Lecture 11

Spatial Effects and Sound Localization

Why is it important?

- Important capability for survival
- Helps us separate simultaneous sound sources
  - Cocktail party, concert, …
- Entertainment!
  - Movie, game, virtual reality, …

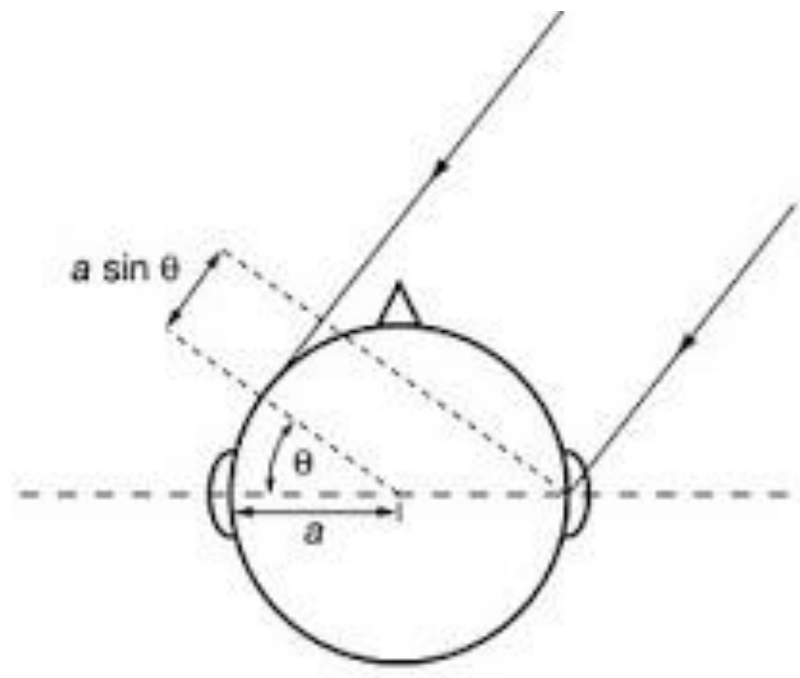
How do humans localize sound?

- Duplex theory by Lord Rayleigh (1907)
  - We localize sound sources based on the minor differences between sounds that the two ears receive

- Two main cues
  - Interaural Intensity/Level Difference (IID or ILD)
  - Interaural Time Difference (ITD)

Illustration

- Shadow effect to left ear (IID)
- Longer sound path to left ear (ITD)
- Azimuth θ: angle of sound source deviates from front

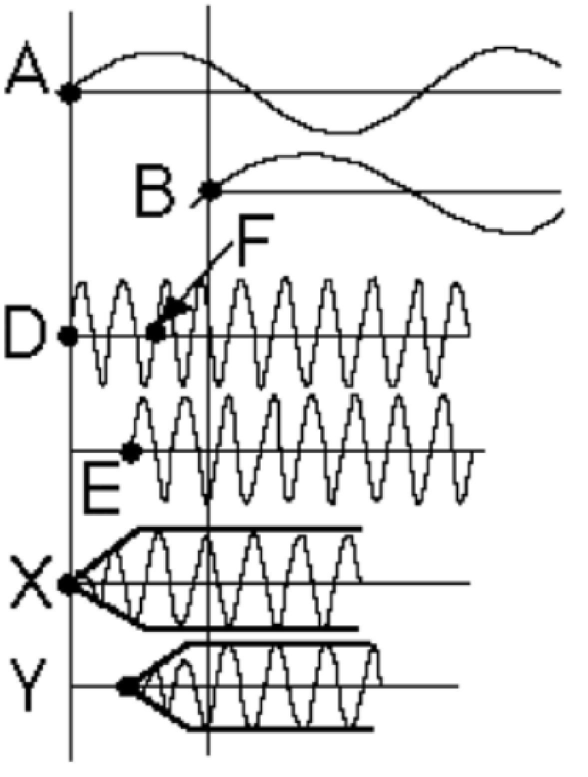


http://interface.cipic.ucdavis.edu/sound/tutorial/psych.html

IID and ITD

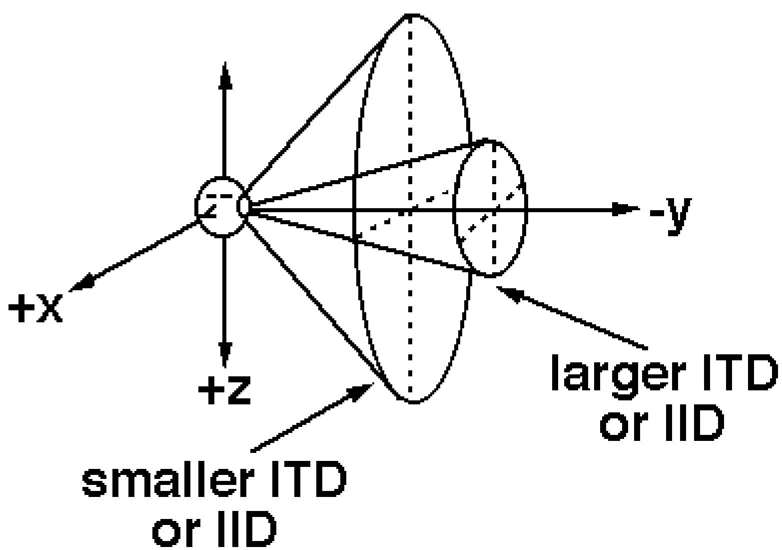
- IID is sensitive for high frequencies (>1.5 kHz)
  - Low frequencies (wavelengths longer than head diameter) pass mostly unharmed around the head
  - High frequencies get attenuated

- ITD is sensitive for low frequencies (<1.5 kHz)
  - ITD for high frequencies are still useful, but are performed on signal envelopes



Cone of Confusion

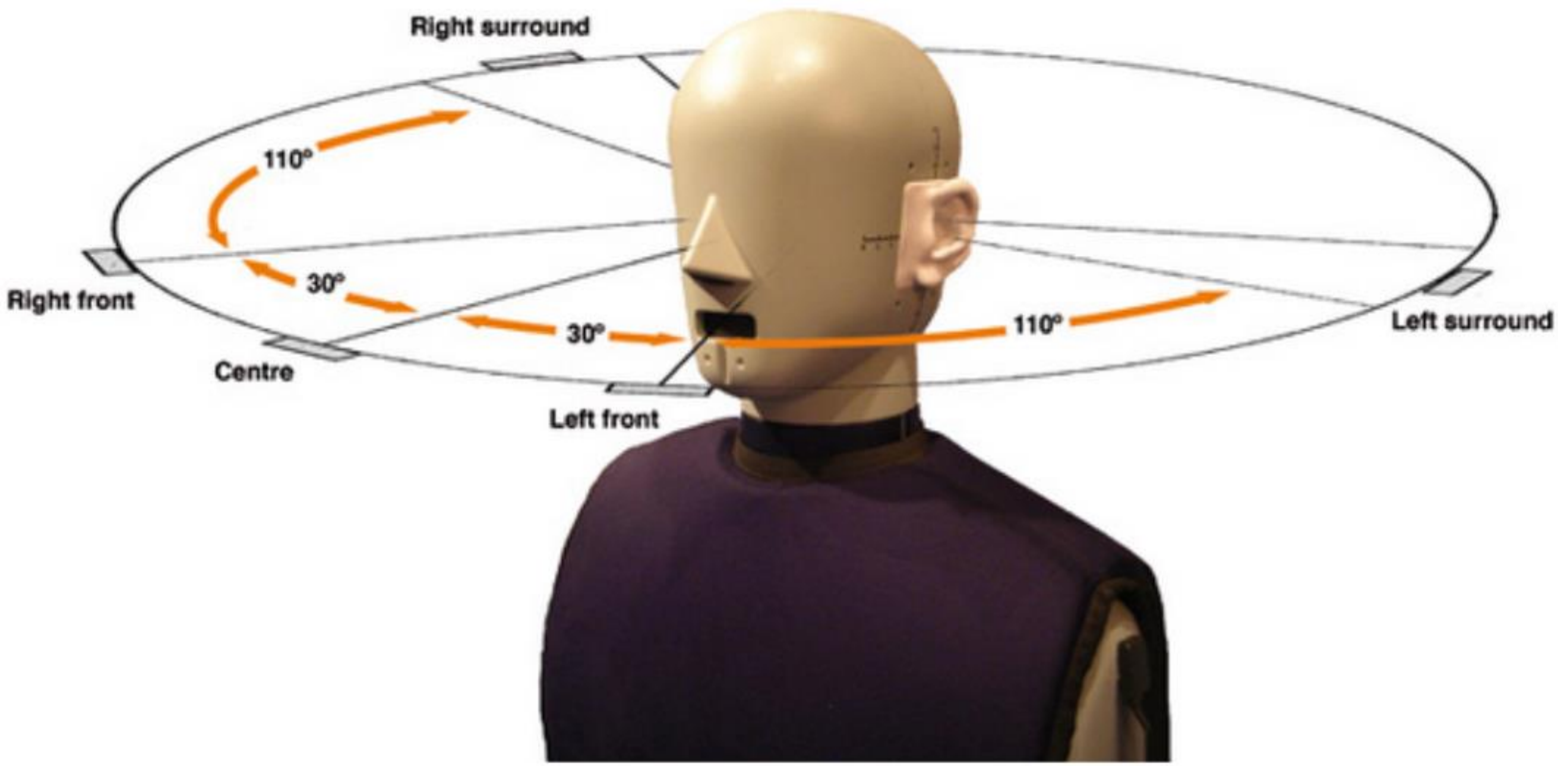
- A sound source located at any point on the surface of the cone produce the same ITD and IID
  - How do we humans resolve the confusion?



http://humansystems.arc.nasa.gov/groups/ACD/projects/dynamic_info.php

On Our Two Ears

- Why are they left/right, instead of up/down?
- Sensitive on azimuth, not sensitive on elevation



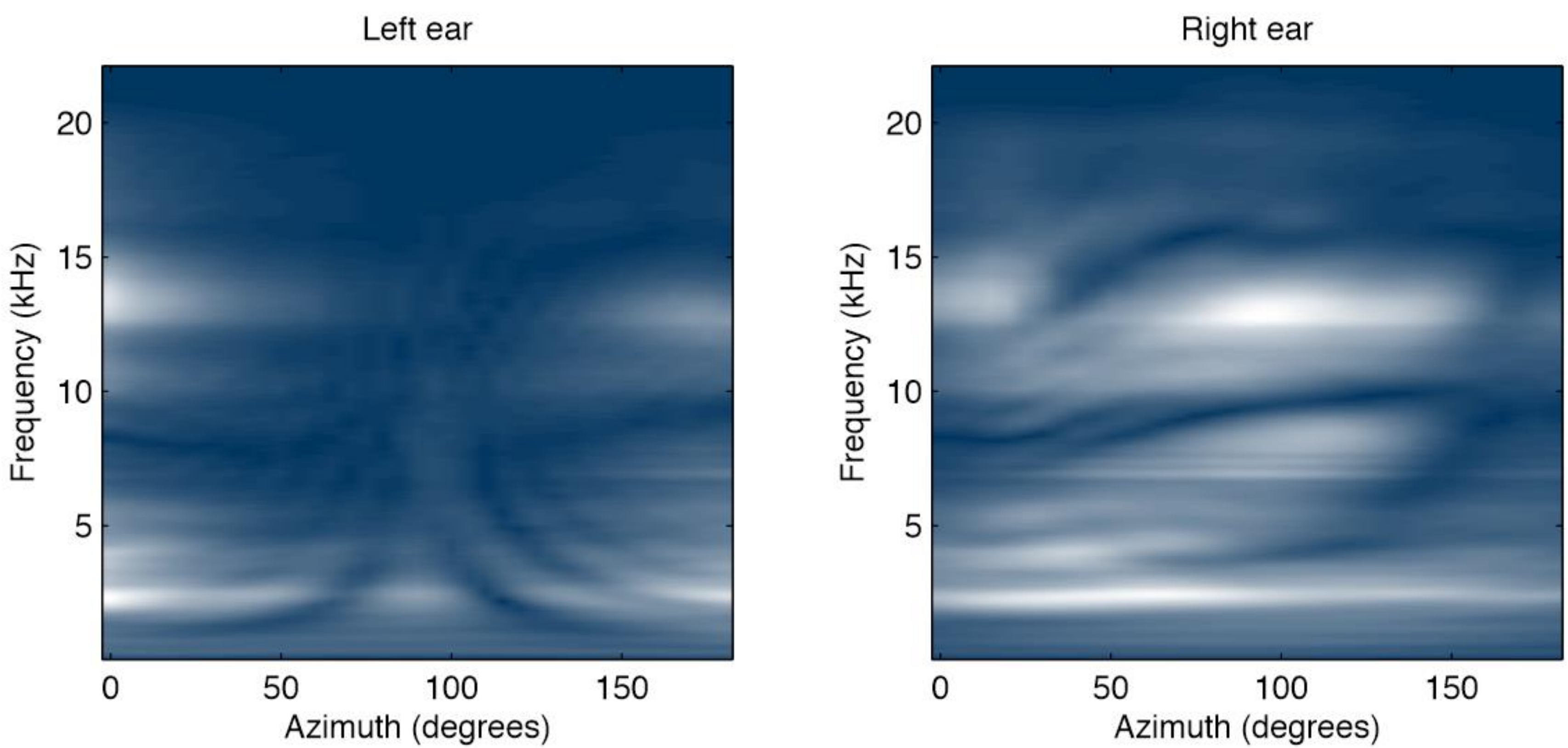
http://www.soundonsound.com/sos/jan08/articles/mp3surround.htm

Head Related Transfer Function (HRTF)

- IID and ITD are insufficient in modeling localization cues
- HRTF is a better model
  - View sound propagation from source to ears as a linear filtering process
  - Sound spectrum is modified by pinnae, head, shoulders, torso, etc., depending on the sound location
  - One transfer function for each location

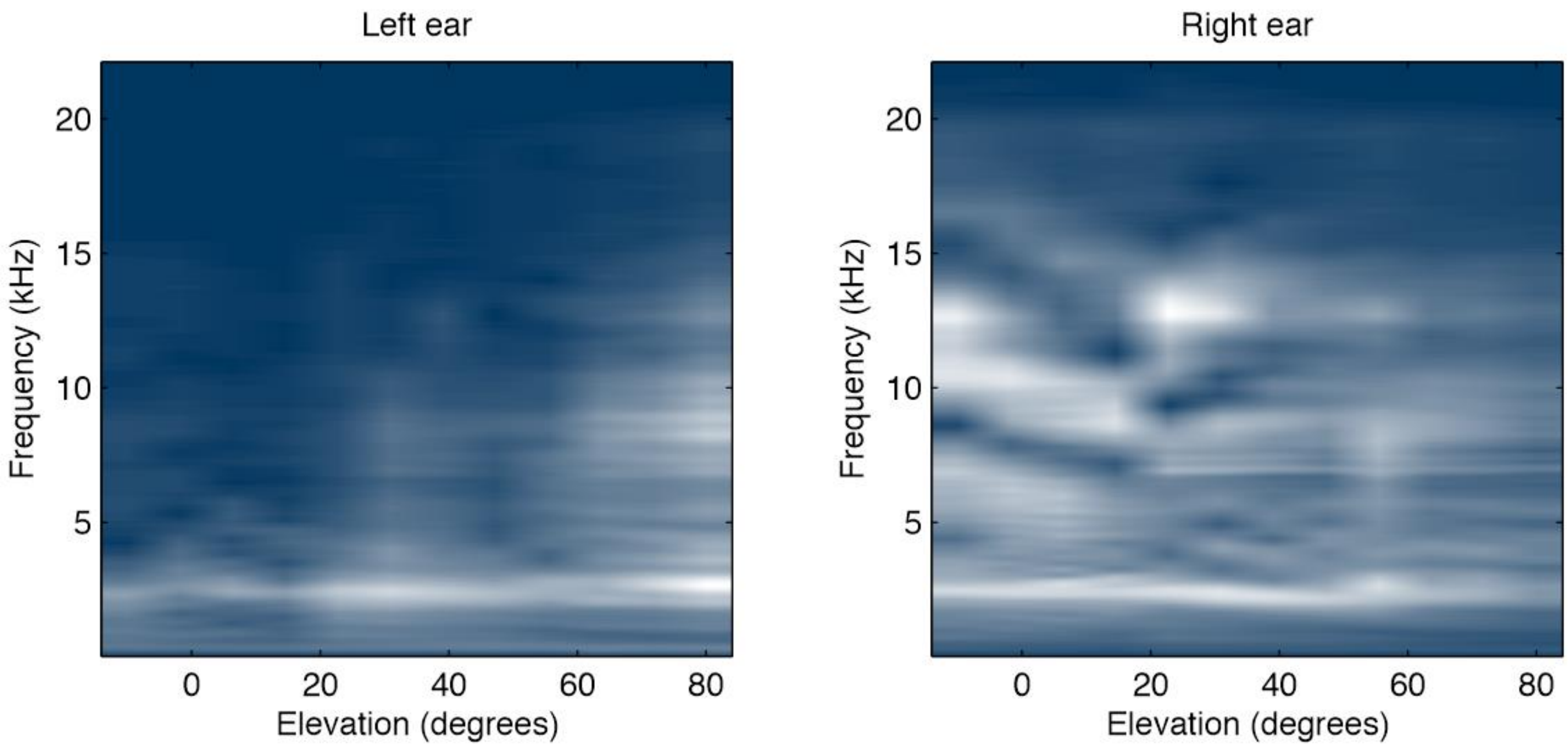
HRTF Illustration

- Move source from front to back (on the right)



HRTF Illustration

- Move source from down to up (on the right)



How to measure HRTF?

- Anechoic chamber
- Dummy head with microphones in ears
- Play sound at different locations
- Measure impulse responses

Two arc spherical positioning system
@ University of Oldenburg

Note: Different people have different body shapes, hence different HRTFs

Applications of Spatial Cues

- Synthesizing spatial effects with headphones
  - Panning
  - Adding time delay
  - Using HRTF

- Sound analysis
  - Sound localization
  - Localization and separation

Amplitude Panning

- Assigns different sound amplitudes for different channels, but the same delay, e.g.,

$$\mathrm{y[left]} = \mathrm{x}\cos\left(\lambda\,\frac{\pi}{2}\right),$$

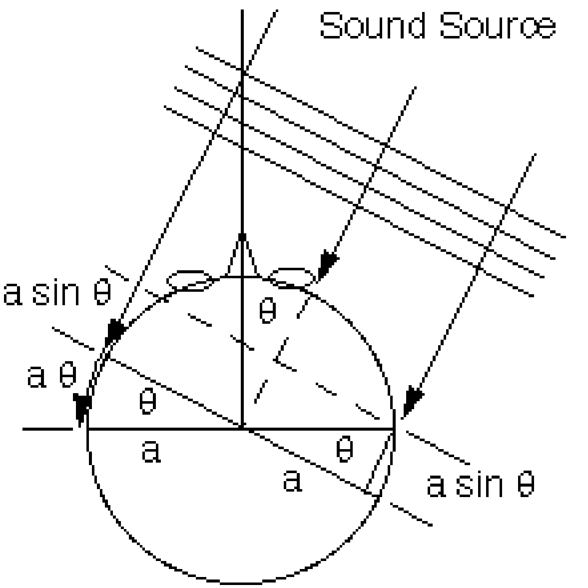
$$\mathrm{y[right]} = \mathrm{x}\sin\left(\lambda\frac{\pi}{2}\right),$$

where $0 \le \lambda \le 1$.

- Only utilizes IID but not ITD
- Simple but not very effective
  - Feels like sound is placed "inside of the head"

Add Time Delay

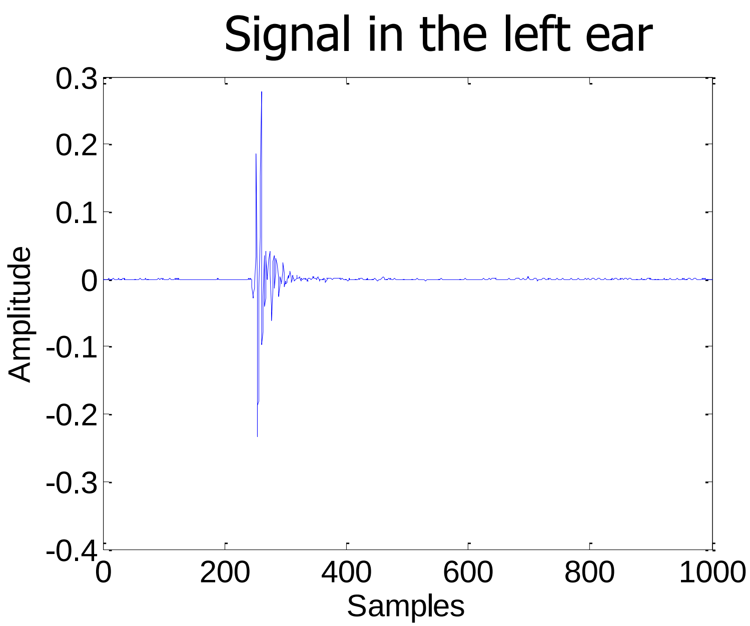
- Add delay based on azimuth, e.g.,
  - $\text{IID (linear amplitude)} = \frac{1-\frac{1}{3}\sin(\theta)}{1+\frac{1}{3}\sin(\theta)}$
  - $\text{ITD} = \frac{a}{c}(\theta + \sin\theta)$

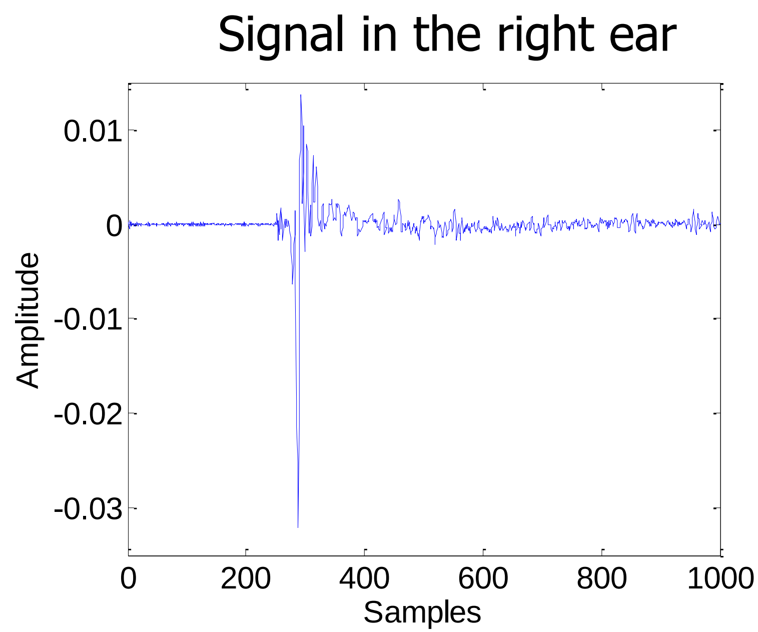


- More realistic than amplitude panning
- Does not consider spectral shaping

Using HRTF

- Convolving signal with corresponding head-related impulse responses (HRIRs)





- In which ear were the above impulse responses measured?

Using HRTF cont.

- Implement the convolution in real time for each channel
  - Each input sample can be viewed as a scaled impulse
  - The HRTF filter has a response to the impulse, i.e., the scaled HRIR
  - The response will affect current and future output signal
  - Simply maintain a buffer for the output signal to hold its current and future samples
  - Update the buffer every time an input signal sample comes
  - Output the current buffer sample
  - Increment the buffer pointer by 1

Issues when using loudspeakers

- Cross-talk: sound from each loudspeaker comes to both ears
  - Panning: comb filtering effects due to sound interference
  - Adding time delays: spreading percept
  - HRTF: cross-talk cancellation
    - Depends on the listener's position
    - Hard to compensate for multiple positions simultaneously

- Room effect: hard to compensate

Sound Localization

- Learn IID/ITD statistical models from training sound locations
- Calculate IID/ITD of sound
- Predict sound location using the statistical models

How to track a moving source?

- Naive way: predict location at each time frame independently

- Better way: consider location history using Kalman filter or hidden Markov models (HMM)
  - Assuming sound source moves continuously

- Additional way: Doppler effect

Localization for Source Separation

- Assumption
  - Sources are at different locations, i.e., having different IID/ITD.
  - Each time-freq bin of the mixture signal's spectrogram mainly comes from only one source.

- Method

DUET algorithm [Yilmaz & Rickard, 2004]

  - Calculate IID/ITD at each time-freq bin
  - Group time-freq bins according to their IID/ITD
  - Binary masking to separate source spectrogram

Anechoic Mixing Model

$$x_k(t) = \sum_{j=1}^{N} a_{kj} s_j(t - \delta_{kj}), \qquad k = 1,2$$

Attenuation coefficients

Time delays

- Without loss of generality, we can set $a_{1j} = 1$ and $\delta_{1j} = 0$ for all $j = 1, \dots, N$. And rename a_{2j} as a_j and δ_{2j} as δ_j, which are relative attenuation and time delay.
- Take STFT:

$$\begin{bmatrix} \hat{x}_1(\tau,\omega) \\ \hat{x}_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ a_1 e^{-i\omega\delta_1} & \dots & a_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} \hat{s}_1(\tau,\omega) \\ \vdots \\ \hat{s}_N(\tau,\omega) \end{bmatrix}$$

How to derive the mask?

$$\begin{bmatrix}\hat{x}_1(\tau,\omega)\\ \hat{x}_2(\tau,\omega)\end{bmatrix} = \begin{bmatrix}1 & \dots & 1\\ a_1 e^{-i\omega\delta_1} & \dots & a_N e^{-i\omega\delta_N}\end{bmatrix}\begin{bmatrix}\hat{s}_1(\tau,\omega)\\ \vdots \\ \hat{s}_N(\tau,\omega)\end{bmatrix}$$

- Assume that each t-f point contains only one source, and its IID and ITD correspond to the location of that source!

$$R_{21}(\tau,\omega) := \frac{\hat{x}_2(\tau,\omega)}{\hat{x}_1(\tau,\omega)} = a_j e^{-i\delta_j\omega}$$

If only source j is active at (τ,ω)

Group T-F Points

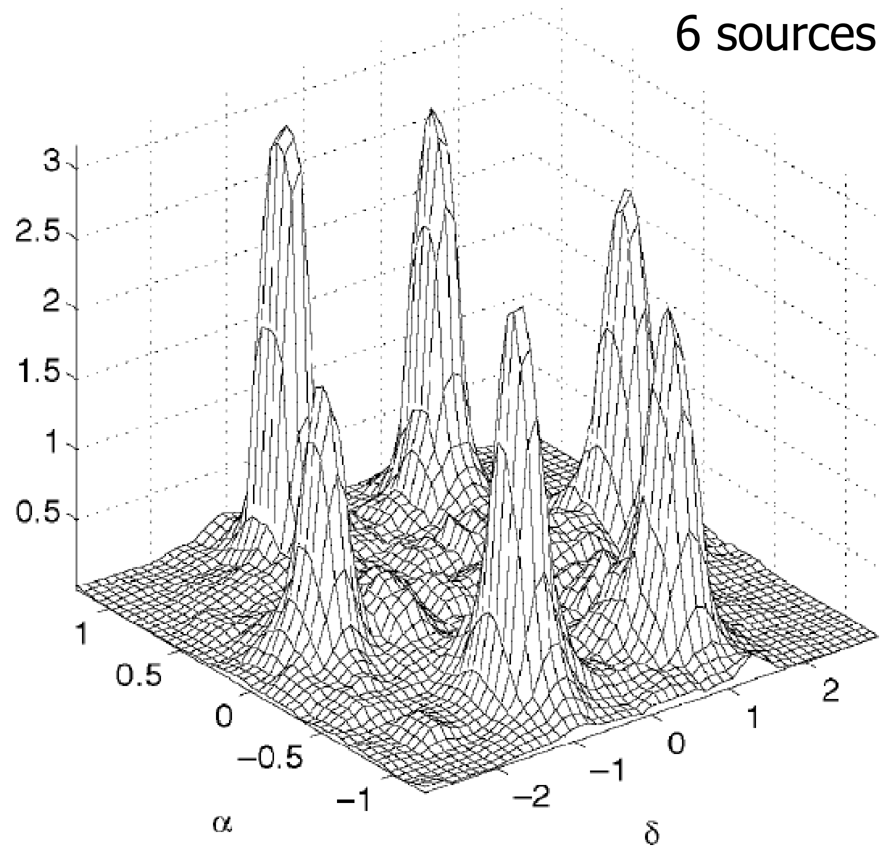
- Therefore, T-F points dominated by the same source have very similar IID and ITD.

$$\tilde{a}(\tau,\omega) := |R_{21}(\tau,\omega)|$$

$$\tilde{\delta}(\tau,\omega) := -\frac{1}{\omega}\angle R_{21}(\tau,\omega)$$

- Plot a 2-D histogram

6 sources



- Here we use symmetric attenuation for better numerical results

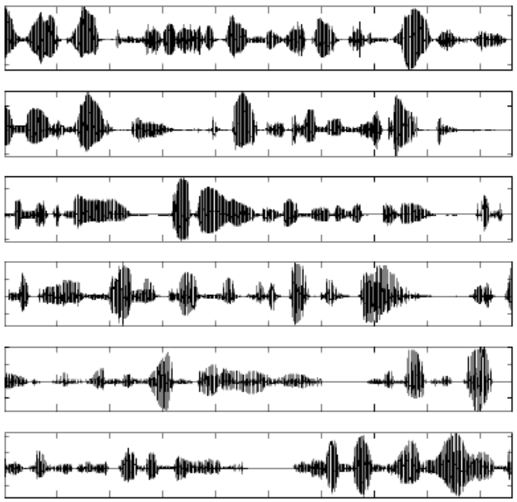
$$\alpha(\tau,\omega) := \tilde{a}(\tau,\omega) - \frac{1}{\tilde{a}(\tau,\omega)}$$

DUET Algorithm

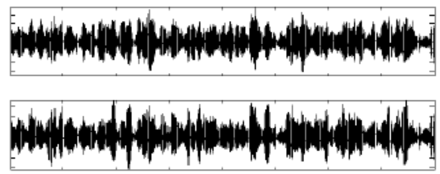
- 1) STFT on both channels
- 2) calculate DUET parameters (i.e. IID and ITD) for each T-F point
- 3) construct a 2-D histogram and locate peaks, where each peak will correspond to a source
- 4) for each peak, construct a binary mask by collecting T-F points whose DUET parameters are close to the peak
- 5) apply the mask to the mixture and do inverse-STFT

Experiments

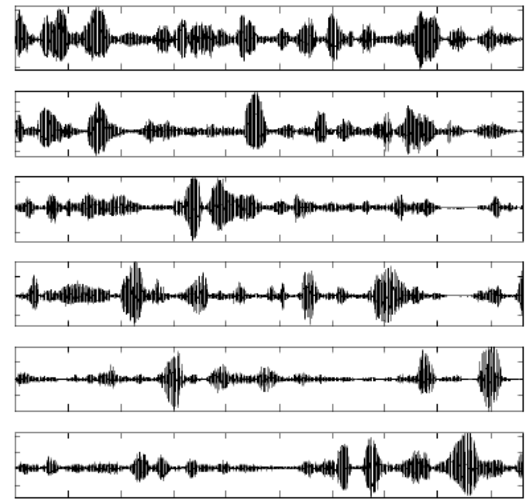
Speech sources



Artificial mixtures using anechoic mixing model



Separated by DUET



source	SIR in (dB)	SIR out (dB)	WDO DUET	WDO 0dB
s_1	-7.29	5.92	0.57	0.80
s_2	-7.29	5.24	0.55	0.78
s_3	-5.08	6.60	0.62	0.81
s_4	-9.29	5.35	0.56	0.69
s_5	-5.03	7.06	0.63	0.81
s_6	-9.28	5.47	0.55	0.66

Experiments in Real Environments

Anechoic room

test	SIR in (dB)	SIR out (dB)
M1 0^o	-2.72	13.67
F1 90^o	-2.05	7.96
M2 180^o	-4.37	13.32
F1 0^o	-9.77	7.97
M1 60^o	-4.30	7.16
F2 90^o	-3.77	5.99
M2 120^o	-5.60	7.05
F3 180^o	-8.59	8.53

Echoic room (reverberation time ~500ms)

test	SIR in (dB)	SIR out (dB)
M1 0^o	-5.20	5.38
M2 90^o	0.07	4.33
F1 180^o	-4.48	6.03

Histograms

Anechoic room

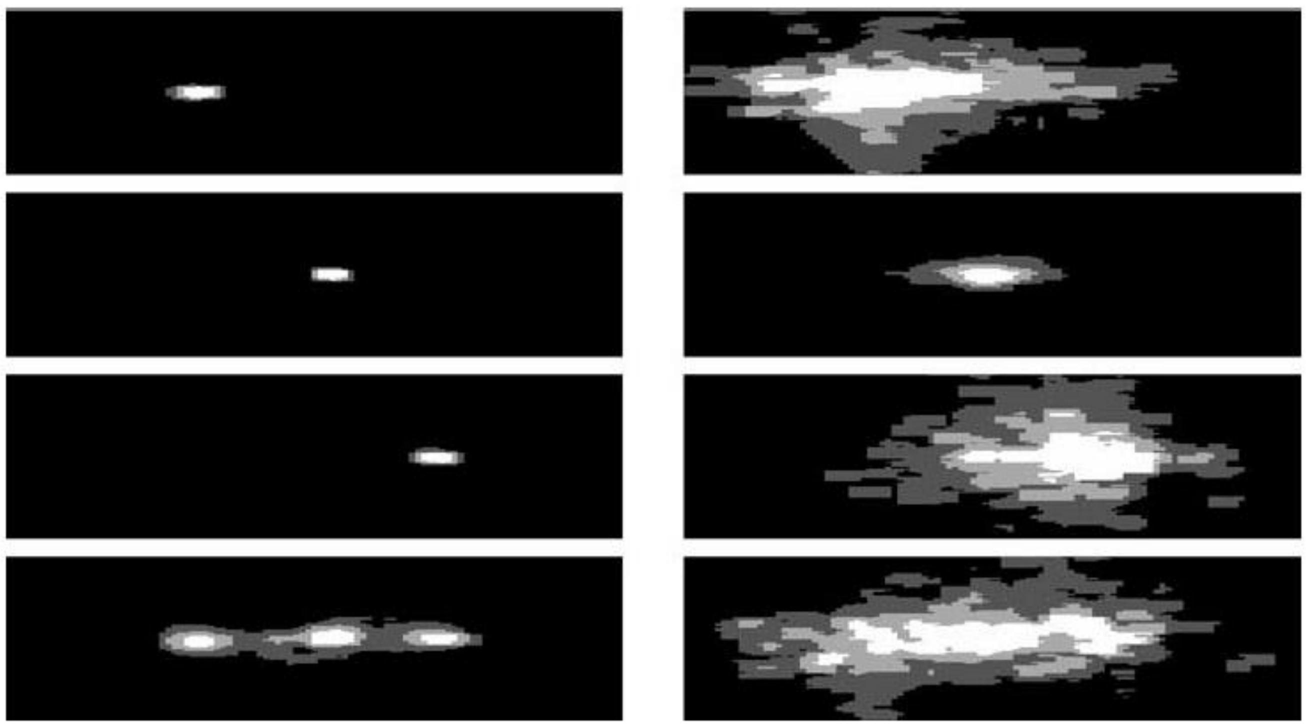
Echoic room (reverberation time ~500ms)

Source 1

Source 2

Source 3

Mixture



Questions

- How to improve this method in reverberant environments?
- How to separate signals of moving sources?
- What if sources move silently occasionally?