Dynamic Sound Field Synthesis for Speech and Music Optimization

Zhenyu Tang
University of North Carolina-Chapel Hill
Department of Computer Science
zytang@cs.unc.edu

Nicolas Morales
University of North Carolina-Chapel Hill
Department of Computer Science
nmorales@cs.unc.edu

Dinesh Manocha
University of Maryland
Department of Computer Science
Department of Electrical & Computer Engineering
dm@cs.umd.edu

http://gamma.cs.unc.edu/SoundField/

ABSTRACT

We present a novel acoustic optimization algorithm to synthesize dynamic sound fields in a static scene. Our approach places new active loudspeakers or virtual sources in the scene so that the dynamic sound field in a region satisfies optimization criteria to improve speech and music perception. We use a frequency domain formulation of sound propagation and reduce the computation of dynamic sound field synthesis to solving a linear least squares problem, and do not impose any constraints on the environment or loudspeakers type, or loudspeaker placement. We highlight the performance on complex indoor scenes in terms of speech and music improvements. We evaluate the performance with a user study and highlight the perceptual benefits for virtual reality and multimedia applications.

KEYWORDS

Sound propagation; acoustic optimization; virtual environments; speech improvement; music reinforcement

ACM Reference Format:
Zhenyu Tang, Nicolas Morales, and Dinesh Manocha. 2018. Dynamic Sound Field Synthesis for Speech and Music Optimization. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240644

1 INTRODUCTION

Recreating an immersive environment that combines both video and audio rendering to simulate the experience of exploring a three-dimensional virtual environment is important for games, virtual/augmented reality (VR/AR), and multimedia applications. Over the last few decades, most of the work has focused on improving the visual fidelity of such environments using multimedia techniques or high quality graphical rendering. Current 3D multimedia or VR content creation tools can generate photo-realistic rendering and also provide capabilities for automatic placement of real or virtual lights. As compared to visual rendering, the state of the art in audio rendering or generation of 3D audio content lags behind. We need better capabilities in terms of algorithms and tools to automatically generate desirable sound fields in virtual environments.

The notion of generating or modifying the sound field is widely studied in the context of sound field synthesis (SFS) for decades [28, 41]. The SFS problem can be formulated as finding the driving signal of a given ensemble of elementary sound sources (usually loudspeakers) such that the superposition of their emitted individual sound fields constitutes a common sound field with given desired properties over an extended area [1]. This problem calls for a new reproduction technique which allows the synthesis of physically correct wave fields of three-dimensional acoustic scenes. Previous works include many audio rendering techniques, where new and artificial wavefronts are synthesized by a large number of active loudspeakers or virtual sources in the environment. The most widely used methods are based on wave-field synthesis, which is based on the Huygens-Fresnel principle, and deals with the use of loudspeaker arrays to control the sound field over an extended area of the environment. In practice, prior methods do not accurately model sound wave propagation or generate reverberation effects. As a result, it is hard to provide guarantees on the performance of current sound field synthesis methods in arbitrary environments.

The sound field is governed by various factors or scene parameters. These include the geometric shape and material representation of the 3D virtual world, the location(s) of audio source(s), the listener location, and input (dry) audio signal(s). Acoustic propagation algorithms simulate the propagation of sound waves through an environment for given source and listener positions and compute the impulse responses (IRs) using geometric or wave-based propagation algorithms. Recent developments in sound propagation and auralization have enabled the generation of environmental acoustic effects and spatial sound at interactive rates for immersive environments. These methods are also used to provide aural cues to a user about the environment and can lead to an improved sense of presence in VR applications [19, 39]. Although these propagation techniques can evaluate the sound field for a given scene configuration or parameters, they have not been used to actively modify or change the sound field to generate desired acoustic effects.

One of the driving applications of our work is to develop techniques that can improve the understanding or perception of speech or music effects. SFS has been shown to be useful in recreating

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240644

acoustic environments for communication (e.g. teleconferencing) and in the entertainment industry based on digital signal processing. With the recent advances in speech understanding and use of voice interfaces for IOT (Internet of Things) devices, there is considerable interest in developing robust SFS methods that can improve the intelligibility of speech in noisy/reverberant environments. Similarly, there is some work on improving the quality of music sound in all types of acoustic environments.

Main Results: We present a novel algorithm for dynamically synthesizing the sound field using a combination of sound propagation and acoustic optimization. Given a static virtual environment with known sound sources, our approach automatically computes the driving signals for a set of active loudspeakers to generate the desired dynamic sound field. We use the frequency domain formulation of the acoustic wave equation and sound propagation, and reduce the dynamic SFS problem to solving a linear least-square system. Our algorithm uses precomputed IRs in the virtual scene for different locations of active loudspeakers. Our approach is general and makes no assumption about the environment, sound sources or their locations. Furthermore, we can provide guarantees on the resulting sound field based on our acoustic optimization. We demonstrate the benefits of our approach in two driving applications:

- **Speech Improvement:** We dynamically synthesize the sound field for speech improvement in indoor scenes. We use the well-known speech transmission index (STI) metric [16] as an indicator to reduce the reverberation effects in an environment using virtual sources. This can be used to improve the quality of far-field speech intelligibility for automated speech recognition (ASR).
- **Music Reinforcement:** We use our dynamic sound field synthesis algorithm for music reinforcement to maintain a desirable frequency transmission in an acoustic environment. Our formulation computes the appropriate frequency component transmission compensation for sources and minimizes the unwanted frequency distortions due to sound propagation in poorly treated acoustic environments.

We have evaluated our algorithm in different indoor scenes. We use ray tracing based geometric propagation algorithm to accurately compute the IRs and combine them with optimization algorithms. We highlight the improvements in the sound fields based on different metrics for speech and music improvement. Overall, we present first set of dynamic SFS algorithms that use sound simulation techniques to modify the sound field to satisfy given metrics.

The rest of the paper is organized as follows. We give an overview of prior work in sound field synthesis and acoustic optimization in Section 2. We introduce our notation and describe the underlying representation used to synthesize the sound field in Section 3. We present our dynamic synthesis algorithm in Section 4 along with the metrics used for speech improvement and music reinforcement. We describe our implementation in Section 5, and highlight the performance on different benchmarks, as well as a perceptual user study in Section 6.

2 RELATED WORK

In this section, we give a brief overview of prior work on SFS, sound propagation, and acoustic optimization.

2.1 Sound Field Synthesis

SFS deals with generating a defined sound field in an extended area that is surrounded by loudspeakers. The idea of SFS was first introduced by Jessel [28] and based on the theoretical assumption of a continuous layer of loudspeakers. In early work on *Ambisonics* [20], several loudspeakers were placed around one location where the sound field was synthesized. This early work on Ambisonics systems evolved into Higher Order Ambisonics (HOA), which accounts for higher order modes [4, 11], and Near-field Compensated Higher Order Ambisonics (NFC-HOA) [11, 12], where sources are assumed to be monopoles that emit omnidirectional waves [35]. One of the limitations of Ambisonics systems is that they can only be used with spherical and circular loudspeaker distributions.

Wave Field Synthesis (WFS) is another popular approach for SFS. It can be regarded as an audio rendering method where the wavefronts originate from virtual sources. Its formulation can be derived from the Rayleigh I integral or the Kirchhoff-Helmholtz integral [6, 41]. While WFS is equivalent to a high-frequency approximation of infinite order HOA [2], it can be applied to any arbitrary convex loudspeaker placement problem. However, dense loudspeaker spacing and the loudspeaker type (monopole, dipole, or linear array type) [14] are needed to compute a solution. Other techniques are based on digital signal processing [26, 27]. In these methods, sound pressure at certain frequencies is matched with the desired sound field by solving for loudspeaker driving signals using least squares techniques. However, the resulting algorithms do not accurately model the room acoustics or reverberation effects in an environment. While complimentary technique of using compensation filters exists [9], it requires a high number of filters (L^2 filters for L loudspeakers) for non-stationary virtual sources as well as tedious measurements.

Our approach for dynamic SFS is complimentary to these techniques and is more general. We try to model the room acoustics using sound propagation and precompute the IRs in preprocessing. Furthermore, we do not impose any constraints on the environment, loudspeaker type, or loudspeaker placement.

2.2 Sound Propagation

Sound propagation methods compute the reflection and diffraction paths from the sound sources to a listener in the virtual environment. Prior algorithms for sound propagation can be classified into two categories: geometric techniques and wave-based techniques. Geometric methods work on the underlying assumption of sound wave propagating in the form of a ray, where the wavelength of the sound is smaller than the size of the obstacles in the environment [19]. These methods include image source methods [3], ray tracing methods [38–40, 43], and beam or frustum tracing methods [8, 18]. They are mostly accurate for higher frequencies and can be used for interactive applications. Wave-based sound propagation methods directly solve the wave equation for sound propagation (see Equation 1). These methods are based on Finite Element Methods (FEM), Boundary Element Methods (BEM), finite-difference time domain (FDTD) approaches [36] and Adaptive Rectangular Decomposition (ARD) methods [32]. Wave-based techniques are accurate, but are only practical for low frequencies and

small scenes. When combining geometric and wave-based methods, a huge amount of precomputation would be required for each scattering object [34].

2.3 Acoustic Optimization

Acoustic optimization methods mainly deal with improving the acoustic characteristics of a space using optimization algorithms by changing the scene parameters. Previous work in acoustic optimization techniques includes modifications to the shape, materials, or topology of the 3D environment. Work in this area includes *Audioptimization*, a framework for optimizing the shape and materials [30], absorbent optimization [37], continuous optimization approaches for material design [31], shape optimization approaches [17, 33], and topology optimization approaches [15]. In the area of SFS, different methods have been proposed for the placement of loudspeakers and microphones [25]. SFS methods are also useful in the field of noise control [42, 44]. Our optimization approach is more general and is complimentary to these acoustic optimization algorithms and uses sound propagation algorithms.

3 SOUND FIELDS

Table 1: Notation and symbols used throughout the paper.

$P(\mathbf{x}, t)$	Sound pressure at point $\mathbf{x}$ at time t
$S(\mathbf{x}, \omega)$	Sound pressure at point $\mathbf{x}$ of angular frequency ω
$D(\mathbf{x}, \omega)$	Driving signal at point $\mathbf{x}$ of angular frequency ω
$\mathrm{IR}(\mathbf{x}_s, \mathbf{x}_l, t)$	Impulse response at time t from point source at Location $\mathbf{x}_s$ to the listener at Location $\mathbf{x}_l$
$G(\mathbf{x}_s, \mathbf{x}_l, \omega)$	Frequency response at ω from point source at Location $\mathbf{x}_s$ to the listener at Location $\mathbf{x}_l$

In this section, we give background on properties of sound field and on sound propagation. These are used in our approach to perform dynamic SFS in an uncontrolled environment with arbitrary loudspeaker distribution.

3.1 Sound Field as 4D Pressure Field

A sound field is defined in a spatial volume $V \subset \mathbb{R}^3$ that has no sources or sinks. Moreover, we assume that sound sources are placed outside the volume, as in prior work in sound field synthesis. For a listener at location $\mathbf{x}_l \in V$ and a source placed at $\mathbf{x}_s$, the sound pressure at time t at the listener induced by the source is denoted by $P(\mathbf{x}_s, \mathbf{x}_l, t)$, which is a 7D pressure field. However, when dealing with multiple sound sources, our goal is to compute the combined sound field at the listener from all the sources. Therefore, we fix $\mathbf{x}_l$ and sum up the sound from all the sources, yielding $P(\mathbf{x}_l, t) = \sum_s P(\mathbf{x}_s, \mathbf{x}_l, t)$, which characterizes our sound field as a 4D pressure field.

3.2 Sound Propagation in Frequency Domain

The process of sound propagation can be described using the wave equation:

$$\frac{\partial^2}{\partial t^2} P(\mathbf{x}, t) - c^2 P(\mathbf{x}, t) = f(\mathbf{x}, t), \tag{1}$$

where c is the speed of sound in a homogeneous medium, which we assume to be $343 m/s$, and $f(\mathbf{x}, t)$ is the forcing term at location $\mathbf{x}$ at time t.

Impulse Response (IR) is the most widely used representation to model sound propagation in the time domain. In practice, an IR sequence is convolved with the source signal sequence to compute the propagated signal and auralization. In this work, we mainly work with the complex frequency domain. Let $\mathrm{IR}(\mathbf{x}_s, \mathbf{x}_l, t)$ denote the IR for source-listener pair $\mathbf{x}_s$ and $\mathbf{x}_l$, the frequency response is the Fourier transform of the IR:

$$G(\mathbf{x}_s, \mathbf{x}_l, \omega) = \mathcal{F}\{\mathrm{IR}(\mathbf{x}_s, \mathbf{x}_l, t)\} = \int_{-\infty}^{\infty} \mathrm{IR}(\mathbf{x}_s, \mathbf{x}_l, t) e^{-i\omega t} \mathrm{d}t, \tag{2}$$

where ω is evenly discretized in a frequency range. Similarly, we transform the sound field to the frequency domain as: $S(\mathbf{x}, \omega) = \mathcal{F}\{P(\mathbf{x}, t)\}$. Let the source signal at location $\mathbf{x}_s$ be given as $D(\mathbf{x}_s, \omega)$, then the propagated sound from all known sources to $\mathbf{x}_l$ can be represented as:

$$S(\mathbf{x}_l, \omega) = \sum_s D(\mathbf{x}_s, \omega) G(\mathbf{x}_s, \mathbf{x}_l, \omega), \tag{3}$$

for all ω in our interested frequency range. This is converted using $\omega = 2\pi f$, with f normally taken from a subset of the human hearing range 20Hz $\sim$ 20000Hz.

3.3 Dynamic Sound Fields

There are many factors that affect the steady state of a sound field, and therefore making the field change continuously. These include: source movement - a change in source location results in changes in sound propagation paths; 3D environment change - when dynamic objects are present in the scene (e.g. a door that might be open or closed) or a change in the environment material, the sound field can also change; source signal change - fast changing source signals will make the sound field more dynamic. In addition, when a source signal becomes zero, it contributes nothing to the sound field and is equivalent to being removed from the system. In this paper, we limit ourselves to static scenes with fixed source locations. We mostly account for changes in the sound field due to the source signal change. Specifically, we place active loudspeaker or virtual sources outside V, which can change the sound field inside V based on certain metrics or criteria. Our two main metrics are based on speech intelligibility and music reinforcement requirements.

4 DYNAMIC SOUND FIELD SYNTHESIS

In this section, we present our dynamic sound field synthesis algorithm in its generalized form and reduce it to an optimization problem. Moreover, we demonstrate the applications of our formulation to two driving applications: speech improvement and music reinforcement. Given an acoustic scene that has some existing static sound sources, we add new loudspeakers that emit constructive or destructive sound signals at multiple locations to change the *existing or original sound field* to *a new sound field* we desire. In the following context, we call these newly added loudspeakers as *active loudspeakers* because they are actively driven by our algorithm.

4.1 Problem Formulation

Given a sound zone $V \subset \mathbb{R}^3$, some known sound sources, and a set of active loudspeakers with known positions, we compute the driving signal for each individual loudspeaker so that the superposition of

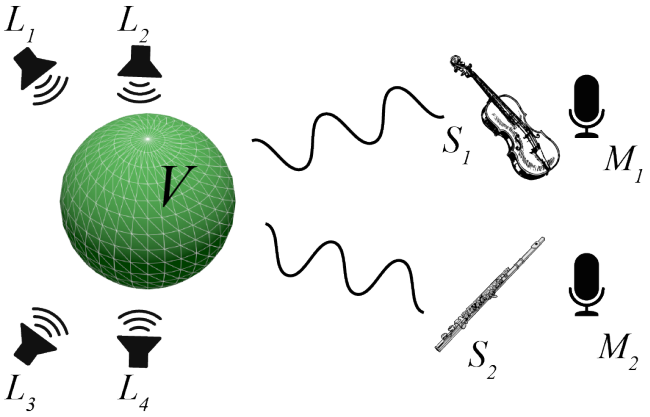


Figure 1: Given a scene with two static sources, S_1 and S_2. Our algorithm manipulates the sound field within V by controlling the source signals at 4 active loudspeakers, L_i.

all propagated signals constitutes a desired sound field S^d over V. Such a setup is also illustrated in Figure 1.

Essentially, without the active loudspeakers, there is only the sound field produced by original sources in the scene. Assume we have N_S known original sources at $\mathbf{x}_s \notin V, s \in \{1, ..., N_S\}$, with $D(\mathbf{x}_s, \omega)$ being the emitted signal at $\mathbf{x}_s$ which can be a dynamic function, the resulting sound pressure at any position $\mathbf{x} \in V$ can be expressed as:

$$S^o(\mathbf{x}, \omega) = \sum_{s=1}^{N_S} D(\mathbf{x}_s, \omega) G(\mathbf{x}_s, \mathbf{x}, \omega). \tag{4}$$

We use S^o to denote the *original sound field*. Equation 4 can be compactly written as:

$$S^o(\mathbf{x}, \omega) = \mathbf{g}^{\mathrm{T}}(\omega; \mathbf{x}) \mathbf{D}(\omega), \tag{5}$$

where $\mathbf{g}(\omega; \mathbf{x}) = [G(\mathbf{x}_1, \mathbf{x}, \omega), ..., G(\mathbf{x}_{N_S}, \mathbf{x}, \omega)]^{\mathrm{T}}$ and $\mathbf{D}(\omega) = [D(\mathbf{x}_1, \omega), ..., D(\mathbf{x}_{N_S}, \omega)]^{\mathrm{T}}$, which are both $N_S \times 1$ complex column vectors.

Next, assume we have N_L active loudspeakers (or virtual sources) at $\mathbf{y}_l \notin V, l \in \{1, ..., N_L\}$, with $D(\mathbf{y}_l, \omega)$ being the emitted signal at $\mathbf{y}_l$ which is driven by our algorithm, the sound field constructed by all active loudspeakers denoted by S^a can be expressed as:

$$S^a(\mathbf{x}, \omega) = \sum_{l=1}^{N_L} D(\mathbf{y}_l, \omega) G(\mathbf{y}_l, \mathbf{x}, \omega), \tag{6}$$

at $\mathbf{x} \in V$. As in Equation 5, we rewrite Equation 6 as:

$$S^a(\mathbf{x}, \omega) = \tilde{\mathbf{g}}^{\mathrm{T}}(\omega; \mathbf{x}) \tilde{\mathbf{D}}(\omega), \tag{7}$$

where $\tilde{\mathbf{g}}(\omega; \mathbf{x}) = [G(\mathbf{y}_1, \mathbf{x}, \omega), ..., G(\mathbf{y}_{N_L}, \mathbf{x}, \omega)]^{\mathrm{T}}$ and $\tilde{\mathbf{D}}(\omega) = [D(\mathbf{y}_1, \omega), ..., D(\mathbf{y}_{N_L}, \omega)]^{\mathrm{T}}$. Finally, we can directly sum up (5) and (7) to get the combined sound field. If our desired or new sound field is $S^d(\mathbf{x}, \omega)$, we want to compute $\tilde{\mathbf{D}}(\omega)$ such that $S^o(\mathbf{x}, \omega) + S^a(\mathbf{x}, \omega) = S^d(\mathbf{x}, \omega)$.

4.2 Sound Field Synthesis: Objective

Our goal is to manipulate the continuous sound field. However, the stated problem cannot be solved analytically. Therefore, we instead select N_M uniformly distributed internal monitor points $\mathbf{p}_m \in V, m \in \{1, ..., N_M\}$, and make the sound field match the desired one at these monitor points, indirectly constraining the continuous sound field. The selection of these monitor points can be based on other principles. To simplify our formulation, we define

$$\begin{aligned} C_m(\omega) &= S^d(\mathbf{p}_m, \omega) - \mathbf{g}^{\mathrm{T}}(\omega; \mathbf{p}_m) \mathbf{D}(\omega), \\ f(X_m, \tilde{\mathbf{D}}(\omega)) &= \tilde{\mathbf{g}}^{\mathrm{T}}(\omega; \mathbf{p}_m) \tilde{\mathbf{D}}(\omega). \end{aligned} \tag{8}$$

This boils down to solving the optimization problem that minimizes the error between our constructed and desired sound fields by choosing the appropriate driving signals. The resulting objective function can be given as:

$$\arg\min_{\tilde{\mathbf{D}}(\omega)} \sum_{m=1}^{N_M} \left[C_m(\omega) - f(X_m, \tilde{\mathbf{D}}(\omega)) \right]^2. \tag{9}$$

4.3 General Solution

Equation (9) can be solved using linear least squares. Since $\tilde{\mathbf{D}}(\omega)$ is an unknown complex vector of length N_L, and we have N_M observations, depending on the relative values of N_L and N_M. Given the linear dependency between active loudspeaker responses, the resulting linear system could be determined, over-determined or under-determined. To deal with the numeric instability of sound propagation algorithms, we tend to choose more loudspeakers than the monitor points. Thus, we turn our linear system into an over-determined system by setting $N_L > N_M$. Moreover, we use ridge regression to enforce a meaningful solution. Let us define the $N_M \times N_L$ frequency response matrix for all loudspeakers

$$\mathbf{Q}(\omega) = \begin{bmatrix} G(\mathbf{y}_1, \mathbf{p}_1, \omega) & \dots & G(\mathbf{y}_{N_L}, \mathbf{p}_1, \omega) \\ \vdots & \ddots & \vdots \\ G(\mathbf{y}_1, \mathbf{p}_{N_M}, \omega) & \dots & G(\mathbf{y}_{N_L}, \mathbf{p}_{N_M}, \omega) \end{bmatrix}, \tag{10}$$

and $\mathbf{C}(\omega) = [C_1(\omega), ..., C_{N_M}(\omega)]^{\mathrm{T}}$. For brevity we omit ω and derive the optimal solution in the least-squares sense as:

$$\tilde{\mathbf{D}} = (\overline{\mathbf{Q}}^{\mathrm{T}} \mathbf{Q} + \lambda \mathbf{I})^{-1} \overline{\mathbf{Q}}^{\mathrm{T}} \mathbf{C}, \tag{11}$$

where $\overline{(\cdot)}$ denotes the complex conjugate of matrices and $\mathbf{I}$ is an identity matrix in the complex domain. The regularization weight λ is typically decided from the experiments or the 3D environment. The regularization term is helpful in constraining the absolute loudspeaker power and making the solution more robust. Note that the right side of Equation (11) can be decoupled so that $(\overline{\mathbf{Q}}^{\mathrm{T}} \mathbf{Q} + \lambda \mathbf{I})^{-1} \overline{\mathbf{Q}}^{\mathrm{T}}$ should only be solved once for the system, and only the observation part $\mathbf{C}$ needs to be updated for specific applications.

4.4 Dynamic SFS for Speech Improvement

One of the driving applications of our work is to improve the speech understandability in an indoor scene. Human speech understanding has been an important task for some smart devices that use Automated Speech Recognition (ASR) [5, 23]. In an indoor environment, even without the presence of mechanical noise, reverberation of the speech signal itself can negatively affect the understanding of spoken phrases [21].

Our formulation is based on the observation that reducing reverberation in the environment can improve the speech intelligibility.

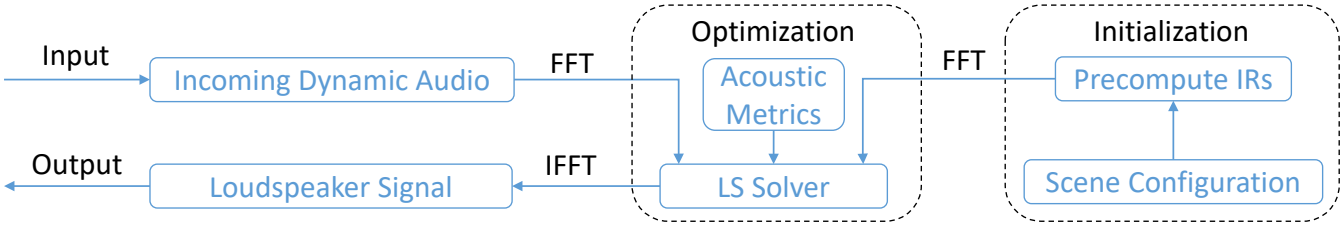


Figure 2: We highlight different stages of our algorithm. The acoustic metric is given by the underlying application. The driving function for loudspeakers are solved from linear systems using complex regularized least-squares (LS).

By using sound field synthesis (equivalently adding virtual sources), we can significantly reduce the reverberation of speech. We are given a 3D environment along with the location of the sound speech sources. Therefore, for a monitor point $\mathbf{p}_m$ in our target sound zone in the 3D environment and a speech signal from $\mathbf{x}_s$, our goal is to model only the direct response and denote it as $G_D(\mathbf{x}_s, \mathbf{p}_m, \omega)$, which only contains the first impulse of $G(\mathbf{x}_s, \mathbf{p}_m, \omega)$. And this impulse can be easily located in the temporal domain. In this case, the desired sound field becomes:

$$S^d(\mathbf{p}_m, \omega) = \sum_{s=1}^{N_S} D(\mathbf{x}_s, \omega) G_D(\mathbf{x}_s, \mathbf{p}_m, \omega). \tag{12}$$

Typically we expect only one of the N_S sources to emit a non-zero signal because it is difficult for someone to listen to two different speech signals at the same time, even if both are very clear. Therefore, we can substitute Equation (12) into Equation (8) and solve for the resulting system.

4.5 Dynamic SFS for Music Reinforcement

A music sound reinforcement system often uses loudspeakers, signal processors, equalizers and amplifiers to distribute live or prerecorded music to the audience. These systems are more sophisticated than modern stereo sound systems at home, and require the user to have a higher level understanding of acoustical signal characteristics to operate [13]. In live music performance, even though the soundtracks are mixed by an expert, as the sound propagates in the environment, the resulting soundtrack tends to experience distortion in its frequencies [10]. In many cases, high frequency signals are attenuated more than low frequency signals. With our dynamic sound field synthesis, we can simulate the propagation effect the environment has on the resulting music soundtrack and negate the distortion. By using sound field synthesis in music reinforcement systems, we can control the transmission of spatial music sound with higher precision.

We use a stage setting to demonstrate the benefits of our approach. During a music performance, input sound streams are captured with one microphone per performer/instrument. Our loudspeakers are located around the ceiling. We want the sound perceived by the audience to have no undesired distortions due to propagation. Therefore, we set the filtered sound field as our desired sound field at each monitor position $\mathbf{p}_m$:

$$S^d(\mathbf{p}_m, \omega) = \sum_{s=1}^{N_S} D(\mathbf{x}_s, \omega) F(\mathbf{x}_s, \omega), \tag{13}$$

where $\mathbf{x}_s$ represents the location of one performer on stage, and F is the frequency dependent filter as tuned by a sound expert for each audio stream. Equation (13) is substituted back into Equation (8) to compute the solution.

4.6 Performance Metrics

We introduce two commonly used metrics we will use in following sections as our metrics for speech and music tasks.

4.6.1 Speech Metric. To measure speech intelligibility quantitatively, we use the STI metric [16] to evaluate the performance. STI is computed from a weighted average of the Modulation Transfer Function (MTF) of an impulse response. MTF can be derived as:

$$m_k(f_m) = \frac{|\int_0^\infty r_k(t)^2 e^{-j2\pi f_m t} dt|}{\int_0^\infty r_k(t)^2 dt}, \tag{14}$$

where $r_k(t)$ is our impulse response filtered to octave band k. The left hand side $m_k(f_m)$ is the modulation transfer ratio at f_m. For evaluating the STI in full range, we use 14 modulation frequencies (0.63Hz to 12.5Hz, 1/3 octave spaced) per band, which gives us 98 samples of $m_k(f_m)$. The STI value is bounded within $[0, 1]$. Larger STI values indicate better speech intelligibility.

4.6.2 Music Metric. To measure the effectiveness of music reinforcement at each listening position, we evaluate the normalized cross-correlation between the actual and desired sound field to evaluate the effect of distortion compensation. Assuming we have obtained the propagated sound field S^p from $S^p(\mathbf{x}, \omega) = S^o(\mathbf{x}, \omega) + S^a(\mathbf{x}, \omega)$, the correlation can be computed as

$$corr(S^p, S^d) = \frac{\sum_\omega \overline{S}^p(\mathbf{x}, \omega) S^d(\mathbf{x}, \omega)}{\sqrt{\sum_\omega |S^p|^2 \sum_\omega |S^d|^2}}. \tag{15}$$

The correlation value will be in the range $[-1, 1]$ and the larger the absolute correlation is, the better our propagated music frequencies match with the desired one.

5 IMPLEMENTATION

In this section, we give details of our implementation. Figure 2 shows our algorithm pipeline, which is explained in detail below.

5.1 Acoustic Scene Configuration

The input to our algorithm is an acoustic scene configuration. A complete scene configuration includes: the acoustic materials of each object in the scene, the geometry of the scene as a 3-D mesh, the locations of loudspeakers as 3D coordinates, and the desired sound field. The first two components are treated as fixed properties of the environment. As indicated in Section 4.3, the locations of loudspeakers have some freedom over the space, so they can

simply be placed at convenient locations near the target region. The computation of the desired sound field depends on the specific application, and we highlight different scenarios in Section 6. For example, we use different metrics for speech improvement and music reinforcement detailed in Section 4.6.

5.2 Monitor Point Sampling & Precomputation

We generate a set of monitor points by uniformly sampling the target sound zone in 3D according to any weighted or probabilistic distribution. This yields monitor points $\mathbf{p}_1, ..., \mathbf{p}_{N_M}$ described in Section 4.2. Then the impulse responses between all pairs of monitor and loudspeaker locations (i.e. $N_M \times N_L$ pairs) are computed using a sound propagation algorithm and subsequently converted to frequency responses. In our current implementation, we use a ray tracing based geometric propagation algorithm. It traces specular and diffuse rays [43] and performs up to 200 bounces to accurately compute the reverberation effects. To approximate low frequency diffraction effects, we model first order diffraction based on the Uniform Theory of Diffraction [39]. Since these computations are performed as a preprocess, we use a sufficient number of ray samples (e.g., 10K) to compute accurate IRs. We use these IRs to compute the solution using our optimization algorithm described in Section 4. We parallelize these computations on a cluster and it can take a few hours for each scene to compute these large number of IRs, depending on the size and complexity of the scene.

5.3 Real-time Computation of Sound Fields

Since our algorithm deals with dynamic sound fields generated using active loudspeakers, we need to monitor and handle existing sources in the scene in real-time. Temporal signals are treated as discrete temporal sound pressure sequences $P(\mathbf{x}, t)$. Because the monitored signal sequences might be very long, we need to segment these signals based on the sampling rate and allowed delay time before processing. For convenience of implementation, we segment any sequence according to our fixed sampling rate 44.1kHz, which is beyond the Nyquist frequency regarding the human hearing range of 20Hz $\sim$ 20kHz. And we perform short-time Fourier transform (STFT) for each segment of length 65536. Note that the segment length can be arbitrary, depending on the allowable processing delay. At each processing step, the optimization problem is formulated, shown as Equation (9), and we solve for a segment. Moreover, we set the active loudspeakers or virtual sources as their driving functions, while the next segment is being prepared. The complex regularized least squares problem in Equation (11) is efficiently solved using the *Eigen* library [22]. In this way we can achieve real-time processing rate for any scene under stable sensing.

6 RESULTS AND ANALYSIS

In this section, we evaluate the performance of our dynamic sound field manipulation algorithm for the two applications described in Section 4. We also demonstrate how the desired sound field is computed based on these scenarios and the metrics. Given the input scene, we do not make any changes to the environment in terms of object positions or the underlying materials. Our goal is to add more virtual sound sources to the environment so that we can change the sound field in a given region.

Table 2: Improvements on the STI metric (Sec 4.6.1) ranging from $[0, 1]$, and a larger value indicates better quality. We observe considerable improvements in the resulting sound fields corresponding to speech sources.

Scene	Number of Loudspeakers (N_L)	Number of Monitors (N_M)	Average STI	
			Before	After
Trinity	14	12	0.525	0.734
Berlin	11	10	0.602	0.724

Table 3: Improvements on the correlation metric (Sec 4.6.2) ranging from $[-1, 1]$, and a larger absolute value indicates better quality. We observe considerable improvements in the resulting sound fields corresponding to music sources.

Scene	Number of Loudspeakers (N_L)	Number of Monitors (N_M)	Average Correlation	
			Before	After
Elmia	14	12	0.039	0.786
Sibenik	12	12	0.040	0.786

6.1 Benchmarks

We used four different 3D environments to evaluate our algorithm. Detailed parameters and results are shown in Table 2 and 3.

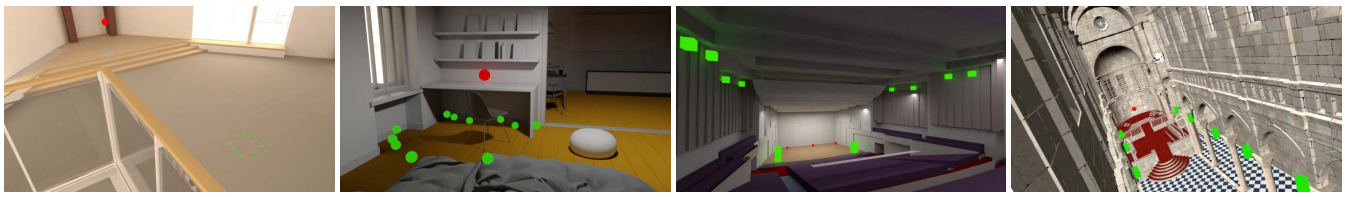
- The Trinity scene comes from direct measurement of a real architecture. This environment (Figure 3(a)) has a long reverberation time, which can considerably affect the speech intelligibility especially when the listener is far from the source. In this scene, a speech sound source in placed on the stage, corresponding to a talking human voice. The listener is assumed to be 10 meters away from that source.
- The Berlin scene (Figure 3(b)) corresponds to a small apartment complex. In this scene, a noise source is placed in the room and a listener is set above the bed in the same room.
- The Elmia scene (Figure 3(c)) is a concert hall with measured acoustic material properties that matches to a real-world scene. In this benchmark, we assume that a live music show is played on the stage, and loudspeakers installed across the hall are used to amplify the music played at stage.
- The Sibenik scene (Figure 3(d)) is modeled from the real-world Sibenik Cathedral. Material properties are mannually assigned to the model. In this scene, a piano is played on the stage in the cathedral and the sound undergo certain distortion in its frequency. Loudspeakers are installed on pillars in the cathedral.

6.2 Speech Improvement

Figure 4 shows the distribution of STI values before and after our optimization algorithm on the Trinity benchmark. Before our optimization, the reverberation effect is significant and the average STI value is 0.525. After we reduce the reverberation using active loudspeakers, the average STI becomes 0.734, whereas human's just noticeable difference (JND) for STI is 0.03 [7]. A higher value of STI indicates higher quality of speech understanding or intelligibility.

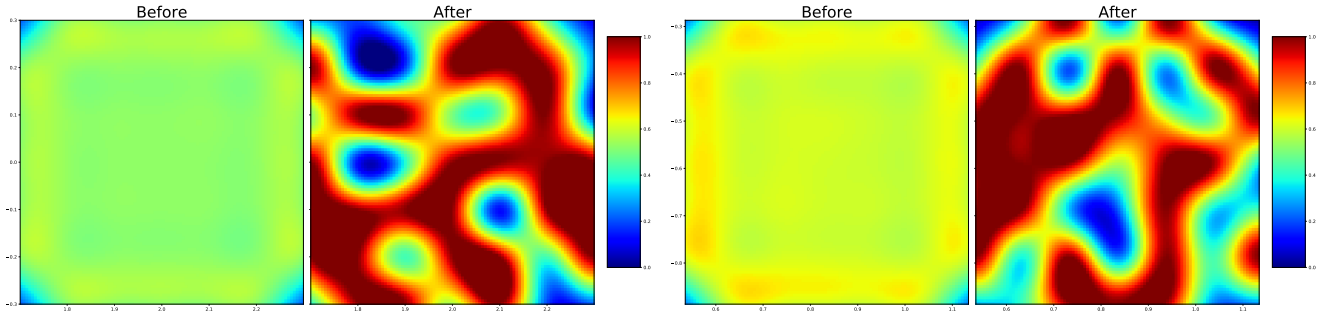
6.3 Music Reinforcement

In the Elmia benchmark, the desired sound field corresponds to the field generated by propagating the music signal from the performer



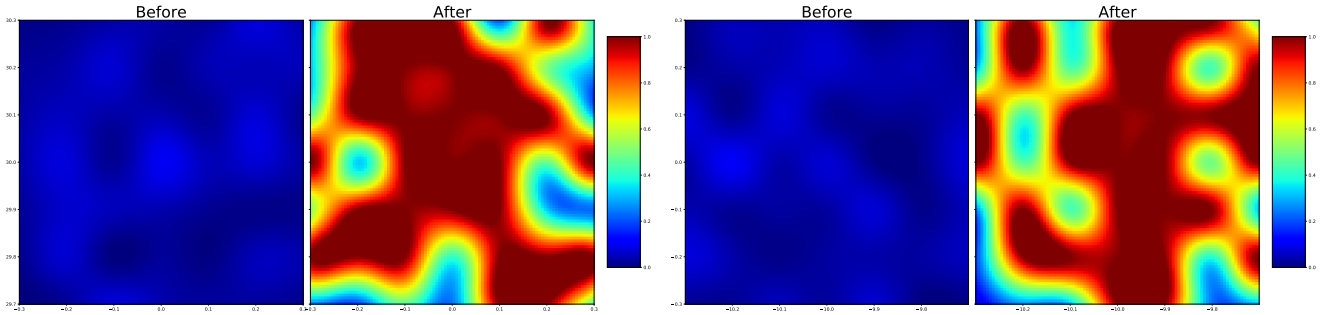
(a) Trinity (b) Berlin (c) Elmia (d) Sibenik

Figure 3: Different benchmarks used to evaluate our dynamic SFS algorithm. We highlight the 3D CAD models with colored sound source and loudspeaker placement: the green spheres represent active loudspeakers; the red spheres represents original sound source(s) in the scene. We drive the signals from active loudspeakers to manipulate the sound field along with original sound sources using the acoustic metrics corresponding to music and speech improvement.



(a) Fields in Trinity scene (b) Fields in Berlin scene

Figure 4: We highlight the STI distribution corresponding to the speech sources in the Trinity and Berlin models. Our optimization algorithm significantly improves the speech understanding as shown by high values (right) as compared to the low values (left) of STI metric. This highlight the benefits of our dynamic SFS algorithm in terms of speech intelligibility, and it makes no assumption about the model or the sound source.



(a) Sound fields in Elmia scene (b) Sound fields in Sibenik scene

Figure 5: We highlight the frequency compensation effects of music sound fields using our dynamic SFS algorithm in two benchmark scenes. The left figure in each scene shows the cross-correlation between the original distorted sound field and the desired sound field, while the right one shows the cross-correlation between our synthesized sound field and the desired sound field. We observe that our optimization algorithm results in sound fields that have the desired sound characteristics in terms of high correlation values (right) over low correlation values (left).

location on the stage with a flat frequency response. Note that in actual performances, users tend to emphasize some components of the frequency, while attenuate the other parts. Our approach can also account for these effects. We show the desired and synthesized sound signal at the listener location in Figure 5.

6.4 User Evaluation

In addition to the numeric results (of sound fields) shown in Section 6.2, we also conducted a user study to evaluate the perceptual benefits of our algorithm for speech improvement.

6.4.1 Study Goal. We aim to demonstrate the effectiveness of our dynamic SFS method to improve speech intelligibility. Moreover, we also compare the perceptual benefits on results generated from the commercial software Era-R developed by *Accusonus Inc,* [24]. Our hypothesis is that our method performs no worse than Era-R.

6.4.2 Study Design. Our study was based on pairwise comparisons. We prepared three reverberant speech clips that were 11 seconds long. By using reverberant clips as the reference, we performed dereverberation on these clips separately using Era-R and our method. Next, we obtained audio clips corresponding to the

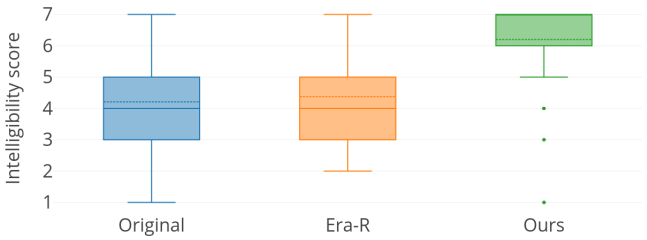


Figure 6: Subject scores for speech intelligibility in the three processing categories: original clip, Era-R [24] and our dynamic SFS algorithm. The score 1 indicates worst intelligibility while 7 indicates the best intelligibility. We observe higher speech intelligibility using our method in these tests.

three categories (original, Era-R, our method) of 3 different speech clips. We randomly ordered the 9 clips for each participant.

6.4.3 Metrics. During the study, participants were asked to listen to audio clips from our test sets. After listening to each one clip, participants were asked to rate the intelligibility for these clips using a 7-point Likert scale, with 1 indicating worst intelligibility, and 7 indicating best intelligibility. Before listening to our test set, they were presented with two extra audio clips (a reverberant clip and a dereverberated clip) for familiarization with such clips and the tasks, as well as the type of sound. The user responses to those two training audio clips were not counted in the evaluation result.

6.4.4 Study Results. We recruited 40 students (17 females), with a mean age of 23.9 (std=2.9), to take the study anonymously. All participants reported to have normal hearing. The study took each participant around 4 minutes to complete. We first performed two-tailed tests on intelligibility scores between ours and other two categories, with the null hypothesis that our method has the same mean score as that from other two categories. Instead, these tests rejected the null hypothesis and showed that the mean intelligibility score from our method was significantly different from all other two categories. Based on that, we performed one-tailed tests with the null hypothesis that our mean score is not higher than the other two. Further tests also rejected this hypothesis, and proved that our method had higher mean intelligibility score than the other two. All tests were run under a significance level of 0.05.

6.5 Benefits and Comparisons

Our approach is different from prior sound field simulation algorithms based on digital signal processing methods. As compared to prior methods, our approach offers the following benefits:

- **Loudspeaker placement:** We do not impose any restrictions on the placement of loudspeakers, except that they need to be closer to the target sound zone than any other sound sources. This gives much more flexibility.
- **Arbitrary domain:** We do not make any assumptions on the size or shape of the geometric model (e.g. rectangular or circular shape), and our approach is applicable to all models. We highlight the performance on many complex models shown in Figure 3.
- **Stability:** Our result is consistent with the multiple-input multiple-output inverse theorem (MINT). But because our IR computation considers the highly unsymmetrical complex acoustic environment, we do not suffer from the location-sensitivity issue in naive MINT implementations [29].
- **Reverberation:** Most existing work on sound field synthesis are limited to direct sound or use simple, pre-computed models of late reverberation. Instead, our approach tends to compute the reverberation effects using accurate sound propagation algorithms. As a result, our approach can reliably model the sound effects in arbitrary domain and account for these effects in terms of dynamic SFS.
- **Dynamic sound signals:** Our approach has the ability to predict the impact of external source signals as we try to manipulate the sound field. This property is useful when the external source locations are known, but the actual audio signal is unpredictable.

7 CONCLUSION, LIMITATIONS, AND FUTURE WORK

We present a novel method of dynamically synthesizing the sound field in a localized 3D environment by adding new loudspeakers (equivalent to virtual sources) to the acoustic scene. We present an optimization algorithm that uses precomputed IRs to compute loudspeaker driving signals to form the desired sound field. As compared to prior sound field synthesis methods, our approach is more general and allows for arbitrary selection and placement of loudspeakers. We highlight the performance of our algorithm on two applications: speech improvement and music reinforcement.

Our approach has some limitations. We assume that an accurate geometric and material model representation is given and do not account for signal sensing delays. Our current implementation is based on geometric ray tracing propagation and may not work well for low-frequency sources. We can improve the accuracy using hybrid wave-numeric methods. The trade-off between precomputation cost in terms of computing all the IRs for source-listener pairs and SFS accuracy remains as future research. Furthermore, our formulation assumes that the input sound sources are static. The resulting algorithms for speech improvement and music reinforcement make some assumptions about these applications.

There are many avenues for future work. In addition to addressing these limitations, we would like to further evaluate our algorithm's performance in various scenarios. It would be useful to combine our algorithm with clustering methods to handle a large number of active loudspeakers or monitoring points in large acoustic spaces. The IR computations between a large number of pairs could be accelerated by exploiting the spatial coherence of the sound field. While some of the metrics used in our formulation (e.g., STI) are based on psycho-acoustic criteria, it would be useful to explore the use of psycho-acoustic metrics to handle large acoustic spaces as part of our optimization algorithm.

ACKNOWLEDGEMENTS

This research was supported by ARO grants W911NF14-1-0437 and W911NF-18-1-0313, NSF grant 1320644, and Intel.

REFERENCES

[1] Jens Ahrens. 2012. *Analytic Methods of Sound Field Synthesis.* Springer Science & Business Media.

[2] Jens Ahrens and Sascha Spors. 2009. On the secondary source type mismatch in wave field synthesis employing circular distributions of loudspeakers. In *Audio Engineering Society Convention 127.* Audio Engineering Society.

[3] Jont B Allen and David A Berkley. 1979. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America* 65, 4 (1979), 943–950.

[4] Jeffrey Stephen Bamford. 1995. *An analysis of ambisonic sound systems of first and second order.* Ph.D. Dissertation. University of Waterloo.

[5] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. 2015. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on.* IEEE, 504–511.

[6] Augustinus J Berkhout, Diemer de Vries, and Peter Vogel. 1993. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America* 93, 5 (1993), 2764–2778.

[7] JS Bradley, R Reich, and SG Norcross. 1999. A just noticeable difference in C 50 for speech. *Applied Acoustics* 58, 2 (1999), 99–108.

[8] Anish Chandak, Christian Lauterbach, Micah Taylor, Zhimin Ren, and Dinesh Manocha. 2008. Ad-frustum: Adaptive frustum tracing for interactive sound propagation. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (2008), 1707–1722.

[9] Etienne Corteel and Rozenn Nicol. 2003. Listening room compensation for wave field synthesis. What can be done?. In *Audio Engineering Society Conference: 23rd International Conference: Signal Processing in Audio Recording and Reproduction.* Audio Engineering Society.

[10] Eugene Czerwinski, Alexander Voishvillo, Sergei Alexandrov, and Alexander Terekhov. 2000. Propagation distortion in sound systems: Can we avoid it? *Journal of the Audio Engineering Society* 48, 1/2 (2000), 30–48.

[11] Jérôme Daniel. 2000. Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia. (2000).

[12] Jérôme Daniel. 2003. Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format. In *Audio Engineering Society Conference: 23rd International Conference: Signal Processing in Audio Recording and Reproduction.* Audio Engineering Society.

[13] Gary Davis and Gary D Davis. 1989. *The sound reinforcement handbook.* Hal Leonard Corporation.

[14] Diemer De Vries. 1996. Sound reinforcement by wavefield synthesis: Adaptation of the synthesis operator to the loudspeaker directivity characteristics. *Journal of the Audio Engineering Society* 44, 12 (1996), 1120–1131.

[15] Maria B Dühring, Jakob S Jensen, and Ole Sigmund. 2008. Acoustic design by topology optimization. *Journal of sound and vibration* 317, 3 (2008), 557–575.

[16] BS EN. 2011. 60268-16: 2011.". *Sound system equipment–Part 16: Objective rating of speech intelligibility by speech transmission index* (2011).

[17] Kazuko Fuchi and Hae Chang Gea. 2013. Room Acoustic Optimization with Variable Thickness Columns. *World Congress on Structural and Multidisciplinary Optimization* (2013).

[18] 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.

[19] Thomas Funkhouser, Nicolas Tsingos, and Jean-Marc Jot. 2003. Survey of Methods for Modeling Sound Propagation in Interactive Virtual Environment Systems. *Presence and Teleoperation* (2003).

[20] Michael A Gerzon. 1973. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society* 21, 1 (1973), 2–10.

[21] Bradford W Gillespie and Les E Atlas. 2002. Acoustic diversity for improved speech recognition in reverberant environments. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, Vol. 1. IEEE, I–557.

[22] Gaël Guennebaud, Benoît Jacob, et al. 2010. Eigen v3. http://eigen.tuxfamily.org. (2010).

[23] Hans-Günter Hirsch and David Pearce. 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW).*

[24] Accusonus Inc. 2017. ERA-R: Single-knob reverb removal plugin. (2017). https://accusonus.com/products/era-r

[25] Hanieh Khalilian, Ivan V Bajić, and Rodney G Vaughan. 2015. Joint optimization of loudspeaker placement and radiation patterns for sound field reproduction. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on.* IEEE, 519–523.

[26] Ole Kirkeby and Philip A Nelson. 1993. Reproduction of plane wave sound fields. *The Journal of the Acoustical Society of America* 94, 5 (1993), 2992–3000.

[27] Ole Kirkeby, Philip A Nelson, Felipe Orduna-Bustamante, and Hareo Hamada. 1996. Local sound field reproduction using digital signal processing. *The Journal of the Acoustical Society of America* 100, 3 (1996), 1584–1593.

[28] M. Jessel. 1973. *Acoustique thÃľorique : propagation et holophonie.* Masson et cie.

[29] Masato Miyoshi and Yutaka Kaneda. 1988. Inverse filtering of room acoustics. *IEEE Transactions on acoustics, speech, and signal processing* 36, 2 (1988), 145–152.

[30] Michael Monks, Byong Mok Oh, and Julie Dorsey. 2000. Audioptimization: Goal-based acoustic design. *Computer Graphics and Applications, IEEE* 20, 3 (2000), 76–90.

[31] Nicolas Morales and Dinesh Manocha. 2016. Efficient wave-based acoustic material design optimization. *Computer-Aided Design* 78 (2016), 83–92.

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

[33] Philip W Robinson, Samuel Siltanen, Tapio Lokki, and Lauri Savioja. 2014. Concert hall geometry optimization with parametric modeling tools and wave-based acoustic simulations. *Building Acoustics* 21, 1 (2014), 55–64.

[34] Atul Rungta, Carl Schissler, Nicholas Rewkowski, Ravish Mehra, and Dinesh Manocha. 2018. Diffraction Kernels for Interactive Sound Propagation in Dynamic Environments. *IEEE transactions on visualization and computer graphics* (2018).

[35] Daniel A Russell, Joseph P Titlow, and Ya-Juan Bemmen. 1999. Acoustic monopoles, dipoles, and quadrupoles: An experiment revisited. *American Journal of Physics* 67, 8 (1999), 660–664.

[36] Shinichi Sakamoto, Ayumi Ushiyama, and Hiroshi Nagatomo. 2006. Numerical analysis of sound propagation in rooms using the finite difference time domain method. *The Journal of the Acoustical Society of America* 120, 5 (2006), 3008–3008.

[37] Kai Saksela, Jonathan Botts, and Lauri Savioja. 2015. Optimization of absorption placement using geometrical acoustic models and least squares. *The Journal of the Acoustical Society of America* 137, 4 (2015), EL274–EL280.

[38] Carl Schissler and Dinesh Manocha. 2016. Adaptive impulse response modeling for interactive sound propagation. In *Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games.* ACM, 71–78.

[39] Carl Schissler and Dinesh Manocha. 2016. Interactive sound propagation and rendering for large multi-source scenes. *ACM Transactions on Graphics (TOG)* 36, 1 (2016), 2.

[40] 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.

[41] Sascha Spors, Rudolf Rabenstein, and Jens Ahrens. 2008. The theory of wave field synthesis revisited. In *124th AES Convention.* 17–20.

[42] Zhenyu Tang and Dinesh Manocha. 2018. Noise Field Control using Active Sound Propagation and Optimization. In *Acoustic Signal Enhancement (IWAENC), 2018 IEEE International Workshop on.* IEEE, to appear.

[43] 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.

[44] Jihui Zhang, Thushara D Abhayapala, Wen Zhang, Prasanga N Samarasinghe, and Shouda Jiang. 2018. Active Noise Control Over Space: A Wave Domain Approach. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 26, 4 (2018), 774–786.