Translation of a Higher Order Ambisonics Sound Scene Based on Parametric Decomposition

Maximilian Kentgens, Andreas Behler, Peter Jax

IEEE ICASSP 2020
From Single Spot Microphone Array Recording...

...to User-Movement-Enabled Immersive Reproduction

\( \theta \) \text{ inclination angle} \\
\( \phi \) \text{ azimuth angle}
**Signal Model**

- Higher Order Ambisonics (HOA) signal $x(\lambda, \mu) \in \mathbb{C}^{(N+1)^2}$ with Spherical Harmonics truncation order $N$:

  $$x(\lambda, \mu) = x_s(\lambda, \mu) + x_a(\lambda, \mu)$$

  $\lambda$: frame index  
  $\mu$: frequency bin

- **Direct part** $x_s(\lambda, \mu)$: variable number of $I(\lambda, \mu) \in \{0, 1, 2, \ldots, (N + 1)^2\}$ plane wave sources:

  $$x_s(\lambda, \mu) = \left[ y(\theta_1, \phi_1) \mid \ldots \mid y(\theta_I(\lambda, \mu), \phi_I(\lambda, \mu)) \right] = \text{array manifold matrix } Y_s$$

- **Ambient part** $x_a(\lambda, \mu)$: spatially diffuse

- Spatial covariance matrix (SCM):

  $$C_x(\lambda, \mu) = \mathbb{E}\{x(\lambda, \mu)x^H(\lambda, \mu)\} = \mathbb{E}\{x_s(\lambda, \mu)x_s^H(\lambda, \mu)\} + \text{diag}(\sigma_a^2)$$
Proposed System Overview

**SCM** Estimation of Spatial Covariance Matrix

$(\theta, \phi)$ Angles: (inclination, azimuth)

$\lambda$ Frame index

$\mu$ Frequency index

$w_i$ Beamforming weights

$y(\theta, \phi)$ Vector of Spherical Harmonics evaluated at $(\theta, \phi)$
**Analysis Stage**

Segregation into multiple direct time-frequency components & ambient residuum

**Variant 1:** Subspace approach using **SORTE & MUSIC**
Adopted from [Politis, Tervo & Pulkki, ICASSP’18]

- Use eigenvalue decomposition to segregate $C_x(\lambda, \mu)$ into
  - direct components subspace $U_s(\lambda, \mu)$
  - residual subspace $U_a(\lambda, \mu)$

- Model order selection: determine number $I(\lambda, \mu)$
  - find gap in eigenvalue sequence using second-order statistic of the eigenvalues (SORTE)

- Use MUSIC to find $I(\lambda, \mu)$ source directions which are local maxima in
  
  $$P^{(MUSIC)}(\theta, \phi) = \frac{1}{\|y^H(\theta, \phi)U_a\|^2}$$
Analysis Stage

Segregation into multiple direct time-frequency components & ambient residuum

- **Variant 2**: Steered Response Power (SRP) Map

  - Source directions: find all local maxima in
  \[
P^{\text{SRP}}(\theta, \phi) = y^H(\theta, \phi) C_x y(\theta, \phi)
  \]

  - Model order selection: \( I = \# \) minima found
### Analysis Stage

- **Beamforming** to extract direct components $i = 1, 2, ..., I$

  $$d_i(\lambda, \mu) = w_i^H(\lambda, \mu)x(\lambda, \mu)$$

  weights $w_i$: distortionless-response constraint in direction $(\theta_i, \phi_i)$ and null constraints for $(\theta_j, \phi_j), j \neq i$

- No semantic relation between direct components in time and frequency!
Synthesis Stage

- **Direct components:** repanning & gain adjustment according to geometric model

\[
\widetilde{x}_d = \begin{bmatrix} y(\theta'_1, \phi'_1) & \ldots & y(\theta'_1, \phi'_I) \end{bmatrix} \cdot \begin{pmatrix} \frac{r_1'}{r_1} & 0 & \ldots & \frac{r_I'}{r_I} \\ 0 & \vdots & \ddots & 0 \end{pmatrix} \cdot \begin{pmatrix} d_1(\lambda, \mu) \\ \vdots \\ d_I(\lambda, \mu) \end{pmatrix}
\]

*a priori* information on source distances \(r_1, r_2, \ldots, r_I\) needed

- **Ambient part:** cancel out direct components from input signal \(x(\lambda, \mu)\)

- In total:

\[
\widetilde{x}(\lambda, \mu) = \widetilde{x}_d(\lambda, \mu) + \widetilde{x}_a(\lambda, \mu)
\]
Discussion

- **Early reflections** not considered in signal model
  - Likely to be modified as if they were direct sound
  - **Precedence effect**: early reflections do not contribute to perceived source direction
Subjective Evaluation

MUSHRA Listening Test for Localization & Overall Audio Quality | 18 Participants

- HOA scene (order N=4) with two sound sources
- Translation to $\vec{p}'_1$ and $\vec{p}'_2$

- Two variants of the proposed algorithm:
  - MUSIC/SORTE
  - SRP-based

- For reference:
  - ORACLE: direction estimates replaced by actual source directions
  - Low anchor (AL): omni-directional sound, 4 kHz low-pass filtered
  - Mid anchor (AM): non-translated input signal $x(\lambda, \mu)$
  - Hidden reference (HR): ground truth at $\vec{p}''_1$ and $\vec{p}''_2$, respectively

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>T60</td>
<td>0.4 s</td>
</tr>
<tr>
<td>Source #1</td>
<td>Pop music</td>
</tr>
<tr>
<td>Source #2</td>
<td>Male speaker</td>
</tr>
<tr>
<td>Noise</td>
<td>Diffuse (−30 dB)</td>
</tr>
<tr>
<td>Sample duration</td>
<td>8 s</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>16 kHz</td>
</tr>
<tr>
<td>SH order N</td>
<td>4</td>
</tr>
<tr>
<td>Frame length</td>
<td>512 samples</td>
</tr>
<tr>
<td>Frame overlap</td>
<td>50%</td>
</tr>
<tr>
<td>Sub-band grouping</td>
<td>Critical bands</td>
</tr>
<tr>
<td>Rec. smoothing</td>
<td>0.1 s</td>
</tr>
<tr>
<td>Decay time</td>
<td>900</td>
</tr>
<tr>
<td>Spatial sampling</td>
<td>990</td>
</tr>
<tr>
<td>Oracle distance $r_o$</td>
<td>2 m</td>
</tr>
</tbody>
</table>
Subjective Evaluation

MUSHRA Listening Test for Localization & Overall Audio Quality | 18 Participants

- $T_{60}$: 0.4s
- Source #1: pop music
- Source #2: male speaker
- Noise: diffuse (−30dB)
- Sample duration: 8s

- HR: hidden reference
- AL: low anchor
- AM: mid anchor (input signal)
- OR: proposed method with oracle information on source directions
- MU: proposed method with MUSIC
- SR: proposed method with SRP
Conclusions

- Novel 3DoF+ system for translation in scene-based HOA content
  - Only single capture device needed

- Excellent subjective performance...
  - ... despite model error due to early reflections not treated differently than direct sound (→ psychoacoustic precedence effect)
  - Median subject ratings in localization between 91% and 99% (SRP method)

- Applications: VR/AR systems, immersive teleconferencing & telepresence