Analysis and Prediction of Protected-Ear Localization

Griffin D. Romigh, Ph.D.
Air Force Research Labs
2610 Seventh St., Area B, Bldg. 441
WPAFB, OH 45433
Overview

• Spatial Hearing and HRTFs
• A Different Approach
• An Efficient Representation
• Applying Bayesian estimation
• Modeling individual differences
• Summary of Contributions
Spatial Hearing

**Interaural Level Difference (ILD)**
- Sound energy is scattered by the head
- Less energy arrives at the far ear
- Results in a level difference at the two ears

- Smaller wavelengths are attenuated more
- Results in larger ILDs at high frequencies
Spatial Hearing

Interaural Level Difference (ILD)

- Behind
- Above
- Front

Lateral Angle

Left

Top

Front

Intraconic Angle, $\phi$

Lateral Angle, $\theta$

ILD, dB

-20
-15
-10
-5
0
5
10
15
20
Spatial Hearing

Interaural Time Difference (ITD)
- Sound arrives at near ear before far ear
- Results in a arrival and phase difference
- Becomes ambiguous at high frequencies
Spatial Hearing

Interaural Time Difference (ITD)

Lateral Angle

-50  0  50

Front  Above  Behind

ITD, samples

-20  -15  -10  -5  0  5  10  15  20

Lateral Angle, $\theta$

Intraconic Angle, $\phi$
Spectral Cues
- High frequency cues due to pinna
- Lower frequency cue due to shoulders
- Perceptually weighted to favor closer ear
Spatial Hearing

Spectral Cues

Frequency, kHz

Level, dB

Front
Above
Behind
Head-Related Transfer Functions:

\[ HRTF_{r, \theta, \phi}(n) \]

\[ ITD_{\theta, \phi} \]

\[ HRTF_{1, \theta, \phi}(n) \]

\[ S[n] \]

\[ y[n] \]

Physical Apparatus

Signal Post-Processing

Desired Quantities
HRTF ↔ Spatial Hearing Cues

- **Level Difference (ILD)**
- **Time Difference (ITD)**

**Right**
- Magnitude Response
- Phase Response

**Left**
- Magnitude Response
- Phase Response

**Interaural Level Difference (ILD)**

**Level Timing**
HRTF ↔ Spatial Hearing Cues

Left
Magnitude Response
Phase Response

Right
Magnitude Response
Phase Response

Level Timing

Interaural
Time Difference (ITD)
Binaural Synthesis
Spatial Auditory Displays:

- Guidance systems
- Hearing Restoration
- Virtual Reality
- Augmented Reality
HRTFs: Idiosyncrasy

By location

By individual

HRTF Magnitude (dB)

Frequency (kHz)

Rear

Top

Front

Log frequency (Hz)

2000 4000 8000 15000

0 50 0 0 -200
HRTFs: Idiosyncrasy

• SADs need Individual HRTFs
• Otherwise:
  1. No sense of elevation
  2. Frequent FB Reversals
  3. Localized “In the Head”

- Brungart et al., 2009
# HRTFs: Spatial Measurement

<table>
<thead>
<tr>
<th>Fixed Spherical Array</th>
<th>Rotating Arc Array</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pros:</strong> Fast (5 – 10 min)</td>
<td><strong>Cons:</strong> Expensive, Permanent</td>
</tr>
<tr>
<td><strong>Cons:</strong> Expensive, Permanent</td>
<td><strong>Pros:</strong> Cheaper, Temporary</td>
</tr>
<tr>
<td></td>
<td><strong>Cons:</strong> Slow (1 – 2 hours)</td>
</tr>
</tbody>
</table>
The Problem

How can we get an HRTF for every spatial angle with as few physical measurements as possible?
## Previous Methods:

### Measurement
- Reciprocity
- Spectral asynchrony

**• Same Equipment**
- Less Time
- Perceptually Equivalent Performance

<table>
<thead>
<tr>
<th>Parallel</th>
<th>Baseline HRTF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• 277 locations</td>
</tr>
<tr>
<td></td>
<td>• 256 taps</td>
</tr>
</tbody>
</table>

| Non-Acoustic Interpolation |

<table>
<thead>
<tr>
<th>Naive</th>
<th>Statistical</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Linear kNN</td>
<td>- Pattern Matching</td>
</tr>
<tr>
<td>- Spherical Basis</td>
<td>- Neural Net</td>
</tr>
</tbody>
</table>

**• Less Equipment**
- Less Time
- Perceptually Equivalent Performance

<table>
<thead>
<tr>
<th>Subjective Selection</th>
<th>Structural Models</th>
<th>Generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Most Externalized</td>
<td>- “Snowman”</td>
<td>- Averaging</td>
</tr>
<tr>
<td>- Vertical Lift</td>
<td>- Anthropometric</td>
<td>- Super Subject</td>
</tr>
</tbody>
</table>

**• Least Equipment**
- Less Time
- Poor Performance
Irrelevant Spectral Details

- Auditory system has limited spectral resolution
- This results in fine spectral details being averaged out
- Most impactful at high frequencies

Maybe we can get away with smoothing the spatial detail
Spatial Representation:

\[ |H_{\theta,\phi}(\omega)| \Leftrightarrow |H_{\omega}(\theta, \phi)| \]
Spherical Harmonics

\[ Y_{nm}(\phi, \theta) = \begin{cases} 
N_n^m P_n^m(\cos \theta) \cos(m\phi) & \text{if } m \geq 0 \\
N_n^m P_n^{-|m|}(\cos \theta) \sin(m\phi) & \text{if } m < 0 
\end{cases} \]

**Associated Legendre Polynomials**

Orthogonal functions for lateral angle
Spherical Harmonics

\[ Y_{nm}(\phi, \theta) = \begin{cases} 
N_n^m P_n^m(\cos \theta) \cos(m\phi) & \text{if } m \geq 0 \\
N_n^m P_n|m|\cos \theta \sin(m\phi) & \text{if } m < 0 
\end{cases} \]

Orthogonal functions in intraconic angle
Spherical Harmonics

Orthonormal basis over the continuous sphere

\[
\int_{\theta = -\pi/2}^{\pi/2} \int_{\varphi = -\pi}^{\pi} Y_{nm}(\varphi, \theta) Y_{n'm'}(\varphi, \theta) = \delta_{nn'} \delta_{mm'}
\]

Allow us to represent any square integrable spherical function with a set of SH coefficients

\[
f(\phi, \theta) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} Y_{nm}(\phi, \theta) C_{nm}
\]

**** We can do Fourier analysis on a sphere ***
Spherical Harmonics
Practical SH Expansion

Re-cast problem into system of linear equations

\[ f = Yc \]

where \[ f = [f(\phi_0, \theta_0), f(\phi_1, \theta_1), \ldots, f(\phi_S, \theta_S)]^T \]

\[ c = [C_{00}, C_{1-1}, C_{10}, C_{11}, \ldots, C_{PP}]^T \]

\[ Y = \begin{bmatrix}
Y_{00}(\phi_1, \theta_1) & Y_{-11}(\phi_1, \theta_1) & \cdots & Y_{PP}(\phi_1, \theta_1) \\
Y_{00}(\phi_2, \theta_2) & Y_{-11}(\phi_2, \theta_2) & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
Y_{00}(\phi_S, \theta_S) & Y_{-11}(\phi_S, \theta_S) & \cdots & Y_{PP}(\phi_S, \theta_S)
\end{bmatrix} \]

Simple least-squares solution

\[ \hat{c} = (Y^TY)^{-1}Y^Tf \]
Truncating the expansion provides spatial smoothing

\[ f(\phi, \theta) = \sum_{n=0}^{P} \sum_{m=-n}^{n} Y_{nm}(\phi, \theta) C_{nm} \]
Localization task
• 8 Subjects
• 250-ms noise bursts
• 245 locations
Recap…

- New SH-based HRTF representation
  - Spatially continuous
  - Reduces irrelevant spatial variation
  - Localization equivalent to full HRTF
  - Reduces # of parameters by 95% w.r.t. baseline HRTF

Can non-individualized HRTFs provide information about a new HRTF measurement?
Bayesian HRTF Model

Model all HRTFs as belonging to the same underlying distribution

\[ f = Yc + n \]

\[ c : \mathcal{N}(m_c, R_{cc}) \]

\[ n : \mathcal{N}(0, \sigma^2 I) \]

Independent

Non-individual information is incorporated through hyper-parameters

\[ R_{cc} = \begin{bmatrix} \sigma_{00}^2 & 0 & \ldots & 0 \\ 0 & \sigma_{11}^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_{PP}^2 \end{bmatrix} \]

\[ m_c = \begin{bmatrix} E[C_{00}] \\ E[C_{-11}] \\ \vdots \\ E[C_{PP}] \end{bmatrix} \]
Bayesian Estimation

Estimation via MMSE Estimator

\[ \hat{c} = E[c|f] = m_c + R_{cc} Y^T (Y R_{cc} Y^T + \sigma^2 I)^{-1} (f - Y m_c) \]

- Estimated SH coefficients for individual
- Difference between individual HRTF and average HRTF at measurement locations

Estimator is based on how the HRTF is different from average…
Bayesian Estimation

Estimation via MMSE Estimator

\[ \hat{c} = E[c|f] = m_c + R_{cc} Y^T (Y R_{cc} Y^T + \sigma^2 I)^{-1} (f - Y m_c) \]

- Average SH coefficients
- Innovations from individualized measurements (bias)

Assuming hyper-parameters are already known…
We have fixed unknown model parameters….

\[ c : \mathcal{N}(m_c, R_{cc}) \]

**Classical Estimation (MVUB)**

\[
\hat{m}_c = \frac{1}{M} \sum_{i=1}^{M} c_i
\]

\[
\hat{\sigma}_j^2 = \frac{1}{M-1} \sum_{i=1}^{M} (c_i[j] - \hat{m}_c[j])^2
\]

\[
R_{cc} = \begin{bmatrix}
\sigma_{00}^2 & 0 & \cdots & 0 \\
0 & \sigma_{11}^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma_{PP}^2
\end{bmatrix}
\]

Assuming we have \(M\) individuals SH coefficients…
Estimating Hyper-parameters

We have fixed unknown model parameters….

\[ c : \mathcal{N}(m_c, R_{cc}) \]

**Classical Estimation (MVUB)**

\[ \hat{m}_c = \frac{1}{M} \sum_{i=1}^{M} c_i \]

\[ \hat{\sigma}_j^2 = \frac{1}{M-1} \sum_{i=1}^{M} (c_i[j] - \hat{m}_c[j])^2 \]

But we can’t measure SH coefficients. We need a way to estimate both simultaneously.

\[ R_{cc} = \begin{bmatrix}
\sigma_{00}^2 & 0 & \cdots & 0 \\
0 & \sigma_{11}^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma_{PP}^2 
\end{bmatrix} \]

Assuming we have \( M \) individuals’ SH coefficients…
Compute parameters and hyper-parameters iteratively

1. Initialize $R_{cc}$ and $m_c$ to arbitrary values

2. Calculate Bayesian estimates of SH coefficients

$$\hat{c} = m_c + R_{cc} Y^T (Y R_{cc} Y^T + \sigma^2 I)^{-1} (f - Y m_c)$$

3. Update estimates of $R_{cc}$ and $m_c$ using new coefficient values

$$\hat{m}_c = \frac{1}{M} \sum_{i=1}^{M} c_i$$

$$\hat{\sigma}_j^2 = \frac{1}{M-1} \sum_{i=1}^{M} (c_{i[j]} - \hat{m}_c[j])^2$$

4. Repeat 2 and 3 until estimates converge
Training the model
- EM based
- 44 subjects
- 274 spatial samples

Testing the model
- Bayesian estimation
- 10 subjects
- varied # of samples

Better reconstruction performance with fewer spatial samples
Computational Performance

Subject 1

Subject 2

Subject 3
Localization Task
- 6 Subjects
- 250-ms noise bursts
- 245 locations

Equivalent performance with as few as 12 measurements
Recap…

- Bayesian HRTF model
  - Models general HRTF distribution as MVN
  - Individualized HRTF represents a single sample

- Bayesian HRTF Estimation
  - Non-individualized HRTFs provide “template”
  - Individualized measurements personalize the template
  - Much fewer measurements are needed (~12 distributed)

How do HRTFs differ amongst individuals?
Further Model Reduction

- Non-individual localization is bad mostly in polar dimension

- Implies inter-subject differences in HRTFs account for polar cue difference

If we can separate out polar cues we might only need to estimate those!
Further Model Reduction

Sectoral coefficients capture mostly intraconic variation
Further Model Reduction

Sectoral coefficients contain most of the inter-subject variance

These coefficients may be all that need to be individualized
Separate individual (Sectoral) and non-individual (Lateral) features.

\[ h = h_{Sec} + h_{Lat} \]
\[ = Y_{Sec}c_{Sec} + Y_{Lat}c_{Lat} \]

Only sectoral coefficients need to be estimated. The rest can be average values.
Sectoral HRTF Model:

Subject 1

Subject 2

Subject 3

Sectoral model does capture the intraconic HRTF features
Estimating the Sectoral HRTF:

Estimate Sectoral HRTF with average lateral coefficients.

\[ \hat{h}_{\text{Sec}} \approx h - Y_{\text{Lat}} \bar{c}_{\text{Lat}} \]
\[ \approx Y_{\text{Sec}} c_{\text{Sec}} \]

Now use Bayesian technique with Sectoral basis functions.

\[ \hat{c}_{\text{Sec}} = E[c|h_{\text{Sec}}] \]
\[ = \bar{c}_{\text{Sec}} + R_{\text{Sec}} Y_{\text{Sec}}^T (Y_{\text{Sec}} R_{\text{Sec}} Y_{\text{Sec}}^T + \sigma^2 I)^{-1} (\hat{h}_{\text{Sec}} - Y_{\text{Sec}} \bar{c}_{\text{Sec}}) \]

Estimated Sectoral HRTF
Why the median plane?

- Bad DC estimate off midline

- Sectoral harmonics contain no energy off the midline at high orders
Perceptual Evaluation

Localization Task
- 6 Subjects
- 250-ms noise bursts
- 245 locations
- HRTFs
  - Full 4th-Order (SH4)
  - 4th-Order Sectoral (SEC4)

- Statistically similar performance with as few as 12 measurements
- No performance difference from Full SH model
Maintains good performance off the midline
Recap…

• **Sectoral HRTF Model**
  - Sectoral coefficients contain large inter-subject variance
  - Only sectoral coefficients need to be individualized
  - The rest of the coefficients can be replaced with average
  - 98% fewer parameters w.r.t. baseline HRTF

• **Median-Plane Estimation**
  - Sectoral harmonics vary mainly in intraconic dimension
  - Values can be estimated from median plane measurements
Thank You
Head-tracking and/or prediction of anthropometric parameters via webcam
Project Ideas

HRTF measurement using a single speaker and a head tracker
HRTF-based sound source localization/segregation from a binaural recording (many recordings available)