3D room reconstruction from sound

Alessio Del Bue and Marco Crocco

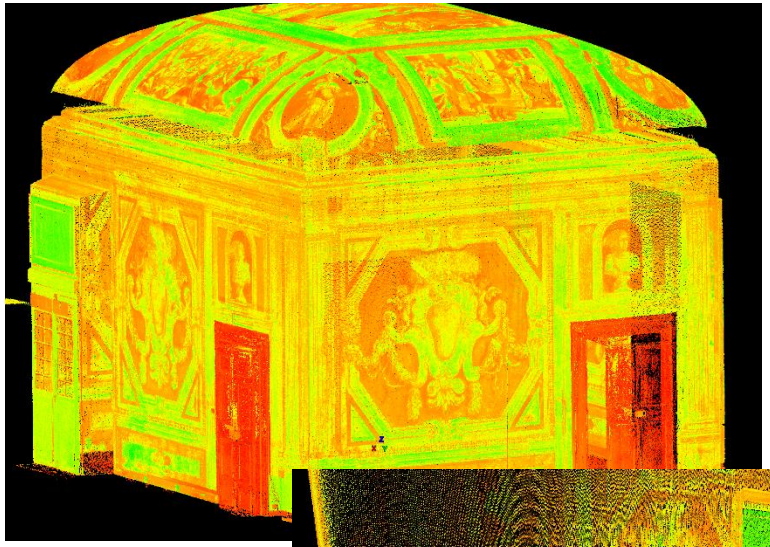
Visual Geometry and Modelling Lab

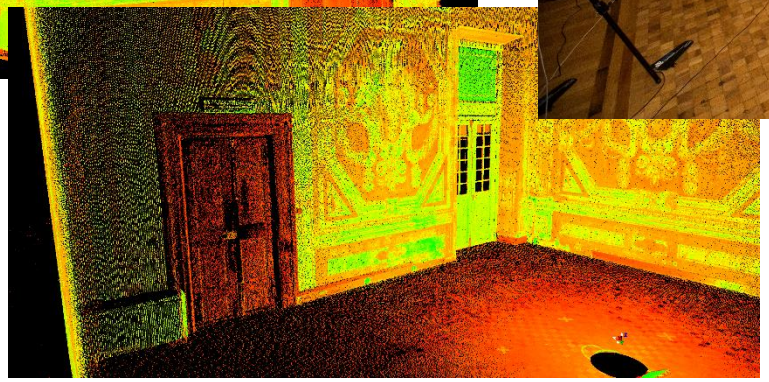
Istituto Italiano di Tecnologia (IIT)

Genova, Italy

3D room reconstruction from sound

It is the problem of estimating the room geometry (position of reflectors) and sensors (microphones) position solely from a set of audio signals emitted from sound sources.





What is about?

A sensing problem

Is it possible to understand the shape of a room from audio signals only?

An inverse problem

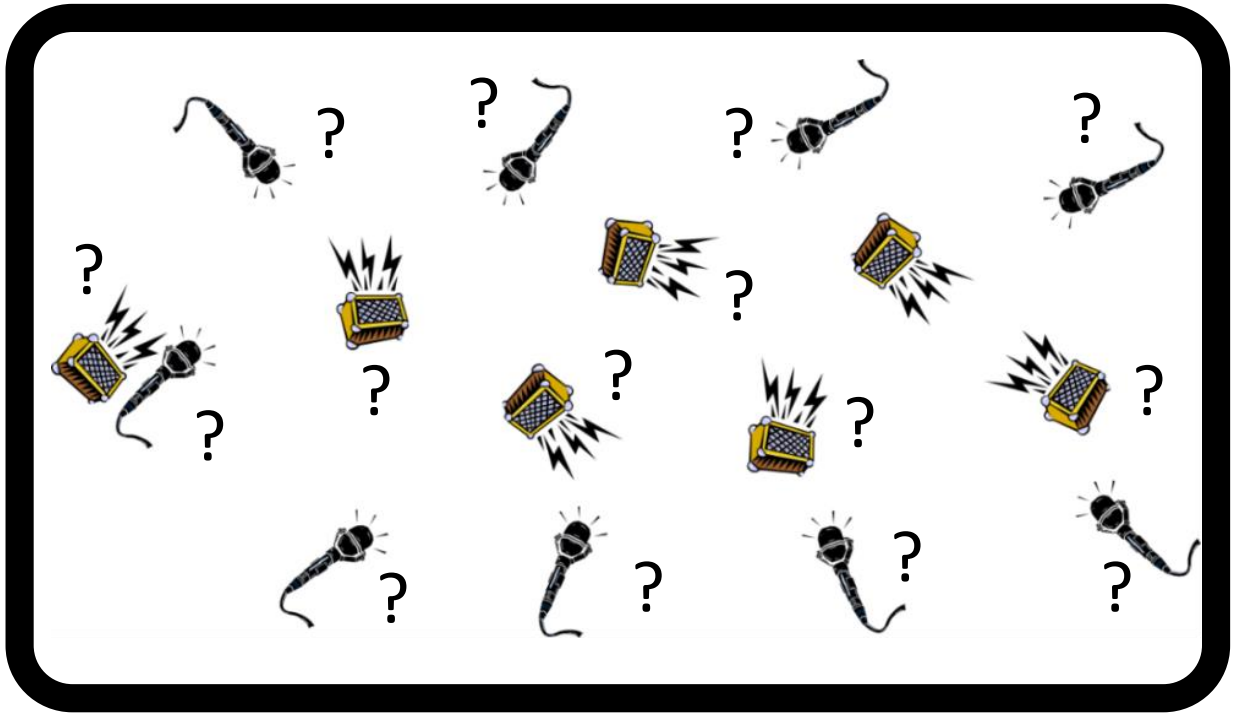
From a set of 1D signals recover the 3D structure of the environment.

An optimization problem

Related to deconvolution, combinatorial problems, multi-dimensional scaling, etc.

What is the aim?

To provide and explain a set of methods for solving the 3D room reconstruction problem with a major emphasis on the most difficult case: **No knowledge about the sensors/audio sources involved and of the nature of the audio signal/room.**



Which is the focus?

This tutorial puts more emphasis on the **geometrical aspects of the room reconstruction problem** rather then the specific signal processing problems.

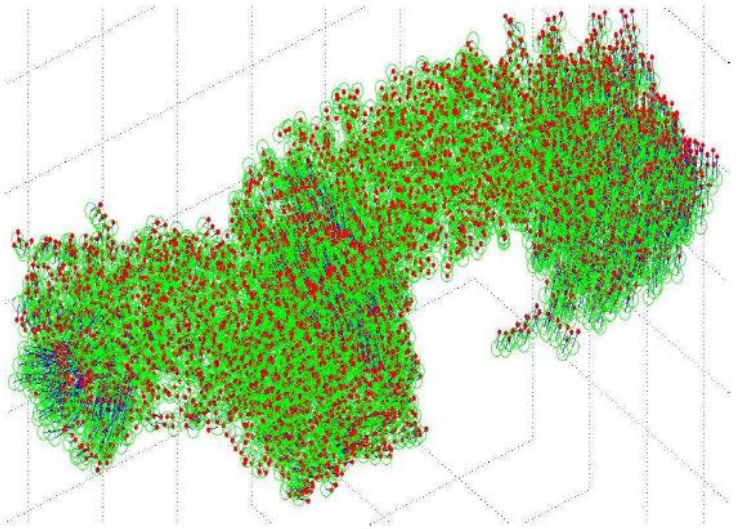
In particular we will provide an **algorithmic pipeline from audio to 3D geometry computation** that can be used in practice. We will show that geometrical reasoning can save from the several pitfalls when searching for the solution.

We will put emphasis on real scenarios and the difficulties in evaluating such methods due to lack of data. To this end, **a real dataset with ground truth** will be made available.

Applications: Sensor network

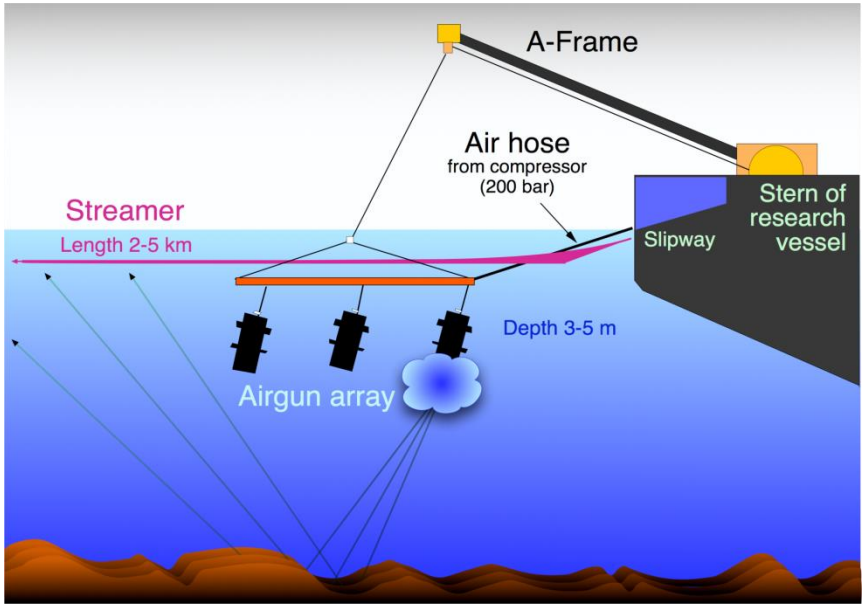
Small networks often are easily deployable and localisation of the sensors is made manually (if the sensors location is accessible) or using custom devices (e.g. GPS)

What happens with >100, >1000 devices?

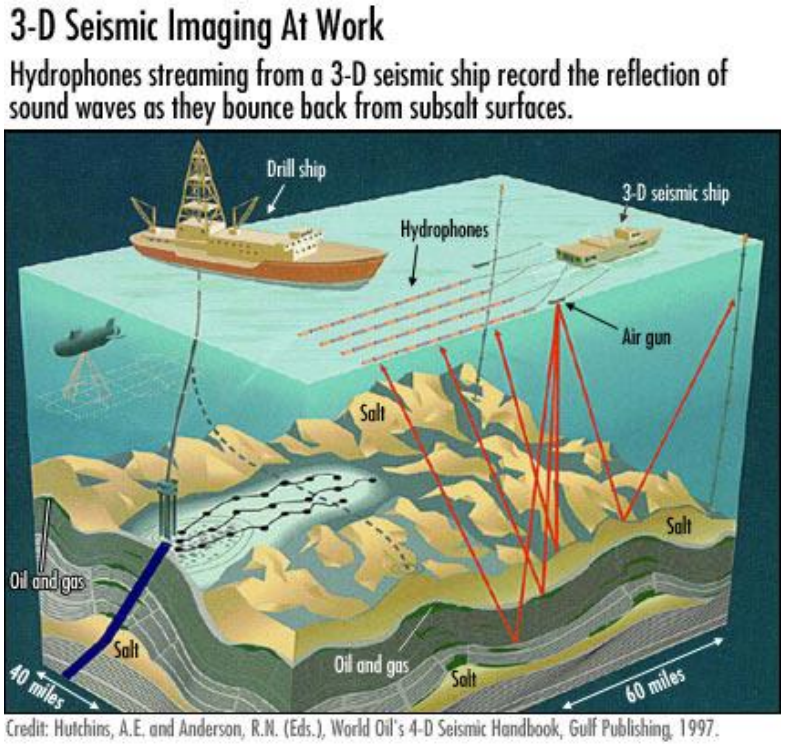


Manual localisation is not feasible anymore. Adding an external localisation device may be too expensive.

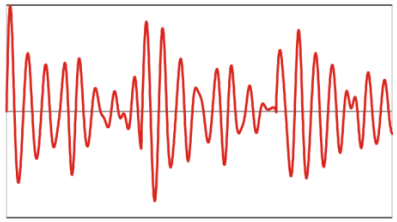
Applications: Sismic imaging

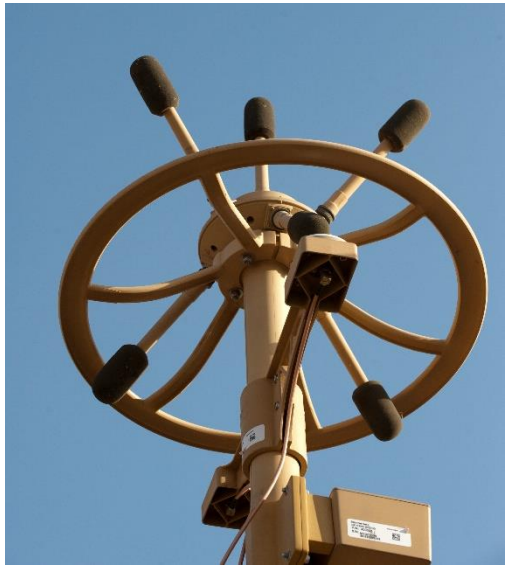


Source: Hannes Grobe, Alfred Wegener Institute



Applications: Sniper detection





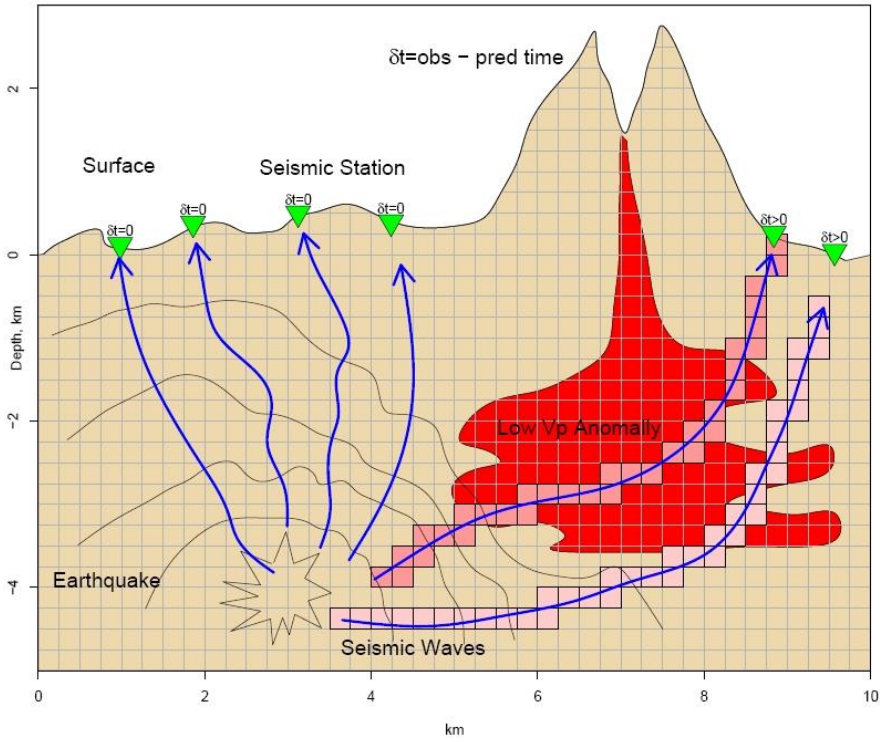
Source Wikipedia

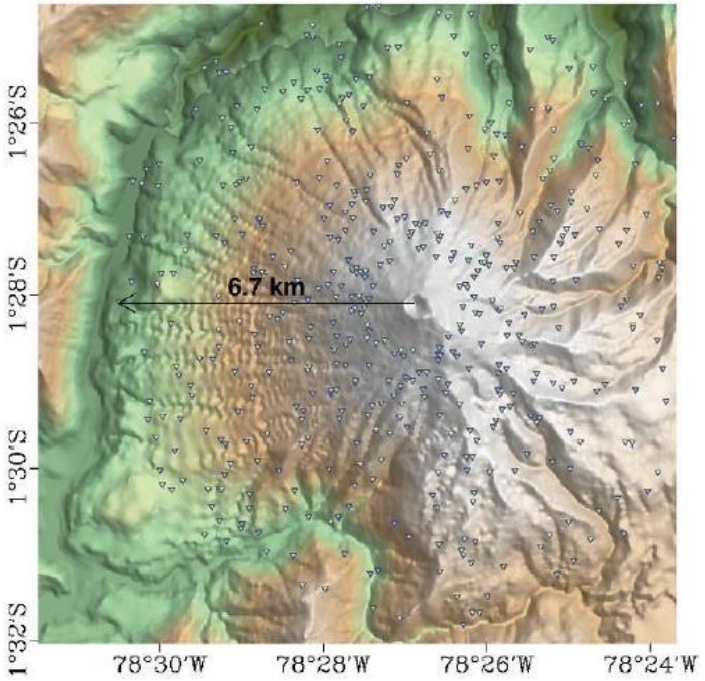
Boomerang is a gunfire locator developed by DARPA and BBN Technologies primarily for use against snipers.

Users receive simultaneous visual and auditory information on the point of fire from an LED 12-hour clock image display panel and speaker mounted inside the vehicle. For example, if someone is firing from the rear, the system announces "Shot, 6 o'clock", an LED illuminates at the 6 o'clock position, and the computer tells the user the shooter's range, elevation and azimuth (source: Wikipedia).

Application: Volcano tomography

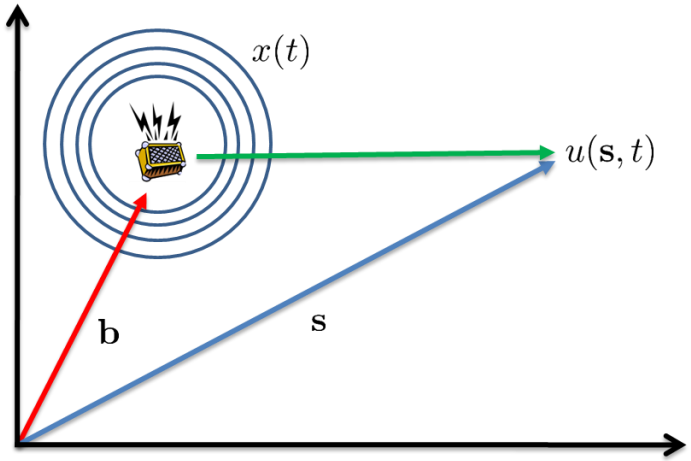
A sensor network composed by 500 sensors deployed in a volcanic area.





The system analyses seismic signals and computes real-time, full-scale, three-dimensional fluid dynamics of the volcano conduit system within the active network.

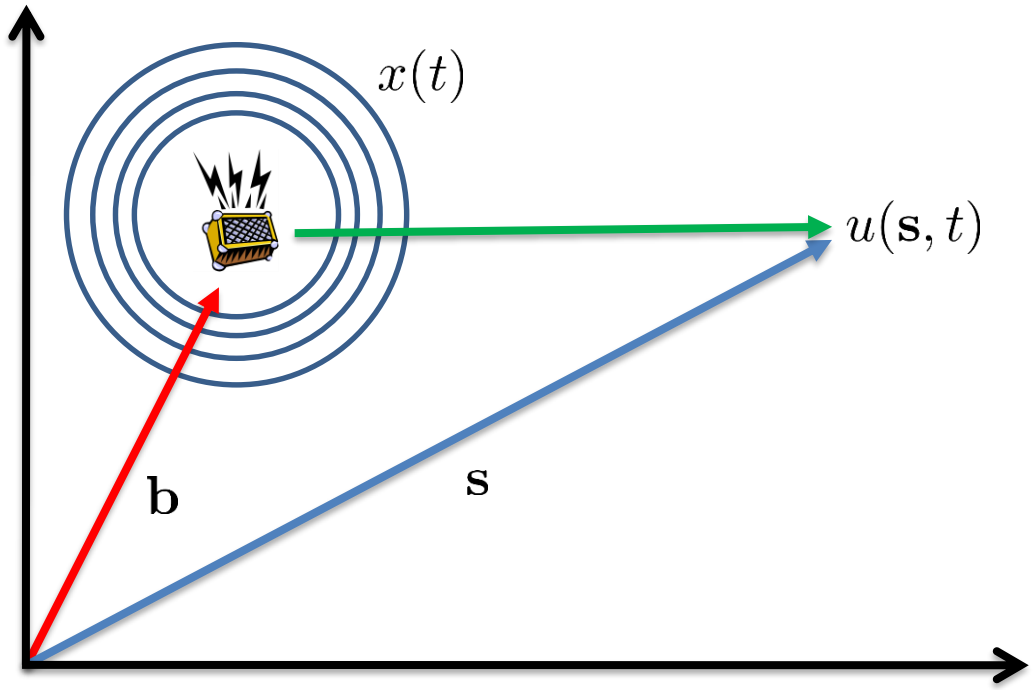
Introduction on
sound propagation and room acoustic



Sound propagation in free space

Wave equation: $\Delta u - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0$

Δ	laplacian operator
c	sound velocity
$u(\mathbf{s}, t)$	pressure field at position **s** and time *t*
$d = \|\mathbf{s} - \mathbf{b}\|_2$	distance
$\tau = d/c$	time of flight

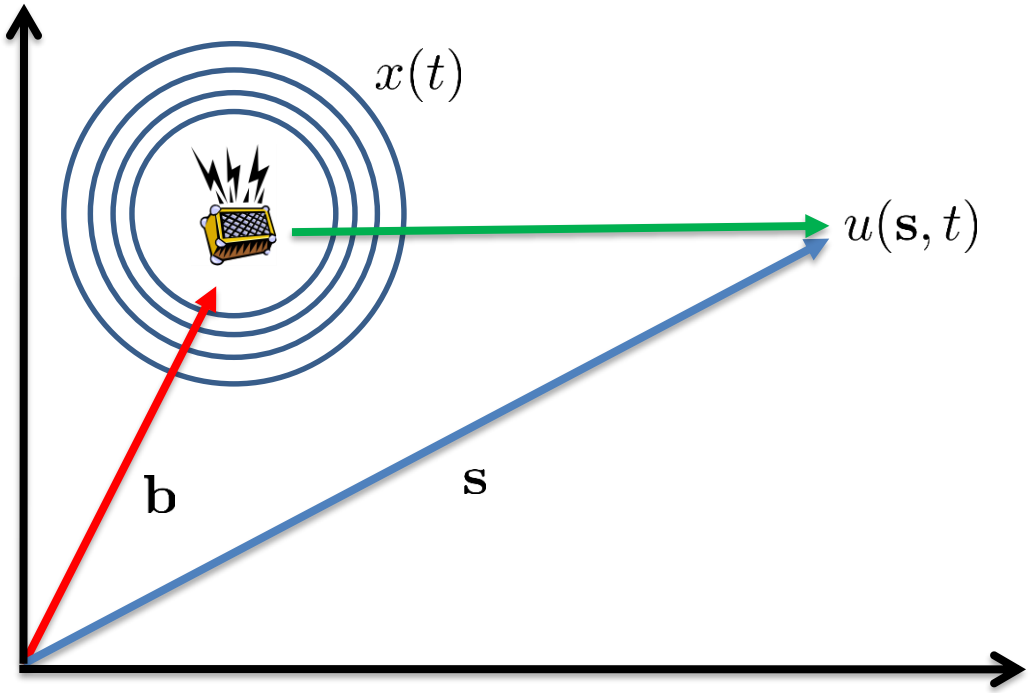


For a point-like omnidirectional source placed at a 3D spatial coordinate and emitting a signal $x(t)$ a solution is given by: $u(\mathbf{s}, t) \propto \frac{x(t-\tau)}{4\pi d}$

Assumptions on sound propagation

$$u(\mathbf{s},t) \propto \frac{x(t-\tau)}{4\pi d}$$

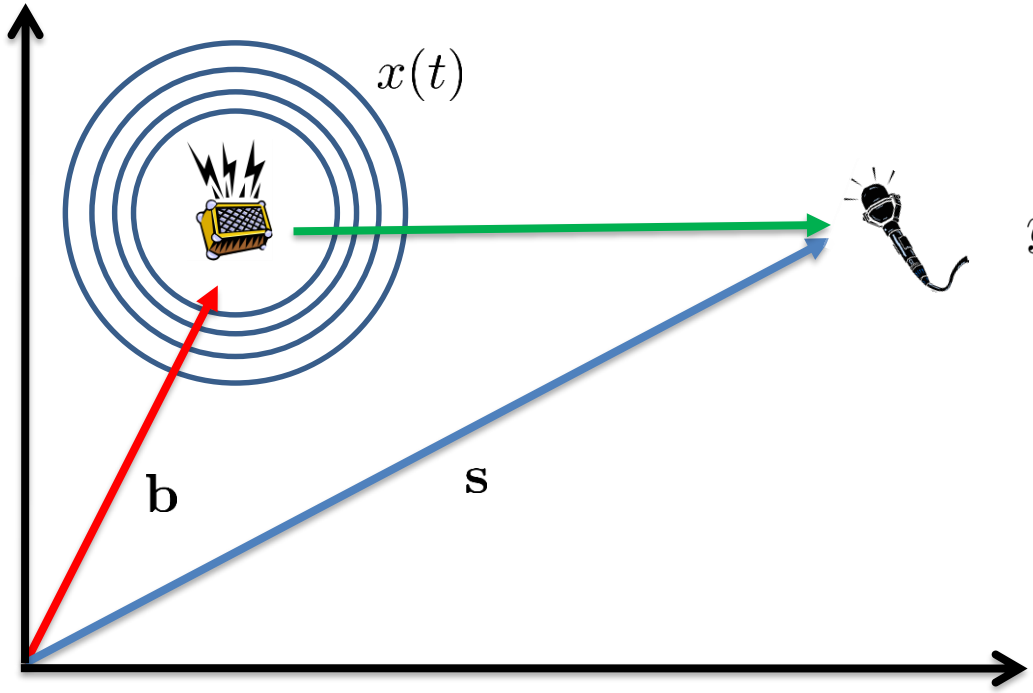
Spherical shape propagation with geometric attenuation



Hypotheses:

Frequency-dependent thermoviscous attenuation and nonlinear acoustic propagation neglected.
For moderate sound intensities and distances < 50 - 100 m the above approximations are acceptable.

Probing the sound: single microphone



$$y(t) = u(\mathbf{s}, t) * h^{mic}(t) + \nu(t)$$

$h^{mic}(t)$: microphone impulse response

$\nu(t)$: ground noise

Hypotheses:

- Point like, omnidirectional microphone.
- Nonlinear effects in the transduction from pressure to voltage signal neglected.

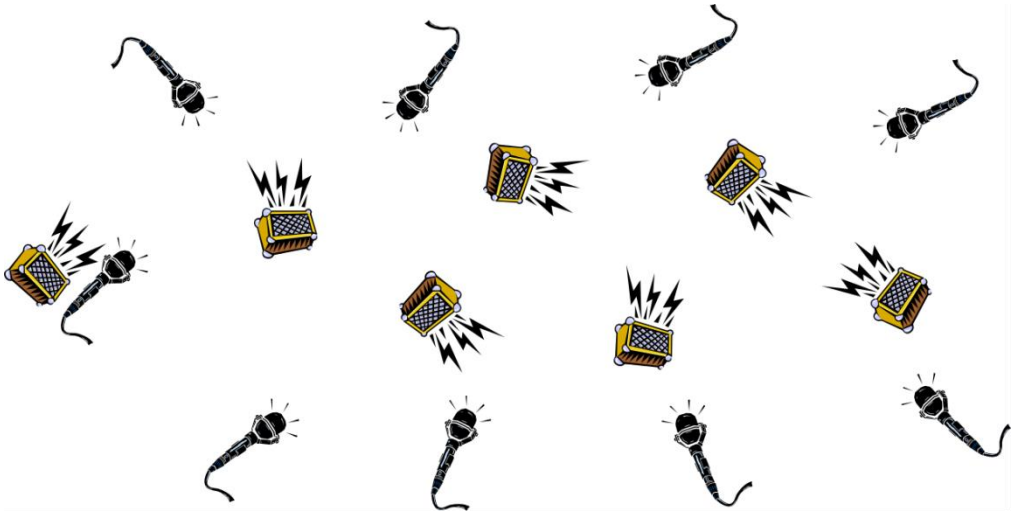
Multiple sources and mics

Sound event *n* at position *m*:

$$u^n(\mathbf{s}_m, t) \propto \frac{x^n(t - \tau^n_{m0})}{4\pi d^n_{m0}}$$

Signal *n* at microphone *m*:

$$y^n_m(t) = u^n(\mathbf{s}_m, t) * h^{mic}_m(t) + \nu_m(t)$$

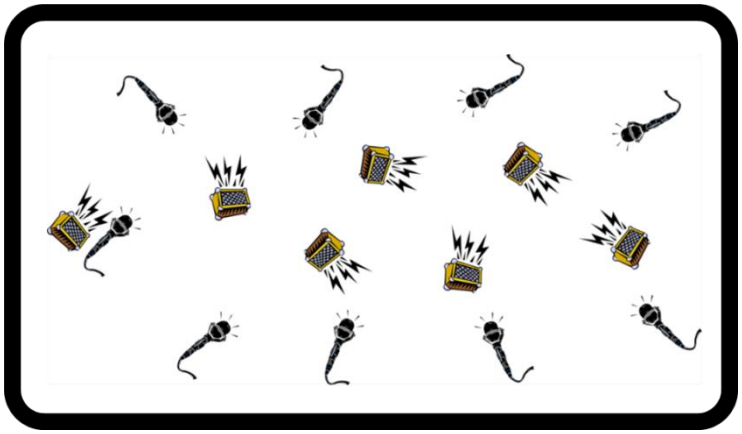


Distance from source *n* to microphone *m*: $d^n_{m0} = \|\mathbf{s}_m - \mathbf{b}^n\|_2$

Time of flight from source *n* to microphone *m*: $\tau^n_{m0} = d^n_{m0}/c$

Sound propagation in a room

In a general sound propagation is subject to **reflection** and **scattering** from objects and surfaces inside the space.



We model sound propagation in a room defining a **Room Impulse Response (RIR)** $h_m^{room,n}(t)$ between each source point *n* and each microphone point *m* such that*:*

$$y_m^n(t) = x^n(t) * h_m^{room,n}(t) * h_m^{mic} + \nu_m(t)$$

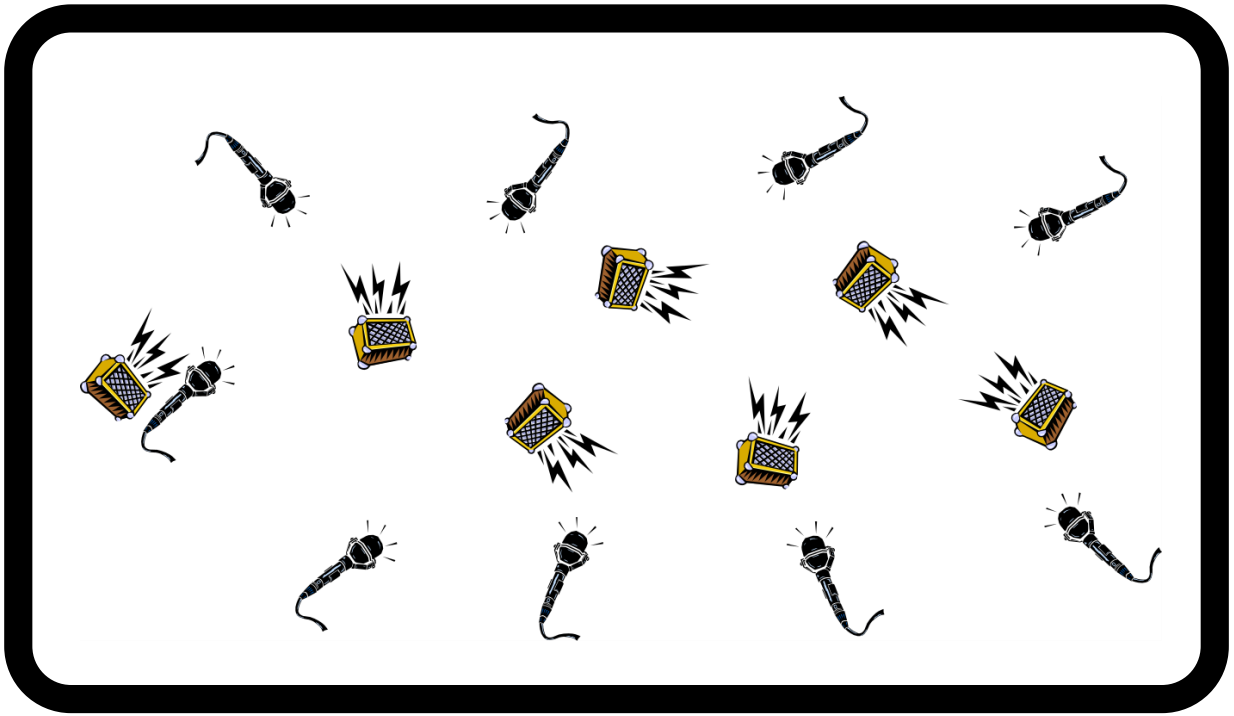
For a general environment RIRs are difficult to model!

Simpler case: planar surfaces

Assume that the boundaries of the environment are given by $k = 1, \cdots, K$ intersecting *K* **planar surfaces** giving a **convex polyhedron.**

$k = 1$

$k = K$

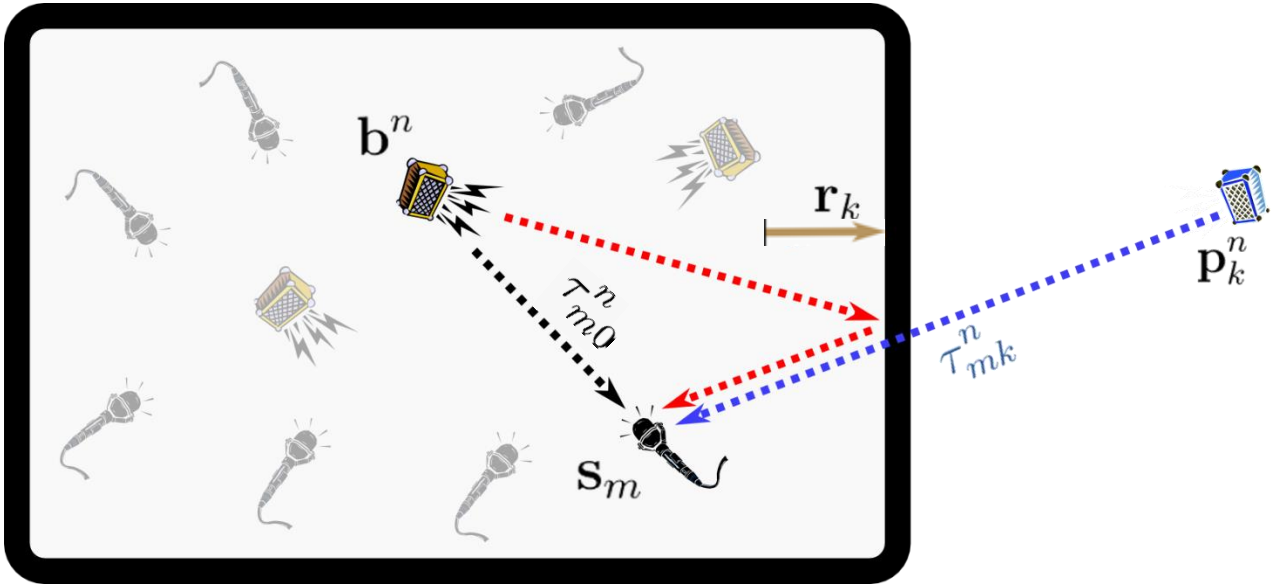


$k = 2$

$k = 3$

The Image Source Method (1)

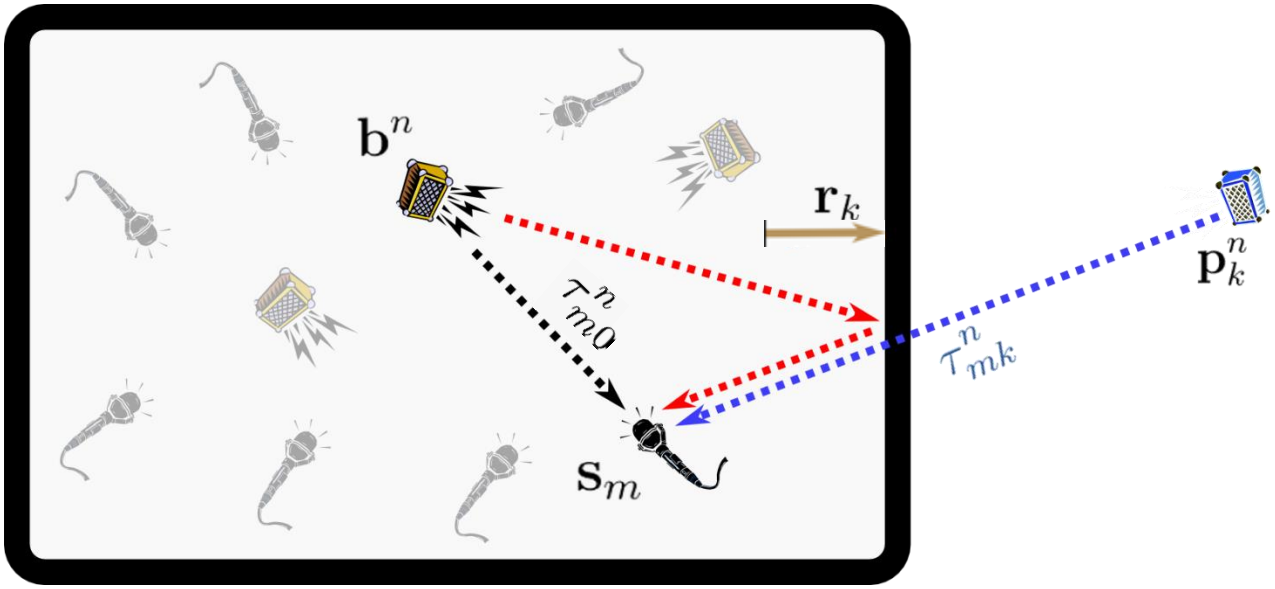
Assume that each planar surface is an infinite plane and that the wall is perfectly rigid



Reflection of source $\mathbf{b}^n$ from planar surface *k* can be modeled as a **virtual source synchronous** with the real one, **emitting a scaled version** of the transmitted signal, and with a **location** $\mathbf{p}^n_k$ **symmetric** to the real source, taking as **symmetry plane the planar surface.**

Allen and Berkley, "Image method for efficiently simulating small-room acoustics." *The Journal of the Acoustical Society of America,* 1979.

The Image Source Method (2)



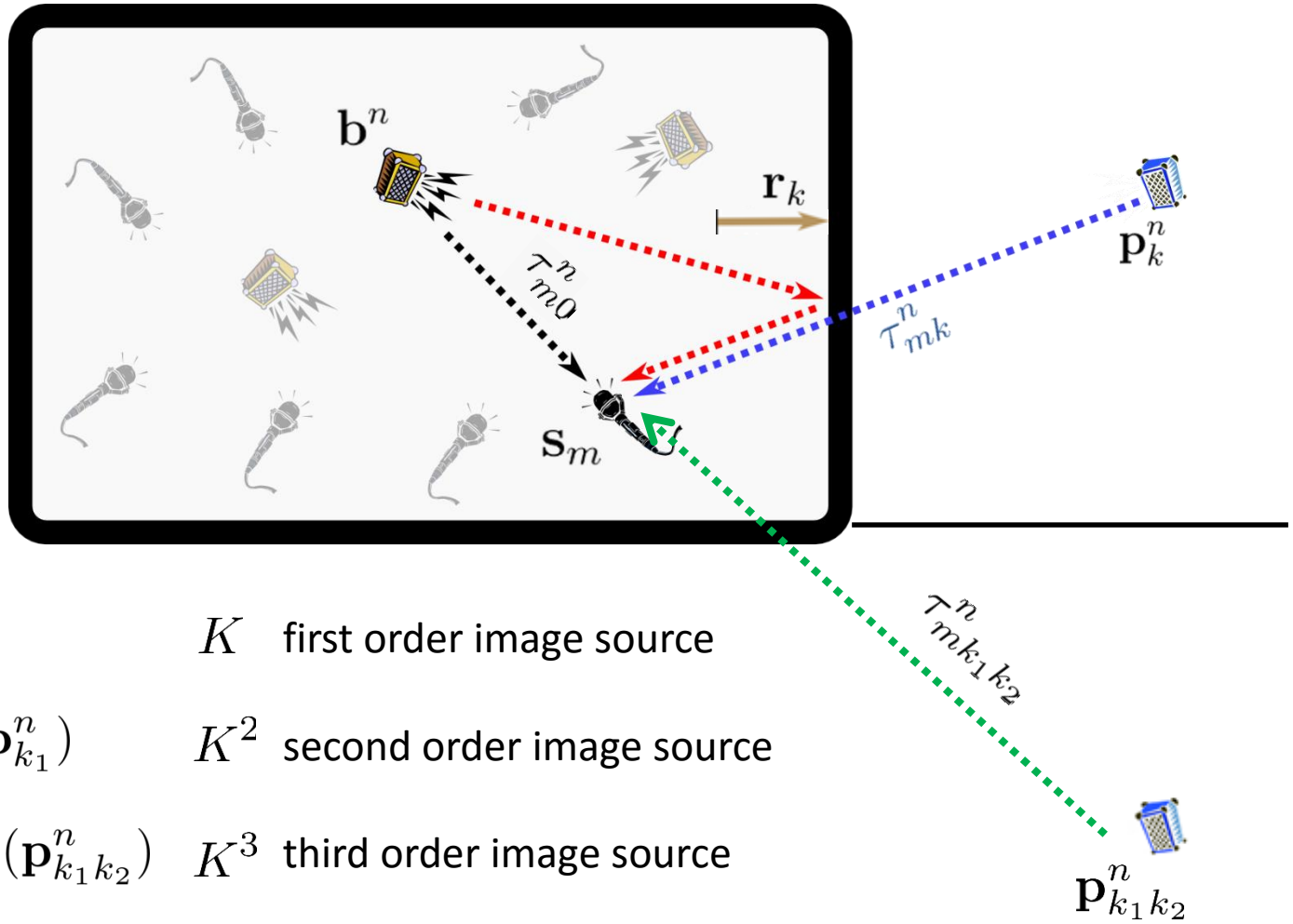
$\mathbf{p}_k^n$ is named as the Image Source since the planar surface act as an acoustic mirror for the real source. The 3D location of the virtual source can be expressed as:

$$\mathbf{p}_k^n = IM_k(\mathbf{b}^n) = \mathbf{b}^n + 2\left(1 - \frac{\mathbf{r}_k^\top \mathbf{b}^n}{\|\mathbf{r}_k\|_2^2}\right)\mathbf{r}_k$$

where $\mathbf{r}_k$ is the vector normal to the *k* planar surface and whose length is equal to the distance of the coordinate origin from the plane.

Higher order reflections

Each virtual source acts as a real source for the other walls generating **higher order virtual sources and relative reflections.**



$$\mathbf{p}^n_k = IM_k(\mathbf{b}^n) \qquad K \quad \text{first order image source}$$

$$\mathbf{p}^n_{k_1k_2} = IM_{k_2}(\mathbf{p}^n_{k_1}) \qquad K^2 \quad \text{second order image source}$$

$$\mathbf{p}^n_{k_1k_2k_3} = IM_{k_3}(\mathbf{p}^n_{k_1k_2}) \qquad K^3 \quad \text{third order image source}$$

RIR modeling with image sources

Each RIR can be modeled as a sequence of Dirac pulses, weighted by amplitude coefficients, whose delays correspond to the propagation delays from all the sources, either real or virtual, and a microphone.

$$h_m^{room,n}(t) = a_{m0}^n \delta(t - \tau_{m0}^n) + \sum a_{mk}^n \delta(t - \tau_{mk}^n) + \sum a_{mk_1k_2}^n \delta(t - \tau_{mk_1k_2}^n) + \cdots$$

Each propagation delay is known as the **Time of Flight (TOF)**

TOF direct path

$$\tau_{m0}^n = \frac{\|\mathbf{b}^n - \mathbf{s}_m\|_2}{c}$$

TOF 1st order reflection

$$\tau_{mk}^n = \frac{\|\mathbf{p}_k^n - \mathbf{s}_m\|_2}{c}$$

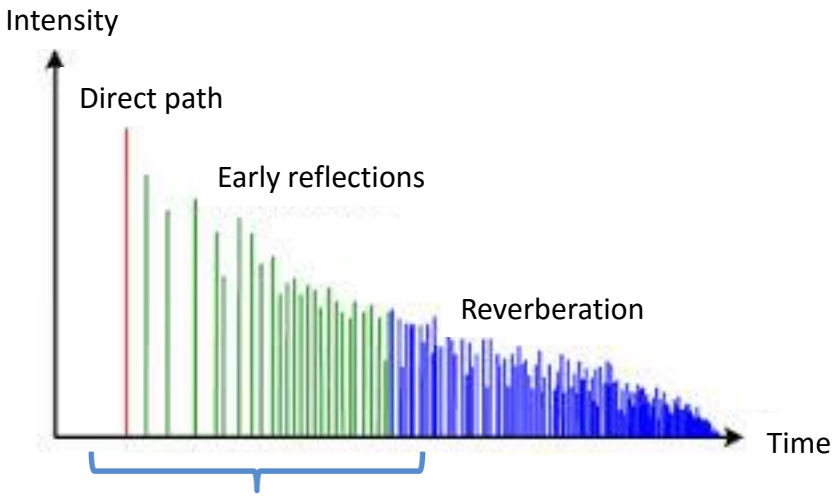
TOF 2nd order reflection

$$\tau_{mk_1k_2}^n = \frac{\|\mathbf{p}_{k_1k_2}^n - \mathbf{s}_m\|_2}{c}$$

Under this model every RIR is intrinsically sparse.

Image Source method validity

- The image source method is **exact only for specific geometries.**
- For a generic convex polyhedron Image Source method is an **approximate model** due to the **finite wall dimensions** implying border effects.
- Moreover **amplitude coefficients** are in general **frequency variant** and depending on the incident angle of the wave.
- Despite this mismatches between model and reality, the Image Source method is able to model remarkably well the first part of the RIR corresponding to the first reflections, i.e. its sparse portion.



Emission times and offset in reception

In general each source has an unknown **emission time** τ_e^n such that:

$$x^n(t) = \hat{x}^n(t - \tau_e^n)$$

Similarly, each microphone, even when synchronized with the other ones, can have an **offset in the reception clock:** τ_m^o embedded in its impulse response as follows:

$$h_m^{mic}(t) = h^{mic}(t - \tau_m^o)$$

Notice that, apart from the offset time, microphone impulse response is the same for all mics.

Final relation between TX and RX signals (1)

It is possible to incorporate emission and offset times into the RIR obtaining a new impulse response:

$$h_m^n(t) = h_m^{room,n}(t - \tau_e^n - \tau_m^o)$$

The microphone impulse response, supposed to be common to all the microphones except for the offset time, can be incorporated into the transmitted signal giving:

$$s^n(t) = h^{mic}(t) * \hat{x}^n(t)$$

obtaining the final relation:

$$y_m^n(t) = h_m^n(t) * s^n(t) + \nu_m(t)$$

Final relation between TX and RX signals (2)

Incorporating the microphone impulse response into the TX signals, except for the offset time allows to preserve the sparsity of $h_m^n(t)$:

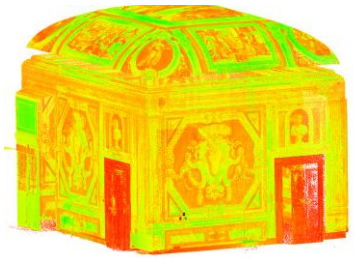
$$h_m^n(t) = \sum_{l=1}^{L} a_{ml}^n \delta(t - \tau_{ml}^{an})$$

Where τ_{ml}^{an} are the absolute Time of Arrival (TOA) of the signals following the different reflection paths. TOAs are defined as follows:

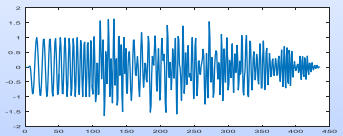
$$\tau_{ml}^{an} = \begin{cases} \tau_{m0}^n + \tau_e^n + \tau_m^o \\ \tau_{mk}^n + \tau_e^n + \tau_m^o \\ \tau_{mk_1k_2}^n + \tau_e^n + \tau_m^o \\ \vdots \end{cases}$$

Notice that the total number of TOAs is arbitrary since it is a truncation of an infinite set. Moreover, by convention, TOAs are increasing with the index *l.*

Estimating Room Geometry

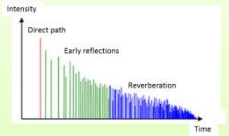


$$y_m^n(t) = h_m^n(t) * s^n(t) + \nu_m(t)$$



Signal acquisition

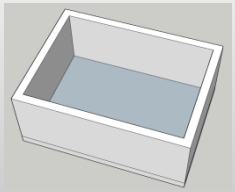
$$\{\tau_e^n\}\,\{\tau_m^o\}\{\tau_{m0}^n\}$$



Time of Flight (TOF) estimation

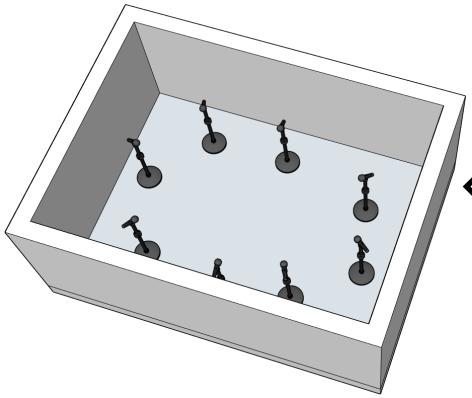
Mic/Sources calibration

$$\{\mathbf{s}_m\}\,\{\mathbf{b}^n\}$$



Room estimation

$$\{\mathbf{r}_k\}$$



NEXT: Time of Arrival estimation