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
**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)

$$
\begin{bmatrix}
  y_1 \\
  \vdots \\
  y_M
\end{bmatrix}
=
\begin{bmatrix}
  G_{1,1} & \cdots & G_{1,S} \\
  \vdots & \ddots & \vdots \\
  G_{M,1} & \cdots & G_{M,S}
\end{bmatrix}
\begin{bmatrix}
  H_1 \\
  \vdots \\
  H_S
\end{bmatrix}
\times x
$$

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

$$
\begin{bmatrix}
  t_1 \\
  \vdots \\
  t_M
\end{bmatrix}
=
\begin{bmatrix}
  G_{1,1} & \cdots & G_{1,S} \\
  \vdots & \ddots & \vdots \\
  G_{M,1} & \cdots & G_{M,S}
\end{bmatrix}
\begin{bmatrix}
  h_1 \\
  \vdots \\
  h_S
\end{bmatrix}
$$
ACOUSTIC MODELS

FLEXIBILITY IN TARGET AND TRANSMISSION MODELS

- Plane Wave
  \[ G(f) = A e^{i(k \cdot r - \omega t)} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = A \delta \left( \frac{n \cdot r}{c} - t \right) \]

- Spherical Wave
  \[ G(f) = \frac{A e^{i(k r - \omega t)}}{r} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = \frac{A}{r} \delta \left( \frac{r}{c} - t \right) \]

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

Convolution Matrix

\[ G(f) = A e^{i(k \cdot r - \omega t)} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = A \delta \left( \frac{n \cdot r}{c} - t \right) \]

\[ G(f) = \frac{A e^{i(k r - \omega t)}}{r} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = \frac{A}{r} \delta \left( \frac{r}{c} - t \right) \]

\[ G(f) = A e^{i(k r - \omega t)} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = A \delta \left( \frac{n \cdot r}{c} - t \right) \]

\[ G(f) = \frac{A e^{i(k r - \omega t)}}{r} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = \frac{A}{r} \delta \left( \frac{r}{c} - t \right) \]

\[ G(f) = A e^{i(k \cdot r - \omega t)} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = A \delta \left( \frac{n \cdot r}{c} - t \right) \]

\[ G(f) = \frac{A e^{i(k r - \omega t)}}{r} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = \frac{A}{r} \delta \left( \frac{r}{c} - t \right) \]

\[ G(f) = A e^{i(k \cdot r - \omega t)} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = A \delta \left( \frac{n \cdot r}{c} - t \right) \]

\[ G(f) = \frac{A e^{i(k r - \omega t)}}{r} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = \frac{A}{r} \delta \left( \frac{r}{c} - t \right) \]

\[ G(f) = A e^{i(k \cdot r - \omega t)} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = A \delta \left( \frac{n \cdot r}{c} - t \right) \]

\[ G(f) = \frac{A e^{i(k r - \omega t)}}{r} \overset{\mathcal{F}^{-1}}{\longrightarrow} g(t) = \frac{A}{r} \delta \left( \frac{r}{c} - t \right) \]
NUMERICAL AUDITORY SCENE SYNTHESIS PROBLEM

FLEXIBLE CONVEX PROGRAM

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

  \[
  \hat{h} = \arg \min_h \|\Gamma h\|_q \text{ s.t. } Gh = t
  \]

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

  \[
  \hat{h} = \arg \min_h \|W(Gh - t)\|_p \text{ s.t. } \|\Gamma_i h\|_{q_i} \leq \gamma_i, \\
  \quad \forall i, i = 1 \ldots I
  \]

- **Convex**, flexible error/regularizer norm, spatio-temporal projection matrices (incorporate perception)
SPATIO-TEMPORAL TRANSFORMS
ALTER SOLUTION SPACE AND/OR FILTER SPECIFICATION

\[ \hat{h} = \arg \min_{h} \|W(Gh - t)\|_p \quad \text{s.t.} \|\Gamma_i h\|_{q_i} \leq \gamma_i, \]
\[ \forall i, i = 1, \ldots, I \]

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

\[ W_t = \begin{bmatrix} F_1 & 0 \\ \vdots & \vdots \\ 0 & F_M \end{bmatrix} \quad F = \begin{bmatrix} f_{1,1} & \cdots & f_{1,N_t} \\ \vdots & \ddots & \vdots \\ f_{N_f,N_t} & \cdots & f_{N_f,N_t} \end{bmatrix} \]

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

\[ W_s = \begin{bmatrix} y_{1,1}I & \cdots & y_{1,M}I \\ \vdots & \ddots & \vdots \\ y_{C,1}I & \cdots & y_{C,M}I \end{bmatrix} \quad Y = \begin{bmatrix} y_{1,1} & \cdots & y_{1,M} \\ \vdots & \ddots & \vdots \\ y_{C,1} & \cdots & y_{C,M} \end{bmatrix} \]

* similar transforms for \( I' \) in paper
NUMERICAL AUDITORY SCENE SYNTHESIS PROBLEM
OPTIONS FOR SYSTEM DESIGN

• Not studied here:
  • Which error norm, \( p \)?
  • How many target points?
  • How many speakers?
  • Which speaker locations?

\[
\hat{h} = \arg\min_h \|W(Gh - t)\|_p \quad \text{s. t.} \quad \|\Gamma_i h\|_{q_i} \leq \gamma_i, \\
\forall i, i = 1 \ldots, I
\]
CASE STUDY 1: THE EFFECT OF CONSTRAINT NORM

EXPERIMENT SETUP

\[ \hat{h} = \arg \min_{h} \| Gh - t \|_2 \quad \text{s.t.} \| \Theta h \|_q \leq \gamma \]

Which constraint norm, q?
Which constraint value, \( \gamma \)?

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

4 systems:
- unconstrained
- \( l_1 \)-norm constraint (\( q = 1 \))
- \( l_2 \)-norm constraint (\( q = 2 \))
- \( l_\infty \)-norm constraint (\( q = \infty \))
**CASE STUDY 1: THE EFFECT OF CONSTRAINT NORM**

**WAVEFIELD AT 500 HZ**

\[
\hat{h} = \arg \min_h \|Gh - t\|_2 \quad \text{s. t.} \|\Gamma h\|_q \leq \gamma
\]

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 = 39 DB FOR ALL CASES

Large gain required at low frequencies

Few active speakers per frequency band

unconstrained

I1-norm

I2-norm

Limited required power

Limited maximum magnitude

I∞-norm
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

unconstrained

\( l_2 \)-norm

\( l_1 \)-norm

\( l_{\infty} \)-norm
\[
\hat{h} = \arg \min_h \|W(Gh - t)\|_2 \quad \text{s.t.} \quad \|\Gamma h\|_q \leq \gamma
\]

Which acoustic model, \(G\)?
Which error transform, \(W\)?
Which acoustic target, \(t\)?
Which constraint transform, \(\Gamma\)?

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

- **3 systems:**
  - 8 speakers, HRIR, unconstrained
  - 2 speakers, spherical wave, unconstrained
  - 2 speakers, HRIR, \(l_\infty\)-norm constraint, ERB-spaced DFT (\(W, \Gamma\))
CASE STUDY 2: PERCEPTUAL ERROR TRANSFORM

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

8 speakers
HRIR unconstrained

2 speakers
spherical wave unconstrained

2 speakers
HRIR $l_{\infty}$-norm (28 dB)
ERB-spaced DFT

exact solution
large timbral errors
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

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!