# ECHO-AWARE signal processing for audio scene analysis

Diego DI CARLO
November 25, 2020

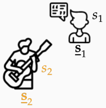
PhD Director: Nancy BERTIN
PhD Supervisor: Antoine DELEFORGE
Jury members: Laurent GIRIN (reviewer)
Simon DOCLO (reviewer)
Fabio ANTONACCI
Renaud SEGUIER
Collaborators: Clément ELVIRA,
Robin SCHEIBLER, Ivan DOKMANIĆ,
Sharon GANNOT, Pinchas TANDEITNIK

Université de Rennes 1, IRISA/INRIA, Panama research group

# Overview

## Echo-aware signal processing for audio scene analysis

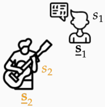
# Introduction

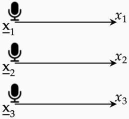



Sound

- produced by sources

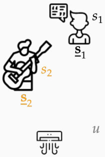
# Scenario

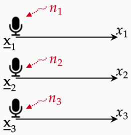



Sound

- produced by sources
- recorded by microphones

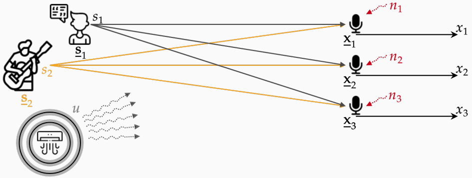



## Sound

- produced by sources
- recorded by microphones
- corrupted by noise

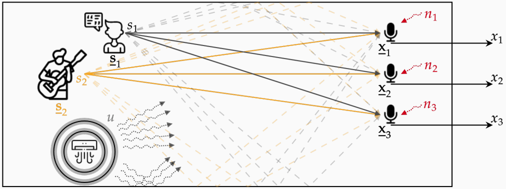
# Scenario




Sound

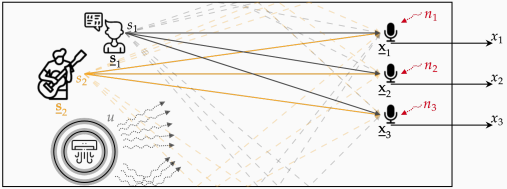
- produced by sources
- recorded by microphones
- corrupted by noise
- propagates in the space

# Scenario




Sound

- produced by sources
- recorded by microphones
- corrupted by noise
- propagates in the room
  ↪ reverberation

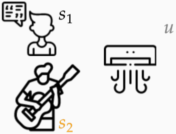



Sound

- produced by sources
- recorded by microphones
- corrupted by noise
- propagates in the room
  ↪ reverberation

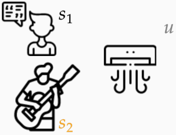
Attention: artificial sound vs **(natural) microphone recordings**

Semantic information



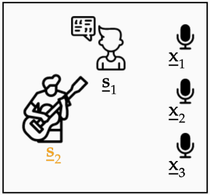

on nature and content

Semantic information




on nature and content

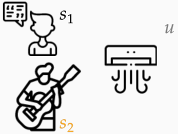
Spatial information




on position and geometry

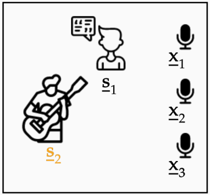
# Echo-aware signal processing for audio scene analysis

Semantic information



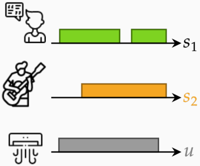

on nature and content

Spatial information



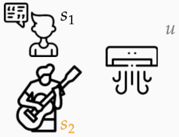

on position and geometry

Temporal information



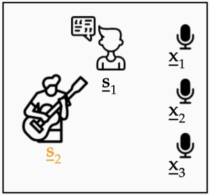

on events activity

Semantic information



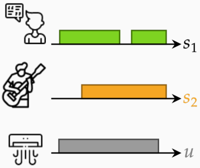

on nature and content

Spatial information




on position and geometry

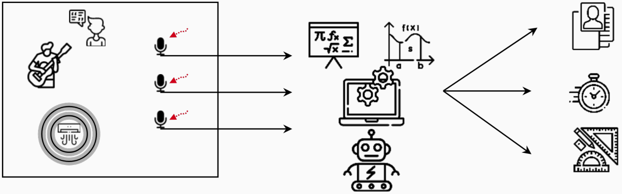
Temporal information



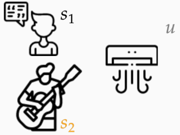

on events activity

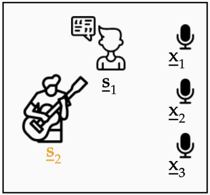
**Audio Scene Analysis**

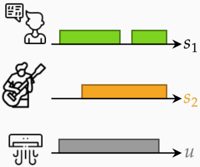
Extraction and organization of all the information in the sound



| Semantic information | Spatial information | Temporal information |
|---|---|---|



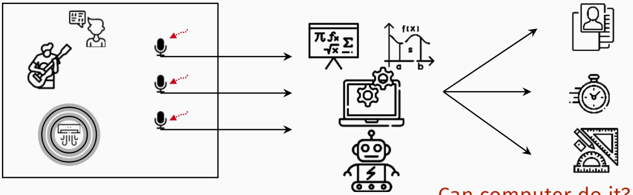



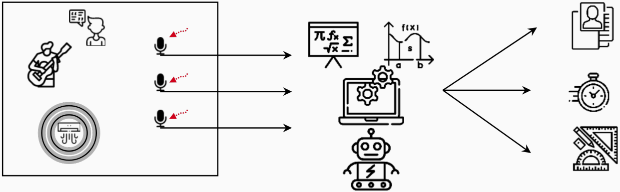

| on nature and content | on position and geometry | on events activity |
|---|---|---|

**Audio Scene Analysis**

Extraction and organization of all the information in the sound

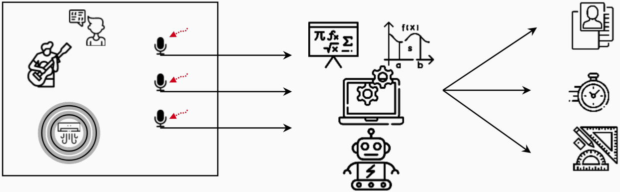


Can computer do it?



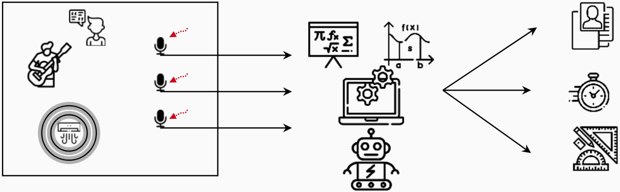
f(x)
s
a
b

# Echo-aware signal processing for audio scene analysis




**Signal Processing**

Mathematical models, frameworks and tools to tackle and solve such problems

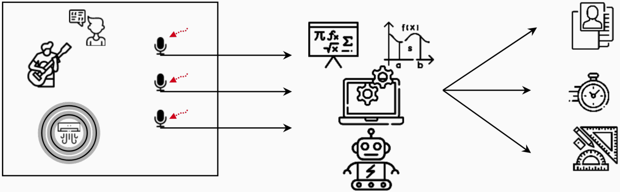



## Signal Processing

Mathematical models, frameworks and tools to tackle and solve such problems

Some (inverse) problems

- Speaker Identification
- Sound Source Separation (SSS)
- Speech Enhancement (SE)
- Automatic Speech Recognition (ASR)
- Sound Source Localization (SSL)
- Room Geometry Estimation (RooGE)
- Voice Activity Detection
- Diarization
- $\mathrm{RT}_{60}$ estimation
- Acoustic Channel Estimation
- Wall Absorption Estimation
- *and many many other*




## Signal Processing

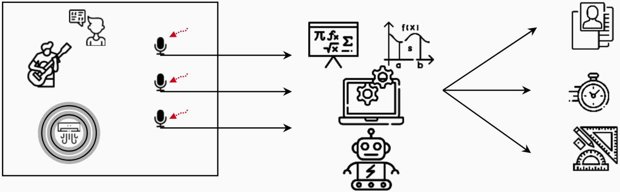
Mathematical models, frameworks and tools to tackle and solve such problems

Some (inverse) problems

- Speaker Identification } Who?
- Sound Source Separation (SSS)
- Speech Enhancement (SE)
- Automatic Speech Recognition (ASR) } What?
- Sound Source Localization (SSL)
- Room Geometry Estimation (RooGE) } Where?

- Voice Activity Detection
- Diarization } When?
- $RT_{60}$ estimation
- Acoustic Channel Estimation
- Wall Absorption Estimation } How?
- *and many many other*

# Echo-aware signal processing for audio scene analysis



**Signal Processing**

Mathematical models, frameworks and tools to tackle and solve such problems

Some (inverse) problems

- Speaker Identification
- Sound Source Separation (SSS)
- Speech Enhancement (SE)
- Automatic Speech Recognition (ASR)
- Sound Source Localization (SSL)
- Room Geometry Estimation (RooGE)

Who?

What?

Where?

- Voice Activity Detection
- Diarization
- $\mathrm{RT}_{60}$ estimation
- Acoustic Channel Estimation
- Wall Absorption Estimation
- *and many many other*

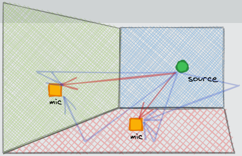
When?

How?

HOW $\xrightarrow{\text{helps}}$ WHERE $\xrightarrow{\text{helps}}$ WHEN $\xrightarrow{\text{helps}}$ WHAT $\xrightarrow{\text{helps}}$ HOW $\xrightarrow{\text{helps}}$ ...

## Acoustic Echoes

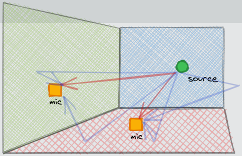
- Elements of the sound propagation
- Standing out for time and strength
- Repetition of a sound but later
- Both outdoor and indoor




Audio signal processing methods

### Acoustic Echoes

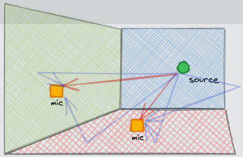
- Elements of the sound propagation
- Standing out for time and strength
- Repetition of a sound but later
- Both outdoor and indoor




Audio signal processing methods

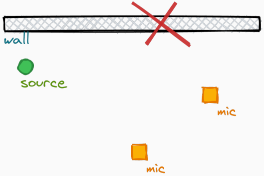
## Acoustic Echoes

- Elements of the sound propagation
- Standing out for time and strength
- Repetition of a sound but later
- Both outdoor and indoor



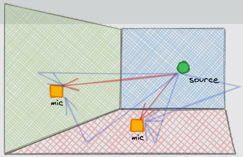

Audio signal processing methods

- ignore it



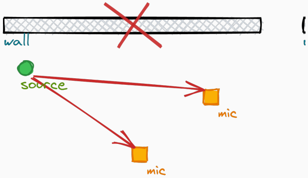
## Acoustic Echoes

- Elements of the sound propagation
- Standing out for time and strength
- Repetition of a sound but later
- Both outdoor and indoor



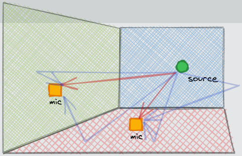

### Audio signal processing methods

- ignore it
- assume it free-field



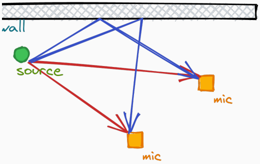
## Acoustic Echoes

- Elements of the sound propagation
- Standing out for time and strength
- Repetition of a sound but later
- Both outdoor and indoor



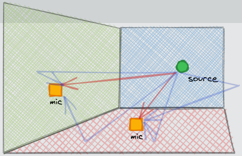

### Audio signal processing methods

- ignore it
- assume it free-field
- model it entirely



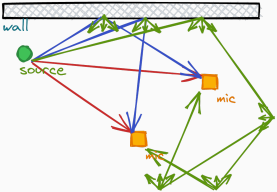
## Acoustic Echoes

- Elements of the sound propagation
- Standing out for time and strength
- Repetition of a sound but later
- Both outdoor and indoor



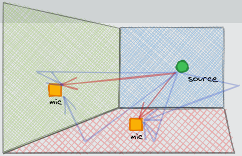

Audio signal processing methods

- ignore it
- assume it free-field
- model it entirely
- model as few reflection



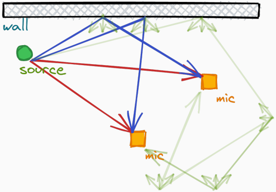
## Acoustic Echoes

- Elements of the sound propagation
- Standing out for time and strength
- Repetition of a sound but later
- Both outdoor and indoor




Audio signal processing methods

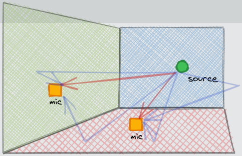
- ignore it
- assume it free-field
- model it entirely
- model as few reflection
- model it as early and late parts



# Echo-aware signal processing for audio scene analysis

## Acoustic Echoes

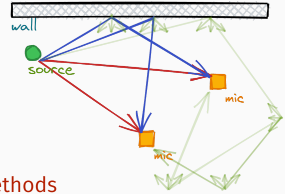
- Elements of the sound propagation
- Standing out for time and strength
- Repetition of a sound but later
- Both outdoor and indoor




Audio signal processing methods

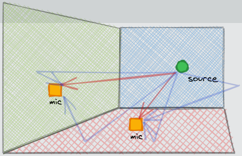
- ignore it
- assume it free-field
- model it entirely
- model as few reflection
- model it as early and late parts

Echo-aware methods



## Acoustic Echoes

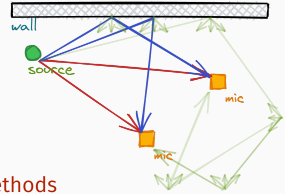
- Elements of the sound propagation
- Standing out for time and strength
- Repetition of a sound but later
- Both outdoor and indoor




### Audio signal processing methods

- ignore it
- assume it free-field
- model it entirely
- model as few reflection
- model it as early and late parts

Echo-aware methods




### Modelling the sound field

- as free field $\Longrightarrow$ reverberation is noise
- entirely $\Longrightarrow$ it is very challenging

# Goals and contributions

Audio Scene Analysis

↓

context and problems

# Goals and contributions

Audio Scene Analysis
↓
context and problems

Signal Processing
↓
models and frameworks

# Goals and contributions

| Audio Scene Analysis | Signal Processing | Acoustic Echoes |
| --- | --- | --- |
| ↓ | ↓ | ↓ |
| context and problems | models and frameworks | better processing |

# Goals and contributions

| Audio Scene Analysis | Signal Processing | Acoustic Echoes |
|---|---|---|
| ↓ | ↓ | ↓ |
| context and problems | models and frameworks | better processing |

**Goals**

1. How to **estimate** acoustic echoes?
2. How to **extend methods** for echo-aware audio scene analysis

# Goals and contributions

| Audio Scene Analysis | Signal Processing | Acoustic Echoes |
|---|---|---|
| ↓ | ↓ | ↓ |
| context and problems | models and frameworks | better processing |

**Goals**

1. How to **estimate** acoustic echoes?
2. How to **extend methods** for echo-aware audio scene analysis

1. Estimation

2. Application

# Goals and contributions

| Audio Scene Analysis | Signal Processing | Acoustic Echoes |
|---|---|---|
| ↓ | ↓ | ↓ |
| context and problems | models and frameworks | better processing |

**Goals**

1. How to **estimate** acoustic echoes?
2. How to **extend methods** for echo-aware audio scene analysis

### 1. Estimation

- Knowledge-based echo estimation
  ↪ `Blaster`
- Learning-based echo estimation
  ↪ `Lantern`

### 2. Application

# Goals and contributions

| Audio Scene Analysis | Signal Processing | Acoustic Echoes |
| --- | --- | --- |
| ↓ | ↓ | ↓ |
| context and problems | models and frameworks | better processing |

**Goals**

1. How to **estimate** acoustic echoes?
2. How to **extend methods** for echo-aware audio scene analysis

### 1. Estimation

- Knowledge-based echo estimation
  ↪ `Blaster`
- Learning-based echo estimation
  ↪ `Lantern`

### 2. Application

- Echo-aware Source Separation
  ↪ `Separake`
- Echo-aware Source Localization
  ↪ `Mirage`
- Echo-aware Speech Enhancement
- Echo-aware Room Geometry Estimation

# Goals and contributions

Audio Scene Analysis
↓
context and problems

Signal Processing
↓
models and frameworks

Acoustic Echoes
↓
better processing

**Goals**

1. How to **estimate** acoustic echoes?
2. How to **extend methods** for echo-aware audio scene analysis

### 1. Estimation

- Knowledge-based echo estimation
  ↪ `Blaster`
- Learning-based echo estimation
  ↪ `Lantern`

### 2. Application

- Echo-aware Source Separation
  ↪ `Separake`
- Echo-aware Source Localization
  ↪ `Mirage`
- Echo-aware Speech Enhancement
- Echo-aware Room Geometry Estimation

### 3. Data:

Echo-aware database → `dEchorate`

# Modeling

# Acoustic Impulse Response

Sound interacts with environment

- it is reflected (specularly and diffusely)
- \+ it is diffracted
- \+ it is absorbers and transmitted
- \+ other physical interaction

# Acoustic Impulse Response

Sound interacts with environment

$$\left.\begin{array}{ll} & \text{it is reflected (specularly and diffusely)} \\ + & \text{it is diffracted} \\ + & \text{it is absorbers and transmitted} \\ + & \text{other physical interaction} \end{array}\right\} = \text{all sound propagation}$$

# Acoustic Impulse Response

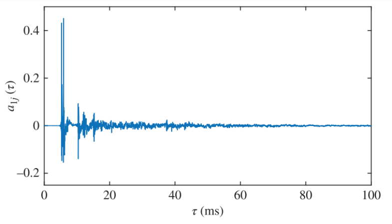
Sound interacts with environment

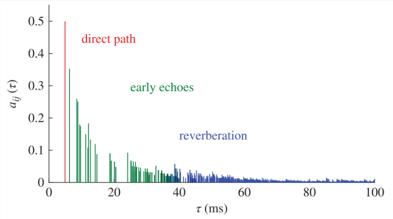
$$\left.\begin{array}{ll} & \text{it is reflected (specularly and diffusely)} \\ + & \text{it is diffracted} \\ + & \text{it is absorbers and transmitted} \\ + & \text{other physical interaction} \end{array}\right\} = \text{all sound propagation}$$

**Sound propagation** process = **source** $\rightarrow$ **filter** $\rightarrow$ **receiver** process

$$x(t) = (a * s)(t)$$ $\leftarrow$ continuous time domain!

The filter $a(t)$ is linear and is called **Acoustic Impulse Response**, (AIR)



# Acoustic Impulse Response

Sound interacts with environment

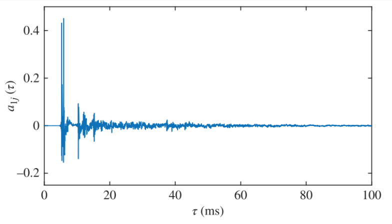
$$\left.\begin{array}{l} \text{it is reflected (specularly and diffusely)} \\ + \text{ it is diffracted} \\ + \text{ it is absorbers and transmitted} \\ + \text{ other physical interaction} \end{array}\right\} = \text{all sound propagation}$$

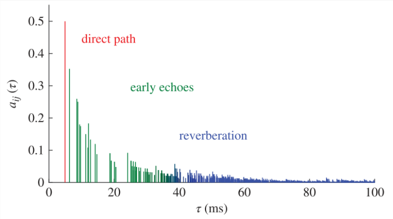
Indoor **Sound propagation** process = **source** $\rightarrow$ **filter** $\rightarrow$ **receiver** process

$$x(t) = (h * s)(t)$$

$\leftarrow$ continuous time domain!

The filter $h(t)$ is linear and is called Room **Impulse Response**, (RIR)



# Acoustic Impulse Response

Sound interacts with environment

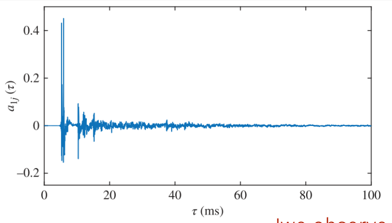
$$\left.\begin{array}{l} \text{it is reflected (specularly and diffusely)} \\ + \text{ it is diffracted} \\ + \text{ it is absorbers and transmitted} \\ + \text{ other physical interaction} \end{array}\right\} = \text{all sound propagation}$$

Indoor **Sound propagation** process = **source** $\rightarrow$ **filter** $\rightarrow$ **receiver** process

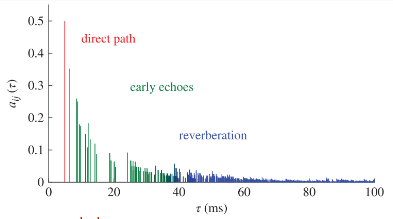
$$x(t) = (h * s)(t)$$

$\leftarrow$ continuous time domain!

The filter $h(t)$ is linear and is called Room **Impulse Response**, (RIR)




$\overset{?}{\approx}$

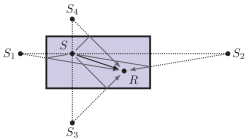



!we observe filters, not delays

# Echoes and Room Impulse Response

RIRs can be modeled with the Image Methods

- specular reflection only
- for cuboid room, **it describes** the sound prop.
- in general, **well models the first part**.




*"playing billiard in a concert hall"*

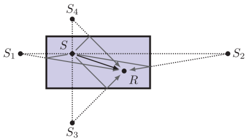
# Echoes and Room Impulse Response

RIRs can be modeled with the Image Methods

- specular reflection only
- for cuboid room, **it describes** the sound prop.
- in general, **well models the first part**.

$$h_i^e(t) = \sum_{r=0}^{R} \alpha_i^{(r)} \delta(t - \tau_i^{(r)})$$

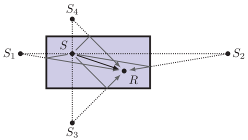
sum of Dirac's delta




*"playing billiard in a concert hall"*

RIRs can be modeled with the Image Methods

- specular reflection only
- for cuboid room, **it describes** the sound prop.
- in general, **well models the first part**.




*"playing billiard in a concert hall"*

$$h_i^e(t) = \sum_{r=0}^{R} \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) \quad \xrightarrow{\text{more realistic}} \quad H_i^e(f) = \sum_{r=0}^{R} \alpha_i^{(r)}(f) \mathrm{e}^{-\mathrm{i}2\pi f \tau_i^{(r)}}$$

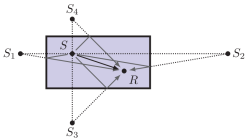
sum of Dirac's delta

sum of filters

# Echoes and Room Impulse Response

RIRs can be modeled with the Image Methods

- specular reflection only
- for cuboid room, **it describes** the sound prop.
- in general, **well models the first part**.




*"playing billiard in a concert hall"*

$$h_i^e(t) = \sum_{r=0}^{R} \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) \quad \xrightarrow{\text{more realistic}} \quad H_i^e(f) = \sum_{r=0}^{R} \alpha_i^{(r)}(f) \mathrm{e}^{-\mathrm{i}2\pi f \tau_i^{(r)}}$$

sum of Dirac's delta

sum of filters

RIRs accounts for

the **geometry** of the room

- Room shape and size
- Mic and Source position
- other objects (eg. reflectors)

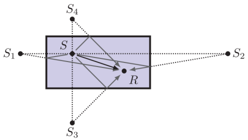
the **acoustic properties** of

- surface materials
- objects materials

# Echoes and Room Impulse Response

RIRs can be modeled with the Image Methods

- specular reflection only
- for cuboid room, **it describes** the sound prop.
- in general, **well models the first part**.




*"playing billiard in a concert hall"*

$$h_i^e(t) = \sum_{r=0}^{R} \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) \quad \xrightarrow{\text{more realistic}} \quad H_i^e(f) = \sum_{r=0}^{R} \alpha_i^{(r)}(f) \mathrm{e}^{-\mathrm{i} 2\pi f \tau_i^{(r)}}$$

sum of Dirac's delta

sum of filters

RIRs accounts for

the **geometry** of the room

- Room shape and size
- Mic and Source position
- other objects (eg. reflectors)

the **acoustic properties** of

- surface materials
- objects materials

**Echoes**

strong and distinct specular reflection

# Acoustic Echo Estimation

## The acoustic echoes retrieval (AER) problem

Estimating early (strong) acoustic reflections:

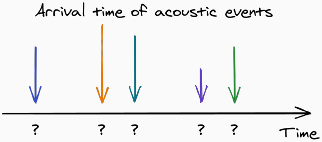
- their time of arrivals $\rightarrow$ TOAs Estimation
- their amplitude

# Acoustic Echo Retrieval

## The acoustic echoes retrieval (AER) problem

Estimating early (strong) acoustic reflections:

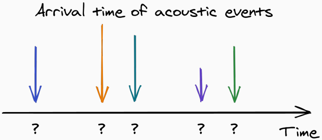
- their time of arrivals → TOAs Estimation
- their amplitude



## The acoustic echoes retrieval (AER) problem

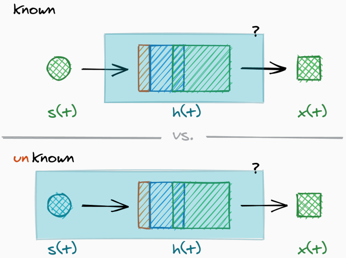
Estimating early (strong) acoustic reflections:

- their time of arrivals → TOAs Estimation
- their amplitude




Approaches

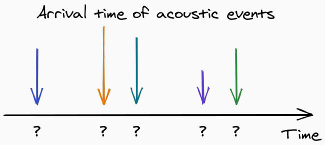
Source signal is



## The acoustic echoes retrieval (AER) problem

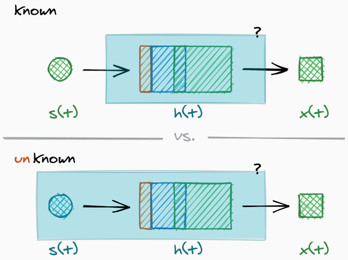
Estimating early (strong) acoustic reflections:

- their time of arrivals → TOAs Estimation
- their amplitude




Approaches

Source signal is




### Active approaches

- **easier** problem
- **intrusive** or specific setup
- Time of Arrival accessible
- single mic

Application Sonar, Calibration, Measurements

### Passive approaches

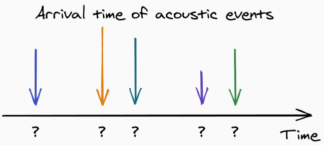
- more difficult problem — blind inverse problem
- passive hearing
- Time Difference of Arrivals only
- only multi-microphones

Application smart speakers, robot

## The acoustic echoes retrieval (AER) problem

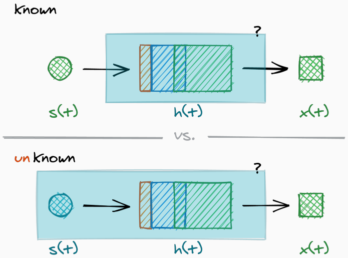
Estimating early (strong) acoustic reflections:

- their time of arrivals → TOAs Estimation
- their amplitude

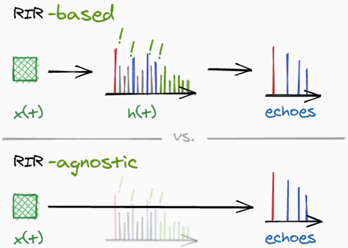



### Approaches

Source signal is



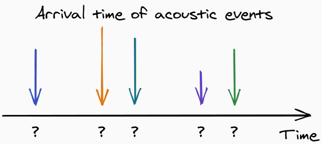

Estimation is



**The acoustic echoes retrieval (AER) problem**

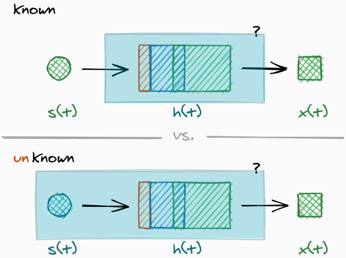
Estimating early (strong) acoustic reflections:

- their time of arrivals → TOAs Estimation
- their amplitude

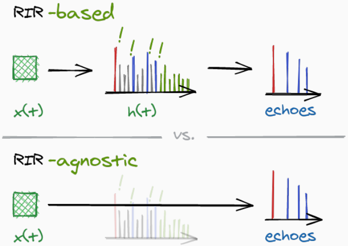



Approaches

Source signal is




Estimation is



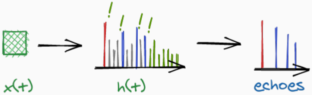

**Scenario:** signal source, only TOAs and passive system

Passive Acoustic Echo Estimation:

# Passive Acoustic Echo Estimation

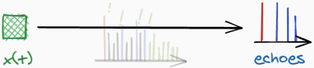
## Passive Acoustic Echo Estimation:

RIR-based approaches




1. SIMO BCE problem $\implies$ RIRs
2. Peak picking $\implies$ Echoes

RIR-agnostic approaches

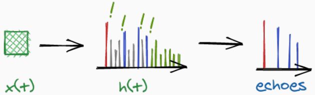



1. Estimation in the space $\left\{\tau_i^{(r)}, \alpha_i^{(r)}\right\}_{i,r}$

   (+ direction of arrivals can be used instead)

## Passive Acoustic Echo Estimation:

### RIR-based approaches

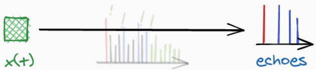



1. SIMO BCE problem $\implies$ RIRs
2. Peak picking $\implies$ Echoes

Pros

- SIMO BCE is well studied
- Good in some scenario

### RIR-agnostic approaches




1. Estimation in the space $\left\{\tau_i^{(r)}, \alpha_i^{(r)}\right\}_{i,r}$

   (+ direction of arrivals can be used instead)

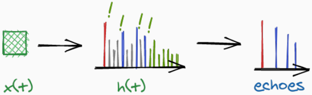
Pros

- No need for full RIRs
- Sub-sampling accuracy
- Low complexity
- Sparsity and Non-negativity are respected

# Passive Acoustic Echo Estimation

## Passive Acoustic Echo Estimation:

### RIR-based approaches




1. SIMO BCE problem $\implies$ RIRs
2. Peak picking $\implies$ Echoes

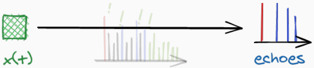
Pros

- SIMO BCE is well studied
- Good in some scenario

Cons

- Full RIR is estimated
- Peak picking
- on-grid estimation

### RIR-agnostic approaches




1. Estimation in the space $\left\{\tau_i^{(r)}, \alpha_i^{(r)}\right\}_{i,r}$

   (+ direction of arrivals can be used instead)

Pros

- No need for full RIRs
- Sub-sampling accuracy
- Low complexity
- Sparsity and Non-negativity are respected

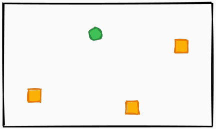
Cons

- echo sorting
- exploratory

## AER as discrete SIMO BCE

Key ingredient – *Cross relation identity*

$$x_i = h_i * s$$

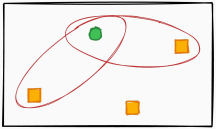


# AER as discrete SIMO BCE

Key ingredient – *Cross relation identity*

$$x_i = h_i * s$$

$$h_2 * x_1 = h_2 * h_1 * s = h_1 * h_2 * s = h_1 * x_2$$

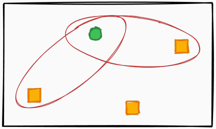


## AER as discrete SIMO BCE

Key ingredient – *Cross relation identity*

$$x_i = h_i * s$$

$$h_2 * x_1 = h_2 * h_1 * s = h_1 * h_2 * s = h_1 * x_2$$



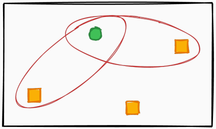
Ideas:

1. Sampled version of $x_1, x_2$ are available $(\mathbf{x_1}, \mathbf{x_2})$

# AER as discrete SIMO BCE

Key ingredient – *Cross relation identity*

$$x_i = h_i * s$$



$$h_2 * x_1 = h_2 * h_1 * s = h_1 * h_2 * s = h_1 * x_2$$

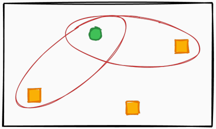
Ideas:

1. Sampled version of $x_1, x_2$ are available $(\mathbf{x}_1, \mathbf{x}_2)$
2. echoes' TOAs $\propto$ sampling frequency

# AER as discrete SIMO BCE

Key ingredient – *Cross relation identity*

$$x_i = h_i * s$$



$$h_2 * x_1 = h_2 * h_1 * s = h_1 * h_2 * s = h_1 * x_2$$

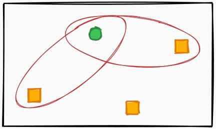
Ideas:

1. Sampled version of $x_1, x_2$ are available $(\mathbf{x}_1, \mathbf{x_2})$
2. echoes' TOAs $\propto$ sampling frequency
3. Find echoes $\to$ **find sparse vectors** $\mathbf{h}_1, \mathbf{h}_2$ of length $L$

# AER as discrete SIMO BCE

Key ingredient – *Cross relation identity*

$$x_i = h_i * s$$



$$h_2 * x_1 = h_2 * h_1 * s = h_1 * h_2 * s = h_1 * x_2$$

Ideas:

1. Sampled version of $x_1, x_2$ are available $(\mathbf{x}_1, \mathbf{x_2})$
2. echoes' TOAs $\propto$ sampling frequency
3. Find echoes $\rightarrow$ **find sparse vectors $\mathbf{h}_1, \mathbf{h}_2$** of length $L$
4. Modeled as **Lasso**-like problem

$$\widehat{\mathbf{h}}_1, \widehat{\mathbf{h}}_2 \in \underset{\mathbf{h}_1, \mathbf{h}_2 \in \mathbf{R}^n}{\arg\min} \|\mathbf{x}_1 * \mathbf{h}_2 - \mathbf{x}_2 * \mathbf{h}_1\|_2^2 + \lambda \mathcal{P}(\mathbf{h}_1, \mathbf{h}_2) \quad \text{s.t.} \quad \mathcal{C}(\mathbf{h}_1, \mathbf{h}_2)$$

$\mathcal{P}(\mathbf{h_1}, \mathbf{h_2}) \longrightarrow$ sparse promoting regularizer

$\mathcal{C}(\mathbf{h_1}, \mathbf{h_2}) \longrightarrow$ non-negativity / anchor constraints

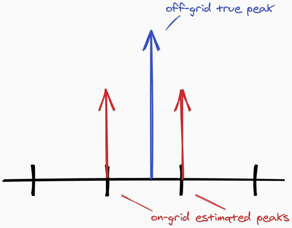
✓ [Tong et al., 1994] ✓ [Lin et al., 2007, Lin et al., 2008] ✓ [Aissa-El-Bey and Abed-Meraim, 2008] ✓ [Kowalczyk et al., 2013] ✓ [Crocco and Del Bue, 2015, Crocco and Del Bue, 2016]

$\mathbf{x_i} * \mathbf{h}_j$ computed as $\mathcal{T}(\mathbf{x}_i)\mathbf{h_j} \in \mathcal{O}(L^2)$

# AER as SIMO BCE: Limitations & bottleneck

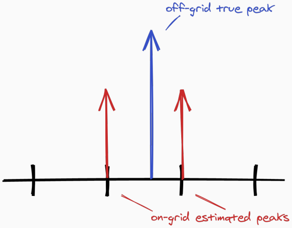
1. Estimation is on-grid

- Sparsity and non-negativity not true "on grid"
- *Body guard* effect [Duval and Peyré, 2017]
  - $\rightarrow$ low recall $\implies$ low accuracy
  - $\rightarrow$ slow convergence



# AER as SIMO BCE: Limitations & bottleneck

1. Estimation is on-grid
   - Sparsity and non-negativity not true “on grid”
   - *Body guard* effect [Duval and Peyré, 2017]
     - $\rightarrow$ low recall $\Longrightarrow$ low accuracy
     - $\rightarrow$ slow convergence

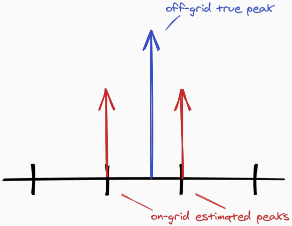



2. Pick Picking

$\rightarrow$ Manually tuned or labeling

# AER as SIMO BCE: Limitations & bottleneck

1. Estimation is on-grid
   - Sparsity and non-negativity not true "on grid"
   - *Body guard* effect [Duval and Peyré, 2017]
     - $\rightarrow$ low recall $\Longrightarrow$ low accuracy
     - $\rightarrow$ slow convergence




2. Pick Picking

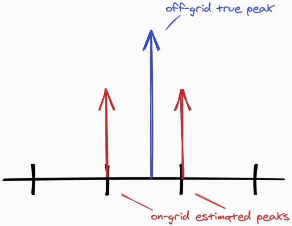
$\rightarrow$ Manually tuned or labeling

3. Increase the sampling frequency, $F_s$

$\rightarrow$ Increase Precision

# AER as SIMO BCE: Limitations & bottleneck

1. Estimation is on-grid
   - Sparsity and non-negativity not true "on grid"
   - *Body guard* effect [Duval and Peyré, 2017]
     - $\rightarrow$ low recall $\Longrightarrow$ low accuracy
     - $\rightarrow$ slow convergence




2. Pick Picking

   $\rightarrow$ Manually tuned or labeling

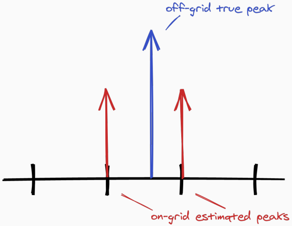
3. Increase the sampling frequency, $F_s$

   $\rightarrow$ Increase Precision

4. Computational bottleneck
   - Bigger vectors and matrices
     - $\longrightarrow$ memory usage

# AER as SIMO BCE: Limitations & bottleneck

1. Estimation is on-grid
   - Sparsity and non-negativity not true "on grid"
   - *Body guard* effect [Duval and Peyré, 2017]
     - $\rightarrow$ low recall $\implies$ low accuracy
     - $\rightarrow$ slow convergence




2. Pick Picking

   $\rightarrow$ Manually tuned or labeling

3. Increase the sampling frequency, $F_s$

   $\rightarrow$ Increase Precision

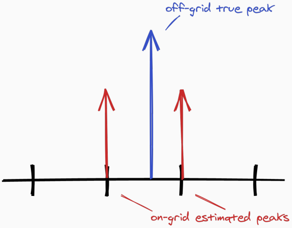
4. Computational bottleneck
   - Bigger vectors and matrices
     - $\longrightarrow$ memory usage
   - Computational complexity: at best $\mathcal{O}(F_s^2)$ per iteration

# AER as SIMO BCE: Limitations & bottleneck

1. Estimation is on-grid
    - Sparsity and non-negativity not true "on grid"
    - *Body guard* effect [Duval and Peyré, 2017]

      $\rightarrow$ low recall $\Longrightarrow$ low accuracy

      $\rightarrow$ slow convergence




2. Pick Picking

   $\rightarrow$ Manually tuned or labeling

3. Increase the sampling frequency, $F_s$

   $\rightarrow$ Increase Precision

4. Computational bottleneck
    - Bigger vectors and matrices

      $\longrightarrow$ memory usage

    - Computational complexity: at best $\mathcal{O}(F_s^2)$ per iteration
    - the higher the sampling frequency, the more ill-conditioned

      $\longrightarrow$ slow convergence

How to solve this?

# AER: Challenges and Contribution

State Of The Art

1. discrete (sparse) SIMO BCE
   based on time-domain XREL
2. Peak-picking

# AER: Challenges and Contribution

State Of The Art

1. discrete (sparse) SIMO BCE
   based on time-domain XREL
2. Peak-picking

$\Longrightarrow$ however

- Estimation in the RIR space
  memory issue
- Echoes are “off-grid”
  accuracy issues and mismatch
- Peak picking and labeling
  tuned and NP-hard

# AER: Challenges and Contribution

State Of The Art

1. discrete (sparse) SIMO BCE based on time-domain XREL
2. Peak-picking

$\implies$ however

- Estimation in the RIR space memory issue
- Echoes are “off-grid” accuracy issues and mismatch
- Peak picking and labeling tuned and NP-hard

$\implies$ we propose

`Blaster`[Di Carlo et al., 2020]

1. Knowledge-based approach
2. BCE + Continuous Dictionary based on XREL
3. Iterative-like approach
4. Inputs:
   - stereo mic recordings
   - # echoes
5. Output: $\tau_i^{(r)}, \alpha_i^{(r)}{}_{i,r}$

`Lantern`[Di Carlo et al., 2019]

1. Learning-based regression
2. Deep Learning used for SSL
3. Inputs: stereo audio feature
4. Output in the TDOA space ($\neq$ Echo space)

# Blaster- Knowledge-based Off-grid AER

Observation 1: the cross relation remains true in the frequency domain

$$\mathcal{F}x_1 \cdot \mathcal{F}h_2({}^{n}/{}_{F_s}) = \mathcal{F}x_2 \cdot \mathcal{F}h_1({}^{n}/{}_{F_s}) \qquad n = 0 \dots N-1$$

Observation 2: $\mathcal{F}\delta_{\text{echo}}$ is known in closed-form

Observation 3: $\mathcal{F}\mathrm{x_i}$ can be (well) approximated by DFT

$$\mathbf{X}_i = \mathsf{DFT}(\mathbf{x}_i) \simeq \mathcal{F}\mathbf{x}_i(nF_s) \qquad n = 0 \dots N-1$$

**Idea: Recover echoes by matching a finite number of frequencies**

$$\underset{h_1,h_2 \in \substack{\text{measure}\\ \text{space}}}{\arg\min} \ \tfrac{1}{2}\|\mathbf{X}_1 \cdot \mathcal{F}h_2(f) - \mathbf{X}_2 \cdot \mathcal{F}h_1(f)\|_2^2 + \lambda\|h_1 + h_2\|_{\text{TV}} \quad \text{s.t.} \begin{cases} h_1(\{0\}) = 1 \\ h_l \geq 0 \end{cases}$$

Looks like a **Lasso** problem, but $\mathcal{F}h_2(f)$ is a continuous function.
Instance of a **BLasso** problem[Bredies and Carioni, 2020]
Solved with Sliding Frank-Wolfe algorithm [Denoyelle et al., 2019]

✓no Toeplitz matrix

✓ Solutions is a train of Dirac

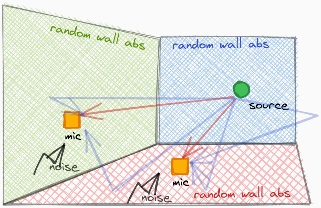
✓ anchor prevents trivial solution

In the manuscript:

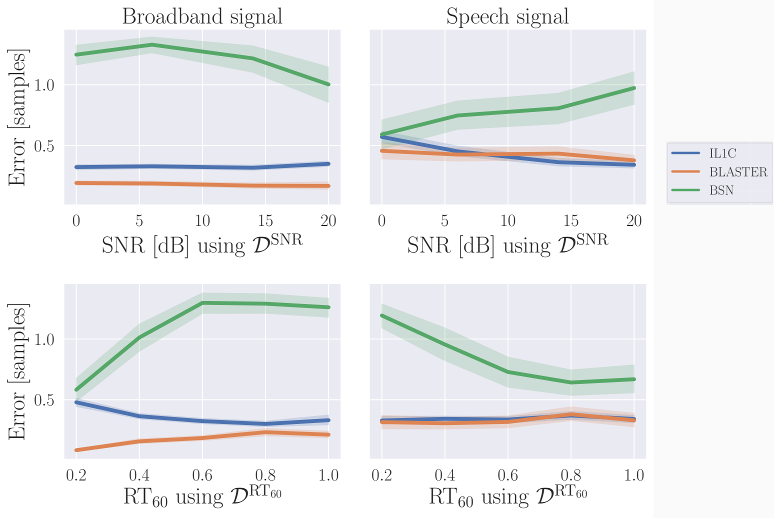
## Methods

- BSN [Lin et al., 2007]
- IL1C: iteratively-weighted $\ell_1$ constraint SIMO BCE [Crocco and Del Bue, 2015]
- Blaster: Proposed off-grid approach

Baseline method are xvalidated on other dataset



# Error per Dataset/Signal while recovering 7 echoes




✓Lower RMSE

✓ Robustness to SNR and $\mathrm{RT}_{60}$

✗ Source signal dependent

# Performance per # of echoes

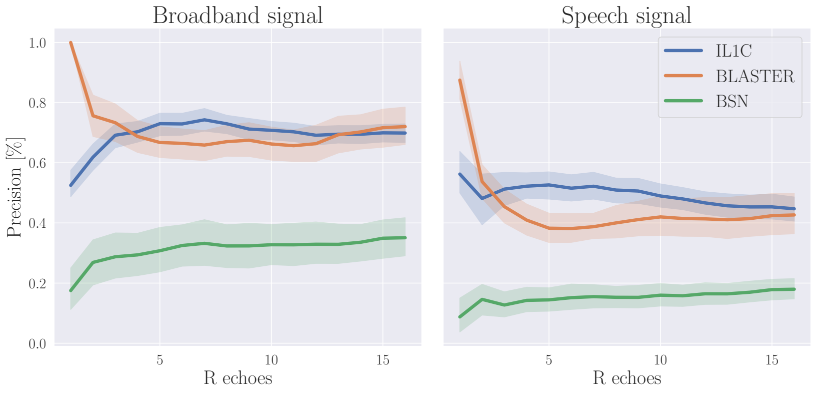



**Figure 1:** $\text{RT}_{60} = 400$ ms and SNR = 20 dB.

✗ Sensitive to # echoes

✗ Sensitive source signal

✓ Good for 2 echoes

# Performance per # of echoes

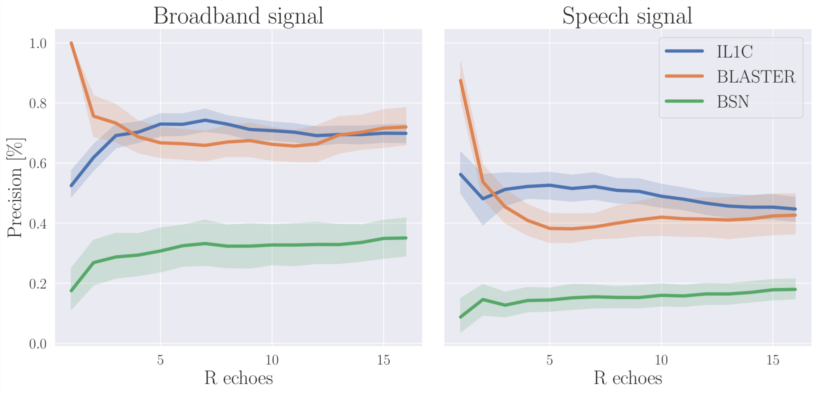



**Figure 1:** $\mathrm{RT}_{60} = 400$ ms and SNR = 20 dB.

✗ Sensitive to # echoes

✗ Sensitive source signal

✓ Good for 2 echoes [Scheibler et al., 2018, Di Carlo et al., 2019]

Observation 1: Mapping from observation to echo is extremely difficult

Later echoes are not considered, they may help

**Observation 1: Mapping from observation to echo is extremely difficult**

Later echoes are not considered, they may help

**Observation 2: We have acoustic simulators**

Acoustic simulators based on ISM + annotation for free

**Observation 1: Mapping from observation to echo is extremely difficult**

Later echoes are not considered, they may help

**Observation 2: We have acoustic simulators**

Acoustic simulators based on ISM + annotation for free

**Observation 3: (Deep) Learning-based methods successful for localization**

Echoes are strongly related to the source position

# Lantern- Data-driven off-grid AER

Observation 1: Mapping from observation to echo is extremely difficult

Later echoes are not considered, they may help

Observation 2: We have acoustic simulators

Acoustic simulators based on ISM + annotation for free

Observation 3: (Deep) Learning-based methods successful for localization

Echoes are strongly related to the source position

Idea: Use Deep Learning for AER

- Extend previous work on source localization for Echo Estimation
- Estimate the first echo TOA
  ↪ simple case, but with important application in SSL

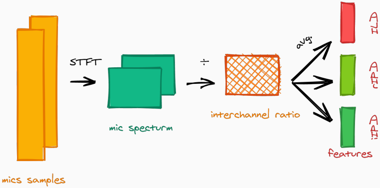
# Lantern- Data & Models

**Input:** features

**Output:** target

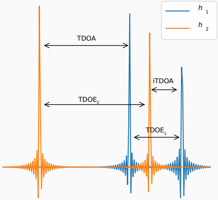
# Lantern- Data & Models

**Input:** features




$\longrightarrow$

**Output:** target




Relative Transfer Function

$$\mathtt{ReTF}[f] = \frac{H_2[f]}{H_1[f]} \approx \underset{t}{\mathtt{avg.}} \left( \frac{X_2[f,t]}{X_1[f,t]} \right)$$

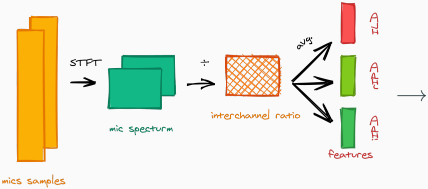
This is the instantaneous ReTF

TDOAs inter and intra reflection

$$\mathcal{V} = \{\mathtt{TDOA}, \mathtt{iTDOA}, \mathtt{TDOE}_1\}$$

First strongest echo $\Leftrightarrow$ close surface

**Input:** features

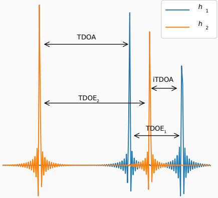



Relative Transfer Function

$$\mathtt{ReTF}[f] = \frac{H_2[f]}{H_1[f]} \approx \underset{t}{\mathtt{avg.}}\left(\frac{X_2[f,t]}{X_1[f,t]}\right)$$

This is the **instantaneous ReTF**

**Output:** target




TDOAs inter and intra reflection

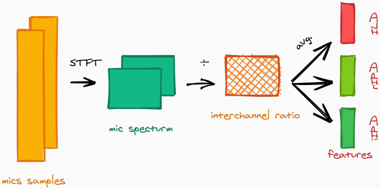
$$\mathcal{V} = \{\mathtt{TDOA}, \mathtt{iTDOA}, \mathtt{TDOE}_1\}$$

First strongest echo ⇔ close surface

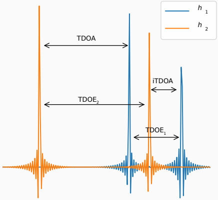
## Model

- Architecture: **CNN** [Chakrabarty and Habets, 2017, Nguyen et al., 2018]

**Input:** features




**Output:** target




Relative Transfer Function

$$\mathtt{ReTF}[f] = \frac{H_2[f]}{H_1[f]} \approx \underset{t}{\mathtt{avg.}} \left( \frac{X_2[f,t]}{X_1[f,t]} \right)$$

This is the **instantaneous ReTF**

TDOAs inter and intra reflection

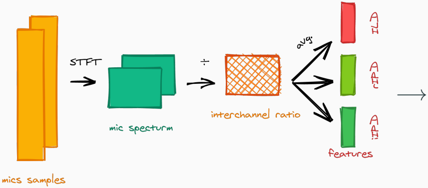
$$\mathcal{V} = \{\mathtt{TDOA}, \mathtt{iTDOA}, \mathtt{TDOE}_1\}$$

First strongest echo ⇔ close surface

## Model

- Architecture: **CNN** [Chakrabarty and Habets, 2017, Nguyen et al., 2018]
- Loss Function:
  1. RMSE (Multi-label regression) on $\mathcal{V}$
  2. Gaussian log-likelihood $\rightarrow \{\mu, \sigma^2\}$
  3. Student log-likelihood $\rightarrow \{\mu, \lambda, \nu\}$

**Input:** features

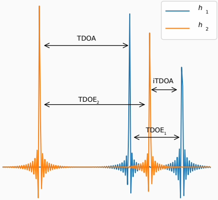



Relative Transfer Function

$$\text{ReTF}[f] = \frac{H_2[f]}{H_1[f]} \approx \underset{t}{\text{avg.}}\left(\frac{X_2[f,t]}{X_1[f,t]}\right)$$

This is the **instantaneous ReTF**

**Output:** target




TDOAs inter and intra reflection

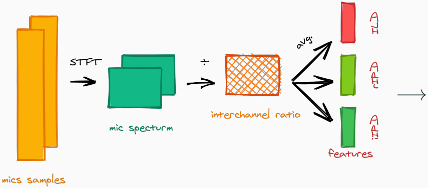
$$\mathcal{V} = \{\text{TDOA}, \text{iTDOA}, \text{TDOE}_1\}$$

First strongest echo ⇔ close surface

## Model

- Architecture: **CNN** [Chakrabarty and Habets, 2017, Nguyen et al., 2018]
- Loss Function:
  1. RMSE (Multi-label regression) on $\mathcal{V}$
  2. Gaussian log-likelihood $\rightarrow \{\mu, \sigma^2\}$
  3. Student log-likelihood $\rightarrow \{\mu, \lambda, \nu\}$
- Virtually Supervised Learning (= data from acoustic simulator)

**Input:** features

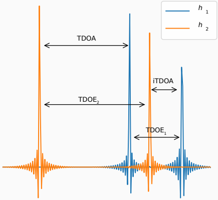



Relative Transfer Function

$$\texttt{ReTF}[f] = \frac{H_2[f]}{H_1[f]} \approx \underset{t}{\text{avg.}}\left(\frac{X_2[f,t]}{X_1[f,t]}\right)$$

This is the **instantaneous ReTF**

**Output:** target




TDOAs inter and intra reflection

$$\mathcal{V} = \{\texttt{TDOA}, \texttt{iTDOA}, \texttt{TDOE}_1\}$$

First strongest echo ⇔ close surface

## Model

- Architecture: **CNN** [Chakrabarty and Habets, 2017, Nguyen et al., 2018]
- Loss Function:
  1. RMSE (Multi-label regression) on $\mathcal{V}$
  2. Gaussian log-likelihood $\rightarrow \{\mu, \sigma^2\}$
  3. Student log-likelihood $\rightarrow \{\mu, \lambda, \nu\}$

  } Generative models ← for data fusion / similar to **MDN** [Bishop, 1994]
- Virtually Supervised Learning (= data from acoustic simulator)

Baseline: GCCPHAT (only TDOA),
$\texttt{MLP}_{\mathcal{V}}$ [Di Carlo et al., 2019]

Proposed: $\texttt{CNN}_{\mathcal{V}}$, $\texttt{CNN}_{\mathcal{V}_{\mathcal{N}}}$, $\texttt{CNN}_{\mathcal{V}_{\mathcal{T}}}$

# `Lantern`- Experiments & Resuls

**Baseline:** `GCCPHAT` (only TDOA),
`MLP`$_{\mathcal{V}}$ [Di Carlo et al., 2019]

**Proposed:** `CNN`$_{\mathcal{V}}$, `CNN`$_{\mathcal{V}_{\mathcal{N}}}$, `CNN`$_{\mathcal{V}_{\mathcal{T}}}$

**Metric:** normalized RMSE
(0 = best fit, 1 = random)

# Lantern- Experiments & Resuls

Baseline: GCCPHAT (only TDOA), $\mathtt{MLP}_{\mathcal{V}}$ [Di Carlo et al., 2019]

Proposed: $\mathtt{CNN}_{\mathcal{V}}$, $\mathtt{CNN}_{\mathcal{V}_{\mathcal{N}}}$, $\mathtt{CNN}_{\mathcal{V}_{\mathcal{T}}}$

Metric: normalized RMSE (0 = best fit, 1 = random)

Train:

- random RT60, random SNR
- broadband source (wn)
- instantaneous RTF

Test: similar to train

# Lantern- Experiments & Resuls

**Baseline:** GCCPHAT (only TDOA), $\text{MLP}_{\mathcal{V}}$ [Di Carlo et al., 2019]

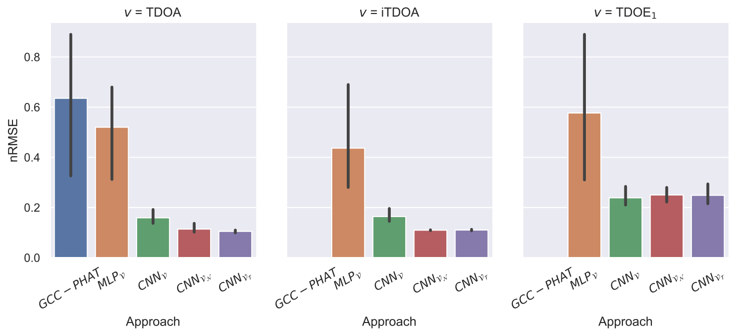
**Proposed:** $\text{CNN}_{\mathcal{V}}$, $\text{CNN}_{\mathcal{V}_{\mathcal{N}}}$, $\text{CNN}_{\mathcal{V}_{\mathcal{T}}}$

**Metric:** normalized RMSE (0 = best fit, 1 = random)

Train:

- random RT60, random SNR
- broadband source (wn)
- instantaneous RTF

Test: similar to train




✓ CNNs outperform GCC-PHAT, MLP

✓ CNNs less variance

✗ Gaussian ∼ Student-T

Baseline: GCCPHAT (only TDOA), $\mathtt{MLP}_{\mathcal{V}}$ [Di Carlo et al., 2019]

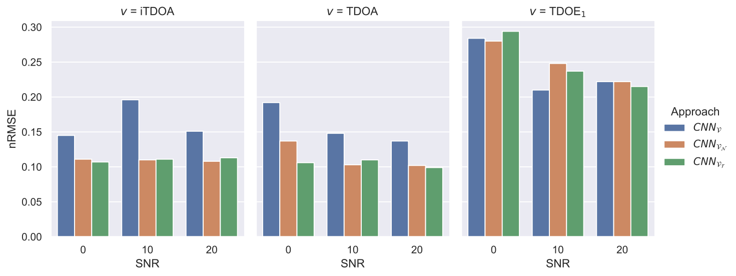
Proposed: $\mathtt{CNN}_{\mathcal{V}}$, $\mathtt{CNN}_{\mathcal{V}_{\mathcal{N}}}$, $\mathtt{CNN}_{\mathcal{V}_{\mathcal{T}}}$

Metric: normalized RMSE (0 = best fit, 1 = random)

Train:

- random RT60, random SNR
- broadband source (wn)
- instantaneous RTF

Test: similar to train




✓ Generative better than Normal

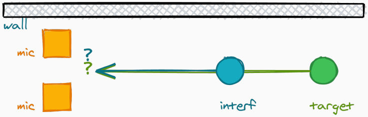
✗ Gaussian ∼ Student-T

✗ Bigger error on TDOE

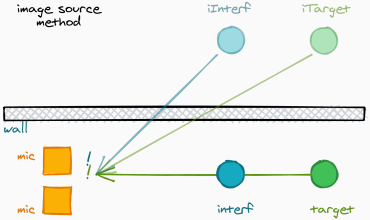
# Echo-aware Application

# Echo-aware Application

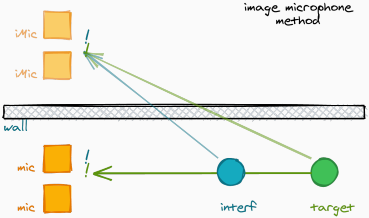
Echoes = same content, different time/direction



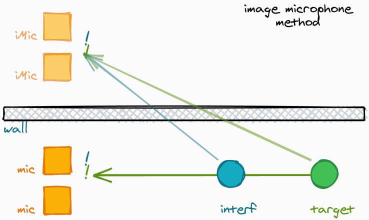
Echoes = same content, different time/direction



Echoes = same content, different time/direction

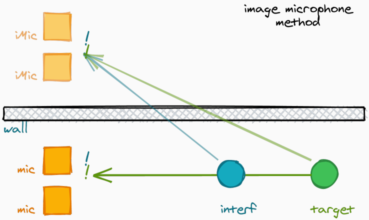


Echoes = same content, different time/direction




Recent literature on echo-aware processing:

Echoes = same content, different time/direction

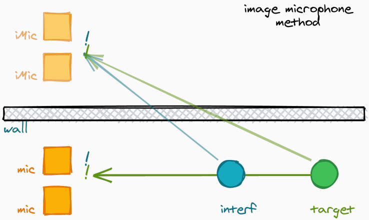



Recent literature on echo-aware processing:

**What?**
Echoes = repetitions

- Sound Source Separation
  [Leglaive et al., 2016]
- Speech Enhancement
  [Flanagan et al., 1993,
  Dokmanić et al., 2015, **?**]

Echoes = same content, different time/direction




Recent literature on echo-aware processing:

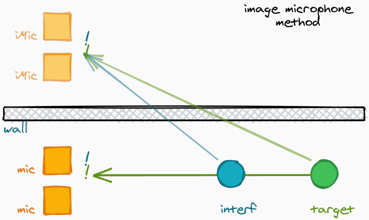
**What?**
Echoes = repetitions

- Sound Source Separation
  [Leglaive et al., 2016]
- Speech Enhancement
  [Flanagan et al., 1993,
  Dokmanić et al., 2015, **?**]

**Where?**
Echoes ∈ indoor propagation

- Sound Source Localization
  [Ribeiro et al., 2010,
  Jensen et al., 2019]
- Microphone Calibration
  [Dokmanić et al., 2015,
  Salvati et al., 2016]
- Room Geometry
  Estimation
  [**?**, Crocco et al., 2017]

Echoes = same content, different time/direction




Recent literature on echo-aware processing:

**What?**
Echoes = repetitions

- Sound Source Separation
  [Leglaive et al., 2016]
- Speech Enhancement
  [Flanagan et al., 1993, Dokmanić et al., 2015, **?**]

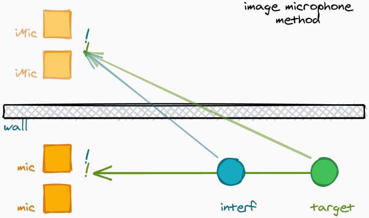
**Where?**
Echoes ∈ indoor propagation

- Sound Source Localization
  [Ribeiro et al., 2010, Jensen et al., 2019]
- Microphone Calibration
  [Dokmanić et al., 2015, Salvati et al., 2016]
- Room Geometry Estimation
  [**?**, Crocco et al., 2017]

**How?**
Echoes ∈ sound propagation

- Blind Channel Estimation
  [Lin et al., 2007, Crocco et al., 2017]
- Acoustic Measurements
  [Eaton et al., 2015, Kuttruff, 2016]

Echoes = same content, different time/direction




Recent literature on echo-aware processing:

**What?**
Echoes = repetitions

- Sound Source Separation
  [Leglaive et al., 2016]
- Speech Enhancement
  [Flanagan et al., 1993,
  Dokmanić et al., 2015, ?]

**Where?**
Echoes ∈ indoor propagation

- Sound Source Localization
  [Ribeiro et al., 2010,
  Jensen et al., 2019]
- Microphone Calibration
  [Dokmanić et al., 2015,
  Salvati et al., 2016]
- Room Geometry
  Estimation
  [?, Crocco et al., 2017]

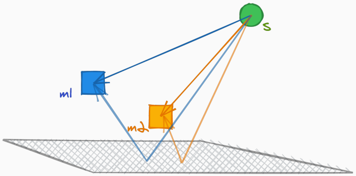
**How?**
Echoes ∈ sound propagation

- Blind Channel Estimation
  [Lin et al., 2007,
  Crocco et al., 2017]
- Acoustic Measurements
  [Eaton et al., 2015,
  Kuttruff, 2016]

# Mirage- Sound Source Locatization with Echoes

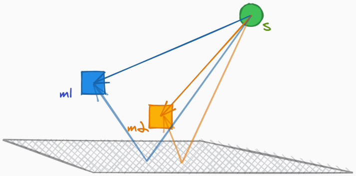
The Picnic Scenario:

- One source
- Two microphones
  - → passive scenario
  - → generalizable to more pairs



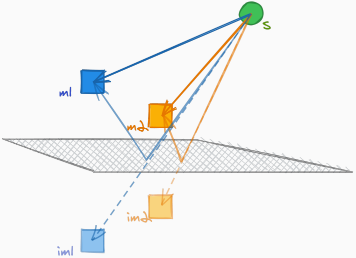
The Picnic Scenario:

- One source
- Two microphones
  - → passive scenario
  - → generalizable to more pairs
- Close to a very reflective surface
  - → First echo = Strongest echo
  - → $\alpha_{\text{picnic}}$ const. $\forall f$
  - → table-top device



## The Picnic Scenario:

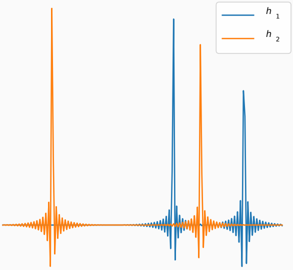
- One source
- Two microphones
  - $\rightarrow$ passive scenario
  - $\rightarrow$ generalizable to more pairs
- Close to a very reflective surface
  - $\rightarrow$ First echo = Strongest echo
  - $\rightarrow$ $\alpha_{\text{picnic}}$ const. $\forall f$
  - $\rightarrow$ table-top device




Each pair is augmented with echoes

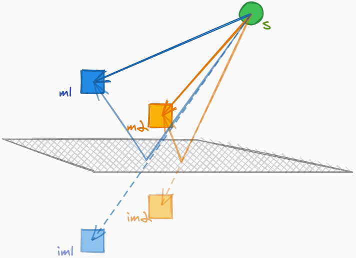
Mirage Array

How to access the *image* microphones?



The Picnic Scenario:

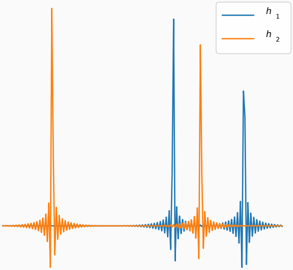
- One source
- Two microphones
  - → passive scenario
  - → generalizable to more pairs
- Close to a very reflective surface
  - → First echo = Strongest echo
  - → $\alpha_{\text{picnic}}$ const. $\forall f$
  - → table-top device




Each pair is augmented with echoes

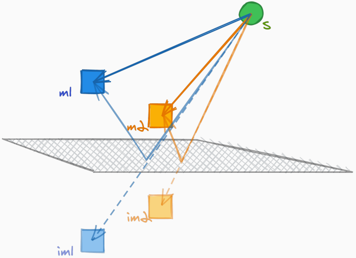
Mirage Array

How to access the *image* microphones?



## The Picnic Scenario:

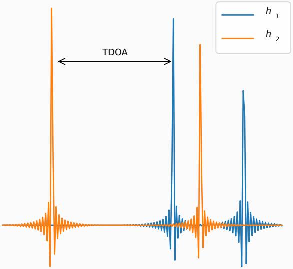
- One source
- Two microphones
  - → passive scenario
  - → generalizable to more pairs
- Close to a very reflective surface
  - → First echo = Strongest echo
  - → $\alpha_{\text{picnic}}$ const. $\forall f$
  - → table-top device




Each pair is augmented with echoes

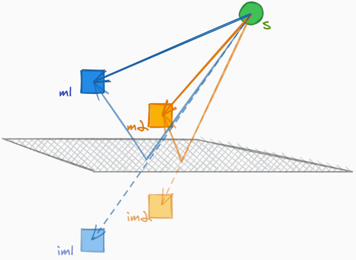
Mirage Array

How to access the *image* microphones?



The Picnic Scenario:

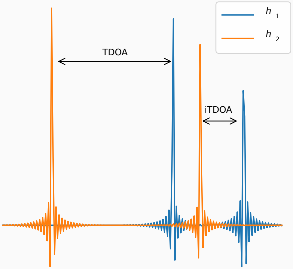
- One source
- Two microphones
  - → passive scenario
  - → generalizable to more pairs
- Close to a very reflective surface
  - → First echo = Strongest echo
  - → $\alpha_{\text{picnic}}$ const. $\forall f$
  - → table-top device




Each pair is augmented with echoes

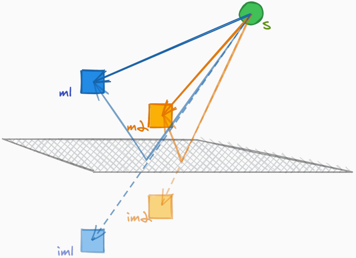
Mirage Array

How to access the *image* microphones?



The Picnic Scenario:

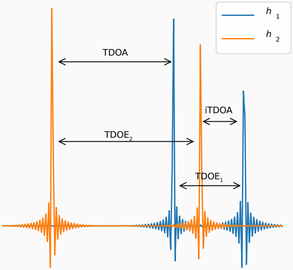
- One source
- Two microphones
  - → passive scenario
  - → generalizable to more pairs
- Close to a very reflective surface
  - → First echo = Strongest echo
  - → $\alpha_{\text{picnic}}$ const. $\forall f$
  - → table-top device




Each pair is augmented with echoes

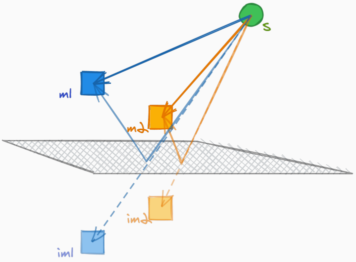
Mirage Array

How to access the *image* microphones?



The Picnic Scenario:

- One source
- Two microphones
  - $\rightarrow$ passive scenario
  - $\rightarrow$ generalizable to more pairs
- Close to a very reflective surface
  - $\rightarrow$ First echo = Strongest echo
  - $\rightarrow$ $\alpha_{\text{picnic}}$ const. $\forall f$
  - $\rightarrow$ table-top device



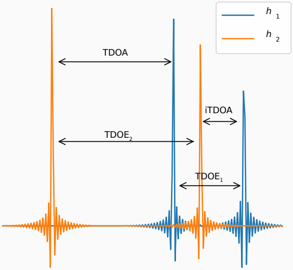

Each pair is augmented with echoes

Mirage Array

How to access the *image* microphones?

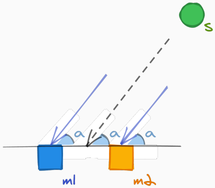
**idea:** use SSL algorithm on it

**recall:** echoes are known



## SSL with 2 microphones

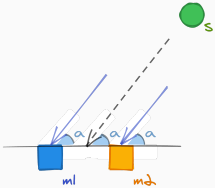
- 1D SSL: only angle of arrival (AOA)
- e.g. GCC-PHAT for TDOA estimation [Knapp and Carter, 1976]
  (known limitation, but good in practice)



# Mirage- Sound Source Locatization with Echoes

SSL with 2 microphones

- 1D SSL: only angle of arrival (AOA)
- e.g. GCC-PHAT for TDOA estimation [Knapp and Carter, 1976]
  (known limitation, but good in practice)

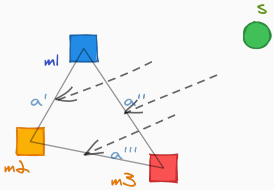



SSL with more microphones

- 2D SSL: azimuth and elevation

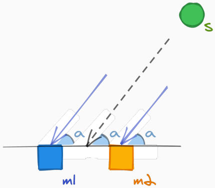
1. For each pair $p$:
   $\text{AOA}_p \leftarrow$ TDOA-based 2-mic-SSL
2. "Fuse" together all the observation
   (Angular spectra, Probability distributions)

- e.g. SRP-PHAT [DiBiase et al., 2001]



## SSL with 2 microphones

- 1D SSL: only angle of arrival (AOA)
- e.g. GCC-PHAT for TDOA estimation [Knapp and Carter, 1976]
  (known limitation, but good in practice)

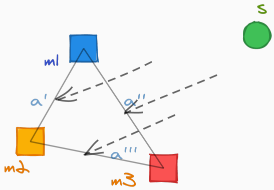



## SSL with more microphones

- 2D SSL: azimuth and elevation

1. For each pair $p$:
   $\text{AOA}_p \leftarrow$ TDOA-based 2-mic-SSL
2. "Fuse" together all the observation (Angular spectra, Probability distributions)

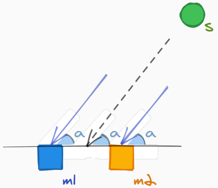
- e.g. SRP-PHAT [DiBiase et al., 2001]




**Baseline:** GCC-PHAT on true microphones

## SSL with 2 microphones

- 1D SSL: only angle of arrival (AOA)
- e.g. GCC-PHAT for TDOA estimation [Knapp and Carter, 1976] (known limitation, but good in practice)

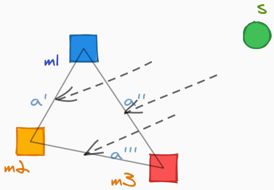



## SSL with more microphones

- 2D SSL: azimuth and elevation

1. For each pair $p$:
   $\text{AOA}_p \leftarrow$ TDOA-based 2-mic-SSL
2. "Fuse" together all the observation (Angular spectra, Probability distributions)

- e.g. SRP-PHAT [DiBiase et al., 2001]

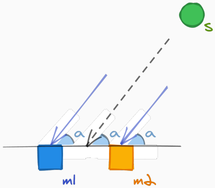



**Baseline:** GCC-PHAT on true microphones

**Proposed Approach:** using Lantern (DNN-based TDOA estimation)

SSL with 2 microphones

- 1D SSL: only angle of arrival (AOA)
- e.g. GCC-PHAT for TDOA estimation [Knapp and Carter, 1976]
  (known limitation, but good in practice)

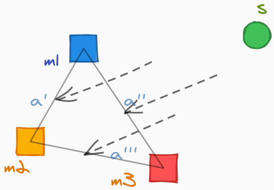



SSL with more microphones

- 2D SSL: azimuth and elevation

1. For each pair $p$:
   $\text{AOA}_p \leftarrow$ TDOA-based 2-mic-SSL
2. "Fuse" together all the observation
   (Angular spectra, Probability distributions)

- e.g. SRP-PHAT [DiBiase et al., 2001]




**Baseline:** GCC-PHAT on true microphones

**Proposed Approach:** using Lantern (DNN-based TDOA estimation)

**issue:** just punctual estimation cannot be "fused"
**solution:** use Lantern in generative mode (mic pos assumed known)

# Mirage- Results

**Data:** virtually generated closeddataset as for Lantern
**Metric:** angular mean error and accuracy (thr=10, 20)

## AOA estimation

- ✓ Similar when wn
- ✗ Huge drop when noise
- ✗ Huge drop when speech and noise

| AOA | Input | ACCURACY $\theta < 10$ | $\theta < 20$ |
|---|---|---|---|
| MIRAGE | wn | 4.10 (77) | 5.97 (97) |
| MIRAGE | wn+n | 5.00 (26) | 9.89 (54) |
| GCC-PHAT | wn | 4.22 (81) | 6.19 (97) |
| GCC-PHAT | wn+n | 4.03 (65) | 5.34 (83) |

# Mirage- Results

**Data:** virtually generated closeddataset as for Lantern
**Metric:** angular mean error and accuracy (thr=10, 20)

## AOA estimation

- ✓ Similar when wn
- ✗ Huge drop when noise
- ✗ Huge drop when speech and noise

| AOA | Input | ACCURACY $\theta < 10$ | $\theta < 20$ |
|---|---|---|---|
| MIRAGE | wn | 4.10 (77) | 5.97 (97) |
| MIRAGE | wn+n | 5.00 (26) | 9.89 (54) |
| GCC-PHAT | wn | 4.22 (81) | 6.19 (97) |
| GCC-PHAT | wn+n | 4.03 (65) | 5.34 (83) |

# Mirage- Results

**Data:** virtually generated closeddataset as for Lantern
**Metric:** angular mean error and accuracy (thr=10, 20)

### AOA estimation

- ✓ Similar when wn
- ✗ Huge drop when noise
- ✗ Huge drop when speech and noise

| AOA | Input | ACCURACY $\theta < 10$ | $\theta < 20$ |
|---|---|---|---|
| MIRAGE | wn | 4.10 (77) | 5.97 (97) |
| MIRAGE | wn+n | 5.00 (26) | 9.89 (54) |
| GCC-PHAT | wn | 4.22 (81) | 6.19 (97) |
| GCC-PHAT | wn+n | 4.03 (65) | 5.34 (83) |
| MIRAGE | sp | 4.83 (63) | 7.26 (82) |
| MIRAGE | sp+n | 4.60 (16) | 9.88 (35) |
| GCC-PHAT | sp | 4.08 (82) | 5.34 (97) |
| GCC-PHAT | sp+n | 4.70 (19) | 8.38 (32) |

# Mirage- Results

**Data:** virtually generated closeddataset as for Lantern
**Metric:** angular mean error and accuracy (thr=10, 20)

## AOA estimation

- ✓ Similar when wn
- ✗ Huge drop when noise
- ✗ Huge drop when speech and noise

| AOA | Input | ACCURACY $\theta < 10$ | ACCURACY $\theta < 20$ |
|---|---|---|---|
| MIRAGE | wn | 4.10 (77) | 5.97 (97) |
| MIRAGE | wn+n | 5.00 (26) | 9.89 (54) |
| GCC-PHAT | wn | 4.22 (81) | 6.19 (97) |
| GCC-PHAT | wn+n | 4.03 (65) | 5.34 (83) |
| MIRAGE | sp | 4.83 (63) | 7.26 (82) |
| MIRAGE | sp+n | 4.60 (16) | 9.88 (35) |
| GCC-PHAT | sp | 4.08 (82) | 5.34 (97) |
| GCC-PHAT | sp+n | 4.70 (19) | 8.38 (32) |

## 2D SSL estimation
(both Az. and El.)

| DoA | | ACCURACY $< 10$ | | ACCURACY $< 20$ | |
|---|---|---|---|---|---|
| | Input | $\theta$ | $\phi$ | $\theta$ | $\phi$ |
| MIRAGE | wn | 4.5 (59) | 3.9 (71) | 6.8 (79) | 5.9 (88) |
| MIRAGE | wn+n | 4.4 (18) | 5.5 (26) | 9.4 (35) | 11.1 (66) |
| MIRAGE | sp | 4.6 (45) | 4.8 (59) | 8.1 (71) | 7.2 (83) |
| MIRAGE | sp+n | 5.2 (17) | 5.9 (12) | 10.7 (38) | 12.3 (43) |

✓ Solved “impossible” localization

✗ Performance depending on echo estimation

# Echo-aware Dataset

# Echo-aware Datasets

Data in audio signal processing

1. typically RIRs
2. necessary for validation/learning
3. real data collection need expertise, equipment and time
4. dataset are for ad hoc setup

# Echo-aware Datasets

Data in audio signal processing

1. typically RIRs
2. necessary for validation/learning
3. real data collection need expertise, equipment and time
4. dataset are for ad hoc setup

$\implies$ simulated data are used instead $\implies$ validating/learning model with model

# Echo-aware Datasets

Data in audio signal processing

1. typically RIRs
2. necessary for validation/learning
3. real data collection need expertise, equipment and time
4. dataset are for ad hoc setup

$\implies$ simulated data are used instead $\implies$ validating/learning model with model

Echo-aware real data in audio signal processing

For SE: strong echoes✓, but not annotated✗, specific array✗
[Szöke et al., 2019, Bertin et al., 2019, Remaggi et al., 2016]

For RooGE: good geom annotation✓, but a few acoustic scenarios✗
[Dokmanić et al., 2013, Crocco et al., 2017, Remaggi et al., 2019]

# Echo-aware Datasets

Data in audio signal processing

1. typically RIRs
2. necessary for validation/learning
3. real data collection need expertise, equipment and time
4. dataset are for ad hoc setup

$\Longrightarrow$ simulated data are used instead $\Longrightarrow$ validating/learning model with model

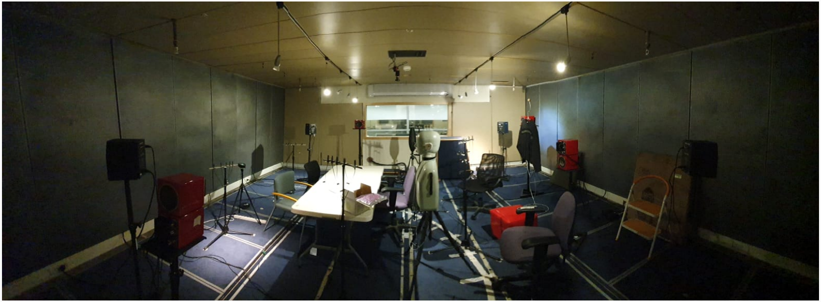
Echo-aware real data in audio signal processing

**For SE:** strong echoes✓, but not annotated✗, specific array✗
[Szöke et al., 2019, Bertin et al., 2019, Remaggi et al., 2016]

**For RooGE:** good geom annotation✓, but a few acoustic scenarios✗
[Dokmanić et al., 2013, Crocco et al., 2017, Remaggi et al., 2019]

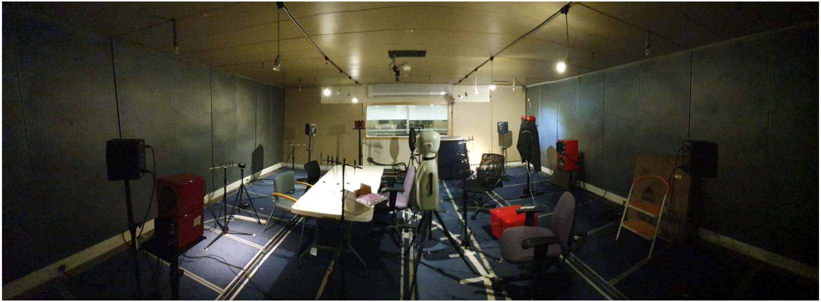
A good echo-aware dataset should allow **SE**, **RooGE** and **AER**
HOW?
signal annotation $\Longleftrightarrow$ geometric annotation

dEchorate: echo-aware dataset



dEchorate: echo-aware dataset

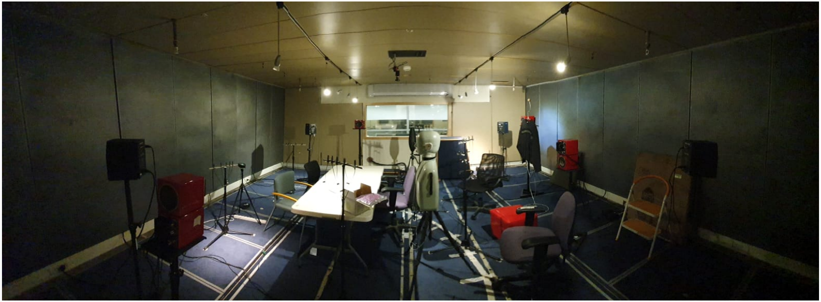
**Recorded:** Acoustic lab of Bar'Ilan (Shoebox)



## dEchorate: echo-aware dataset

**Recorded:** Acoustic lab of Bar'Ilan (Shoebox)

**Annotated:** manually



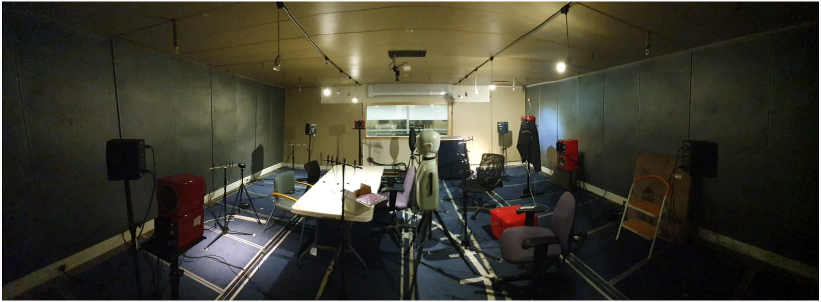
# dEchorate: realization

## dEchorate: echo-aware dataset

**Recorded:** Acoustic lab of Bar'Ilan (Shoebox)

**Annotated:** manually

**Collaboration:** prof. Sharon Gannot and ing. Pinchas Tandeitnik



# dEchorate: realization

## dEchorate: echo-aware dataset

**Recorded:** Acoustic lab of Bar'Ilan (Shoebox)

**Annotated:** manually

**Collaboration:** prof. Sharon Gannot and ing. Pinchas Tandeitnik

Key features:

- revolving panels (different $\mathrm{RT}_{60}$ and echo prominency)
- 6 nULA with 5 mics and 4 sound sources
- geometry annotated & echo annotated
- measured RIRs $\leftarrow$ (matching) $\rightarrow$ simulated RIRs

# dEchorate: annotation

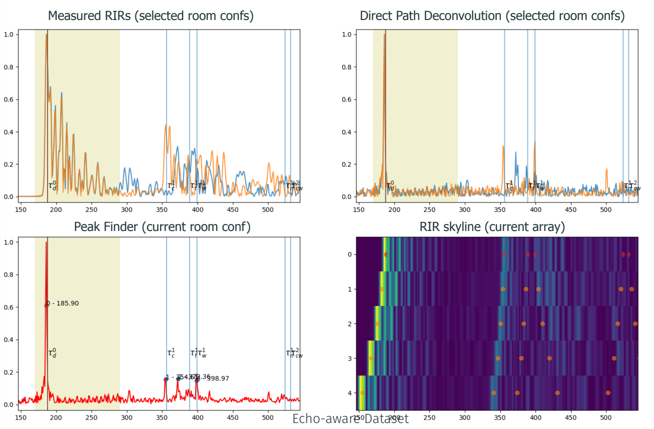
1. **RIR estimation** with **chirps signal** [Farina, 2007, Szöke et al., 2019]

# dEchorate: annotation

1. **RIR estimation** with **chirps signal** [Farina, 2007, Szöke et al., 2019]
2. **IPS calibration** with beacon $\longrightarrow$ mic and src positioning ($\pm 2$ cm)

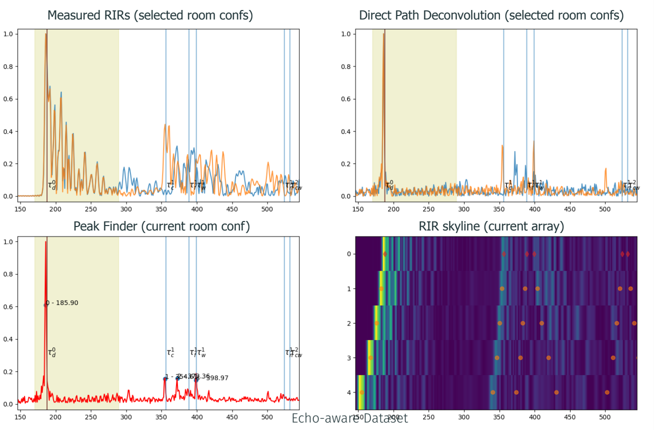
# dEchorate: annotation

1. **RIR estimation** with **chirps signal** [Farina, 2007, Szöke et al., 2019]
2. **IPS calibration** with beacon $\longrightarrow$ mic and src positioning ($\pm 2$ cm)
3. **GUI for annotation**
   Skyline, Matched Filter, Assisted Peak Picking



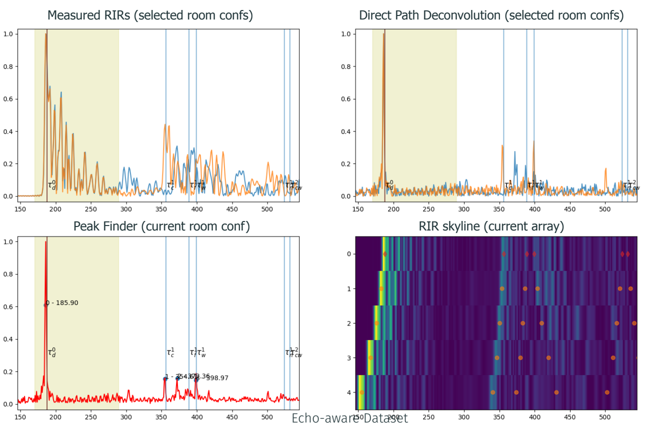
# dEchorate: annotation

1. **RIR estimation** with **chirps signal** [Farina, 2007, Szöke et al., 2019]
2. **IPS calibration** with beacon $\longrightarrow$ mic and src positioning ($\pm 2$ cm)
3. **GUI for annotation**
   Skyline, Matched Filter, Assisted Peak Picking
4. Refined position with **Least Square optimization**

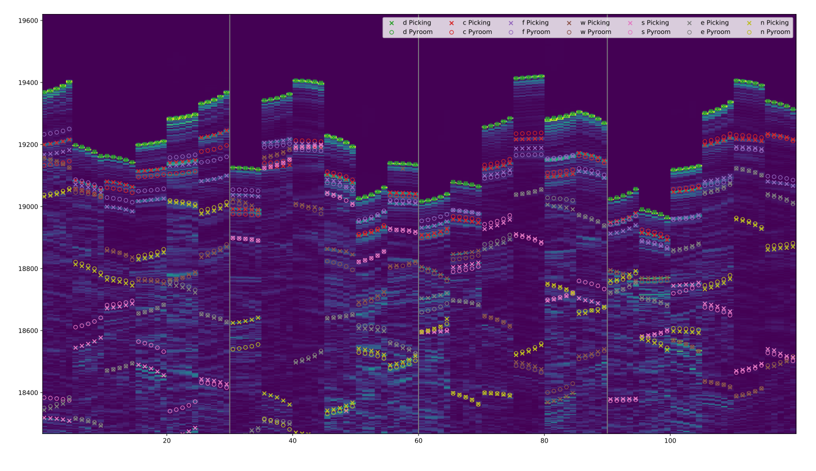


# dEchorate: annotation

1. **RIR estimation** with **chirps signal** [Farina, 2007, Szöke et al., 2019]
2. **IPS calibration** with beacon $\longrightarrow$ mic and src positioning ($\pm 2$ cm)
3. **GUI for annotation**
   Skyline, Matched Filter, Assisted Peak Picking
4. Refined position with **Least Square optimization**
5. **iterate including ceiling** (perfectly flat)



# dEchorate— Annotation




RIR Skyline showing

- absolute value of stacked RIRs as a figure
- × manual echo annotation
- ∘ *matching* echo annotation for `ISM simulator`

### Room Geometry Estimation (RooGE)

Estimate shape, volume or reflector position from signal (or form TOAs).

# Room Geometry Estimation (RooGE) with `dEchorate`

**Room Geometry Estimation (RooGE)**

Estimate shape, volume or reflector position from signal (or form TOAs).

If TOAs annotation (label and value) is available $\implies$ Image Source Inversion

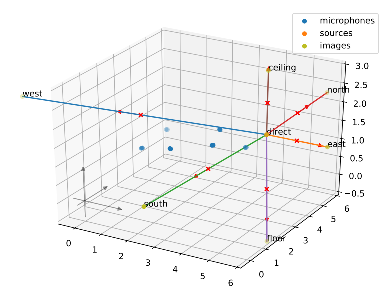
For each wall/label:

**Room Geometry Estimation (RooGE)**

Estimate shape, volume or reflector position from signal (or form TOAs).

If TOAs annotation (label and value) is available $\implies$ Image Source Inversion
For each wall/label:

1. TOA $\rightarrow$ image source position via 3D multilateration



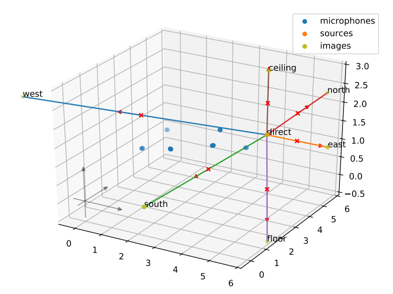
**Room Geometry Estimation (RooGE)**

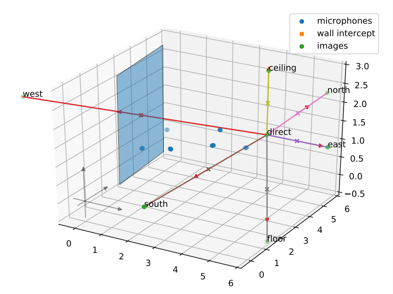
Estimate shape, volume or reflector position from signal (or form TOAs).

If TOAs annotation (label and value) is available $\implies$ Image Source Inversion
For each wall/label:

1. TOA $\rightarrow$ image source position via 3D multilateration
2. image source position $\rightarrow$ reflector estimation via geometric reasoning

other methods differ for priors and setup [Filos et al., 2011, Antonacci et al., 2012, Crocco et al., 2017]



# Room Geometry Estimation (RooGE) with dEchorate

**Room Geometry Estimation (RooGE)**

Estimate shape, volume or reflector position from signal (or form TOAs).

If TOAs annotation (label and value) is available $\implies$ Image Source Inversion
For each wall/label:

1. TOA $\rightarrow$ image source position via 3D multilateration
2. image source position $\rightarrow$ reflector estimation via geometric reasoning

other methods differ for priors and setup [Filos et al., 2011, Antonacci et al., 2012, Crocco et al., 2017]

| source id | 1 | | 2 | | 3 | | 4 | |
|---|---|---|---|---|---|---|---|---|
| wall | DE | AE | DE | AE | DE | AE | DE | AE |
| west | 0.74 | 8.99 | 4.59 | 8.32 | 5.89 | 5.75 | **0.05** | **2.40** |
| east | **0.81** | **0.08** | 0.9 | 0.50 | *69.51* | *55.70°* | 0.31 | 0.21 |
| south | 3.94 | *16.08°* | **0.18** | 1.77 | *14.37* | *18.55°* | 0.82 | **1.65** |
| north | 1.34 | 0.76 | 1.40 | 8.94 | **0.63** | **0.17** | 2.08 | 1.38 |
| floor | **5.19** | **1.76** | 7.27 | 2.66 | 7.11 | 2.02 | 5.22 | 1.90 |
| ceiling | 1.16 | 0.28 | 0.67 | 0.76 | **0.24** | 1.16 | 0.48 | **0.26** |

Distance Error (DE) [cm] and Angular Error (AE), **best**, *outliers*

# Echo-aware Speech Enhancement with dEchorate

**Speech Enhancement (SE)**

Improve the quality of a target sound source w.r.t.:

| interfering sources | + | noise | + | reverberation |
|---|---|---|---|---|
| ↓ | | ↓ | | ↓ |
| source separation | | denoising | | dereverberation<br>room equalization |

**Speech Enhancement (SE)**

Improve the quality of a target sound source w.r.t.:

| interfering sources | + | noise | + | reverberation |
|---|---|---|---|---|
| ↓ | | ↓ | | ↓ |
| source separation | | denoising | | dereverberation<br>room equalization |

SE via **linear spatial filtering**

$$\mathbf{h}\mathbf{s} + \mathbf{n} = \mathbf{x} \in \mathbb{C}^I \quad \longrightarrow \quad \mathbf{w}^{\mathsf{H}} \in \mathbb{C}^I \quad \longrightarrow \quad \mathbf{w}^{\mathsf{H}}\mathbf{x} \approx \mathbf{s}$$

# Echo-aware Speech Enhancement with dEchorate

**Speech Enhancement (SE)**

Improve the quality of a target sound source w.r.t.:

| interfering sources | + | noise | + | reverberation |
|---|---|---|---|---|
| ↓ | | ↓ | | ↓ |
| source separation | | denoising | | dereverberation<br>room equalization |

SE via **linear spatial filtering**

$$\mathbf{h}\mathbf{s} + \mathbf{n} = \mathbf{x} \in \mathbb{C}^I \quad \longrightarrow \quad \mathbf{w}^{\mathsf{H}} \in \mathbb{C}^I \quad \longrightarrow \quad \mathbf{w}^{\mathsf{H}}\mathbf{x} \approx \mathbf{s}$$

- **target is distortionless** (vs. Multichannel Wiener Filtering)
- Partial *steering vectors* from **geometry** (if anechoic $\Longrightarrow$ DS based on AOA)
- many variant, e.g. enhance or null multiple sources [Gannot et al., 2017]

## Speech Enhancement (SE)

Improve the quality of a target sound source w.r.t.:

| interfering sources | + | noise | + | reverberation |
|---|---|---|---|---|
| ↓ | | ↓ | | ↓ |
| source separation | | denoising | | dereverberation<br>room equalization |

SE via **linear spatial filtering**

$$\mathbf{h}\mathbf{s} + \mathbf{n} = \mathbf{x} \in \mathbb{C}^I \quad \longrightarrow \quad \mathbf{w}^{\mathsf{H}} \in \mathbb{C}^I \quad \longrightarrow \quad \mathbf{w}^{\mathsf{H}}\mathbf{x} \approx \mathbf{s}$$

- **target is distortionless** (vs. Multichannel Wiener Filtering)
- Partial *steering vectors* from **geometry** (if anechoic $\implies$ DS based on AOA)
- many variant, e.g. enhance or null multiple sources [Gannot et al., 2017]

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbb{E}\left\{\left\|\mathbf{w}^{\mathsf{H}}\mathbf{x}\right\|_2^2\right\} \quad \text{s.t.} \quad \mathbf{w}^{\mathsf{H}}\mathbf{h} = 1$$

Reducing output energy + distortionless ⇔ reduce any noise

Closed-form solution, but it requires:

| | Noise covariance matrix | Steering Vectors |
|---|---|---|
| DS | - | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{DP}}$ | Noise | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{ReTF}}$ | Noise | Relative Transfer Function |
| $\mathtt{MVDR}_{\mathtt{Rake}}$ | Noise | Relative Early Echoes |

## Echo-aware Speech Enhancement

Closed-form solution, but it requires:

| | Noise covariance matrix | Steering Vectors |
|---|---|---|
| DS | - | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{DP}}$ | Noise | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{ReTF}}$ | Noise | Relative Transfer Function |
| $\mathtt{MVDR}_{\mathtt{Rake}}$ | Noise | Relative Early Echoes |
| $\mathtt{MVDR}_{\mathtt{DP+Late}}$ | Noise + Late Diffusion | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{ReTF+Late}}$ | Noise + Late Diffusion | Relative Transfer Function |
| $\mathtt{MVDR}_{\mathtt{Rake+Late}}$ | Noise + Late Diffusion | Relative Early Echoes |

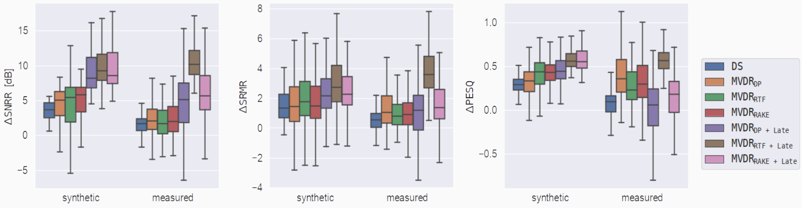
Closed-form solution, but it requires:

| | Noise covariance matrix | Steering Vectors |
|---|---|---|
| DS | - | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{DP}}$ | Noise | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{ReTF}}$ | Noise | Relative Transfer Function |
| $\mathtt{MVDR}_{\mathtt{Rake}}$ | Noise | Relative Early Echoes |
| $\mathtt{MVDR}_{\mathtt{DP+Late}}$ | Noise + Late Diffusion | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{ReTF+Late}}$ | Noise + Late Diffusion | Relative Transfer Function |
| $\mathtt{MVDR}_{\mathtt{Rake+Late}}$ | Noise + Late Diffusion | Relative Early Echoes |

Closed-form solution, but it requires:

| | Noise covariance matrix | Steering Vectors |
|---|---|---|
| DS | - | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{DP}}$ | Noise | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{ReTF}}$ | Noise | Relative Transfer Function |
| $\mathtt{MVDR}_{\mathtt{Rake}}$ | Noise | Relative Early Echoes |
| $\mathtt{MVDR}_{\mathtt{DP+Late}}$ | Noise + Late Diffusion | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{ReTF+Late}}$ | Noise + Late Diffusion | Relative Transfer Function |
| $\mathtt{MVDR}_{\mathtt{Rake+Late}}$ | Noise + Late Diffusion | Relative Early Echoes |

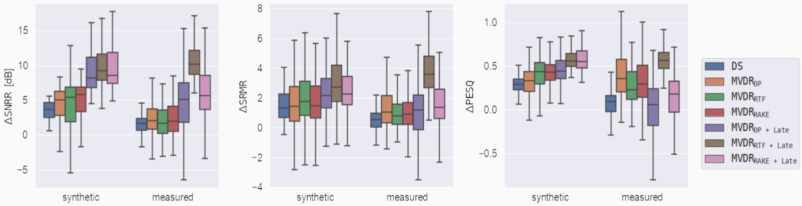
Comparison on dEchorate data



Closed-form solution, but it requires:

| | Noise covariance matrix | Steering Vectors |
|---|---|---|
| DS | - | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{DP}}$ | Noise | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{ReTF}}$ | Noise | Relative Transfer Function |
| $\mathtt{MVDR}_{\mathtt{Rake}}$ | Noise | Relative Early Echoes |
| $\mathtt{MVDR}_{\mathtt{DP+Late}}$ | Noise + Late Diffusion | Direct Path (AOA) |
| $\mathtt{MVDR}_{\mathtt{ReTF+Late}}$ | Noise + Late Diffusion | Relative Transfer Function |
| $\mathtt{MVDR}_{\mathtt{Rake+Late}}$ | Noise + Late Diffusion | Relative Early Echoes |

Comparison on dEchorate data

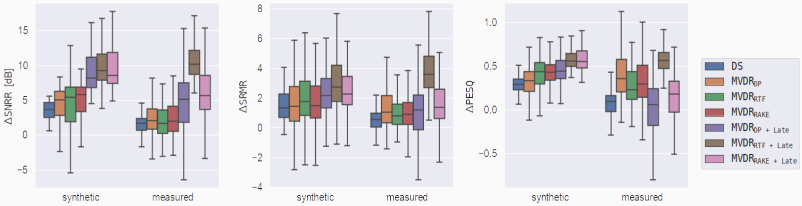


# Echo-aware Speech Enhancement

Closed-form solution, but it requires:

| | Noise covariance matrix | Steering Vectors |
|---|---|---|
| DS | - | Direct Path (AOA) |
| $\mathrm{MVDR}_{\mathrm{DP}}$ | Noise | Direct Path (AOA) |
| $\mathrm{MVDR}_{\mathrm{ReTF}}$ | Noise | Relative Transfer Function |
| $\mathrm{MVDR}_{\mathrm{Rake}}$ | Noise | Relative Early Echoes |
| $\mathrm{MVDR}_{\mathrm{DP+Late}}$ | Noise + Late Diffusion | Direct Path (AOA) |
| $\mathrm{MVDR}_{\mathrm{ReTF+Late}}$ | Noise + Late Diffusion | Relative Transfer Function |
| $\mathrm{MVDR}_{\mathrm{Rake+Late}}$ | Noise + Late Diffusion | Relative Early Echoes |

Comparison on dEchorate data



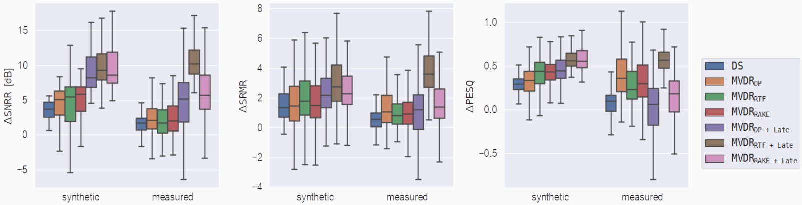

- *In theory* echo-aware ws. echo-agnostic?
  → Better than Direct Path, but ReTF and Rake are comparable

# Echo-aware Speech Enhancement

Closed-form solution, but it requires:

| | Noise covariance matrix | Steering Vectors |
|---|---|---|
| DS | - | Direct Path (AOA) |
| $\mathrm{MVDR}_{\mathrm{DP}}$ | Noise | Direct Path (AOA) |
| $\mathrm{MVDR}_{\mathrm{ReTF}}$ | Noise | Relative Transfer Function |
| $\mathrm{MVDR}_{\mathrm{Rake}}$ | Noise | Relative Early Echoes |
| $\mathrm{MVDR}_{\mathrm{DP+Late}}$ | Noise + Late Diffusion | Direct Path (AOA) |
| $\mathrm{MVDR}_{\mathrm{ReTF+Late}}$ | Noise + Late Diffusion | Relative Transfer Function |
| $\mathrm{MVDR}_{\mathrm{Rake+Late}}$ | Noise + Late Diffusion | Relative Early Echoes |

Comparison on dEchorate data




- *In theory* echo-aware ws. echo-agnostic?
  $\rightarrow$ Better than Direct Path, but ReTF and Rake are comparable
- Spatial filtering on Synthetic data vs. Measured data?
  $\rightarrow$ Rake suffer from mismatch, but better than DP

# Conclusion

# Summary of contributions

1. How to estimate echoes?

2. How to use echoes?

# Summary of contributions

1. How to estimate echoes?

`Blaster`: knowledge-based

- ✓ non-neg.ty and sparsity
- ✓ low RMSE by super-resolution
- ✗ dep. on source and # echoes
- ✗ vanilla solver

2. How to use echoes?

# Summary of contributions

1. How to estimate echoes?

2. How to use echoes?

`Blaster`: knowledge-based

- ✓ non-neg.ty and sparsity
- ✓ low RMSE by super-resolution
- ✗ dep. on source and # echoes
- ✗ vanilla solver

`Lantern`: Learning-based

- ✓ promising results
- ✗ only 2 echoes, how more?
- ✗ dep. on source

# Summary of contributions

1. How to estimate echoes?

`Blaster`: knowledge-based
- ✓ non-neg.ty and sparsity
- ✓ low RMSE by super-resolution
- ✗ dep. on source and # echoes
- ✗ vanilla solver

`Lantern`: Learning-based
- ✓ promising results
- ✗ only 2 echoes, how more?
- ✗ dep. on source

both:
- ✓ off-grid & parameter-free methods
- ✗ only synthetic & stereo data

2. How to use echoes?

# Summary of contributions

1. How to estimate echoes?

**Blaster**: knowledge-based
- ✓ non-neg.ty and sparsity
- ✓ low RMSE by super-resolution
- ✗ dep. on source and # echoes
- ✗ vanilla solver

**Lantern**: Learning-based
- ✓ promising results
- ✗ only 2 echoes, how more?
- ✗ dep. on source

both:
- ✓ off-grid & parameter-free methods
- ✗ only synthetic & stereo data

2. How to use echoes?

**Mirage** Echo-aware SSL
- ✓ Impossible 2D SSL
- ✓ Easy to extend

# Summary of contributions

## 1. How to estimate echoes?

**Blaster**: knowledge-based
- ✓ non-neg.ty and sparsity
- ✓ low RMSE by super-resolution
- ✗ dep. on source and # echoes
- ✗ vanilla solver

**Lantern**: Learning-based
- ✓ promising results
- ✗ only 2 echoes, how more?
- ✗ dep. on source

**both**:
- ✓ off-grid & parameter-free methods
- ✗ only synthetic & stereo data

## 2. How to use echoes?

**Mirage** Echo-aware SSL
- ✓ Impossible 2D SSL
- ✓ Easy to extend

**Separake** Echo-aware SSS
- ✓ echo boost SSS
- ✓ easy integration in NMF

# Summary of contributions

## 1. How to estimate echoes?

`Blaster`: knowledge-based
- ✓ non-neg.ty and sparsity
- ✓ low RMSE by super-resolution
- ✗ dep. on source and # echoes
- ✗ vanilla solver

`Lantern`: Learning-based
- ✓ promising results
- ✗ only 2 echoes, how more?
- ✗ dep. on source

both:
- ✓ off-grid & parameter-free methods
- ✗ only synthetic & stereo data

## 2. How to use echoes?

`Mirage` Echo-aware SSL
- ✓ Impossible 2D SSL
- ✓ Easy to extend

`Separake` Echo-aware SSS
- ✓ echo boost SSS
- ✓ easy integration in NMF

both: dep. on echo estimation

only synthetic data

# Summary of contributions

### 1. How to estimate echoes?

**Blaster**: knowledge-based
- ✓ non-neg.ty and sparsity
- ✓ low RMSE by super-resolution
- ✗ dep. on source and # echoes
- ✗ vanilla solver

**Lantern**: Learning-based
- ✓ promising results
- ✗ only 2 echoes, how more?
- ✗ dep. on source

both:
- ✓ off-grid & parameter-free methods
- ✗ only synthetic & stereo data

### 2. How to use echoes?

**Mirage** Echo-aware SSL
- ✓ Impossible 2D SSL
- ✓ Easy to extend

**Separake** Echo-aware SSS
- ✓ echo boost SSS
- ✓ easy integration in NMF

both: dep. on echo estimation

only synthetic data

### 3. Where to find echoes?

**dEchorate**: dataset
- ✓ dataset for AER, SE and RooGE
- ✓ geom labels $\leftrightarrow$ echo labels
- ✗ few inconsistency
- ✗ shoebox, linear array and directional

# Summary of contributions

## 1. How to estimate echoes?

`Blaster`: knowledge-based
- ✓ non-neg.ty and sparsity
- ✓ low RMSE by super-resolution
- ✗ dep. on source and # echoes
- ✗ vanilla solver

`Lantern`: Learning-based
- ✓ promising results
- ✗ only 2 echoes, how more?
- ✗ dep. on source

both:
- ✓ off-grid & parameter-free methods
- ✗ only synthetic & stereo data

## 2. How to use echoes?

`Mirage` Echo-aware SSL
- ✓ Impossible 2D SSL
- ✓ Easy to extend

`Separake` Echo-aware SSS
- ✓ echo boost SSS
- ✓ easy integration in NMF

both: dep. on echo estimation
only synthetic data

## 3. Where to find echoes?

`dEchorate`: dataset
- ✓ dataset for AER, SE and RooGE
- ✓ geom labels ↔ echo labels
- ✗ few inconsistency
- ✗ shoebox, linear array and directional

`dEchorate`: Validation
- ✓ on RooGE
- ✓ on echo-aware SE
- ✗ echo-SE similar to ReTF-SE
- ✗ echo-SE suffer for mismatch

# Echo-aware perspective

Directions for future work:

# Echo-aware perspective

Directions for future work:

- on **estimation**

  **Blaster** extended to multichannel [Tukuljac, 2020] and ReTF [Doclo and Moonen, 2002]

  use high-level priors ($RT_{60}$, DRR, Diffusion) [Badeau, 2019]

  **Lantern** extended physic-based learning

  more than 2 echoes: network or learning

# Echo-aware perspective

Directions for future work:

- on **estimation**
  - Blaster extended to multichannel [Tukuljac, 2020] and ReTF [Doclo and Moonen, 2002]
  - use high-level priors ($RT_{60}$, DRR, Diffusion) [Badeau, 2019]
  - Lantern extended physic-based learning
  - more than 2 echoes: network or learning
- on **application**
  - validation on real data (eg. smart home speaker)
  - use dEchorate data for validation
  - other Audio Scene Analysis problem
  - other field of echoes (Seismology, Underwater acoustic, Volcano tomography)
- on dEchorate
  - Synthetic to Real RIRs
  - popularization and divulgation of these data
  - write how to and how not echo-aware dataset

# Echo-aware perspective

Directions for future work:

- on **estimation**

  Blaster extended to multichannel [Tukuljac, 2020] and ReTF [Doclo and Moonen, 2002]

  use high-level priors ($RT_{60}$, DRR, Diffusion) [Badeau, 2019]

  Lantern extended physic-based learning

  more than 2 echoes: network or learning

- on **application**
  - validation on real data (eg. smart home speaker)
  - use dEchorate data for validation
  - other Audio Scene Analysis problem
  - other field of echoes (Seismology, Underwater acoustic, Volcano tomography)

- on dEchorate
  - Synthetic to Real RIRs
  - popularization and divulgation of these data
  - write how to and how not echo-aware dataset

- Echo estimation $\Longleftrightarrow$ Audio Analysis

# List of publications and artifacts

## Publications

- Estimation
  - `Lantern` [Di Carlo et al., 2019]
  - `dEchorate` [Di Carlo et al., 2020]
- Application
  - `Mirage` [Di Carlo et al., 2019]
  - `Separake` [Scheibler et al., 2018]
- Data
  - `dEchorate` (Unpublished)
- Other
  - Signal Processing CUP 2019 [Deleforge et al., 2019]
  - LOCATA Challenge 2019 [Lebarbenchon et al., 2018]
  - Collaboration with Honda [Di Carlo and Deleforge, ]

## Code

`dEchorate`: GUI and code for `dEchorate` (and RooGE)

`Risotto`: library for Relative Transfer Function

`Brioche`: library for echo-aware Spatial filtering

`pyMBSSLocate`: MBSSLocate in Python

`Separake`: Multichannel NMF in Python

Aissa-El-Bey, A. and Abed-Meraim, K. (2008).
Blind simo channel identification using a sparsity criterion.
In *2008 IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*, pages 271–275. IEEE.

Antonacci, F., Filos, J., Thomas, M. R., Habets, E. A., Sarti, A., Naylor, P. A., and Tubaro, S. (2012).
Inference of room geometry from acoustic impulse responses.
*IEEE Transactions on Audio, Speech, and Language Processing*, 20(10):2683–2695.

Badeau, R. (2019).
Common mathematical framework for stochastic reverberation models.
*The Journal of the Acoustical Society of America*, 145(4):2733–2745.

Bertin, N., Camberlein, E., Lebarbenchon, R., Vincent, E., Sivasankaran, S., Illina, I., and Bimbot, F. (2019).
Voicehome-2, an extended corpus for multichannel speech processing in real homes.
*Speech Communication*, 106:68–78.

Bishop, C. M. (1994).
Mixture density networks.

Bredies, K. and Carioni, M. (2020).
**Sparsity of solutions for variational inverse problems with finite-dimensional data.**
*Calculus of Variations and Partial Differential Equations*, 59(1):14.

Chakrabarty, S. and Habets, E. A. (2017).
**Broadband doa estimation using convolutional neural networks trained with noise signals.**
In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 136–140. IEEE.

Crocco, M. and Del Bue, A. (2015).
**Room impulse response estimation by iterative weighted l 1-norm.**
In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1895–1899. IEEE.

Crocco, M. and Del Bue, A. (2016).
**Estimation of tdoa for room reflections by iterative weighted l 1 constraint.**
In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3201–3205. IEEE.

Crocco, M., Trucco, A., and Del Bue, A. (2017).
**Uncalibrated 3d room geometry estimation from sound impulse responses.**
*Journal of the Franklin Institute*, 354(18):8678–8709.

Deleforge, A., Di Carlo, D., Strauss, M., Serizel, R., and Marcenaro, L. (2019).
**Audio-based search and rescue with a drone: Highlights from the ieee signal processing cup 2019 student competition [sp competitions].**
*IEEE Signal Processing Magazine*, 36(5):138–144.

Denoyelle, Q., Duval, V., Peyré, G., and Soubies, E. (2019).
**The sliding frank–wolfe algorithm and its application to super-resolution microscopy.**
*Inverse Problems*, 36(1):014001.

Di Carlo, D. and Deleforge, A.
**Hri-jf collaboration - final phase ii deliverable.**
Technical report, Inria Nancy - Grand Est.

Di Carlo, D., Deleforge, A., and Bertin, N. (2019).
**Mirage: 2d source localization using microphone pair augmentation with echoes.**
In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 775–779. IEEE.

Di Carlo, D., Elvira, C., Deleforge, A., Bertin, N., and Gribonval, R. (2020).
**Blaster: An off-grid method for blind and regularized acoustic echoes retrieval.**
In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 156–160. IEEE.

DiBiase, J. H., Silverman, H. F., and Brandstein, M. S. (2001).
Robust localization in reverberant rooms.
In *Microphone Arrays*, pages 157–180. Springer.

Doclo, S. and Moonen, M. (2002).
Gsvd-based optimal filtering for single and multimicrophone speech enhancement.
*IEEE Transactions on signal processing*, 50(9):2230–2244.

Dokmanić, I., Parhizkar, R., Walther, A., Lu, Y. M., and Vetterli, M. (2013).
Acoustic echoes reveal room shape.
*Proceedings of the National Academy of Sciences*, 110(30):12186–12191.

Dokmanić, I., Scheibler, R., and Vetterli, M. (2015).
Raking the cocktail party.
*IEEE journal of selected topics in signal processing*, 9(5):825–836.

Duval, V. and Peyré, G. (2017).
Sparse regularization on thin grids i: the lasso.
*Inverse Problems*, 33(5):055008.

Eaton, J., Gaubitch, N. D., Moore, A. H., and Naylor, P. A. (2015).
The ace challenge—corpus description and performance evaluation.
In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE.

Evers, C. and Naylor, P. A. (2018).
Acoustic slam.
*IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1484–1498.

Farina, A. (2007).
Advancements in impulse response measurements by sine sweeps.
In *Audio Engineering Society Convention 122*. Audio Engineering Society.

Filos, J., Canclini, A., Thomas, M. R., Antonacci, F., Sarti, A., and Naylor, P. A. (2011).
Robust inference of room geometry from acoustic measurements using the hough transform.
In *2011 19th European Signal Processing Conference*, pages 161–165. IEEE.

Flanagan, J. L., Surendran, A. C., and Jan, E.-E. (1993).
Spatially selective sound capture for speech and audio processing.
*Speech Communication*, 13(1-2):207–222.

Gannot, S., Vincent, E., Markovich-Golan, S., and Ozerov, A. (2017).
**A consolidated perspective on multimicrophone speech enhancement and source separation.**
*IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730.

Jensen, J. R., Saqib, U., and Gannot, S. (2019).
**An em method for multichannel toa and doa estimation of acoustic echoes.**
In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 120–124. IEEE.

Knapp, C. and Carter, G. (1976).
**The generalized correlation method for estimation of time delay.**
*IEEE transactions on acoustics, speech, and signal processing*, 24(4):320–327.

Kowalczyk, K., Habets, E. A., Kellermann, W., and Naylor, P. A. (2013).
**Blind system identification using sparse learning for tdoa estimation of room reflections.**
*IEEE Signal Processing Letters*, 20(7):653–656.

Kreković, M., Dokmanić, I., and Vetterli, M. (2016).
**Echoslam: Simultaneous localization and mapping with acoustic echoes.**
In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11–15. Ieee.

- Kuttruff, H. (2016).
  *Room acoustics.*
  CRC Press.
- Lebarbenchon, R., Camberlein, E., Di Carlo, D., Gaultier, C., Deleforge, A., and Bertin, N. (2018).
  Evaluation of an open-source implementation of the srp-phat algorithm within the 2018 locata challenge.
  *Proc. of LOCATA Challenge Workshop-a satellite event of IWAENC.*
- Leglaive, S., Badeau, R., and Richard, G. (2016).
  Multichannel audio source separation with probabilistic reverberation priors.
  *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2453–2465.
- Lin, Y., Chen, J., Kim, Y., and Lee, D. D. (2007).
  Blind sparse-nonnegative (bsn) channel identification for acoustic time-difference-of-arrival estimation.
  In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 106–109. IEEE.

Lin, Y., Chen, J., Kim, Y., and Lee, D. D. (2008).
Blind channel identification for speech dereverberation using l1-norm sparse learning.
In *Advances in Neural Information Processing Systems*, pages 921–928.

Nguyen, Q., Girin, L., Bailly, G., Elisei, F., and Nguyen, D.-C. (2018).
Autonomous sensorimotor learning for sound source localization by a humanoid robot.

Remaggi, L., Jackson, P. J., Coleman, P., and Wang, W. (2016).
Acoustic reflector localization: novel image source reversion and direct localization methods.
*IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(2):296–309.

Remaggi, L., Jackson, P. J., and Wang, W. (2019).
Modeling the comb filter effect and interaural coherence for binaural source separation.
*IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2263–2277.

Ribeiro, F., Ba, D., Zhang, C., and Florêncio, D. (2010).
**Turning enemies into friends: Using reflections to improve sound source localization.**
In *2010 IEEE International Conference on Multimedia and Expo*, pages 731–736. IEEE.

Salvati, D., Drioli, C., and Foresti, G. L. (2016).
**Sound source and microphone localization from acoustic impulse responses.**
*IEEE Signal Processing Letters*, 23(10):1459–1463.

Scheibler, R., Di Carlo, D., Deleforge, A., and Dokmanić, I. (2018).
**Separake: Source separation with a little help from echoes.**
In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6897–6901. IEEE.

Szöke, I., Skácel, M., Mošner, L., Paliesek, J., and Černockỳ, J. H. (2019).
**Building and evaluation of a real room impulse response dataset.**
*IEEE Journal of Selected Topics in Signal Processing*, 13(4):863–876.

Tong, L., Xu, G., and Kailath, T. (1994).
**Blind identification and equalization based on second-order statistics: A time domain approach.**
*IEEE Transactions on information Theory*, 40(2):340–349.

Tukuljac, H. P. (2020).
*Sparse and Parametric Modeling with Applications to Acoustics and Audio.*
PhD thesis, École polytechnique fédérale de Lausanne.