# Perceptual Spatial Audio Recording, Simulation, and Rendering

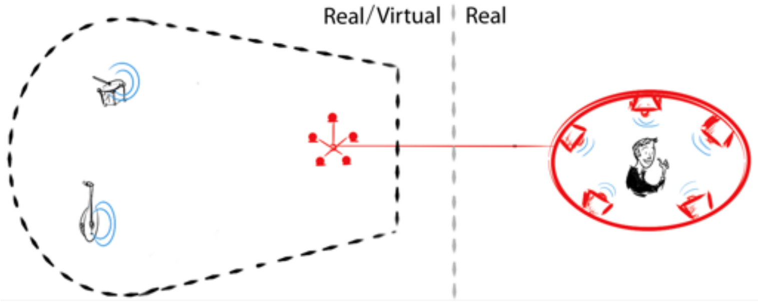
Enzo De Sena

Institute of Sound Recording, University of Surrey, UK

Apple R&D Talk, 13 Aug 2019

# Objective

- Making listener feel transported to a different auditory scene, which can be
  - a real recorded one (live music performance, sporting event..)
  - a virtual one (video games, VR/AR, architectural acoustics..)



# Outline

- Perceptual Soundfield Recording and Reproduction
  - Limitations of physical-based models
  - Localization uncertainty of phantom sources
  - Perceptual Soundfield Reconstruction
- Perceptual Simulation of Room Acoustics
  - Limitations of physical-based models
  - Scattering Delay Network
- Conclusions

# Acknowledgements

**Joint work with:**
Prof Zoran Cvetkovič (King's College London)
Prof Huseyin Hacihabiboglu (METU)
Prof Julius O. Smith (Stanford University)
Prof Toon van Waterschoot (KU Leuven)
Prof Marc Moonen (KU Leuven)
Dr Niccoló Antonello (KU Leuven)
Stojan Djordjevic (University of Surrey)
Ashley Andrew-Jones (University of Surrey)
Ege Erdem (METU)

**Funding from:**
EPSRC, EU Commission, FWO, KU Leuven research funds

# About this talk

- Interrupt me!
- Details and maths left to references (at the end)
- Demos after this talk

# Outline

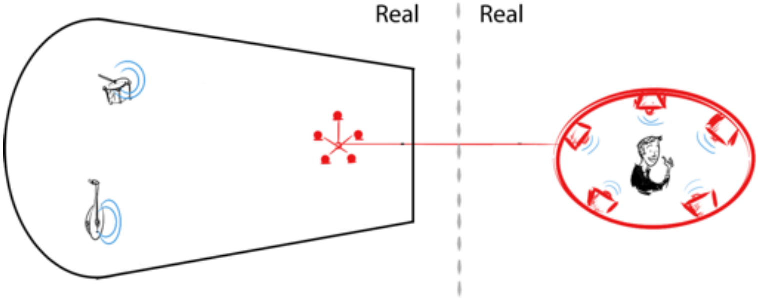
Perceptual Soundfield Recording and Reproduction
Limitations of physical-based models
Localization uncertainty of phantom sources
Perceptual Soundfield Reconstruction

Perceptual Simulation of Room Acoustics

Conclusions

# Perceptual Spatial Audio Recording

- Let's start from the case of a real sound scene to be recorded



# Physical and cross-talk cancellation methods

| | SFR | Multichannel | 2-Channel |
|---|---|---|---|
| **Channel count** | 50+ | $< 10$ | 2 |
| **Equipment Load** | High | Commercially viable | Low |
| **Psychoacoustics** | None | Required | Critical |
| **Sweet Spot** | Large | Medium, small group | Small, individual |

- Sound Field Reconstruction (SFR) provide mathematically elegant solution (e.g. HOA, WFS)...
  - but large number of loudspeakers: $r = \frac{c}{f} \frac{N}{2e\pi}$, e.g. $f = 10$ kHz, $r = 0.1$ m $\Rightarrow N = 56$
- 2-channel (cross-talk cancellation) methods, only two channels...
  - but small sweet spot (e.g. [Rose et al., 2002] report $\approx 3$ cm)
- We'll focus on multichannel systems with limited equipment load, which need to leverage somehow psychoacoustics effects

# Reproduction of plane waves

- Let's simplify: reproduction of a plane wave
- Assume for now that plane wave direction, $\theta_s$, is known
- Relevant case for spatial audio objects (MPEG-H)
- The plane wave could represent e.g. a single sound source or a wall reflection
- If we solve this, summation of plane waves trivial (linearity)

# Reproduction of plane waves

- Let's simplify: reproduction of a plane wave
- Assume for now that plane wave direction, $\theta_s$, is known
- Relevant case for spatial audio objects (MPEG-H)
- The plane wave could represent e.g. a single sound source or a wall reflection
- If we solve this, summation of plane waves trivial (linearity)

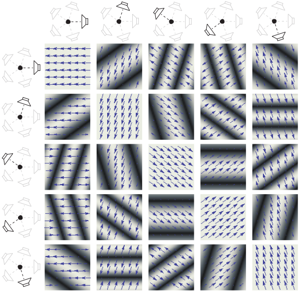
## Reproduced plane wave should be:

1. perceived in correct direction (low localization error)
2. easy to localize (low localization uncertainty)

- in the largest possible area (large sweet spot)

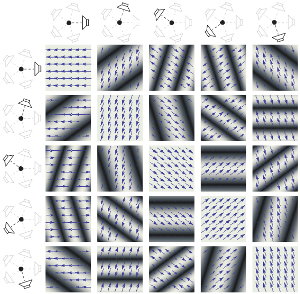
# How many loudspeakers to use to reproduce plane wave?

- **Question**: should we use $> 2$ loudspeakers for each source?
- Active intensity (AI) fields for plane waves



# How many loudspeakers to use to reproduce plane wave?

- **Question**: should we use $> 2$ loudspeakers for each source?
- Active intensity (AI) fields for plane waves

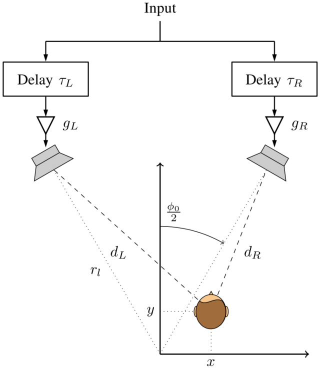


- Fluctuation speed depends on angle between loudspeaker pair
- **Answer**: use only the two loudspeakers closest to direction of plane wave [De Sena et al., 2013]
- This reduces problem to good ol' stereophonic reproduction

# Frequency-independent inter-channel differences

- What should we do with these two loudspeakers?

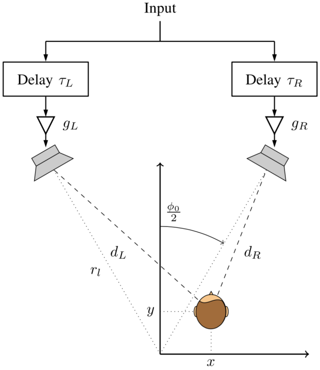
- Consider frequency independent inter-channel time differences (ICTD) and level differences (ICLD)
- ICTD/ICLDs lead to low coloration [Spors et al., 2013], which is most important attribute for sound quality [Rumsey et al., 2005]



# Frequency-independent inter-channel differences

- What should we do with these two loudspeakers?

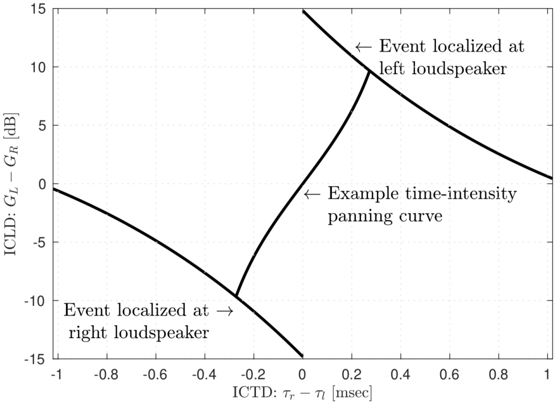
- Consider frequency independent inter-channel time differences (ICTD) and level differences (ICLD)
- ICTD/ICLDs lead to low coloration [Spors et al., 2013], which is most important attribute for sound quality [Rumsey et al., 2005]




- As long as ICTD below echo threshold, listeners will perceive a fused “phantom source” (summing localization effect)

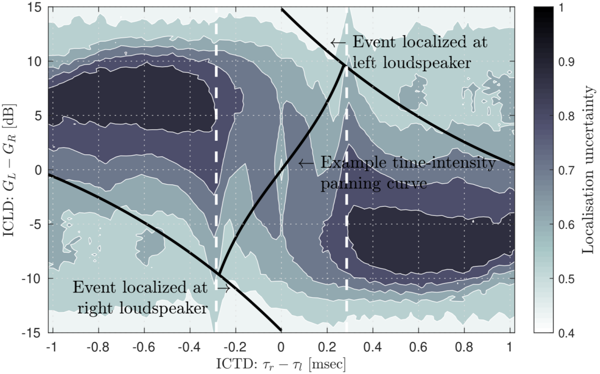
# Position of phantom source

- Position of phantom source depends on ICTD/ICLD pair
- Same position can be achieved with different ICTD/ICLD pair
- One can use e.g. intensity only (most commercial sound recordings), time only, or time-intensity



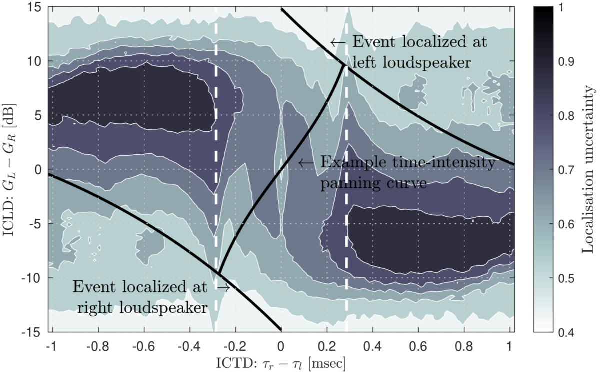
# Not all ICTD/ICLD pairs are created equal

- ICTD/ICLD pairs lead to different localization uncertainty
- Computational model in [De Sena et al., 2019]:



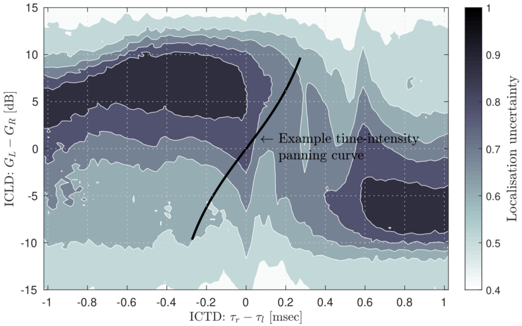
# Not all ICTD/ICLD pairs are created equal

- ICTD/ICLD pairs lead to different localization uncertainty
- Computational model in [De Sena et al., 2019]:




- Inconsistent ICTD/ICLD lead to high uncertainty
- The vertical bands correspond to cases where 2 replicates at one ear, but only 1 at the other

# Localization uncertainty in off-center positions




- Listener moves 10 cm to the right, then entire plot moves (approximately) to the right
- Now intensity methods lie in area with high uncertainty!
- Time-intensity largely avoids this area

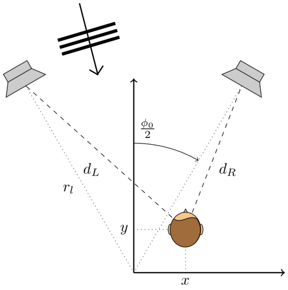
# What is happening?

- Useful to define "relative" ICTD/ICLD as observed by the listener:

$$\mathrm{RICLD} \approx \mathrm{ICLD} - \frac{x}{r_l}\frac{20\sin\left(\frac{\phi_0}{2}\right)}{\log_e(10)} ,$$

$$\mathrm{RICTD} \approx \mathrm{ICTD} - x\frac{2}{c}\sin\left(\frac{\phi_0}{2}\right) .$$

where $\phi_0$ base angle, $x$ lateral displacement and $c$ speed of sound



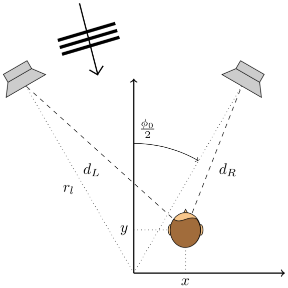
# What is happening?

- Useful to define "relative" ICTD/ICLD as observed by the listener:

$$\text{RICLD} \approx \text{ICLD} - \frac{x}{r_l}\frac{20\sin\left(\frac{\phi_0}{2}\right)}{\log_e(10)} \ ,$$

$$\text{RICTD} \approx \text{ICTD} - x\frac{2}{c}\sin\left(\frac{\phi_0}{2}\right) \ .$$

  where $\phi_0$ base angle, $x$ lateral displacement and $c$ speed of sound




- E.g. consider ICTD = 0 ms and ICLD = 5 dB (left leading)
- RICTD = −0.29 and RICLD = 4.78, which are contradicting
- Adding a small ICTD will delay the onset of contradicting RICTD/RICLD pairs

# Parametrization of ICTD (time-delay microphone array)

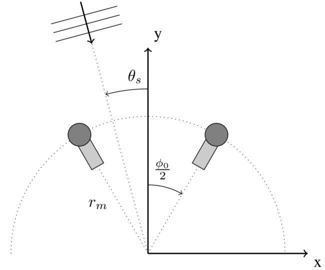
- Convenient now to specify ICTD and ICLD functions of $\theta_s$, including a parameter taking into account how much we rely on ICLD compared to ICLD (time-intensity trade-off)

# Parametrization of ICTD (time-delay microphone array)

- Convenient now to specify ICTD and ICLD functions of $\theta_s$, including a parameter taking into account how much we rely on ICLD compared to ICLD (time-intensity trade-off)

- Let the ICTD be defined according to the delay that would be observed on two spatially separated microphones as in figure:

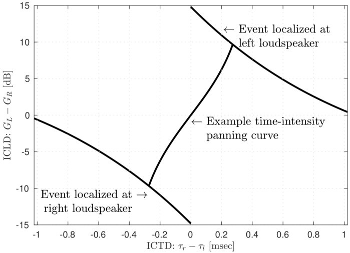
$$\text{ICTD}(\theta_s, r_m) = 2\frac{r_m}{c}\sin\left(\frac{\phi_0}{2}\right)\sin\theta_s$$




where $r_m$ is the array radius

- This parametrization is convenient since it allows to easily extend to the case of recording with circular arrays

# Parametrization of ICLDs

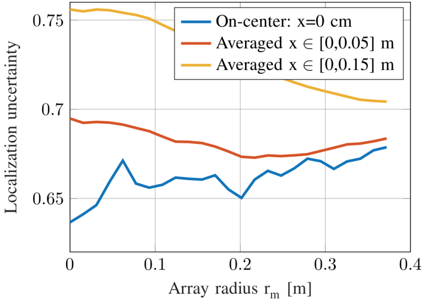



- Psychoacoustic curves give only extreme positions
- Could use different curves, for instance [De Sena et al., 2013]:

$$\mathrm{ICLD}(\theta_s, r_m) = 20 \log_{10} \frac{\sin\left(\frac{\phi_0}{2} + \beta(r_m) + \theta_s\right)}{\sin\left(\frac{\phi_0}{2} + \beta(r_m) - \theta_s\right)}$$

  where $\beta(r_m)$ is a parameter used to fit the extrema
- With this parametrization, a higher $r_m$ leads to more reliance on ICTDs and lower ICLDs

# Localization uncertainty as a function of array radius




- Larger radii lead to:
  - higher uncertainty for observer in the center
  - lower uncertainty for observer away from the center
- Help reconcile long-stanging debate between academia (preferring intensity methods) and sound engineering community (also using time-intensity methods)

# Choosing array radius parameter

- Trade-off between center and off-center
- If we don't know how far the listener will move, then avoid vertical bands mentioned before, which leads to

$$r_m = r_h \frac{\cos\left(\theta_e - \frac{\phi_0}{2}\right) + \frac{\phi_0}{2} + \theta_e - \frac{\pi}{2}}{2\sin^2\left(\frac{\phi_0}{2}\right)}$$

  where $\theta_e$ is angle of ear and $r_h$ is head radius
- Interestingly, larger head, means larger array!
- Examples:
  - $\phi_0 = 60^\circ$, $r_h = 9$ cm and $theta_e = 100^\circ$, then $r_m = 0.19$ cm
  - $\phi_0 = \frac{360^\circ}{5} = 72^\circ$, $r_h = 9$ cm and $theta_e = 100^\circ$, then $r_m = 0.16$ cm.

# More complex situations

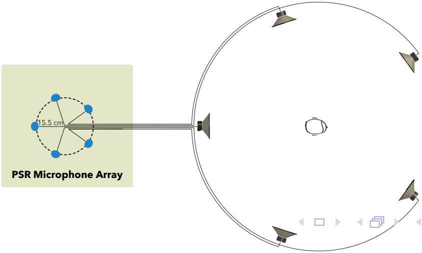
- So far we assumed we know the direction of the plane wave
- Possible approach is to estimate direction of arrival (DOA) and then artificially add ICTD/ICLD
- If multiple incoming waves, can estimated DOAs in time windows (see e.g. Dirac/SDM/SIRR)

# Perceptual Soundfield Reconstruction

- Another approach is to connect each microphone with loudspeaker
- Design the microphone directivity pattern to approximate ICLD($\theta_s, r_m$) [De Sena et al., 2013]
- This makes DOA estimation unnecessary!

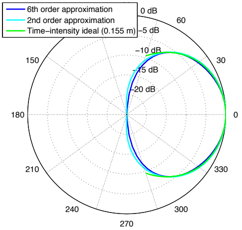
## Perceptual Soundfield Reconstruction (PSR) Array

- 5 channels, uniformly distributed, 15.5 cm radius (optimal according to



# Microphone directivity that approximates ICLD($\theta_s$, $r_m$)

- First-order microphones (e.g. cardioid, hypercardioid) not sufficiently directive for this purpose
- Second-order already sufficient (e.g. differential microphone array [De Sena et al., 2011])




## Results of PSR formal listening experiments:

- Comparable performance in the center of the array...
- but larger sweet-spot

# PSR Extensions

- PSR recently extended to third dimension using extrapolation from Eigenmike [Erdem et al., 2019]
- Time-intensity in the vertical dimension leads to a perceived improvement in stability of sweet spot [Andrew-Jones, 2019]

# Outline

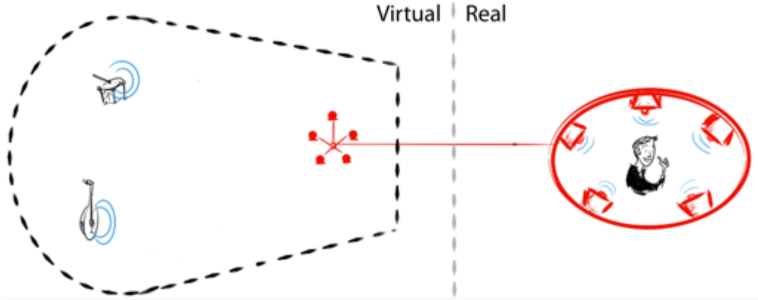

Perceptual Soundfield Recording and Reproduction

Perceptual Simulation of Room Acoustics
- Limitations of physical-based models
- Scattering Delay Network

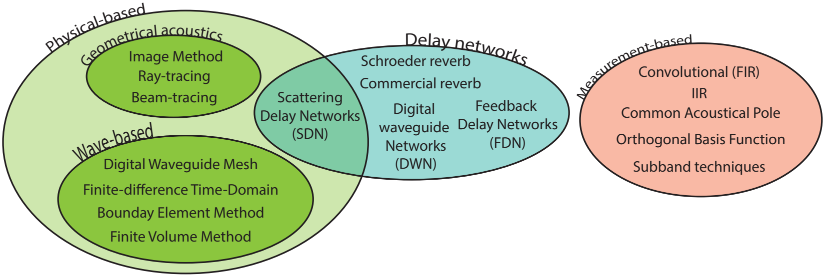
Conclusions

# Perceptual Simulation of Room Acoustics

1. Simulate virtual room acoustics
2. *Virtual* recording and *real* reproduction (simulate microphone array as described in first part of talk)



# Overview




- Overview of more than 50 years of room acoustic simulation in [Välimäki *et al.*, 2012], [Välimäki *et al.*, 2016] and [Hacıhabiboğlu *et al.*, 2017]
- Wave-based models are the most accurate ones

# Rendering of dynamic scenes with wave models

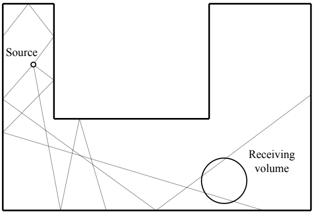
- In a complete wave model of a room:
  - sources and listeners can be moved
  - *spatialized* using microphone arrays or "virtual dummy head"

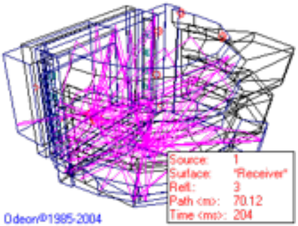
## Example: How expensive is a wave-based model?

- Audio bandwidth = 20 kHz ≈ 1.27 cm wavelength
- Spatial samples every 0.63 cm or less
- 3.65 × 5.8 × 2.4 m room requires > 200 million grid points
- 3D finite difference model requires one multiply and 6 additions per grid point ⇒ 70 billion FLOPS at $F_s = 50$ kHz
- 30 × 15 × 6 m concert hall requires > *3 quadrillion FLOPS*

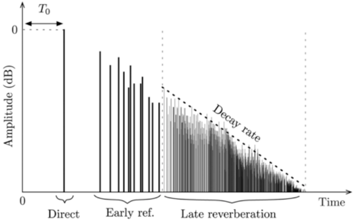
# Geometric Models

- Geometric acoustics models have lower complexity
- Source emits rays in all directions
- Specular reflections (diffraction also possible)
- Build impulse response by recording time and amplitude at receiver
- Choice of receiver size and number of rays is critical



# Room Impulse Response (RIR)




## RIR components:

- Direct line-of-sight
- Early reflections: relatively sparse first echoes
- Late reverberation: so densely populated with echoes that it is best to characterise the response *statistically*.

# Rendering of dynamic scenes with geometric models

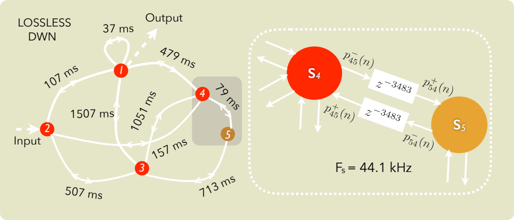
- When source moves recalculate RIR
- Still need to run a convolution with anechoic sound sample

Example:

- $T_{60} = 2$ s, $F_s = 50$ kHz: convolution requires 5 *billion* FLOPS
- Three sources and two listening points (ears) $\Rightarrow$ 60 billion FLOPS
- 20 dedicated CPUs clocked at 3 Gigahertz
- FFT convolution is faster, if throughput delay is tolerable (and there are low-latency algorithms)

- If physical accuracy not needed, perceptual methods provide better option

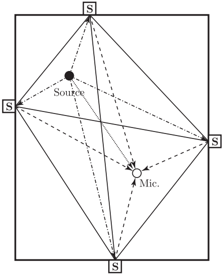
# Digital waveguide networks (DWN)




- Network of bi-directional delay lines connected at scattering junctions [Smith, 1985]
- Can be interpreted as network of acoustic tubes
- **Question:** How to set parameters (delay line lengths, network connections, scattering matrix..)?

# Scattering delay network (SDN) [De Sena et al., 2015]

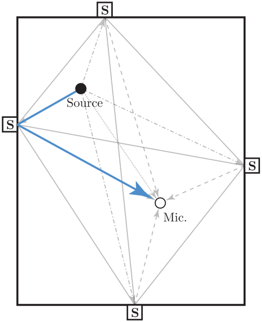
- Design DWN based on characteristics of a physical room



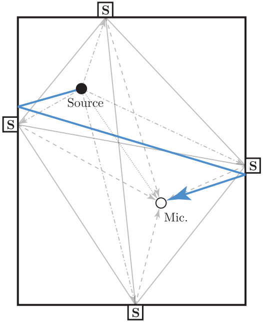

- Position nodes at first-order reflection points
- Fully connected DWN network
- Mono-directional lines for source-junction and junction-mic

# SDN: approximation of geometric acoustics

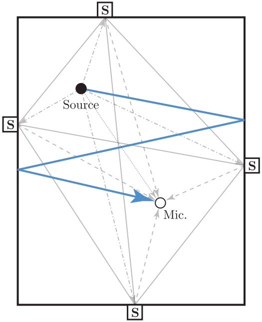
- Correct rendering of LOS and first-order reflections in time, amplitude and direction
- Approximation of second and higher-order reflections, less important perceptually




I-order reflection



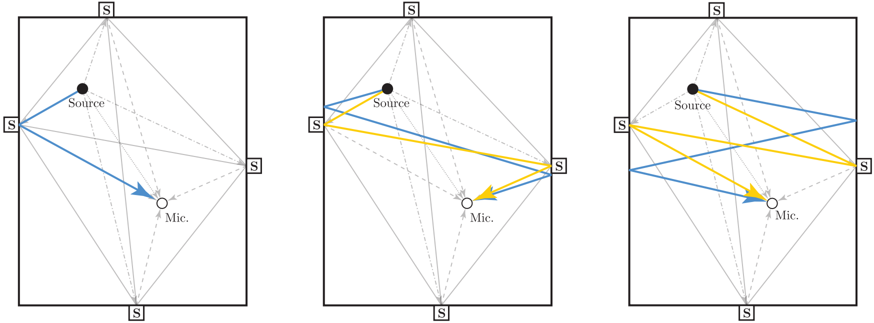

II-order reflection




Another II-order reflection

# SDN: approximation of geometric acoustics

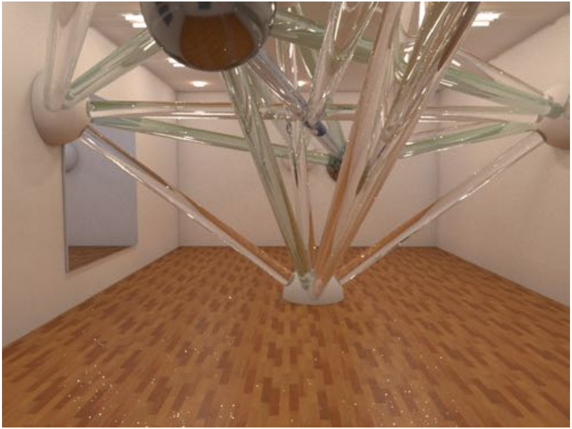
- Correct rendering of LOS and first-order reflections in time, amplitude and direction
- Approximation of second and higher-order reflections, less important perceptually




I-order reflection II-order reflection Another II-order reflection

# SDN: alternative interpretation

- Can also be interpreted as model of network of acoustic tubes



# Advantages

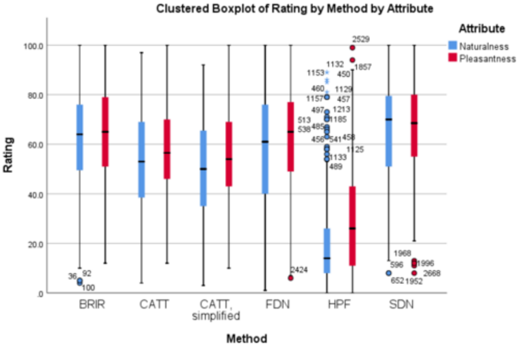
- Less resources spent for less important part of room impulse response (late reverberant tail)
- Also, not shown here:
  - similar frequency-dependent RT60 to full-scale models
  - similar echo density to full-scale models
  - sufficient modal density
  - axial resonant modes of room well approximated
- Orders of magnitude faster than convolution (alone!)
- All parameters of model derived from physical properties

## Advantages w.r.t. other delay networks:

- No need for hands-on parameters tuning
- Physical interpretation $\Rightarrow$ spatialisation possible, e.g. using microphone array as defined in the first part of the talk

# Perceptual evaluation [Djordjevic, 2019]

- Headphone-based (binaural) comparison (28 subjects)
- Higher pleasantness ($p < 0.001$) and naturalness ($p < 0.001$) than comparable delay-network based method



# Outline

Perceptual Soundfield Recording and Reproduction

Perceptual Simulation of Room Acoustics

Conclusions

# Conclusions

- Physical methods for spatial audio require significant resources
  - Recording and reproduction: many loudspeakers
  - Room Acoustics Simulation: high computational complexity
- Known perceptual effects allow to reduce requirements
  - Recording and reproduction: exploit summing localization effect and small ICTDs to achieve larger sweet spot
  - Room Acoustics Simulation: spend more resources for important perceptual features

Thanks for your attention!
(demos to come)

Thanks for your attention!
(demos to come)

Questions?

# Further Reading

**Spatial Sound Overview**

H. Hacihabiboglu, E. De Sena, Z. Cvetkovic, J. Johnston, J. O. Smith III, "Perceptual Spatial Audio Recording, Simulation, and Rendering: An overview of spatial-audio techniques based on psychoacoustics," IEEE Signal Processing Magazine, 34(3), 36-54, 2017. F. Rumsey, *Spatial audio,* Focal Press, 2012.

S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, F. Zotter "Spatial sound with loudspeakers and its perception: A review of the current state," *Proc. IEEE*, **101**(9):1920–1938, 2013.

**Physical Sound-Field Reconstruction**

Mark A Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.,* **53**(1):1004–1025, 2005.

J. Daniel, "Spatial sound encoding including near field effect," *Proc. AES 23rd Int. Conf.,* paper #16, 2003.

T. Betlehem, T. D. Abhayapala, "Theory and design of sound field reproduction in reverberant rooms," *J. Acoust. Soc. Amer.,* **117**(4):2100–2111, 2005.

M. Kolundzija, C. Faller, M. Vetterli, "Reproducing Sound Fields Using MIMO Acoustic Channel Inversion," *J. Audio Eng. Soc.*, **59**(10):727–734, 2011.

A. J. Berkhout, D. de Vries, P. Vogel, "Acoustic Control by Wave Field Synthesis," *J. Acoust. Soc. Amer.,* **93**(5):2764–2778, 1993.

E. Hulsebos, D. de Vries, E. Bourdillat, "Improved microphone array configurations for auralization of sound fields by wave-field synthesis," *J. Audio Eng. Soc.* **50**(10):779-790, 2002.

J. Ahrens, S. Spors, "A Modal Analysis of Spatial Discretization of Spherical Loudspeaker Distributions Used for Sound Field Synthesis," *IEEE TrASLP*, **20**(9):2564–2574, 2012.

J. Ahrens, S. Spors, "Sound Field Reproduction Using Planar and Linear Arrays of Loudspeakers," *IEEE TrASLP*, **18**(8):2038–2050, 2010.

# Further Reading

**Perceptual Sound Field Reconstruction**


J. D. Johnston, Y. H. Lam, “Perceptual soundfield reconstruction,” *109th AES Conv.*, paper #2399, 2000.

H.-K. Lee, F. Rumsey, “Investigation into the effect of interchannel crosstalk in multichannel microphone technique,” *118th AES Conv.*, paper #6405, 2005.

M. Williams, G. Le Du, “Microphone array analysis for multi- channel sound recording,” *107th AES Conv.*, paper #4997, 1999.

M. Williams, G. Le Du, “Miultichannel sound recording: Multichannel Microphone Array Design (MMAD),” 2010.

N. V. Franssen, *Stereophony*, Eindhoven, The Netherlands: Philips Research Laboratories, 1964.

L. Simon, R. Mason, F. Rumsey,“Localisation curves for a regularly-spaced octagon loudspeaker array,” *127th AES Conv.*, paper #7915, 2010.

S. P. Lipshitz, “Stereo Microphone Techniques... Are the Pursuits Wrong?,” *J. Audio Eng. Soc.*, **34**(9):716–744, 1986.

H. Fletcher, *Speech and Hearing in Communication*, New York, NY, USA: van Nostrand, 1953.

Y. Ando, K. Kurihara, “Nonlinear response in evaluating the subjective diffuseness of sound fields,” *J. Acoust. Soc. Amer.*, **80**(3)833–836, 1986.

E. De Sena, H. Hacihabiboglu, Z. Cvetkovic, “Analysis and Design of Multichannel Systems for Perceptual Sound Field Reconstruction,” *IEEE TrASLP.*, **21**(8):1653–1665, 2013.

E. De Sena, Z. Cvetkovic, “A Computational Model for the Estimation of Localisation Uncertainty,” *Proc. ICASSP*, pp. 388–392, 2013.

# Further Reading


E. De Sena, *Analysis, Design and Implementation of Multichannel Audio Systems,* PhD Thesis, King's College London, 2013.

V. Pulkki, "Virtual Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, **45**(6):456–466, 1997.

V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding," *J. Audio Eng. Soc.*, **55**(6):503–516, 2007.

E. De Sena, H. Hacıhabiboğlu, and Z. Cvetković, "On the design and implementation of higher-order differential microphones," IEEE Trans. on Audio, Speech and Language Process., vol. 20, no. 1, pp 162-174, Jan. 2012.

E. De Sena, H. Hacıhabiboğlu, Z. Cvetković, and J. O. Smith III "Efficient Synthesis of Room Acoustics via Scattering Delay Networks," IEEE/ACM Trans. Audio Speech Language Process., vol. 23, no. 9, pp 1478 - 1492, Sept. 2015.

E. De Sena, Niccoló Antonello, Marc Moonen, and Toon van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," IEEE/ACM Trans. Audio Speech Language Process., vol. 23, no. 4, Apr. 2015.

Rose, P. Nelson, B. Rafaely, and T. Takeuchi, "Sweet spot size of virtual acoustic imaging systems at asymmetric listener locations," J. Acoust. Soc. Amer., vol. 112, no. 5, pp. 1992- 2002, 2002

F. Rumsey, S. Zielinski, R. Kassier, S. Bech, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality", J. Acoust. Soc. Amer., 2005

Stojan Djordjevic, B.Sc. Thesis, University of Surrey, 2019

Ashley Andrew-Jones, B.Sc. Thesis, University of Surrey, 2019

E De Sena, Z Cvetkovic, H Hacihabiboglu, M Moonen, T van Waterschoot, "Localization Uncertainty in Time-Intensity Stereophonic Reproduction", arXiv preprint arXiv:1907.11425 (submitted to IEEE TrASLP), 2019.

E Erdem, E De Sena, H Hacihabiboglu, Z Cvetkovic, "Perceptual Soundfield Reconstruction in Three Dimensions

via Sound Field Extrapolation", ICASSP, 2019