Overview

• Human perception of sound and space
  • ITD, IID, HRTFs, and all that

• 3D audio
  • Measuring HRTFs
  • Synthesizing 3D audio

• Virtual audio
  • Synthesizing virtual audio
What is 3D audio?

• Fooling a listener that a sound is coming from a specific location around them

• Two ways to get it:
  • Easy: Using headphones
  • Hard: Using speakers
What is virtual audio?

• Modeling the effects of being in a virtual environment
  • Includes 3D audio effects
  • Also includes room effects
  • Also includes additional environmental effects
Why bother?

• Entertainment
  • Immersive gaming, 3D movies, virtual worlds, ...

• Practical
  • Help listeners parse more audio streams simultaneously
  • Help users localize multiple sources
    • e.g. pilot discussions in place cockpits
  • For grabbing people’s attention
    • E.g. in auditory display interfaces
A bit of hearing theory

• In order to synthesize 3D audio we need to know how to fool the human ear

• What are the cues that we need to use?
  • And how do we implement them?

• Lots of levels of complexity
On having two ears

• Why are our ears on the sides of our head?
• Why not one on the chin and one on the forehead?
Fundamentally different than vision

- Unlike our eyes that directly perceive 3D, our ears have to get that “computed” in the brain

- Special neural circuits in the Superior Olivary Complex (SOC) compare signals from both ears
The Duplex Theory

• Formulated by Lord Raleigh (1907)

• A listener’s ears receive a sound with some minor differences which act as localization cues

• The two main cues
  • Interaural Time Differences (ITD)
  • Interaural Intensity/Level Differences (IID, or ILD)
Interaural Time Differences (ITD)

- Simplest possible cue
  - Relative time difference between a sound reaching our ears
    - Sounds familiar?
ITD tradeoffs

- Perceiving ITDs is increasingly more unreliable with higher frequencies
  - Historically the cutoff was set to 1.5kHz (any guess why?)

- But we also perform ITD with the envelopes of sounds that we hear
  - So we use higher frequencies as well
How we will model it

- We can simulate ITDs with delays
  - Similar idea to the mic array steering vector
- There will be an upper limit to the delay
  - What is it?
One more thing

- The precedence effect (a.k.a. Haas effect)

- Up to 40msec delays register as an ITD
  - More than that and we form echo percepts

![Approximate delay time to left channel (msec)](image-url)
Interaural Intensity Differences (IID)

- For wavelengths smaller than the listener’s head we observe sound absorption
  - High frequencies get attenuated
  - Low frequencies pass mostly unharmed

- Level differences in high frequencies are a very strong cue to help us localize sounds
  - They are called IIDs, or ILDs
    - For intensity or level
IID tradeoffs

- IIDs mostly apply to wavelengths shorter than the head of the listener
  - About a 1.5kHz cutoff
  - Lower frequencies diffract around the head

- IIDs work better when the sound source is off the plane between the two ears
  - Otherwise there is no relative head shadowing
    - What’s an example location?
How can we model it?

- Easy to model using gain between ears
  - The “panpot” model
    - Ignores frequency dependencies (more later)
- Can be implemented as a filter
“Lateralization”

- ITDs and IIDs tend to produce lateralization
  - The percept of a sound on the axis between ears
  - “Inside the head” effect

- Useful for studying perception
  - But not quite 3D sound
Combining ITDs and IID

- We can very simply combine both cues
  - This will give us a rudimentary 3D system

- Each ear gets a filter
  - Filter imposes a time delay for ITD
  - And a gain factor for the ILD

- Demo!
Cones of confusion

- There are parts of space that will result in the same ITD and IID values
  - We cannot distinguish sounds from these locations
    - At least not well
    - In real-life we resolve that by moving our heads

![Diagram of cones of confusion](image)
Zoological intermission

- The Barn Owl
  - Hunts through hearing in the dark

- Can shape its face to funnel sound towards its ears

- Has asymmetrical ears
  - Can use ITDs for horizontal, and IIDss for vertical localization
Entomological intermission

• The Ormia Ochracea
  • Finds host crickets through hearing
  • Very good at localization!

• Ears are 0.5mm close
  • How does it use ITD/IID?

• Coupled eardrums create new cues
  • Currently used as model for new mics
One cue to rule them all!

- ITDs and ILDs can be insufficient
  - Very simple model of environment
  - Our ears adapt to localize and are in fact a lot smarter

- Head Related Transfer Functions (HRTFs)
  - Incorporating more, and finer cues for localization
What to HRTFs capture?

- Many effects relating to our body
  - Funneling by the ears, reflections off our shoulders, sound absorption from head, effects from hair, ...
- They also incorporate ITDs and ILDs

![HRTF of sound from the right](image)
How do they look like?

- Sweep from front to back (right side)
How do they look like?

- Sweep from front to back (right side)
How do they look like?

- Sweep from down to up on the right
How good are HRTFs?

• Each person has a different head/torso shape
  • We often just use average HRTFs
    • They won’t work for everyone
    • Being average helps in this case!

• But how to we get HRTFs?
Solution 1: Binaural recordings

- Use a dummy head to make 3D recordings
- Or stick microphones in your ears
  - (but please don’t stick anything in your ears!!)
Solution 2: Measure real HRTFs

- If we measure real HRTFs we can then use them on arbitrary sounds to make 3D audio
  - Just apply them as filters to generate left/right/signals

- Two ways to measure HRTFs
  - Measure a dummy head’s HRTF
    - Should be an “average” set
  - Measure your own HRTFs
    - You then have a personalized copy
How do we measure HRTFs?

- Same process as measuring room responses
  - Setup microphones in dummy of human subject
  - Play MLS from different locations
  - For each location measure the transfer function
    - You should remove the speaker/mic functions though

- Pro tip
  - You should do that in an anechoic chamber
    - Why?
In math

• We record:

\[ y_{\theta,\phi}[t] = h_{\theta,\phi}[t] * x[t] \]

\[ Y_{\theta,\phi}[\omega] = H_{\theta,\phi}[\omega]X[\omega] \]

• We deconvolve with:

\[ H_{\theta,\phi}[\omega] = X^*[\omega]Y_{\theta,\phi}[\omega] \]

• We remove speaker/mic responses
  • Use inverse filters of these responses
    • How do we measure these?
One complication

- This requires some serious lab space
One more complication

- We measure the transfer function from the source location to inside the ear

- What will convolution with an HRTF give us?
  - How do we reproduce it to sound as being 3D?
Synthesizing 3D audio

- Pick a location to position a source
  - Usually azimuth/elevation

- Select appropriate filters from HRTF set
  - Note that there is left/right symmetry so there is no need to keep all of the HRTFs

- Filter sound to model 3D effects
  - What about moving sounds?
Fast convolution reminder

- Convolution can be sped up significantly using the FFT
  - Perform convolution in the frequency domain
  - Complexity drops to $2N \log_2 N$

\[ z = x \ast y \iff \text{DFT}(z) = \text{DFT}(x) \odot \text{DFT}(y) \]

- But is this useful for our case?
  - No, results in very large FFTs, doesn’t allow for changing filters

- Using the STFT for convolution instead
  - Convolve each STFT frame with the desired filter at that time
Overlap-add fast convolution

• Similar to spectrograms
  • Step 1: Make frames
    • Zero pad to accommodate convolution’s output length
    • Hop size == frame size
    • Do not window
  • Step 2: Convolve frames using FFTs
    • i.e. multiply complex spectra
      • Multiply each STFT frame with the DFT of the desired filter
  • Step 3: Invert back to time
    • Use overlap and add!
    • Do not window
Usual problems

• Response mismatch
  • People with funny head shapes
  • Poor reproduction (e.g. bad headphones, MP3s)

• Front/back confusion
  • Really prominent for many people

• Head movements
  • Chance the relative angle of a source
Compensating for head movement

- We can track the listener’s head movements
  - Using a simple sensor on the headphones
  - Or using computer vision to measure head pose

- This allows us to find the angle between the virtual source and the rotated use head
  - One drawback: Time lag
  - One advantage: We can resolve localization ambiguities
    - We use head movements to deal with ambiguities
What about speakers

- We need to perform crosstalk cancellation
  - Use “negative” signals to construct HRTF filtering

- What are the complications here?

Figure 5.13. Illustration of crosstalk cancellation theory. Consider a symmetrically placed listener and two loudspeakers. The cross-talk signal paths describe how the left speaker will be heard at the right ear by way of the path $R_{ct}$, and the right speaker at the left ear via path $L_{ct}$. The direct signals will have an overall time delay $t$, and the cross-talk signals will have an overall time delay $t + t'$. Using crosstalk cancellation techniques, one can eliminate the $L_{ct}$ path by mixing a 180-degree phase-inverted and $t'$-delayed version of the $L_{ct}$ signal along with the $L$ signal, and similarly for $R_{ct}$.
Complications with speaker systems

- **Head movements**
  - We need to compensate for moving ears!
  - Not trivial to cater to multiple people simultaneously
    - E.g. you won’t get 3D sound in a movie theater

- **Room effects**
  - Speaker output gets convolves with room
    - and speakers ...
  - Difficult to compensate for all that
Moving towards virtual sound

• 3D sound models source-to-ear effects
  • Created 3D percept, but this is not the whole story

• There are more cues that we use to localize
  • Movement cues, distance cues, context cues, ...

• Proper virtual audio also models these cues
Movement cues

- Moving sources exhibit an additional important cue for localization
  - The Doppler effect
Modeling the doppler effect

- Variable delay lines
  - We can read off a delay line with interpolation
    - Sort of like changing the sample rate

- Tricky to get good interpolation
  - More later in the semester
Distance cues

• We can also perceive how far a sound is

• Static cues
  • Level, amount of reverberation

• Dynamic cues
  • Change of source angle by head translation
And some more context cues

• Room acoustics
  • Sounds in different parts of a room sound different

• We can use HRTF filter on all the reflections
  • Overkill, but makes a difference

• And we know how to do that now! :)
Virtual sound can be complicated

- Lots effects that combine
  - Not completely clear which are necessary
    - Depends on usage scenario
    - Also not fully clear how they all interact

- Still an open problem
  - But sounds pretty good as is
Surround sound

• Potentially simpler approach
  • Localization takes place using multiple speakers
  • Optionally one can use sophisticated filtering

• Common setups
  • 5.1 /7.1 sets
  • Stereo “surround”
    • Avoid like the plague!
    • Ruins stereo imaging

A virtual acoustic room setup
Theater surround sound

- Front channel for dialog
  - Ensures consistent localization

- Side and rear channels for FX
  - Also ambience sounds

- One of Dolby’s claims to fame
Recap

• Some of the basics of 3D perception

• HRTFs
  • How to measure them
  • How to use them

• Additional ties for virtual audio

• Surround sound
Reading material

• 3D Sound for Virtual Reality and Multimedia
  • http://human-factors.arc.nasa.gov/publications/
    Begault_2000_3d_Sound_Multimedia.pdf
Next lab

• Let’s make some 3D sounds!
  • Remember to bring your headphones/earphones
    • You won’t be able to hear the results otherwise