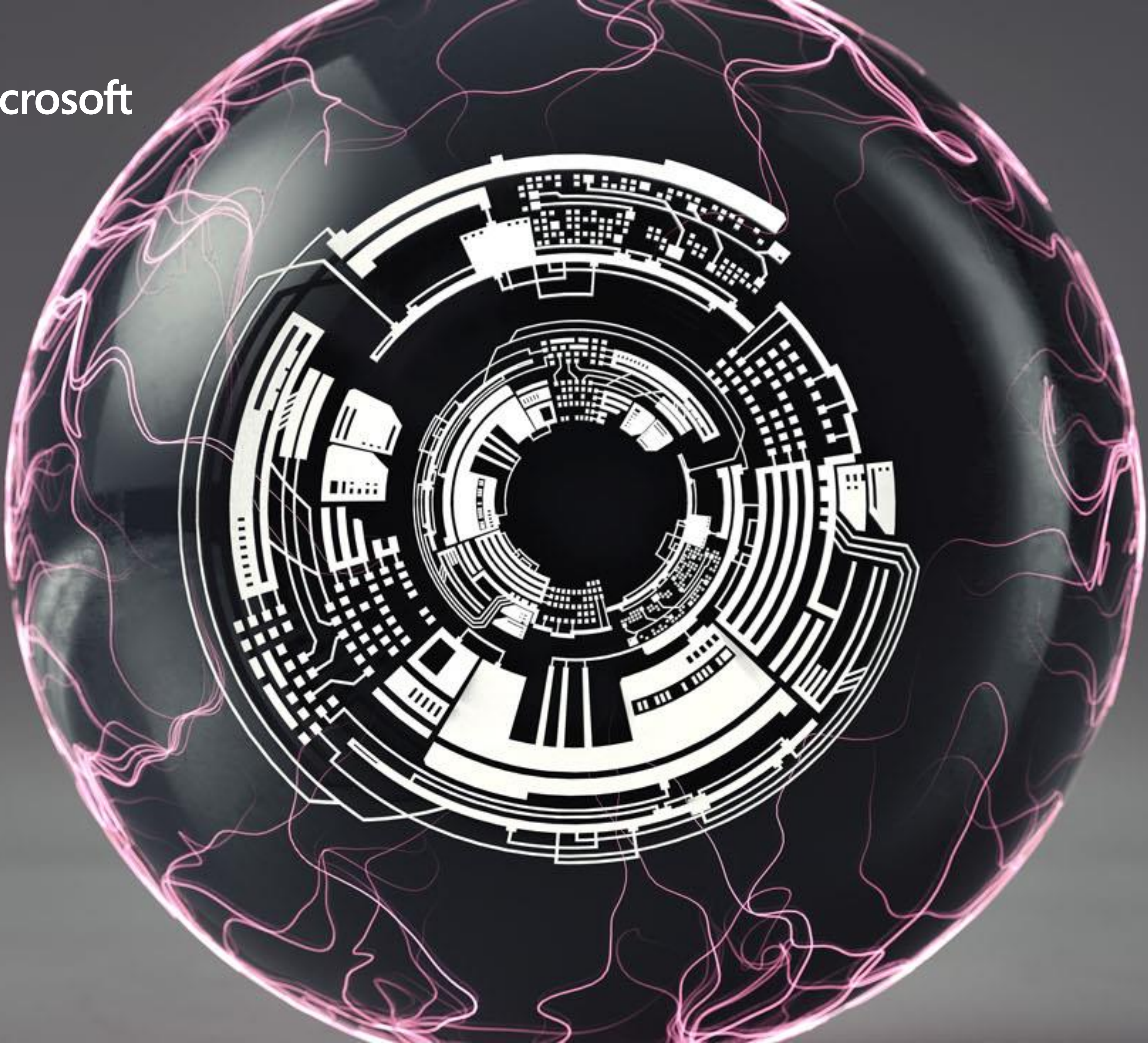
Microsoft

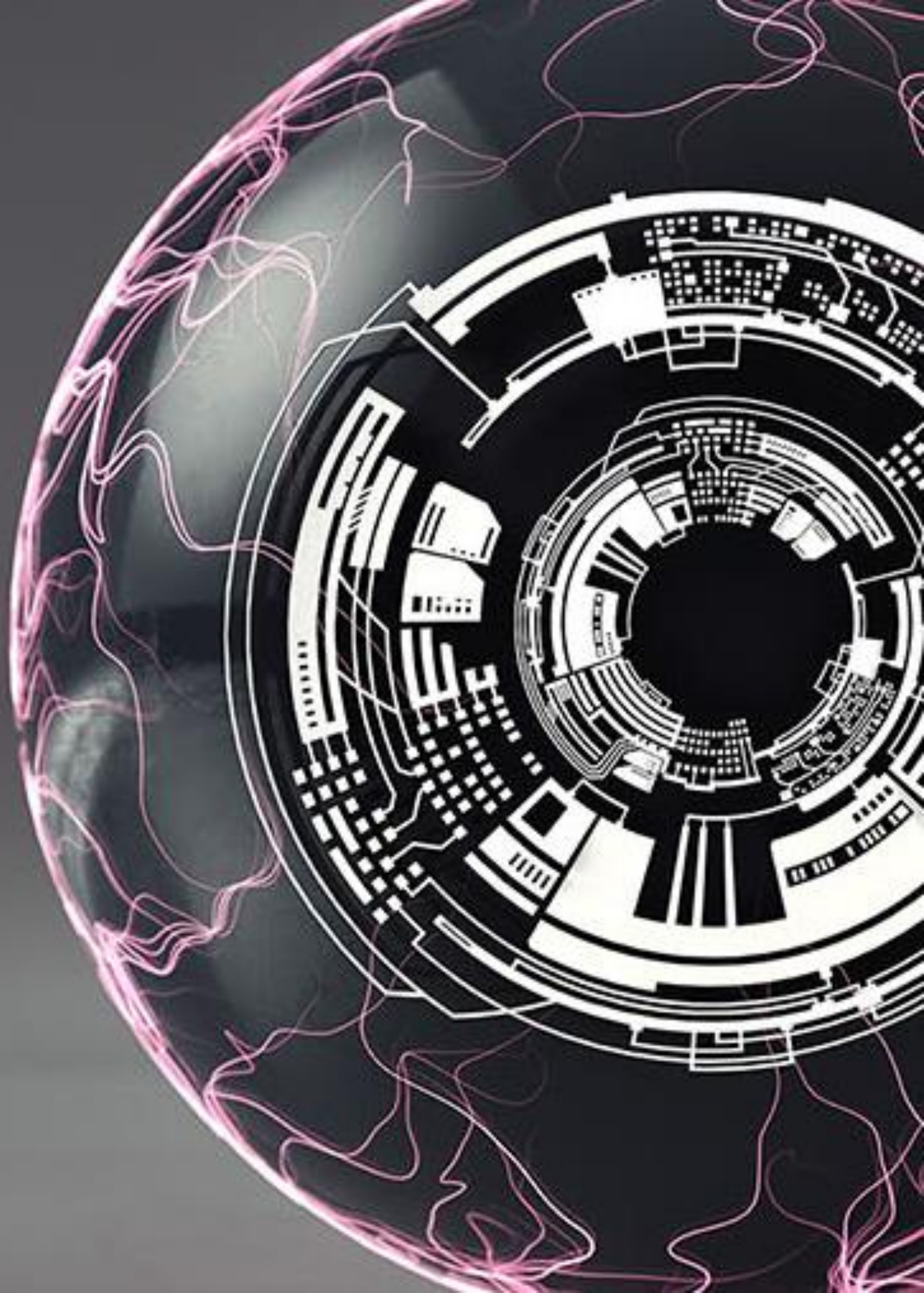


Microsoft Research Faculty Summit 2016

Microsoft

Spatial Audio research at Microsoft

Hannes Gamper
MSR Labs



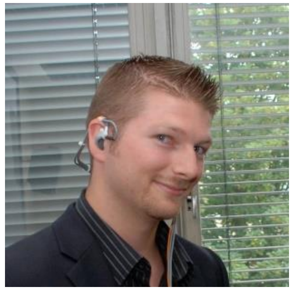
Collaborators and contributors

Audio and Acoustics Research Group in MSR Labs

Hannes Gamper
Microsoft Research

David Johnston
Microsoft Research

Ivan Tashev
Microsoft Research



Mark R. P. Thomas
Dolby Laboratories

Jens Ahrens
Chalmers University, Sweden

Interns: Piotr Bilinski, Archontis Politis, Keith Godin

The exceptional engineering teams in HoloLens, Kinect, and Windows we had the honour to work with

Introduction

VR & AR devices

Oculus Rift

Samsung Gear VR

Microsoft Hololens

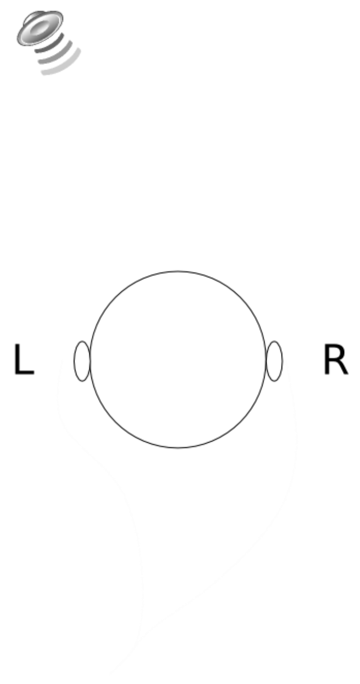
Definition: Spatial audio

- Within audible frequency and dynamic range
- Delivered to one or both ears
- Contains auditory localisation cues:
  - Interaural time and level differences
  - Spectral cues
  - Reverberation
  - Dynamic and multimodal cues
  - (expectation and experience)

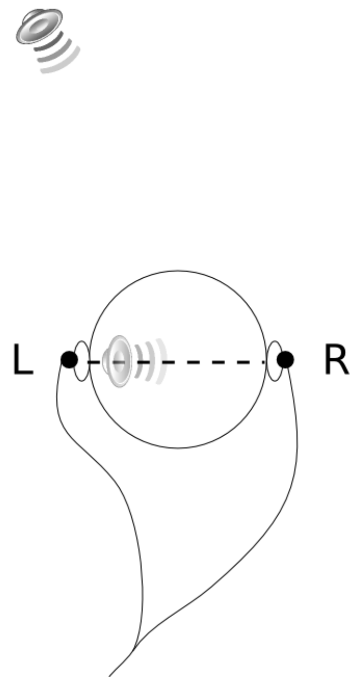
(Interaural time and level differences, Spectral cues) } Head-related transfer function (HRTF)

Spatial audio rendering

Spatial audio rendering

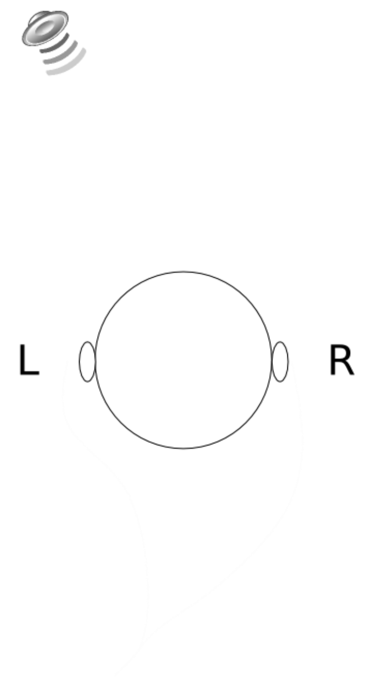


Real sound source

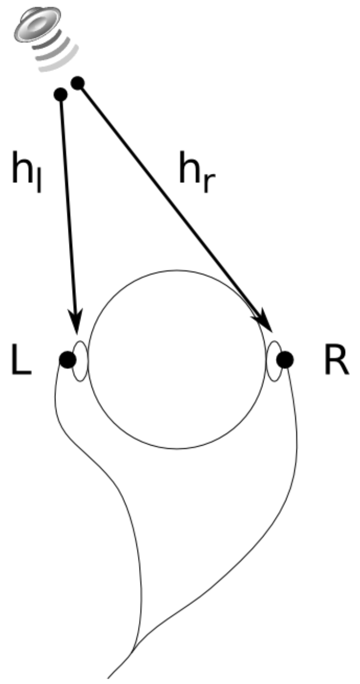


Virtual sound source

Spatial audio rendering

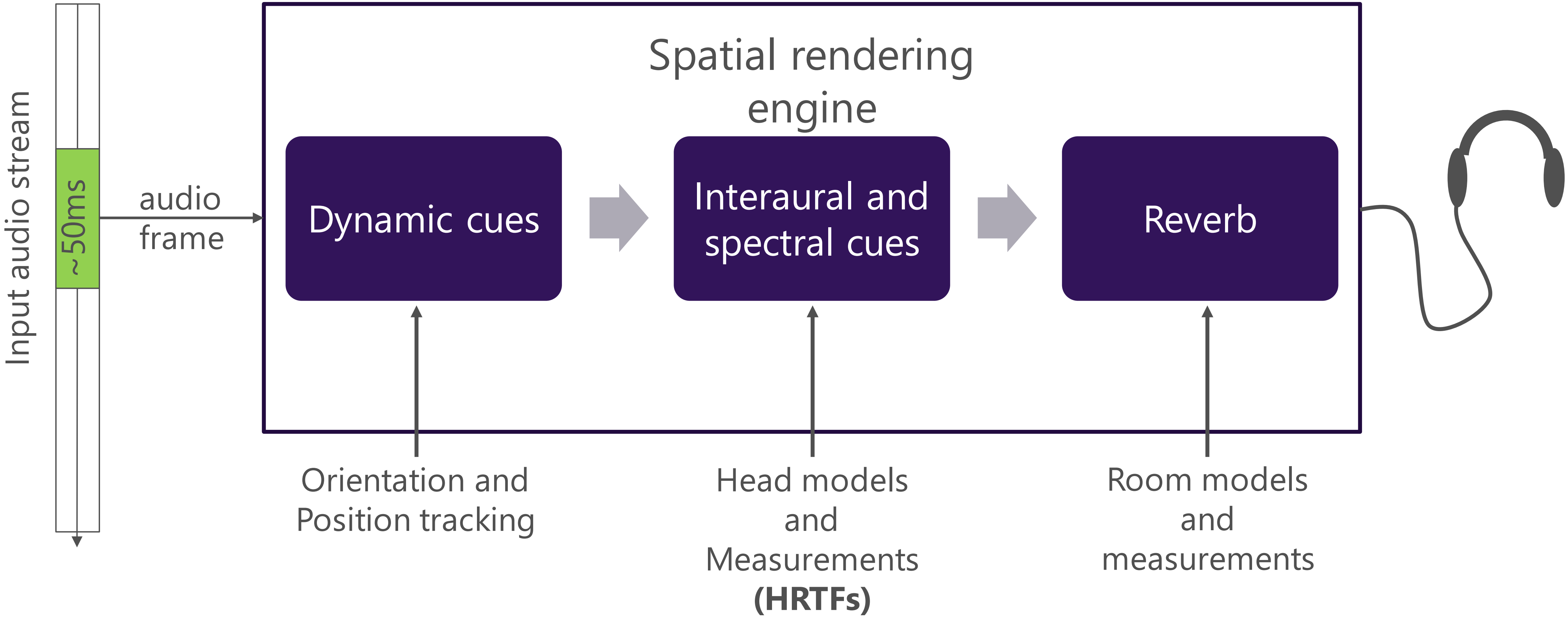


Real sound source

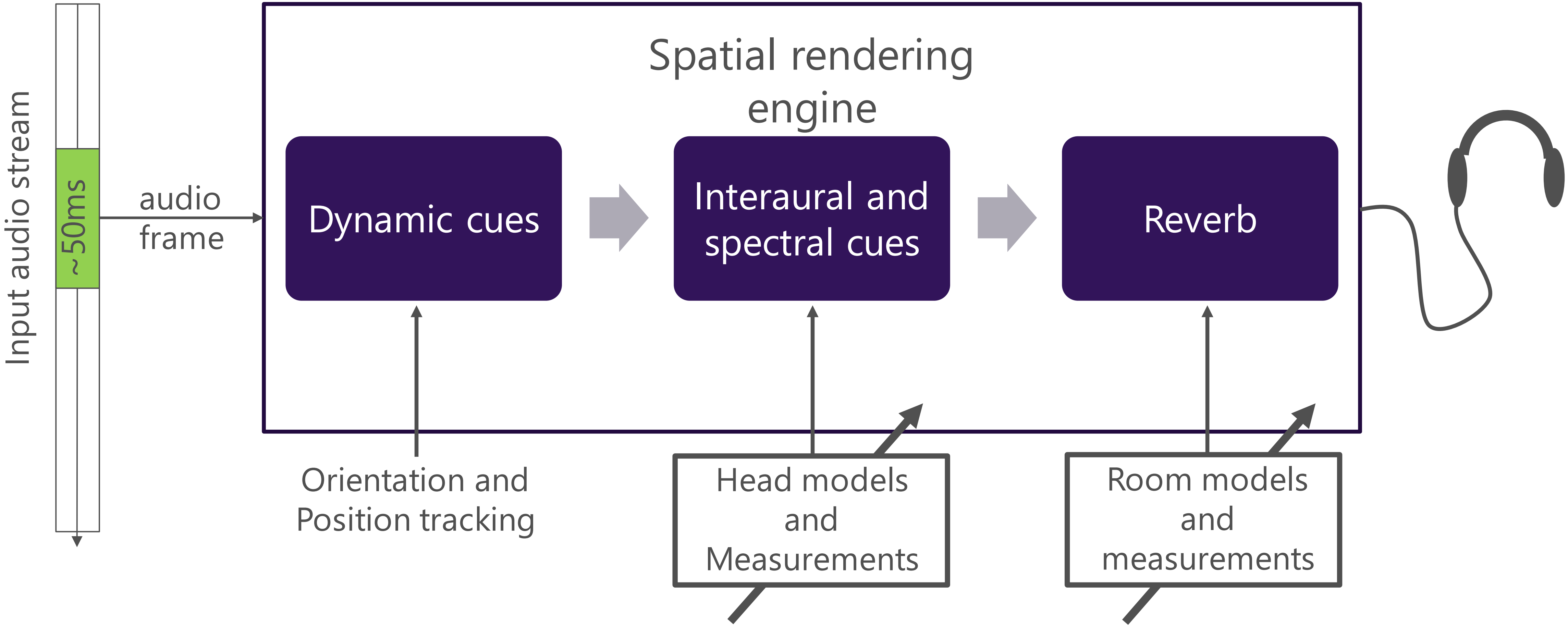


Spatial audio rendering via head-related transfer functions (HRTFs)

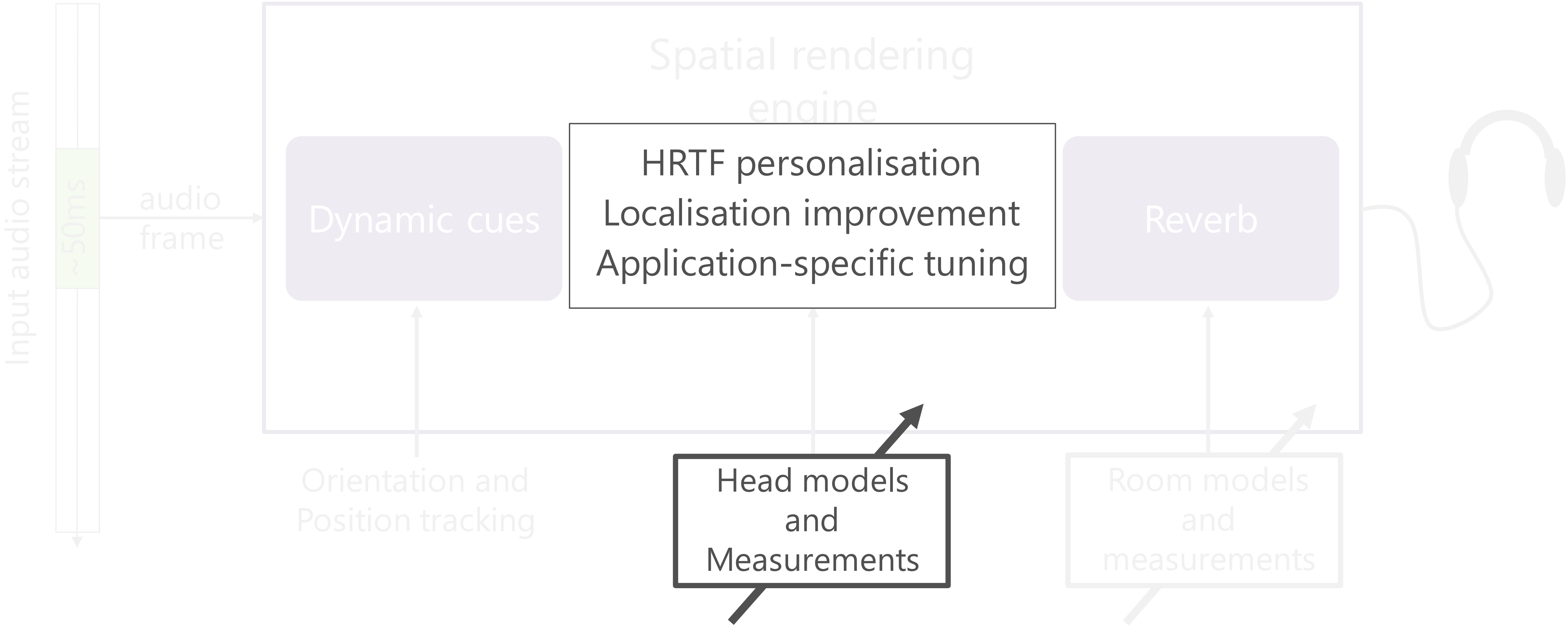
Spatial audio rendering



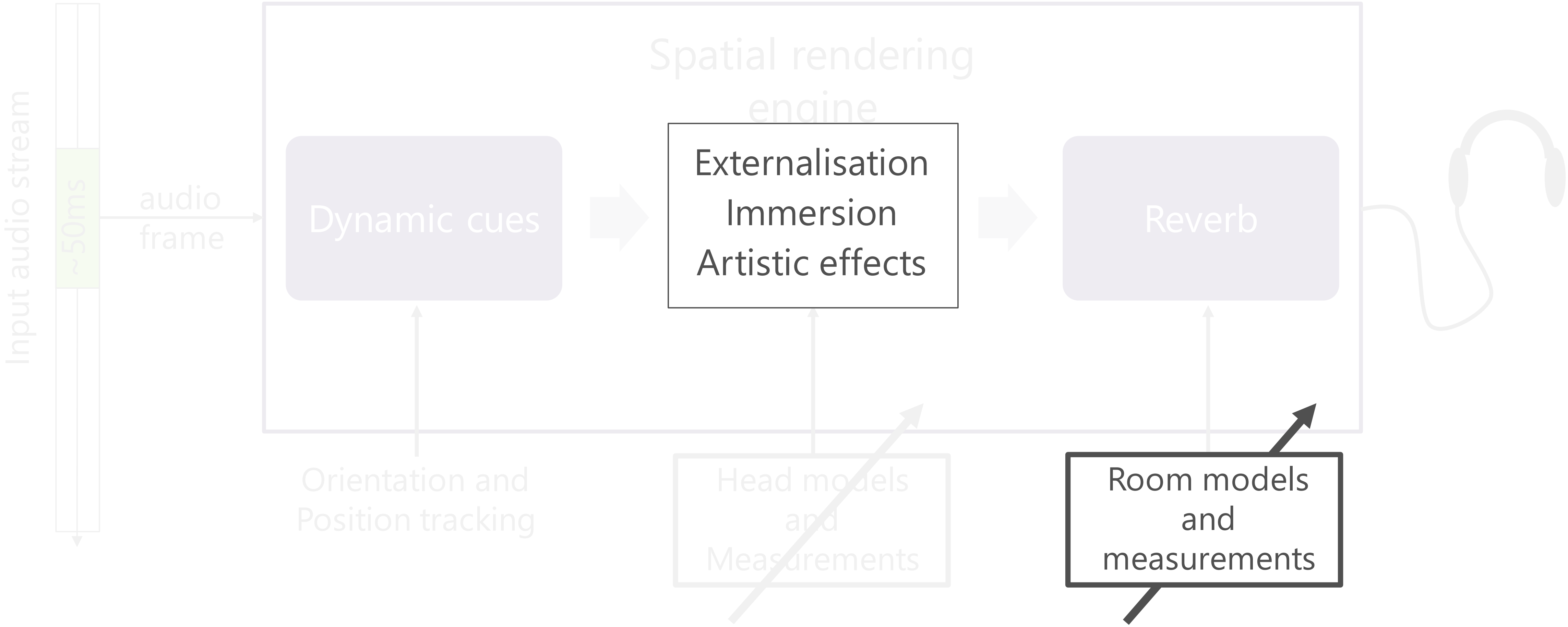
Spatial audio rendering



Spatial audio rendering



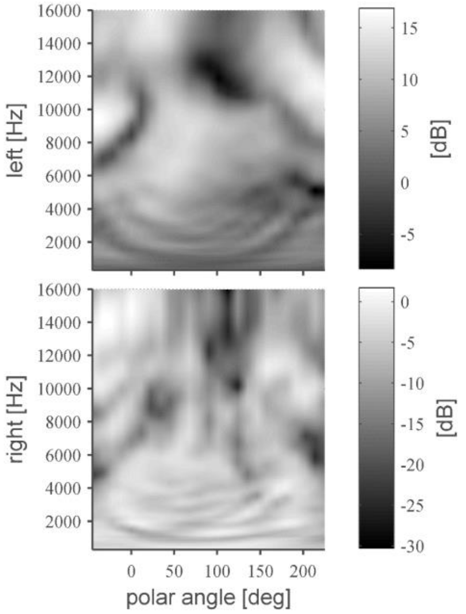
Spatial audio rendering



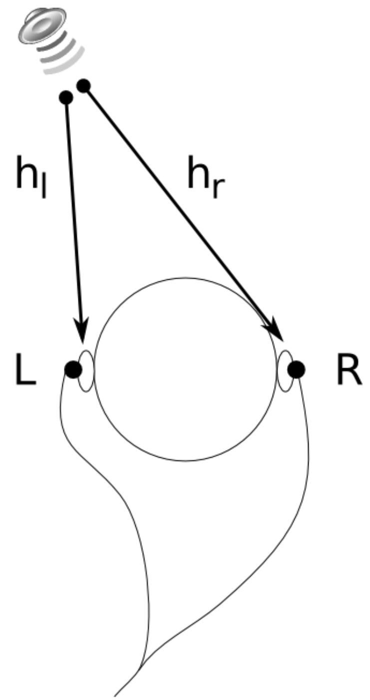
HRTF measurement and personalisation

Rendering framework

Personalisation



Rendering



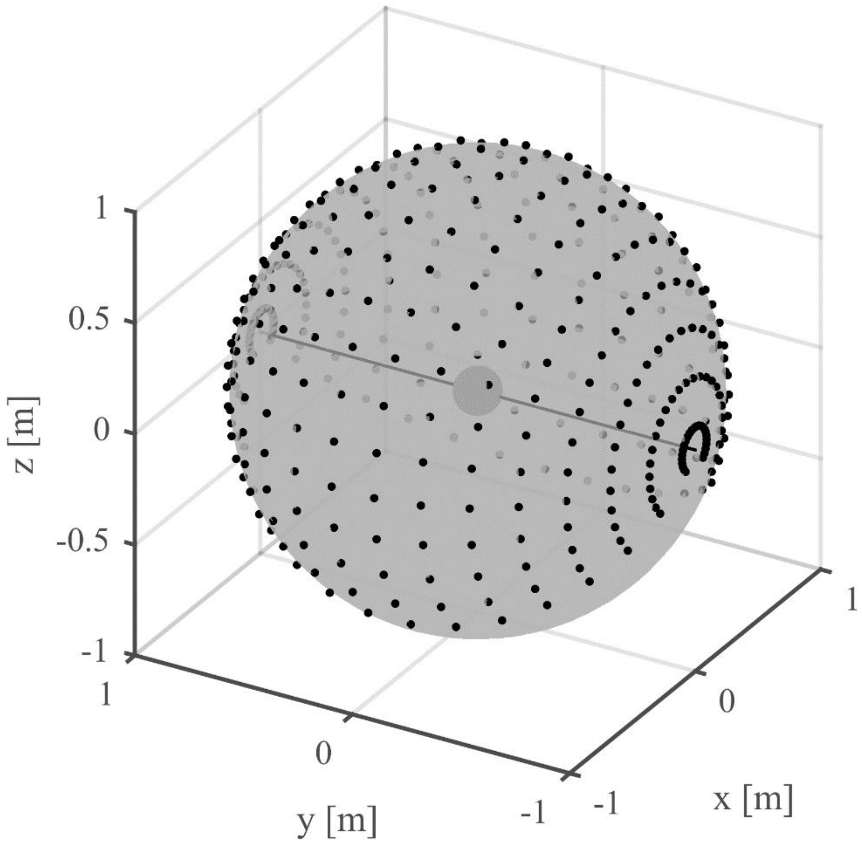
Anthropometry
3-D scan, photo, questionnaire, measurements

HRTFs
Measured, modelled

Spatial sound
Hololens, Windows 10, Cities Unlocked

HRTF measurement

HRTF measurement rig



Measurement locations

MSR HRTF database

- ~200 subjects
- HRTFs measured at 400 locations
- High resolution 3D head scans
- Direct anthropometrics measurements
  - Head width, depth, height, etc.
- Questionnaire
  - Hat size, shirt size, jeans size, etc.

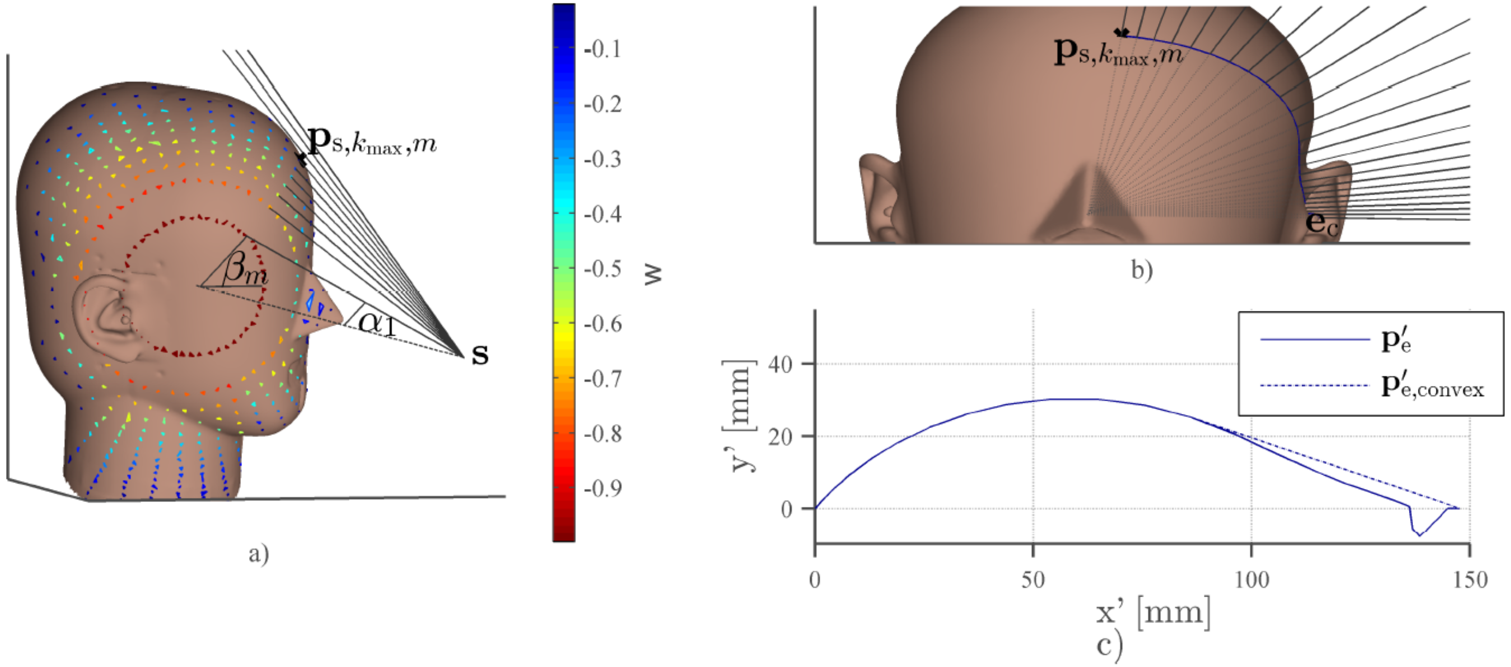
3-D head scan

Measurement tools

Direct estimation

Trace acoustic propagation from source positions to ear entrances.

Good results with high-resolution scan

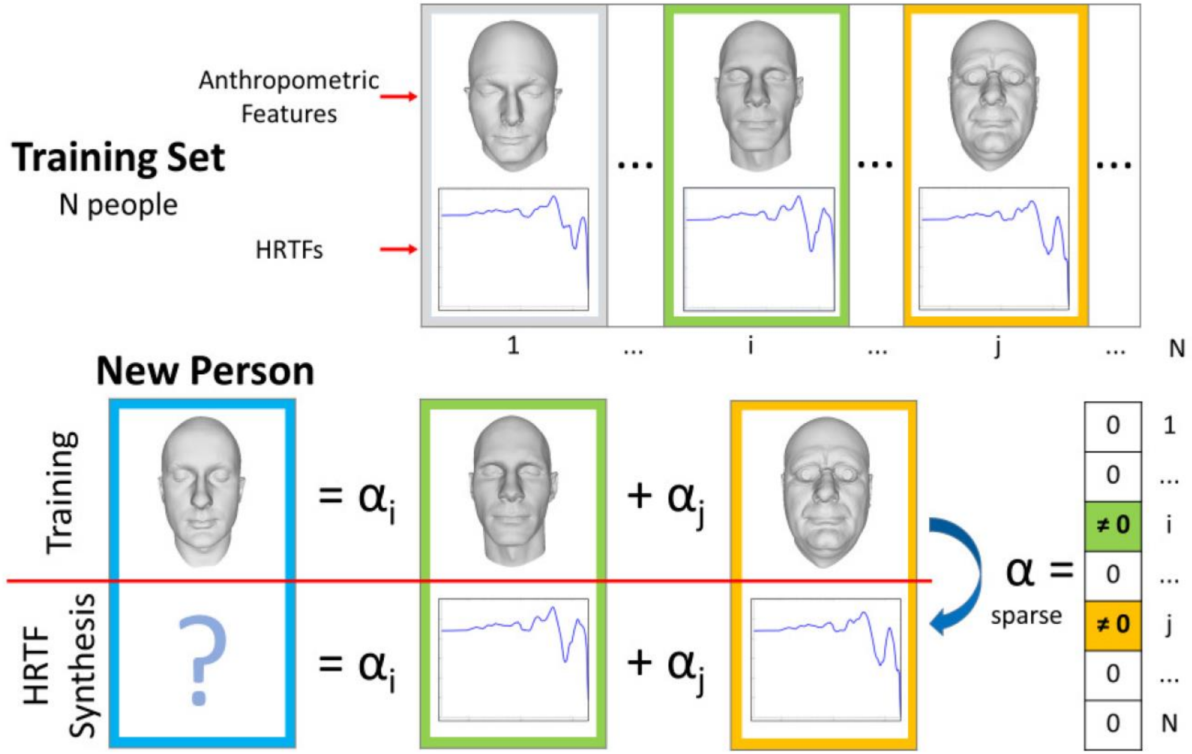


Gamper, H.; Thomas, M. & Tashev, I. (2015). "Estimation of multipath propagation delays and interaural time differences from 3-D head scans." *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP).*

Anthropometry-based personalisation

Given database of anthropometric features, represent a new candidate's features as a **sparse combination** α of people in the database.

Combine HRTF magnitude spectra with **same weights** α to synthesize personalized HRTF.

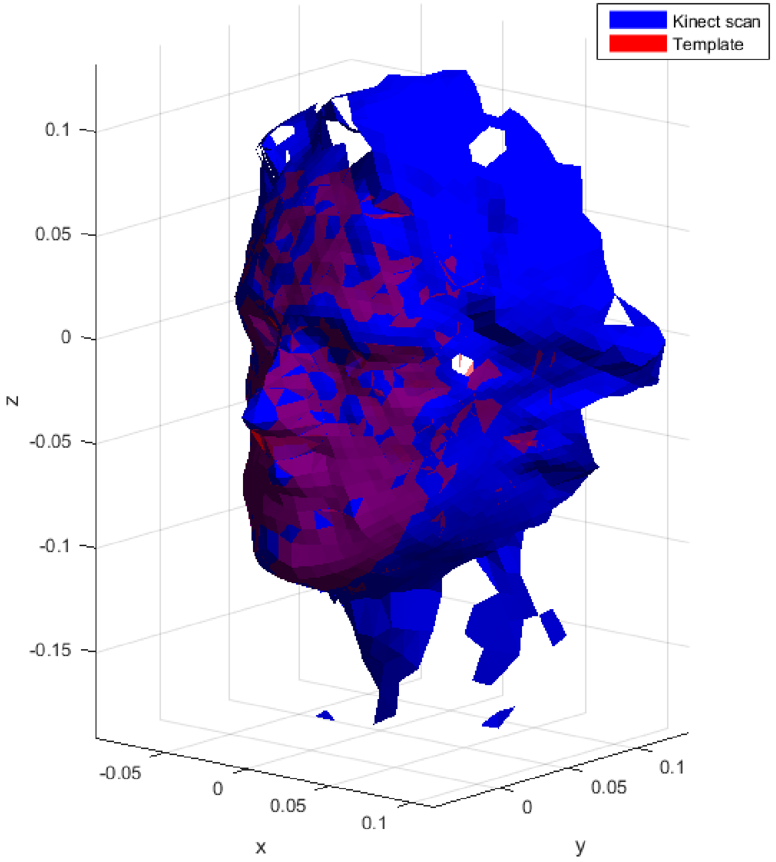


P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," *ICASSP*, 2014.

Model-based personalisation

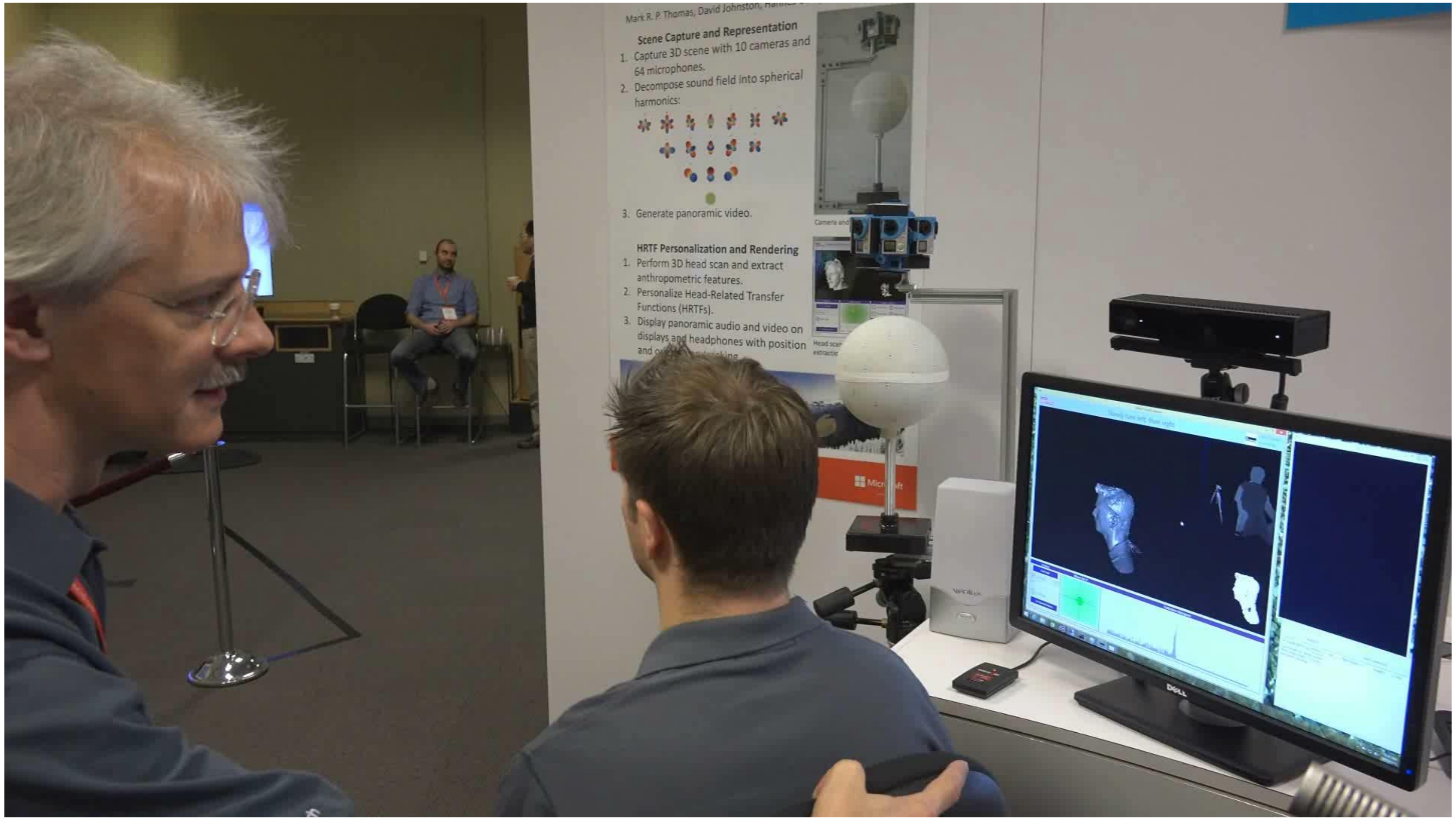
Given single (Kinect) depth image, fit average face to scan.

Map geometric deformation to acoustic features.



Average face fitted to depth image.

HRTF personalisation demonstration



Rendering approaches

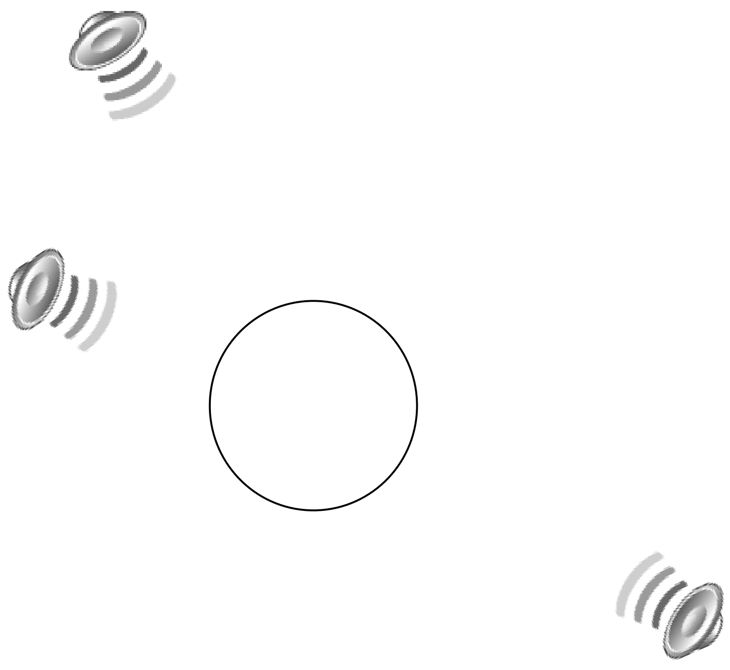
Object-based rendering

Render each source individually

Provides full 3-D control

Complexity increases ~linearly with #sources

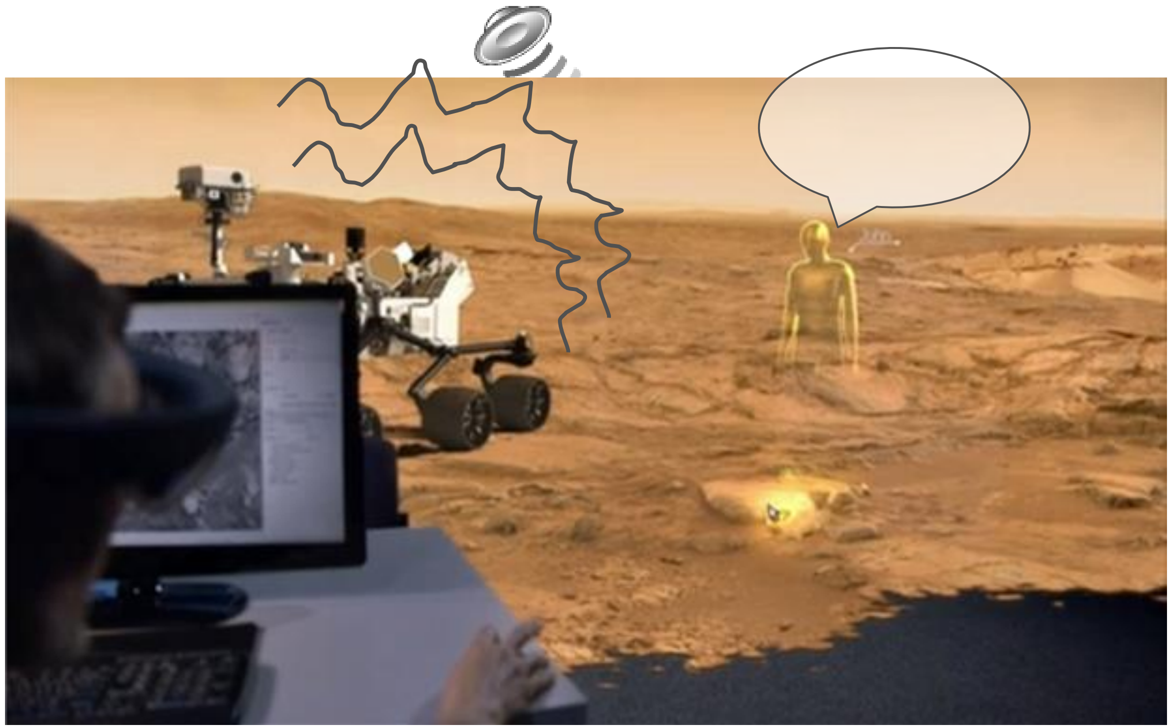
Suitable for synthetic (AR/VR) scenes
→ e.g., Hololens



Object-based rendering

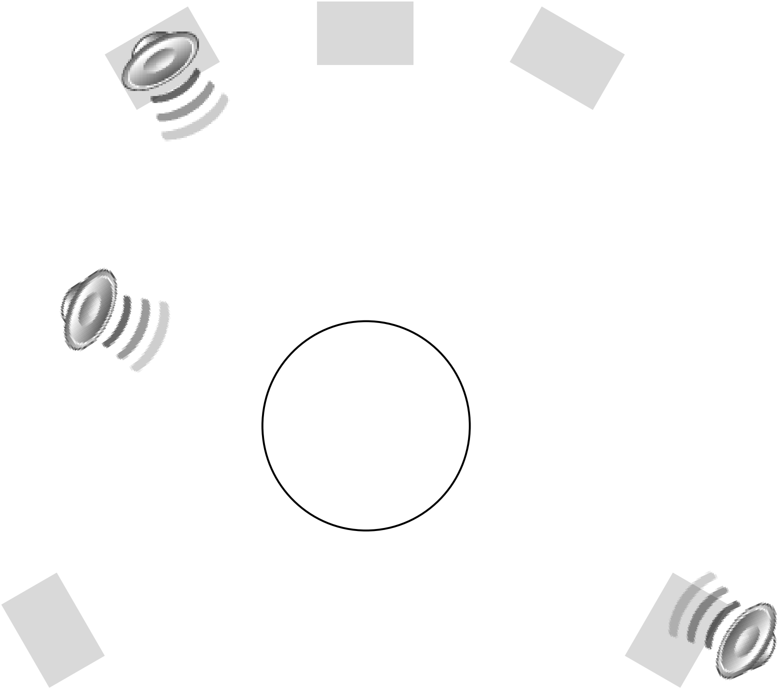
Render each source individually

sou



Channel-based rendering

d



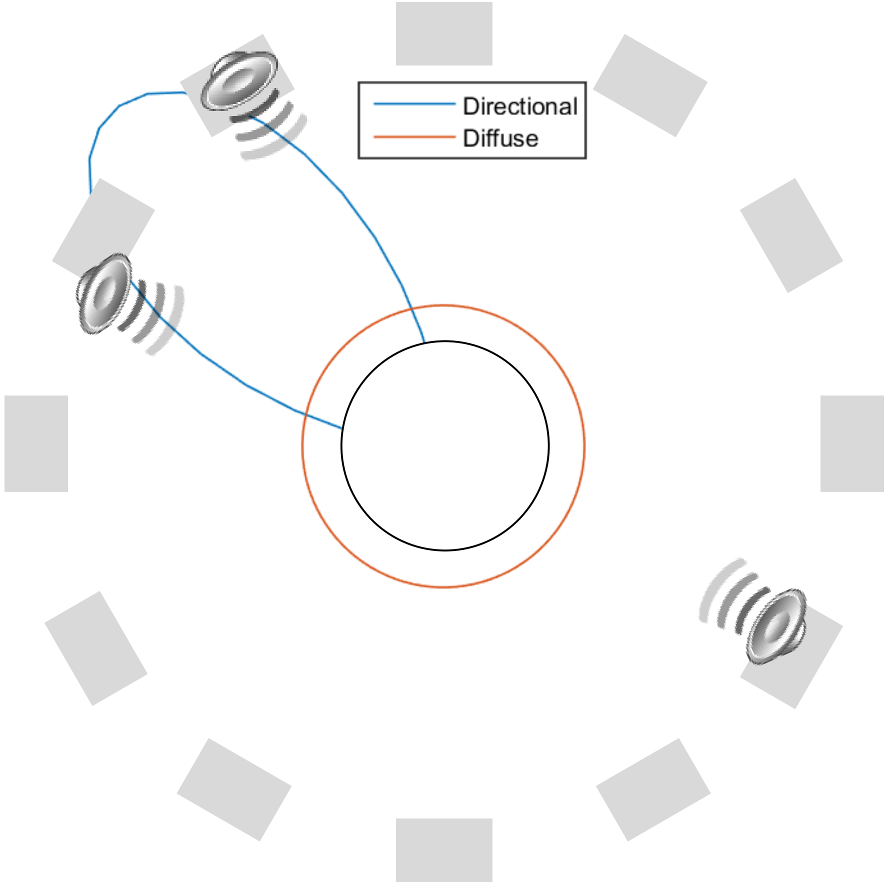
Parametric approaches

Render (fixed) number of virtual speakers

Based on psychoacoustics

Constant complexity

Suitable for spatial recordings
→ Directional Audio Coding (DirAC)*



*V. Pulkki, "Spatial sound reproduction with directional audio coding," J. Audio Eng. Soc., vol. 55, no. 6, pp. 503-516, June 2007.

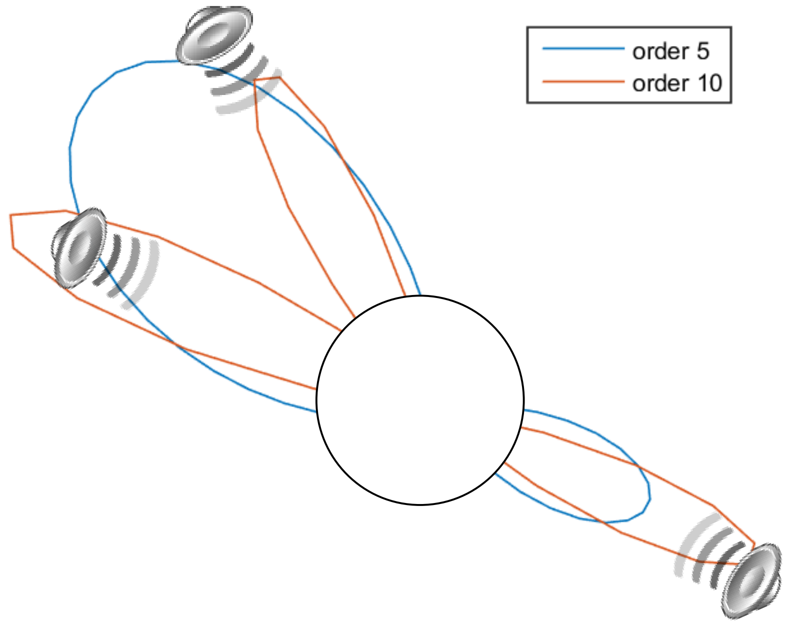
Modal rendering

Render fixed spherical order

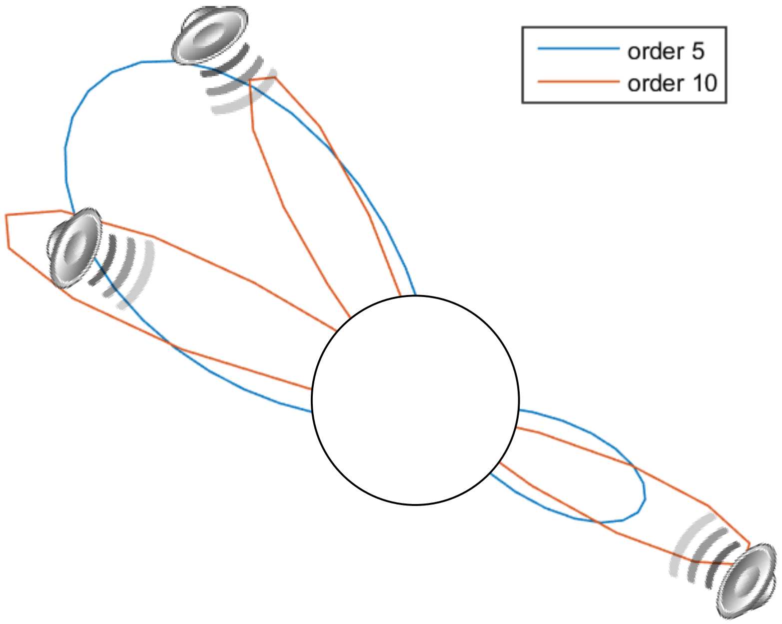
De facto media standard

Variable complexity

Suitable for spatial recordings (e.g., Ambisonics)
→ e.g., Ambisonics



Modal rendering



Sound field capture

16-ch. 4.5″
spherical mic. array

64-ch. 200mm
spherical mic. array

16-ch. 4.5″
cylindrical mic. array

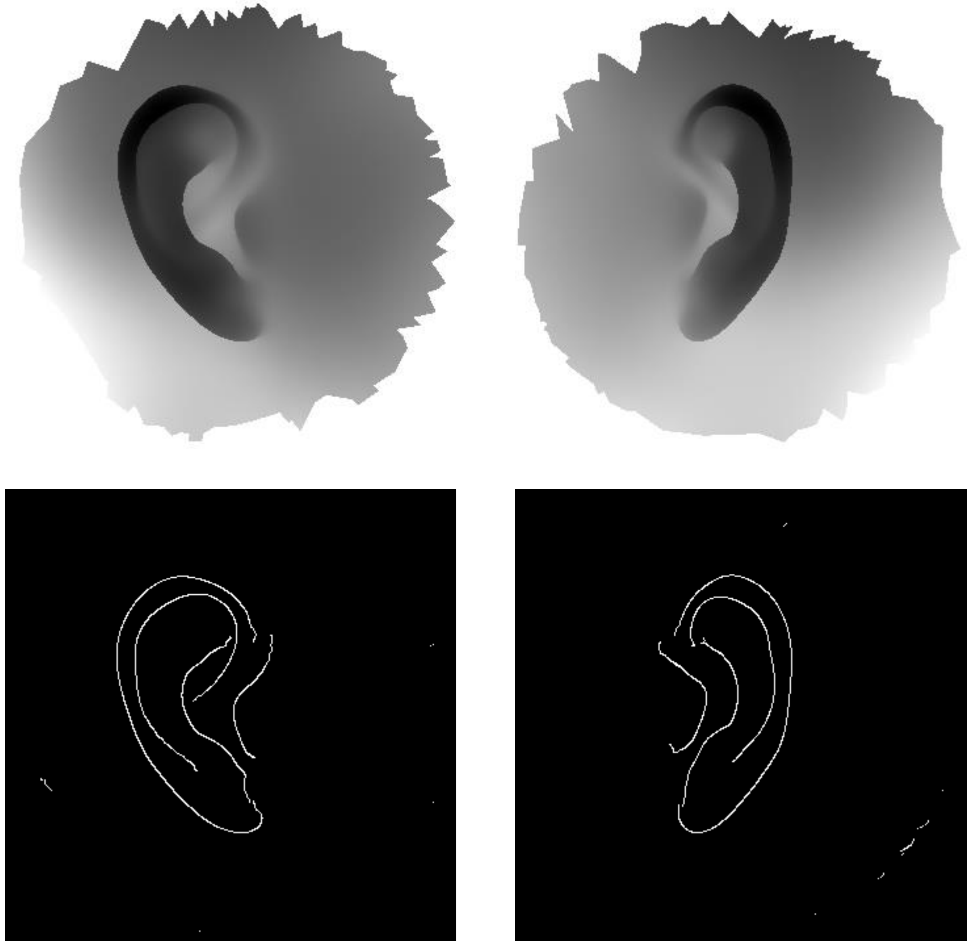
Future outlook

Improve rendering engine

Continue HRTF personalization efforts

Collect user feedback

Study elevation perception
(intern: Vani Rajendran)



Pinna scans.

HRTFs: Application-specific tuning

AR vs. VR

Object-based rendering vs. modal rendering

Dealing with constraints, expectation

Reverb and room modelling

AR vs. VR

Object-based rendering vs. modal rendering

Dealing with constraints, expectation

Conclusion

Spatial audio is key component of AR/VR experience

Growing number of devices/applications/users

Many open research questions – we have only scratched the surface!

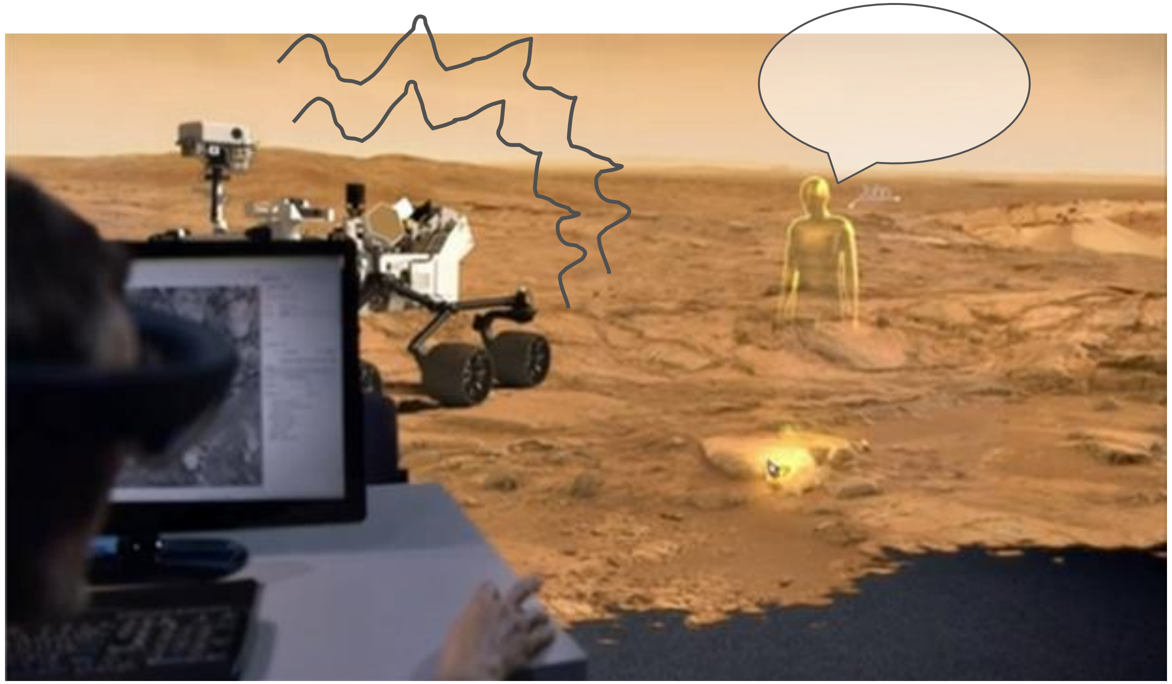
Thank you!

Microsoft

© Copyright Microsoft Corporation. All rights reserved.

Backup slides

AR & VR scenes



Microsoft Research
Faculty Summit 2016

Microsoft

Comparison: vision vs. hearing

	Vision*	Hearing
Frequency range	430 – 770 THz (1 octave)	20 – 20000 Hz (10 octaves)
Wavelength	700-390 m^{-9}	17 - 0.017 m
Dynamic range	~140 dB	~140 dB
Spatial resolution	~1 arc minute	~5-20 degrees
Temporal resolution	~1/25 s	~10-20 µs
Field of view	130° vertical, 200° horizontal	4π steradians
Energy	Up to 1000 W/m^2 in a daylight	Pain threshold 10^{-5} W/m^2

*Source: Wikipedia

→ Sound is a low energy phenomenon with wavelengths comparable to objects surrounding us.

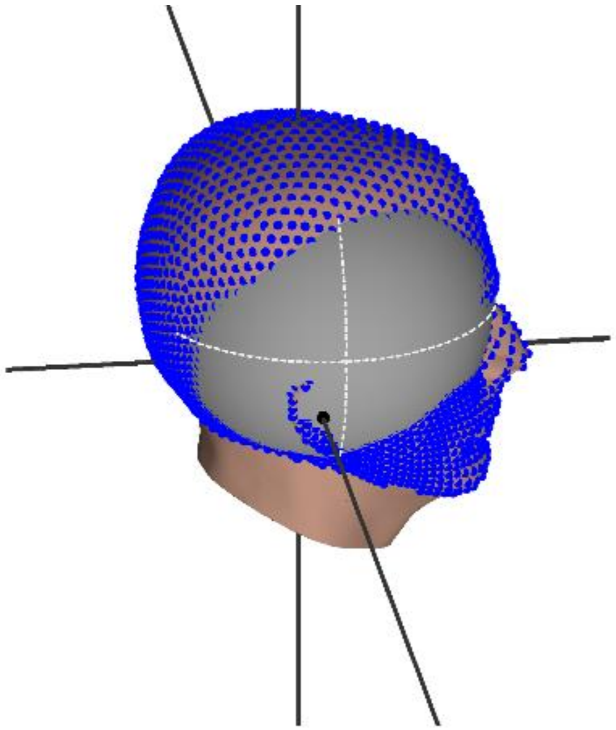
→ Human hearing has high temporal/low spatial resolution, unlimited field of view.

→ Both senses are head-locked!

Interaural time delay modelling

Fit sphere to scan to parameterise ITD models.

Should work with noisier scans (e.g., Kinect)



Sphere fitted to 3-D head scan.

Gamper, H.; Thomas, M. & Tashev, I. (2015). "Anthropometric parameterisation of a spherical scatterer ITD model with arbitrary ear angles." *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).*