Mixed Source Sound Field Translation for Virtual Binaural Application with Perceptual Validation

Lachlan Birnie*, Thushara Abhayapala*, Vladimir Tourbabin†, and Prasanga Samarasinghe*
*Audio & Acoustic Signal Processing Group, The Australian National University, Canberra, Australia
†Facebook Reality Labs, Redmond, Washington, USA

***Abstract*—Non-interactive and linear experiences like cinema film offer high quality surround sound audio to enhance immersion, however, the perspective is usually fixed to the recording microphone position. With the rise of virtual reality, there is a demand for recording and recreating real-world experiences that allow users to move throughout the reproduction. Sound field translation achieves this by building an equivalent environment of virtual sources to recreate the recording spatially. However, the technique remains to restrict the maximum distance a user can translate away from the recording microphone's perspective due to the discrete sampling by commercial higher order microphones only being capable of recording an acoustic sweet-spot. In this paper, we propose a method for binaurally reproducing a microphone recording in a virtual application that allows the user to freely translate their body further beyond the recording position. The method incorporates a mixture of near-field and far-field sources in a sparsely expanded virtual environment to maintain a perceptually accurate reproduction. We perceptually validate the method through a Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) experiment. Compared to the planewave benchmark, the proposed method offers both improved source localizability and robustness to spectral distortions at translated listening positions. A cross-examination with numerical simulations demonstrated that the sparse expansion relaxes the inherent sweet-spot constraint, leading to the improved localizability for sparse environments. Additionally, the proposed method is seen to better reproduce the intensity and binaural room impulse response spectra of near-field environments, further supporting the perceptual results.**

***Index Terms*—Sound field translation/navigation, virtual-reality, binaural synthesis, MUSHRA, higher order microphone.**

I. Introduction

Virtual reality devices will provide a novel framework for people to interact with each other at a higher social bandwidth through immersive audio and visual reproductions of the real-world [1], [2]. For example, in the future a person may be able to experience a live concert or orchestral performance through a virtual reproduction in their own home [3]. To complete the immersive experience, the listener/viewer should be allowed to explore/navigate and interact with the virtual reproduction [4]. Subsequently, methods to accurately record and model the perceptual change in visual and auditory information as the user moves are required to reconstruct a perceptually equivalent experience. Camera arrays have been used to capture visual information at multiple points-of-view for use in virtual reproductions [5]. Similarly, microphones distributed about an environment can record the spatial auditory scene from multiple points-of-view [6]. However, hardware and feasibility restrictions limit the continuous space that can be recorded, and as a result, acoustic reproductions are usually spatially confined to the perspective of each microphone. [7].

We are interested in binaurally reconstructing an acoustic environment that is perceptually equivalent to a recorded target environment, in a way that enables a listener to seamlessly move about the acoustic reproduction in a virtual/augmented reality application. Likewise, we are interested in using *sound field translation* to shift the three-dimensional space-varying acoustic field of a recorded environment about a listener (or reference point), such that the listener experiences the auditory sensation of moving seamlessly through the original sound field. Equivalently, the task of sound field translation can also be achieved by shifting the listener's binaural perspective through a recorded/modeled continuous three-dimensional sound field.

A continuous sound field can be modeled by matching the pressure and normal particle velocity over a bounding surface that surrounds the source free acoustic region of interest [8]. In practice, the equivalent source method is commonly utilized for estimation/reconstruction of the sound field exterior/interior to a single-layer surface [9]. This method simplifies the sound field estimation/reconstruction by a discretized superposition of virtual point-sources or planewaves throughout the bounding surface [10]–[12]. Additionally, the equivalent source method can be further simplified when considering a sparse acoustic environment that contains only a few sound sources [13]. Typically, the method's equivalent virtual sources are estimated from microphones distributed about the target sound field's bounding surface. Many microphones are required to adequately sample the boundary of larger acoustic environments. This often leads to an infeasibility of traditional equivalent source methods for the desired goal of capturing larger real-world environments for listeners to move throughout. Therefore, alternative approaches towards sound field translation, and thereby sound field estimation/reconstruction, that utilize fewer and more practical microphones are desired. Methods that extend the spatial capabilities of smaller commercial microphone arrays are of most interest.

Recently, there have been two key sound field translation approaches towards extending the auditory range of a spatial microphone recording that a listener can move within. These

This research is supported by Facebook Reality Labs, and an Australian Government Research Training Program (RTP) Scholarship

The ethical aspects of this research have been approved by the Australian National University Human Research Ethics Committee (Protocol 2019/767).

are an interpolation-based [14] and an extrapolation-based approach [15]. First, the interpolation-based approach records an acoustic environment with a distributed grid of multiple higher order microphones [16], [17]. During reconstruction, the sound field is continuously interpolated between each microphone, such that the listener is able to seamlessly move about the interior region recorded by the microphone grid. Good coloration and localization performance is expected from this interpolation-based approach [14]. However, the microphone grid may not be feasible for all real-world scenarios due to the large spatial, hardware, and synchronization costs associated with implementation [18]. Furthermore, listeners are typically confined to only moving within the boundaries of the microphone grid [19]; and sound sources within the grid are difficult to handle and may cause comb-filtering spectral distortions [20]. Methods that alleviate these drawbacks and allow the listener to translate beyond the microphone grid have been proposed. However, they usually require additional localization and separation of direct sound field components [21], [22]. Consequently, a large open area is required for recording, which may not always be feasible.

On the other hand, the extrapolation-based approach utilizes a single higher order microphone recording; and the listener moves about an estimation of the sound field that is exterior to the microphone [23]. This overcomes many of the hardware and spatial drawbacks of the interpolation-based approach. Because a single higher order microphone is utilized, the audio and visual capture system can occupy a single seat in the audience of a live event, which causes less obstruction and allows for more impromptu recordings.

Many extrapolation-based sound field translation methods have been developed, such as Ambisonic [23], [24], harmonic re-expansion [25], discrete source [26], and point-source distribution [27]. One of the most popular extrapolation-based methods which we consider to be the benchmark in this paper is the planewave method [28]. In this method, similar to the equivalent source method, the sound field within the higher order microphone region is equivalently matched by a superposition of discrete virtual planewave sources [29]. However, in addition, the planewave sound field translation method also aims to estimate the sound field exterior to the microphone region in a perceptually accurate manner. Overall an unbounded sound field is reconstructed from the recording. The listener can then perceptually move about the reproduction by a phase shifting technique applied to translate the equivalent virtual planewave sound field [26], [28].

In practice, however, most extrapolation-based approaches, including the planewave method, are constrained by the inherent properties of the higher order microphone [23]. Hardware limitations result in a discretely sampled sound field recording that is confined to a finite region [30]. This results in a truncated (finite-order) expansion of the recorded sound field that is governed by both the upper target frequency band and the microphone's radius [31]. As a result, accurate reproduction is only possible when the listener moves within a small acoustic sweet-spot of a few centimeters which is defined by the commercial microphone's size [32]. Attempting to move beyond this inherent sweet-spot region, even after extrapolation, results in spectral distortions [33]–[35], degraded source localization [23], [36], and a poor perceptual experience.

In this paper, we propose an alternative virtual source model for an extrapolation-based sound field translation method that enables both a near-field and far-field propagation mixture. With the proposed source model we are able to binaurally recreate a recorded sound field exterior to (or translated away from) the recording microphone with sufficient perceptual accuracy for human reception. Specifically, we are addressing recordings of commercial higher order microphones [32], [37], [38]. We emphasize that we are mainly interested in perceptual accuracy, not numerical accuracy; and that the proposed method takes liberties with its implementation that are unconventional for typical sound field estimation/reconstruction. We are concerned with the task of modeling/reconstructing a spatially recorded sound source. We note that modeling of complex acoustic environments and reverberation is beyond the scope of this paper due to its equal difficulty as a separate problem that requires an explicit solution [39].

The proposed sound field translation method builds upon the benchmark planewave method which we review in Section II. We add an additional distribution of near-field virtual sources to the planewave method's far-field virtual source distribution (Section III). This allows the proposed method to compensate for near-field effects of recorded sound sources, which attributes to better sound field reproduction [40]. We note that near-field effects are not managed well by the far-field virtual sources in the planewave method [41]. Furthermore, we do not use source localization processing or prior knowledge of the recorded source position to achieve a near-field model; which is typically the case for existing translation methods [22], [42]–[44]. Alternatively, we apply a sparse assumption to the proposed sound field translation method using a L1-norm regularization [45] (Section III-C). Sparse solutions have been used in several microphone array applications including source localization and the equivalent source method [46], [47]. The sparse expansion allows the proposed method to activate near-field virtual sources for a recorded near-field target source; and activate far-field virtual sources for a recorded far-field target source. Moreover, the sparse expansion helps to extrapolate the truncated sound field recording [48]–[50].

We initially proposed the near-field far-field source mixture in [51] without any substantial verification of the method's perceptual performance. Furthermore, in many virtual/augmented reality applications, the user is usually stimulated through both auditory and visual sensory modalities at the same time. The interaction between the two modalities and its consequences on the user's perception of the audio-visual scene are very complicated and may significantly differ from the audio-only case. Hence, in this paper we study the perceptual aspects of the source mixture through an audio-visual listening test with human participants in Section IV. We utilize a MUltiple Stimulus with Hidden Reference and Anchor (MUSHRA) [52], [53] framework adapted for use in a virtual environment to provide listeners with both an auditory and visual reference of the reconstructed target sound source [18], [43]. We compare four sound field translation methods with differing virtual source models and expansion techniques. We test the methods

for the reproduction of human speech and music against the metrics of source localizability and robustness to spectral distortions. We show that the proposed method offers greater perceptual accuracy and a more immersive experience for listeners moving throughout an expanded virtual reproduction.

In Section V, we investigate our perceptual experiment results against numerical simulations of the extrapolated pressure and intensity fields. We show that the proposed method's reconstruction better matches the pressure and intensity of the original environment beyond the microphone's sweet-spot. Furthermore, we study the proposed method's robustness to moderate reverberation noise with the image source method against translation distance and reverberation time (Sec. V-D). We provide our concluding remarks and discuss future research directions in Section VI.

II. Problem Formulation and the Planewave Sound Field Translation Method

In this section, we formulate the problem of reconstructing a recorded real-world experience such that a listener is able to perceptually move through the acoustic reproduction. We first present the process of recording a general sound field with a commercial higher order microphone. We then review the planewave sound field translation method presented in [28], which segments the reproduction into three parts. First, a virtual acoustic environment is built from a superposition of planewave sources. Second, planewave driving signals are estimated from the recording to model an equivalent acoustic environment. Third, a listener is placed inside the virtual equivalent environment and binaural signals are rendered as they move. We provide a discussion on the perceptual shortcomings of this planewave translation method at the end.

A. Sound Field Capture

Consider a real-world acoustic environment that contains multiple sound sources, for example, a musical performance with many instruments. Let the origin $\boldsymbol{o}$ denote the center of the environment's listening space, such as a seat in the middle of an audience. Each sound source is positioned at $\boldsymbol{z} = (r, \theta, \phi)$ with respect to $\boldsymbol{o}$, where $\theta \in [0, \pi]$ is the elevation angle downwards from the z-axis, and $\phi \in [0, 2\pi)$ is the azimuth angle counterclockwise from the x-axis. For a listener in the audience at position $\boldsymbol{d}$, the true sound they experience in the real-world can be described by

$$\overset{\text{(real)}}{\{\text{l,r}\}}P(k, \boldsymbol{d}) = \sum_{\mu=1}^{U} H_{\{\text{l,r}\}}(k, \boldsymbol{z}_\mu; \boldsymbol{d}) \times s_\mu(k), \tag{1}$$

where $\overset{\text{(real)}}{\{\text{l,r}\}}P(k, \boldsymbol{d})$ is the pressure at the listener's left and right ear, $H_{\{\text{l,r}\}}(k, \boldsymbol{z}; \boldsymbol{d})$ is the transfer function between each sound source and the listener's ears, or simply the Head-Related Transfer Function (HRTF) when the listener is in a free-field space without any reflections, $s_\mu(k)$ is the sound signal of the μ^{th} source, $\mu = (1, \cdots, U)$, $k = 2\pi f/c$ is the wave number, f is the frequency, and c is the speed of sound. From here on, we assume H to be the free-field HRTF for simplicity.

The aim is to record and reproduce the target real-world auditory experience of (1) for every possible listening position. The homogeneous sound field that encompasses every arbitrary listening position $\boldsymbol{x}$, where $|\boldsymbol{x}| < |\boldsymbol{z}|$, can be expressed through a spherical harmonic decomposition of [8]

$$P(k, \boldsymbol{x}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \alpha_{nm}(k) j_n(k|\boldsymbol{x}|) Y_{nm}(\hat{\boldsymbol{x}}), \tag{2}$$

where $|\cdot| \equiv ||\cdot||_2 \equiv r$, $\hat{\cdot} \equiv (\theta, \phi)$, n and m are index terms denoting spherical harmonic order and mode, respectively, $j_n(\cdot)$ are the spherical Bessel functions of the first kind, $Y_{nm}(\cdot)$ are the set of spherical harmonic basis functions, and $\alpha_{nm}(k)$ are the sound field's coefficients which completely describe the source-free acoustic environment centered about $\boldsymbol{o}$ when $\alpha_{nm}(k)$ is known for all $n \in [0, \infty)$.

In practice, the target real-world acoustic environment can be recorded with an N^{th} order microphone, by estimating the sound field's $\alpha_{nm}(k)$ coefficients for a finite set of $n \in [0, N]$. Consider a N^{th} order microphone centered at $\boldsymbol{o}$, such as an open or rigid spherical [32] (or planar [54], [55]) microphone array. The microphone array consists of $q = (1, \cdots, Q)$ pressure sensors that enclose the spherical acoustic region (listening space) of radius $|\boldsymbol{x}_Q|$ to be recorded. The sound field within this region can be estimated with [56]

$$\alpha_{nm}(k) \approx \sum_{q=1}^{Q} w_q \frac{P(k, \boldsymbol{x}_q) Y^*_{nm}(\hat{\boldsymbol{x}}_q)}{b_n(k|\boldsymbol{x}_Q|)}, \quad n \in [0, N], \tag{3}$$

where w_q are a set of suitable sampling weights [57], and $b_n(\cdot)$ is the rigid baffle equation [8].

However, commercial N^{th} order microphones can only record a small acoustic region ($|\boldsymbol{x}_Q| < 0.05$ m [32]) due to the hardware complexity and size constraint trade-offs with the spatial sampling Nyquist theorem [30]. The microphone's truncation order is restricted by the limited number of sensors, such that $Q \geq (N+1)^2$. Furthermore, the microphone's recording region and frequency range are balanced by the $N = \lceil k|\boldsymbol{x}_Q| \rceil$ rule [58]. These two microphone properties define a maximum $|\boldsymbol{x}_Q|$ inside which the sound field is effectively of order $\leq N$. For example, a $Q = 32$ sensor microphone and a desired upper frequency limit of 5200 Hz would define a $N = 4^{\text{th}}$ order recording region of maximum size $|\boldsymbol{x}_Q| = 0.042$ m. The sound field within this region can be recorded and reconstructed accurately up to 5200 Hz. However, attempting to reconstruct the sound field beyond $|\boldsymbol{x}_Q|$ requires higher orders $> N$ which are unknown, and this results in truncation error that degrades perceptual accuracy.

When reconstructing (1) from the recording, the left and right ear signals for the listener can be reassembled in the spherical harmonic domain by [59], [60]

$$\overset{\text{(mic)}}{\{\text{l,r}\}}P(k, \boldsymbol{o}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} H^{nm}_{\{\text{l,r}\}}(k) \times \alpha_{nm}(k), \tag{4}$$

where $H^{nm}_{\{\text{l,r}\}}(k)$ are the spherical harmonic decomposition coefficients of the HRTF $H_{\{\text{l,r}\}}(k, \boldsymbol{z}; \boldsymbol{o})$. In the reproduction (4), truncation forces the listener to the fixed auditory perspective of the microphone at $\boldsymbol{o}$. If the listener attempts to move then

they would immediately translate beyond the $|\boldsymbol{x}_Q|$ boundary and begin to experience spectral distortions, degraded source localization performance, and a loss in perceptual immersion.

The objective of this paper is to relax this sweet-spot spatial constraint when reconstructing the target sound field of a commercial microphone recording, and to build an equivalent virtual environment that allows for a listener to move about the acoustic space with a sustained perceptual immersion. For the remainder of this section we review the planewave sound field translation method that we consider to be the baseline method for enabling listener navigation.

B. Planewave Distribution

The planewave sound field translation method aims to construct a virtual acoustic environment that is perceptually equivalent to the real-world recording. The building block of this virtual environment is the planewave source, whose sound field is modeled as

$$P(k, \boldsymbol{x}) = \frac{e^{-ik\hat{\boldsymbol{y}}\cdot\boldsymbol{x}}}{4\pi}, \tag{5}$$

where $\hat{\boldsymbol{y}}$ denotes the planewave's incident direction. It is known that any acoustic free field can be modeled by an infinite superposition of planewaves [29]. Therefore, the equivalent virtual environment is constructed from a spherical distribution of virtual planewave sources, expressed as

$${}^{(\text{pw})}P(k, \boldsymbol{x}) = \int \psi(k, \hat{\boldsymbol{y}}; \boldsymbol{o}) \frac{e^{-ik\hat{\boldsymbol{y}}\cdot\boldsymbol{x}}}{4\pi} d\hat{\boldsymbol{y}}, \tag{6}$$

where $\psi(k, \hat{\boldsymbol{y}}; \boldsymbol{o})$ denotes the driving function of the planewave distribution as observed at $\boldsymbol{o}$. If the driving function is modeled correctly then the planewave distribution can recreate the acoustic environment, such that ${}^{(\text{pw})}P(k, \boldsymbol{x}) = {}^{(\text{real})}P(k, \boldsymbol{x})$. To achieve this, the driving function needs to be estimated/expanded from the recorded $\alpha_{nm}(k)$ coefficients, which we describe next.

C. Planewave Expansion

The sound field about $\boldsymbol{o}$ due to a single virtual planewave (5) can be expressed by the decomposition of [8]

$$\frac{e^{-ik\hat{\boldsymbol{y}}\cdot\boldsymbol{x}}}{4\pi} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} (-i)^n Y_{nm}^*(\hat{\boldsymbol{y}}) j_n(k|\boldsymbol{x}|) Y_{nm}(\hat{\boldsymbol{x}}). \tag{7}$$

Additionally, the driving function centered at $\boldsymbol{o}$ can also be expressed in terms of a harmonic decomposition, given as

$$\psi(k, \hat{\boldsymbol{y}}; \boldsymbol{o}) = \sum_{n'=0}^{\infty} \sum_{m'=-n'}^{n'} \beta_{n'm'}(k) Y_{n'm'}(\hat{\boldsymbol{y}}), \tag{8}$$

where $\beta_{nm}(k)$ are the spherical harmonic decomposition coefficients of $\psi(k, \hat{\boldsymbol{y}}; \boldsymbol{o})$ which describe the sound field about the planewave distribution. Substituting both (7) and (8) into (6) gives the planewave distribution's sound field in spherical harmonics, as

$${}^{(\text{pw})}P(k, \boldsymbol{x}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \underbrace{(-i)^n \beta_{nm}(k)}_{\alpha_{nm}(k)} j_n(k|\boldsymbol{x}|) Y_{nm}(\hat{\boldsymbol{x}}). \tag{9}$$

From (9), the relationship between the $\beta_{nm}(k)$ coefficients and the recorded $\alpha_{nm}(k)$ coefficients can be extracted. Rearranging this relationship for $\beta_{nm}(k) = i^n \alpha_{nm}(k)$, expresses a planewave distribution that is equivalent to the recorded environment. Substituting this relationship back into (8), gives a closed-form expansion for a planewave driving function that matches the recording,

$$\psi(k, \hat{\boldsymbol{y}}; \boldsymbol{o}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} i^n \alpha_{nm}(k) Y_{nm}(\hat{\boldsymbol{y}}). \tag{10}$$

Synthesizing a virtual environment with this driving function through (6) produces a sound field that is equivalent to the recording. However, the recording (3) is only an approximation of the real environment, and therefore (10) is also an approximate, such that ${}^{(\text{pw})}P(k, \boldsymbol{x}) \equiv {}^{(\text{mic})}P(k, \boldsymbol{x}) \approx {}^{(\text{real})}P(k, \boldsymbol{x})$.

D. Planewave Auralization

A listener inside the equivalent planewave distribution experiences a spatial reproduction of the recorded environment. Binaural signals can reconstruct the recording at the distribution center by exchanging the listener's HRTF into (6), giving

$${}^{(\text{pw})}_{\{\text{l,r}\}}P(k, \boldsymbol{o}) = \int \psi(k, \hat{\boldsymbol{y}}; \boldsymbol{o}) H_{\{\text{l,r}\}}(k, \hat{\boldsymbol{y}}; \boldsymbol{o}) d\hat{\boldsymbol{y}}. \tag{11}$$

Note that $H_{\{\text{l,r}\}}(k, \hat{\boldsymbol{y}}; \boldsymbol{o})$ is rotated with the listener's looking direction such that the binaural signals are updated when the listener turns their head. Furthermore, the listener can move perceptually about the reproduction using the planewave sound field translation method. The sound heard by the listener when translated to $\boldsymbol{x} = [\boldsymbol{o} + \boldsymbol{d}] \equiv \boldsymbol{d}$ can be derived from (6) as

$$\begin{aligned} {}^{(\text{pw})}P(k, [\boldsymbol{o} + \boldsymbol{d}]) &= \int \psi(k, \hat{\boldsymbol{y}}; \boldsymbol{o}) \frac{e^{-ik\hat{\boldsymbol{y}}\cdot[\boldsymbol{o}+\boldsymbol{d}]}}{4\pi} d\hat{\boldsymbol{y}} \\ &= \int \psi(k, \hat{\boldsymbol{y}}; \boldsymbol{o}) e^{-ik\hat{\boldsymbol{y}}\cdot\boldsymbol{d}} \frac{e^{-ik\hat{\boldsymbol{y}}\cdot\boldsymbol{o}}}{4\pi} d\hat{\boldsymbol{y}}. \end{aligned} \tag{12}$$

It is observed from (12) that the translation in space differs only by a phase shift in the planewave driving function. Therefore, applying the translational phase shift of [41]

$$\psi(k, \hat{\boldsymbol{y}}; \boldsymbol{d}) = \psi(k, \hat{\boldsymbol{y}}; \boldsymbol{o}) \times e^{-ik\hat{\boldsymbol{y}}\cdot\boldsymbol{d}}, \tag{13}$$

to the binaural signals in (11), allows for the listener to dynamically move their acoustic perspective by

$${}^{(\text{pw})}_{\{\text{l,r}\}}P(k, \boldsymbol{d}) = \int \psi(k, \hat{\boldsymbol{y}}; \boldsymbol{d}) H_{\{\text{l,r}\}}(k, \hat{\boldsymbol{y}}; \boldsymbol{o}) d\hat{\boldsymbol{y}}. \tag{14}$$

In practice, the virtual planewave distribution (6) can be realized with a discrete set of planewave sources,

$${}^{(\text{pw})}P(k, \boldsymbol{x}) \approx \sum_{\ell=1}^{L} w_\ell \psi(k, \hat{\boldsymbol{y}}_\ell; \boldsymbol{o}) \frac{e^{-ik\hat{\boldsymbol{y}}_\ell\cdot\boldsymbol{x}}}{4\pi} \tag{15}$$

where $\ell = (1, \cdots, L)$ index each virtual planewave, L is the total number of sources, and w_ℓ are a set of suitable sampling weights. Similarly, the listener's binaural signals (14) can be realized from the discrete planewave distribution with

$${}^{(\text{pw})}_{\{\text{l,r}\}}P(k, \boldsymbol{d}) \approx \sum_{\ell=1}^{L} w_\ell \psi(k, \hat{\boldsymbol{y}}_\ell; \boldsymbol{d}) H_{\{\text{l,r}\}}(k, \hat{\boldsymbol{y}}_\ell; \boldsymbol{o}). \tag{16}$$

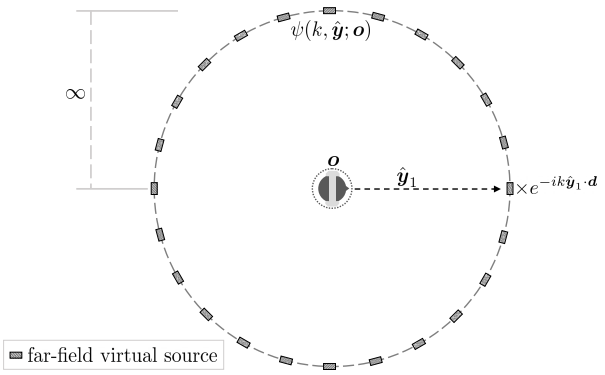


Fig. 1: Illustration of the equivalent virtual planewave distribution. The listener's perspective is fixed at the distribution center $\boldsymbol{o}$, where a phase shift applied to the planewave driving function translates the sound field about the listener. Note that this illustrates a two-dimensional cross section, in practice the source distribution is a sphere about the listener.

We illustrate this planewave method to sound field translation in Fig. 1. The reproduction is expressed by many discrete planewave signals that are known continuously throughout the virtual environment. Therefore, the method does not explicitly limit the amount the listener can translate. However, (16) uses $\psi(k, \hat{\boldsymbol{y}}; \boldsymbol{o})$ which is estimated through (3) and (10). As a result, the recording's N^{th} order truncation inherently remains, and the listener's movement is still implicitly limited.

E. Planewave Sound Field Translation Method Discussion

The virtual planewave expansion enables listener translation, however, some shortcomings are still exhibited in the listener's perception:

- As mentioned, the planewave method inherits truncation artifacts through an over-approximation of (10), and the listener's movement remains inherently restricted inside the virtual reproduction. As the listener translates further away from the recording's sweet-spot, they begin to experience spectral distortions, a loss in source localization, and poorer perceptual accuracy.
- The planewave expansion has difficulties in synthesizing near-field sound sources due to its far-field source model.
- The planewave auralization (16) is degraded by the HRTF perspective begin fixed to the virtual distribution center. We discuss this further in Sec. III-F.

In the next section we propose an alternative sound field translation model to address the above shortcomings.

III. Mixedwave Sound Field Translation Method

In this section, we define a virtual source that models both a near-field and far-field propagation, which we will refer to as a *mixedwave* source. We then build a virtual distribution of mixedwave sources and expand a real-world recording into an equivalent sound field. Additionally, we also propose a sparse method for expanding a virtual source distribution that alleviates some of the spatial restrictions imposed by the truncated recording.

A. Source for Near-Field and Far-Field Mixture

Here, we define the virtual source that will be the building block for our proposed method. Consider a near-field point-source at $\boldsymbol{y}$, where the driving signal of the source with respect to itself is denoted $\dot{\psi}(k, \boldsymbol{y})$. We can express the driving function observed at a position $\boldsymbol{x}$ with [8]

$$\psi(k, \boldsymbol{y}; \boldsymbol{x}) = \dot{\psi}(k, \boldsymbol{y}) \frac{e^{ik|\boldsymbol{y}-\boldsymbol{x}|}}{|\boldsymbol{y}-\boldsymbol{x}|}. \tag{17}$$

Evaluating (17) when $\boldsymbol{x} = \boldsymbol{o}$ gives the driving function observed by a receiver/microphone, as

$$\psi(k, \boldsymbol{y}; \boldsymbol{o}) = \dot{\psi}(k, \boldsymbol{y}) \frac{e^{ik|\boldsymbol{y}|}}{|\boldsymbol{y}|}. \tag{18}$$

Rearranging (18) gives an expression for the source signal in terms of the source's distance and the driving function observed by the receiver,

$$\dot{\psi}(k, \boldsymbol{y}) = \psi(k, \boldsymbol{y}; \boldsymbol{o}) |\boldsymbol{y}| e^{-ik|\boldsymbol{y}|}. \tag{19}$$

Substituting (19) back into (17) provides the driving function observed at any arbitrary point $\boldsymbol{x}$ in terms of the function observed by the receiver/microphone, expressed as

$$\psi(k, \boldsymbol{y}; \boldsymbol{x}) = \underbrace{\psi(k, \boldsymbol{y}; \boldsymbol{o}) |\boldsymbol{y}| e^{-ik|\boldsymbol{y}|}}_{\dot{\psi}(k, \boldsymbol{y})} \frac{e^{ik|\boldsymbol{y}-\boldsymbol{x}|}}{|\boldsymbol{y}-\boldsymbol{x}|}. \tag{20}$$

We note that the $|\boldsymbol{y}| e^{-ik|\boldsymbol{y}|}$ term can be seen to have redefined the point-source from being a function with respect to itself, to being a function with respect to $\boldsymbol{o}$. This allows us to observe the source distribution at $\boldsymbol{o}$ with a microphone and estimate the sound at any translated position $\boldsymbol{x}$.

Additionally, the constant term has the property of [31]

$$\lim_{|\boldsymbol{y}| \to \infty} |\boldsymbol{y}| e^{-ik|\boldsymbol{y}|} \frac{e^{ik|\boldsymbol{y}-\boldsymbol{x}|}}{4\pi|\boldsymbol{y}-\boldsymbol{x}|} = \frac{e^{-ik\hat{\boldsymbol{y}} \cdot \boldsymbol{x}}}{4\pi}, \tag{21}$$

which allows for a mixture of near-field and far-field virtual source distributions to be modeled with this building block. We define this building block as the mixedwave source,

$$P(k, \boldsymbol{x}) = |\boldsymbol{y}| e^{-ik|\boldsymbol{y}|} \frac{e^{ik|\boldsymbol{y}-\boldsymbol{x}|}}{4\pi|\boldsymbol{y}-\boldsymbol{x}|}. \tag{22}$$

In the spherical harmonic domain,

$$|\boldsymbol{y}| e^{-ik|\boldsymbol{y}|} \frac{e^{ik|\boldsymbol{y}-\boldsymbol{x}|}}{4\pi|\boldsymbol{y}-\boldsymbol{x}|} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} ik|\boldsymbol{y}| e^{-ik|\boldsymbol{y}|} h_n(k|\boldsymbol{y}|) Y_{nm}^*(\hat{\boldsymbol{y}}) j_n(k|\boldsymbol{x}|) Y_{nm}(\hat{\boldsymbol{x}}), \tag{23}$$

where $h_n(\cdot)$ is the spherical Hankel function of the first kind. We note that spherical Hankel functions also have

$$\lim_{|\boldsymbol{y}| \to \infty} ik|\boldsymbol{y}| e^{-ik|\boldsymbol{y}|} h_n(k|\boldsymbol{y}|) = (-i)^n, \tag{24}$$

to correspond with (21). We can observe from (24) that when the mixedwave source is placed in the far-field, the definition of (23) will match that of the planewave source (7). This property then allows for both a near-field sound propagation

to be modeled by a mixedwave distribution with a small radius, and a far-field sound propagation to be modeled by a mixedwave distribution with a large radius. We will use this near-field and far-field distribution of mixedwave sources as the basis of our proposed sound field translation method next.

B. Mixedwave Method for Sound Field Translation

Following the planewave translation method, our proposed mixedwave translation method is also broken into three parts.

1) Mixedwave Distribution: We propose constructing a virtual equivalent sound field from two concentric spherical distributions of mixedwave sources. The first virtual sphere is placed in the near-field with a radius of $R_{(\text{nf})}$, and the second sphere is placed at $R_{(\text{ff})}$ in the far-field, such that

$$\begin{aligned} {}^{(\text{mw})}P(k,\boldsymbol{x}) &= \\ &\int \psi(k,R_{(\text{nf})}\hat{\boldsymbol{y}};\boldsymbol{o})R_{(\text{nf})}e^{-ikR_{(\text{nf})}}\frac{e^{ik|R_{(\text{nf})}\hat{\boldsymbol{y}}-\boldsymbol{x}|}}{4\pi|R_{(\text{nf})}\hat{\boldsymbol{y}}-\boldsymbol{x}|}d\hat{\boldsymbol{y}} \\ +&\int \psi(k,R_{(\text{ff})}\hat{\boldsymbol{y}};\boldsymbol{o})R_{(\text{ff})}e^{-ikR_{(\text{ff})}}\frac{e^{ik|R_{(\text{ff})}\hat{\boldsymbol{y}}-\boldsymbol{x}|}}{4\pi|R_{(\text{ff})}\hat{\boldsymbol{y}}-\boldsymbol{x}|}d\hat{\boldsymbol{y}}, \end{aligned} \tag{25}$$

where, $\psi(k,R\hat{\boldsymbol{y}};\boldsymbol{o})$, $R\in\{R_{(\text{nf})},R_{(\text{ff})}\}$, are the driving functions of the two mixedwave distributions co-centered at $\boldsymbol{o}$.

2) Mixedwave Expansion: Following the procedure in Section II-C, we can decompose the $\psi(k,R\hat{\boldsymbol{y}};\boldsymbol{o})$ driving function into spherical harmonic aperture coefficients of $\beta_{n'm'}(k,R)$, expressed as

$$\psi(k,R\hat{\boldsymbol{y}};\boldsymbol{o})=\sum_{n'=0}^{\infty}\sum_{m'=-n'}^{n'}\beta_{n'm'}(k,R)Y_{n'm'}(\hat{\boldsymbol{y}}). \tag{26}$$

We substitute both (26) and (23) into (25) to extract the relationship between $\beta_{nm}(k)$ and $\alpha_{nm}(k)$, given as

$$\beta_{nm}(k,R)=\frac{\alpha_{nm}(k)}{ikRe^{-ikR}h_n(kR)}. \tag{27}$$

Finally, we substitute (27) back into (26) to derive a closed-form expansion for the mixedwave driving functions in terms of the recorded coefficients,

$$\psi(k,R\hat{\boldsymbol{y}};\boldsymbol{o})=\sum_{n=0}^{N}\sum_{m=-n}^{n}\frac{\alpha_{nm}(k)}{ikRe^{-ikR}h_n(kR)}Y_{nm}(\hat{\boldsymbol{y}}). \tag{28}$$

We use the recorded coefficients $\alpha_{nm}(k)$ of the target real-world environment with (28) to estimate the driving functions of the near-field and far-field virtual distributions, such that ${}^{(\text{mw})}P(k,\boldsymbol{x})\equiv{}^{(\text{mic})}P(k,\boldsymbol{x})\approx{}^{(\text{real})}P(k,\boldsymbol{x})$.

3) Mixedwave Auralization: Consider a listener inside the virtual mixedwave distribution at the translated position $\boldsymbol{x}=[\boldsymbol{o}+\boldsymbol{d}]\equiv\boldsymbol{d}$, $|\boldsymbol{d}|<R_{(\text{nf})}$, as shown in Fig. 2. We render the left and right binaural signals by applying the mixedwave driving function to the HRTF based on the listener's translated position, given as [23]

$${}^{(\text{mw})}_{\{\text{l,r}\}}P(k,\boldsymbol{d})=\int\psi(k,R\hat{\boldsymbol{y}};\boldsymbol{o})H_{\{\text{l,r}\}}(k,R\hat{\boldsymbol{y}};\boldsymbol{d})d\boldsymbol{y}, \tag{29}$$

where $R\hat{\boldsymbol{y}};\boldsymbol{d}$ denotes the position of the mixedwave source with respect to the listener at $\boldsymbol{d}$, which is given by $(\boldsymbol{y}-\boldsymbol{d})$. We note that $H_{\{\text{l,r}\}}(k,R\hat{\boldsymbol{y}};\boldsymbol{d})$ is rotated with the listener's looking direction such that the binaural signals are updated when the listener turns their head.

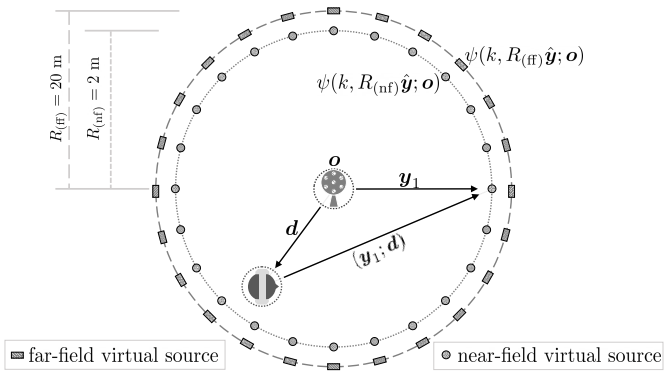


Fig. 2: The equivalent mixedwave virtual source distribution. The listener is translated to $\boldsymbol{d}$, and the vectors $(\boldsymbol{y}_\ell;\boldsymbol{d})$ are updated with the HRTF to auralize an immersive reproduction. Note that this illustrates a two-dimensional cross section, in practice the source distribution is a sphere about the listener.

Once again, a set of discrete sources can be used to practically realize the virtual mixedwave distributions, given as

$$\begin{aligned} {}^{(\text{mw})}P(k,\boldsymbol{x})&\approx \\ &\sum_{\ell=1}^{L}w_\ell\psi(k,R_{(\text{nf})}\hat{\boldsymbol{y}}_\ell;\boldsymbol{o})R_{(\text{nf})}e^{-ikR_{(\text{nf})}}\frac{e^{ik|R_{(\text{nf})}\hat{\boldsymbol{y}}_\ell-\boldsymbol{x}|}}{4\pi|R_{(\text{nf})}\hat{\boldsymbol{y}}_\ell-\boldsymbol{x}|} \\ +&\sum_{\ell=1}^{L}w_\ell\psi(k,R_{(\text{ff})}\hat{\boldsymbol{y}}_\ell;\boldsymbol{o})R_{(\text{ff})}e^{-ikR_{(\text{ff})}}\frac{e^{ik|R_{(\text{ff})}\hat{\boldsymbol{y}}_\ell-\boldsymbol{x}|}}{4\pi|R_{(\text{ff})}\hat{\boldsymbol{y}}_\ell-\boldsymbol{x}|}, \end{aligned} \tag{30}$$

where the near-field and far-field distributions each contain L sources. Similarly, we realize the mixedwave auralization within the discrete virtual distributions by

$${}^{(\text{mw})}_{\{\text{l,r}\}}P(k,\boldsymbol{d})=\sum_{\ell=1}^{2L}w_\ell\psi(k,\boldsymbol{y}_\ell;\boldsymbol{o})H_{\{\text{l,r}\}}(k,\boldsymbol{y}_\ell;\boldsymbol{d}), \tag{31}$$

where $|\boldsymbol{y}_\ell|=R_{(\text{nf})}$ for $\ell\in[1,L]$, and $|\boldsymbol{y}_\ell|=R_{(\text{ff})}$ for $\ell\in[L+1,2L]$, and $\boldsymbol{y}_\ell;\boldsymbol{d}$ is the propagation direction of the ℓ^{th} mixedwave source with respect to the translated listener. Unlike the planewave method, the maximum distance a listener can translate within the mixedwave environment is restricted by $R_{(\text{nf})}$. However, we suspect that $R_{(\text{nf})}$ can be selected to match the size of a small real-world room that is recorded.

C. Sparse Expansion Methods

The closed-form expansion constructs a virtual environment that is equivalent to the original recording. However, the expansion distributes energy $\psi(k,\boldsymbol{y}_\ell;\boldsymbol{o})$ throughout all virtual sources. This causes an over-approximation of the truncated recording's underlying spatial artifacts. As a result, the amount a listener can translate before experiencing a loss in immersion is still inherently restricted by the recording's truncation. Furthermore, it is believed that modeling fewer virtual sources from propagation directions that are similar to the original environment will lead to better perceptual immersion [51].

For these reasons, we propose a sparse constrained expansion method for constructing our virtual mixedwave environment.

The coefficients $\alpha_{nm}(k)$ observed at the center of a virtual distribution can be expressed in matrix form as

$$\boldsymbol{A}\boldsymbol{\psi} = \boldsymbol{\alpha}, \tag{32}$$

where $\boldsymbol{\alpha} = [\alpha_{00}(k), \alpha_{1-1}(k), \cdots, \alpha_{NN}(k)]^{\mathrm{T}}$ are the recorded coefficients, $\boldsymbol{\psi} = [\psi(k, \boldsymbol{y}_1; \boldsymbol{o}), \cdots, \psi(k, \boldsymbol{y}_{\mathcal{L}}; \boldsymbol{o})]$ are the $\mathcal{L}$ equivalent virtual source driving signals, and $\boldsymbol{A}$ is the $(N+1)^2$ by $\mathcal{L}$ expansion matrix. The entries of $\boldsymbol{A}$ are given by $(-i)^n Y_{nm}^*(\hat{\boldsymbol{y}}_\ell)$ for a planewave expansion (10), and $ik|\boldsymbol{y}_\ell|e^{-ik|\boldsymbol{y}_\ell|}h_n(k|\boldsymbol{y}_\ell|)Y_{nm}^*(\hat{\boldsymbol{y}}_\ell)$ where $\mathcal{L} = 2L$ for the two source distributions of a mixedwave expansion (28). We assume $L > (N+1)^2$ for the under-determined case.

We construct a sparse source distribution by solving the linear regression problem (32) using Iteratively Reweighted Least Squares (IRLS) [45]. In brief, the IRLS approach replaces the ℓ^{p}-objective function (where $0 < \mathrm{p} \leq 1$)

$$\min_{\boldsymbol{\psi}} |\boldsymbol{\psi}|_{\mathrm{p}}^{\mathrm{p}} \qquad \text{subject to } \boldsymbol{A}\boldsymbol{\psi} = \boldsymbol{\alpha}, \tag{33}$$

with a weighted ℓ^2-norm,

$$\min_{\boldsymbol{\psi}} \sum_{i=1}^{\mathcal{L}} w_i \boldsymbol{\psi}_i^2 \qquad \text{subject to } \boldsymbol{A}\boldsymbol{\psi} = \boldsymbol{\alpha}, \tag{34}$$

where $w_i = |\boldsymbol{\psi}_i^{(\nu-1)}|^{p-2}$ are the weights computed from the previous iterate $\boldsymbol{\psi}^{(\nu-1)}$. The next iterate is given by

$$\boldsymbol{\psi}^{(\nu)} = Q_\nu \boldsymbol{A}^{\mathrm{T}} \left(\boldsymbol{A} Q_\nu \boldsymbol{A}^{\mathrm{T}}\right)^{-1} \boldsymbol{\alpha}, \tag{35}$$

where Q_ν is the diagonal matrix with $1/w_i = |\boldsymbol{\psi}_i^{(\nu-1)}|^{2-p}$. Other regularization techniques can also be utilized, such as the Least-Absolute Shrinkage and Selection Operator (Lasso) [61], [62], and we direct the reader to [63] for further information in regards to compressive sensing.

D. Selecting Mixedwave Spherical Shell Radius

Here we briefly provide some insight into how we select the near-field and far-field radius of each spherical shell of virtual mixedwave sources. The ideal inner shell radius would match the distance of any recorded near-field sound source. However, we do not propose using any source localization in this paper, and therefore we do not know the source's distance. Instead, we select the near-field shell radius based on the desired maximum translation distance, because the listener cannot move beyond the inner most shell. For this reason we select a 2 m radius for the inner shell. This radius is suitable for keeping the virtual sources in the near-field for a scenario where the listener is using a virtual reality device inside a 2-by-2 m space. As for the outer spherical shell, it can be placed anywhere so long as it is in the far-field. Therefore, we simply select a 20 m radius. The mixedwave distribution can be customized further with additional spherical shells and/or more virtual sources per shell. We note that for a traditional implementation of sparsity in sound field estimation/reconstruction, the virtual sources should be placed at every possible target source location [39]. That is, many spherical shells of different radius and many virtual sources per shell. However, given limited time and computation power it is desirable to utilize the minimum number of virtual sources required to obtain a perceptually accurate result. Additional shells only provide minor benefit, as they can only be placed outside the inner most shell to give more radii options for the sparse expansion. Therefore, we have found that a single near-field shell is sufficient for good perceptual performance. Expanding more virtual sources per shell gives the sparse expansion more opportunity to match the true sound direction of arrival, however, at a trade-off with computation cost. As such, we found $L = 36$ virtual sources per shell to be an adequate trade-off between angular resolution and computation cost for the perceptual experiment in Sec. IV.

E. Mixedwave Sound Field Translation Method Discussion

Continuing our discussion on the planewave method's shortcomings in Sec. II-E, we give the following comments:

- Sparsely expanding the virtual source distribution (32) with IRLS is expected to further enhance the perceptual immersion for a listener, as they should experience more localized virtual sources. Additionally, the sparsity relaxes the spatial sweet-spot restriction and over-approximation issue stemming from the closed-form expansion used by the planewave method. These properties are demonstrated by experiment in Section IV and by simulation in Section V.
- The mixedwave distribution can easily synthesize near-field sound sources. The modified point-source (22) can model a spherical-wave propagation by simply positioning the mixedwave source in the near-field.
- Below we discuss how unlike the planewave method, the mixedwave auralization (31) translates the HRTF with the listener. Intuitively, this attribute is expected to result in greater perceptual immersion.

F. Translation of the HRTF

In the following we discuss and clarify the differences of HRTF implementation between the planewave and mixedwave translation methods. We observe from (16) that the HRTF in the planewave auralization only updates when the listener rotates their head, and that the HRTF does not change when the listener translates. That is, the vector $(\hat{\boldsymbol{y}}_\ell; \boldsymbol{o})$ only changes $H_{\{\mathrm{l,r}\}}(k, \hat{\boldsymbol{y}}_\ell; \boldsymbol{o})$ when the listener rotates. Consider a virtual planewave that is approximating a target sound source 1 m directly in front of the listener. If the listener translates away from the origin, the virtual planewave phase is updated with (13) to render the change in sound at the new listening position. However, the HRTF does not update with this translation as $(\hat{\boldsymbol{y}}_\ell; \boldsymbol{o})$ is unchanged (assuming the listener did not rotate). As a result, the HRTF still models binaural time and level differences for a sound that is directly in front of the listener, even though the listener may have translated to the left of the source. This is inconsistent with reality. Instead, the HRTF should update when the perceived sound direction changes as the listener translates, such that the HRTF models $(\hat{\boldsymbol{y}}_\ell; \boldsymbol{d})$. However, by definition of a infinitely distanced planewave

$(\hat{\boldsymbol{y}}_\ell; \boldsymbol{d}) \equiv (\hat{\boldsymbol{y}}_\ell; \boldsymbol{o})$, and therefore the planewave method's HRTF perspective is fixed at the origin. In contrast, the finitely distanced mixedwave source allows for the vector $(\boldsymbol{y}_\ell; \boldsymbol{d})$ to update with the relative angle due to listener translation. Therefore, the HRTF in the mixedwave auralization (31) gives the binaural time and level differences that match the listener's movement. This benefit of the mixedwave source is expected to improve perceptual accuracy, which we will examine through binaural spectra in Sec. V-B3.

IV. Perceptual Experiment

Our aim is to maintain the immersion for a listener inside an acoustic reproduction. Therefore, it is of crucial importance, foremost, that we evaluate the proposed method against the planewave benchmark in a perceptual listening experiment. This section details the perceptual experiment system we have implemented and presents the statistical results at the end.

A. Experiment Methodology

1) Compared Methods: We conducted a MUSHRA perceptual experiment to compare four translation methods. In total the experiment presented six signals:

- *Reference / hidden reference:* Signals of the true free-field transfer function between the target real-world point-source and the translated listener, given by (1).
- *Anchor:* Signals of the truncated recording that is fixed spatially to the microphone's position (4). Sound field rotation is still rendered, but no translation is processed. This is a similar anchor to the three-degrees-of-freedom used in [43].
- *Benchmark / planewave closed-form (PW-CF):* Signals rendered from a virtual planewave distribution (16) that are expanded through the closed-form expression (10).
- *Planewave IRLS (PW-IRLS):* Signals from a IRLS (Section III-C) sparsely expanded planewave distribution.
- *Mixedwave closed-form (MW-CF):* Signals rendered from a virtual mixedwave distribution (31) that are expanded through the closed-form expression (28).
- *Proposed method / mixedwave IRLS (MW-IRLS):* Signals from a IRLS sparsely expanded mixedwave distribution.

The perceptual experiment comprised of four separate MUSHRA tests evaluated on the attribute of either *source localization* or *basic audio quality*, for a source signal of either *human speech* or *music*. The source localization test asked listeners to score on the perceived direction of the sound-source, the source width, and the sound field sparseness with respect to both a visual-reference and the reference signal. Whereas, the basic audio quality test asked listeners to score against the reference for spectral distortions and other audible processing artifacts. The speech signal was a male utterance taken from the TIMIT dataset [64], and the music signal was rock-music containing bass guitar, electric guitar, and drums with cymbals. Both signals did not exhibit reverberation. In total the scores of 17 participants were collected for the speech sound-source, and 11 scores for the music sound-source. The recording microphone was shown in the virtual environment, and listeners were informed that the further they translate, the greater the differences they should perceive between methods. We asked the listeners to score while accounting for each method's performance over a 1 m by 1 m reproduction space.

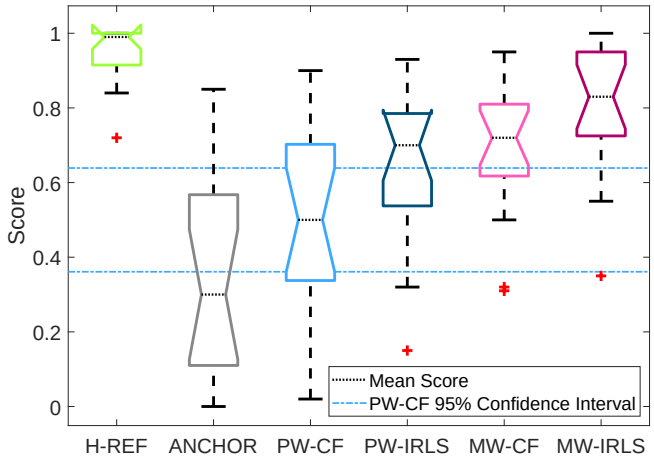
2) Experiment System: We used an Oculus Rift along with a pair of Beyerdynamic DT 770 pro headphones to track the listener and provide a visual reference of the true sound source. We used the HRTFs of the FABIAN head and torso simulator [65] from the HUTUBS dataset [66], [67] for auralization. The HRTFs were rotated for each test signal by multiplying the HRTF coefficients with Wigner-D functions [68]. Signals were processed at a frame size of 4096 with 50% overlap and a 16 kHz sampling frequency, due to hardware constraints and the computational costs of the real-time experiment.

3) Virtual Environment: We simulated the target real-world auditory experience with a single free-field point-source in order to generate a true experiment reference signal for the listener at every position. We constructed a virtual environment with $\boldsymbol{o}$ placed at the center, and the xy-plane 1.25 m above the ground to align with a listener's head while sitting. We modeled the *true target* sound source with a static point-source at $(1, 0, 0)$ m. By *true*, we signify that the sound field generated by this point-source is denoted as the target real-world auditory experience we record and reproduce. Visually, the participants were presented with an infinitely flat landscape to best match the free-field acoustic environment they are experiencing. Two visual references were presented for the higher order microphone (at $\boldsymbol{o}$), and the target sound source.

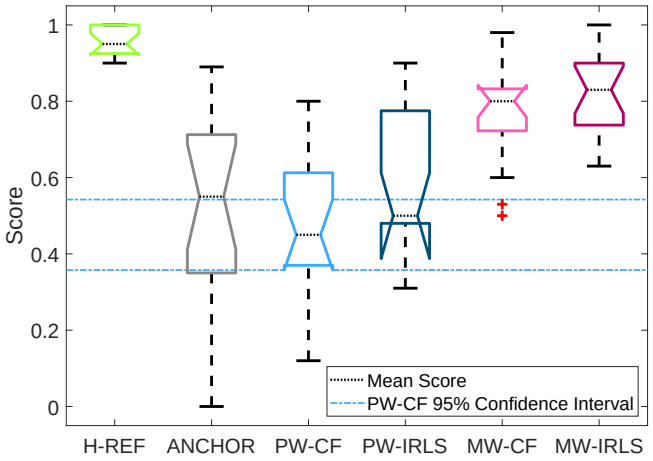
Additionally, we also simulated the process of recording the truncated sound field of the target point-source. We used a 4^{th} order rigid spherical microphone array centered at $\boldsymbol{o}$. Microphone sensors were distributed at Fliege positions [69] with 0.042 m radius to best represent a commercial microphone [32]. Recordings were generated by convolving the sound source's signal with the microphone's impulse response. The $\alpha_{nm}(k)$ coefficients were extracted with (3) before being expanded into virtual distributions.

The planewave distribution consisted of $L = 36$ virtual sources at Fliege positions [69]. This selection was made as a trade-off with computation complexity. However, adding more planewaves is not expected to improve source localization performance, as the distribution already over-samples the 4^{th} order recording [15], [34], [36]. Similarly, the mixedwave distribution consisted of two sets of $L = 36$ virtual sources at the same Fliege positions. The first set was distributed in the near-field at $R_{(\text{nf})} = 2$ m, and the second was placed at $R_{(\text{ff})} = 20$ m in the far-field.

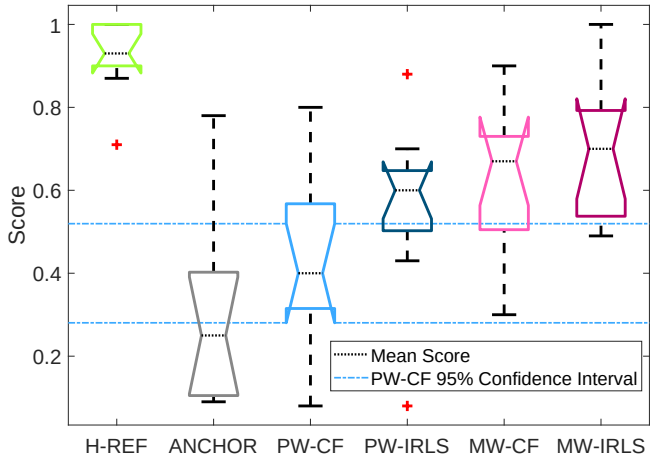
4) Experiment Auralization: The reference was rendered by convolving (in frequency domain) the sound source signal with the target source-to-listener HRTF. For the anchor (4), the signals were convolved by multiplying $\alpha_{nm}(k)$ with the spherical HRTF-coefficients directly [60]. The planewave method signals were rendered with the convolution of the HRTF at $\boldsymbol{o}$ and the phase-shifted driving function (16). The phase-shift was updated with the Oculus head position to render perceptual translation. For the mixedwave methods, the HRTFs were reconstructed between each source and the listener's translated position. Binaural signals were then rendered with



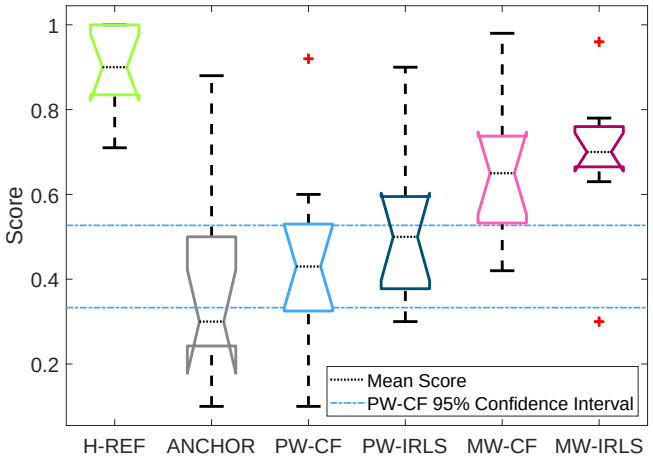
(a) Source localization scores with speech sound-source.



(b) Basic audio quality scores with speech sound-source.



(c) Source localization scores with music sound-source.



(d) Basic audio quality scores with music sound-source.

Fig. 3: Box plot of perception experiment scores for source localization (a) and (c), and basic audio quality (b) and (d). Each box bounds the interquartile range (IQR) with the center bar indicating the median score, and the whiskers extend to a maximum of $1.5 \times$ IQR. The v-shaped notches in the box refer to the 95% confidence interval. When the notches between two boxes do not overlap it can be concluded with 95% confidence that the true medians differ.

the convolution of the mixedwave driving function and the $\boldsymbol{y}_\ell$-to-$\boldsymbol{d}$ HRTF (31).

B. Experiment Results

1) Box Plot: Figure 3 shows the perceptual scores of the translation methods for all four MUSHRA tests. We discuss the results of these scores through an analysis of variance (ANOVA) examination. During the following discussion we use a Tukey-Kramer multiple comparison test with 95% confidence to determine if two methods show a statistically significant difference. That is, a p-value score of $p_{\text{val}} < 0.05$ rejects the ANOVA null hypothesis, and suggests that the difference in score was unlikely to have occurred by chance. We also provide the F-statistic denoted by $F_{(\cdot,\cdot)}$ for completeness. Note that a Lilliefors test (where $p_{\text{val}} > 0.01$) determined that our collected scores were normally distributed, except for the speech-source anchors.

2) One-Factor ANOVA Results: We used a one-factor ANOVA to determine if any of the translation methods performed significantly different in each of the perception tests. For speech localization (Fig. 3a), both MW-CF and MW-IRLS showed a significant improvement in score ($F_{(3,64)} = 6.2, p_{\text{val}} < 0.001$) compared to the PW-CF benchmark. Similar results ($F_{(3,64)} = 16.25, p_{\text{val}} < 0.001$) are shown for speech quality (Fig. 3b), where the mixedwave methods were found to be significantly different to PW-IRLS in addition to the benchmark. In the music sound-source tests (Fig. 3c and Fig. 3d), only MW-IRLS showed significantly improved means over the benchmark, while MW-CF did not. However, MW-CF was still observed to perform well for music localization in Fig. 3c and music quality in Fig. 3d as indicated by the significant median scores.

3) Two-Factor ANOVA Results: We performed a two-factor ANOVA to compare the effects of source-type (planewave

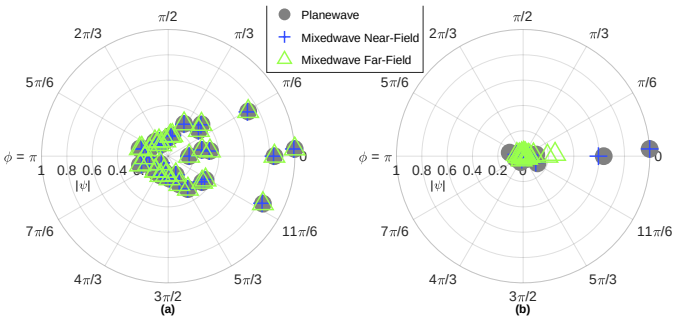


Fig. 4: Normalized activity/magnitude $|\psi|$ of each virtual source averaged over time and frequency for the: (a) closed-form expansions, and (b) sparse IRLS expansions. Note the virtual sources are reconstructing a near-field source at $\phi = 0$.

and mixedwave) and expansion-type (closed-form and IRLS). In all four tests ($p_{\text{val}} \leq 0.008$), mixedwave source distributions were found to score higher means than planewave distributions. Whereas, a significant difference in expansion-type was only found in the speech sound-source tests, with IRLS showing better scores. For music localization ($F_{(1,40)} = 3.36, p_{\text{val}} = 0.074$) and music quality ($F_{(1,40)} = 1.07, p_{\text{val}} = 0.307$), no significant difference was found between closed-form and sparse expansions. Lastly, no interaction effects ($p_{\text{val}} \geq 0.382$) between virtual source-type and expansion-type were found.

4) Summary and Discussion: The proposed MW-IRLS method showed an improvement against the PW-CF benchmark in the perceptual criteria of source localization and audio quality for both a speech and music source. Furthermore, MW-CF also received higher mean scores when reconstructing human speech, and higher median scores for music. When comparing virtual expansion-types, the IRLS expansion was seen to have better quality robustness and localizability for a speech source, but not a music source. This may be explained by the IRLS matching the sparseness of a single human's speech, but not the natural sound of music which is normally generated by multiple sound-sources. Nonetheless, this paper focuses on the modeling of secondary virtual sources. No interaction effect between the source model and expansion-type was found. This indicates that the strong perceptual results achieved by mixedwave methods were not dependent on the expansion-type, and are instead an outcome of the near-field and far-field virtual source mixture. In Section V-B we conduct a simulation analysis on the sound fields used in this experiment to gain further insight on properties that may have influenced these strong perceptual results.

C. The Equivalent Virtual Sources

In Figure 4 we show how the closed-form and sparse expansions utilize the virtual source distributions for the perceptual experiment. For the closed-form expansions (Fig. 4a) we observe that many virtual sources are active in a cone shape towards the direction of the target sound source. Intuitively this behavior would perceptually obscure the location of the reproduced sound source. In contrast, the sparse expansions (Fig. 4b) have fewer active virtual sources that are aligned with the target azimuth direction. This supports the previous source localization results, where the IRLS expansions scored higher than their closed-form counterparts. Furthermore, we observe that both the MW-IRLS and PW-IRLS expansions have activated virtual sources at the same locations. However, the MW-IRLS expansion is able to selectively activate the near-field mixedwave source to match the near-field target source. This suggests that the difference in perceptual results between MW-IRLS and PW-IRLS is predominantly attributed to the modeling of the mixedwave virtual source.

V. Simulation Analysis

In this section, we firstly simulate the same free-field virtual environment that was used in the perception test (Section IV-A3). We examine the pressure and intensity fields to identify factors that may correlate with the perceptual performance results. Secondly, we also examine each sound field translation method in an environment with minor reverberation and few strong acoustic reflections using the image source method.

A. Error Metrics

We define the pressure error (PE) and intensity magnitude error (IME) between the true and reproduced sound field as

$$\left(\text{PE} = \frac{|P - \tilde{P}|^2}{|P|^2}, \quad \text{IME} = \frac{|\boldsymbol{I} - \tilde{\boldsymbol{I}}|^2}{|\boldsymbol{I}|^2}\right) \times 100(\%). \tag{36}$$

where $\boldsymbol{I} = \frac{1}{2}\text{Re}\,(P\boldsymbol{V}^*)$, $\boldsymbol{V}^*$ is the conjugated sound field velocity, and $\tilde{\cdot}$ denotes the reconstructed field. We also study the intensity direction error (IDE) which is linked to human perception of sound source localization [70]. The IDE, which is denoted as the acute angle between the true recorded and reproduced intensity field, is expressed as

$$\text{IDE} = \arccos\left(\frac{\boldsymbol{I} \cdot \tilde{\boldsymbol{I}}}{|\boldsymbol{I}| \cdot |\tilde{\boldsymbol{I}}|}\right) / \pi \times 100(\%). \tag{37}$$

Additionally, for intensity fields, we also illustrate the true and reproduced intensity unit vector difference by $\boldsymbol{I}/|\boldsymbol{I}| - \tilde{\boldsymbol{I}}/|\tilde{\boldsymbol{I}}|$.

B. Pressure and Intensity of Perception Test Environment

1) Pressure and Intensity Fields: We discuss each sound field translation method's pressure field (Fig. 5) and intensity field (Fig. 6) reconstruction performance together. The sound field for the target sound source at $(1, 0, 0)$ m that was recorded and reproduced virtually in the perceptual experiment is shown in Fig. 5a1 and Fig. 6a1. The reconstruction of the recorded 4^{th} order pressure field and intensity field of the target source is given in Fig. 5b1 and Fig. 6b1, respectively. Immediately we observe the consequence of truncation in the measured pressure field, where the distinct near-field wavefront of the target source is lost in the reconstruction of the microphone array recording. As expected, the measurement is seen to be localized spatially within the microphone array, illustrated by the sweet-spot within the measurement PE (Fig. 5b2). Similarly, the measurement's intensity field is also seen to be concentrated about the microphone sweet-spot.

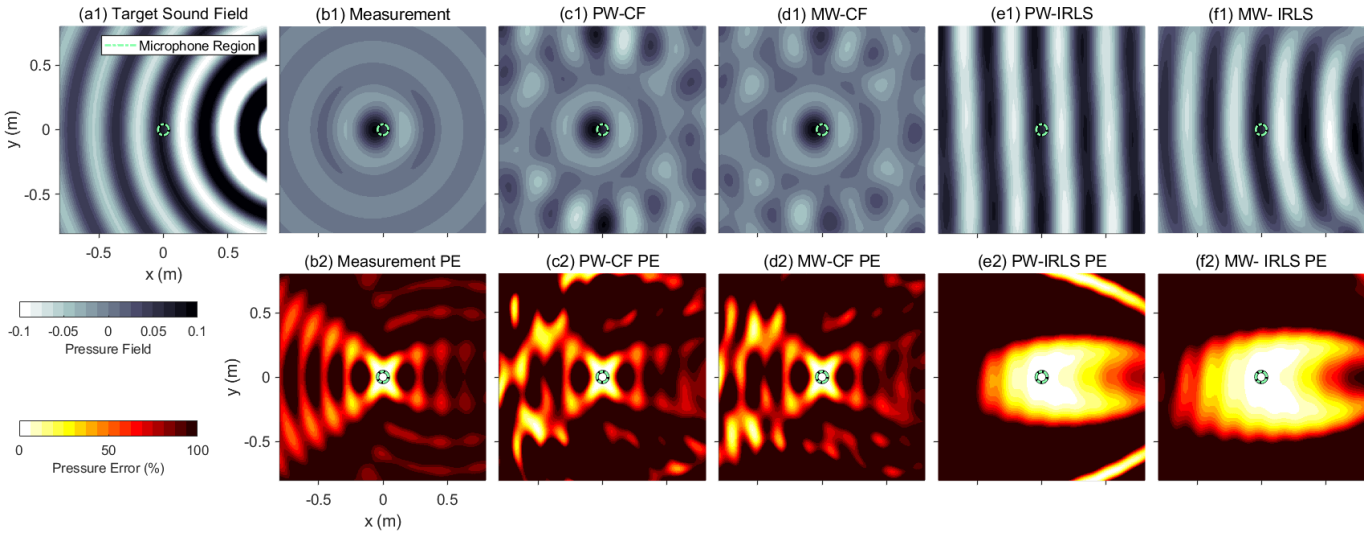


Fig. 5: Pressure fields and pressure errors of the target sound-source and sound field translation method reconstructions at 1000 Hz in the xy-plane, where the target point-source is located at $(1, 0, 0)$ m.

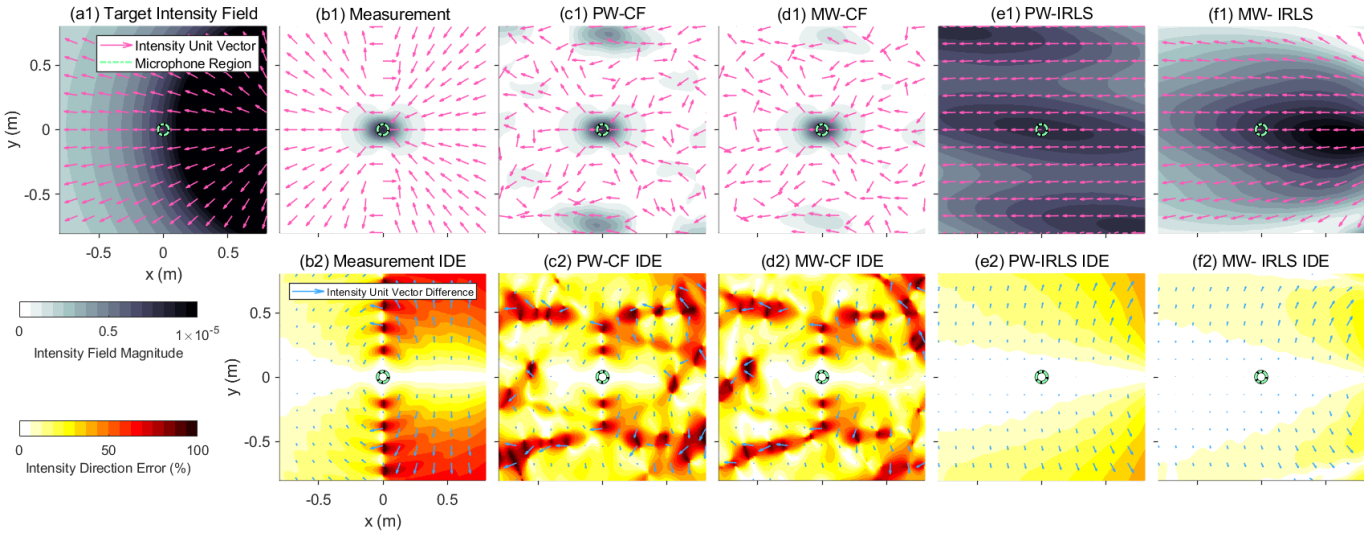


Fig. 6: Intensity fields and intensity direction errors of the target sound-source and sound field translation method reconstructions at 1000 Hz in the xy-plane, where the target point-source is located at $(1, 0, 0)$ m.

Beyond the sweet-spot, truncation error is seen to degrade the measurement's pressure and intensity accuracy, leading to the perceptual artifacts we wish to resolve by extrapolating an equivalent virtual source environment.

We observe that the PW-CF method experiences the same sweet-spot behaviors as the truncated measurement, where once again the reproduced pressure (Fig. 5c) and intensity (Fig. 6c) fields are localized to the microphone's region. A similar result is also obtained by the MW-CF method, supporting that the sweet-spot is produced from the closed-form expansion over-approximating the truncated measurement. The PW-CF and MW-CF intensity fields are also seen to be non-uniform throughout the virtual reconstruction. It is expected that this may be a dominant factor contributing to the poorer perceptual evaluations of the closed-form expansions.

In contrast, the PW-IRLS and MW-IRLS reproductions show better pressure (Fig. 5e/f) and intensity (Fig. 6e/f) results. As intended, the IRLS expansions are seen to relax the sweet-spot constraint and extend the region of reproduction accuracy. This is believed to aid the perceptual stability of the reproduction as the listener translates further from the original recording position. Furthermore, the sparsely expanded intensity fields are shown to be more uniform, leading to better IDE results and likely contributing to stronger perceptual evaluations. We observe that the MW-IRLS method produces a better reconstruction than the PW-IRLS method. This is due to the mixedwave virtual source model producing a spherical near-field wavefront in the pressure field (Fig. 5f1), which also leads to better IDE performance (Fig. 6f2). Finally, we take note of the discrepancy between the MW-CF method's good perceptual evaluations and its poor sound field reconstruction in comparison to the PW-IRLS method. We attribute this discrepancy to the difference in the binaural rendering of each sound field, which we will further examine in Section V-B3.

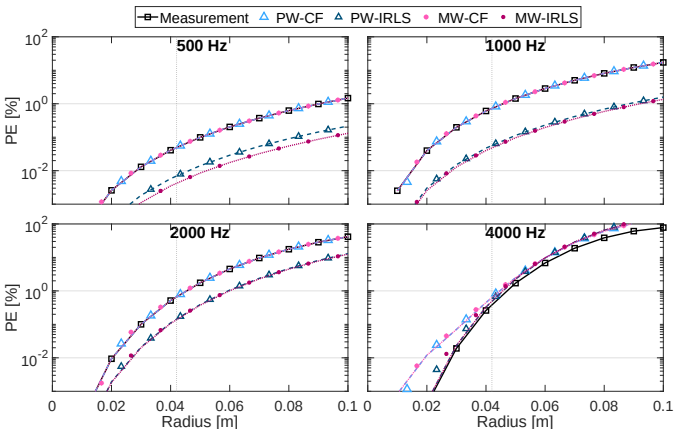


Fig. 7: Average pressure error of reconstruction over a spherical surface of varying radius at four frequencies for the measured and reproduced sound fields.

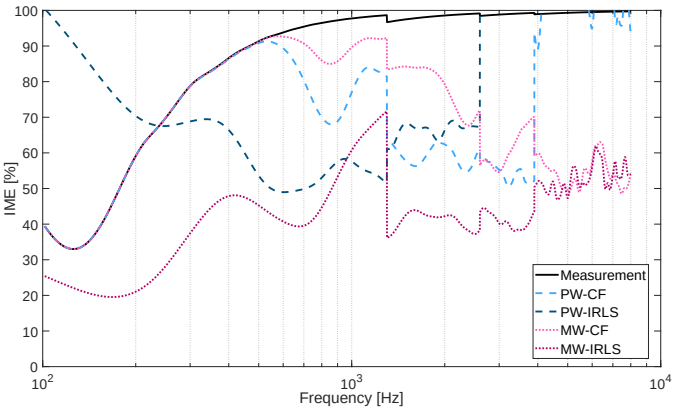


Fig. 8: Average intensity magnitude error of reconstruction over a 0.8 m spherical surface plotted against frequency.

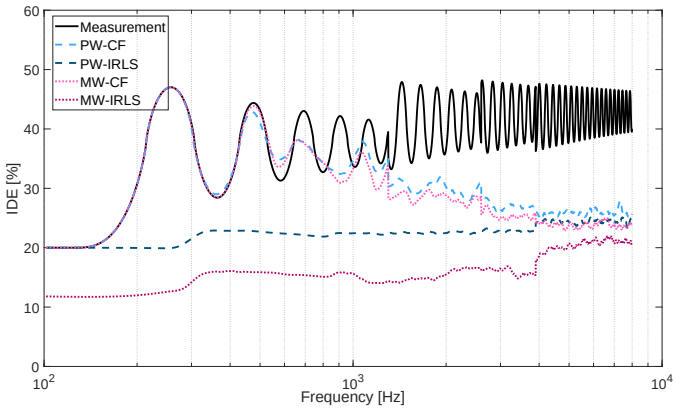


Fig. 9: Average intensity direction error of reconstruction over a 0.8 m spherical surface plotted against frequency.

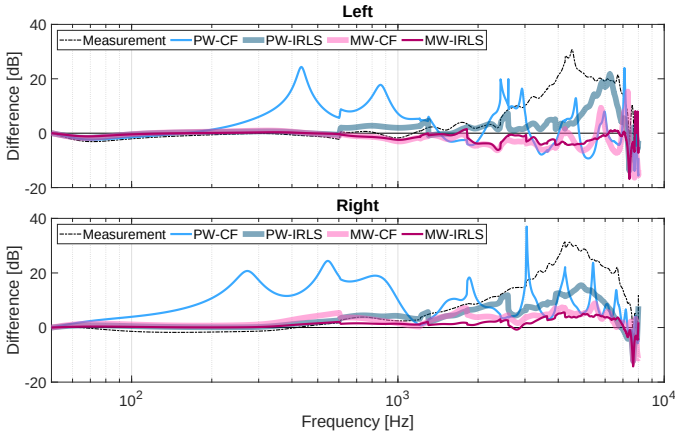


Fig. 10: BRIR spectral difference between the true (reference) and reproduced signals for a translated listener positioned at $(0, 0.5, 0)$ m facing parallel to the x-axis $(\theta = \pi/2, \phi = 0)$.

2) Pressure and Intensity Error: We present the averaged PE at various translation distances in Fig. 7. A clear difference in performance is observed at the lower frequencies, where the two IRLS expansions (PW-IRLS and MW-IRLS) are seen to better reproduce the pressure field throughout a 0.1 m region. This result corroborates with the prior sweet-spot observations, where the IRLS expansions are able to relax spatial constraints. On the other hand, the closed-form expansions are shown to match the PE of the measurement, further illustrating that the PW-CF and MW-CF methods over-approximate the recording's truncation artifacts.

All methods are observed to have poor IME at the translation of 0.8 m in Fig. 8. At higher frequencies both MW-CF and MW-IRLS have lower error than their planewave counterparts. However, the IME is still poor, and it is difficult to know if this behavior contributed to perceptual results. Additionally, large spikes in error are found when the microphone's truncation increases between the $\lceil k|\boldsymbol{x}_Q|\rceil$ frequency bands. It may be possible to smooth the activation of each band to further improve perceptual stability.

The IDE shows clearer results at the 0.8 m translation in Fig. 9. The MW-IRLS reproduction is seen to strongly match the direction of the true sound-source's intensity across the full frequency range. This intensity alignment is expected to have contributed to the perceptual results of the MW-IRLS method. This is in contrast to the PW-CF benchmark which is seen to follow the recording's poor IDE at lower frequencies.

3) BRIR Response: Here we measure the system response of each sound field translation method by recording, expanding, and auralizing a sine-sweep signal for a translated listener positioned at $(0, 0.5, 0)$ m. The listener's looking direction is fixed forward parallel to the x-axis $(\theta = \pi/2, \phi = 0)$ such that the sound-source is roughly 1 m from and $60°$ to the right of the listener. This system response can be seen as the binaural room impulse response (BRIR) of the translated listener inside the reconstructed virtual environment. Furthermore, unlike the prior pressure and intensity results, the rendering of the listener's HRTF is included in the BRIR result. Therefore, the BRIR provides insight into how well each translation method reconstructs head reflections that match the original listening experience. In Fig. 10 we give the spectra difference of BRIR between each translation method compared to the reference (the free-field point source convolved with HRTF) for this single listener position and orientation. The BRIR spectral results show the most substantial difference between

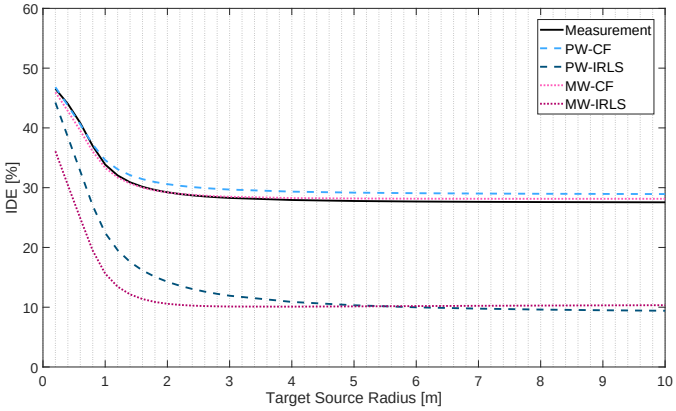


Fig. 11: Average intensity direction error of each method's reconstruction over a 0.8 m spherical surface at 1000 Hz given a target point-source at different recorded distances.

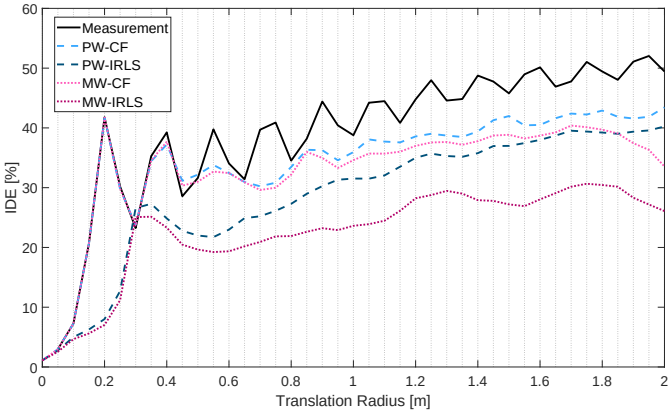


Fig. 12: Average intensity direction error of reconstruction over a spherical surface of varying radius at 1000 Hz in an environment with minor reverberation $T_{60} = 190$ ms.

the PW-CF benchmark and the mixedwave methods thus far. Below 1000 Hz the PW-CF BRIR is seen to deviate significantly from the reference. At higher frequencies all translation methods show some divergence from the reference, however, the mixedwave methods still exhibit smaller BRIR distortions. Between the two closed-form expansions, the MW-CF shows significantly better BRIR accuracy. This is likely due to the finitely distanced mixedwave sources allowing for a more accurate HRTF translation (discussed in Section III-F). Overall, the MW-IRLS method most accurately reconstructs the sound heard by a translated listener in the true environment. As such, the BRIR results suggest that the MW-IRLS offers greater perceptual accuracy, which is in agreement with the perceptual experiment results.

C. Varying Distance of the Recorded Target Sound-Source

The mixedwave virtual source model improves upon the planewave model in the case of reconstructing a near-field target source. The mixedwave method's performance should not be significantly worse than the planewave method when the target source is in the far-field. Figure 11 shows each sound field translation method's IDE when reconstructing a target point-source recorded at different distances. Indeed, we observe the MW-IRLS method to improve upon the PW-IRLS method for a near-field target source. Best performance occurs in the sparse expansion when the target source distance is close to the mixedwave virtual source inner-sphere radius at 2 m. It is expected that performance for near-field target sources at < 1 m distances can be improved by reducing the inner-sphere radius of the mixedwave model, however, at the expense of a reduced maximum translation distance. Furthermore, we observe that the MW-IRLS method is comparable to the PW-IRLS method for far-field target sources. This suggests that the mixedwave source provides near-field benefits to sound field translation without any likely trade-off to far-field reproduction performance.

D. Robustness to Environment Reverberation

Here, we briefly examine each sound field translation method in an environment with minor reverberation and two strong reflections. Our intention is to investigate each translation method's robustness to reverberant noise. We want to reiterate that we do not consider using the proposed method for modeling highly reverberant environments in this paper. If early reflections and diffuse reverberation are present in the original recording, then we would intuitively recommend that these properties of the sound are carefully handled separately during reproduction. In such a case, we expect that the sound source's direct path can be isolated and processed with the proposed sound field translation method. Moreover, the reverberation can be isolated and removed to be processed separately or be substituted by an artificial environment during reproduction. Nonetheless, for this paper we consider the proposed method to be applied in large open environments where reverberation is negligible.

1) The Reverberant Environment: was modeled by a rectangular room with minor reverberation using the image source method [71]. Room dimensions were 8 m, 6 m, 3 m and wall reflection coefficients were $(0.65, 0.65, 0.6, 0.6, 0.4, 0.6)$, to give a reverberation time of $T_{60} = 190$ ms using the Sabine formula [72]. We centered the higher order microphone in the room, and placed the target true sound source at $(1, 0, 0)$ m with respect to this microphone. Additionally, we placed two mirrored point sources at positions $(1, 6, 0)$ m and $(1, \text{-}6, 0)$ m with an amplitude gain of 0.975 to model two strong acoustic reflections.

2) Intensity Direction Error Against Translation Radius: is given in Fig. 12 for the reverberant environment. We observe that both the PW-IRLS and MW-IRLS methods perform better than their closed-form counterparts for shorter translation distances. This suggests that sparsity is the main contributor to short translation performance. At larger translations (>0.5 m), however, the MW-IRLS shows significantly better intensity reconstruction than the PW-IRLS method, indicating that the mixedwave source provides greater perceptual performance

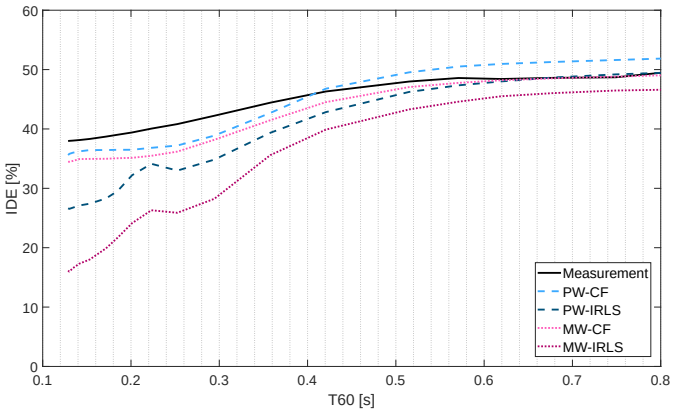


Fig. 13: Average intensity direction error of reconstruction over a 1 m spherical surface at 1000 Hz in an environment with minor reverberation of varying T_{60} reverberation time.

compared to the planewave source. This result aligns with the conclusions drawn from the perceptual experiment. Furthermore, we observe that the MW-IRLS IDE score remains between approximately 20% to 30% throughout a 2 m translation. This result is comparable to the 20% IDE score achieved in the free-field environment (Fig. 9), suggesting that the MW-IRLS method is somewhat robust to minor reverberation.

3) Intensity Direction Error Against Reverberation Time: is shown in Fig. 13. For this result we have adjusted the image source method wall reflection coefficients of our considered environment to obtain various T_{60} reverberation times. The room size and strong reflecting mirror-sources remain unchanged. We clearly observe that the proposed MW-IRLS method performs best in low reverberation. The contrasting IDE performance between the MW-IRLS and the PW-IRLS methods demonstrates the mixedwave source's better intensity field reconstruction in low reverberant environments. At higher reverberation (T_{60}>400 ms) all sound field translation methods converge to the performance of the truncated measurement. This suggests that both the benchmark and the proposed methods are susceptible to highly reverberant environments, and therefore, perceptual performance is expected to be poor during reconstruction.

VI. Conclusion

Virtual reality technology enhances acoustic real-world reproductions by allowing listeners to perceptually move about the environment. At this time, however, the benchmark planewave method towards sound field translation is still limited by inherited microphone constraints. Furthermore, the planewave source model is restricted to the far-field, which results in the listener's HRTF perspective being fixed during translation. As a result, immersion in the planewave environment is degraded by poor source localizability and audible spectral distortions. We have proposed an alternative source model for sound field translation that enables a sparse virtual environment to contain a mixture of near-field and far-field sources. We compared this proposed mixedwave method against the planewave benchmark through a perceptual MUSHRA experiment and cross-examined the results with numerical simulations. For human speech reproduction, the mixedwave source model improved the perceptual source localizability and audio quality. Both the closed-form and sparsely expanded mixedwave reproductions were found to provide a more immersive experience. Similar results were also found for a music sound source. The sparse expansion was shown to help enlarge the reproduction sweet-spot, and activate the near-field virtual sources to match the near-field target source. The closed-form expansion also benefited from the mixedwave source model, as the finitely distanced sources allowed the relative angle in the HRTF to update with listener movement. This was illustrated by the lower BRIR spectral error achieved by the mixedwave binaural rendering. Finally, the proposed method better matched the target intensity direction at differing frequencies, translation distances, and under low reverberation noise; which further corroborates our perceptual experiment results.

We note that this paper focuses on the perceptual effects of modeling a near-field far-field mixture for sound field translation. As such, we have only studied an over-simplified acoustic environment to gain clear insight into the perceptual attributes of the mixedwave method. Future work is required to achieve a complete virtual application with satisfying performance and human interactivity. Firstly, the modeling of multiple sound sources should be investigated. Secondly, algorithms to separate the direct sound of target sources from a reverberant environment are required. This then enables the complex reverberant environment to be treated explicitly with its own model. Or, alternatively, it might be that replacing the reverberation with a synthetic environment provides a more satisfying experience for the listener. Thirdly, algorithms to estimate and model the directivity of the sound source, and using high resolution personalized HRTFs are both required for more accurate perception. Finally, for virtual applications it may be desirable to add synthetic objects to the environment, and therefore a method to update the perceptual experience to match the addition of objects that were not present during the original recording is required. It is important to evaluate the perceptual attributes of each component in the virtual application, which is what we have provided for the modeling of a target sound source with the proposed mixedwave sound field translation method.

VII. Acknowledgment

The authors would like to thank Zamir Ben-Hur for guidance in developing the perceptual test, and Shawn Featherly for the development of the perceptual test Unity application.

References

[1] P. Dodds, S. Amengual Garí, W. Brimijoin, and P. Robinson, "Auralization systems for simulation of augmented reality experiences in virtual environments," in *Audio for Virtual, Augmented and Mixed Realities: Proc. ICSA 2019; 5th Intl. Conf. on Spatial Audio*, 2019, pp. 29–34.

[2] Y. Suzuki *et al.*, "3d spatial sound systems compatible with human's active listening to realize rich high-level kansei information," *Interdisciplinary information sciences*, vol. 18, no. 2, pp. 71–82, 2012.

[3] J. G. Tylka and E. Y. Choueiri, "Models for evaluating navigational techniques for higher-order ambisonics," in *Proc. Meetings on Acoust.* ASA, 2017, vol. 30, p. 050009.

[4] S. Amengual Garí, C. Schissler, R. Mehra, S. Featherly, and P. Robinson, "Evaluation of real-time sound propagation engines in a virtual reality framework," in *Proc. Intl. Audio Eng. Soc. Conf. on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[5] M. Ziegler *et al.*, "Immersive virtual reality for live-action video using camera arrays," *IBC, Amsterdam, Netherlands*, 2017.

[6] D. Rivas Méndez, C. Armstrong, J. Stubbs, M. Stiles, and G. Kearney, "Practical recording techniques for music production with six-degrees of freedom virtual reality," in *Audio Eng. Soc. Conv. 145*. Audio Engineering Society, 2018.

[7] C. D. Salvador, S. Sakamoto, J. Trevino, and Y. Suzuki, "Spatial accuracy of binaural synthesis from rigid spherical microphone array recordings," *Acoust. Sci. Technol.*, vol. 38, no. 1, pp. 23–30, 2017.

[8] E.G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*, Academic Press, London, UK, 1999.

[9] G. H. Koopmann, L. Song, and J. B. Fahnline, "A method for computing acoustic fields based on the principle of wave superposition," *J. Acoust. Soc. Amer.*, vol. 86, no. 6, pp. 2433–2438, 1989.

[10] A. Sarkissian, "Method of superposition applied to patch near-field acoustic holography," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 671–678, 2005.

[11] M. E. Johnson, S. J. Elliott, K. H. Baek, and J. Garcia-Bonito, "An equivalent source technique for calculating the sound field inside an enclosure containing scattering objects," *J. Acoust. Soc. Amer.*, vol. 104, no. 3, pp. 1221–1231, 1998.

[12] N. P. Valdivia and E. G. Williams, "Study of the comparison of the methods of equivalent sources and boundary element methods for near-field acoustic holography," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 3694–3705, 2006.

[13] E. Fernandez-Grande, A. Xenaki, and P. Gerstoft, "A sparse equivalent source method for near-field acoustic holography," *J. Acoust. Soc. Amer.*, vol. 141, no. 1, pp. 532–542, 2017.

[14] J. G. Tylka and E. Y. Choueiri, "Fundamentals of a parametric method for virtual navigation within an array of ambisonics microphones," *J. Audio Eng. Soc.*, vol. 68, no. 3, pp. 120–137, 2020.

[15] J. G. Tylka and E. Y. Choueiri, "Performance of linear extrapolation methods for virtual sound field navigation," *J. Audio Eng. Soc.*, vol. 68, no. 3, pp. 138–156, 2020.

[16] N. Mariette and B. Katz, "Sounddelta—large scale multi-user audio augmented reality," in *Proc. of the EAA Symposium on Auralization*, 2009, pp. 15–17.

[17] J. G. Tylka and E. Y. Choueiri, "Domains of practical applicability for parametric interpolation methods for virtual sound field navigation," *J. Audio Eng. Soc.*, vol. 67, no. 11, pp. 882–893, 2019.

[18] E. Patricio, A. Ruminski, A. Kuklasinski, L. Januszkiewicz, and T. Zernicki, "Toward six degrees of freedom audio recording and playback using multiple ambisonics sound fields," in *Audio Eng. Soc. Conv. 146*. Audio Engineering Society, 2019.

[19] P. Samarasinghe, T. Abhayapala, and M. Poletti, "Wavefield analysis over large areas using distributed higher order microphones," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 3, pp. 647–658, 2014.

[20] J. G. Tylka and E. Choueiri, "Soundfield navigation using an array of higher-order ambisonics microphones," in *Proc. Intl. Audio Eng. Soc. Conf. on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2016.

[21] Y. Wang and K. Chen, "Translations of spherical harmonics expansion coefficients for a sound field using plane wave expansions," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. 3474–3478, 2018.

[22] O. Thiergart, G. Del Galdo, M. Taseska, and E. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 12, pp. 2583–2594, 2013.

[23] J. G. Tylka and E. Choueiri, "Comparison of techniques for binaural navigation of higher-order ambisonic soundfields," in *Audio Eng. Soc. Conv. 139*. Audio Engineering Society, 2015.

[24] M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich, "A 3d ambisonic based binaural sound reproduction system," in *Proc. of 24th Intl. Audio Eng. Soc. Conf. on Multichannel Audio, The New Reality*. Audio Engineering Society, 2003.

[25] D. Menzies and M. Al-Akaidi, "Ambisonic synthesis of complex sources," *J. Audio Eng. Soc.*, vol. 55, no. 10, pp. 864–876, 2007.

[26] T. Pihlajamaki and V. Pulkki, "Synthesis of complex sound scenes with transformation of recorded spatial sound in virtual reality," *J. Audio Eng. Soc.*, vol. 63, no. 7/8, pp. 542–551, 2015.

[27] E. Fernandez-Grande, "Sound field reconstruction using a spherical microphone array," *J. Acoust. Soc. Amer.*, vol. 139, no. 3, pp. 1168–1178, 2016.

[28] F. Schultz and S. Spors, "Data-based binaural synthesis including rotational and translatory head-movements," in *Proc. of 52nd Intl. Audio Eng. Soc. Conf. on Sound Field Control-Engineering and Perception*. Audio Engineering Society, 2013.

[29] R. Duraiswami, Z. Li, D. N. Zotkin, E. Grassi, and Nail A. G., "Plane-wave decomposition analysis for spherical microphone arrays," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2005, pp. 150–153.

[30] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025, 2005.

[31] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 697–707, 2001.

[32] MH Acoustics, "Em32 eigenmike microphone array release notes (v17. 0)," *25 Summit Ave, Summit, NJ 07901, USA*, 2013.

[33] N. Hahn and S. Spors, "Modal bandwidth reduction in data-based binaural synthesis including translatory head-movements," in *Proc. German Annu. Conf. Acoust.(DAGA)*, 2015, pp. 1122–1125.

[34] N. Hahn and S. Spors, "Physical properties of modal beamforming in the context of data-based sound reproduction," in *Audio Eng. Soc. Conv. 139*. Audio Engineering Society, 2015.

[35] A. Kuntz and R. Rabenstein, "Limitations in the extrapolation of wave fields from circular measurements," in *Eur. Signal Process. Conf.*, 2007, pp. 2331–2335.

[36] F. Winter, F. Schultz, and S. Spors, "Localization properties of data-based binaural synthesis including translatory head-movements," in *Proceedings of the Forum Acusticum, Krakow, Poland*, 2014, vol. 31.

[37] Zylia Sp. z o.o., "ZYLIA ZM-1 Microphone," *https://www.zylia.co/zylia-zm-1-microphone.html*, accessed: Feb. 2020.

[38] VisiSonics Corporation, "Visisonics 5/64 audio/visual camera," *https://visisonics.com/564avcamera/*, accessed: Feb. 2020.

[39] S. Koyama and L. Daudet, "Sparse representation of a spatial sound field in a reverberant environment," *IEEE J. Selected Topics Signal Process.*, vol. 13, no. 1, pp. 172–184, 2019.

[40] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," in *Proc. of 23rd Intl. Audio Eng. Soc. Conf. on Signal Processing in Audio Recording and Reproduction*. Audio Engineering Society, 2003.

[41] D. Menzies and M. Al-Akaidi, "Nearfield binaural synthesis and ambisonics," *J. Acoust. Soc. Amer.*, vol. 121, no. 3, pp. 1559–1563, 2007.

[42] K. Wakayama, J. Trevino, H. Takada, S. Sakamoto, and Y. Suzuki, "Extended sound field recording using position information of directional sound sources," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 185–189.

[43] A. Plinge, S. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. Habets, "Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information," in *Proc. Intl. Audio Eng. Soc. Conf. on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

[44] M. Kentgens, A. Behler, and P. Jax, "Translation of a higher order ambisonics sound scene based on parametric decomposition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 151–155.

[45] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *IEEE Intl. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 3869–3872.

[46] P. Simard and J. Antoni, "Acoustic source identification: Experimenting the $\ell 1$ minimization approach," *Appl. Acoust.*, vol. 74, no. 7, pp. 974–986, 2013.

[47] E. Fernandez-Grande and A. Xenaki, "Compressive sensing with a spherical microphone array," *J. Acoust. Soc. Amer.*, vol. 139, no. 2, pp. EL45–EL49, 2016.

[48] S. Emura, "Sound field estimation using two spherical microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 101–105.

[49] Y. Hu, P. N. Samarasinghe, T. D. Abhayapala, and G. Dickins, "Modeling characteristics of real loudspeakers using various acoustic models: Modal-domain approaches," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 561–565.

[50] Y. Maeno, Y. Mitsufuji, and T. D. Abhayapala, "Mode domain spatial active noise control using sparse signal representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 211–215.

[51] L. Birnie, T. Abhayapala, P. Samarasinghe, and V. Tourbabin, "Sound field translation methods for binaural reproduction," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 140–144.

[52] ITU Radiocommunication Assembly, "ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," October 2015.

[53] Z. Ben-Hur, D. L. Alon, B. Rafaely, and R. Mehra, "Loudness stability of binaural sound with spherical harmonic representation of sparse head-related transfer functions," *EURASIP J. Audio, Speech, Music Process.*, vol. 2019, no. 1, pp. 5, 2019.

[54] H. Chen, T. D. Abhayapala, and W. Zhang, "Theory and design of compact hybrid microphone arrays on two-dimensional planes for three-dimensional soundfield analysis," *J. Acoust. Soc. Amer.*, vol. 138, no. 5, pp. 3081–3092, 2015.

[55] VisiSonics Corporation, "Visisonics audio/visual planar array," *https://visisonics.com/audio-visual-planar-array/*, accessed: Feb. 2020.

[56] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 2, pp. II–1949.

[57] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, 2005.

[58] R. A. Kennedy, P. Sadeghi, T. D. Abhayapala, and H. M. Jones, "Intrinsic limits of dimensionality and richness in random multipath fields," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2542–2556, 2007.

[59] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized hrtf fitting using spherical harmonics," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2009, pp. 257–260.

[60] W. Zhang, T. D. Abhayapala, R. A. Kennedy, and R. Duraiswami, "Insights into head-related transfer function: Spatial dimensionality and continuous representation," *J. Acoust. Soc. Amer.*, vol. 127, no. 4, pp. 2347–2357, 2010.

[61] G. N. Lilis, D. Angelosante, and G. B. Giannakis, "Sound field reproduction using the lasso," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 8, pp. 1902–1912, 2010.

[62] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[63] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[64] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic data consortium*, vol. 10, no. 5, pp. 0, 1993.

[65] A. Lindau, T. Hohn, and S. Weinzierl, "Binaural resynthesis for comparative studies of acoustical environments," in *Audio Eng. Soc. Conv. 122*. Audio Engineering Society, 2007.

[66] F. Brinkmann *et al.*, "A cross-evaluated database of measured and simulated hrtfs including 3d head meshes, anthropometric features, and headphone impulse responses," *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 705–718, 2019.

[67] F. Brinkmann *et al.*, "The hutubs head-related transfer function (hrtf) database," *[online]. http://dx.doi.org/10.14279/depositonce-8487*, accessed: Feb. 2020.

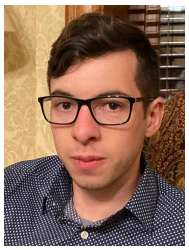
[68] B. Rafaely and M. Kleider, "Spherical microphone array beam steering using wigner-d weighting," *IEEE Signal Process. Lett.*, vol. 15, pp. 417–420, 2008.

[69] J. Fliege and U. Maier, "The distribution of points on the sphere and corresponding cubature formulae," *IMA J. Numer. Anal.*, vol. 19, no. 2, pp. 317–334, 1999.

[70] M. Shin, P. A. Nelson, F. M. Fazi, and J. Seo, "Velocity controlled sound field reproduction by non-uniformly spaced loudspeakers," *Journal of Sound and Vibration*, vol. 370, pp. 444–464, 2016.

[71] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[72] E. A. Lehmann, A. M. Johansson, and S. Nordholm, "Reverberation-time prediction method for room impulse responses simulated with the image-source model," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2007, pp. 159–162.



Lachlan Birnie received the B.E. (Hons.) degree with double major in electronic and communication systems, mechanical and material systems, in 2017, from the Australian National University (ANU), Canberra, Australia; where he is currently working towards the Ph.D. degree in spatial audio signal processing. His research interests include spatial sound field capture and reproduction using higher order microphone and loudspeaker arrays, and binaural reproduction for virtual reality applications.

Thushara Abhayapala (Senior Member, IEEE) received the B.E. degree in engineering and the Ph.D. degree in telecommunications engineering from the Australian National University (ANU), Canberra, Australia, in 1994 and 1999, respectively. Currently he is a Professor at the ANU. He has held a number of leadership positions including Deputy Dean of the ANU College of Engineering and Computer Science (2015–2019), Head of the ANU Research School of Engineering (2010–2014), and Leader of the Wireless Signal Processing Program with the National ICT Australia from 2005–2007. His research interests include areas of spatial audio and acoustic signal processing, and multichannel signal processing. Among many contributions, he is one of the first researchers to use spherical harmonic based Eigen-decomposition in microphone arrays and to propose the concept of spherical microphone arrays, and was one of the first to show the fundamental limits of spatial sound field reproduction using arrays of loudspeakers and spherical harmonics. He has active collaborations with a number of companies. He was an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and was a Member of the Audio and Acoustic Signal Processing Technical Committee (2011–2016) of the IEEE Signal Processing Society. He is a Fellow of Engineers Australia (IEAust).

Vladimir Tourbabin (M'16) received the B.Sc. degree (summa cum laude) in materials science and engineering in 2005, the M.Sc. degree (cum laude) in electrical and computer engineering in 2011, and the Ph.D. degree in electrical and computer engineering in 2016. All three degrees are from Ben-Gurion University of the Negev, Israel. After graduation, he joined the General Motors' Advanced Technical Center in Israel, to work on microphone array processing solutions for speech recognition. Since 2017, Dr. Tourbabin is with Facebook Reality Labs Research (formerly known as Oculus Research) working on research and advanced development of audio signal processing technologies for augmented and virtual reality applications.

Prasanga Samarasinghe (Senior Member, IEEE) received the B.E. (Hons.) degree in electronic and electrical engineering from the University of Peradeniya, Peradeniya, Sri Lanka, in 2009, and the Ph.D. degree from the Australian National University (ANU), Canberra, Australia, in 2014. She is currently a Research Fellow with the College of Engineering and Computer Science, ANU. Her research interests include spatial sound recording and reproduction, spatial noise cancellation, and array optimization using compressive sensing techniques.