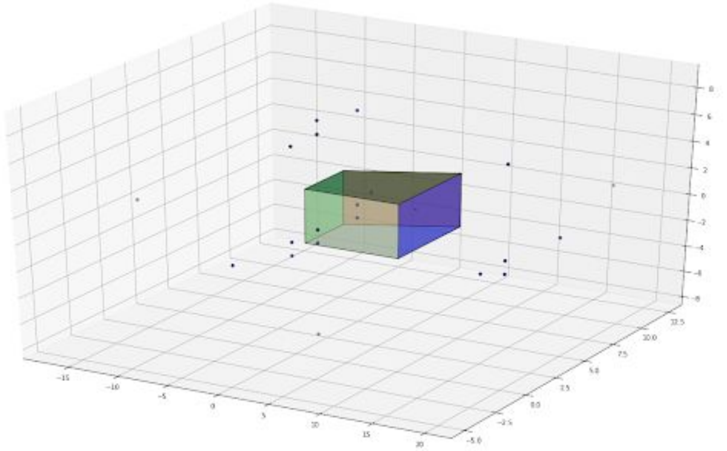
Sound Source Localization



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

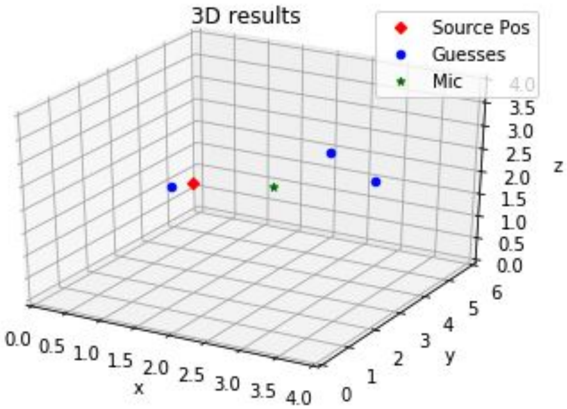
The Problem

Situation:

- Talker localization with microphone arrays
- Objective is to have a speech localization system with the **highest accuracy**.

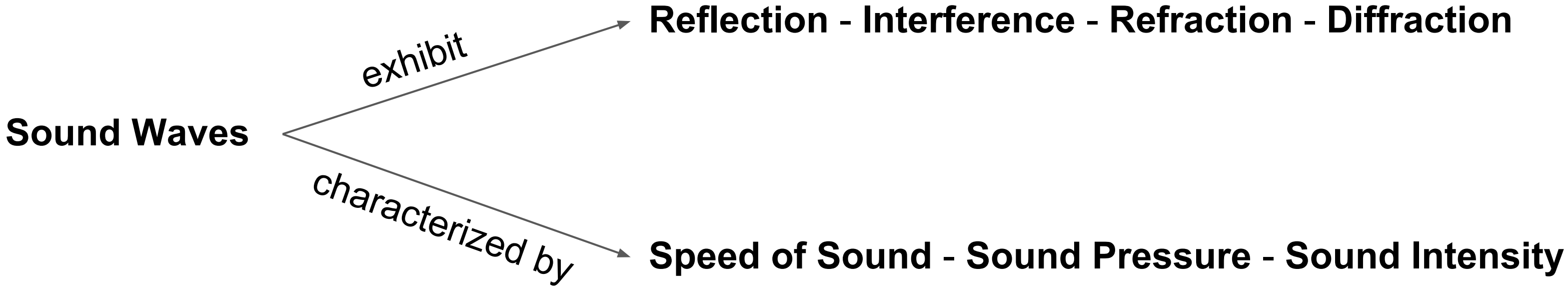
Assumptions:

- Position of microphones & room dimensions are known
- Closed room
- Everything is static



Sound propagation

- Sound propagates through air as a longitudinal wave
- The speed of sound is determined by the properties of the air.
- At 20°C sound propagates at a speed of 343.0 m/s



Sound Waves

exhibit

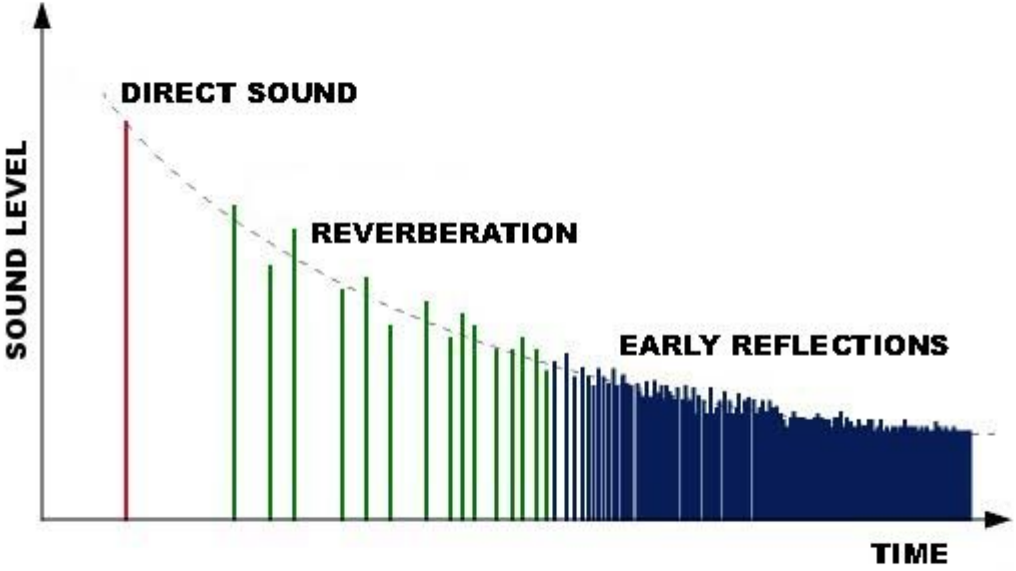
Reflection - Interference - Refraction - Diffraction

characterized by

Speed of Sound - Sound Pressure - Sound Intensity

Room Impulse Response

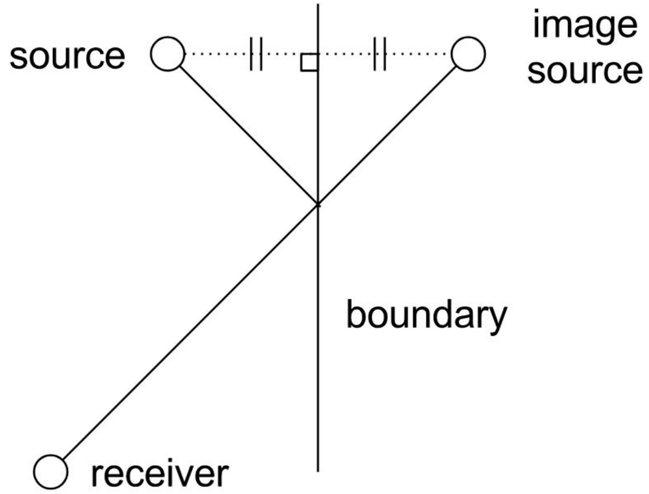
- Room impulse responses are transfer functions between the sound source and microphone.
- To recover the original sound source, the received microphone signal can be convolved with the inverse of the room impulse response function.
- Room impulse responses have the following components:
  - ➢ **Propagation delay** : the length in time the sound travels from the source to the listener
  - ➢ **Direct sound** : in the line of sight, the direct sound is a peak corresponding to the shortest travel path
  - ➢ **Early reflections part**
    - ○ First reflection (usually the reflection from the ground)
    - ○ Second and other reflections
  - ➢ **Reverberation Tail part** : more reflections still clearly distinguishable



Reference : https://commons.wikimedia.org/wiki/File:Acoustic_room_impulse_response.jpeg

Image Source Model

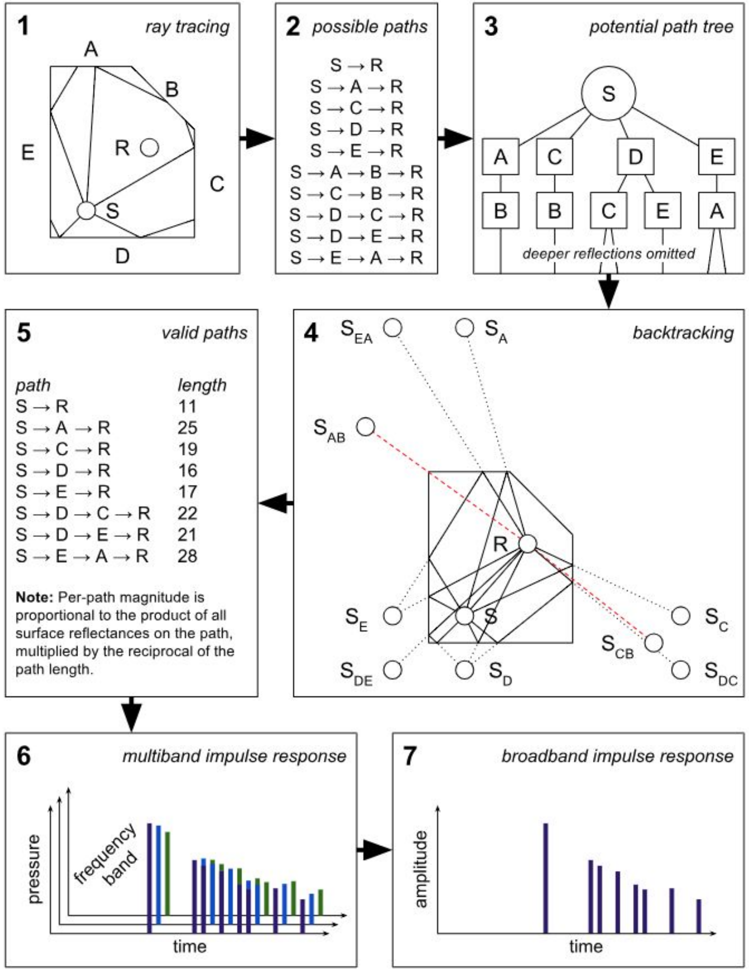
- Aims at finding the purely **specular reflection** paths between a source and a receiver.
- Assumes that sound propagates only along straight lines or **rays**.
- Sound energy travels at a **fixed speed**, corresponding to the speed of sound, along these rays.
- The energy in each ray **decreases with $1/r^2$**, where r is the total distance that the ray has travelled



Reference : https://reuk.github.io/wayverb/image_source.html

Implementation of the Image-Source Method

- Large number of random rays are traced from the source
- At each reflection point, the receiver is checked to see whether it is visible
- The complexity of ray tracing grows linearly rather than exponentially with reflection depth
- Each unique path found in this way is used to generate an image source sequence
- The majority of surface sequences are not checked, so the Image Source process is fast



2 possible paths

S → R
S → A → R
S → C → R
S → D → R
S → E → R
S → A → B → R
S → C → B → R
S → D → C → R
S → D → E → R
S → E → A → R

5 valid paths

path	length
S → R	11
S → A → R	25
S → C → R	19
S → D → R	16
S → E → R	17
S → D → C → R	22
S → D → E → R	21
S → E → A → R	28

Note: Per-path magnitude is proportional to the product of all surface reflectances on the path, multiplied by the reciprocal of the path length.

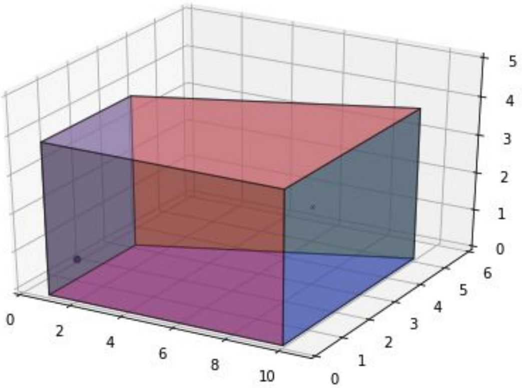
Reference : https://reuk.github.io/wayverb/image_source.html

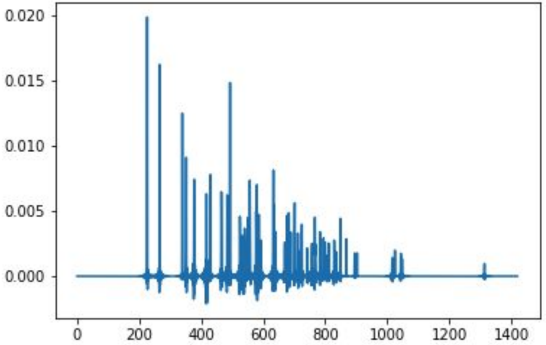
Pyroomacoustics

‘A software package for rapid development and testing of audio array processing algorithms’

Three main components

- A Python object-oriented interface to construct different simulation scenarios in 2D and 3D rooms
- Image Source Model implementation for general polyhedral rooms to generate room impulse responses and simulate the propagation between sources and receivers
- Reference implementations of popular algorithms for beamforming, direction finding, and adaptive filtering (SRP, MUSIC, CSSM, WAVES, FRIDA)





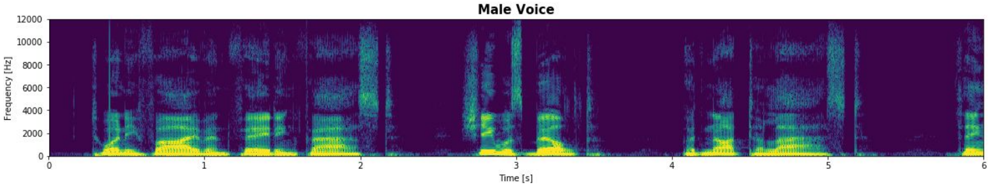
Analysis of the Spectrogram

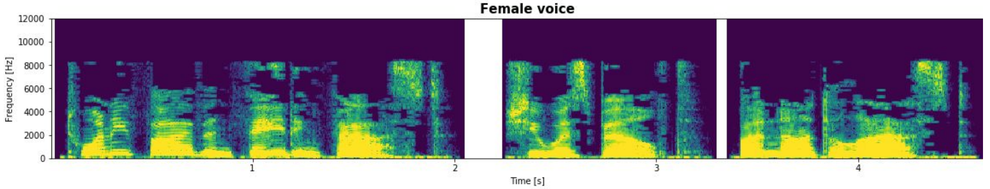
A **spectrogram** is a visual representation of the spectrum of frequencies of sound as they vary with time

Man vs Woman

- Women's voice is of high frequencies, it reaches frequencies up to 8000 Hz but is mainly around 2000 Hz

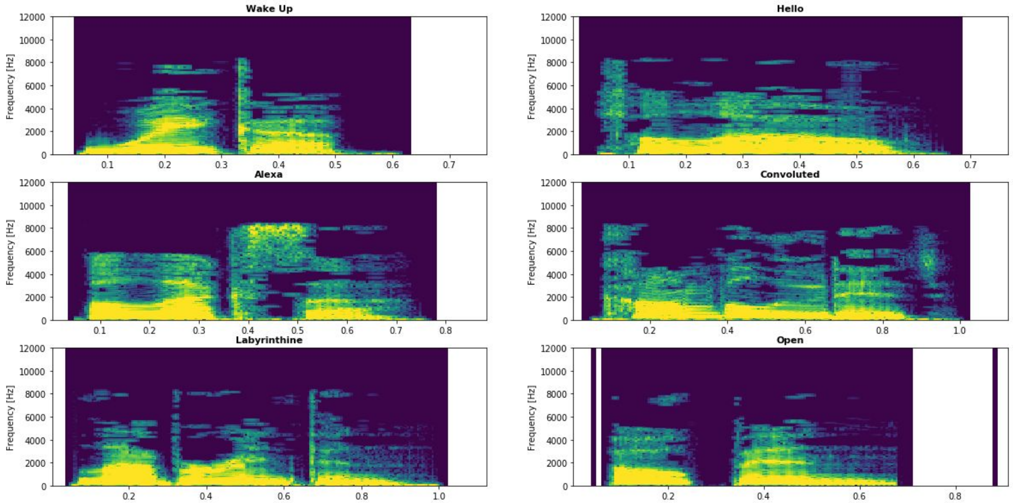
- Men's voice is of lower frequencies, it reaches up to 4000 Hz and is mainly around 800 Hz


Male Voice


Female voice

Wake Up Words

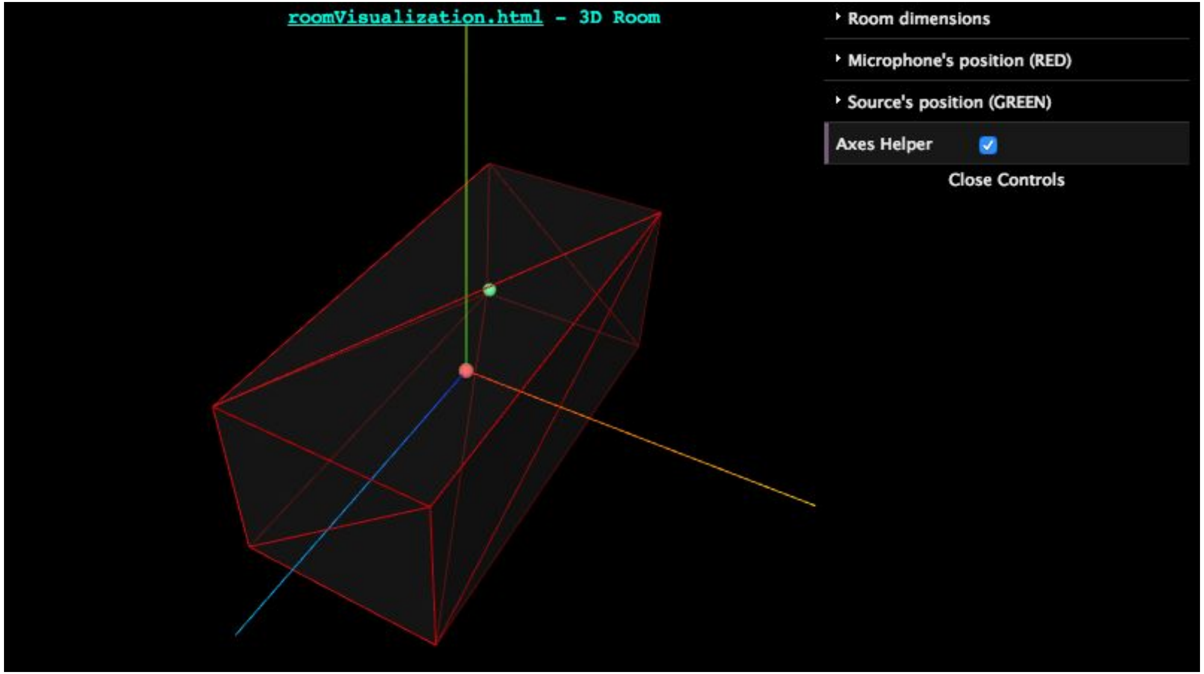
- Consonants like ***c***, ***h*** and ***x*** have high frequencies
- Vowels like ***o*** have lower frequencies
- There are similarities between the vowels and consonants spectrogram : we can observe that the spectrogram of ***c*** contains a part of the spectrogram of ***e***, this because when you pronounce ***c*** you are also pronouncing the letter ***e***
- *Wobble* sounds are the carrier of the words in general



Spectrogram of different wake up words pronounced by a woman

An Interactive Platform

An interactive platform where users can choose between different SSL algorithms and play some sort of hide and seek game where we need to figure out what was the location of the sound source fixed by the user.



roomVisualization.html - 3D Room

Sound Source Localization Techniques

Three types of sound source localization approaches:

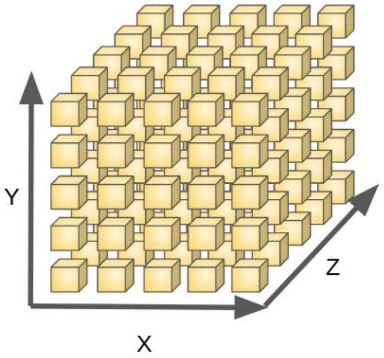
- TDOA-Based Locators
- High-Resolution Spectral-Estimation-Based Locators
- Steered-Beamformer-Based Locators

A Failed Attempt

SSL using grid search

- Create a grid in the room where each element represents a sound source position. For each of these position compute the RIR between the sound source and the fixed microphone

- Compute the RIR of the unknown new position and perform a cross correlation with all the different RIRs in the above dictionary and select the most similar ones.



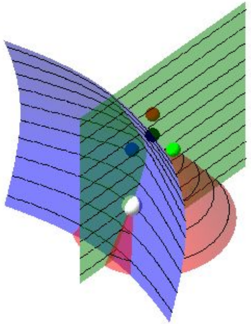
Because of the **curse of dimensionality**, this method has no intuition and fails miserably at performing correctly !

~~SSL using Grid Search~~

TDOA-Based SSL

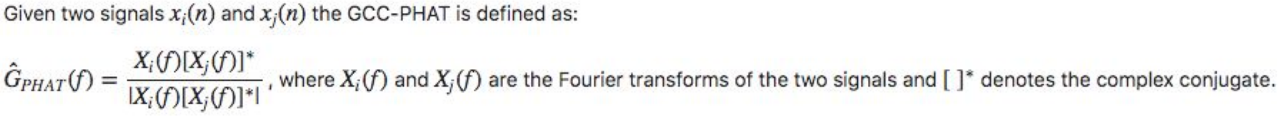
Two-step procedure

- Time delay estimation (**TDE**) of the speech signals relative to pairs of spatially separated microphones is performed.
- Microphone positions are then used to generate hyperbolic curves which are then intersected in some optimal sense to arrive at a source location estimate



⇨ **TDE** is the key to the effectiveness of localizers within this genre

↪ Estimation is done using **GCC-PHAT** Cross-Correlation



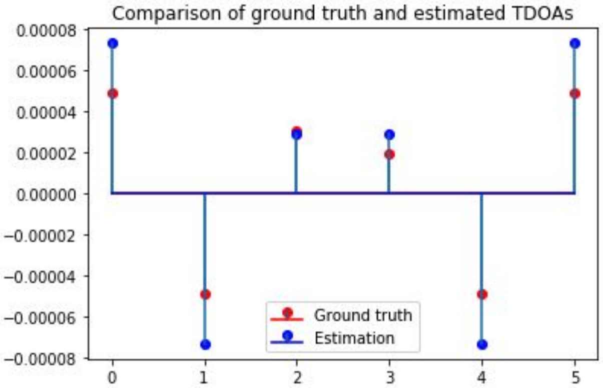
Given two signals $x_i(n)$ and $x_j(n)$ the GCC-PHAT is defined as:

$$\hat{G}_{PHAT}(f) = \frac{X_i(f)[X_j(f)]^*}{|X_i(f)[X_j(f)]^*|},$$ where $X_i(f)$ and $X_j(f)$ are the Fourier transforms of the two signals and $[\ ]^*$ denotes the complex conjugate.

Example : ML-TDOA Based SSL

Goal : find the position ***q*** that minimizes the ***Least Square Criterion :*** $E(\boldsymbol{q}) = \sum_{i=1}^{M} (\hat{\tau}_i - T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}))^2.$

- The TDOA estimate $\boldsymbol{\tau}_{\mathbf{i}}$ is computed using the GCC-PHAT
- The function ***T()*** represents the ground truth TDOA between the microphones at position $\mathbf{p}_{\mathbf{i1}}$ and $\mathbf{p}_{\mathbf{i2}}$ and the source is at position **q**
- The location estimate is then found from : $\hat{\boldsymbol{q}}_s = \underset{q}{\operatorname{argmin}}\, E(\boldsymbol{q})$



SRP-PHAT SSL

- Localizes a single source from a frame of data received at the microphones
- Uses the Steered Response Power with the PHAse Transform (**SRP-PHAT**) as the functional
- The true source location will have the **maximum SRP-PHAT value**

SRP-PHAT algorithm consists in a grid-search procedure that evaluates the objective function on a grid G of candidate source locations to estimate the spatial location of the sound source, as the point of the grid that provides the maximum SRP:

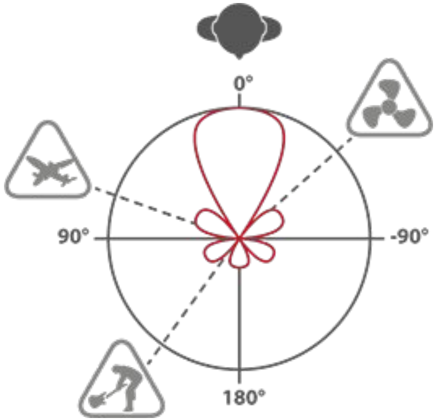
$$\hat{\mathbf{x}}_s = \arg\max_{\mathbf{x}\in\mathcal{G}} P(\mathbf{x}).$$

Steered-Beamformer-Based Locators

- Focused beamformer which steers the array to various locations and **searches for a peak** in output power

- Steered response obtained using output of a delay-and-sum beamformer (i.e. a conventional beamformer)

↪ Apply time shifts to the array signals to compensate for the propagation delays in the arrival of the source signal at each microphone



Next Steps

- Benchmark algorithms
- Experiments using new setups; polyhedral rooms (source behind a wall..)
- Try implementing the WASPAA paper
- Create visualization of the SSL algorithms
- Try to run everything interactively