Spatial audio

Simon Leglaive

2D-3D Image & Sound @ CentraleSupélec - 2020/2021

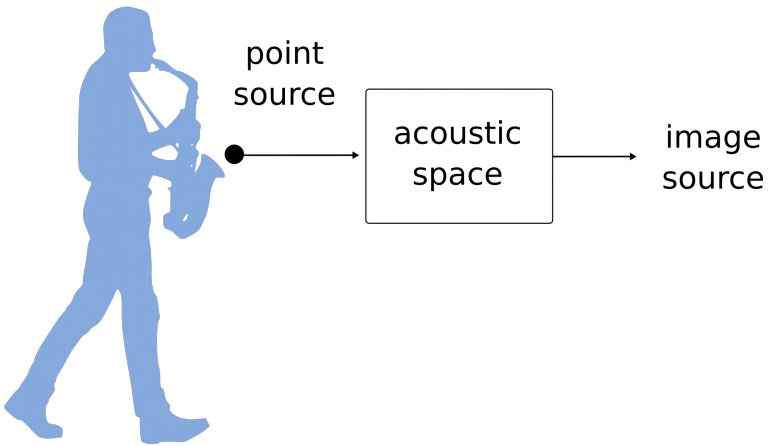
Today

- Sound propagation in free-field and rooms
- Room impulse response
- Source separation based on spatial diversity

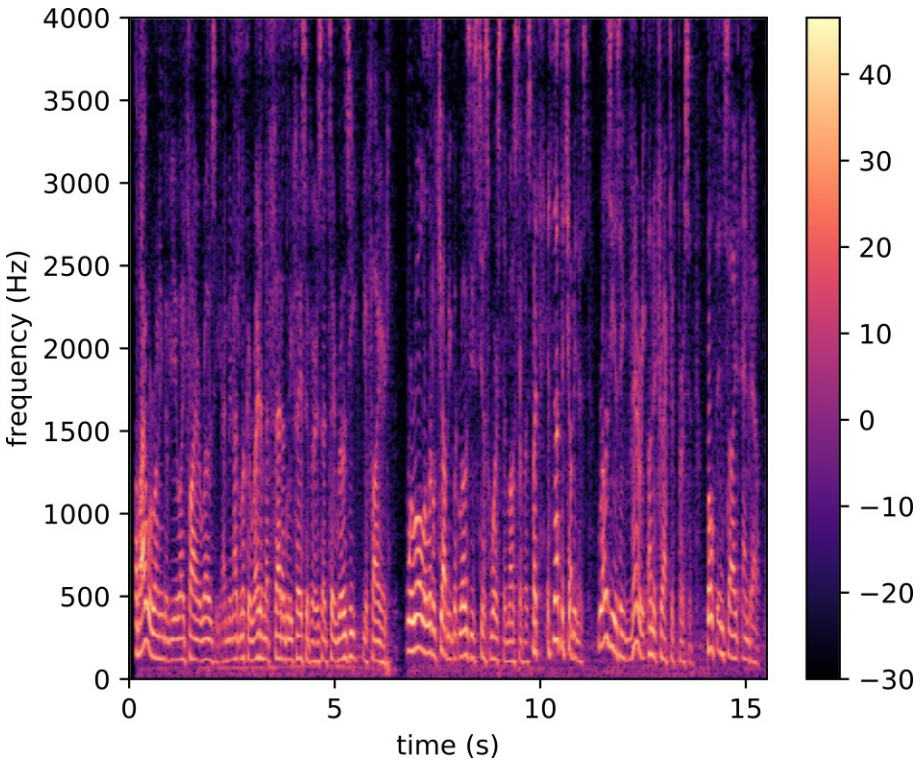
Introduction

The acoustic space

- There can be no sound without a medium allowing for the propagation of acoustic waves.
- Let's consider a sound source in a room (e.g. a speaker or musical instrument)
  - The signal recorded by a microphone does not only characterize the sound source.
  - It corresponds to the image of the sound source as it is heard through the acoustic space of the recording medium.
  - This source image depends on the **the recording environment**, the **position of the source** and the **position of the microphone(s)**.

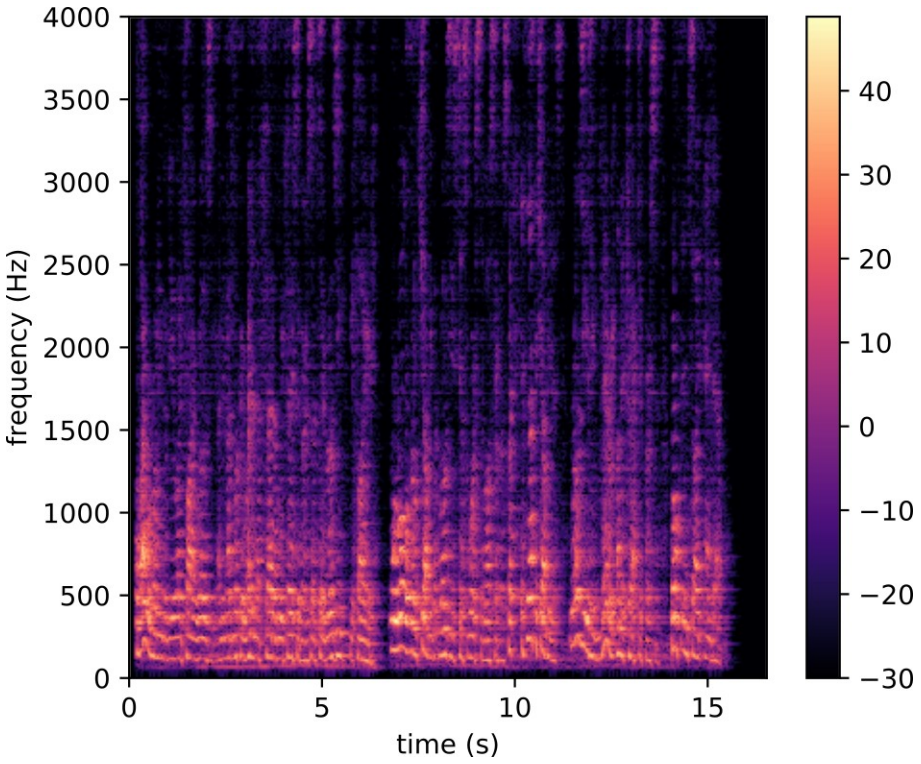


Example of the room effect



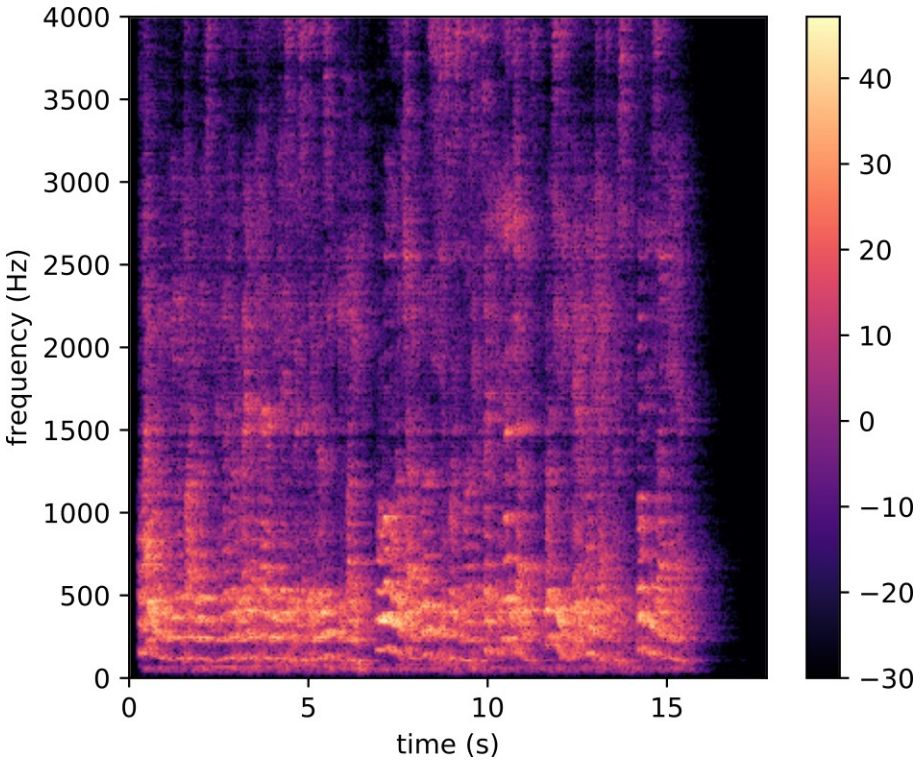
- Original signal

Example of the room effect



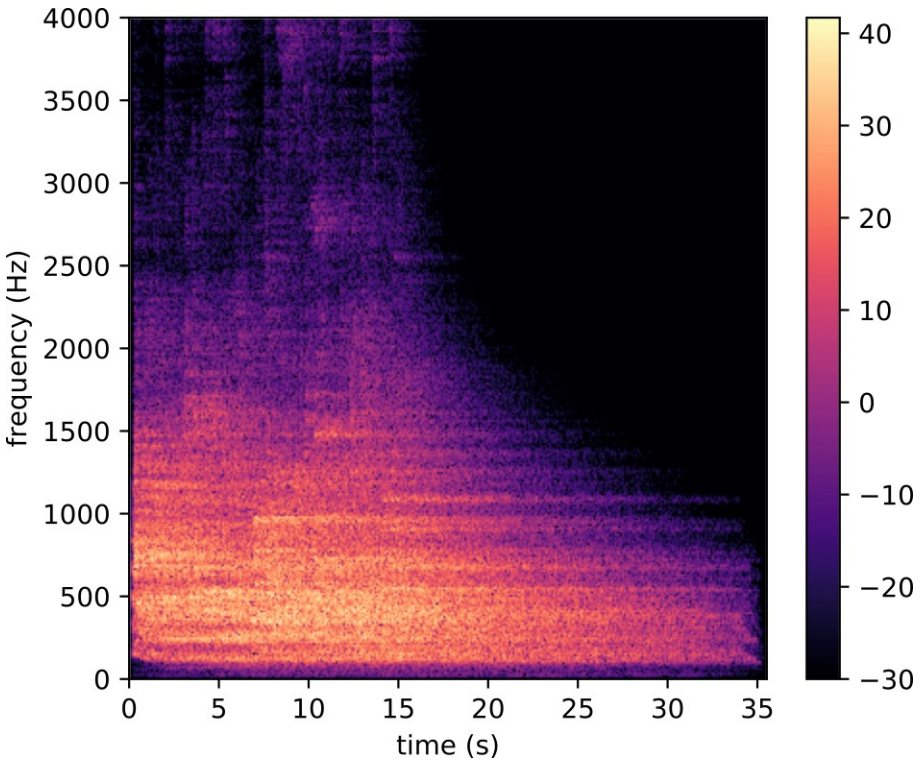
- In a bathroom

Example of the room effect



- In a concert hall

Example of the room effect



- In the Inchindown oil tanks in Scotland.
- World's longest reverberation.

The acoustic space has a great influence on the recorded audio signal.

It may for instance affect the intelligibility of a speech signal,
making extraction of information challenging for
machine listening systems.

Auralization

- In the previous example, we used the principle of **auralization**, i.e. the process of **creating a virtual rendition of a sound in a space**.
- It is used for instance in music studios, or in architectural acoustics to simulate how a concert hall will sound before building it.
- For auralization, you need:
  1. An anechoic sound, i.e. a sound recorded in a room with (almost) no reflections.
  2. The effect of the room (reflections from the walls, floor, ceiling, objects, etc.) which is captured in what is called the **room impulse response**.
- The auralized signal corresponds to the **convolution** of the anechoic sound with the room impulse response (RIR).

Convolution (reminder)

The convolution between two sequences $u(t)$ and $v(t)$ is defined for all $t \in \mathbb{Z}$ by

$$[u \star v](t) = \sum_{\tau \in \mathbb{Z}} u(\tau)v(t-\tau) = \sum_{\tau \in \mathbb{Z}} v(\tau)u(t-\tau).$$

Commutativity:

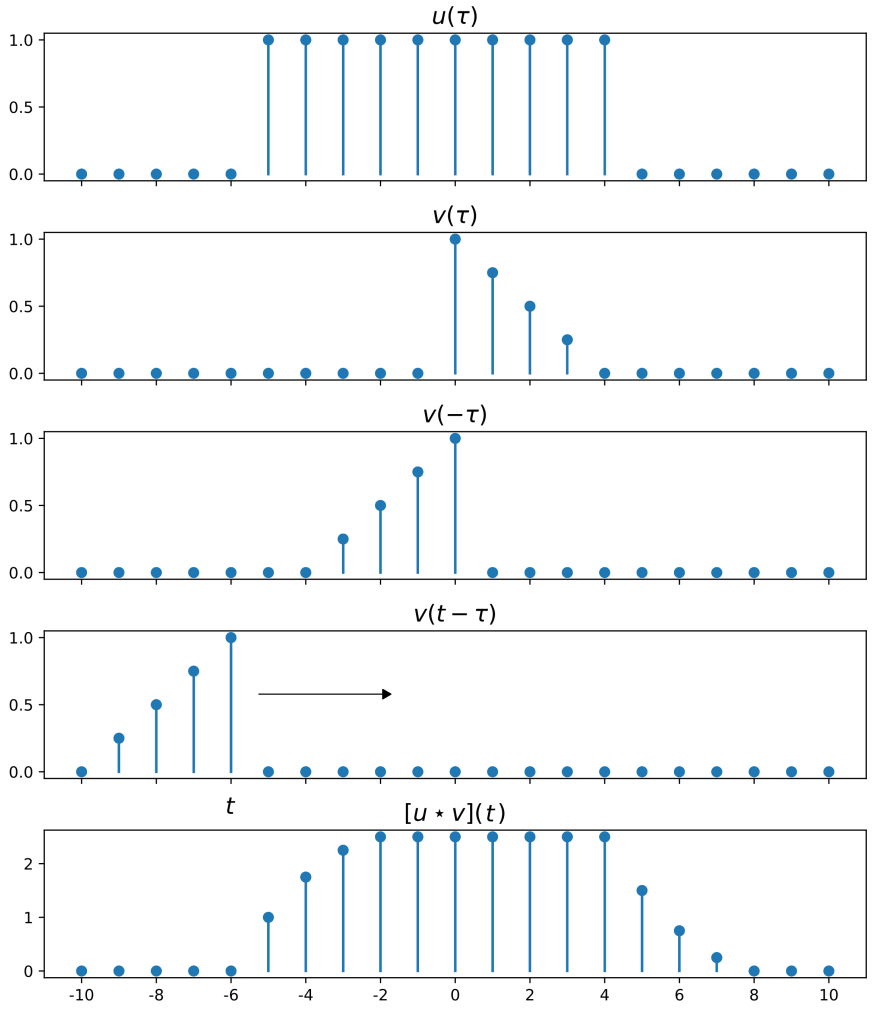
$$[u \star v](t) = [v \star u](t)$$

Associativity:

$$[(u \star v) \star w](t) = [u \star (v \star w)](t)$$

Linearity:

$$[u \star (\lambda_1 v + \lambda_2 w)](t) = \lambda_1 [u \star v](t) + \lambda_2 [u \star w](t)$$



Sound in free field

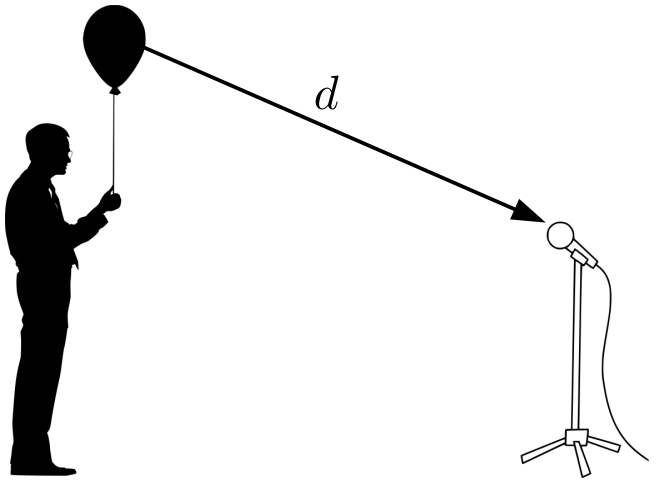
Free field propagation

- Open air environment, without any obstacle.
- Let $s(t)$ be the **source signal** and $x(t)$ the **microphone signal**, defined for $t \in \mathbb{N}$.
- We have the following relationship:

$$x(t) = \frac{1}{\sqrt{4\pi}d} s\left(t - \frac{d}{c} f_s\right),$$

where

- d is the source-to-microphone distance (in m);
- $c = 343$ is the sound speed (in m/s at 20°C);
- d/c is the time of arrival (in s);
- f_s is the sampling rate (in Hz).



- At the microphone, the source signal is simply **attenuated** and **delayed**.

Free field propagation

- We can rewrite the microphone signal as:

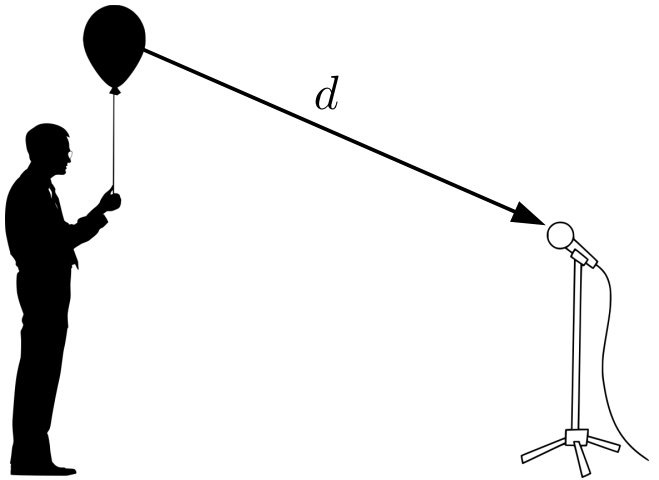
$$x(t) = [h \star s](t),$$

where

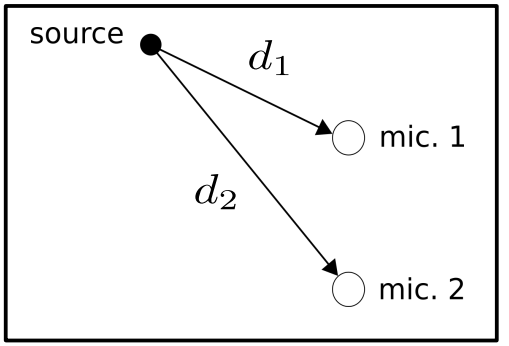
$$h(t) = \frac{1}{\sqrt{4\pi}d}\delta\left(t - \frac{d}{c}f_s\right),$$

and

$$\delta(t) = \begin{cases} 1 & t = 0 \\ 0 & t \neq 0 \end{cases}.$$



Multi-microphone anechoic recording



- In a multi-microphone recording, we have different attenuation and delay factors for each microphone:

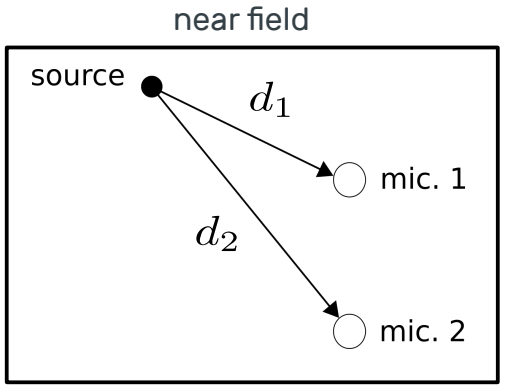
$$x_1(t) = \frac{1}{\sqrt{4\pi} d_1} s\left(t - \frac{d_1}{c} f_s\right) \qquad x_2(t) = \frac{1}{\sqrt{4\pi} d_2} s\left(t - \frac{d_2}{c} f_s\right)$$

- Without loss of generality, we can absorb the attenuation and delay parameters at the first microphone into the definition of the source signal (change of variable):

$$x_1(t) = s(t) \qquad x_2(t) = \boxed{\frac{d_1}{d_2}} s\left(t - \boxed{\frac{d_2 - d_1}{c}} f_s\right)$$

- The **level ratio** and the **time difference of arrival** (TdoA) between the microphones convey information about the **position of the source**.

Far-field case



In the far-field case, we assume that the source-to-microphone distances are large compared with the inter-microphone distances such that:

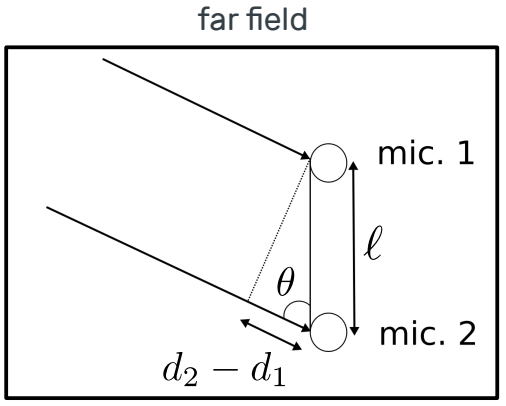
- The level ratio is almost equal to one:

$$d_1/d_2 \approx 1$$

- The TDoA is given by:

$$\mathrm{TDoA} = \frac{d_2 - d_1}{c} = \frac{\ell}{c}\cos(\theta)$$

where ℓ is the inter-microphone distance (in m).



Practical activity #1

Anechoic auralization

Sound in rooms

Room impulse response

- The room can be considered as a **causal**, **linear** and **time-invariant** system (if we neglect changes in temperature, pressure, etc.), so we still have:

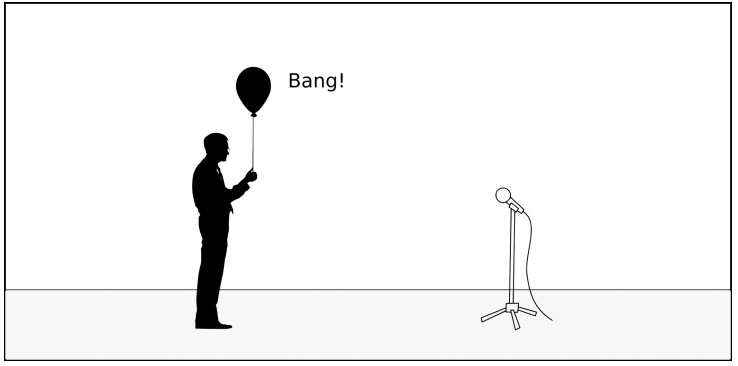
$$x(t) = [h \star s](t),$$

where $h(t) = 0$ for $t < 0$ due to causality.

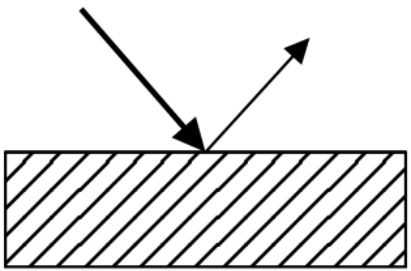
- $h(t)$ is called the **room impulse response** (RIR), it characterizes the **acoustic path** between 2 points in the room (the source and the microphone locations).

It is called the RIR because it is the response of the room when the excitation (source) signal is a dirac delta function:

$$x(t) = \sum_{\tau=0}^{+\infty} h(\tau)\delta(t-\tau) = h(t).$$

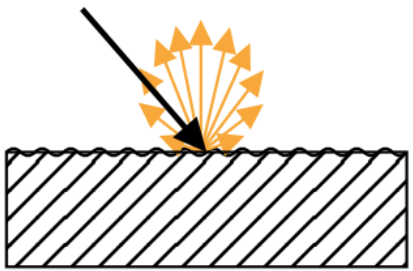


- In rooms, sound propagation is more complicated than a simple attenuation and delay.
- When a sound wave propagates in a room, in encounters surfaces with which it interacts in different ways.



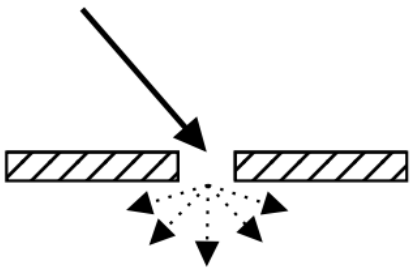
Specular Reflection

The direction of the reflected wave is symmetrical to the direction of the incident wave with respect to the surface normal,



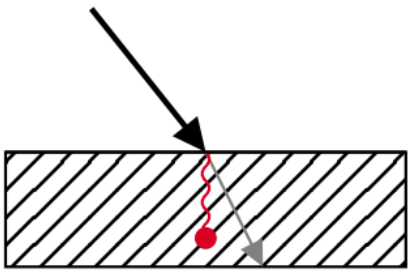
Diffuse Reflection

The wave may be reflected in many directions depending on its wavelength and the dimension and irregularities of the surface.



Diffraction

The wave is diffracted in a way that depends on its wavelength, the shape of the obstacle or opening, its material and the angle of incidence.



Refraction and Absorption

The absorption ratio depends on the material and the angle of incidence.

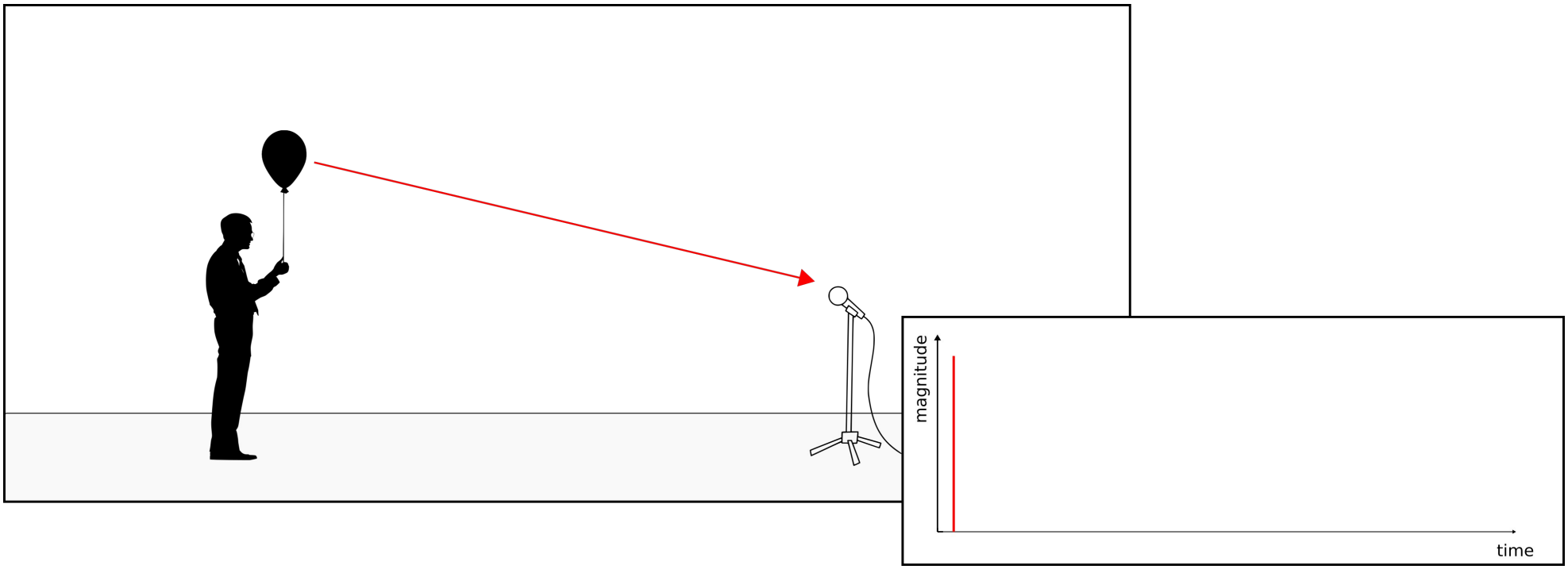
Image credits: Diego Di Carlo's Ph.D. Thesis "Echo-aware signal processing for audio scene analysis", 2020.

Reverberation in rooms

- When a source emits sound in a room, many successive reflections typically occur before the sound power becomes negligible, this is called **reverberation**.
- This induces **multiple propagation paths between the source and the microphone**, each with a different delay and attenuation factor.

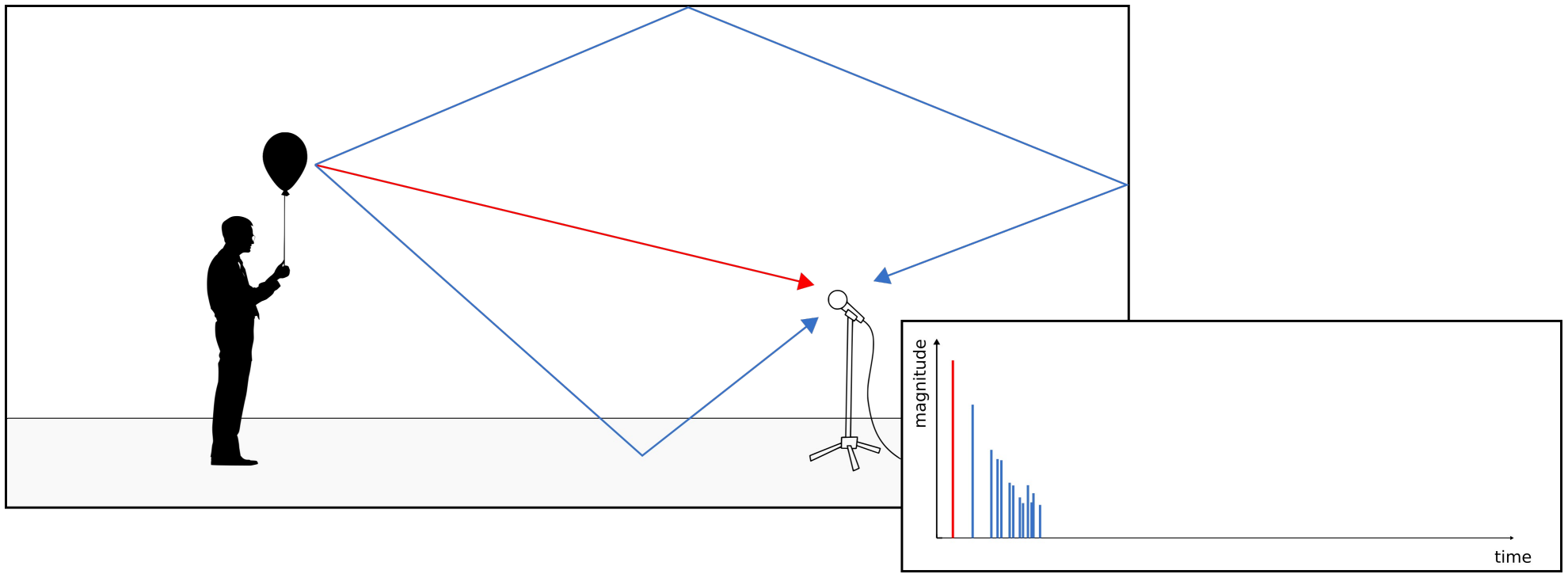
Schematic illustration

Direct path



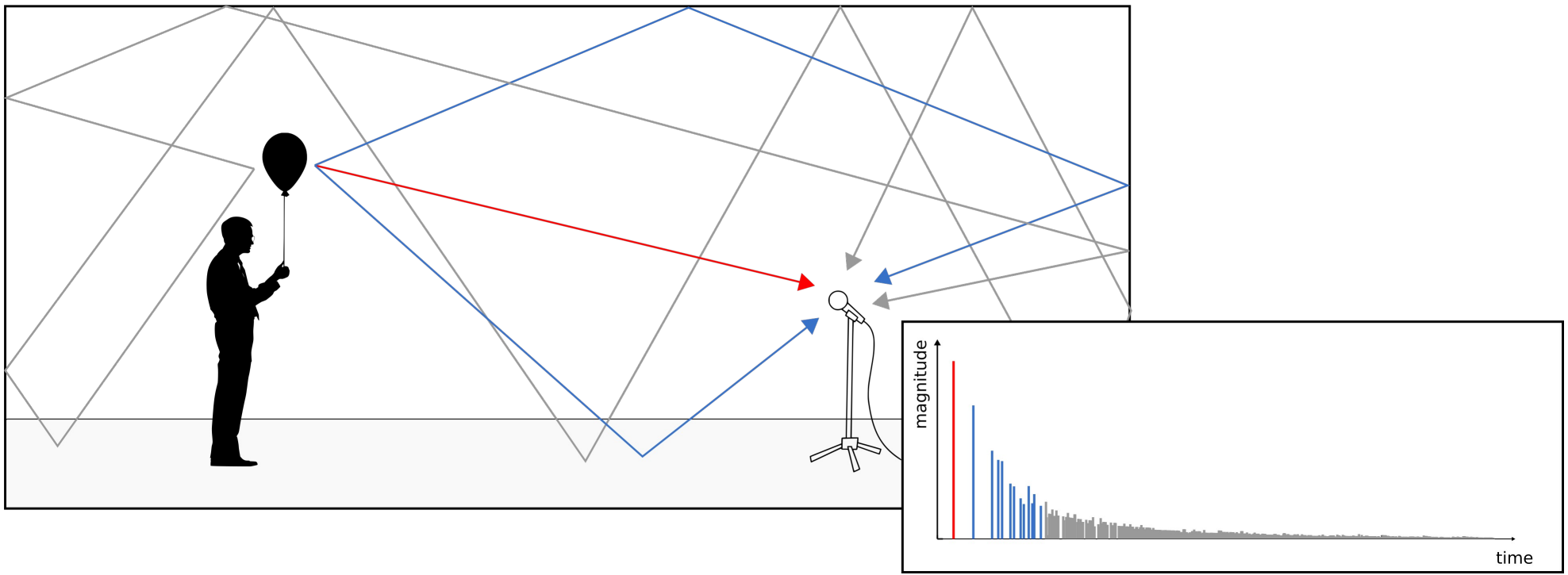
Schematic illustration

Early echoes

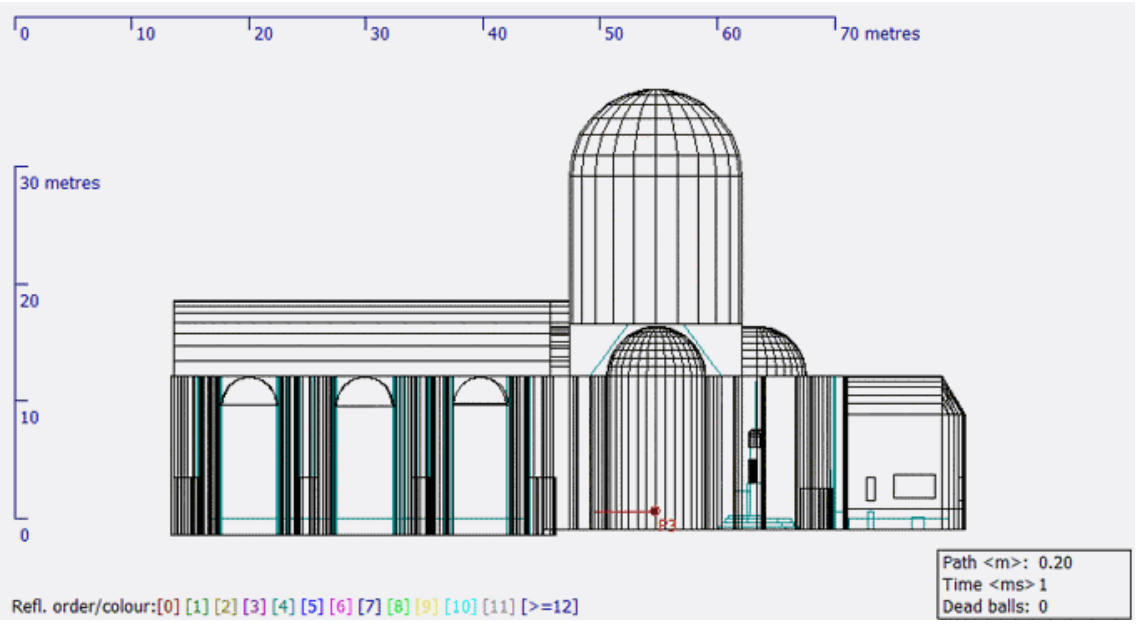


Schematic illustration

Late reverberation



Simulating room acoustics



- Simulation of the acoustics in the Church of the Redentore in Venice.
- It is possible to simulate room impulse responses based on the geometry of the room.

Image credit: Braxton Boren/American Univ.

Capturing room acoustics

Notre Dame, Paris, after the fire on April 15, 2019.

Image credit: B. Katz and M. Pardoen/CNRS

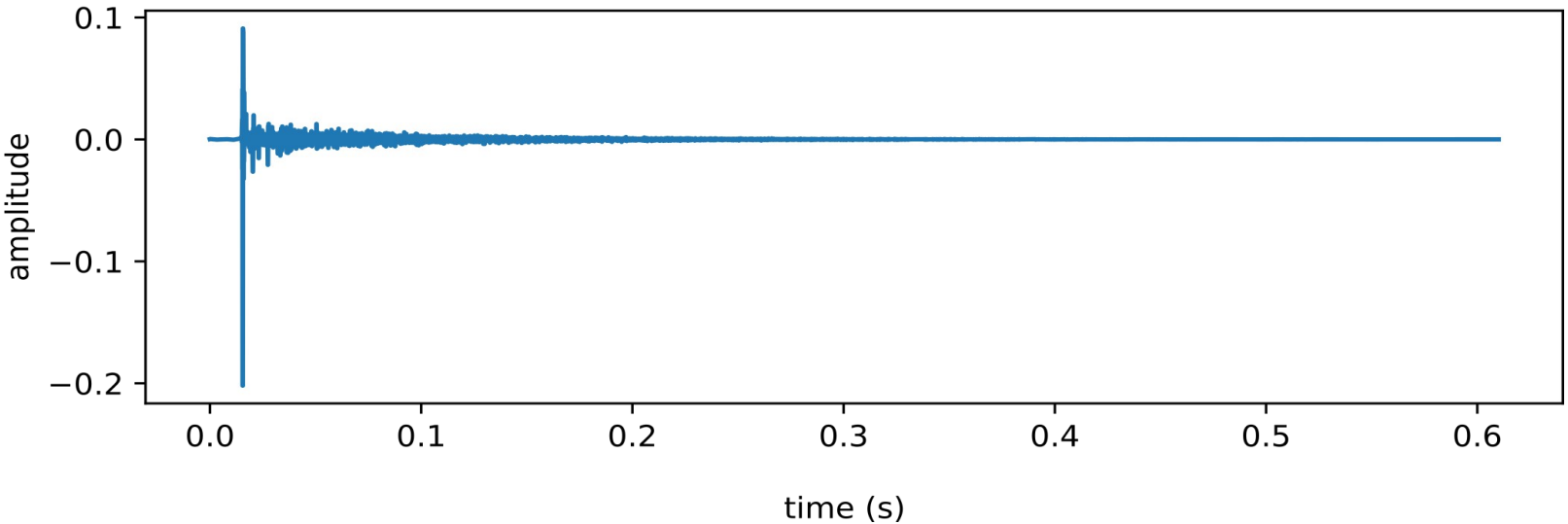
- Brian Katz took acoustic measurements of room impulse responses in Notre Dame, in 2013.
- These data may help restoring the acoustics of the cathedral.

B. Katz installation to record Notre Dame's acoustics.

Image credit: B. Katz/CNRS

More info: https://www.sciencenews.org/article/notre-dame-cathedral-fire-legendary-acoustics-restoration

Characterization of RIRs



Thousands of coefficients, which can be described by **three main properties**:

- Reverberation time (T60)
- Direct-to-reverberant ratio (DRR)
- Direct-to-early ratio (DER)

Reverberation time

- The **reverberation time** is the time it takes for the sound energy to **decrease by 60 dB** after extinction of the source.
- It depends on the size and absorption coefficient of the room materials, and obstacles in the room.
- Between 0.1 and 0.8 seconds for domestic or office rooms, up to 2 seconds for concert halls, and 75 seconds in the Inchindown oil tanks!
- In the late 1890s, Sabine established a relationship between the T60 (in s) of a room, its volume (in m^3), its total surface area (in m^2), and the average absorption coefficient of the room surfaces:

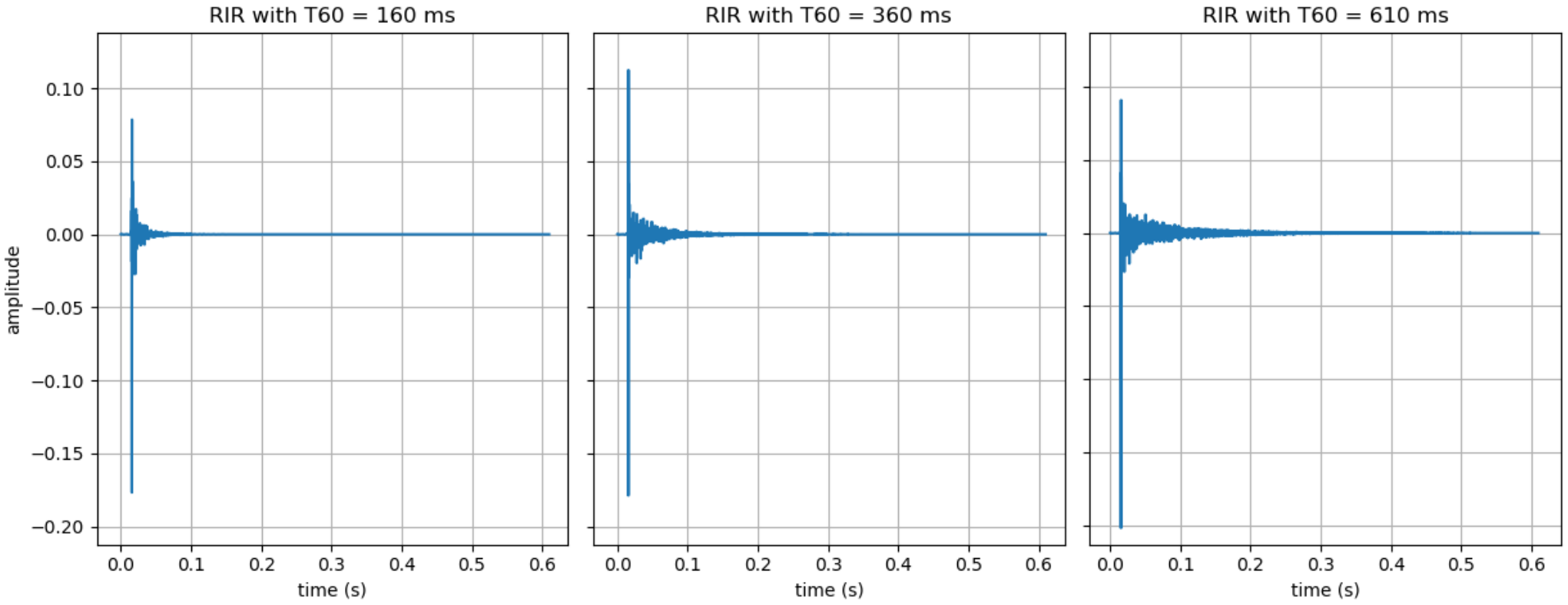
$$\mathrm{T60} = \frac{24\ln(10)}{c}\frac{V}{Sa} \approx 0.1611\frac{V}{Sa}$$

volume

average absorption coefficient

total surface area

RIRs with different T60s



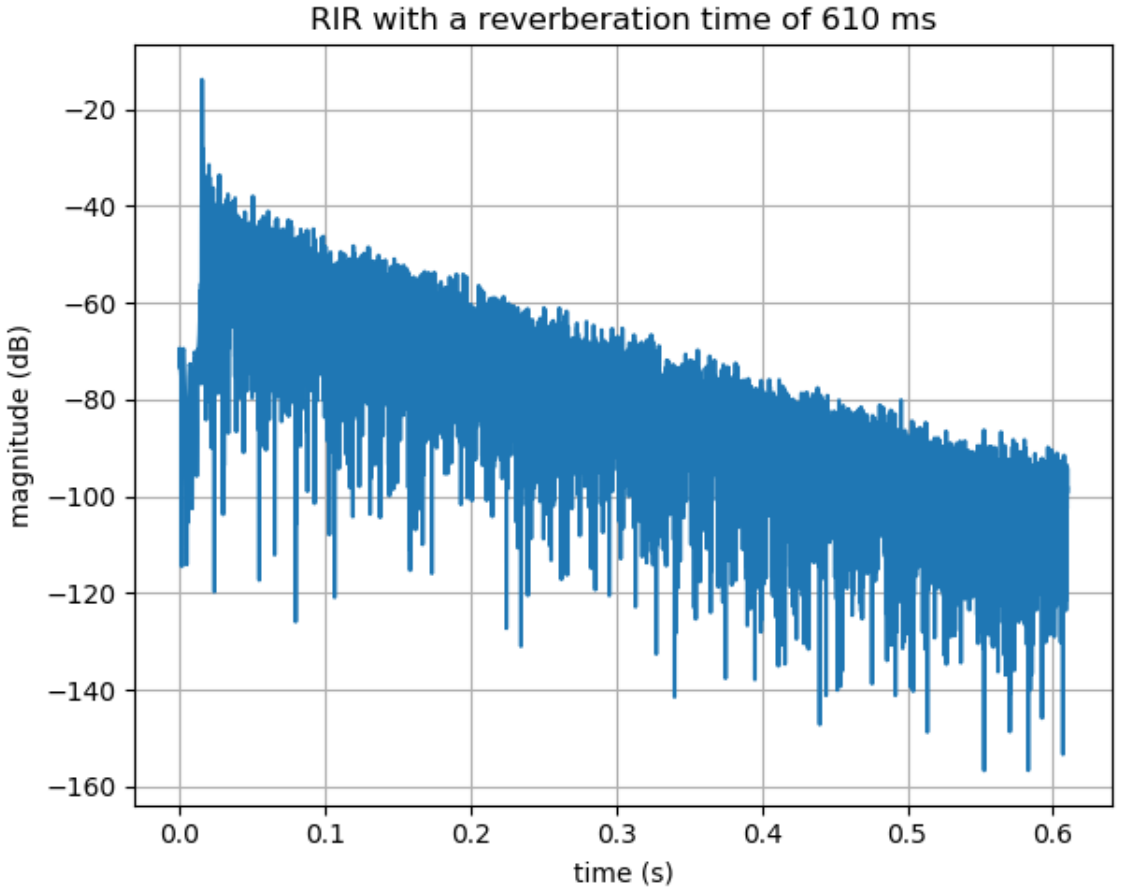
Late reverberation

- The **late reverberation energy decays exponentially** in time.
- Late reverberation can be modeled as a white Gaussian noise with an exponentially decaying envelope:

$$h_{\text{late}}(t) = w(t)e^{-t/\tau}$$

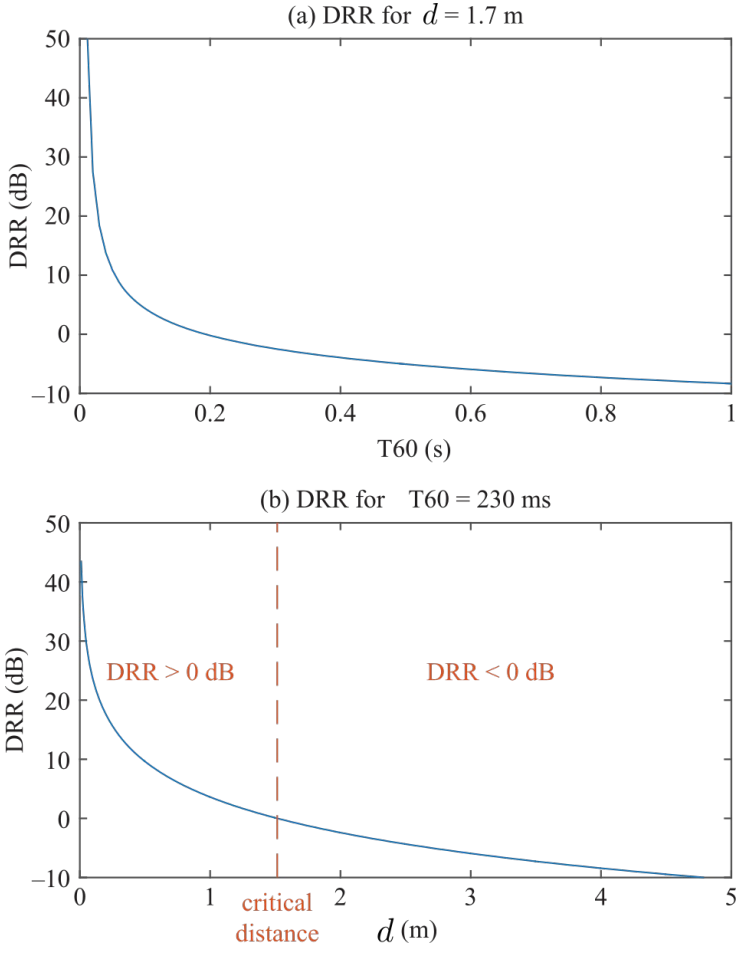
where

- $w(t)$ is a white Gaussian noise
- $\tau = \dfrac{T_{60} f_s}{3\ln(10)}$



Direct-to-reverberant ratio

- The DRR is the **power ratio of the direct and reflected sounds**.
- It varies with the size and the absorption of the room, but also with the distance between the source and the microphone.
- The distance beyond which the power of indirect sound becomes larger than that of direct sound is called the **critical distance**.



DRR as a function of the T60 and the source distance, assuming no obstacle between the source and mic.

image credit: E. Vincent, S. Gannot, T. Virtanen, "Acoustics - Spatial properties", in "*Audio source separation and speech enhancement*", *Wiley,* 2018, 978-1-119-27989-1.

Direct-to-early ratio

- The DER is the ratio of the power of direct path and early echoes.
- It quantifies the modification of the power spectrum of the signal induced by early echoes
- It is low when the microphone and/or the source is close to an obstacle such as a table or a window, and higher otherwise.

Practical activity #2

Source separation based on spatial cues