A UNIFIED APPROACH TO NUMERICAL AUDITORY SCENE SYNTHESIS USING LOUDSPEAKER ARRAYS

JOSHUA ATKINS, ISMAEL NAWFAL, DANIELE GIACOBELLO

SEPTEMBER 4, 2014

MOTIVATION

SPATIAL SOUND SYSTEMS

- **Spatial sound overview**
  - Physical Reconstruction: wave-field synthesis (WFS), near-field compensated higher-order ambisonics (NFC-HOA)
    - Issues: not flexible in speaker arrangement, challenging for full-band audio
  - Interpolation: vector base amplitude panning (VBAP)
    - Issues: not flexible in speaker arrangement, sources located on surface of array, coloration of sources
  - Numerical Optimization: equivalent source method (ESM), mode-matching, crosstalk cancellation
    - Issues: no inclusion of perception, filter design is left as separate problem

- **Proposal in this work (Numerical Auditory Scene Synthesis)**
  - Goal: correct reproduction of perceived auditory scene (not wave field)
  - Convex numerical framework: flexible speaker layouts, listener positions, and error-norms
  - Inherently broadband: time-domain filter generation
  - Spatio-temporal projection: include perception, spatial error distribution

OPTIMIZATION FRAMEWORK

PROBLEM STATEMENT

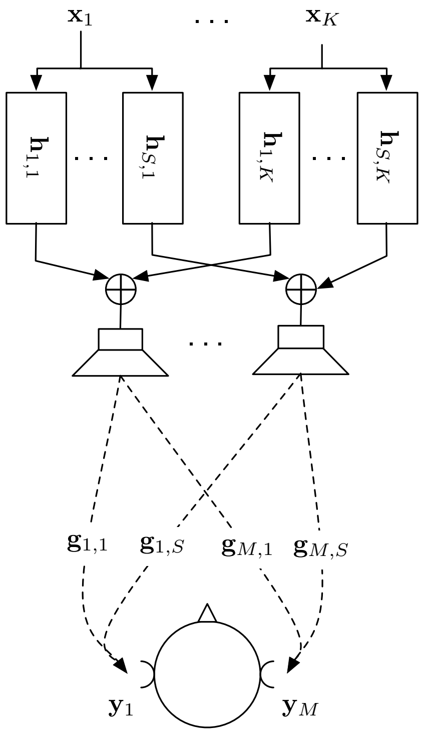
problem: design filters to best approximate response at target locations
assumptions: 1 source, M target points, S speakers

- reproduction: **y** (signal at target points), **G** (acoustic impulse responses, convolution matrices), **H** (unknown filters, convolution matrices), **x** (signal)

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{1,1} & \cdots & \mathbf{G}_{1,S} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{M,1} & \cdots & \mathbf{G}_{M,S} \end{bmatrix} \begin{bmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_S \end{bmatrix} \mathbf{x}$$

- design: **t** (desired impulse response at target points), **h** (unknown filters)

$$\begin{bmatrix} \mathbf{t}_1 \\ \vdots \\ \mathbf{t}_M \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{1,1} & \cdots & \mathbf{G}_{1,S} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{M,1} & \cdots & \mathbf{G}_{M,S} \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_S \end{bmatrix}$$



ACOUSTIC MODELS

FLEXIBILITY IN TARGET AND TRANSMISSION MODELS

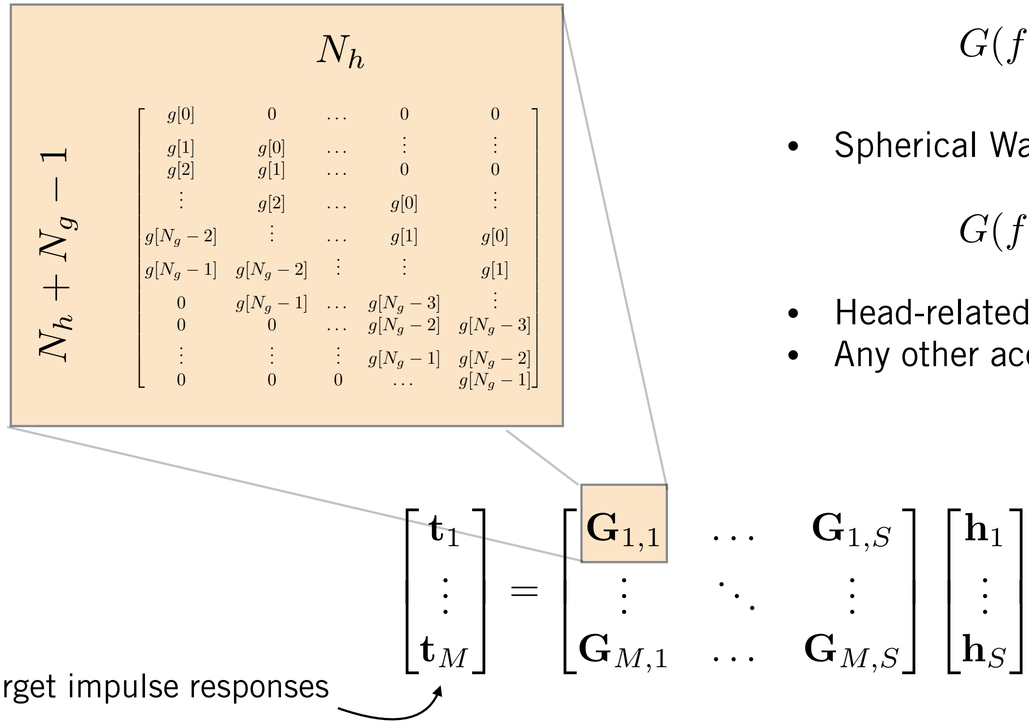
- Plane Wave

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

- Spherical Wave

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

- Head-related impulse response (HRIR, BRIR)
- Any other acoustic impulse response



Convolution Matrix

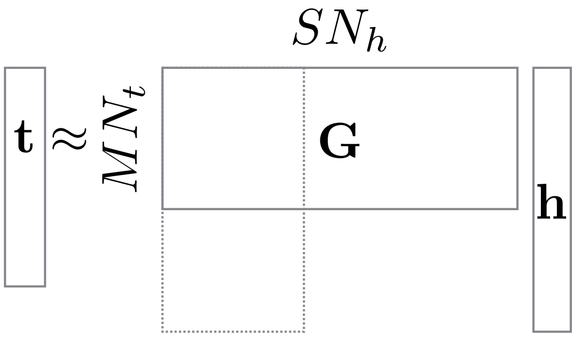
$$\begin{array}{cc} & N_h \\ N_h + N_g - 1 & \begin{bmatrix} g[0] & 0 & \dots & 0 & 0 \\ g[1] & g[0] & \dots & \vdots & \vdots \\ g[2] & g[1] & \dots & 0 & 0 \\ \vdots & g[2] & \dots & g[0] & \vdots \\ g[N_g-2] & \vdots & \dots & g[1] & g[0] \\ g[N_g-1] & g[N_g-2] & \vdots & \vdots & g[1] \\ 0 & g[N_g-1] & \dots & g[N_g-3] & \vdots \\ 0 & 0 & \dots & g[N_g-2] & g[N_g-3] \\ \vdots & \vdots & \vdots & g[N_g-1] & g[N_g-2] \\ 0 & 0 & 0 & \dots & g[N_g-1] \end{bmatrix} \end{array}$$

$$\begin{bmatrix} \mathbf{t}_1 \\ \vdots \\ \mathbf{t}_M \end{bmatrix} = \begin{bmatrix} \boxed{\mathbf{G}_{1,1}} & \dots & \mathbf{G}_{1,S} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{M,1} & \dots & \mathbf{G}_{M,S} \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_S \end{bmatrix}$$

target impulse responses

NUMERICAL AUDITORY SCENE SYNTHESIS PROBLEM

FLEXIBLE CONVEX PROGRAM



- **Underdetermined** ($S\,N_h > M\,N_t$, full rank)
  choose one of many exact solutions:

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\boldsymbol{\Gamma}\mathbf{h}\|_q \quad \text{s.t.} \quad \mathbf{G}\mathbf{h} = \mathbf{t}$$

- **Overdetermined** ($S\,N_h < M\,N_t$, full rank) and/or uncertainty
  approximate solution:

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{W}(\mathbf{G}\mathbf{h} - \mathbf{t})\|_p \;\; \text{s.t.} \|\boldsymbol{\Gamma}_i\mathbf{h}\|_{q_i} \leq \gamma_i, \quad \forall i,\; i = 1\ldots, I$$

- Convex, flexible error/regularizer norm, spatio-temporal projection matrices (incorporate perception)

SPATIO-TEMPORAL TRANSFORMS

ALTER SOLUTION SPACE AND/OR FILTER SPECIFICATION

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \ \|\mathbf{W}(\mathbf{G}\mathbf{h} - \mathbf{t})\|_p \ \ \mathrm{s.t.} \|\mathbf{\Gamma}_i \mathbf{h}\|_{q_i} \leq \gamma_i, \quad \forall i,\ i = 1\ldots, I$$

- **Time-frequency transform** (DFT, filter banks, and time/frequency weighting, averaging, interpolation)

$$\mathbf{W}_{\mathbf{t}} = \begin{bmatrix} \mathbf{F}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{F}_M \end{bmatrix} \qquad \mathbf{F} = \begin{bmatrix} f_{1,1} & \cdots & f_{1,N_t} \\ \vdots & \ddots & \vdots \\ f_{N_f,N_t} & \cdots & f_{N_f,N_t} \end{bmatrix}$$

- **Space-wavenumber transform** (spherical/cylindrical harmonics and spatial weighting, averaging, …)

$$\mathbf{W}_s = \begin{bmatrix} y_{1,1}\mathbf{I} & \cdots & y_{1,M}\mathbf{I} \\ \vdots & \ddots & \vdots \\ y_{C,1}\mathbf{I} & \cdots & y_{C,M}\mathbf{I} \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,M} \\ \vdots & \ddots & \vdots \\ y_{C,1} & \cdots & y_{C,M} \end{bmatrix}$$

* similar transforms for Γ in paper

NUMERICAL AUDITORY SCENE SYNTHESIS PROBLEM

OPTIONS FOR SYSTEM DESIGN

Which acoustic model, **G**?

Which error transform, **W**?

Which constraint norm, q?

Which constraint value, γ?

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{W}(\mathbf{G}\mathbf{h}-\mathbf{t})\|_p \quad \text{s.t.} \|\boldsymbol{\Gamma}_i\mathbf{h}\|_{q_i} \le \gamma_i, \quad \forall i,\ i = 1\ldots, I$$

Which acoustic target, **t**?

Which constraint transform, ***Γ***?

- Not studied here:
  - Which error norm, p?
  - How many target points?
  - How many speakers?
  - Which speaker locations?

CASE STUDY 1: THE EFFECT OF CONSTRAINT NORM

EXPERIMENT SETUP

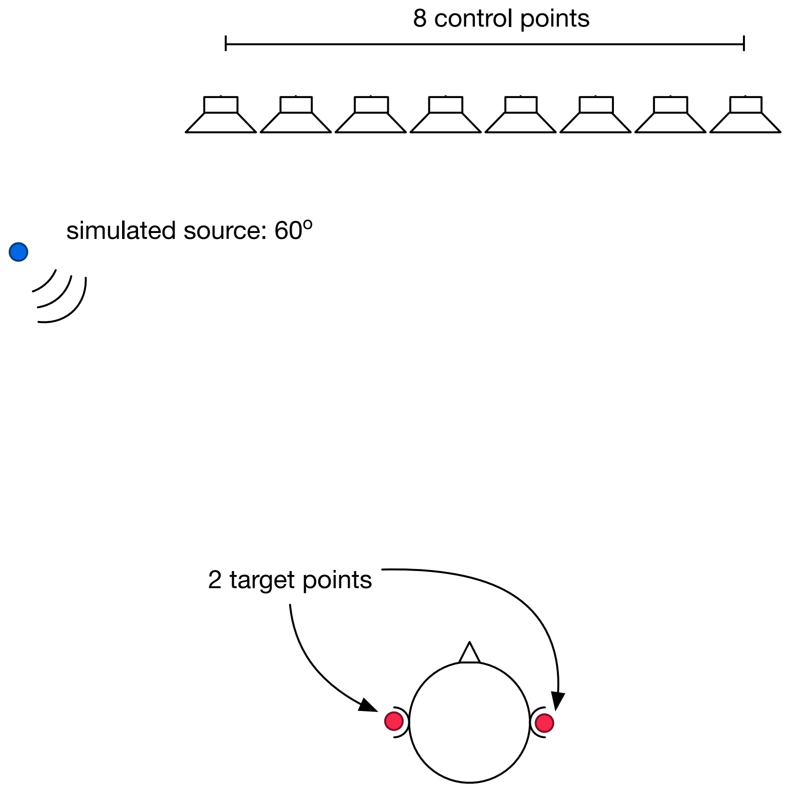
$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \ \|\mathbf{G}\mathbf{h} - \mathbf{t}\|_2 \ \ \mathrm{s.t.} \|\mathbf{\Gamma}\mathbf{h}\|_q \leq \gamma$$

Which constraint norm, q?

Which constraint value, γ?

- spherical wave acoustic model (**G**, **t**)
- l_2-norm error (p = 2)
- DFT projection matrix (***Γ***)
- filter length = 1024, modeling delay = 100

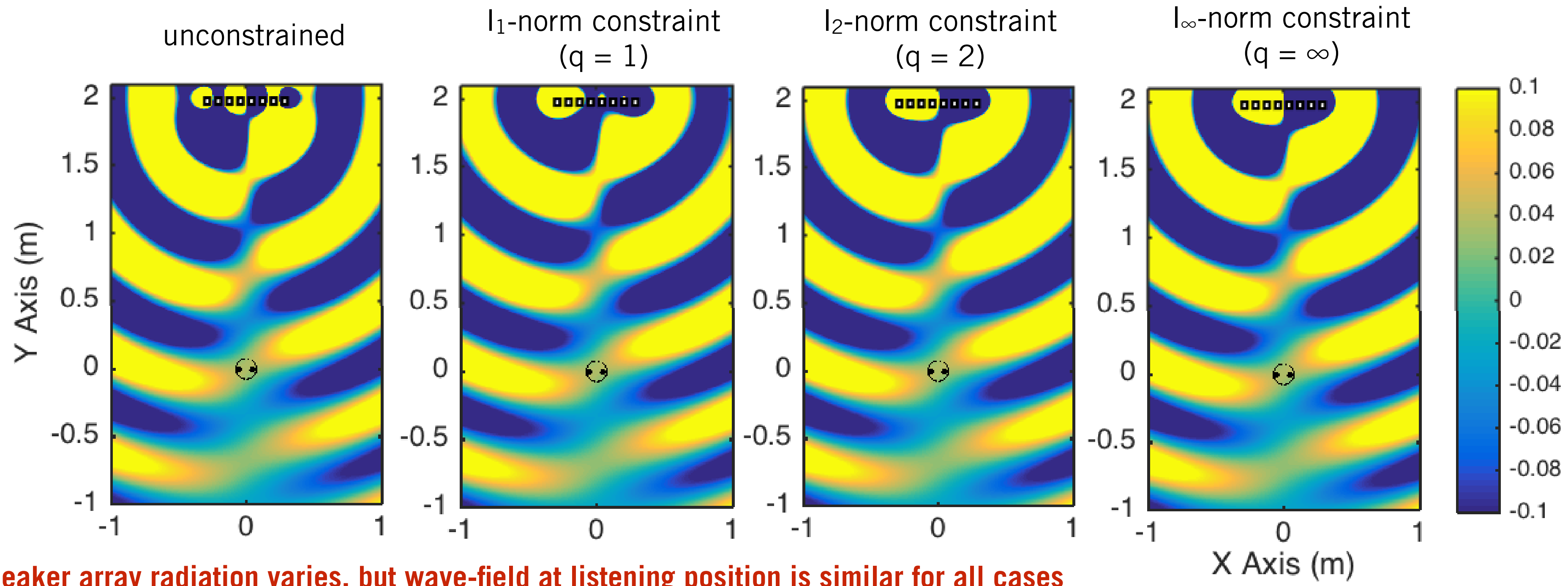
- **4 systems:**
  - unconstrained
  - l_1-norm constraint (q = 1)
  - l_2-norm constraint (q = 2)
  - l_∞-norm constraint (q = ∞)



CASE STUDY 1: THE EFFECT OF CONSTRAINT NORM

WAVEFIELD AT 500 HZ

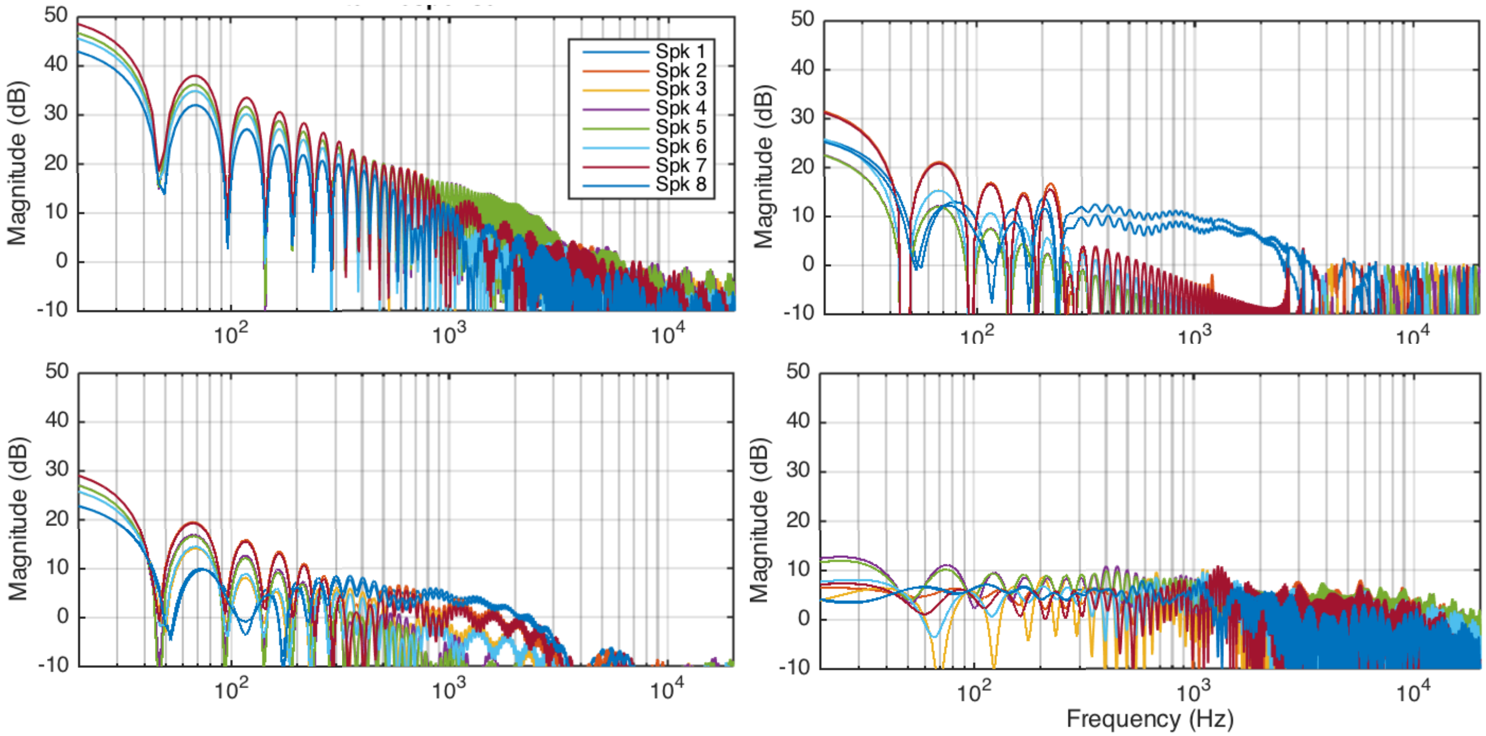
$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \ \|\mathbf{G}\mathbf{h} - \mathbf{t}\|_2 \ \ \mathrm{s.t.} \|\mathbf{\Gamma}\mathbf{h}\|_q \leq \gamma$$



speaker array radiation varies, but wave-field at listening position is similar for all cases

CASE STUDY 1: THE EFFECT OF CONSTRAINT NORM

BROADBAND FILTER RESPONSE, FIXED L2-NORM=39DB FOR ALL CASES



Large gain required at low frequencies

Few active speakers per frequency band

unconstrained

l_1-norm

l_2-norm

l_∞-norm

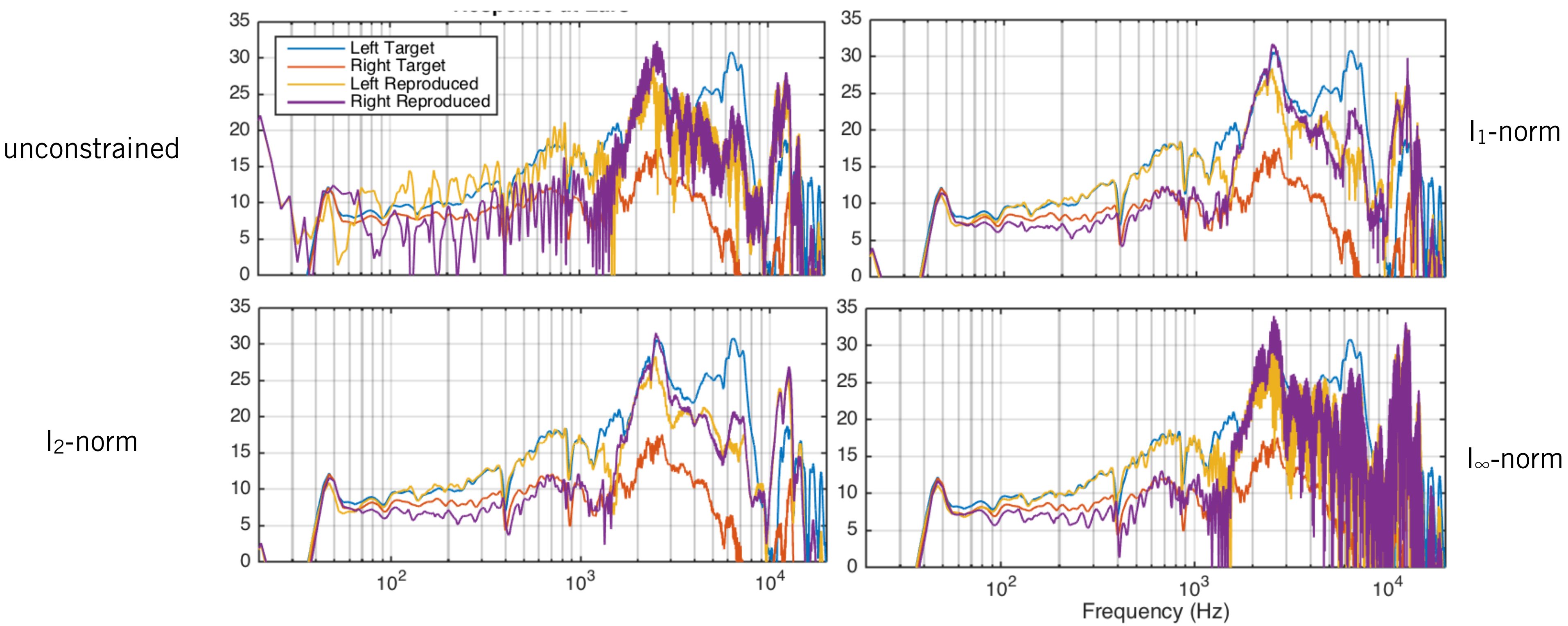
Limited required power

Limited maximum magnitude

CASE STUDY 1: THE EFFECT OF CONSTRAINT NORM

RESPONSE SIMULATED AT EAR DRUM REFERENCE (TARGET POINTS)

All fail above 1-2 kHz, high frequency coloration



CASE STUDY 2: PERCEPTUAL ERROR TRANSFORM

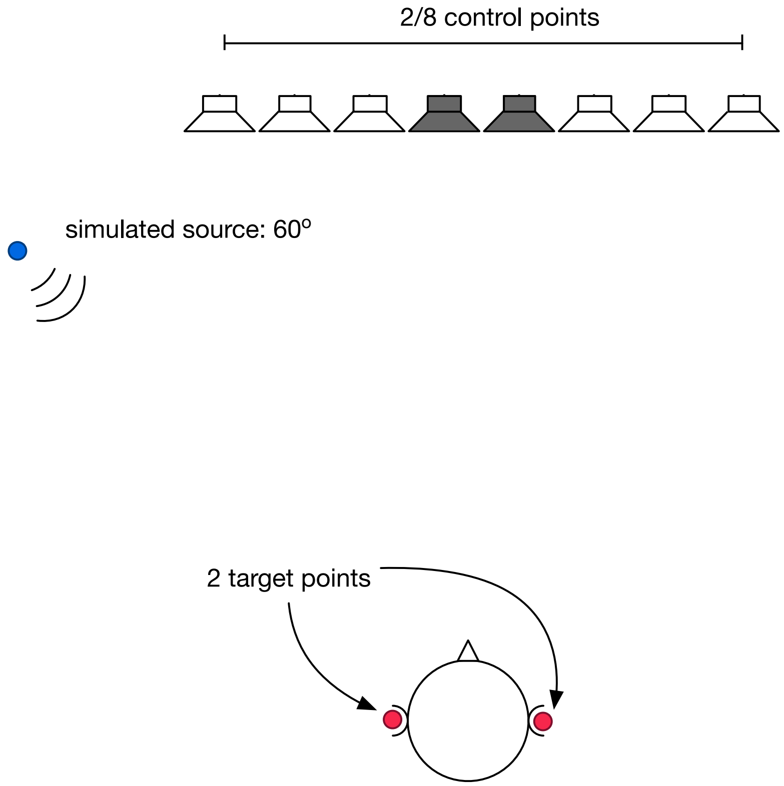
EXPERIMENTAL SETUP

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

Which acoustic model, **G**?
Which error transform, **W**?
Which acoustic target, **t**?
Which constraint transform, ***Γ***?

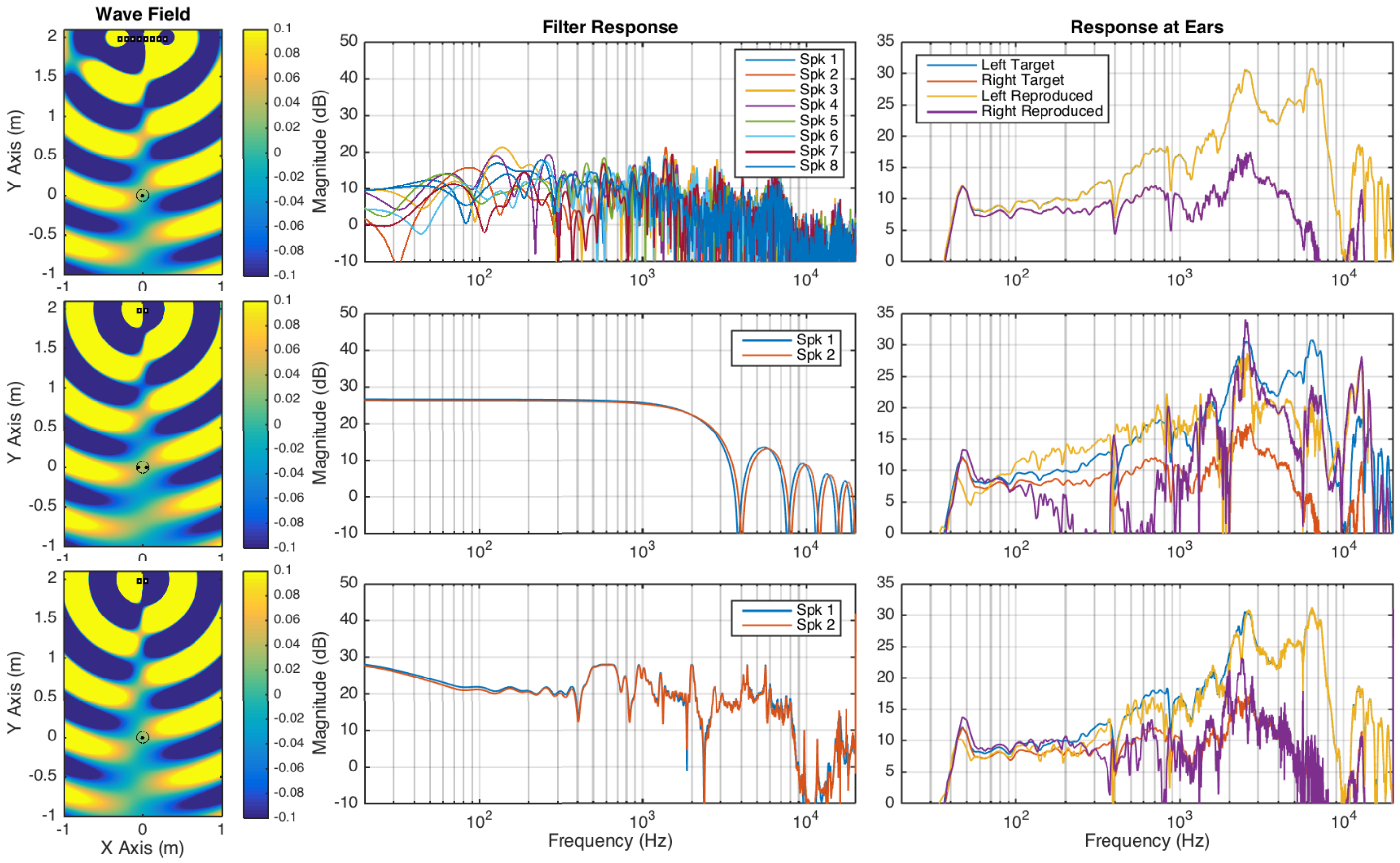
- l_2-norm error (p = 2)
- filter length = 1024, modeling delay = 100

- **3 systems:**
  - 8 speakers, HRIR, unconstrained
  - 2 speakers, spherical wave, unconstrained
  - 2 speakers, HRIR, l_∞-norm constraint, ERB-spaced DFT (**W**, ***Γ***)



CASE STUDY 2: PERCEPTUAL ERROR TRANSFORM

WAVEFIELD (500 HZ), FILTER RESPONSE, AND RESPONSE AT EAR DRUM



8 speakers
HRIR
unconstrained

exact solution

2 speakers
spherical wave
unconstrained

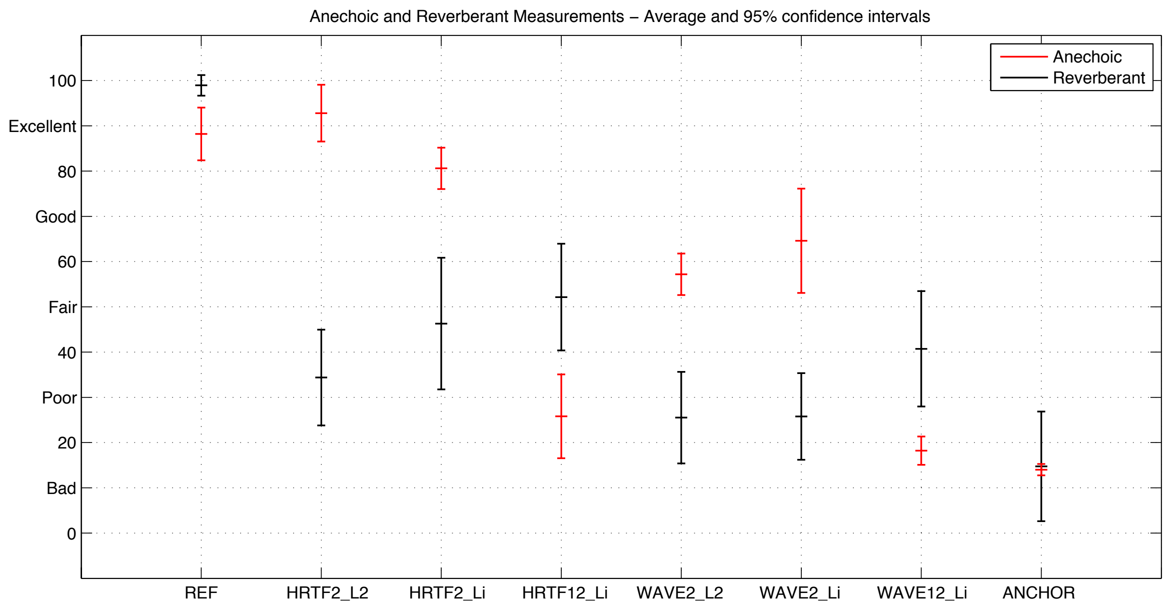
large timbral errors

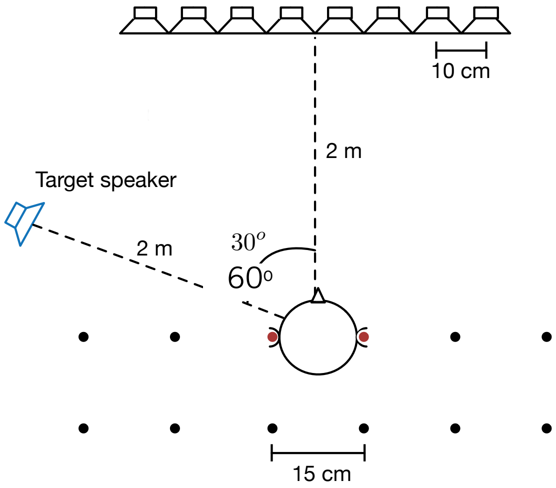
2 speakers
HRIR
l_∞-norm (28 dB)
ERB-spaced DFT

very close response

IN PRACTICE: SOME EARLY PERCEPTUAL RESULTS

FROM: AES 55TH CONFERENCE ON SPATIAL AUDIO





HRIR target is preferred, multipoint (overdetermined) does better in reverberant scenario

[1] Ismael Nawfal, Joshua Atkins, Daniele Giacobello, Stephen Nimick. “Perceptual Evaluation of Numerical Auditory Scene Synthesis Using Loudspeaker Arrays.” Proceedings of the 55th Convention of the Audio Engineering Society. August 2014.

CONCLUSION

…

- Numerical Auditory Scene Synthesis
  - Flexible spatial rendering method for generating time-domain **broadband filters**
  - Can be used with **arbitrary loudspeaker arrays**
  - **Convex** program guarantees achievable solution
  - Spatio-temporal transform matrices allow for simple inclusion of **perceptual constraints**

- Analysis
  - Showed effect of filter constraint norm on resulting system
    - easily prefer sparse loudspeaker activations or limit maximum gain applied to loudspeaker array
  - Simple perceptual constraints: ERB-spaced transform, HRIR target/acoustic model
    - outperforms spherical wave assumption in objective & subjective tests

THANKS!