IEEE Signal Processing Society

IEEE

2018 IEEE International Conference on Acoustics, Speech and Signal Processing

15–20 April 2018 • Calgary, Alberta, Canada

Signal Processing and Artificial Intelligence: Changing the World

(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

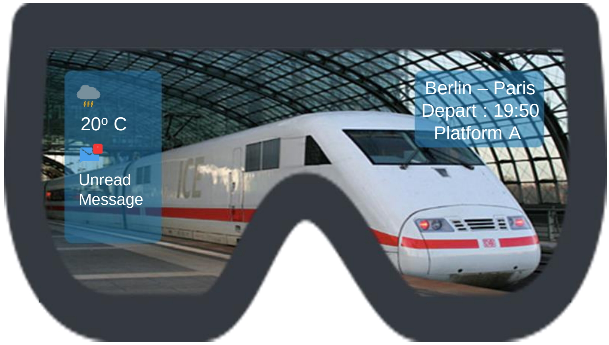
- Summary of key Techniques
- Spatial Audio Tools
- Emerging Applications of VR/AR Audio
- Challenges and Future Research Trends

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.

- Google cardboard
- Samsung Gear VR
- Oculus Rift

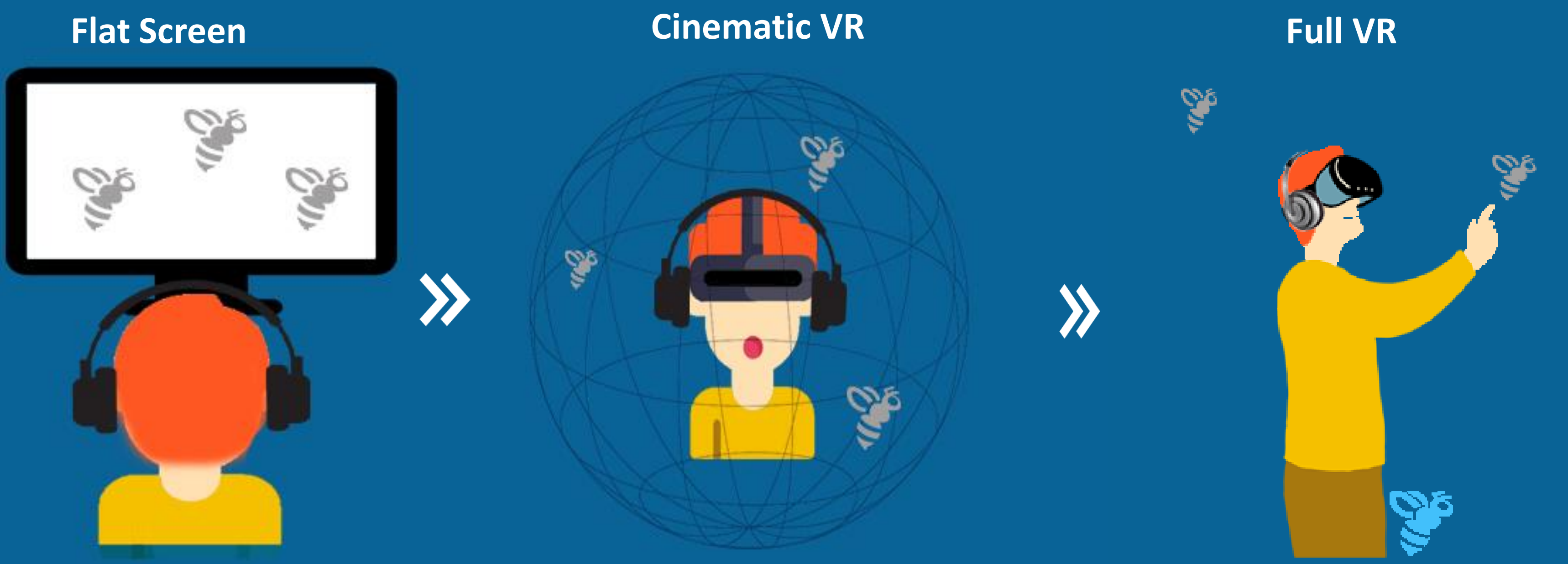
Augmented Reality (AR):
In a real world environment whose elements are augmented (overlays) by computer-generated (CG) sensory input (sound, video, data). However, the real-world content and the CG content are not respond/react to each other.

- Google Glass
- Bose AR
- Microsoft Hololens

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

- Magic Leap
- Meta 2
- HTC Vive Pro

From PC Flat Screen to Full 360 VR Experience



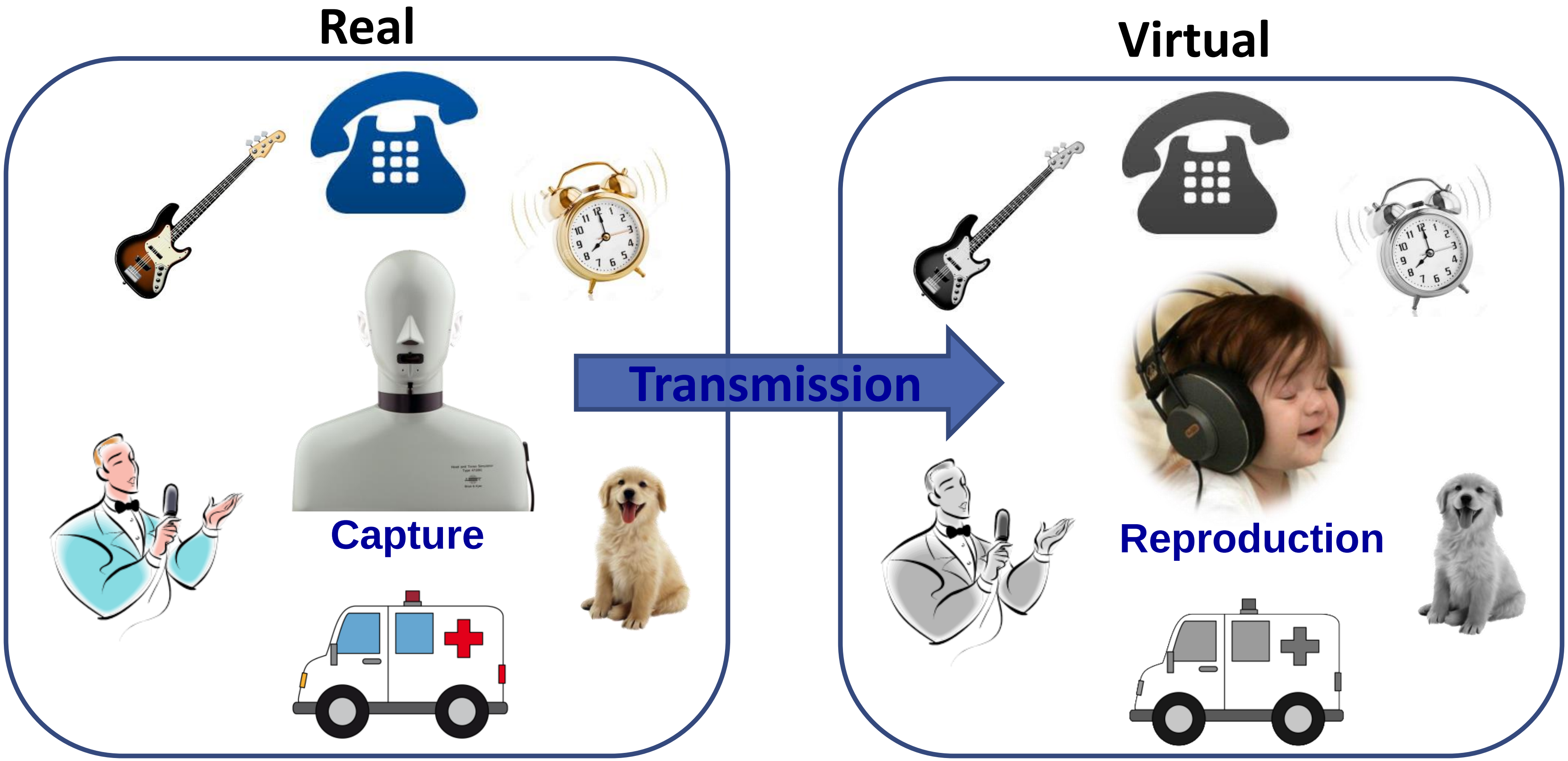
- Pre-rendering
- 0 DoF

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

- Changing environment & interaction
- 6 DoF

Illustration by Santi, image credit: Freepik.com

From Real Sound to Virtual Reality Audio



Real

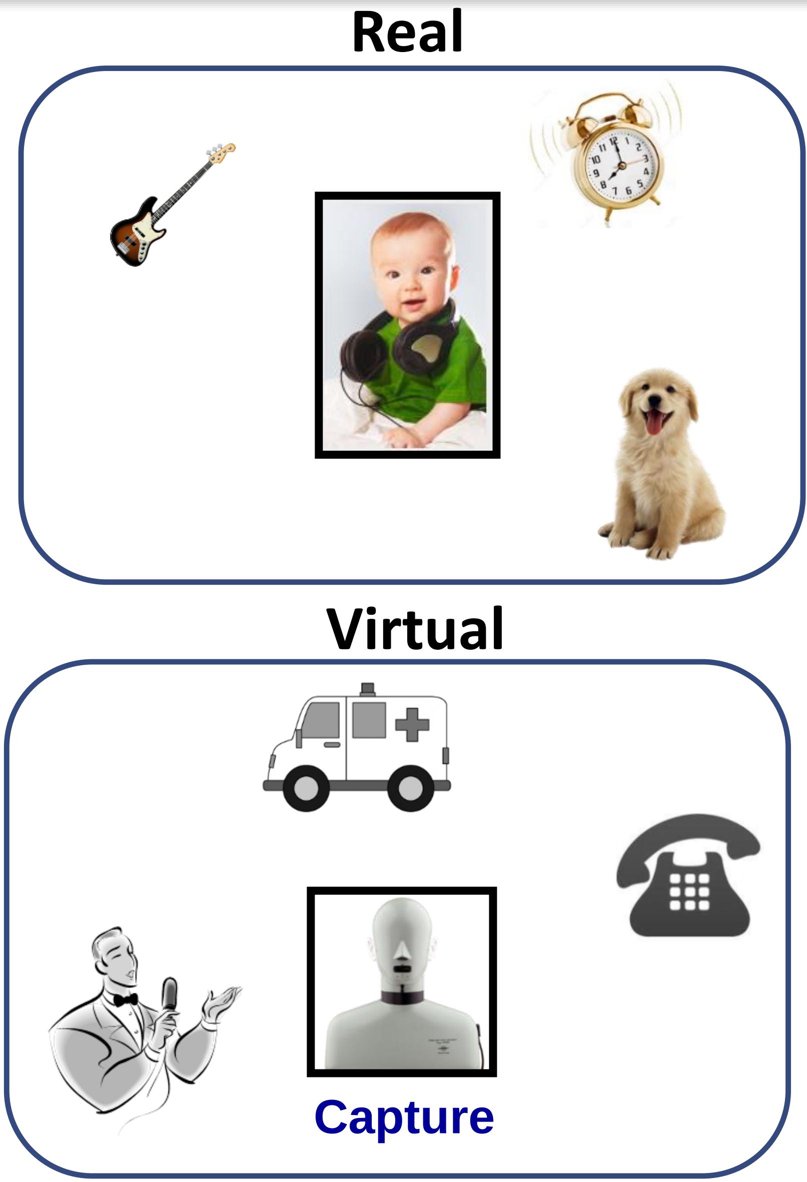
Virtual

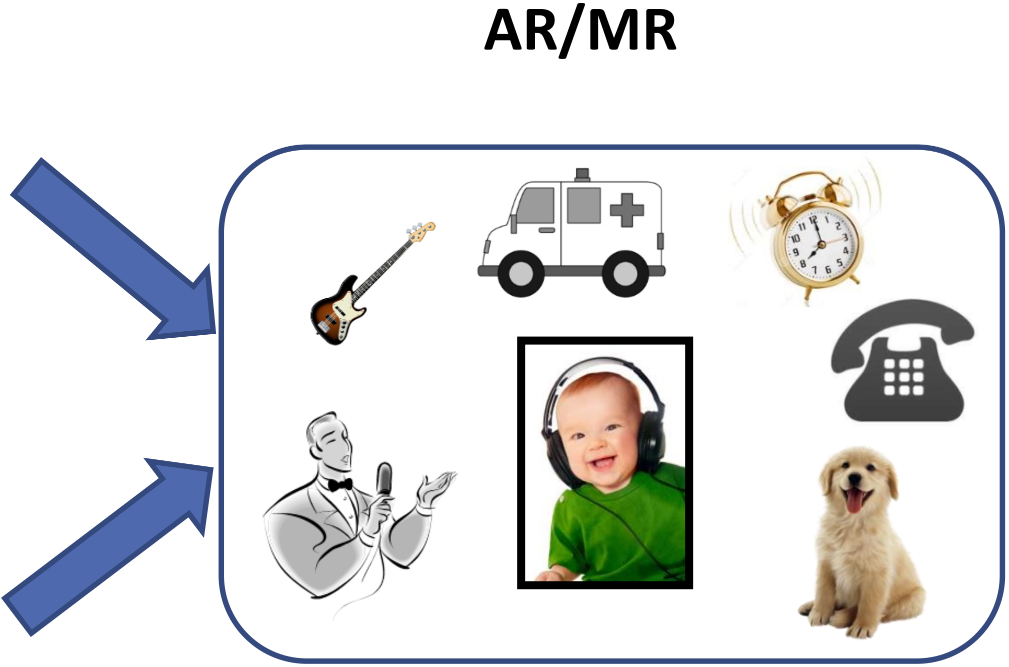
Transmission

Capture

Reproduction

Natural Listening in 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

Samsung VR Gear

HTC Vive

- With integrated headphones and built-in speakers

Samsung HMD Odyssey

HTC Vive Delux

Oculus Go

Microsoft Hololens

Wearable playback devices for VR, AR/MR

Closed Back Headphones - Good isolation & bass - Block off from environment	**Opened Back Headphones** - Less isolation & sound leakage - Good environmental awareness - spacious	**In-Ear Monitors** - Excellent isolation and frequency responses - Block off environmental noise	**Intra-concha / supra aural Earphones** - Poor isolation/ response variances - Some environmental awareness - Lightweight	**Built-in Speaker** - Poor isolation & bass - Leakage

- ***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?

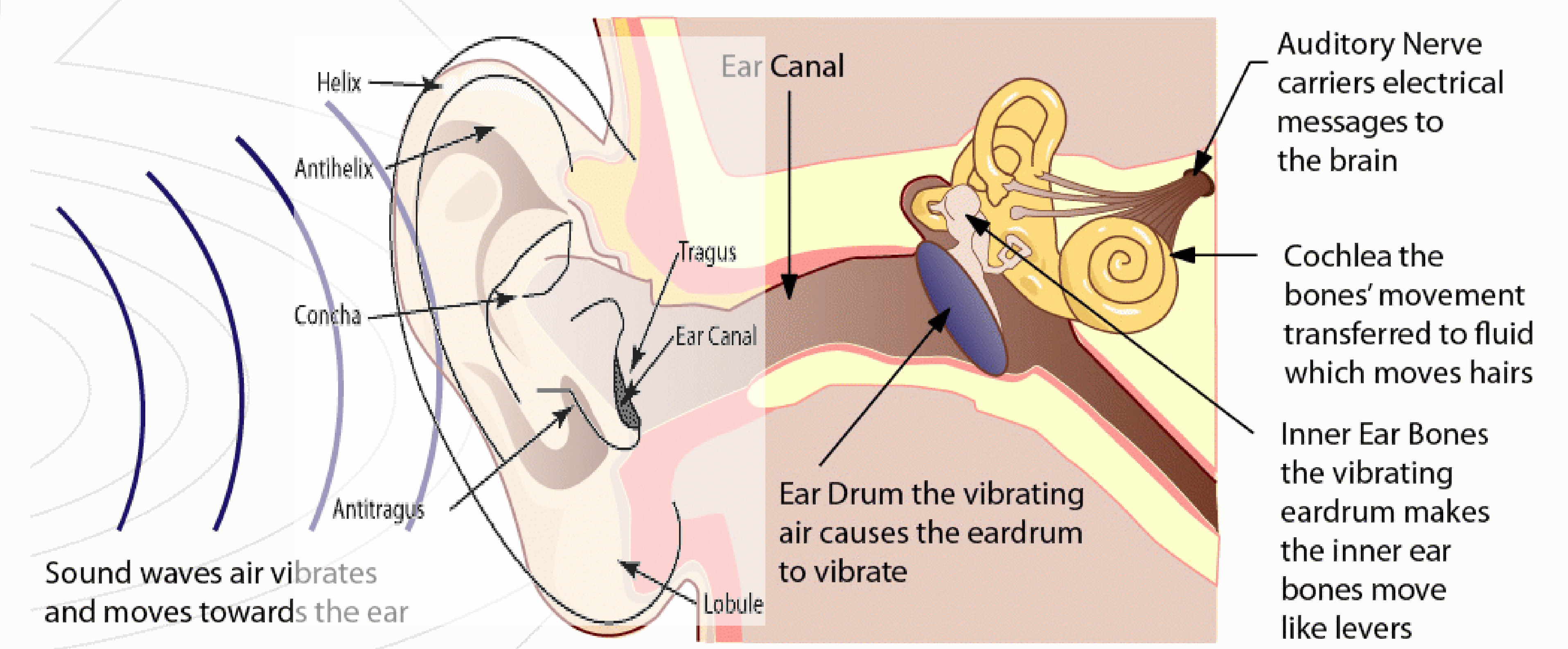

Helix
Antihelix
Concha
Antitragus
Tragus
Ear Canal
Lobule
Ear Canal
Sound waves air vibrates and moves towards the ear
Ear Drum the vibrating air causes the eardrum to vibrate
Auditory Nerve carriers electrical messages to the brain
Cochlea the bones' movement transferred to fluid which moves hairs
Inner Ear Bones the vibrating eardrum makes the inner ear bones move like levers

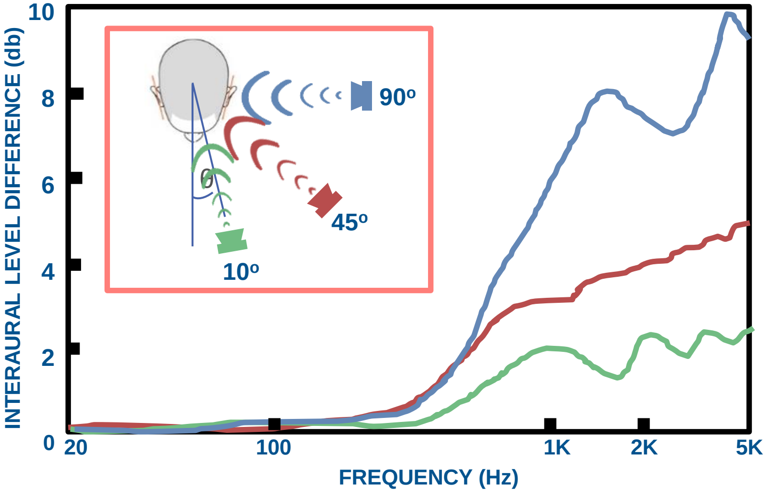
Image Source: http://www.soundproofingcompany.com/soundproofing101/what-is-sound/

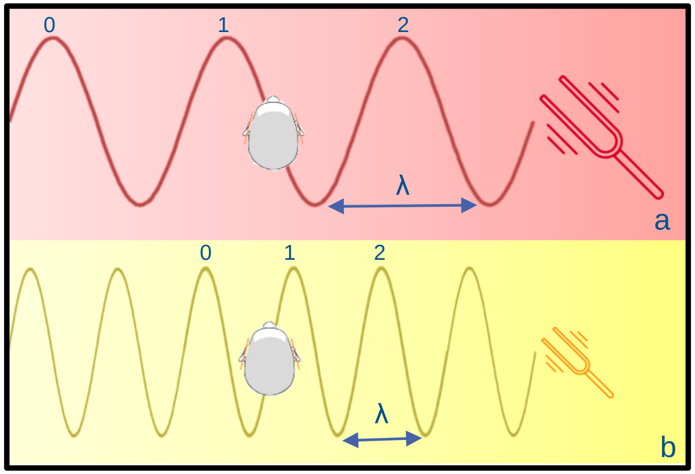
Primary Auditory Cues:

- 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

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 (˜22cm) > 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

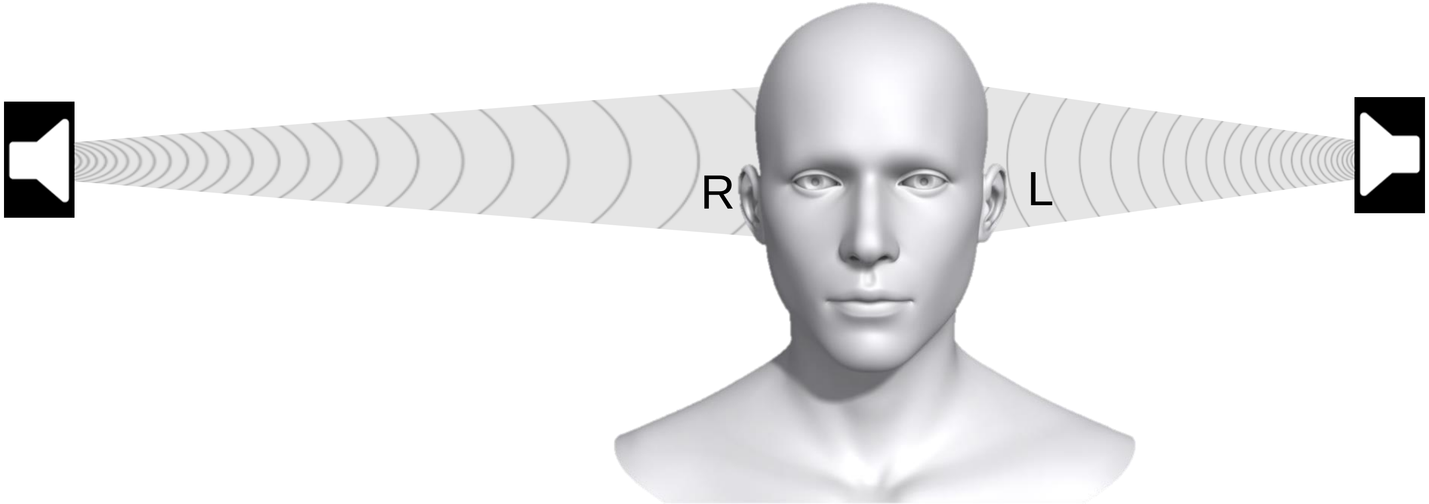


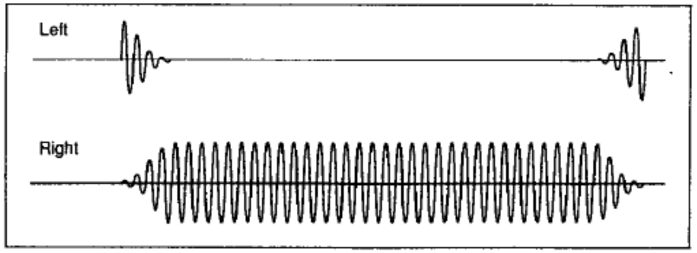


Pictures modified from [W. M. Hartmann, 1999]

Precedence (Law of 1st 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)

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

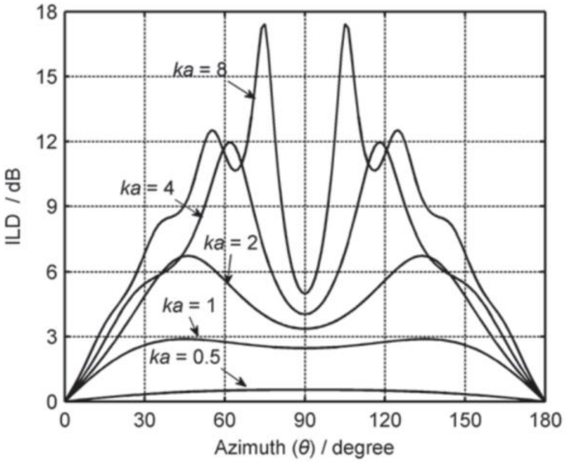
Equation for Interaural Level Difference (ILD)

- 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.

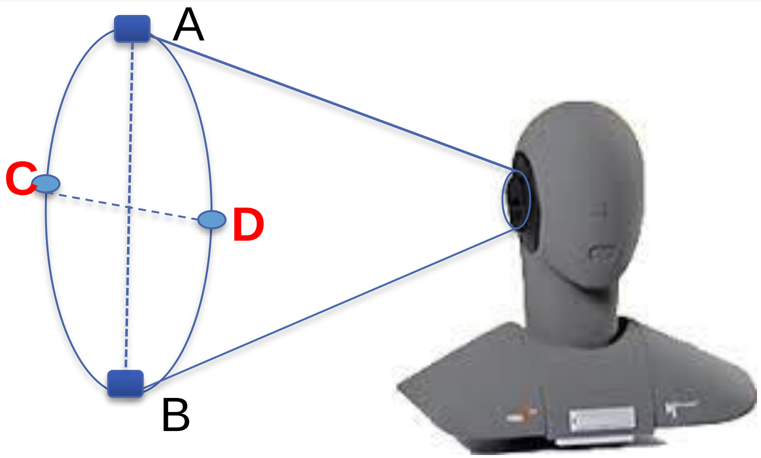


[Xie, 2013]

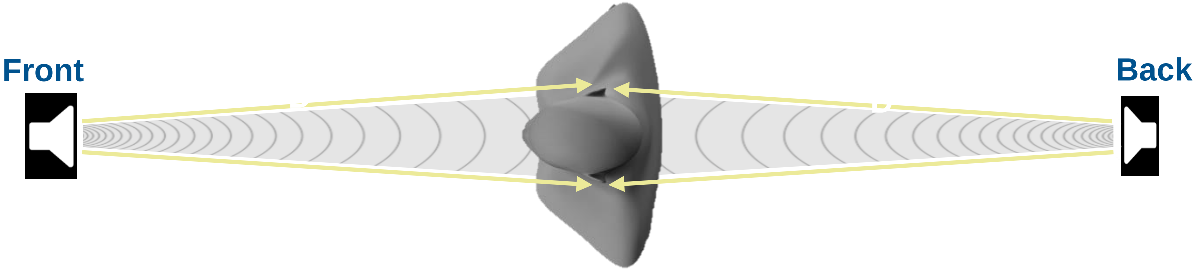
Cone of confusion and dynamic cues

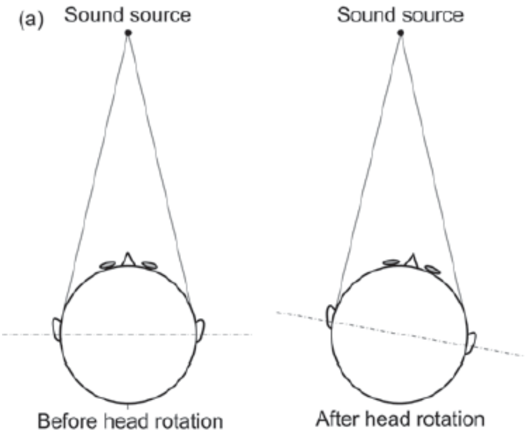
- **Similar ITD and ILD due to:**
  - **Cone of confusion**

    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?



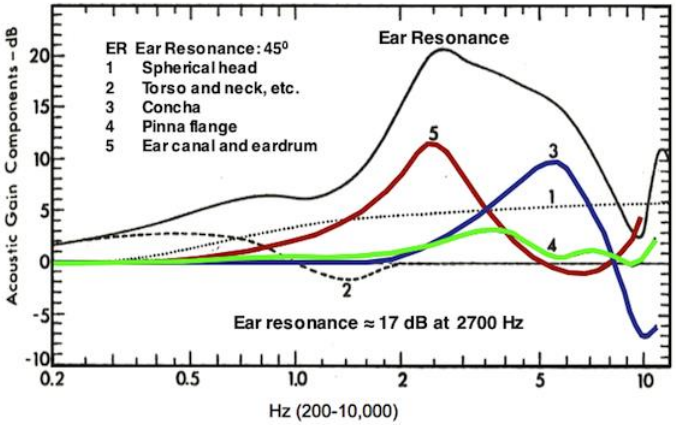


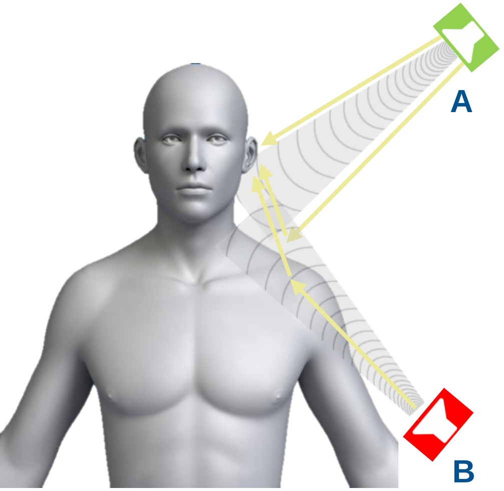
- **We need another sound localization cue!**

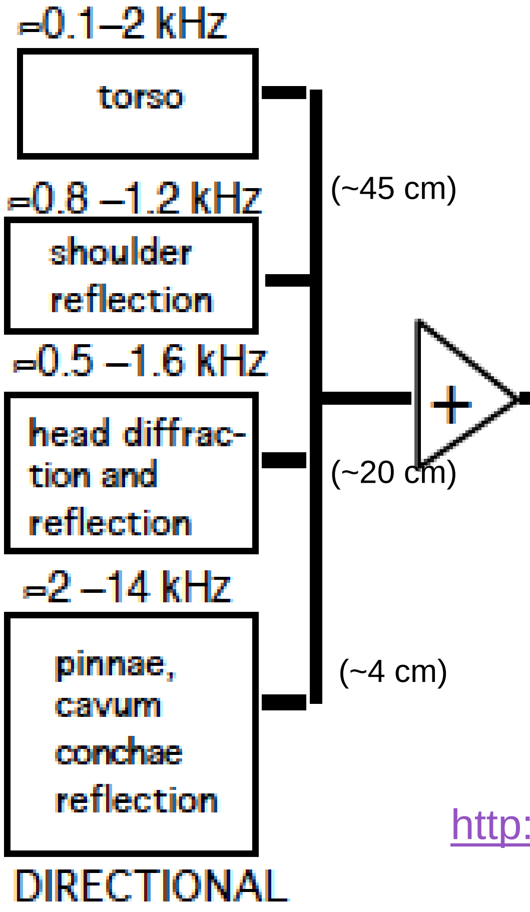
- **Head rotation as a dynamic cue can help resolve this**

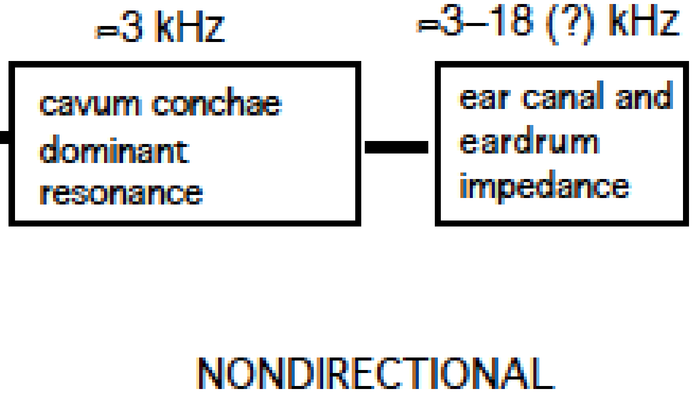
Modeling of Sound Scattering (Human body & ears)

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





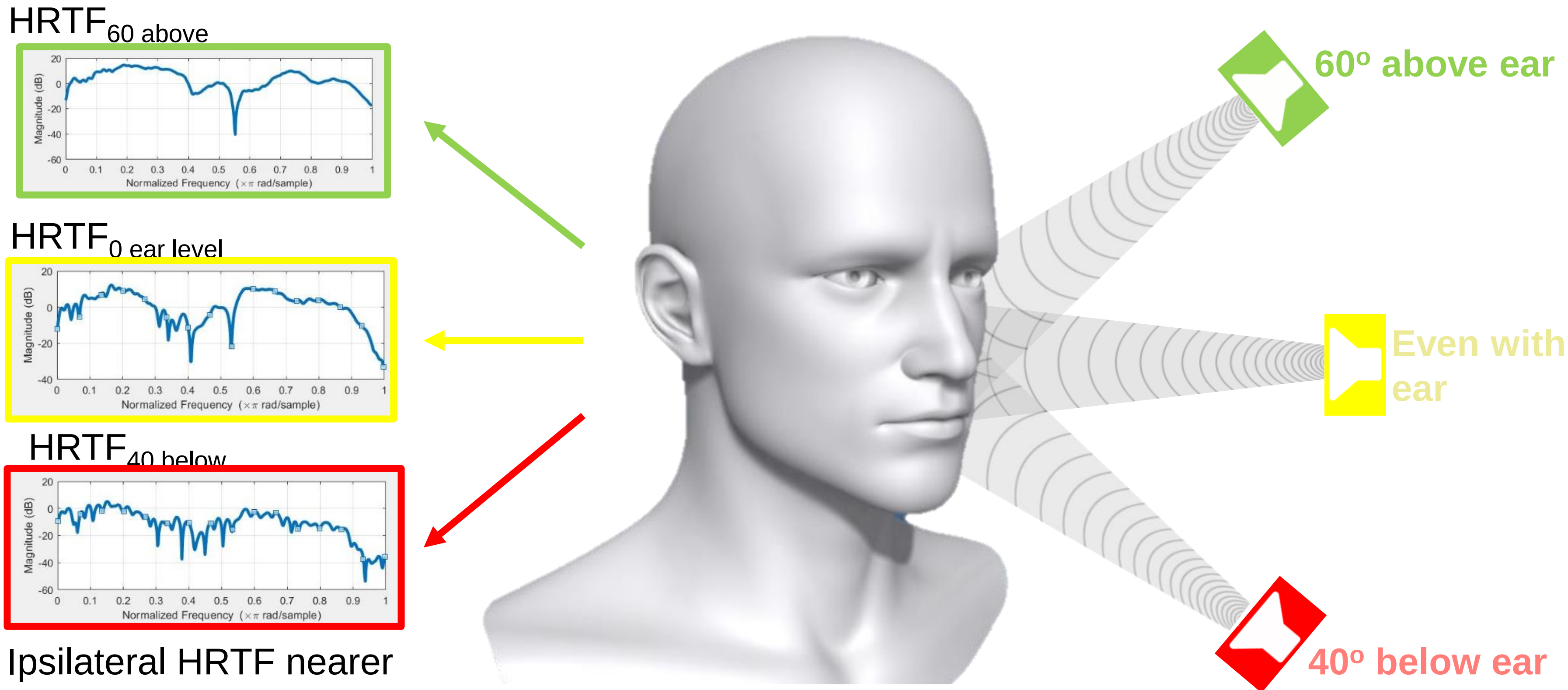



≈3 kHz
cavum conchae dominant resonance
≈3–18 (?) kHz
ear canal and eardrum impedance
NONDIRECTIONAL

Plot from 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?***

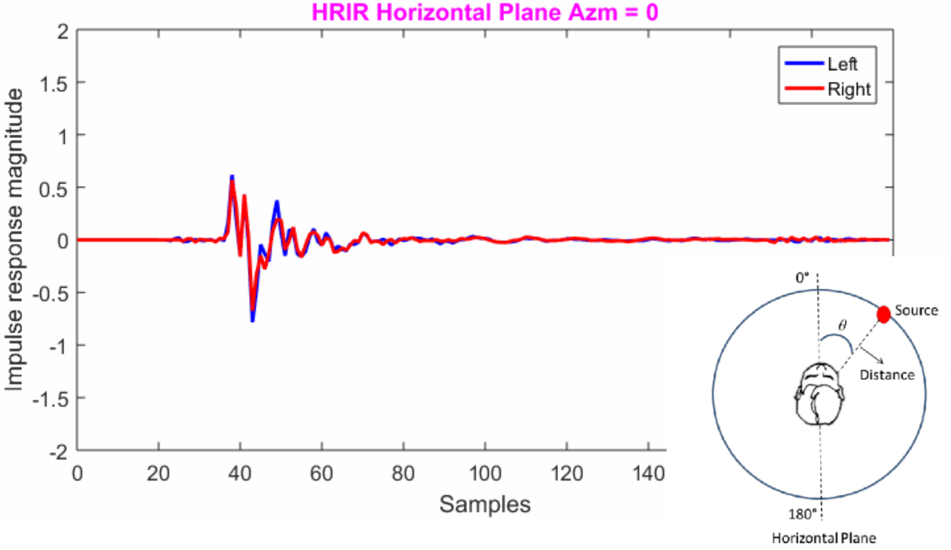


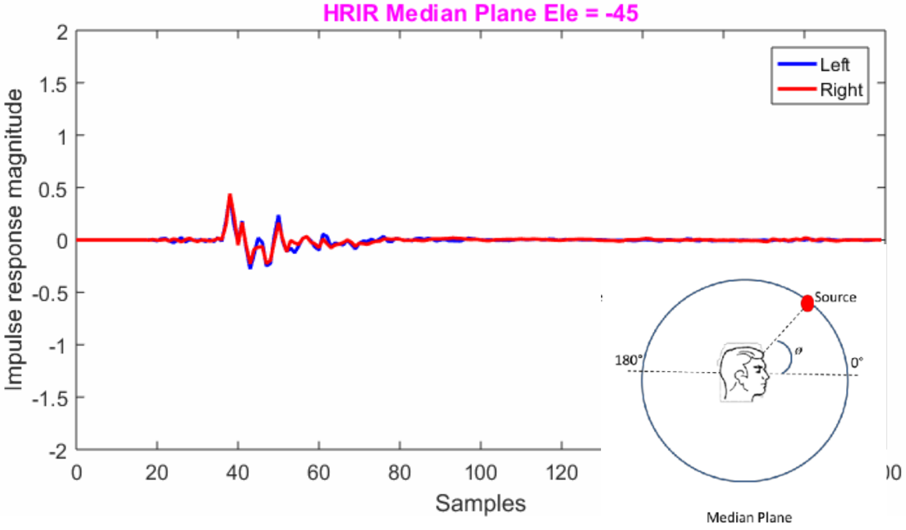
Ipsilateral HRTF nearer to source (shown above)

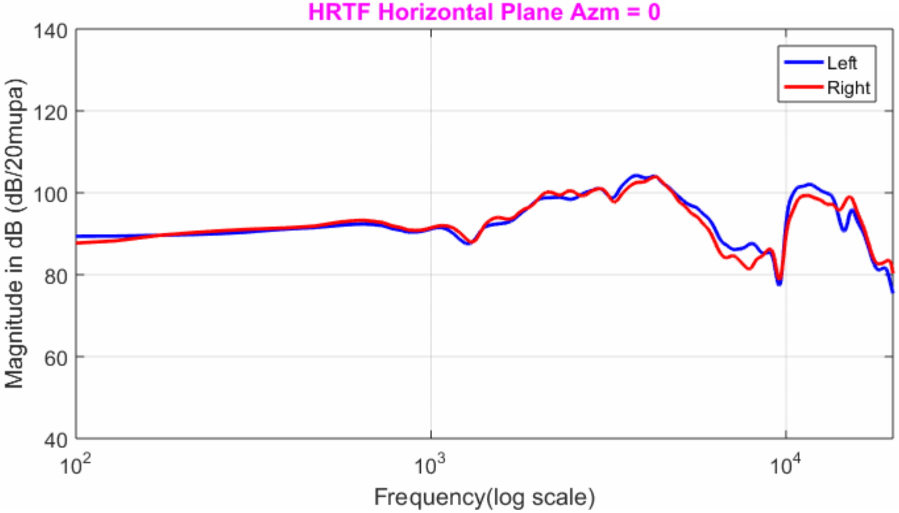
Animation of HRIR/HRTF: Database from CIPIC (S03)

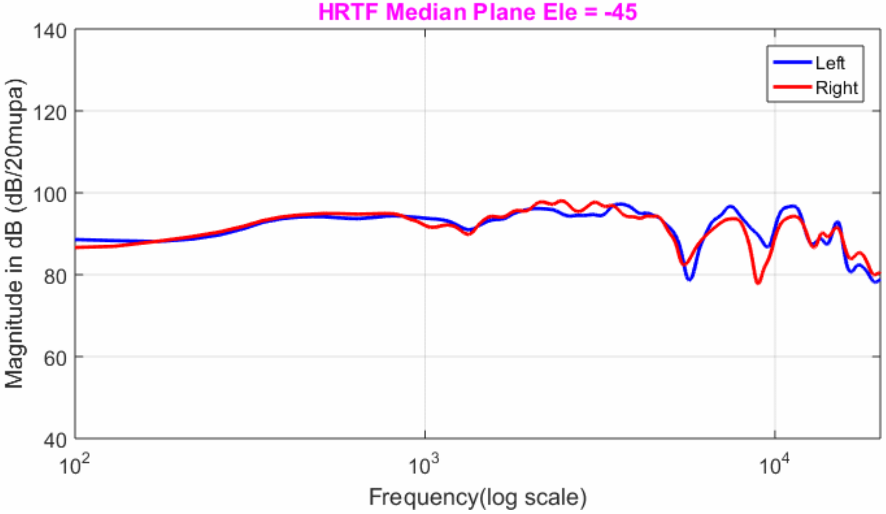
Azimuth

Elevation


HRIR Horizontal Plane Azm = 0

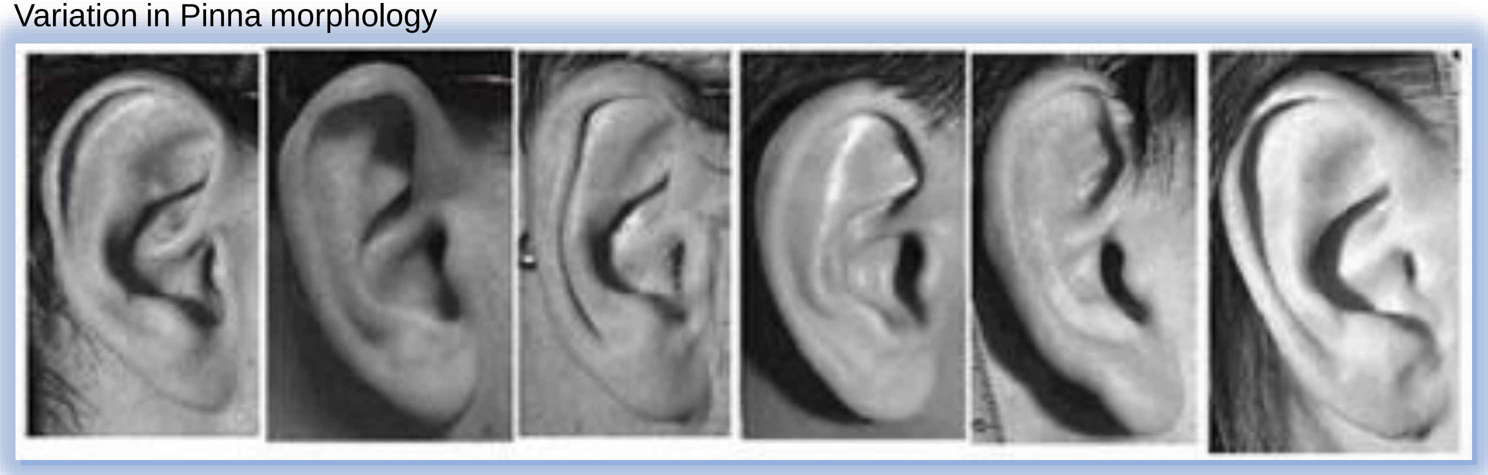

HRIR Median Plane Ele = -45


HRTF Horizontal Plane Azm = 0


HRTF Median Plane Ele = -45

Individual Sound Filtering (Earprint)

Variation in Pinna morphology

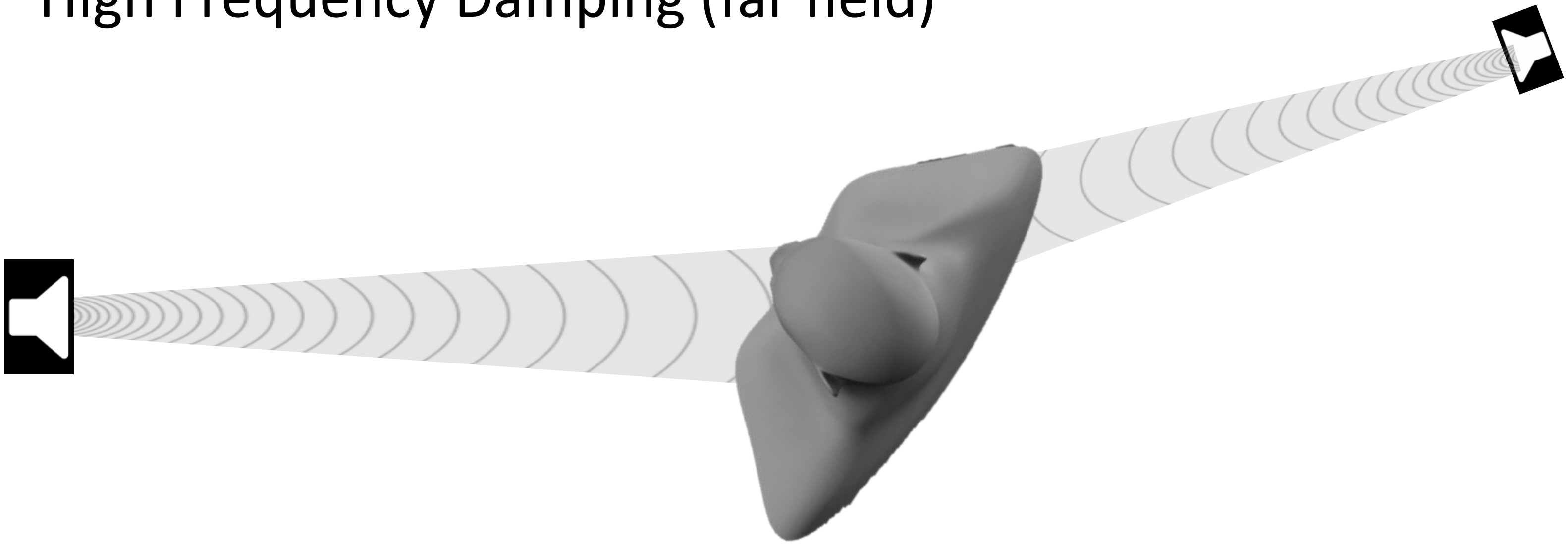


Pinna of human subjects taken from the CIPIC database

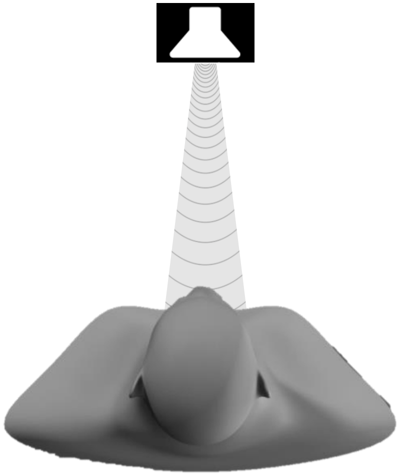
- Human pinna is found to be as **idiosyncratic as the fingerprint**
- Scattering wave around ears are different.
- HRTFs are highly individual and differs substantially from one subject to the other.
- For perfect 3D audio playback, **individualized recordings/HRTFs** and **individualized headphone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

Distance Localization Cue

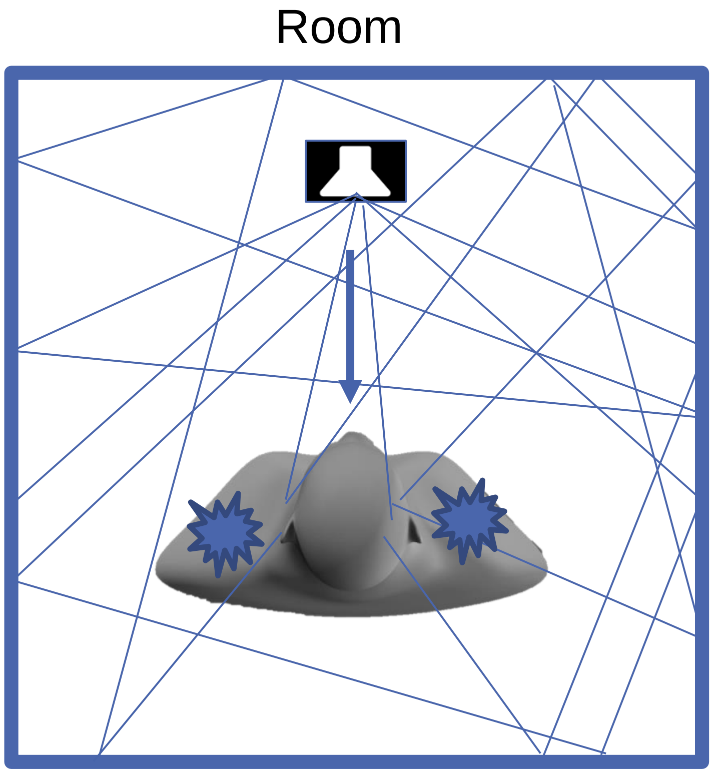
- Loudness
  - Familiar sound sources
  - Moving sound sources
- Initial Time Delay
- Ratio of Direct and Reverberant energy
- Motion Parallax (near field)
- ILD (near field)
- High Frequency Damping (far field)



Reverberation



Spatialization (Anechoic)
only solves direct sound propagation



Reverberation (Ambience)
Provides indirect audio cues

Design of Spatial Audio Rendering System

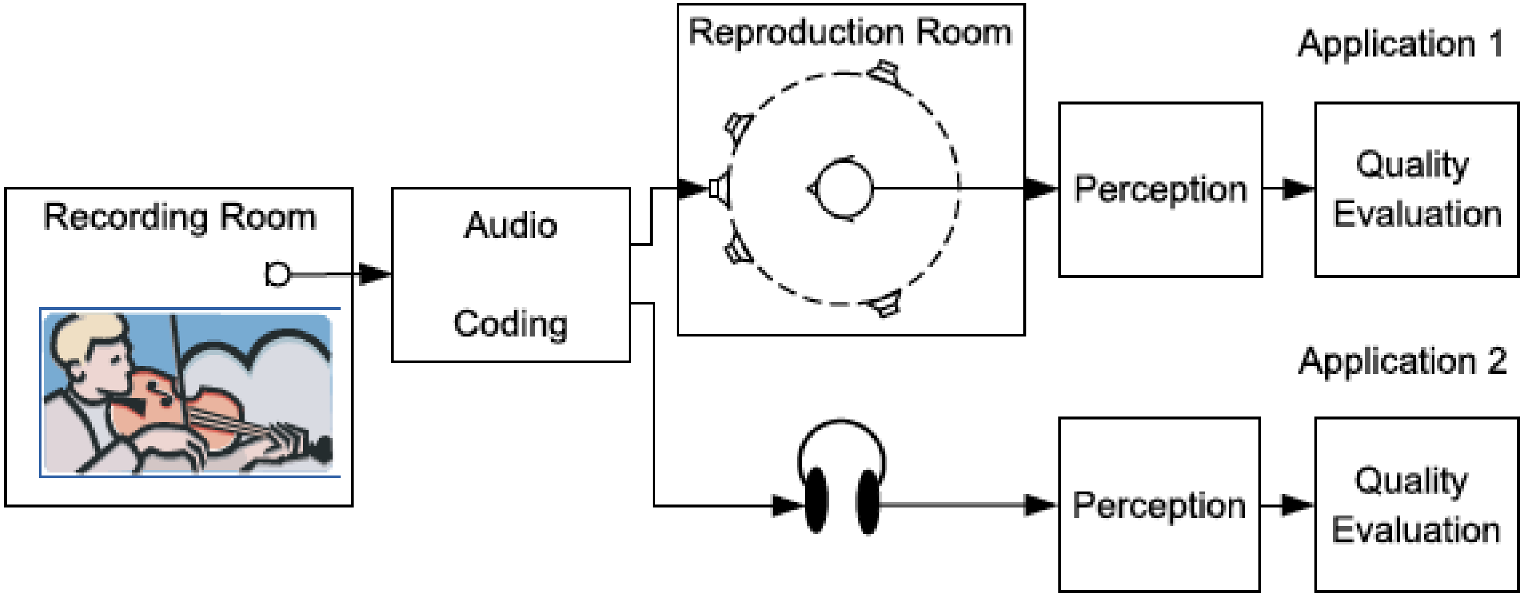
- Spatial audio rendering is concerned with the linking of physics and auditory perceptual effects.
- Not to overly rely on complex mathematical tools; just a tool for analysis.
- A highly accurate design of spatial audio processing system may not be required for plausible perceptual performance.
- Allow 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

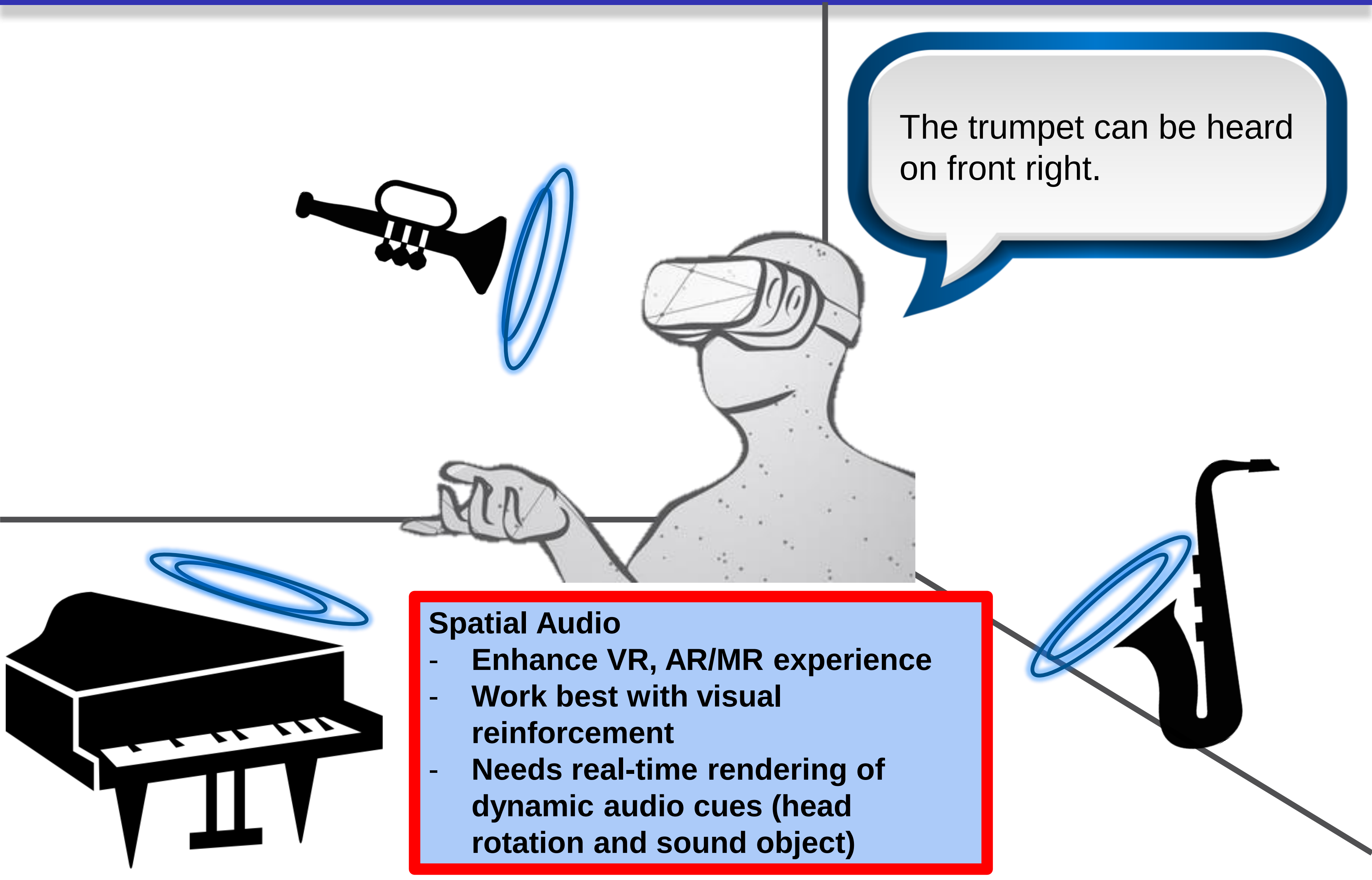
[Schoeffler et al.,2015]

Key Standards/tests for evaluation of 3D sound perception

S. No.	Standard name/test	Title	Remarks
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 **MU**lti-**S**timuli test with **H**idden **R**eference and **A**nchor (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 360º SPATIAL AUDIO



Spatial Audio

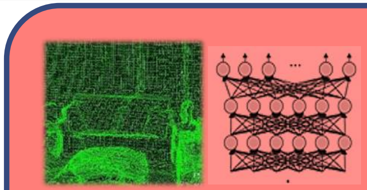
- **Enhance VR, AR/MR experience**
- **Work best with visual reinforcement**
- **Needs real-time rendering of dynamic audio cues (head rotation and sound object)**

Spatial Audio Technologies for Immersive VR/AR/MR

Module B topics

Spatial Audio Formats

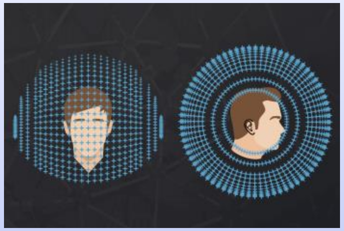
- Object, Ambisonics
- Parametric processing



Environment Estimation

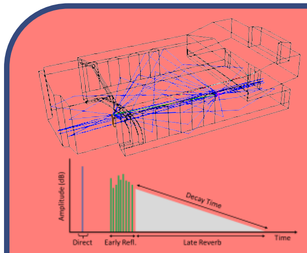
- Depth camera
- Reverberation fingerprint
- Machine learning

Module C topics



Individualized Binaural Rendering

- Individualized HRTFs
- Equalization



Environment Rendering

- Wave based
- Geometrical based
- Perceptual based

Y AXIS YAW
X AXIS PITCH
Z AXIS ROLL
Y AXIS UP/DOWN
X AXIS LEFT/RIGHT
Z AXIS FRONT/BACK

Dynamic Binaural Synthesis

- Head tracking
- Position tracking

Virtual & Physical Sound Fusion

- Adaptive equalization
- Hear-through processing

References in Module A

- Minnaar, P., Plogsties, J., Olesen, S.K., Christensen, F. and Møller, H., The interaural time difference in binaural synthesis. In *Audio Engineering Society Convention 108*, 2000
- Xie, B., *Head-related transfer function and virtual auditory display*. J. Ross Publishing, 2013
- Katz, B.F. and Noisternig, M., A comparative study of interaural time delay estimation methods. *The Journal of the Acoustical Society of America*, *135*(6), pp.3530-3540, 2014.
- Rumsey F et. al., “On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality,” Journal of the Acoustical Society of America 118, 968 ,2005.
- Schoeffler, M., Silzle, A. and Herre, J.,. Evaluation of Spatial/3D Audio: Basic Audio Quality Versus Quality of Experience. *IEEE Journal of Selected Topics in Signal Processing*, *11*(1), pp.75-88, 2017.
- Bech, S. and Zacharov, N., *Perceptual audio evaluation-Theory, method and application*. John Wiley & Sons. 2007.
- Begault D.R, *3-D Sound for Virtual Reality and Multimedia*, Academic Press, 1994.
- Blauert J Ed, The Theory of Binaural Listening, Springer, 2013.
- Lyon, Human and Machine Hearing- Extracting Meaning from Sound, Cambridge University Press, 2018.
- W. M. Hartmann, “How we localize sound,” Physics Today, 52(11), 24,1999.

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

Medium

Playback system

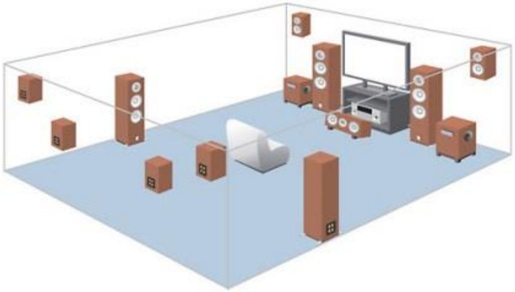
Receiver

Human

Channel

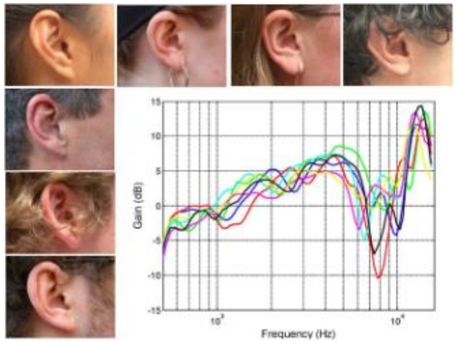
Object

Scene

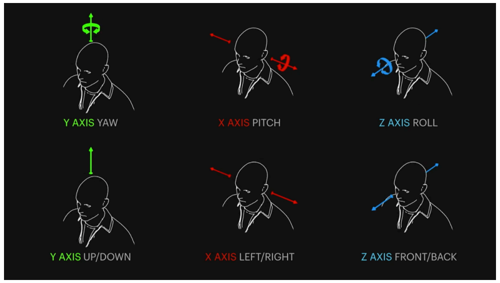


Loudspeakers

Headphones

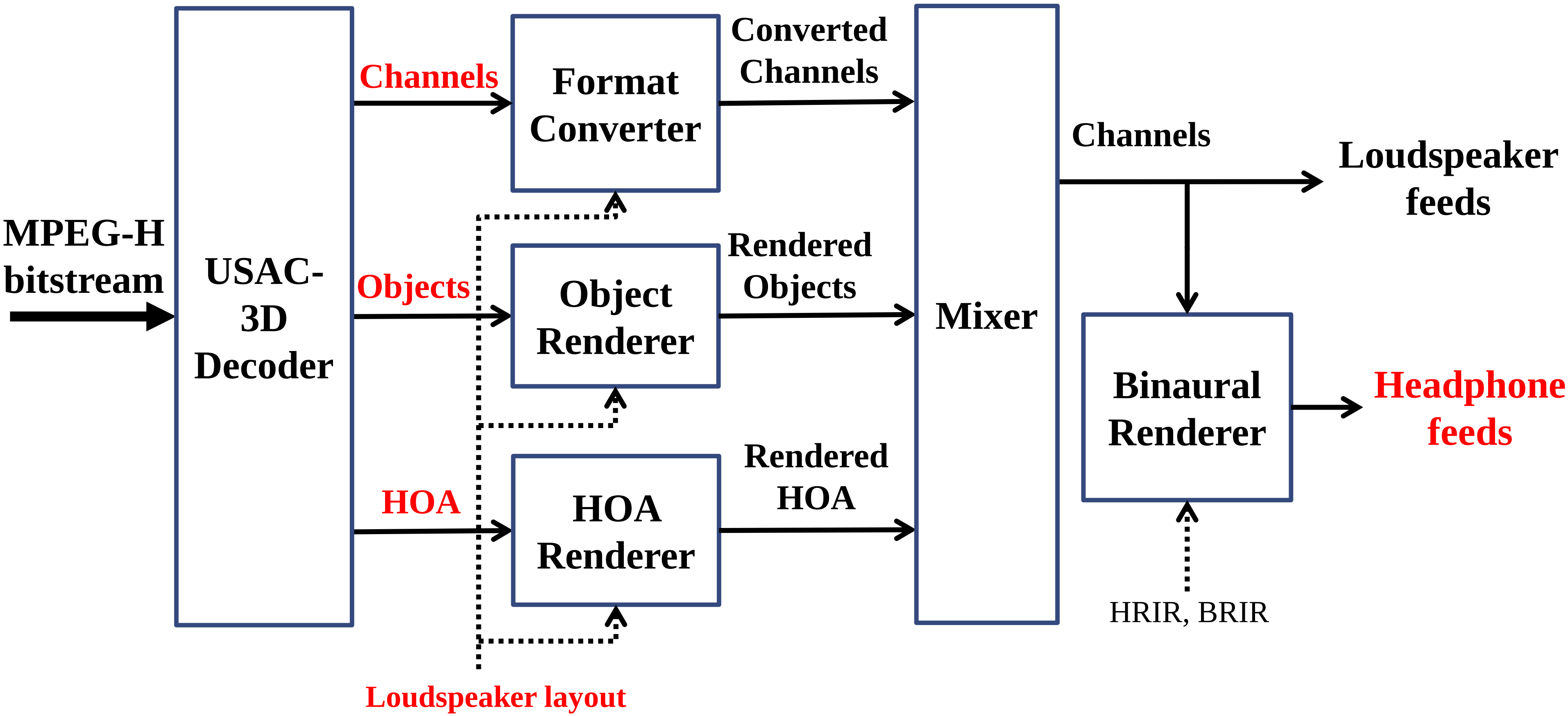


Ears



Movements

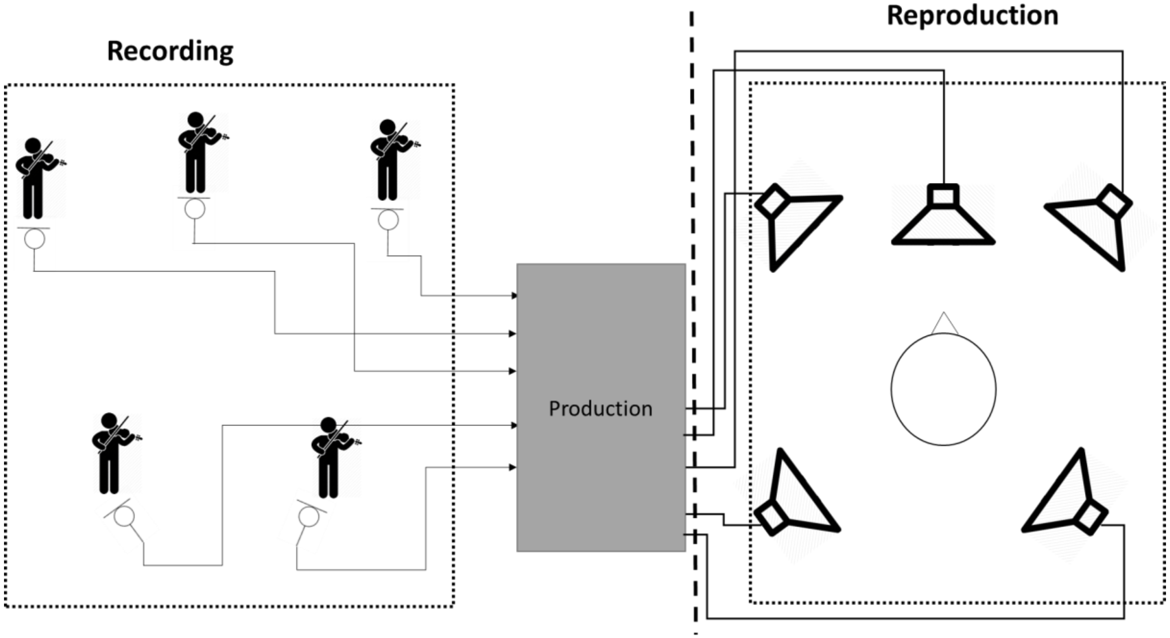
MPEG-H 3D audio standard (2015)



[Herre, 2013]

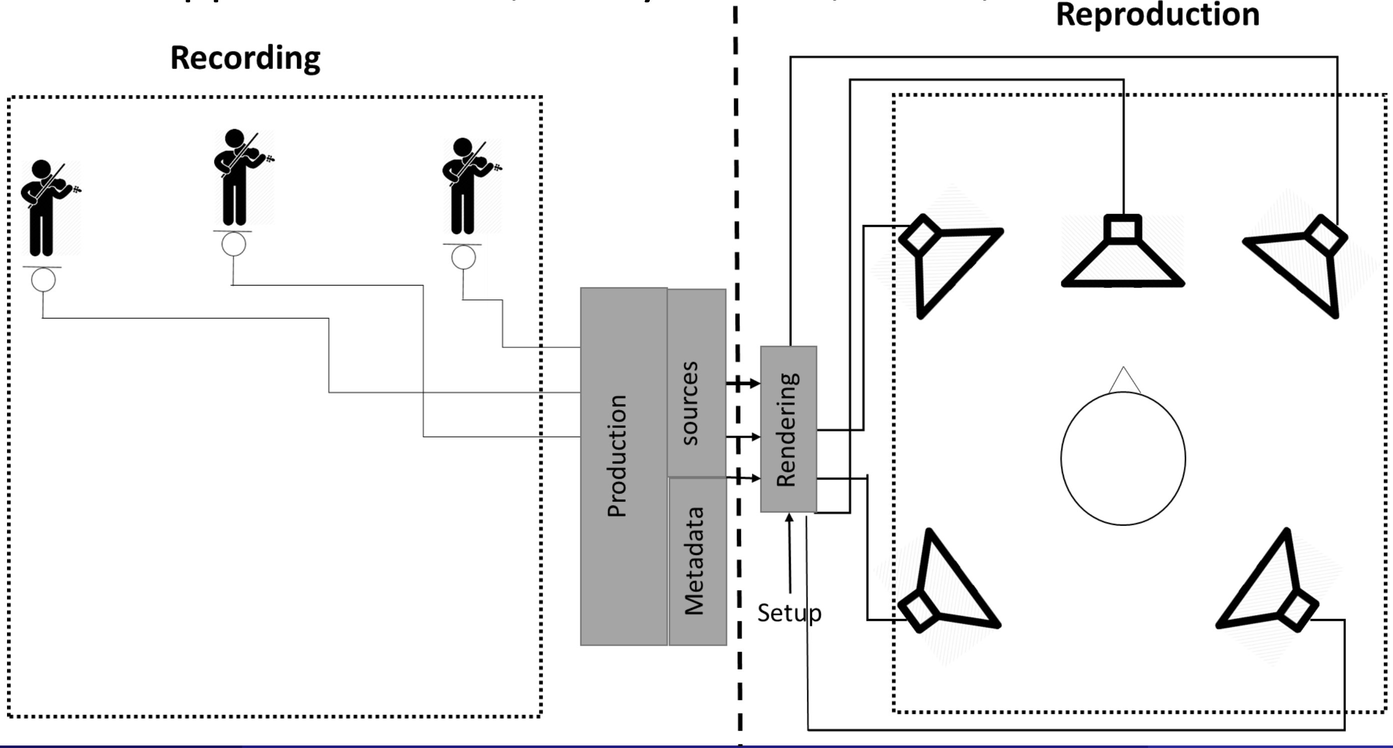
Channel-based audio

- 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

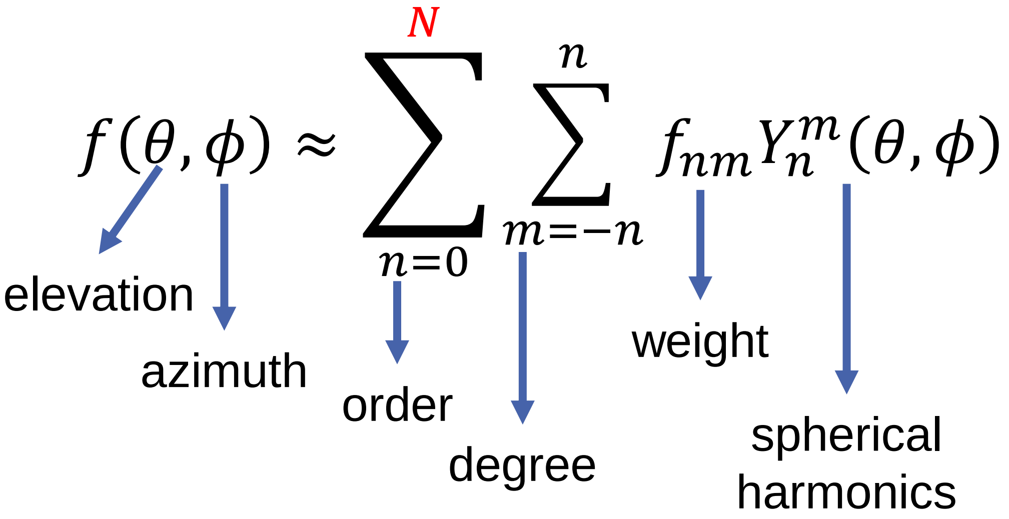


Scene-based audio: ambisonics basics

- 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



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

[Rafaely, 2015]

Spherical harmonics

$$\boxed{Y_n^m(\theta,\phi)=\sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}P_n^m(\cos\theta)e^{im\phi}}$$

Y_0^0

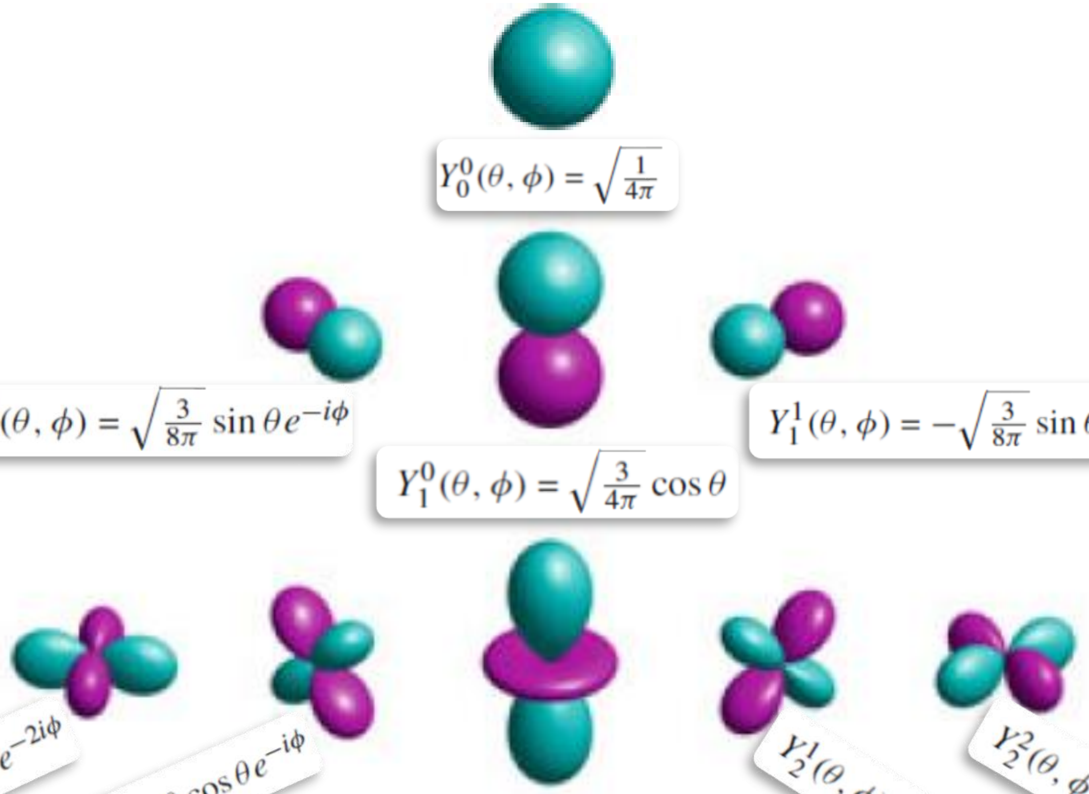
$Im\{Y_1^{-1}\}$ Y_1^0 $Re\{Y_1^1\}$

$Im\{Y_2^{-2}\}$ $Im\{Y_2^{-1}\}$ Y_2^0 $Re\{Y_2^1\}$ $Re\{Y_2^2\}$

0th order ***n*** **= 0**

1st order ***n*** **= 1**

2nd order ***n*** **= 2**



$Y_0^0(\theta,\phi)=\sqrt{\frac{1}{4\pi}}$

$Y_1^{-1}(\theta,\phi)=\sqrt{\frac{3}{8\pi}}\sin\theta e^{-i\phi}$

$Y_1^0(\theta,\phi)=\sqrt{\frac{3}{4\pi}}\cos\theta$

$Y_1^1(\theta,\phi)=-\sqrt{\frac{3}{8\pi}}\sin\theta e^{i\phi}$

$Y_2^{-2}(\theta,\phi)=\sqrt{\frac{15}{32\pi}}\sin^2\theta e^{-2i\phi}$

$Y_2^{-1}(\theta,\phi)=\sqrt{\frac{15}{8\pi}}\sin\theta\cos\theta e^{-i\phi}$

$Y_2^0(\theta,\phi)=\sqrt{\frac{5}{16\pi}}(3\cos^2\theta-1)$

$Y_2^1(\theta,\phi)=-\sqrt{\frac{15}{8\pi}}\sin\theta\cos\theta e^{i\phi}$

$Y_2^2(\theta,\phi)=\sqrt{\frac{15}{32\pi}}\sin^2\theta e^{2i\phi}$

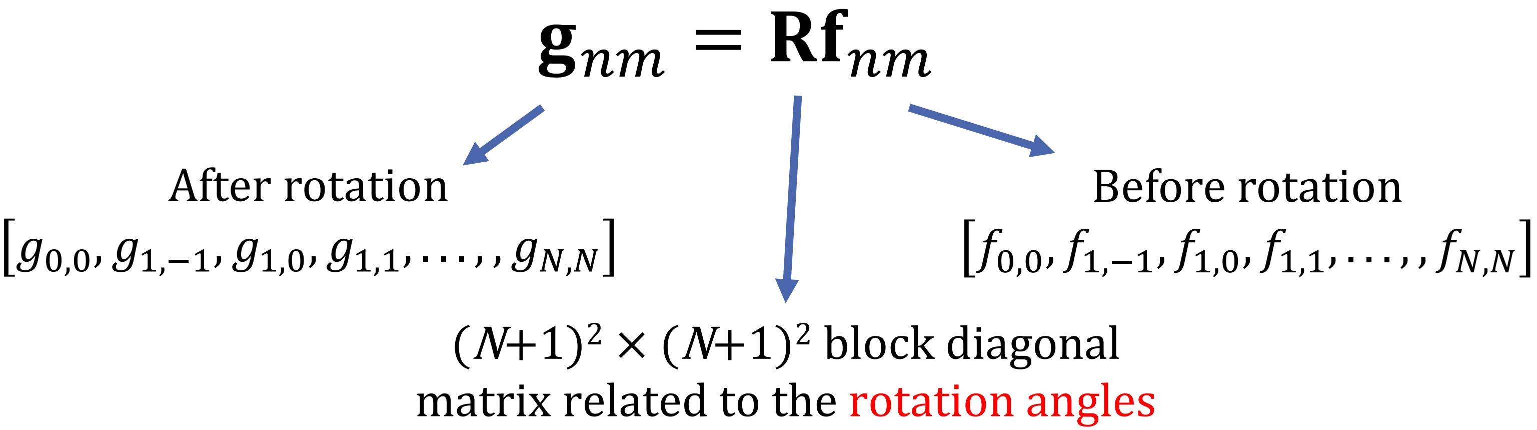
Spherical harmonic weights

- **Spherical harmonic weights**

$$f_{nm} = \int_0^{2\pi}\int_0^{\pi} f(\theta,\phi)\left[Y_n^m(\theta,\phi)\right]^* \sin\theta d\theta d\phi$$

→ what the ambisonics microphone records directly or indirectly

- **Rotation in spherical harmonic domain**



$$\mathbf{g}_{nm} = \mathbf{R}\mathbf{f}_{nm}$$

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

Loudspeaker signals

Ambisonic signals

For any layouts

For **regular** layouts

$$\mathbf{C}\mathbf{s}(t) = \mathbf{A}(t) \rightarrow \mathbf{s}(t) = \left(\mathbf{C}^{\mathrm{T}}\mathbf{C}\right)^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{A}(t) = \frac{1}{J}\mathbf{C}^{\mathrm{T}}\mathbf{A}(t)$$

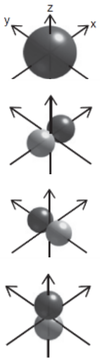
Encoding spherical harmonic matrix related to loudspeaker positions

- 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

[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 $\mathrm{Y} = \frac{1}{K}\sum_{k=1}^{K} s_k\left[\sin\phi_k \cos\theta_k\right]$

z-directional $Z = \frac{1}{K}\sum_{k=1}^{K} s_k\left[\sin\theta_k\right]$

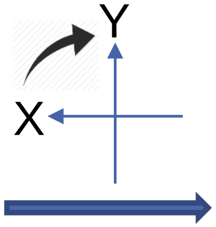
- Reproduction/decoding to regular layout

Loudspeaker signal $p_j = \frac{1}{J}\begin{bmatrix} W & X & Y & Z \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \cos\phi_j \cos\theta_j \\ \sin\phi_j \cos\theta_j \\ \sin\theta_j \end{bmatrix}$

- Rotation (e.g., azimuth rotation by θ)

$$\begin{bmatrix} W' \\ X' \\ Y' \\ Z' \end{bmatrix} = R\begin{bmatrix} W \\ X \\ Y \\ Z \end{bmatrix} \qquad R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & 0 \\ 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R\,|_{\theta=90^\circ} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

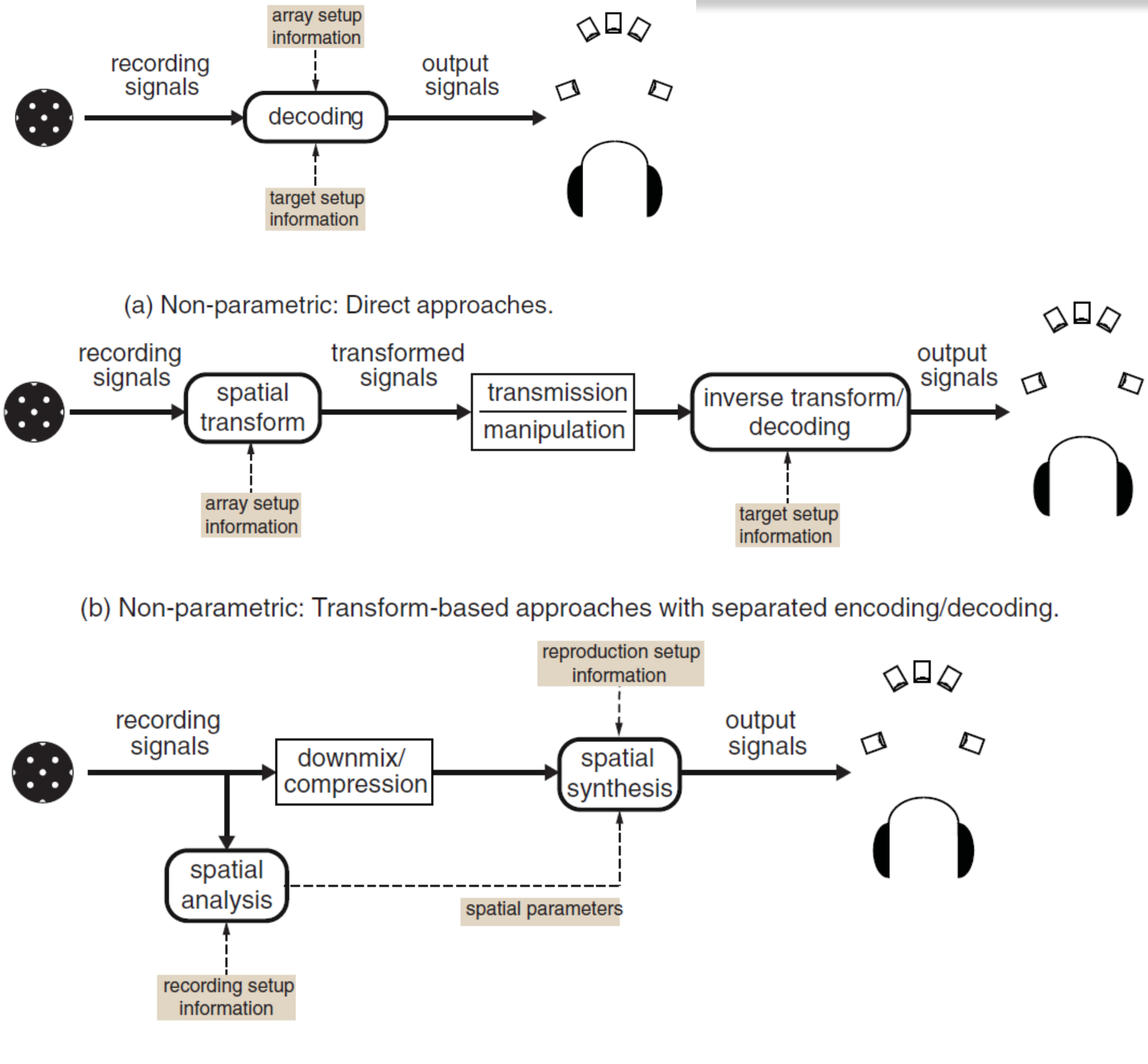


$$\begin{bmatrix} W' \\ X' \\ Y' \\ Z' \end{bmatrix} = \begin{bmatrix} W \\ -Y \\ X \\ Z \end{bmatrix}$$

An overview and comparison

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

Parametric spatial audio processing (PSAP)



(a) Non-parametric: Direct approaches.

(b) Non-parametric: Transform-based approaches with separated encoding/decoding.

(c) Parametric approaches.

Characteristics:

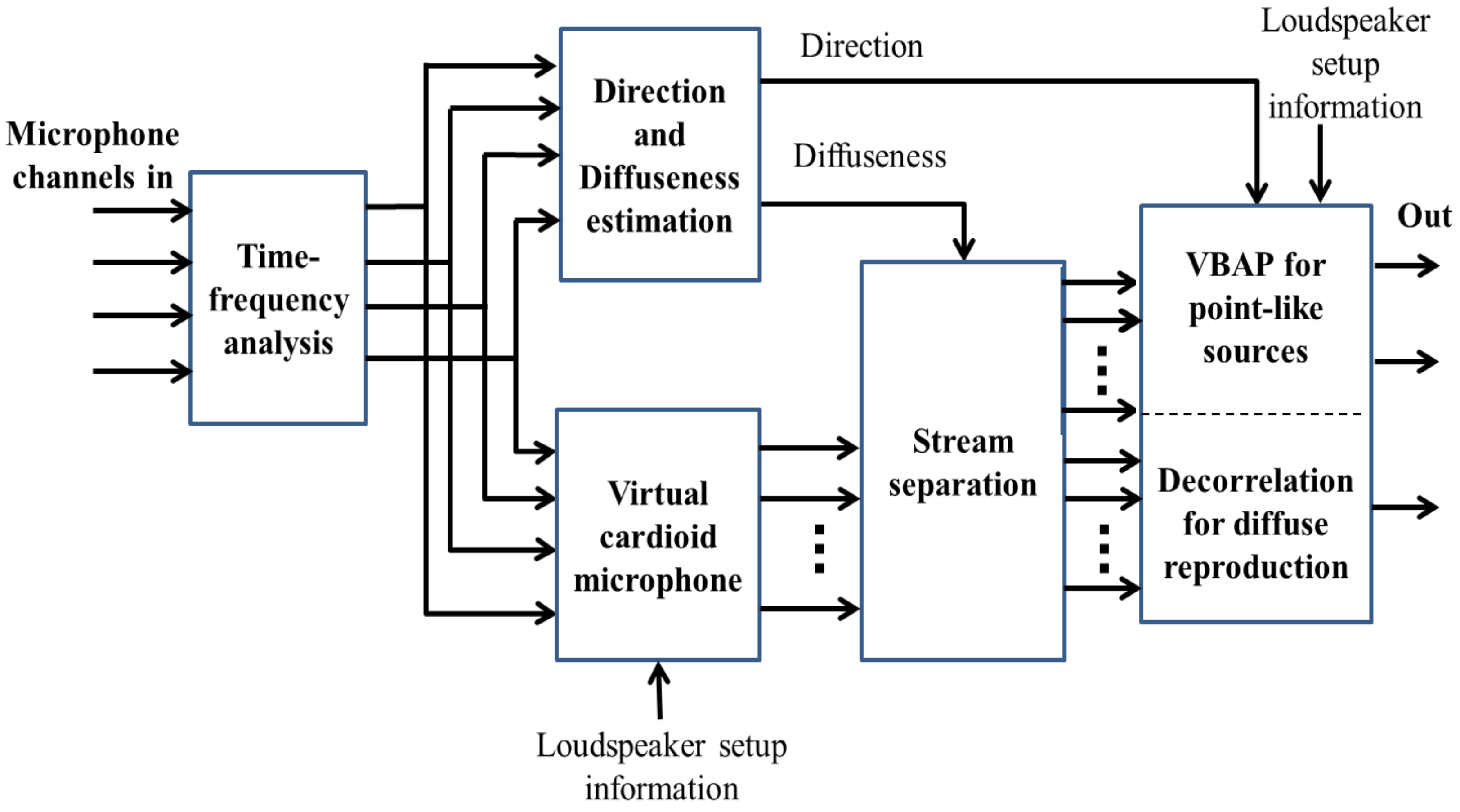
- **Flexible**
- **Effective**
- **Efficient**

[Pulkki, 2018]

Parametric sound field/scene models

Audio content type	Directional sound	Diffuse sound	Parameters	Related techniques
Channel-based	One/ multiple	Yes	ICTD, ICLD	Spatial audio scene coding (SASC) / Primary-ambient extraction (PAE)
	Multiple	No	Azimuth, (elevation)	Blind source separation (BSS)
	Multiple	No	ICTD, ICLD, ICC	MPEG Spatial audio Coding (SAC)
Object-based	Multiple	No	Azimuth, elevation, (distance)	MPEG Spatial audio object coding (SAOC)
Scene-based (ambisonics)	Multiple	Yes	Azimuth, elevation, Diffuseness	Directional audio coding (DirAC)
Scene-based (mic array)	Multiple	No	Azimuth, (elevation)	BSS
	Multiple	Yes	Azimuth, elevation	Spatial filtering

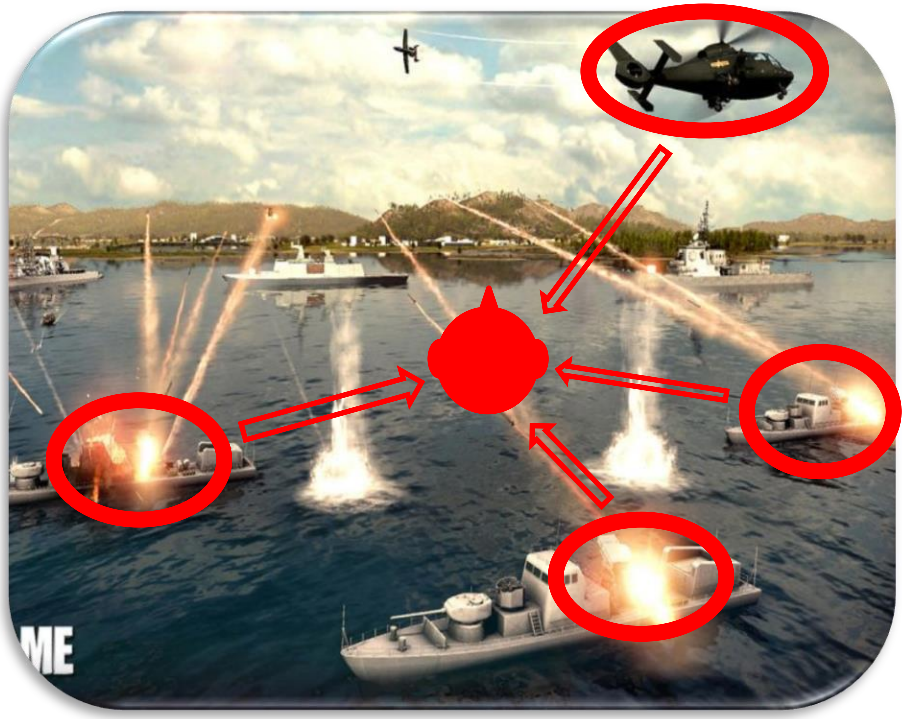
PSAP example: Directional audio coding (DirAC)


Microphone channels in
Time-frequency analysis
Direction and Diffuseness estimation
Direction
Diffuseness
Loudspeaker setup information
Virtual cardioid microphone
Loudspeaker setup information
Stream separation
VBAP for point-like sources
Decorrelation for diffuse reproduction
Out

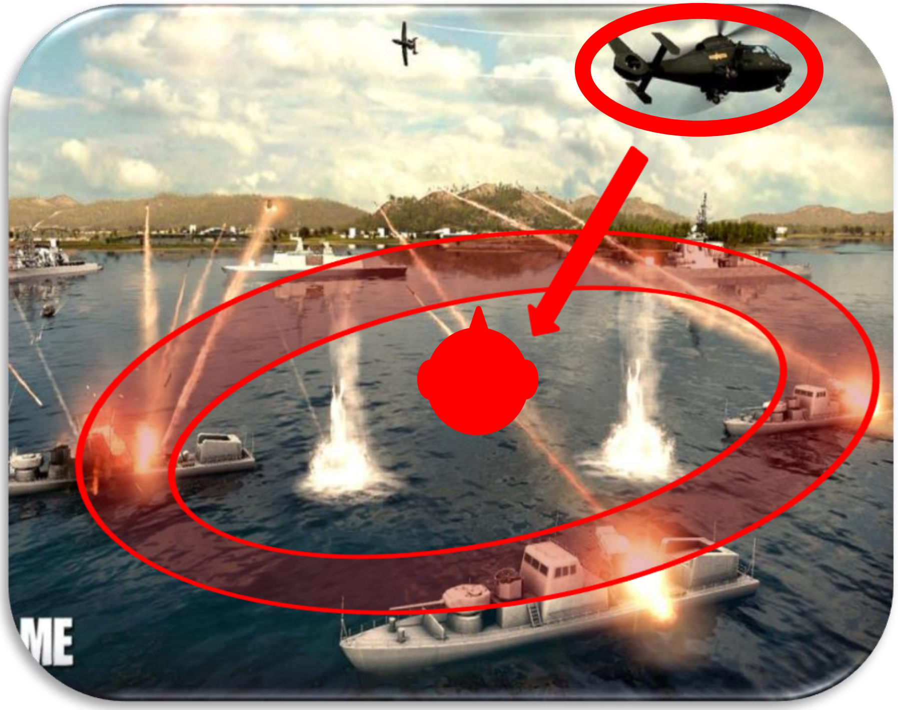
[Pulkki, 2007]

PSAP example: sound scene decomposition

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



“Sum of sources”



“Sum of primary and ambient components”

Blind source separation

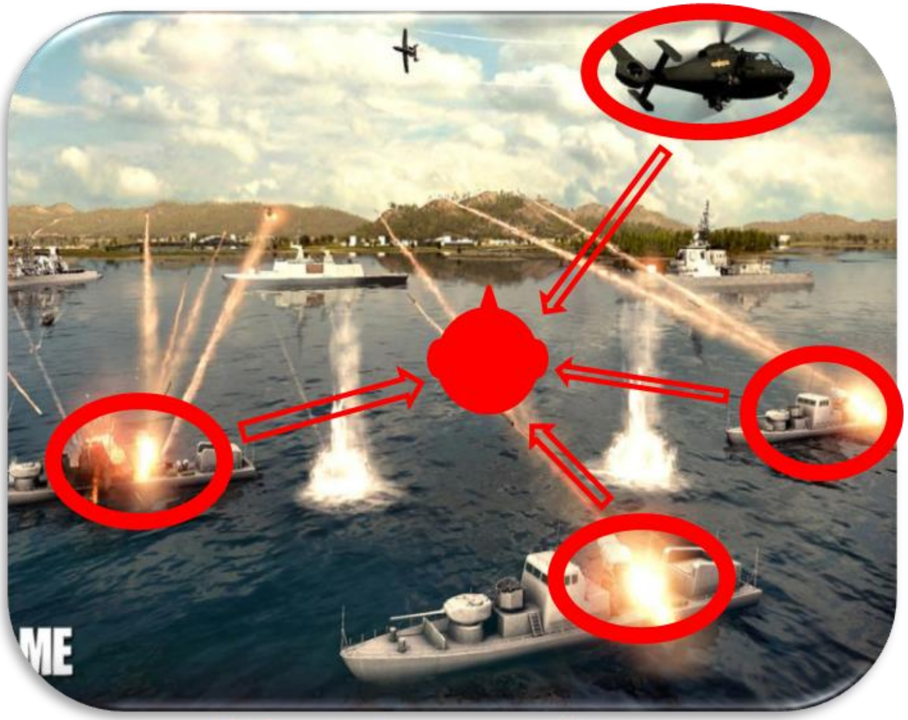
Primary ambient extraction

[Sunder, 2015]

https://www.vg247.com/2014/04/10/wargame-red-dragon-screenshots-show-off-warships-in-the-rts/

Sound scene decomposition: BSS

Objective:
to extract the K sources from M mixtures



"Sum of sources"

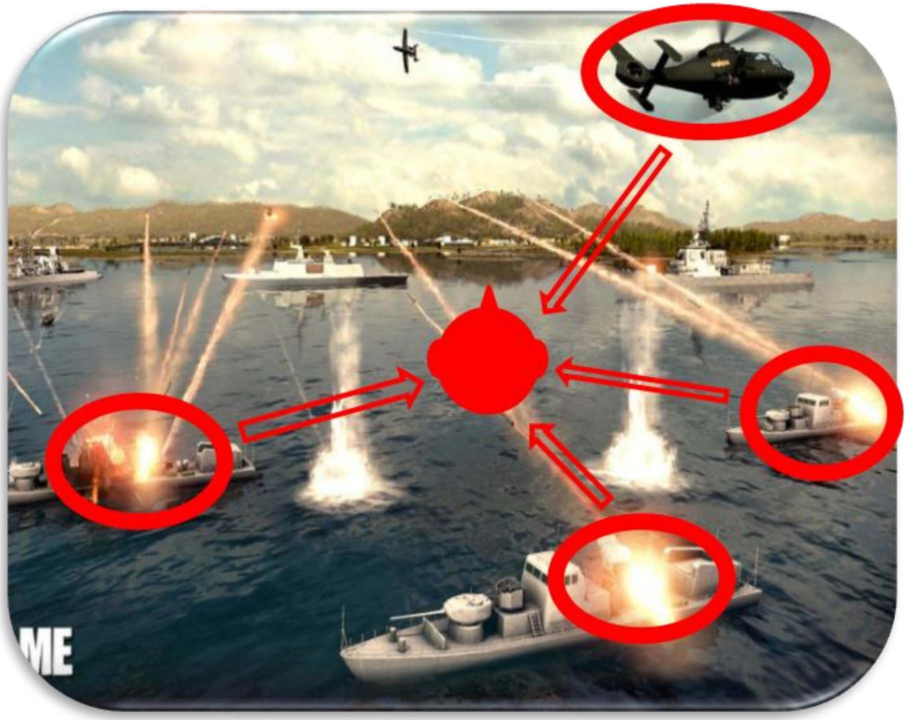
Mixtures = function (gain, source, time difference, model error)

$$x_m(n) = \sum_{k=1}^{K} g_{mk} s_k(n - \tau_{mk}) + e_m(n), \qquad \forall m \in \{1, 2, \ldots, M\}$$

Sound scene decomposition: BSS

Objective:
to extract the K sources from M mixtures

Case		Typical techniques
M = K		ICA
M > K		ICA with PCA, Least-squares
M < K	M > 2	ICA with sparse solutions
	M = 2	Time-frequency masking
	M = 1	NMF, CASA



"Sum of sources"

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

Sound scene decomposition: PAE

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

"Sum of primary and ambient components"

Mixtures = primary component + ambient component

$$x_m(n) = p_m(n) + a_m(n)$$

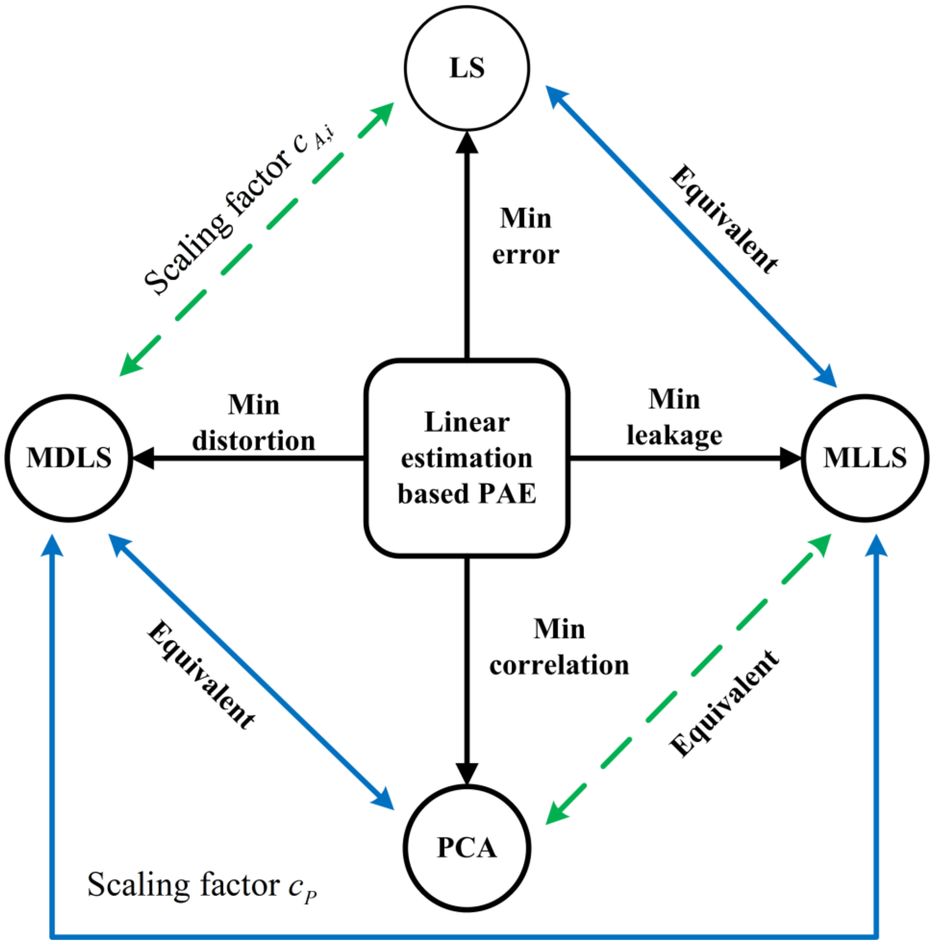
Sound scene decomposition: PAE

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

"Sum of primary and ambient components"

Case		Typical techniques
Basic model	Channel-wise	Time frequency masking
	Combine channels	Linear estimation (PCA, LS), Ambient spectrum estimation
Complex model		Time/phase shifting, Classification, Sub-band, Pairing up two channels, etc.

PAE: Linear estimation for stereo signals



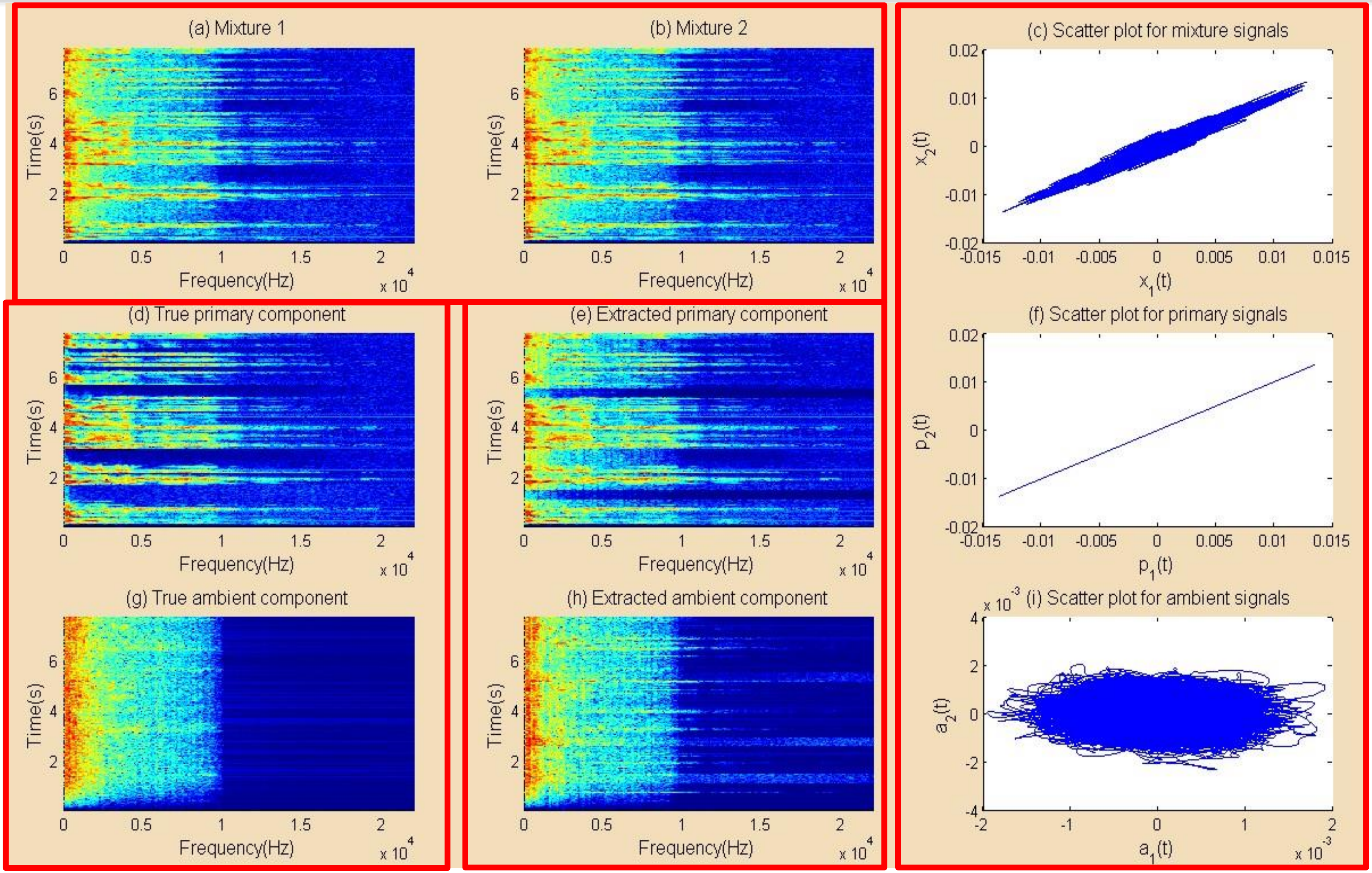
$$\begin{bmatrix} \hat{p}_0(n) \\ \hat{p}_1(n) \\ \hat{a}_0(n) \\ \hat{a}_1(n) \end{bmatrix} = \begin{bmatrix} w_{\mathrm{P}0,0} & w_{\mathrm{P}0,1} \\ w_{\mathrm{P}1,0} & w_{\mathrm{P}1,1} \\ w_{\mathrm{A}0,0} & w_{\mathrm{A}0,1} \\ w_{\mathrm{A}1,0} & w_{\mathrm{A}1,1} \end{bmatrix} \begin{bmatrix} x_0(n) \\ x_1(n) \end{bmatrix}$$

Objectives 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

[He, 2014]

PAE: an example from least-squares



B.2 Binaural rendering for VR/AR/MR

- VR/AR/MR audio aims to deliver an interactive immersive listening experience in a virtual/augmented world

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

An illustration

Natural
Cinematic / 360
Synthetic / Full

Key Technology	Ambisonics	Object-based audio
Direction rendering	HRTF	HRTF
Distance rendering	Amplitude adjustment	3D modeling, Amplitude adjustment
Reverberation	Fixed	3D modeling, Early reflection modeling
Interaction	Ambisonic rotation	3D modeling, Low pass filtering
Performance & Complexity	Medium	High (no. of sources)

http://superpowered.com/3d-spatialized-audio-virtual-reality

Other VR/AR/MR audio effects

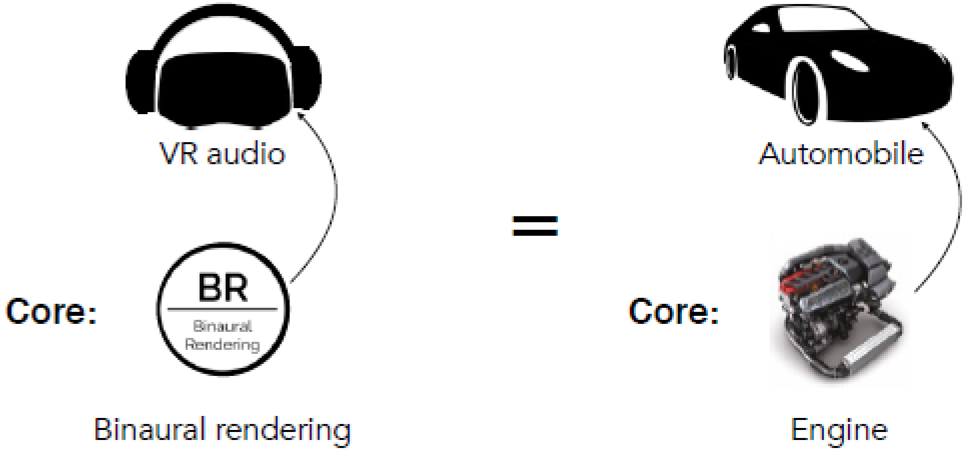
- Area sources (source with width)
- Source with directivity
- Sound transport time
- Non-spatialized audio
- Audio effects

https://developer.oculus.com/documentation/audiosdk/latest/concepts/audio-intro-mixing/

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

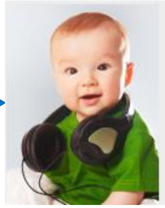
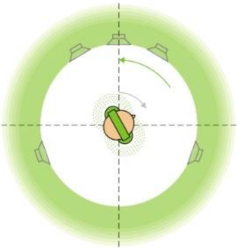
Binaural rendering recreates all the listening cues for both ears using headphones

- **Direction rendering**
- **Distance rendering**
- **Environment rendering**
- **Interaction**



[Image courtesy: GAUDIO, 2016]

Challenges and solutions

Listening	Headphone listening	Signal processing	Natural listening
Source		Virtualization of source & environment	
Medium		Equalization	
Receiver		Individualization	
		Tracking	

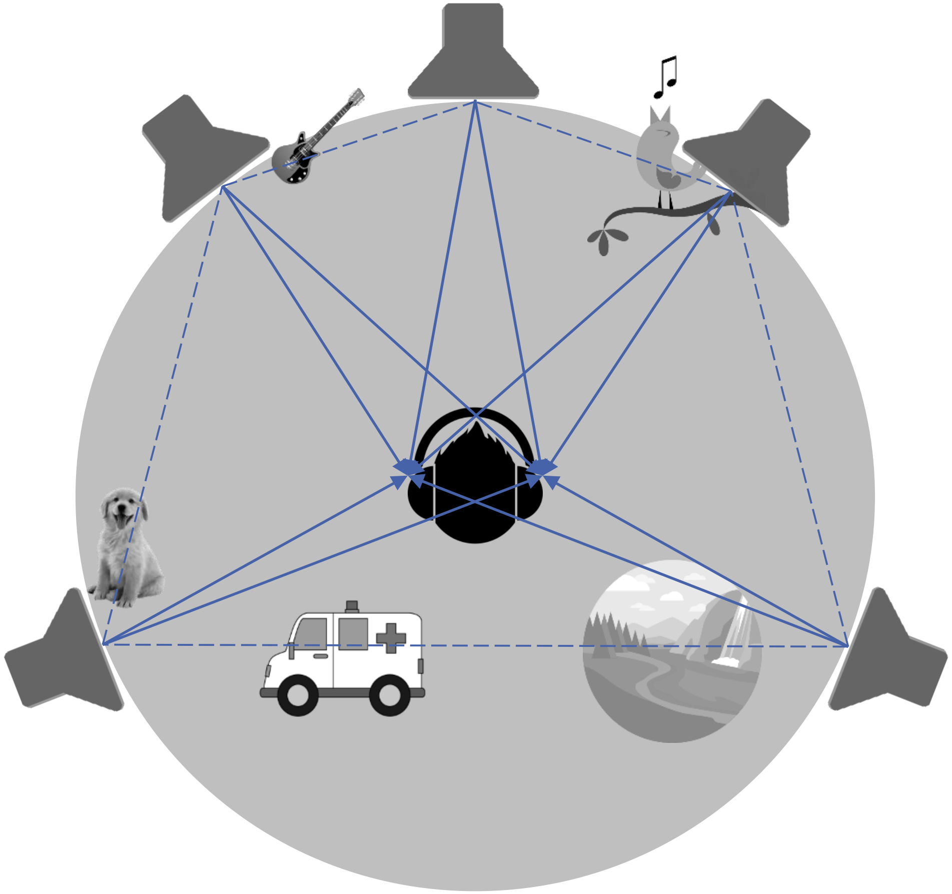
[Begault, 2000]

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

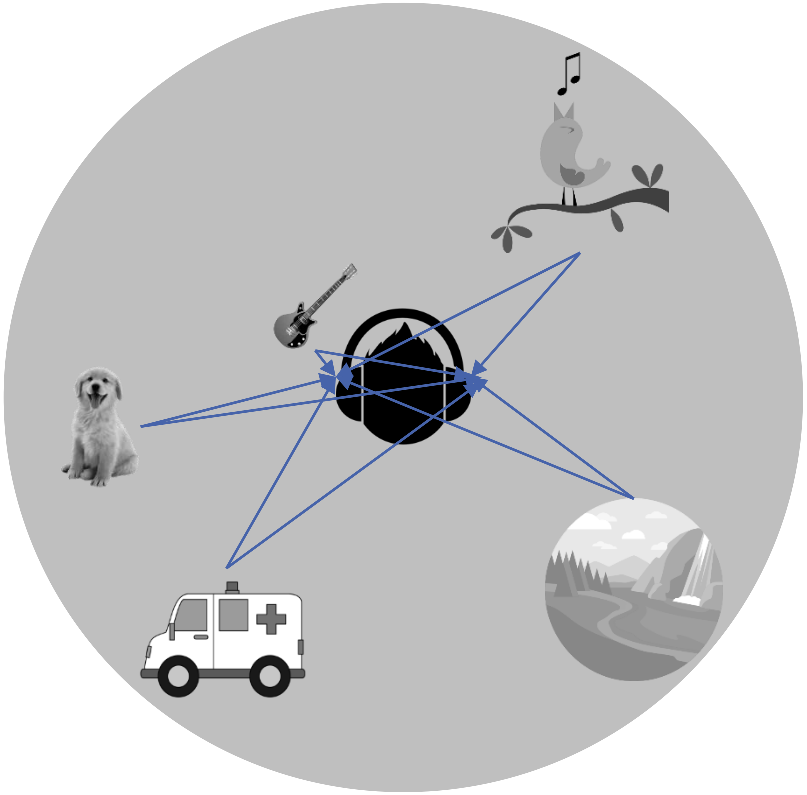
Original scene



Perceived scene

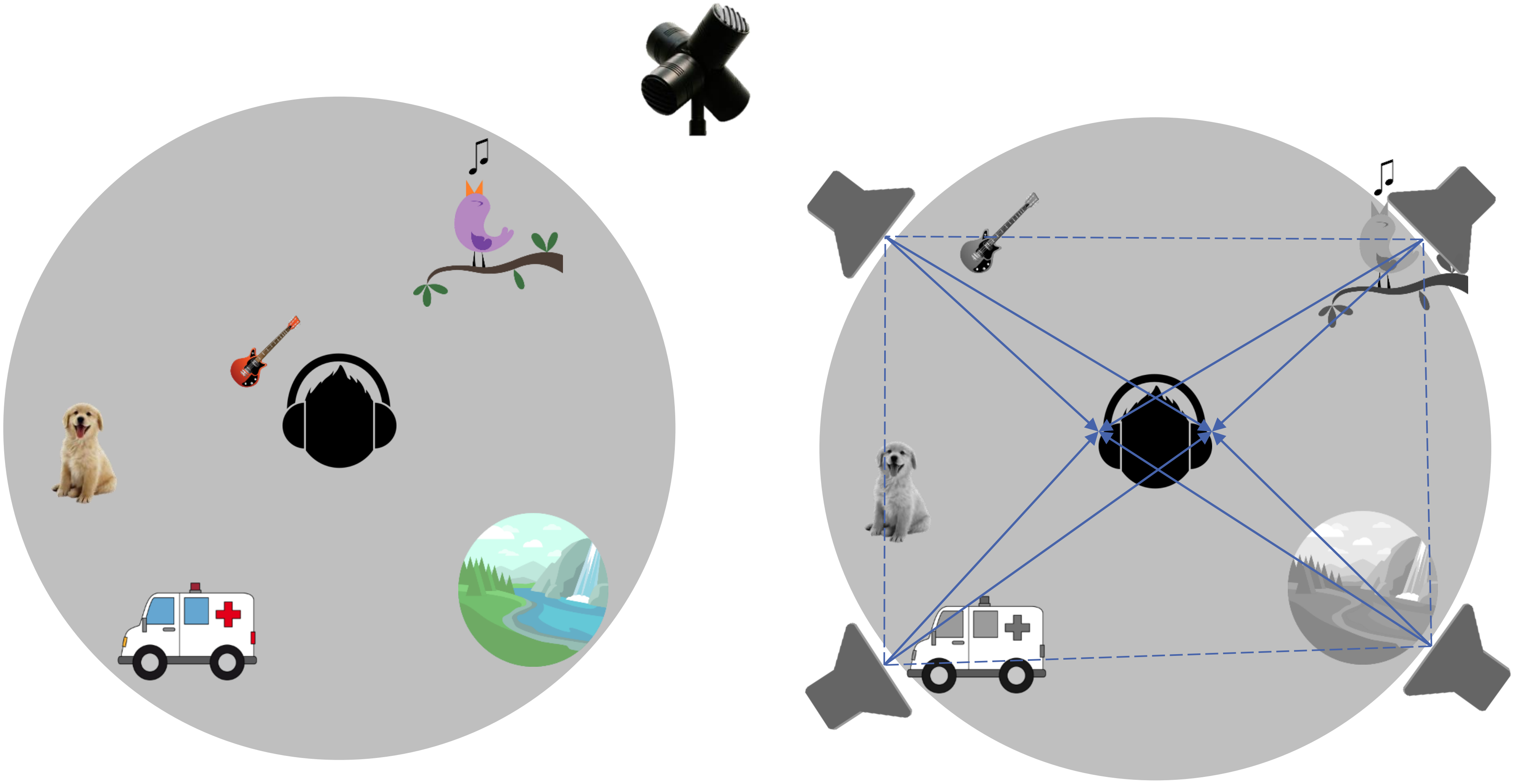
Binaural rendering of object-based audio

Original scene



Perceived scene

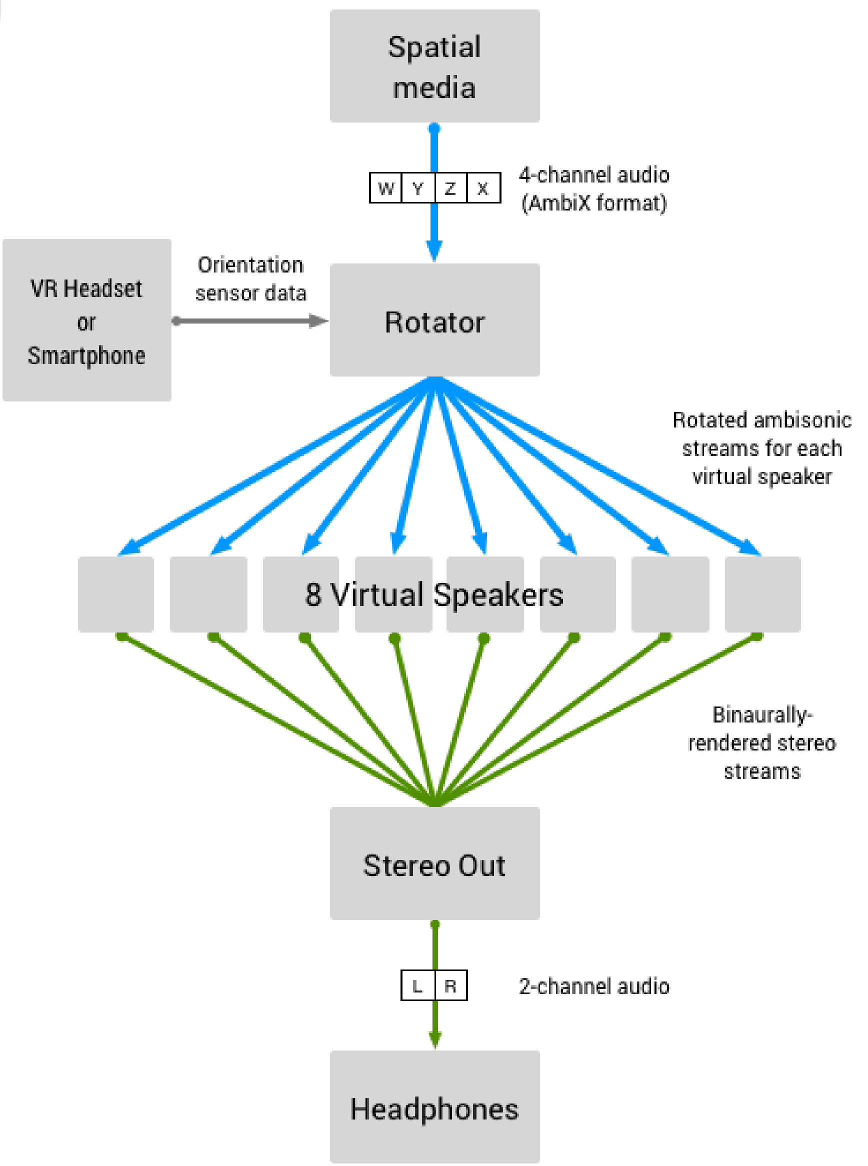
Binaural rendering of scene-based audio

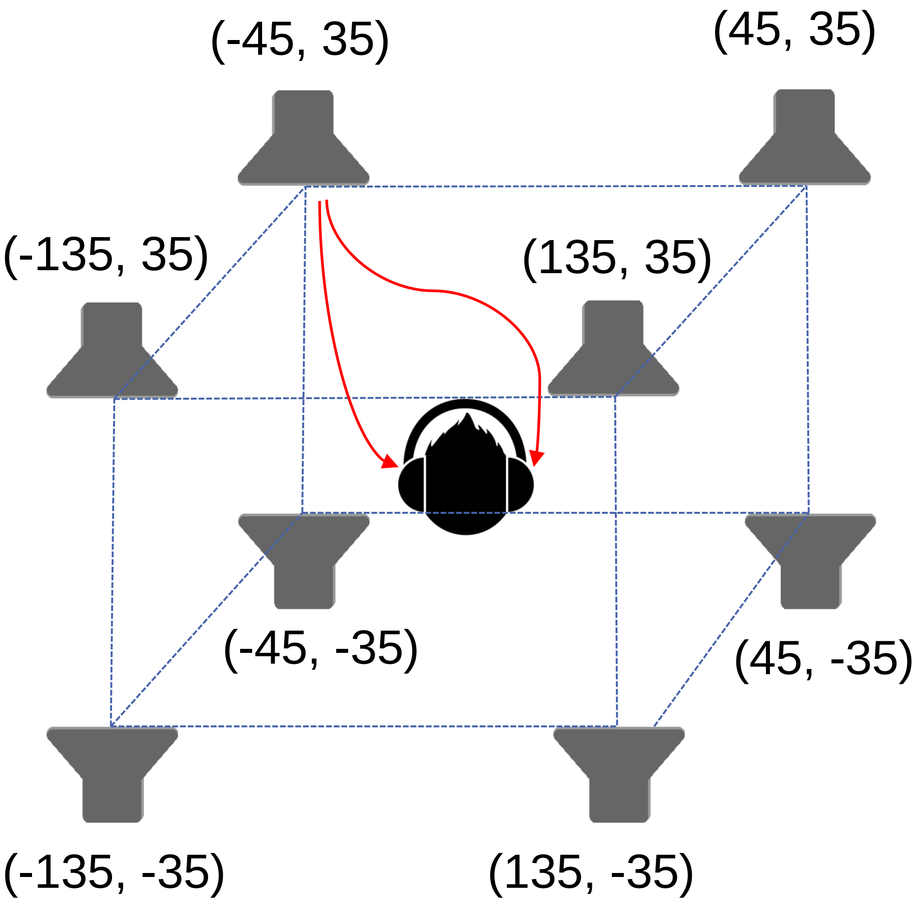


Original scene

Perceived scene

Google Omnitone

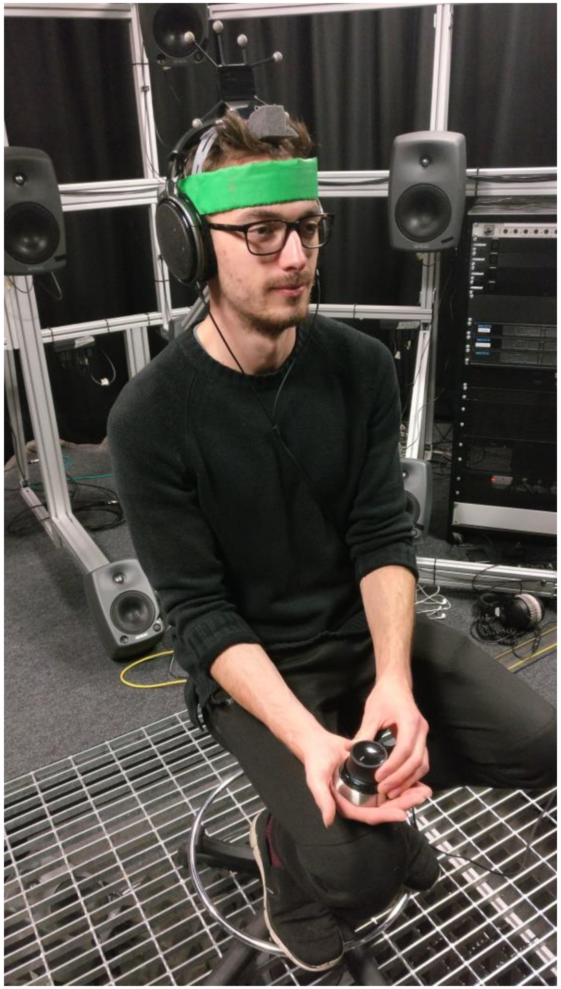




Required orders for ambisonics based binaural rendering

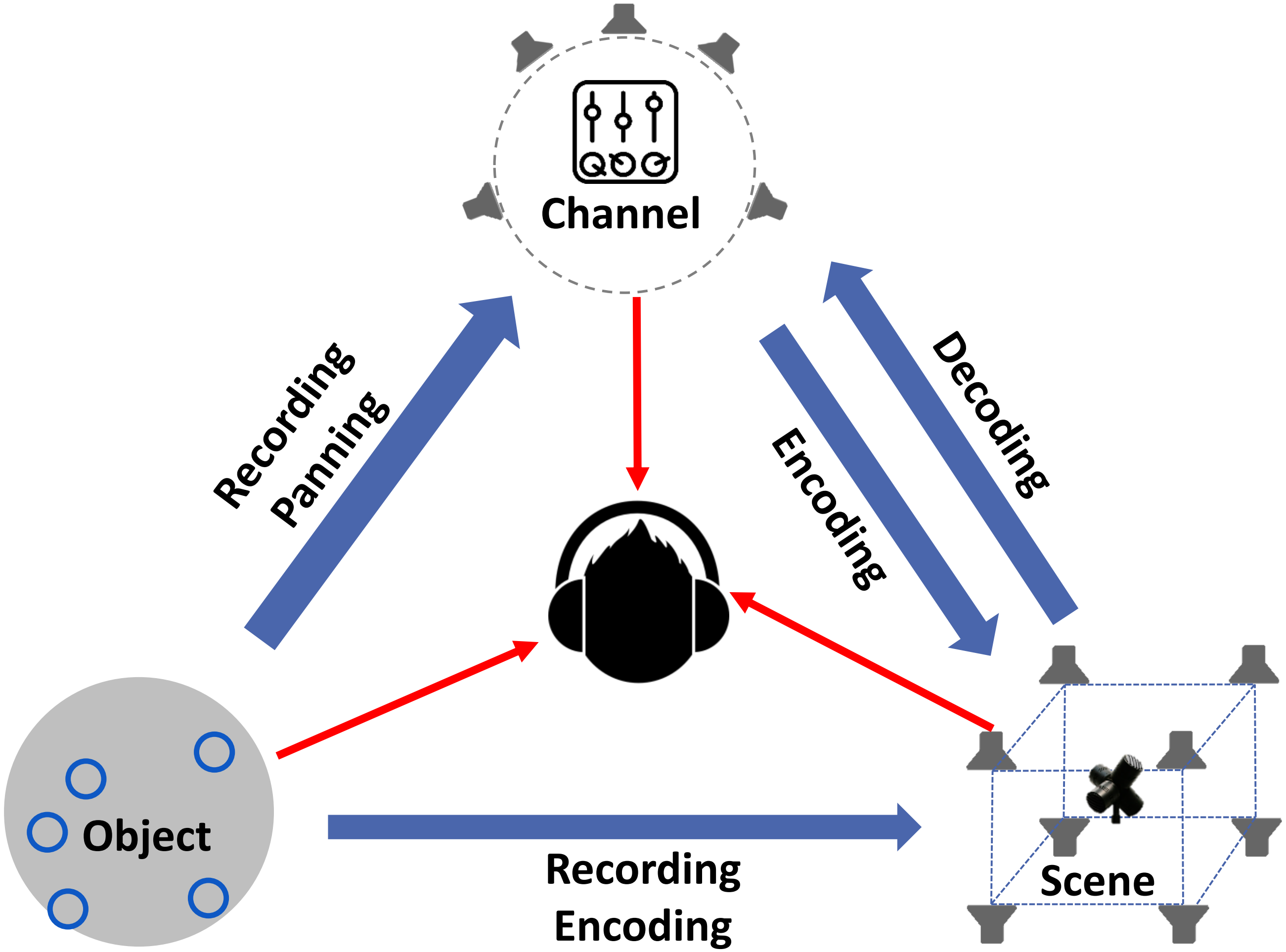
Ambisonic order	Average localization error
1st	24°
3rd	17°
5th	15°

- Using **generic HRTFs with head tracking,** performance might differ with individualized HRTFs
- Significant improvement found by increasing from 1st order to 3rd order.
- Little advantage found in 5th order over 3rd order.

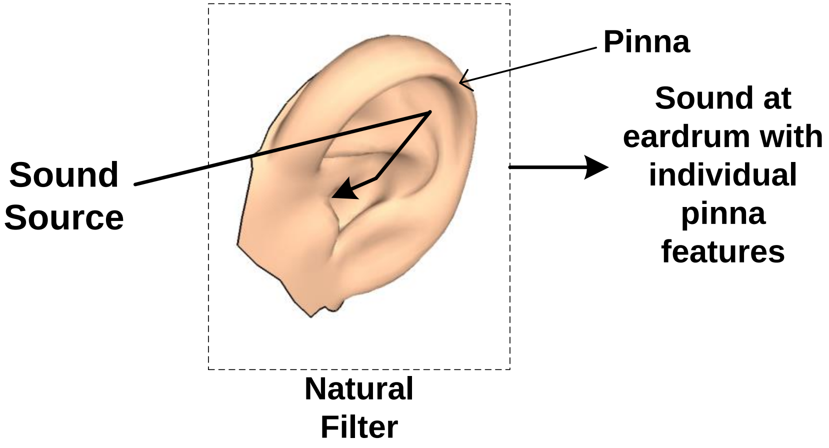


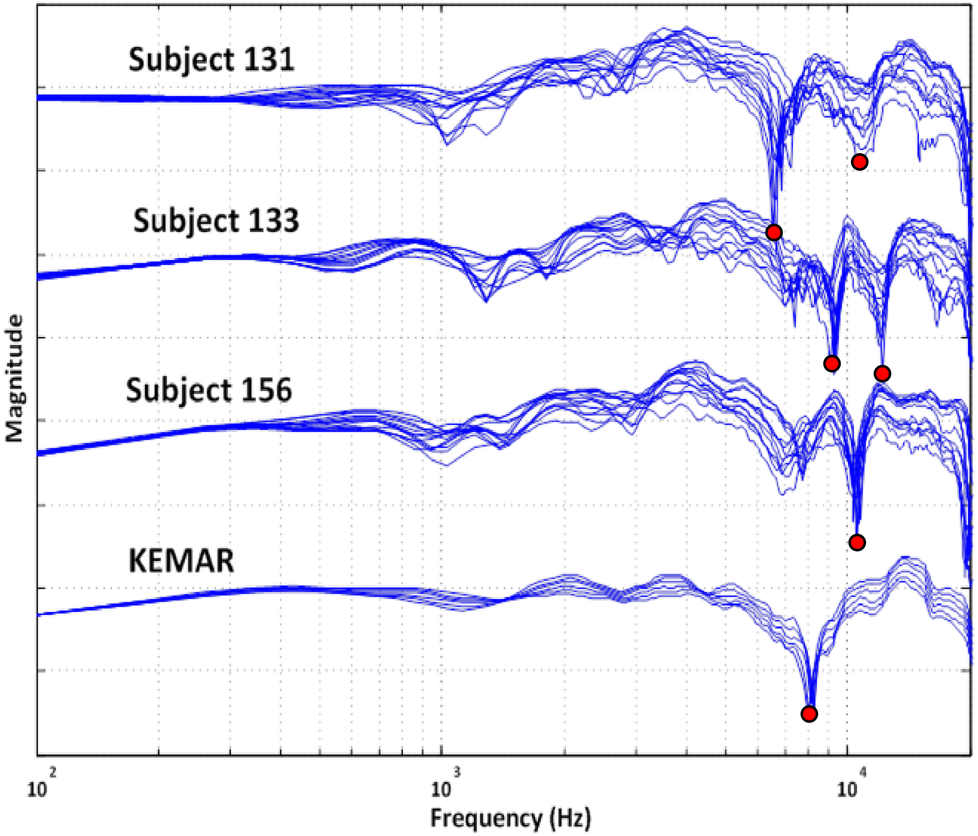
[Thresh, 2017]

Audio format transformation



B.3 HRTF Individualization



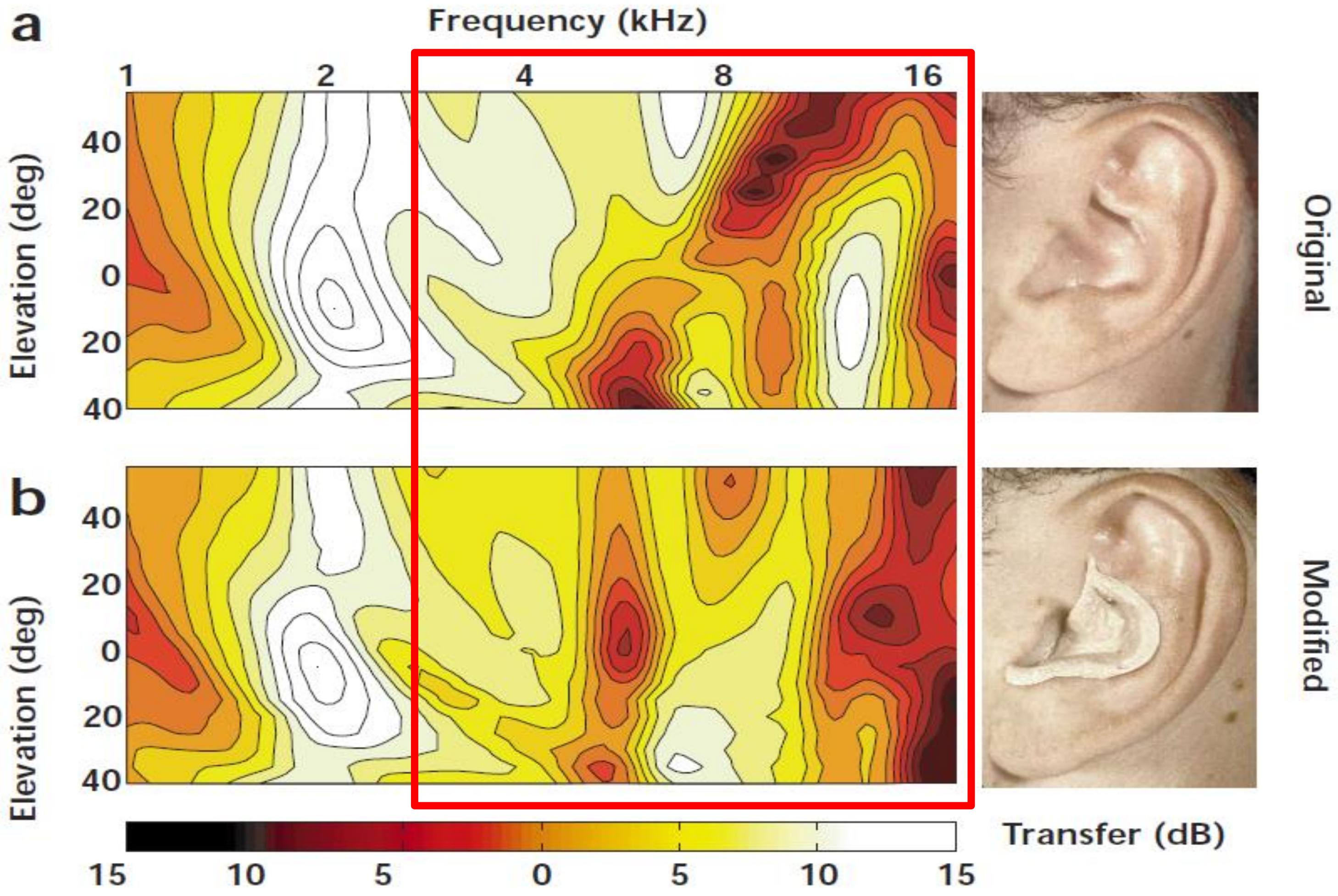


(c)

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**

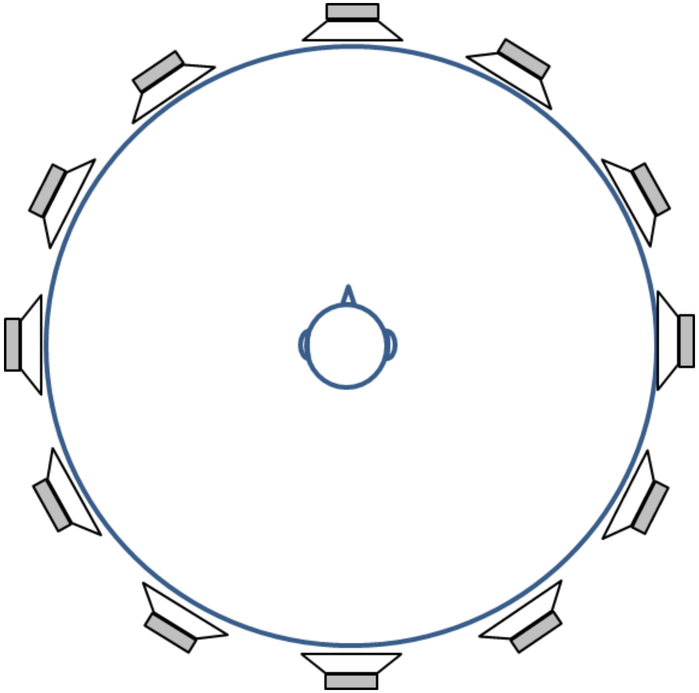
HRTF measurement techniques with human subjects

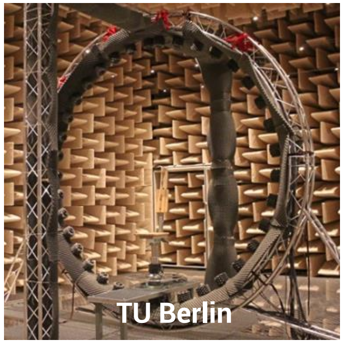
Representative references	Microphone type	Number of loudspeaker	Loudspeaker movement	Subject posture	Subject (head) movement	Subject tracking	Excitation signal	Performance evaluation	Approximate duration*
Møller, 1995 Algazi, 2001	Binaural	1-N	Discrete positions across azimuth and elevation	Sit on a normal chair	Not allowed	No	Sweep or maximum length sequence (MLS)	Reference technique	1+ hours
Carpentier, 2014	Binaural	1	Discrete positions across elevation	Sit on a chair on the turntable	Not allowed	Yes	Sweep	No	1+ hours
Majdak, 2007	Binaural	22	No	Sit on a chair on the turntable	Not allowed	Yes	Multiple exponential sweep method (MESM)	Objective	30 minutes
Bilinski, 2014	Binaural	16	Discrete positions across elevation	Sit on a normal chair	Not allowed and fixed mechanically	Yes	MESM	No	30 minutes
Bomhardt, 2017	Binaural	64	No	Stand on a turntable	Not allowed and fixed mechanically	No	MESM	No	10 minutes
Pollow, 2012	Binaural	40	No	Stand on a turntable	Not allowed and fixed mechanically	No	MESM	Subjective	10 minutes
Zotkin, 2006	32 channels	1	Across the two ears	Sit on a normal chair	Not allowed	No	Sweep	Objective	30 minutes
Fukudome, 2007	Binaural	1	Move/rotate vertically	Sit on a chair on the turntable	Not allowed	No	MLS	Objective	1+ hours
Pulkki, 2010	Binaural	1	Discrete positions across elevation, and continuous rotation across azimuth	Sit on a normal chair	Not allowed	No	Sweep	Objective	1+ hours
Enzner, 2008	Binaural	1	Discrete positions across elevation	Sit on a chair on the turntable	Not allowed	No	White noise or perfect sweep	Objective	30 minutes
Enzner, 2009	Binaural	4	Few discrete positions across azimuth/ elevation	Sit on a chair on the turntable	Not allowed	No	White noise	Objective	10 minutes
He, 2016 Li, 2017	Binaural	1	Few discrete positions across azimuth/ elevation	Sit on a rotatable chair	Free movement across azimuth/ elevation	Yes	White noise, or perfect sweep	Objective	30 minutes
Reijniers, 2017	Binaural	1	Few discrete positions across azimuth	Sit on a normal chair	Free movement across azimuth/ elevation	Yes	Sweep	Objective and subjective	30 minutes
He, 2018	Binaural	1	Few discrete positions across azimuth/ elevation	Sit on a rotatable chair	Free movement across azimuth/ elevation	Yes	White noise	Objective and subjective	30 minutes

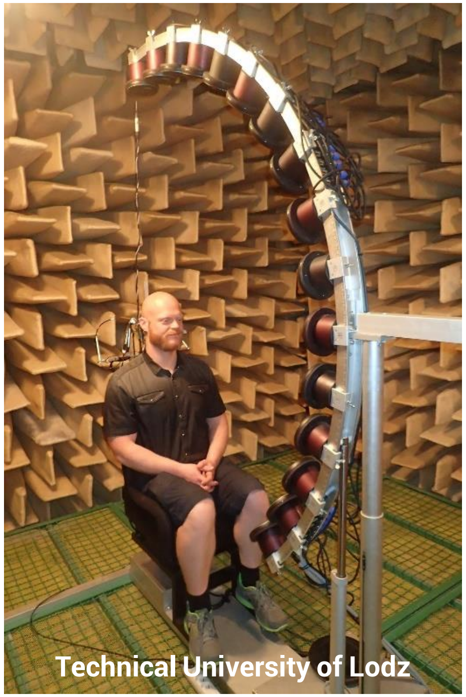
Existing HRTF acquisition techniques

Discrete stop-and-go HRTF acquisition

- ✓ Fixed measurement setups (Multiple loudspeakers play one-by-one)
- ❑ Tedious and time consuming especially for human subjects.







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:

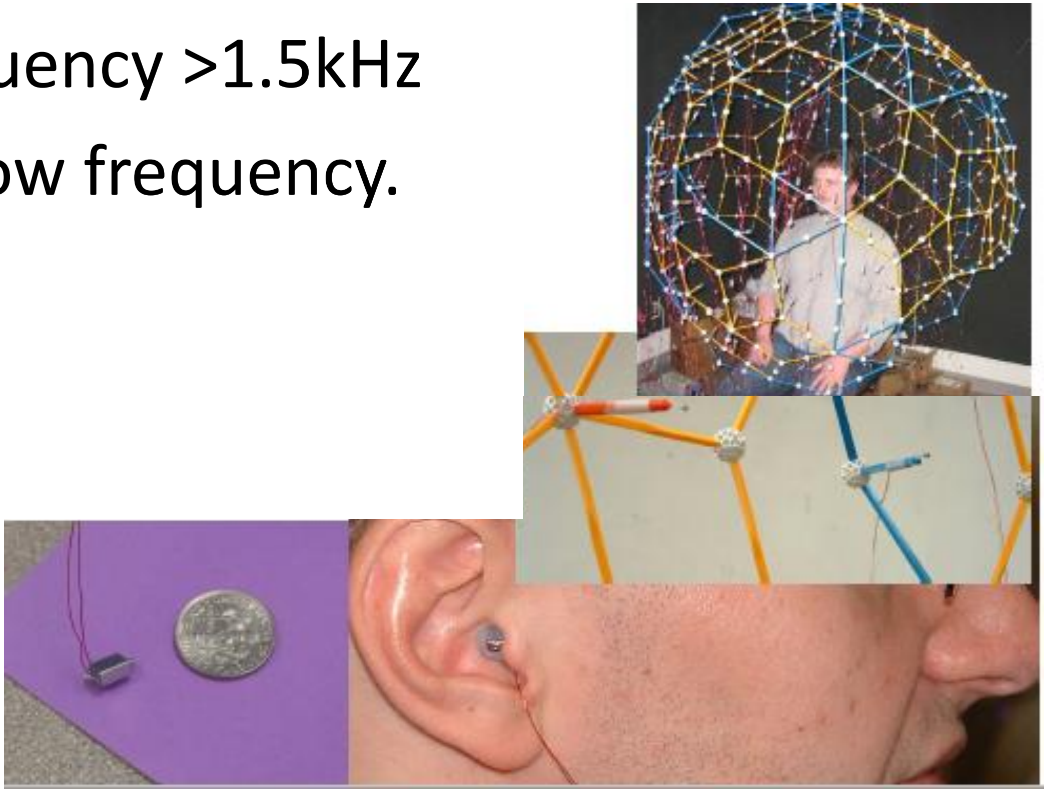
- Virtualization of 16 loudspeakers in rooms
- Head-tracking in 2D
- Individualized HRTFs via measurement
- Individualized headphone equalization
- Up-mixer
- Bit-stream decoding: Dolby, DTS, Auro-3D formats

Summary of popular HRTF databases

Databases	(Subjects, Directions)	Measuring Conditions and Features	
IRCAM France http://recherche.ircam.fr/equipes/salles/listen	(51, 187)	Far field: 1.95m Source: Log sine sweep Blocked ear canal	Length: 8192-pt / 512-pt Fs = 44.1 kHz Anechoic room
CIPIC , UC Davis https://www.ece.ucdavis.edu/cipic/spatial-sound/hrtf-data/	(45,1250)	Far field: 1m Source: Golay code Blocked ear canal	Length: 200-pt Fs = 44.1 kHz Non-anechoic room
Tohoku University, Japan http://www.ais.riec.tohoku.ac.jp/lab/db-hrtf	(3,454)	Far field: 1.2m Source: Time stretched pulse Blocked ear canal	Length: 512-pt Fs = 44.1kHz Anechoic room
Nagoya University, Japan http://www.sp.m.is.nagoya-u.ac.jp/HRTF/database.html	(96,72)	Far field: 1.52m Source: Time stretched pulse Not entirely block	Length: 512-pt Fs = 48kHz Non-anechoic room
Austrian Academy of Sciences http://www.kfs.oeaw.ac.at/index.php?option=com_content&view=article&id=608:ari-hrtf-database&catid=158:resources-items&Itemid=606&lang=en	(70,1550)	Far field: 1.2m Source: exponential sweep signal Blocked ear canal	Length: 2400-pt/256-pt Fs = 48kHz Semi-anechoic room
TU Berlin https://depositonce.tu-berlin.de/handle/11303/6153.2	(FABIAN,11950)	Far field: 1.7m Source: Sine sweep Blocked Ear canal	Length: 256-pt Fs = 44.1 KHz
MIT Lab http://sound.media.mit.edu/resources/KEMAR.html	(KEMAR,710)	Far field: 1.4m Source: MLS Ear simulator	Length: 512-pt Fs = 44.1 kHz Anechoic room
Oldenburg University (0.8m,3m) http://medi.uni-oldenburg.de/hrir/html/documentation.html	(HATS,365)	Far field: 0.8 – 3m Source: MIRS In-the ear and behind the ear	Fs = 48kHz Anechoic room/offices
SDAC, KAIST (0.2,0.6,1m) http://sdac.kaist.ac.kr/research/index.php?mode=area&act=DownHRTFDatabase	(HATS, 100)	Far field: 1m Source: White noise	Length: 200-pt Fs = 44.1kHz
RIEC University (1.5 m) http://www.riec.tohoku.ac.jp/pub/hrtf/hrtf_data.html	(105,865)	Far field: 1.5 m Source: Time stretched pulse Blocked ear canal	Length: 512-pt Fs = 48 kHz
Xie (Chinese Human subject database) (1.5m) https://link.springer.com/article/10.1007/s11433-007-0018-x	(52,493)	Far Field: 1.5 m Source: MLS Blocked ear canal	Length: 512 Fs = 44.1 KHz
DSP Lab @ NTU (0.35,0.45,0.50,0.60,0.75,0.8,1,1.4m) http://eeewebc.ntu.edu.sg/dsplab/ewsgan/resource.html	(HATS + 3 subjects, 600)	Far field: 1-1.4m & Near field: 0.35m-0.8m Source: MLS Block era canal/Ear simulator	Length: 512-pt / 256-pt Fs = 44.1 kHz Anechoic room

Reciprocal HRTF measurement technique

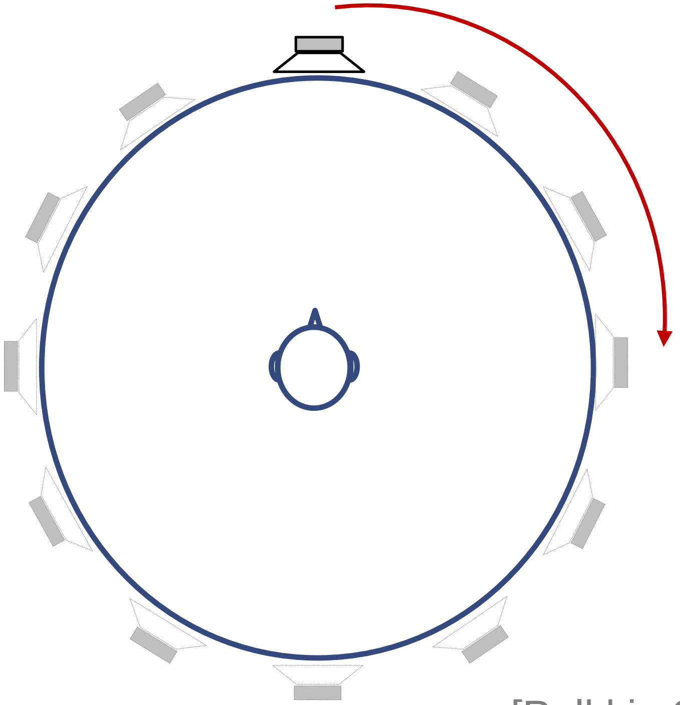
- Placing micro-speakers inside ear canal
- Spherical microphone array surround subject
- Measure HRTF at frequency >1.5kHz
- Use HRTF of HATS at low frequency.



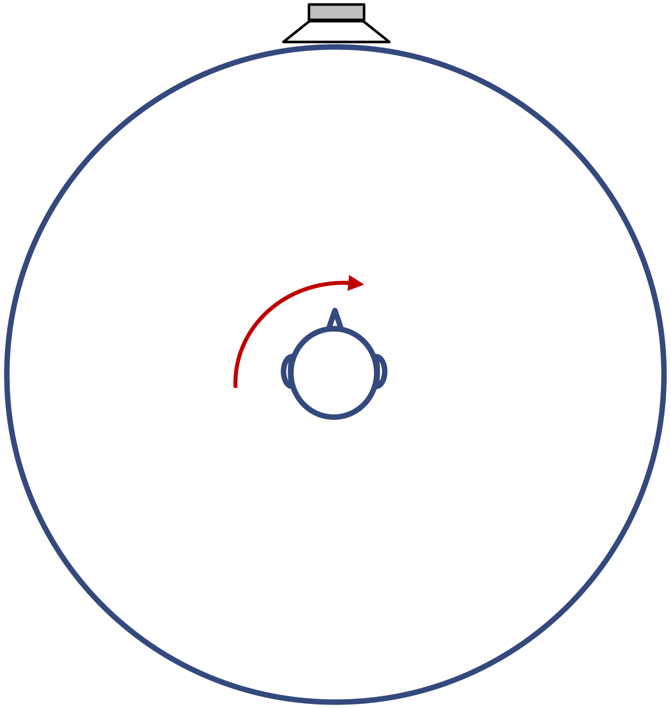
[Zotkin, 2006]

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



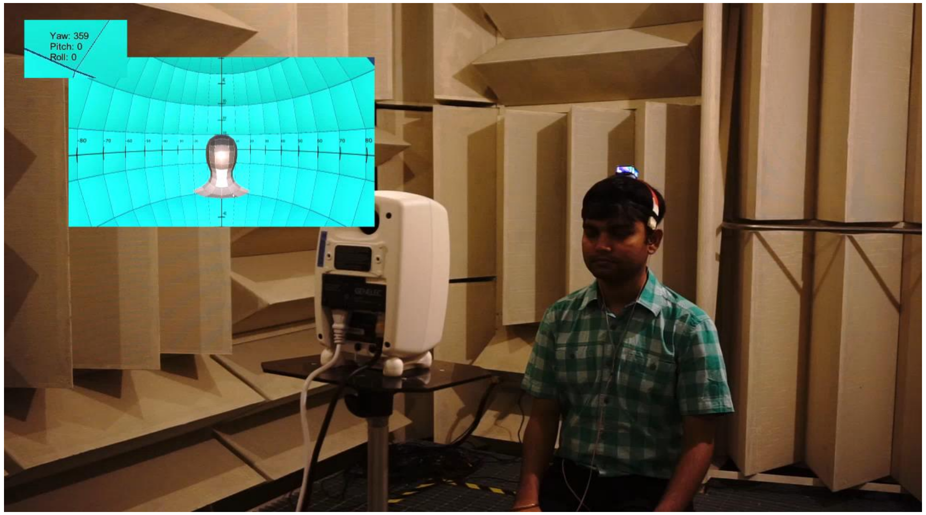
[Pulkki, 2010]

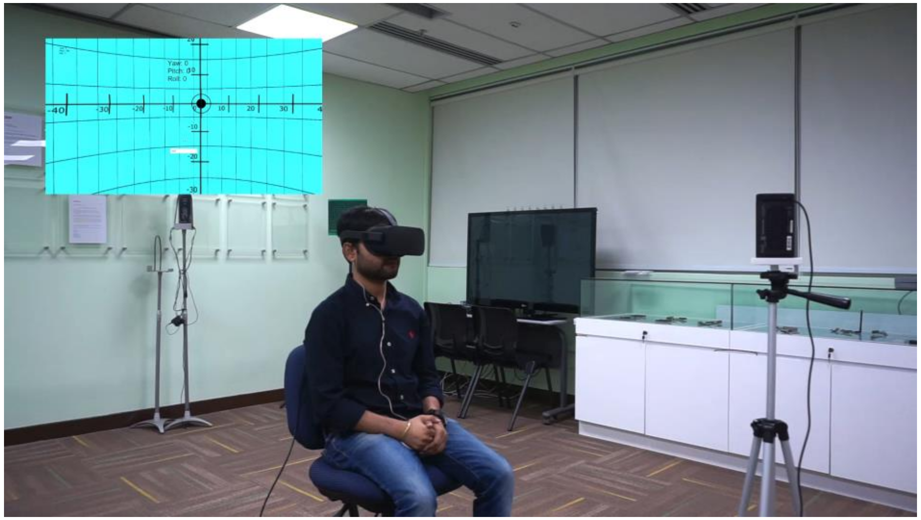


[Enzner , 2008]

Fast HRTF measurement system

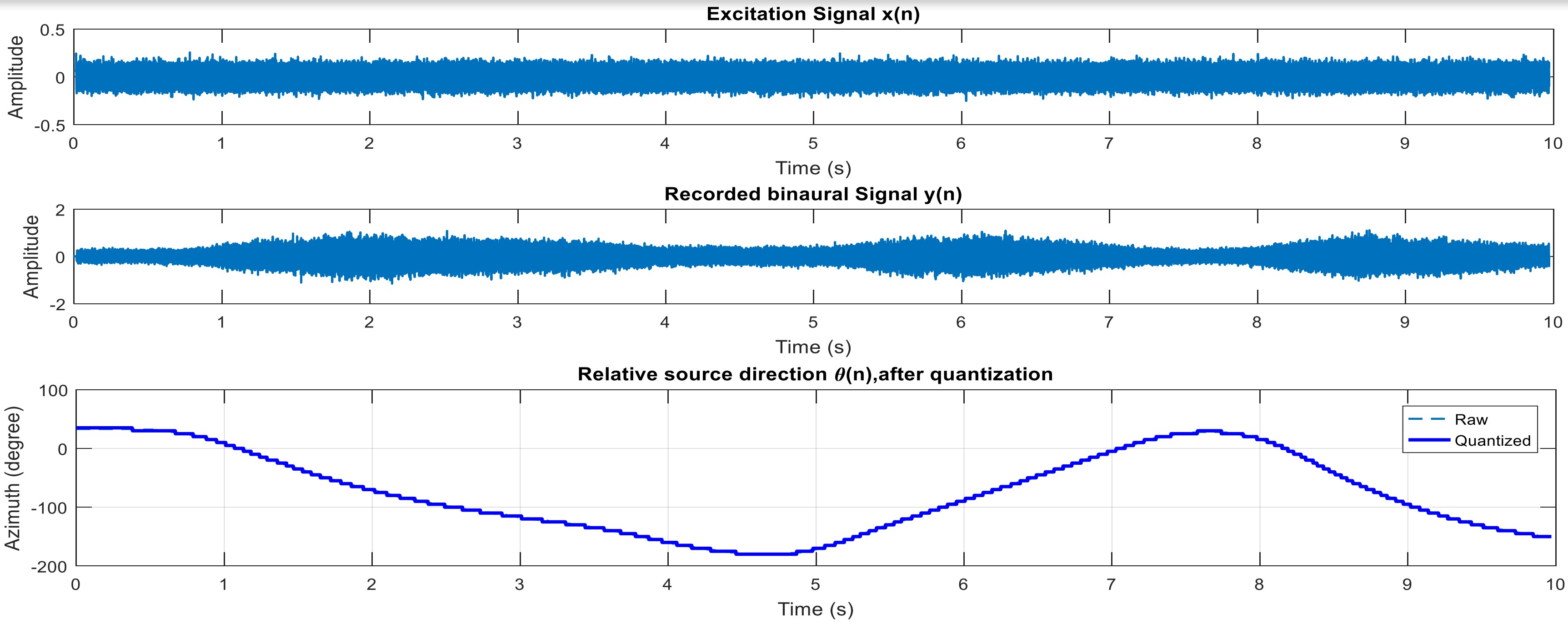
- **Head tracking allows free movements** in azimuth / elevation
- **A fixed loudspeaker** continuously emitting **broadband** signal
- Binaural recording at listener’ ears and **synchronized with directional movements**
- **Visual display** feedbacks the movement pattern



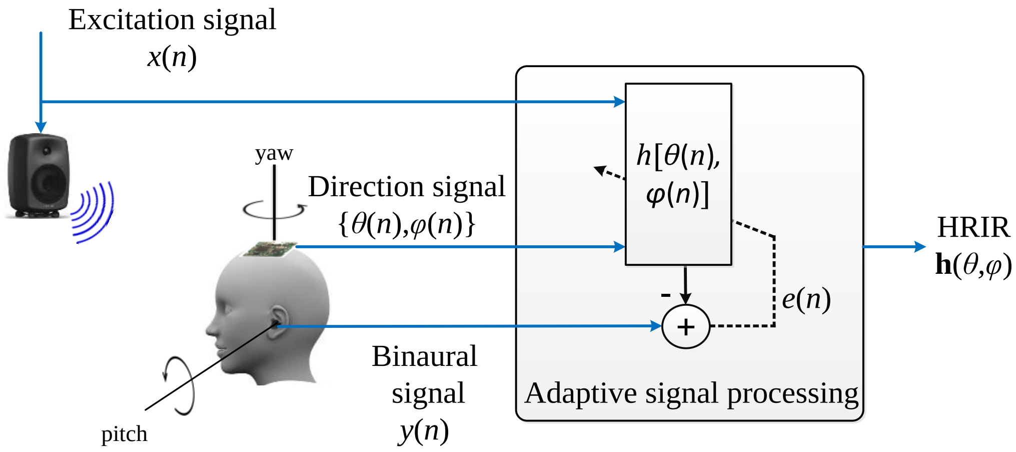


[He, 2015; Ranjan, 2016]

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

$$\underset{\text{Measured Signal}}{y(n)} = \underbrace{\boldsymbol{h}^{\mathrm{T}}[\theta(n),\varphi(n)]}_{\text{Time varying HRIR at } \theta(n),\varphi(n)} \underbrace{\mathbf{x}(n)}_{\text{Excitation Signal}} + \underbrace{v(n)}_{\text{Measurement Noise}}$$

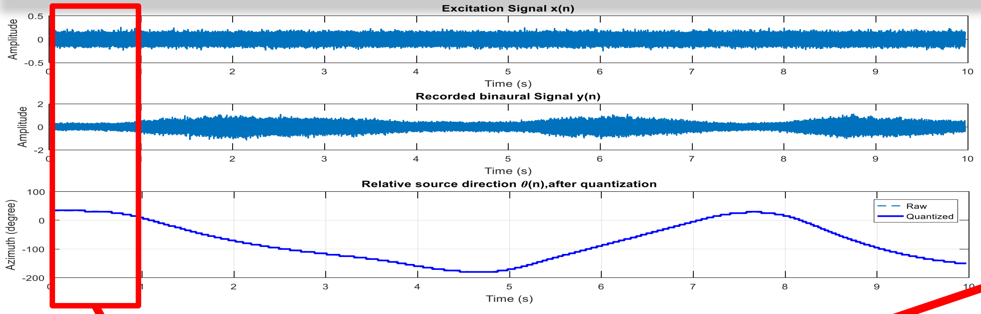
Discretization of the continuous directions

$$y(n) = \boldsymbol{H}^{\mathrm{T}}[\mathrm{d}(n)]\mathbf{x}(n) + v(n)$$

$$\boldsymbol{H}_{K\times M} = \begin{bmatrix} \mathbf{h}(\theta_1, \varphi_1) & \cdots & \mathbf{h}(\theta_1, \varphi_M) \\ \vdots & \mathbf{h}(\theta_k, \varphi_m) & \vdots \\ \mathbf{h}(\theta_K, \varphi_1) & \cdots & \mathbf{h}(\theta_K, \varphi_M) \end{bmatrix}$$

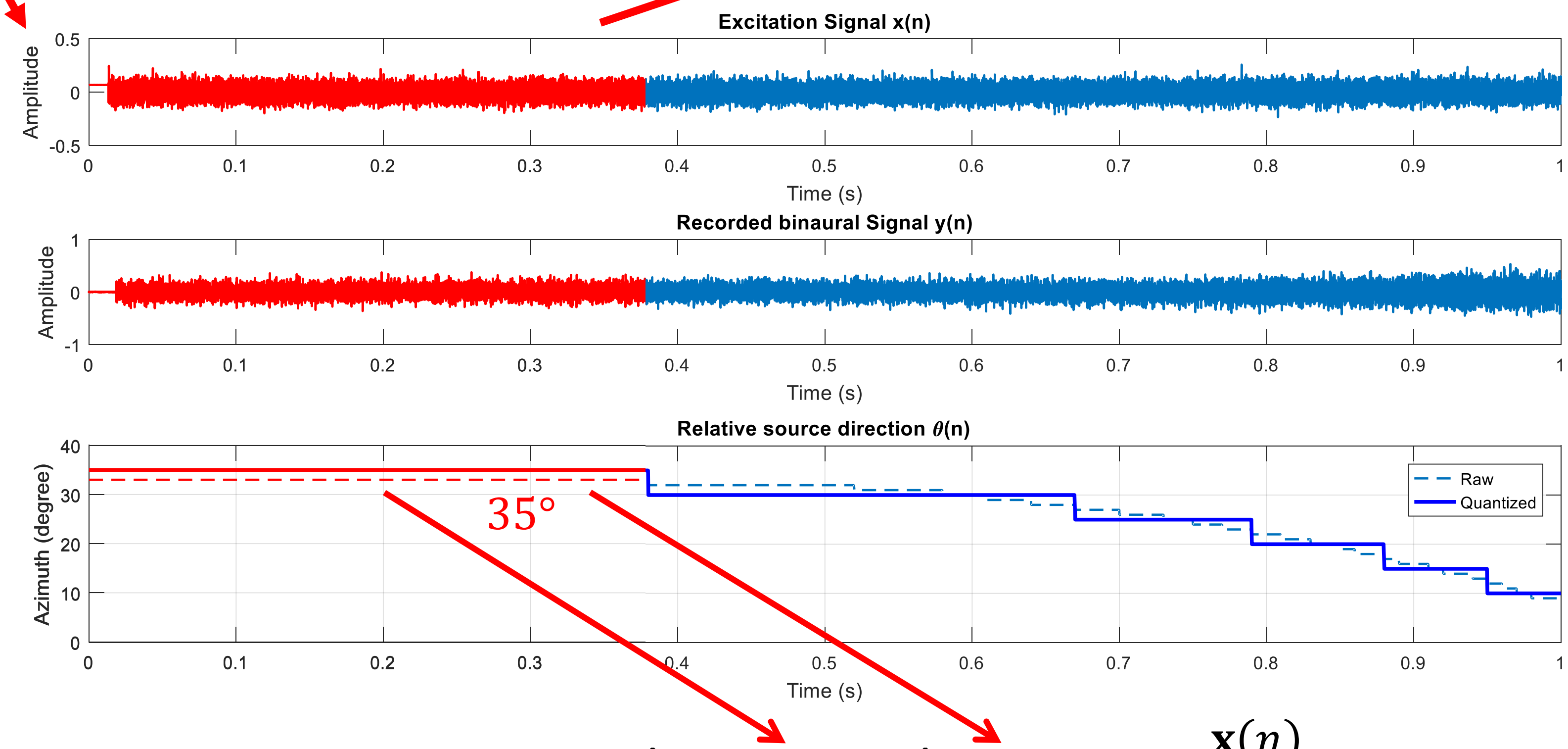
$\boldsymbol{K} \times \boldsymbol{M}$ discrete HRIRs to be estimated

Adaptive filter for dynamic HRIR estimation



	40°	① 35°	30°	25°	

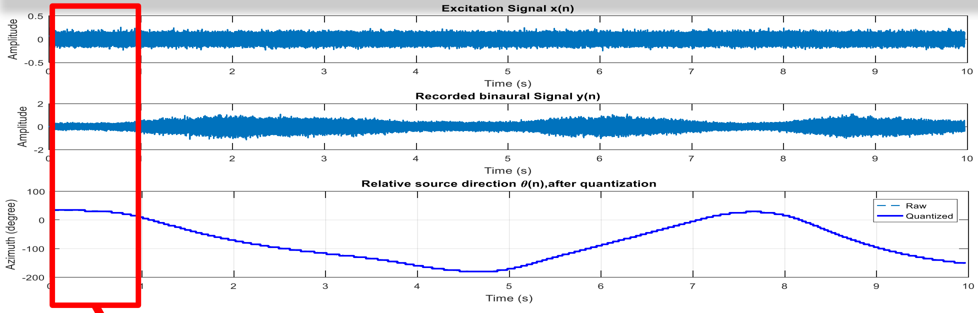
HRTF estimation results matrix



Progressive based NLMS

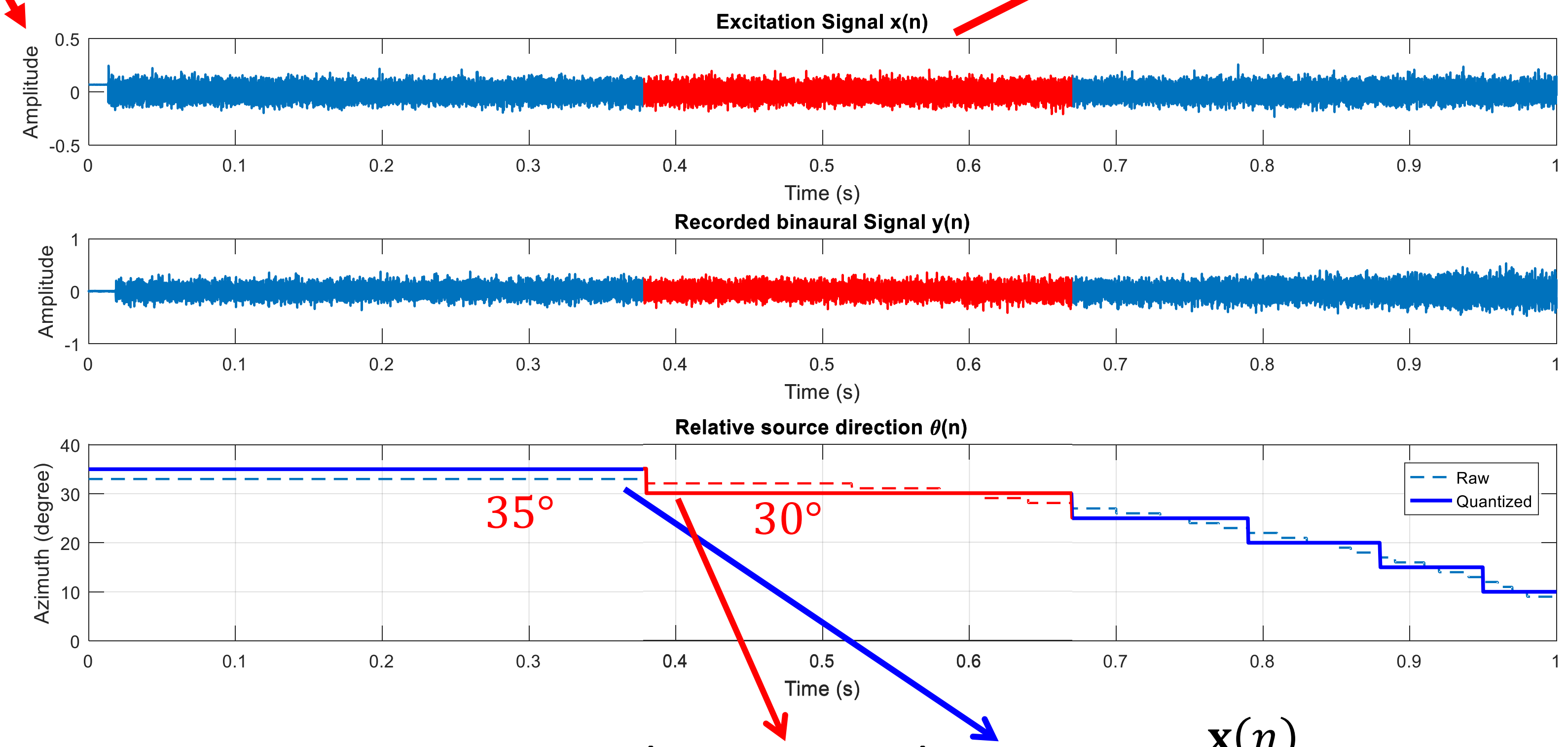
$$\hat{\mathbf{h}}_{n+1}(35°) = \hat{\mathbf{h}}_n(35°) + \mu \frac{\mathbf{x}(n)}{\|\mathbf{x}(n)\|_2^2} e(n)$$

Adaptive filter for dynamic HRIR estimation



	40°	① 35°	① 30°	25°	

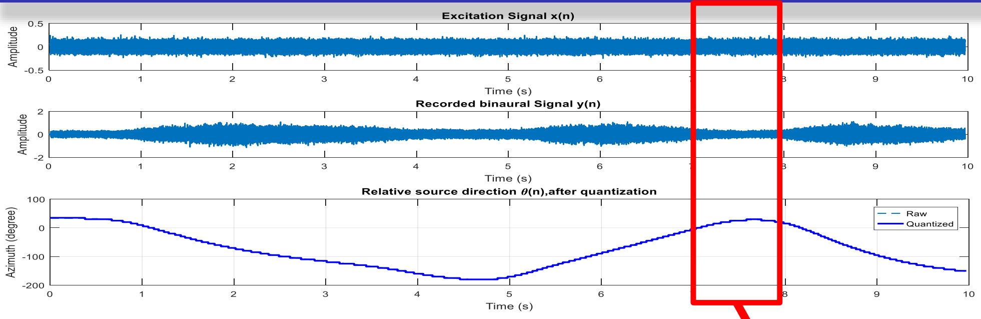
HRTF estimation results matrix



Progressive based NLMS

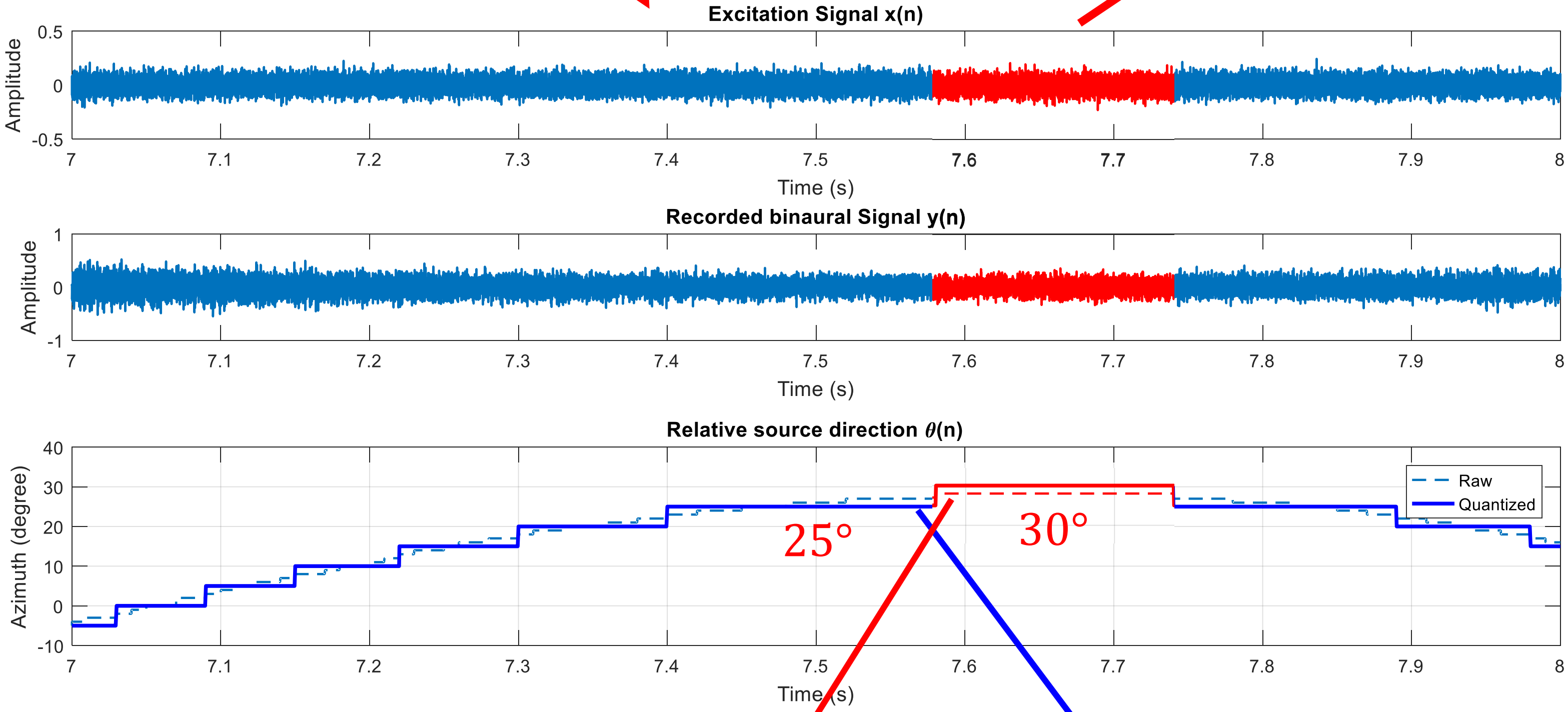
$$\hat{\mathbf{h}}_{n+1}(30°) = \hat{\mathbf{h}}_{n}(35°) + \mu \frac{\mathbf{x}(n)}{\|\mathbf{x}(n)\|_2^2} e(n)$$

Adaptive filter for dynamic HRIR estimation



	40°	① 35°	① 30° ②	25°	

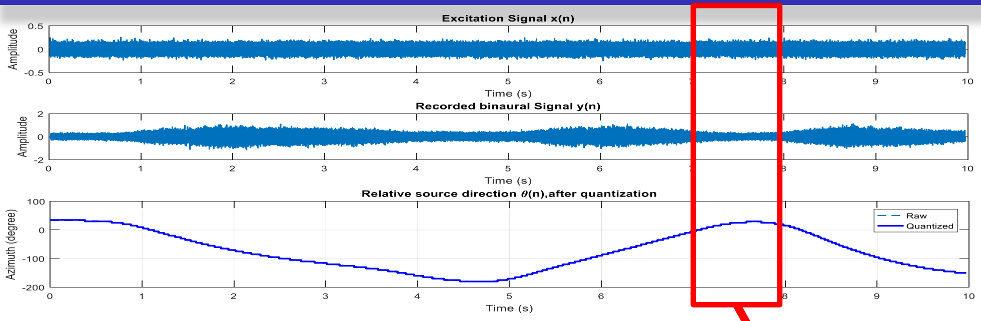
HRTF estimation results matrix



Progressive based NLMS

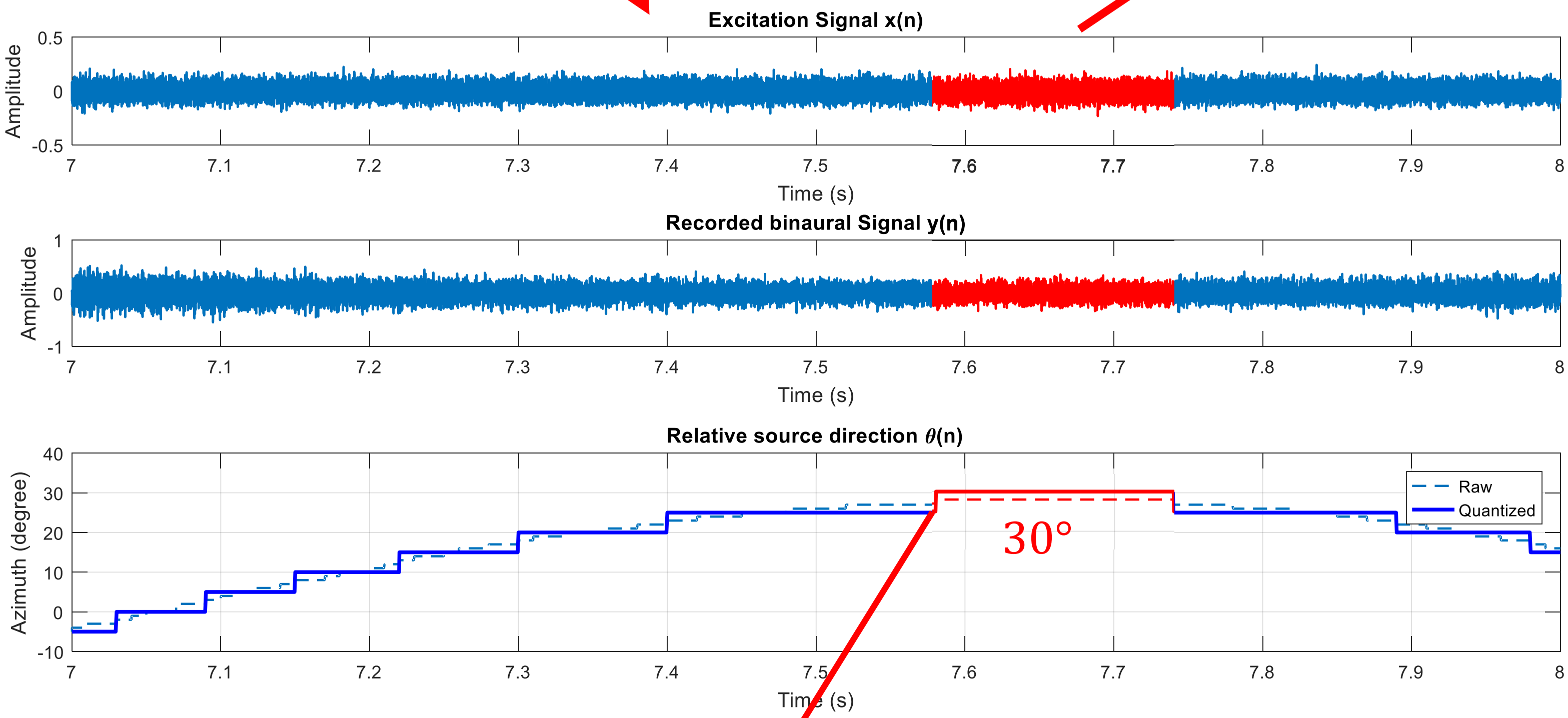
$$\hat{\mathbf{h}}_{n+1}(30^\circ, \textcircled{2}) = \hat{\mathbf{h}}_n(25^\circ) + \mu \frac{\mathbf{x}(n)}{\|\mathbf{x}(n)\|_2^2} e(n)$$

Adaptive filter for dynamic HRIR estimation



	40°	① 35°	① 30° ②	25°	

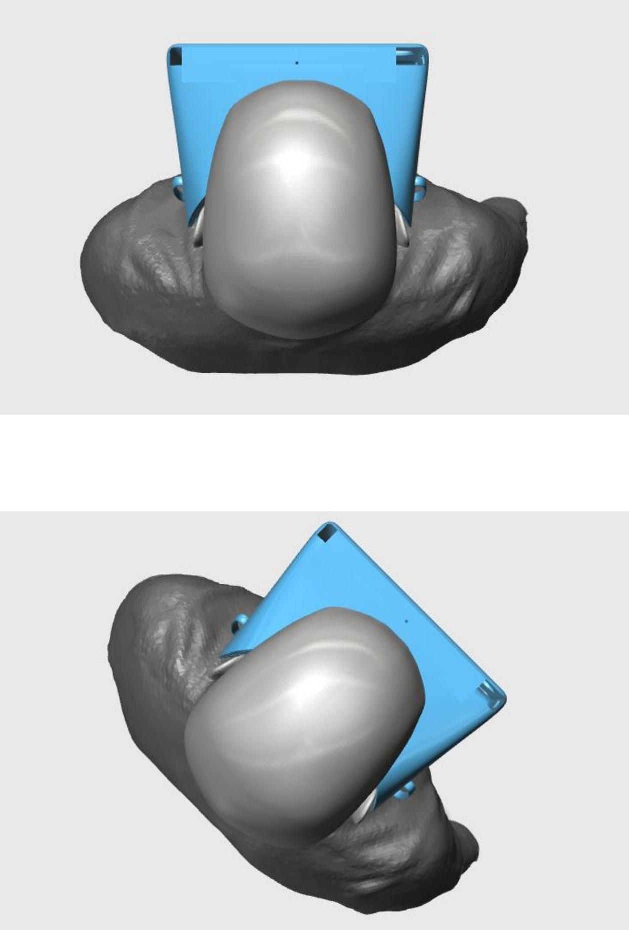
HRTF estimation results matrix



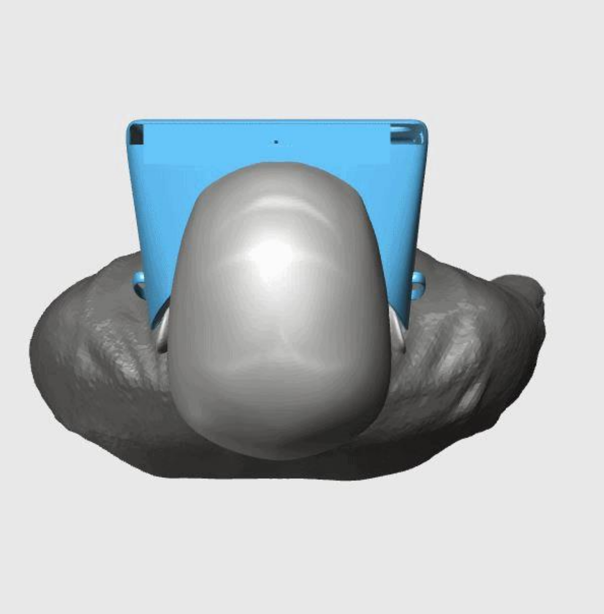
Activation based NLMS

$$\hat{\mathbf{h}}_{n+1}(30°, \text{②}) = \hat{\mathbf{h}}_{n}(30°, \quad) + \mu \frac{\mathbf{x}(n)}{\|\mathbf{x}(n)\|_2^2} e(n)$$

(1) Static vs dynamic measurement

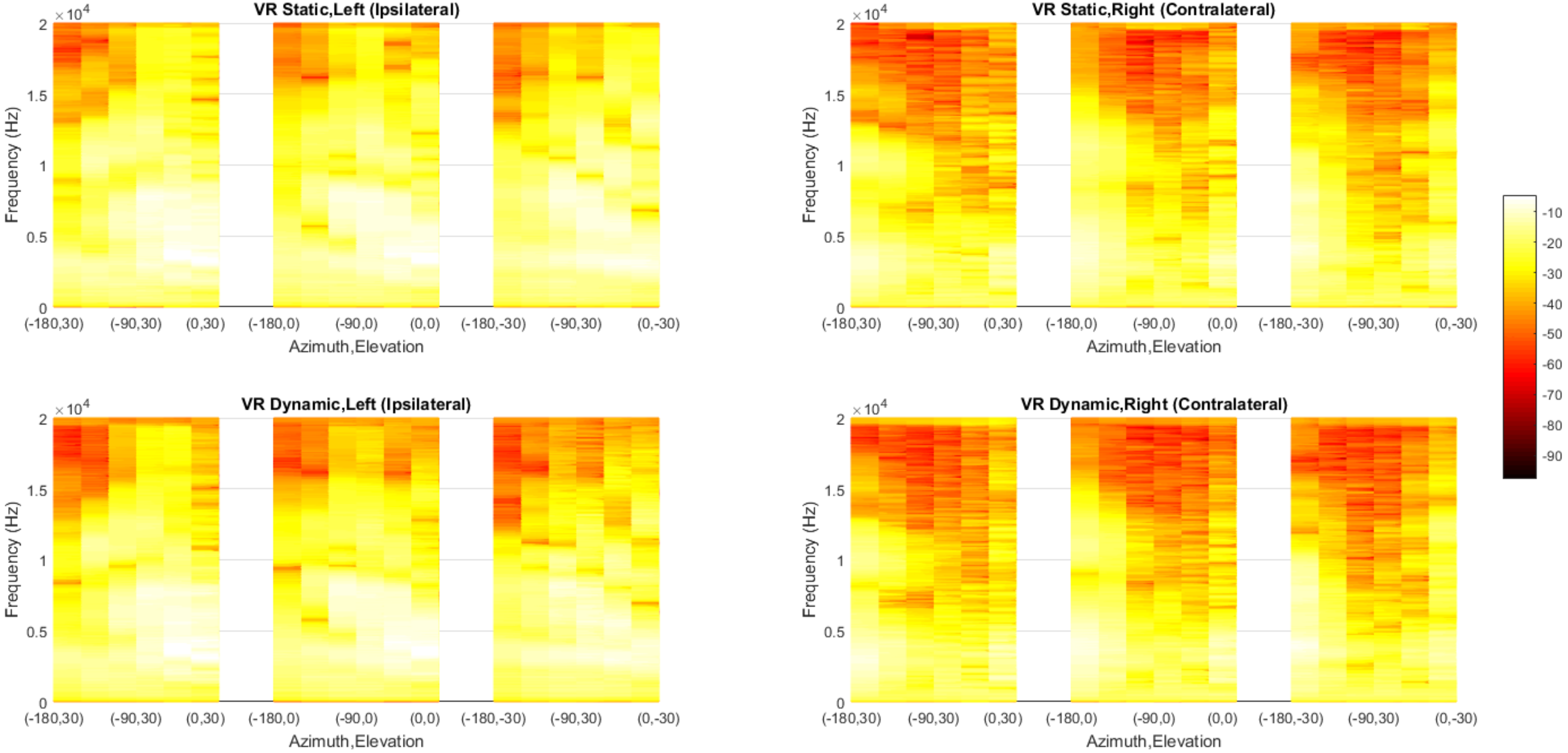


Static Measurement

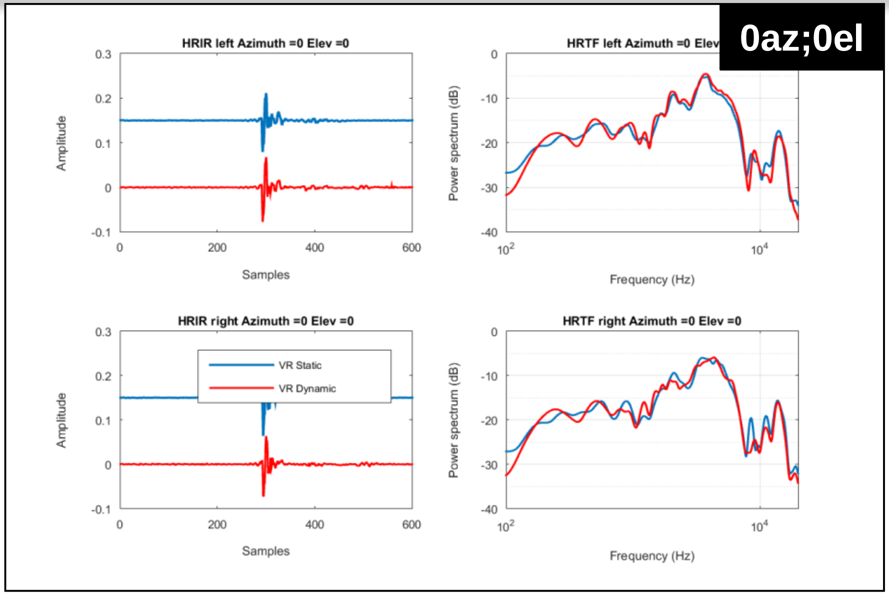


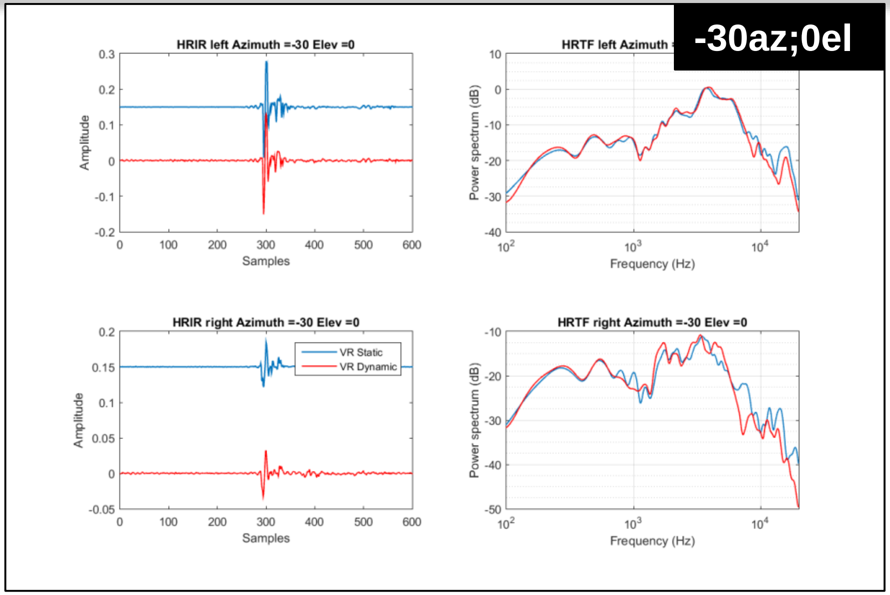
Dynamic Measurement

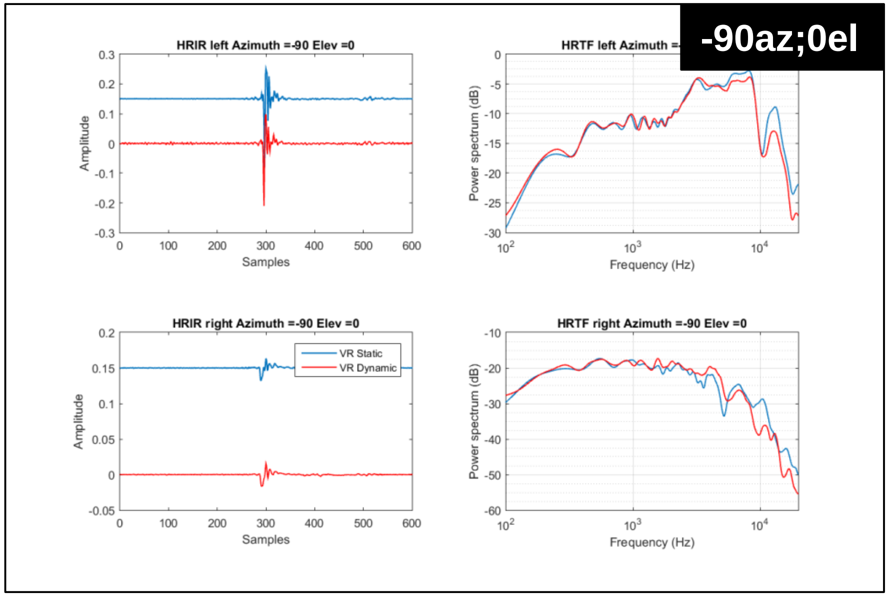
Static vs dynamic acquisition: Spectrogram

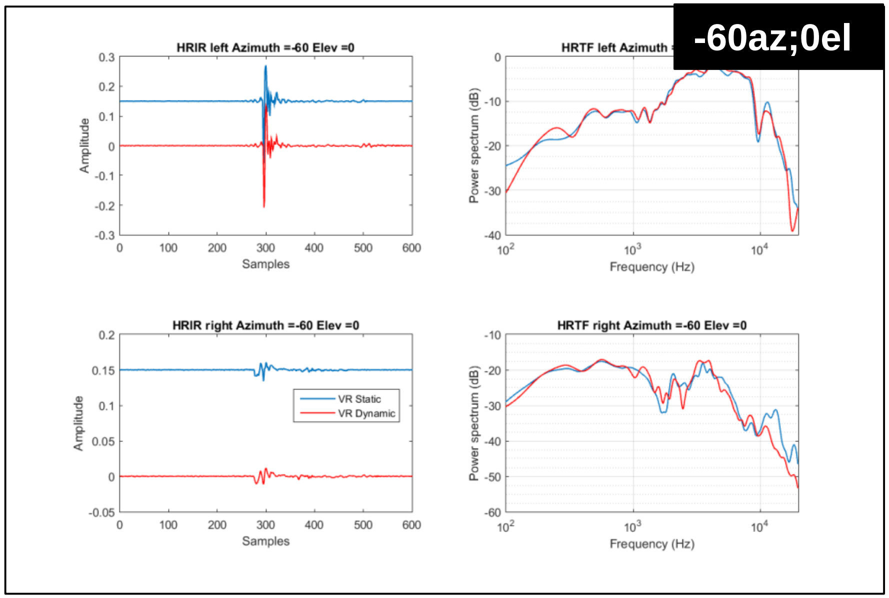


Static vs dynamic acquisition: HRIR/HRTF

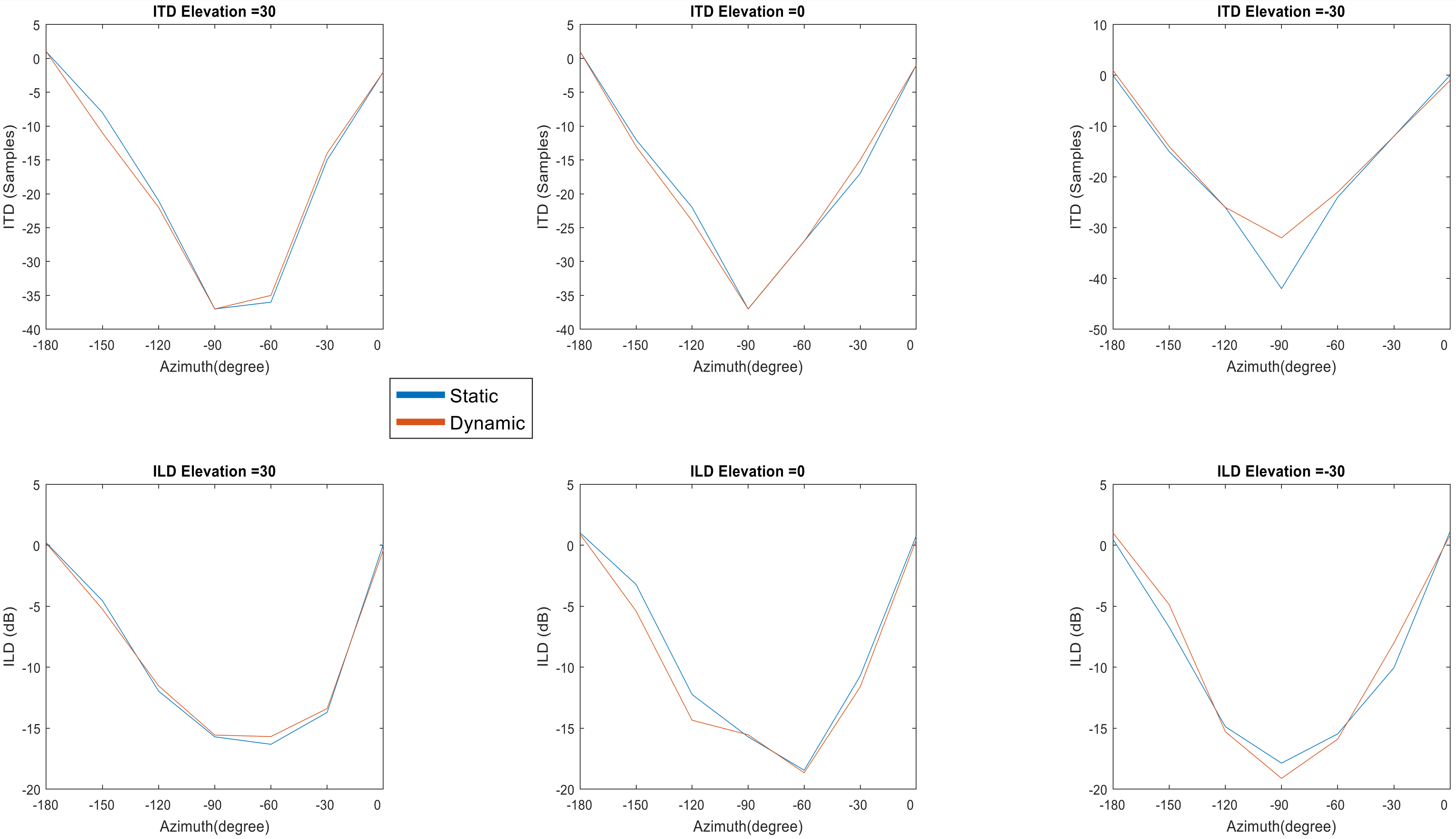

0az;0el


-30az;0el

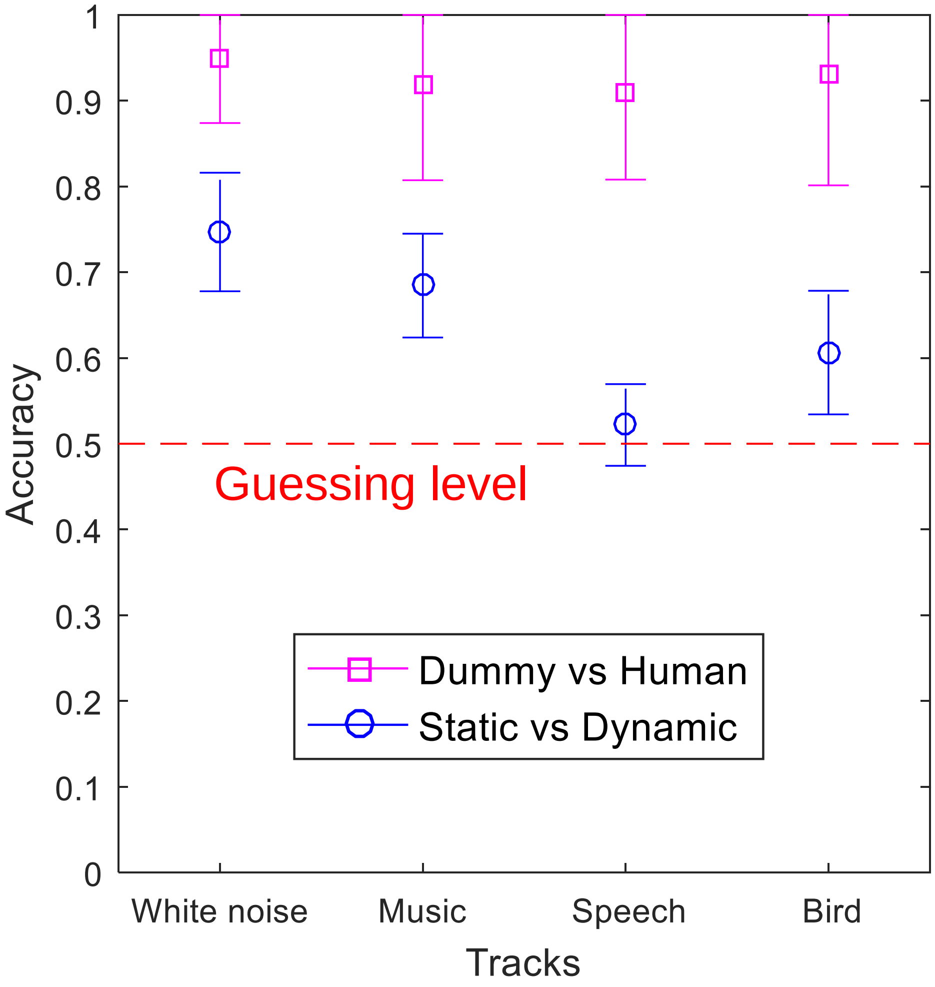

-90az;0el


-60az;0el

Static vs dynamic acquisition: ITD and ILD

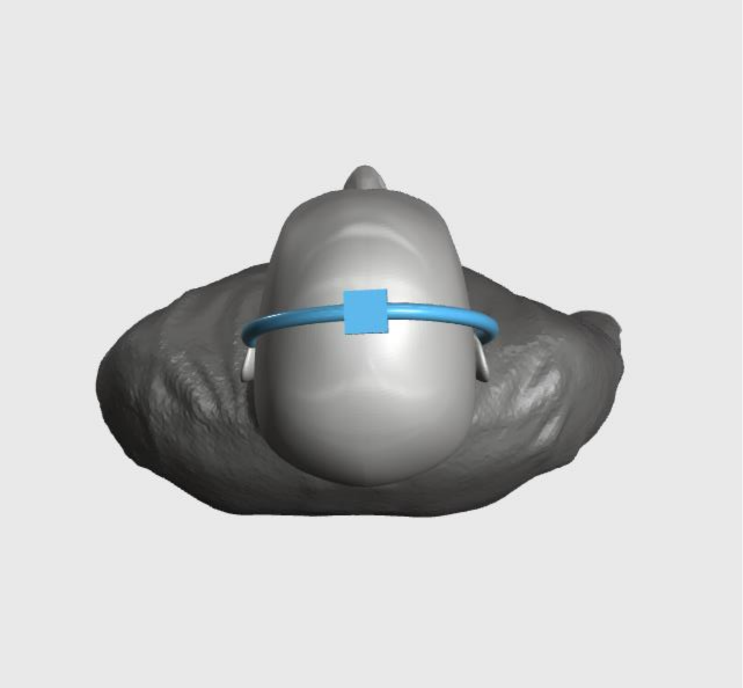


Static vs dynamic acquisition: Identification accuracy

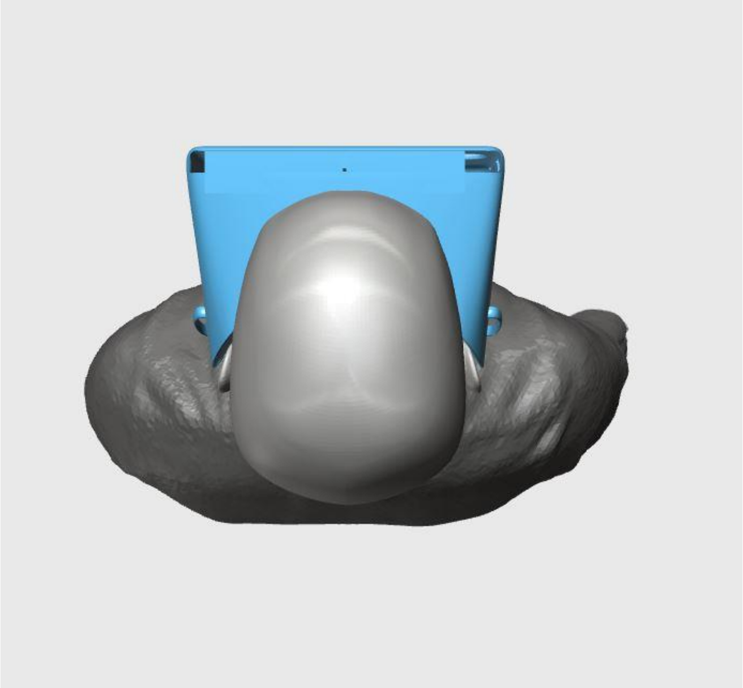


- ➢ Over 90% accuracy in identifying difference from human and dummy head.
- ➢ Only 50-75% accuracy in identifying static HRIR and dynamic HRIR.
- ➢ Most subjects reported that they need to focus to make the selection on static vs dynamic HRTF.

(2) Head tracker vs. VR Oculus Rift

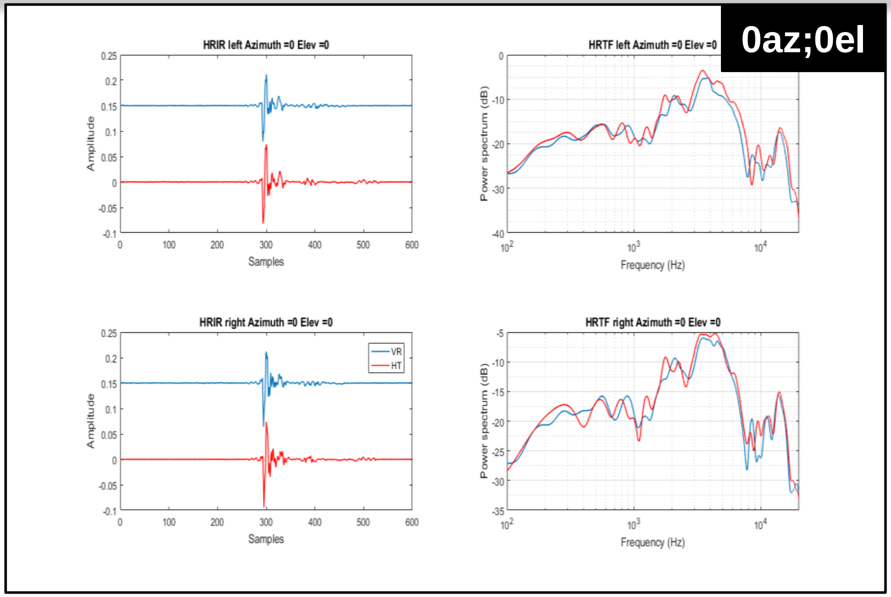


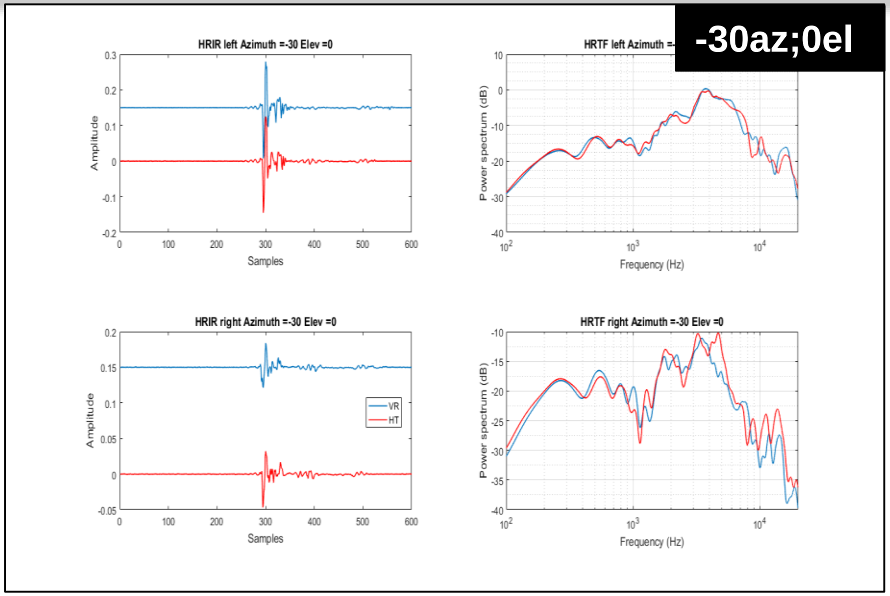
HRTF Acquisition using head-tracker

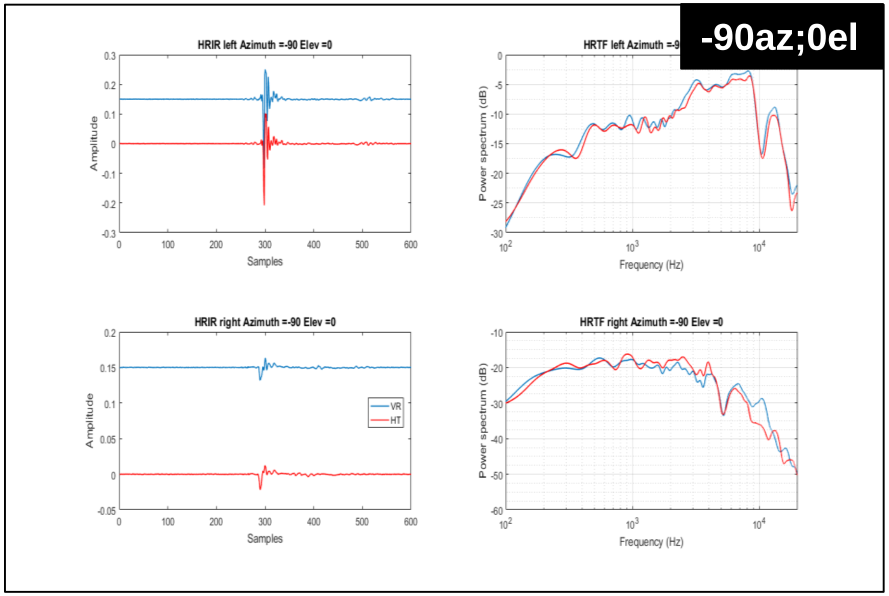


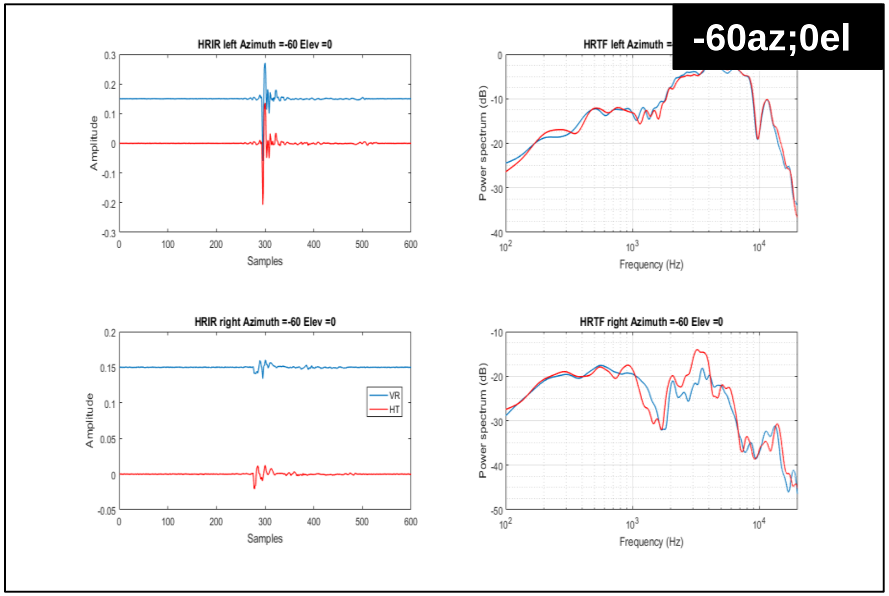
HRTF Acquisition using Oculus Rift

Head Tracker vs VR: HRIR/HRTF

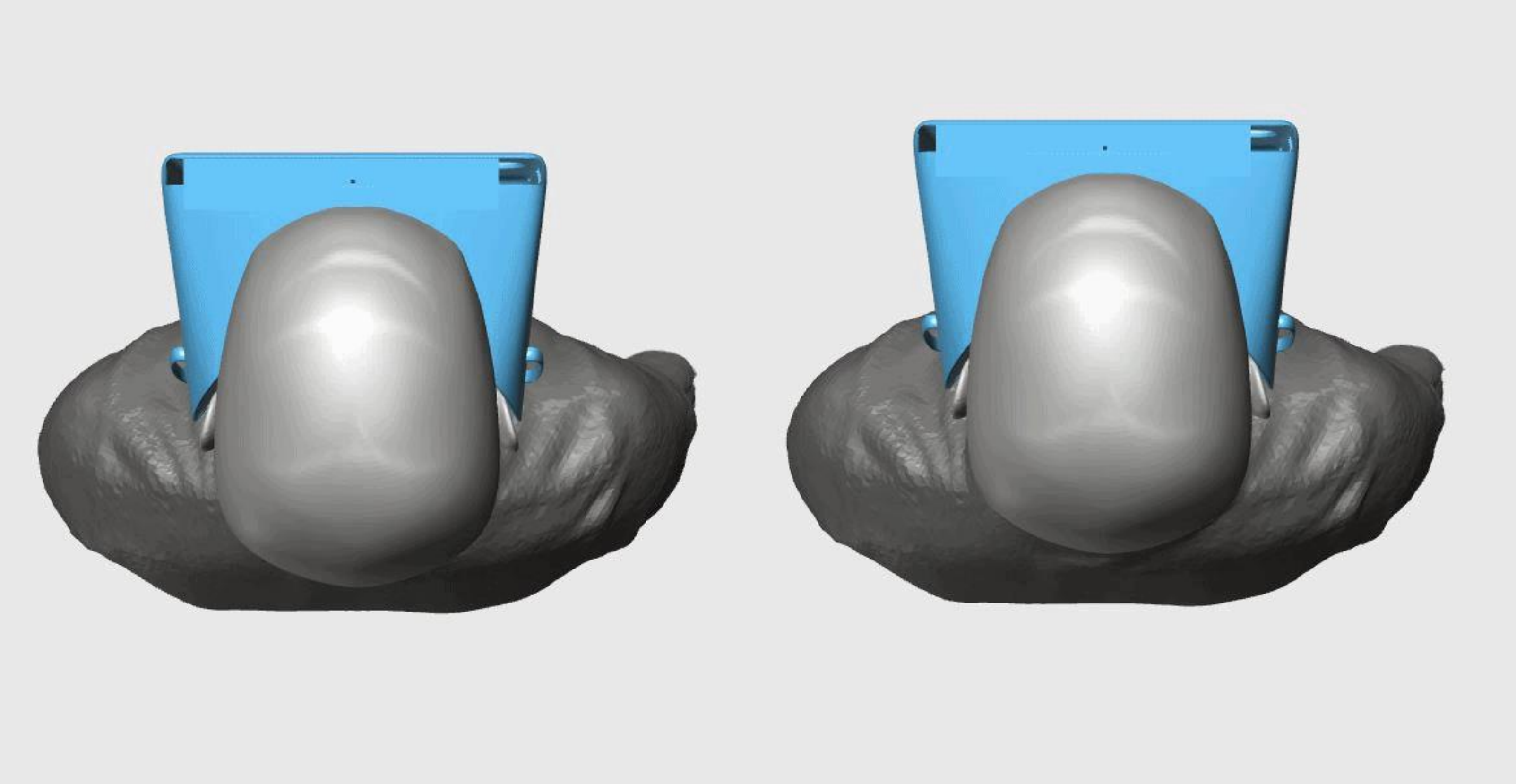

0az;0el


-30az;0el


-90az;0el


-60az;0el

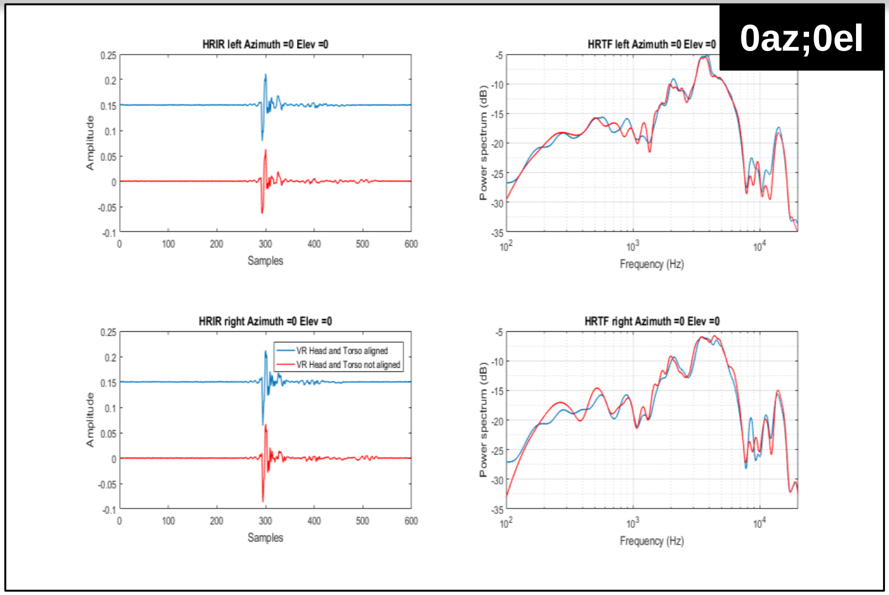
(3) Head Above Torso Orientation (HATO)

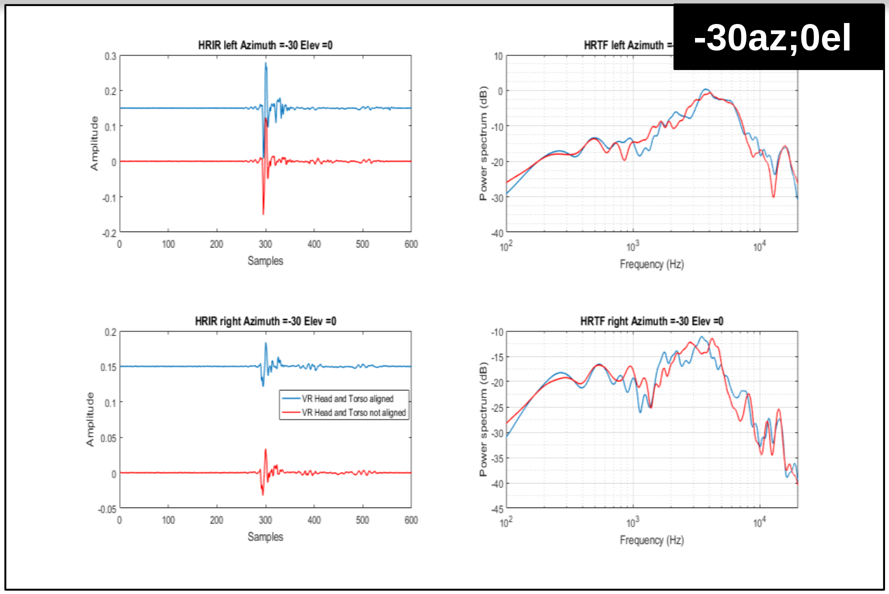


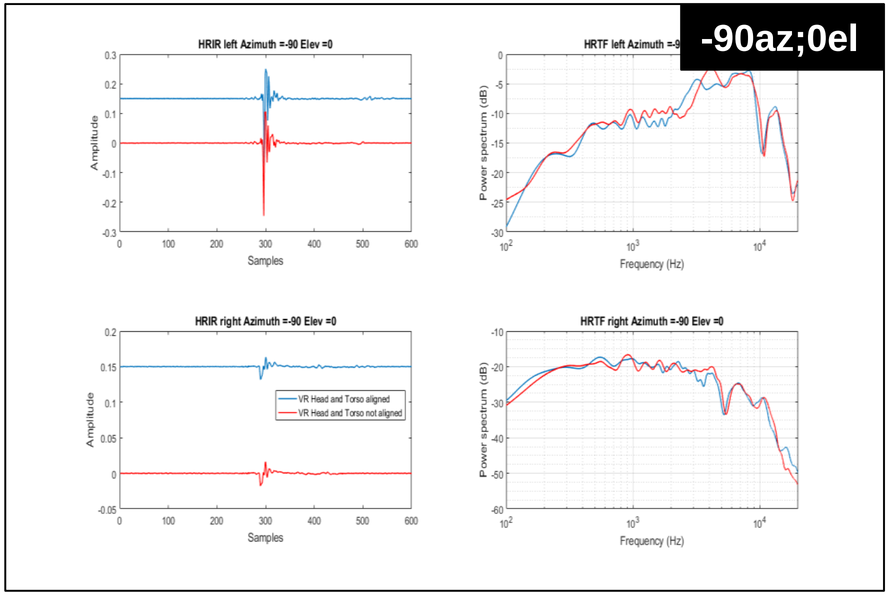
Head Above Torso Orientation Aligned

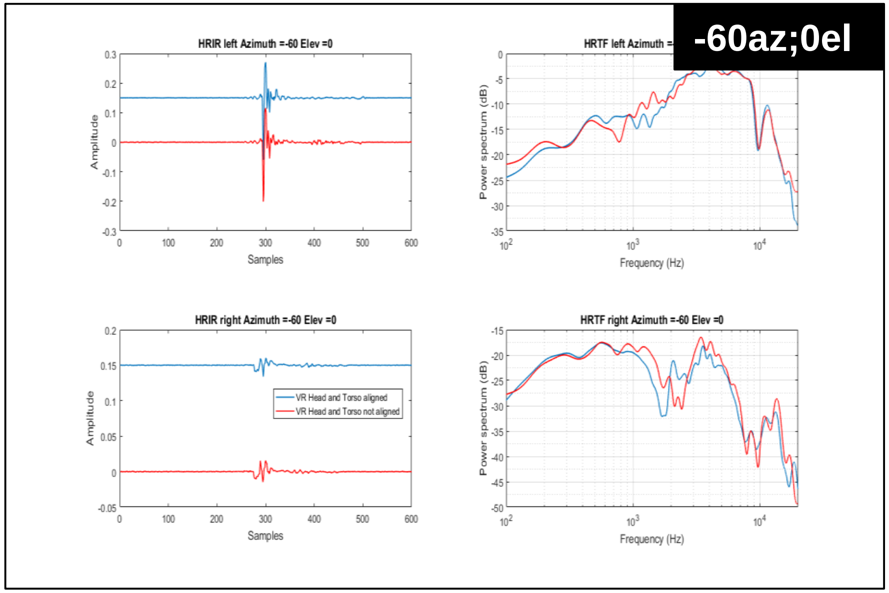
Head Above Torso Orientation Not-aligned (Torso Fixed)

HATO Aligned vs Not-aligned: HRIR/HRTF

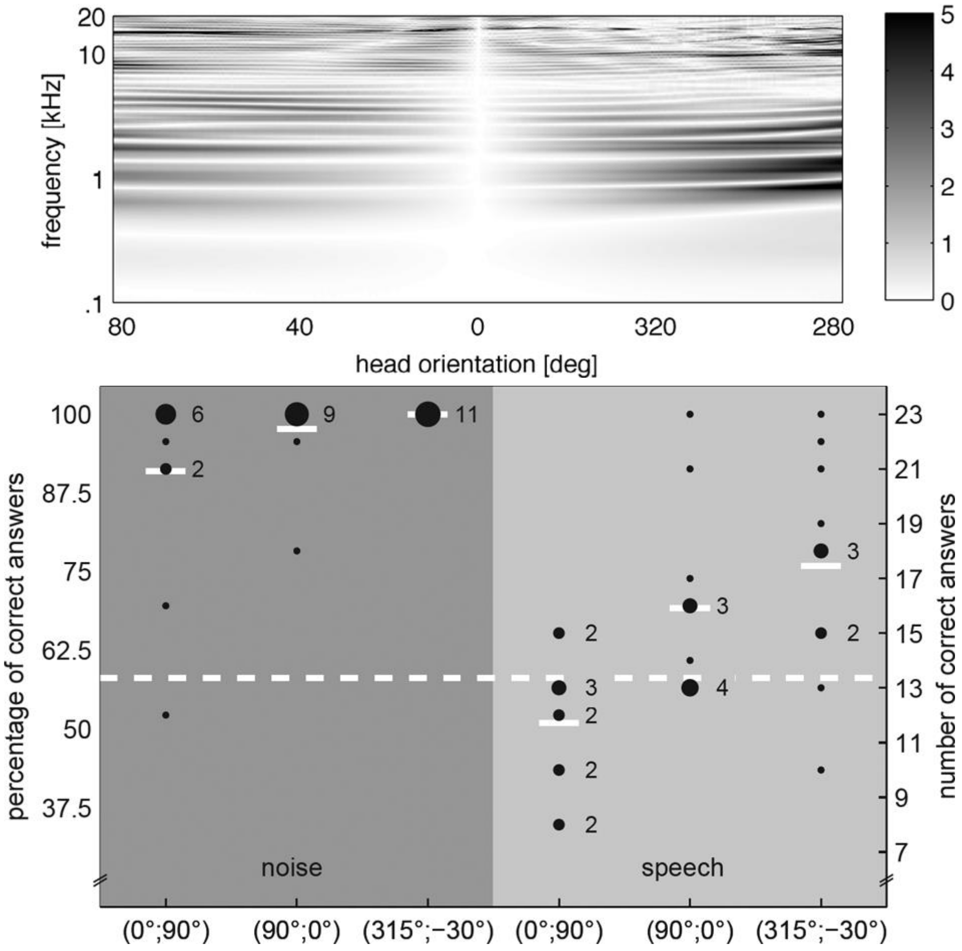








HATO difference on HRTF: physical and perceptual



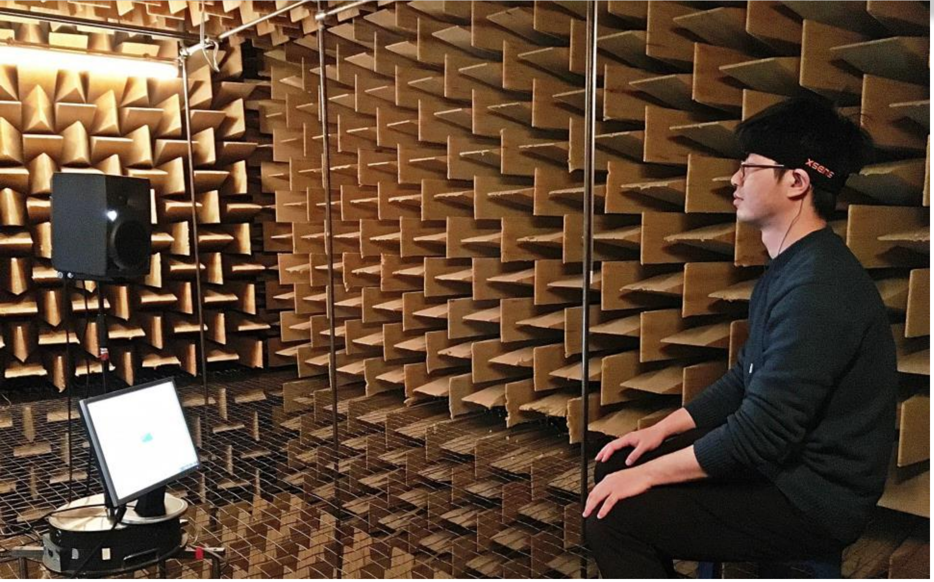
- ➢ More difference exists in
  - middle frequency
  - contralateral ear (280 to 360 degree)
- ➢ Easily distinguishable with broadband signals, but less difference with speech signals
- ➢ Subjects mentioned coloration and localization difference between 2 types of HRTFs

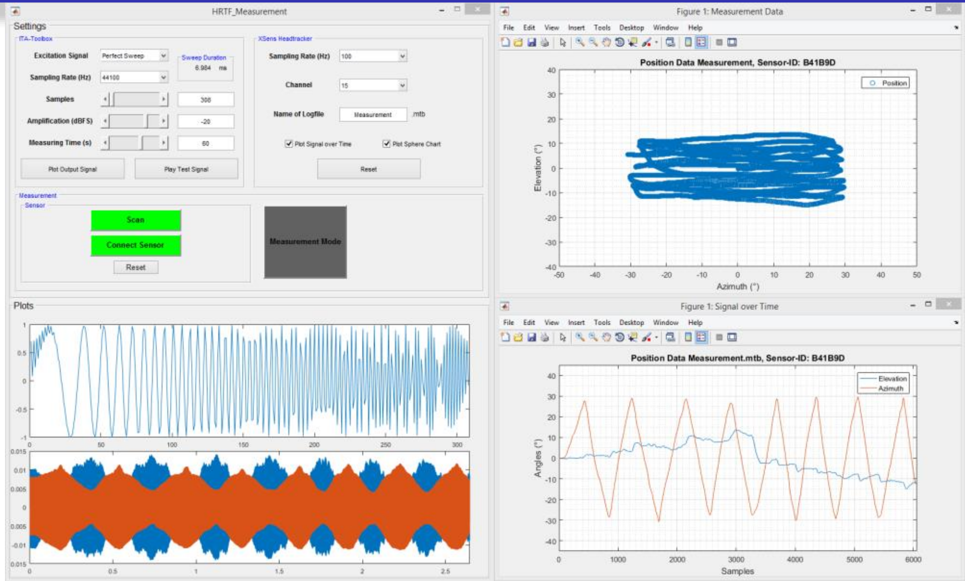
[Brinkmann, 2015]

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




HRTF_Measurement
Settings
ITA-Toolbox
XSens Headtracker
Plot Output Signal
Play Test Signal
Reset
Measurement
Sensor
Scan
Connect Sensor
Reset
Measurement Mode
Plots
Figure 1: Measurement Data
File Edit View Insert Tools Desktop Window Help
Position Data Measurement, Sensor-ID: B41B9D
Position
Elevation (°)
Azimuth (°)
Figure 1: Signal over Time
File Edit View Insert Tools Desktop Window Help
Position Data Measurement.mtb, Sensor-ID: B41B9D
Elevation
Azimuth
Angles (°)
Samples

Excitation Signal: Perfect Sweep

Sweep Duration: 6.984 ms

Sampling Rate (Hz): 44100

Samples: 308

Amplification (dBFS): -20

Measuring Time (s): 60

Sampling Rate (Hz): 100

Channel: 15

Name of Logfile: Measurement .mtb

☑ Plot Signal over Time ☑ Plot Sphere Chart

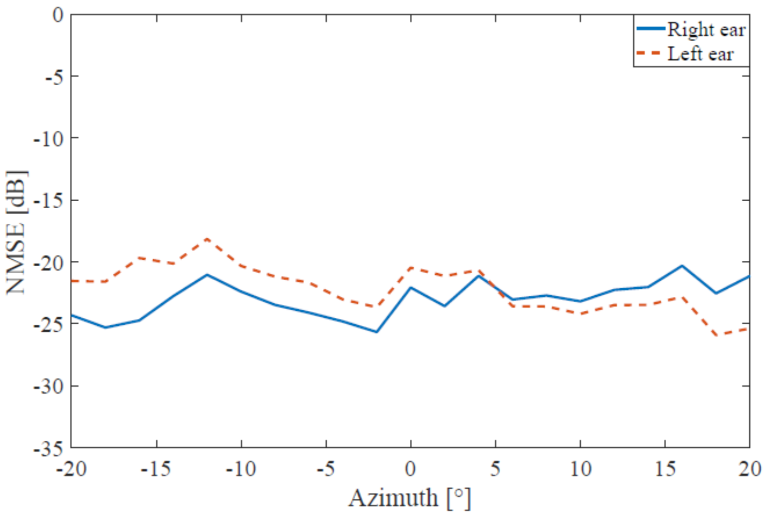


Figure 6: NMSE results for 1D head movements in azimuth

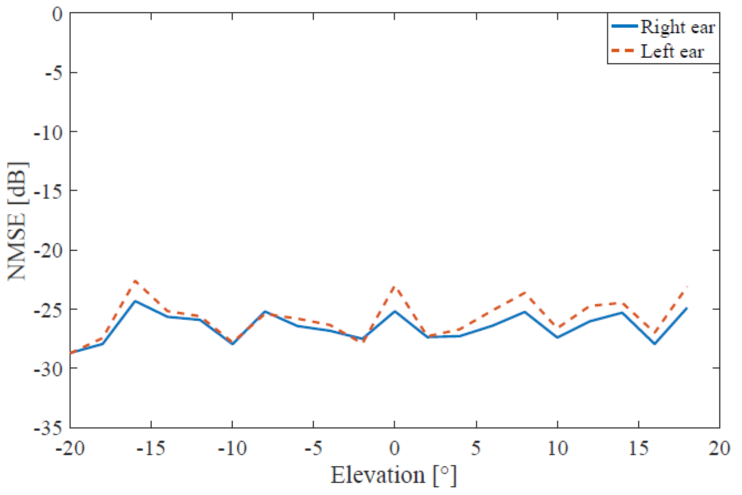
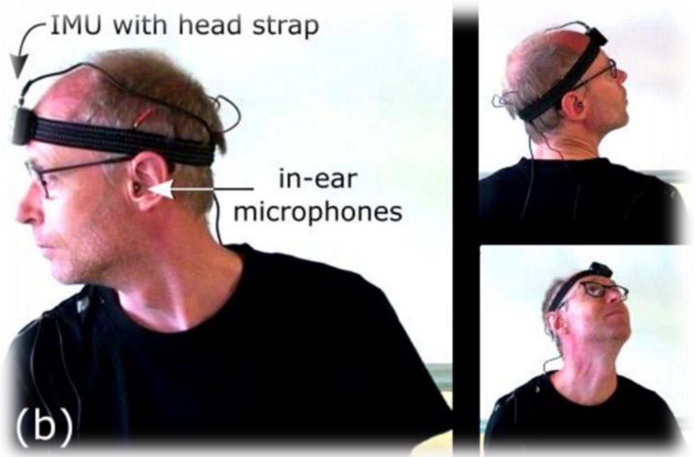


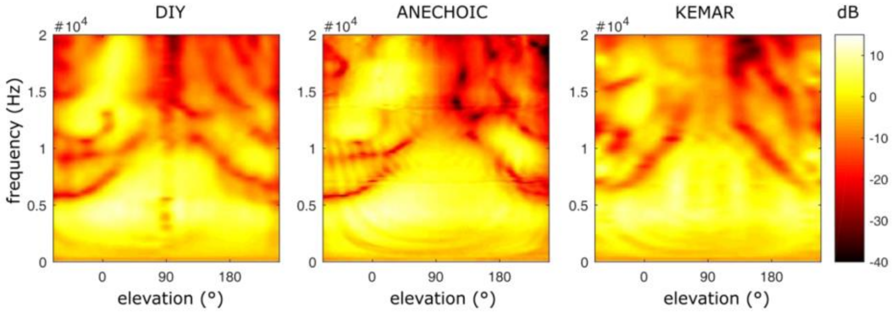
Figure 7: NMSE results for 1D head movements in elevation

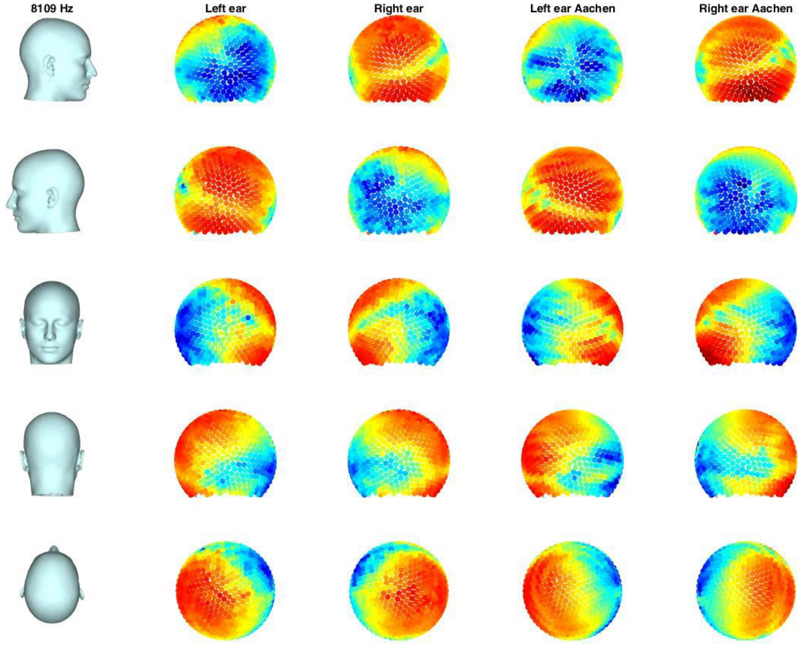
[Li, 2017]

Earfish's fast HRTF measurement system

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







https://www.earfish.eu/

[Reijniers, 2017]

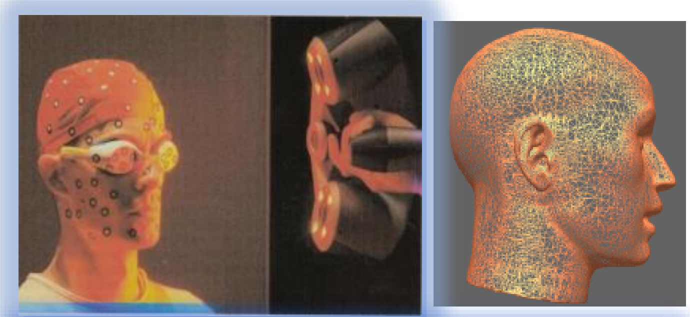
HRTF interpolation

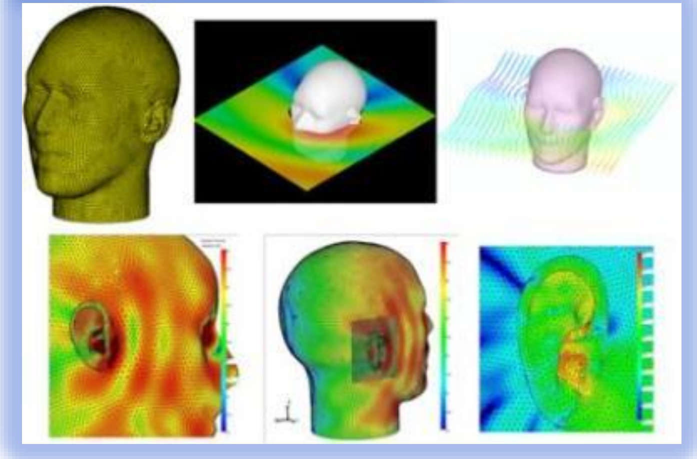
Directional resolution: measurement < rendering

Methods (domain):

- Directional $$\hat{H}_X = \sum_{i=1}^{I} w_i H_i$$
  - minimum phase, magnitude/phase [Nam, 2008; Jot, 1995]
  - Use nearest 2, 3, 4 directions [Xie, 2013; Gamper, 2013]
  - Better performance with denser input, lower frequency, ipsilateral ear [Christensen, 1999]
- Spectral basis (e.g., PCA) [Martens, 1987]
- Spatial basis (e.g., spherical harmonics) [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

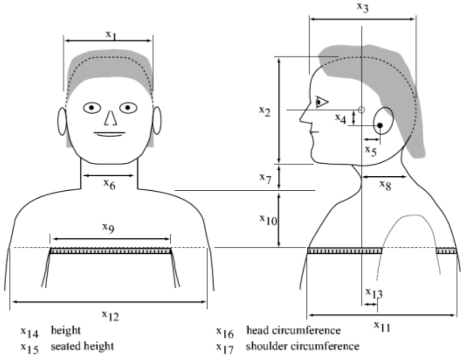


Figure 2: *Head and torso measurements*

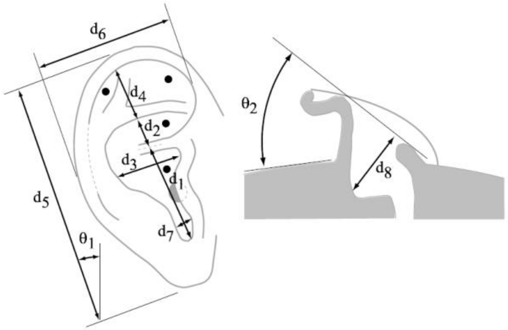
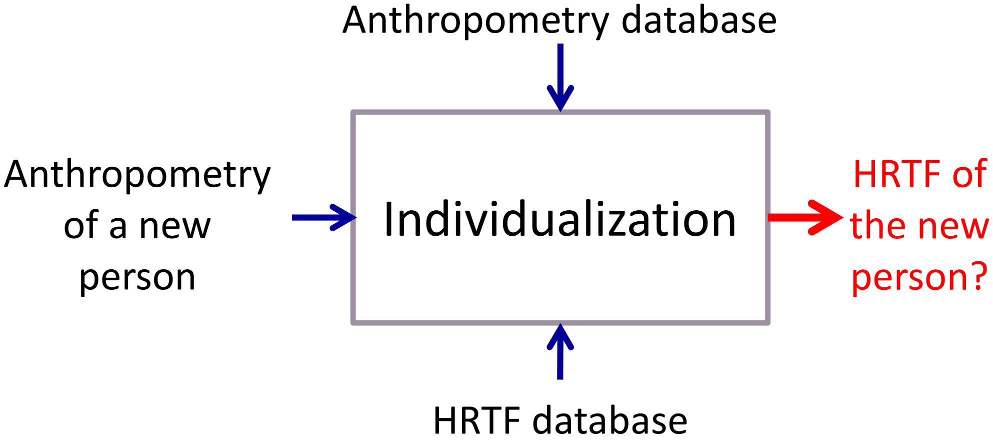


Figure 3: *Pinna measurements*



Anthropometry database

Anthropometry of a new person

Individualization

HRTF of the new person?

HRTF database

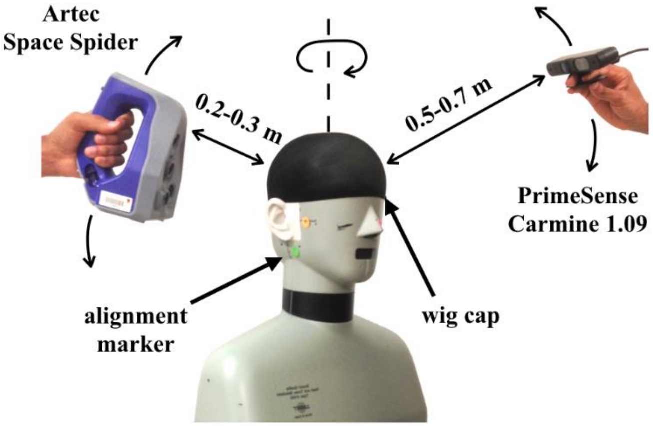
HRTF database with 3D models

ITA HRTF-database

48 subjects, head and ear models (FMRI)

http://www.akustik.rwth-aachen.de/go/id/lsly

[Bomhardt, 2016]



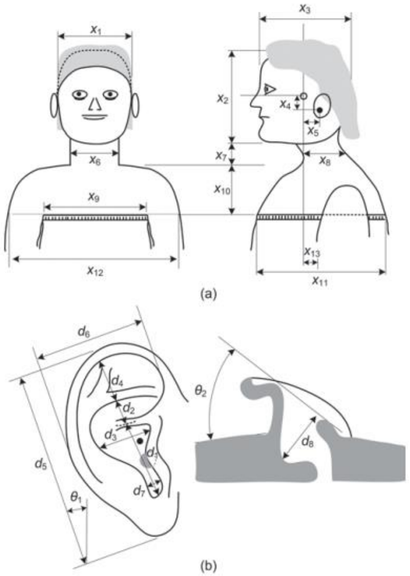
Princeton University 3D3A Lab (using 3D camera sensor and blue light)

http://www.princeton.edu/3D3A/HRTFMeasurements.ht

[Sridhar, 2017]

HRTF databases with anthropometry

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



Head width: strong correlation with ITD

[Middlebrooks, 1999; Xie, 2007]

Significant anthropometric parameters

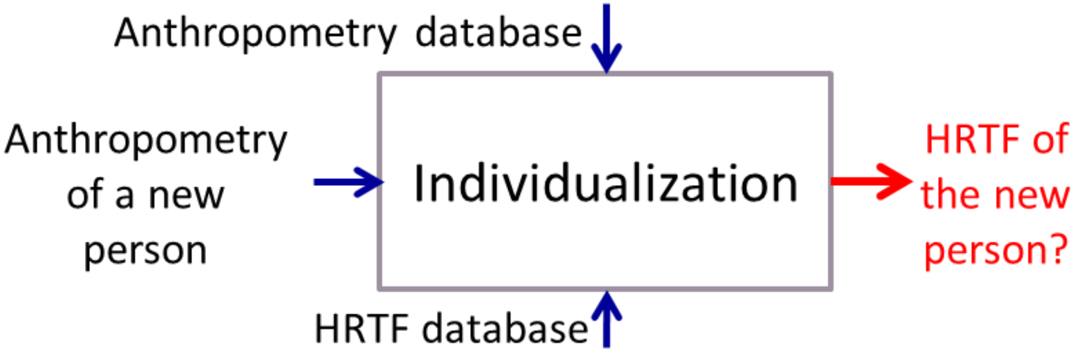
- Distance between ear and shoulder, breadth of head and back vertex; breadth and depth of cavum conchae and rotation of ears [Fels, 2004]
- head depth, pinna offset back, cavum concha, width, fossa height, pinna height, pinna width, pinna rotation angle and pinna flare angle

[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 [Zotkin, 2003]
  - Linear sparse representation of A features [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]



Anthropometry database

Anthropometry of a new person

Individualization

HRTF of the new person?

HRTF database

Performance on machine learning methods

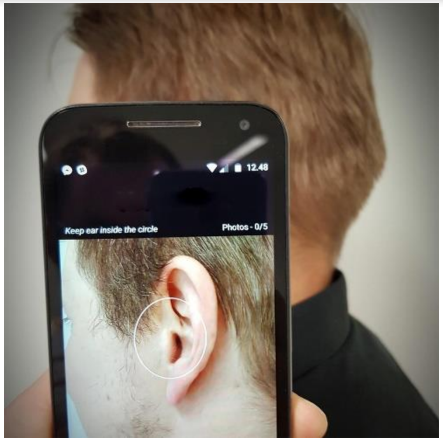
Method	Mean spectral distortion SD (dB)
PCA + NN [Zhou, 2008]	<3
SVD + RBF NN [Li, 2013]	~3
Isomap + NN [Grijalva, 2016]	4.6
NN [Favek, 2017]	3

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.

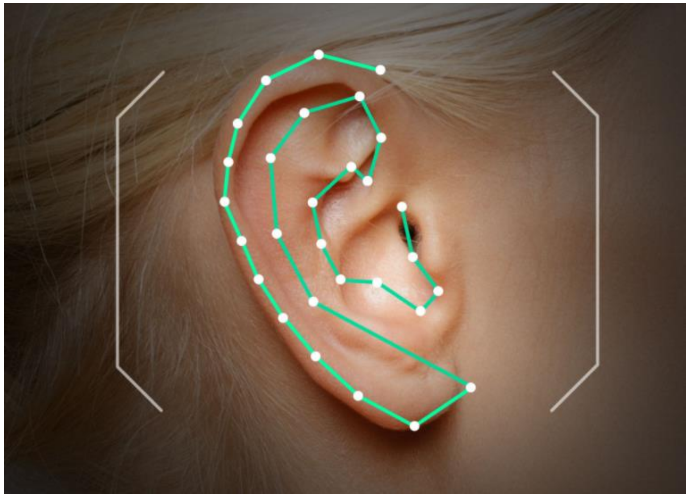
Commercial examples

Microsoft Windows 10 & Hololens



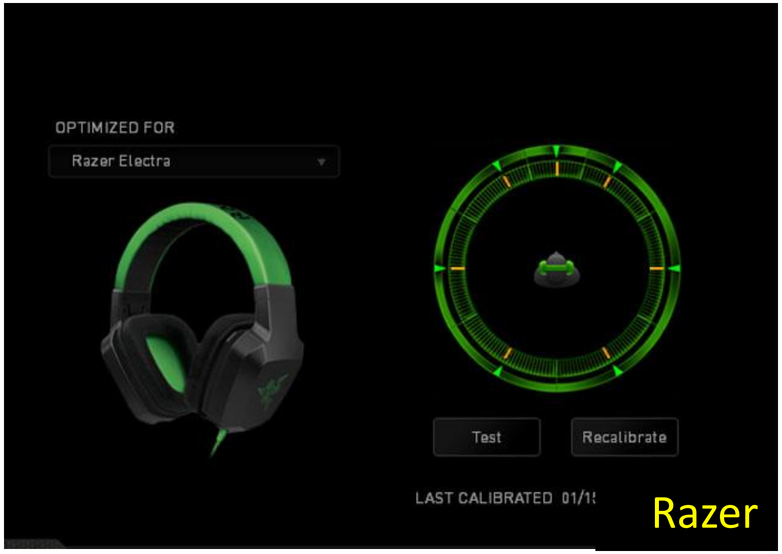
IDA audio

3D Sound labs

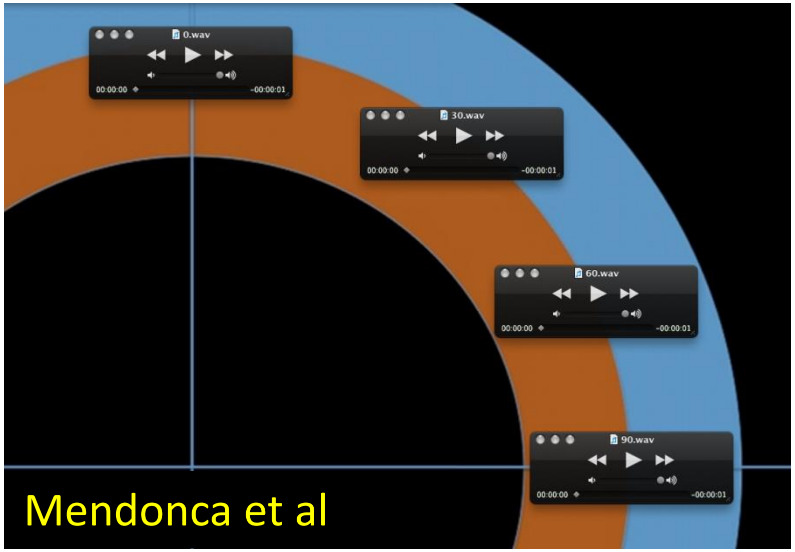


Creative Labs Super X-FI

Individualization: training/tuning via listening



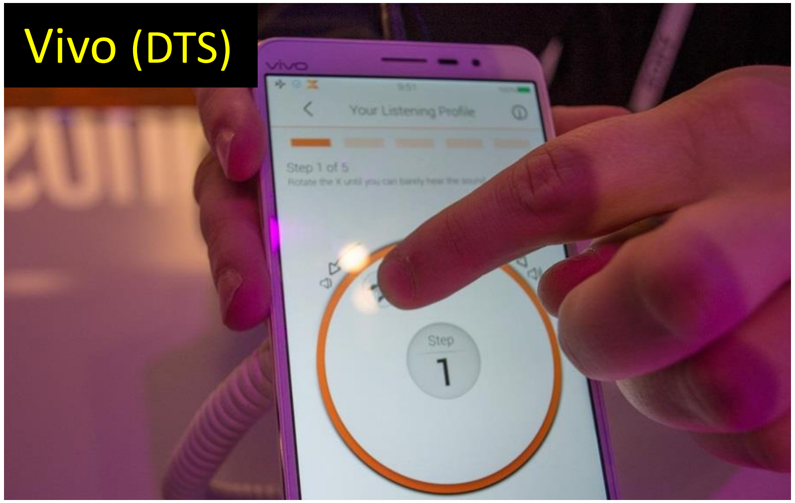
Razer



Mendonca et al

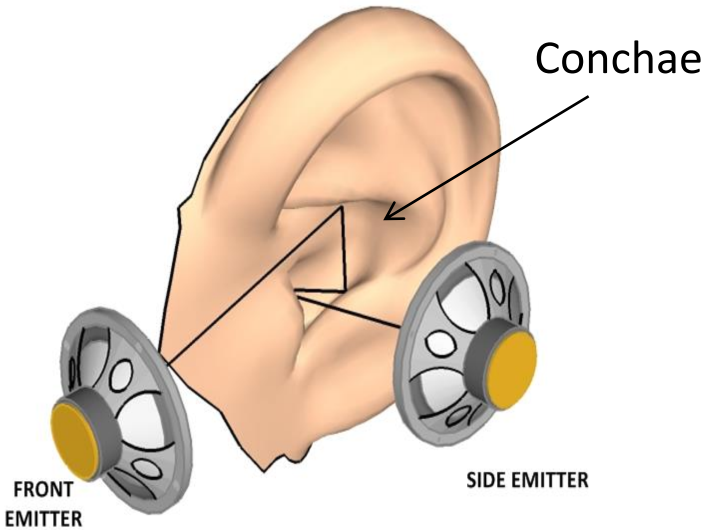
PERSONALIZE YOUR LISTENING EXPERIENCE
DIAL DOWN THE VOLUME UNTIL YOU CAN BARELY HEAR THE SOUND. THEN PRESS NEXT
STEP 3

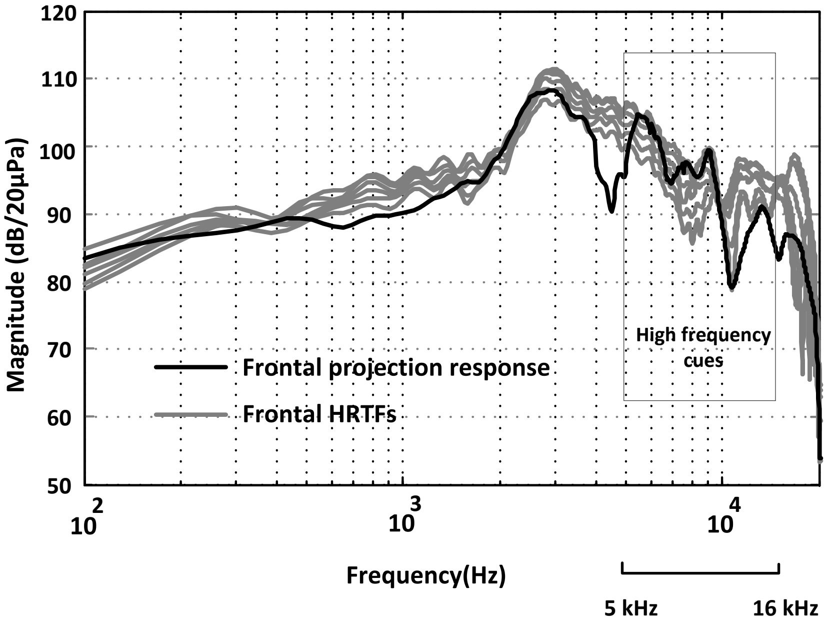
DTS



Vivo (DTS)

Individualization: headphone projection

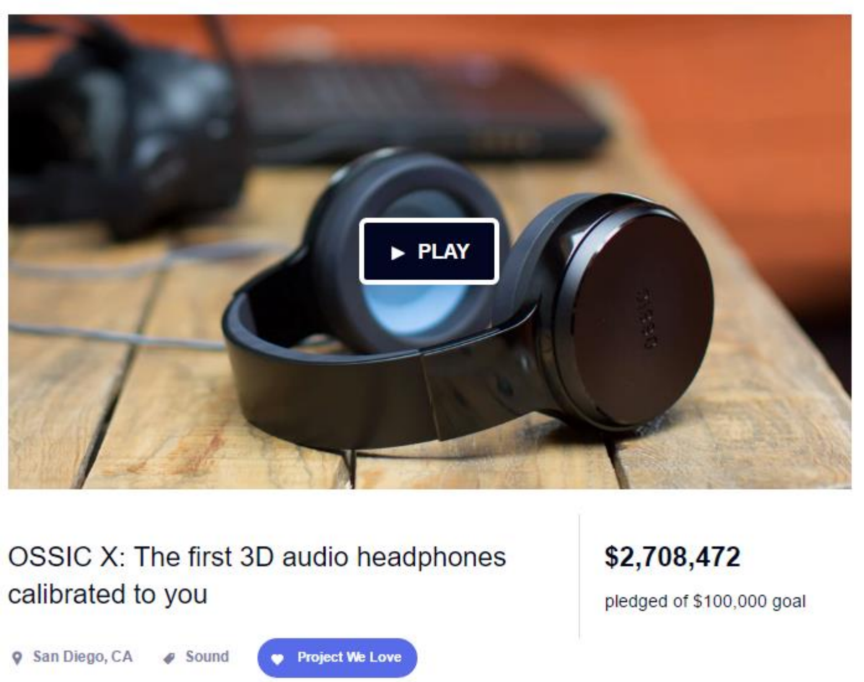




- No additional measurements and listening experiments required
- Reduce front-back confusion by > 50%;
- Zero user effort, plug and play (automatic during playback)

[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

San Diego, CA · Sound · Project We Love

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

Techniques		Resources	User contribution	Performance
Acoustic	Static	5	5	5
	Dynamic	5	4	4
Anthropometric	3D model	4	2	3
	Features	2	2	2
Listening	Training	1	3	2
	Tuning	1	3	3
Headphone Projection		3	1	3
Non-individualized HRTFs		0	0	1

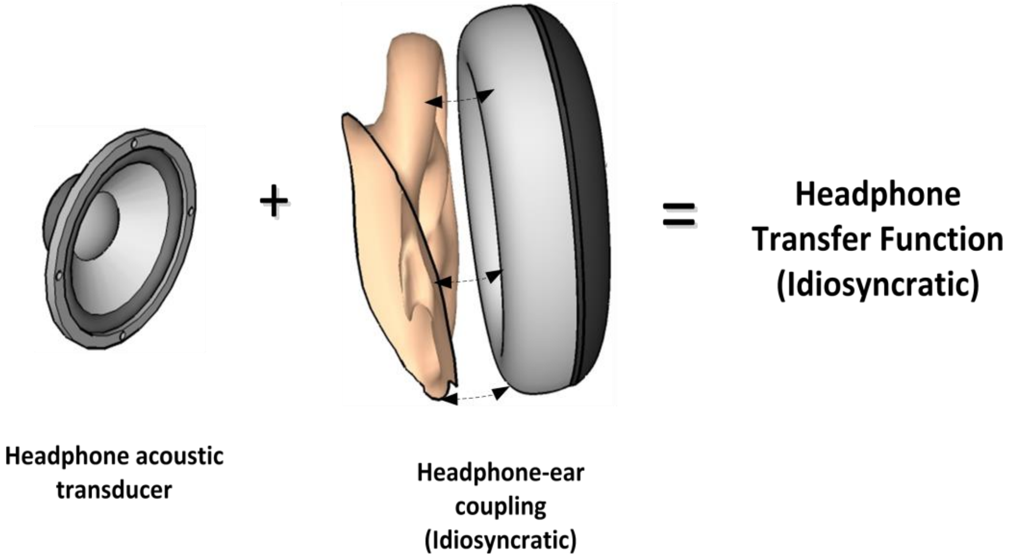
⚠ **The numbers are for illustration of our qualitative relative opinion purpose only**

- 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


+
=
Headphone Transfer Function (Idiosyncratic)
Headphone acoustic transducer
Headphone-ear coupling (Idiosyncratic)

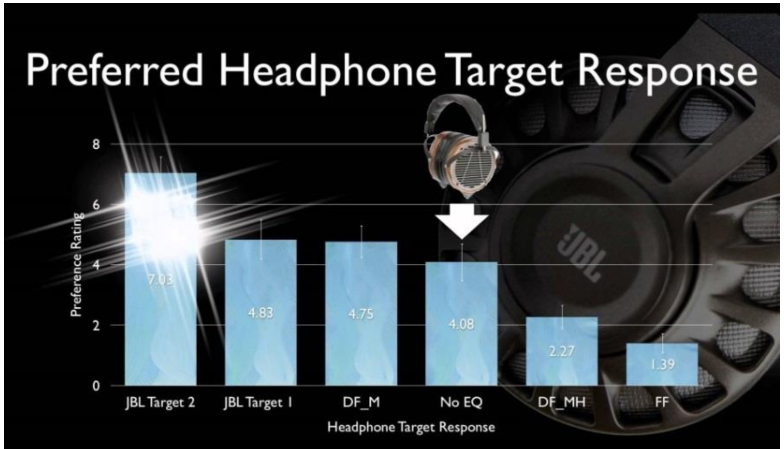
Breakdown of headphone transfer function (HPTF)

[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



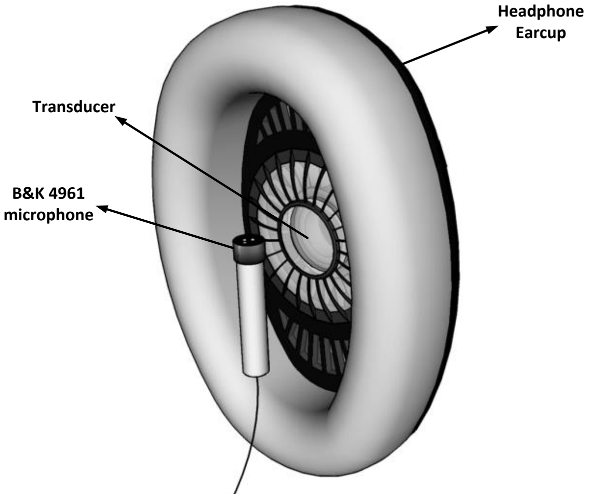
Source from http://seanolive.blogspot.sg/2014/01/the-perception-and-measurement-of.html

[Olive, 2013]

Non-decoupled equalization for binaural

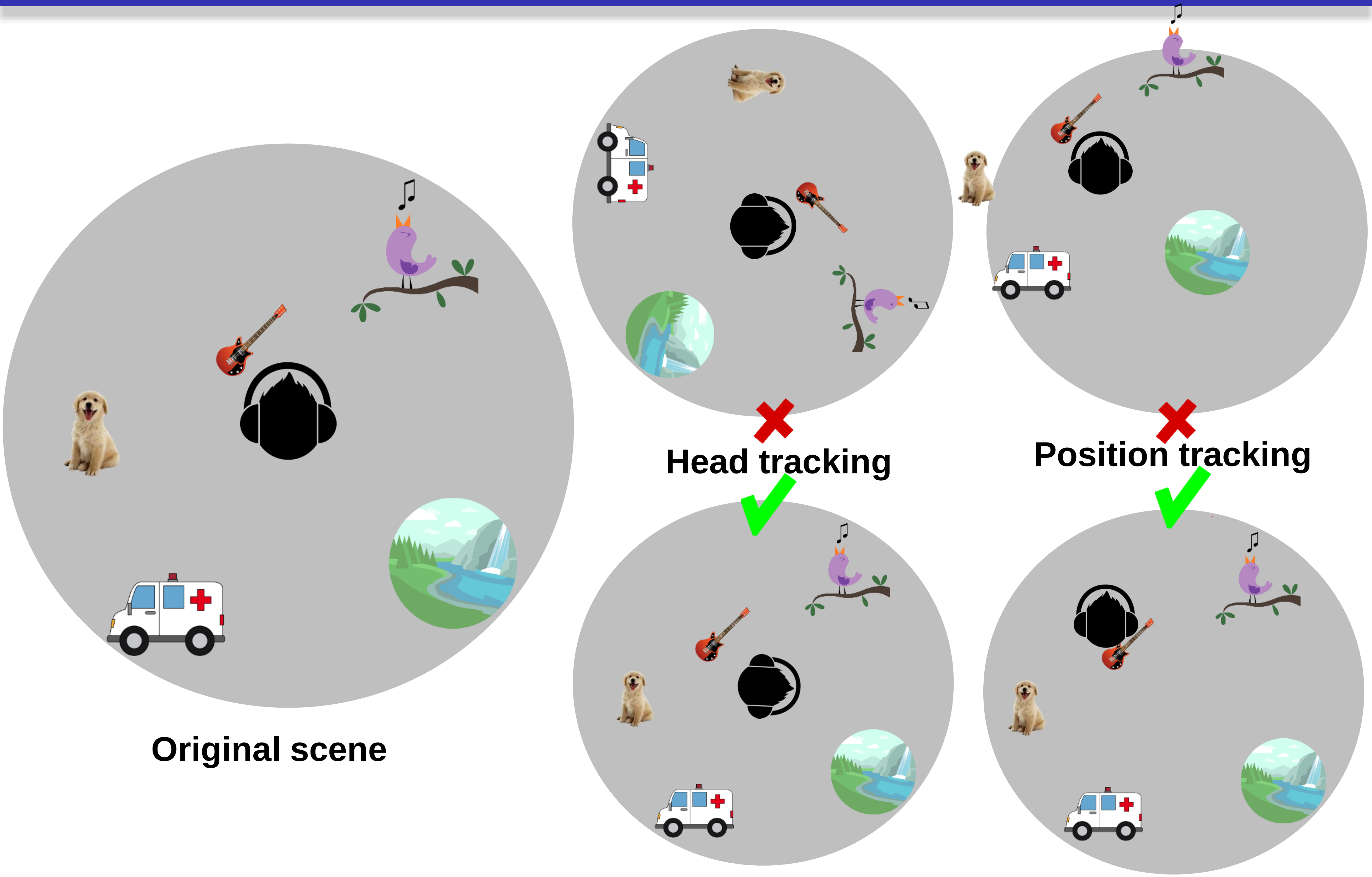
Aim: Spectrum at eardrum is the individual HRTF features

- 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



[Pralong, 1996; Kulkarni, 2000; Larcher, 1998; Sunder, 2013; Kirkeby, 1999; Norcross, 2004; Lindau, 2012; Gomez-Bolanos, 2016]

B.5 Movement Tracking

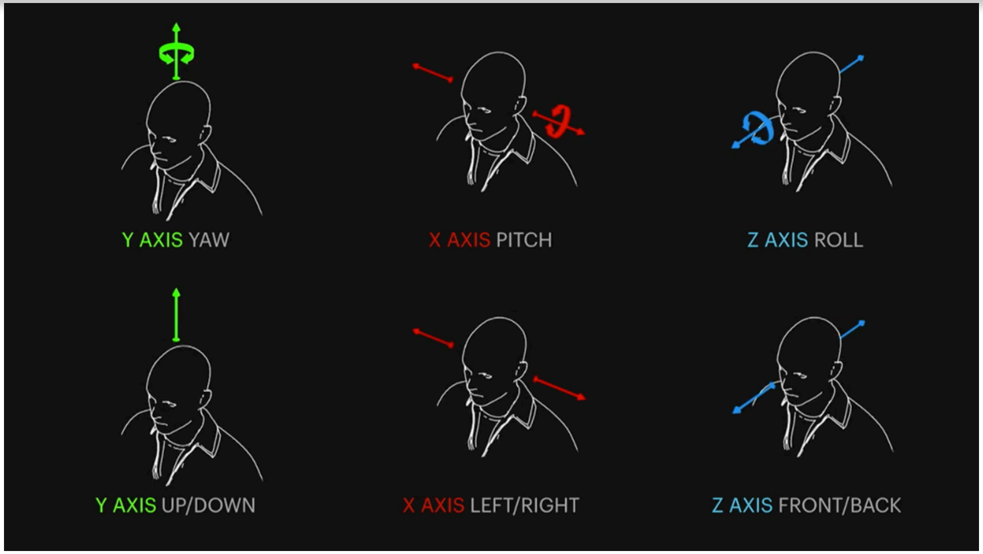


Original scene

Head tracking

Position tracking

6 DOF movement tracking



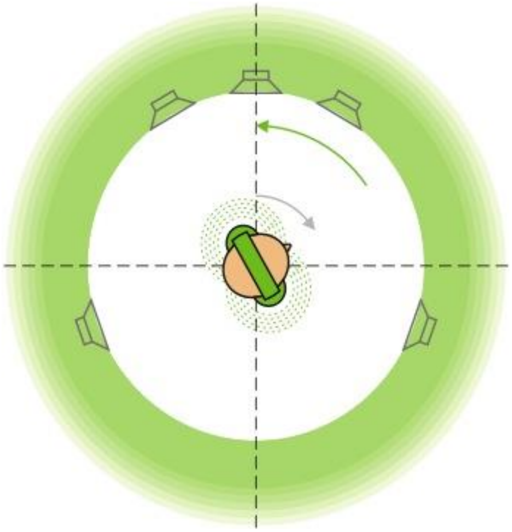
- Track all 6 DOF ideally but could be simplified
- Positional tracking
  - Camera based, laser based techniques
  - Affect direction, distance, diffraction perception
  - Perceptual effects on localization accuracy and latency need more investigation

https://developers.facebook.com/videos/f8-2017/surround-360-beyond-stereo-360-cameras/

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 [Sandvad, 1996]
- Concern of head tracking latency: <100ms (variation high)

Source from http://3dsoundlabs.com/en/



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

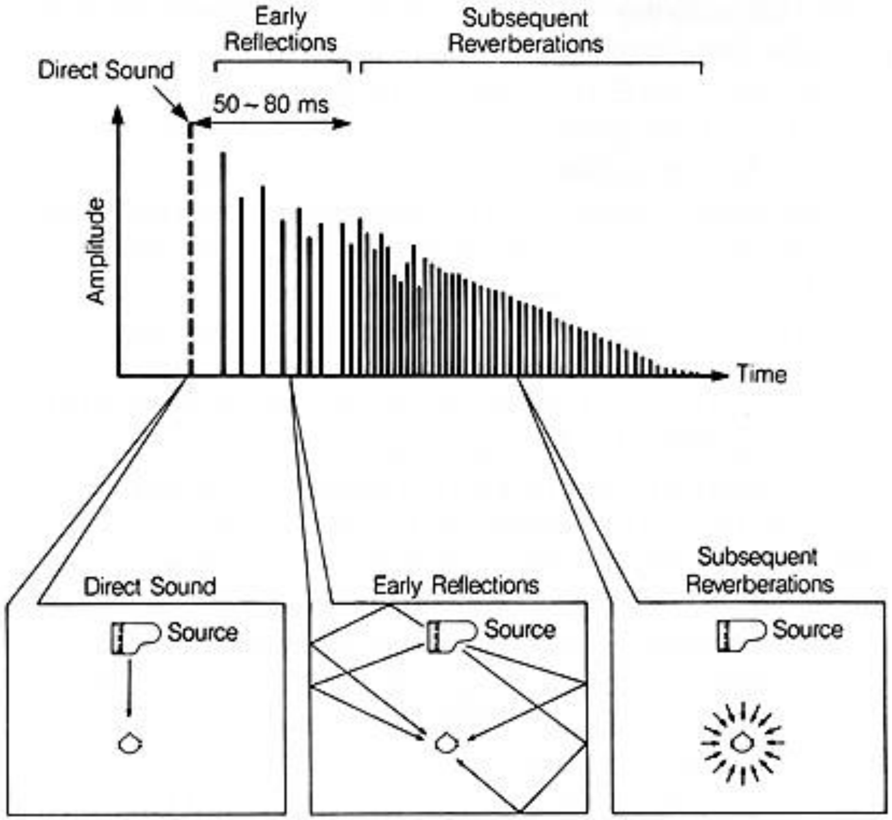
Head tracking improves source localization

- **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]

B.6 Environment rendering

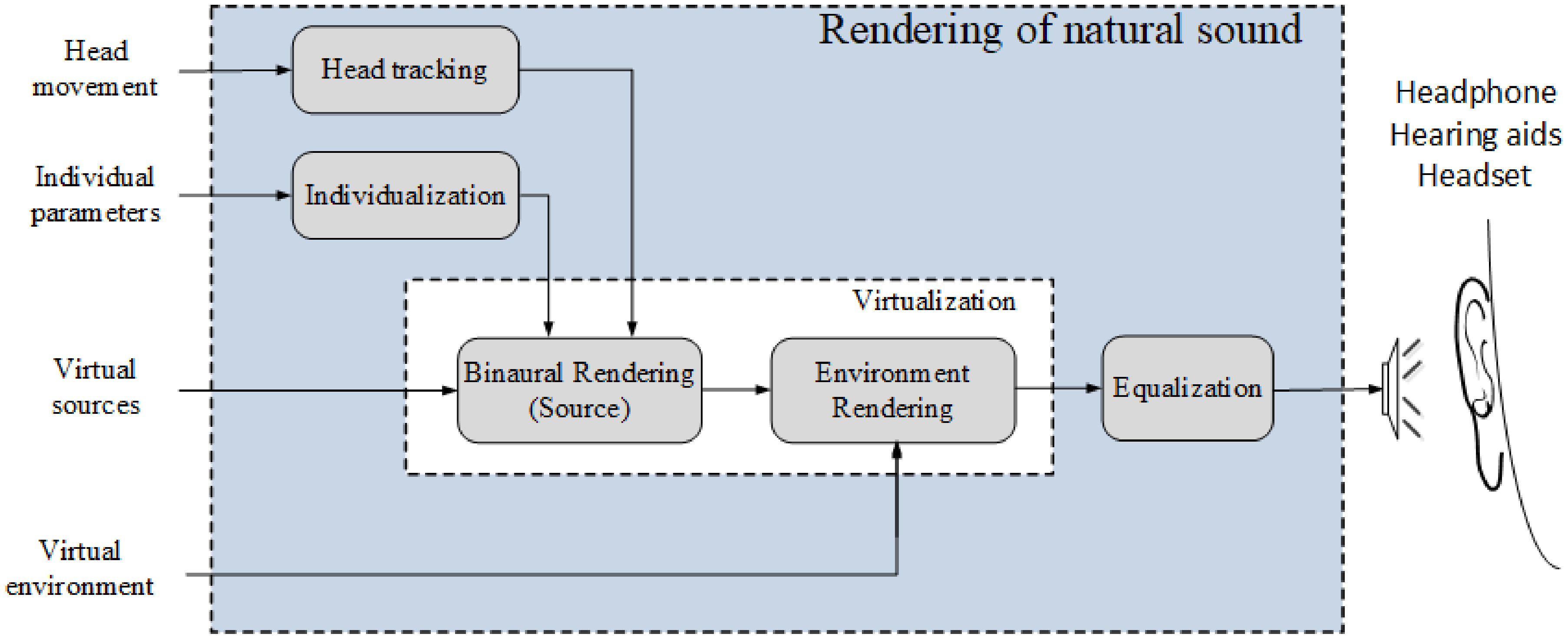
- **Apply artificial reverberation to binaural rendering**
  - Externalization of the sound sources, and enhance depth perception;
  - Rendering of the sound environment.

More on this during the 2nd half of this tutorial



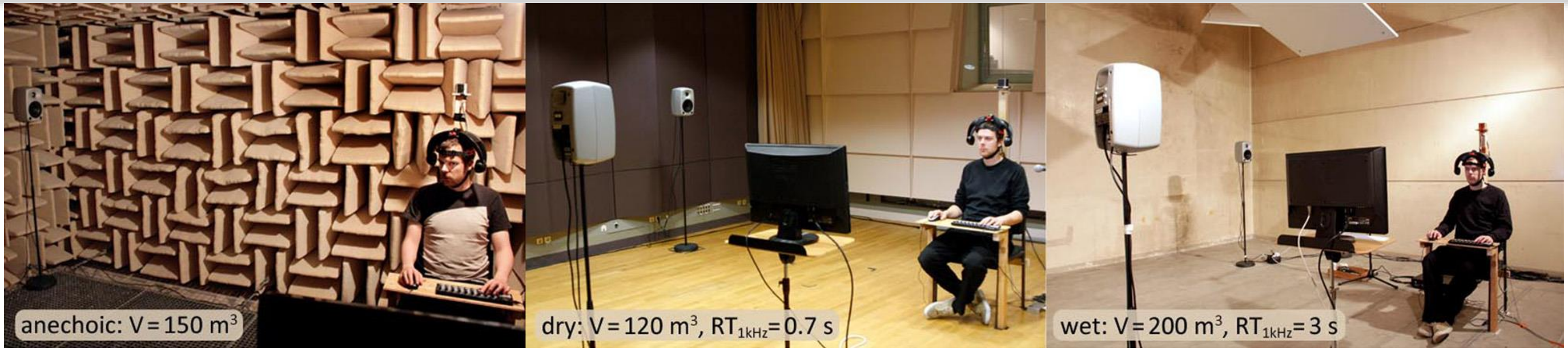
Source from
http://www.torgny.biz/Recording%20sound_2.htm

B.7 Integration



[Sunder, 2015]

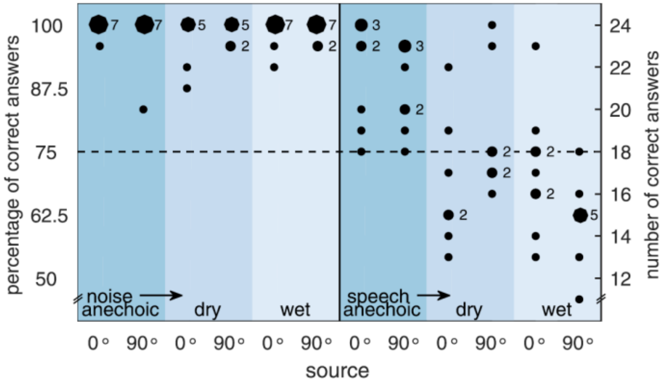
Authenticity of individual dynamic binaural synthesis


anechoic: V = 150 m³
dry: V = 120 m³, RT1kHz = 0.7 s
wet: V = 200 m³, RT1kHz = 3 s

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**:

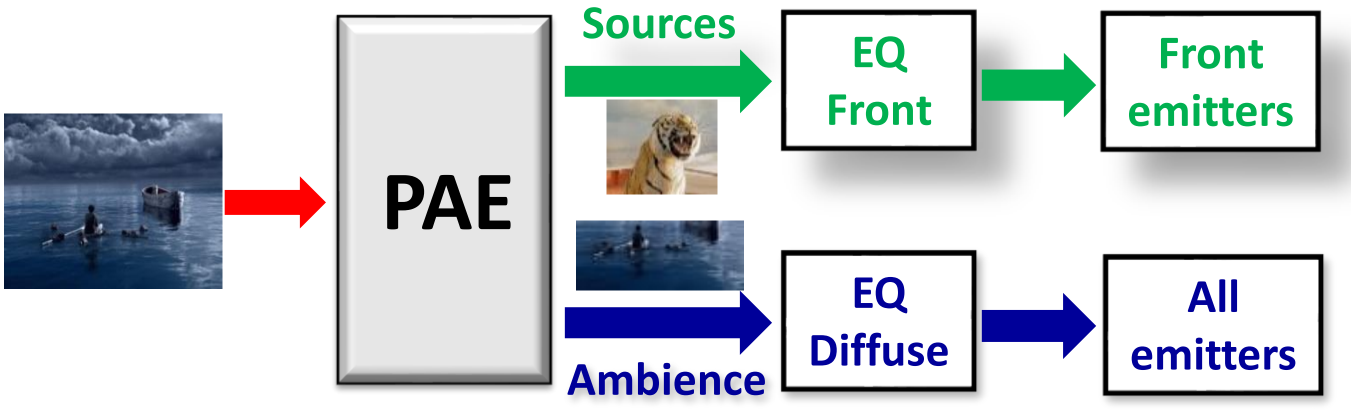
- Pink noise (all) > speech signals (half).
- Anechoic > reverberant, for speech.
- Coloration > localization.
- Dynamic > static.



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

[Brinkmann, 2017]

3D Audio Headphone: an example



PAE

Sources

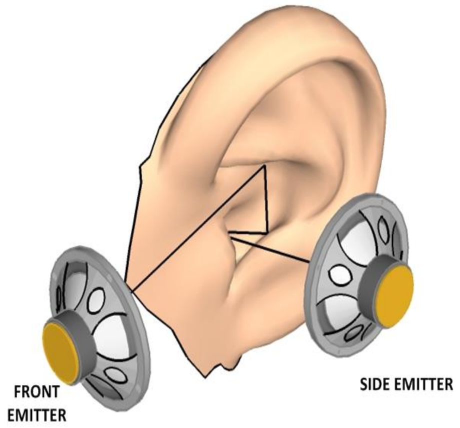
EQ Front

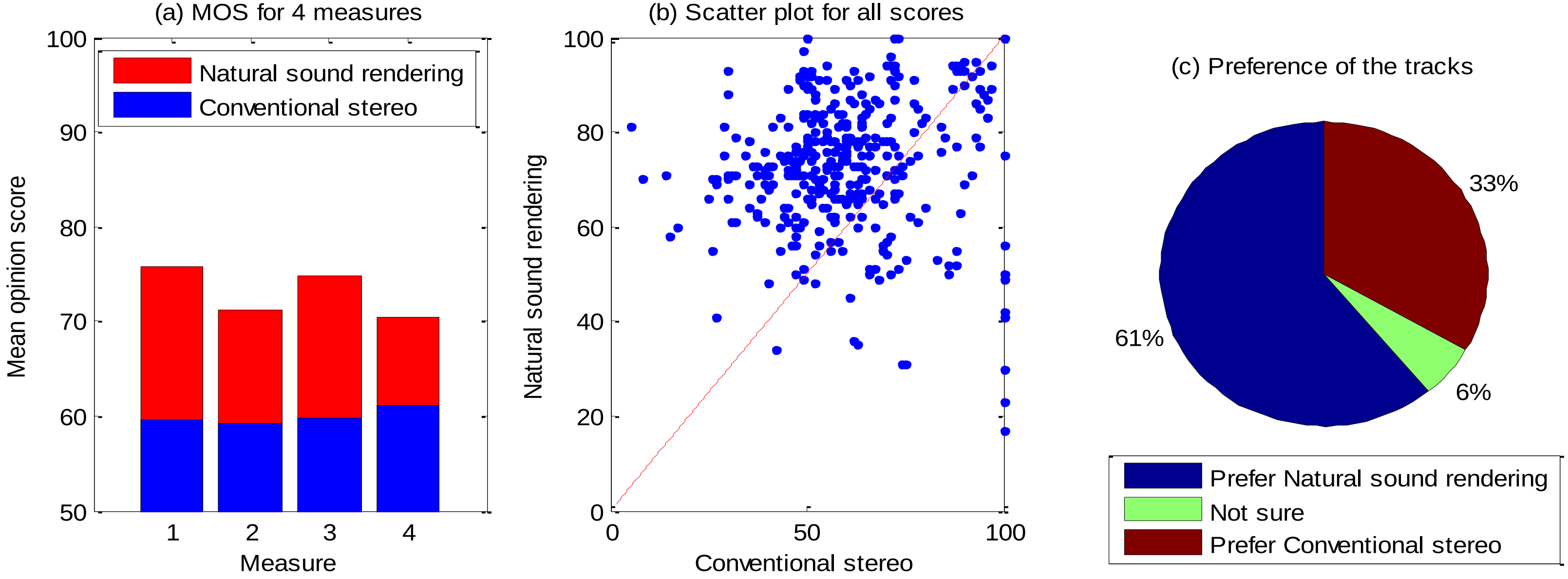
Front emitters

Ambience

EQ Diffuse

All emitters





(a) MOS for 4 measures

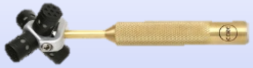
(b) Scatter plot for all scores

(c) Preference of the tracks

- **A grade higher in 4 measures**: Sense of direction, externalization, ambience, and timbral quality; **more preferred**.

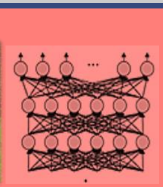
[Sunder, 2015]

Spatial Audio Technologies for Immersive VR, AR/MR



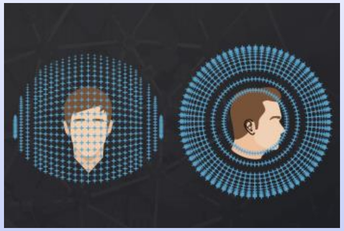
Spatial Audio Formats

- Object, Ambisonics
- Parametric processing



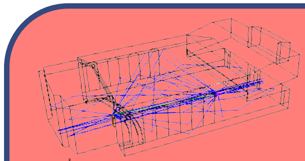
Environment Estimation

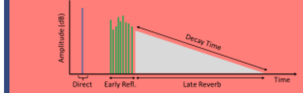
- Depth camera
- Reverberation fingerprint
- Machine learning



Individualized Binaural Rendering

- Individualized HRTFs
- Equalization





Environment Rendering

- Wave based
- Geometrical based
- Perceptual based

Y AXIS YAW
X AXIS PITCH
Z AXIS ROLL
Y AXIS UP/DOWN
X AXIS LEFT/RIGHT
Z AXIS FRONT/BACK

Dynamic Binaural Synthesis

- Head tracking
- Position tracking

Virtual & Physical Sound Fusion

- Adaptive equalization
- Hear-through processing

General references on binaural rendering

- D. R. Begault, *3-D sound for virtual reality and multimedia*: AP Professional, 2000.
- S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proc. IEEE,* vol. 101, no. 9, pp. 1920-1938, Sep. 2013.
- C. Faller and F. Baumgarte, “Binaural cue coding—part II: schemes and applications,” *IEEE Trans. Speech Audio Process.,* vol. 11, no. 6, pp. 520–531, Nov. 2003.
- V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *Signal Processing Magazine, IEEE,* vol. 28, no. 1, pp. 33-42, Jan. 2011.
- R. Nicol, *Binaural Technology*: AES, 2010.
- K. Sunder, J. He, E. L. Tan, and W. S. Gan, “Natural sound rendering for headphones,” IEEE Signal Processing Magazine, vol. 32, no. 2, pp. 100-113, Mar. 2015.
- L. Thresh, C. Armstrong and G. Kearney, “A Direct Comparison of Localisation Performance When Using First, Third and Fifth Order Ambisonics For Real Loudspeaker And Virtual Loudspeakear Rendering,” AES 143, New York, Oct.017.
- S. Xu, Z. Li, and G. Salvendy, "Individualization of head-related transfer function for three-dimensional virtual auditory display: a review," in Virtual Reality, ed: Springer, 2007, pp. 397-407.
- Fabian Brinkmann, Alexander Lindau, and Stefan Weinzierl, “On the authenticity of individual dynamic binaural synthesis,” The Journal of the Acoustical Society of America 142, 1784 (2017); doi: 10.1121/1.5005606

General references on binaural rendering

- Paul M. Hoffman, “Relearning sound localization with new ears,” nature neuroscience, volume 1 no. 5, september 1998
- S. Carlile (2014) The plastic ear and perceptual relearning in auditory spatial perception. Front. Neurosci. 8:237. doi: 10.3389/fnins.2014.00237
- D. Arteaga, “Introduction to Ambisonics,” Technical report, 2015
- B. Rafaely, Fundamentals of Spherical Array Processing, Springer, 2015
- J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, “MPEG-H audio–the new standard for universal spatial/3D audio coding,” J. Audio Eng. Soc., vol. 62, no. 12, pp. 821–830, Dec. 2013.
- K. Sunder, E. L. Tan, and W. S. Gan, "Individualization of binaural synthesis using frontal projection headphones," *J. Audio Eng. Soc.,* vol. 61, no. 12, pp. 989-1000, Dec. 2013.
- W. S. Gan and E. L. Tan, "Listening device and accompanying signal processing method," US Patent 2014/0153765 A1, 2014.
- 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
- V. Pulkki, "Spatial sound reproduction with directional audio coding," J. Audio Eng. Soc., vol. 55, no.6, pp. 503-516, Jun. 2007.
- A. Hyvärinen, J. Karhunen, and E. Oja, Independent component analysis. New York: John Wiley & Sons, 2004.
- O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Processing, vol. 52, no.7, pp. 1830-1847, Jul. 2004.
- T. Virtanen, "Sound source separation in monaural music signals," PhD Thesis, Tampere University of Technology, 2006.
- E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation," *IEEE Signal Processing Magazine,* vol. 31, no. 3, pp. 107-115, 2014.
- D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications.* NJ: Wiley-IEEE Press, 2006.
- J. NiKunen et al. “Binaural rendering of microphone array captures based on source separation,” Speech Communication, 2015.
- J. Merimaa, M. M. Goodwin, and J. M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Proc. 123rd Audio Engineering Society Convention,* New York, Oct. 2007.
- J. He, E. L. Tan, and W. S. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," *IEEE/ACM Trans. Audio, Speech, and Language Processing,* vol. 22, no.2, pp. 505-517, 2014.

References on parametric spatial audio processing

- J. Thompson, B. Smith, A. Warner, and J. M. Jot, “Direct-diffuse decomposition of multichannel signals using a system of pair-wise correlations,” in *Proc. 133rd Audio Eng. Soc. Conv.*, San Francisco, 2012.
- J. He, E. L. Tan, and W. S. Gan, “Primary-ambient extraction using ambient spectrum estimation for immersive spatial audio reproduction,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 9, pp. 1431-1444, Sept. 2015.
- J. He, W. S. Gan, and E. L. Tan, “Time-shifting based primary-ambient extraction for spatial audio reproduction,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 10, pp. 1576-1588, Oct. 2015.
- M. M. Goodwin and J. M. Jot, "Binaural 3-D audio rendering based on spatial audio scene coding," in Proc. 123rd Audio Engineering Society Convention, New York, Oct. 2007.
- J. Breebaart and E. Schuijers, "Phantom materialization: a novel method to enhance stereo audio reproduction on headphones," IEEE Trans. Audio, Speech, and Language Processing, vol. 16, no.8, pp. 1503-1511, Nov. 2008.
- C. Avendano and J. M. Jot, "A frequency-domain approach to multichannel upmix," J. Audio Eng. Soc., vol. 52, no.7/8, pp. 740-749, Jul. 2004.
- C. Faller, "Multiple-loudspeaker playback of stereo signals," J. Audio Eng. Soc., vol. 54, no.11, pp. 1051-1064, Nov. 2006.
- F. Menzer and C. Faller, "Stereo-to-binaural conversion using interaural coherence matching," in Proc. 128th Audio Engineering Society Convention, London, UK, May 2010.

References on HRTF measurement methods

- H. Møller, M. Sorensen, D. Hammershøi, and C. Jensen, "Head-related transfer functions of human subjects," J. Audio Eng. Soc., vol. 43, no. 5, pp.300–321, May 1995.
- V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in Proc. WASPAA, New Paltz, NY, USA, Oct. 2001.
- T. Carpentier, H. Bahu, M. Noisternig, O. Warusfel. "Measurement of a head-related transfer function database with high spatial resolution," Presetned in 7th Forum Acusticum(EAA), Krakow, Poland, Sep 2014.
- P. Majdak, P.Balazs, and B. Laback. "Multiple exponential sweep method for fast measurement of head-related transfer functions," J. Audio Eng. Soc., vol. 55, no. 7/8, pp. 623–637, Jul./Aug. 2007.
- P. Bilinski, J. Ahrens, M. R. P. Thomas, I. Tashev, and J. C. Plata, "HRTF magnitude synthesis via sparse representation of anthropometric features," in Proc. ICASSP, Florence, Italy, pp. 4501-4505, May 2014.
- R. Bomhardt, M. Klein, and J. Fels, "A high-resolution head-related transfer function and three-dimensional ear model Database," Proc. Mtgs. Acoust. 29, 050002 (2016); doi: 10.1121/2.0000467
- M. Pollow, B. Masiero, P. Dietrich, J. Fels, and M. Vorländer, "Fast measurement system for spatially continuous individual HRTFs," in Proc. Ambisonics, York, UK (2012).
- D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov, "Fast head-related transfer function measurement via reciprocity," J. Acoust. Soc. Amer., vol. 120, pp. 2202–22015, Oct. 2006.
- K. Fukudome, T. Suetsugu, T. Ueshin, R. Idegami, K. Takeya, "The fast measurement of head related impulse responses for all azimuthal directions using the continuous measurement method with a servo-swiveled chair," Applied Acoustics, 68(8):864–884, 2007.

References on HRTF measurement methods

- V. Pulkki, M. V. Laitinen, and V. P. Sivonen, “HRTF measurements with a continuously moving loudspeaker and swept sines,” in Proc. 128th AES Conv., London, UK, May 2010.
- G. Enzner, “Analysis and optimal control of LMS-type adaptive filtering for continuous azimuth acquisition of head related impulse responses,” in Proc. ICASSP, Las Vegas, NV, Apr. 2008.
- G. Enzner. “3D-continuous-azimuth acquisition of head-related impulse responses using multichannel adaptive filtering,” in Proc. WASPAA, New Paltz, NY, Oct. 2009.
- J. He, R. Ranjan, and W. S. Gan, “Fast continuous HRTF acquisition with unconstrained movements of human subjects,” in Proc. ICASSP, Shanghai, China, Mar. 2016, pp. 321-325.
- R. Ranjan, J. He, and W. S. Gan, “Fast continuous acquisition of HRTF in 2D for human subjects with unconstrained random head movements,” in Proc. AES Headphone conference, Aalborg, Denmark, Aug. 2016.
- Li, and, J. Peissig, “Fast Estimation of 2D Individual HRTFs with Arbitrary Head Movements,” Proc. 2017 22nd International Conference on Digital Signal Processing (DSP), London, Aug. 2017
- R. Braun, S. Li, and J. Peissig, “A Measurement System for Fast Estimation of 2D Individual HRTFs with Arbitrary Head Movements,” Proc. 4th International Conference on Spatial Audio, Graz, Austria, Sept. 2017.
- J. Reijniers, B. Partoens, and H. Peremans, “DIY measurement of your personal HRTF at home: low-cost, fast and validated,” in Proc. 143rd AES Conv., New York, Oct 2017.

References on HRTF interpolation

- Christensen F., MØller H., and Minnaar P., et al. (1999). “Interpolating between head-related transfer functions measured with low-directional resolution,” in AES 107th Convention, New York, USA, Preprint: 5047
- Wightman F.L., and Kistler D.J. (1992a). “The dominant role of low-frequency interaural time difference in sound localization,” J. Acoust. Soc. Am. 91(3), 1648-1661.
- Cheng C.I., and Wakefield G.H. (1999). “Spatial frequency response surfaces (SFRS’S): an alternative visualization and interpolation technique for head relation transfer functions (HRTF’s),” in AES 16th International Conference, Rovaniemi, Finland
- Gamper, H., 2013. Head-related transfer function interpolation in azimuth, elevation, and distance. J. Acoust. Soc. Am., 134(6), 547-553.
- Martens W.L. (1987). “Principal component analysis and resynthesis of spectral cues to perceived direction,” in Proceeding of the International Computer Music Conference, San Francisco, CA, USA, 274-281
- Zhong X.L., and Xie B.S. (2005b). “Spatial characteristics of head related transfer function,” Chinese Physics Letter 22(5), 1166-1169
- J. Nam, M. Kolar, and J. S. Abel, “On the minimum-phase nature of head-related transfer functions,” in Proc. Audio Eng. Soc. 125th Conv., San Francisco, CA, 2008.
- J.-M. Jot, V. Larcher, and O. Warusfel, “Digital signal processing issues in the context of binaural and transaural stereophony,” in Proc. Audio Eng. Soc. 98th Conv., Paris, France, Feb. 1995.
- Evans M.J., Angus J.A.S., and Tew A.I. (1998a). “Analyzing head-related transfer function measurements using surface spherical harmonics,” J. Acoust. Soc. Am. 104 4), 2400-2411.

References on anthropometry features for HRTF

- Fels, J., Buthmann, P., & Vorländer, M. (2004). Head-related transfer functions of children. *Acta Acustica united with Acustica*, *90*(5), 918-927.
- Burkhard, M. D., & Sachs, R. M. (1975). Anthropometric manikin for acoustic research. The Journal of the Acoustical Society of America, 58(1), 214-222.
- Middlebrooks, J. C. (1999). Individual differences in external-ear transfer functions reduced by scaling in frequency. The Journal of the Acoustical Society of America, 106(3), 1480-1492.
- Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The cipic hrtf database. In Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the (pp. 99-102). IEEE.
- Nishino, T., Inoue, N., Takeda, K., & Itakura, F. (2007). Estimation of HRTFs on the horizontal plane using physical features. Applied Acoustics, 68(8), 897-908
- Xie, B., Zhong, X., Rao, D., & Liang, Z. (2007). Head-related transfer function database and its analyses. Science in China Series G: Physics Mechanics and Astronomy, 50(3), 267-280.
- Bilinski,Piotr; Ahrens, Jens; Thomas, Mark R.P; Tashev, Ivan; Platt,John C (2014). "HRTF MAGNITUDE SYNTHESIS VIA SPARSE REPRESENTATION OF ANTHROPOMETRIC FEATURES". IEEE ICASSP, Florence, Italy: 4468–4472.
- Zhang, M., Kennedy, R., Abhayapala, T., and Zhang, W., "Statistical method to identify key anthropometric parameters in HRTF individualization," in Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on, pp. 213–218, IEEE, 2011.

References on anthropometry features for HRTF

- H., Zhou, L., Ma, H., and Wu, Z., “HRTF personalization based on artificial neural network in individual virtual auditory space,” Applied Acoustics, 69(2), pp. 163 – 172, 2008.
- Li, L. and Huang, Q., “HRTF personalization modeling based on RBF neural network,” in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3707–3710, 2013.
- Grijalva, F., Martini, L., Florencio, D., and Goldenstein, S., “A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(3), pp. 559–570, 2016.
- H. M. Fayek, et al., “On Data-Driven Approaches to Head-Related Transfer Function Personalization,” in AES 143, New York, Oct. 2017
- M. Rothbucher, M. Durkovic, H. Shen, and K. Diepold, "HRTF customization using multiway array analysis," in Proc. 18th European Signal Processing Conference (EUSIPCO'10), Aalborg, August 2010, pp. 229-233.
- D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03), New York, Oct. 2003, pp. 157-160.
- R. Sridhar, J. G. Tylka, and E. Y. Choueiri. A database of head-related transfer function and morphological measurements. In Audio Engineering Society Convention 143, October 2017.
- R. Bomhardt, M. Klein, and J. Fels, “A high-resolution head-related transfer function and three-dimensional ear model Database,” Proc. Mtgs. Acoust. 29, 050002 (2016).

References on listening based HRTF individualization

- J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," J. Acoust. Soc. Amer., vol. 106, no. 3, pp. 1480-1492, Sep. 1999.
- K. J. Fink and L. Ray, " Individualization of head related transfer functions using principal
- component analysis,” Applied Acoustics, 87 (2015) 162–173.
- A. Bondu, S. Busson, V. Lemaire, and R. Nicol, "Looking for a relevant similarity criterion for HRTF clustering: a comparative study," in Proc. 120th Audio Engineering Society Convention, Paris, France, May 2006.
- Y. Iwaya, “Individualization of head-related transfer functions with tournament-style listening test: Listening with other’s ears,” Acoust. Sci. & Tech. 27, 6 (2006)
- A. Harma et al, “Personalization of headphone spatialization based on the relative localization error in an auditory gaming interface,” AES Convention, 2012
- E. Schwenker, and G. Romigh, “An Evolutionary Algorithm Approach to Customization of Non-Individualized Head Related Transfer Functions,” AES Convention, 2014
- C. Mendonca, “A Review on Auditory Space Adaptations to Altered Head-Related Cues,” Frontiers in Neuroscience, vol. 8, article 219 (2014).
- F. Klein, and S. Werner, “Auditory Adaptation to Non-Individual HRTF Cues in Binaural Audio Reproduction,” JAES, 2016
- C. Mendonca, J. A. Santos, G. Campos, P. Dias, and J. Vieira, “On the adaptation to non-individualised HRTF auralisations: a longitudinal study,” AES 45th International Conference, Helsinki, Finland, 2012 March.

Key References on equalization

- S. Olive, T. Welti, and E. McMullin, "Listener Preferences for Different Headphone Target Response Curves," in *Proc. 134th Audio Engineering Society Convention*, Rome, Italy, May 2013.
- K. Sunder, E. L. Tan, and W. S. Gan, "Individualization of binaural synthesis using frontal projection headphones," *J. Audio Eng. Soc.,* vol. 61, no. 12, pp. 989-1000, Dec. 2013.
- H. Møller, D. Hammershoi, C. B. Jensen, and M. F. Sorensen, "Transfer characteristics of headphones measured on human ears," *J. Audio Eng. Soc.,* vol. 43, no. 4, pp. 203-217, Apr. 1995.
- V. Larcher, J. M. Jot, and G. Vandernoot, "Equalization methods in binaural technology," in *Proc. 105th Audio Engineering Society Convention,* SanFrancisco, Sep. 1998.
- A. Kulkarni and H. S. Colburn, "Variability in the characterization of the headphone transfer-function," *J. Acoust. Soc. Amer.,* vol. 107, no. 2, pp. 1071-1074, Feb. 2000.
- H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Design criteria for headphones," *J. Audio Eng. Soc.,* vol. 43, no. 4, pp. 218-232, Apr. 1995.
- D. Pralong and S. Carlile, “The Role of Individualized Headphone Calibration for the Generation of High Fidelity Virtual Auditory Space,” J. Acoust. Soc. Am., vol. 100, no. 6, pp. 3785–3793 (1996 Dec.).
- O. Kirkeby and P. A. Nelson, “Digital Filter Design for Inversion Problems in Sound Reproduction,” J. Audio Eng. Soc., vol. 47, pp. 583–595 (1999 Jul./Aug.).
- S. G. Norcross, G. A. Soulodre, and M. C. Lavoie, “Subjective Investigations of Inverse Filtering,” J. Audio Eng. Soc, vol. 52, pp. 1003–1028 (2004 Oct.).
- A. Lindau and F. Brinkmann, “Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings,” J. Audio Eng. Soc., vol. 60, pp. 54–62 (2012 Jan./Feb.).
- J. GOMEZ-BOLANOS, A. Makivirta, and V. Pulkki, “Automatic Regularization Parameter for Headphone Transfer Function Inversion,” J. Audio Eng. Soc., Vol. 64, No. 10, 2016 October

References on head tracking

- Wallach H. (1940). “The role of head movement and vestibular and visual cue in sound localization,” J.Exp. Psychol. 27(4), 339-368.
- Wenzel E.M., Arruda M., and Kistler D.J., et al. (1993a). “Localization using nonindividualized head-related transfer functions,” J. Acoust. Soc. Am. 94(1), 111-123.
- Wenzel E.M. (1995). “The relative contribution of interaural time and magnitude cues to dynamic sound localization,” in Proceedings of the IEEE 1995 Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 80-83.
- Wenzel E.M. (1996). “What perception implies about implementation of interactive virtual acoustic environments,” in AES 101st Convention, Los Angeles, CA, U.S.A., Preprint: 4353.
- Wenzel E.M. (1991). “Localization in Virtual Acoustic Displays,” Presence 1 (1), 80–107.
- Wightman F.L., and Kistler D.J. (1999). “Resolution of front-back ambiguity in spatial hearing by listener and source movement,” J. Acoust. Soc. Am. 105(5), 2841-2853.
- Sandvad J. (1996). “Dynamic aspects of Auditory virtual environments,” in AES 100th Convention, Copenhagen, Denmark, Preprint 4226.
- Saviojia L., Huopaniemi J., and Lokki T., et al. (1999). “Creating interactive virtual acoustic environments,” J. Audio. Eng. Soc. 47(9), 675-705.
- Horbach U., Karamustafaoglu A., and Pellegrini R., et al. (1999). “Design and applications of a data-based auralization system for surround sound,” in AES 106th Convention, Munich, Germany, Preprint: 4976.
- D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," J. Audio Eng. Soc., vol. 49, no. 10, pp. 904-916, Oct. 2001.
- Etienne Hendrickx, Peter Stitt, Jean-Christophe Messonnier, Jean-Marc Lyzwa, Brian FG Katz, and Catherine de Boishéraud, “Influence of head tracking on the externalization of speech stimuli for nonindividualized binaural synthesis,” The Journal of the Acoustical Society of America 141, 2011 (2017); doi: 10.1121/1.4978612

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

C.1 Augmented/Mixed Reality: overview

Augmented/Mixed reality is enhancing the way we experience the real world

Wearable AR/MR devices:

Meta

HoloLens

Magic leap

AR/MR applications:

Navigation

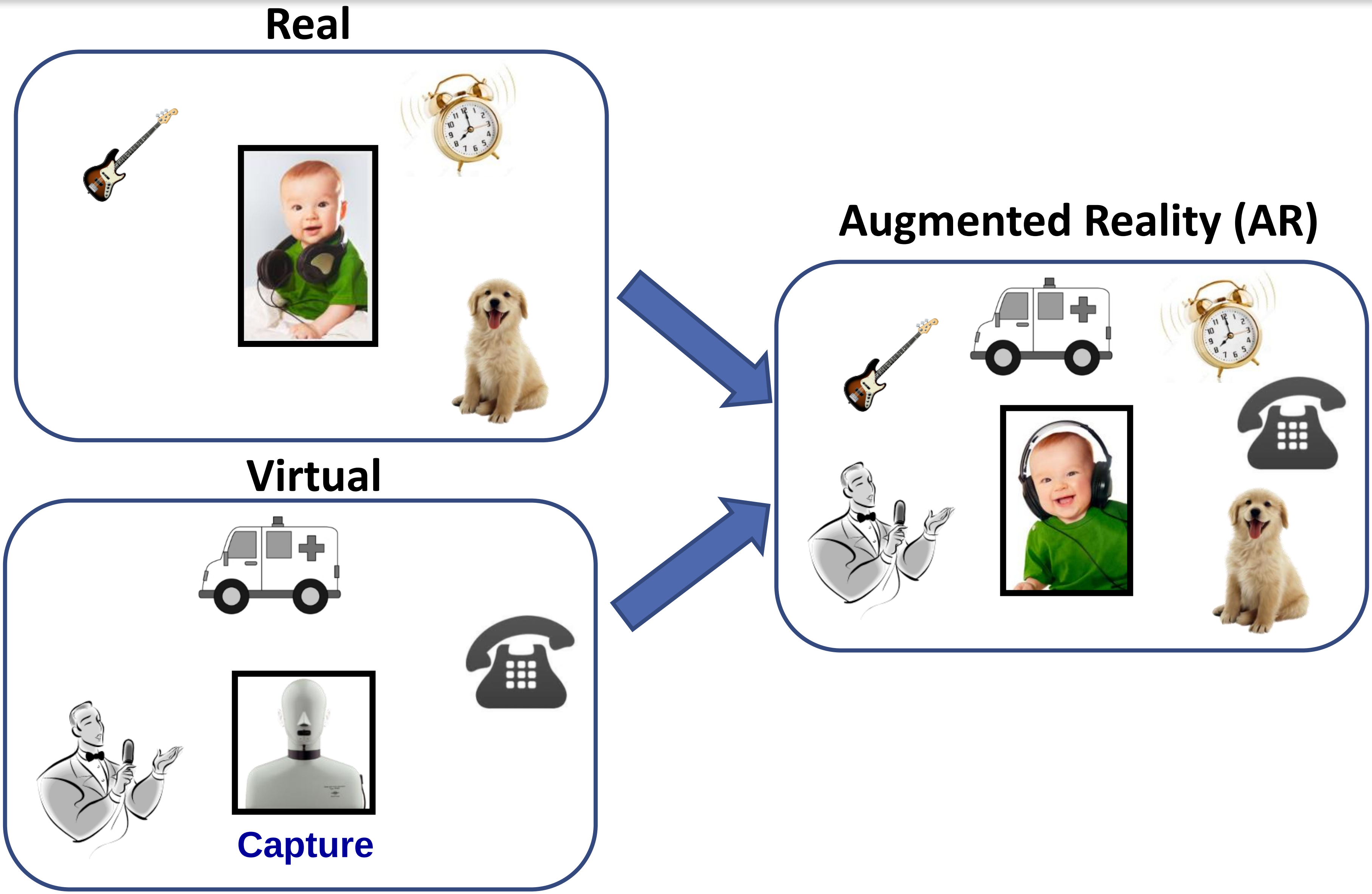
VR and AR world

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



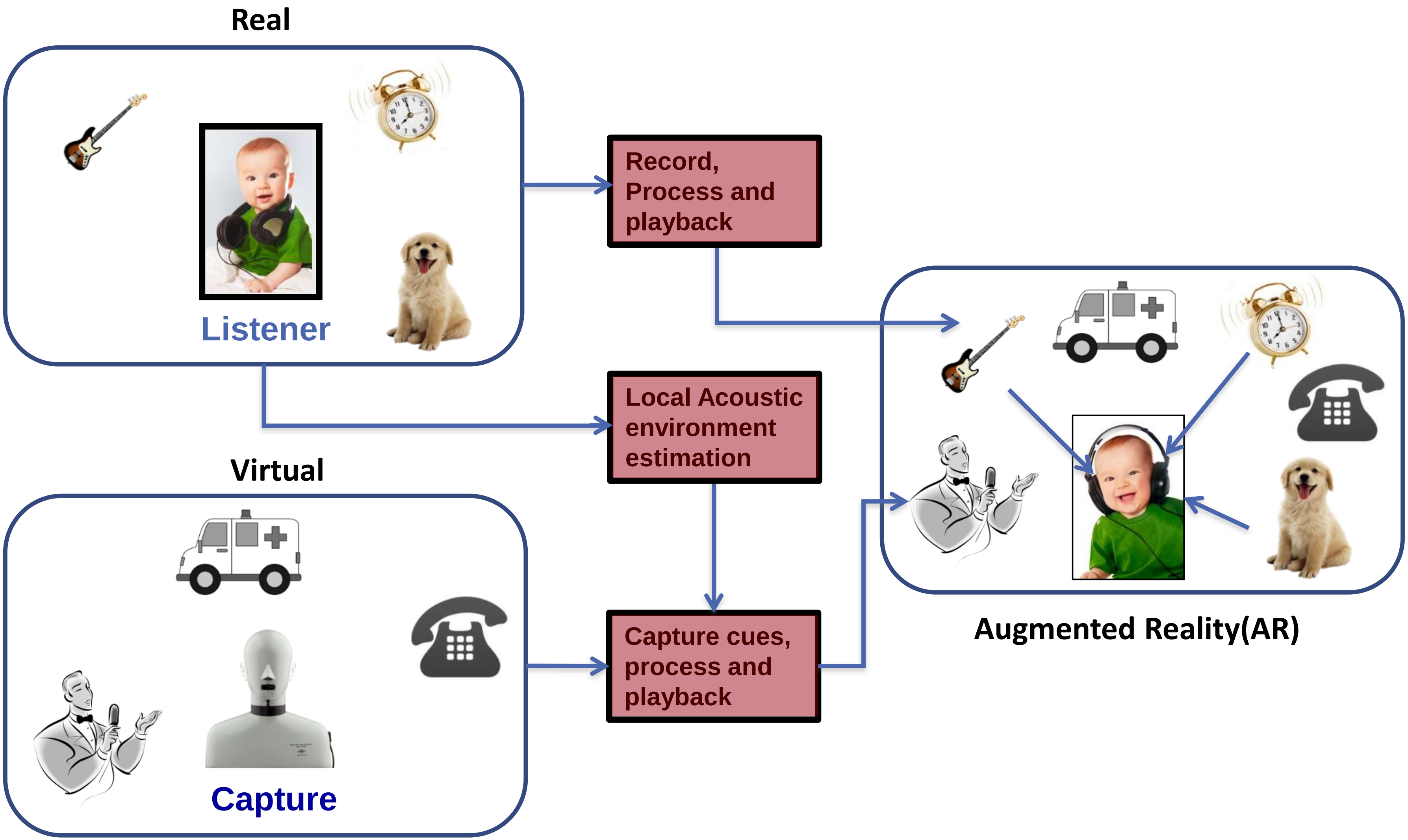
Real

Virtual

Capture

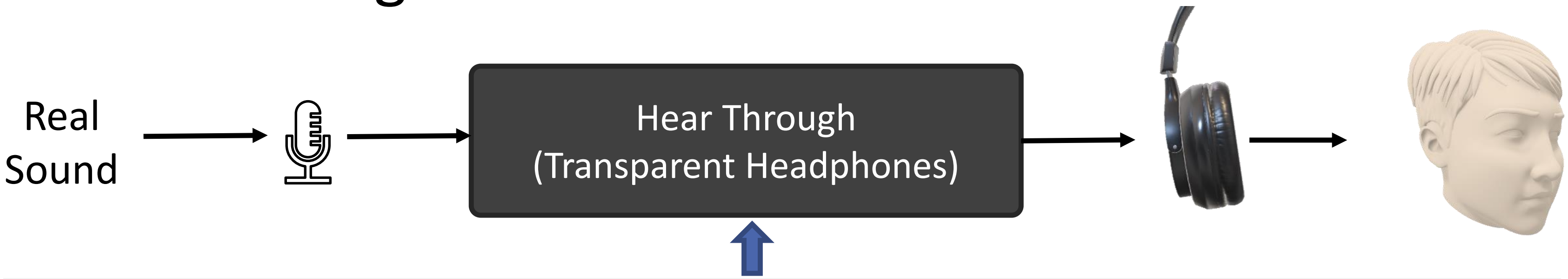
Augmented Reality (AR)

C.3 Signal Processing Techniques in NAL


Real
Listener
Record, Process and playback
Local Acoustic environment estimation
Virtual
Capture
Capture cues, process and playback
Augmented Reality(AR)

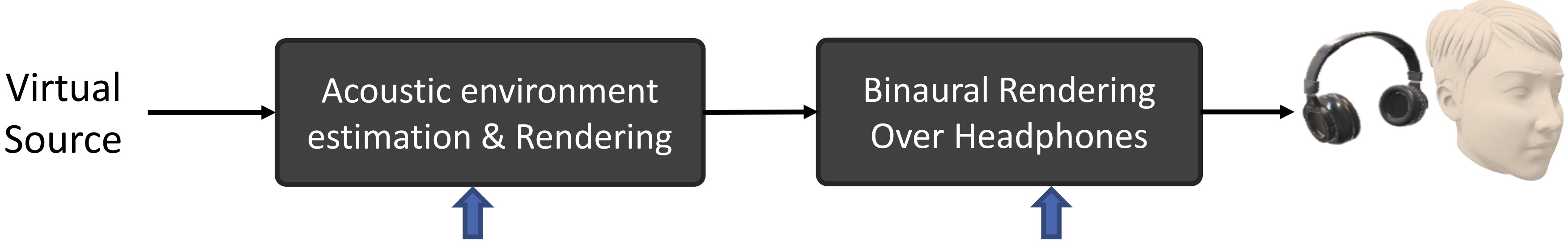
Natural Augmented Listening: 3 Major Components

- Hear through of real sounds



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)

Type II –
Open-back over
ear headset

- Allows natural sound to pass through
- Show best externalization
- Privacy issue due to leakage
- Poor bass for speakers

Closed Headphones

Type III –
Closed In-ear
headset

Type IV –
Closed-back over
ear headset

- Blocks most of the natural sounds
- Introduces occlusion effect
- Transparent hearing using electrical hear through

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[ǂ]
  - Closed headphones block most of the natural sounds

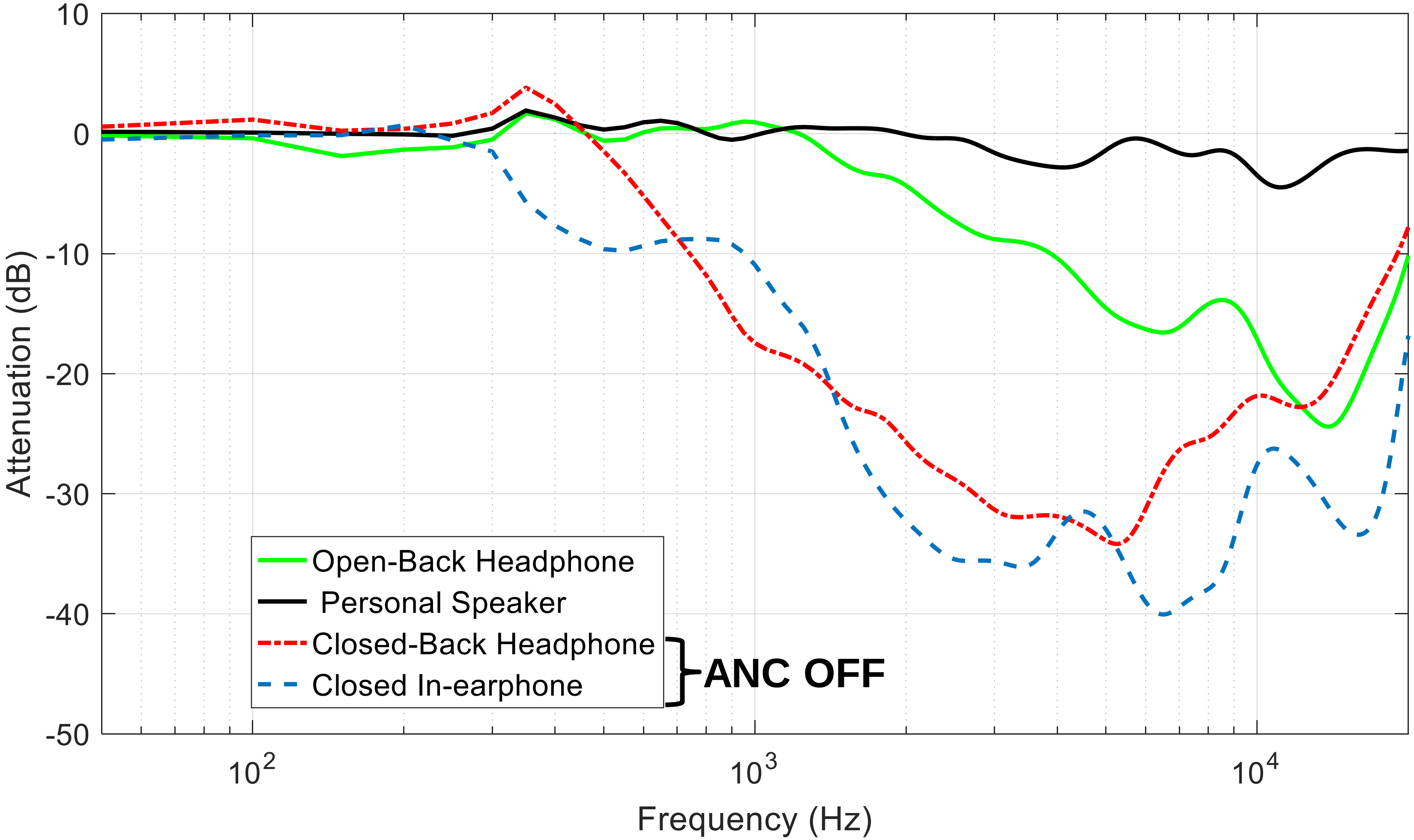
- Headphones Isolation obtained by measuring the speaker response at subject's ears with headphones, $H_{with\ hp}(f)$ and without headphones, $H_{ref}(f)$:

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

[ǂ] Open-back headphones attenuates poorly in higher frequencies [See next slide]

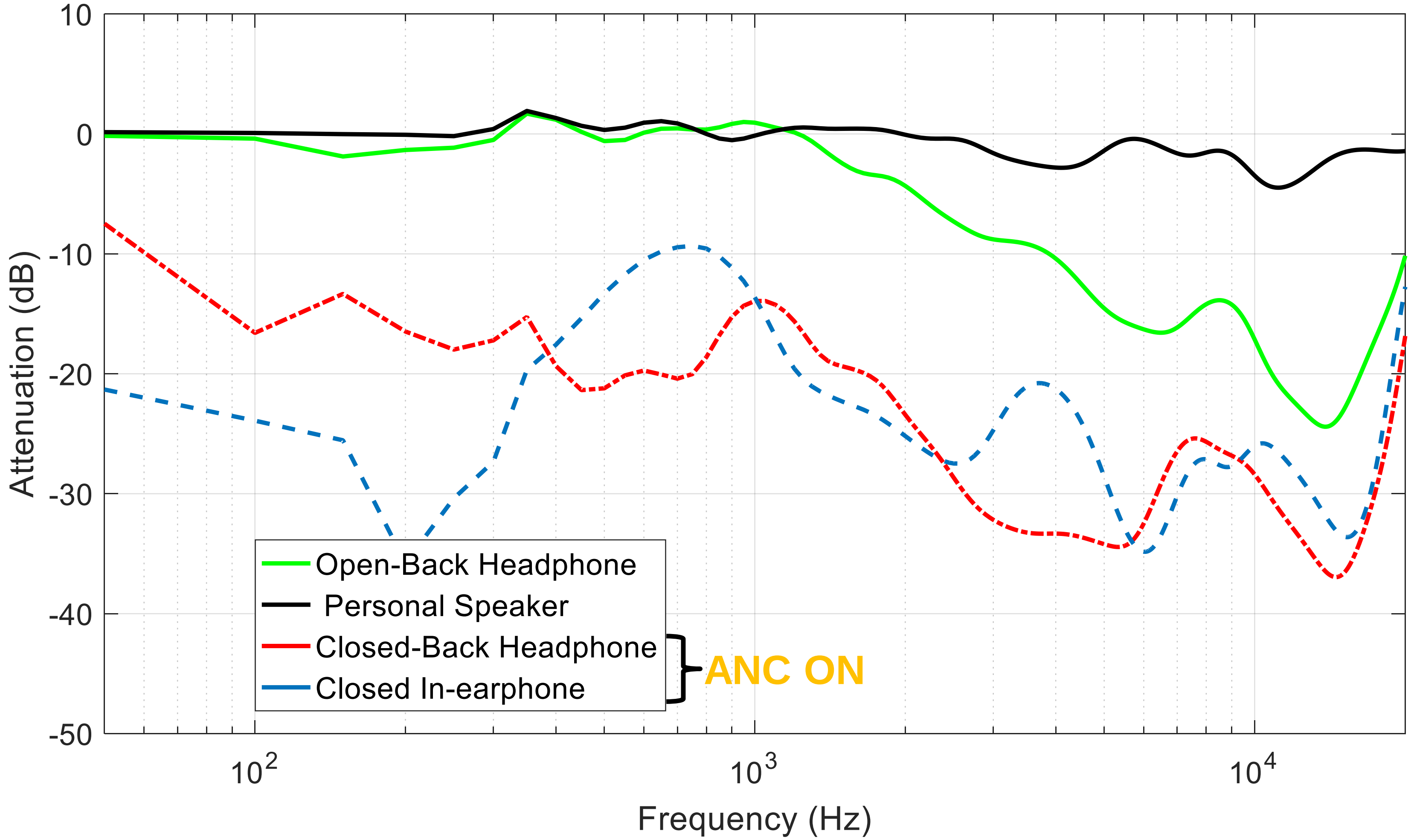
Headphones Isolation Characteristics

Attenuation curves for 4 different types of headphones

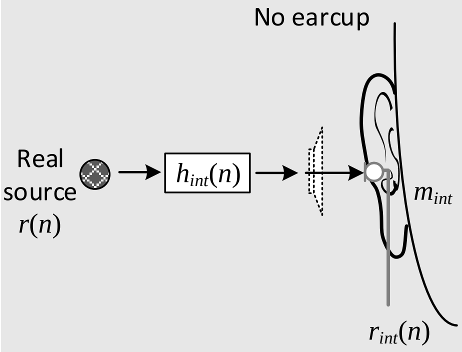
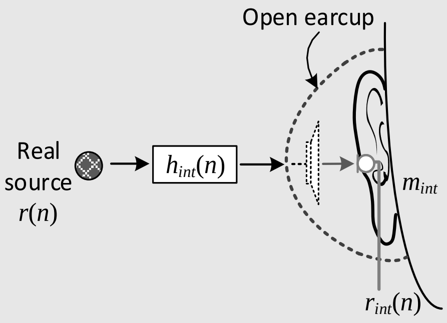
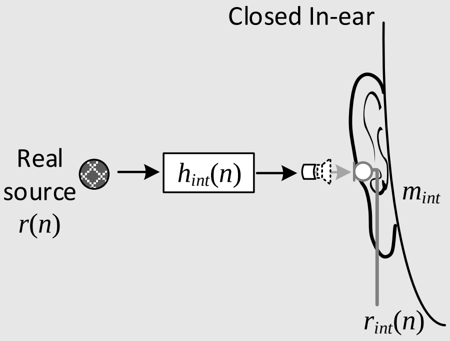
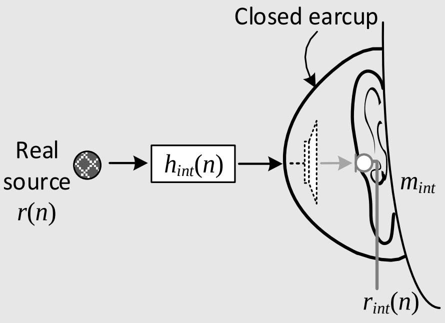
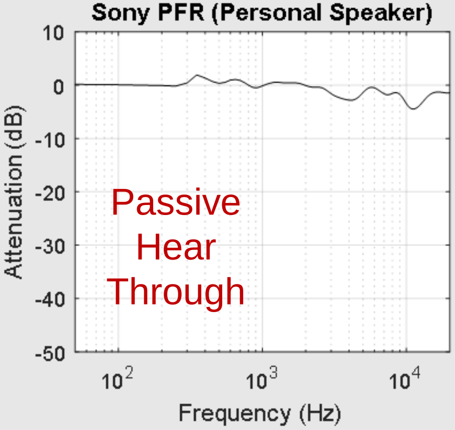
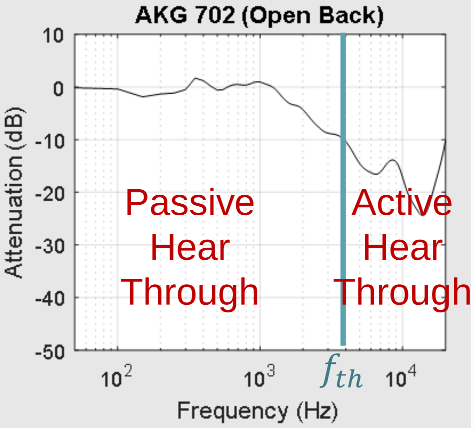
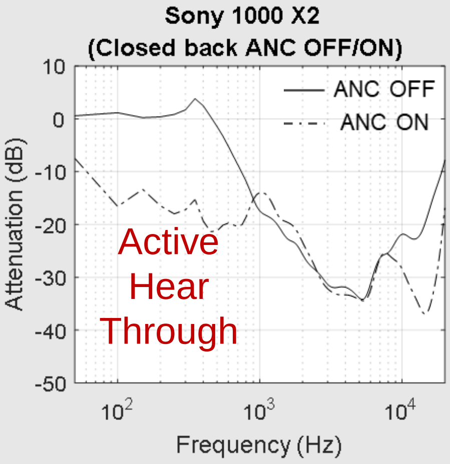
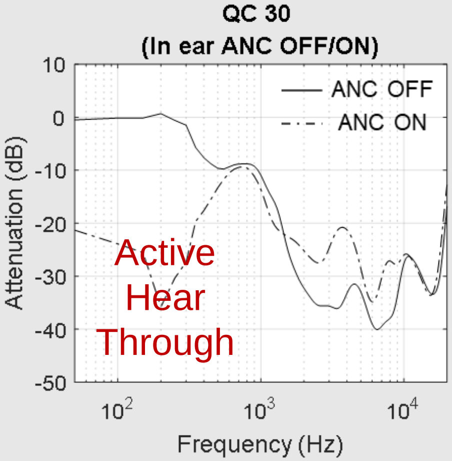


Headphones Isolation Characteristics

Attenuation curves for 4 different types of headphones

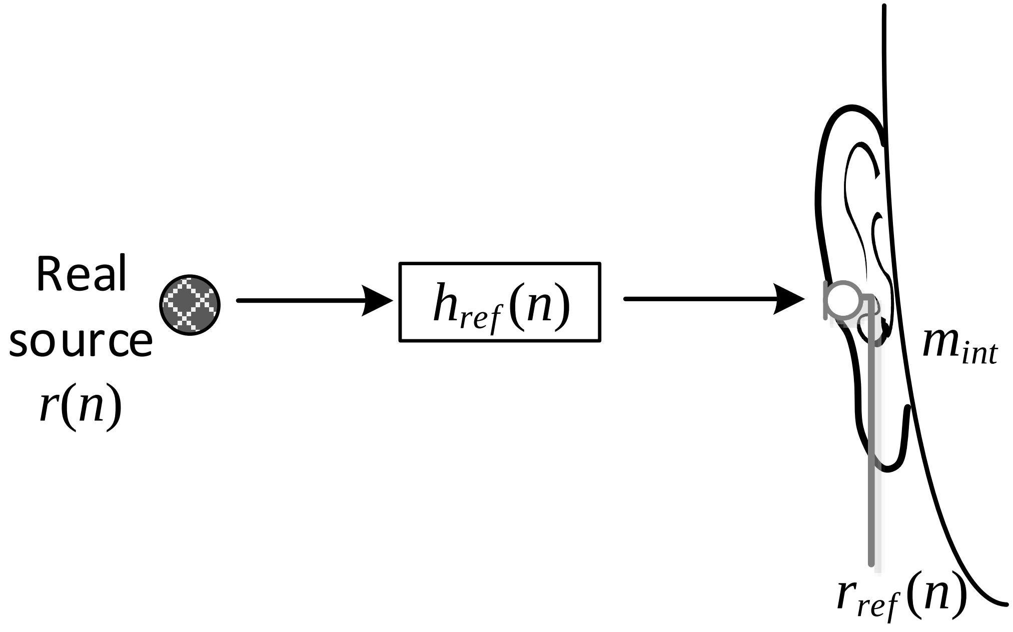


Headphone Isolation - Summary

Type I (Personal Speaker)	Type II (Open-back over ear headset)	Type III (Closed in-ear headset)	Type IV (Closed-back over ear headset)
			
$A(f) \cong 1$	$A(f) = \begin{cases} \cong 1; & f < f_{th} \\ \ll 1; & f \ge f_{th} \end{cases}$	$A(f) \ll 1$	$A(f) \ll 1$
			

Active Hear Through Mode

- Open ear scenario (Reference)



Signal at m_{int} :

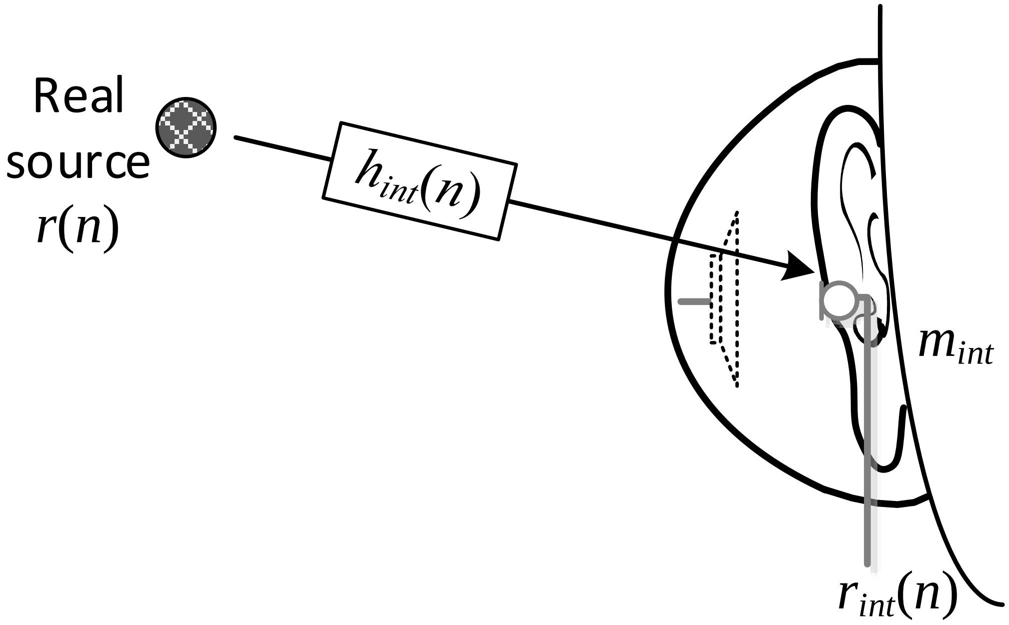
$$\underbrace{r(n) * h_{ref}(n)}$$

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)



Signal at m_{int} :

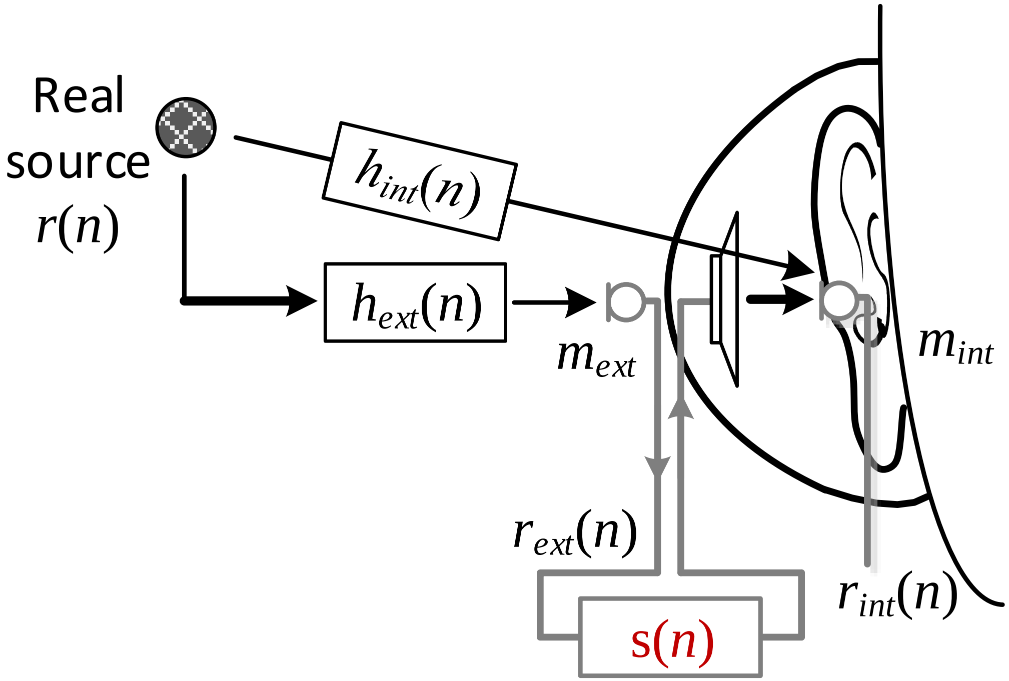
$$\underbrace{r(n) * h_{int}(n)}$$

Leaked real signal, $r_{int}(n)$ captured with headphones

$h_{int}(n)$: Impulse response at m_{int} measured with headphones

Active Hear Through Mode

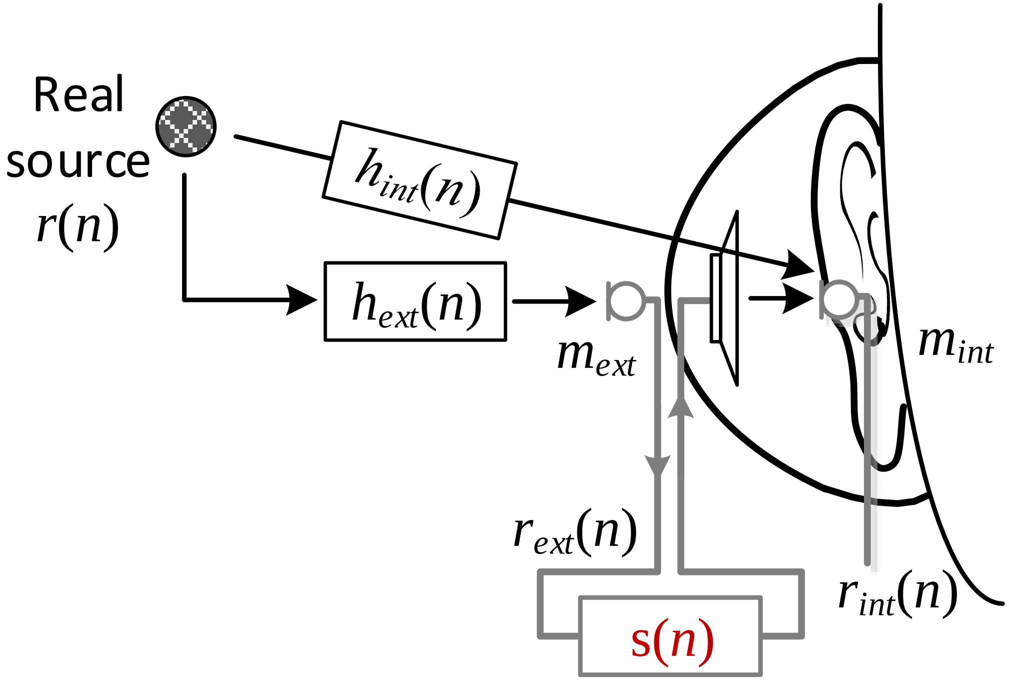
- Equalized/Compensated real signal (After EQ)



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} :

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

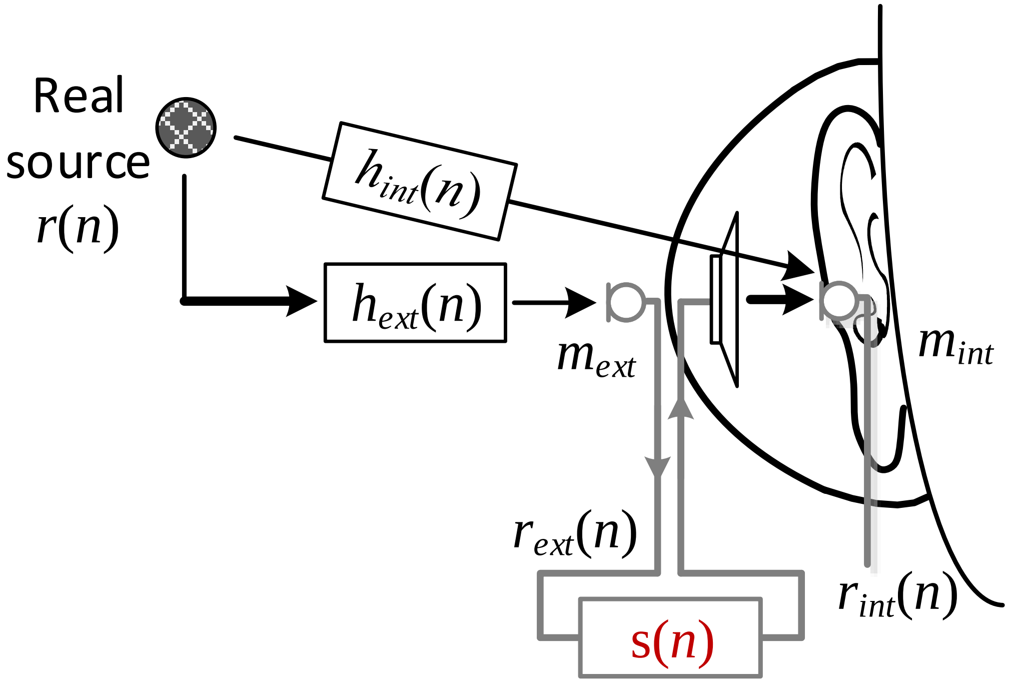
Processed real signal, $\hat{r}_{ref}(n)$ after headphone playback

$h_{ext}(n)$: 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} :

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

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.,* $\mathrm{E}[r_{int}(n)] \ll \mathrm{E}\left[\hat{r}_{ref}(n)\right]$

$$\underbrace{r(n) * h_{int}(n)}_{r_{int}(n) \approx 0} + \underbrace{r_{ext}(n) * s(n) * h_{hp}(n)}_{\hat{r}_{ref}(n) \equiv r_{ref}(n)}$$

$$\Downarrow$$

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

$$\Downarrow$$

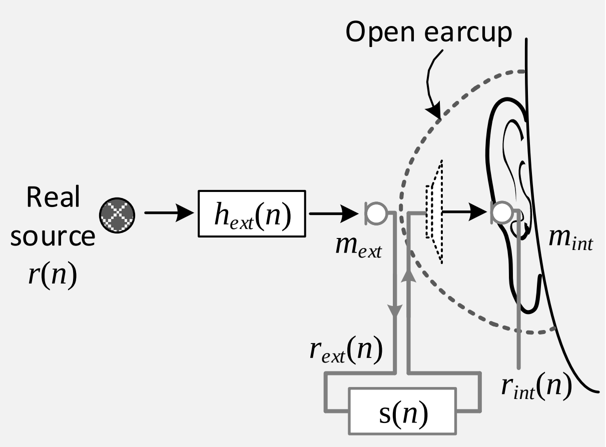
$$S(f) = \boxed{\frac{H_{ref}(f)}{H_{ext}(f)}} \times \boxed{\frac{1}{H_{hp}(f)}}$$

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)

Active Hear Through Mode - Summary

Type II (Open-back)



(1) EQ designed as *high-pass filter*

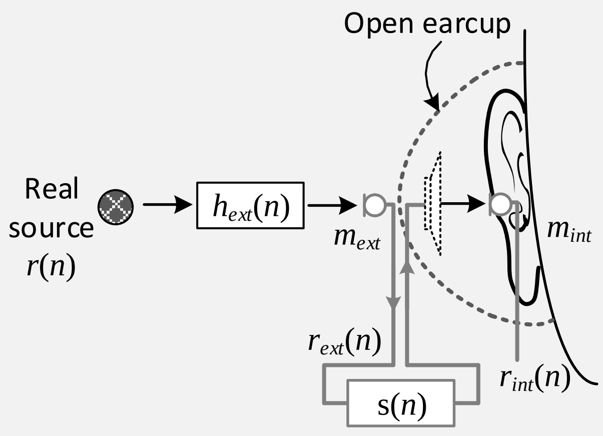
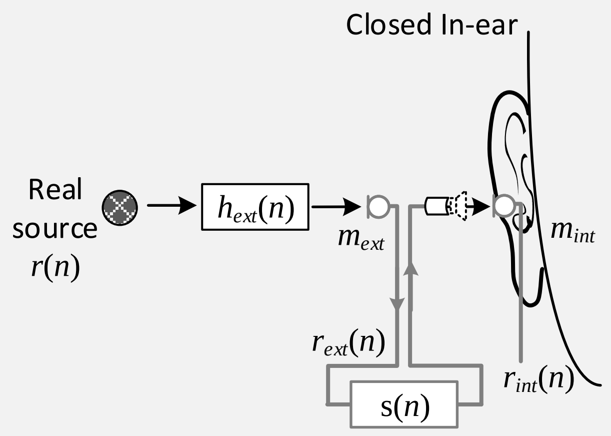
(2) EQ requires *pinnae cues* to be embedded

(3) Strong Comb effect due to poor attenuation

(4) Delay between leaked and processed real signal in high frequency

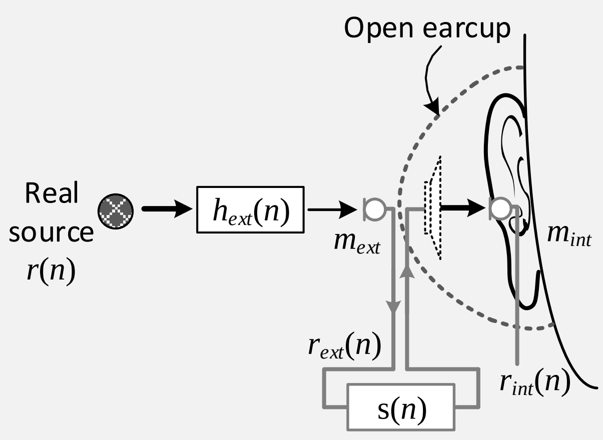
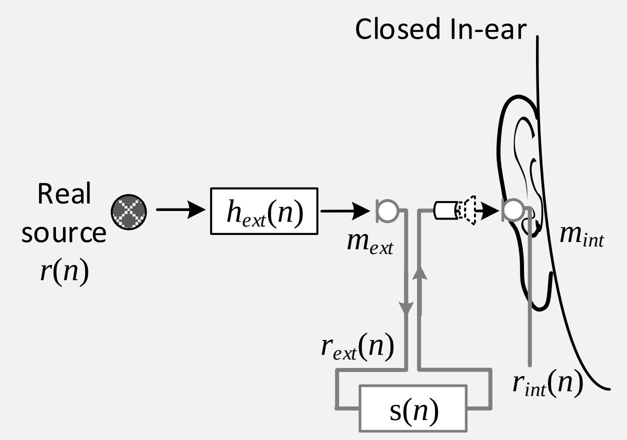
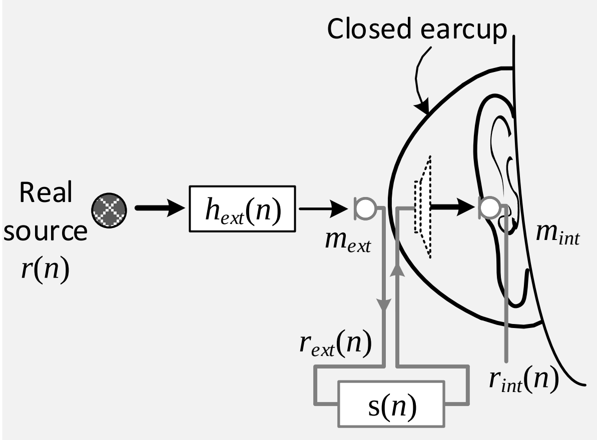
Active Hear Through Mode - Summary

Type II (Open-back)	Type III (Closed In-ear)

 Open earcup
Real source r(n)
h_ext(n)
m_ext
m_int
r_ext(n)
r_int(n)
s(n)
	 Closed In-ear
	Real source r(n)
	h_ext(n)
	m_ext
	m_int
	r_ext(n)
	r_int(n)
	s(n)

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

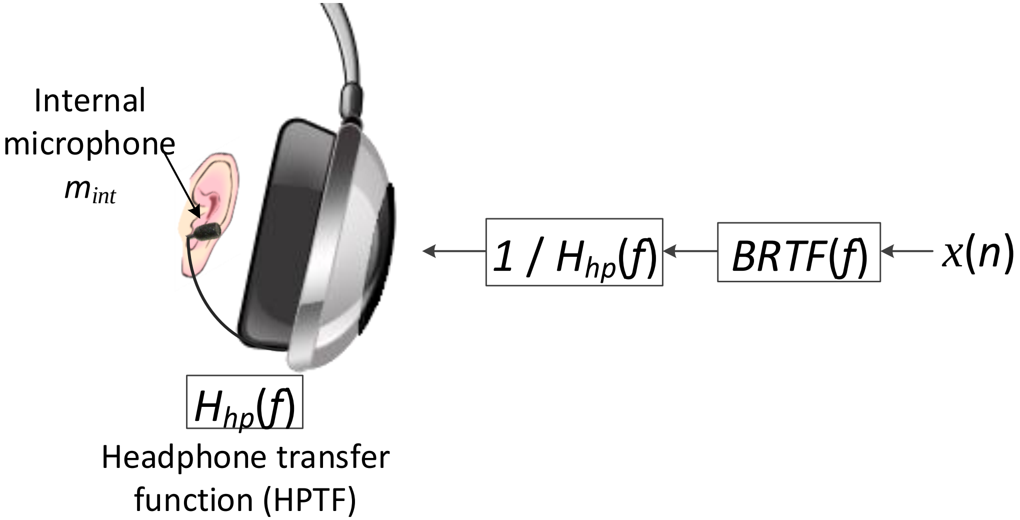
Active Hear Through Mode - Summary

Type II (Open-back)	Type III (Closed In-ear)	Type IV (Closed-back)
 Open earcup Real source $r(n)$ $h_{ext}(n)$ m_{ext} m_{int} $r_{ext}(n)$ $r_{int}(n)$ $s(n)$	 Closed In-ear Real source $r(n)$ $h_{ext}(n)$ m_{ext} m_{int} $r_{ext}(n)$ $r_{int}(n)$ $s(n)$	 Closed earcup Real source $r(n)$ $h_{ext}(n)$ m_{ext} m_{int} $r_{ext}(n)$ $r_{int}(n)$ $s(n)$
(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	(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	(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


Internal microphone m_{int}
$1 / H_{hp}(f)$
$BRTF(f)$
$x(n)$
$H_{hp}(f)$
Headphone transfer function (HPTF)

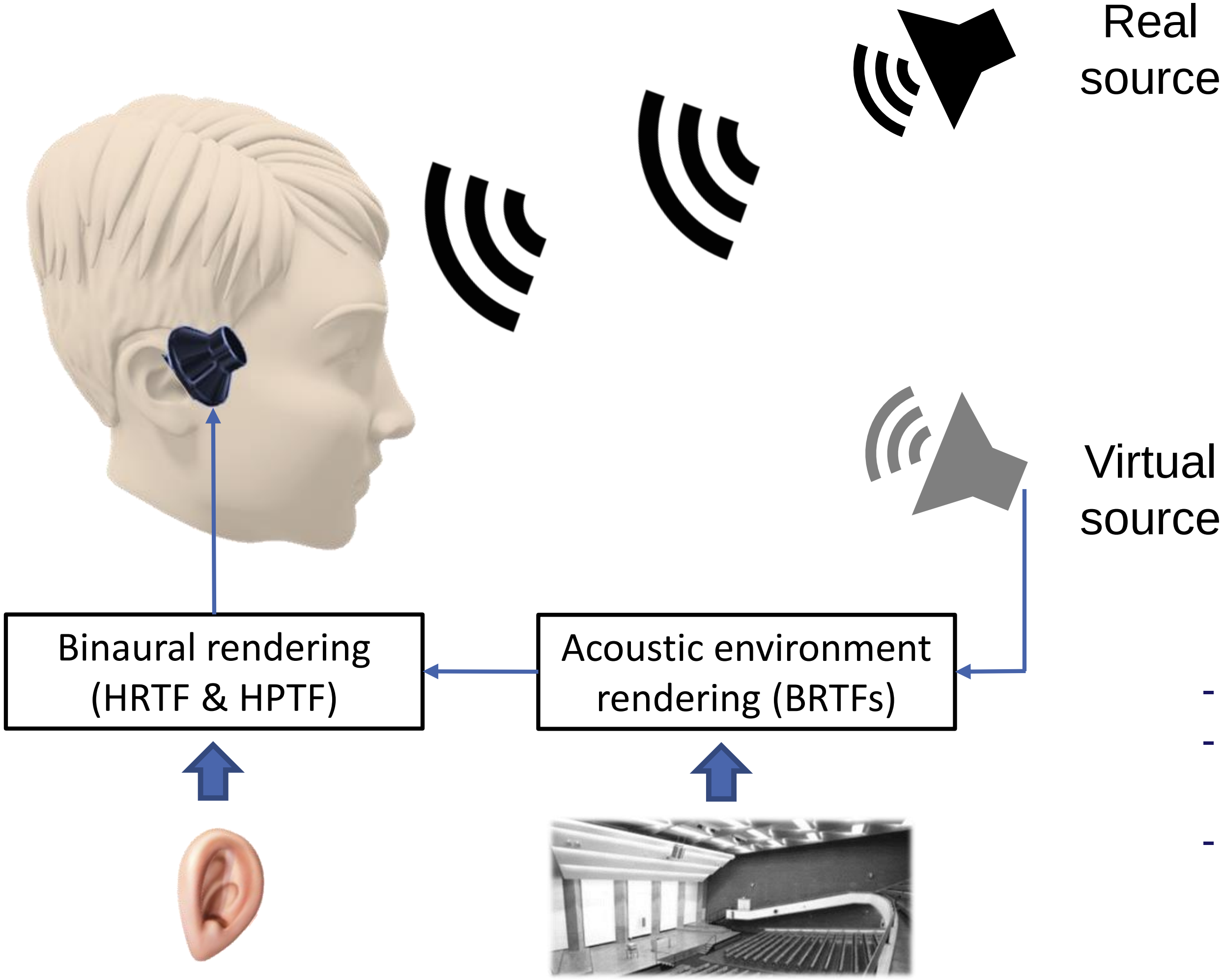
- Virtual monaural signal convolves with Binaural room transfer function
- Individual HPTF effect must be removed using equalization filter:
  - Direct inversion of HPTF
  - Using an adaptive algorithm like FxLMS

[Bouchard, 2006]

[Kuo and Morgan, 1995]

Type I – Open ear Design

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


Real source
Virtual source
Binaural rendering (HRTF & HPTF)
Acoustic environment rendering (BRTFs)

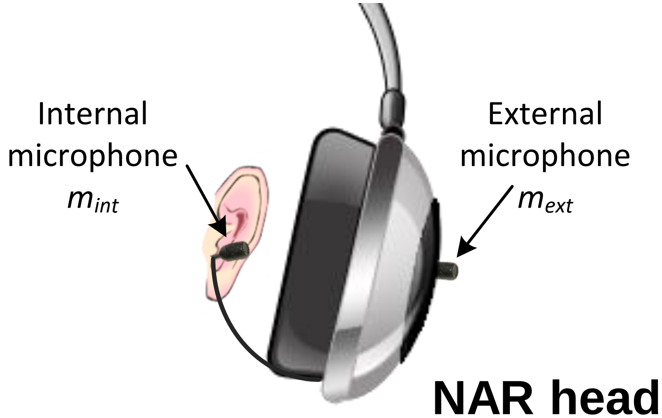
HoloLens with open ear design

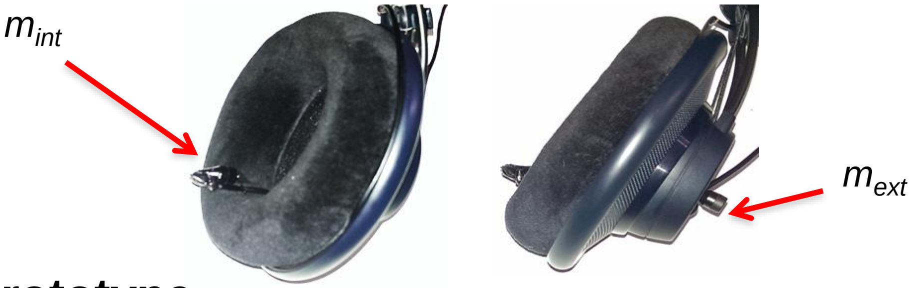
- Ideal hear through
- HRTF adapted to estimated head width
- Fixed Room Reverb options (small, medium, large)

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)





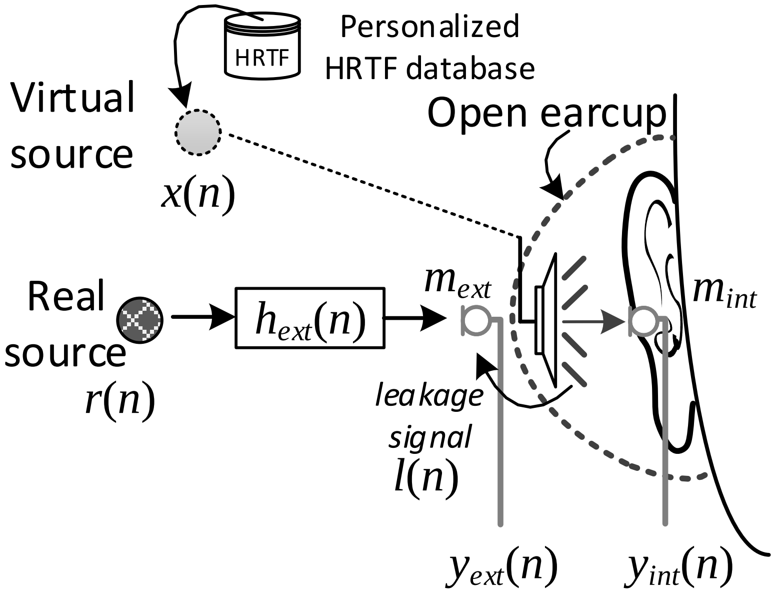
NAR headset prototype

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**

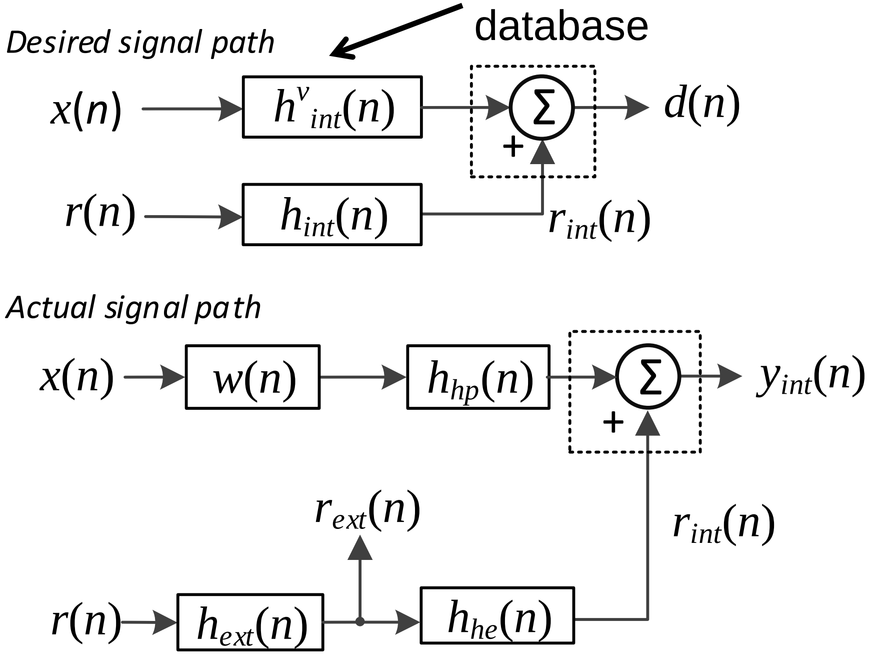
Augmented reality mode – **virtual sound reproduction** in the **presence of external signals**



$l(n)$: Leakage from headphone to external microphone, m_{ext}

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

Personalized HRTF selected from database



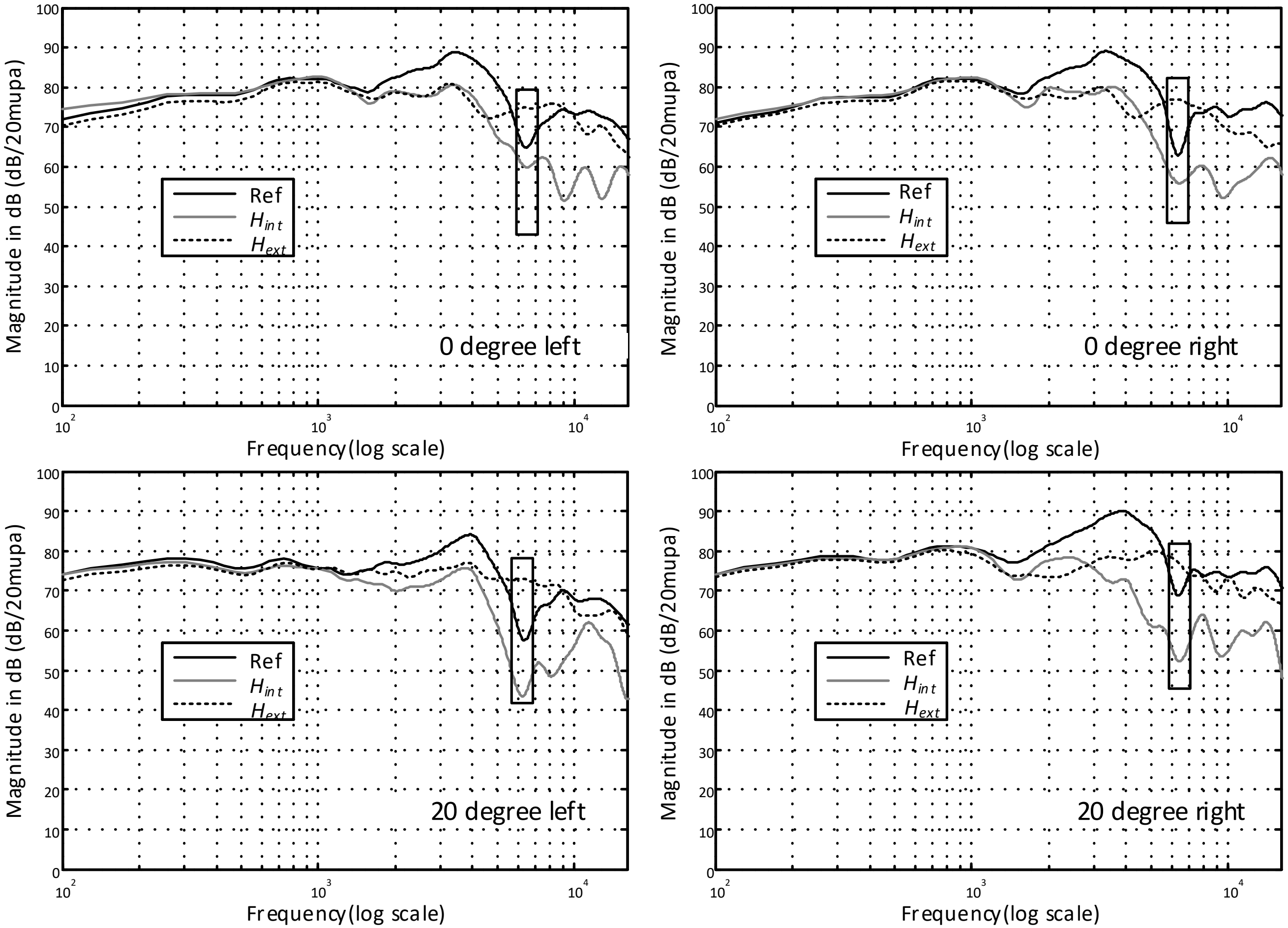
$$x_{int}(n) = w(n) * h_{hp}(n) * x(n)$$

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

[Ranjan and Gan, 2015]

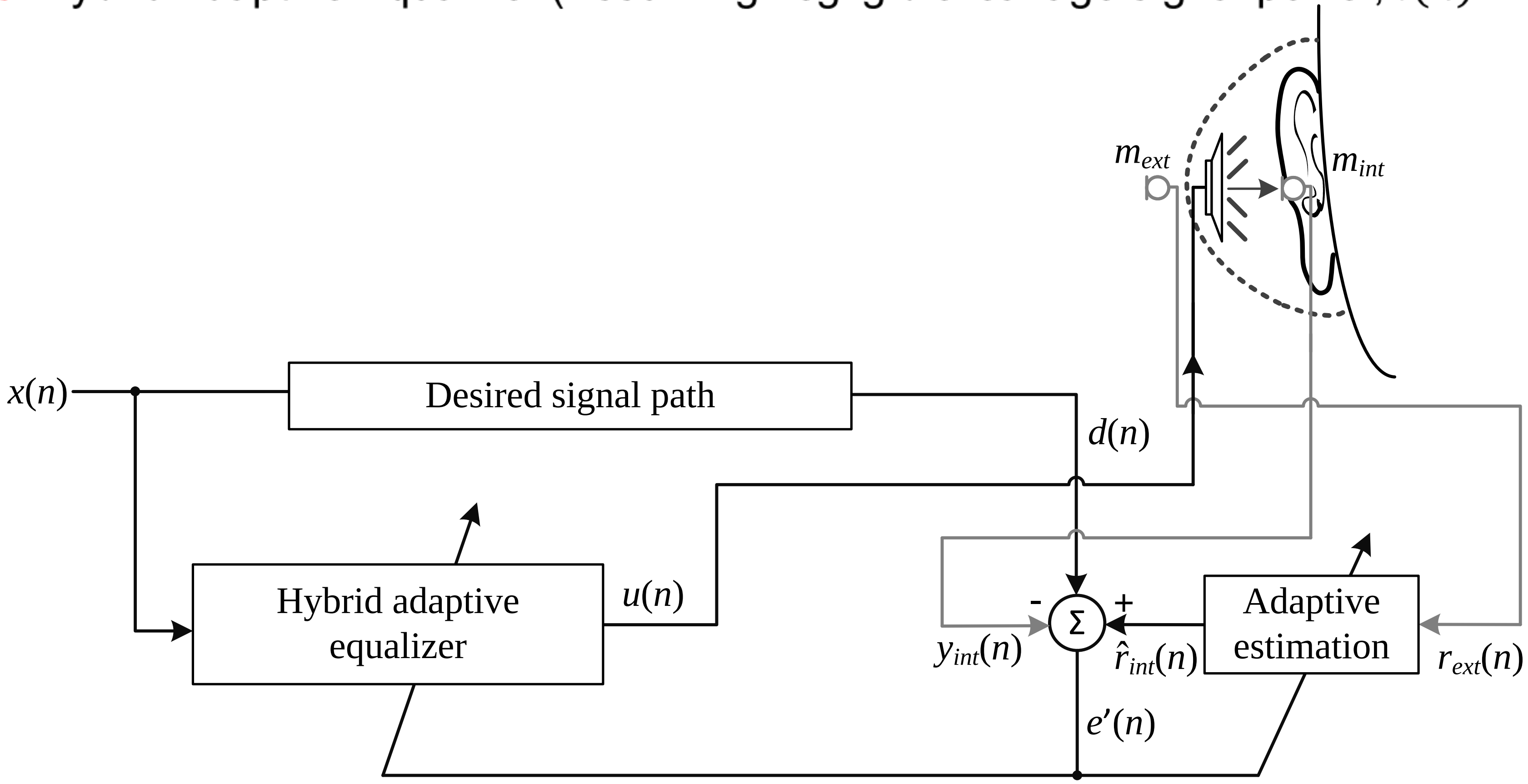
Type II – *int* and *ext* Transfer Functions

- $H_{int}(z)$ is an approximate HRTF with additional headphone effects
- $H_{ext}(z)$ contains all individual related characteristics and environment minus the pinnae specific notch and headphone shell reflections



Type II – Mixing virtual augmented signal with real ext source

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

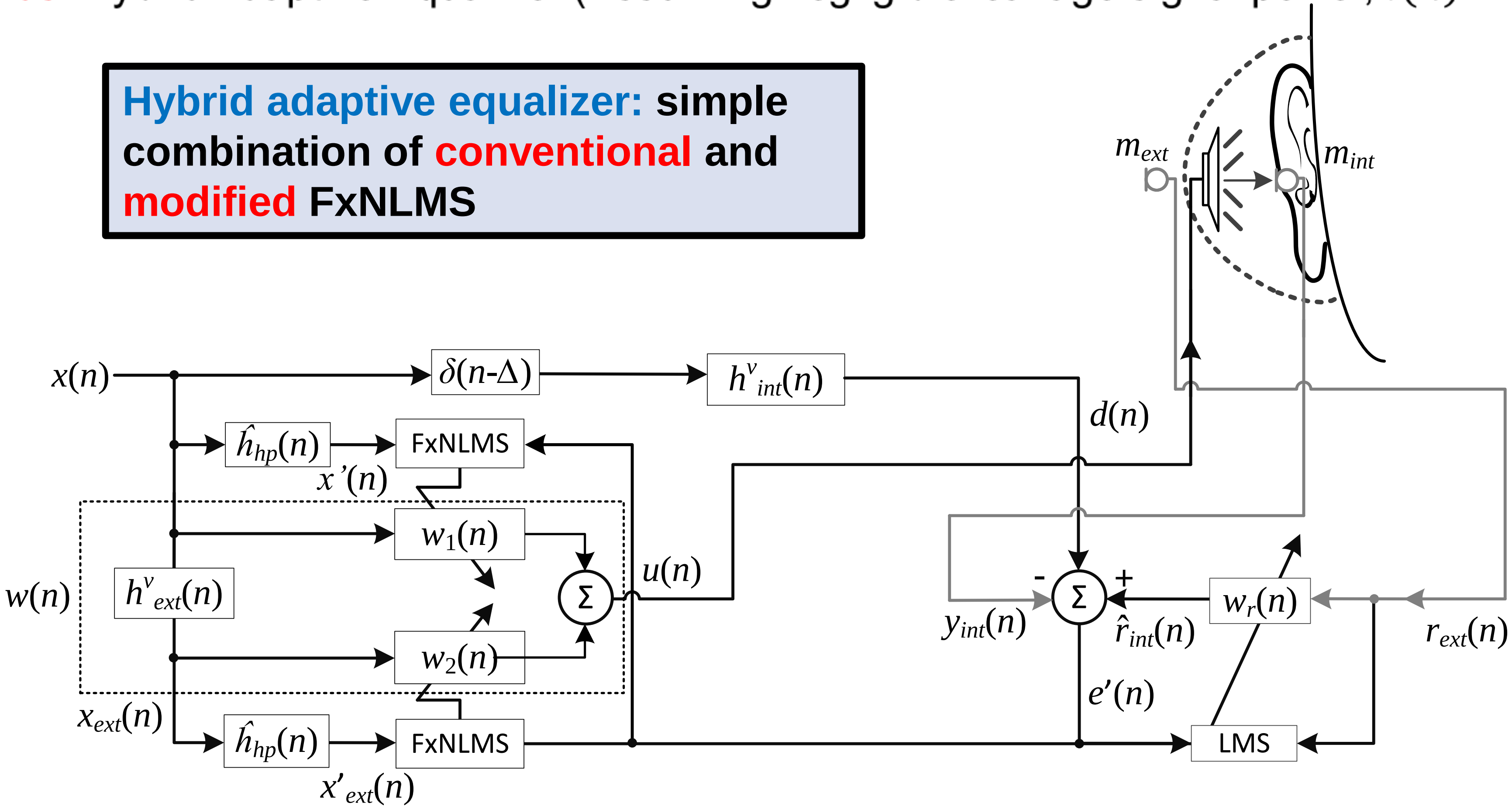


[Ranjan and Gan, 2015]

Type II – Mixing virtual augmented signal with real ext source

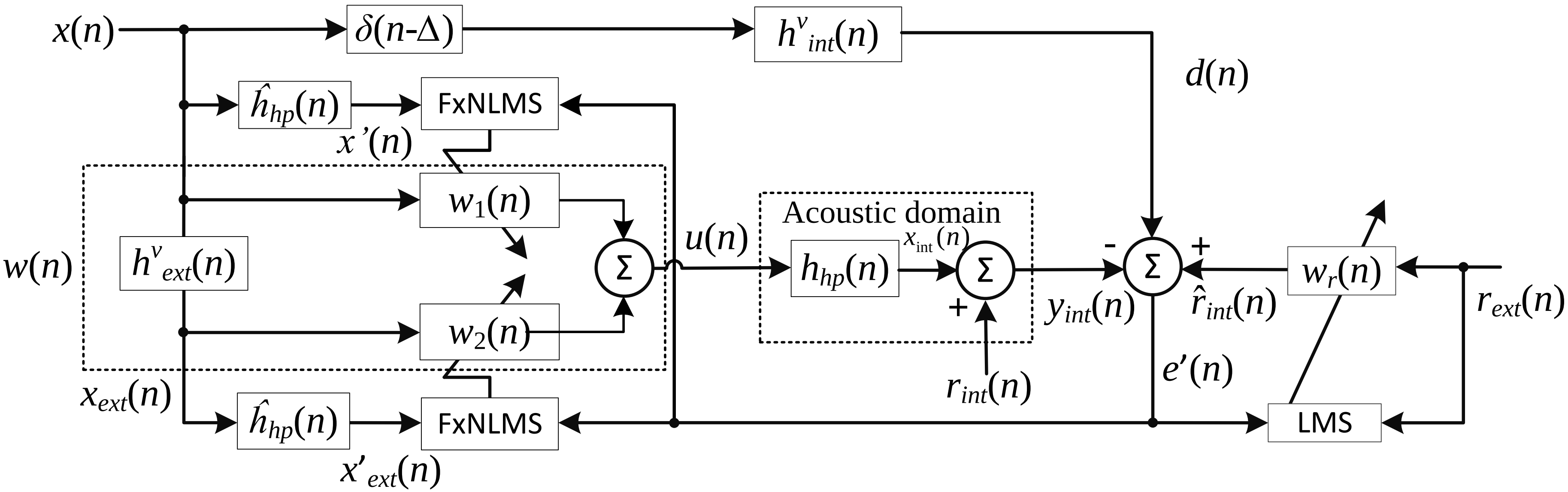
Augmented reality mode – virtual sound reproduction in the presence of external sounds: Hybrid Adaptive Equalizer (Assuming negligible leakage signal power, $l(n) = 0$)

Hybrid adaptive equalizer: simple combination of conventional and modified FxNLMS



Type II – Mixing virtual augmented signal with real ext source

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



$w_1(n)$: Adaptive filter corresponding to *conventional FXNLMS*

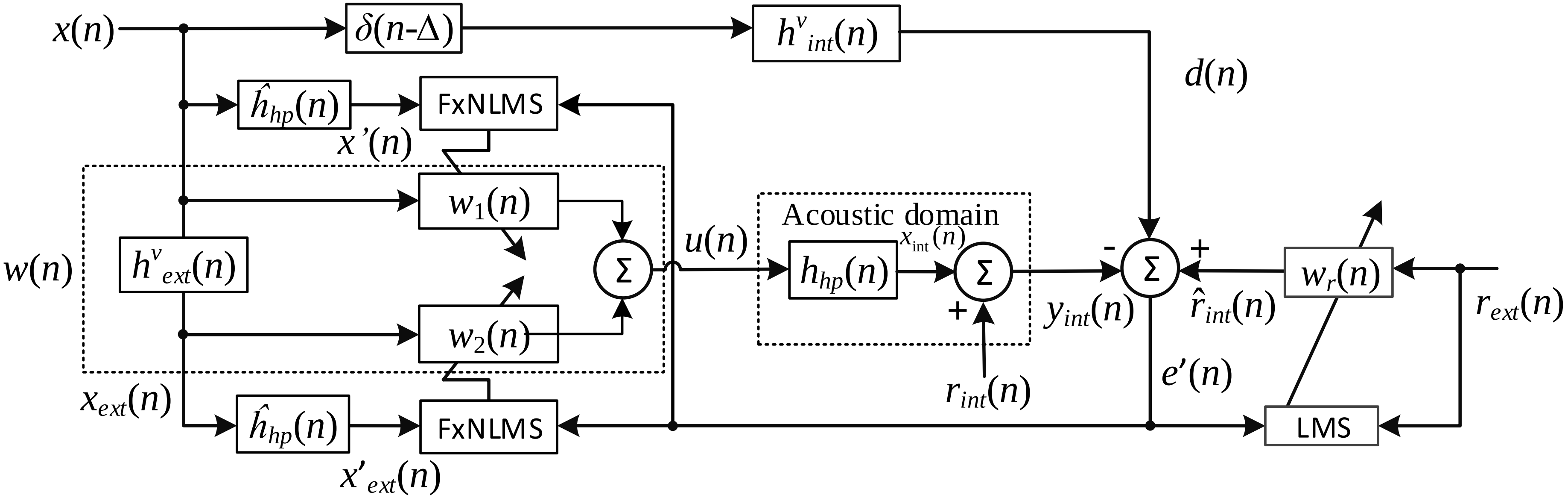
- Slower convergence rate due to presence of secondary path transfer function

$w_2(n)$: Adaptive 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)

Type II – Mixing virtual augmented signal with real ext source

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 :

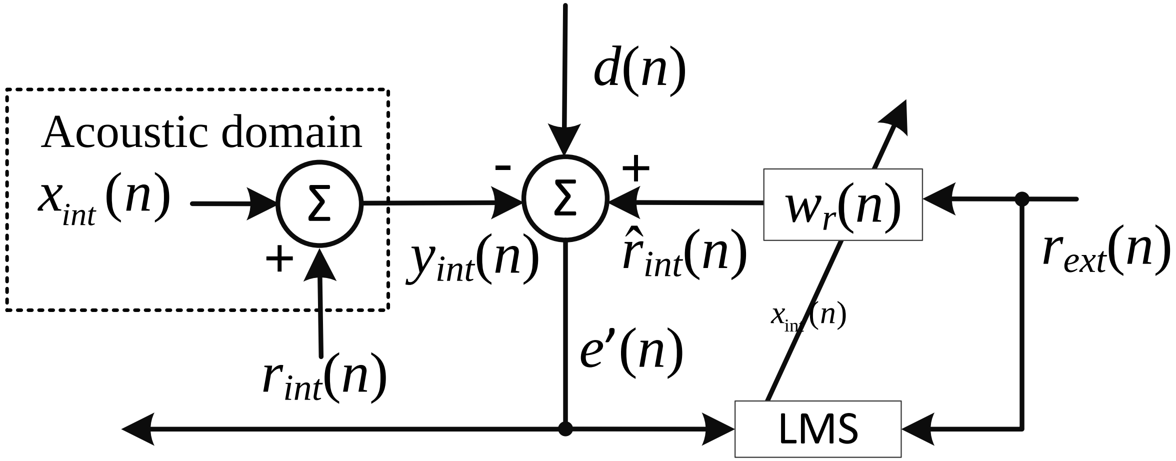
$$w(n) = w_1(n) + h^v{}_{ext}(n) * w_2(n)$$

$$\Big\Downarrow f$$

$$W(f) = W_1(f) + H^v_{ext}(f)W_2(f)$$

- 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,

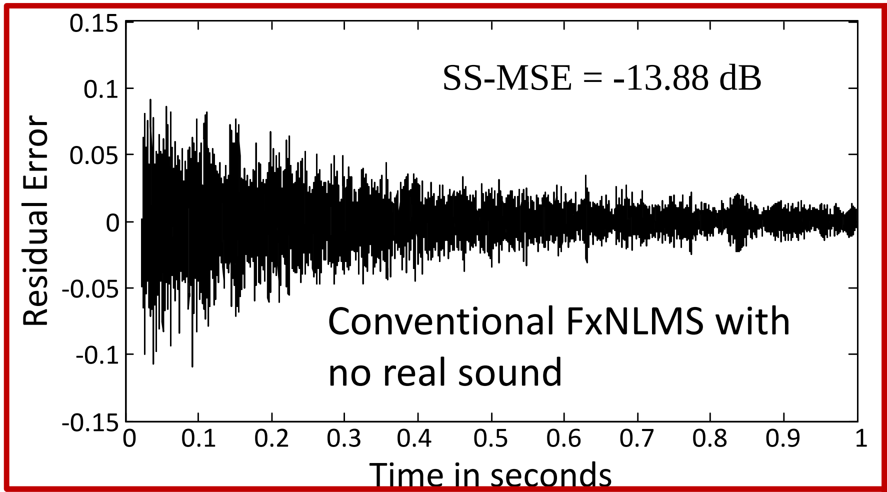
$$x_{int}(n) = h_{hp}(n) * u(n)$$

Error signal:

0

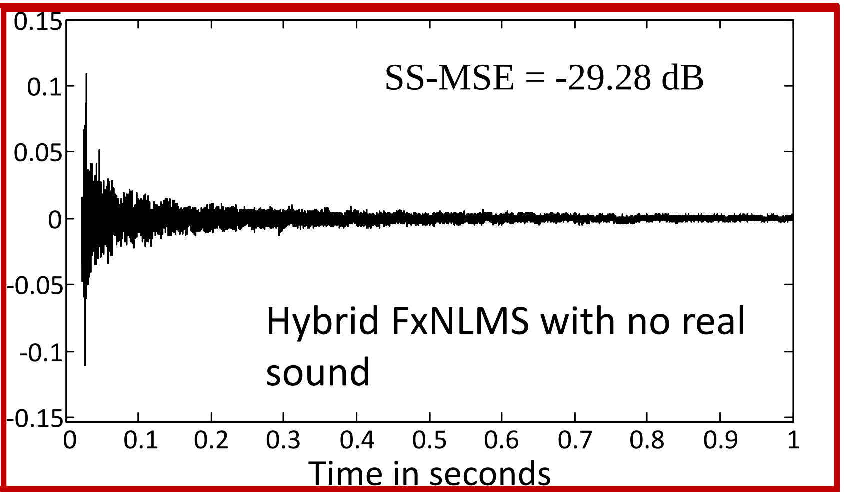
$$\begin{aligned} e'(n) &= \{d(n) + \hat{r}_{int}(n)\} - y_{int}(n) \\ &= \{d(n) - x_{int}(n)\} + \{-(r_{int}(n) - \hat{r}_{int}(n))\} \\ &= e_v(n) + e_r(n) \end{aligned}$$

Type II – Results



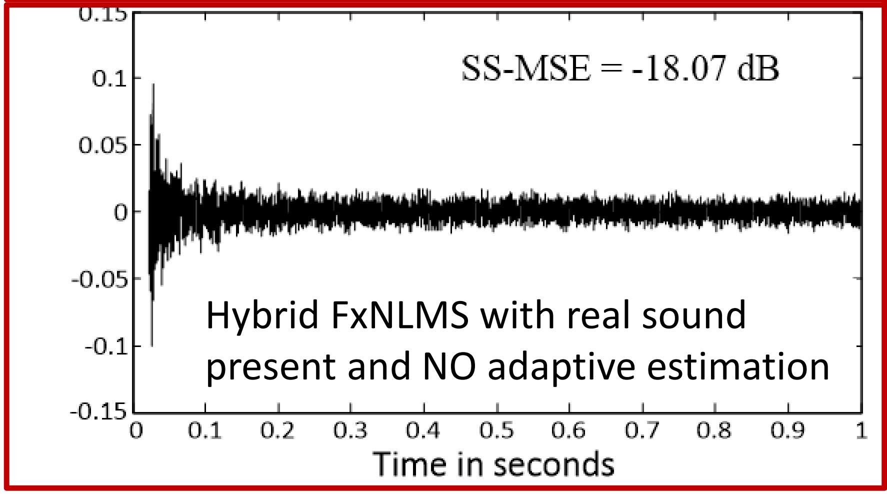
SS-MSE = -13.88 dB

Conventional FxNLMS with no real sound



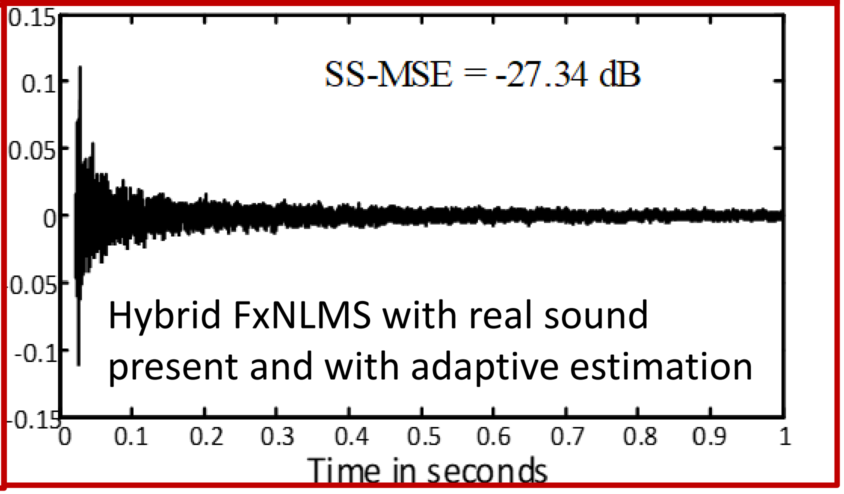
SS-MSE = -29.28 dB

Hybrid FxNLMS with no real sound



SS-MSE = -18.07 dB

Hybrid FxNLMS with real sound present and NO adaptive estimation



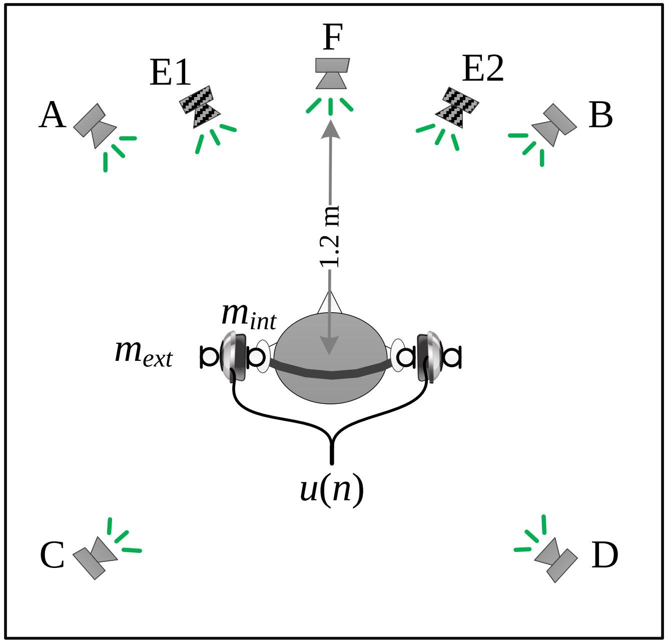
SS-MSE = -27.34 dB

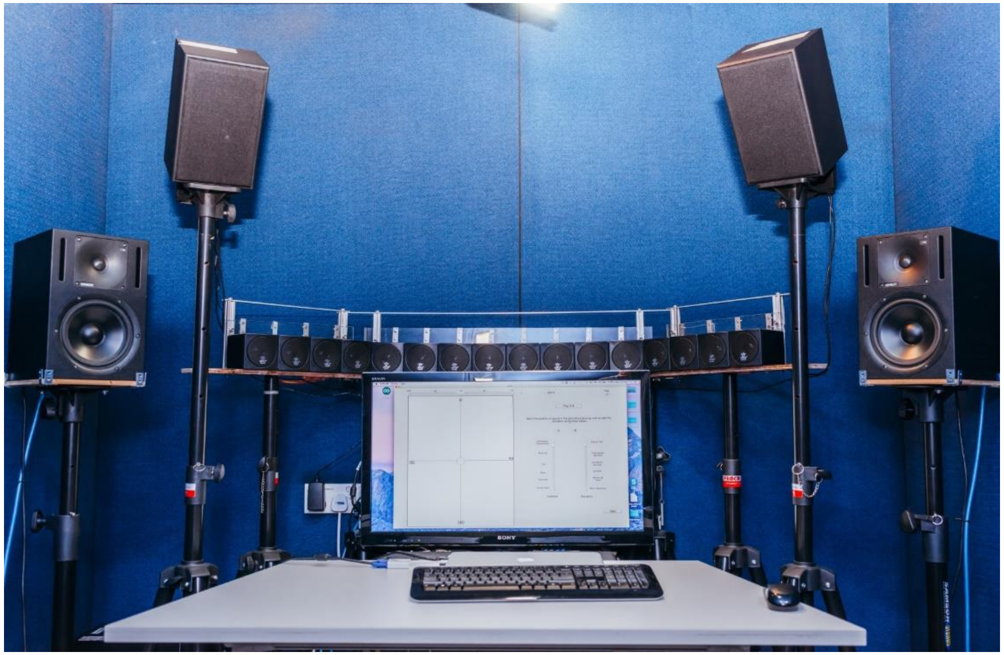
Hybrid FxNLMS with real sound present and with adaptive estimation

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

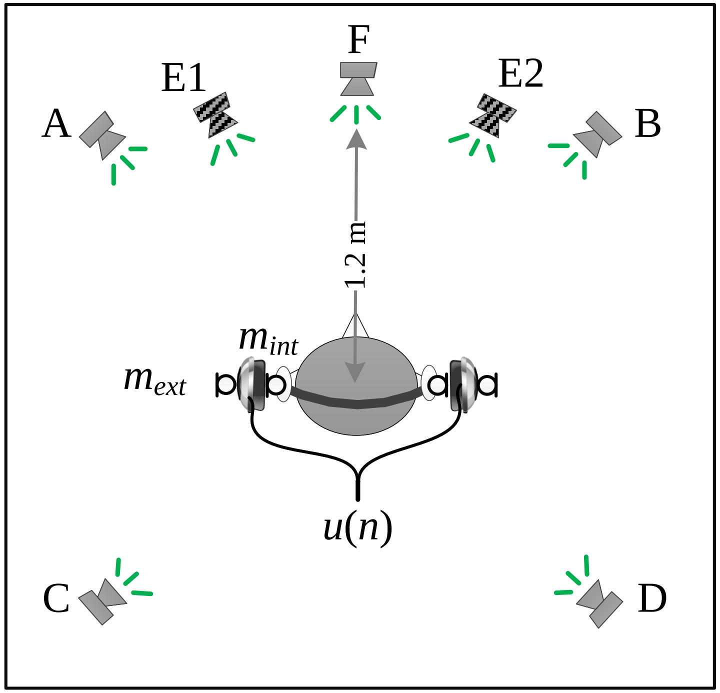




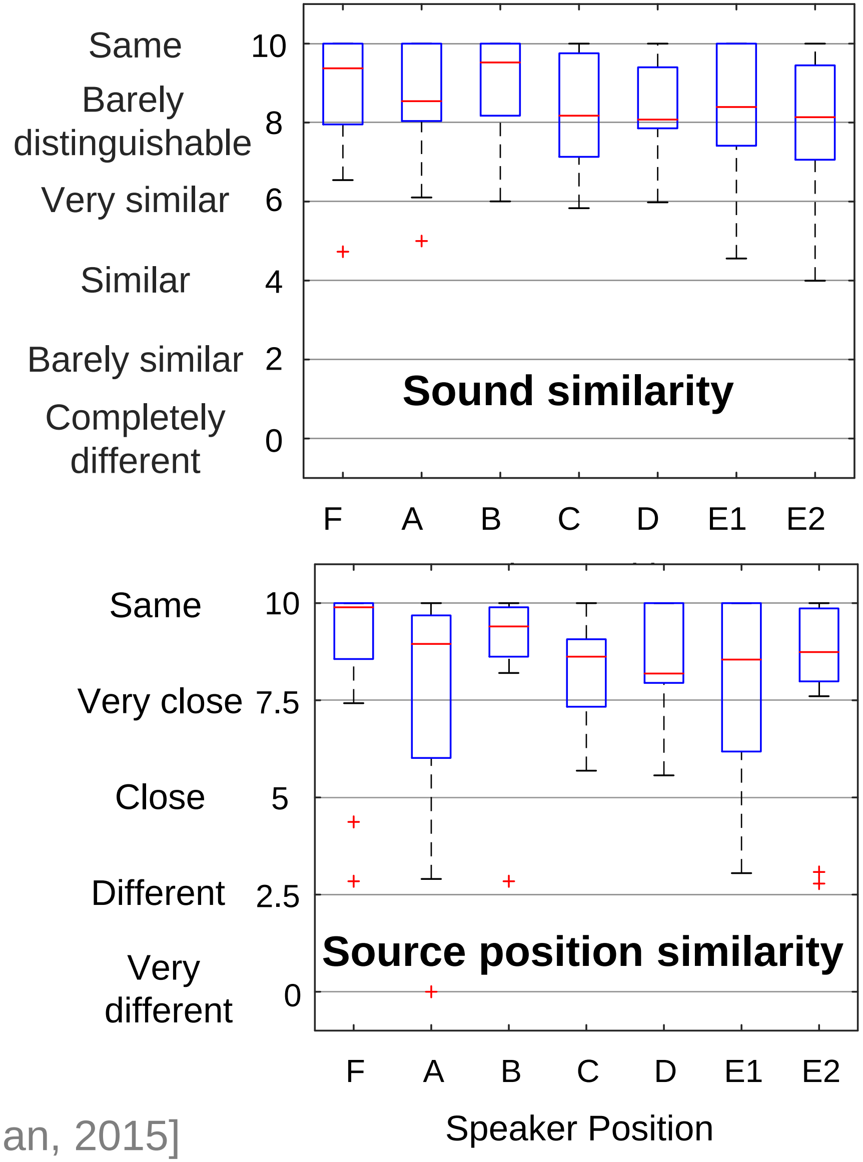
Listening Test Set up (▨ : Elevated speaker; ⧗ : Azimuth speaker)

Type II – Listening Test Results

- Subjective test evaluation for **sound similarity** and **source position similarity**



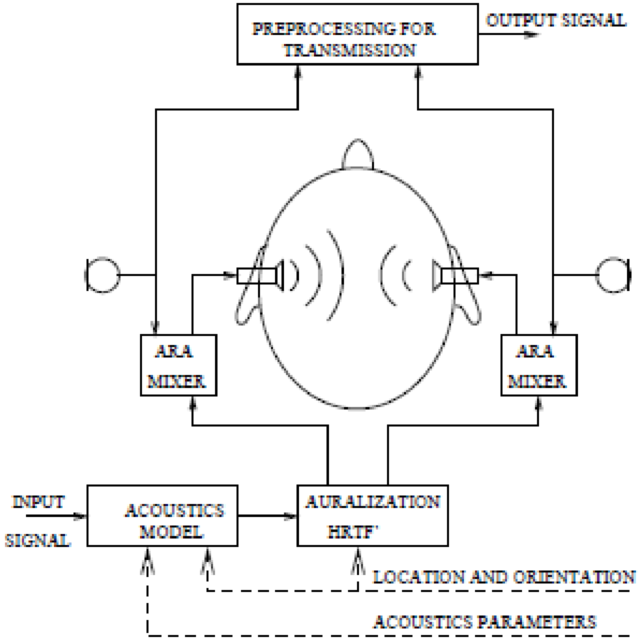
[Ranjan and Gan, 2015]



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

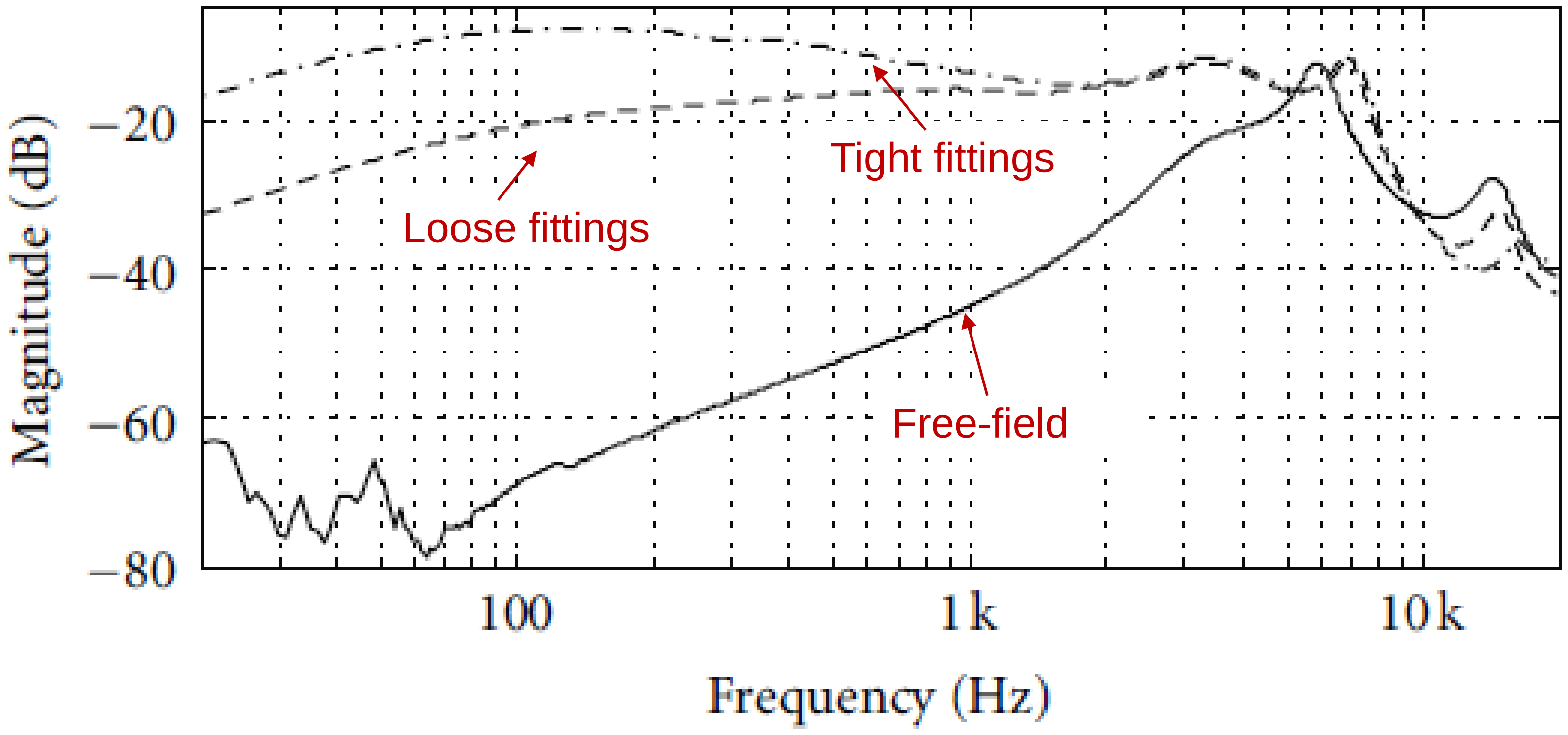


ARA headset system diagram

[Härmä *et al* 2004; Tikander *et al* 2008]

Type III – (Headphone Fittings)

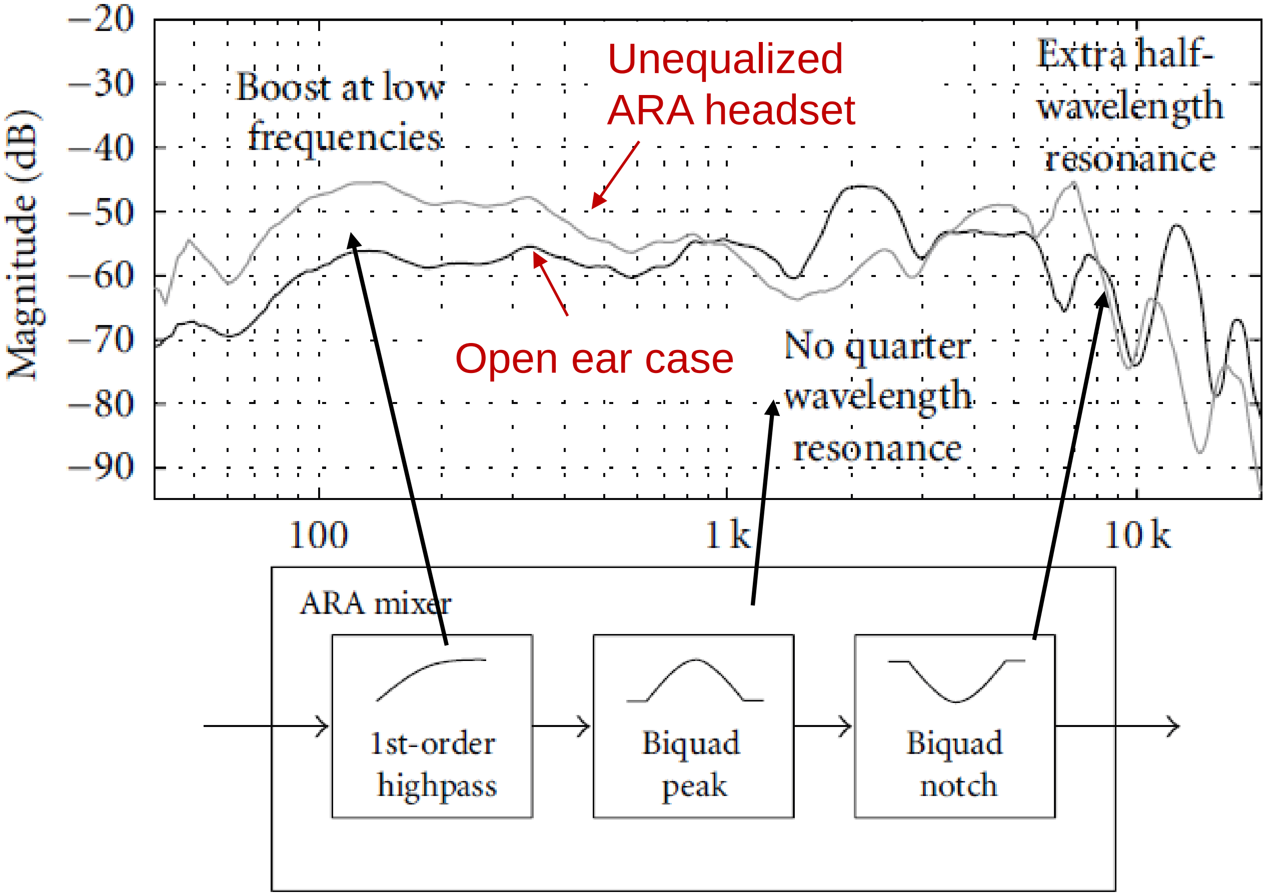
- Difficult to predict the headphone response if loosely fitted



[Rämö and Välimäki, 2012]

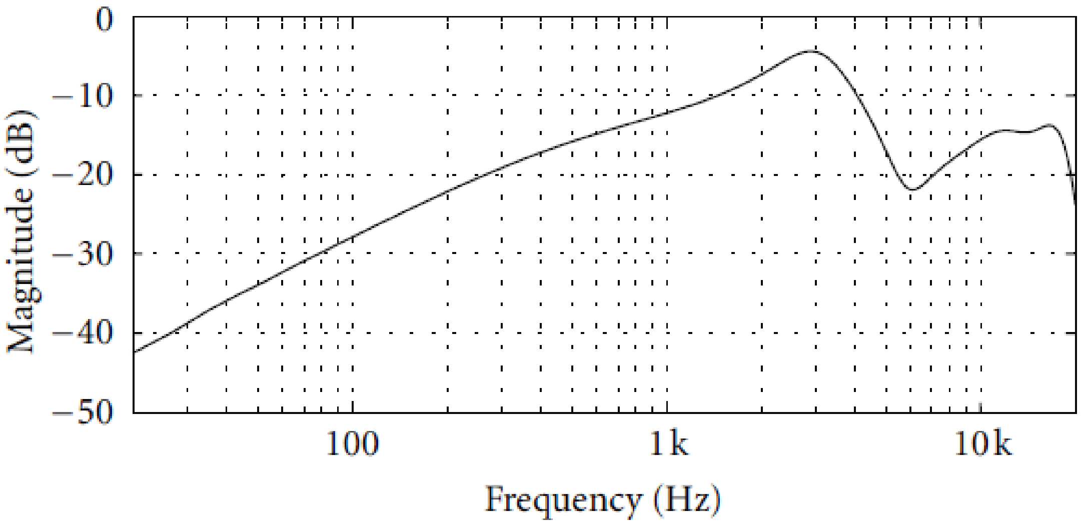
Type III – Hear Through Equalizer

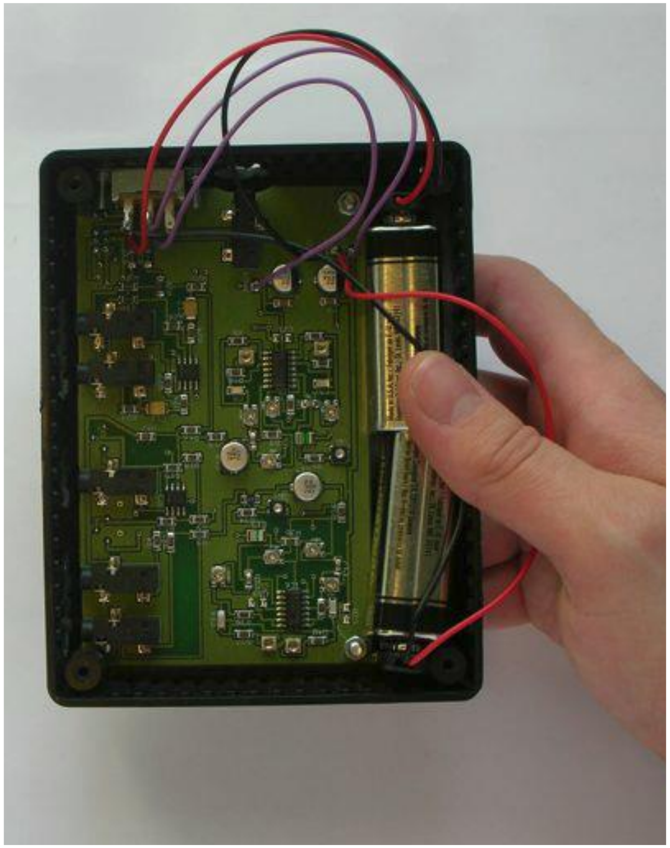
- Closed in-ear phones modify the ear canal resonance due to blocked ear canal (pressure chamber principle)



Type III – ARA Headset Equalizer

- Generic ARA equalization based on 4 individual measurements

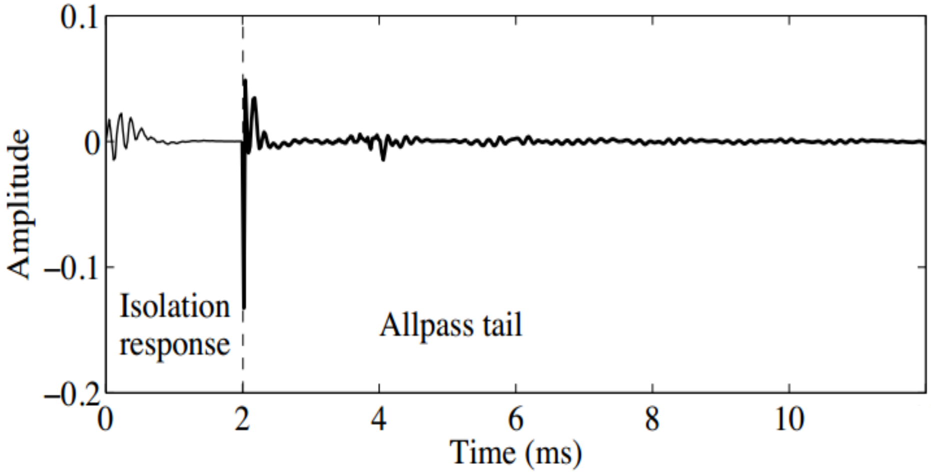




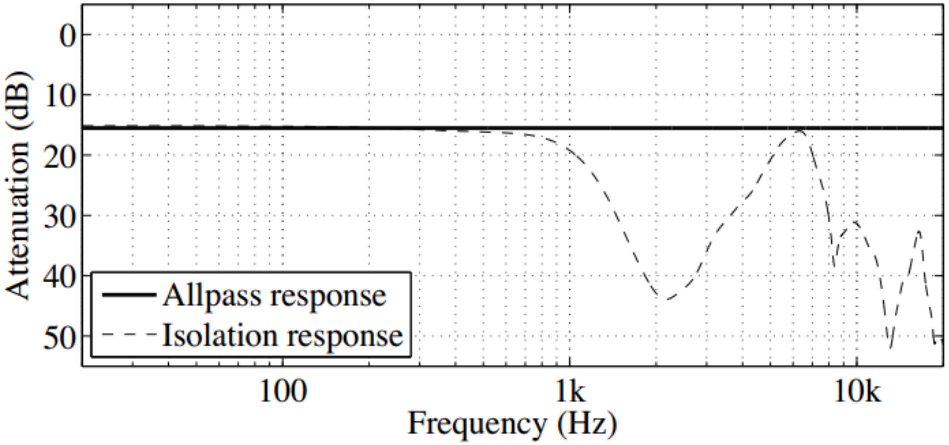
[Rämö and Välimäki, 2012]

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:



Extended impulse response with all pass tail

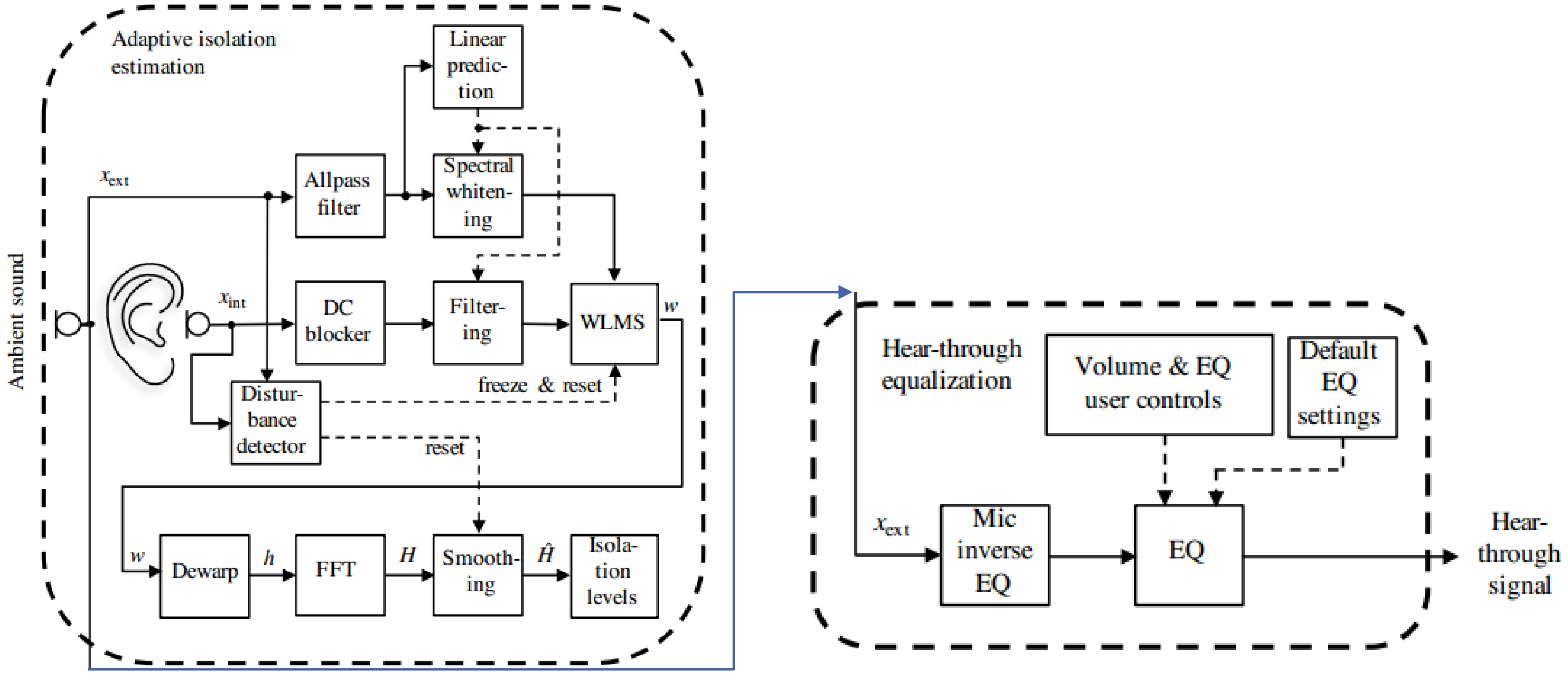


All pass system magnitude

[Rämö and Välimäki, 2014]

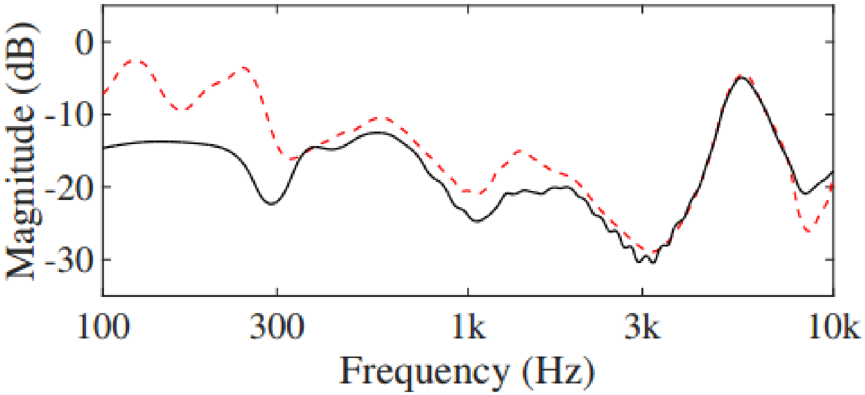
Type III – Adaptive Hear Through

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

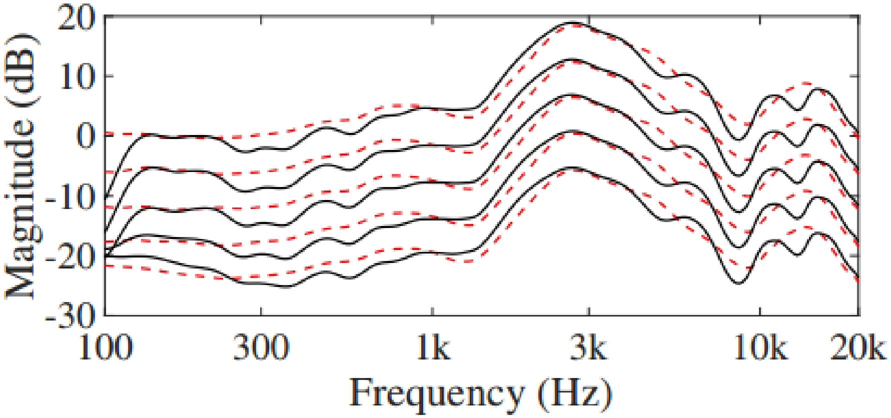


[Juho *et al*, 2016]

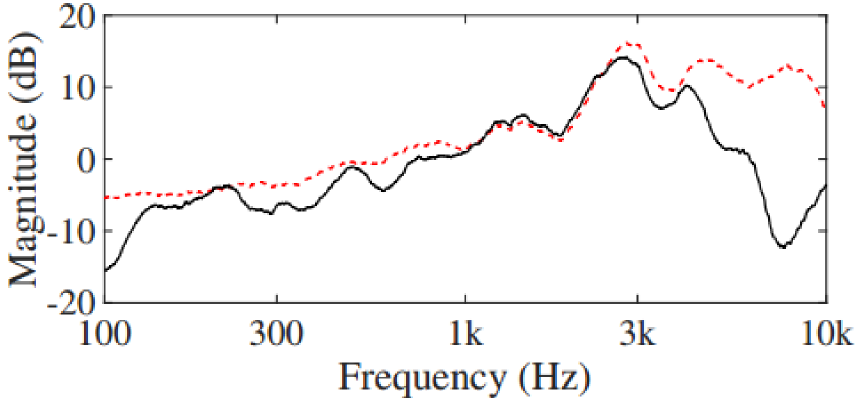
Type III – Adaptive Hear Through - Results



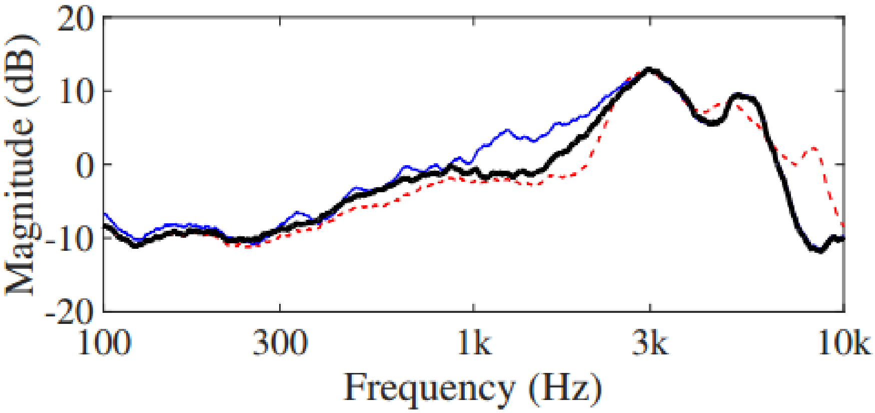
Measured Isolation (dashed red) Vs Estimated Isolation Curve (solid black)



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



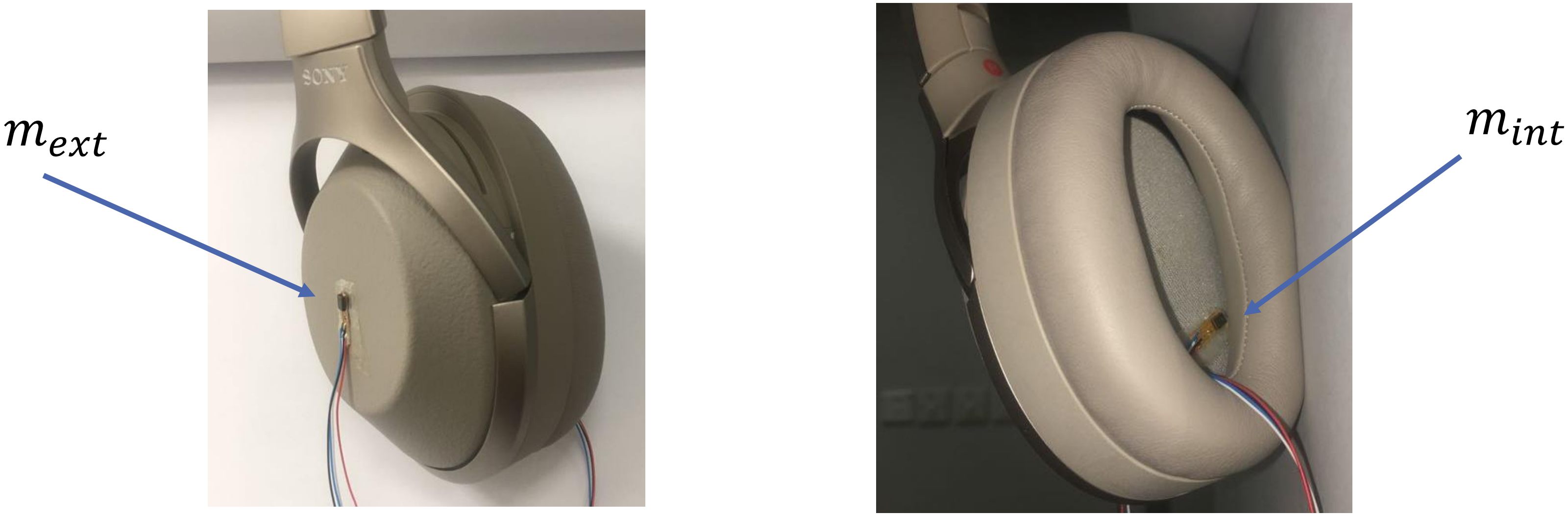
Measured Isolation (dashed red) Vs Estimated Isolation Curve (solid black) for non-ideal direction (75^o)



Effect of automatic EQ on poor headphone fitting. Fixed EQ (Solid blue) Vs Automatic EQ (Solid Black)

Type IV – Closed-back Design

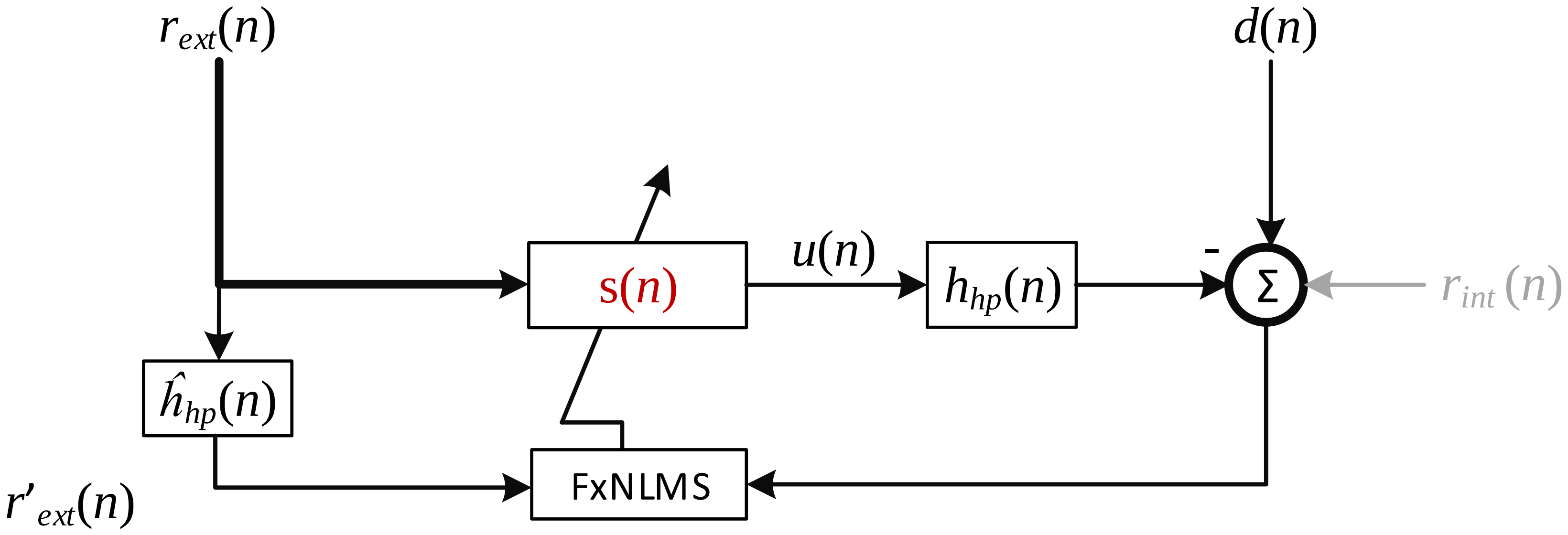
- Closed back ear cup design with internal and external microphones



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



$r_{int}(n)$: 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.

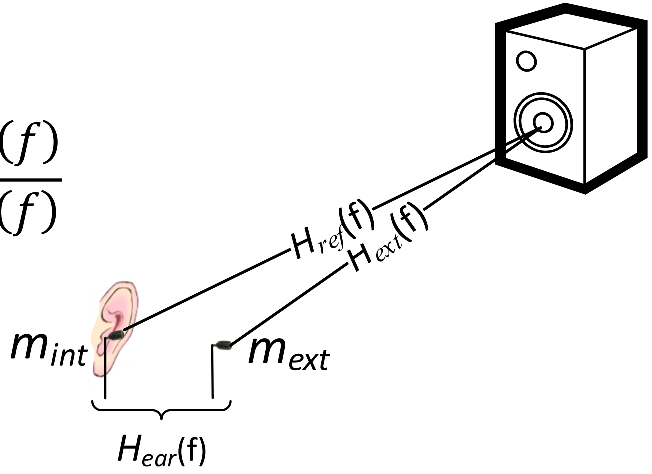
Type IV – Closed-Back Hear Through using FxLMS

Rewriting ideal EQ filter:

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

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

$$H_{ear}(f) = \frac{H_{ref}(f)}{H_{ext}(f)}$$

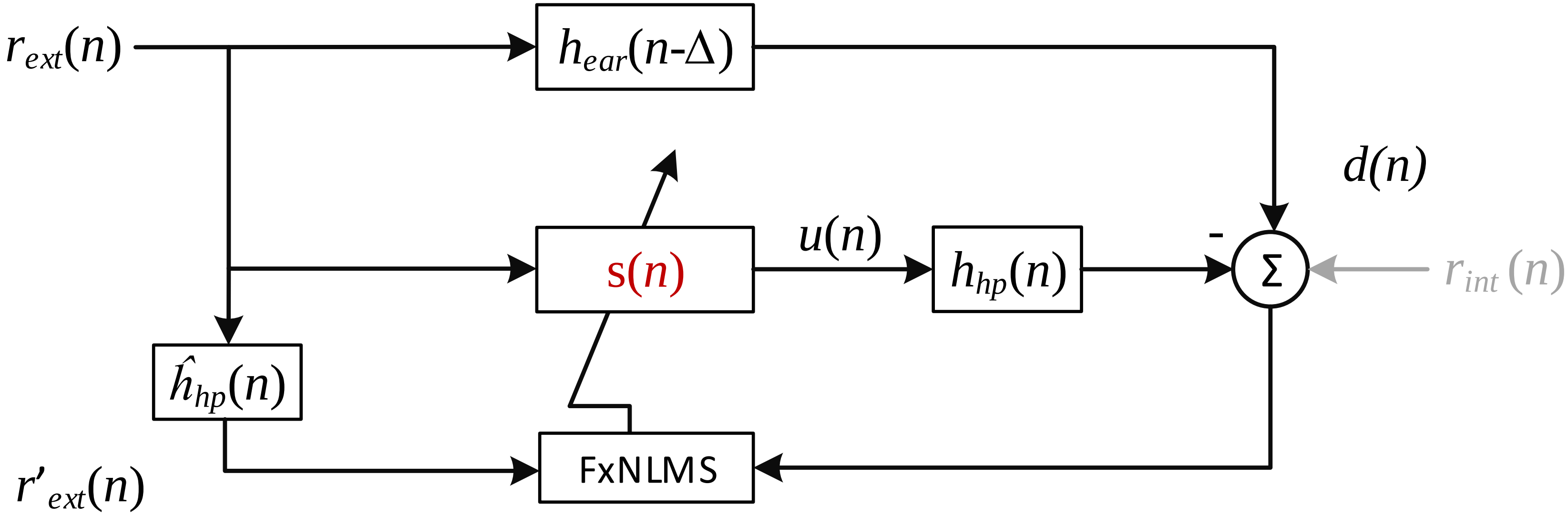


Rewriting desired signal:

$$\begin{aligned} d(n) &= r(n) * h_{ref}(n) \\ &= r(n) * h_{ext}(n) * h_{ear}(n) \\ &= r_{ext}(n) * h_{ear}(n) \end{aligned}$$

Desired signal can be equivalently expressed as $r_{ext}(n)$ passing through a directional filter $h_{ear}(n)$

Type IV – Closed-Back Hear Through using FxLMS



Δ: Minimum estimated group delay of secondary path

Un-equalized response: $H_{ext}(f)H_{hp}(f) + H_{int}(f)$

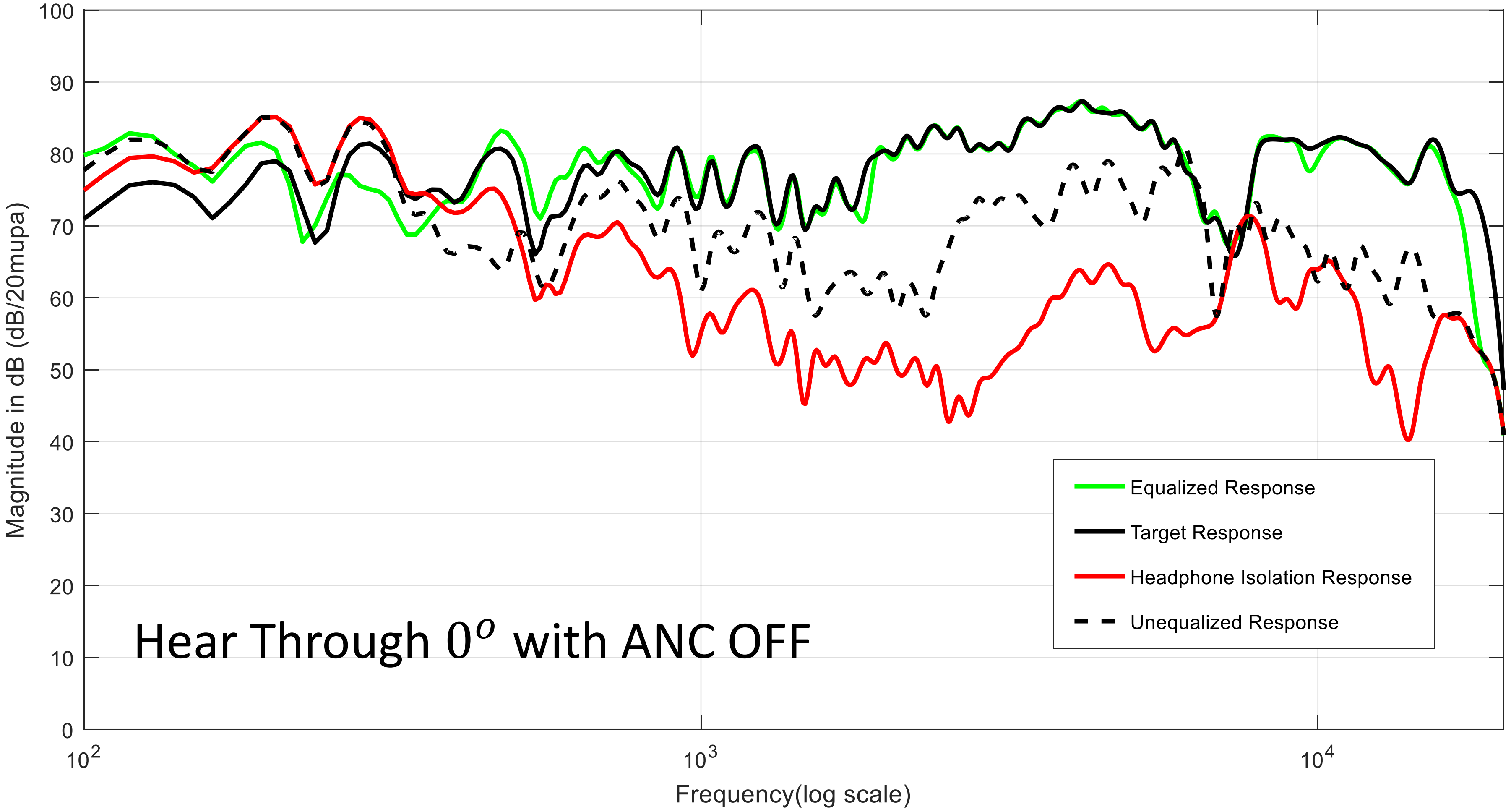
Equalized Transfer Function: $H_{ext}(f)S(f)H_{hp}(f) + H_{int}(f)$

Closed-back hear through EQ can be computed:

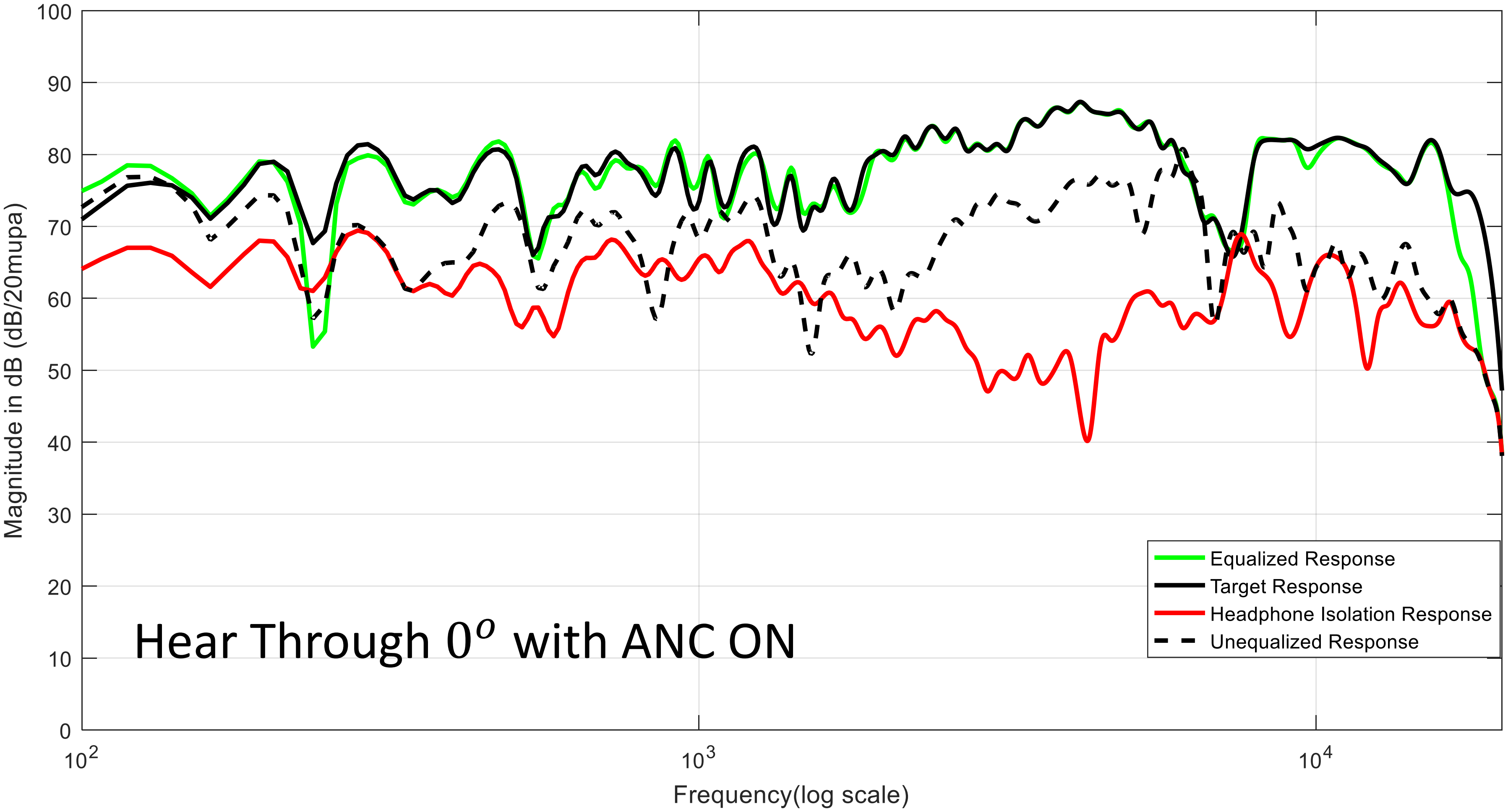
(1) If Directional transfer function, $h_{ear}(n)$ is known (can be modelled), And

(2) Introducing a minimum delay in primary path to ensure EQ can converge

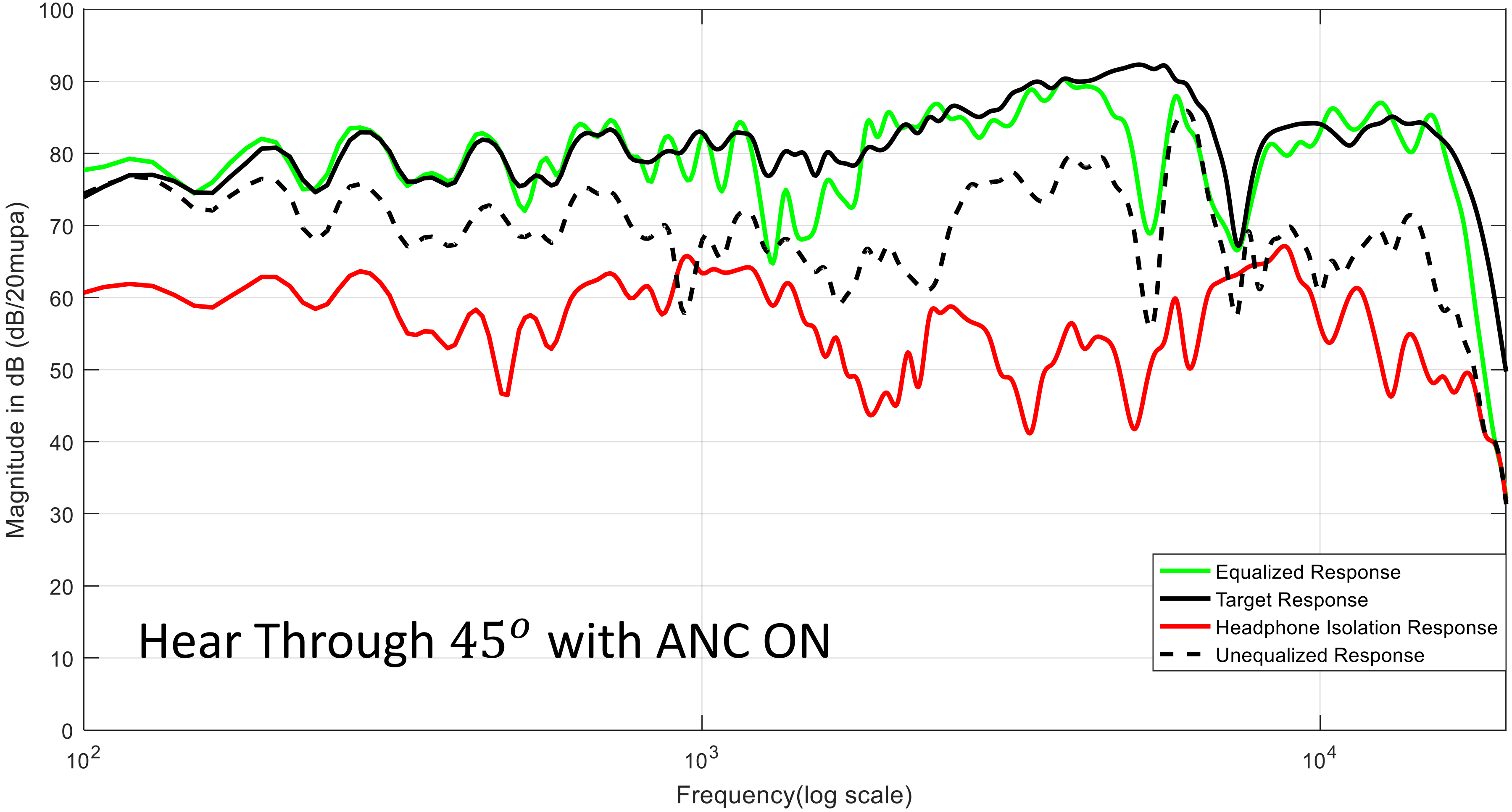
Type IV – Closed-Back Hear Through Results



Type IV – Closed-Back Hear Through Results



Type IV – Closed-Back Hear Through Results



What’s next…

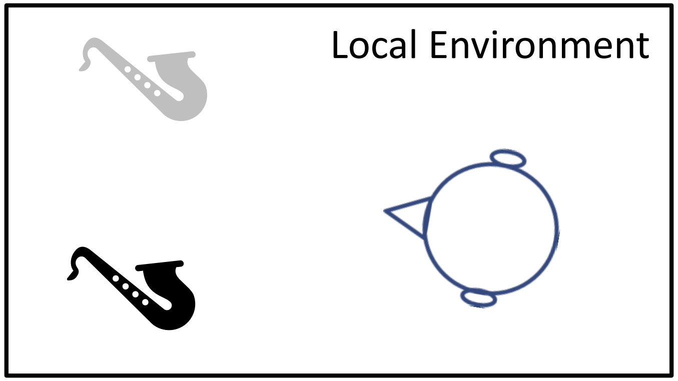
- Hear Through EQ may also vary for different source incident directions
    - One fixed average EQ Vs group of EQs
- Diffuse sound field or multiple real sound sources scenarios
- Headphone isolation can be highly idiosyncratic (especially for closed-back headphones)
- Perceptual evaluation of localization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

Natural Listening in AR/MR - Summary

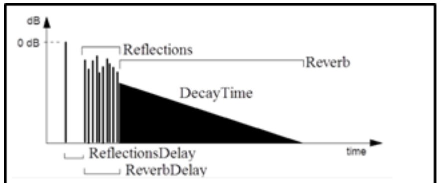
	Type I – Open ear	Type II – Open-back Over ear	Type III – Closed in-ear	Type IV – Closed-back Over ear
Real sound reproduction	Heard as is – No processing required	Only higher frequencies may be compensated	Recorded, processed for entire spectrum	
Characteristics	Natural listening, No obstruction	Natural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

C.6 Acoustic Environment Estimation and Rendering

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

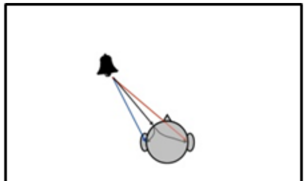


- Acoustic environment characteristics must be captured and embedded into the binaural playback



Local acoustic environment

+

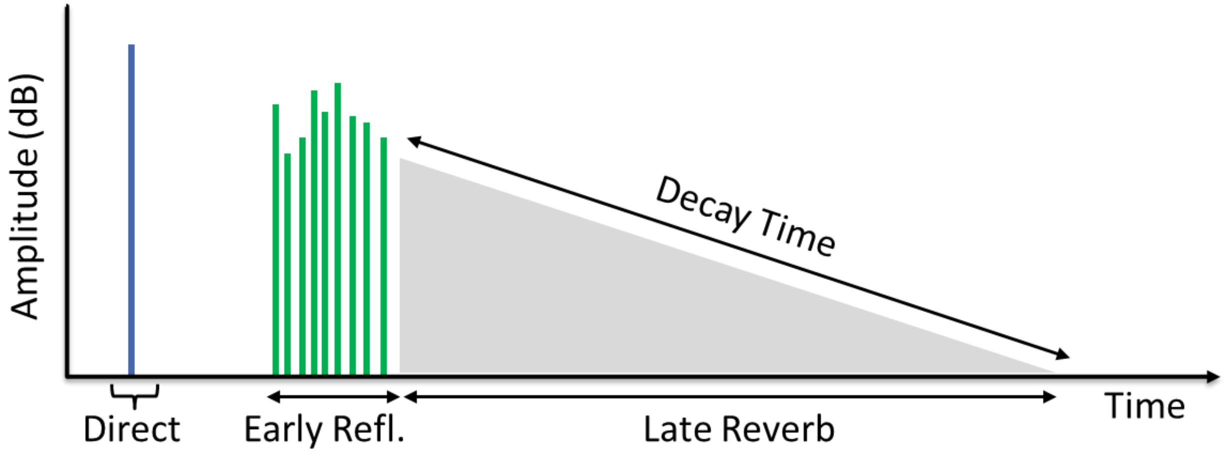


Binaural render

= Natural Listening AR audio

Acoustic Environment – Room Impulse Response

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



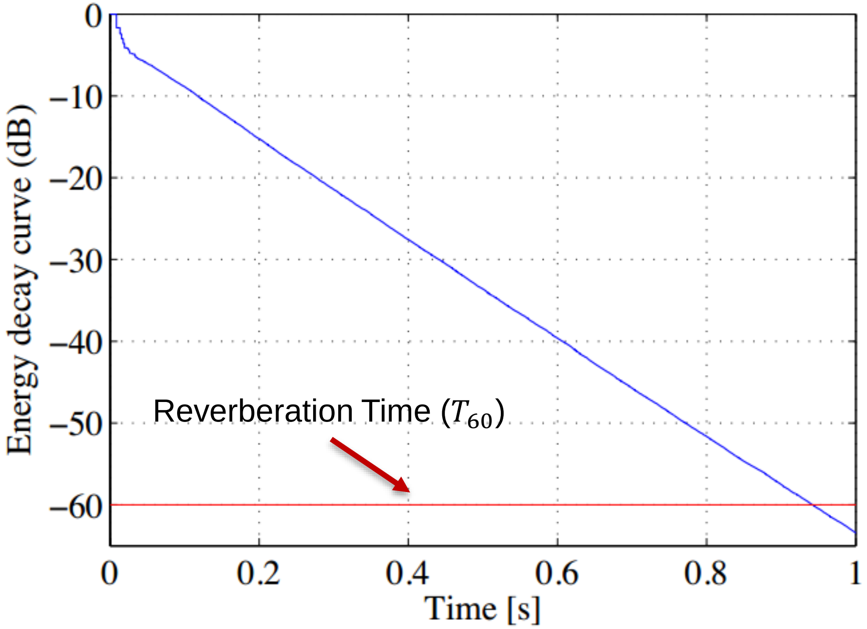
- Three major components of RIR:
  - Direct Sound: Straight path between Source and Receiver
  - Early Reflections: Sparse first few reflections from source to receiver
  - Reverberations: densely populated reflections (best described statistically)

RIR – Energy Decay Curve

- Energy decay curve (EDC): signal energy remaining in RIR at time *t*

$$EDC(t) = \frac{\int_t^{\infty} h^2(\tau)\mathrm{d}\,\tau}{\int_0^{\infty} h^2(\tau)\mathrm{d}\,\tau}$$

- Reverberation Time (T_{60}): Time when EDC crosses -60 dB



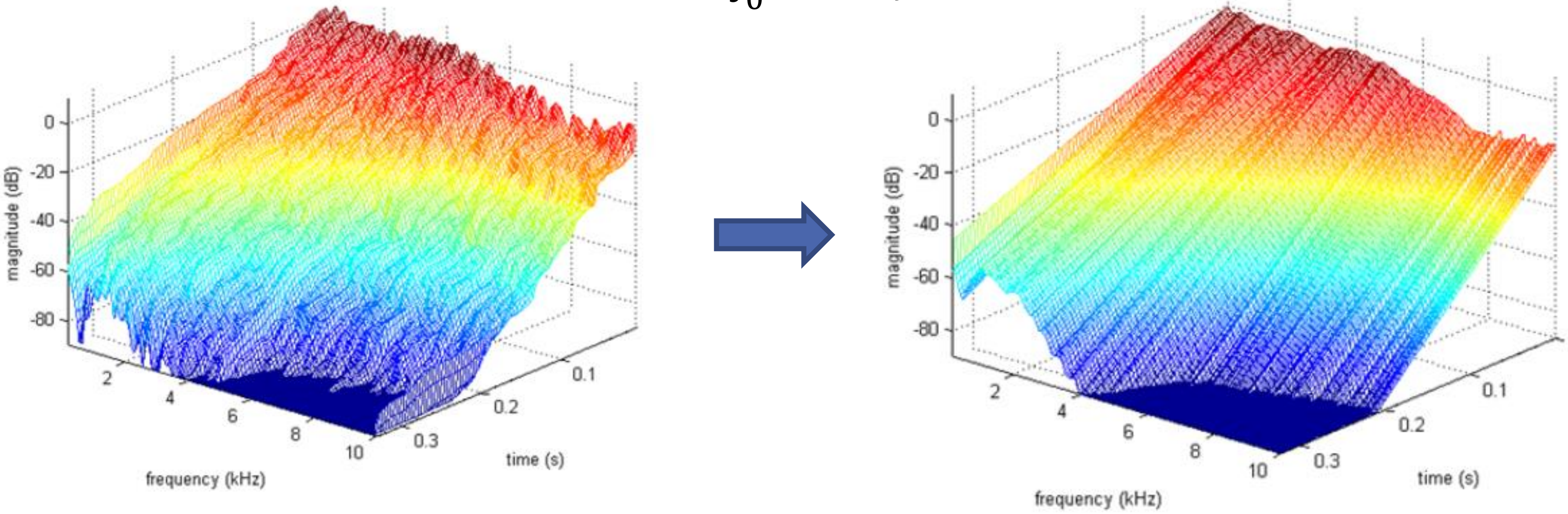
RIR – Energy Decay Relief

- Energy Decay Relief (EDR): EDC generalized to frequency bands

[Jot, 1992]

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

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



Measured EDR

Modelled EDR

[Jot and Lee, 2016]

RIR – Schroeder frequency

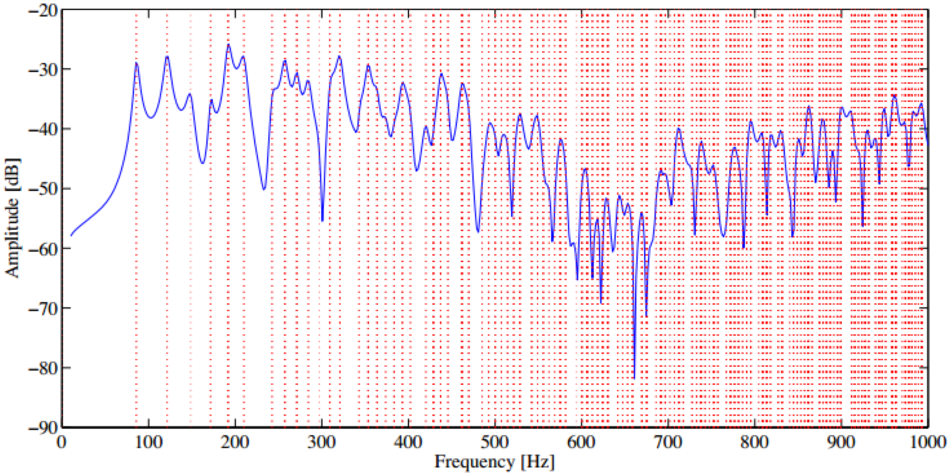
- Frequency response of RIR can be divided into two regions:
  - Sparsely distributed low frequency resonant modes
  - Densely populated resonance modes
- Schroeder frequency defined as transition frequency between the two regions:

[Schroeder, 1962]

$$F_C = 2000\sqrt{\frac{T_{60}}{V}}$$

Example:

Bathroom V = 10 m^3, T_{60} = 0.35s
Fc = 374 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
- Solve Helmholtz Wave Equation

$$\frac{\partial^2 p}{\partial t^2} - c^2 \nabla^2 p = F(\mathbf{x}, t)$$

Wave Based Methods Summary	
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

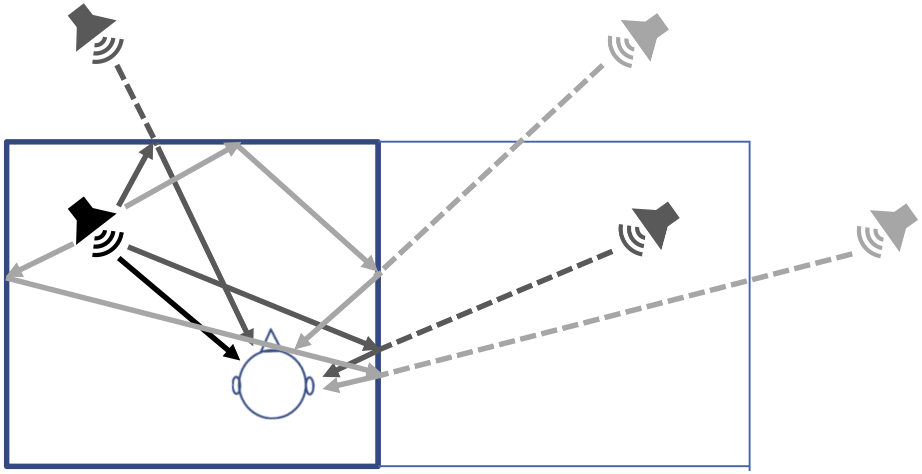
If they are the most accurate, then why don't we usually use these methods for real-time acoustics simulation of any environment?

- 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 [Raghuvanshi *et al*, 2009,2010]
- Current limitations:
  - Static Scenes and high memory requirement
  - Low frequencies up to 1.5 kHz for medium sized room

Geometrical Acoustics: Image source method

- Source reflections are created using image equivalence

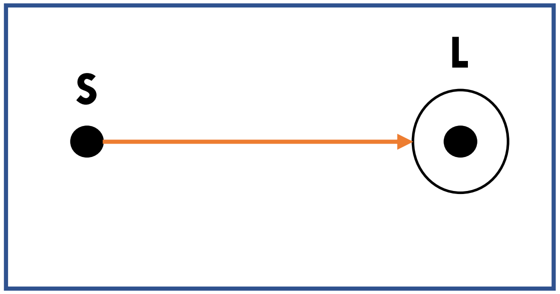
- Start from Source -> Reflect against all rigid walls -> Check for listener visibility -> Reflect image sources -> And so on...



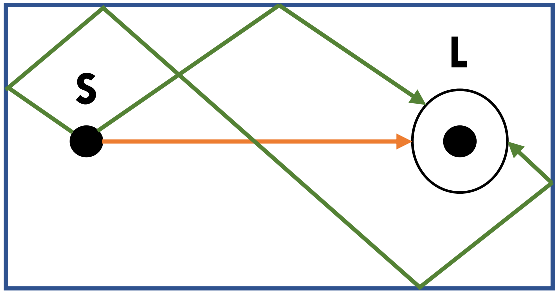
- Accurate but number of image sources increases exponentially after first few order of reflections -> Truncated

- Wall surfaces are assumed to be smooth i.e., only specular reflections are allowed

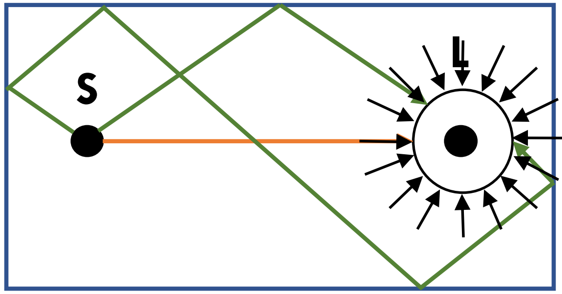
Geometrical Acoustics: Ray Tracing



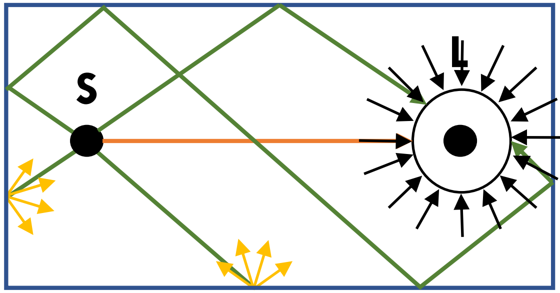
Direct sound path between source and listener
(distance attenuation)



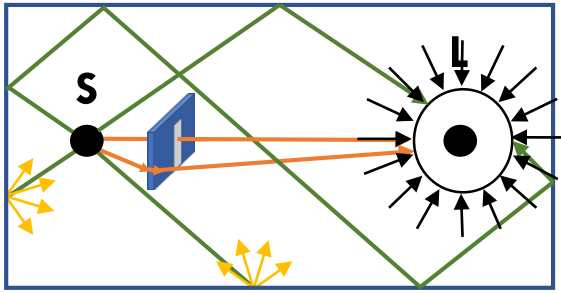
Early/Specular reflections
(multiple sound paths due to reflecting surfaces)



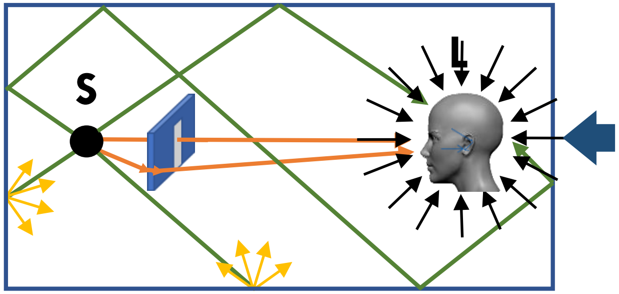
Reverberations
(Have no direction & densely populated)



Diffuse reflections
(Scattering due to roughness of the surfaces)



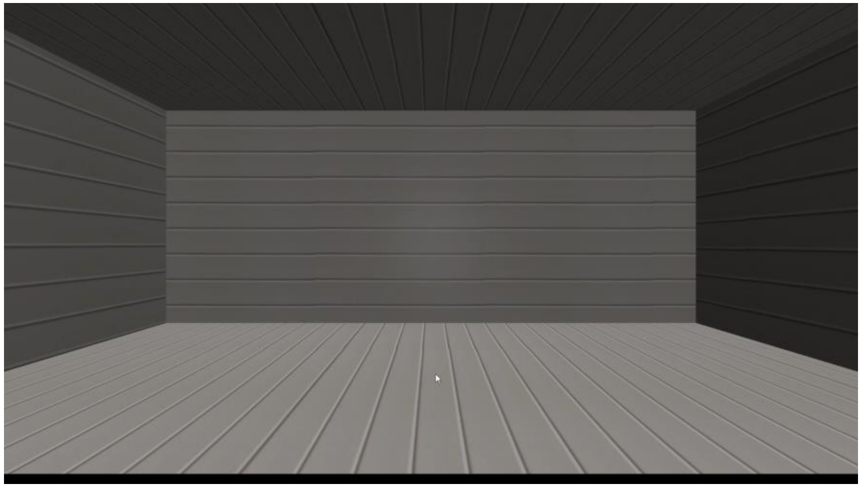
Diffraction (occlusion effect)
(Sound reflects around object edges and changes phase)



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

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

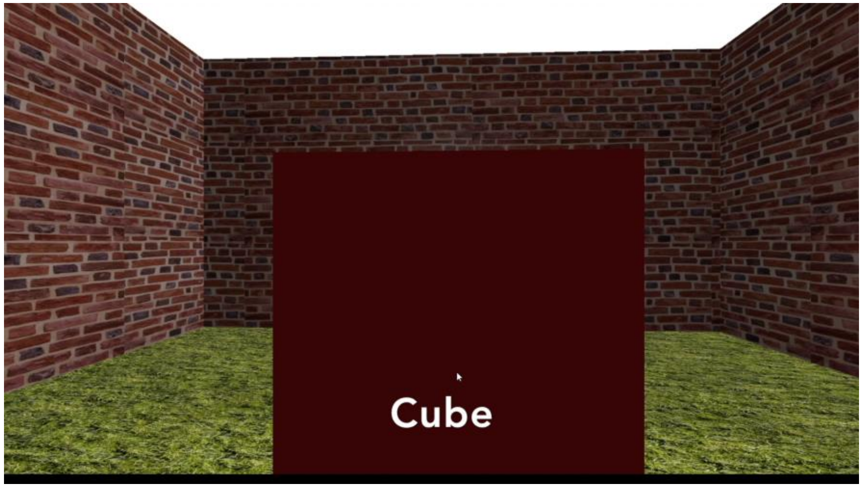
Demo scenes



Room materials effect

Acoustics reflections of different material surfaces produces realistic sound effects.

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



Occlusion effect

[Source: Immerzen Labs Pte. Ltd. Singapore]

Perceptual Methods

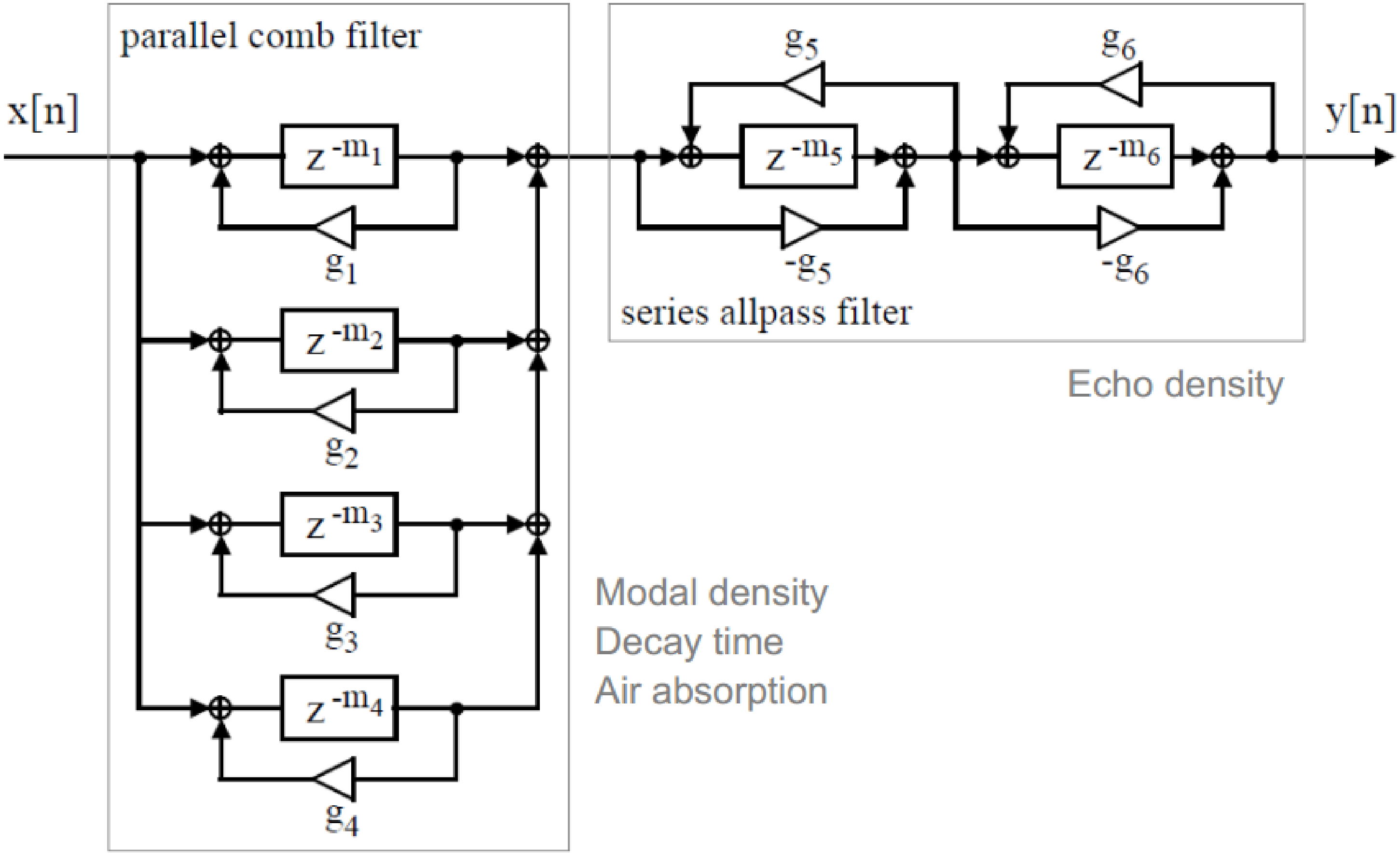
- There are too much details in physical methods
- If physical accuracy not required, perceptual methods are better alternative
- To simulate what is perceptually important NOT physically

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



Echo density

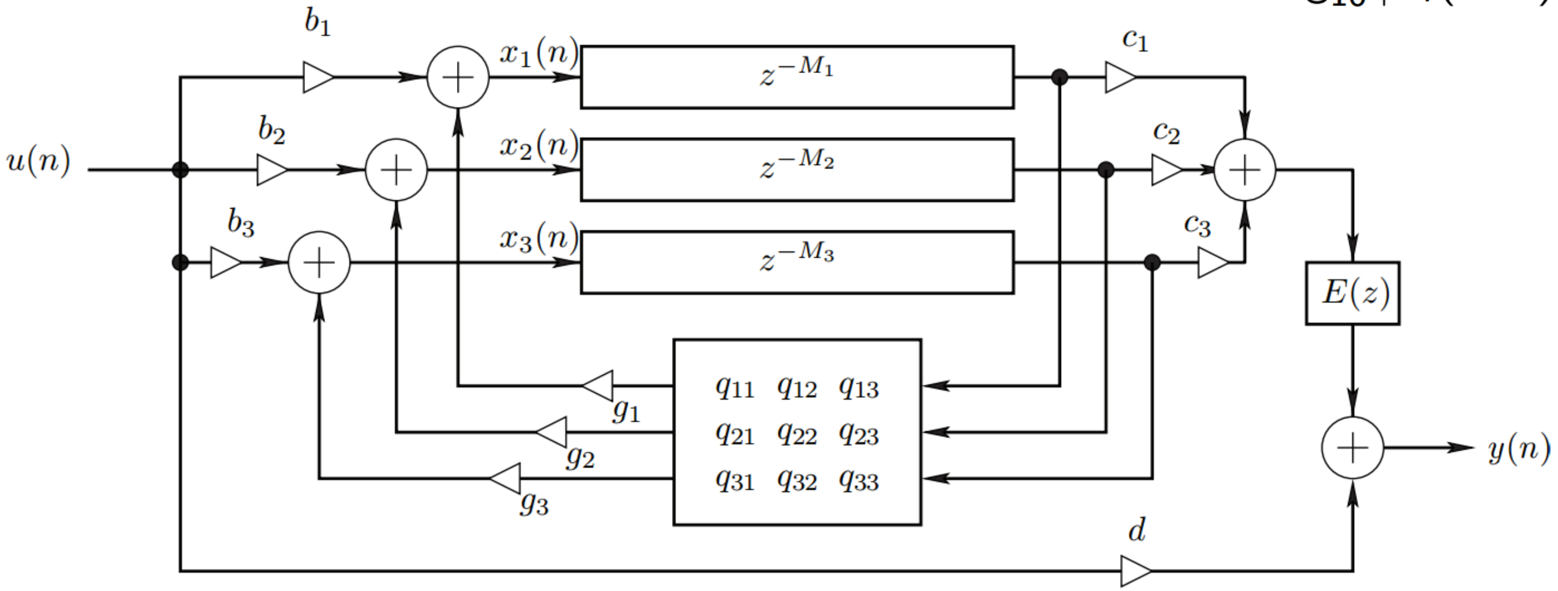
Modal density
Decay time
Air absorption

[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 $T_{60}(f)$

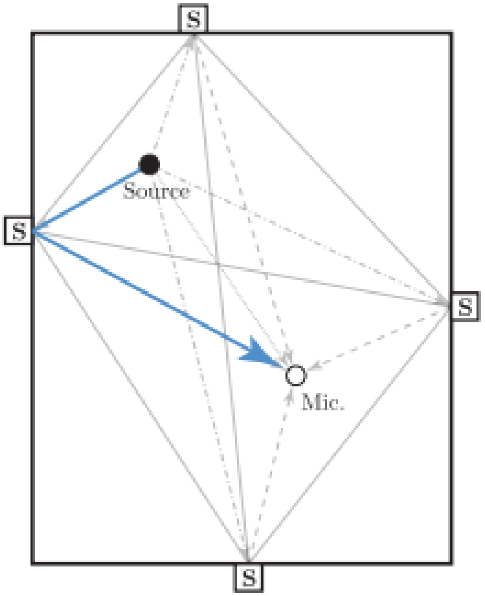
$$20\log_{10}|G_i(e^{j\omega t})| = -60\frac{M_i T}{T_{60}(\omega)}$$



[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

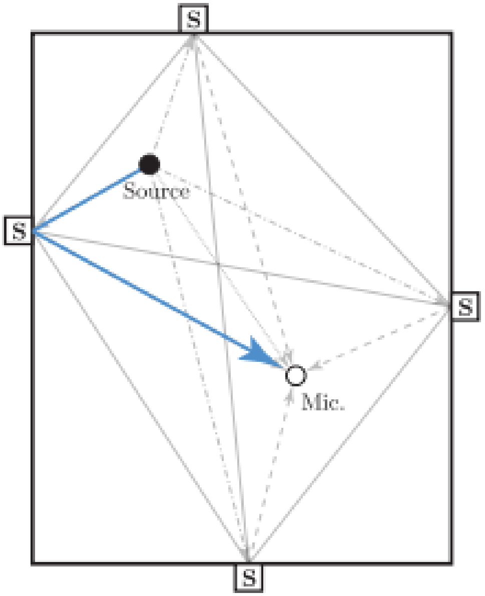


I-order reflection

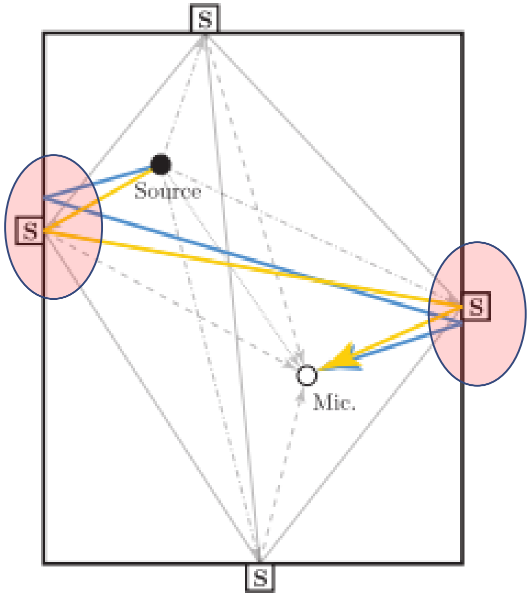
[Karjalainen *et al.*, 2005; Sena *et al.*, 2015]

Scattering Delay Network

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



I-order reflection

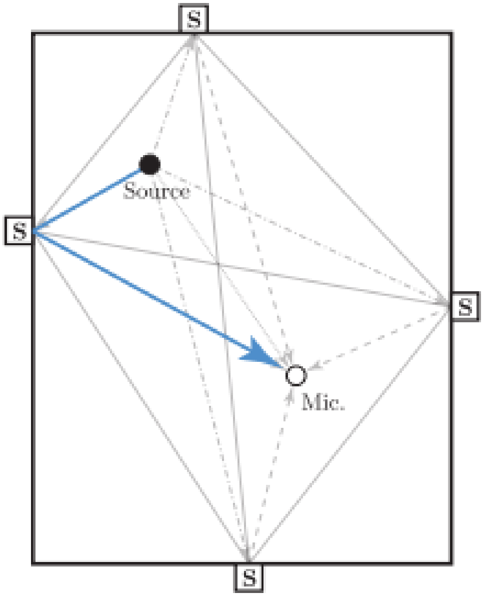


II-order reflection

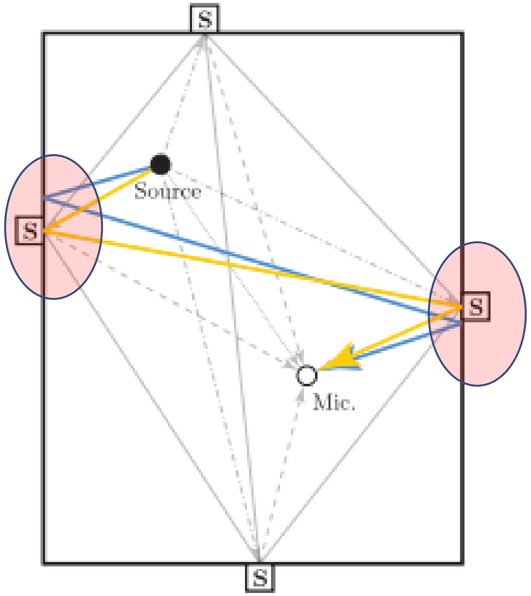
[Karjalainen *et al.*, 2005; Sena *et al.*, 2015]

Scattering Delay Network

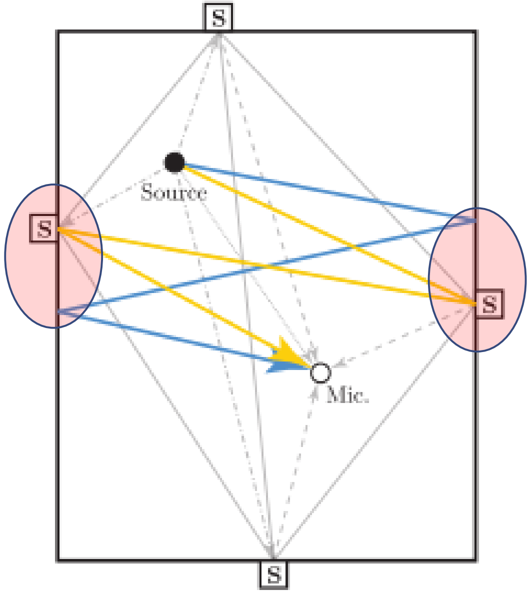
- Approach in between delay networks and physical models
- One node per reflecting surface
- Approximation of image source method



I-order reflection



II-order reflection



Another II-order reflection

[Karjalainen *et al.*, 2005; Sena *et al.*, 2015]

Environment Rendering Methods - summary

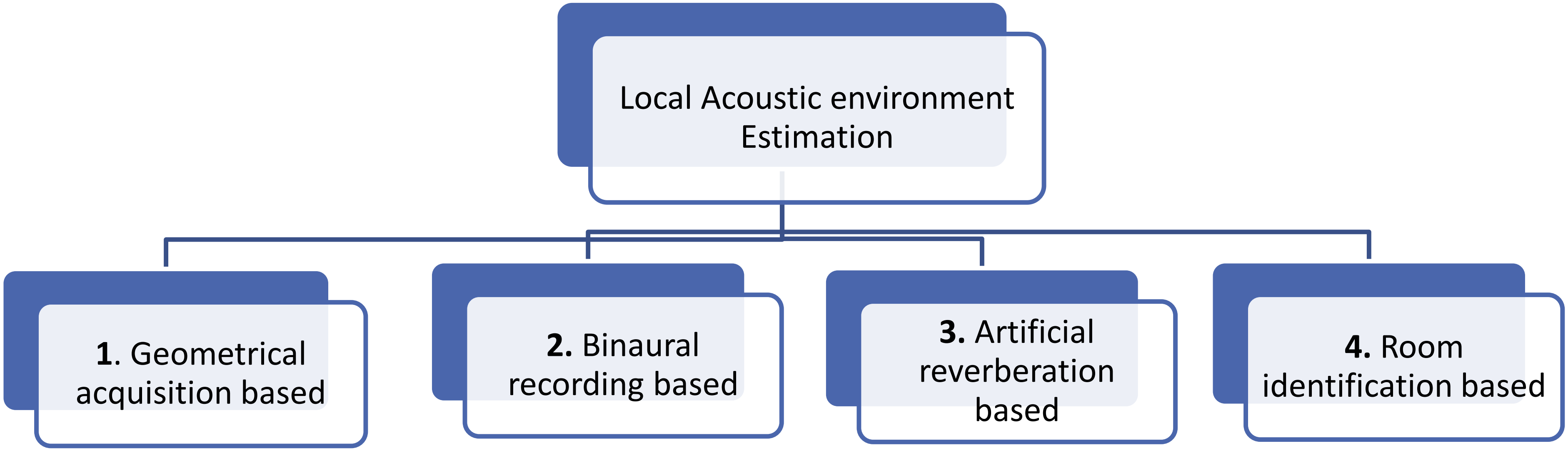
- Wave Methods:
  - Infeasible for high frequencies
- Geometrical Acoustics
  - Image source:
    - Only possible for early reflections. Usually combined with ray tracing for accuracy
  - Ray Tracing:
    - High frequency approximation
    - Choice of number of rays and size of source & listener is critical
    - 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

Method	Speed/Load	Accuracy	Interactivity
Wave Based	Slow, Very high	Excellent	Yes
Geometrical Acoustics	Very fast, High	Very good at high frequency, medium at low frequency	Yes
Hybrid Wave(LF) + GA(HF)	Fast, High	Very good	Yes
FDN	Very Fast, Low	Poor	No
SDN	Very Fast, Low	Medium	Yes

Acoustic Environment Estimation



Local Acoustic environment Estimation

1. Geometrical acquisition based

2. Binaural recording based

3. Artificial reverberation based

4. Room identification based

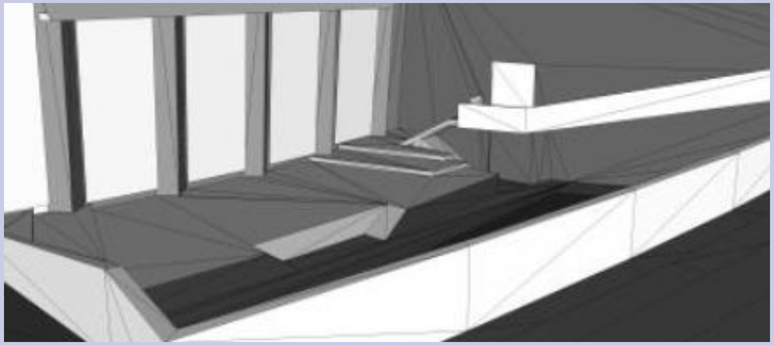
Geometry Acquisition and Acoustics Processing

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

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

Geometry Acquisition

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



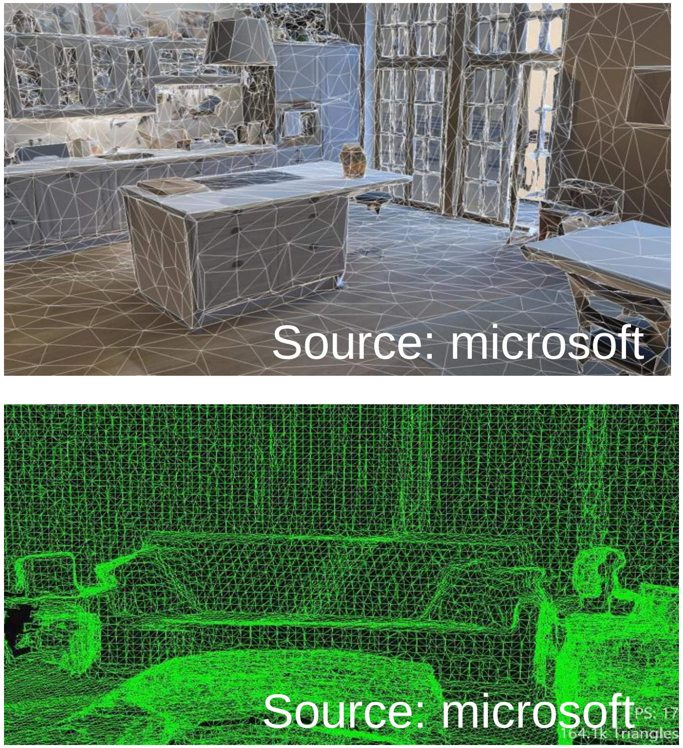
Geometrical Acoustics Processing

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*


Source: microsoft

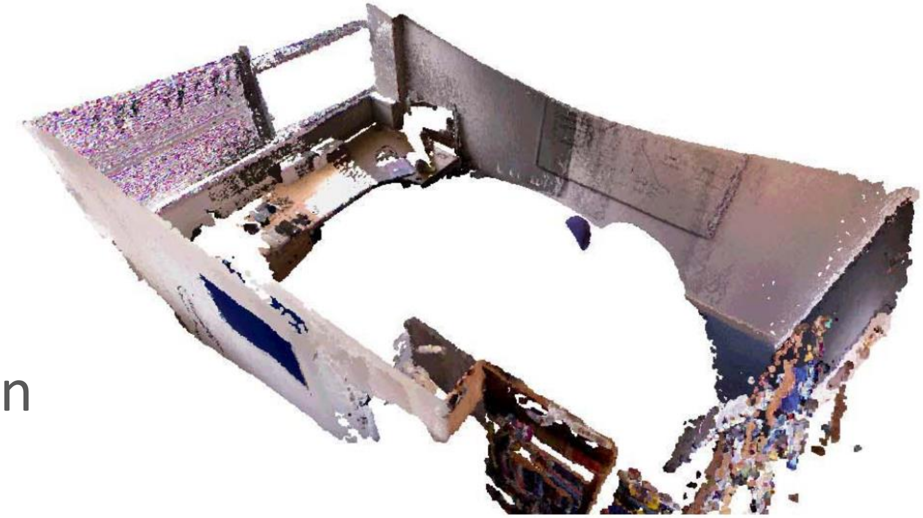
Source: microsoft

3D spatial mesh of environment

* 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]

Surface Recognition for Acoustical Simulation

- 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

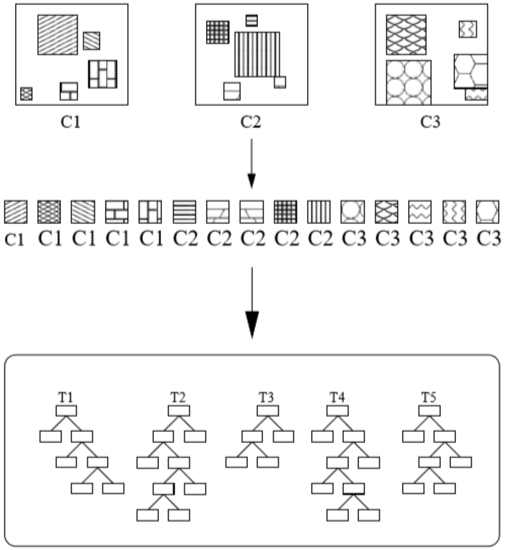
[David, *et al* 2012]

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

[Milos *et al*, 2013]

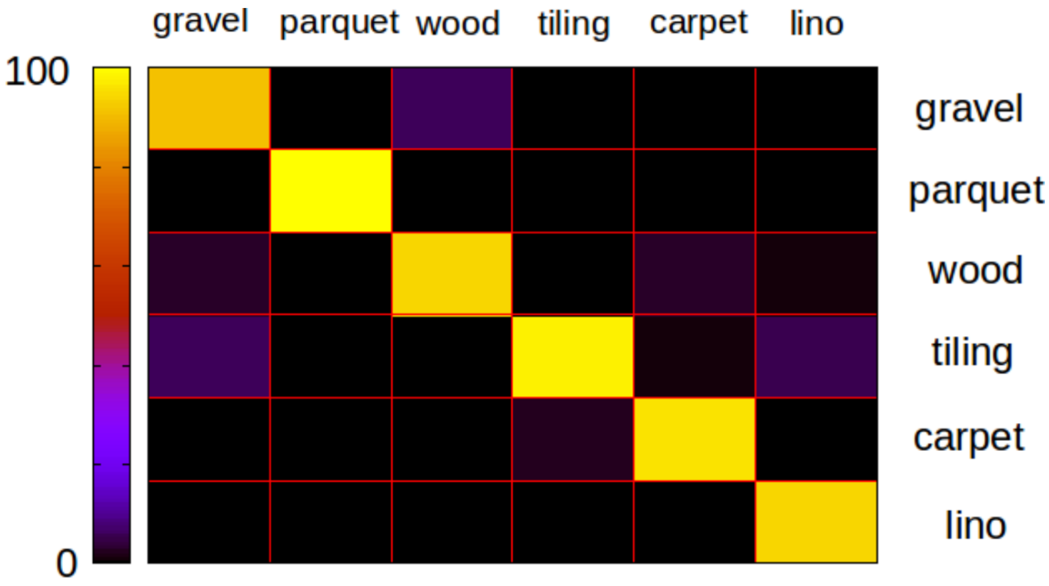
Surface Recognition for Acoustical Simulation

- 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

[M. Raphael *et al*, 2005]



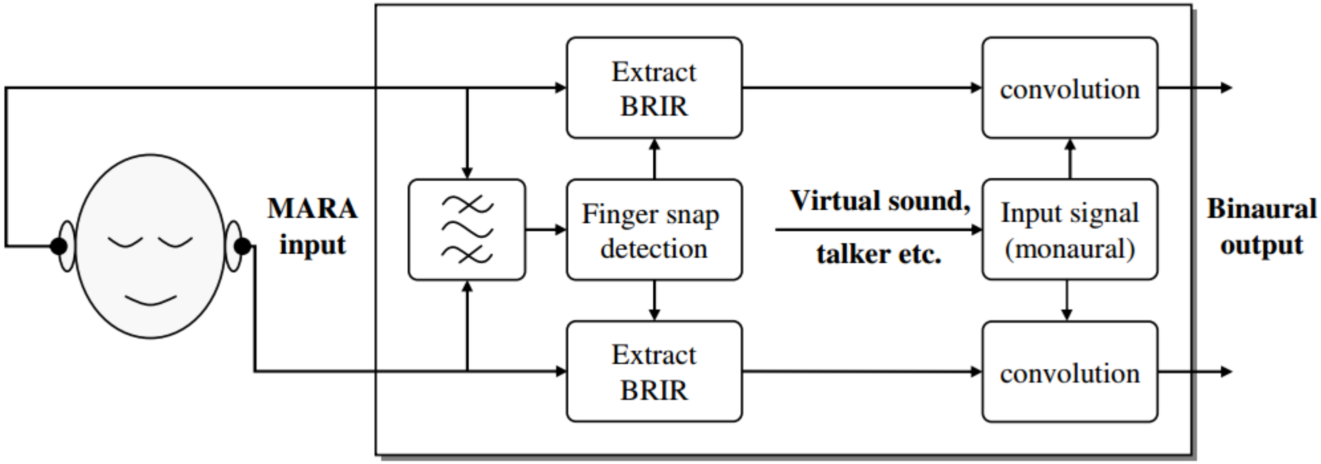
Confusion matrix for different ground materials

[David, *et al*, 2012]

Acquisition using "finger snaps or claps"

- 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

[Hannes, and Lokki, 2009]



Extracted BRIRs response will be colored due to non-flat snap signal spectrum. Other excitation methods with more flat spectrum can be used

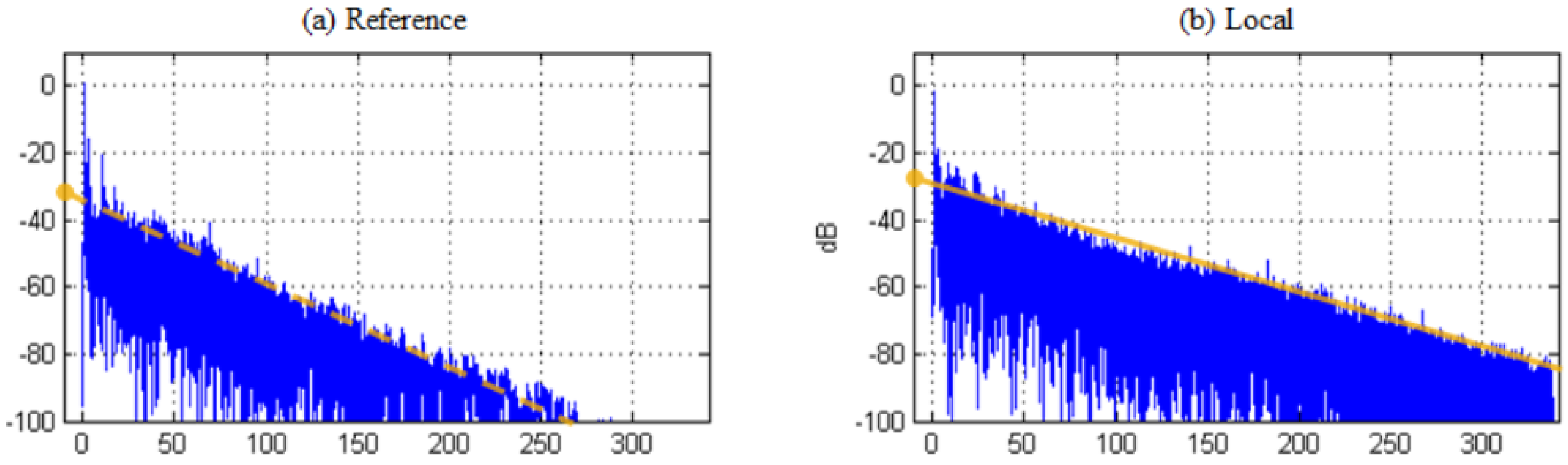
Local Environment adaption – Statistical Approach

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

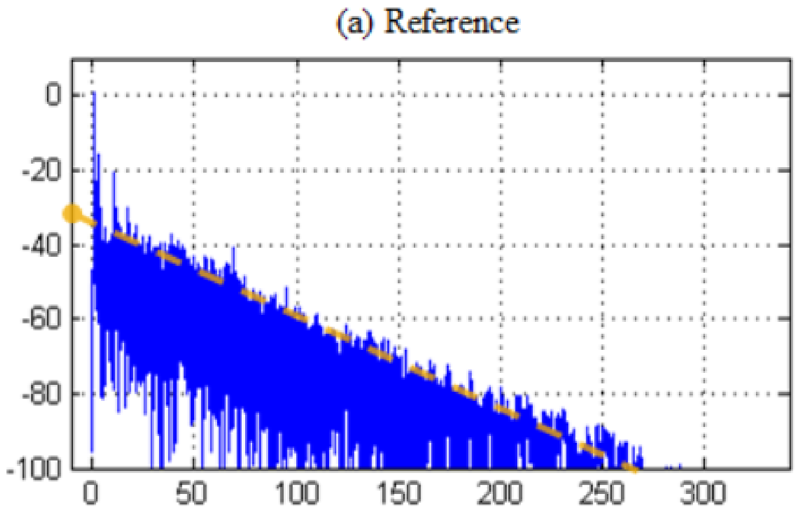
- Reverberation fingerprint of a room :
  - Reverberation Time,$T_{60}(f)$: derived as linear curve fitting on modelled EDR extrapolated back to time of emission $EDR(0, f)$
  - Room Volume, V: Estimated from initial power spectrum $P(f)$ ($\propto 1/V$)
- **Advantage**: Just information of frequency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

[Jot and Lee, 2016]

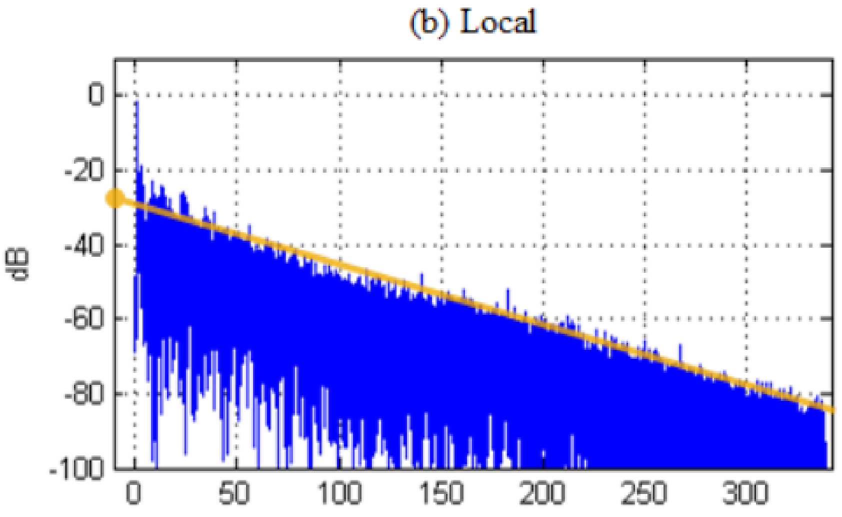
Using Reverberation Fingerprint to match local room



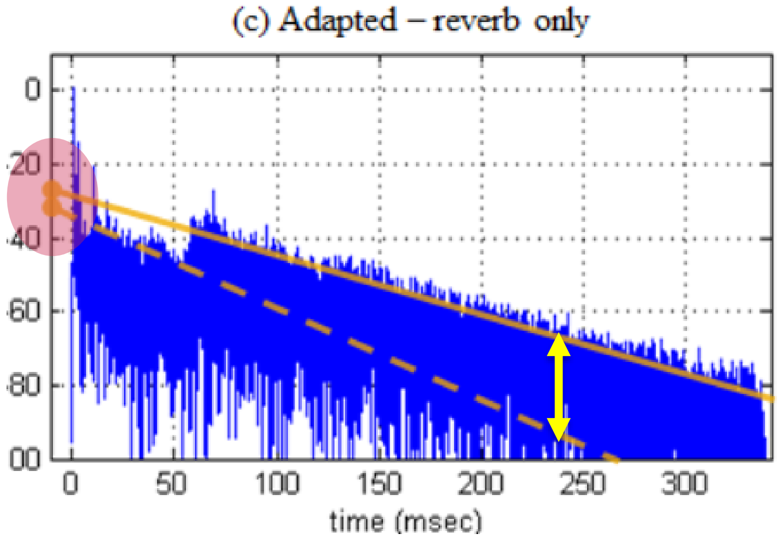
Using Reverberation Fingerprint to match local room



(a) Reference



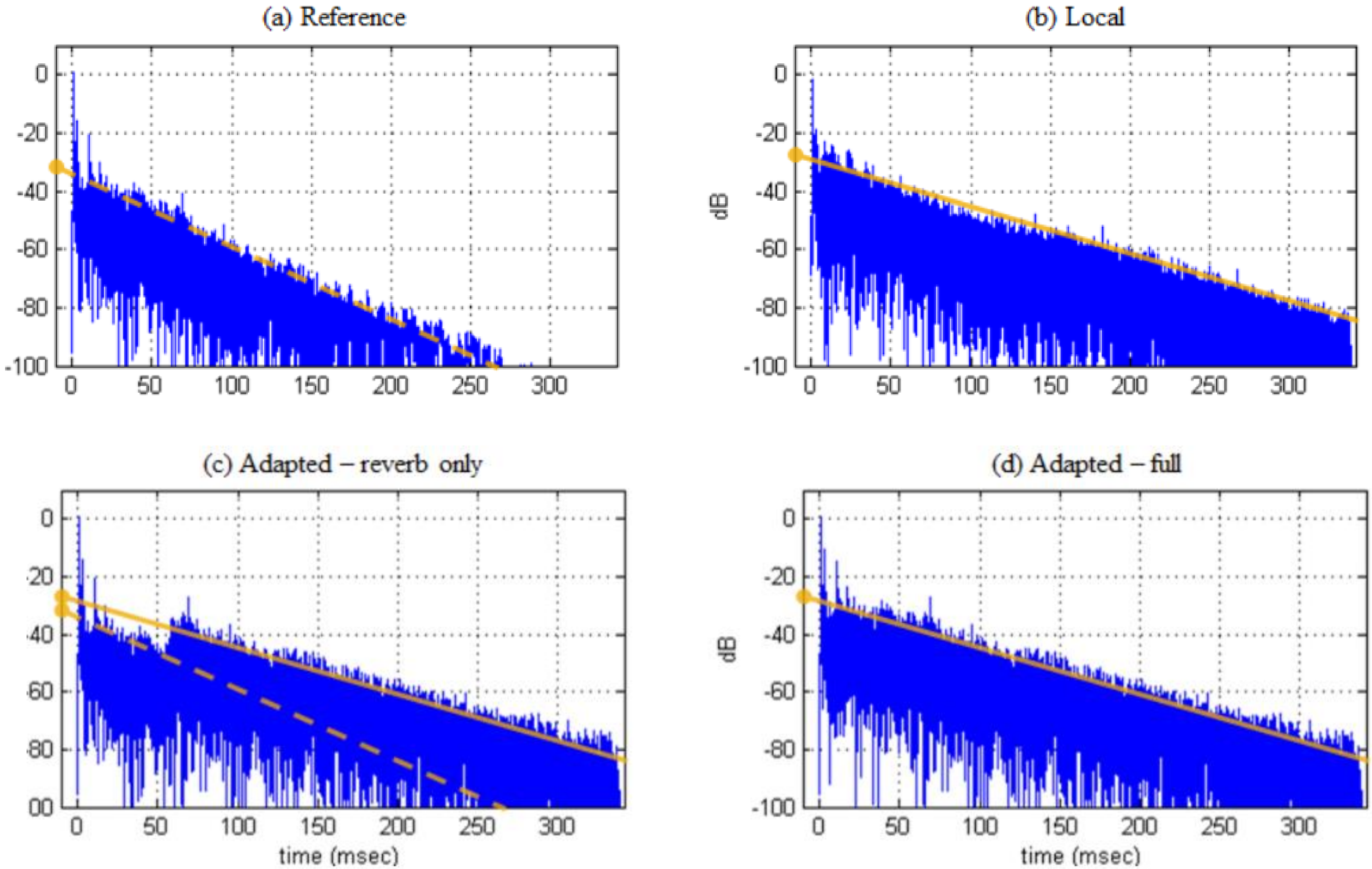
(b) Local



(c) Adapted – reverb only

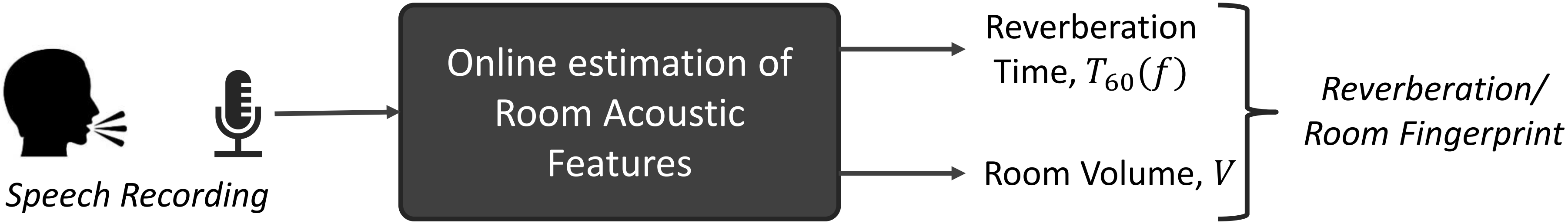
- Initial power offset $\frac{V_{ref}}{V_{local}}$
- Correction of time-frequency envelope using per frequency dB offset

Using Reverberation Fingerprint to match local room



How to obtain Reverberation Fingerprint?

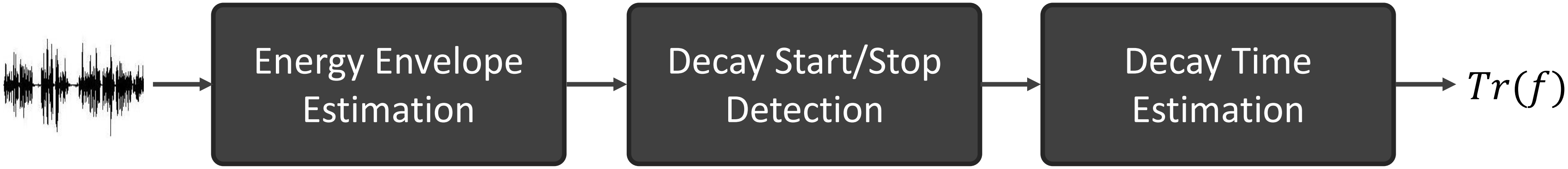
- In an augmented environment, user can be in a space characterized by different acoustic properties
- On-the-fly acquisition using existing audio signals to define *Reverberation Fingerprint* to be further used for rendering
- Blind estimation of room acoustic parameters of a unknown environment using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



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

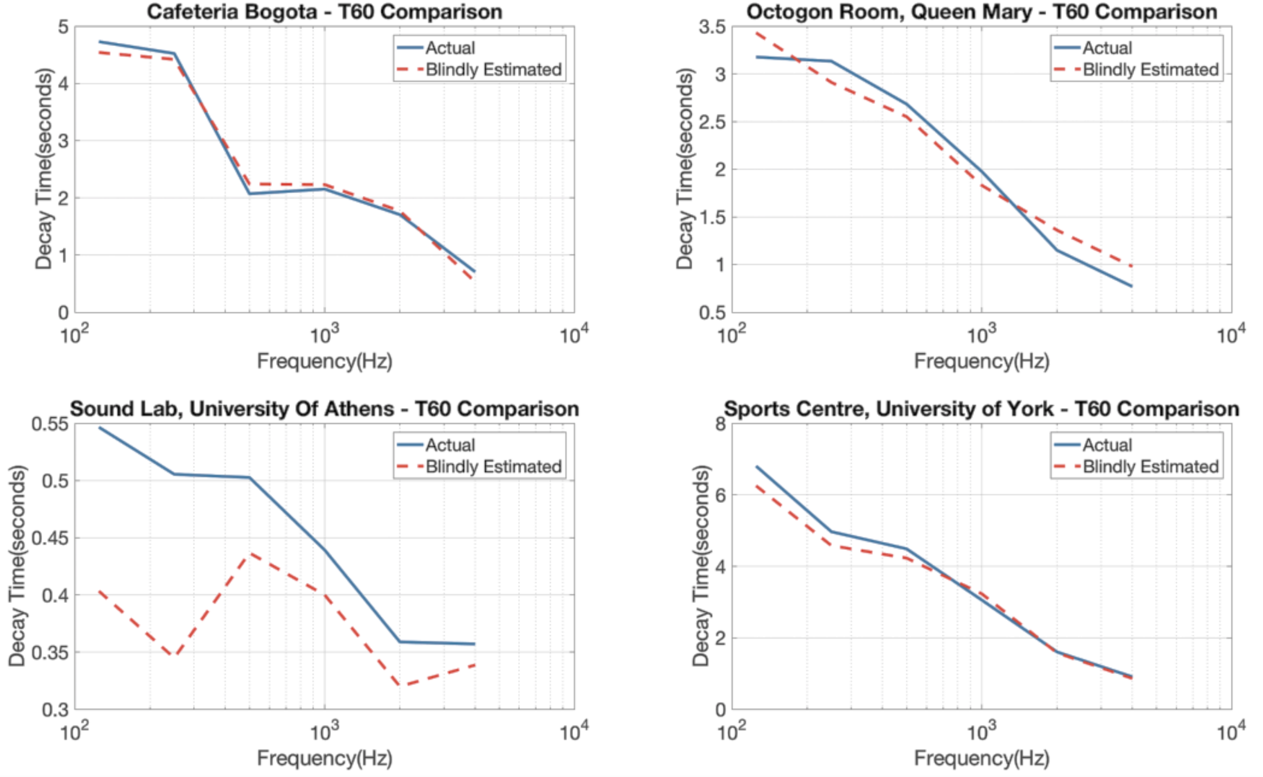


Energy Envelope Estimation

Decay Start/Stop Detection

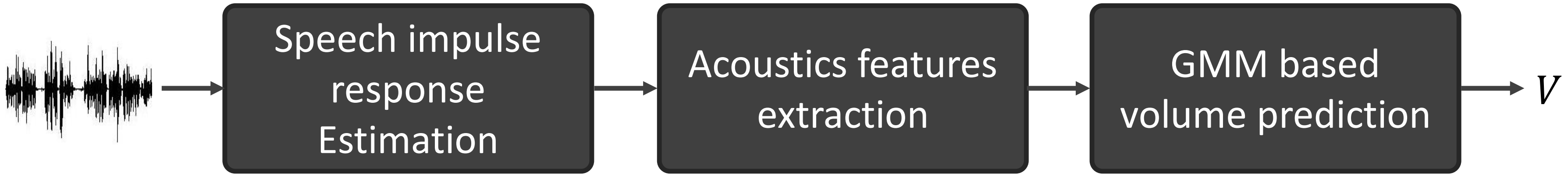
Decay Time Estimation

$Tr(f)$



[Murgai *et al*, 2017]

Room Volume Prediction



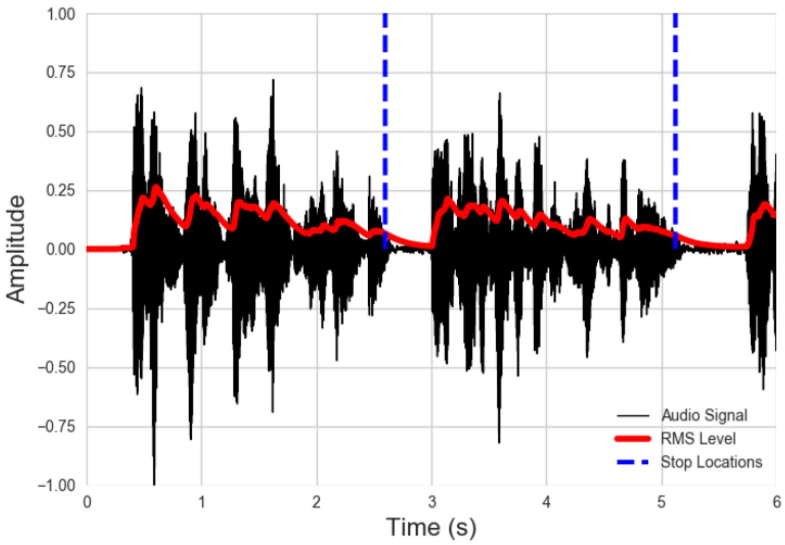
Speech impulse response Estimation

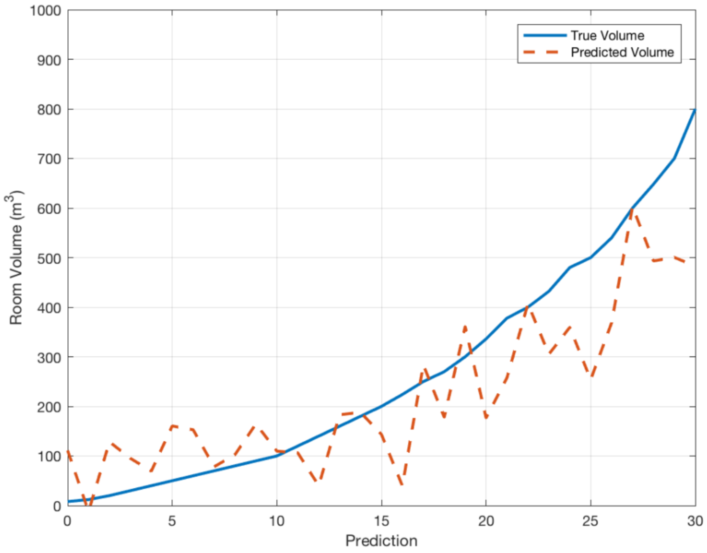
Acoustics features extraction

GMM based volume prediction

V

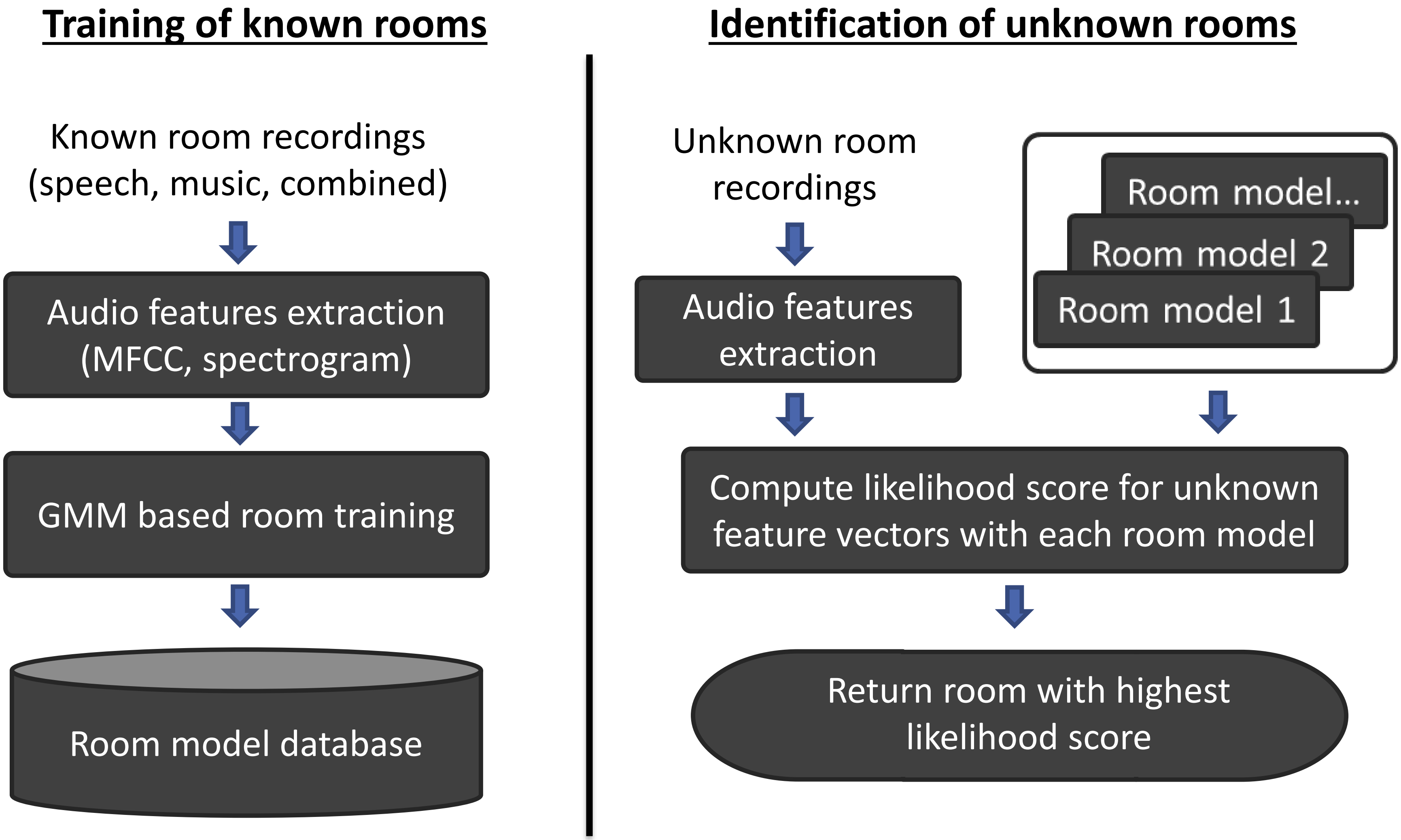
T60, C80, C50, density of early reflections, kurtosis (t), density of low frequency room modes, and kurtosis (f).





[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

Room model…

Room model 2

Room model 1

Compute likelihood score for unknown feature vectors with each room model

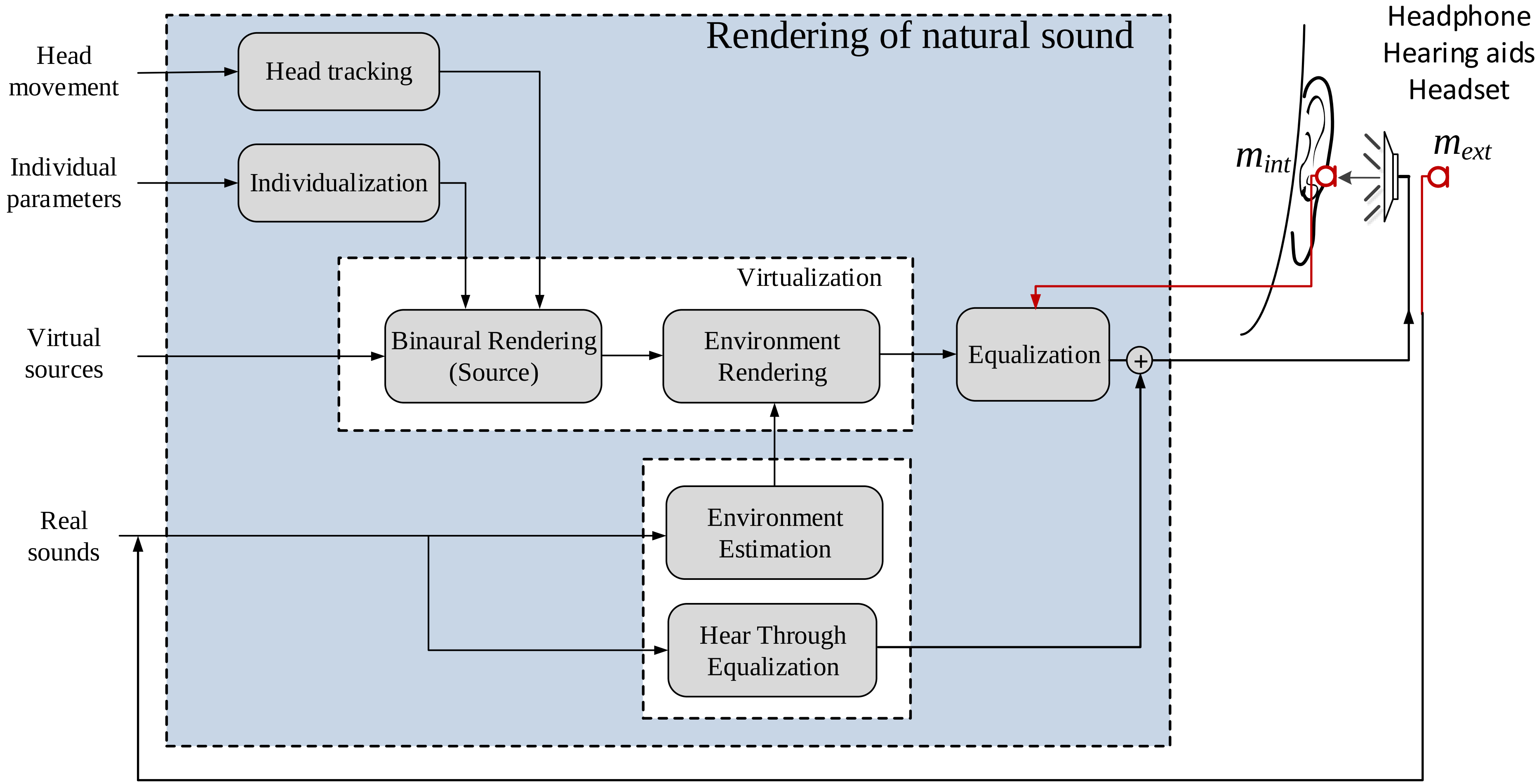
Return room with highest likelihood score

[P. Nils *et al*, 2016]

Environment Estimation and Rendering - Summary

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

C.7 Signal Processing Techniques Integration for AR/MR



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

- R. T. Azuma, “A survey of augmented reality,”Presence, vol. 6, pp. 355–385, 1997.
- T. Nilsen, S. Linton, and J. Looser, “Motivations for augmented reality gaming,” in Proc. FUSE, 2004, vol. 4, pp. 86–93.
- M. Billinghurst and H. Kato, “Collaborative augmented reality,” Commun. ACM, vol. 45, pp. 64–70, 2002.T. Miyashita, et al., “An augmented reality museum guide,” in Proc. 7th IEEE/ACM Int. Symp. Mixed Augment. Real., 2008, pp. 103–106.

Augmented/Mixed Reality Audio

- R. W. Lindeman, H. Noma, and P. G. de Barros, “Hear-through and mic-through augmented reality: Using bone conduction to display spatialized audio,” inProc. 6th IEEE and ACM Int. Symp. Mixed Augment. Real., 2007, pp. 1–4.
- A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki,and J. Hiipakkaet al., “Augmented reality audio for mobile and wearable appliances,”J. Audio Eng. Soc., vol. 52, pp. 618–639, 2004.
- M. Tikander, M. Karjalainen, and V. Riikonen, “An augmented reality audio headset,” in Proc. 11th Int. Conf. Digital Audio Effects (DAFx-08), Espoo, Finland, 2008.

Key References

Augmented/Mixed Reality Audio

- J. Rämö and V. Välimäki, “Digital augmented reality audio headset,” J. Elect. Comput. Eng., vol. 2012, p. 13, 2012, Article ID 457374.
- T. Lokki, H. Nironen, S. Vesa, L. Savioja, A. Härmä, and M. Karjalainen, “Application scenarios of wearable and mobile augmented reality audio,” in Proc. Audio Eng. Soc. Conv. 116, 2004.
- D. W. Schobben and R. M. Aarts, “Personalized multi-channel headphone sound reproduction based on active noise cancellation,”Acta Acusti. United Acust., vol. 91, pp. 440–450, 2005.
- J. Rämö, “Evaluation of an augmented reality audio headset and mixer,” Ph.D. dissertation, Helsinki Univ. of Technol., Espoo, Finalnd, 2009.
- V. Välimäki, A. Franck, J. Rämö, H. Gamper, and L. Savioja, “Assisted Listening Using a Headset,” IEEE Signal Processing Magazine, vo. 32, no. 2, pp. 92-99, Mar. 2015.
- Liski, Juho, Riitta Väänänen, Sampo Vesa, and Vesa Välimäki. "Adaptive Equalization of Acoustic Transparency in an Augmented-Reality Headset." In *Audio Engineering Society Conference: 2016 AES International Conference on Headphone Technology*. Audio Engineering Society, 2016.
- Rämö, Jussi, and Vesa Välimäki. "An allpass hear-through headset." *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 2014.

Key References

Augmented/Mixed Reality Audio

- Rämö, Jussi, Vesa Välimäki, and Miikka Tikander. "Live sound equalization and attenuation with a headset." Audio Engineering Society Conference: 51st International Conference: Loudspeakers and Headphones. Audio Engineering Society, 2013.
- R. Ranjan and W.-S. Gan, "Applying active noise control technique for augmented reality headphones," in Proc. Internoise, Melbourne, Australia, 2014.
- R. Ranjan and W.-S. Gan, "Natural Listening over Headphones in Augmented Reality Using Adaptive Filtering Techniques," IEEE/ACM Transactions On Audio, Speech, And Language Processing, Vol. 23, No. 11, November 2015
- R. Ranjan, and W.-S. Gan. "Adaptive Equalization of Natural Augmented Reality Headset Using Non-stationary Virtual Signals." *Audio Engineering Society Conference: 2016 AES International Conference on Headphone Technology*. Audio Engineering Society, 2016.
- R. Ranjan. *"3D audio reproduction: natural augmented reality headset and next generation entertainment system using wave field synthesis."* Thesis Dissertation 2016.

Key References

Augmented/Mixed Reality Audio

- F. Brinkmann and A. Lindau, “On the effect of individual headphone compensation in binaural synthesis,”Fortschritte der Akustik: Tagungsband d. 36. DAGA, pp. 1055–1056, 2010.
- M. Bouchard, S. G. Norcross, and G. A. Soulodre, “Inverse filtering design using a minimal-phase target function from regularization,” in Proc. Audio Eng. Soc. Conv. 121, 2006.
- S. T. Neely and J. B. Allen, “Invertibility of a room impulse response,” J. Acoust. Soc. Amer., vol. 66, pp. 165–169, 1979.
- S. M. Kuo and D. Morgan, Active noise control systems: algorithms and DSP implementations: John Wiley & Sons, Inc., 1995.

Environment Estimation and Rendering

- M. Vorlander, *Auralization – Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality,* Springer-Verlag, 2008.
- M. Kleiner, B.-I. Dalenback, P. Svensson,“Auralization – An Overview,” *J. Audio Eng. Soc.*, **41**(11):861–875, 1993.
- M. Vorl¨ander, D. Schr¨oder, S. Pelzer, and F. Wefers, “Virtual reality for architectural acoustics,” *J. of Building Performance Simulation*, **8**, (1):15–25, 2015.

Key References

Environment Estimation and Rendering

- V. Valimaki, J. D. Parker, L. Savioja, J. O. Smith, J. S. Abel,“Fifty years of artificial reverberation,”*IEEE TrASLP.,***20**(5):1421–1448, 2012.
- E. De Sena et al, “On the Modeling of Rectangular Geometries in Room Acoustic Simulations,” *IEEE TrASLP*, **23**(4), April 2015.
  H. Kuttru, “Room Acoustics,” SPON Press, 2000.
- B. Allen, D. A. Berkley, ”Image method for efficiently simulating small-room acoustics, *J. Acoust. Soc. Amer.,* **65**(4):943–950, 1979.
- J. Borish, “Extension of the image model to arbitrary polyhedra,” *J. Acoust. Soc. Am.*, **75**(6):1827-1836, 1984
- A. Krokstad, S. Strom and S. Sorsdal, “Calculating the acoustical room response by the use of a ray tracing technique,” J. of Sound and Vibration, **8**(1):118-125, 1968
- T. Funkhouser, N. Tsingos, I. Carlbom, G. Elko, M. Sondhi, J. E. West, G. Pingali, P. Min, A. Ngan, “A beam tracing method for interactive architectural acoustics,” *J. Acoust. Soc. Amer.,* **115**(2):740–756, 2004.
- M. Karjalainen, C. Erkut, “Digital waveguides versus finite difference structures: Equivalence and mixed modeling,” *EURASIP J. Applied Sign. Proc.,* **2004**(7):978–989, 2004.
- J.-M. Jot, A. Chaigne, “Digital delay networks for designing artificial reverberators,” *104th AES Conv.*, paper #3030, 1991.
- D. Rocchesso, J. O. Smith, ”Circulant and elliptic feedback delay networks for artificial reverberation, *IEEE TrSAP*, **5**(1):51–63, 1997.

Key References

Environment Estimation and Rendering

- M. Karjalainen, P. Huang, J. O. Smith, “Digital Waveguide Networks for Room Response Modeling and Synthesis,” *118th AES Convention*, paper #6394, 2005.
- W. G. Gardner, “Reverberation Algorithms” in “Applications of Digital Signal Processing to Audio and Acoustics,” edited by M. Kahrs and K. Brandenburg, Kluwer Academic, 1998
- E. De Sena, H. Hacihabiboglu, Z. Cvetkovic, J. O. Smith, “Efficient Synthesis of Room Acoustics via Scattering Delay Networks,” *IEEE TrASLP,* 2015.
- J.-M. Jot, ``An analysis/synthesis approach to real-time artificial reverberation,'' in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, San Francisco*, (New York), pp. II.221-II.224, IEEE Press, 1992.
- M. R. Schroeder, ``Natural-sounding artificial reverberation,'' *Journal of the Audio Engineering Society*, vol. 10, no. 3, pp. 219-223, 1962.
- Raghuvanshi, N., Narain, R., And Lin, M. C, “Efficient And Accurate Sound Propagation Using Adaptive Rectangular Decomposition.” *IEEE Transactions On Visualization And Computer Graphics 15*, 5, 789–801, 2009.
- Raghuvanshi, N., Snyder, J., Mehra, R., Lin, M., & Govindaraju, N., “Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes.” *ACM Transactions on Graphics (TOG)*, *29*(4), 68. 2010.
- Markovic, Milos, So/ren K. Olesen, and Dorte Hammershoi. "Three-dimensional point-cloud room model for room acoustics simulations." Proceedings of Meetings on Acoustics ICA2013. Vol. 19. No. 1. ASA, 2013.

Key References

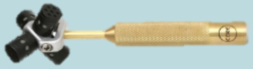
Environment Estimation and Rendering

- Aspöck, Lukas, and Michael Vorländer. "Room geometry acquisition and processing methods for geometrical acoustics simulation models."Proceedings of 9th Iberian Congress on Acoustics: EuroRegio 2016
- Filliat, David, et al. "RGBD object recognition and visual texture classification for indoor semantic mapping." *Technologies for Practical Robot Applications (TePRA), IEEE International Conference on*. IEEE, 2012.
- Maree, Raphael, et al. "Random subwindows for robust image classification." *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE, 2005.
- Gamper, Hannes, and Tapio Lokki. "Instant BRIR acquisition for auditory events in audio augmented reality using finger snaps." Proceedings of the International Workshop on the Principles and Applications of Spatial Hearing (IWPASH). 2009.
- Jot, Jean-Marc, and Keun Sup Lee. "Augmented Reality Headphone Environment Rendering." AES International Conference on Audio for Virtual and Augmented Reality. Audio Engineering Society, 2016.
- Murgai, Prateek, Mark Rau, and Jean-Marc Jot. "Blind Estimation of the Reverberation Fingerprint of Unknown Acoustic Environments." Audio Engineering Society Convention 143. Audio Engineering Society, 2017.
- Peters, Nils, Howard Lei, and Gerald Friedland. "Room identification using acoustic features in a recording." U.S. Patent No. 9,449,613. 20 Sep. 2016.

Module D
Summary and Future Trends

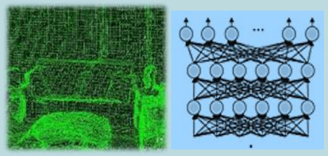
1. Summary of key Techniques
2. Spatial Audio Tools
3. Emerging (potential) Applications of VR/AR Audio
4. Challenges and Future Research Trends

D.1 Spatial Audio Technologies for Immersive VR, AR/MR



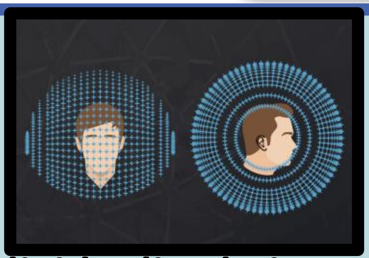
Spatial Audio Formats

- Object, Ambisonics
- Parametric processing



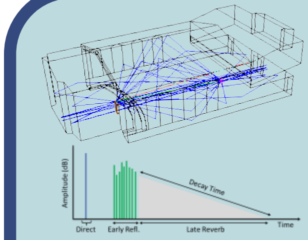
Environment Estimation

- Depth camera
- Reverberation fingerprint
- Machine learning



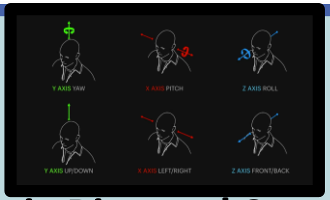
Individualized Binaural Rendering

- Individualized HRTFs
- Equalization



Environment Rendering

- Wave based
- Geometrical based
- Perceptual based



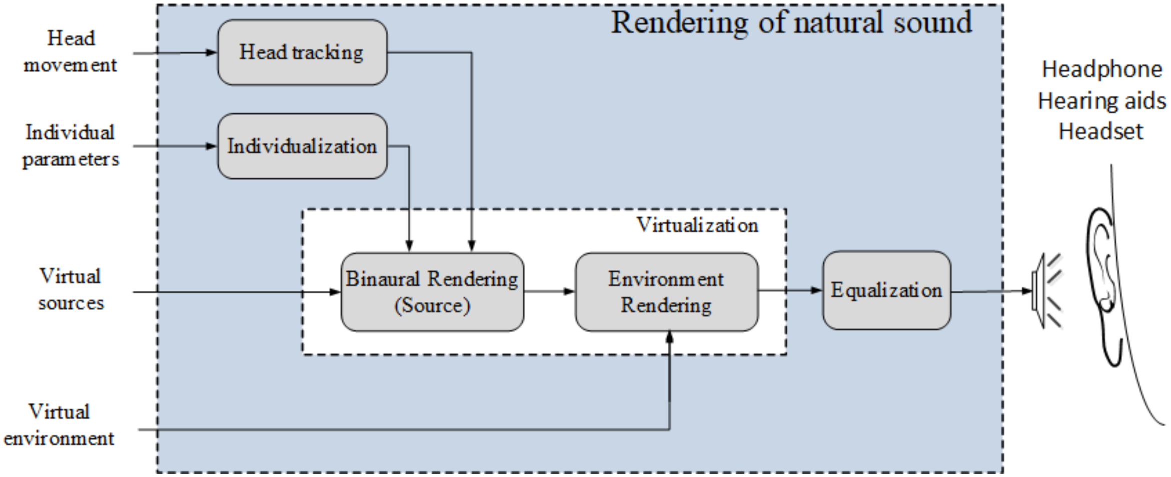
Dynamic Binaural Synthesis

- Head tracking
- Position tracking

Virtual & Physical Sound Fusion

- Adaptive equalization
- Hear-through processing

Summary: Different Listening Modes for VR and AR/MR



Head movement

Individual parameters

Virtual sources

Virtual environment

Rendering of natural sound

Head tracking

Individualization

Virtualization

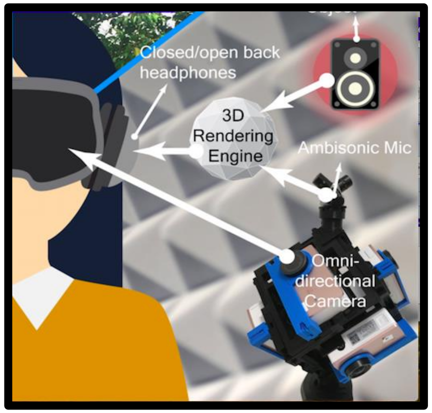
Binaural Rendering (Source)

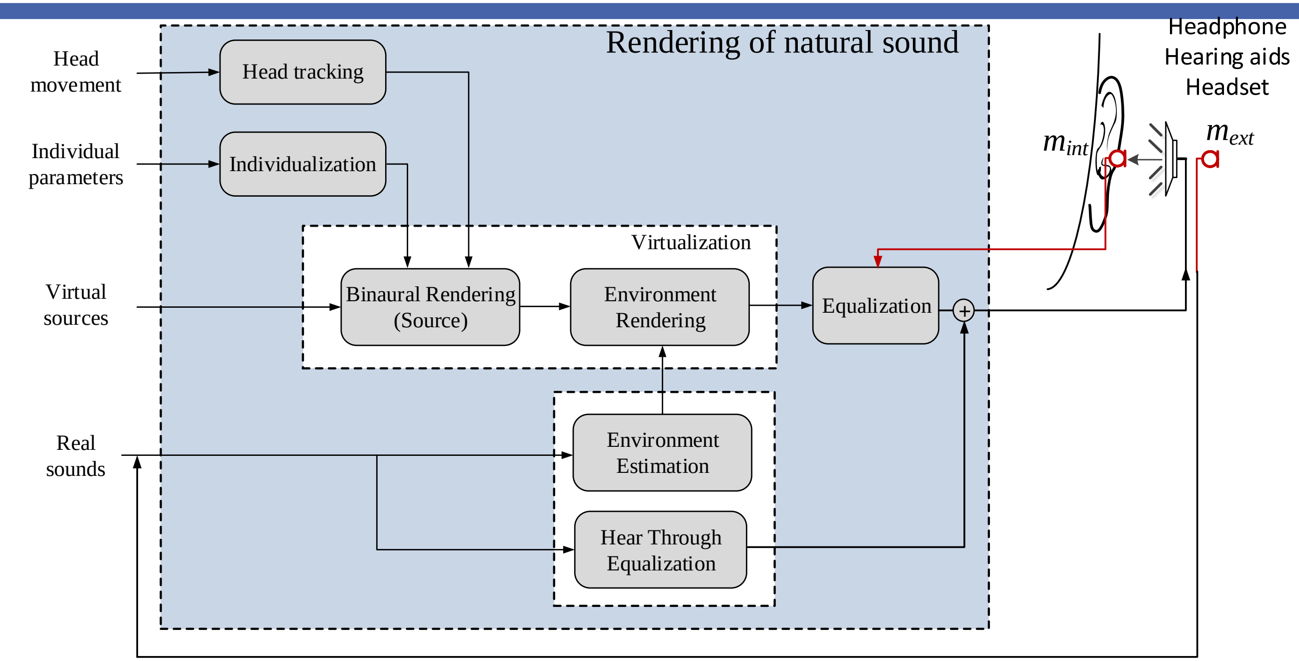
Environment Rendering

Equalization

Headphone
Hearing aids
Headset

VR





Head movement

Individual parameters

Virtual sources

Real sounds

Rendering of natural sound

Head tracking

Individualization

Virtualization

Binaural Rendering (Source)

Environment Rendering

Equalization

Environment Estimation

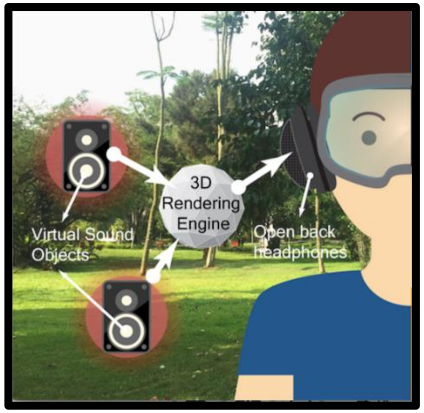
Hear Through Equalization

Headphone
Hearing aids
Headset

m_{int}

m_{ext}

AR/MR



D.2 Binaural rendering tools

S. No.	Tool Name	Common Features						Unique features
		Input format			Inclusion of custom HRTF allowed	Auralization		
		HOA	FOA	Obj.		Artificial reverberation	Physics based simulation	
1.	Steam Audio [1]	X	X		X	X	X	Audio occlusion effects, real time sound propagation
2.	Resonance Audio [2]		X			X	X	Near field effects, occlusions, sound source directivity
3.	Dyasonics [3]		X			X		Real time motion tracked binaural playback while editing
4.	Oculus [4]		X			X	X	Volumetric sources, near field rendering
5.	Gaudio SOL [5]		X	X		X		Supports FOA and unlimited number of virtual speakers using proprietary GA5 format
6.	Super-powered [6]		X	X		X		low latency VR and mobile solution
7.	Nx Audio [7]	X	X		X	X		supports mix to other formats such as 5.1, 7.0 allows tuning of HRTF using head measurements, EQ calibration for headphones supported
8.	DearVR [8]			X		X		Allows 3D sound object design inside VR
9.	Real Space 3D [9]		X		X	X		allows custom HRTF tuning using anthropometric measurements and user selected HRTF playback in real time
10.	Dirac VR [10]		X		X	X		Headphone Optimization tech, Dynamic HRTF incorporating relative head-torso movement

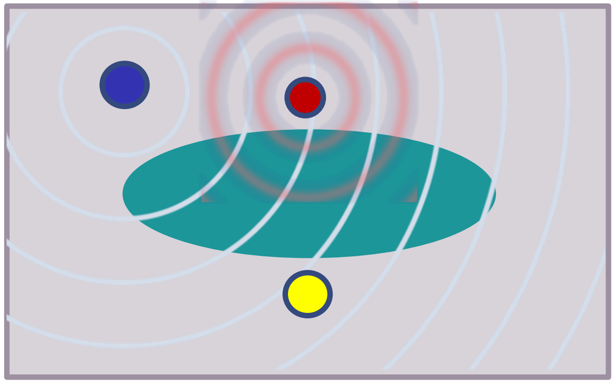
Spatial Audio Tool References

Companies	Websites
STEAM® AUDIO	[1] SteamAudio - https://valvesoftware.github.io/steam-audio/
Resonance Audio by Google	[2] Resonance audio- https://developers.google.com/resonance-audio/develop/unity/developer-guide
DYSONICS rondo360	[3] Dyasonics Rondo360- https://dysonics.com/rondo360/
Oculus Audio SDK	[4] Oculus- https://developer.oculus.com/downloads/package/oculus-audio-sdk-plugins/1.1.0/
GAUDIO	[5] Gaudio - https://www.gaudiolab.com/resources/download/works
superpowered Cross-Platform Audio SDK for Android, iOS and Wearables	[6] Superpowered- http://superpowered.com/3d-audio-spatializer-ambisonics-vr-audio
WAVES nx 3D audio for any headphones	[7] Nx Audio- https://www.waves.com/nx
dearVR 3D audio reality engine	[8] DearVR http://dearvr.com/
RealSpace3D	[9] Realspace 3D- https://realspace3daudio.com/
Dirac	[10] Dirac VR https://www.dirac.com/dirac-vr

D.3 Emerging Applications of Spatial Audio for VR, AR/MR

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



Airport

- Usually large hall
- Significant ambient sound

Bedroom

- Usually small-medium sized room
- Quiet environment

Meeting Room

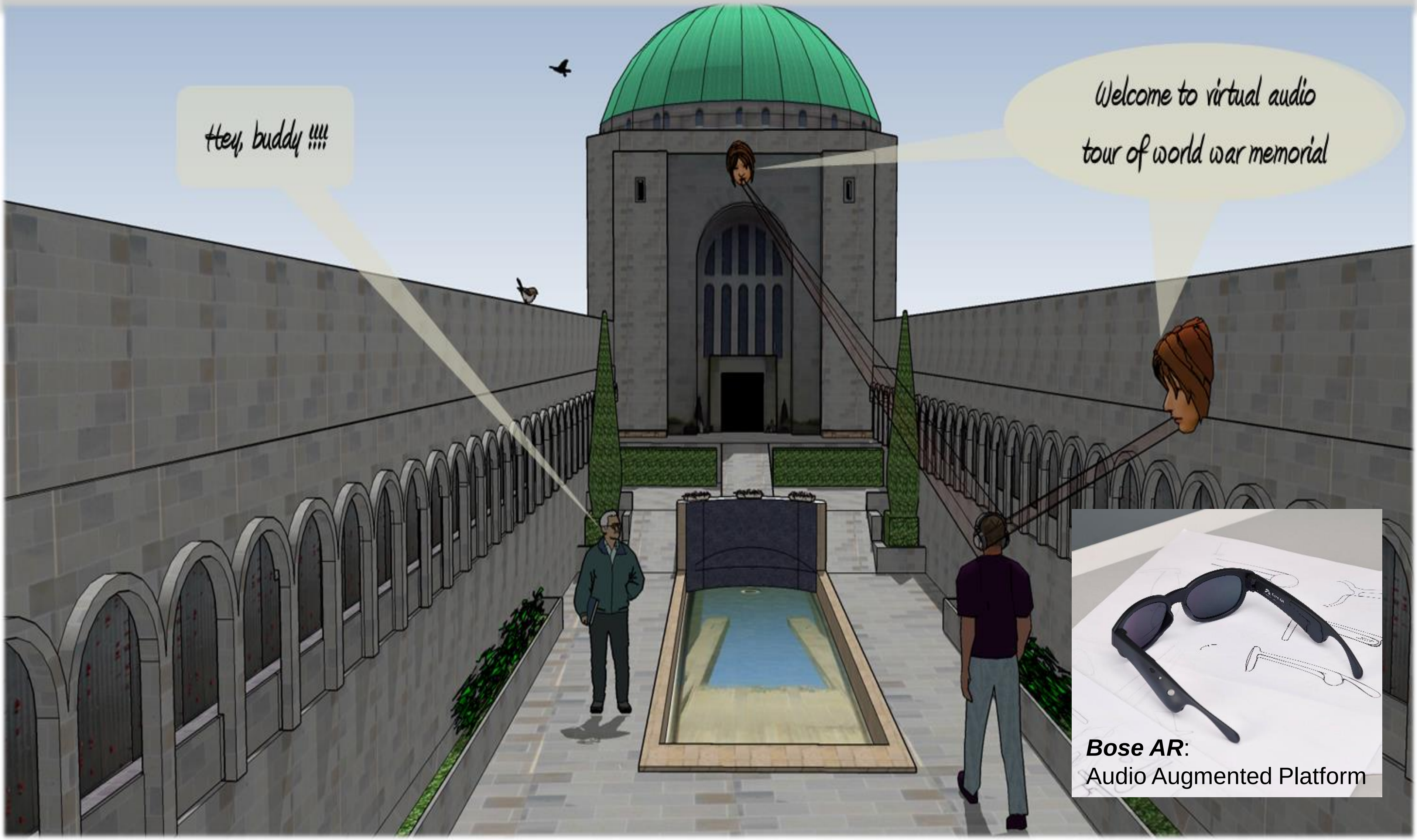
- Enclosed room with reverberant characteristics

Required Audio Technology

- Environment estimation
- Adaptive Filtering
- ANC mode may be required if loud external noise
- Reverberation Rendering
- Dynamic Binaural audio

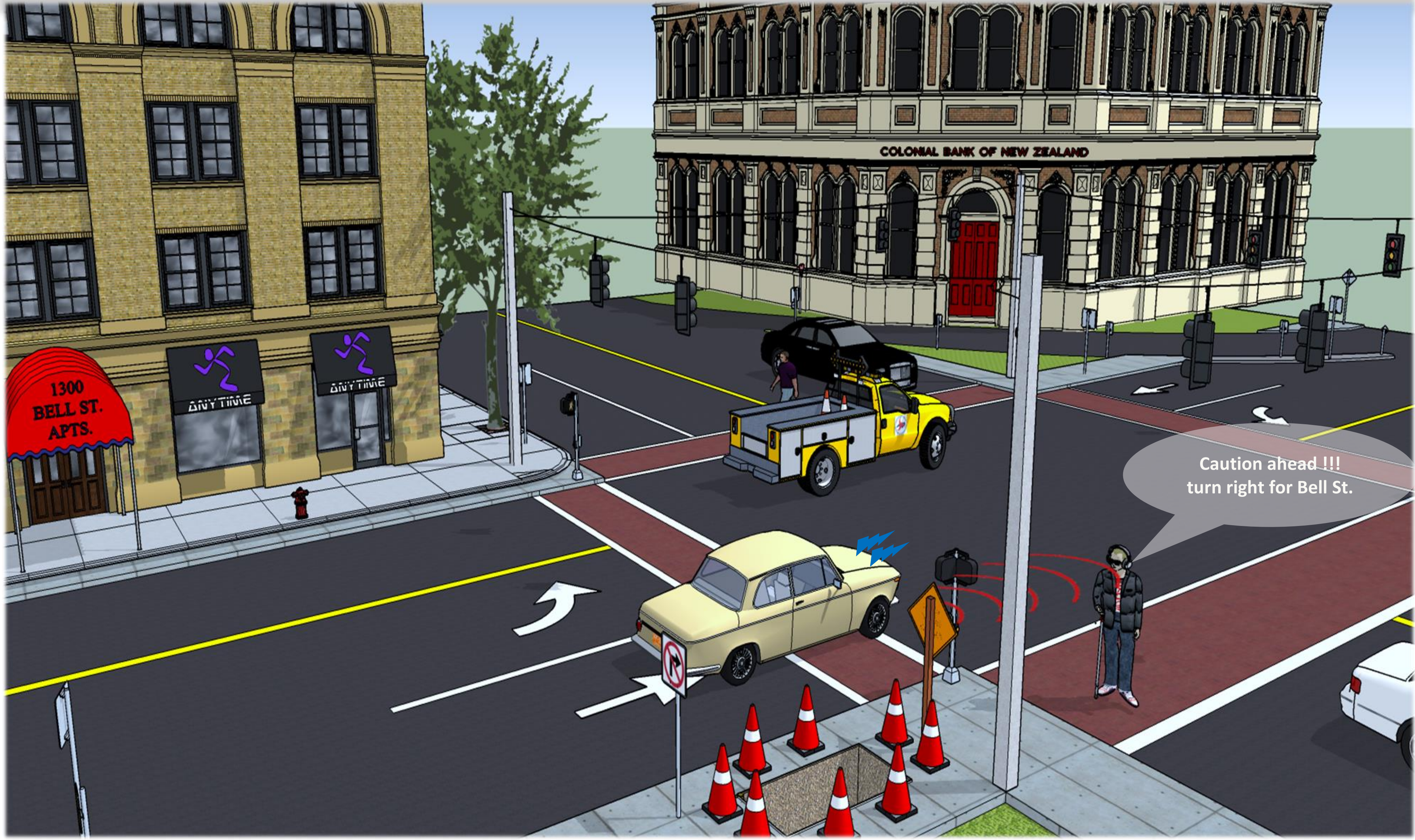
Illustration by Santi, image credit: Freepik.com

2) Augmented Audio Tour

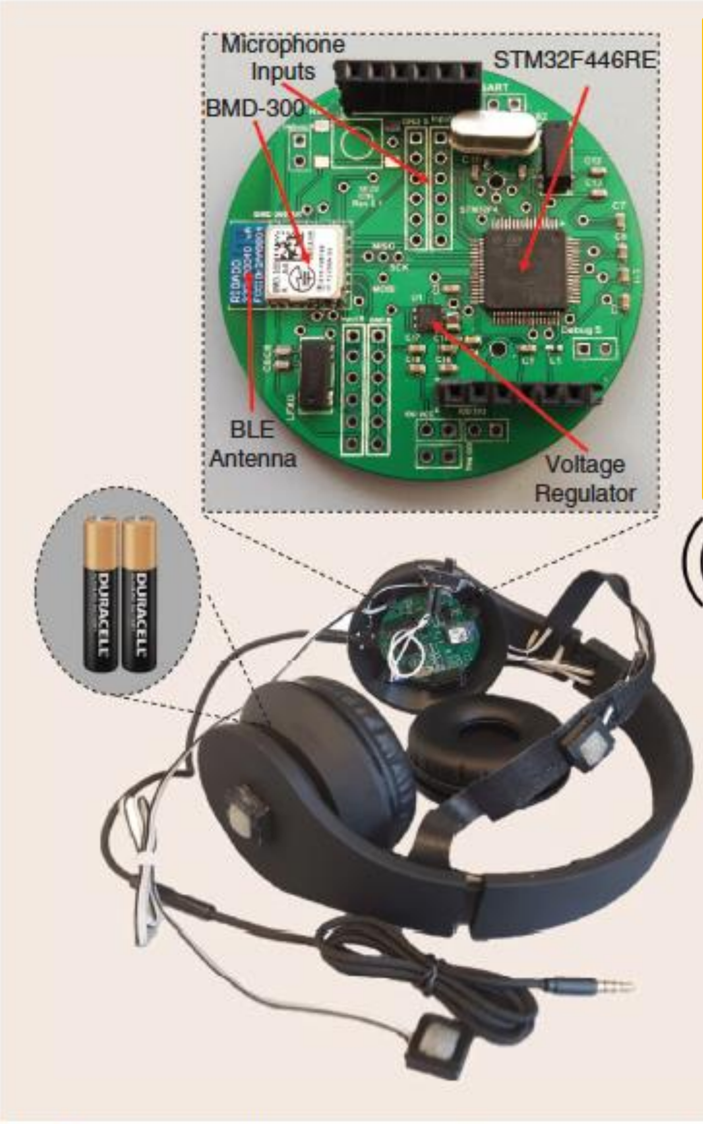


Bose AR:
Audio Augmented Platform

3) Augmented Assistive listening for Visually Impaired


1300 BELL ST. APTS.
ANYTIME
ANYTIME
COLONIAL BANK OF NEW ZEALAND
Caution ahead !!! turn right for Bell St.

Safety Headphones



Assistive Listening is also needed for the normal person!

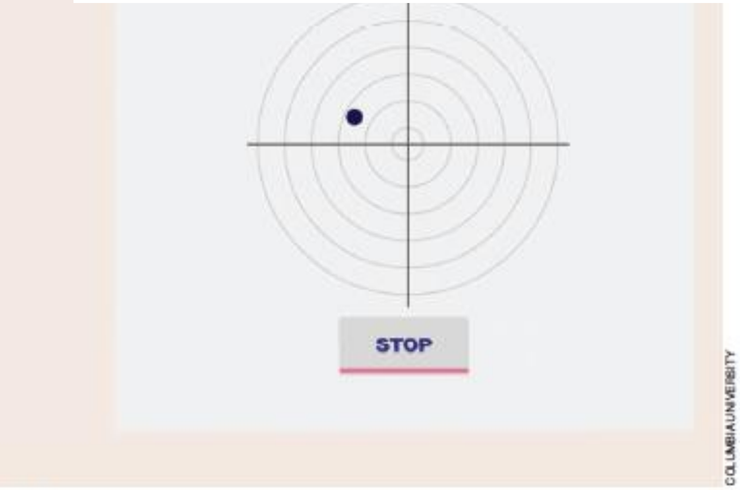


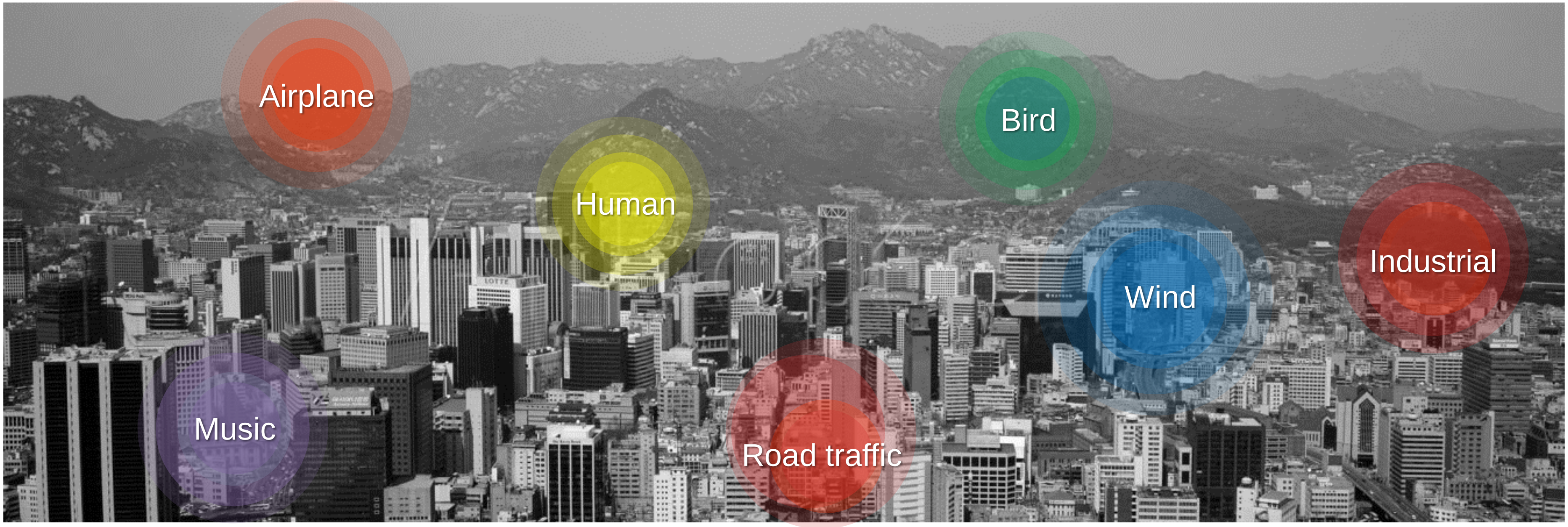
FIGURE 4. The prototype intelligent pedestrian safety headset and connected smartphone.

Fred Jiang
Columbia University, USA

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

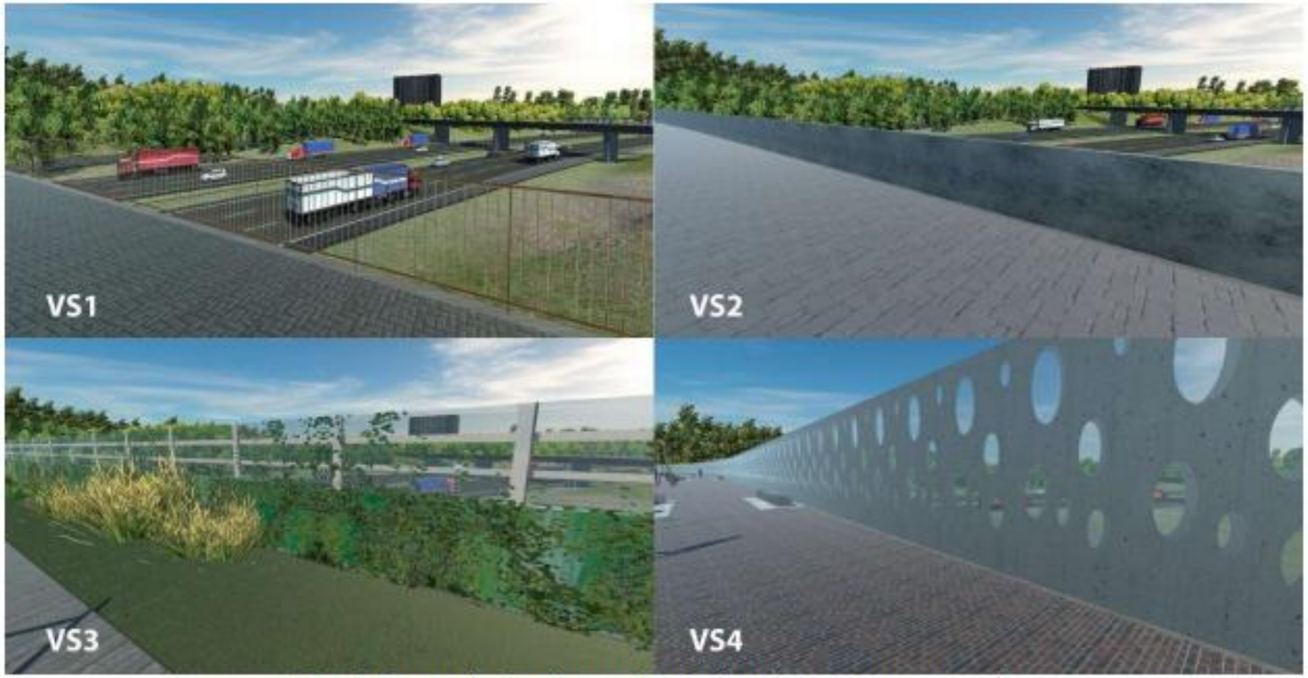


Fig. 5. Four noise barriers and sound environments proposed.
VS1: common rail, no sound barrier. VS2: concrete opaque, 1.2 m barrier. VS3: concrete vegetated with upper part in glass, 2 m barrier. VS4: concrete opaque with oval windows, 3 m barrier.

[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

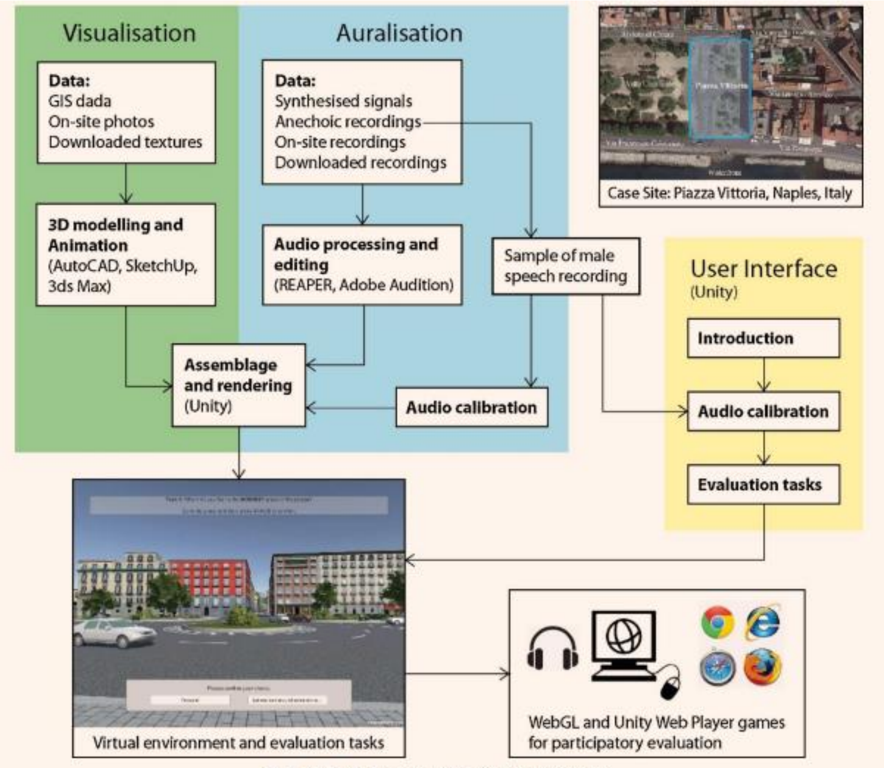
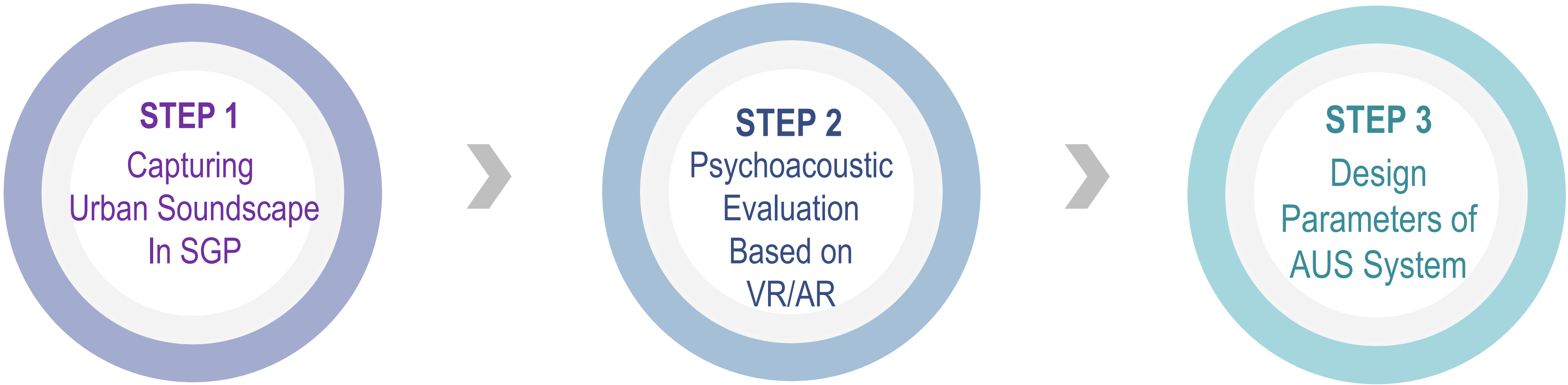


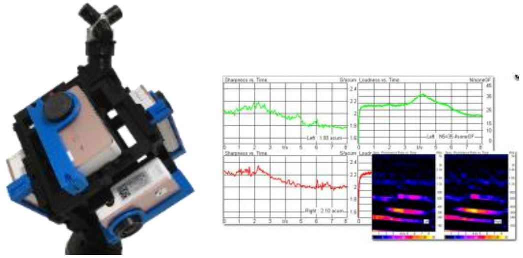
Fig. 1. Workflow of the development of the demonstrator tool.

[Jiang, 2018]

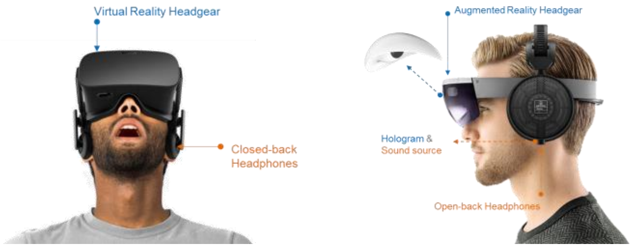
Overview of Our Augmented Urban Soundscape

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


STEP 1
Capturing Urban Soundscape In SGP
STEP 2
Psychoacoustic Evaluation Based on VR/AR
STEP 3
Design Parameters of AUS System



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



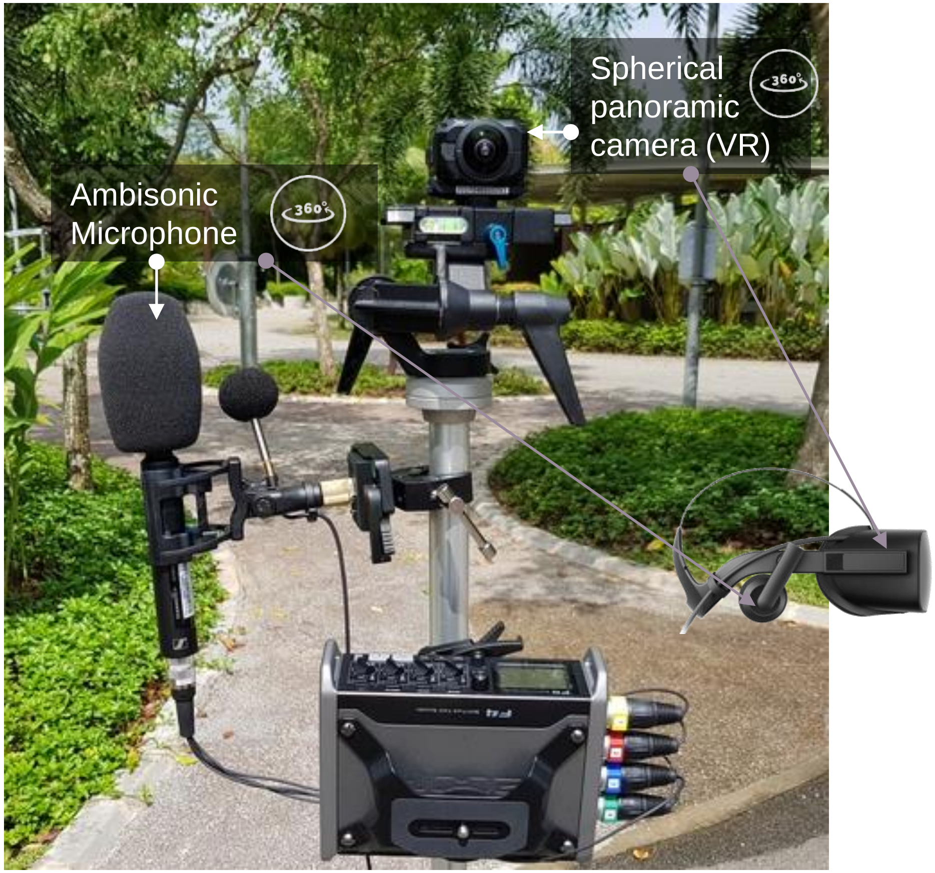
- Psychoacoustic evaluation using VR and AR
- Developing optimal masking algorithm

Outdoor Soundscapes
Soundscaping Units
Traffic Noise

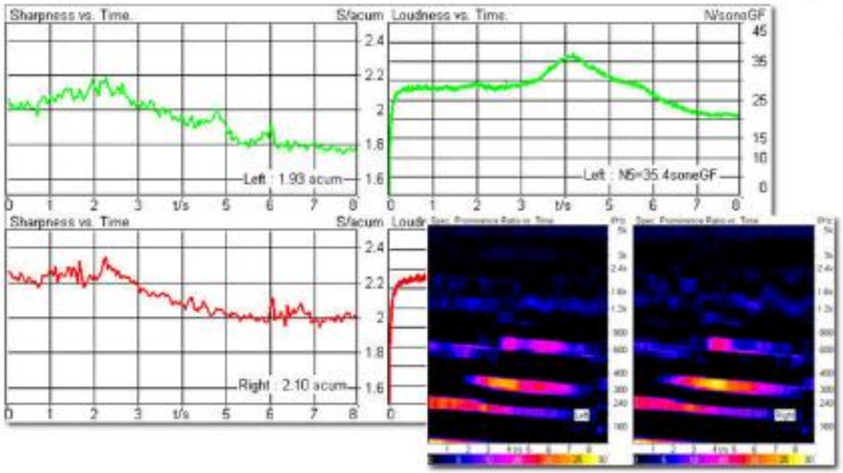
- Design parameters for AUS algorithm
- Aid in design of AUS in Phase 2

Capturing of Urban Soundscape

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



- **Psychoacoustic indicators**

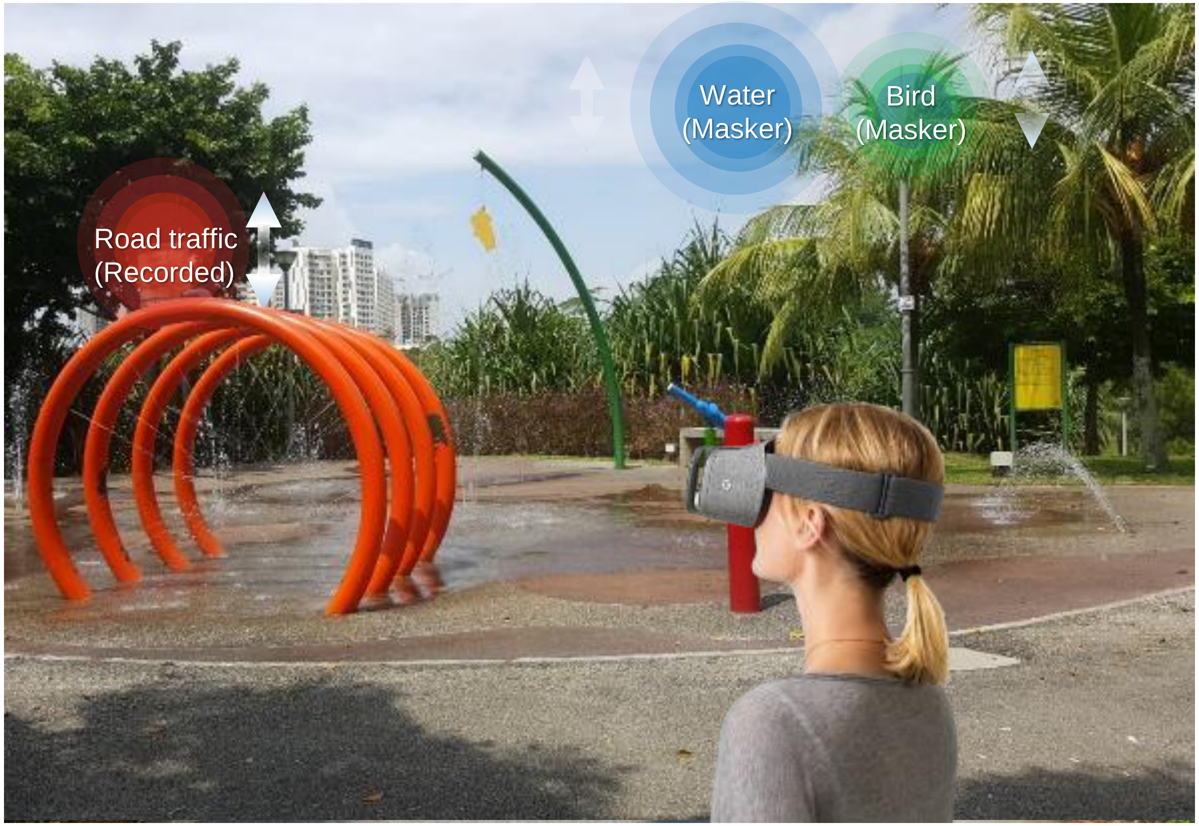


- Loudness
- Sharpness
- Roughness
- Fluctuation strength
- Tonality
- …

Psychoacoustic Evaluation using VR

- **VR scenario**

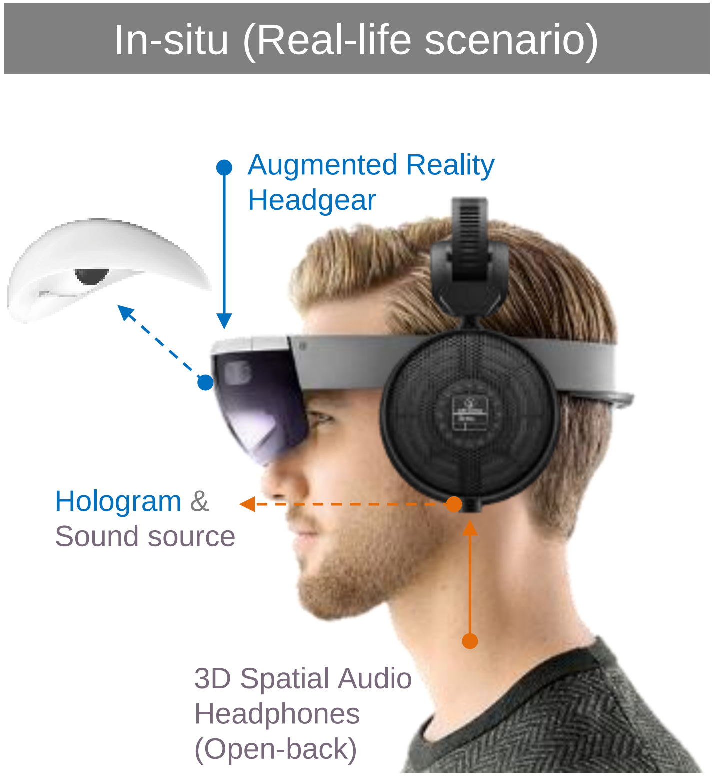
Laboratory condition (Controlled)

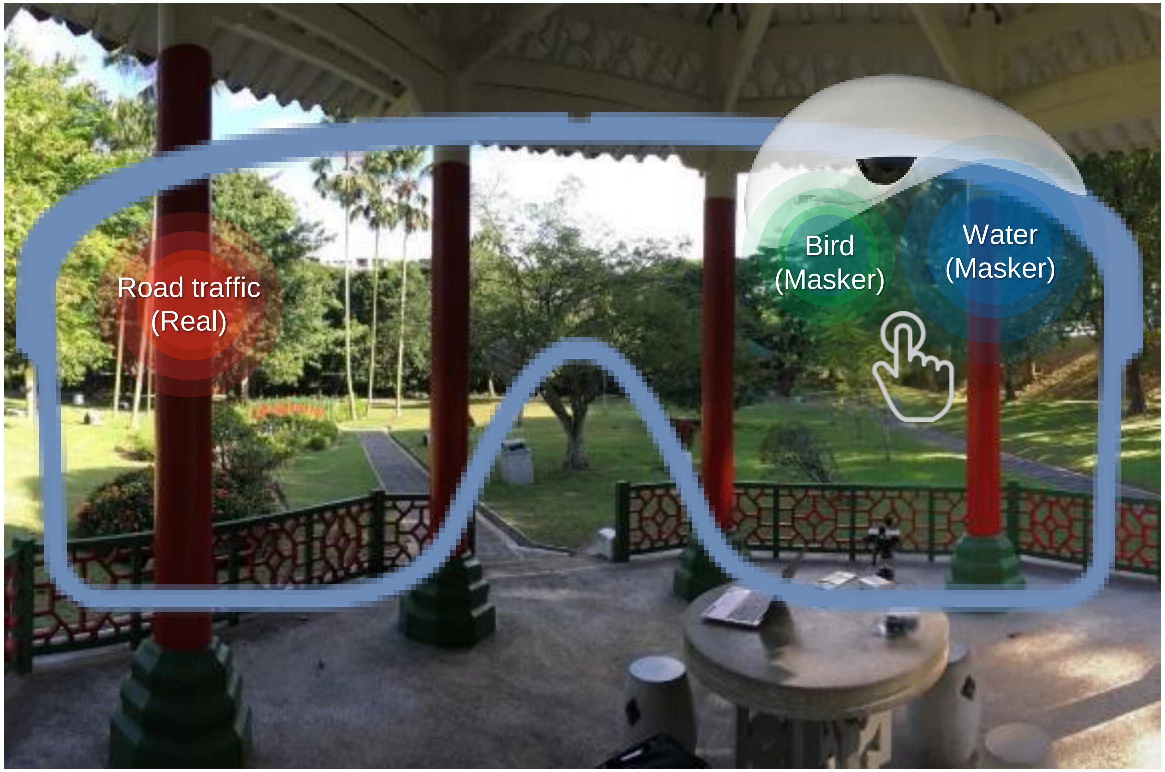


Video Demo: Spatial Audio for VR

Psychoacoustic Evaluation using AR

- **AR scenario**


In-situ (Real-life scenario)
Augmented Reality Headgear
Hologram & Sound source
3D Spatial Audio Headphones (Open-back)



Video Demo: Spatial Audio for AR/MR

Augmented Reality
(Soundscape with static & movable
masker in Yunnan Garden)

Summary of Audio Techniques for Soundscapes

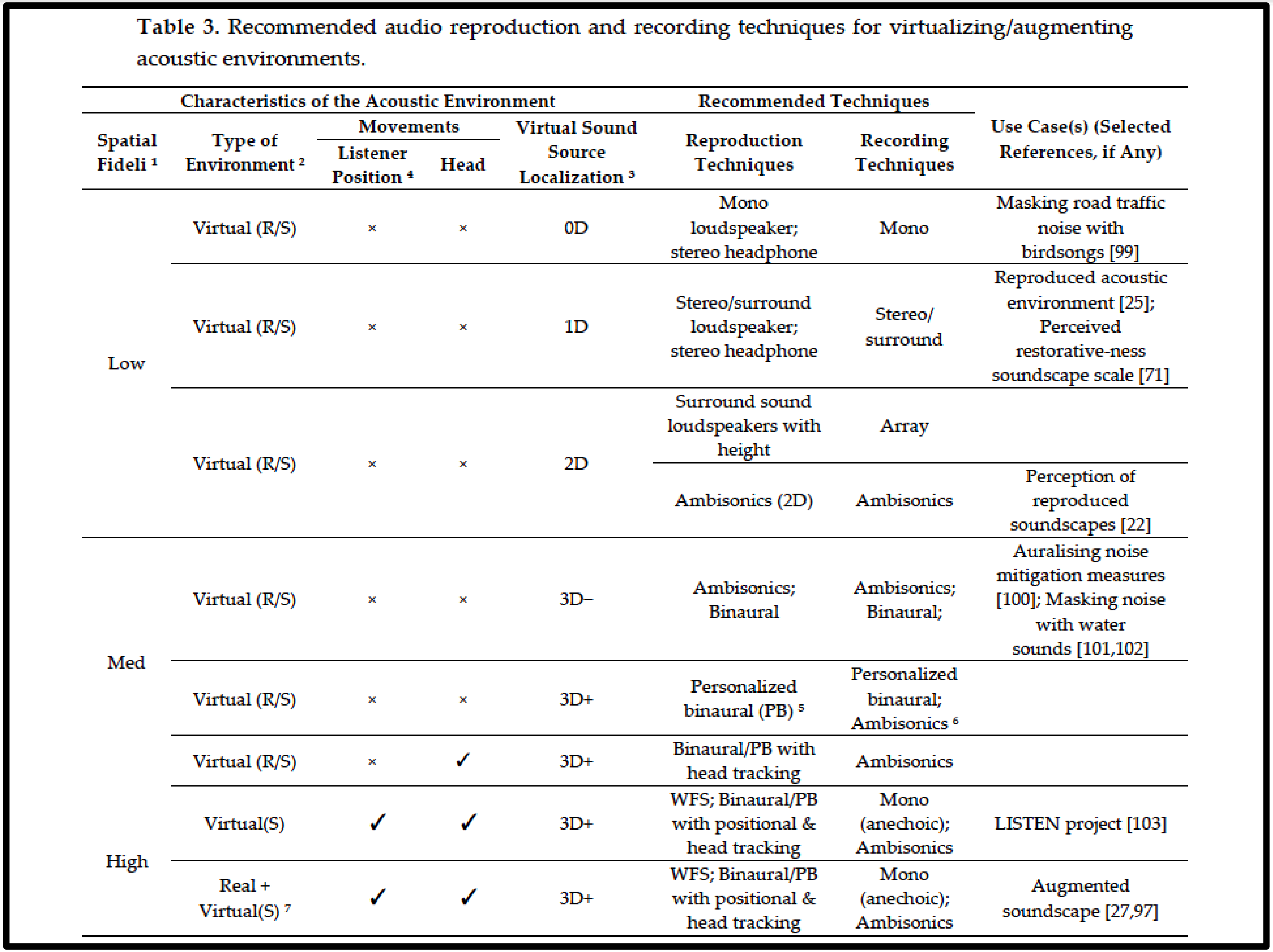


Table 3. Recommended audio reproduction and recording techniques for virtualizing/augmenting acoustic environments.

Characteristics of the Acoustic Environment					Recommended Techniques		
Spatial Fideli [1]	Type of Environment [2]	Movements		Virtual Sound Source Localization [3]	Reproduction Techniques	Recording Techniques	Use Case(s) (Selected References, if Any)
		Listener Position [4]	Head				
Low	Virtual (R/S)	×	×	0D	Mono loudspeaker; stereo headphone	Mono	Masking road traffic noise with birdsongs [99]
	Virtual (R/S)	×	×	1D	Stereo/surround loudspeaker; stereo headphone	Stereo/ surround	Reproduced acoustic environment [25]; Perceived restorative-ness soundscape scale [71]
	Virtual (R/S)	×	×	2D	Surround sound loudspeakers with height	Array	
					Ambisonics (2D)	Ambisonics	Perception of reproduced soundscapes [22]
Med	Virtual (R/S)	×	×	3D−	Ambisonics; Binaural	Ambisonics; Binaural;	Auralising noise mitigation measures [100]; Masking noise with water sounds [101,102]
	Virtual (R/S)	×	×	3D+	Personalized binaural (PB) [5]	Personalized binaural; Ambisonics [6]	
	Virtual (R/S)	×	✓	3D+	Binaural/PB with head tracking	Ambisonics	
High	Virtual(S)	✓	✓	3D+	WFS; Binaural/PB with positional & head tracking	Mono (anechoic); Ambisonics	LISTEN project [103]
	Real + Virtual(S) [7]	✓	✓	3D+	WFS; Binaural/PB with positional & head tracking	Mono (anechoic); Ambisonics	Augmented soundscape [27,97]

[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

Holy Grail in Spatial Sound

*“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

- Sanchez, G.M.E., Van Renterghem, T., Sun, K., De Coensel, B. and Botteldooren, D., 2017. Using Virtual Reality for assessing the role of noise in the audio-visual design of an urban public space. *Landscape and Urban Planning*, *167*, pp.98-107.
- Hong, J.Y., He, J., Lam, B., Gupta, R. and Gan, W.S., 2017. Spatial audio for soundscape design. Recording and reproduction. *Applied Sciences*, *7*(6), p.627.
- Jiang, L., Masullo, M., Maffei, L., Meng, F. and Vorländer, M., 2018. A demonstrator tool of web-based virtual reality for participatory evaluation of urban sound environment. *Landscape and Urban Planning*, *170*, pp.276-282.

Selected Authors’ Publications on Spatial Audio

- W.S. Gan and J.W Choi, “A Special Issue on Spatial Audio,” Applied Science, 2017, www.mdpi.com/journal/applsci/special_issues/spatial_audio
- J. Y. Hong, J. He, B. Lam, R. Gupta, and W. S. Gan, “Spatial Audio for Soundscape Design: Recording and Reproduction,” Applied Science, vol. 7, no. 6: 627, Jun 2017.
- K. Sunder, W.S. Gan, “Individualization of head-related transfer functions in the median plane using frontal projection headphones,” Journal of Audio Engineering Society, vol 64, no. 12, pp. 1026–1041, Dec 2016.
- R. Ranjan, W.S. Gan, “Natural Listening over Headphones in Augmented Reality using Adaptive Filtering Techniques,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.11, pp 1988-2002, Nov 2015.
- K. Sunder, W.S. Gan, E.L. Tan, “Modeling distance-dependent individual head-related transfer functions in the horizontal plane using frontal projection headphones,” Journal of Acoustical Society of American, Vol 138, No. 1, pp 150-171, July 2015,
- J. He, W.S. Gan, E.L. Tan, "Primary-Ambient Extraction Using Ambient Spectrum Estimation for Immersive Spatial Audio Reproduction," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.23, no.9, pp.1431,1444, Sept. 2015.
- J. He, W.S. Gan, E.L. Tan, "Time-Shifting Based Primary-Ambient Extraction for Spatial Audio Reproduction," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.10, pp.1576,1588, Oct. 2015.
- K. Sunder, J. He, E.L. Tan, W.S. Gan, “Natural Sound Rendering for Headphones: Integration of Signal Processing Techniques,” IEEE Signal Processing Magazine, Vol 32, No. 2, pp 100-113, March 2015.

Selected Authors’ Publications on Spatial Audio

- J. He, W.S. Gan,” Primary-Ambient Extraction Using Ambient Phase Estimation with a Sparsity Constraint,” IEEE Signal Processing Letter, Vol 22, No. 8, August 2015.
- J. He, E.L. Tan, W.S. Gan, “Linear Estimation Based Primary-Ambient Extraction for Stereo Audio Signals,” IEEE/ACM Transaction on Audio, Speech, and Language Processing, Vol 22, No.2, pp 505-517, Feb 2014.
- K. Sunder, E.L. Tan, W.S. Gan, “Individualization of Binaural Synthesis Using Frontal Projection Headphones,” Journal of Audio Engineering Society, Vol 61, No. 12, pp989-1000, 2013.
- J. He, R. Ranjan, W.S. Gan, “Fast Continuous HRTF Acquisition with Unconstrained Movements of Human Subjects,” International Conference on Acoustic, Speech, and Signal Processing, 20-25 March 2016, Shanghai, China.
- J. He, “Spatial audio reproduction with primary ambient extraction,” SpringerBriefs in Signal Processing. DOI: 10.1007/978-981-10-1551-9. Springer, Singapore, 2017.
- J. He, W. S. Gan, and E. L. Tan, “On the preprocessing and postprocessing of HRTF individualization based on sparse representation of anthropometry features,” in Proc. ICASSP, Brisbane, Australia, Apr. 2015, pp. 639-643.
- R. Ranjan, J. He, and W. S. Gan, “Fast continuous acquisition of HRTF in 2D for human subjects with unconstrained random head movements,” in Proc. AES Headphone conference, Aalborg, Denmark, Aug. 2016.
- R. Ranjan, and W.-S. Gan, "Adaptive Equalization of Natural Augmented Reality Headset Using Non-stationary Virtual Signals," in Audio Engineering Society International Conference on Headphone Technology, 2016.

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)

http://eeewebc.ntu.edu.sg/dsplab/ewsgan/ICASSP2018.html

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

2nd version of real-time iHRTF ACQ-Spatial Audio Rendering Unit