Fundamental and Technological Limitations of Immersive Audio Systems

Chris Kyriakakis

Presented by: Jay A. Patel
CS 598 KN, Spring 2005

Background

- Emerging media systems output mixed media in real time
- Immersive systems
  - Synthesize multi-modal perceptions unavailable in current physical environment
  - Seamless blend of visual and/or aural information
  - Significant area of research in imaging and video processing, but not audio

Immersive Audio

- Goal: Accurate spatial reproduction of sound
  - The human ear-brain can localize sounds in 3-D environment
    - Time of arrival differences: 7 microseconds
  - Perception based on multiple cues, including:
    - Level and time differences
    - Direction-dependent frequency-response based on
      - Reflections in outer ear, head, and torso
      - Or, Head-Related Transfer function (HRTF)
    - Timbre: differentiating sounds of same pitch and volume

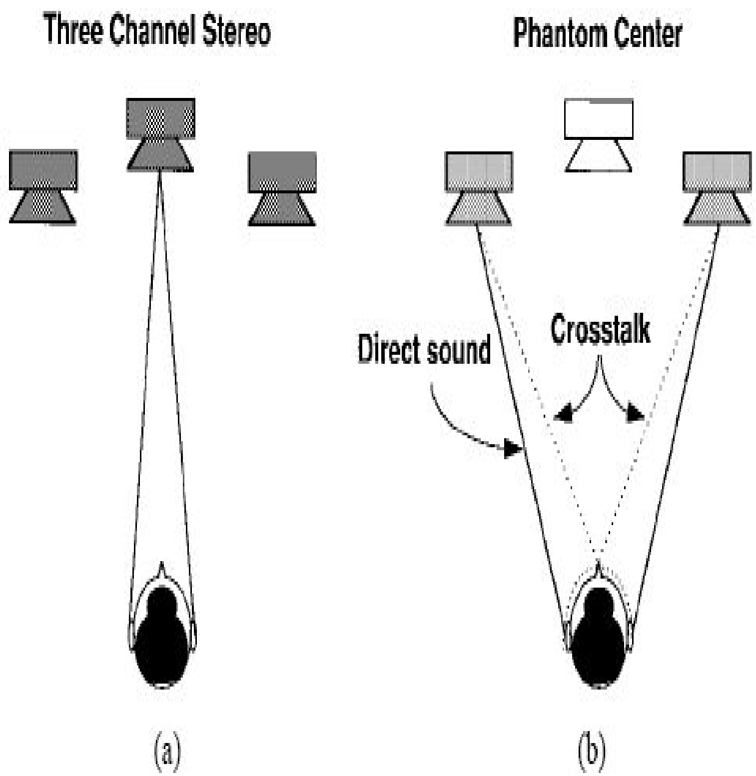
Limitations

- Physical Laws
  - Sound propagation and attenuation
  - Perception of spatial attributes:
    - Direction, distance, room space, source size, etc.
  - Complicated by need to alleviate HRTFs
- Technological Considerations
  - Auralization: numerical modeling of sound
  - Hardware limitations

“Suspension of Disbelief”

- Each listener judges sound quality subjectively
  - Apparent source width,
  - listener envelope,
  - clarity, and
  - vision (position)
    - Mismatch between aurally perceived and visually observed
      - Professional sound designers: 4-degree offset
      - Average person: 15-degree offse

Two-Channel Stereo



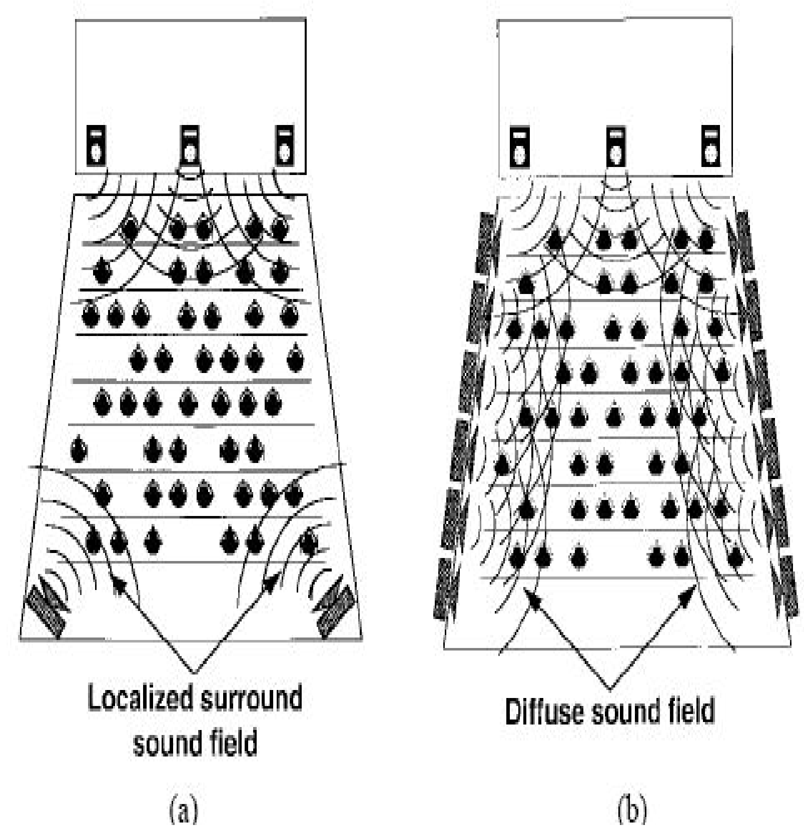
- Stereo: Greek word meaning "solid"
  - Two-channel comes from Phonographs
- Stereophony started in UK by Blumlein
- Feltcher, Steinber and Snow at Bell Labs in US
  - Actually, 3-channel
  - First demonstration in 1934: Philly Orchestra in DC

Quadraphonic Stereo

- Stereophonic falls short of true 3-D sound
- Added info about direct and reverberant sound fields
  - Clever encoding schemes (limitation: phonograph)
- Problems reproducing sound to the side
- Failure:
  - technical glitches
  - marketplace

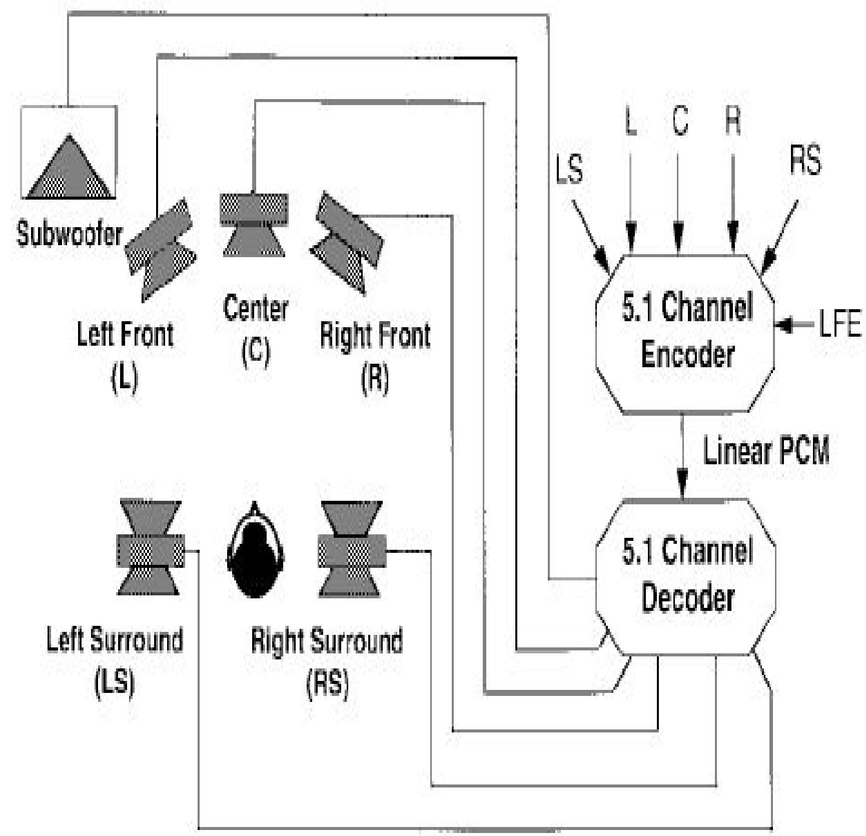
Multichannel Surround

- 20th Century Fox: 1950s
  - The Enemy: Television
- 3-channel stereophonic + two monophonic, rear speakers
- Perfect for people in center
- Problems for the off-centered
  - Array of side speakers
  - Until the 1970s



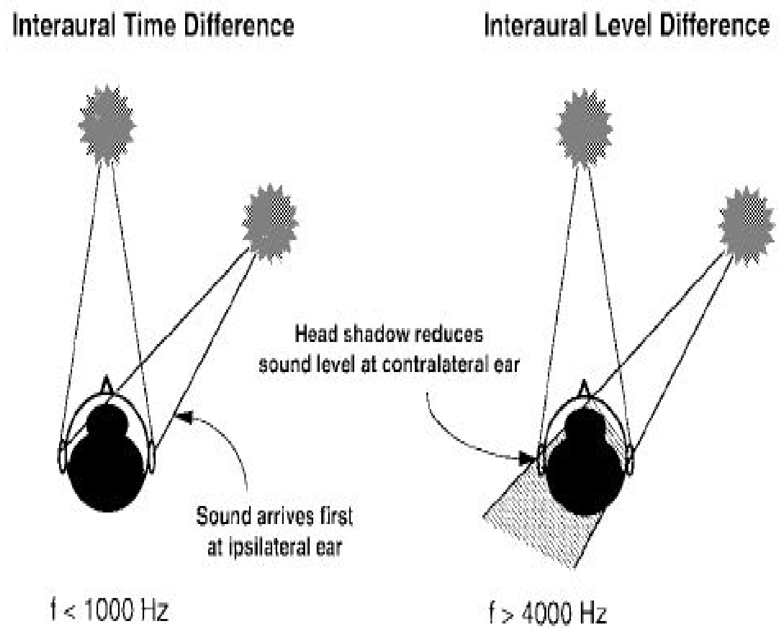
Multichannel Surround (Today)

- Dolby
  - Stereo (mid 1970s)
    - Encoded 4 channels into 2
    - L, C, R, Mono surround
  - Stereo Digital (1992)
    - No encoding
    - 5 discrete channels
    - L, C, R, IL, IR (LFE)
  - AC-3 compression
    - Added LFE: 1 – 20 Hz



Spatial Audio

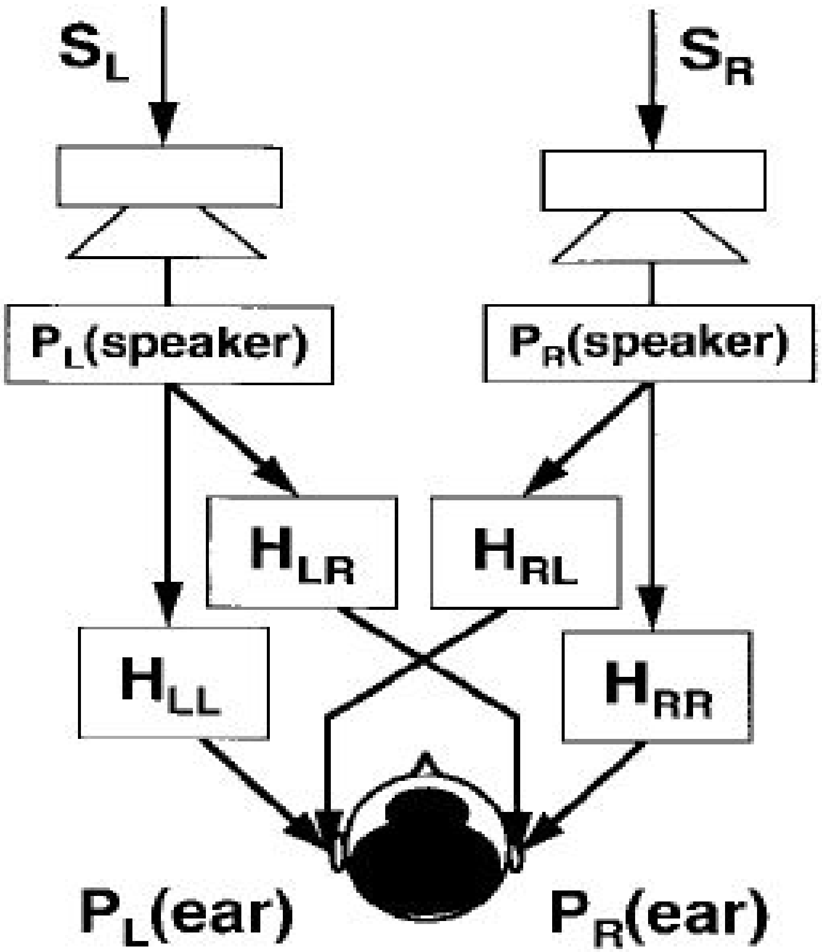
- Human hearing based on differences of time and level
- Horizontal Plane:
  - ITD: 20 Hz – 1 kHz
  - ILD: 4 kHz – 20 kHz
- Vertical Plane:
  - HRTFs
  - Individualism



3-D Audio: Additional Challenges

- HRTFs: distinct and specialized for each person
  - Averaging or modeling based on “good localizers” of sound
  - Extensive amount of lab experiments
- Computational Limitations
  - Avg. impulse duration: 3s
  - Sampled at 48 kHz
    - Requires 13 Gflops/channel
- Cross-talk cancellation

3-D Audio Rendering: Cross Talk Cancellation



$$P_L(\text{speaker}) = H_{LL}S_L + H_{RL}S_R$$
$$P_R(\text{speaker}) = H_{LR}S_L + H_{RR}S_R$$

$$P_L(\text{ear}) = P_L(\text{speaker})$$
$$P_R(\text{ear}) = P_R(\text{speaker}).$$

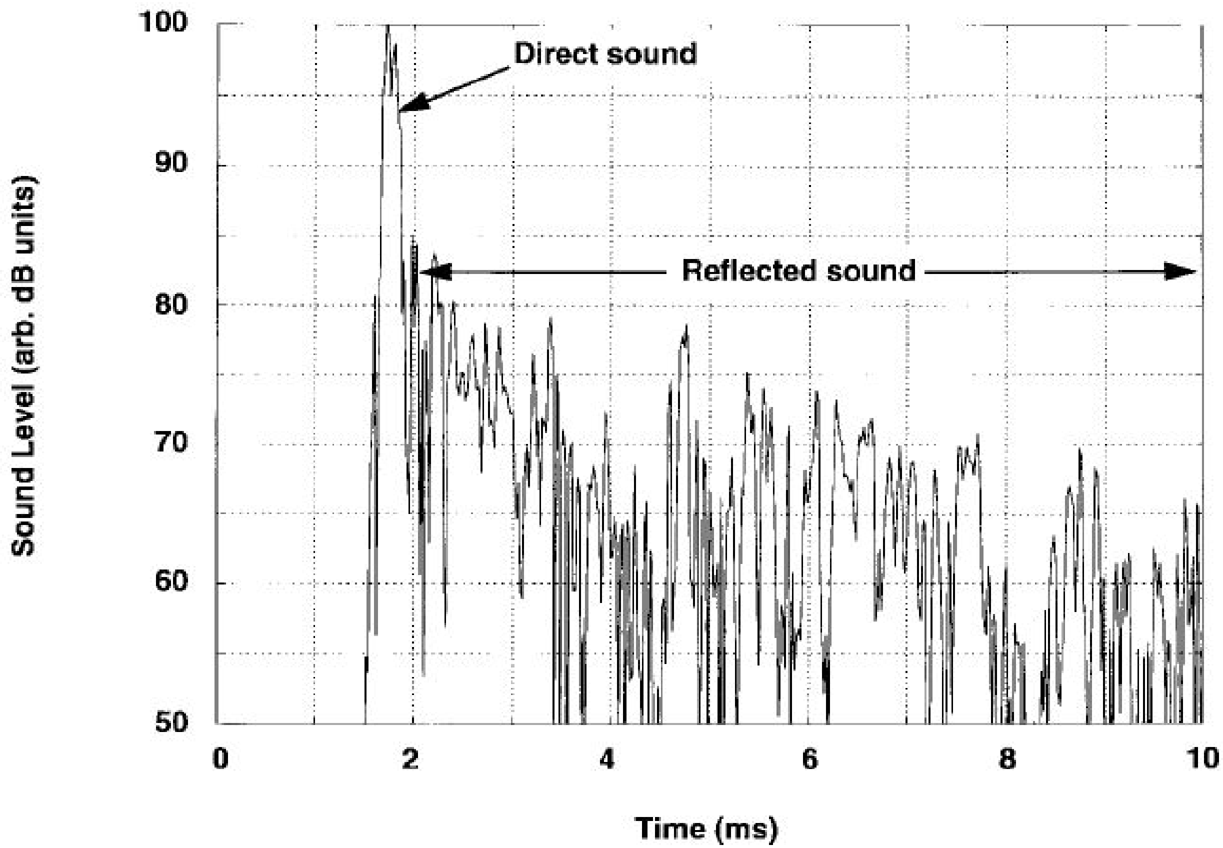
$$S_L = \frac{H_{RR}P_L(\text{ear}) - H_{RL}P_R(\text{ear})}{H_{LL}H_{RR} - H_{LR}H_{RL}}$$
$$S_R = \frac{H_{LL}P_R(\text{ear}) - H_{LR}P_L(\text{ear})}{H_{LL}H_{RR} - H_{LR}H_{RL}}.$$

Audio for Desktop Applications

- User-imposed limitations:
  - Two speakers (mostly)
- Problem: Small rooms
    - Early reflections: biggest source of errors
    - Maximum when difference is less than 15dB (within 15ms)
    - Solution: Near-field monitoring
      - Direct sound is dominant as users are close to speakers
      - Again, problems: Strong reflections from other close objects
- Problem: Low frequency anomalies

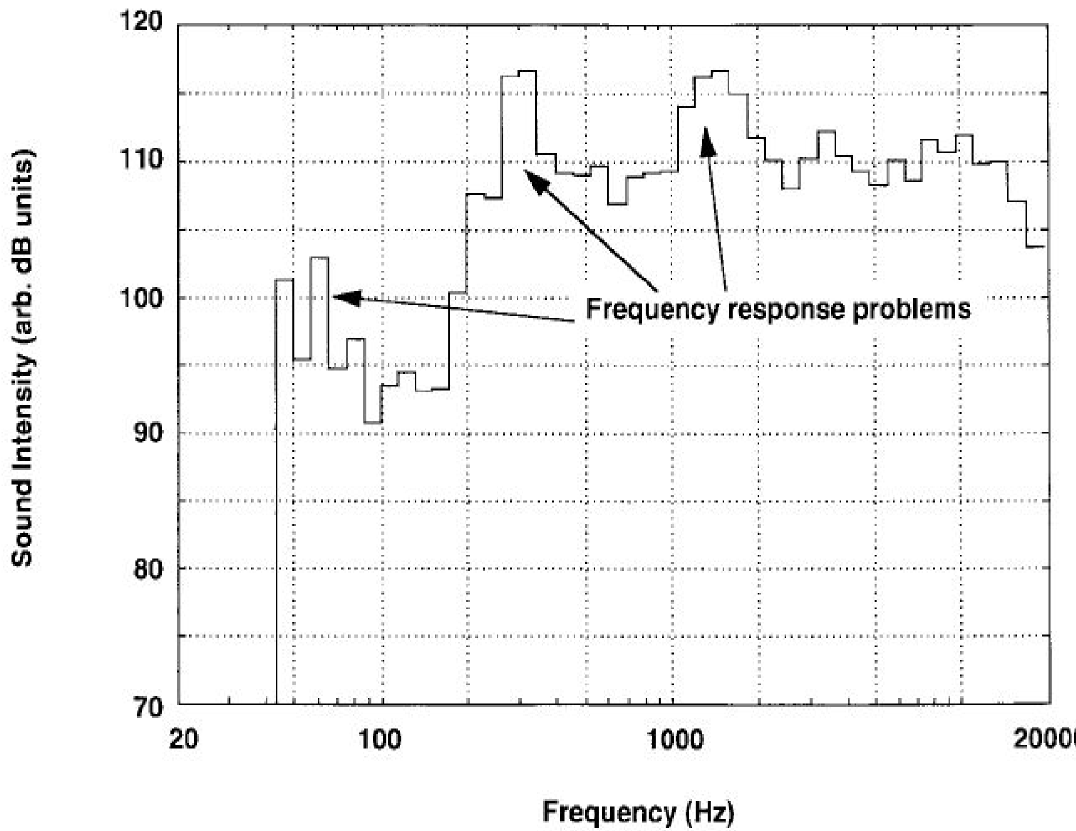
Small Room Reflection

- Noticeable when dB difference: < 15 dB



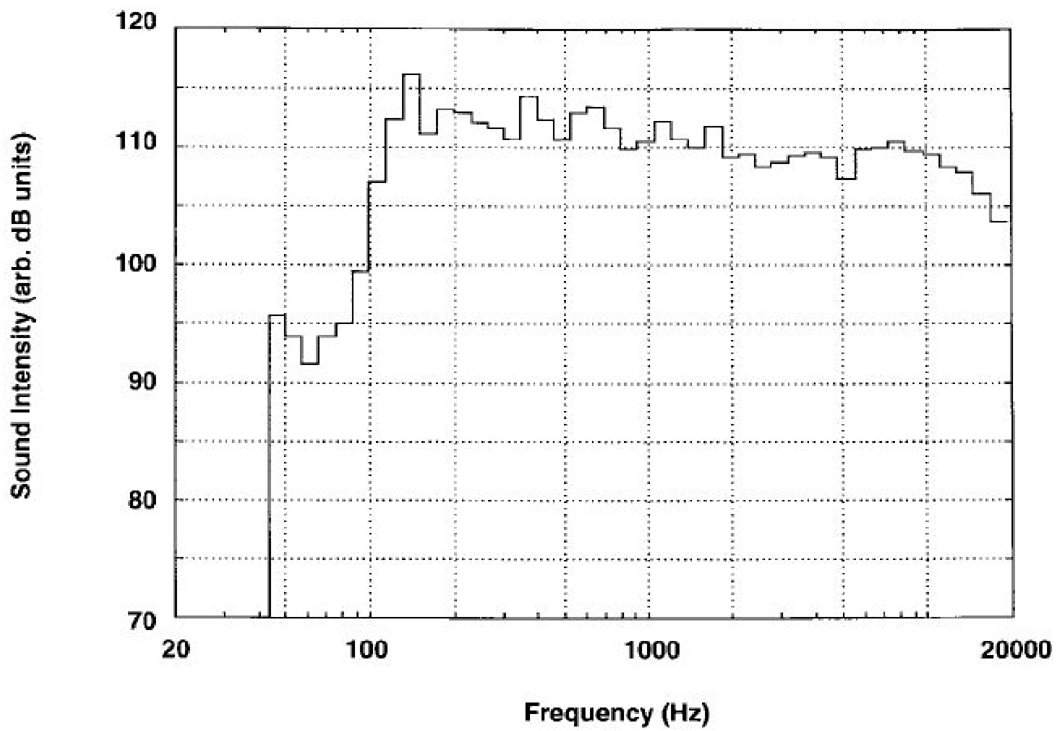
Frequency-Response Problem

- Low Frequencies: Physical size of room
- High Frequencies: Interaction with large objects



Direct-Path Dominant System

- Includes compensation for frequency anomalies
- Uses Sub-woofer (placed farther away)



Location Considerations

- Monitor = No center speaker
  - However, there exists exactly **one** “sweet spot”
  - Create a phantom image, originating from center
  - Problem: user moves
    - One of the two speaker dominates (Precedence effect)
    - For non-stationary user, system must know location

Vision-Based Solutions

- Use Computer Vision to detect user
  - Adapt the “sweet spot”
  - Monitor the user movement in the lateral plane