Efficient Physics Based Simulation of Auditory Environments and Interactive 3D Sources for VR and AR: Theory, Tools and Workflows.

Ramani Duraiswami, Professor, University of Maryland
Co-Founder, VisiSonics Corporation
Virtual and Augmented Reality

VR/AR

Bets by Facebook, Google, Sony, Microsoft, Samsung, HTC

The way you interact with the world via your senses is smartly intermediated by technology

Predicted to reach $120B in 5 yrs

The next platform?

Personal Computers, 1990s
Networked PCs, 200x
Mobile, 201x
VR/AR: 202x
VR/AR

Virtual Reality
- Create artificial world that the user believes is real
- Simulation, Gaming and Entertainment

Augmented Reality
- Insert objects and information into the real world
- Training, Surgery, Entertainment, Pokemon Go, ...

Enablers
- Hardware: Moore’s law, displays, graphics, tracking,
- Software: Engines for creating virtual worlds
- Visual Perception: Improved latency, using persistence
Sensing the world visually and auditorily

- Vision and audition are stand-off senses
- Vision - foveated detailed view
- Audition - broad knowledge of general surroundings and communication
- Complementary!
- Same amount of real estate devoted to audio and video in the brain!
- Many interconnections, including to the motor areas
- Unless the world is rendered consistently for the experience the brain experiences fatigue
Problems needed to be solved in AR/VR

• Create virtual world
  – Gaming
  – Mixed reality experience

• Augment experiences
  – Operating system commands
  – Virtual assistants

• Video postproduction
  – Advertising
  – Feature movies

• Telepresence
  – Remote sports
  – Remote tourism
  – Teleconferencing

• Record multimodal experiences

• Experience rooms/materials
HOW WE PERCEIVE THE WORLD AUDITORILY
How do we perceive sound location?

- Naïve time and level difference at ears are not sufficient to describe our ability
- Other physics mechanisms necessary to explain
  - Scattering of sound
    - Off our bodies
    - Off the environment
  - Purposive Motion

Interaural time delays can only locate sounds up to a “cone of confusion”. For a given time delay, the source can lie on a surface of constant Time Delay:

$$|x-x_L| - |x-x_R| = c \delta t$$

A hyperboloid of revolution
Audible Sound Scattering

- Sound wavelengths comparable to human dimensions and dimensions of spaces we live in.
- \( f \lambda = c \)
- When \( \lambda \gg a \) wave is unaffected by object
  \( \lambda \sim a \)
  behavior of scattered wave is complex and diffraction effects are important.
- \( \lambda \ll a \)
  wave behaves like a ray
- the audio that reaches our ears does so after scattering off objects of similar size to wavelength

wavelengths are comparable to our rooms, bodies, and features
Not an accident but evolutionary selection!

![Graph showing frequency vs. wavelength with various dimensions labeled]
Human spatial localization ability

One may wonder how good the human ability to localize sound is. Experiments show it is quite good.
ACHIEVING ACCURATE 3D AUDIO
Approach: “Render” Sound Correctly

• Get sound right at entrances to the ear canals
  • Model audio propagation from sources in scene to ear
• Do what graphics and vision did –
  • Move from emulation to approximate simulation
  • Use physics based models
  • Render not only objects but scenes
• Simplify based on
  • knowledge of what is perceptible
  • Understand neural computation models used by users for achieving multimodal tasks
  • Personalize rendering to particular users
  • Level of detail based on available computing power
Mathematical modeling of propagation and scattering

- Could mathematically solve sound propagation

Wave equation:
\[
\frac{\partial^2 p'}{\partial t^2} = c^2 \left( \frac{\partial^2 p'}{\partial x^2} + \frac{\partial^2 p'}{\partial y^2} + \frac{\partial^2 p'}{\partial z^2} \right) = c^2 \nabla^2 p'
\]

Fourier Transform from Time to Frequency Domain

\[
P(x, y, z, w) = \int_{-\infty}^{\infty} p'(x, y, z, t) e^{-i\omega t} dt
\]

Helmholtz equation:
\[
\nabla^2 P + k^2 P = s \delta(x - x')
\]

Boundary conditions:

Sound-hard boundaries:
\[
\frac{\partial P}{\partial n} = 0
\]

Sommerfeld radiation condition
\[
\lim_{r \to \infty} r \left( \frac{\partial P}{\partial r} - ikP \right) = 0
\]

- Not feasible to do in real-time with today’s algorithms / supercomputers in real-time
Accurate Approximate Scattering

• Linear systems can be exactly characterized by impulse response (IR)
  – Knowing IR, can compute response to general source by convolution

• Full impulse response for a given source location and environment is the Binaural Room Impulse Response (BRIR)

• Goal develop an approximate compositional model for BRIR
  – combines the room impulse response (which has directional reflections and diffraction),
  – scattering off the person by directional sounds impinging on the body, via corresponding directional Head Related Impulse Responses (HRIR)
  – HRTF and RTF are Fourier transforms of the Impulse response
  – Convolution is cheaper in the Fourier domain
Breaking up the BRIR Filter

- Filter is extremely long and computation expensive to construct.
- Focus on perceptually important bits and make them accurately. Simplify.
- Convolution is linear – simplify by breaking up filter into pieces
- Early reflections are more important and time separated
  - Convolved with HRTF of appropriate direction
  - Important for determining range
- Later reflections are a continuum
  - Important for “spaciousness,” “envelopment,” “warmth,” etc.
- Create early reflections filter on the fly
  - Reflections of up to 6th order (depending on computational resources)
    - These are convolved with the HRTF corresponding to the direction
- Tail of room impulse response is approximated depending on room to be perceptually pleasing
Why this works

• Interactions change received sound waves
  • Scattering of body and ears
    – Bodies ~ 50 cm
    – Heads ~ 25 cm
    – Ears ~ 4 cm
    – Not much multiple scattering
  • Scattering off surroundings
    – Rooms ~ 2m – 10m
    – More multiple scattering
    – Larger sizes => lower frequencies

Because of this separation of scales we can model these effects relatively independently.
Approximation of the BRIR

-- Start with the pre-computed tail of IR (reflections $n^{th}$ order and up)
-- Quickly compute the reflections of order $0-(n-1)$ for current geometry
-- (Reflection of order 0 is just the direct arrival)
-- Stick them onto this generic tail
-- parts that are perceptually important are updated in real time
Psychophysics and Personalization

• Beyond HRTFs
• Weighting and integration of sound-localization cues by human listeners.
  – Task dependent
  – Understand low level Cues (gross ITD, ILD)
  – Neural computation strategies

Copyright © 2016 VisiSonics
CUSTOMIZING THE ENGINE
Head Related Transfer Function

- Scattering causes frequency dependent amplification/attenuation
  - Effects can be of the order of tens of dB
  - Encodes location
HRTF Personalization

• Humans have different sizes and shapes
• Ear shapes are very individual as well
• If ear shapes and body sizes are different
  – Properties of scattered wave are different
  – HRTFs will be very individual
• Need individual HRTFs for creating virtual audio
Typically HRTFs are measured

Measurements from CIPIC, Sydney and AFRL are shown

Many other labs also measure HRTFs directly
Reciprocity Principle:

If sound waves are excited at a point A, the pressure at any other point B coincides, in magnitude and phase, with the pressure that would be observed at A if the sound source were placed at B.
Patented Fast Approach

• Turned out headphone drivers
• Array of tiny microphones
• Send out a highpass signal and measure received signal
• Use analytical anthropometric representation for low frequencies and compose
Compensating for low frequencies

• Goal: measure HRTFs over 20Hz-20 kHz

• Good low-frequency measurements are problematic,
  – large loudspeakers necessary to emit energy at low frequencies

• Solution: Use anthropometric model for HRTFs below 1kHz

• Low frequency HRTFs well represented by such models

• Perceptually they even sound better!

• Over 1000 HRTFs measured in this way and tested
Comparisons

• Direct vs. Reciprocal (Zotkin et al. 2006. JASA)


Other research on personalizing HRTFs

- Computations from meshes via fast-multipole accelerated BEM
- Machine learning from measured HRTFs and anthropometry
- Fundamental feature analyses
- Goal: to personalize the audio for everybody in the world
Room impulse response

• Room reflections that reflect the space, as well as occlusions are very important
• Many graphics researchers working in 3D audio have focused on this part of the rendering as it allows use of ray-tracing codes already developed.
• However, perceptually only leading few orders of reflections are important
• We use a fast to compute approximation model to convert room to a union of boxes to compute approximate reflections
Speeding up authoring

• Union of shoebox model generation for arbitrary scenes via bounded volume hierarchies
• Interfaces with material data if available
THE REALSPACE™ 3D ENGINE FOR VR/AR GAMING AND VR VIDEO
• Scalable, efficient – provides return on computing power
  – Combination of frequency domain and time domain processing
  – FFTs, multithreading, parallel processing, perceptual approximation

• Dynamic realism through interplay of perceptual cues
  – Head Tracking
  – Propagation and Room Scattering (Room Impulse Responses)
  – Body Scattering (Head Related Transfer Functions)

• Extensible

Shipping on various platforms

http://www.realspace3daudio.com
Game and Audio Engines Supported

**Standalone Engines**
- Android
- Linux
- Microsoft

**Audio Middleware**
- Wwise
- Fabric

**Forthcoming**
- Lumberyard
- Consoles

**Game Engines**
- Unreal Engine
- Unity
RS3D Audio Unity Plugin
Game Engine Workflows

Editor with rich authoring and scripting environment
Compiles to efficient run time executable
Hotkeys for changing sounds:
(1) Crushin' by AudionautiX.com
(2) Cool Ride by AudionautiX.com
(3) Pop Tune by AudionautiX.com
(4) Creaking Doors & Gate
(5) Crackling Campfire
(6) Footsteps
(7) Washing Hands
(8) Voice: News & Weather
(9) Voice: Speakerphone
(0) Voice: Shakespearean Insults

Hotkeys for changing reflection presets:
(E) No Reflections
(R) Tasteful Reflections
(T) Considerable Reflections
(Y) Maximum Reflections

( ) Toggle extra GUI
Video Workflows

• Standalone Editor – Project Chorizo
  – Import video and audio stems, and place sounds via keyframes and sound path
  – Publish to object video or ambisonics

• Editor can be slaved to DAWs or video editors

• Export to HMDs, smartphone VR and YouTube
Project Chorizo: Engine for Video Workflows
RealSpace 3D addresses audio requirements for developers at all levels of VR/AR
CAPTURE SOLUTIONS
Capture and Telepresence

- Place microphones at a remote location (e.g. concert hall)
- Replay spatialized audio at a remote location
- Must play it for many users
- Use rendering algorithms/representations
• sound at a point can be represented in terms of the local point-eigenfunctions of the Helmholtz equation

\[ \psi_{in}(k; r) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} A^n_m R^n_m(k; r), \]

\[ R^n_m(k; r) = j_n(kr)Y^n_m(\theta, \varphi), \]

– Expand solutions in series, but truncate at \( p \) terms causing an error \( \varepsilon_p \)

\[ |\varepsilon_p(s, r)| \lesssim \exp \left\{ -\frac{1}{3} \left[ 2 \frac{p - kR}{(kR)^{1/3}} \right]^{3/2} \right\} = \delta_p, \quad kR \gg 1. \]

– Error depends on frequency
  • For a given sound of wavenumber \( k \) this gives us minimum order for sensible representation
• Analysis of solutions of the Helmholtz equation in our book
  – Elsevier, 2005
• What do these basis functions look like?
Spherical Harmonics

\[ Y_n^m(\theta, \varphi) = (-1)^m \sqrt{\frac{2n + 1}{4\pi}} \frac{(n - |m|)!}{(n + |m|)!} P_n^{|m|}(\cos \theta) e^{im\varphi}, \]

\[ n = 0, 1, 2, \ldots; \quad m = -n, \ldots, n. \]

Zonal \hspace{1cm} Tesseral \hspace{1cm} Sectorial

\[ Y_n^m(\theta, \varphi) = Y_n^m(\theta, \varphi). \]

\[ P_0^0(\theta, \varphi) = \text{const} = \frac{1}{\sqrt{4\pi}}. \]
Yet another representation (Plane Waves)

- Soundfield at a point in can be expressed as a sum of plane waves.
  - Integral over a unit sphere at the point
  - Decomposes any sound field into a set of planewaves of various strengths

- Connected to spherical representation

  \[ \psi_{in}(\mathbf{r}) = \frac{1}{4\pi} \int_{S_u} e^{ik\mathbf{s} \cdot \mathbf{r}} \mu_{in}(\mathbf{s}) dS(\mathbf{s}), \]

  - In practice these integrals are evaluated via quadrature

  \[ e^{ik\mathbf{s} \cdot \mathbf{r}} = 4\pi \sum_{n=0}^{\infty} \sum_{m=-n}^{n} i^n Y_n^{-m}(\mathbf{s}) R_n^m(\mathbf{r}), \quad R_n^m(\mathbf{r}) = \frac{i^{-n}}{4\pi} \int_{S_u} e^{ik\mathbf{s} \cdot \mathbf{r}} Y_n^m(\mathbf{s}) dS(\mathbf{s}), \]

  - Approximation error in this case is related to error in the quadrature

  \[ \int_{S_u} F(\mathbf{s}) dS = \sum_{j=0}^{L_Q-1} F(\mathbf{s}_j) w_j, \quad F(\mathbf{s}) = \sum_{n=0}^{p-1} \sum_{m=-n}^{n} C_n^m Y_n^m(\mathbf{s}), \]

  - Quadrature error formula relates \( L_Q \) to \( p \)
Capture to Representation

- VR/AR 3D hardware capture solution
- VisiSonics’ Spherical Microphone Array
  - Real-time 3D Audio Capture
  - 64 microphone 360 degree capture
  - Competitors capture only in one plane
  - We capture to High Order Ambisonics and plane-waves
- Enabling “Best Seat in the House” recording and Real-time Telepresence
- Also enable real-time spatial audio visualization

Customers in Automotive, Defense, R&D
Patented digital microphone chain architecture
Sound Travel Time: 17ms, Distance: 5.70 meters
OTHER SOLUTIONS FROM VISISONICS
Other solutions not discussed in this talk

- Capture to high-order ambisonics and beamforming
- Ambisonics from microphones placed on objects that are non-spherical
- Digital Microphones for arbitrary microphone array layouts
- Noise visualization and 3D sound-field analysis tools for listening spaces and automobiles
VisiSonics

- Provides 3D audio creation and capture
- Software fits into content creator workflows
- Software is a component in capture hardware
- Become the 3D Audio Company for VR/AR
- Personalize the experience for every end-user
Thank you...