# A UNIFIED APPROACH TO NUMERICAL AUDITORY SCENE SYNTHESIS USING LOUDSPEAKER ARRAYS

*Joshua Atkins, Ismael Nawfal, Daniele Giacobello*

Beats Electronics, LLC, 8600 Hayden Place, Culver City, CA 90232
{Josh.Atkins, Ismael.Nawfal, Daniele.Giacobello}@beatsbydre.com

## ABSTRACT

In this work we address the problem of simulating the spatial and timbral cues of a given sound event, or *auditory scene*, using an array of loudspeakers. We first define the problem with a general numerical framework that encompasses many known techniques from physical acoustics, crosstalk cancellation, and acoustic control. In contrast to many previous approaches, the system described in this work is inherently broadband as it jointly designs a set of spatio-temporal filters while allowing for constraints in other domains. With this framework we show similarities and differences between known techniques and suggest some new, unexplored methods. In particular, we focus on perceptually motivated choices for the cost function and regularization. These methods are then compared by implementing the systems on a linear array of loudspeakers and evaluating the timbral and spatial qualities of the system using objective metrics.

***Index Terms***— spatial audio, crosstalk cancellation, binaural hearing, equivalent source method, mode-matching

## 1. INTRODUCTION

The control of the sound field produced from an array of loudspeakers has many interesting applications in acoustics such as room correction, spatial sound reproduction, active noise control, assisted reverberation, quiet zone generation, and focused sound reproduction [1]. While seemingly different problems, many of these disparate goals are tackled in the literature by solving the same type of numerical optimization problem. The conventional techniques aim to generate a set of filters for each loudspeaker in an array by minimizing the $\ell_2$-norm of the error between the reproduced field and a target response at a few listening positions. Many times this is done in a narrow-band sense, focusing on a single frequency and the generation of a broadband filter is left as a separate optimization problem.

The literature on the spatial audio reproduction problem is divided into four areas: attempts at accurate reproduction of a wave-field (e.g., wave-field synthesis (WFS) [2] or near-field compensated higher-order ambisonics (NFC-HOA) [3]), attempts at accurate binaural reproduction at a particular listening position (e.g., crosstalk cancellation [4] or loudspeaker-binaural rendering [5]), attempts to reproduce the perceptual attributes of a sound field using heuristic approaches (e.g., vector-base amplitude panning (VBAP) [6]), and numerical approaches to reconstructing a sound-field (e.g., mode-matching for HOA [7] and the equivalent source method (ESM) [8]).

While the WFS and HOA approaches can accurately reproduce a wave-field up to a spatial aliasing frequency in a specific region in space using analytic solutions to the wave equation, they are limited in usefulness due to strict constraints on both loudspeaker locations and the region in which the wave-field is accurately reproduced [9]. In contrast, numerical approaches allow flexibility in both the loudspeaker location and reproduction region while also supporting flexible models of acoustical propagation between loudspeakers and intended listening locations [1]. The mode-matching approach for HOA reproduction is a numerical framework that allows for flexible loudspeaker layouts by minimizing error in the wavenumber-domain [7]. Similarly, the WFS analog for numerical reproduction, formulated by discretizing the Kirchhoff-Helmholtz integral in the spatial domain, is formatted in the ESM problem [8]. While not explicitly viewed as the same problem in the past, crosstalk cancellation and numerical sound field control share the same framework, differing only in their models of acoustical propagation and target.

In this paper we provide a unified numerical framework for recreating the spatial and timbral characteristics of a virtual auditory scene for one or more human listeners given an array of loudspeakers. We define this problem as *numerical auditory scene synthesis* (NASS) to differentiate the approach from sound field synthesis where the goal is exact reproduction of a physical wave-field in a spatial region. The differences are subtle in some cases, but the NASS viewpoint allows for the incorporation of perceptual considerations in the evaluation and generation of the auditory scene, a necessity when the number or position of loudspeakers does not meet strict requirements necessary for non-numerical approaches. The NASS approach here allows for both clear links between many disparate areas of spatial audio reproduction and leads to many relevant solutions that have been unexplored in the past. While the most generic NASS optimization problem would incorporate a model of both the peripheral and central auditory system [10], we focus this work on convex constraints and cost functions that are known to have unique solutions leading to feasible systems. The goal is then to find objective criteria that fit within this mathematical framework yet map well to perceptual features like localization, loudness, timbre, and spatial extent. In contrast to much of the previous work in sound-field reproduction, we consider the design in a broadband sense and jointly design a set of spatio-temporal filters while still allowing for constraints in other domains.

In Section 2, we review various models of acoustical propagation. The unifying numerical method, which allows for flexible constraints and arbitrary spatio-temporal transforms, is then presented in Section 3. In Section 4, we consider the perceptual and acoustical implications of various spatio-temporal transforms. In Section 5, we present a case study to provide a proof of concept of the flexibility of the framework presented. In particular, we use different extensions of the NASS framework to render a source using a few common loudspeaker arrangements.

## 2. MODELS OF ACOUSTICAL PROPAGATION

In this work, we consider the model of acoustical propagation abstractly; the model can be an analytic one such as a plane-wave, spherical-wave, or multi-pole source, a measured anechoic head-related impulse response (HRIR), a measured HRIR in a room (i.e. a binaural room impulse response (BRIR)), a measured loudspeaker response in a room or anechoic setting or any other measured acoustic impulse response.

In the analytic cases we can evaluate solutions to the acoustic wave equation under specific boundary conditions [11]. For a plane-wave source in an anechoic setting (free-field), the response at any point in space can be described by

$$G(f) = Ae^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \xrightarrow{\mathcal{F}^{-1}} g(t) = A\delta\left(\frac{\mathbf{n}\cdot\mathbf{r}}{c} - t\right), \qquad (1)$$

where $k = \frac{\omega}{c}$ is the wavenumber, $\mathbf{k} = k\mathbf{n}$ is the wavenumber vector with unit vector $\mathbf{n}$ pointing in the direction of propagation, $c$ is the speed of sound in the medium (approx. 343 m/s in air), $\omega = 2\pi f$ is the frequency ($f$ is in Hertz), $i$ is the imaginary number, $\mathbf{r}$ is the vector pointing from the origin to the evaluation point, $\delta(\cdot)$ is the Dirac delta function, and $\mathcal{F}^{-1}$ represents the inverse Fourier transform. For a monopole source in an anechoic setting (free-field), the response at any point in space is described by a spherically spreading wave

$$G(f) = \frac{Ae^{i(kr-\omega t)}}{r} \xrightarrow{\mathcal{F}^{-1}} g(t) = \frac{A}{r}\delta\left(\frac{r}{c} - t\right), \qquad (2)$$

where $r = |\mathbf{r}-\mathbf{r}'|$ is the distance from source to evaluation point and $\mathbf{r}'$ is the vector pointing from the origin to the source location. More general spatial responses can be described by a multi-pole source, described in free-field as as sum of spatially distributed monopole sources.

In practice, these impulse responses are sampled at discrete time steps, $n$, and thus can have delays that are fractions of a given sample rate [12]. In the case of HRIRs, BRIRs, and other measured impulse responses, we can simply refer to the measured response as $\mathbf{g} = [g[n], \ldots, g[n+N_g-1]]$ where $N_g$ is the length of the measured impulse response.

Taking the impulse response from a source $s$ to a location $m$ as $\mathbf{g}_{ms}$, we can write the signal at point $m$, $y_m[n]$, as generated from a set of $S$ sources, $x_s[n]$, as

$$y_m[n] = \sum_{s=1}^{S}\sum_{k=1}^{N_g-1} g_{ms}[k]x_s[n-k]. \qquad (3)$$

In matrix form, this can be written as

$$\mathbf{y} = \mathbf{G}\mathbf{x}, \qquad (4)$$

where

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \cdots & \mathbf{G}_{1S} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{M1} & \cdots & \mathbf{G}_{MS} \end{bmatrix}, \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_S \end{bmatrix}, \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_M \end{bmatrix}.$$

The channel matrix $\mathbf{G}$ is composed of matrices $\mathbf{G}_{ms}$, each representing the $(N_g+N_x-1)\times N_x$ dimensional acyclic convolution matrix of an individual source-to-receiver impulse response $\mathbf{g}_{ms}$. Each loudspeaker signal is $\mathbf{x}_s = [x_s[n], \ldots, x_s[n-(N_x-1)]]^T$.

To simplify the notation, we assume the filters of each set to be of same length, $N_{\mathbf{x}_s} = N_x$, $N_{\mathbf{g}_{ms}} = N_g$, $\forall m$, $\forall s$. However, the same results and conclusions in the next section apply to sets of filters of nonuniform length.

## 3. OPTIMIZATION FRAMEWORK

In the reminder of the paper, without loss of generality, we consider reproducing only one spatial source $v[n]$ through the $S$ loudspeaker set. We represent the desired sound field at the $m$-th point as $\bar{\mathbf{y}}_m = \mathbf{T}_m\mathbf{v}$, where $\mathbf{T}_m$ is the $(N_t+N_v-1)\times N_v$ acyclic convolution matrix of $\mathbf{t}_m \in \mathbb{R}^{N_t}$; $\mathbf{t}_m$ is the so-called target response designed according to the application. We can then consider a set of $S$ filters used to equalize the single spatial source in each loudspeaker $\mathbf{x}_s = \mathbf{H}_s\mathbf{v}$ where $\mathbf{H}_s$ is the $(N_h+N_v-1)\times N_v$ acyclic convolution matrix of $\mathbf{h}_s \in \mathbb{R}^{N_h}$. If we consider the propagation model in (4), we can rewrite the reproduced and desired signal, respectively, as

$$\mathbf{y} = \mathbf{G}\begin{bmatrix} \mathbf{H}_1\mathbf{v} \\ \vdots \\ \mathbf{H}_S\mathbf{v} \end{bmatrix} \quad \text{and} \quad \bar{\mathbf{y}} = \begin{bmatrix} \mathbf{T}_1\mathbf{v} \\ \vdots \\ \mathbf{T}_M\mathbf{v} \end{bmatrix}.$$

Thus, given our goal is to match desired and reproduced signals at the target points, i.e., $\hat{\mathbf{y}} \approx \mathbf{y}$. We can rewrite our problem as $\mathbf{T} \approx \mathbf{G}\mathbf{H}$. Given the particular structure of the convolution matrices, the problem becomes

$$\mathbf{G}\mathbf{h} \approx \mathbf{t}. \qquad (5)$$

We can now consider the optimization problem associated with finding a set of $S$ filters $\mathbf{h}_s \in \mathbb{R}^{N_h}$ from a set of observed $MS$ acoustic path models $\mathbf{g}_{ms} \in \mathbb{R}^{N_g}$ so that the reproduction error of the target function at each measurement point, $\mathbf{t}_m \in \mathbb{R}^{N_t}$, is minimized; $N_t = N_g + N_h - 1$ is used to ensure the linearity of the convolution.

An exact solution to (5) can be found in the case $S = 2M$, assuming $N_t = 2N_h - 1$, as is noted in the MINT method [13]. That is the case of $\mathbf{G}$ being a square matrix and thus the system having a unique solution, provided that $\mathbf{G}$ is full rank. Minimum norm solutions are always possible when $SN_h \geq MN_t$ and the matrix has full row rank, a condition that can be assumed for the convolution matrices considered here. In this case, the system (5) is said to be *underdetermined* and thus has infinitely many solutions, so we seek a particular solution that minimizes the $\ell_p$-norm of the solution vector. The optimization problem becomes

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{\Gamma}\mathbf{h}\|_q \ \text{ s.t. } \ \mathbf{G}\mathbf{h} = \mathbf{t}, \qquad (6)$$

where $\mathbf{\Gamma}$ is an optional spatio-temporal transform and $\|\cdot\|_q$ is defined as $\|\mathbf{x}\|_q = (\sum_{n=1}^{N} |x(n)|^q)^{\frac{1}{q}}$ and represents the $\ell_q$-norm.

Exploring the neighborhood of the minimum norm solution (6) by relaxing the constraint $\mathbf{G}\mathbf{h} = \mathbf{t}$ and determining an approximate solution is of general interest since perfect multichannel inversion is difficult to achieve when spatial robustness and possible perturbations of the measurement are considered [14]. This is also the case when $SN_h < MN_t$, when the system is *overdetermined*, and the condition $\mathbf{G}\mathbf{h} = \mathbf{t}$ cannot be fulfilled. In both cases the optimization problem can be written as

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{W}(\mathbf{G}\mathbf{h}-\mathbf{t})\|_p \ \text{ s.t. } \ \begin{array}{l} \|\mathbf{\Gamma}_i\mathbf{h}\|_{q_i} \leq \gamma_i, \\ \forall i,\ i = 1\ldots, I \end{array} \qquad (7)$$

where the matrices $\mathbf{\Gamma}_i$ and $\mathbf{W}$ represent linear projections, or transformations in a given domain, the implications of which will be discussed in the next section.

## 4. SPATIO-TEMPORAL TRANSFORMS

We consider here the options for spatio-temporal transforms that can be applied through $\mathbf{W}$ or $\mathbf{\Gamma}$ in (7) along with their acoustical and perceptual implications. Spatial transforms can include spatial weighting or averaging, spatial interpolation or extrapolation, wavenumber-domain transformations (e.g., spherical and cylindrical harmonics), wavenumber-domain interpolation or extrapolation, and wavenumber-domain weighting or averaging (some of which have been explored before, e.g., [11]). Temporal transforms can include the uniformly or non-uniformly spaced discrete Fourier transform (DFT), filter banks (including the auditory filter bank), temporal averaging or weighting, frequency averaging or weighting, and time or frequency interpolation or extrapolation (some of which have been explored before, e.g., [1]). Any transform choice can incorporate multiple space and time transforms, but we will analyze each domain separately in the following discussion. Note that the transform $\mathbf{W}$ can alter the rank and numerical tractability of the problem (7).

The transforms $\mathbf{W}$ and $\mathbf{\Gamma}$ operate on a set of time domain vectors stacked in order of their respective spatial positions. Thus the temporal only transforms, $\mathbf{W_t}$ and $\mathbf{\Gamma_t}$, are block diagonal matrices

$$\mathbf{W_t} = \begin{bmatrix} \mathbf{F}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{F}_M \end{bmatrix}, \mathbf{\Gamma_t}_i = \begin{bmatrix} \mathbf{F}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{F}_S \end{bmatrix} \tag{8}$$

where, for ease of explanation, the temporal transform, $\mathbf{F}$, is assumed to be the same for each impulse response. The most obvious choice of $\mathbf{F}$ is the uniformly spaced DFT matrix which allows for operations in the frequency domain. A closely related choice is for $\mathbf{F}$ to be a row vector containing the DFT corresponding to a single frequency $k$, which can be used to set up a set of constraints for each frequency, $\mathbf{\Gamma}_k = \mathbf{F}_k$. A simple frequency-dependent weighting can be applied to the DFT matrix to incorporate the perceptual non-linear frequency scaling, or an auditory filter bank can be applied instead which would have a similar effect [15].

The spatial transforms take on a more complicated form. Given some matrix $\mathbf{Y}$ which maps the spatial points ($S$ or $M$ depending on the transform) to some new domain with $C$ points, a transform matrix, can be constructed as

$$\mathbf{W_s} = \begin{bmatrix} y_{11}\mathbf{I} & \cdots & y_{1M}\mathbf{I} \\ \vdots & \ddots & \vdots \\ y_{C1}\mathbf{I} & \cdots & y_{CM}\mathbf{I} \end{bmatrix}, \mathbf{\Gamma_s}_i = \begin{bmatrix} y_{11}\mathbf{I} & \cdots & y_{1S}\mathbf{I} \\ \vdots & \ddots & \vdots \\ y_{C1}\mathbf{I} & \cdots & y_{CS}\mathbf{I} \end{bmatrix}$$

where $y_{ij}$ is the value of $\mathbf{Y}$ in the $i^{\text{th}}$ row and $j^{\text{th}}$ column and $\mathbf{I}$ is the identity matrix of size $N_t \times N_t$ for $\mathbf{W_s}$ and $N_h \times N_h$ for $\mathbf{\Gamma_s}_i$. Common choices for $\mathbf{Y}$ include the wavenumber-domain transforms created by the discrete spherical harmonics transform or the discrete cylindrical harmonics transform which can serve to distribute the reproduction error away from the center of a loudspeaker array when used in the cost function through $\mathbf{W}$ [9]. These transforms have the requirement that the spatial points be located on the surface of a sphere or cylinder, but appropriate radial variation can be included by solving the appropriate exterior or interior problem [11].

## 5. EXPERIMENTAL ANALYSIS

It is clear now that the optimization framework presented in (7) allows a great deal of flexibility in the design of the set of filters $\hat{\mathbf{h}}$ for NASS. The many applications and interactions of each of the transforms presented in Section 4 are too numerous to consider fully in this work.

For experimental evaluation, we chose two case studies. In the first study, we considered a spherical wave propagation model (2) and we designed time domain filters while applying constraints in the frequency domain using the uniformly spaced DFT, a simple case of temporal transform. In the second study, instead of spherical wave propagation, we used anechoic HRIR measurements for both the target and propagation model and we applied a non-uniformly spaced DFT that allowed us to minimize the error in a more perceptually relevant manner.

We considered the filter design problem in rendering a source at a given angle using a uniform linear loudspeaker array (ULA) with both 8 and 2 loudspeakers. We chose $N_g = 8192$, $N_h = 1024$, and $N_t = N_g + N_h - 1 = 9215$, as defined in Section 3. The target vectors were zero padded to guarantee causality [14] as $\mathbf{t}_m = [0, \ldots, 0, t_m[0], \ldots, t_m[N_t - D - 1]]^T$, where $D = 100$ is the number of leading zeros. We estimated the set of filters $\{\mathbf{h}_0 \ldots \mathbf{h}_{S-1}\}$ to render a plane-wave source arriving from 60° to the listener's left. The primary listening position was defined to be 2 m from the ULA and all evaluations were carried out with $M = 2$ points located at the left and right ears. The distance between drivers in the ULA was 10 cm, which corresponds to a minimum and maximum loudspeaker span of 3° and 20° for the 8 loudspeaker case, and a spatial aliasing frequency of 1.7 kHz.

### 5.1. Case Study 1: The Effect of Constraint Norm

We focus our attention on the optimization problem assuming a $\ell_2$-norm criterion on the unweighted cost function and impose a frequency domain constraint on the solution vector. While other norms have been considered, especially $\ell_1$-norm and $\ell_\infty$-norm to minimize the frequency domain response error measure in other applications [16], a thorough analysis of how these improve the perceptual quality of the result has not been shown in the literature, and we leave these cases open to further investigation. The problem in (7) is then written as

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \ \|\mathbf{G}\mathbf{h} - \mathbf{t}\|_2 \ \text{ s. t. } \|\mathbf{\Gamma}\mathbf{h}\|_q \leq \gamma, \tag{9}$$

where $\mathbf{\Gamma}$ is a block diagonal matrix composed of $S$ DFT matrices as defined in (8) and $\mathbf{t} = \left[\mathbf{t}_1^T, \ldots, \mathbf{t}_M^T\right]^T$.

The cases when $q = 1, 2, \infty$ are of particular interest, especially given their physical meaning in the frequency domain. These problems are convex and can be solved efficiently using, e.g., interior point methods [17]. Results for these three cases are plotted in Figure 1. It is important to note the effect of ill-conditioning on the solution in the unconstrained case (Figure 1a), where the filters obtained are physically unfeasible, requiring a large boost at the lower frequencies.

When the $\ell_2$-norm is minimized, the overall energy of the filters is constrained. Thus, the parameter $\gamma$ as an important physical interpretation representing the maximum square root of the energy that can be output by the system, and thus $\gamma = \sqrt{E_{\text{max}}}$. In Figure 1c the results for $\gamma = 39$ dB are shown. It is clear that the filters require less total energy than the unconstrained solution ($\|\mathbf{\Gamma}\mathbf{h}\|_1 = 67$ dB and $\|\mathbf{\Gamma}\mathbf{h}\|_\infty = 32$ dB for this design).

While the $\ell_2$-norm constrains the maximum energy flowing through the system, the $\ell_\infty$-norm allows us to define the maximum possible absolute value of the estimated filters in the frequency domain, which is particularly relevant when determining physically meaningful solutions considering the loudspeaker output. This minimax type of solution engenders a relatively flat spectrum. In Figure 1d the results for $\gamma = 6$ dB are shown where it is clear that the maximum output is limited to this threshold ($\|\mathbf{\Gamma}\mathbf{h}\|_2 = 39$ dB and $\|\mathbf{\Gamma}\mathbf{h}\|_1 = 77$ dB).

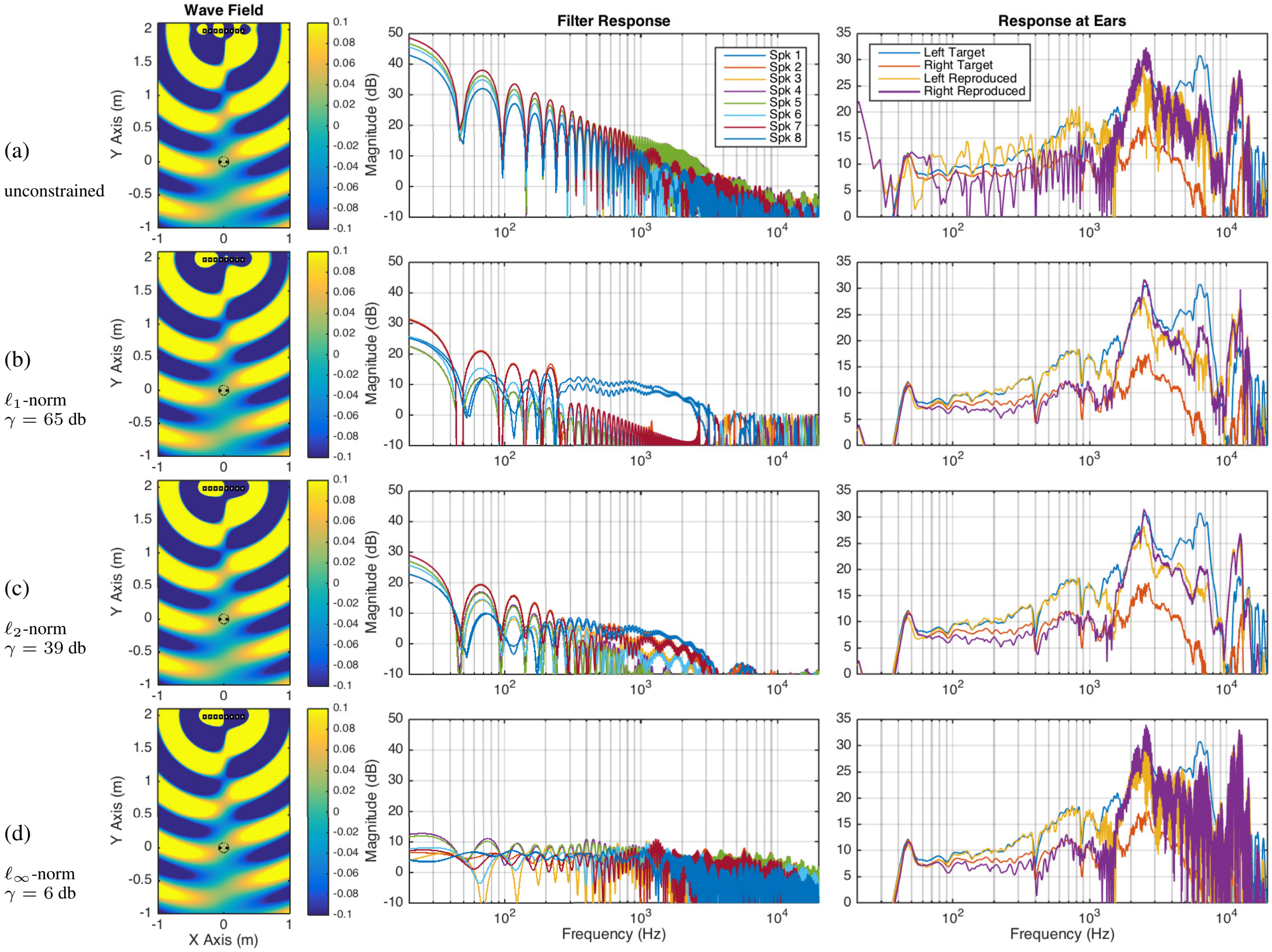

**Fig. 1**: Simulation results of a plane-wave source rendered at 60° using an $S = 8$ loudspeaker array with $M = 2$ (black dots) for different values of of $q$ and $\delta$ in (9). The wave field is plotted at 500 Hz and the response at the ears is plotted using a measured anechoic HRIR.

The $\ell_1$-norm is a powerful convex tool often associated with minimizing the *sparsity* of a vector, i.e., the so-called $\ell_0$-norm. This type of optimization is relevant if the constraint is applied in either the time or frequency domain. In the time domain, we will obtain filters that are *sparser*, and often shorter, allowing for computationally efficient NASS. In the frequency domain, the importance of the sparse criterion is less obvious. When $\|\mathbf{\Gamma h}\|_1$ is minimized, the number of active DFT bins is minimized across all loudspeakers (i.e., the frequencies and the loudspeakers are considered jointly). In Figure 1b, with $\gamma = 65$ dB we found a sparse solution where only two loudspeakers were used to reproduce the field across most of the mid-frequency bands and four loudspeakers for the low frequencies ($\|\mathbf{\Gamma h}\|_2 = 39$ dB and $\|\mathbf{\Gamma h}\|_\infty = 34$ dB). Notice that, if we operate on a bin-by-bin basis where $\mathbf{\Gamma}_k$ is the DFT row vector discussed in Section 4, the problem minimizes the number of active loudspeakers at each frequency independently and (9) becomes a generalization of [18].

### 5.2. Case Study 2: Perceptual Error

It is clear from Figure 1 that the spherical wave propagation model can only match the target HRIR at the listening position well up to 1.1 kHz due to both spatial aliasing issues and the mismatch between the analytic models and the real-world listening scenario. Given the poor results, we modified the system in (9) by including two perceptual alterations. First, the impulse responses of the propagation model, $\mathbf{G}$ and target functions, $\mathbf{t}$, were changed to actual measured HRIRs, guaranteeing a matching between the acoustic situation encountered by a listener in the reproduced sound field. Secondly, we applied a weighting using a non-uniformly spaced DFT matrix (instead of the uniformly spaced matrix in Section 5.1) with points linearly spaced along the equivalent rectangular bandwidth (ERB) scale between 20 Hz and 20 kHz. The problem in (7) is then

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \ \|\mathbf{W}(\mathbf{G}\mathbf{h} - \mathbf{t})\|_2 \ \text{ s. t. } \|\mathbf{\Gamma h}\|_q \leq \gamma, \qquad (10)$$

where both $\mathbf{W}$ and $\mathbf{\Gamma}$ are the ERB-spaced DFT matrix transformations. Figure 2 shows three examples to outline the effects of these two modifications. Figure 2a displays the unconstrained solution to (10) with $S = 8$ loudspeakers to show that the use of both the HRIR target and acoustic model in the underdetermined case leads to both realizable filters and perfect reconstruction of the target HRIR at the listening position. Figure 2b shows what happens with an unconstrained spherical wave propagation model with $S = 2$ loudspeakers in the overdetermined case. It is clear from the response at the ears that the system will have impaired timbral response in the 100 Hz to 1 kHz range. Figure 2c shows the result of solving (10) with

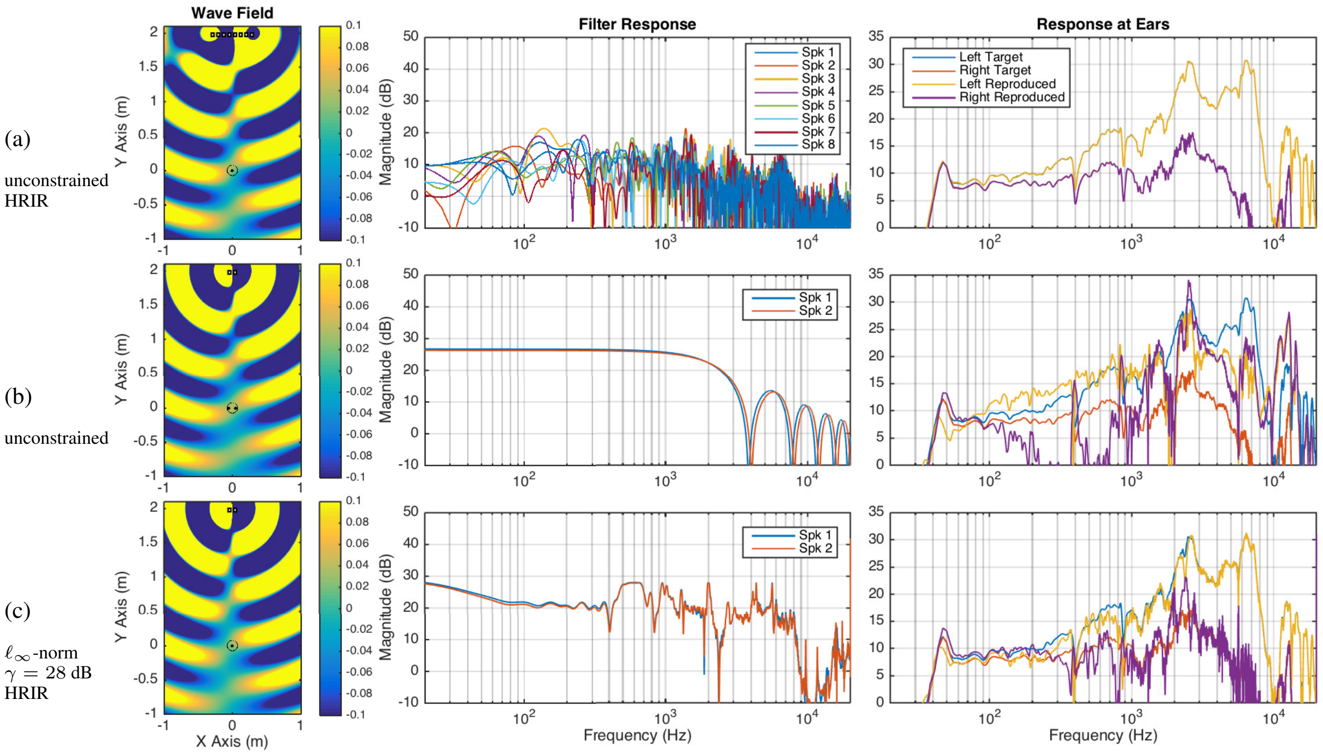

**Fig. 2**: Simulation results of a plane-wave source rendered at 60° using (a) $S = 8$ with HRIR target an propagation model, (b) $S = 2$ with spherical-wave propagation model, and (c) $S = 2$ with HRIR target and propagation model. The wave field is plotted at 500 Hz.

$\gamma = 28$ dB (the same $\ell_\infty$-norm as Figure 2b). Here it is clear that the final response at the ears of the listener matches the target response quite closely across the full bandwidth.

As a final analysis of the experiments presented in Figures 1 and 2 it is worth noting that all techniques show minimal differences in the wave fields at the listening position, a commonly presented evaluation metric in the literature. However, the response at the ears highlights large predicted perceived differences between each solution which match our informal listening experiments, motivating the use of both broadband design and analysis techniques when comparing spatial audio methods.

## 6. CONCLUSIONS

We have presented a unified framework to reproduce an auditory scene through loudspeaker arrays. This framework allowed us to encompass several approaches well known in the literature and explore new techniques in a systematic way including perceptually relevant solutions. We focused on numerical methods for reproduction where the $\ell_2$-norm between desired and reproduced sound fields is minimized. This problem, often ill-conditioned, was solved using different types of regularization based on the $\ell_q$-norm, with $q = 1, 2, \infty$, leading to solutions with different physical meanings. In particular, when $q = 1$, we were able to find sparse solutions, effectively reducing the number of active loudspeakers to reproduce a given field. When $q = 2$ and $q = \infty$, we were able to reduce the maximum energy flowing in the system and effectively bound the frequency response of the generated filters, respectively. Two simple perceptually relevant modifications were highlighted as an example of the flexibility of this framework: modification of the acoustic models to use the HRIR and projection of the error and constraints into an ERB-spaced frequency domain.

## REFERENCES

[1] M. R. Bai, et al., *Acoustic array systems*, Wiley, 2013.

[2] A. J. Berkhout, et al. "Acoustic control by wave field synthesis," *J. ASA*, vol. 93, pp. 2764, 1993.

[3] J. Daniel, "Spatial sound encoding including near field effect," *23rd AES Conf.*, 2003.

[4] D. B. Ward, "Joint least squares optimization for robust acoustic crosstalk cancellation," *IEEE Trans. SAP*, vol. 8, no. 2, pp. 211–215, 2000.

[5] I. Nawfal and J. Atkins, "Binaural reproduction over loudspeakers using a modified target response," *Proc. ICAD*, 2014.

[6] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. AES*, vol. 45, no. 6, pp. 456–466, 1997.

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

[8] G. H. Koopmann, et al., "A method for computing acoustic fields based on the principle of wave superposition," *J. ASA*, vol. 86, no. 6, pp. 2433–2438, 1989.

[9] J. Daniel, et al., "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," *AES Conv.*, 2003.

[10] J. Blauert, Ed., *The technology of binaural listening*, Springer, 2013.

[11] E. G. Williams, *Fourier acoustics*, Academic Press, 1999.

[12] V. Välimäki and T. I. Laakso, "Principles of fractional delay filters," *Proc. ICASSP*, 2000.

[13] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. ASSP*, vol. 36, no. 2, pp. 145–152, 1988.

[14] J. O. Jungmann, et. al., "Combined acoustic mimo channel crosstalk cancellation and room impulse response reshaping," *IEEE Trans. ASL*, vol. 20, no. 6, pp. 1829–1842, 2012.

[15] J. Huopaniemi and J. O. Smith III, "Spectral and time-domain preprocessing and the choice of modeling error criteria for binaural digital filters," in *16th AES Conf.*, 1999.

[16] S. Yan and Y. Ma, "A unified framework for designing FIR filters with arbitrary magnitude and phase response," *Digital Sig. Proc.*, vol. 14, no. 6, pp. 510 – 522, 2004.

[17] S. J. Wright, *Primal-dual interior-point methods*, SIAM, 1997.

[18] G.N. Lilis, et al., "Sound field reproduction using the lasso," *IEEE Trans. ASL*, vol. 18, no. 8, pp. 1902–1912, 2010.