(T11) Natural and Augmented Listening for VR and AR/MR

Woon-Seng Gan*, Jianjun He*, Rishabh Ranjan+, Rishabh Gupta*

* Digital Signal Processing Laboratory
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
{ewsgan, rishabh007}@ntu.edu.sg

* Maxim Integrated Product Inc
San Jose, California, USA
Jianjun.He@maximintegrated.com

* Immerzen Labs Pte. Ltd.,
Singapore
rishabh@immerzenlabs.com
# Outline of Tutorial

### Module A: Introduction
- Definition of VR, AR/MR
- Fundamentals in Human Listening and Spatial Audio
- Brief Overview of Perceptual Evaluation
- Why VR, AR/MR needs Immersive Spatial Audio
- Outline of Following Modules

### Module B: Binaural 3D Audio for VR, AR/MR
- Overview of 3D Audio Reproduction
- Binaural Rendering for VR/AR/MR
- HRTF Individualization (including measurements)
- Equalization
- Movement Tracking
- Environment Rendering
- Integrated System
- Conclusion

### Module C: Augmented/Mixed Reality 3D Audio
- Types of Augmented/Mixed Reality Audio
- Natural Listening in AR/MR: An Overview
- Signal Processing Techniques in NAL
- Hear Through of Real Sound
- Virtual Sound Augmented with Real Sound
- Acoustic Environment Estimation and Rendering
- Integrated System
- Conclusion

### Module D: Summary and Future Trends
- Summary of key Techniques
- Spatial Audio Tools
- Emerging Applications of VR/AR Audio
- Challenges and Future Research Trends
Module A

Introduction

1. Definition of VR, AR/MR
2. Fundamentals in Human Listening and Spatial Audio
3. Brief Overview of Perceptual Evaluation
4. Why VR, AR/MR needs Spatial Audio?
5. Outline of Following Modules

Physics of Sound Propagation + Psychophysics of Auditory Perception
Definitions of VR, AR/MR

**Virtual Reality (VR):**
Immersive multimedia (or computer-simulated reality) to replicate an environment that simulates a physical presence in real or imaginary world. Allow user to interact in the VR world.

**Augmented Reality (AR):**
In a real world environment whose elements are augmented (overlays) by computer-generated (CG) sensory input (sound, video, data). However, the real-world content and the CG content are not respond/react to each other.

**Mixed Reality (MR):**
Merging of real and virtual worlds to produce new environments (physical and virtual objects co-exist and interact in real time).

- Google Glass
- Bose AR
- Microsoft Hololens
- Magic Leap
- Meta 2
- HTC Vive Pro

- Google cardboard
- Samsung Gear VR
- Oculus Rift
From PC Flat Screen to Full 360 VR Experience

Flat Screen
- Pre-rendering
- 0 DoF

Cinematic VR
- Real-time rendering; but environment is still pre-rendered
- 3 DoF

Full VR
- Changing environment & interaction
- 6 DoF

Illustration by Santi, image credit: Freepik.com
From Real Sound to Virtual Reality Audio

Real

Capture

Transmission

Virtual

Reproduction
Natural Listening in AR/MR

Real

Virtual

Capture

AR/MR

See through field of vision:
30 to 90 degree (virtual)

Audio field of listening:
360 degree

Microsoft Hololens
Playback device for current VR, AR/MR Headgear

- Without integrated headphones

- With integrated headphones and built-in speakers

- Samsung VR Gear
- HTC Vive

- Samsung HMD Odyssey
- HTC Vive Delux
- Oculus Go
- Microsoft Hololens
### Wearable playback devices for VR, AR/MR

<table>
<thead>
<tr>
<th>Closed Back Headphones</th>
<th>Opened Back Headphones</th>
<th>In-Ear Monitors</th>
<th>Intra-concha / supra aural Earphones</th>
<th>Built-in Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Good isolation &amp; bass&lt;br&gt;• Block off from environment</td>
<td>• Less isolation &amp; sound leakage&lt;br&gt;• Good environmental awareness&lt;br&gt;• spacious</td>
<td>• Excellent isolation and frequency responses&lt;br&gt;• Block off environmental noise</td>
<td>• Poor isolation/response variances&lt;br&gt;• Some environmental awareness&lt;br&gt;• Lightweight</td>
<td>• Poor isolation &amp; bass&lt;br&gt;• Leakage</td>
</tr>
</tbody>
</table>

**Which type of playback devices should be used for VR, AR/MR?**

- **VR** requires isolation of the real sound to get immersed in virtual sound.
- **AR/MR** requires Transparent Listening to blend virtual with real sound.
A.2 Overview of Human Auditory Model

- How do we hear?
- Binaural cues for localization of single source
- Cone of confusion and head movements
- Spectral cues
- HRTF definition
How do we hear?

Primary Auditory Cues:
- Interaural Level Difference
- Interaural Time Difference
- **Monoaural Spectral Cues (pinna)**
- Torso and Body reflection & diffraction
- Environmental (Direct/Reverberation Ratio)
- **Head Motion**
- Familiarity with sound source

Image Source: http://www.soundproofingcompany.com/soundproofing101/what-is-sound/
Binaural cues for localization of single source

- Compare sound received at two ears
  - **Interaural Level Differences (ILD)**
    - Effective for high frequencies above 1.5 kHz
    - Head size (~22 cm) > wavelength
    - Smallest detectable ILD = 0.5 dB

- **Interaural Time Differences (ITD)**
  - Effective for low frequencies below 1.5 kHz
  - Rayleigh’s duplex theory of ILD and ITD
  - Smallest detectable ITD = 13 μs

Pictures modified from [W. M. Hartmann, 1999]
Precedence (Law of 1\textsuperscript{st} Wavefront) Effect

- First wavefront determines localization
- Used in sound reinforcement system;
- But when played back form headphones, the effect is very different.
## Equations for Interaural time difference (ITD)

<table>
<thead>
<tr>
<th>S/No.</th>
<th>Technique Name</th>
<th>ITD formula</th>
<th>Parameter definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Woodworth Formula and extensions [Minnaar, 2000]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Original:</strong> $ITD = \frac{a}{c}(\sin\theta + \theta), 0 \leq \theta \leq \pi/2$</td>
<td></td>
<td>$\alpha$ - radius of sphere $c$ - speed of sound $\theta$ - azimuth angle $\phi$ - elevation angle</td>
</tr>
<tr>
<td></td>
<td><strong>Extension 1:</strong> $ITD = \frac{a}{c}[\arcsin(\cos\phi\sin\theta) + \cos\phi\sin\theta]$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Extension 2:</strong> $ITD = \frac{a}{c}(\sin\theta + \theta)\cos\phi$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Interaural Phase Delay [Blauert, 1997; Xie, 2013]</td>
<td>$ITD_p(\theta, f) = \frac{\Delta\psi}{2\pi f} = -\frac{\psi_L - \psi_R}{2\pi f}$</td>
<td>$\psi_L, \psi_R$ is the phase of sound pressure for left ear and right ear respectively and is the frequency at which ITD is calculated</td>
</tr>
<tr>
<td>3.</td>
<td>Interaural Cross correlation (IACC) and related methods [Katz, 2014]</td>
<td>$IACC(\theta, \tau) = \frac{\int p_L(\theta, t)p_R(\theta, t + \tau)dt}{\sqrt{\int_{t_1}^{t_2} p_L^2(\theta, t)dt \int_{t_1}^{t_2} p_R^2(\theta, t)dt}}$</td>
<td>$p_L(\theta, t)p_R(\theta, t)$ measured HRIR for left and right ear $\theta$ incident angle, $t_1=0$ $t_2 = \text{max of the lengths of } p_L(\theta, t) \text{ and } p_R(\theta, t)$</td>
</tr>
<tr>
<td></td>
<td><strong>Method 1:</strong> Max IACC $ITD(\theta) = \arg\max IACC(\theta, \tau),</td>
<td>$</td>
<td>\tau</td>
</tr>
<tr>
<td></td>
<td><strong>Method 2:</strong> Centroid of IACC $ITD(\theta) = C_\tau(IACC(\theta, \tau))$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Group delay Methods [Minnaar, 2000]</td>
<td>$ITD = IGD_0 = \text{abs} \left( \tau_g(0)<em>{\text{left}} - \tau_g(0)</em>{\text{right}} \right)$</td>
<td>$\tau_g(0)_{\text{left/right}}$ - group delay for excess phase component of HRTFs for left/right channel</td>
</tr>
</tbody>
</table>
The equation for ILD is given by:

\[
ILD(r, \theta, \phi, f) = 20 \log \left| \frac{P_R(r, \theta, \phi, f)}{P_L(r, \theta, \phi, f)} \right|
\]

\(P_L(r, \theta, \phi, f), P_R(r, \theta, \phi, f)\) are the freq-domain sound-pressures at left and right ears.

If head and two ears are approximated by rigid sphere and two opposite points on spherical surface, the pressures can be calculated as scattering solutions to rigid head.

[Xie, 2013]
Cone of confusion and dynamic cues

- Similar ITD and ILD due to:
  - Cone of confusion
  Sources A & B; Sources C & D have identical ITD and ILD
  - Media Plane (extreme case of cone of confusion)

*How can we tell if the sound is in front or behind?*

- We need another sound localization cue!

- Head rotation as a dynamic cue can help resolve this
Modeling of Sound Scattering (Human body & ears)

- Sound interacts with torso, head, external ears and arrives at the two ear canals: Scattering Effect
- Provide **filtering cues for localization**

![Diagram showing scattering effect and acoustic gain components](http://hearinghealthmatters.org/waynesworld/2014/human-ear-canal-viii/#refmark-1)
Head-Related Transfer Function (HRTF)

- HRTFs encode filter characteristics for a sound arriving from a specific direction.
- Many high-frequency details due to pinna scattering.
- **How we measure or generate HRTF?**

![HRTF Graphs]

- **HRTF 60 above ear**
- **HRTF 0 ear level**
- **HRTF 40 below ear**

Ipsilateral HRTF nearer to source (shown above)
Animation of HRIR/HRTF: Database from CIPIC (S03)

Azimuth

HRIR Horizontal Plane Azm = 0

Elevation

HRIR Median Plane Ele = -45

HRTF Horizontal Plane Azm = 0

HRTF Median Plane Ele = -45
Natural and augmented listening for VR, AR/MR: Module A

- Human pinna is found to be as idiosyncratic as the fingerprint.
- Scattering wave around ears are different.
- HRTFs are highly individual and differs substantially from one subject to the other.
- For perfect 3D audio playback, individualized recordings/HRTFs and individualized headphone equalization are required.
Distance Localization Cue

- Loudness
  - Familiar sound sources
  - Moving sound sources
- Initial Time Delay
- Ratio of Direct and Reverberant energy
- Motion Parallax (near field)
- ILD (near field)
- High Frequency Damping (far field)
Reverberation

Spatialization (Anechoic) only solves direct sound propagation

Reverberation (Ambience) Provides indirect audio cues
Spatial audio rendering is concerned with the linking of physics and auditory perceptual effects.

Not to overly rely on complex mathematical tools; just a tool for analysis.

A highly accurate design of spatial audio processing system may not be required for plausible perceptual performance.

Allow some degrees of mathematical errors and measurement errors.
A.3 Perceptual Quality Evaluation

- Aim and overall process flow for evaluation of sound quality
- Key aspects of perceptual quality assessment
- Key standards and protocols
Aims and process flow

• Aim of listening tests is to determine whether the recorded or reproduced sound recreate the similar “acoustic sensation” for the listener as the original event.

Picture from [Schoeffler et al., 2015]
Key aspects for perceptual quality assessment

- Experimental design
- Selection of listening panel
- Test methods
- Attributes
- Program material
- Reproduction devices
- Listening conditions
- Statistical representation of data
- Presentation of results

[Schoeffler et al., 2015]
<table>
<thead>
<tr>
<th>S. No.</th>
<th>Standard name/test</th>
<th>Title</th>
<th>Remarks</th>
</tr>
</thead>
</table>
| 1      | ITU-R BS.1116-3    | Methods for subjective assessment of small impairments | • Double blind triple stimuli with hidden reference  
• Uses a test form with an open given external reference and a **five point scale**.  
• The test is designed to emphasize **small differences** between test items and reference. |
| 2      | ITU-R BS.1534 (MUSHRA Test) | Method of subjective assessment of intermediate quality level of audio systems | • Double blind **MUlti-Stimuli test with Hidden Reference and Anchor (MUSHRA)** with **continuous scale**  
• Hundred point scale with five verbal descriptor labels used. |
| 3      | ITU-R WP6C (under progress) | Multi stimuli method for quality evaluation | • Will have no open reference, to make it applicable to all the test cases where a reference is not defined.  
• Planned to include additional attributes and an ideal profiling method, which aims at finding out how close products are. |
| 4.     | ABX Test (force-choice testing to detect any perceptual difference between two stimuli in double-blind trials) | | • Subject is presented with two category of known stimuli (A and B) and ask to identify category of unknown stimuli (X)  
• If X cannot be identified with a low p-value, then no perceptual difference between A and B. |

[Bech and Zacharov, 2007]
Experiencing VR, AR/MR with 360° SPATIAL AUDIO

The trumpet can be heard on front right.

Spatial Audio
- Enhance VR, AR/MR experience
- Work best with visual reinforcement
- Needs real-time rendering of dynamic audio cues (head rotation and sound object)
Spatial Audio Technologies for Immersive VR/AR/MR

Module A
- Natural and augmented listening for VR, AR, MR

Module B
- Spatial Audio Formats
  - Object, Ambisonics
  - Parametric processing
- Individualized Binaural Rendering
  - Individualized HRTFs
  - Equalization
- Dynamic Binaural Synthesis
  - Head tracking
  - Position tracking

Module C
- Environment Estimation
  - Depth camera
  - Reverberation fingerprint
  - Machine learning
- Environment Rendering
  - Wave based
  - Geometrical based
  - Perceptual based
- Virtual & Physical Sound Fusion
  - Adaptive equalization
  - Hear-through processing
References in Module A

Module B
Binaural 3D audio for VR, AR/MR

1. Overview of 3D Audio Reproduction
2. Binaural Rendering for VR/AR/MR
3. HRTF Individualization (including measurements)
4. Equalization
5. Movement Tracking
6. Environment Rendering
7. Integrated System
8. Conclusion
B.1 Source-medium-receiver of spatial audio reproduction

Immersive Audio

“being there”

Source
Audio content

Channel

Object

Scene

Medium
Playback system

Loudspeakers

Headphones

Receiver
Human

Ears

Movements

WS Gan, JJ He, R Ranjan, R Gupta

Natural and augmented listening for VR, AR/MR: Module B

16th Apr. 2018

B.2 /100
MPEG-H 3D audio standard (2015)

MPEG-H bitstream → USAC-3D Decoder → Format Converter → Converted Channels → Mixer → Binaural Renderer → Headphone feeds

MPEG-H bitstream → USAC-3D Decoder → Object Renderer → Converted Channels → Mixer → Binaural Renderer → Headphone feeds

MPEG-H bitstream → USAC-3D Decoder → HOA Renderer → Rendered HOA → Mixer → Binaural Renderer → Headphone feeds

Hoa → HOA Renderer → Rendered HOA → Mixer → Binaural Renderer → Headphone feeds

Converter

Channels

Objects

HOA

Loudspeaker layout

MPEG-H bitstream

[Herre, 2013]
Channel-based audio

- Audio sources are mixed for target setup/channels, like stereo, 5.1, 7.1, 9.1, 22.2, etc.
- Channels are stored/transmitted
- Channels are reproduced by target setup
- Pros: Legacy content (music/movies), direct playback
- Cons: not flexible to playback system mismatch, sub-optimal performance
Object-based audio

- Audio object = audio source + metadata
- Audio object is stored/transmitted
- Audio object is rendered into mix by receiver to actual setup at playback time
- Agnostic to playback configuration, compromise-free object rendering
- Personalization
- Industrial support: MPEG-H, Dolby ATMOS, DTS:X, Auro-3D
Scene-based audio: ambisonics basics

- Assume a sound field = superposition of **plane waves**
  - Recording/Encoding: sound sources/objects
  - Reproduction/decoding: loudspeakers (to find the weights)

- Any spatial function (e.g., plane wave) on the unit-sphere
  = infinite sum of spherical harmonics (SH)
  ≈ finite sum of SH with \( N \) orders

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

- **f**: function
- **\( \theta \)**: elevation
- **\( \phi \)**: azimuth
- **n**: order
- **m**: degree
- **\( Y_{n}^{m} \)**: spherical harmonics
- **f_{nm}**: weight

[Rafaely, 2015]
Spherical harmonics

\[ Y^m_n(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P^m_n(\cos \theta) e^{im\phi} \]

0\(^{th}\) order \( n = 0 \)

\[ Y^0_0(\theta, \phi) = \sqrt{\frac{1}{4\pi}} \]

1\(^{st}\) order \( n = 1 \)

\[ Y^{-1}_1(\theta, \phi) = \sqrt{\frac{3}{8\pi}} \sin \theta e^{-i\phi} \quad Y^0_1(\theta, \phi) = \sqrt{\frac{3}{4\pi}} \cos \theta \quad Y^1_1(\theta, \phi) = -\sqrt{\frac{3}{8\pi}} \sin \theta e^{i\phi} \]

2\(^{nd}\) order \( n = 2 \)

\[ Y^{-2}_2(\theta, \phi) = \sqrt{\frac{15}{32\pi}} \sin^2 \theta e^{-2i\phi} \quad Y^{-1}_2(\theta, \phi) = \sqrt{\frac{15}{8\pi}} \sin \theta \cos \theta e^{-i\phi} \quad Y^0_2(\theta, \phi) = \sqrt{\frac{5}{16\pi}} (3 \cos^2 \theta - 1) \quad Y^1_2(\theta, \phi) = -\sqrt{\frac{15}{32\pi}} \sin \theta \cos \theta e^{2i\phi} \quad Y^2_2(\theta, \phi) = \sqrt{\frac{15}{8\pi}} \sin^2 \theta e^{2i\phi} \]
Spherical harmonic weights

- Spherical harmonic weights
  \[ f_{nm} = \int_0^{2\pi} \int_0^\pi f(\theta, \phi) [Y_n^m(\theta, \phi)]^* \sin \theta d\theta d\phi \]
  → what the ambisonics microphone records directly or indirectly

- Rotation in spherical harmonic domain
  \[ g_{nm} = Rf_{nm} \]
  After rotation
  \[ [g_{0,0}, g_{1,-1}, g_{1,0}, g_{1,1}, \ldots, g_{N,N}] \]
  Before rotation
  \[ [f_{0,0}, f_{1,-1}, f_{1,0}, f_{1,1}, \ldots, f_{N,N}] \]
  \((N+1)^2 \times (N+1)^2\) block diagonal matrix related to the rotation angles
  → Sound field does not need to be recorded again!
Ambisonic decoding/reproduction

- The above **physical decoding** technique
  - assumes coherent sum of loudspeaker signals and reproduce original velocity
  - works well only for **low** frequency with a **small** sweet spot

- Other techniques include **psychoacoustic decoding**
  - assumes incoherent sum of the loudspeaker signals and reproduces the original energy
  - Works better at **higher** frequency

\[
Cs(t) = A(t) \rightarrow s(t) = \left( C^T C \right)^{-1} C^T A(t) = \frac{1}{J} C^T A(t)
\]

Encoding spherical harmonic matrix related to loudspeaker positions

[Arteaga, 2015]
Ambisonics: B-format

- **Recording/encoding**
  - omnidirectional \( W = \frac{1}{K} \sum_{k=1}^{K} s_k \left[ \frac{1}{\sqrt{2}} \right] \)
  - x-directional \( X = \frac{1}{K} \sum_{k=1}^{K} s_k \left[ \cos \phi_k \cos \theta_k \right] \)
  - y-directional \( Y = \frac{1}{K} \sum_{k=1}^{K} s_k \left[ \sin \phi_k \cos \theta_k \right] \)
  - z-directional \( Z = \frac{1}{K} \sum_{k=1}^{K} s_k \left[ \sin \theta_k \right] \)

- **Reproduction/decoding to regular layout**
  - Loudspeaker signal \( p_j = \frac{1}{J} \left[ W \ X \ Y \ Z \right] \left[ \begin{array}{c} \frac{1}{\sqrt{2}} \\ \cos \phi_j \cos \theta_j \\ \sin \phi_j \cos \theta_j \\ \sin \theta_j \end{array} \right] \)

- **Rotation (e.g., azimuth rotation by \( \theta \))**

\[
\begin{bmatrix}
W' \\
X' \\
Y' \\
Z'
\end{bmatrix} = R \begin{bmatrix}
W \\
X \\
Y \\
Z
\end{bmatrix}
\]

\[
R = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & \cos \theta & -\sin \theta & 0 \\
0 & \sin \theta & \cos \theta & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]

\[
R|_{\theta=90^\circ} = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\]
### An overview and comparison

<table>
<thead>
<tr>
<th>Audio content format</th>
<th>Channel-based</th>
<th>Object-based</th>
<th>Scene-based</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advantages</strong></td>
<td>Easy to set up; no processing for the matched playback configurations</td>
<td>Flexible for arbitrary playback configuration; accurate sound image; enable interactivity</td>
<td>Flexible for arbitrary playback configuration; full 3D sound image</td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td>Difficult to fit in different playback configurations; 3D sound image limited</td>
<td>High transmission or storage; high computation complexity</td>
<td>Require a large number of speakers placed on the surface of a sphere</td>
</tr>
<tr>
<td><strong>Status</strong></td>
<td><strong>Legacy audio format, still dominant</strong></td>
<td><strong>Emerging audio format used in movies/games</strong></td>
<td><strong>Adopted in VR/AR/MR</strong></td>
</tr>
<tr>
<td><strong>Desired reproduction system</strong></td>
<td>Stereo and multichannel surround sound system</td>
<td>Amplitude panning, WFS, binaural, transaural rendering</td>
<td>Ambisonics</td>
</tr>
</tbody>
</table>
Parametric spatial audio processing (PSAP)

Characteristics:
- Flexible
- Effective
- Efficient

[Chavez, 2018]
<table>
<thead>
<tr>
<th>Audio content type</th>
<th>Directional sound</th>
<th>Diffuse sound</th>
<th>Parameters</th>
<th>Related techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel-based</td>
<td>One/multiple</td>
<td>Yes</td>
<td>ICTD, ICLD</td>
<td>Spatial audio scene coding (SASC) / Primary-ambient extraction (PAE)</td>
</tr>
<tr>
<td></td>
<td>Multiple</td>
<td>No</td>
<td>Azimuth, elevation</td>
<td>Blind source separation (BSS)</td>
</tr>
<tr>
<td></td>
<td>Multiple</td>
<td>No</td>
<td>ICTD, ICLD, ICC</td>
<td>MPEG Spatial audio Coding (SAC)</td>
</tr>
<tr>
<td>Object-based</td>
<td>Multiple</td>
<td>No</td>
<td>Azimuth, elevation, distance</td>
<td>MPEG Spatial audio object coding (SAOC)</td>
</tr>
<tr>
<td>Scene-based (ambisonics)</td>
<td>Multiple</td>
<td>Yes</td>
<td>Azimuth, elevation, Diffuseness</td>
<td>Directional audio coding (DirAC)</td>
</tr>
<tr>
<td>Scene-based (mic array)</td>
<td>Multiple</td>
<td>No</td>
<td>Azimuth, elevation</td>
<td>BSS</td>
</tr>
<tr>
<td></td>
<td>Multiple</td>
<td>Yes</td>
<td>Azimuth, elevation</td>
<td>Spatial filtering</td>
</tr>
</tbody>
</table>
PSAP example: Directional audio coding (DirAC)

[Link to Pulkki, 2007]
PSAP example: sound scene decomposition

**Aim:** to obtain useful information about the original sound scene from given mixtures, and facilitate natural sound rendering.

blind source separation

"Sum of sources"

Primary ambient extraction

"Sum of primary and ambient components"


[Sunder, 2015]
Sound scene decomposition: BSS

Objective: to extract the K sources from M mixtures

Mixtures = function (gain, source, time difference, model error)

\[ x_m(n) = \sum_{k=1}^{K} g_{mk} s_k(n - \tau_{mk}) + e_m(n), \quad \forall m \in \{1, 2, \ldots, M\} \]
### Sound scene decomposition: BSS

**Objective:** to extract the $K$ sources from $M$ mixtures

<table>
<thead>
<tr>
<th>Case</th>
<th>Typical techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M = K$</td>
<td>ICA</td>
</tr>
<tr>
<td>$M &gt; K$</td>
<td>ICA with PCA, Least-squares</td>
</tr>
<tr>
<td>$M &lt; K$</td>
<td></td>
</tr>
<tr>
<td>$M &gt; 2$</td>
<td>ICA with sparse solutions</td>
</tr>
<tr>
<td>$M = 2$</td>
<td>Time-frequency masking</td>
</tr>
<tr>
<td>$M = 1$</td>
<td>NMF, CASA</td>
</tr>
</tbody>
</table>

**Examples:**

- ICA : Independent component analysis
- PCA : Principal component analysis
- NMF : Non-negative matrix factorization;
- CASA: Computational auditory scene analysis

"Sum of sources"
Sound scene decomposition: PAE

Objective: to extract the primary and ambient components from M mixtures

Mixtures = primary component + ambient component

\[ x_m(n) = p_m(n) + a_m(n) \]
Sound scene decomposition: PAE

**Objective:**
To extract the primary and ambient components from $M$ ($M = 2$, stereo) mixtures

<table>
<thead>
<tr>
<th>Case</th>
<th>Typical techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic model</td>
<td>Time frequency masking</td>
</tr>
<tr>
<td>Channel-wise</td>
<td></td>
</tr>
<tr>
<td>Combine channels</td>
<td>Linear estimation (PCA, LS), Ambient spectrum estimation</td>
</tr>
<tr>
<td>Complex model</td>
<td>Time/phase shifting, Classification, Sub-band, Pairing up two channels, etc.</td>
</tr>
</tbody>
</table>
PAE: Linear estimation for stereo signals

Objective and relationships of four linear estimation based PAE approaches.

- **Blue** solid lines represent the relationships in the primary component;
- **Green** dotted lines represent the relationships in the ambient component.
- **MLLS**: minimum leakage LS
- **MDLS**: minimum distortion LS

\[
\begin{bmatrix}
\hat{p}_0(n) \\
\hat{p}_1(n) \\
\hat{a}_0(n) \\
\hat{a}_1(n)
\end{bmatrix} =
\begin{bmatrix}
w_{P0,0} & w_{P0,1} \\
w_{P1,0} & w_{P1,1} \\
w_{A0,0} & w_{A0,1} \\
w_{A1,0} & w_{A1,1}
\end{bmatrix}
\begin{bmatrix}
x_0(n) \\
x_1(n)
\end{bmatrix}
\]

[He, 2014]
PAE: an example from least-squares
VR/AR/MR audio aims to deliver an interactive immersive listening experience in a virtual/augmented world.

<table>
<thead>
<tr>
<th>Use case</th>
<th>Cinematic / 360 (Video/Streaming)</th>
<th>Synthetic / Full (Game/App)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual Position</td>
<td>Static</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Real position</td>
<td>Static</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Tracking</td>
<td>Head orientation</td>
<td>Head / + body</td>
</tr>
<tr>
<td>Source directions</td>
<td>Dynamic</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Source distances</td>
<td>Static</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Reverberation</td>
<td>Static</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Diffraction</td>
<td>Static</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Doppler effect</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Deliver</td>
<td>Coded content</td>
<td>Coded content + rendering engine</td>
</tr>
<tr>
<td>Common format</td>
<td>Scene (Ambisonics)</td>
<td>Object (better performance), Scene</td>
</tr>
<tr>
<td>Real sound</td>
<td>Presented naturally in AR/MR</td>
<td></td>
</tr>
</tbody>
</table>
### An illustration

<table>
<thead>
<tr>
<th>Natural</th>
<th>Cinematic / 360</th>
<th>Synthetic / Full</th>
</tr>
</thead>
</table>

#### Key Technology

<table>
<thead>
<tr>
<th>Direction rendering</th>
<th>Ambisonics</th>
<th>Object-based audio</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRTF</td>
<td></td>
<td>HRTF</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distance rendering</th>
<th>Amplitude adjustment</th>
<th>3D modeling, Amplitude adjustment</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Reverberation</th>
<th>Fixed</th>
<th>3D modeling, Early reflection modeling</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Ambisonic rotation</th>
<th>3D modeling, Low pass filtering</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Performance &amp; Complexity</th>
<th>Medium</th>
<th>High (no. of sources)</th>
</tr>
</thead>
</table>

http://superpowered.com/3d-spatialized-audio-virtual-reality
Other VR/AR/MR audio effects

- Area sources (source with width)
- Source with directivity
- Sound transport time
- Non-spatialized audio
- Audio effects

Binaural rendering is the soul of VR/AR/MR audio

Binaural rendering recreates all the listening cues for both ears using headphones

- Direction rendering
- Distance rendering
- Environment rendering
- Interaction

[Image courtesy: GAUDIO, 2016]
## Challenges and solutions

<table>
<thead>
<tr>
<th>Listening</th>
<th>Headphone listening</th>
<th>Signal processing</th>
<th>Natural listening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td><img src="image" alt="Virtualization" /></td>
<td>Virtualization of source &amp; environment</td>
<td><img src="image" alt="Natural listening" /></td>
</tr>
<tr>
<td>Medium</td>
<td><img src="image" alt="Equalization" /></td>
<td>Equalization</td>
<td><img src="image" alt="Natural listening" /></td>
</tr>
<tr>
<td>Receiver</td>
<td><img src="image" alt="Individualization" /></td>
<td>Individualization</td>
<td><img src="image" alt="Natural listening" /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Tracking" /></td>
<td>Tracking</td>
<td><img src="image" alt="Natural listening" /></td>
</tr>
</tbody>
</table>

[Begault, 2000]
Binaural rendering for 3 types of formats

- Channel based
- Object based
- Scene based
  - Ambisonics
  - Other microphone array recording
  - Binaural recording: not suitable for VR/AR/MR
Binaural rendering of channel-based audio

Original scene

Perceived scene
Binaural rendering of object-based audio

Original scene

Perceived scene
Binaural rendering of scene-based audio

Original scene

Perceived scene
Google Omnitone

**Spatial media**

VR Headset or Smartphone

Orientation sensor data

**Rotator**

8 Virtual Speakers

Rotated ambisonic streams for each virtual speaker

**Stereo Out**

2-channel audio

Headphones

4-channel audio (AmbiX format)
Using **generic HRTFs with head tracking**, performance might differ with individualized HRTFs.

- Significant improvement found by increasing from $1^{st}$ order to $3^{rd}$ order.
- Little advantage found in $5^{th}$ order over $3^{rd}$ order.

---

<table>
<thead>
<tr>
<th>Ambisonic order</th>
<th>Average localization error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1^{st}$</td>
<td>24°</td>
</tr>
<tr>
<td>$3^{rd}$</td>
<td>17°</td>
</tr>
<tr>
<td>$5^{th}$</td>
<td>15°</td>
</tr>
</tbody>
</table>

[Thresh, 2017]
Audio format transformation

Channel

Recording Panning

Decoding

Encoding

Object

Scene

Recording Encoding
B.3 HRTF Individualization

Variation of HRTFs (Idiosyncratic)

[Xu, 2007; Carlile, 2014]
Highly individualized ear response

[Hoffman, 1998]
Overview of HRTF individualization techniques

To obtain individualized HRTF/perception

- **Acoustical measurements**
  - Stop-and-go *static* measurements
  - Fast *dynamic* measurements

- **Anthropometric measurements**
  - Numeric simulation based on 3D *models*
  - Data-driven approaches based on *features*

- **Listening and evaluation**
  - Tuning HRTF set based on perception
  - Training to adapt to new HRTFs

- **Multi-driver headphone sound projection**
## HRTF measurement techniques with human subjects

<table>
<thead>
<tr>
<th>Representative references</th>
<th>Microphone type</th>
<th>Number of loudspeaker movement</th>
<th>Loudspeaker movement</th>
<th>Subject posture</th>
<th>Subject (head) movement</th>
<th>Subject tracking</th>
<th>Excitation signal</th>
<th>Performance evaluation</th>
<th>Approximate duration*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Møller, 1995 Algazi, 2001</td>
<td>Binaural</td>
<td>1-N</td>
<td>Discrete positions across azimuth and elevation</td>
<td>Sit on a normal chair</td>
<td>Not allowed</td>
<td>No</td>
<td>Sweep or maximum length sequence (MLS)</td>
<td>Reference technique</td>
<td>1+ hours</td>
</tr>
<tr>
<td>Carpentier, 2014</td>
<td>Binaural</td>
<td>1</td>
<td>Discrete positions across elevation</td>
<td>Sit on a chair on the turntable</td>
<td>Not allowed</td>
<td>Yes</td>
<td>Sweep</td>
<td>No</td>
<td>1+ hours</td>
</tr>
<tr>
<td>Majdak, 2007</td>
<td>Binaural</td>
<td>22</td>
<td>No</td>
<td>Sit on a chair on the turntable</td>
<td>Not allowed</td>
<td>Yes</td>
<td>Multiple exponential sweep method (MESM)</td>
<td>Objective</td>
<td>30 minutes</td>
</tr>
<tr>
<td>Bilinski, 2014</td>
<td>Binaural</td>
<td>16</td>
<td>Discrete positions across elevation</td>
<td>Sit on a normal chair</td>
<td>Not allowed and fixed mechanically</td>
<td>Yes</td>
<td>MESM</td>
<td>No</td>
<td>30 minutes</td>
</tr>
<tr>
<td>Bomhardt, 2017</td>
<td>Binaural</td>
<td>64</td>
<td>No</td>
<td>Stand on a turntable</td>
<td>Not allowed and fixed mechanically</td>
<td>No</td>
<td>MESM</td>
<td>No</td>
<td>10 minutes</td>
</tr>
<tr>
<td>Pollow, 2012</td>
<td>Binaural</td>
<td>40</td>
<td>No</td>
<td>Stand on a turntable</td>
<td>Not allowed and fixed mechanically</td>
<td>No</td>
<td>MESM</td>
<td>Subjective</td>
<td>10 minutes</td>
</tr>
<tr>
<td>Zotkin, 2006</td>
<td>Binaural</td>
<td>1</td>
<td>Across the two ears</td>
<td>Sit on a normal chair</td>
<td>Not allowed</td>
<td>No</td>
<td>Sweep</td>
<td>Objective</td>
<td>30 minutes</td>
</tr>
<tr>
<td>Fukudome, 2007</td>
<td>Binaural</td>
<td>1</td>
<td>Move/rotate vertically</td>
<td>Sit on a chair on the turntable</td>
<td>Not allowed</td>
<td>No</td>
<td>MLS</td>
<td>Objective</td>
<td>1+ hours</td>
</tr>
<tr>
<td>Pulkki, 2010</td>
<td>Binaural</td>
<td>1</td>
<td>Discrete positions across elevation, and continuous rotation across azimuth</td>
<td>Sit on a normal chair</td>
<td>Not allowed</td>
<td>No</td>
<td>Sweep</td>
<td>Objective</td>
<td>1+ hours</td>
</tr>
<tr>
<td>Enzner, 2008</td>
<td>Binaural</td>
<td>1</td>
<td>Discrete positions across elevation</td>
<td>Sit on a chair on the turntable</td>
<td>Not allowed</td>
<td>No</td>
<td>White noise or perfect sweep</td>
<td>Objective</td>
<td>30 minutes</td>
</tr>
<tr>
<td>Enzner, 2009</td>
<td>Binaural</td>
<td>4</td>
<td>Few discrete positions across azimuth/ elevation</td>
<td>Sit on a chair on the turntable</td>
<td>Not allowed</td>
<td>No</td>
<td>White noise</td>
<td>Objective</td>
<td>10 minutes</td>
</tr>
<tr>
<td>He, 2016 Li, 2017</td>
<td>Binaural</td>
<td>1</td>
<td>Few discrete positions across azimuth/ elevation</td>
<td>Sit on a rotatable chair</td>
<td>Free movement across azimuth/ elevation</td>
<td>Yes</td>
<td>White noise, or perfect sweep</td>
<td>Objective</td>
<td>30 minutes</td>
</tr>
<tr>
<td>Reijniers, 2017</td>
<td>Binaural</td>
<td>1</td>
<td>Few discrete positions across azimuth</td>
<td>Sit on a normal chair</td>
<td>Free movement across azimuth/ elevation</td>
<td>Yes</td>
<td>Sweep</td>
<td>Objective and subjective</td>
<td>30 minutes</td>
</tr>
<tr>
<td>He, 2018</td>
<td>Binaural</td>
<td>1</td>
<td>Few discrete positions across azimuth/ elevation</td>
<td>Sit on a rotatable chair</td>
<td>Free movement across azimuth/ elevation</td>
<td>Yes</td>
<td>White noise</td>
<td>Objective and subjective</td>
<td>30 minutes</td>
</tr>
</tbody>
</table>
Existing HRTF acquisition techniques

- **Discrete stop-and-go HRTF acquisition**
  - Fixed measurement setups (Multiple loudspeakers play one-by-one)
  - Tedious and time consuming especially for human subjects.

![Fixed measurement setups](image1)

- *Tohoku univ. Japan*
- *TU Berlin*
- *Technical University of Lodz*
Individualization: acoustical measurements

Air Force Research Laboratory, US

Nagaoka University of Technology, Japan

ISVR, University of Southampton, UK

South China University of Technology, China

Tohoku University, Japan
Example: Smyth Realiser

Key features:
- Virtualization of 16 loudspeakers in rooms
- Head-tracking in 2D
- Individualized HRTFs via measurement
- Individualized headphone equalization
- Up-mixer
- Bit-stream decoding: Dolby, DTS, Auro-3D formats
<table>
<thead>
<tr>
<th>Databases</th>
<th>(Subjects, Directions)</th>
<th>Measuring Conditions and Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRCAM France</td>
<td>(51, 187)</td>
<td>Far field: 1.95m Source: Log sine sweep Blocked ear canal</td>
</tr>
<tr>
<td>CIPIC, UC Davis</td>
<td>(45,1250)</td>
<td>Far field: 1m Source: Golay code Blocked ear canal</td>
</tr>
<tr>
<td>Tohoku University, Japan</td>
<td>(3,454)</td>
<td>Far field: 1.2m Source: Time stretched pulse Blocked ear canal</td>
</tr>
<tr>
<td>Nagoya University, Japan</td>
<td>(96,72)</td>
<td>Far field: 1.52m Source: Time stretched pulse Not entirely block</td>
</tr>
<tr>
<td>Austrian Academy of Sciences</td>
<td>(70,1550)</td>
<td>Far field: 1.2m Source: exponential sweep signal Blocked ear canal</td>
</tr>
<tr>
<td>TU Berlin</td>
<td>(FABIAN,11950)</td>
<td>Far field: 1.7m Source: Sine sweep Blocked Ear canal</td>
</tr>
<tr>
<td>MIT Lab</td>
<td>(KEMAR,710)</td>
<td>Far field: 1.4m Source: MLS Ear simulator</td>
</tr>
<tr>
<td>Oldenburg University</td>
<td>(HATS,365)</td>
<td>Far field: 0.8 – 3m Source: MIRS In-the ear and behind the ear</td>
</tr>
<tr>
<td>SDAC, KAIST (0.2,0.6,1m)</td>
<td>(HATS, 100)</td>
<td>Far field: 1m Source: White noise</td>
</tr>
<tr>
<td>RIEC University</td>
<td>(105,865)</td>
<td>Far field: 1.5 m Source: Time stretched pulse Blocked ear canal</td>
</tr>
<tr>
<td>Xie (Chinese Human subject database) (1.5m)</td>
<td>(52,493)</td>
<td>Far Field: 1.5 m Source: MLS Blocked ear canal</td>
</tr>
<tr>
<td>DSP Lab @ NTU (0.35,0.45,0.50,0.60,0.75,0.8,1,1.4m)</td>
<td>(HATS + 3 subjects, 600)</td>
<td>Far field: 1.1.4m &amp; Near field: 0.35m-0.8m Source: MLS Block ear canal/Ear simulator</td>
</tr>
</tbody>
</table>
Reciprocal HRTF measurement technique

- Placing micro-speakers inside ear canal
- Spherical microphone array surround subject
- Measure HRTF at frequency >1.5kHz
- Use HRTF of HATS at low frequency.

[Zotkin, 2006]
Existing HRTF acquisition techniques

- **Fast Continuous HRTF acquisition**
  - Moving loudspeaker or subject using continuous excitation method
  - Require rotating facility (e.g., turntable) for constant speed movement

![Diagram showing continuous movement of loudspeaker and subject](Pulkki, 2010)

[![Diagram showing turntable](Enzner, 2008)]
Fast HRTF measurement system

- Head tracking allows free movements in azimuth / elevation
- A fixed loudspeaker continuously emitting broadband signal
- Binaural recording at listener’s ears and synchronized with directional movements
- Visual display feedbacks the movement pattern

[He, 2015; Ranjan, 2016]
Adaptive filter for dynamic HRIR estimation

- Dynamically varying HRIRs
- Corresponding directions known
- HRIRs at neighboring directions show similarity and continuity
Signal model for Fast HRTF measurement system

\[ y(n) = h^T[\theta(n), \varphi(n)] x(n) + v(n) \]

Discretization of the continuous directions

\[ y(n) = H^T[d(n)]x(n) + v(n) \]

\[ H_{K \times M} = \begin{bmatrix} h(\theta_1, \varphi_1) & \ldots & h(\theta_1, \varphi_M) \\ \vdots & \ddots & \vdots \\ h(\theta_K, \varphi_1) & \ldots & h(\theta_K, \varphi_M) \end{bmatrix} \]

\( K \times M \) discrete HRIRs to be estimated
Adaptive filter for dynamic HRIR estimation

Progressive based NLMS
\[ \hat{h}_{n+1}(35°) = \hat{h}_n(35°) + \mu \frac{x(n)}{\|x(n)\|^2} e(n) \]
Progressive based NLMS

\[ \hat{h}_{n+1}(30°) = \hat{h}_n(35°) + \mu \frac{x(n)}{||x(n)||_2^2} e(n) \]
Adaptive filter for dynamic HRIR estimation

Progressive based NLMS

\[
\hat{h}_{n+1}(30^\circ, 2) = \hat{h}_n(25^\circ) + \mu \frac{x(n)}{||x(n)||^2_2} e(n)
\]
Adaptive filter for dynamic HRTIR estimation

\[ \hat{h}_{n+1}(30^\circ, 2) = \hat{h}_n(30^\circ, 2) + \mu \frac{x(n)}{||x(n)||^2_2} e(n) \]

Activation based NLMS

![Graphs showing excitation signal, recorded binaural signal, and relative source direction over time.](image-url)
(1) Static vs dynamic measurement

Static Measurement

Dynamic Measurement
Static vs dynamic acquisition: Spectrogram

- VR Static, Left (Ipsilateral)
- VR Static, Right (Contra lateral)

- VR Dynamic, Left (Ipsilateral)
- VR Dynamic, Right (Contra lateral)
Static vs dynamic acquisition: HRIR/HRTF

- **0°;0°**
  - HRIR left Azimuth: 0° Elev: 0°
  - HRIR right Azimuth: 0° Elev: 0°
  - HRTF left Azimuth: 0° Elev: 0°
  - HRTF right Azimuth: 0° Elev: 0°

- **-30°;0°**
  - HRIR left Azimuth: -30° Elev: 0°
  - HRIR right Azimuth: -30° Elev: 0°
  - HRTF left Azimuth: -30° Elev: 0°
  - HRTF right Azimuth: -30° Elev: 0°

- **-90°;0°**
  - HRIR left Azimuth: -90° Elev: 0°
  - HRIR right Azimuth: -90° Elev: 0°
  - HRTF left Azimuth: -90° Elev: 0°
  - HRTF right Azimuth: -90° Elev: 0°

- **-60°;0°**
  - HRIR left Azimuth: -60° Elev: 0°
  - HRIR right Azimuth: -60° Elev: 0°
  - HRTF left Azimuth: -60° Elev: 0°
  - HRTF right Azimuth: -60° Elev: 0°
Static vs dynamic acquisition: ITD and ILD

ITD Elevation = 30

ITD Elevation = 0

ITD Elevation = -30

ILD Elevation = 30

ILD Elevation = 0

ILD Elevation = -30

Azimuth (degree)

ITD (Samples)

ILD (dB)
Static vs dynamic acquisition: Identification accuracy

- Over 90% accuracy in identifying difference from human and dummy head.
- Only 50-75% accuracy in identifying static HRIR and dynamic HRIR.
- Most subjects reported that they need to focus to make the selection on static vs dynamic HRTF.

![Graph showing accuracy for different tracks](image)

- Tracks: White noise, Music, Speech, Bird
- Accuracy: Y-axis
- Guessing level: Dashed line at 0.5
(2) Head tracker vs. VR Oculus Rift

HRTF Acquisition using head-tracker

HRTF Acquisition using Oculus Rift
Head Tracker vs VR: HRIR/HRTF

0az;0el

-30az;0el

-90az;0el

-60az;0el
(3) Head Above Torso Orientation (HATO)

Head Above Torso Orientation Aligned

Head Above Torso Orientation Not-aligned (Torso Fixed)
Natural and augmented listening for VR, AR/MR: Module B

HATO Aligned vs Not-aligned: HRIR/HRTF

- 0az;0el
- 30az;0el
- 90az;0el
- 60az;0el
More difference exists in
• middle frequency
• contralateral ear (280 to 360 degree)

Easily distinguishable with broadband signals, but less difference with speech signals

Subjects mentioned coloration and localization difference between 2 types of HRTFs

[Brinkmann, 2015]
Key observations on fast HRTF measurement system

- Allows fast HRTF measurement with unconstrained movements

- For static vs dynamic measurement:
  - Good match in spectrum, ITD, and ILD
  - Perceptual difference: low identification accuracy

- Differences in VR/AR gear must be compensated

- HATO aligned and not-aligned are more obvious in contralateral HRIRs and in middle frequency.
Similar system from Leibniz University Hannover

**Figure 6:** NMSE results for 1D head movements in azimuth

**Figure 7:** NMSE results for 1D head movements in elevation

[Li, 2017]
Earfish’s fast HRTF measurement system

- A single fixed loudspeaker continuously emitting an excitation signal
- Response recorded at listener’s ears and synchronized with dynamic head movements
- Use sinesweep, but algorithms unknown

https://www.earfish.eu/

[Reijniers, 2017]
HRTF interpolation

Directional resolution: measurement < rendering

Methods (domain):

- **Directional**  
  \[
  \hat{H}_x = \sum_{i=1}^{I} w_i H_i
  \]
  - minimum phase, magnitude/phase
  - Use nearest 2, 3, 4 directions
  - Better performance with denser input, lower frequency, ipsilateral ear

- **Spectral basis** (e.g., PCA)

- **Spatial basis** (e.g., spherical harmonics)

[Christensen, 1999]  
[Xie, 2013; Gamper, 2013]  
[Nam, 2008; Jot, 1995]  
[Martens, 1987]  
[Evans, 1998]
Individualization: anthropometric measurements

Numeric simulation
1. 3D head and ear model construction via 3D scan/video/images
2. Solving of acoustic equation
3. Numerical methods: FEM, BEM, FDTD
4. How to make it more efficient
HRTF database with 3D models

ITA HRTF-database
48 subjects, head and ear models (FMRI)
http://www.akustik.rwth-aachen.de/go/id/lsly

[Bomhardt, 2016]

Princeton University 3D3A Lab
(using 3D camera sensor and blue light)
http://www.princeton.edu/3D3A/HRTFMeasurements.ht

[Sridhar, 2017]
## HRTF databases with anthropometry

<table>
<thead>
<tr>
<th>HRTF database</th>
<th>Year</th>
<th># subjects</th>
<th>Region</th>
<th># Directions</th>
<th># anthropometry features</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIPIC</td>
<td>2001</td>
<td>40+</td>
<td>Western</td>
<td>1250</td>
<td>37</td>
</tr>
<tr>
<td>Nishino et al</td>
<td>2007</td>
<td>86</td>
<td>Japanese</td>
<td>72</td>
<td>9</td>
</tr>
<tr>
<td>Xie et al</td>
<td>2007</td>
<td>52</td>
<td>Chinese</td>
<td>72</td>
<td>17</td>
</tr>
<tr>
<td>TUM LDV</td>
<td>2013</td>
<td>35</td>
<td>Western</td>
<td>2160</td>
<td>8</td>
</tr>
<tr>
<td>Microsoft Research</td>
<td>2014</td>
<td>250+</td>
<td>Global</td>
<td>512</td>
<td>45+</td>
</tr>
</tbody>
</table>

Head width: strong correlation with ITD

Significant anthropometric parameters
- Distance between ear and shoulder, breadth of head and back vertex; breadth and depth of cavum conchae and rotation of ears
- Head depth, pinna offset back, cavum concha, width, fossa height, pinna height, pinna width, pinna rotation angle and pinna flare angle

---

[Middlebrooks, 1999; Xie, 2007]

[Fels, 2004]

[Zhang, 2011]
Techniques using anthropometric features

- A diverse database of HRTFs and Anthropometric (A) features
- Apply relation among A features in HRTFs
  - Select HRTF set based on the closest A features
  - Linear sparse representation of A features
  - [Zotkin, 2003]
  - [Blinski, 2014; He, 2015]
- Train relation between A features and HRTF
  - Transform HRTF into a different domain using, e.g., PCA, SVD, Least-squares, spherical harmonics, NMF, etc.
  - Select anthropometric features
  - Training using multiple linear regression, ANN, DNN, SVM, etc.
  - Direct relation via frequency scaling, resonant frequency
  - [Zotkin, 2003; Zhou, 2008; Li, 2013; Fayek, 2017]
## Performance on machine learning methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean spectral distortion SD (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA + NN [Zhou, 2008]</td>
<td>&lt;3</td>
</tr>
<tr>
<td>SVD + RBF NN [Li, 2013]</td>
<td>~3</td>
</tr>
<tr>
<td>Isomap + NN [Grijalva, 2016]</td>
<td>4.6</td>
</tr>
<tr>
<td>NN [Favek, 2017]</td>
<td>3</td>
</tr>
</tbody>
</table>

Note: Mean SD was computed differently.

1. **Why NN still not good?** Maybe still lack a big and diverse HRTF and anthropometry dataset
2. **Standard method to obtain anthropometric features?**
3. **How about perceptual performance?** SD is not a good criteria and HRTF can be simplified to remove perceptually irrelevant details.
Commercial examples

- Microsoft Windows 10 & Hololens
- IDA audio
- 3D Sound labs
- Creative Labs Super X-Fi
Individualization: training/tuning via listening

Razer

Mendonca et al

DTS

Vivo (DTS)
Individualization: headphone projection

- No additional measurements and listening experiments required
- Reduce front-back confusion by > 50%;
- Zero user effort, plug and play (automatic during playback)

[Sunder, 2013; Sunder, 2015]
Example: OSSIC X Multi-driver Headphone

**OSSIC X:** The first 3D audio headphones calibrated to you

$2,708,472

pledged of $100,000 goal

---

**HRTF ANATOMY CALIBRATION**

OSSIC X instantly calibrates to your head and torso calibration, without any lab needed. This enables incredibly accurate sound placement for higher level of sound quality and immersion.

**INTEGRATED HEAD TRACKING**

By incorporating head tracking into the OSSIC X, sounds will appear to come from outside your head and stay fixed in space, enabling a higher sense of acoustic presence.

**MULTIDRIVER ARRAY**

Eight individual drivers work in tandem to play back sound to the correct portion of your ear. This allows your unique ear shape to naturally interact with the 3D sound field the same way it does in real life.
## Comparison of HRTF individualization techniques

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Resources</th>
<th>User contribution</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acoustic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Static</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Dynamic</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td><strong>Anthropometric</strong></td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3D model</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Features</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Listening</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Tuning</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>Headphone Projection</strong></td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Non-individualized HRTFs</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The numbers are for illustration of our qualitative relative opinion purpose only:

- 5/4/3/2/1/0: Very High / High / Medium / Low / Very Low / No
- **Resources** include hardware, software, database, etc.
- **User contribution** includes user’s time and efforts
- The **actual performance** must be evaluated in psychoacoustic experiments
**B.4 Equalization**

**Headphone is not acoustically transparent:**
- Headphone colors the input sound spectrum;
- Affects the free-field characteristics of the sound pressure at the ear

![Diagram showing the breakdown of headphone transfer function (HPTF)]

Breakdown of headphone transfer function (HPTF)

[Møller, 1995]
Decoupled equalization for stereo

**Aim: Emulate the reproduction in a reference field**

- **Free-field:**
  - Target: free-field front loudspeaker response

- **Diffuse-field and other reference curves:**
  - Target: response of diffuse-field, or a reference room
  - Lesser inter-individual variability

[Olive, 2013]
Non-decoupled equalization for binaural

**Aim: Spectrum at eardrum is the individual HRTF features**

- Conventional headphone: removing HPTF
  - EQ = 1/HPTF
  - Dependent on individual pinna feature and repositioning

- Projection headphone: preserving individualized HPTF
  - EQ = 1/free-field HPTF
  - No headphone-ear coupling

- Inversion requires regularization

References:

B.5 Movement Tracking

Original scene

Head tracking

Position tracking
6 DOF movement tracking

- Track all 6 DOF ideally but could be simplified
- Positional tracking
  - Camera based, laser based techniques
  - Affect direction, distance, diffraction perception
  - Perceptual effects on localization accuracy and latency need more investigation

Head tracking

- Head movement information is tracked by a sensor (e.g., accelerometer, gyroscope, magnetometer, camera)
- Adapt to the changes of sound scene with respect to head movements
- Cross-fading is required to ensure smooth perception
- Scene update rate: 50ms or lower
- Concern of head tracking latency: <100ms (variation high)

Source from http://north-america.beyerdynamic.com

[Sandvad, 1996]
Head tracking improves source localization

- **Reduce front-back confusion (FBC)**, especially with non-individualized HRTF

  [Wenzel 1993a, 1995; Sandvad, 1996; Horbach, 1999; Wightman, 1999]

- **Improve externalization** for front and rear sources, especially using non-individualized HRTFs

  [Hendrickx, 2017]

- Reduce FBC from 50% to 28%, **more** than reverberation and individualized HRTF.

  [Begault, 2001]

- **Enhances the realism** of the virtual acoustic environment as a whole

  [Wenzel, 1991; Saviojia 1999]
Apply artificial reverberation to binaural rendering

- Externalization of the sound sources, and enhance depth perception;
- Rendering of the sound environment.

More on this during the 2\textsuperscript{nd} half of this tutorial

Source from [http://www.torgny.biz/Recording%20sound_2.htm](http://www.torgny.biz/Recording%20sound_2.htm)
B.7 Integration

[Sunder, 2015]
Authenticity of individual dynamic binaural synthesis

Compare real sound and virtual binaural synthesis with individualized HRTF/BRIR and HpTF measurement, allow head movements, using ABX test and Spatial Audio Quality Index (SAQI) test.

In terms of the **detections rates:**
- **Pink noise (all) > speech signals (half).**
- **Anechoic > reverberant, for speech.**
- **Coloration > localization.**
- **Dynamic > static.**

Suggesting for every audio content, if sufficient care taken for acquisition, postprocessing, and rendering, authentic binaural synthesis can be achieved.

[Brinkmann, 2017]
A grade higher in 4 measures: Sense of direction, externalization, ambience, and timbral quality; more preferred.

[Sunder, 2015]
Spatial Audio Technologies for Immersive VR, AR/MR

Spatial Audio Formats
- Object, Ambisonics
- Parametric processing

Environment Estimation
- Depth camera
- Reverberation fingerprint
- Machine learning

Individualized Binaural Rendering
- Individualized HRTFs
- Equalization

Environment Rendering
- Wave based
- Geometrical based
- Perceptual based

Dynamic Binaural Synthesis
- Head tracking
- Position tracking

Virtual & Physical Sound Fusion
- Adaptive equalization
- Hear-through processing
General references on binaural rendering

Paul M. Hoffman, “Relearning sound localization with new ears,” nature neuroscience, volume 1 no. 5, september 1998


Xie, B., 2013. Head-related transfer function and virtual auditory display. J. Ross Publishing
References on parametric spatial audio processing

- V. Pulkki, S. Delikaris-Manias, and A. Politis (Edited), “Parametric time-frequency domain spatial audio,” Wiley, 2018
References on parametric spatial audio processing

References on HRTF measurement methods

References on HRTF measurement methods

References on HRTF interpolation

References on anthropometry features for HRTF

References on anthropometry features for HRTF

References on listening based HRTF individualization

- Y. Iwaya, “Individualization of head-related transfer functions with tournament-style listening test: Listening with other’s ears,” Acoust. Sci. & Tech. 27, 6 (2006)
Key References on equalization

References on head tracking

Module C
Augmented/Mixed Reality Audio in Headsets

1. Types of Augmented/Mixed Reality Audio
2. Natural Listening in AR/MR: An overview
3. Signal Processing Techniques in Natural Augmented Listening
4. Hear Through of Real Sound
5. Virtual Sound Augmented with Real Sound
6. Acoustic Environment Estimation and Rendering
7. Conclusion
Augmented/Mixed reality is enhancing the way we experience the real world

Wearable AR/MR devices:

- Meta
- HoloLens
- Magic leap

AR/MR applications:

- Navigation
- VR and AR world
- Gaming
What is Augmented/Mixed Reality Audio

- A layer of augmented digital information
- Usually tagged with location based digital audio information playback
- Spatial audio superimposed with real sounds
- Interaction between real and augmented audio
C.2 Natural Listening in Augmented/Mixed Reality

Real

Virtual

Augmented Reality (AR)

Capture
C.3 Signal Processing Techniques in NAL

Real

Listener

Record, Process and playback

Local Acoustic environment estimation

Capture cues, process and playback

Virtual

Augmented Reality (AR)

Capture

Real and Virtual Listener

Local Acoustic environment

Capture cues
Natural Augmented Listening: 3 Major Components

- **Hear through of real sounds**

  Real Sound → **Hear Through** (Transparent Headphones) → Headphones

  **Headphones** might need to be equipped with **external microphone(s)** to record real sounds (to be equalized & playback)

- **Virtual sounds augmented with real sounds seamlessly**

  Virtual Source → **Acoustic environment estimation & Rendering** → **Binaural Rendering Over Headphones**

  Built in **sensors** to capture and estimate the **local acoustic environment** (for environment rendering) → Built in **internal microphone(s)** to capture **individual cues** for **binaural rendering**
Which headphones to choose for NAL?

Augmented content delivery methods varies based on design/choice of headsets

**Open Headphones**
- **Type I** – Personal speaker (No earcup)
  - Allows natural sound to pass through
  - Show best externalization
  - Privacy issue due to leakage
  - Poor bass for speakers

- **Type II** – Open-back over ear headset

**Closed Headphones**
- **Type III** – Closed In-ear headset
  - Blocks most of the natural sounds
  - Introduces occlusion effect
  - Transparent hearing using electrical hear through

- **Type IV** – Closed-back over ear headset
Headset Modes of Operation

(A) Hear Through mode
- Only real sound source present

(B) Virtual Reality mode
- Only virtual sound source present

(C) Augmented Reality mode
- Both real and virtual sound source present with natural fusion of two

(D) Enhanced/Mixed Reality mode
- Both real and virtual sound source present with selective control of the real sound
- Only applicable for closed headset design
C.4 Hear Through of Real Sound

- AR headsets should allow the **direct sounds** coming from physical sources for **acoustical transparency**
  - Open headphones allow most of the natural sounds to pass through unattenuated†
  - Closed headphones block most of the natural sounds

- Headphones Isolation obtained by measuring the speaker response at subject’s ears with headphones, $H_{\text{with \, hp}}(f)$ and without headphones, $H_{\text{ref}}(f)$:

$$\text{Attenuation\,[dB]}\, , A(f) = 20\log_{10} \left| \frac{H_{\text{with \, hp}}(f)}{H_{\text{ref}}(f)} \right|$$

† Open-back headphones attenuates poorly in higher frequencies [See next slide]
Attenuation curves for 4 different types of headphones

- Open-Back Headphone
- Personal Speaker
- Closed-Back Headphone
- Closed In-earphone

**ANC OFF**
Headphones Isolation Characteristics

Attenuation curves for 4 different types of headphones
# Headphone Isolation - Summary

<table>
<thead>
<tr>
<th>Type I (Personal Speaker)</th>
<th>Type II (Open-back over ear headset)</th>
<th>Type III (Closed in-ear headset)</th>
<th>Type IV (Closed-back over ear headset)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No earcup</td>
<td>Open earcup</td>
<td>Closed In-ear</td>
<td>Closed earcup</td>
</tr>
<tr>
<td>$h_{int}(n)$</td>
<td>$h_{int}(n)$</td>
<td>$h_{int}(n)$</td>
<td>$h_{int}(n)$</td>
</tr>
<tr>
<td>$r(n)$</td>
<td>$r(n)$</td>
<td>$r(n)$</td>
<td>$r(n)$</td>
</tr>
<tr>
<td>$m_{int}$</td>
<td>$m_{int}$</td>
<td>$m_{int}$</td>
<td>$m_{int}$</td>
</tr>
</tbody>
</table>

$A(f) \approx 1$

$A(f) = \begin{cases} 
\approx 1; & f < f_{th} \\
\ll 1; & f \geq f_{th} 
\end{cases}$

$A(f) \ll 1$

$A(f) \ll 1$

---

**Graphs:**

- **Sony PFR (Personal Speaker):**
  - Passive Hear Through
  - Attenuation (dB) vs Frequency (Hz)

- **AKG 702 (Open Back):**
  - Passive Hear Through
  - Active Hear Through
  - Attenuation (dB) vs Frequency (Hz)

- **Sony 1000 X2 (Closed back ANC OFF/ON):**
  - Active Hear Through
  - Attenuation (dB) vs Frequency (Hz)

- **QC 30 (In ear ANC OFF/ON):**
  - Active Hear Through
  - Attenuation (dB) vs Frequency (Hz)
Active Hear Through Mode

- Open ear scenario (Reference)

Signal at $m_{int}$:

$r(n) * h_{ref}(n)$

Reference real signal, $r_{ref}(n)$ captured without headphones

$h_{ref}(n)$: Impulse response at $m_{int}$ measured without headphones
**Active Hear Through Mode**

- **Attenuated/Leaked real signal (No EQ)**

**Diagram:**
- Real source \( r(n) \)
- Impulse response at \( m_{int} \) \( h_{int}(n) \)
- Leakage real signal, \( r_{int}(n) \) captured with headphones

**Equation:**
\[
\text{Signal at } m_{int} : \quad r(n) * h_{int}(n)
\]

**Explanation:**
- \( h_{int}(n) \): Impulse response at \( m_{int} \) measured with headphones
- \( r(n) \): Real source signal
- \( r_{int}(n) \): Leaked real signal captured with headphones
Active Hear Through Mode

- Equalized/Compensated real signal (After EQ)

Complete acoustical transparency can be achieved by recording, processing, and playback of real sound at an external microphone.
Active Hear Through Mode

- Equalized/Compensated real signal (After EQ)

Signal at $m_{int}$:

$$r(n) * h_{int}(n) + r_{ext}(n) * s(n) * h_{hp}(n)$$

Processed real signal, $\hat{r}_{ref}(n)$ after headphone playback

- $h_{ext}(n)$: Impulse response at $m_{ext}$ measured without headphones
- $s(n)$: Hear through EQ filter
Active Hear Through Mode

- Equalized/Compensated real signal (After EQ)

Signal at $m_{int}$:

$$r(n) * h_{int}(n) + r_{ext}(n) * s(n) * h_{hp}(n)$$

Processed real signal, $\hat{r}_{ref}(n)$ after headphone playback

Active Hear through EQ design factors:

1. Leaked real signal must be strongly isolated i.e., $r_{int}(n) \approx 0$
2. Processed real signal should follow reference real signal i.e., $\hat{r}_{ref}(n) \equiv r_{ref}(n)$
3. Minimum electrical delay between leaked and processed real signal
Active Hear Through Mode

Assuming energy of leaked real signal much lesser than that of processed real signal i.e., $E[r_{int}(n)] \ll E[\hat{r}_{ref}(n)]$

$$
\begin{align*}
    r(n) * h_{int}(n) + r_{ext}(n) * s(n) * h_{hp}(n) \\
    r_{int}(n) \approx 0 & \quad & \hat{r}_{ref}(n) \equiv r_{ref}(n)
\end{align*}
$$

\[ \Downarrow \]

$$
H_{ext}(f)S(f)H_{hp}(f) = H_{ref}(f)
$$

\[ \Downarrow \]

$$
S(f) = \frac{H_{ref}(f)}{H_{ext}(f)} \times \frac{1}{H_{hp}(f)}
$$

- Hear-through EQ must account for difference between transfer function $H_{ext}(f)$ and $H_{ref}(f)$
- Headphones non-flat response should be equalized while playing back (same as Headphone EQ)
Type II (Open-back)

- **Real source** $r(n)$
  - $h_{ext}(n)$
  - $m_{ext}$
  - $r_{ext}(n)$
  - $s(n)$
  - $r_{int}(n)$
  - $m_{int}$

1. EQ designed as **high-pass filter**
2. EQ requires **pinnae cues** to be embedded
3. Strong Comb effect due to poor attenuation
4. Delay between leaked and processed real signal in high frequency
### Active Hear Through Mode - Summary

<table>
<thead>
<tr>
<th>Type II (Open-back)</th>
<th>Type III (Closed In-ear)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="ActiveHearThroughMode.png" alt="Diagram" /></td>
<td><img src="ActiveHearThroughMode.png" alt="Diagram" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Type II (Open-back)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>EQ designed as <em>high-pass filter</em></td>
</tr>
<tr>
<td>(2)</td>
<td>EQ requires <em>pinnae cues</em> to be embedded</td>
</tr>
<tr>
<td>(3)</td>
<td>Strong Comb effect due to poor attenuation</td>
</tr>
<tr>
<td>(4)</td>
<td>Delay between leaked and processed real signal in high frequency</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Type III (Closed In-ear)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Best suited for EQ if tightly fitted as <em>pinnae cues are preserved</em> in $r_{ext}(n)$</td>
</tr>
<tr>
<td>(2)</td>
<td>Occlusion produces <em>unnatural listening</em> of real sound</td>
</tr>
<tr>
<td>(3)</td>
<td>Loose fittings result in poor isolation</td>
</tr>
<tr>
<td>(4)</td>
<td>Delay between leaked and processed real signal</td>
</tr>
</tbody>
</table>
### Active Hear Through Mode - Summary

<table>
<thead>
<tr>
<th>Type II (Open-back)</th>
<th>Type III (Closed In-ear)</th>
<th>Type IV (Closed-back)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Diagram" /></td>
<td><img src="image2" alt="Diagram" /></td>
<td><img src="image3" alt="Diagram" /></td>
</tr>
<tr>
<td>Real source $r(n)$</td>
<td>Real source $r(n)$</td>
<td>Real source $r(n)$</td>
</tr>
<tr>
<td>$h_{ext}(n)$</td>
<td>$h_{ext}(n)$</td>
<td>$h_{ext}(n)$</td>
</tr>
<tr>
<td>$m_{ext}$</td>
<td>$m_{ext}$</td>
<td>$m_{ext}$</td>
</tr>
<tr>
<td>$m_{int}$</td>
<td>$m_{int}$</td>
<td>$m_{int}$</td>
</tr>
<tr>
<td>$r_{ext}(n)$</td>
<td>$r_{ext}(n)$</td>
<td>$r_{ext}(n)$</td>
</tr>
<tr>
<td>$r_{int}(n)$</td>
<td>$r_{int}(n)$</td>
<td>$r_{int}(n)$</td>
</tr>
<tr>
<td>$s(n)$</td>
<td>$s(n)$</td>
<td>$s(n)$</td>
</tr>
</tbody>
</table>

- **Type II (Open-back):**
  1. EQ designed as high-pass filter
  2. EQ requires pinnae cues to be embedded
  3. Strong Comb effect due to poor attenuation
  4. Delay between leaked and processed real signal in high frequency

- **Type III (Closed In-ear):**
  1. Best suited for EQ if tightly fitted as pinnae cues are preserved in $r_{ext}(n)$
  2. Loose fittings result in poor isolation
  3. Delay between leaked and processed real signal
  4. Occlusion produces unnatural listening of real sound

- **Type IV (Closed-back):**
  1. Need EQ for entire spectrum
  2. Strong isolation irrespective of headphone fitting
  3. Open ear canal listening but pinnae cues need to be embedded

Both Type III & IV design additionally gives us more control over real sounds. Real sounds can either be 1) fully blocked 2) selectively passed or 3) completely hear through.
C.5 Virtual Reality Mode

Binaural Synthesis using headphones playback

- Virtual monaural signal convolves with Binaural room transfer function
- Individual HPTF effect must be removed using equalization filter:
  - Direct inversion of HPTF
  - Using an adaptive algorithm like FxLMS

\[
\frac{1}{H_{hp}(f)} \cdot BRTF(f) = x(n)
\]

[Houchard, 2006]
[Kuo and Morgan, 1995]
Type I – Open ear Design

- Most simplest form of design as real sounds reach unaltered allowing natural fusion

- Binaural rendering (HRTF & HPTF)
- Acoustic environment rendering (BRTFs)
- HoloLens with open ear design
  - Ideal hear through
  - HRTF adapted to estimated head width
  - Fixed Room Reverb options (small, medium, large)
Type II – Open-back Design

A headset structure with two pairs (int/ext) binaural microphones attached to the earcups.

- Headset equipped with 2 pairs of binaural microphones
- Adaptive Headphone equalization for virtual augmented sounds
- Natural mixing of real and virtual sources (Passive hear through)

Internal microphone used as error microphone to adapt the virtual sound at ear canal to natural sound. External microphone used as reference microphone to capture real sounds.

[Ranjan and Gan, 2015]
Type II – Augmented Reality Mode

Augmented reality mode – virtual sound reproduction in the presence of external signals

Aim: To reproduce virtual sources as if they sound similar to physical sources, without being affected by external sounds

Personalized HRTF selected from database

Desired signal path

Actual signal path

$x(n)$ $\rightarrow h_{\text{int}}(n)$ $\rightarrow \sum d(n)$

$r(n)$ $\rightarrow h_{\text{int}}(n)$ $\rightarrow \sum y_{\text{int}}(n)$

$x(n)$ $\rightarrow w(n)$ $\rightarrow h_{\text{hp}}(n)$ $\rightarrow \sum y_{\text{int}}(n)$

$r(n)$ $\rightarrow h_{\text{ext}}(n)$ $\rightarrow h_{\text{he}}(n)$ $\rightarrow \sum r_{\text{int}}(n)$

$x_{\text{int}}(n) = w(n) \ast h_{\text{hp}}(n) \ast x(n)$

$W(f) = \frac{H_{\text{int}}^v(f)}{H_{\text{hp}}(f)}$

[Source: Ranjan and Gan, 2015]
Type II – \textit{int} and \textit{ext} Transfer Functions

- \(H_{\text{int}}(z)\) is an approximate HRTF with additional headphone effects
- \(H_{\text{ext}}(z)\) contains all individual related characteristics and environment minus the pinnae specific notch and headphone shell reflections
Augmented reality mode – virtual sound reproduction in the presence of external sounds: Hybrid Adaptive Equalizer (Assuming negligible leakage signal power, \( l(n) = 0 \))

\[
x(n) \xrightarrow{\text{Desired signal path}} u(n) \xrightarrow{\text{Hybrid adaptive equalizer}} d(n) \xrightarrow{\text{Adaptive estimation}} r_{\text{ext}}(n)
\]

\[ y_{\text{int}}(n) + \hat{r}_{\text{int}}(n) \]

[Source: Ranjan and Gan, 2015]
Augmented reality mode – virtual sound reproduction in the presence of external sounds: Hybrid Adaptive Equalizer (Assuming negligible leakage signal power, $l(n) = 0$)

**Type II** – Mixing virtual augmented signal with real ext source

**Hybrid adaptive equalizer**: simple combination of conventional and modified FxNLMS
Augmented reality mode – virtual sound reproduction in the presence of external sounds: Hybrid Adaptive Equalizer (Assuming negligible leakage signal power, \( l(n) = 0 \))

**Type II** – Mixing virtual augmented signal with real ext source

**Adaptation filter corresponding to conventional FxNLMS**
- Slower convergence rate due to presence of secondary path transfer function

**Adaptation filter corresponding to Modified FxNLMS**
- Faster convergence rate by introducing spatial filter, \( h^v_{ext}(n) \) in the secondary path but slightly higher steady state error (shorter filter taps)
Augmented reality mode – virtual sound reproduction in the presence of external sounds: Hybrid Adaptive Equalizer (Assuming negligible leakage signal power, \( l(n) = 0 \))

**Hybrid adaptive filters:**

\[
\begin{align*}
    w(n) &= w_1(n) + h^v_{ext}(n) * w_2(n) \\
    W(f) &= W_1(f) + H^v_{ext}(f)W_2(f)
\end{align*}
\]

- Spatial information retained in \( h^v_{ext}(n) \) results in faster convergences and smaller MSE using hybrid adaptive filters.
Type II – Adaptive estimation

Augmented superimposed signal:

\[ y_{int}(n) = x_{int}(n) + r_{int}(n) , \]
where,

\[ x_{int}(n) = h_{hp}(n) \ast u(n) \]

Error signal:

\[
e'(n) = \{ d(n) + \hat{r}_{int}(n) \} - y_{int}(n) \\
= \{ d(n) - x_{int}(n) \} + \{ -(r_{int}(n) - \hat{r}_{int}(n)) \} \\
= e_v(n) + e_r(n)
\]
Hybrid FxNLMS with adaptive estimation works equally well even in the presence of real sounds reproducing the virtual sources as close as possible to real sources.
Type II – Listening test setup

- 7 loudspeakers: 5 in horizontal plane and 2 in median plane

Listening Test Set up (겠습니다: Elevated speaker; ┩ธุ: Azimuth speaker)
Type II – Listening Test Results

- Subjective test evaluation for sound similarity and source position similarity

[Diagram showing speaker positions and sound similarity evaluation]

[Graphs showing sound similarity and source position similarity]

[Source: Ranjan and Gan, 2015]
Type III – Closed In ear headset design

- Use of closed in-earphones with external mic to capture the real sound, process, and playback
- Basic idea is to relay the external sounds unaltered with minimum latency (<1ms)

Left and middle: ARA headset (Philips SHN2500)
Right: Prototype ARA mixer

ARA headset system diagram

[Härmä et al 2004; Tikander et al 2008]
Difficult to predict the headphone response if loosely fitted

[Rämö and Välimäki, 2012]
Closed in-ear phones modify the ear canal resonance due to blocked ear canal (pressure chamber principle).

- Unequalized ARA headset
- Open ear case

Diagram showing the frequency response with and without equalization.
Type III – ARA Headset Equalizer

- Generic ARA equalization based on 4 individual measurements

[Rämö and Välimäki, 2012]
Type III – Hear Through Mode using an All Pass

- Extend the attenuated response $h(n)$ at ear drum with an all pass tail to make the entire spectrum flat:

All pass system magnitude

Extended impulse response with all pass tail

[Rämö and Välimäki, 2014]
**Type III – Adaptive Hear Through**

- Adaptive equalization of acoustical transparency for In-ear headphones to estimate isolation curve online and apply hear through EQ

![Diagram of Adaptive Hear Through](image)

[Juho et al, 2016]
Type III – Adaptive Hear Through - Results

Measured Isolation (dashed red) Vs Estimated Isolation Curve (solid black)

Target Open ear responses (dashed red) Vs equalized HT responses (solid black)

Measured Isolation (dashed red) Vs Estimated Isolation Curve (solid black) for non-ideal direction (75°)

Effect of automatic EQ on poor headphone fitting. Fixed EQ (Solid blue) Vs Automatic EQ (Solid Black)
Type IV – Closed-back Design

- Closed back ear cup design with internal and external microphones

\[ m_{\text{ext}} \quad m_{\text{int}} \]

**Type IV Design Prototype using Sony MDR 1000 XZ**

- **Internal microphone** used as error microphone for virtual sound adaption as well as hear through. **External microphone** used as reference microphone to capture real sounds and for hear through coupled with internal microphone.
Type IV – Closed-Back Hear Through using FxLMS

\[ r_{ext}(n) \]

\[ \hat{h}_{hp}(n) \]

\[ r'_{ext}(n) \]

\[ FxNLMS \]

\[ s(n) \]

\[ u(n) \]

\[ h_{hp}(n) \]

\[ \Sigma \]

\[ d(n) \]

\[ r_{int}(n) \]

- **r_{int}(n):** Real signal attenuated through headphones at \( m_{int} \)
- **d(n):** Desired reference real signal,
- **r_{ext}(n):** Real signal captured at \( m_{ext} \)
- **s(n):** Hear Through EQ filter

Hear through EQ filter is **tuned to minimize the error** signal due to difference between desired reference signal and processed EQ signal.
Type IV – Closed-Back Hear Through using FxLMS

Rewriting ideal EQ filter:

\[ S(f) = \frac{H_{\text{ref}}(f)}{H_{\text{ext}}(f) H_{\text{hp}}(f)} = \frac{H_{\text{ear}}(f)}{H_{\text{hp}}(f)} \]

Accounts for **directional pinnae cues**. Can be modelled by measuring transfer functions at two microphones.

Rewriting desired signal:

\[ d(n) = r(n) * h_{\text{ref}}(n) \]
\[ = r(n) * h_{\text{ext}}(n) * h_{\text{ear}}(n) \]
\[ = r_{\text{ext}}(n) * h_{\text{ear}}(n) \]

Desired signal can be equivalently expressed as \( r_{\text{ext}}(n) \) passing through a directional filter \( h_{\text{ear}}(n) \).
Type IV – Closed-Back Hear Through using FxLMS

\[
\sum \delta(n)u(n) - s(n)\hat{h}_\text{hp}(n) = d(n)
\]

\[
r_{\text{ext}}(n) \rightarrow h_{\text{ear}}(n-\Delta) \rightarrow s(n) \rightarrow u(n) \rightarrow h_{\text{hp}}(n) \rightarrow \sum \rightarrow r_{\text{int}}(n)
\]

\[
\hat{h}_\text{hp}(n)
\]

\[
r_{\text{prim}}(n)
\]

\[
\Delta: \text{Minimum estimated group delay of secondary path}
\]

Un-equalized response:

\[
H_{\text{ext}}(f)H_{\text{hp}}(f) + H_{\text{int}}(f)
\]

Equalized Transfer Function:

\[
H_{\text{ext}}(f)S(f)H_{\text{hp}}(f) + H_{\text{int}}(f)
\]

**Closed-back hear through EQ can be computed:**

1. If Directional transfer function, \(h_{\text{ear}}(n)\) is known (can be modelled), And
2. Introducing a minimum delay in primary path to ensure EQ can converge
Type IV – Closed-Back Hear Through Results

Hear Through $0^\circ$ with ANC OFF

Magnitude in dB (dBA/20mPa)

Frequency (log scale)

- Equalized Response
- Target Response
- Headphone Isolation Response
- Unequalized Response
Type IV – Closed-Back Hear Through Results

Hear Through \(0^\circ\) with ANC ON
Type IV – Closed-Back Hear Through Results

Hear Through $45^\circ$ with ANC ON
What’s next...

- Hear Through EQ may also vary for different source incident directions
  - One fixed average EQ Vs group of EQs
- Diffuse sound field or multiple real sound sources scenarios
- Headphone isolation can be highly idiosyncratic (especially for closed-back headphones)
- Perceptual evaluation of localization and timbre quality of hear through mode
- Subjective impression of comb effect due to leakage of real signal
## Natural Listening in AR/MR - Summary

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Type I – Open ear</th>
<th>Type II – Open-back Over ear</th>
<th>Type III – Closed in-ear</th>
<th>Type IV – Closed-back Over ear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real sound reproduction</td>
<td>Heard as is – No processing required</td>
<td>Only higher frequencies may be compensated</td>
<td>Recorded, processed for entire spectrum</td>
<td>Non-natural listening, No fitting issue, pinna cues to be embedded for active HT</td>
</tr>
<tr>
<td>Natural listening, No obstruction</td>
<td>Natural until mid-frequency, pinna cues preserved if passive HT, comb-effect if active HT</td>
<td>Non-natural listening, strong comb-effect due to fittings issue, pinna cues preserved for active HT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virtual sound reproduction</td>
<td>Personal micro-speaker used – Open ear listening</td>
<td>Over the ear emitters used – Open ear listening</td>
<td>In ear emitters used – Blocked ear canal effect</td>
<td>Over the ear emitters used – Open ear listening</td>
</tr>
<tr>
<td>Characteristics</td>
<td>Low volume, poor bass, No universal EQ,</td>
<td>High volume, personalized headphone EQ</td>
<td>Generic EQ</td>
<td>Personalized headphone EQ</td>
</tr>
<tr>
<td>Leakage, natural mixing, good externalization, <strong>only AR/MR</strong></td>
<td>High volume, proper mixing required, poor externalization, <strong>VR/AR/MR</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
C.6 Acoustic Environment Estimation and Rendering

- Critical for virtual objects to sound discernible from real sounds in an augmented reality environment (ARE)

- Acoustic environment characteristics must be captured and embedded into the binaural playback

\[
\text{Local acoustic environment} + \text{Binaural render} = \text{Natural Listening AR audio}
\]
Acoustic Environment – Room Impulse Response

- Characterized by room impulse response (RIR), which accounts for sense of environment to listener:

  - Three major components of RIR:
    - Direct Sound: Straight path between Source and Receiver
    - Early Reflections: Sparse first few reflections from source to receiver
    - Reverberations: densely populated reflections (best described statistically)
RIR – Energy Decay Curve

- Energy decay curve (EDC): signal energy remaining in RIR at time $t$
  \[ EDC(t) = \frac{\int_{t}^{\infty} h^2(\tau) \, d\tau}{\int_{0}^{\infty} h^2(\tau) \, d\tau} \]

- Reverberation Time ($T_{60}$): Time when EDC crosses -60 dB
**Energy Decay Relief (EDR):** EDC generalized to frequency bands

- Used to calculate frequency dependent reverberation time using linear curve fitting

\[
EDR(t, f) = \frac{\int_t^\infty h^2(\tau, f) d\tau}{\int_0^\infty h^2(\tau, f) d\tau}
\]

- **Measured EDR**
- **Modelled EDR**

[Jot, 1992]

[Jot and Lee, 2016]
RIR – Schroeder frequency

- Frequency response of RIR can be divided into two regions:
  - Sparsely distributed low frequency resonant modes
  - Densely populated resonance modes
- Schroeder frequency defined as transition frequency between the two regions:
  \[ F_C = 2000 \sqrt{\frac{T_{60}}{V}} \]

Example:

Bathroom \( V = 10 \) m\(^3\), \( T_{60} = 0.35 \)s
\[ F_C = 374 \text{ Hz} \]
Environment (RIR) rendering approaches

- **Physics based rendering:** Akin to simulating visual reality. Use computer aided model of environment to compute impulse response.
  - **Wave Based theoretical methods:** Numerically solve wave equations for sound using FEM, BEM, FDTD etc. Very close to what would we measure.
  - **Geometrical Acoustics:** Discretize sound waves as rays and use geometrical approximation of wave equation, image source, ray tracing, beam tracing etc.

- **Perceptual based rendering:** Synthesizes an impulse response with perceptual impression similar to real IR. Artificial reverberation based model of real IR.
Wave Based methods

- Highly accurate method and closest to what we measure from physical world
  \[ \frac{\partial^2 p}{\partial t^2} - c^2 \nabla^2 p = F(x, t) \]
- Solve Helmholtz Wave Equation

<table>
<thead>
<tr>
<th>Wave Based Methods Summary</th>
</tr>
</thead>
</table>
| **Finite-Difference Time-Domain (FDTD)** | Acoustic space is discretized in **uniform spaced and shaped mesh**  
  Second order partial derivatives using finite differences in time domain  
  Straightforward and simple to implement |
| **Finite Element Method (FEM)** | Volume of the acoustic space is discretized into **arbitrary shape and size**  
  Wave equation is solved numerically using PDEs  
  **Closed/interior areas** are best solved using FEM  
  More accurate than FDTD but computationally more demanding |
| **Boundary Element Method (BEM)** | **Discretize only boundary** of acoustic domain and sound propagation is defined at the boundaries  
  Surface integral of the pressure and its derivatives are solved  
  **Not limited to closed space** modelling unlike volume based methods |
Wave Based methods

If they are the most accurate, then why don’t we usually use these methods for real-time acoustics simulation of any environment?

- Because they are Compute intensive $\propto f^4$
- Most of the cost spent on high frequencies, where we don’t care about so much details
- Too expensive for real-time computation and some approximation is required
- Recent fast techniques shows significant speedups incorporating moving sound sources and listeners

Current limitations:
- Static Scenes and high memory requirement
- Low frequencies up to 1.5 kHz for medium sized room

[Raghuvanshi et al, 2009,2010]
**Geometrical Acoustics: Image source method**

- Source reflections are created using image equivalence

- Start from Source -> Reflect against all rigid walls -> Check for listener visibility -> Reflect image sources -> And so on...

- Accurate but number of image sources increases exponentially after first few order of reflections -> Truncated

- Wall surfaces are assumed to be smooth i.e., only specular reflections are allowed
**Geometrical Acoustics: Ray Tracing**

- **Direct sound path between source and listener**
  (distance attenuation)

- **Early/Specular reflections**
  (multiple sound paths due to reflecting surfaces)

- **Reverberations**
  (Have no direction & densely populated)

- **Immersive audio output**
  (Binaural rendering applied to provide natural listening)

- **Diffraction (occlusion effect)**
  (Sound reflects around object edges and changes phase)

- **Diffuse reflections**
  (Scattering due to roughness of the surfaces)

Sound waves (aka. *approximated as sound rays*) bounces off with walls and objects (represented as triangulated 3D mesh) and reaches listener’s ears accounting for human head acoustics model.
Demo scenes

Acoustics reflections of different material surfaces produces realistic sound effects.

Acoustics reflections of occluding objects (diffraction) gives the impression of real-life situations.

Room materials effect

Occlusion effect

[Source: Immerzen Labs Pte. Ltd. Singapore]
Perceptual Methods

- There are too much details in physical methods
- If physical accuracy not required, perceptual methods are better alternative
- To simulate what is perceptually important NOT physically
What factors are perceptually important

- **Early Reflections**
  - Spaciousness, envelopment and apparent source width
  - Dependent on source and listener position and orientation
  - Image source or Ray Tracing is used widely

- **Late Reflections**
  - Reverberation Time, $T_{60}(f)$ -> gives impression of size
  - Direct to Reverberant Ratio -> Affects source-listener distance perception
  - Echo density -> Tells about texture information of environment
  - Modal density -> Necessary for natural sounding reverb
  - Can be modelled stochastically
Schroeder Reverberator

- First digital reverberator using comb and all-pass filter

![Diagram of Schroeder Reverberator](image)

- Parallel comb filter
- Series all-pass filter

- Modal density
- Decay time
- Air absorption

Echo density

[Schroeder, 1962; Gardner, 1998]
Feedback Delay Network (FDN)

- Generalized version of Schroeder Reverberator
- Design methodology:
  1. Design lossless prototype with infinite reverberation time
  2. Add losses (absorption) to each delay unit to obtain desired $T_{60}(f)$

\[
20 \log_{10} |G_i(e^{j\omega t})| = -60 \frac{M_i T}{T_{60}(\omega)}
\]

[Jot and Chaigne, 1991]
Scattering Delay Network

- Approach in between delay networks and physical models
- One node per reflecting surface
- Approximation of image source method

[Karjalainen et al., 2005; Sena et al., 2015]
Scattering Delay Network

- Approach in between delay networks and physical models
- One node per reflecting surface
- Approximation of image source method

[Scattering Delay Network diagram]

[Karjalainen et al., 2005; Sena et al., 2015]
Scattering Delay Network

- Approach in between delay networks and physical models
- One node per reflecting surface
- Approximation of image source method

[Karjalainen et al., 2005; Sena et al., 2015]
Environment Rendering Methods - summary

- **Wave Methods:**
  - Infeasible for high frequencies

- **Geometrical Acoustics**
  - Image source:
    - Only possible for early reflections. Usually combined with ray tracing for accuracy
  - Ray Tracing:
    - High frequency approximation
    - Choice of number of rays and size of source & listener is critical
    - One may not be able to find all reflections

- **Perceptual Methods**
  - Late reflections can be modelled stochastically using FDN/SDN

Hybrid method: **Wave** (Low frequency) + **GA** (High Frequency, Early & Late Reflections) / **Perceptual Methods** (Late Reflections)
## Environment Rendering: Summary

<table>
<thead>
<tr>
<th>Method</th>
<th>Speed/Load</th>
<th>Accuracy</th>
<th>Interactivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wave Based</td>
<td>Slow, Very high</td>
<td>Excellent</td>
<td>Yes</td>
</tr>
<tr>
<td>Geometrical Acoustics</td>
<td>Very fast, High</td>
<td>Very good at high frequency, medium at low frequency</td>
<td>Yes</td>
</tr>
<tr>
<td>Hybrid Wave(LF) + GA(HF)</td>
<td>Fast, High</td>
<td>Very good</td>
<td>Yes</td>
</tr>
<tr>
<td>FDN</td>
<td>Very Fast, Low</td>
<td>Poor</td>
<td>No</td>
</tr>
<tr>
<td>SDN</td>
<td>Very Fast, Low</td>
<td>Medium</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Acoustic Environment Estimation

Local Acoustic Environment Estimation

1. Geometrical acquisition based
2. Binaural recording based
3. Artificial reverberation based
4. Room identification based
Geometry Acquisition and Acoustics Processing

- Geometry acquisition provides 3D mesh of the real space to be used by GA methods

**Geometry Acquisition**
- 3D map plan/model database
- Manual measurements
- 3D depth scanning with semantic estimation

**Geometrical Acoustics Processing**
- Geometry Triangulation
- Geometry artefacts repair
- Acoustics simplification (LOD based)
- Real-time Ray/Beam tracing

For AR, dynamic changing scenes need to be captured instantly, processed, and rendered in real-time
Using 3D scanning

- 3D depth sensing technology can provide an approximate 3D mesh of local environment geometry

* 3D scanning devices*

* Consists of stereo vision and depth sensing technologies
Current challenges for GA processing

- Geometry scanned are usually not perfectly closed
  - Cameras are usually placed in center of room and thus, cannot capture hidden objects/surfaces in space
  - Dynamic moving camera can solve this issue partially

- Holes and gaps in the scanned mesh must be repaired

- Acoustics processing doesn’t require as much detail as in visual processing
  - Mesh must be simplified and acoustically insignificant details can be removed

[Milos et al, 2013; Lukas and Vorlander, 2016]
Acoustic properties of surfaces (walls) is quite critical in perception of environmental type.

Geometry acquisition provides very detailed surfaces’ data -> Use it to identify surface type.

This depth information combined with RGB data can be used by material recognition algorithms.

Can also be used for complex surfaces like porous materials, rough surfaces etc. allowing more natural sound phenomena like scattering.

[David, et al 2012]

[Milos et al, 2013]
Surface Recognition for Acoustical Simulation

- Machine learning based approach applied to vision
- Based on extremely randomized trees approach using sub images for robust image classification

Random trees based on sub images

Confusion matrix for different ground materials

Finger snaps or claps used as excitation to capture instant BRIRs using microphones on MARA headset

BRIRs extraction applied on windowed samples of band-pass filtered (1.5-3 kHz) microphones signal after finger snap detection

Extracted BRIRs response will be colored due to non-flat snap signal spectrum. Other excitation methods with more flat spectrum can be used
Local Environment adaption – Statistical Approach

- Diffuse reverberation model (independent of source and listener) as *Reverberation fingerprint*

- Reverberation fingerprint of a room:
  - Reverberation Time, $T_{60}(f)$: derived as linear curve fitting on modelled EDR extrapolated back to time of emission $EDR(0, f)$
  - Room Volume, $V$: Estimated from initial power spectrum $P(f) (\propto 1/V)$

- **Advantage**: Just information of frequency dependent reverberation time and room volume required

[Jot and Lee, 2016]
Using Reverberation Fingerprint to match local room

(a) Reference

(b) Local
Using Reverberation Fingerprint to match local room

- Initial power offset $\frac{V_{ref}}{V_{local}}$
- Correction of time-frequency envelope using per frequency dB offset
Using Reverberation Fingerprint to match local room
How to obtain Reverberation Fingerprint?

- In an augmented environment, user can be in a space characterized by different acoustic properties
- On-the-fly acquisition using existing audio signals to define *Reverberation Fingerprint* to be further used for rendering
- **Blind estimation** of room acoustic parameters of an unknown environment using speech signal

![Diagram](image)

Speech Recording → Online estimation of Room Acoustic Features → Reverberation Time, $T_{60}(f)$, Room Volume, $V$ → Reverberation/Room Fingerprint

[Murgai et al, 2017]
Reverberation Time Estimation

Energy Envelope Estimation → Decay Start/Stop Detection → Decay Time Estimation → $T_r(f)$

- Cafeteria Bogota - T60 Comparison
  - Actual vs. Blindly Estimated

- Octagon Room, Queen Mary - T60 Comparison
  - Actual vs. Blindly Estimated

- Sound Lab, University Of Athens - T60 Comparison
  - Actual vs. Blindly Estimated

- Sports Centre, University of York - T60 Comparison
  - Actual vs. Blindly Estimated

[Murgai et al, 2017]
Room Volume Prediction

Speech impulse response estimation → Acoustics features extraction → GMM based volume prediction

T60, C80, C50, density of early reflections, kurtosis (t), density of low frequency room modes, and kurtosis (f).

[Graphs showing waveforms and volume prediction]

[Murgai et al, 2017]
Room Identification using Acoustic Features

Training of known rooms

- Known room recordings (speech, music, combined)
- Audio features extraction (MFCC, spectrogram)
- GMM based room training
- Room model database

Identification of unknown rooms

- Unknown room recordings
- Audio features extraction
- Compute likelihood score for unknown feature vectors with each room model
- Return room with highest likelihood score

[Reference: P. Nils et al, 2016]
## Environment Estimation and Rendering - Summary

<table>
<thead>
<tr>
<th>Environment Estimation Methods</th>
<th>Output of Environment Estimation method</th>
<th>Suitable Environment Rendering approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometrical acquisition</td>
<td>3D geometry mesh with semantic information</td>
<td>Geometrical acoustics (Image source + Ray Tracing) or Wave + GA methods</td>
</tr>
<tr>
<td>Binaural Recording</td>
<td>Binaural room impulse response (BRIR)</td>
<td>BRIR convolution</td>
</tr>
<tr>
<td>Artificial Reverberation</td>
<td>Reverberation fingerprint</td>
<td>FDN/SDN</td>
</tr>
<tr>
<td>Room Identification using acoustic features</td>
<td>Room models database</td>
<td>Pre-stored RIR convolution</td>
</tr>
</tbody>
</table>

**Module C**

WS Gan, JJ He, R Ranjan, R Gupta  
Natural and augmented listening for VR, AR, MR  
16th Apr. 2018  
C.86 /95
C.7 Signal Processing Techniques Integration for AR/MR

Rendering of natural sound

Head movement
- Head tracking

Individual parameters
- Individualization

Virtual sources
- Binaural Rendering (Source)
  - Environment Rendering
  - Equalization
    - Environment Estimation
    - Hear Through Equalization

Real sounds

Headphone
Hearing aids
Headset

m_{int}

m_{ext}
Summary

- Acoustic Transparent hearing using passive/active hear through
- Headphone equalization for virtual sound rendering with adaptive estimation for real sounds
- Environment rendering
- Acoustic environment Estimation
Key References

Augmented/Mixed Reality Overview


Augmented/Mixed Reality Audio

Augmented/Mixed Reality Audio

Key References

Augmented/Mixed Reality Audio


Key References

Augmented/Mixed Reality Audio


Environment Estimation and Rendering

Key References

Environment Estimation and Rendering

Key References

Environment Estimation and Rendering

Key References

Environment Estimation and Rendering


Module D
Summary and Future Trends

1. Summary of key Techniques
2. Spatial Audio Tools
3. Emerging (potential) Applications of VR/AR Audio
4. Challenges and Future Research Trends
D.1 Spatial Audio Technologies for Immersive VR, AR/MR

**Spatial Audio Formats**
- Object, Ambisonics
- Parametric processing

**Depth camera**
- Reverberation fingerprint
- Machine learning

**Individualized Binaural Rendering**
- Individualized HRTFs
- Equalization

**Environment Estimation**
- Depth camera
- Reverberation fingerprint
- Machine learning

**Spatial Audio Formats**
- Object, Ambisonics
- Parametric processing

**Environment Rendering**
- Wave based
- Geometrical based
- Perceptual based

**Dynamic Binaural Synthesis**
- Head tracking
- Position tracking

**Virtual & Physical Sound Fusion**
- Adaptive equalization
- Hear-through processing
Summary: Different Listening Modes for VR and AR/MR

VR

- Head movement
- Individual parameters
- Virtual sources
- Virtual environment
- Head tracking
- Individualization
- Virtualization
- Binaural Rendering (Source)
- Environment Rendering
- Equalization
- Headphone Hearing aids Headset

AR/MR

- Head movement
- Individual parameters
- Virtual sources
- Real sounds
- Head tracking
- Individualization
- Virtualization
- Binaural Rendering (Source)
- Environment Rendering
- Equalization
- Environment Estimation
- Hear Through Equalization
- m_{int}
- m_{ext}
- Headphone Hearing aids Headset
## D.2 Binaural rendering tools

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Tool Name</th>
<th>Input format</th>
<th>Common Features</th>
<th>Unique features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HOA</td>
<td>FOA</td>
<td>Obj.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Audio occlusion effects, real time sound propagation**
- **Near field effects, occlusions, sound source directivity**
- **Real time motion tracked binaural playback while editing**
- **Volumetric sources, near field rendering**
- **Supports FOA and unlimited number of virtual speakers using proprietary GA5 format**
- **Low latency VR and mobile solution**
- **Supports mix to other formats such as 5.1, 7.0 allows tuning of HRTF using head measurements, EQ calibration for headphones supported**
- **Allows 3D sound object design inside VR allows custom HRTF tuning using anthropometric measurements and user selected HRTF playback in real time**
- **Headphone Optimization tech, Dynamic HRTF incorporating relative head-torso movement**
## Spatial Audio Tool References

<table>
<thead>
<tr>
<th>Companies</th>
<th>Websites</th>
</tr>
</thead>
<tbody>
<tr>
<td>SteamAudio</td>
<td>[1] SteamAudio - <a href="https://valvesoftware.github.io/steam-audio/">https://valvesoftware.github.io/steam-audio/</a></td>
</tr>
</tbody>
</table>
1. Spatial Audio Communication & Collaboration
2. Augmented Audio Tour
3. Augmented Assistive Listening for Visually Impaired
4. Soundscape Studies
1) Communication & Collaboration using VR and AR/MR

**Meeting Room**
- Enclosed room with reverberant characteristics

**Airport**
- Usually large hall
- Significant ambient sound

**Bedroom**
- Usually small-medium sized room
- Quiet environment

**AR/MR mode**

**Required Audio Technology**
- Environment estimation
- Adaptive Filtering
- ANC mode may be required if loud external noise
- Reverberation Rendering
- Dynamic Binaural audio
2) Augmented Audio Tour

Bose AR: Audio Augmented Platform

Welcome to virtual audio tour of world war memorial

Hey, buddy !!!!
3) Augmented Assistive listening for Visually Impaired

Caution ahead !!! turn right for Bell St.
Safety Headphones

Assistive Listening is also needed for the normal person!

Picture from IEEE Signal Processing Magazine, March 2018
4) What is Soundscape?

- Paradigm shift in urban sound evaluations from ‘Noise control’ into ‘Soundscape design’

**Soundscape (ISO/TC 43 SC1: DIS 12913-1)**

- Acoustic environment as perceived or experienced and/or understood by people, in context
- The challenge is to create good and health-promoting soundscapes in urban environments.
Selected previous studies

- Sound field behind different noise barriers calculated and auralized through VR

Fig. 5. Four noise barriers and sound environments proposed.

VS1: common rail, no sound barrier. VS2: concrete opaque, 1.2 m barrier. VS3: concrete vegetated with upper part in glass, 2 m barrier. VS4: concrete opaque with oval windows, 3 m barrier.

[Sanchez, 2017]
Selected previous studies

- Use of web-based PC and VR spatial sound tool for auralization of soundscapes
- Study found no statistical difference between evaluation in-situ and VR

[Jiang, 2018]
Overview of Our Augmented Urban Soundscape

**Aim**: To develop AR/VR design tools for soundscape design and evaluation

**STEP 1**
Capturing Urban Soundscape In SGP

- 3D Audio-Visual Recording
- Capturing pleasant masker sounds
- Analyzing psychoacoustic indicators

**STEP 2**
Psychoacoustic Evaluation Based on VR/AR

- Psychoacoustic evaluation using VR and AR
- Developing optimal masking algorithm

**STEP 3**
Design Parameters of AUS System

- Design parameters for AUS algorithm
- Aid in design of AUS in Phase 2
Capturing of Urban Soundscape

- **3D audio-video for VR**

  - Ambisonic Microphone
  - Spherical panoramic camera (VR)

  - Psychoacoustic indicators
  - Loudness
  - Sharpness
  - Roughness
  - Fluctuation strength
  - Tonality
  - ...

---
Psychoacoustic Evaluation using VR

VR scenario

- Laboratory condition (Controlled)
- VR Headgear
- 3D Spatial Audio
- Road traffic (Recorded)
- Water (Masker)
- Bird (Masker)
Video Demo: Spatial Audio for VR
Psychoacoustic Evaluation using AR

- **AR scenario**

  **In-situ (Real-life scenario)**

  - Augmented Reality Headgear
  - Hologram & Sound source
  - 3D Spatial Audio Headphones (Open-back)
  - Road traffic (Real)
  - Bird (Masker)
  - Water (Masker)
Augmented Reality
(Soundscape with static & movable masker in Yunnan Garden)
### Summary of Audio Techniques for Soundscapes

<table>
<thead>
<tr>
<th>Characteristics of the Acoustic Environment</th>
<th>Recommended Techniques</th>
<th>Use Case(s) (Selected References, if Any)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spatial Fidelity</strong></td>
<td><strong>Type of Environment</strong></td>
<td><strong>Movements</strong></td>
</tr>
<tr>
<td>Virtual (R/S)</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Virtual (R/S)</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Virtual (R/S)</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Med</td>
<td>Virtual (R/S)</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Virtual (R/S)</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Virtual (R/S)</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Virtual(S)</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Virtual(S)</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Hong, 2017]
D.4 Challenges of Spatial Audio for VR, AR/MR

- **Audio format** for VR/AR/MR (ambisonics vs object)
- **Audio reproduction system** (headphones vs speakers)
- **Low cost and effective HRTF individualization** method (including measurements) for consumer adoption
- **Basic Audio Quality** (Spatial and Timbre quality) vs **Overall Listening Experience** using Spatial Audio in VR/AR/MR
- **Distance rendering** (including near-field)
- **Latency** in dynamic binaural rendering
D.4 Challenges of Spatial Audio for VR, AR/MR

- Plausible hear through of real sound in AR/MR
- Real-time interaction of virtual audio in dynamic real environment (AR/MR), including the efficient methods for estimating environment acoustics in real-time (indoor and outdoor)

- How AI/machine learning can help:
  - Audio scene recognition for making informed decision
  - Individualization of HRTFs using photos
  - Environment estimation
  - Assisted listening
“The holy grail in truly immersive 3D sound is real-time customized spatial audio that is calibrated to the anatomical measurements of one’s ears and uses head-tracking technology to update the soundscape as one moves their head around. “It really becomes real to you, and vivid, if it feels like you’ve been immersed in a new, living acoustic reality;” “You feel like you’re somewhere else.”

From sound installation artist, Gabe Liberti
Key References


Selected Authors’ Publications on Spatial Audio

Acknowledgement

- Kaushik Sunder (Principal Scientist, Embody VR)
- Santi Peksi (Project Officer, NTU)
- Nguyen Duy Hai (Project Officer, NTU)
- Lokesh Dhakar (Co-founder, Immerzen Labs)
- Ronak Bajaj (Co-founder, Immerzen Labs)

**Tutorial (T-11) Companion Website**
(Contains supplementary and updated materials of this tutorial)

ICASSP ‘18 Demo

• Title: *An fast iHRTF Acquisition and Immersive 3D Audio Headset for Virtual and Augmented Reality* (ID #21)

• Date/Time: *Wednesday, April 18th, 13:30pm-15:30pm*

• Venue: *Exhibit Hall Foyer*

1st version of the iHRTF ACQ unit demoed in ICASSP 2017