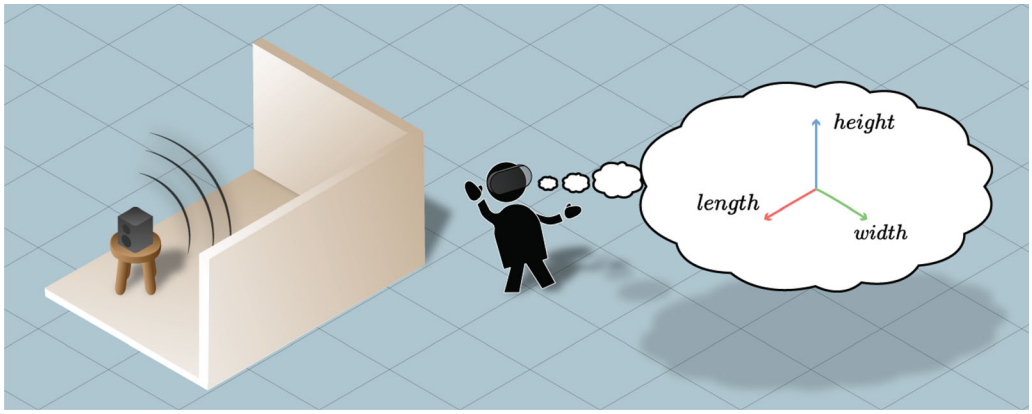


Figure 1. Illustration of our system for AR displays. With a single audio source inside a physical room, our learning-based approach automatically predicts 3D geometry by 'hearing,' allowing for rapid, behind-the-wall room geometry prediction despite optical occlusions.

correspondence between sound reverberation in a room and the room's spatial layout.

The fundamental principle supporting our method is the fact that sounds propagate across space, creating a characteristic *reverb*. This reverb can serve as the signature of the room and encodes information that can be used to calculate the distances between sound-reflecting surfaces (assuming that the location of the sound and the recording microphone(s) are known). One can capture a room's characteristic reverb by using a microphone to record the response to an impulse signal (Dirac delta function) produced by a speaker. The result is a signal that captures the physical interaction between sound pressure waves and the room surfaces [15]. It is possible to estimate the distances between objects and walls using an impulse response (IR) to reconstruct the 3D layout of the room environment. However, this requires prior knowledge of the physical location of the speaker and the microphone(s) used. In order to decode the relationship between the room's reverb signature and the 3D environment, we built and trained two deep learning models which predict the size of simple box-shaped rooms from IRs recorded without explicit knowledge of the speaker or microphone location. Our first model was trained using a large open-source dataset of simulated audio IRs collected inside rooms with diverse sizes and acoustic conditions. Our second model was trained with a smaller dataset that we generated using Unity to simulate the condition where the microphone is outside of the room.

Figure 2 visualizes our computational pipeline that converts an IR from a stereo microphone recording to a 3D reconstruction of a reverberating physical environment. First, the IR is obtained by deconvolving a stereo recording of a stimulus signal played by a loudspeaker in the room. Next, we calculate the difference between the recordings of each stereo channel to yield a third signal to consider human perception of stereoscopic disparity, including difference in amplitude, distance of sound traveled, resonances, and anti-resonances. Then, we convert these three signals into a Mel-spectrogram (MS), a 2D time-frequency representation of reverberant energy, which we use as an image input to a CNN model which decodes the room's 3D dimensions. Simulated experiments demonstrate our method's accuracy and precision in predicting various three-dimensional spaces.

We envision our work to open new possibilities of rapidly establishing physical scene perception for displaying real objects and environments in AR. Our framework applies to highly-occluded or remote scenes, without the currently tedious camera-based scanning process. Furthermore, it exploits audio signal processing hardware and software used with consumer-level AR displays, for multi-modal spatial scene content detection in augmented reality and autonomous driving. We also believe that, in order to be able to carry out the 3D scene reconstruction task, our model must embed the input signal into a generalized spatial representation that could be used in future work for other downstream tasks, such as discrete sound event localization or acoustic imaging, to name a couple of examples.

2. Related work

Interior spaces are part of everyday life. Hence, it is not surprising that much research has been invested toward understanding how to virtually recreate spaces from differing modalities such as optical and aural sources (e.g. light rays and sound waves). Below we classify previous work based on their respective focus area.

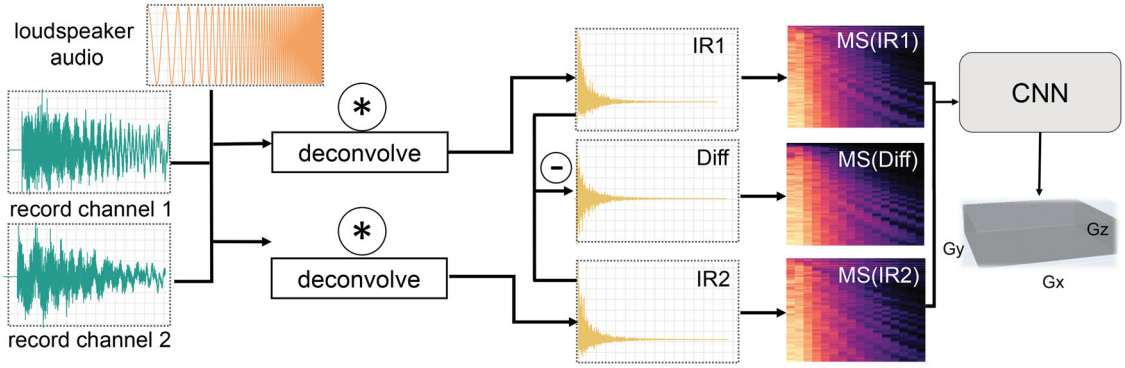


Figure 2. Our pipeline takes a two-channel recording from a stereo microphone. A deconvolution operation between these recordings and the loudspeaker signal yields the impulse response (IR) each channel. Then, the Mel-spectrograms of channels 1, 2, and the differences (Diff) between the channels are calculated to incorporate amplitude, distance, resonance, and anti-resonance disparities (Section 3). Finally, the Mel-spectrograms images are stacked and used as the input features for our CNN, which predicts the room geometry.

2.1. Visual-audio learning

Many assume that we perceive the world around us in a visual fashion. However, there are other important modalities that shape the way in which we sense our space. For example, it has been shown that we can estimate spatial dimensions by audio and echolocation [16]. This also applies to augmented reality experiences [17]. Hence, there has been continuing research interest on the intersection of audio and visual modalities in virtual environments [18]. Among such work, we briefly overview approaches most related to our work, which are recent deep learning methods that focus on this emerging area.

Previously, [19] proposed a method for learning sound associated with images by using a CNN-based neural network. Additionally, [20] proposed a method to estimate depth from the sound. They use real-world data of coupled audio and visual recordings. While they do manage to estimate depth, they do not investigate the problem of full 3D room geometry. Finally, [21] demonstrated how to estimate depth based on RGB input and a supplemental audio echo response. Other work focused on identifying acoustic properties of scene geometries [22].

2.2. Scene inference

Scene inference is an area of interest for multiple research communities, such as vision, robotics, and immersive reality. The main goal of this research is to create a virtual representation of a real-world input scene which could be used for visualization on mixed-reality devices. Such representations vary from low-quality scene reconstruction to high-fidelity semantic 3D modeling [23]. Hence, a diverse set of methods has been proposed. Since geometry can be intuitively described visually, a common approach for scene inference is via images and related inputs. In such cases, single images [24,25], multi-view images [26], depth maps [27,28] are used in conjunction with camera locations and computational machinery such as deep neural networks to reconstruct real-world 3D scenes virtually.

In addition to images, researchers have looked into other modalities for understanding spatial surroundings. In robotics, [29] demonstrates a method that uses sonar for classifying 3D objects. Sonar has been shown to be better than vision for estimating specific geometrical proprieties for scene reconstruction [30]. Intermingling modalities may improve scene inference results. For example, [31] demonstrated that using both audio and visual sensing allows rapid floorplan reconstruction from video.

Our work focuses on sound as a means to estimate 3D spatial structures, since sound propagates in a manner different from light, allowing inference of features beyond line-of-sight. Earlier work proposed an approach for estimating the shape of 2D polygonal acoustic spaces via IRs [32]. Researchers have also shown that it is possible to use an IR to measure the time it takes for sound reflections from walls to reach a microphones, and use this measurement to triangulate sound sources [33–35]. Other methods either require use of sophisticated signal processing on recordings by multi-microphone arrays [36], or must reduce the complexity via strong assumptions to make the problem solvable using the wall and object reflections that can be captured by an IR [37,38].

Recently, researchers have proposed using audio recordings of speech to train deep neural networks that are able to estimate room volume [39], and IR measurements to estimate room geometry [40]. However, these methods were either developed on closed-source data and software, did not study the condition where

the microphone is outside the room, worked with signals exclusively in the time domain, or did not make use of stereo recordings, which are a better presentation of human listening and can lead to better model performance. Our paper shows that stereo signals (commonly built-in in AR and VR hardware) and their difference can be used for accurate room geometry predictions in conditions where the microphone is inside or immediately outside the room.

3. Method

This section is structured as follows. In Section 3.1.1 we define our room geometry formulation. Next, Section 3.1.2 describes the physics of audio wave propagation in a room, the mathematical relationship between spatial and acoustic components in an enclosed space, and the signal processing principles needed to acquire signals that explain this relationship. The resulting acoustic pressure equation is non-unique, demonstrating why it is impossible to obtain room geometry analytically using only the acoustic wave equation. Subsequently, Section 3.1.3 explains how an IR can be used for audio-only geometric reconstruction, and in Section 3.1.4 we explain what a Mel-spectrogram time-frequency feature-map is. After this, in Section 3.2 we introduce the two IR datasets that we process to generate inputs for our deep learning model. The first dataset simulates stereo IR recordings inside a room, while the second dataset simulates stereo IR recordings outside a room. Finally, Section 3.3 presents our convolutional neural network (CNN) architecture that uses Mel-spectrograms to estimate room geometry (see Figure. 2 for the illustration of our data-model pipeline).

3.1. Definitions

3.1.1. Room geometry

In the scope of this paper, we consider 3D indoor, closed, empty, box-shaped rooms. We assume rooms to contain a sound-permeable microphone that does not reflect or occlude audio waves. We also assume audio signals are produced by omnidirectional speakers, which a microphone is able to record via a stereo/two-channel signal. Formally, we define the room geometry $\mathbf{G}$ with:

$$\mathbf{G} := (G_x, G_y, G_z), \tag{1}$$

where G_x, G_y, and G_z indicate the room's width, height, and length in meters, respectively. Considering real-world room spaces, we assume each dimension to be within a range $G_i \in [G_{i,\min}, G_{i,\max}]$, for $i \in \{x, y, z\}$.

Note that different building materials (e.g. metal vs. wood) may introduce varied signal propagation-affecting behaviours, such as absorption, reflectiveness, and transmission of signals. Our system considers rooms with walls that exhibit these sound-reflective properties with different absorption coefficients. Additionally, we assume the speed of sound is approximately constant, with room temperature varying within a small range (see Section 3.2).

3.1.2. Acoustic propagation

An acoustic signal can be measured by a microphone membrane due to the changes in acoustic pressure that it causes in a medium, such as air [41]. The acoustic signal could be measured either inside the room or from the outside if the sound resonance leaks out of the room. Signal propagation is governed by the acoustic wave equation, which provides a framework for calculating how the acoustic pressure changes across space (i.e. a specific position $\mathbf{P}$) and time: $A(\mathbf{P}, t)$. Given initial conditions for source position and time inside room geometry $\mathbf{G}$:

$$A(\mathbf{P}, 0) = A_m, \tag{2}$$

we can use the acoustic wave equation to compute the acoustic pressure A_m [41]:

$$\nabla^2 A_m - \frac{1}{c^2}\frac{\partial^2 A_m}{\partial^2 t} = 0, \tag{3}$$

where ∇^2 is the Laplace operator, c is the speed of sound in the medium, and t is time.

In a room, this displacement results from the sum of spherical waves originating from sound sources and reflecting surfaces [42]. Recording the wave displacement results in a signal that contains information about the Time of Arrival (TOA) from the sound source(s) and wall reflections to the microphone, although with different amplitudes. It is possible to use the TOA to decode the distance between sound source, early (i.e. louder) reflections, and the microphone, but this requires appropriate information about the location of the microphone in order to do a simple triangulation. Hence, the wave propagation functions and the TOA information in an impulse response measurement are unable to describe a direct relationship between room $\mathbf{G}$ and A without reference points. It is, therefore, necessary to use another method to map wave propagation to $\mathbf{G}$.

Note that our goal is to obtain $\mathbf{G}$ via the acquired A_m. However, Equation (3) is a differential function which may have multiple solutions (i.e. non-unique), making it impossible to obtain the $\mathbf{G}$ analytically. Thus we devise a data-driven approach via a deep neural network and narrow the signal into a particular type of time-frequency features, as discussed below.

3.1.3. Impulse response and signal acquisition

Due to the infinite number of possibilities for the location and type of A_s (the sound source amplitude), solving Equation (3) with a data-driven approach involves the discrete sampling of data points that represent a finite number of A_s. Thus, we focus on a specific type of audio signal: a sine wave with a frequency that grows logarithmically over 1 second with a sampling rate of 16 kHz to fill a physical space with acoustic energy at all audible frequencies. In the datasets we used, this signal is a logarithmic sine sweep (or chirp), starting at a low frequency of 20 Hz and reaching a high frequency of 8 kHz. This stimulus captures the range of frequencies where information-rich signals, like speech, are audible to the average person [43].

A recording of the sine-sweep signal in a room can yield the *Room Impulse Response* (RIR), which is a transfer function implicitly parameterized by the room acoustics. To obtain the RIR from the sine-sweep recording, one must deconvolve the RIR using:

$$I(t, A_m, A_s) = \mathbb{F}^{-1}\left(\frac{\mathbb{F}(A_m)}{\mathbb{F}(A_s)}\right), \tag{4}$$

where $\mathbb{F}$ / $\mathbb{F}^{-1}$ are the Fast Fourier Transform and the inverse Fast Fourier Transform, respectively. Assuming that A_s is a signal with energy at all audible frequency components (i.e. a chirp signal), I approximates a Dirac delta function being played in the room. Note that Equation (4) only uses the speaker and microphone audio signals A_m, A_s without requiring explicit knowledge of the room geometry.

3.1.4. Mel-frequency spectrograms (MFS)

Performing data-driven computations directly on the impulse responses I that Equation (4) yields is suboptimal for calculating the room estimation. As shown empirically in our results (Section 4.3), using such a 1D signal (i.e.time-series only) with a large number of highly-correlated time-based features can easily lead to model over-fitting. Therefore, it is better to transform the signal into a time-frequency representation that breaks down the signal into orthogonal frequency bands, reduces temporal redundancy, and converts the time-series signal into a 2D representation, similar to images used to train state-of-the-art computer vision neural networks [44]. To achieve this transformation, we convert every I datapoint into a Mel-spectrogram. This transformation is described by [45]:

$$\mathbb{M} \cdot \mathbb{F}(I(t, A_m, A_s)), \tag{5}$$

where $\mathbb{M}$ computes the dot product between a bank of Mel-filters and the Short-time Fourier transform of the I. In our pipeline, each I is converted into a Mel-spectrogram using a sliding window of 128 ms that moves at a resolution of 64 ms. The resulting Mel-spectrograms have 128 logarithmically-spaced frequency bins/Mel-filters centered between 20 Hz and 8 kHz.

3.2. Dataset

Our model considers conditions where the microphones may be located inside or outside the target room. To this end, we carry out two experiments with different datasets. The first one is a publicly available dataset with RIR simulated via the image method [46], where both the recording stereo microphone and the omnidirectional sound source are inside the room. The second dataset, which we created, simulates similar conditions but with the microphone located outside of the room instead. The **G** of the rooms simulated in both datasets are used as labels for their corresponding RIR signals.

3.2.1. Microphones inside a room

The Big Impulse Response Dataset (BIRD) [46] consists of RIRs of simulated rooms that were generated using the Image Method [47]. The dataset includes 100,000 rooms randomized with lengths and widths between 5.0 m and 15.0 m, and heights between 3.0 m and 4.0 m. Across data points, the wall absorption coefficient is uniformly varied between 0.2 and 0.8, the speed of sound is uniformly varied between $340.0\frac{m}{s}$ and $355.0\frac{m}{s}$ (to reflect a narrow range of diverse temperatures inside rooms), and the distance between microphones is uniformly varied between 0.01 m and 0.3 m. Sound sources and microphones are placed randomly in rooms, ensuring a minimal distance of 0.5 m between microphones and sound sources. For each room, there are 4 impulse responses measured by 2 microphones, totalling 800,000 impulse response signals.

3.2.2. Microphones outside of a room

We are also interested in predicting room size when the microphone is outside of the room (i.e. sound-based geometric estimation of occluded spaces), which motivates a second custom dataset; in order for a sound played inside a room to reach the microphone outside, audio signals must traverse the environment, interact with surfaces, and encode spatial geometric information before reaching the outside and the microphone. The microphone and speaker in this dataset are acoustically transparent (i.e.do not reflect acoustic signals) and do not interfere with the room reverberations. In contrast with the BIRD dataset, in this second dataset, variation of wall reflecting materials was not considered; thus, sounds reflect from walls with the same attenuation across all data points.

Audio wave propagation behaviour is simulated using the Resonance Audio Package for Unity, a sound synthesis library with natural sound and occlusion effects that imitate sound wave propagation and reverberation behaviours. In this package, the reflection of audio waves is calculated with ray tracing, which allows for the computation of realistic acoustic behaviour.

To generate RIRs for a given room, we played the sine-sweep signal and deconvolved the RIR using Equation (4) on the original sine sweep signal and the recorded audio. We generated 25,000 data points of two-channel RIRs with different room geometry information **G**. Each room environment was randomly generated to match BIRD's [46] **G** specifications: closed rectangular prisms with lengths and widths between 5.0 m and 15.0 m, and heights between 3.0 m and 4.0 m. For every data point, we recorded audio from a randomly placed omnidirectional speaker. The speaker was always positioned within the room, and we used a stereo microphone placed outside of the room to record the propagated audio signal. In all cases, if we consider the (x, y, z) coordinates (0, 0, 0) to be the center of the room, the microphone was placed at a height of 0 and an x-position of $(-G_x/2 - 1)$, or 1 meter out from the center of a wall in the x-direction. The forward/look-at direction of each microphone was the center of its respective room.

3.3. Deep neural networks for learning room geometry

Given the time-frequency mapping of Equation (5), we model our problem as a 2D image to 3D vector mapping $\mathbb{M} \cdot \mathbb{F}(I(t, A_m, A_s)) \rightarrow \mathbf{G}$. To implicitly reveal the mapping, we devise a data-driven approach using a convolutional neural network (CNN).

3.3.1. Network architecture

Our model architecture has a 2D input with a depth of three channels. Each channel is a Mel-spectrogram of shape 128 frequency bins by 256 time bins for a total of 32,768 features per Mel-spectrogram image. To make RIRs of different durations fit into 256 time bins, we resized the number of time-bins using a bilinear interpolation resizing method. Across all RIRs, the Mel-spectrogram always had 128 frequency bins, so interpolation on the frequency axis was never needed. We leverage alternating layers of 2D convolutional and max pooling operations to perform time-frequency convolution and downsampling on Mel-spectrogram features (Figure 3). These layers are followed by dense layers and an L1 regularization penalty to reduce overfitting. The final output is a vector of size 3, predicting G_x, G_y, G_z. We used the rectified linear unit (relu) as the activation function after all convolution and dense transform operations in the neural network. Figure 3 gives an overview of our CNN.

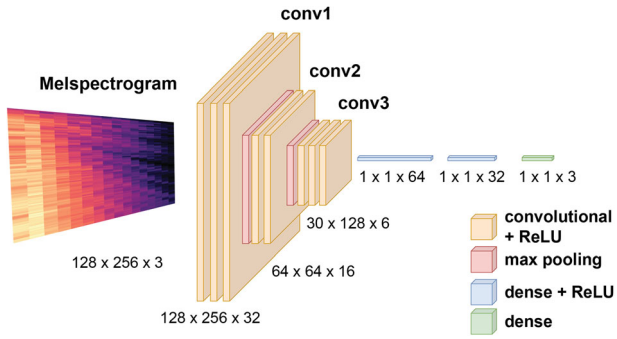


Figure 3. Our deep neural network architecture, which contains three convolutional 2D layers followed by three dense layers. The model is trained to predict room geometries for signals recorded by a stereo microphone inside and outside a room. The network input is a Mel-spectrum signal, while the output prediction is a vector containing the 3D shape of the room.

3.3.2. Loss function

We used Mean Squared Error (MSE) as our loss function to train the network. MSE is a widely used metric for regression tasks, which is calculated as the element-wise average of the squared differences between the ground truth and predicted values, (G_x, G_y, G_z) and $(\hat{G}_x, \hat{G}_y, \hat{G}_z)$, respectively: $\frac{1}{n}\sum_{i=1}^{n}(G_i - \hat{G}_i)^2$, where n is the number of training samples used to compute the MSE. This function is easy to interpret for our purposes, since its square root is the absolute error between the target and the values predicted by the model.

3.3.3. Training and implementation

The goal of the training process is to find a neural network with a set of weights that minimizes the loss function from Section 3.3.2. For all experiments, we use a learning rate of 0.01. Deep networks are prone to overfitting, hence, we use an L1 regularization penalty of 0.01. We also used the stochastic gradient descent-based Adam optimizer [48] with an epsilon of 1e−7 to seek the lowest MSE for room geometry **G** regression.

We trained our neural network architecture on Mel-spectrograms generated from both datasets (Sections 3.2.2 and 3.2.1) separately and without prior transfer learning, resulting in two models: BIRD-CNN, using BIRD-only data simulating microphones inside the room, and Outside-CNN, using only our custom data simulating microphones outside of the room. Mel-spectrogram inputs were generated using the same procedure for both models Section 3.1.4.

We also carried out ablation studies (discussed further in Section 4.3) using the same architecture. To

test the effect of audio signal length on model performance (Section 4.2), we trained five additional CNNs on Mel-spectrograms extracted from audio files of different durations from our dataset. In our ablation study, a model was again trained for each dataset, but on raw impulse response time-series (1D time-only representation) rather than Mel-spectrograms (2D time-frequency representation). All models and their data specifications are denoted in Table 1. All model training was done on a single NVIDIA GTX3090 GPU using Keras [49].

4. Evaluation

To evaluate the quality of our method, we first present our model performance metrics in Section 4.1. We approach the evaluation from the perspective of how accurately our models may predict a room's geometry with purely audio information. Furthermore, as our input Mel-spectrograms are generated from recorded audio waves, a temporal signal, the minimum amount of data needed to develop our model primarily determines how fast/responsively the system can achieve the scene perception aim. Consequently, we perform a pressure test on the recorded audio duration vs. scene perception accuracy in Section 4.2.

4.1. Predicting room geometry

Evaluation metrics We evaluate our model by measuring the accuracy and precision of its predictions. Given a ground truth geometry $\mathbf{G}$ and the corresponding model prediction $\hat{\mathbf{G}}$, we can quantify the accuracy and precision using the mean and standard deviation of the ratios of $\mathbf{G}$ and $\hat{\mathbf{G}}$. Specifically, given a ratio $r_i = \hat{G}_i/G_i$ for $i \in \{x, y, z\}$, we measure accuracy with the *percentage error* of the model predictions from ground truth:

$$\epsilon_i = (\mu_i - 1) \times 100\%, \tag{6}$$

where μ_i is the mean of r_i. We measure precision with σ_i, the standard deviation of r_i.

If the percentage error, ϵ_i, equals zero, the model has no bias toward under- or over-estimating the true room dimensions. Similarly, if the precision metric, σ_i, equals zero, the model has low variation in the prediction.

Research questions We train separate models to evaluate our method's effectiveness for audio-based AR scene perception. Our research questions are: *First*, do the captured audio signals sufficiently carry the spatial information required to understand the physical scene geometry? This can be assessed by measuring $\mathbf{G}$ regression accuracy in BIRD-CNN and Outside-CNN. *Second*, can our method robustly apply to arbitrary listening conditions despite optical and physical occlusion/obstruction of a recording device (e.g. a wall between a remote sensing microphone and the scene of interest)? The answer determines the potential practical AR display applications since users may position themselves in arbitrary locations while seeking to obtain knowledge of the physical space. *Third*, is the model biased toward a given condition? Comparing the performance variances across the two conditions above may reveal the answer: hypothetically, if the two conditions show close to identical performance, the data may indicate a lack of bias.

Hypotheses We hypothesize that the Outside-CNN may perform worse than BIRD-CNN due to acoustic occlusion properties of the wall placed between the sound source and the microphone, where the wall could act as a filter and dampen the transmitted audio signal. We considered that sounds that are able to transmit through a wall would be dominated by low frequencies (resonant with the large room geometry). As a result, the missing high-frequency information in this outside-room condition may compromise the useful room geometry information. On the other hand, because our simulated data does not randomize microphone positions or wall absorption/reflectiveness (as the BIRD data does), it may be easier for the neural network to overfit our generated dataset's features. In addition, our dataset includes impulse responses that are longer than 1 second in duration, whereas the BIRD dataset is limited to 1-second impulse responses, which may limit the useful information for training.

Results We quantify the performance of our models using the metrics defined above, and test the hypotheses proposed. We evaluate our model on the test partition of our dataset which is comprised of 10% of the total data. Table 2 shows the percentage error of predictions ϵ, the spread of prediction errors σ, as well as the model's computed MSE losses. Figures 4 and 5 visualize the results: While the BIRD-CNN model has a lower ϵ across all predictions, and thus predicts more closely to ground truth values (i.e. higher accuracy), the Outside-CNN exhibits a lower σ, and thus has less variance (i.e. higher precision) than the BIRD-CNN model. Compared with a baseline condition that learns the average shape of the datasets, our method demonstrates significant enhancement of precision (low σ).

Analysis The experimental results indicate that our method predicts room geometry $\mathbf{G}$ with reasonable accuracy using only audio signals from a stereo listener. Our accuracy and precision metrics ϵ and σ for both CNNs are approximately equal to the ideal values (0 ± 0), for $\mathbf{G}$ and all G_i. Notably, the results also suggest that not only can this be done using audio acquired from within an enclosed room of interest, but that similar results are achievable by using audio acquired from outside

Table 1. Input feature and training parameter specifications (number of data points per training split) for all models trained according to our CNN architecture.

Model	Input image feature (128 × 256 × 3 shape)	Train	Test	Validation
BIRD-CNN	MFS of IR from BIRD	320,000	40,000	40,000
Outside-CNN	MFS of IR from our dataset	20,000	2,500	2,500
Outside-0.01	MFS of first 0.01 sec of IR from our dataset	20,000	2,500	2,500
Outside-0.1	MFS of first 0.1 sec of IR from our dataset	20,000	2,500	2,500
Outside-0.5	MFS of first 0.5 sec of IR from our dataset	20,000	2,500	2,500
Outside-1	MFS of first 1.0 sec of IR from our dataset	20,000	2,500	2,500
BIRD-Raw	Reshaped IR from BIRD	320,000	40,000	40,000
Outside-Raw	Reshaped IR from our dataset	20,000	2,500	2,500

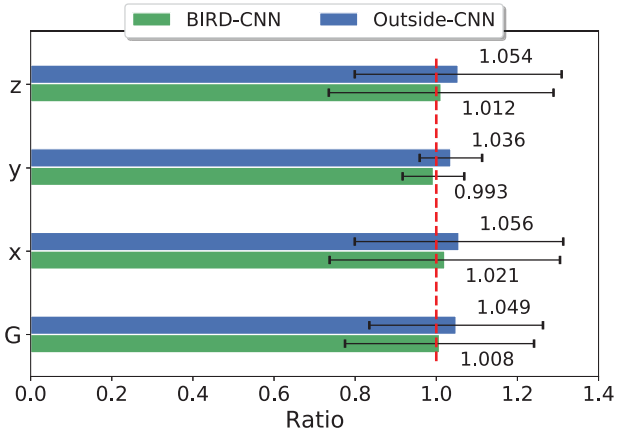


Figure 4. Evaluation of our approach with our ratio metrics μ and σ. The bars denote the μ scores (rounded to the first decimal). The error bars denote the σ scores. The vertical dashed red line indicates the ideal μ of 1.0.

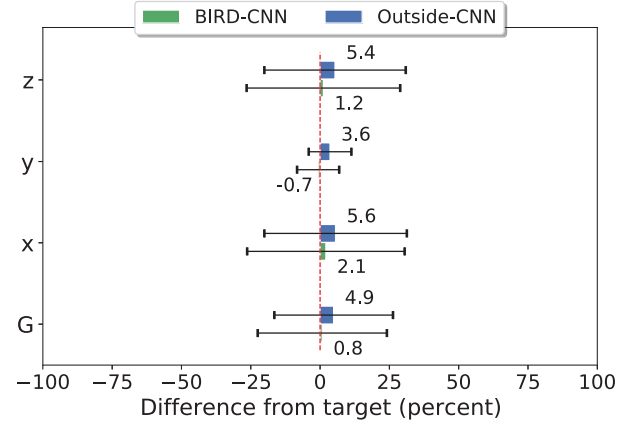


Figure 5. Evaluation of our approach with our ϵ and σ metrics in percentages. The bars denote the ϵ value (rounded to the first decimal). The error bars denote the σ in percentages. The vertical dashed red line indicates the ideal ϵ of 0.

the environment, as observed from the marginal difference between Outside-CNN and BIRD-CNN metrics (Figure 6).

The percentage error metric, ϵ, supports the initial hypothesis that, on average, the BIRD-CNN can predict room geometries more accurately than the Outside-CNN

Table 2. Results for predicting room geometries via a single acoustic source.

	(a) BIRD-CNN results				(b) Outside-CNN results			
	MSE	ϵ(%)	σ	σ_0	MSE	ε(%)	σ	σ_0
G	4.0	0.8	.23↓	.30	3.9	4.9	.21↓	.32
G_x	6.0	2.1	.28↓	.35	5.9	5.6	.26↓	.38
G_y	0.1	−0.7	.08	.08	0.1	3.6	.08	.08
G_z	6.0	1.2	.28↓	.35	5.7	5.4	.26↓	.39

Notes: (a) and (b) show the results with microphones inside (BIRD) and outside the room, respectively. A lower MSE indicates predictions match the ground truth more accurately. σ_0 shows the standard deviation of the baseline condition which learns and predicts the average shape of the datasets.

can. All CNNs were able to predict the G_y most accurately of the dimensions, which can be attributed to the smaller range of G_y values (3–4 meters) as opposed to the other dimensions (5–15 meters), and thus better coverage of data for training across data splits. Additionally, the lower σ results for Outside-CNN may be attributed to the more limited parameter variation in our room simulations as opposed to the BIRD's randomized range of data, as described in Section 3.2.

4.2. Data duration vs. quality

The audio length needed for the method is an essential factor for the responsiveness of the proposed framework: shorter recording requirement allows for faster perception, especially in dynamic environments. However, it may potentially compromise the accuracy. We aim to measure the effects between the two confounding factors. To this end, we train our CNN model using Mel-spectrograms generated from audio files truncated to only the first 0.01, 0.1, 0.5, and 1-seconds and observe the changes in prediction accuracy.

As seen in Figure 7, the MSE and σ (lower means decreased error variances and thus better prediction accuracy) of the model decrease as the duration of recorded audio increases. This supports our hypothesis that longer raw recordings of the signal convolved by the room contain more useful information for the model. Investigating the optimal responsiveness-quality balancing is an interesting future direction.

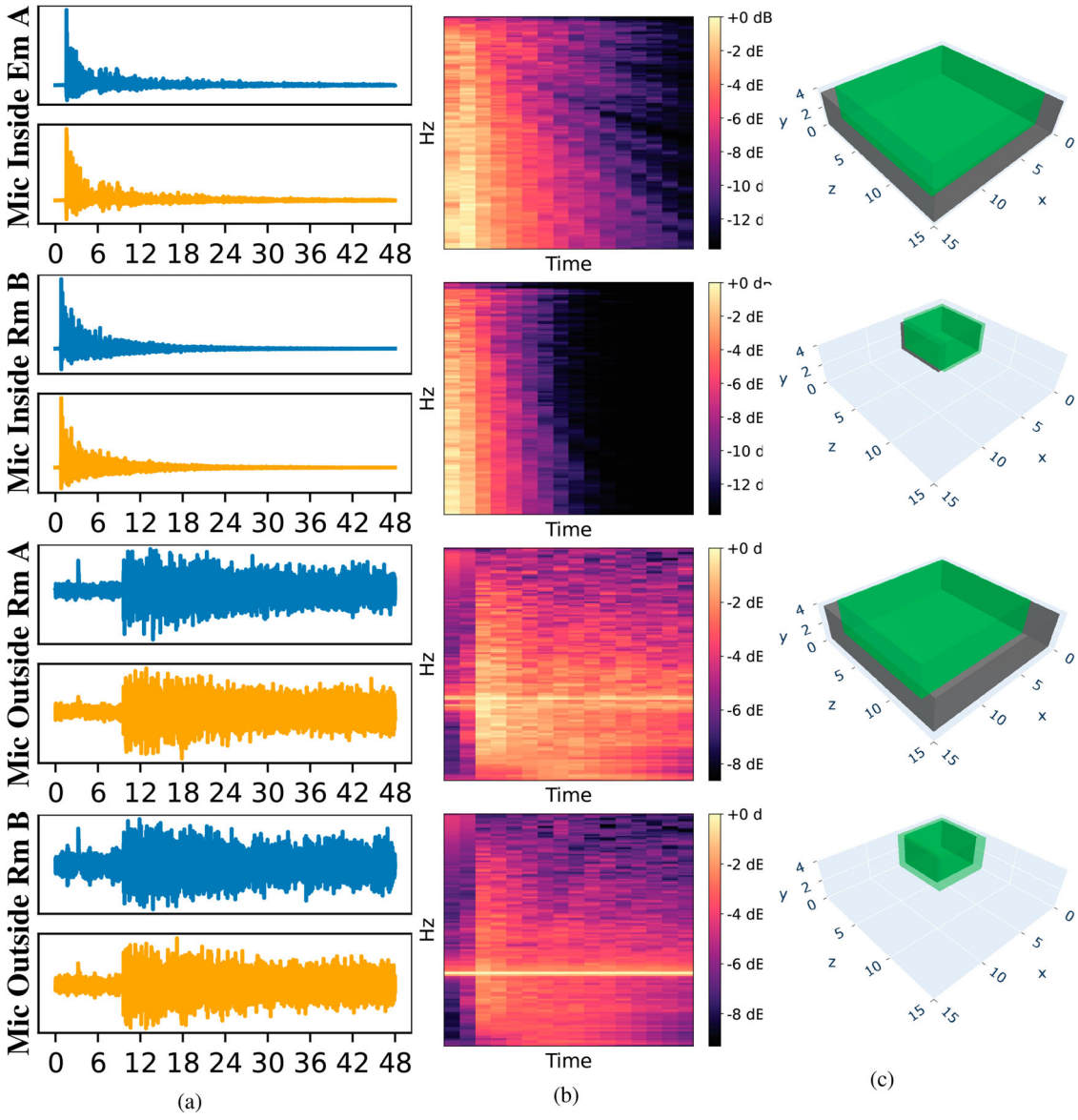


Figure 6. Example results using our method for interior microphone (top two rows) and exterior microphone (bottom two rows) for large (Rm A) and small rooms (Rm B). (a): two-channel IR plots (blue: channel 1, orange: channel 2), which demonstrate amplitude vs. time (in hundredths of a second) at a sampling rate of 16 kHz. (b): Mel-spectrogram of the first IR channel (blue plot) showing the frequency in Hz vs. time. The colormap describes the measure of decibels in each frequency band at a point in time, from 0 to 8,192 Hz (bottom to top on the *y*-axis) and 0 to 0.48 seconds (left to right on the *x*-axis). (c): Prediction visualizations, which demonstrate the differences between the ground truth and our method's prediction. The black colour denotes ground truth room geometry, and the green colour denotes predicted geometry. (a) Impulse response. (b) Mel-Spectrogram and (c) Prediction visualization.

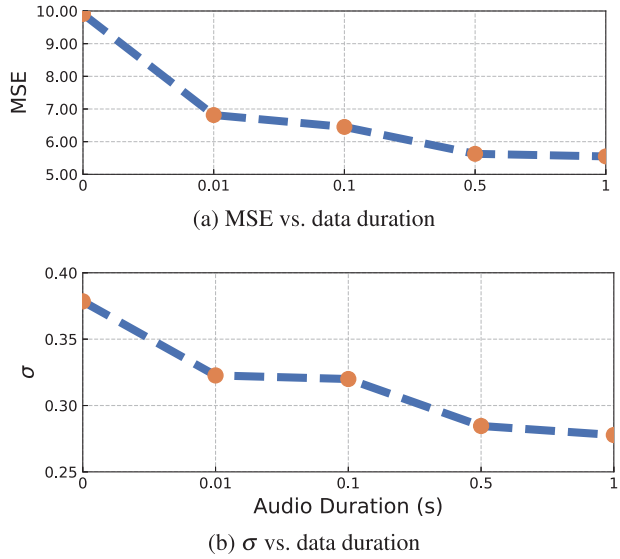
4.3. Ablation study

Table 3 shows that training our CNN architecture on raw RIRs instead of Mel-Spectrograms yields a lower bias toward under- or over-estimating the room geometry (i.e. higher accuracy), as indicated by the smaller ϵ across both datasets. However, when comparing σ results in Table 3, Table 2(a,b), we can see that the networks trained on the raw RIRs were less precise due to the higher prediction variation. This is most apparent between the CNNs which use the RIRs of our dataset, since OUTSIDE-CNN exhibits a geometry ratio standard deviation of 0.214, and the ablation experiment's CNN has a standard deviation of 0.283.

Table 3. Ablation study results.

	MSE	ϵ(%)	σ
BIRD-Raw	4.227	0.3	0.240
Outside-Raw	6.529	0.4	0.283

The results from this experiment show that like BIRD-CNN and Outside-CNN, the accuracy and precision are



(a) MSE vs. data duration

(b) σ vs. data duration

Figure 7. (a,b) shows MSE/σ changes as a function of audio wave duration used for training. (a) MSE vs. data duration and (b) σ vs. data duration.

approximately equal to the ideal values (0 ± 0), suggesting that a CNN trained on raw RIRs only is robust enough to predict room geometry.

5. Discussion

We presented a machine learning-based method for discovering room geometries with invisible acoustic signals. The method simply requires a single source emitting audio for a short duration.

Our method faces several limitations. First, our method was tested in simulated settings. Extensions to real-world use cases are possible but require further experimentation, as the model may require learning of ambient noise. Second, our data generation approach assumes walls do not possess modular absorption coefficients. Changing the absorption coefficient of surface materials would allow for the simulation of different types of walls, such as wood, concrete, and brick. Third, our model assumes that the audio source in each room is an omnidirectional speaker; a speaker playing a sound in a specific direction will result in different sound propagation throughout the room. Lastly, the method only predicts room geometries but not interior objects, such as furniture/human locations and sizes. For the purposes of our experiment, we used only simple geometry. Expanding our simulated dataset to include rooms with populations of different objects and arrangements of objects would further improve our dataset for training a model toward real world applications. Future experiments could be performed to evaluate how different combinations of wall and object material properties affect room geometry prediction from outside of a room. Exploring high dimensional output with varied neural network design and datasets is an interesting future direction.

6. Conclusion

Acquiring, perceiving, and understanding the physical environment has remained an essential problem for accurately displaying augmented reality scenes. It serves as the data foundation for 3D modeling, interacting with virtual objects, and physics-based animation. In practice, scene reconstruction for AR displays has been broadly limited by the tedious process of scanning room geometries via hand or eye-worn cameras. Additionally, the process also faces inherent barriers, especially occlusions from walls and interior objects. In this paper, we present the first attempt at addressing this problem from the acoustic perspective. To this end, we develop an end-to-end framework that only requires a single speaker and an arbitrarily placed microphone (inside or outside the room). The framework is driven by our hybrid synthetic dataset, along with a signal-tailored machine learning approach. With simulated environments, we demonstrate our system's effectiveness and accuracy in precisely predicting a broad range of 3D room geometries. Although validated only with simulated environments such as ambient noise, we envision our method to open new possibilities and directions in the field of multi-modal scene perception and visualization for AR. We envision that further development along with addressing the limitations may introduce new applications beyond this scope, including assistive technologies to people with visual impairments, and privacy protection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Defense Sciences Office, DARPA [PTG] and NSF [2232817] and [2225861].

Notes on contributors

Benjamin S. Liang is a first year Ph.D. Student at New York University. His research interests lie in VR/AR and human perception.

Andrew S. Liang is a first year Masters student at Columbia University. His research interests lie in computational neuroscience and VR/AR.

Iran Roman is a Post-doctoral researcher at New York University. His research interests include developing machines that can listen to music and speech like humans. He holds a Ph.D. from Stanford University in Computer-based Music Theory and Acoustics.

Tomer Weiss is a professor at the Informatics Department in the New Jersey Institute of Technology. He holds a Ph.D. from University of California, Los Angeles. He is broadly interested in real-time optimization for visual applications, AR/VR/XR, multiagent simulation, 3D content creation, and machine learning.

Budmonde Duinkharjav is a Ph.D. candidate at New York University. His research is focused around studying the relationship between what we observe in our surroundings and how its perception affects our behavior and ability to perform visual tasks in the context of computer graphics applications. More broadly, he is interested in how research on human perception can be leveraged to augment and improve computer graphics systems to aide in our daily lives.

Juan Pablo Bello is a Professor of Music Technology and Computer Science & Engineering at New York University. He holds a Ph.D. from Queen Mary, University of London. He is the director of the Music and Audio Research Lab (MARL), where he leads research on sound and music informatics.

Qi Sun is an assistant professor at New York University, where he leads the Immersive Computing Lab. He obtained his Ph.D. at Stony Brook University. His research interests lie in perceptual computer graphics, computational cognition, VR/AR, and visual optics.

References

[1] Behringer R., Klinker G. and Mizell D., *Augmented Reality: Placing artificial objects in real scenes* (AK Peters/CRC Press, New York, 1999).

[2] Luo X., Huang J.B., Szeliski R., Matzen K. and Kopf J., ACM Trans. Graph. (ToG) **39** (4), 71–1 (2020).

[3] Han L., Zheng T., Zhu Y., Xu L. and Fang L., IEEE. Trans. Vis. Comput. Graph. **26** (5), 2012–2022 (2020).

[4] Sicat R., Li J., Choi J., Cordeil M., Jeong W.K., Bach B. and Pfister H., IEEE. Trans. Vis. Comput. Graph. **25** (1), 715–725 (2018).

[5] Faccio D., Velten A. and Wetzstein G., Nat. Rev. Phys. **2** (6), 318–327 (2020).

[6] O'Toole M., Lindell D.B. and Wetzstein G., Nature **555** (7696), 338–341 (2018).

[7] Lee C.Y., Badrinarayanan V., Malisiewicz T. and Rabinovich A., Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy (2017).

[8] Pintore G., Mura C., Ganovelli F., Fuentes-Perez L., Pajarola R. and Gobbetti E., Comp. Grap. Forum **39** (2), 667–699 (2020). doi:10.1111/cgf.14021.

[9] Roman I.R. and Bello J.P., Detection and Classification of Acoustic Scenes and Events 2021 (2021).

[10] Politis A., Adavanne S., Krause D., Deleforge A., Srivastava P. and Virtanen T., Detection and Classification of Acoustic Scenes and Events 2021 (2021).

[11] Lee J.Y., Kim Y., Lee S., Cho W. and Kim S.C., IEEE. Sens. J. **19** (24), 12316–12324 (2019).

[12] Chen W., Xu J., Zhao X., Liu Y. and Yang J., IEEE Trans. Indust. Electron. **68** (7), 6042–6052 (2020).

[13] Farina A., Martignon P., Capra A. and Fontana S. Illusions in Sound, 22nd Audio Engineering Society-United Kingdom (AES-UK) Conference, Cambridge, UK (2007).

[14] A. Inoue, Yatabe K., Oikawa Y. and Ikeda Y.ACM SIGGRAPH 2017 Posters, Los Angeles (2017), pp. 1–2.

[15] Cecchi S., Carini A. and Spors S., Appl. Sci. **8** (1), 16 (2018).

[16] Flanagin V.L., Schörnich S., Schranner M., Hummel N., Wallmeier L., Wahlberg M., Stephan T. and Wiegrebe L., J. Neurosci. **37** (6), 1614–1627 (2017).

[17] Fabre T., Verhulst A., Balandra A., Sugimoto M. and Inami M. 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). Puglia, Italy (2021), pp. 320–328.

[18] Andreasen A., Geronazzo M., Nilsson N.C., Zovnercuka J., Konovalov K. and Serafin S., IEEE. Trans. Vis. Comput. Graph. **25** (5), 1876–1886 (2019).

[19] Owens A., Wu J., McDermott J.H., Freeman W.T. and Torralba A., European Conference on Computer Vision, Amsterdam, The Netherlands (2016), pp. 801–816.

[20] Chen Z., Hu X. and Owens A. Conference on Robot Learning (CoRL), London, UK. 2021.

[21] Gao R., Chen C., Al-Halah Z., Schissler C. and Grauman K., European Conference on Computer Vision. Glasgow, UK (2020), pp. 658–676.

[22] Colombo M. IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), 2021, pp. 487–490.

[23] Chen K., Lai Y.K. and Hu S.M., Comput. Visual Media **1** (4), 267–278 (2015).

[24] Dahnert M., Hou J., Nießner M. and Dai A., Adv. Neural. Inf. Process. Syst. **34**, 8282–8293 (2021).

[25] Weiss T., Nakada M. and Terzopoulos D., Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, Hawaii, USA (2017), pp. 41–47.

[26] Liu C., Schwing A.G., Kundu K., Urtasun R. and Fidler S., Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition, Boston, MA (2015), pp. 3413–3421.
[27] Choe J., Im S., Rameau F., Kang M. and Kweon I.S., Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada (2021), pp. 16086–16095.
[28] Wu Y.C., Chan L. and Lin W.C., 2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Beijing, China (2019), pp. 26–36.
[29] Beigi M.M. and Zell A., 2008 IEEE International Conference on Robotics and Automation, Pasadena, CA (2008), pp. 3270–3275.
[30] Frank N., Wolf L., Olshansky D., Boonman A. and Yovel Y., 2020 IEEE International Conference on Computational Photography (ICCP), Cluj-Napoca, Romania (2020), pp. 1–12.
[31] Purushwalkam S., Gari S.V.A., Ithapu V.K., Schissler C., Robinson P., Gupta A. and Grauman K., Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy (2020), pp. 1183–1192.
[32] Dokmanić I., Lu Y.M. and Vetterli M., 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic (2011), pp. 321–324.
[33] Kuster M., J. Acoust. Soc. Am. **124** (2), 982–993 (2008).
[34] Antonacci F., Filos J., Thomas M.R., Habets E.A., Sarti A., Naylor P.A. and Tubaro S., IEEE. Trans. Audio. Speech. Lang. Process.**20** (10), 2683–2695 (2012).
[35] Antonacci F., Sarti A. and Tubaro S., 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX (2010), pp. 2822–2825.
[36] Tervo S. and Tossavainen T., 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan (2012), pp. 513–516.
[37] Moore A.H., Brookes M. and Naylor P.A., 21st European Signal Processing Conference (EUSIPCO 2013), Marrakech, Morocco (2013), pp. 1–5.
[38] Markovic D., Antonacci F., Sarti A. and Tubaro S., 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY (2013), pp. 1–4.
[39] Genovese A.F., Gamper H., Pulkki V., Raghuvanshi N. and Tashev I.J., ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK (2019), pp. 231–235.
[40] Yu W. and Kleijn W.B., IEEE/ACM. Trans. Audio. Speech. Lang. Process. **29**, 436–447 (2021).
[41] Morse P.M. and Ingard K.U., *Theoretical Acoustics* (Princeton University Press, New Jersey, 1986).
[42] Kuttruff H., *Room Acoustics* (CRC Press, Boca Raton, 2016).
[43] Moore B.C., *An introduction to the psychology of hearing* (Brill, Leiden, 2012).
[44] Khan S., Rahmani H., Shah S.A.A. and Bennamoun M., Synth. Lect. Comput. Vision **8** (1), 1–207 (2018).
[45] Logan B., International Symposium on Music Information Retrieval, Plymouth, MA (2000).
[46] Grondin F., Lauzon J.S., Michaud S., Ravanelli M. and Michaud F., arXiv preprint arXiv:2010.09930 (2020).
[47] Allen J.B. and Berkley D.A., J. Acoust. Soc. Am.**65** (4), 943–950 (1979).
[48] Kingma D.P. and Ba J., arXiv preprint arXiv:1412.6980 (2014).
[49] Chollet F., Keras **7** (8), T1 (2015). https://keras.io/k.