

# Echo-aware signal processing for audio scene analysis

---

Diego Di Carlo

December 3, 2020

**PhD supervisors:** Antoine Deleforge  
Nancy Bertin

**Jury members:** Simon Doclo (reviewer)  
Laurent Girin (reviewer)  
Fabio Antonacci (examiner)  
Renaud Seguier (examiner, president)

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

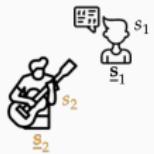
## **Introduction**

---

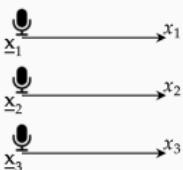
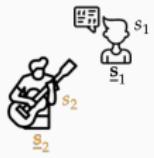
# Echo-aware signal processing for audio scene analysis

## Sound

- produced by **sources**



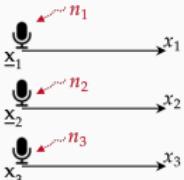
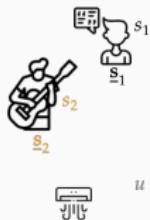
# Echo-aware signal processing for audio scene analysis



## Sound

- produced by **sources**
- recorded by (array of) **microphones**

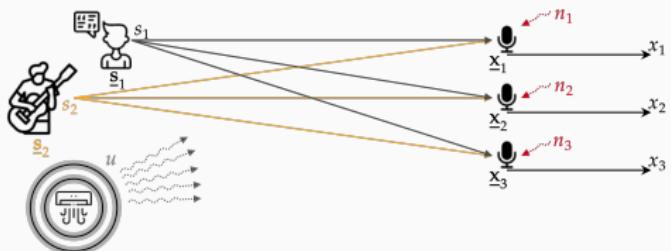
# Echo-aware signal processing for audio scene analysis



## Sound

- produced by **sources**
- recorded by (array of) **microphones**
- corrupted by **noise**

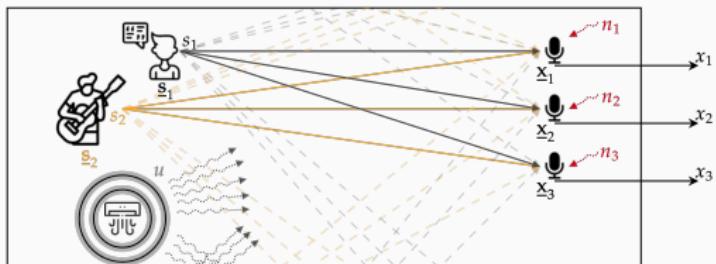
# Echo-aware signal processing for audio scene analysis



## Sound

- produced by **sources**
- recorded by (array of) **microphones**
- corrupted by **noise**
- propagates in the **space**

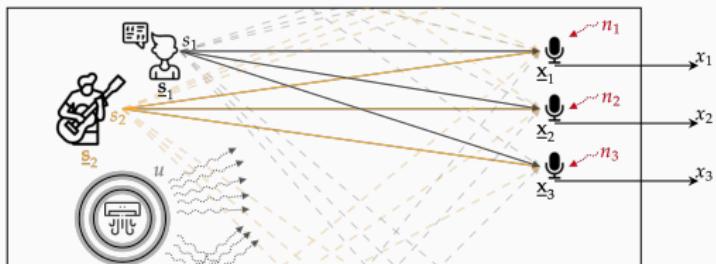
# Echo-aware signal processing for audio scene analysis



## Sound

- produced by **sources**
- recorded by (array of) **microphones**
- corrupted by **noise**
- propagates in the **space**
- interacts with the **room**  
     $\hookrightarrow$  **reverberation**

# Echo-aware signal processing for audio scene analysis



## Sound

- produced by **sources**
- recorded by (array of) **microphones**
- corrupted by **noise**
- propagates in the **space**
- interacts with the **room**  
     $\hookrightarrow$  **reverberation**

$\Sigma = \text{Audio Scene}$

## Echo-aware signal processing for audio scene analysis

# Echo-aware signal processing for audio scene analysis

Semantic information



on nature and content

# Echo-aware signal processing for audio scene analysis

Semantic information



on nature and content

Spatial information



on position and geometry

# Echo-aware signal processing for audio scene analysis

Semantic information



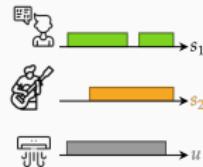
on nature and content

Spatial information



on position and geometry

Temporal information



on events activity

# Echo-aware signal processing for audio scene analysis

Semantic information



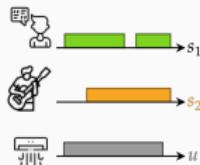
on nature and content

Spatial information



on position and geometry

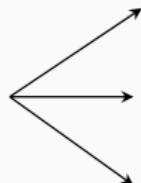
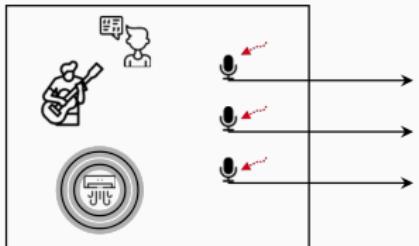
Temporal information



on events activity

## Audio Scene Analysis

Extraction and organization of all the information in the sound



# Echo-aware signal processing for audio scene analysis

Semantic information



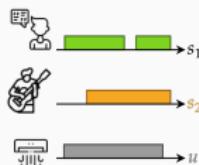
on nature and content

Spatial information



on position and geometry

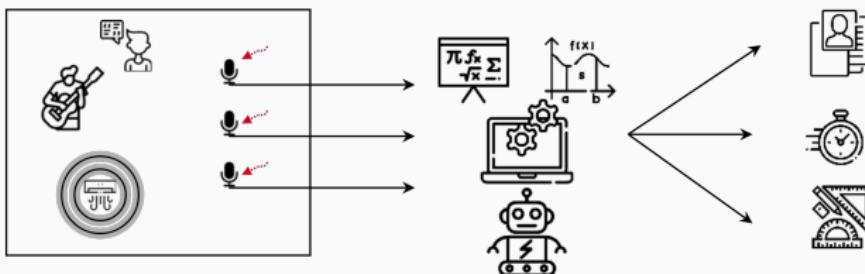
Temporal information



on events activity

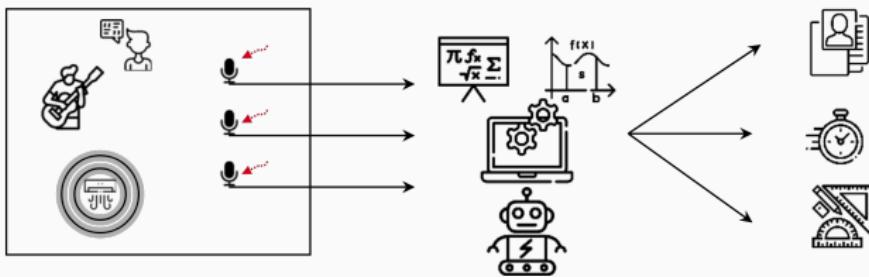
## Audio Scene Analysis

Extraction and organization of all the information in the sound

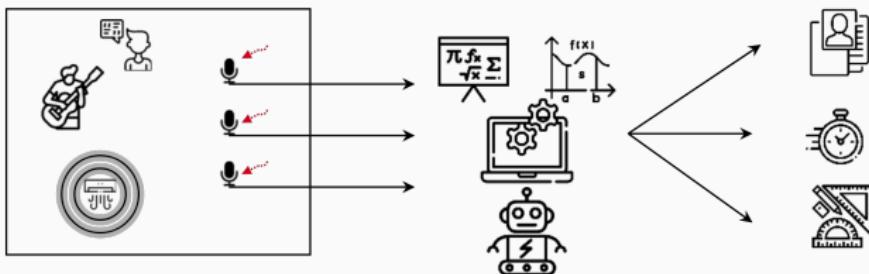


Can computers do it?

## Echo-aware signal processing for audio scene analysis



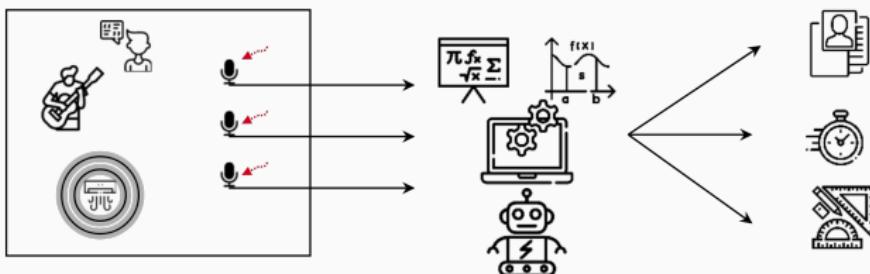
# Echo-aware signal processing for audio scene analysis



## Signal Processing

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

# Echo-aware signal processing for audio scene analysis



## Signal Processing

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

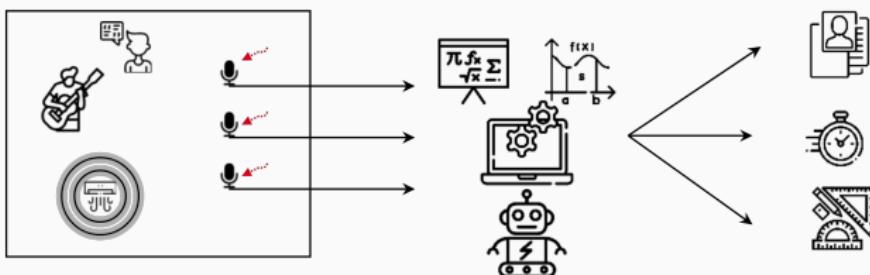
- Sound Source Separation
- Speech Enhancement
- Sound Source Localization
- Room Geometry Estimation

{ What?  
Where? }

- Voice Activity Detection
- Reverberation level estimation
- Acoustic Channel Estimation
- ...

{ When?  
How? }

# Echo-aware signal processing for audio scene analysis



## Signal Processing

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

- Sound Source Separation
  - Speech Enhancement
  - Sound Source Localization
  - Room Geometry Estimation
- { What?  
Where? }

- Voice Activity Detection
- Reverberation level estimation
- Acoustic Channel Estimation
- ...

{ When?  
How? }

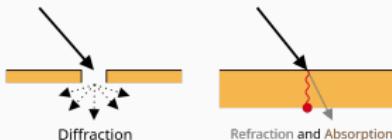
HOW  $\xrightarrow{\text{helps}}$  WHERE  $\xrightarrow{\text{helps}}$  WHAT  $\xrightarrow{\text{helps}}$  HOW  $\xrightarrow{\text{helps}}$  ...

# Echo-aware signal processing for audio scene analysis

**Sound interacts with indoor environment:**

it is reflected  
specularly and diffusely

- + it is absorbed,
- + it is transmitted,
- + it is diffracted, etc.

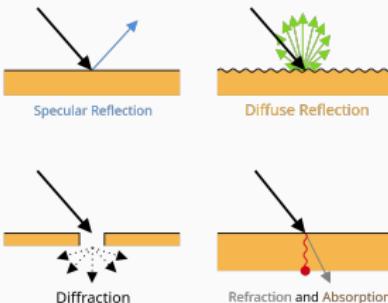


# Echo-aware signal processing for audio scene analysis

Sound interacts with indoor environment:

- it is reflected  
specularly and diffusely
- + it is absorbed,
- + it is transmitted,
- + it is diffracted, etc.

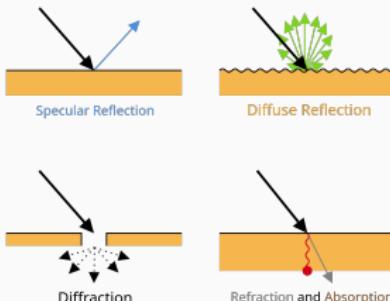
} = reverberation



# Echo-aware signal processing for audio scene analysis

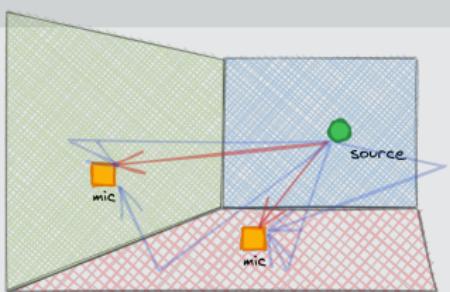
Sound interacts with indoor environment:

- it is reflected  
specularly and diffusely
- + it is absorbed,  
+ it is transmitted,  
+ it is diffracted, etc.
- } = reverberation



## Acoustic Echoes: early specular reflection

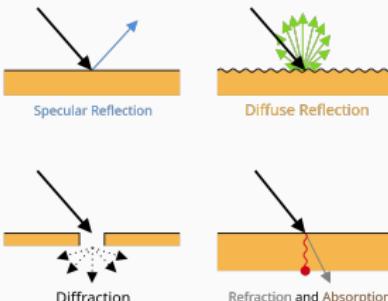
- Reflections “standing out” w.r.t. reverberation
- Copy of a sound but later
  - same content
  - delay  $\Leftrightarrow$  travelled distance



# Echo-aware signal processing for audio scene analysis

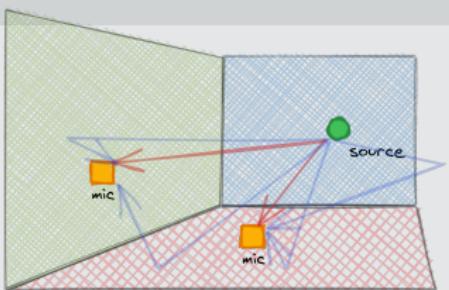
Sound interacts with indoor environment:

- it is reflected  
specularly and diffusely
- + it is absorbed,  
+ it is transmitted,  
+ it is diffracted, etc.
- } = reverberation



## Acoustic Echoes: early specular reflection

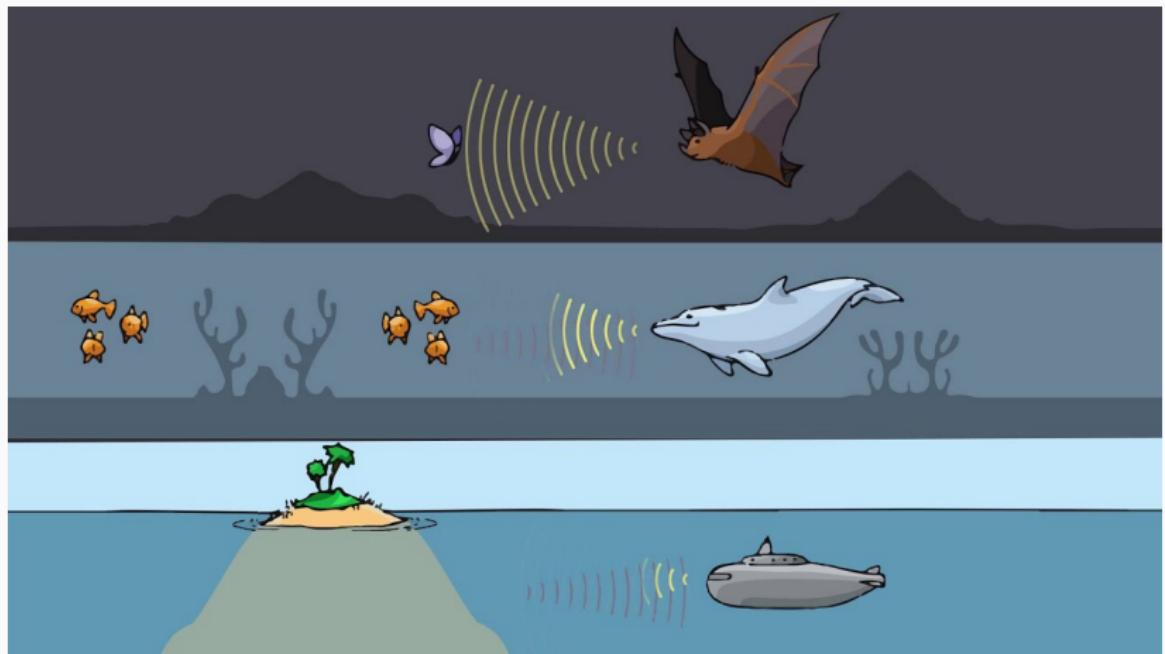
- Reflections “standing out” w.r.t. reverberation
- Copy of a sound but later
  - same content
  - delay  $\Leftrightarrow$  travelled distance



! idea: leverage this

## Echo-aware signal processing for audio scene analysis

Everyday examples: bats, dolphins and sonars



(© Skin & Bones)

## Echo-aware signal processing for audio scene analysis

In audio signal processing, sound propagation is typically

- **ignored**  $\Rightarrow$  simple processing but reverberation is noise
- **fully modeled** and estimated  $\Rightarrow$  very challenging

# Echo-aware signal processing for audio scene analysis

In audio signal processing, sound propagation is typically

- **ignored**  $\Rightarrow$  simple processing but reverberation is noise
- **fully modeled** and estimated  $\Rightarrow$  very challenging

## Echo-aware methods

- explicitly account for some acoustic reflections to boost the performance
- attractive alternative between ignoring reverberation and modelling it entirely

*Turning Enemies into Friends:  
Using reflections to improve sound source localization.*

[Ribeiro et al., 2010]

## Outline and contributions

---

**Thesis title:**

## Outline and contributions

---

**Thesis title:**

Echo-aware



better processing

## Outline and contributions

---

**Thesis title:**

Echo-aware



better processing

Signal Processing



models and frameworks

## Outline and contributions

**Thesis title:**

Echo-aware                      Signal Processing                      for Audio Scene Analysis  
↓                                  ↓                                  ↓  
better processing                models and frameworks                context and problems

# Outline and contributions

**Thesis title:**

Echo-aware                      Signal Processing                      for Audio Scene Analysis  
↓                                  ↓                                  ↓  
better processing                models and frameworks                context and problems

**Thesis contribution:**

1. How to estimate them?
2. How to use them?
3. Where to find them?

## Outline and contributions

### Thesis title:



## **Thesis contribution:**

- ## 1. How to estimate them?      2. How to use them?

- Learning-based method
  - Analytical method

### 3. Where to find them?

## Outline and contributions

## Thesis title:

```
graph TD; A[Echo-aware] --> B[better processing]; C[Signal Processing] --> D["models and frameworks"]; E["for Audio Scene Analysis"] --> F["context and problems"]
```

## **Thesis contribution:**

1. How to estimate them?
  2. How to use them?

- Learning-based method
  - Analytical method
  - Source Localization
  - Speech Enhancement
  - Source Separation(in the )
  - Room Geometry Estimation(in the )

- ### 3. Where to find them?

## Outline and contributions

### Thesis title:

```
graph TD; A[Echo-aware Signal Processing] --> B[better processing]; C[Signal Processing models and frameworks] --> D["models and frameworks"]; E[context and problems for Audio Scene Analysis] --> F["context and problems"]
```

### **Thesis contribution:**

- **Learning-based method**
  - **Analytical method**
  - **Source Localization**
  - **Speech Enhancement**
  - **Source Separation(in the )**
  - **Room Geometry Estimation(in the )**

## **Problem Statement**

---

# Signal model

For one source and  $I$  microphones:

$$\tilde{x}_i(t) = (\tilde{h}_i * \tilde{s})(t) + \tilde{n}_i(t) \quad i \in I$$

mic. signal  $\leftarrow$

source signal

noise term

⚠ continuous-time convolution

# Signal model

For one source and  $I$  microphones:

$$\tilde{x}_i(t) = (\tilde{h}_i * \tilde{s})(t) + \tilde{n}_i(t) \quad i \in I$$

mic. signal ←      → source signal  
                        noise term  
                        ⚠ continuous-time convolution

## Room Impulse Response (RIR)

- linear filtering effect of the sound propagation (reverberation)
- acoustic response of a room to a (perfect) impulsive sound
- depends on spatial properties (room geometry, mic/src position)  
→ one RIR for each microphone and source pair

## Signal model

For one source and  $I$  microphones:

$$\tilde{x}_i(t) = (\tilde{h}_i * \tilde{s})(t) + \tilde{n}_i(t) \quad i \in I$$

mic. signal ←  
source signal →  
noise term →  
⚠ continuous-time convolution

### Room Impulse Response (RIR)

- linear filtering effect of the sound propagation (reverberation)
- acoustic response of a room to a (perfect) impulsive sound
- depends on spatial properties (room geometry, mic/src position)  
→ one RIR for each microphone and source pair

In the Short Time Fourier Transform (STFT) domain:

$$X_i[f, t] = H_i[f]S[f, t] + N_i[f, t]$$

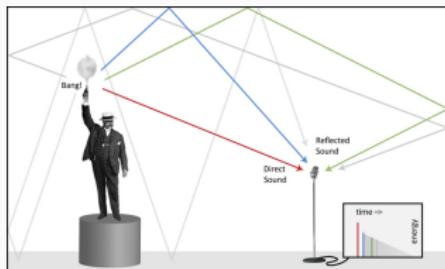
where  $X[f, t], H[f, t], S[f, t], N[f, t] \in \mathbb{C}$

Room Transfer Function (RTF)

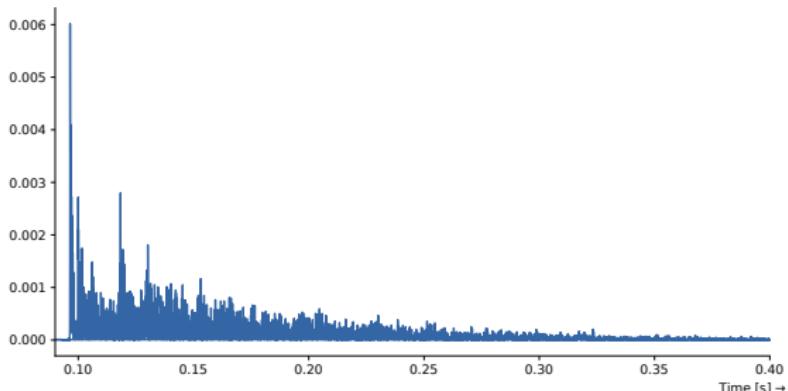
## Elements of RIR

$$\tilde{x}_i(t) = (\tilde{h}_i * \tilde{s})(t) + \tilde{n}(t)$$

$$\tilde{h}_i(t) = \tilde{h}_i^{\text{d}}(t) + \tilde{h}_i^{\text{e}}(t) + \tilde{h}_i^{\text{lrev}}(t)$$



(© acousticshonoursproject)

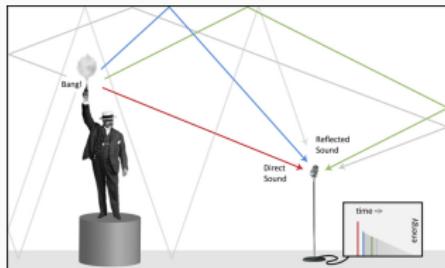


## Elements of RIR

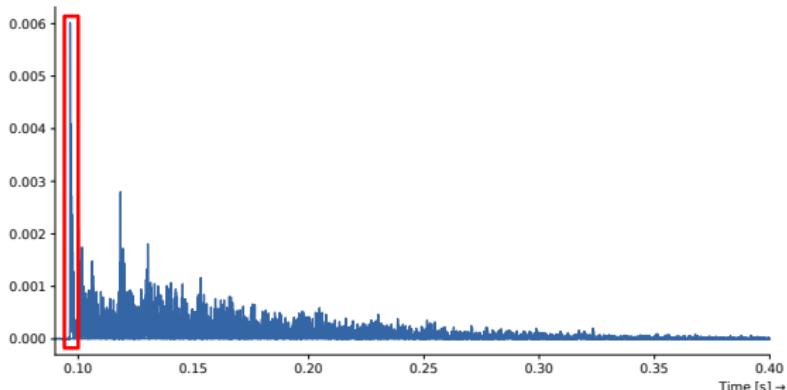
$$\tilde{x}_i(t) = (\tilde{h}_i * \tilde{s})(t) + \tilde{n}(t)$$

$$\tilde{h}_i(t) = \tilde{h}_i^d(t) + \tilde{h}_i^e(t) + \tilde{h}_i^{lrev}(t)$$

- $\tilde{h}_i^d(t)$  = direct path



(© acousticshonoursproject)

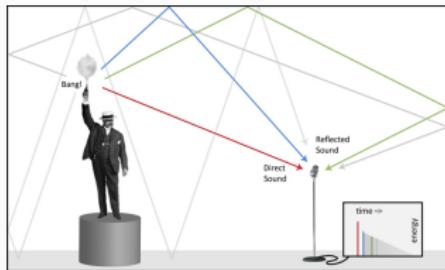


## Elements of RIR

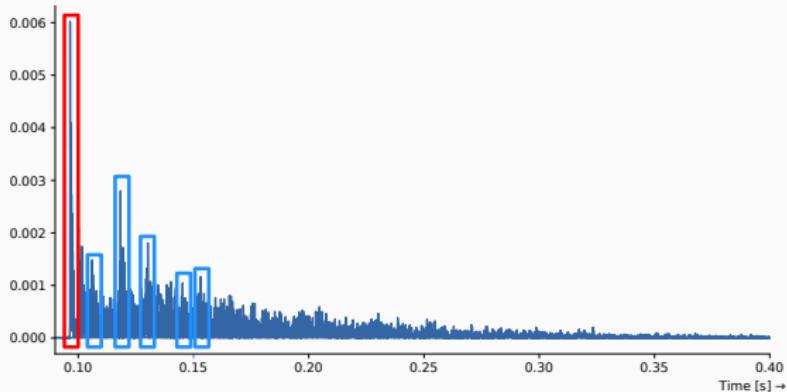
$$\tilde{x}_i(t) = (\tilde{h}_i * \tilde{s})(t) + \tilde{n}(t)$$

$$\tilde{h}_i(t) = \tilde{h}_i^d(t) + \tilde{h}_i^e(t) + \tilde{h}_i^{lrev}(t)$$

- $\tilde{h}_i^d(t)$  = direct path
- $\tilde{h}_i^e(t)$  = early reflection



(© acousticshonoursproject)

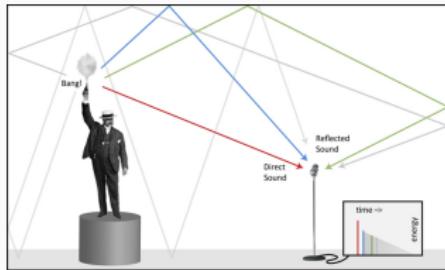


## Elements of RIR

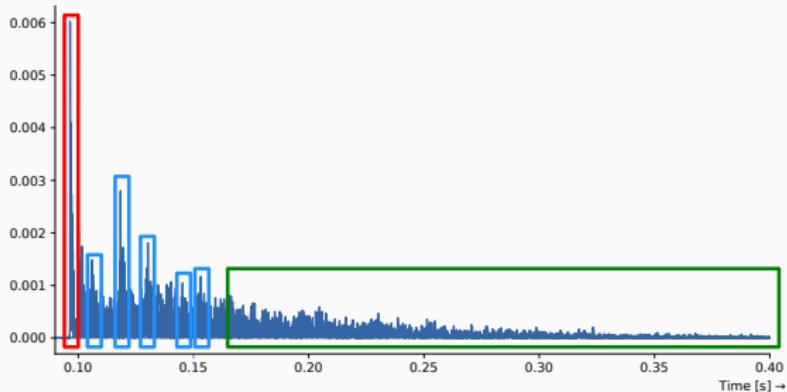
$$\tilde{x}_i(t) = (\tilde{h}_i * \tilde{s})(t) + \tilde{n}(t)$$

$$\tilde{h}_i(t) = \tilde{h}_i^d(t) + \tilde{h}_i^e(t) + \tilde{h}_i^{lrev}(t)$$

- $\tilde{h}_i^d(t)$  = direct path
- $\tilde{h}_i^e(t)$  = early reflection
- $\tilde{h}_i^{lrev}(t)$  = late reverberation



(© acousticshonoursproject)



## Problem Statement

Echoes can be modeled as sum of Dirac's delta functions:

$$\tilde{h}_i(t) = \tilde{h}_i^d(t) + \tilde{h}_i^e(t) + \varepsilon_i(t) \approx \sum_{r=0}^R \alpha_i^{(r)} \delta\left(t - \tau_i^{(r)}\right) + \underbrace{\varepsilon_i(t)}_{\text{models later echoes, reverberation and other.}}$$

# Problem Statement

Echoes can be modeled as sum of Dirac's delta functions:

$$\tilde{h}_i(t) = \tilde{h}_i^d(t) + \tilde{h}_i^e(t) + \varepsilon_i(t) \approx \sum_{r=0}^R \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) + \underbrace{\varepsilon_i(t)}_{\text{models later echoes, reverberation and other.}}$$

## Goal: Acoustic Echo Retrieval (AER)

Estimate  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$  from the microphone signals  $\{x_i\}_i$

# Problem Statement

Echoes can be modeled as sum of Dirac's delta functions:

$$\tilde{h}_i(t) = \tilde{h}_i^d(t) + \tilde{h}_i^e(t) + \varepsilon_i(t) \approx \sum_{r=0}^R \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) + \underbrace{\varepsilon_i(t)}_{\text{models later echoes, reverberation and other.}}$$

## Goal: Acoustic Echo Retrieval (AER)

Estimate  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$  from the microphone signals  $\{x_i\}_i$

## Challenges:

# Problem Statement

Echoes can be modeled as sum of Dirac's delta functions:

$$\tilde{h}_i(t) = \tilde{h}_i^d(t) + \tilde{h}_i^e(t) + \varepsilon_i(t) \approx \sum_{r=0}^R \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) + \underbrace{\varepsilon_i(t)}_{\text{models later echoes, reverberation and other.}}$$

## Goal: Acoustic Echo Retrieval (AER)

Estimate  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$  from the microphone signals  $\{x_i\}_i$

## Challenges:

- RIRs depend on the scene geometry (room, source and mic position)

# Problem Statement

Echoes can be modeled as sum of Dirac's delta functions:

$$\tilde{h}_i(t) = \tilde{h}_i^d(t) + \tilde{h}_i^e(t) + \varepsilon_i(t) \approx \sum_{r=0}^R \alpha_i^{(r)} \delta\left(t - \tau_i^{(r)}\right) + \underbrace{\varepsilon_i(t)}_{\text{models later echoes, reverberation and other.}}$$

## Goal: Acoustic Echo Retrieval (AER)

Estimate  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$  from the microphone signals  $\{x_i\}_i$

## Challenges:

- RIRs depend on the scene geometry (room, source and mic position)
- Big under-modelling error (late reverberation and external noise)

# Problem Statement

Echoes can be modeled as sum of Dirac's delta functions:

$$\tilde{h}_i(t) = \tilde{h}_i^d(t) + \tilde{h}_i^e(t) + \varepsilon_i(t) \approx \sum_{r=0}^R \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) + \underbrace{\varepsilon_i(t)}_{\text{models later echoes, reverberation and other.}}$$

## Goal: Acoustic Echo Retrieval (AER)

Estimate  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$  from the microphone signals  $\{x_i\}_i$

## Challenges:

- RIRs depend on the scene geometry (room, source and mic position)
- Big under-modelling error (late reverberation and external noise)
- In reality:  $\alpha_i^{(r)} \delta(t) \rightarrow (\alpha_i^{(r)} * \delta)(t)$  due to
  - frequency-dependent air attenuation, wall absorption, ...
  - sampling process

## **Acoustic Echo Estimation**

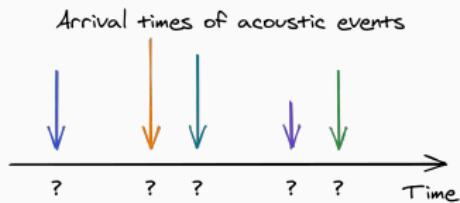
---



# Acoustic Echo Retrieval

Estimating early (strong) reflections from microphones recordings, i.e.,

$$\{\tilde{x}_i\}_i \rightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$





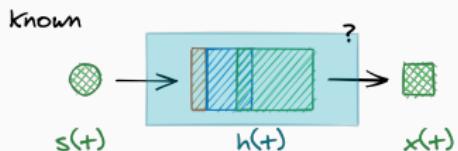
# Acoustic Echo Retrieval

Estimating early (strong) reflections from microphones recordings, i.e.,

$$\{\tilde{x}_i\}_i \rightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$



Two scenarios:



- 🔊 **intrusive** or specific setups
- ⌚ **non-blind** problem
  - (Applications: sonar, measurements, etc.)



# Acoustic Echo Retrieval

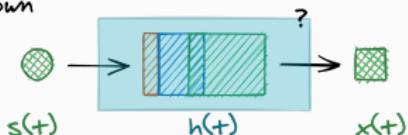
Estimating early (strong) reflections from microphones recordings, i.e.,

$$\{\tilde{x}_i\}_i \rightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$



Two scenarios:

Known

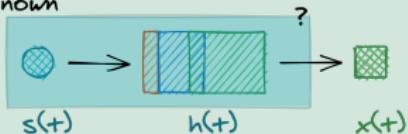


🔊 intrusive or specific setups

👁️ non-blind problem

(Applications: sonar, measurements, etc.)

unknown



🔊 passive and more common setups

👁️ blind inverse problem (harder)

(Applications: recording on smart speakers, etc.)



# Acoustic Echo Retrieval

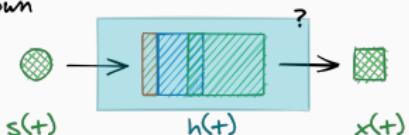
Estimating early (strong) reflections from microphones recordings, i.e.,

$$\{\tilde{x}_i\}_i \rightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$



**Two scenarios:**

Known

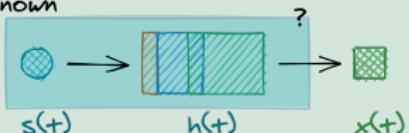


🔊 **intrusive** or specific setups

👁️ **non-blind** problem

(Applications: sonar, measurements, etc.)

unknown



🔊 **passive** and more common setups

👁️ **blind inverse** problem (harder)

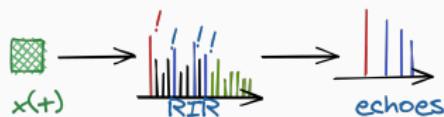
(Applications: recording on smart speakers, etc.)

**Our case:** one source and passive microphone array



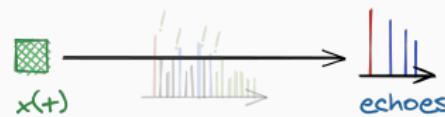
# Passive Acoustic Echo Retrieval

## RIR-based approaches



1. Discrete optimization  $\Rightarrow$  RIRs
2. Peak picking  $\Rightarrow$  Echoes

## RIR-agnostic approaches

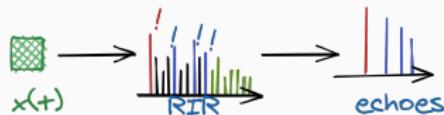


1. Direct estimation of  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}$  e.g., with maximum-likelihood



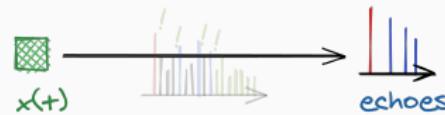
# Passive Acoustic Echo Retrieval

## RIR-based approaches



1. Discrete optimization  $\Rightarrow$  RIRs
2. Peak picking  $\Rightarrow$  Echoes

## RIR-agnostic approaches

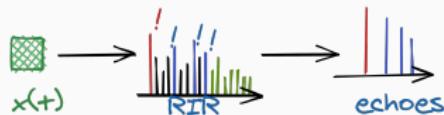


1. Direct estimation of  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}$  e.g., with maximum-likelihood



# Passive Acoustic Echo Retrieval

## RIR-based approaches

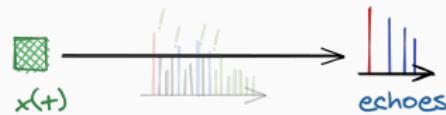


1. Discrete optimization  $\Rightarrow$  RIRs
2. Peak picking  $\Rightarrow$  Echoes

- ✓ BCE is well and known studied
- ✓ reasonably good for some application

[Crocco and Del Bue, 2016]

## RIR-agnostic approaches

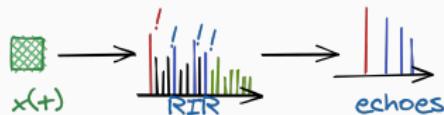


1. Direct estimation of  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}$  e.g., with maximum-likelihood



# Passive Acoustic Echo Retrieval

## RIR-based approaches

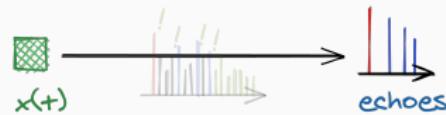


1. Discrete optimization  $\Rightarrow$  RIRs
2. Peak picking  $\Rightarrow$  Echoes

- ✓ BCE is well and known studied
- ✓ reasonably good for some application  
[Crocco and Del Bue, 2016]

- ✗ Full RIRs need to be estimated
- ✗ Peak picking has hyperparameters
- ✗ Issues due to *discrete estimation*

## RIR-agnostic approaches

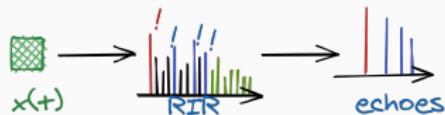


1. Direct estimation of  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}$  e.g., with maximum-likelihood



# Passive Acoustic Echo Retrieval

## RIR-based approaches

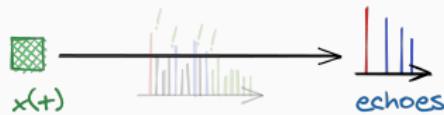


1. Discrete optimization  $\Rightarrow$  RIRs
2. Peak picking  $\Rightarrow$  Echoes

- ✓ BCE is well and known studied
- ✓ reasonably good for some application  
[Crocco and Del Bue, 2016]

- ✗ Full RIRs need to be estimated
- ✗ Peak picking has hyperparameters
- ✗ Issues due to *discrete estimation*

## RIR-agnostic approaches



1. Direct estimation of  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}$  e.g., with maximum-likelihood

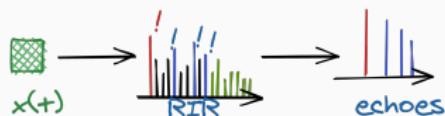
- ✓ No full RIRs & no peak picking
  - lower complexity
  - less hyperparameters

- ✗ exploratory 🌍  
(few works on audio)



# Passive Acoustic Echo Retrieval

## RIR-based approaches



1. Discrete optimization  $\Rightarrow$  RIRs
2. Peak picking  $\Rightarrow$  Echoes

- ✓ BCE is well and known studied
- ✓ reasonably good for some application  
[Crocco and Del Bue, 2016]

- ✗ Full RIRs need to be estimated
- ✗ Peak picking has hyperparameters
- ✗ Issues due to *discrete estimation*

## RIR-agnostic approaches



1. Direct estimation of  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}$  e.g., with maximum-likelihood

- ✓ No full RIRs & no peak picking
  - lower complexity
  - less hyperparameters

- ✗ exploratory 🌍  
(few works on audio)

**Proposed approach** RIR-agnostic & continuous:

1. Analytical approach
2. Learning-based approach

# (Discrete) RIR-based methods: the State of the Art



## Key ingredient – *Cross relation identity*

Signal model

$$x_1 = h_1 \star x$$

$$x_2 = h_2 \star x$$



# (Discrete) RIR-based methods: the State of the Art

## Key ingredient – *Cross relation identity*

Convolving with filters:

$$\textcolor{orange}{h_2} \star x_1 = \textcolor{orange}{h_2} \star h_1 \star x$$

$$\textcolor{orange}{h_1} \star x_2 = \textcolor{orange}{h_1} \star h_2 \star x$$



# (Discrete) RIR-based methods: the State of the Art

## Key ingredient – *Cross relation identity*

Commutativity of convolution:

$$h_2 \star x_1 = h_2 \star h_1 \star x$$

$$h_1 \star x_2 = \underbrace{h_2 \star h_1}_{\text{commutativity}} \star x$$

# (Discrete) RIR-based methods: the State of the Art



## Key ingredient – *Cross relation identity*

Subtraction

$$\begin{aligned} h_2 \star x_1 &= h_2 \star h_1 \star x \\ h_1 \star x_2 &= h_2 \star h_1 \star x \end{aligned} \quad \left. \right\} \rightarrow \textcolor{orange}{x_1 \star h_2 - x_2 \star h_1 = 0}$$

# (Discrete) RIR-based methods: the State of the Art



## Key ingredient – *Cross relation identity*

Subtraction

$$\begin{aligned} h_2 \star x_1 &= h_2 \star h_1 \star x \\ h_1 \star x_2 &= h_2 \star h_1 \star x \end{aligned} \quad \left. \right\} \rightarrow x_1 \star h_2 - x_2 \star h_1 = 0$$

Ideas:

1. Echo TOAs  $\propto$  sampling frequency
2. Find echoes  $\rightarrow$  find sparse non-negative vectors  $h_1, h_2$  of length  $L$
3. Modeled as Lasso-like problem

# (Discrete) RIR-based methods: the State of the Art



## Key ingredient – *Cross relation identity*

Subtraction

$$\begin{aligned} h_2 \star x_1 &= h_2 \star h_1 \star x \\ h_1 \star x_2 &= h_2 \star h_1 \star x \end{aligned} \quad \left. \right\} \rightarrow x_1 \star h_2 - x_2 \star h_1 = 0$$

Ideas:

1. Echo TOAs  $\propto$  sampling frequency
2. Find echoes  $\rightarrow$  find sparse non-negative vectors  $h_1, h_2$  of length  $L$
3. Modeled as Lasso-like problem

$$\hat{h}_1, \hat{h}_2 \in \arg \min_{h_1, h_2 \in \mathbf{R}^n} \|x_1 \star h_2 - x_2 \star h_1\|_2^2 + \lambda \mathcal{P}(h_1, h_2) \quad \text{s.t.} \quad \mathcal{C}(h_1, h_2)$$

$\mathcal{P}(h_1, h_2) \rightarrow$  sparse promoting regularizer       $\mathcal{C}(h_1, h_2) \rightarrow$  constraints e.g. nonnegativity anchor



# (Discrete) RIR-based methods: the State of the Art

## Key ingredient – *Cross relation identity*

Subtraction

$$\begin{aligned} h_2 * x_1 &= h_2 * h_1 * x \\ h_1 * x_2 &= h_2 * h_1 * x \end{aligned} \quad \left. \right\} \rightarrow x_1 * h_2 - x_2 * h_1 = 0$$

Ideas:

1. Echo TOAs  $\propto$  sampling frequency
2. Find echoes  $\rightarrow$  find sparse non-negative vectors  $h_1, h_2$  of length  $L$
3. Modeled as Lasso-like problem

$$\hat{h}_1, \hat{h}_2 \in \arg \min_{h_1, h_2 \in \mathbf{R}^n} \|x_1 * h_2 - x_2 * h_1\|_2^2 + \lambda \mathcal{P}(h_1, h_2) \quad \text{s.t.} \quad \mathcal{C}(h_1, h_2)$$

$\mathcal{P}(h_1, h_2) \rightarrow$  sparse promoting regularizer       $\mathcal{C}(h_1, h_2) \rightarrow$  constraints e.g. nonnegativity anchor

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



## Proposed approach: analytical & continuous

 C. Elvira.

**Observation 1:** the cross-relation remains true in the **continuous** frequency domain

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

## Proposed approach: analytical & continuous



 C. Elvira.

**Observation 1:** the cross-relation remains true in the **continuous frequency domain**

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

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



## Proposed approach: analytical & continuous

 C. Elvira.

**Observation 1:** the cross-relation remains true in the **continuous frequency domain**

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

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

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

$$\mathbf{X}_i = \text{DFT}(x_i) \simeq \mathcal{F}\tilde{x}_i(nF_s) \quad n = 0 \dots N - 1$$



## Proposed approach: analytical & continuous

 C. Elvira.

**Observation 1:** the cross-relation remains true in the **continuous** frequency domain

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

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

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

$$\mathbf{X}_i = \text{DFT}(x_i) \simeq \mathcal{F}\tilde{x}_i(nF_s) \quad n = 0 \dots N - 1$$

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

~ **Lasso**, but  $\mathcal{F}h_i(f)$  is a continuous function → **BLasso** [Azais et al., 2015]

✓ No huge matrix

✓ Solutions are trains of Dirac

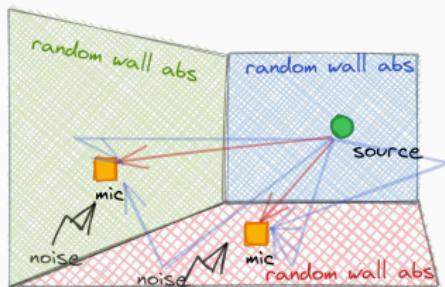
✓ No peak picking

✓ Perfect in noiseless & synthetic case

# 实验结果

## Syntetic Dataset at 16 kHz

- 2 microphones, 1 sound source (noise and speech)
- shoebox with random geometry
- $\mathcal{D}^{\text{SNR}}$ : SNR  $\in [0, 20]$  dB, RT<sub>60</sub> = 400 ms
- $\mathcal{D}^{\text{RT60}}$ : RT<sub>60</sub> = [100, 1000] ms, SNR = 20 dB



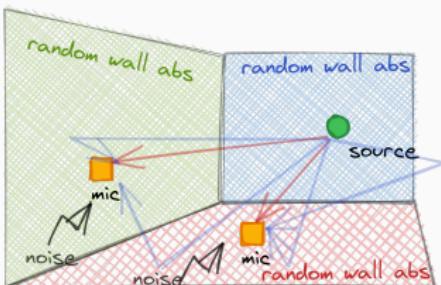
<sup>1</sup>[Lin et al., 2007]

<sup>2</sup>[Crocco and Del Bue, 2015]

# 实验结果

## Syntetic Dataset at 16 kHz

- 2 microphones, 1 sound source (noise and speech)
- shoebox with random geometry
- $\mathcal{D}^{\text{SNR}}$ : SNR  $\in [0, 20]$  dB, RT<sub>60</sub> = 400 ms
- $\mathcal{D}^{\text{RT60}}$ : RT<sub>60</sub> = [100, 1000] ms, SNR = 20 dB



**Baselines:** discrete RIR-based methods based on LASSO

- BSN: Blind, Sparse and Non-negative<sup>1</sup>
- IL1C: iteratively-weighted  $\ell_1$  constraint<sup>2</sup> → State of the Art

hyperparameters and peak-picking tuned via cross-validation

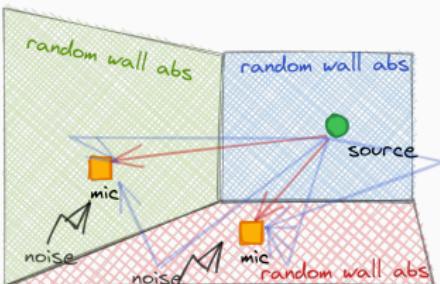
<sup>1</sup>[Lin et al., 2007]

<sup>2</sup>[Crocco and Del Bue, 2015]

# 实验结果

## Syntetic Dataset at 16 kHz

- 2 microphones, 1 sound source (noise and speech)
- shoebox with random geometry
- $\mathcal{D}^{\text{SNR}}$ : SNR  $\in [0, 20]$  dB, RT<sub>60</sub> = 400 ms
- $\mathcal{D}^{\text{RT60}}$ : RT<sub>60</sub> = [100, 1000] ms, SNR = 20 dB



**Baselines:** discrete RIR-based methods based on LASSO

- BSN: Blind, Sparse and Non-negative<sup>1</sup>
- IL1C: iteratively-weighted  $\ell_1$  constraint<sup>2</sup> → State of the Art

hyperparameters and peak-picking tuned via cross-validation

**Proposed method:** Blind and Sparse Technique for Echo Retrieval (**Blaster**)

<sup>1</sup>[Lin et al., 2007]

<sup>2</sup>[Crocco and Del Bue, 2015]

# ¶ Precision per # of echoes

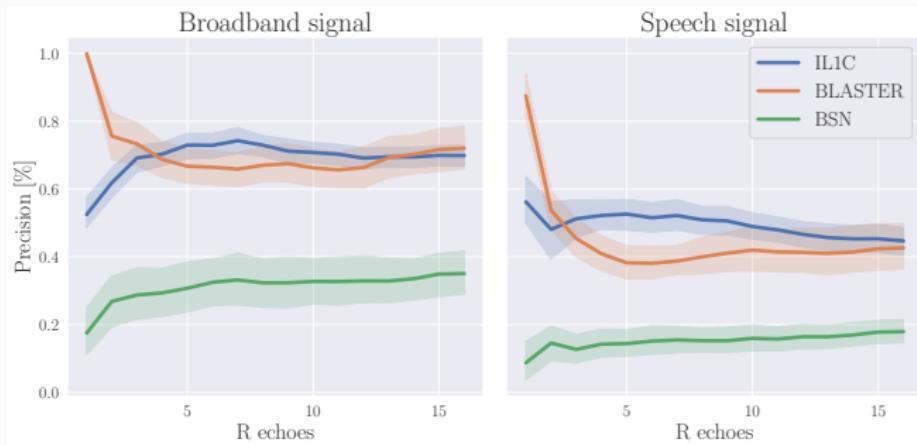


**Metric:** **Precision** = how many estimated echoes are correct (within 2 samples)

# ¶ Precision per # of echoes



**Metric:** Precision = how many estimated echoes are correct (within 2 samples)

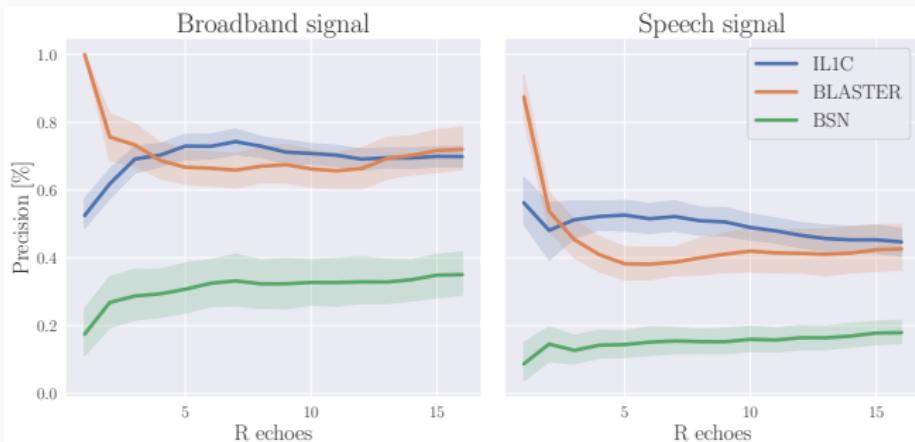


( $RT_{60} = 400$  ms and SNR = 20 dB.)

# Precision per # of echoes



**Metric:** Precision = how many estimated echoes are correct (within 2 samples)



( $RT_{60} = 400$  ms and SNR = 20 dB.)

✗ Sensitive  
to # echoes

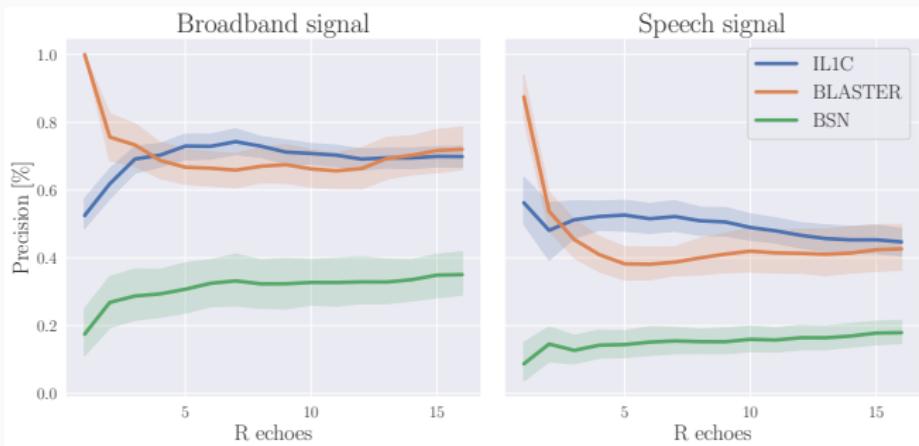
✗ Sensitive  
source signal

✓ Good for  
2 echoes

# ⚠ Precision per # of echoes



**Metric:** Precision = how many estimated echoes are correct (within 2 samples)



( $RT_{60} = 400$  ms and SNR = 20 dB.)

✗ Sensitive  
to # echoes

✗ Sensitive  
source signal

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

# ⚠ Error per Dataset/Signal while recovering 7 echoes

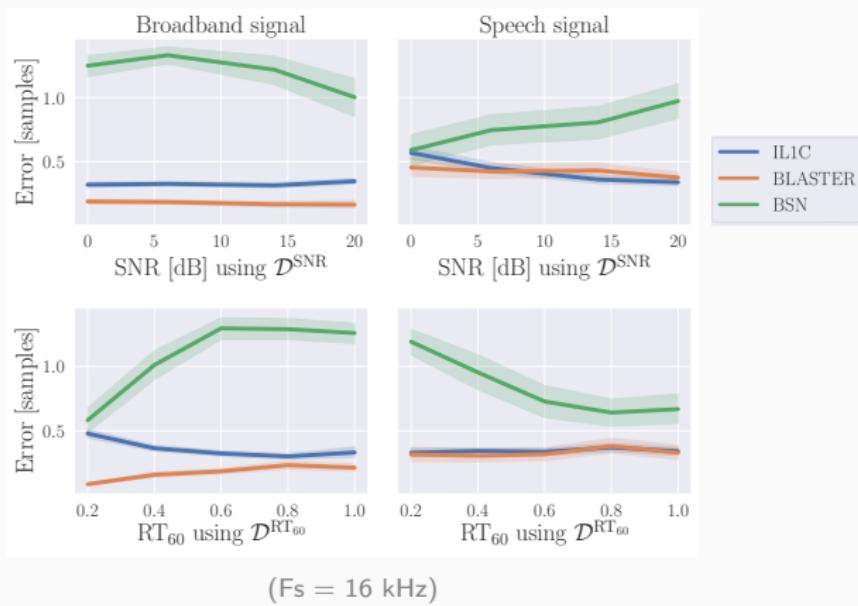


Metric: **RMSE** on matched echoes = error on the correct guess

# ⚠ Error per Dataset/Signal while recovering 7 echoes



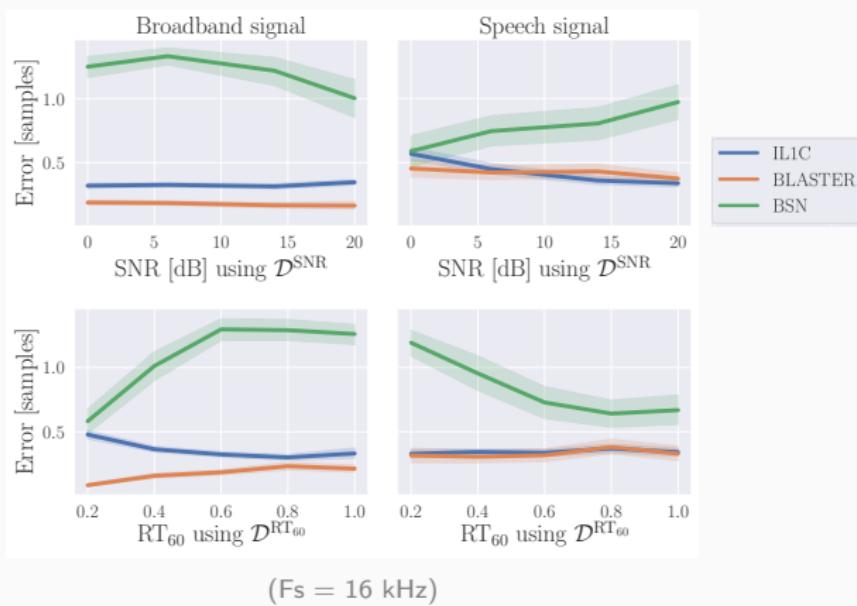
Metric: **RMSE** on matched echoes = error on the correct guess



# ⚠ Error per Dataset/Signal while recovering 7 echoes



Metric: RMSE on matched echoes = error on the correct guess



✓ Lower RMSE

✓ Robustness  
to SNR and RT<sub>60</sub>

✗ Source signal  
dependent



## Learning-based approach



## Learning-based approach

### Idea:

1. Use **virtually supervised deep** learning models
2. Estimate first echo (**simple but important**) (➡ used in the next section)
3. Only 2 microphones attending 1 sound source



## Learning-based approach

### Idea:

1. Use **virtually supervised deep** learning models
2. Estimate first echo (**simple but important**) (⬅ used in the next section)
3. Only 2 microphones attending 1 sound source

### Motivations:

- $x_i \rightarrow \tau_i^{(r)}$  is difficult, while  $\tau_i^{(r)} \rightarrow x_i$  "is not"  
→ acoustic simulators: mic/src/room geometry →  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}$ ,  $\tilde{h}_i$ ,  $\tilde{x}_i$



## Learning-based approach

### Idea:

1. Use **virtually** supervised **deep** learning models
2. Estimate first echo (**simple but important**) (◀ used in the next section)
3. Only 2 microphones attending 1 sound source

### Motivations:

- $x_i \rightarrow \tau_i^{(r)}$  is difficult, while  $\tau_i^{(r)} \rightarrow x_i$  "is not"  
→ acoustic simulators: mic/src/room geometry →  $\{\tau_i^{(r)}, \alpha_i^{(r)}\}$ ,  $\tilde{h}_i$ ,  $\tilde{x}_i$
- Acoustic simulator are "simple", versatile and fast  
→ allow to create large dataset



## Learning-based approach

### Inputs:

Interchannel level and phase difference features from

$$R[f] = \text{avg.} \frac{X_2[f, t]}{X_1[f, t]} \approx \text{avg.} \frac{H_2[f] S[f, t]}{H_1[f] S[f, t]}$$

≈ the relative transfer function → remove source dependency



## Learning-based approach

**Inputs:**

Interchannel level and phase difference features from

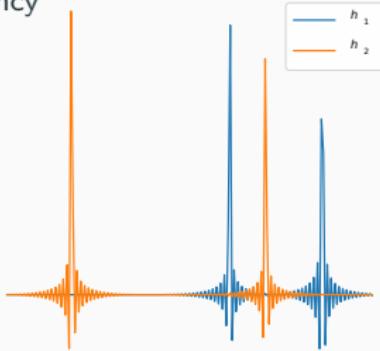
$$R[f] = \text{avg.} \frac{X_2[f, t]}{X_1[f, t]} \approx \text{avg.} \frac{H_2[f] S[f, t]}{H_1[f] S[f, t]}$$

≈ the relative transfer function → remove source dependency

**Outputs:**

Inter and intra Time Differences of Arrivals (TDOAs)

HP: close-surface scenario: first ⇔ strongest echo





## Learning-based approach

**Inputs:**

Interchannel level and phase difference features from

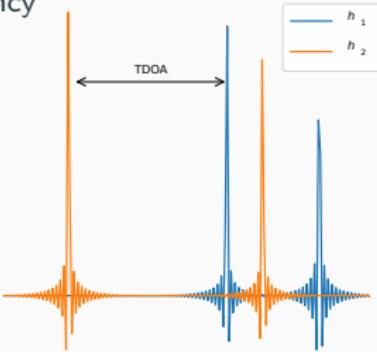
$$R[f] = \text{avg.} \frac{X_2[f, t]}{X_1[f, t]} \approx \text{avg.} \frac{H_2[f] S[f, t]}{H_1[f] S[f, t]}$$

≈ the relative transfer function → remove source dependency

**Outputs:**

Inter and intra Time Differences of Arrivals (TDOAs)

HP: close-surface scenario: first ⇔ strongest echo





## Learning-based approach

### Inputs:

Interchannel level and phase difference features from

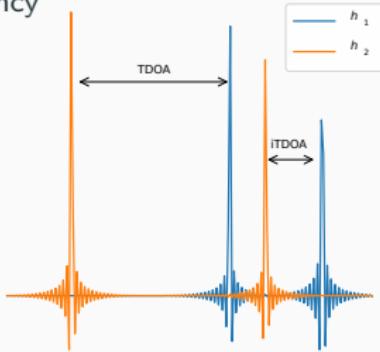
$$R[f] = \text{avg.} \frac{X_2[f, t]}{X_1[f, t]} \approx \text{avg.} \frac{H_2[f] S[f, t]}{H_1[f] S[f, t]}$$

≈ the relative transfer function → remove source dependency

### Outputs:

Inter and intra Time Differences of Arrivals (TDOAs)

HP: close-surface scenario: first ⇔ strongest echo





## Learning-based approach

### Inputs:

Interchannel level and phase difference features from

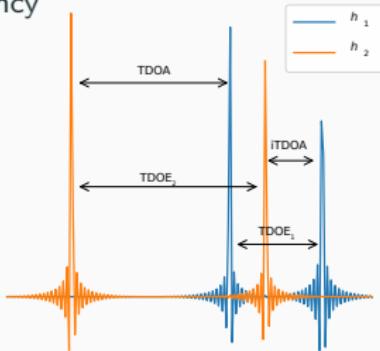
$$R[f] = \text{avg.} \frac{X_2[f, t]}{X_1[f, t]} \approx \text{avg.} \frac{H_2[f] S[f, t]}{H_1[f] S[f, t]}$$

≈ the relative transfer function → remove source dependency

### Outputs:

Inter and intra Time Differences of Arrivals (TDOAs)

HP: close-surface scenario: first ⇔ strongest echo





# Learning-based approach

## Inputs:

Interchannel level and phase difference features from

$$R[f] = \text{avg. } \frac{X_2[f, t]}{X_1[f, t]} \approx \text{avg. } \frac{H_2[f] S[f, t]}{H_1[f] S[f, t]}$$

≈ the relative transfer function → remove source dependency

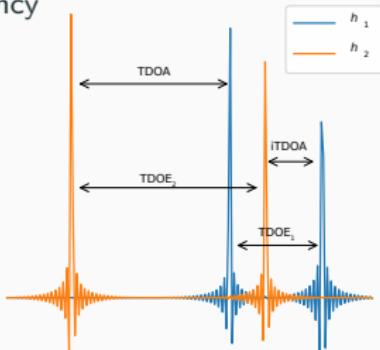
## Outputs:

Inter and intra Time Differences of Arrivals (TDOAs)

HP: close-surface scenario: first ⇔ strongest echo

## Loss Functions

1. RMSE (Multi-label regression) → TDOAs
2. Gaussian log-likelihood →  $\{\mu_\tau, \sigma_\tau^2\} \forall \tau \in \text{TDOAs}$
3. Student log-likelihood →  $\{\mu_\tau, \lambda_\tau, \nu_\tau\} \forall \tau \in \text{TDOAs}$





# Learning-based approach

## Inputs:

Interchannel level and phase difference features from

$$R[f] = \text{avg. } \frac{X_2[f, t]}{X_1[f, t]} \approx \text{avg. } \frac{H_2[f] S[f, t]}{H_1[f] S[f, t]}$$

≈ the relative transfer function → remove source dependency

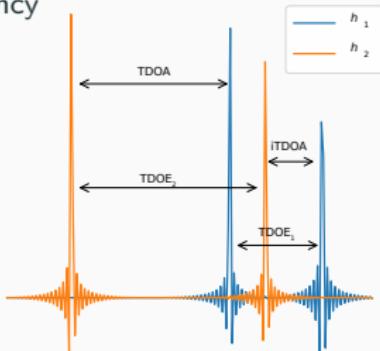
## Outputs:

Inter and intra Time Differences of Arrivals (TDOAs)

HP: close-surface scenario: first ⇔ strongest echo

## Loss Functions

1. RMSE (Multi-label regression) → TDOAs
2. Gaussian log-likelihood →  $\{\mu_\tau, \sigma_\tau^2\} \forall \tau \in \text{TDOAs}$
3. Student log-likelihood →  $\{\mu_\tau, \lambda_\tau, \nu_\tau\} \forall \tau \in \text{TDOAs}$



**Architectures:** MLP, CNN [Chakrabarty and Habets, 2017, Nguyen et al., 2018]

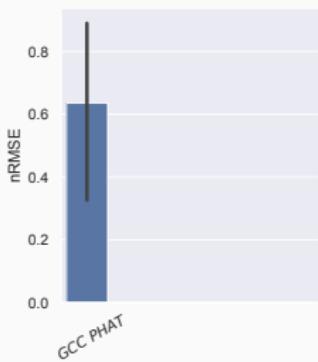


## 实验结果 Experimental results

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

**Baseline:** GCC PHAT [Knapp and Carter, 1976]

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



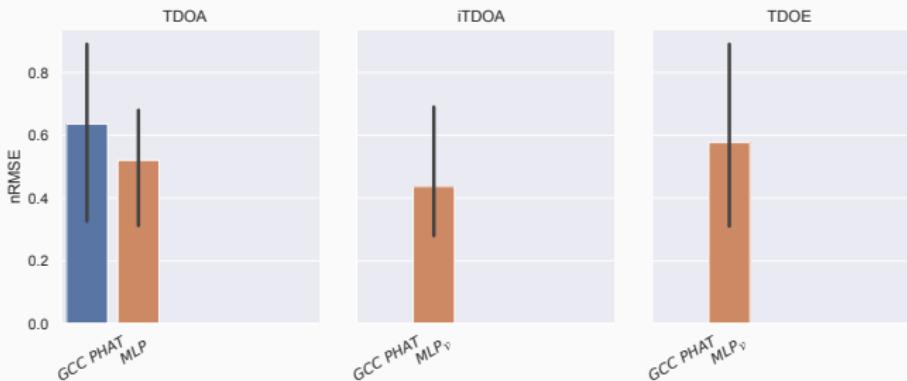


## 实验结果 Experimental results

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

**Baseline:** GCC PHAT [Knapp and Carter, 1976]

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



**Observation:**

- ✓ MLP outperforms GCC PHAT on TDOA estimation

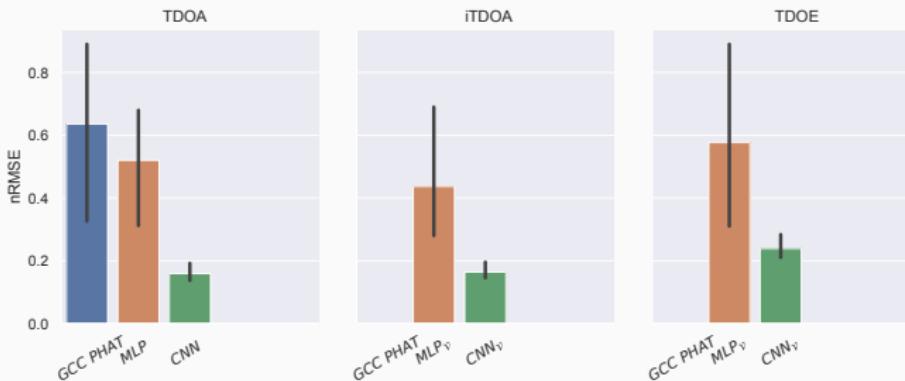


## 实验结果 Experimental results

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

**Baseline:** GCC PHAT [Knapp and Carter, 1976]

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



### Observation:

- ✓ MLP outperforms GCC PHAT on TDOA estimation
- ✓ CNN outperforms MLP (lower error and smaller variance)

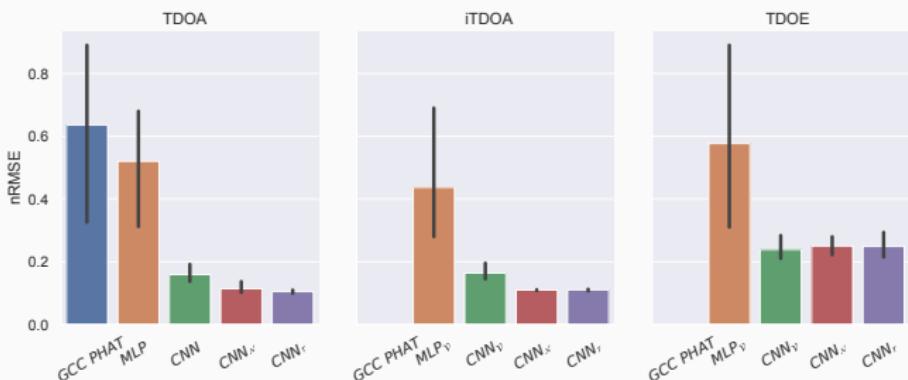


## 实验结果

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

**Baseline:** GCC PHAT [Knapp and Carter, 1976]

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



### Observation:

- ✓ MLP outperforms GCC PHAT on TDOA estimation
- ✓ CNN outperforms MLP (lower error and smaller variance)
- ✓  $\text{CNN}_{\mathcal{N}}$  and  $\text{CNN}_{\mathcal{T}}$  outperform CNN (lower error and smaller variance)
- ✗ TDOA between DP and 1<sup>st</sup> echo more difficult

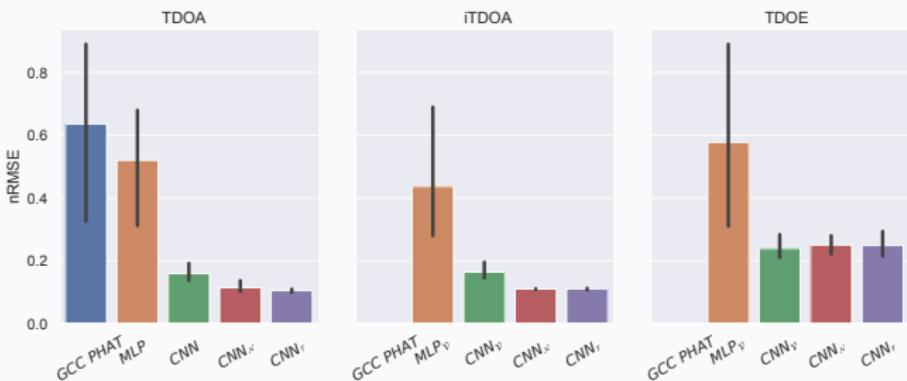


## 实验结果

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

**Baseline:** GCC PHAT [Knapp and Carter, 1976]

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



### Observation:

- ✓ MLP outperforms GCC PHAT on TDOA estimation
- ✓ CNN outperforms MLP (lower error and smaller variance)
- ✓  $\text{CNN}_{\mathcal{N}}$  and  $\text{CNN}_{\mathcal{T}}$  outperform CNN (lower error and smaller variance)
- ✗ TDOA between DP and 1<sup>st</sup> echo more difficult

More echoes  
 Real data

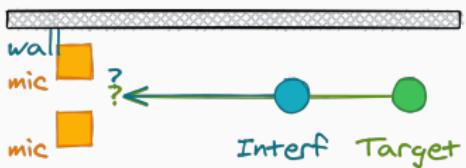
## **Echo-aware Application**

---



# Echo-aware Applications

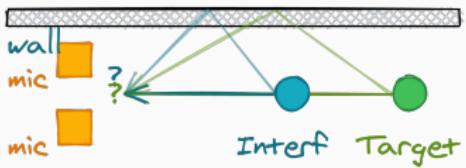
Echoes: same content, different time/direction





# Echo-aware Applications

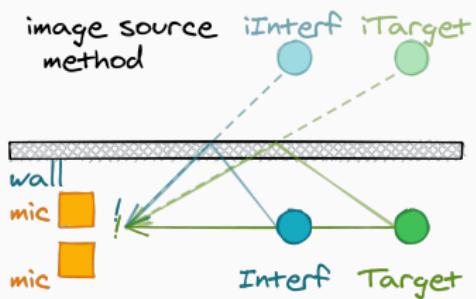
Echoes: same content, different time/direction





## Echo-aware Applications

Echoes: same content, different time/direction





# Echo-aware Applications

Echoes: same content, different time/direction

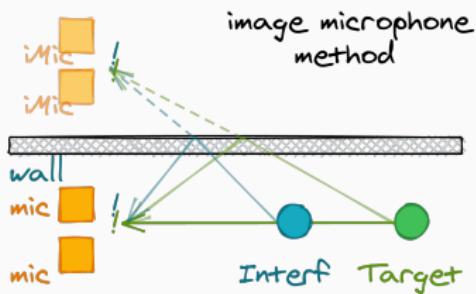
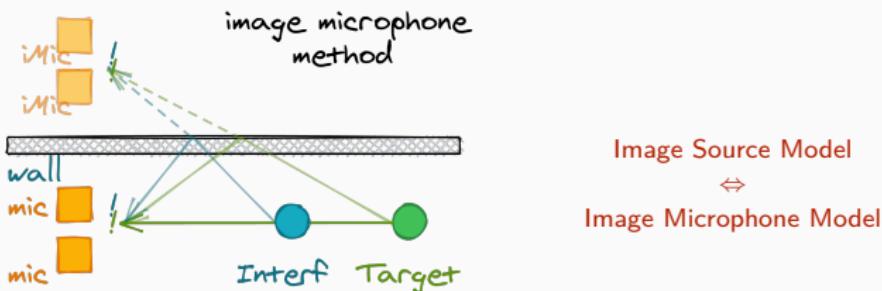


Image Source Model  
↔  
Image Microphone Model



# Echo-aware Applications

Echoes: same content, different time/direction



## What?

Echoes = copy

- Sound Source Separation  
[Leglaive et al., 2016]
- Speech Enhancement  
[Flanagan et al., 1993,  
Dokmanić et al., 2015,  
Kowalczyk, 2019]

(In-depth literature review in .)



# Echo-aware Applications

Echoes: same content, different time/direction

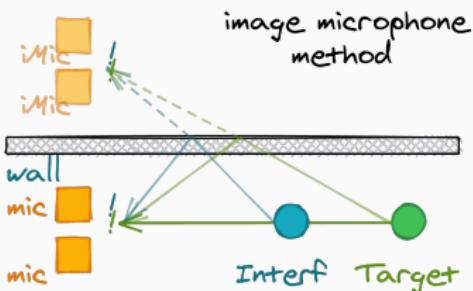


Image Source Model  
↔  
Image Microphone Model

## What?

Echoes = copy

- Sound Source Separation  
[Leglaive et al., 2016]
- Speech Enhancement  
[Flanagan et al., 1993,  
Dokmanić et al., 2015,  
Kowalczyk, 2019]

## Where?

Echoes  $\leftarrow$  image

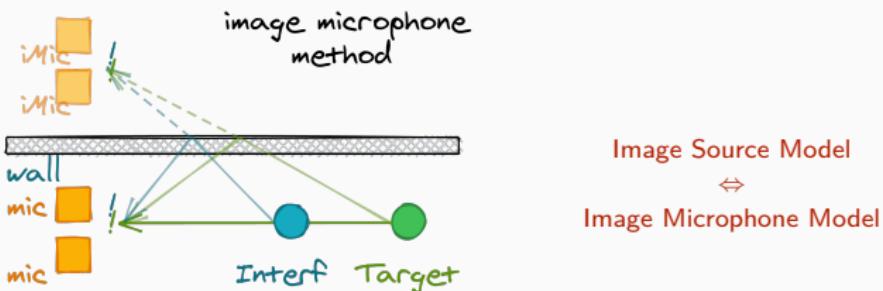
- Sound Source Localization  
[Ribeiro et al., 2010,  
Jensen et al., 2019]
- Room Geometry  
Estimation  
[Antonacci et al., 2012,  
Crocco et al., 2017]

(In-depth literature review in .



# Echo-aware Applications

Echoes: same content, different time/direction



## What?

Echoes = copy

- Sound Source Separation [Leglaive et al., 2016]
- Speech Enhancement [Flanagan et al., 1993, Dokmanić et al., 2015, Kowalczyk, 2019]

## Where?

Echoes  $\leftarrow$  image

- Sound Source Localization [Ribeiro et al., 2010, Jensen et al., 2019]
- Room Geometry Estimation [Antonacci et al., 2012, Crocco et al., 2017]

## How?

Echoes  $\in$  sound propagation

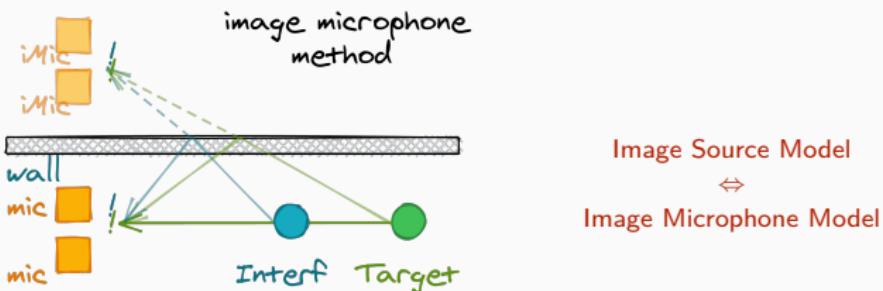
- Blind Channel Estimation [Lin et al., 2007, Crocco et al., 2017]
- Acoustic Measurements [Eaton et al., 2015, Kuttruff, 2016]

(In-depth literature review in .



# Echo-aware Applications

Echoes: same content, different time/direction



## What?

Echoes = copy

- Sound Source Separation [Leglaive et al., 2016]
- Speech Enhancement [Flanagan et al., 1993, Dokmanić et al., 2015, Kowalczyk, 2019]

## Where?

Echoes  $\leftarrow$  image

- Sound Source Localization [Ribeiro et al., 2010, Jensen et al., 2019]
- Room Geometry Estimation [Antonacci et al., 2012, Crocco et al., 2017]

## How?

Echoes  $\in$  sound propagation

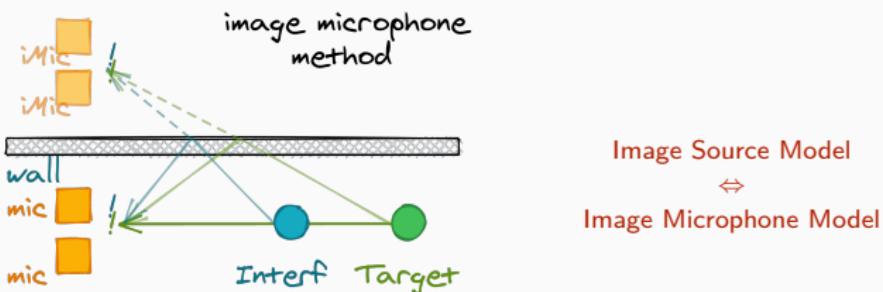
- Blind Channel Estimation [Lin et al., 2007, Crocco et al., 2017]
- Acoustic Measurements [Eaton et al., 2015, Kuttruff, 2016]

(In-depth literature review in .



# Echo-aware Applications

Echoes: same content, different time/direction



## What?

Echoes = copy

- Sound Source Separation [Leglaive et al., 2016]
- **Speech Enhancement** [Flanagan et al., 1993, Dokmanić et al., 2015, Kowalczyk, 2019]

## Where?

Echoes  $\leftarrow$  image

- **Sound Source Localization** [Ribeiro et al., 2010, Jensen et al., 2019]
- Room Geometry Estimation [Antonacci et al., 2012, Crocco et al., 2017]

## How?

Echoes  $\in$  sound propagation

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

(In-depth literature review in .

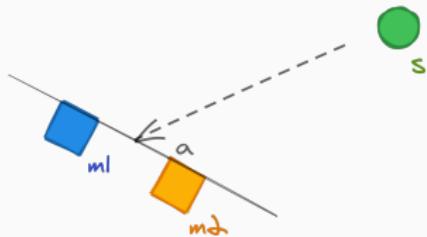
# Sound Source Localization (SSL)

(common knowledge) 

We do not consider here distance estimation.

## SSL with 2 microphones

- Only one angle of arrival (AOA) 
- Can be approximated from TDOA using e.g. GCC PHAT<sup>1</sup>  
(known limitation, but good in practice)<sup>2</sup>



<sup>1</sup> [Knapp and Carter, 1976]

<sup>2</sup> [DiBiase et al., 2001]

<sup>3</sup> [Lebarbenchon et al., 2018]

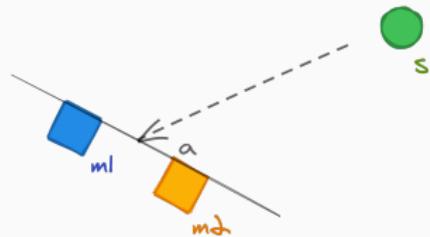
# Sound Source Localization (SSL)

(common knowledge) 

We do not consider here distance estimation.

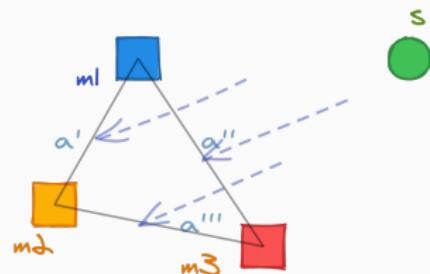
## SSL with 2 microphones

- Only one angle of arrival (AOA) ↗
- Can be approximated from TDOA using e.g. GCC PHAT<sup>1</sup>  
(known limitation, but good in practice)<sup>2</sup>



## SSL with more microphones

- Two Direction of Arrival (DoA): azimuth (↔) and elevation (↑)
- AOA at each pair can be “fused” together (e.g., angular spectra in SRP-PHAT<sup>2</sup>)  
(known limitation, but good in practice)<sup>3</sup>



<sup>1</sup> [Knapp and Carter, 1976]

<sup>2</sup> [DiBiase et al., 2001]

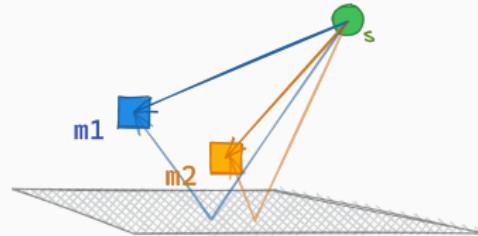
<sup>3</sup> [Lebarbenchon et al., 2018]

# Sound Source Localization with Echoes



## The Picnic Scenario:

- One source
- Two microphones
  - passive scenario
  - generalizable to any array geometry

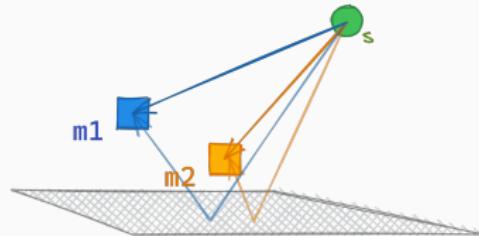


# Sound Source Localization with Echoes



## The Picnic Scenario:

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

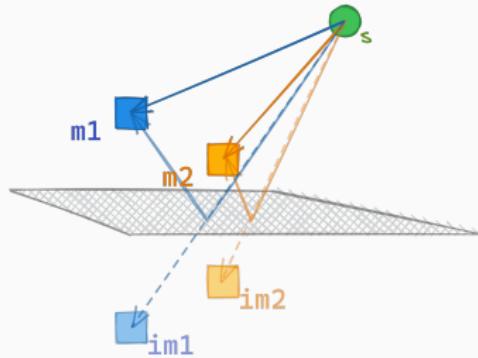


# Sound Source Localization with Echoes



## The Picnic Scenario:

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



Each pair is augmented with echoes

## Mirage Array

(Microphone Array Augmentation with Echoes)

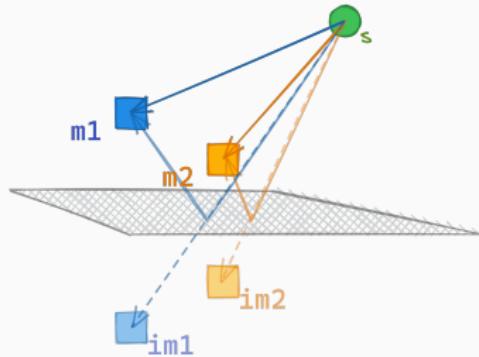
How to access the *image* microphones?



# Sound Source Localization with Echoes

## The Picnic Scenario:

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

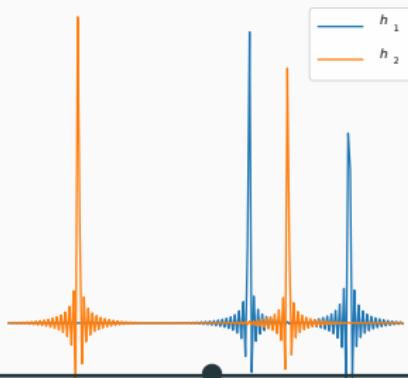


Each pair is augmented with echoes

## Mirage Array

(Microphone Array Augmentation with Echoes)

How to access the *image* microphones?

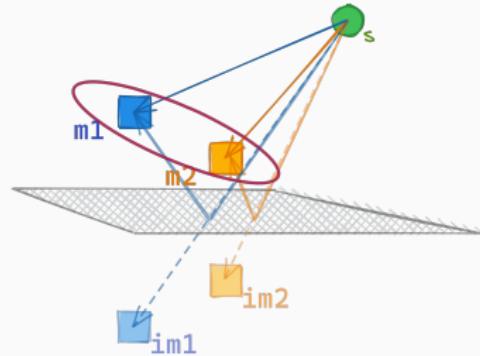




# Sound Source Localization with Echoes

## The Picnic Scenario:

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

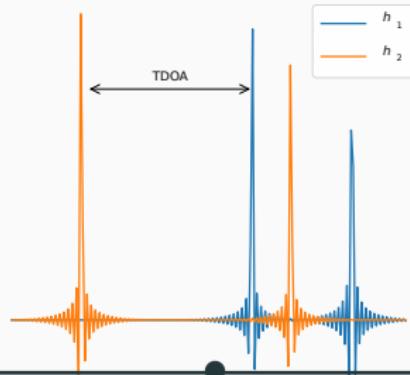


Each pair is augmented with echoes

## Mirage Array

(Microphone Array Augmentation with Echoes)

How to access the *image* microphones?

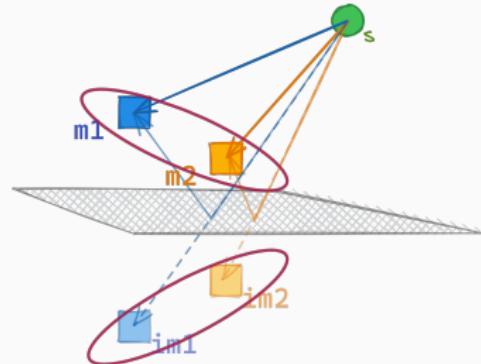




# Sound Source Localization with Echoes

## The Picnic Scenario:

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

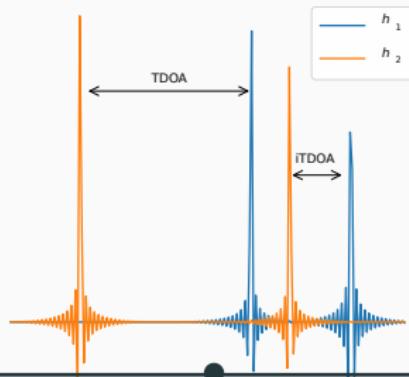


Each pair is augmented with echoes

## Mirage Array

(Microphone Array Augmentation with Echoes)

How to access the *image* microphones?

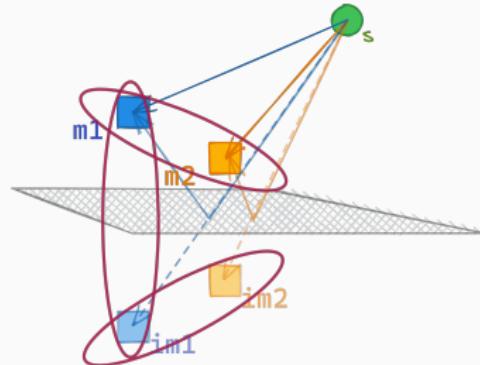




# Sound Source Localization with Echoes

## The Picnic Scenario:

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

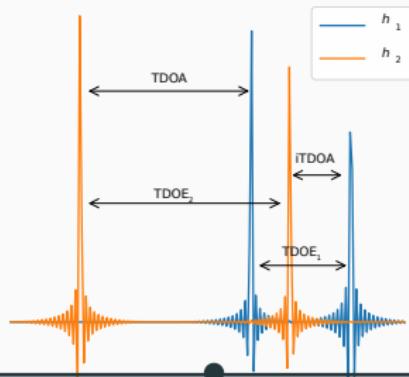


Each pair is augmented with echoes

## Mirage Array

(Microphone Array Augmentation with Echoes)

How to access the *image* microphones?



# Sound Source Localization **with Echoes**



## Proposed Approach:

1. estimate **TDOAs** using proposed learning-based approach (MLP)
2. fuse together the estimation (SRP-PHAT-like algorithm<sup>1</sup>),
  - the error on a validation set provides measure of uncertainty
  - microphone positions are known

<sup>2</sup> [DiBiase et al., 2001]

<sup>1</sup> [Knapp and Carter, 1976]

# Sound Source Localization **with Echoes**



## Proposed Approach:

1. estimate **TDOAs** using proposed learning-based approach (MLP)
2. fuse together the estimation (SRP-PHAT-like algorithm<sup>1</sup>),
  - the error on a validation set provides measure of uncertainty
  - microphone positions are known

**Baseline:** GCC PHAT on true microphones<sup>2</sup>

<sup>2</sup> [DiBiase et al., 2001]

<sup>1</sup> [Knapp and Carter, 1976]



## 实验结果

**Proposed:** MLP with **Mirage**

**Baseline:** GCC PHAT<sup>1</sup>

**Data:** 200 synthetic stereophonic recordings for close-surface scenario

**Metric:** accuracy in % ( $<10^\circ$ ,  $<20^\circ$ ) (↳ also MSE in the □)

AOA ↘	Input	ACCURACY	
		$\alpha < 10^\circ$	$\alpha < 20^\circ$
Mirage	wn	77	97
GCC PHAT	wn	81	97

### Observation

- ✓ Comparable to baseline when white noise source in noiseless case

# 实验结果



**Proposed:** MLP with **Mirage**

**Baseline:** GCC PHAT<sup>1</sup>

**Data:** 200 synthetic stereophonic recordings for close-surface scenario

**Metric:** accuracy in % ( $<10^\circ$ ,  $<20^\circ$ ) (↳ also MSE in the □)

AOA ↘	Input	ACCURACY	
		$\alpha < 10^\circ$	$\alpha < 20^\circ$
Mirage	wn	77	97
Mirage	wn+n	26	54
GCC PHAT	wn	81	97
GCC PHAT	wn+n	65	83

---

## Observation

- ✓ Comparable to baseline when white noise source in noiseless case

# 实验结果



**Proposed:** MLP with **Mirage**

**Baseline:** GCC PHAT<sup>1</sup>

**Data:** 200 synthetic stereophonic recordings for close-surface scenario

**Metric:** accuracy in % ( $<10^\circ$ ,  $<20^\circ$ ) (↳ also MSE in the □)

AOA ↕	Input	ACCURACY	
		$\alpha < 10^\circ$	$\alpha < 20^\circ$
Mirage	wn	77	97
Mirage	wn+n	26	54
GCC PHAT	wn	81	97
GCC PHAT	wn+n	65	83
Mirage	sp	63	82
GCC PHAT	sp	82	97

## Observation

- ✓ Comparable to baseline when white noise source in noiseless case

# 实验结果



**Proposed:** MLP with **Mirage**

**Baseline:** GCC PHAT<sup>1</sup>

**Data:** 200 synthetic stereophonic recordings for close-surface scenario

**Metric:** accuracy in % ( $<10^\circ$ ,  $<20^\circ$ ) (↳ also MSE in the □)

AOA ↘	Input	ACCURACY	
		$\alpha < 10^\circ$	$\alpha < 20^\circ$
Mirage	wn	77	97
Mirage	wn+n	26	54
GCC PHAT	wn	81	97
GCC PHAT	wn+n	65	83
Mirage	sp	63	82
Mirage	sp+n	16	35
GCC PHAT	sp	82	97
GCC PHAT	sp+n	19	32

## Observation

- ✓ Comparable to baseline when white noise source in noiseless case
- ✗ Does not generalize to noisy and speech data

# 实验结果



**Proposed:** MLP with **Mirage**

**Baseline:** GCC PHAT<sup>1</sup>

**Data:** 200 synthetic stereophonic recordings for close-surface scenario

**Metric:** accuracy in % ( $<10^\circ$ ,  $<20^\circ$ ) (↳ also MSE in the □)

AOA ↕	Input	ACCURACY	
		$\alpha < 10^\circ$	$\alpha < 20^\circ$
Mirage	wn	77	97
Mirage	wn+n	26	54
GCC PHAT	wn	81	97
GCC PHAT	wn+n	65	83
Mirage	sp	63	82
Mirage	sp+n	16	35
GCC PHAT	sp	82	97
GCC PHAT	sp+n	19	32

DoA ↗	Input	ACCURACY	
		$\theta \leftrightarrow$	$\phi \leftrightarrow$
Mirage	wn	59	71
Mirage	wn+n	18	26
Mirage	sp	45	59
Mirage	sp+n	17	12

## Observation

- ✓ Comparable to baseline when white noise source in noiseless case
- ✗ Does not generalize to noisy and speech data
- ✓ Take “impossible” localization

# ⚠ Experimental results



**Proposed:** MLP with **Mirage**

**Baseline:** GCC PHAT<sup>1</sup>

**Data:** 200 synthetic stereophonic recordings for close-surface scenario

**Metric:** accuracy in % ( $<10^\circ$ ,  $<20^\circ$ ) (↳ also MSE in the □)

AOA ↕	Input	ACCURACY	
		$\alpha < 10^\circ$	$\alpha < 20^\circ$
Mirage	wn	77	97
Mirage	wn+n	26	54
GCC PHAT	wn	81	97
GCC PHAT	wn+n	65	83
Mirage	sp	63	82
Mirage	sp+n	16	35
GCC PHAT	sp	82	97
GCC PHAT	sp+n	19	32

DoA ↗	Input	ACCURACY	
		$\theta \leftrightarrow$	$\phi \leftrightarrow$
Mirage	wn	59	71
Mirage	wn+n	18	26
Mirage	sp	45	59
Mirage	sp+n	17	12

## Observation

- ✓ Comparable to baseline when white noise source in noiseless case
- ✗ Does not generalize to noisy and speech data
- ✓ Take “impossible” localization
- ⚠ Performance depending on echo estimator (work in progress)

## **Echo-aware Dataset**

---



## Echo-aware datasets

⚠ Everything so far was a simulation

**Echo-aware database requires:**

- annotation of the echoes
- annotation of the geometry
- should cover a vast number of echo-aware applications
- expertise in signal processing, acoustics
- proper recording devices



# dEchorate

## Characteristics of dEchorate

- different room configurations and RT60 ( $\rightarrow$  flipping wall panels)
- $6 \text{ array} \times 5 \text{ mics} \times 4 \text{ sources} \times 11 \text{ wall conf.} = 1320 \text{ annotated RIRs at } 48 \text{ kHz}$
- geometry annotation  $\Leftrightarrow$  echo annotation in the RIRs
- real RIRs  $\Leftrightarrow$  synthetic RIRs
- application to Acoustic Echo Retrieval, Room Geometry Estimation, Speech Enhancement, ...
- silence, chirps, speech, noise, diffuse bubble noise for 64 GB

( prof Gannot, ing. Tandeitnik)





# dEchorate

## Characteristics of dEchorate

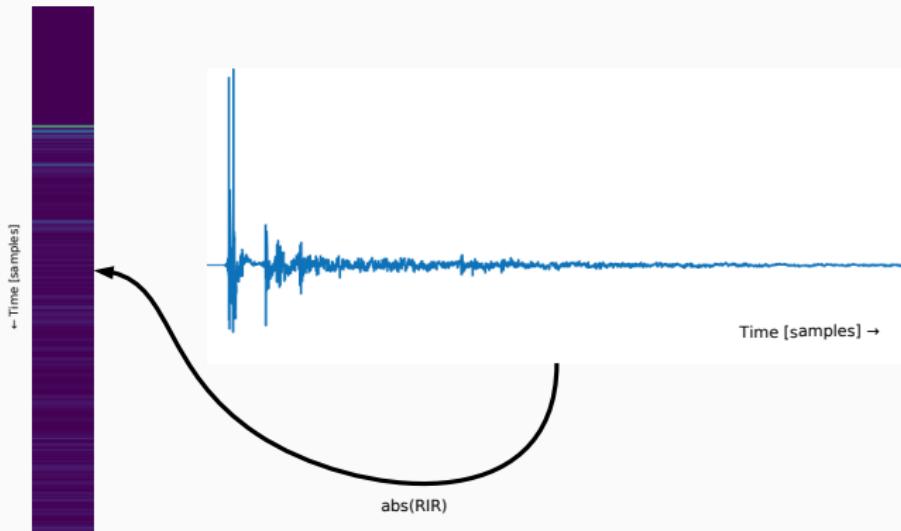
- different room configurations and RT60 ( $\rightarrow$  flipping wall panels)
- 6 array  $\times$  5 mics  $\times$  4 sources  $\times$  11 wall conf. = **1320 annotated RIRs** at 48 kHz
- geometry annotation  $\Leftrightarrow$  echo annotation in the RIRs
- real RIRs  $\Leftrightarrow$  synthetic RIRs
- application to Acoustic Echo Retrieval, Room Geometry Estimation, Speech Enhancement, ...
- silence, chirps, speech, noise, diffuse bubble noise for 64 GB

( prof Gannot, ing. Tandeitnik)





## dEchorate: the skyline view



- each column correspond to the absolute values of one RIR

## dEchorate: the skyline view



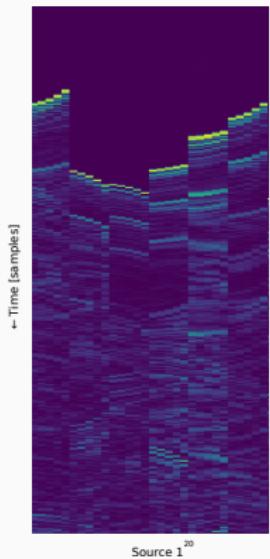
- each column correspond to the absolute values of one RIR

## dEchorate: the skyline view



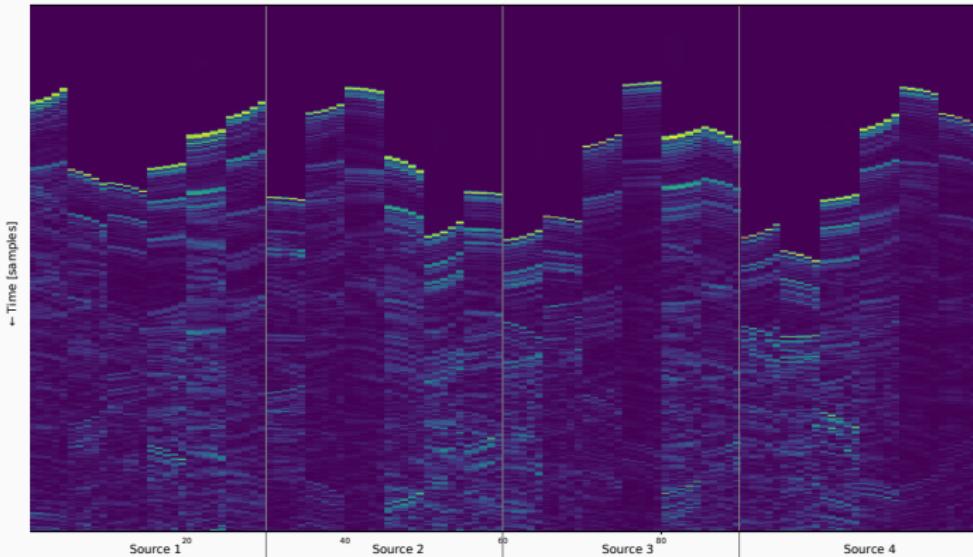
- each column correspond to the absolute values of one RIR
- a block of 5 columns corresponds to one array

## dEchorate: the skyline view



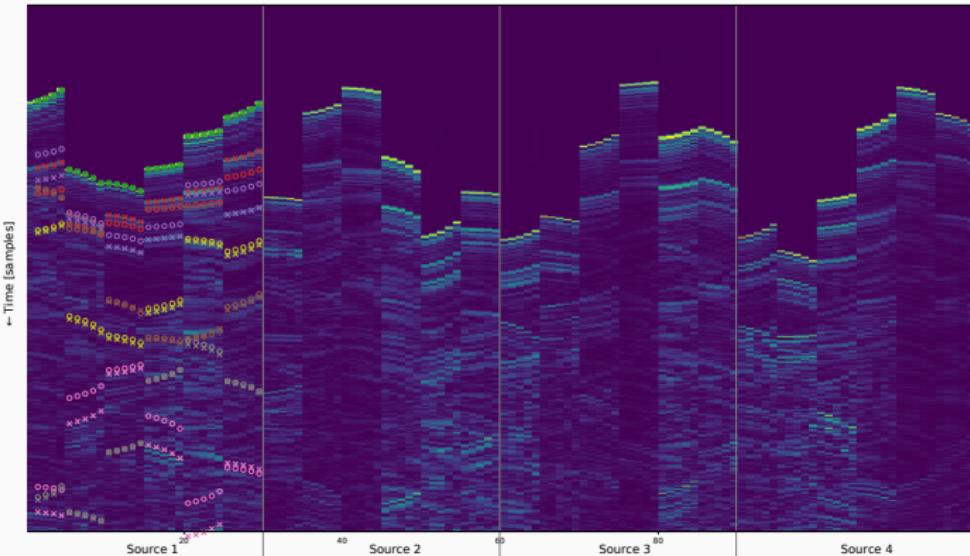
- each column correspond to the absolute values of one RIR
- a block of 5 columns corresponds to one array
- a block of 30 columns corresponds to 6 arrays for 1 sound source

## dEchorate: the skyline view



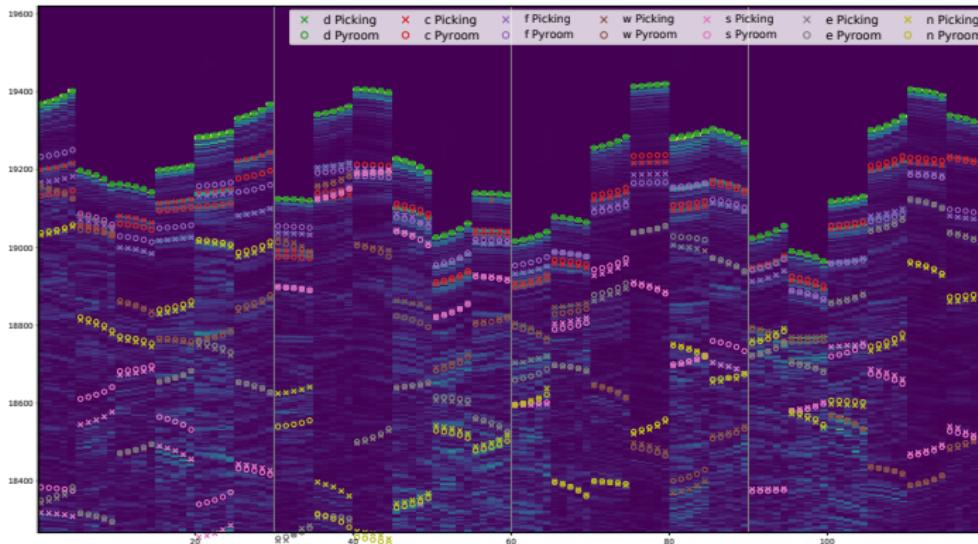
- each column correspond to the absolute values of one RIR
- a block of 5 columns corresponds to one array
- a block of 30 columns corresponds to 6 arrays for 1 sound source
- × corresponds to manual echo location, ◯ to geometric annotation

## dEchorate: the skyline view



- each column correspond to the absolute values of one RIR
- a block of 5 columns corresponds to one array
- a block of 30 columns corresponds to 6 arrays for 1 sound source
- × corresponds to manual echo location, ◊ to geometric annotation

## dEchorate: the skyline view



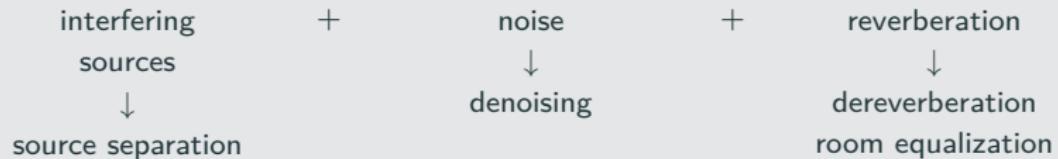
- each column correspond to the absolute values of one RIR
- a block of 5 columns corresponds to one array
- a block of 30 columns corresponds to 6 arrays for 1 sound source
- $\times$  corresponds to manual echo location,  $\circ$  to geometric annotation



# Speech Enhancement with dEchorate

## Speech Enhancement (SE)

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





# Speech Enhancement with dEchorate

## Speech Enhancement (SE)

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





# Speech Enhancement with dEchorate

## Speech Enhancement (SE)

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



SE via **linear spatial filtering** in the STFT domain

$$\mathbf{X}[f, t] = \mathbf{H}[f]\mathbf{S}[f, t] + \mathbf{N}[f, t] \in \mathbb{C}^I \rightarrow \mathbf{W}^H[f] \in \mathbb{C}^I \rightarrow \mathbf{W}^H[f]\mathbf{X}[f, t] \approx \mathbf{S}[f, t]$$

- **target is distortionless** (vs. Multichannel Wiener Filtering)
- many variant, e.g. enhance or null multiple sources [Gannot et al., 2017]



# Speech Enhancement with dEchorate

## Speech Enhancement (SE)

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



SE via **linear spatial filtering** in the STFT domain

$$\mathbf{X}[f, t] = \mathbf{H}[f]\mathbf{S}[f, t] + \mathbf{N}[f, t] \in \mathbb{C}^I \rightarrow \mathbf{W}^H[f] \in \mathbb{C}^I \rightarrow \mathbf{W}^H[f]\mathbf{X}[f, t] \approx \mathbf{S}[f, t]$$

- **target is distortionless** (vs. Multichannel Wiener Filtering)
- many variant, e.g. enhance or null multiple sources [Gannot et al., 2017]

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

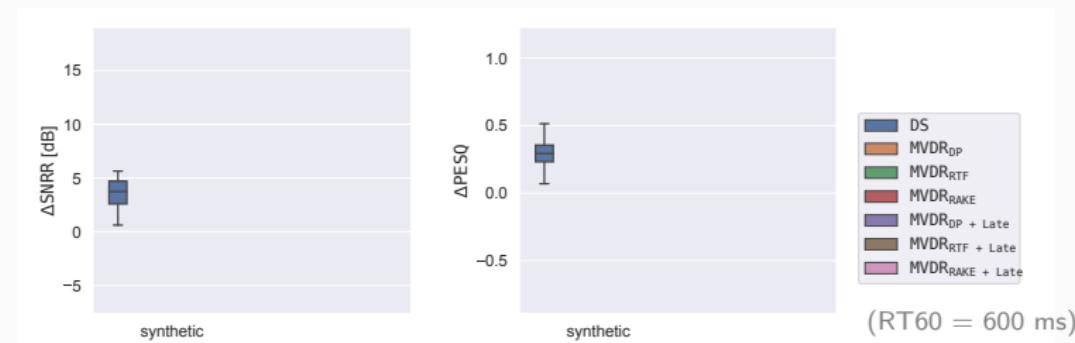
Reducing output energy + distortionless  $\Leftrightarrow$  reduce any uncorrelated noise

# Speech Enhancement with dEchorate



Methods	Noise covariance matrix	RIRs
DS	-	Direct path (AOA)

**Metrics:** Signal to Noise and Reverberant Ratio (SNRR) and Speech Quality (PESQ)

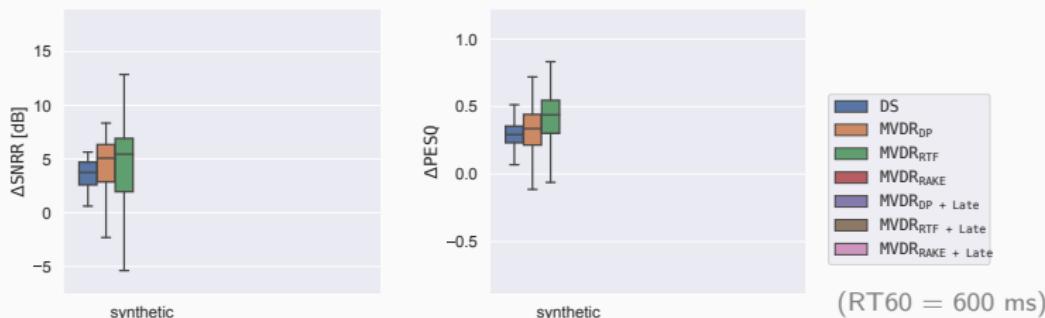


# Speech Enhancement with dEchorate



Methods	Noise covariance matrix	RIRs
DS	-	Direct path (AOA)
MVDR <sub>DP</sub>	Noise	Direct path (AOA)
MVDR <sub>ReTF</sub> <sup>1</sup>	Noise	Relative Transfer Function

**Metrics:** Signal to Noise and Reverberant Ratio (SNRR) and Speech Quality (PESQ)



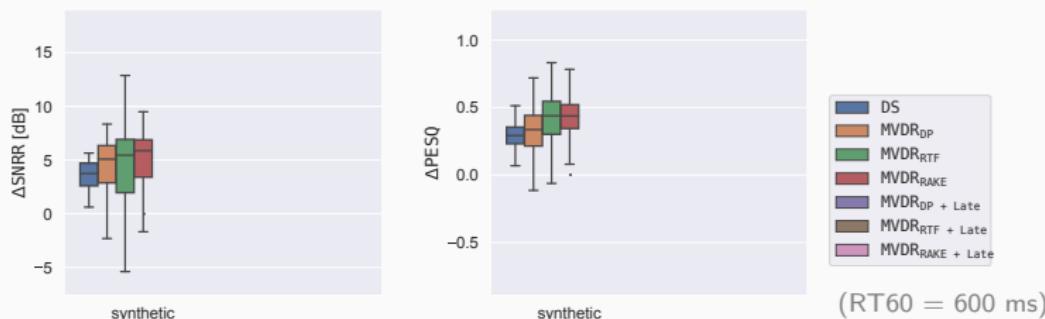
<sup>1</sup>Using [Markovich-Golan et al., 2018],



# Speech Enhancement with dEchorate

Methods	Noise covariance matrix	RIRs
DS	-	Direct path (AOA)
MVDR <sub>DP</sub>	Noise	Direct path (AOA)
MVDR <sub>ReTF</sub> <sup>1</sup>	Noise	Relative Transfer Function
MVDR <sub>Rake</sub> <sup>2</sup>	Noise	4 strongest echoes per channel

Metrics: Signal to Noise and Reverberant Ratio (SNRR) and Speech Quality (PESQ)



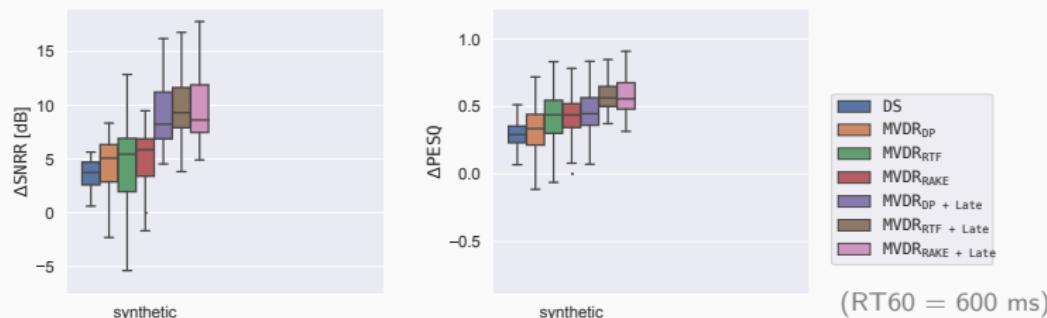
<sup>1</sup>Using [Markovich-Golan et al., 2018], <sup>2</sup>Using [Kowalczyk, 2019],



# Speech Enhancement with dEchorate

Methods	Noise covariance matrix	RIRs
DS	-	Direct path (AOA)
$MVDR_{DP}$	Noise	Direct path (AOA)
$MVDR_{ReTF}^1$	Noise	Relative Transfer Function
$MVDR_{Rake}^2$	Noise	4 strongest echoes per channel
$MVDR_{DP+Late}$	Noise + Late Diffusion <sup>3</sup>	Direct path (AOA)
$MVDR_{ReTF+Late}^1$	Noise + Late Diffusion <sup>3</sup>	Relative Transfer Function
$MVDR_{Rake+Late}^2$	Noise + Late Diffusion <sup>3</sup>	4 strongest echoes per channel

Metrics: Signal to Noise and Reverberant Ratio (SNRR) and Speech Quality (PESQ)



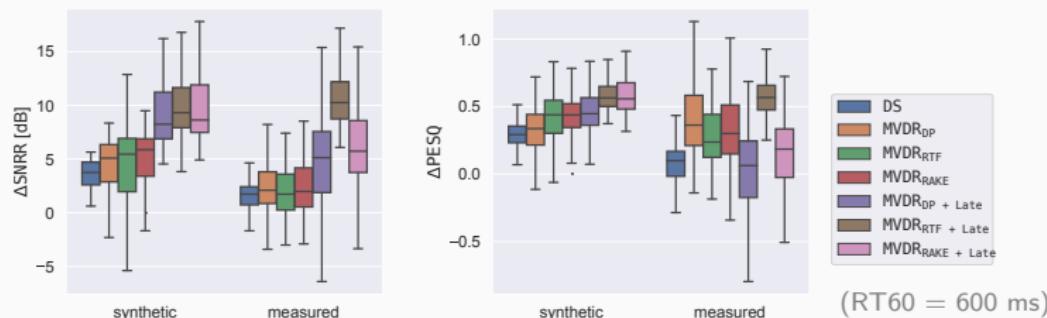
<sup>1</sup>Using [Markovich-Golan et al., 2018], <sup>2</sup>Using [Kowalczyk, 2019], <sup>3</sup>Using [Schwartz et al., 2016],



# Speech Enhancement with dEchorate

Methods	Noise covariance matrix	RIRs
DS	-	Direct path (AOA)
$MVDR_{DP}$	Noise	Direct path (AOA)
$MVDR_{ReTF}^1$	Noise	Relative Transfer Function
$MVDR_{Rake}^2$	Noise	<b>4 strongest echoes per channel</b>
$MVDR_{DP+Late}$	Noise + Late Diffusion <sup>3</sup>	Direct path (AOA)
$MVDR_{ReTF+Late}^1$	Noise + Late Diffusion <sup>3</sup>	Relative Transfer Function
$MVDR_{Rake+Late}^2$	Noise + Late Diffusion <sup>3</sup>	<b>4 strongest echoes per channel</b>

Metrics: Signal to Noise and Reverberant Ratio (SNRR) and Speech Quality (PESQ)



<sup>1</sup>Using [Markovich-Golan et al., 2018], <sup>2</sup>Using [Kowalczyk, 2019], <sup>3</sup>Using [Schwartz et al., 2016],

## Conclusion

---

## Echoes of contributions

---

# Echoes of contributions

## How to estimate them?

In passive stereo scenario:

- Analytical method
  - ✓ direct estimation
  - ✗ depends on source and # echoes
- Learning-based method
  - ✓ estimation of first echo' TDOAs
  - ✗ only on synthetic data and noise source

# Echoes of contributions

## How to estimate them?

In passive stereo scenario:

- Analytical method
  - ✓ direct estimation
  - ✗ depends on source and # echoes
- Learning-based method
  - ✓ estimation of first echo' TDOAs
  - ✗ only on synthetic data and noise source

## How to use them?

- Source Localization
  - ✓ 2D DoA estimation with 2 mic
  - ✗ depends on the echo estimator
- Speech Enhancement
  - ✓ in theory early echoes helps
  - ✗ ... need to be accurately estimated
- Source Separation ↵
- Room Geometry Estimation ↵

# Echoes of contributions

## How to estimate them?

In passive stereo scenario:

- Analytical method
  - ✓ direct estimation
  - ✗ depends on source and # echoes
- Learning-based method
  - ✓ estimation of first echo' TDOAs
  - ✗ only on synthetic data and noise source

## How to use them?

- Source Localization
  - ✓ 2D DoA estimation with 2 mic
  - ✗ depends on the echo estimator
- Speech Enhancement
  - ✓ in theory early echoes helps
  - ✗ ... need to be accurately estimated
- Source Separation ↵
- Room Geometry Estimation ↵

## 3. Where to find them?

- **dEchorate**  
Echo-aware database for both estimation and application
  - ✓ echo annotation ⇔ geometry annotation
  - ✓ synthetic ⇔ real RIRs

## Echo-aware perspective

---

Directions for future work:

## Echo-aware perspective

Directions for future work:

- ▶ **on estimation**
  - develop theoretical guarantees for off-grid acoustic echo retrieval
  - for DNN: extended physics-based learning or other learning paradigm

## Echo-aware perspective

Directions for future work:

- ▶ **on estimation**
  - develop theoretical guarantees for off-grid acoustic echo retrieval
  - for DNN: extended physics-based learning or other learning paradigm
  
- ▶ **on application**
  - other field of echoes: Seismology, Underwater acoustic, Volcanology, etc.

## Echo-aware perspective

Directions for future work:

- ▶ **on estimation**
  - develop theoretical guarantees for off-grid acoustic echo retrieval
  - for DNN: extended physics-based learning or other learning paradigm
- ▶ **on application**
  - other field of echoes: Seismology, Underwater acoustic, Volcanology, etc.
- ▶ **on dEchorate**
  - Benchmark data for echo-aware algorithms
  - Synthetic to Real RIRs (style transfer, new type of acoustic simulator)

## Echo-aware perspective

Directions for future work:

- ▶ **on estimation**
  - develop theoretical guarantees for off-grid acoustic echo retrieval
  - for DNN: extended physics-based learning or other learning paradigm
- ▶ **on application**
  - other field of echoes: Seismology, Underwater acoustic, Volcanology, etc.
- ▶ **on dEchorate**
  - Benchmark data for echo-aware algorithms
  - Synthetic to Real RIRs (style transfer, new type of acoustic simulator)
- ▶ **“close the loop”:** echo estimation  $\Leftrightarrow$  audio analysis
  - in the thesis only  $\Rightarrow$

## List of publications and assets

- On estimation
  - deep learning method in [Di Carlo et al., 2019]
  - **Blaster**: analytical method in [Di Carlo et al., 2020]
- On applications
  - **Mirage**: sound source localization in [Di Carlo et al., 2019]
  - **Separake**: sound source separation in [Scheibler et al., 2018]
- On data
  - **dEchorate**: database (journal in progress)
- Other
  - Signal Processing CUP 2019 [Deleforge et al., 2019]
  - LOCATA Challenge 2019 [Lebarbenchon et al., 2018]
  - Collaboration with Honda Research Group on multichannel **Mirage**

### Code

- **dEchorate**: GUI and code for **dEchorate**
- **Risotto**: ReTF estimation
- **Brioche**: echo-aware Spatial filtering
- **pyMBSSLocate**: MBSSLocate in Python
- **Separake**: Multichannel NMF in Python

## List of publications and assets

- On estimation
  - deep learning method in [Di Carlo et al., 2019]
  - **Blaster**: analytical method in [Di Carlo et al., 2020]
- On applications
  - **Mirage**: sound source localization in [Di Carlo et al., 2019]
  - **Separake**: sound source separation in [Scheibler et al., 2018]
- On data
  - **dEchorate**: database (journal in progress)
- Other
  - Signal Processing CUP 2019 [Deleforge et al., 2019]
  - LOCATA Challenge 2019 [Lebarbenchon et al., 2018]
  - Collaboration with Honda Research Group on multichannel **Mirage**

### Code

- **dEchorate**: GUI and code for **dEchorate**
- **Risotto**: ReTF estimation
- **Brioche**: echo-aware Spatial filtering
- **pyMBSSLocate**: MBSSLocate in Python
- **Separake**: Multichannel NMF in Python

**Thank you!**

## References i

-  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.
-  Azais, J.-M., De Castro, Y., and Gamboa, F. (2015).  
**Spike detection from inaccurate samplings.**  
*Applied and Computational Harmonic Analysis*, 38(2):177–195.
-  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.

## References ii

-  Crocco, M. and Del Bue, A. (2015).  
**Room impulse response estimation by iterative weighted  $\ell_1$ -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  $\ell_1$  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.

## References iii

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

## References iv

-  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.
-  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.

## References v

-  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. (2019).  
**Raking early reflection signals for late reverberation and noise reduction.**  
*The Journal of the Acoustical Society of America*, 145(3):EL257–EL263.
-  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.
-  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.*
-  Leglaise, 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.
-  Markovich-Golan, S., Gannot, S., and Kellermann, W. (2018).  
**Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function.**  
In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2499–2503. IEEE.
-  Nguyen, Q., Girin, L., Bailly, G., Elisei, F., and Nguyen, D.-C. (2018).  
**Autonomous sensorimotor learning for sound source localization by a humanoid robot.**
-  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.

## References viii

-  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.
-  Schwartz, O., Gannot, S., and Habets, E. A. (2016).  
**Joint estimation of late reverberant and speech power spectral densities in noisy environments using frobenius norm.**  
In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1123–1127. IEEE.
-  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.