Perceptual Evaluation of Numerical Auditory Scene Synthesis Using Loudspeaker Arrays

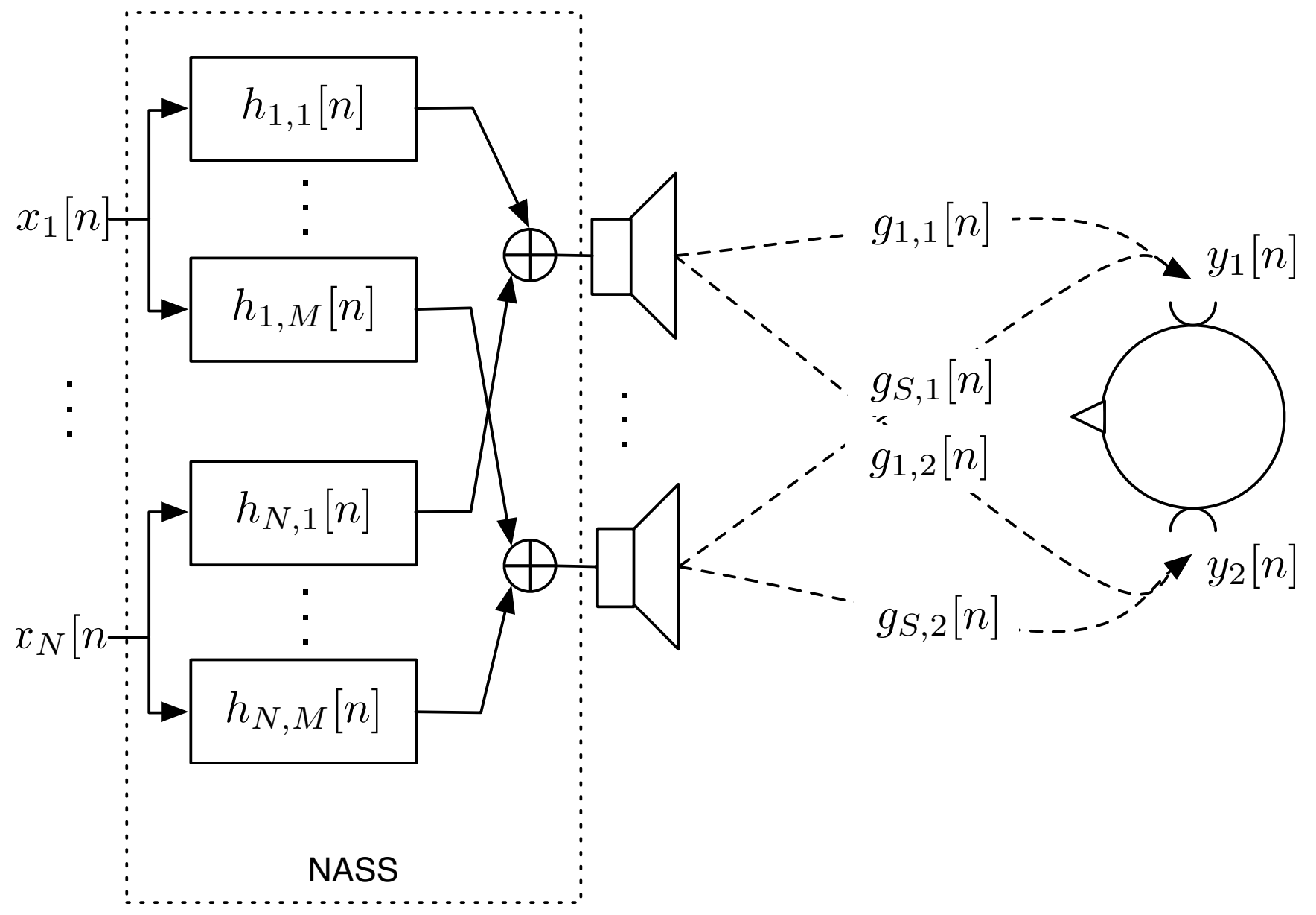
Ismael Nawfal, Joshua Atkins, Daniele Giacobello, and Stephen Nimick
Beats Electronics

Contact Information:
Beats Electronics
8600 Hayden Place
Culver City, CA, 90232, USA
Email: `ismael@beatsbydre.com`

Introduction

- There are many methods used to achieve a spatial sound field, such as Loudspeaker Binaural Rendering (LBR) (1), Wave-field Synthesis (WFS) (2), Vector-base Amplitude Panning (VBAP) (3), Higher Order Ambisonics (HOA) (4), and Equivalent Source Method (ESM) (5).
- There is limited literature on the perceptual evaluation of spatial sound synthesis methods (6).
- We introduced numerical auditory scene synthesis (NASS) in (7); a flexible numerical method that allows for broadband filter design and the incorporation of perceptual error.
- We present evaluations of timbral and spatial quality using variations of the NASS method for the task of simulating a single source outside the aperture of an 8 speaker array.

1 Methodology



NASS system for simulating binaural sources over loudspeakers with N input sources and S loudspeakers and M=2 target points.

- N_g, N_h, N_t: lengths of the acoustic path, filter, and desired response, respectively.
- D, S, M: modeling delay, number of speakers, and number of target points, respectively.
- $\mathbf{Z}$ and $\mathbf{W}$ represent spatio-temporal transforms.
- p, q, δ represent the cost function norm, constraint norm, and constraint threshold, respectively.

$$\mathbf{t}_L = [\underbrace{0,\dots,0,}_{D} t_L[0],\dots,t_L[N_t-1],\underbrace{0,\dots,0}_{N_{h-1}}]$$
$$\mathbf{t}_R = [\underbrace{0,\dots,0,}_{D} t_R[0],\dots,t_R[N_t-1],\underbrace{0,\dots,0}_{N_{h-1}}]$$
$$\mathbf{t} = [t_L, t_R]^T$$

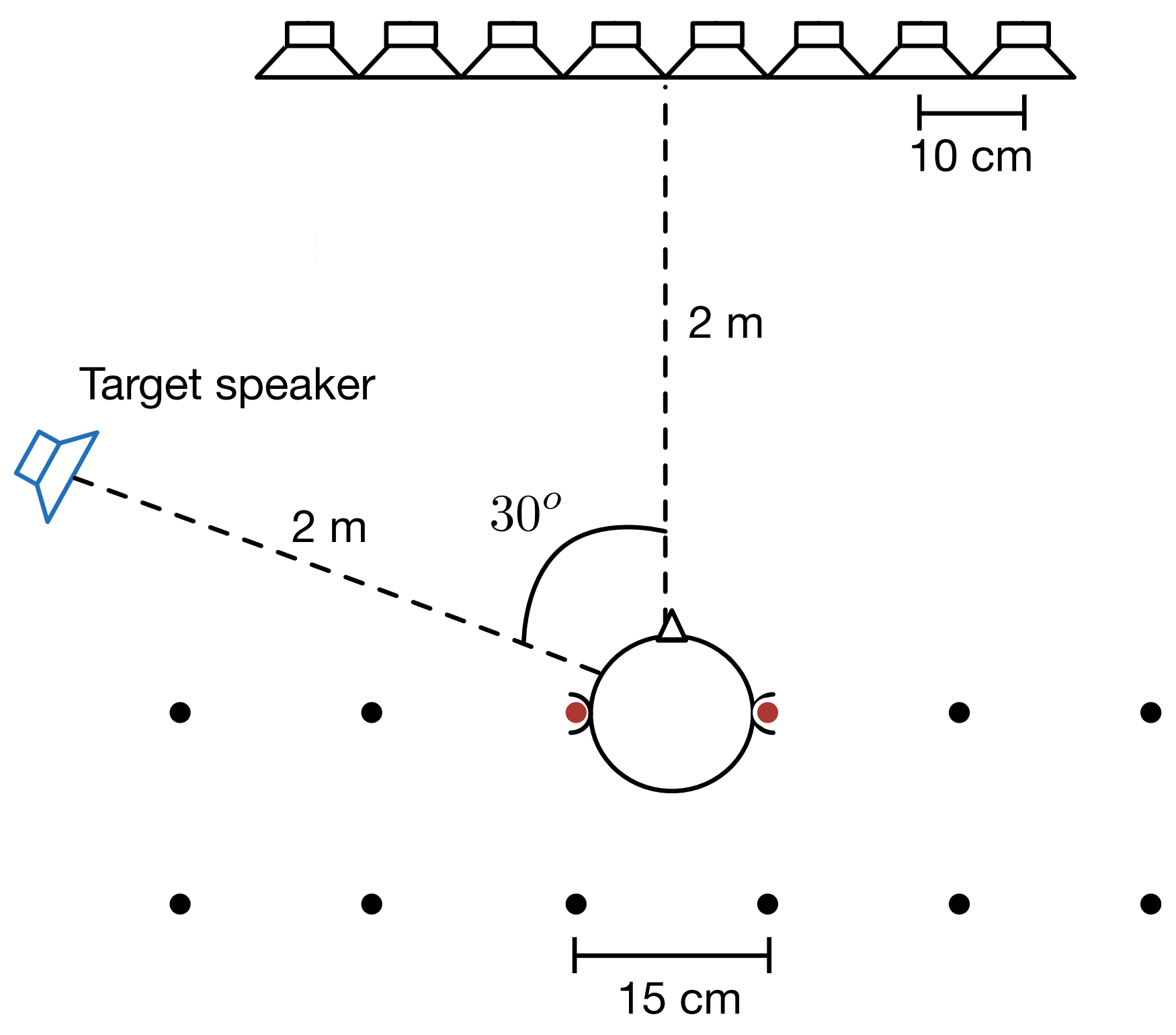
Underdetermined Case ($SN_h \geq MN_t$, Rank($\mathbf{G}$) $= MN_t$)

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{h}\|_q \;\; \text{s.t.} \;\; \underset{\substack{\text{Acoustic}\\\text{IR}}}{\mathbf{G}}\mathbf{h} = \underset{\substack{\text{Target}\\\text{Response}}}{\mathbf{t}}$$

Overdetermined Case ($SN_h < MN_t$, Rank($\mathbf{G}$) $= SN_h$)

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\underset{\substack{\text{Spatio-temporal}\\\text{Projection}}}{\mathbf{W}}(\underset{\substack{\text{Acoustic}\\\text{IR}}}{\mathbf{G}}\mathbf{h} - \underset{\substack{\text{Target}\\\text{Response}}}{\mathbf{t}})\|_p \;\; \text{s.t.} \|\underset{\text{Projection}}{\mathbf{Z}}\mathbf{h}\|_q \leq \underset{\text{Constraint}}{\delta}$$

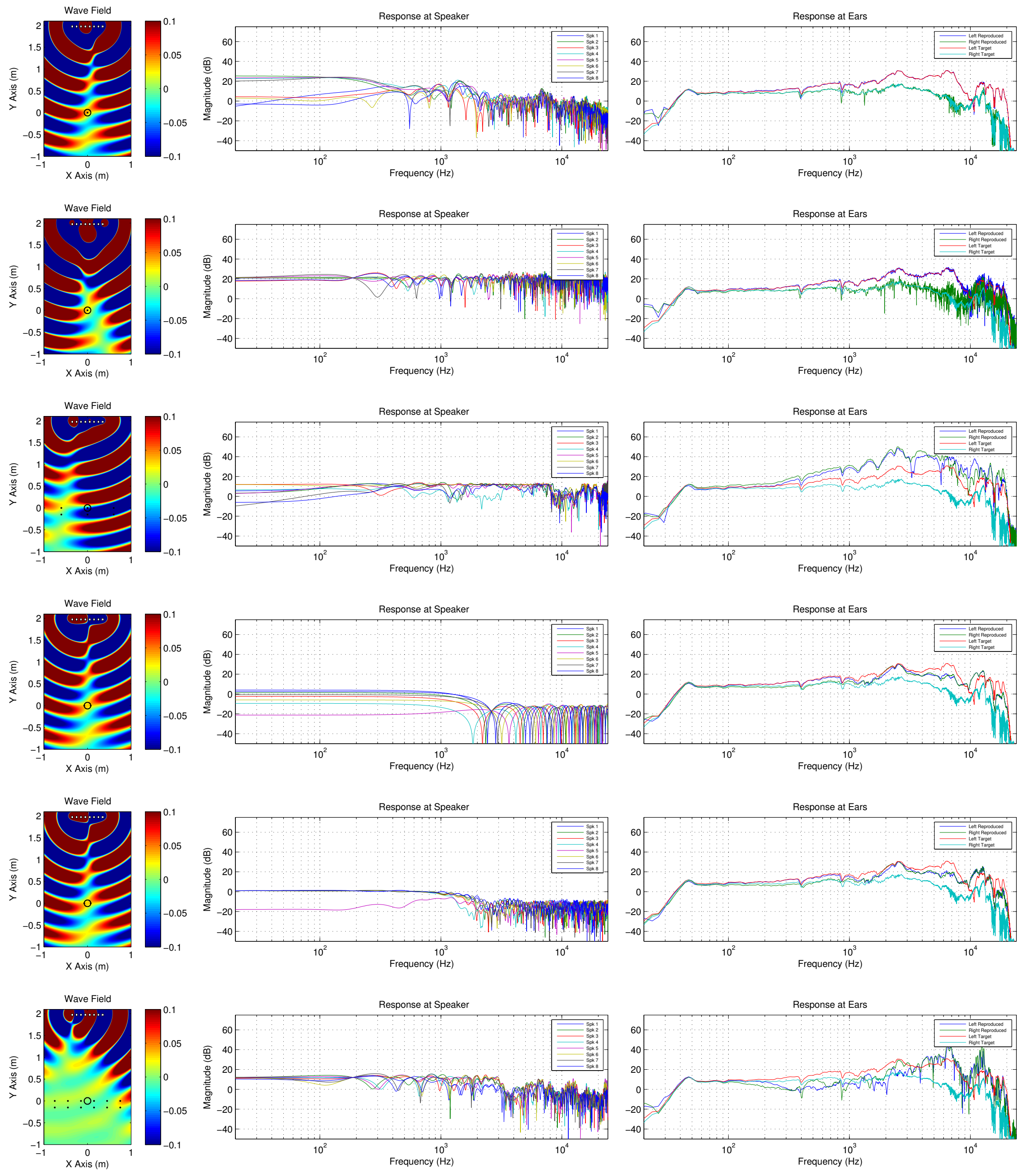
2 Evaluation



Measurement and simulation setup.

- Filters designed for 8 channel uniform linear array.
- $\mathbf{G}$ and $\mathbf{t}$ are represented by measured HRTF or a spherical wave propagation model.
- The following HRTFs and spherical wave based configurations were evaluated:
  - HRTF, $q = 2$, $M = 2$ (HRTF2_L2)
  - HRTF, $q = \infty$, $M = 2$ (HRTF2_Li)
  - HRTF, $q = \infty$, $M = 12$, $p = 2$, $\delta = 12$ dB (HRTF12_Li)
  - Spherical Wave, $q = 2$, $M = 2$ (WAVE2_L2)
  - Spherical Wave, $q = \infty$, $M = 2$ (WAVE2_Li)
  - Spherical Wave, $q = \infty$, $M = 12$, $p = 2$, $\delta = 12$ dB (WAVE12_Li)
- In all cases, N_g = N_h = 256, D = 100, and N_t = 411.

2.1 Objective Evaluation

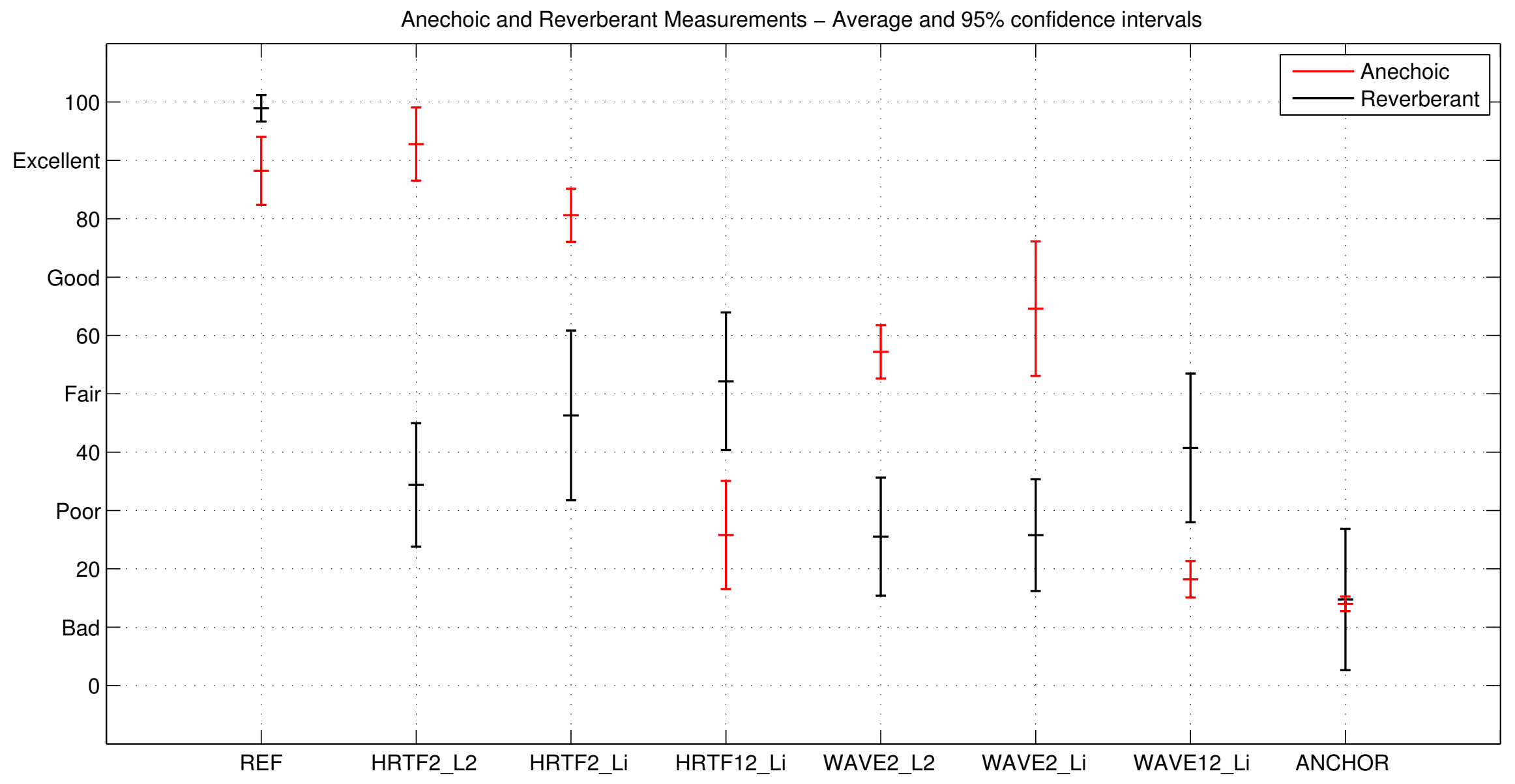


From top to bottom: HRTF2_L2, HRTF2_Li, HRTF12_Li, WAVE2_L2, WAVE2_Li, and WAVE12_Li. The graphs represent, from left to right, the wave field at 500 Hz, the filter frequency response, and the response at the ears.

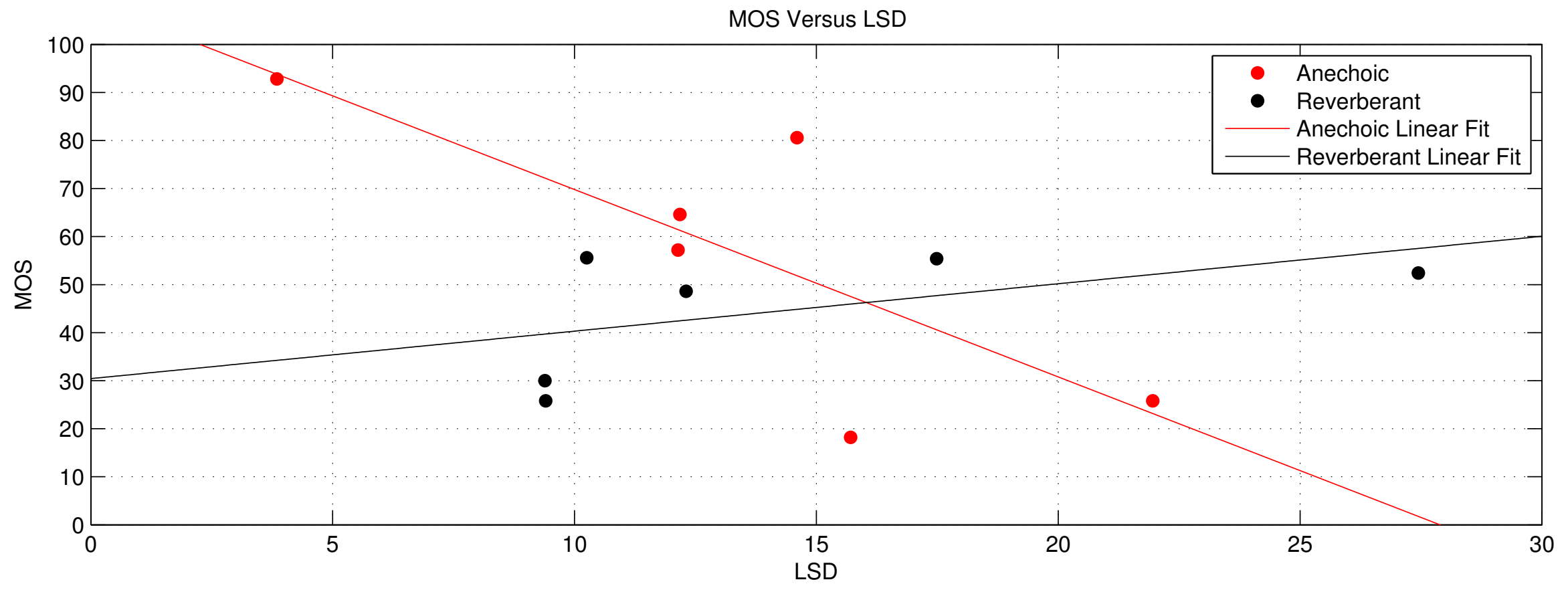
- Underdetermined cases are not spatially robust; the filters are optimized for the center position.
- The HRTF underdetermined cases closely match the expected ear responses at the central position.
- Spherical wave methods, though generating the expected acoustical waveform, don't achieve the desired responses.
- In overdetermined cases, filters are optimized for a larger spatial region resulting in increased error.

2.2 Subjective Evaluation

- 13 listeners; 9 experts and 4 naïve.
- Five audio excerpts were evaluated: castanets, pink noise, music, male voice, and female voice.
- Two tasks:
  - Array and reference speaker in anechoic room.
  - Array and reference speaker in reverberant room.
- Anchor is decorrelated and low-pass filtered.
- MUSHRA evaluations conducted on headphones.



MUSHRA results for evaluated spatial reproduction methods.



Log spectral distortion vs. MOS. Correlation Coefficients: -0.78 (anechoic) and 0.53 (reverberant).

- HRTF-based methods tended to perform better.
- Underdetermined cases performed better in anechoic cases while overdetermined cases performed better in reverberant cases.
- MOS and LSD show a strong relationship during anechoic simulation, but weak for reverberant.

3 Conclusion

- HRTF outperforms spherical wave representation.
- Mismatch between anechoic algorithm design and deployment in a real room.
- Perceptually relevant metrics should be used.
- Future work compares the proposed and conventional crosstalk-based spatial rendering and optimizes the number of speakers and filter length.

References

(1) I. Nawfal and J. Atkins, "Binaural reproduction over loudspeakers using a modified target response," *Proc. ICAD*, 2014.

(2) A.J.Berkhout, et. al., "Acoustic control by wave field synthesis," *J. ASA*, vol. 93, 1993.

(3) V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. AES*, vol. 45, no. 6, 1997.

(4) M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. AES*, vol. 53, no. 11, 2005.

(5) G. H. Koopmann, et. al., "A method for computing acoustic fields based on the principle of wave superposition," *J. ASA*, vol. 86, no. 6, 1989.

(6) F. Rumsey, et al., "QESTRAL (part 1): quality evaluation of spatial transmission and reproduction using an artificial listener," *Proc. 125th AES Conv.*, 2008.

(7) J. Atkins, I. Nawfal, and D. Giacobello, "A unified approach to numerical auditory scene synthesis using loudspeaker arrays," *Proc. EUSIPCO*, 2014.