Echo-aware signal processing for audio scene analysis

Diego Di Carlo

December 1, 2020

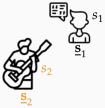
PhD supervisors: Antoine Deleforge
Nancy Bertin

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

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

Introduction

Echo-aware signal processing for audio scene analysis



Sound

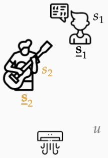
- produced by sources

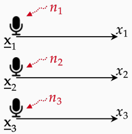
Echo-aware signal processing for audio scene analysis

Sound

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

Echo-aware signal processing for audio scene analysis

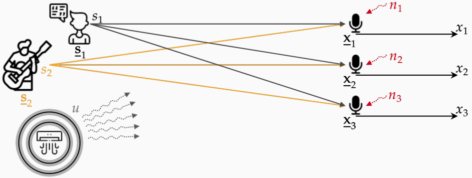




Sound

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

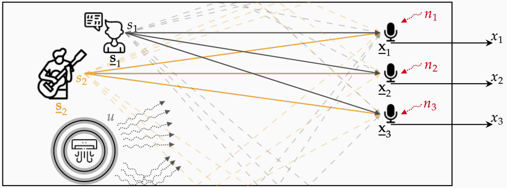
Echo-aware signal processing for audio scene analysis



Sound

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

Echo-aware signal processing for audio scene analysis

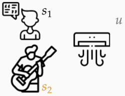


Sound

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

Echo-aware signal processing for audio scene analysis

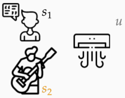
Semantic information



on nature and content

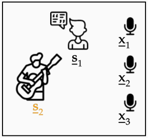
Echo-aware signal processing for audio scene analysis

Semantic information



on nature and content

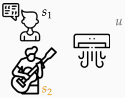
Spatial information



on position and geometry

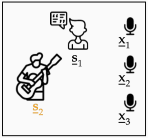
Echo-aware signal processing for audio scene analysis

Semantic information



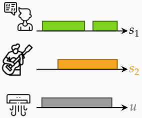
on nature and content

Spatial information



on position and geometry

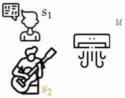
Temporal information

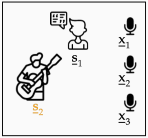


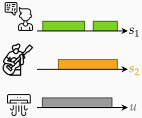
on events activity

Echo-aware signal processing for audio scene analysis

Semantic information | Spatial information | Temporal information



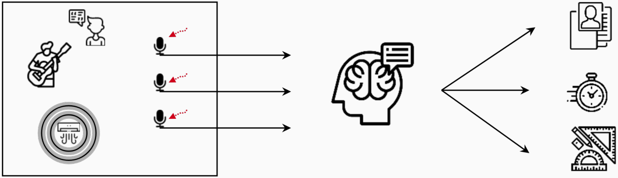




on nature and content | on position and geometry | on events activity

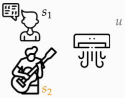
Audio Scene Analysis

Extraction and organization of all the information in the sound



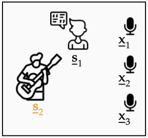
Echo-aware signal processing for audio scene analysis

Semantic information



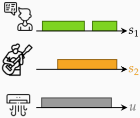
on nature and content

Spatial information



on position and geometry

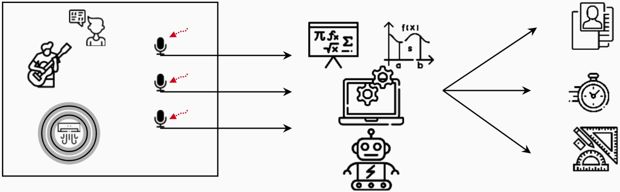
Temporal information



on events activity

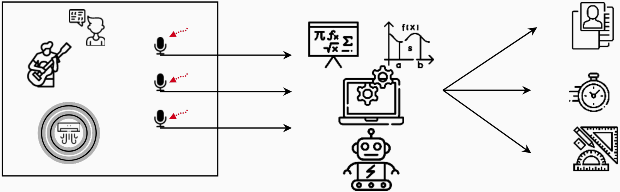
Audio Scene Analysis

Extraction and organization of all the information in the sound

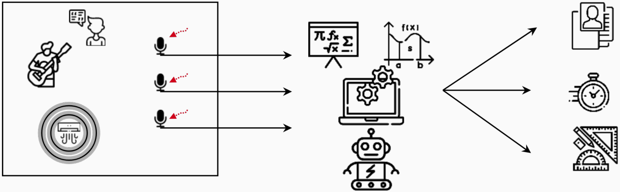


Can computer do it?

Echo-aware signal processing for audio scene analysis



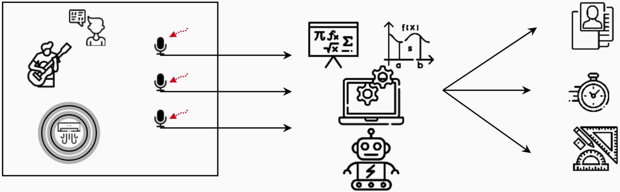
Echo-aware signal processing for audio scene analysis



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

What?

- Sound Source Localization
- Room Geometry Estimation

Where?

- Voice Activity Detection

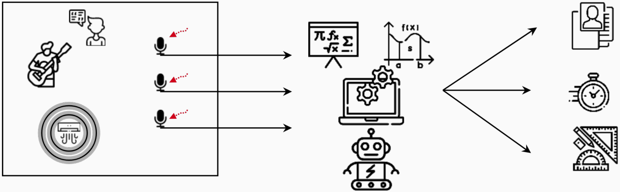
When?

- Reverberation level estimation
- Acoustic Channel Estimation

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

} What?

- Sound Source Localization
- Room Geometry Estimation

} Where?

- Voice Activity Detection

} When?

- Reverberation level estimation
- Acoustic Channel Estimation

} How?

- ...

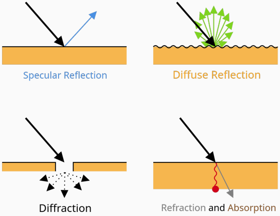
$$\text{HOW} \xrightarrow{\text{helps}} \text{WHERE} \xrightarrow{\text{helps}} \text{WHEN} \xrightarrow{\text{helps}} \text{WHAT} \xrightarrow{\text{helps}} \text{HOW} \xrightarrow{\text{helps}} \dots$$

Echo-aware signal processing for audio scene analysis

Sound interacts with indoor environment:

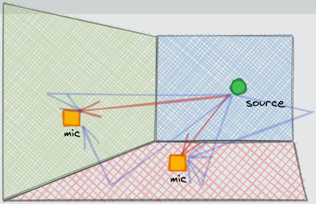
it is reflected
specularly and diffusely
+ it is absorbed,
+ it is transmitted,
+ it is diffracted, ect.

= reverberation



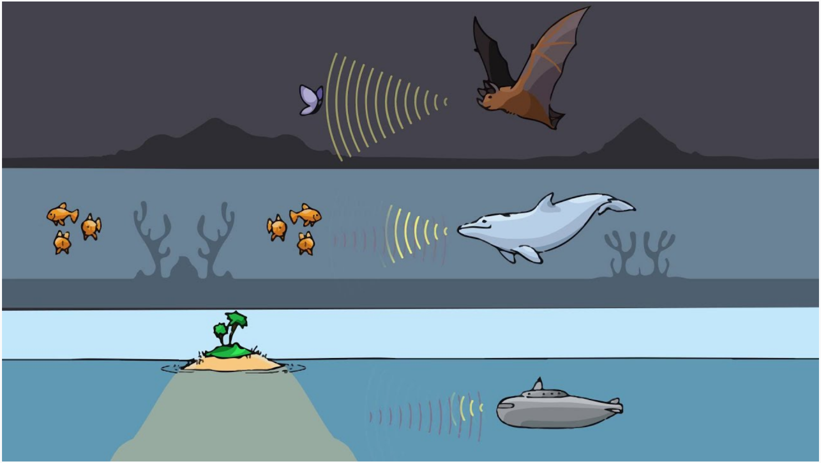
Acoustic Echoes: distict specular reflection

- Specular reflection standing out for time and strength
- Repetition of a sound but after
  - same content
  - delay ⇔ distance



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 $\implies$ simple processing but reverberation = noise
- fully modeled and estimated $\implies$ very challenging

Echo-aware methods

- explicitly account for some acoustic reflection to boost the performances
- attractive alternative between ignoring reverberation and model it entirely

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

[Ribeiro et al., 2010]

Outline and contributions

Thesis title

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

Thesis content:

1. How to estimate them?
   - Learning-based method
   - Analytical method

2. How to use them?
   - Source Localization
   - Source Separation
   - Speech Enhancement
   - Room Geometry Estimation

3. Where to find them?
   - `dEchorate`
     Echo-aware database for both estimation and application

Outline and contributions

Thesis title

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

Thesis content:

1. How to estimate them?
   - Learning-based method **1a**
   - Analytical method **2a**
2. How to use them?
   - Source Localization **2a**
   - Source Separation(not today)
   - Speech Enhancement **3b**
   - Room Geometry Estimation(not today)
3. Where to find them?
   - `dEchorate` **3a**
     Echo-aware database for both estimation and application

Problem Statement

Signal model

For one source and I microphones:

mic. signal

source signal

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

noise term

⚠ continuous-time convolution

Room Impulse Response (RIR)

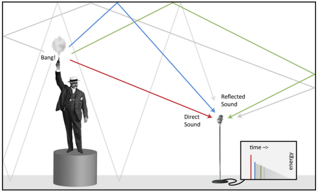
- linear filtering effect of the sound propagation (reverberation)
- acoustic response of a room to a (prefect) impulsive sound
- depends on spatial properties (room geometry, mic/src position)
- $\tilde{h}_i \neq \tilde{h}_j$

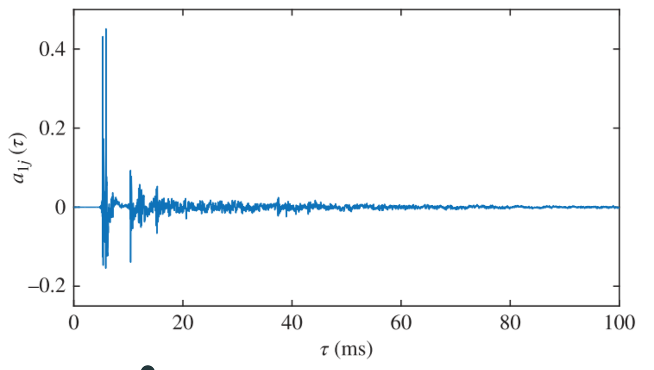
Signal model

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

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

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





Problem Statement

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

models later echoes, reverberation and other.

$$\tilde{h}_i(t) = \tilde{h}_i^{\mathsf{d}}(t) + \tilde{h}_i^{\mathsf{e}}(t) + \varepsilon_i(t) \approx \sum_{r=0}^{R} \alpha_i^{(r)} \delta(t - \tau_i^{(r)}) + \varepsilon_i(t)$$

Goals: Acoustic Echo Retrieval (AER)

Estimated $\left\{\tau_i^{(r)}, \alpha_i^{(r)}{}_i\right\}_{i,r}$ for the microphone signal $\{x_i\}_i$

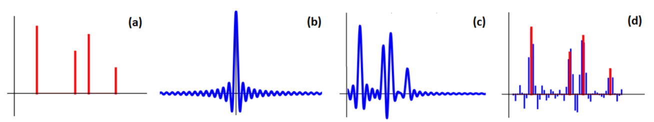
Problem Statement

Goals: Acoustic Echo Retrieval (AER)

Estimated $\left\{\tau_i^{(r)}, \alpha_i^{(r)}{}_i\right\}_{i,r}$ for the microphone signal $\{x_i\}_i$

Challenges:

- RIRs depend on the scene geometry (room, source and mic position)
- Big under-modelling error (late reverberation and noise)
- $\alpha_i^{(r)}$ are distorted:
  - due to air attenuation, wall absorption:
  - $\alpha_i^{(r)} \to \alpha_i^{(r)}(t) \implies$ echo model is sum of filters
  - due to sampling process [Tukuljac et al., 2018]



(Courtesy of Helena Tukuljac [Tukuljac et al., 2018])

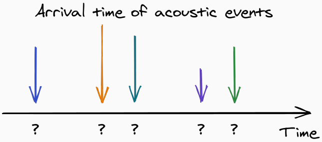
⚠ sampling breaks sparsity and non-negativity

Acoustic Echo Estimation

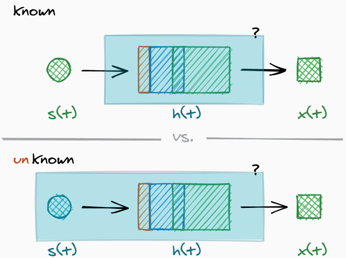
Acoustic Echo Retrieval

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

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



Scenarios: the source signal is

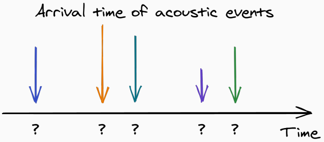


Our case: signal source and passive system of (I microphones)

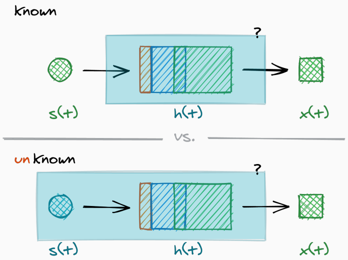
Acoustic Echo Retrieval

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

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



Scenarios: the source signal is

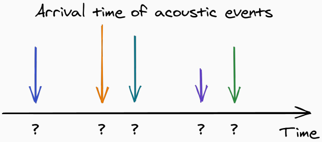


Our case: signal source and passive system of (I microphones)

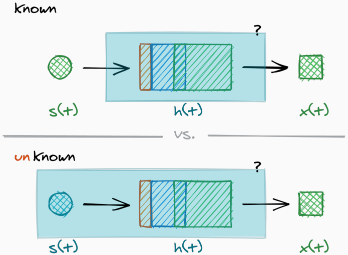
Acoustic Echo Retrieval

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

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



Scenarios: the source signal is

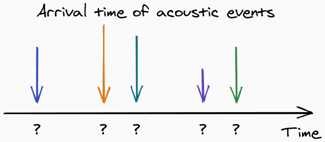


Our case: signal source and passive system of (I microphones)

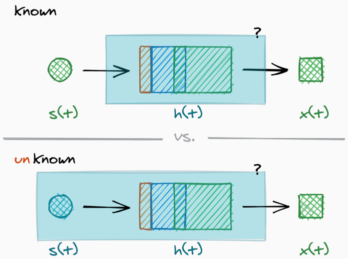
Acoustic Echo Retrieval

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

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



Scenarios: the source signal is



Active

- **non-blind** problem
- **intrusive** or specific setups
  (**Application:** sonar, calibration, measurements, etc.)

Passive

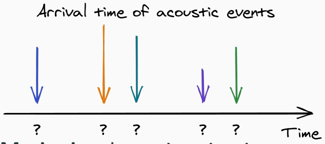
- **blind inverse** problem (harder)
- **passive** and more common setups
  (**Applications:** recording on smart speakers, laptop, etc.)

Our case: signal source and passive system of (I microphones)

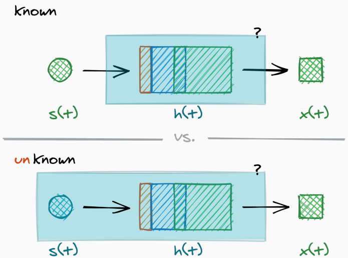
Acoustic Echo Retrieval

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

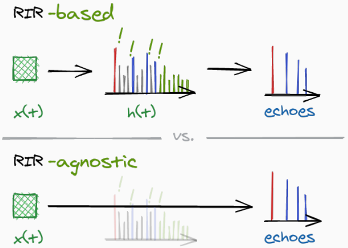
$$\{\tilde{x}_i\}_i \longrightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$



Scenarios: the source signal is



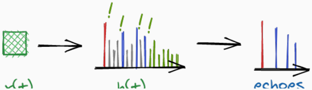
Methods: the estimation is



Our case: signal source and passive system of (I microphones)

Passive Acoustic Echo Retrieval

RIR-based approaches



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

✓ BCE is well and known studied

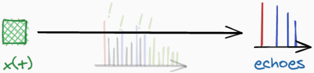
✓ reasonably good for some application
[Crocco and Del Bue, 2016]

✗ Full RIRs need to be estimated

✗ Peak picking has hyperparameters

✗ Issues due to *discrete* estimation

RIR-agnostic approaches



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

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

✓ Sparsity, Non-negativity are respected

✗ exploratory
(no standard solver, few works on audio)

Proposed appoach RIR-agnostic & off-grid:

1. Learning-based approach
2. analytical approach

Proposed approach: learning-based & off-grid

Idea: (Deep) Learning-based AER

1. Use virtually supervised deep learning models
2. Estimate first echo (simple but important)
   ([See Section Application])
3. Only 2 microphones

Motivations:

- This *direct* mapping is difficult, the *inverse* "is not"
  → acoustic simulators: mic/src/room geometry $\longrightarrow$ $\left\{\tau_i^{(r)}, \alpha_i^{(r)}\right\}$, $\tilde{h}_i$, $\tilde{x}_i$
- Acoustic simulator are "simple", versatile and fast
  → many data
- This approach is successful in *Sound Source Localization*
  → position is related to echoes
  [Kataria et al., 2017, Nguyen et al., 2018, Perotin et al., 2019] ⚠ Not only DNN

Proposed approach: models

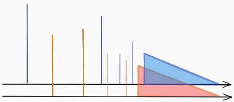
Inputs: Interchannel level and phase difference features[1] from

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

$\approx$ the relative transfer function

$\rightarrow$ remove source dependency

Output: Inter and intra arrival delays



4 TOA
↓
3 Time Difference of Arrivals (**TDOAs**)[1]

HP: first $\Leftrightarrow$ strongest echo

[1] $\mathtt{ILD} = \log|R|$, $\mathtt{IPD} = \arg R/|R|$

Proposed approach: models

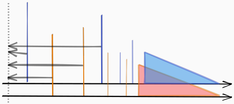
Inputs: Interchannel level and phase difference features[1] from

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

$\approx$ the relative transfer function

$\rightarrow$ remove source dependency

Output: Inter and intra arrival delays



4 TOA

$\downarrow$

3 Time Difference of Arrivals (**TDOAs**)[1]

HP: first $\Leftrightarrow$ strongest echo

[1] $\texttt{ILD} = \log|R|$, $\texttt{IPD} = \arg R/|R|$

Proposed approach: models

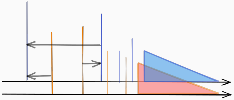
Inputs: Interchannel level and phase difference features[1] from

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

$\approx$ the relative transfer function

$\rightarrow$ remove source dependency

Output: Inter and intra arrival delays



4 TOA

$\downarrow$

3 Time Difference of Arrivals (**TDOAs**)[1]

HP: first $\Leftrightarrow$ strongest echo

[1] $\texttt{ILD} = \log|R|$, $\texttt{IPD} = \arg R/|R|$

Proposed approach: models

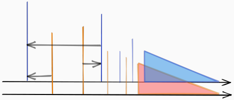
Inputs: Interchannel level and phase difference features[1] from

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

$\approx$ the relative transfer function

$\rightarrow$ remove source dependency

Output: Inter and intra arrival delays



4 TOA

$\downarrow$

3 Time Difference of Arrivals (**TDOAs**)[1]

HP: first $\Leftrightarrow$ strongest echo

- Architecture: `MLP`, `CNN` [Chakrabarty and Habets, 2017, Nguyen et al., 2018]
- Loss Function:
  1. RMSE (Multi-label regression) $\rightarrow$ TDOAs
  2. Gaussian log-likelihood $\rightarrow \{\mu_\tau, \sigma_\tau^2\}\ \forall \tau \in$ TDOAs
  3. Student log-likelihood $\rightarrow \{\mu_\tau, \lambda_\tau, \nu_\tau\}\ \forall \tau \in$ TDOAs

[1] $\texttt{ILD} = \log|R|$, $\texttt{IPD} = \arg R/|R|$

Proposed approach: models

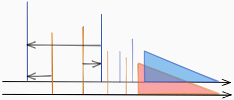
Inputs: Interchannel level and phase difference features[1] from

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

$\approx$ the relative transfer function

$\rightarrow$ remove source dependency

Output: Inter and intra arrival delays



4 TOA
↓
3 Time Difference of Arrivals (**TDOAs**)[1]

HP: first ⇔ strongest echo

- Architecture: `MLP`, `CNN` [Chakrabarty and Habets, 2017, Nguyen et al., 2018]
- Loss Function:
  1. RMSE (Multi-label regression) $\rightarrow$ TDOAs
  2. Gaussian log-likelihood $\rightarrow \{\mu_\tau, \sigma^2_\tau\}\ \forall \tau \in$ TDOAs
  3. Student log-likelihood $\rightarrow \{\mu_\tau, \lambda_\tau, \nu_\tau\}\ \forall \tau \in$ TDOAs

[1] $\texttt{ILD} = \log|R|$, $\texttt{IPD} = \arg R/|R|$

Proposed approach: models

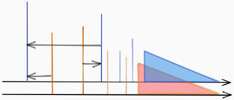
Inputs: Interchannel level and phase difference features[1] from

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

$\approx$ the relative transfer function

$\rightarrow$ remove source dependency

Output: Inter and intra arrival delays



4 TOA

$\downarrow$

3 Time Difference of Arrivals (**TDOAs**)[1]

HP: first $\Leftrightarrow$ strongest echo

- Architecture: `MLP`, `CNN` [Chakrabarty and Habets, 2017, Nguyen et al., 2018]
- Loss Function:
  1. RMSE (Multi-label regression) $\rightarrow$ TDOAs
  2. Gaussian log-likelihood $\rightarrow \{\mu_\tau, \sigma_\tau^2\}\ \forall \tau \in$ TDOAs
  3. Student log-likelihood $\rightarrow \{\mu_\tau, \lambda_\tau, \nu_\tau\}\ \forall \tau \in$ TDOAs

  } Good for data fusion. Similar to **MDN** [Bishop, 1994]
- Data:
  - Virtually-supervised learning (= data from acoustic simulator)
  - white-noise as source signal + AWGN of 0, 10, 20 dB
  - 2 microphone in close-surface scenario

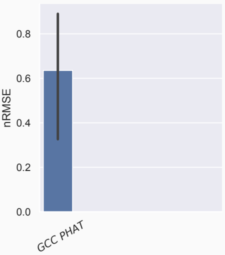
[1] $\text{ILD} = \log|R|$, $\text{IPD} = \arg R/|R|$

Experimental results

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

Baseline: GCC PHAT [Knapp and Carter, 1976]

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

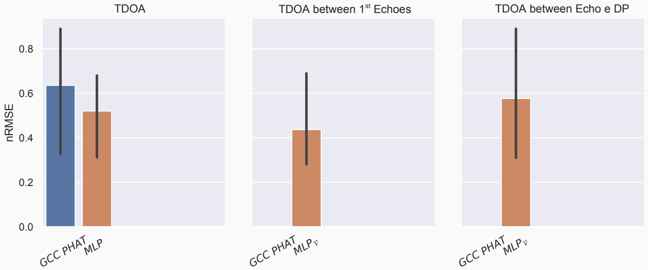


Experimental results

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

Baseline: GCC PHAT [Knapp and Carter, 1976]

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



Observation:

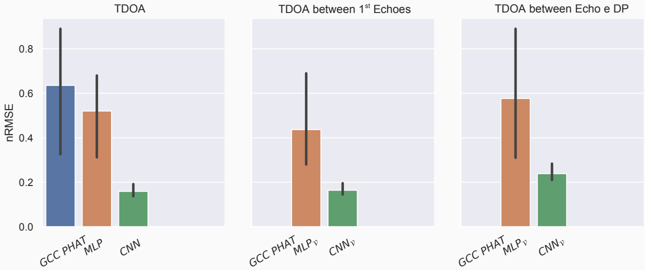
✓ MLP outperforms GCC PHAT on TDOA estimation

Experimental results

Proposed Method: `MLP`, `CNN`, $\mathtt{CNN}_{\mathcal{N}}$, $\mathtt{CNN}_{\mathcal{T}}$

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

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



Observation:

✓ `MLP` outperforms `GCC PHAT` on TDOA estimation

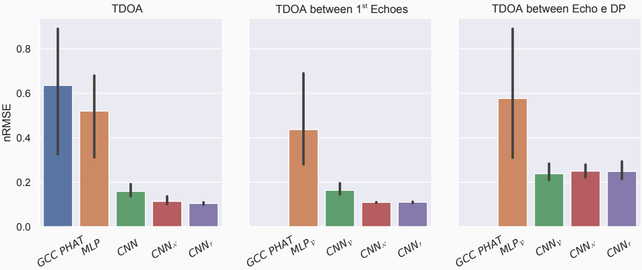
✓ `CNN` outperforms `MLP` (lower error and smaller variance)

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:

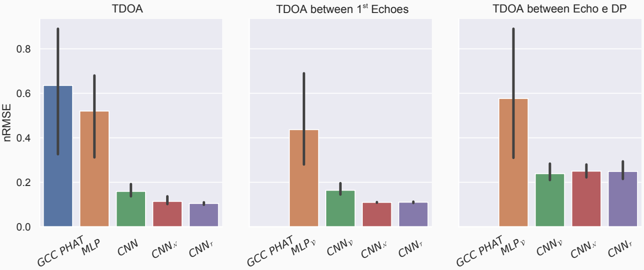
- ✓ MLP outperforms GCC PHAT on TDOA estimation
- ✓ CNN outperforms MLP (lower error and smaller variance)
- ✓ $\text{CNN}_{\mathcal{N}}$ and $\text{CNN}_{\mathcal{T}}$ outperform CNN (lower error and smaller variance)
- ✗ TDOA between DP and 1$^{\text{st}}$ echo more difficult

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)
- ✓ $\text{CNN}_{\mathcal{N}}$ and $\text{CNN}_{\mathcal{T}}$ outperform CNN (lower error and smaller variance)
- ✗ TDOA between DP and 1$^{\text{st}}$ echo more difficult
- ✗ In general, only the first echo on white noise

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

Key ingredient – *Cross relation identity*

$$\begin{cases} \tilde{x}_1 &= \tilde{h}_1 * \tilde{s} \\ \tilde{x}_2 &= \tilde{h}_2 * \tilde{s} \end{cases}$$

$$\tilde{h}_2 * \tilde{x}_1 = \tilde{h}_2 * \tilde{h}_1 * \tilde{s} = \tilde{h}_1 * \tilde{h}_2 * \tilde{s} = \tilde{h}_1 * \tilde{x}_2$$

Ideas:

1. Sampled version of $\tilde{x}_1, \tilde{x}_2$ are available: x_1, x_2
2. Echo TOAs $\propto$ sampling frequency
3. Find echoes $\rightarrow$ **find sparse non-negative vectors** h_1, h_2 of length L
4. Modeled as **Lasso**-like problem

$= \texttt{Toeplitz}(x_i) h_j \in \mathcal{O}(L^2)$

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

$\mathcal{P}(h_1, h_2) \longrightarrow$ sparse promoting regularizer $\qquad \mathcal{C}(h_1, h_2) \longrightarrow$ 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 & off-grid

C. Elvira.

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

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

Proposed approach: analytical & off-grid

C. Elvira.

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

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

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

Proposed approach: analytical & off-grid

C. Elvira.

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

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

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

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

$$\mathbf{X}_i = \mathtt{DFT}(x_i) \simeq \mathscr{F}\tilde{x}_i(nF_s) \qquad n = 0 \dots N-1$$

Proposed approach: analytical & off-grid

C. Elvira.

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

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

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**

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

Idea: Recover echoes by matching a finite number of frequencies

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

$\sim$ **Lasso**, but $\mathcal{F}h_2(f)$ is a continuous function $\rightarrow$ **BLasso** [Bredies and Carioni, 2020]

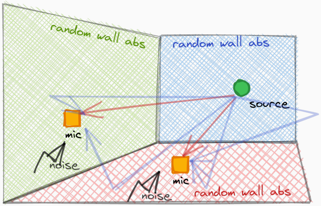
✓No huge matrix ✓ Solutions is a train of Dirac ✓ anchor prevents trivial solution

Experimental results

✓Promising results on noiseless data with RIRs macthing the echo-model

Syntetic Dataset at 16 kHz

- 2 microphones, 1 sound source
- Shoebox with random geometry
- 2 signals: broadband and speech
- $\mathcal{D}^{\text{SNR}}$: $SNR \in [0, 20]$ dB, $\text{RT}_{60} = 400$ ms
- $\mathcal{D}^{\text{RT60}}$: $\text{RT}_{60} = [100, 1000]$ ms, $SNR = 20$ dB



Baseline: discrete RIR-based methods based on LASSO

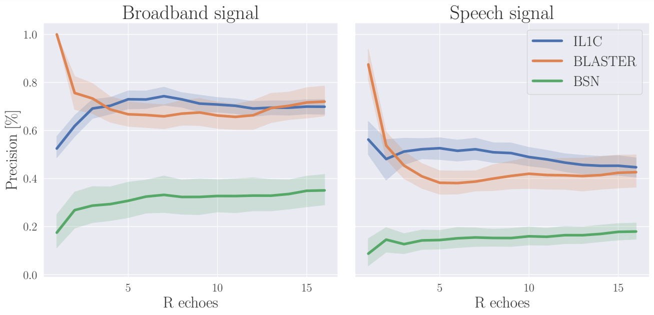
- BSN: Blind, Sparse and Non-negative [Lin et al., 2007]
- IL1C: iteratively-weighted ℓ_1 constraint [Crocco and Del Bue, 2015]
  ↪ State of the Art

hyperparameters and peak-picking tuned via cross-validation

Proposed method: off-grid rir-agnostic based on BLasso
Blind and Sparse Technique for Echo Retrieval (`Blaster`)

Performance per # of echoes

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



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

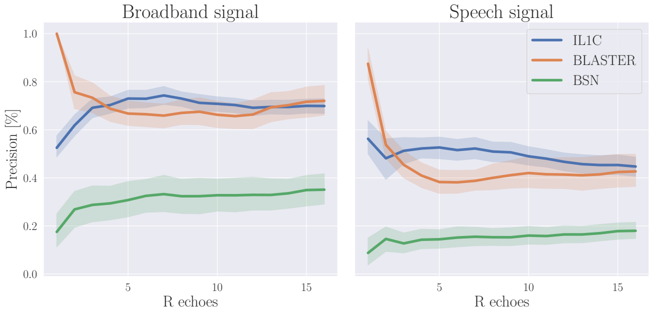
✗ Sensitive to # echoes

✗ Sensitive source signal

✓ Good for 2 echoes

Performance per # of echoes

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



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

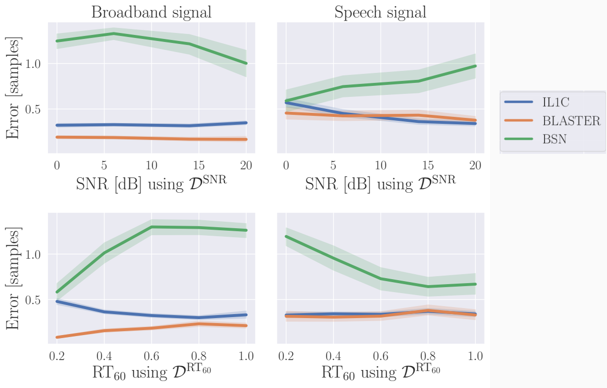
✗ Sensitive to # echoes

✗ Sensitive source signal

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

Error per Dataset/Signal while recovering 7 echoes

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



(Fs = 16 kHz)

✓Lower RMSE

✓ Robustness to SNR and RT_{60}

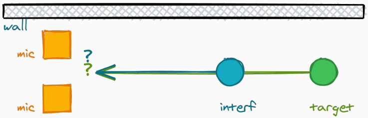
✗ Source signal dependent

Echo-aware Application

Echo-aware Application

Echoes = same content, different time/direction

Image Source Model
⇔
Image Microphone Mode



Some 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 ← image

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

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echoes = same content, different time/direction

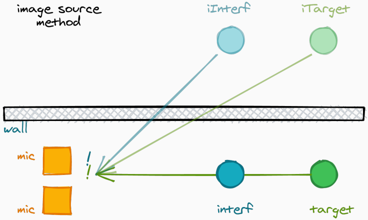


Image Source Model
⇔
Image Microphone Mode

Some 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 ← image

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

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echoes = same content, different time/direction

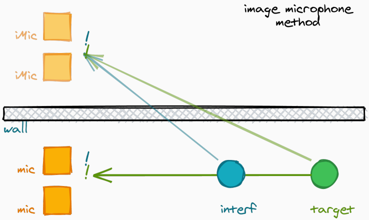


Image Source Model
⇔
Image Microphone Mode

Some 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 ← image

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

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echoes = same content, different time/direction

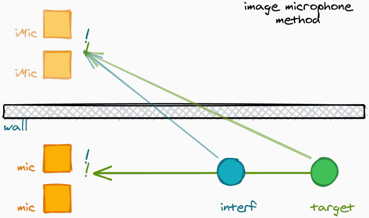


Image Source Model
⇔
Image Microphone Mode

Some 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 ← image

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

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echoes = same content, different time/direction

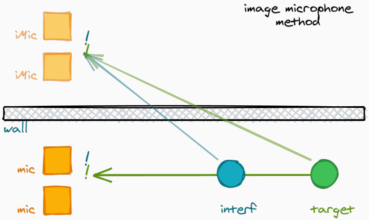


Image Source Model
⇔
Image Microphone Mode

Some 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 ← image

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

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echoes = same content, different time/direction

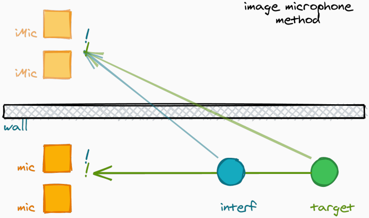


Image Source Model
⇔
Image Microphone Mode

Some 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 ← image

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

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echoes = same content, different time/direction

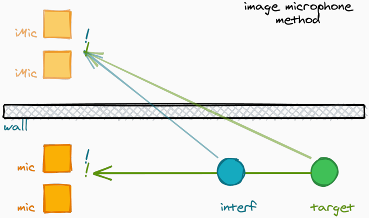


Image Source Model
⇔
Image Microphone Mode

Some 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 ← image

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

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echoes = same content, different time/direction

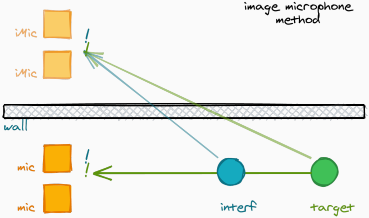


Image Source Model
⇔
Image Microphone Mode

Some 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 ← image

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

How?
Echoes ∈ sound propagation

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

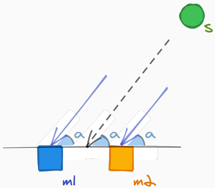
Sound Source Localization (SSL)

(common knowledge)

We do not consider here distance estimation.

SSL with 2 microphones

- Only angle of arrival (AOA)
- can be approximated from TDOA using e.g. GCC PHAT[1]
  (known limitation, but good in practice)



[2] [DiBiase et al., 2001]

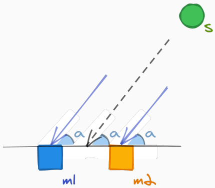
[1] [Knapp and Carter, 1976]

Sound Source Localization (SSL) (common knowledge)

We do not consider here distance estimation.

SSL with 2 microphones

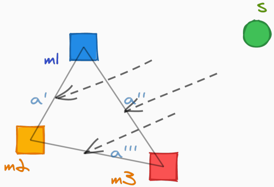
- Only angle of arrival (AOA)
- can be approximated from TDOA using e.g. GCC PHAT[1]
  (known limitation, but good in practice)



SSL with more microphones

- Only Directon of Arrival (DoA): azimuth (↔) and elevation (↕)

1. AOA for each pair can be “fuse” together (e.g. angular spectra in SRP-PHAT[2])
   (known limitation, but good in practice)



---

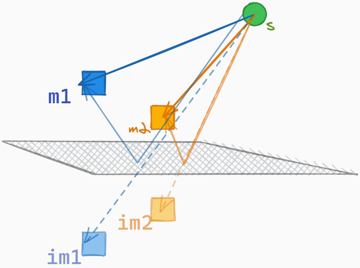
[2] [DiBiase et al., 2001]

[1] [Knapp and Carter, 1976]

Sound Source Localization with Echoes

The Picnic Scenario:

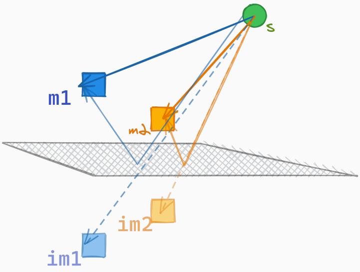
- One source
- Two microphones
  - → passive scenario
  - → generalizable to any array geometry



Sound Source Localization with Echoes

The Picnic Scenario:

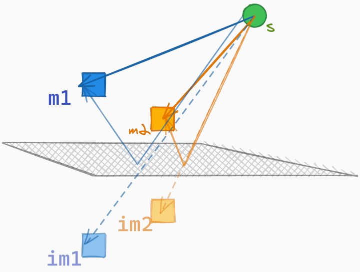
- One source
- Two microphones
  - → passive scenario
  - → generalizable to any array geometry
- Close to a very reflective surface
  - → First echo = Strongest echo
  - → α_{picnic} const. $\forall f$
  - → table-top device



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
  - → α_{picnic} const. $\forall f$
  - → table-top device

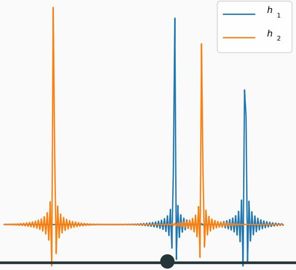


Each pair is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

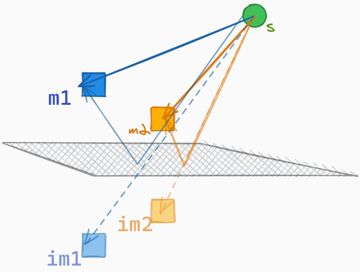
How to access the *image* microphones?



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
  - → α_{picnic} const. $\forall f$
  - → table-top device

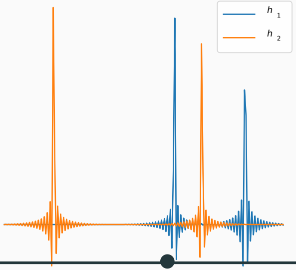


Each pair is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

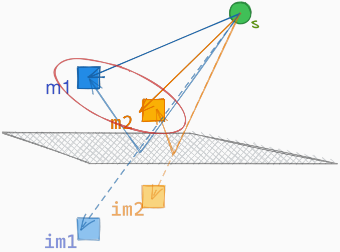
How to access the *image* microphones?



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
  - → α_{picnic} const. $\forall f$
  - → table-top device

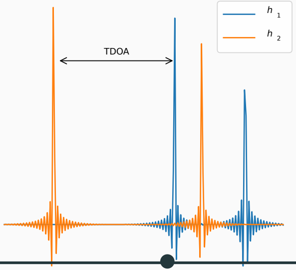


Each pair is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

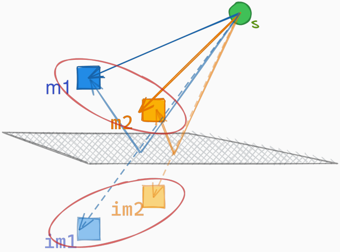
How to access the *image* microphones?



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
  - → α_{picnic} const. $\forall f$
  - → table-top device

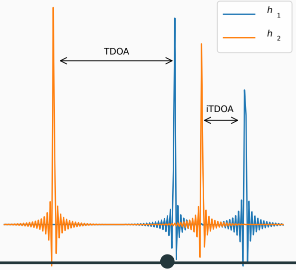


Each pair is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

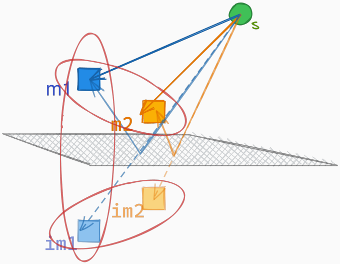
How to access the *image* microphones?



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
  - → α_{picnic} const. $\forall f$
  - → table-top device

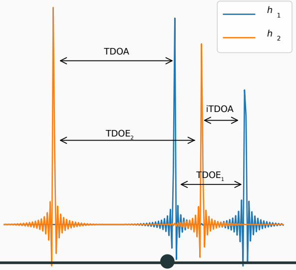


Each pair is augmented with echoes

Mirage Array

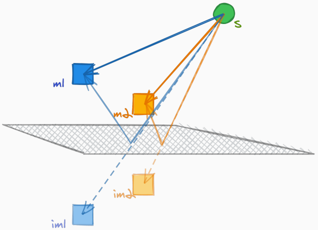
(Microphone Array Augmetation with Echoes)

How to access the *image* microphones?



Sound Source Localization with Echoes

Idea: DoA estimate on the MIRAGE array.
Recall: these TDOAs are the same of the DNN-based method



Proposed Approach:

1. use proposed `MLP` model for TDOAs estimation
2. fuse together the estimation ...
   - of the `Mirage` array (similar to SRP-PHAT[1])
   - knowing the position of the microphones;
   - use the error on a validation set as measure of uncertainty.

Baseline: `GCC PHAT`on true microphones[2]

2 [DiBiase et al., 2001]
1 [Knapp and Carter, 1976]

Experimental results

Proposed: MLP with Mirage

Baseline: GCC PHAT[1]

Data: 200 synthetic stereophonic recordings for close-surface scenario
Metric: accuracy in % (<10°, <20°) (also error in the manuscript)

AOA	Input	ACCURACY $\alpha < 10°$	$\alpha < 20°$
Mirage	wn	77	97
GCC PHAT	wn	**81**	**97**

Observation

✓ comparable to baseline when white noise source in noiseless case

Experimental results

Proposed: MLP with Mirage

Baseline: GCC PHAT[1]

Data: 200 synthetic stereophonic recordings for close-surface scenario
Metric: accuracy in % (<10°, <20°) (also error in the manuscript)

AOA	Input	ACCURACY $\alpha < 10°$	$\alpha < 20°$
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

Experimental results

Proposed: MLP with Mirage

Baseline: GCC PHAT[1]

Data: 200 synthetic stereophonic recordings for close-surface scenario

Metric: accuracy in % (<10°, <20°) (also error in the manuscript)

AOA	Input	ACCURACY $\alpha < 10°$	$\alpha < 20°$
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

Experimental results

Proposed: MLP with Mirage

Baseline: GCC PHAT[1]

Data: 200 synthetic stereophonic recordings for close-surface scenario

Metric: accuracy in % (<10°, <20°) (also error in the manuscript)

AOA	Input	ACCURACY $\alpha < 10°$	$\alpha < 20°$
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

✗ not generalize to noisy and speech data

Experimental results

Proposed: MLP with Mirage

Baseline: GCC PHAT[1]

Data: 200 synthetic stereophonic recordings for close-surface scenario

Metric: accuracy in % (<10°, <20°) (also error in the manuscript)

AOA	Input	ACCURACY $\alpha < 10°$	ACCURACY $\alpha < 20°$
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	< 10° θ	< 10° ϕ	< 20° θ	< 20° ϕ
Mirage	wn	**59**	**71**	**79**	**88**
Mirage	wn+n	18	26	35	66
Mirage	sp	45	59	71	83
Mirage	sp+n	17	12	38	43

Observation

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

Experimental results

Proposed: MLP with Mirage

Baseline: GCC PHAT[1]

Data: 200 synthetic stereophonic recordings for close-surface scenario
Metric: accuracy in % ($<10°$, $<20°$) (also error in the manuscript)

AOA	Input	ACCURACY $\alpha < 10°$	ACCURACY $\alpha < 20°$
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 $< 10°$ θ	ACCURACY $< 10°$ ϕ	ACCURACY $< 20°$ θ	ACCURACY $< 20°$ ϕ
Mirage	wn	**59**	**71**	**79**	**88**
Mirage	wn+n	18	26	35	66
Mirage	sp	45	59	71	83
Mirage	sp+n	17	12	38	43

Observation

- ✓ comparable to baseline when white noise source in noiseless case
- ✗ not generalize to noisy and speech data
- ✓ Solved "impossible" localization
- ⚠ Performance depending on echo estimation methods

27

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 and
- proper recording devices.

dEchorate

Characteristics of dEchorate

- different room configurations and RT60 (→ flipping wall panels)
- 6 array × 5 mics × 4 sources × 11 wall conf. = **1320 annotated RIRs** at 48 kHz
- geometry annotation ⇔ echo annotation in the RIRs
- real RIRs ⇔ 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)

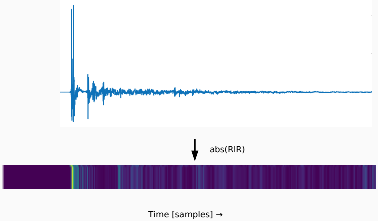
dEchorate

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

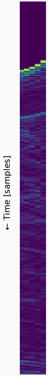


- each column correspond to the absolute values of one RIR

dEchorate: the skyline view

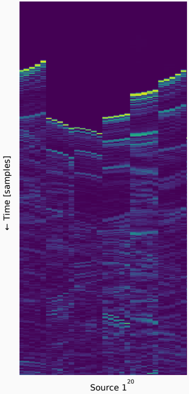
- each column correspond to the absolute values of one RIR
- every 5 columns corresponds to one array

dEchorate: the skyline view



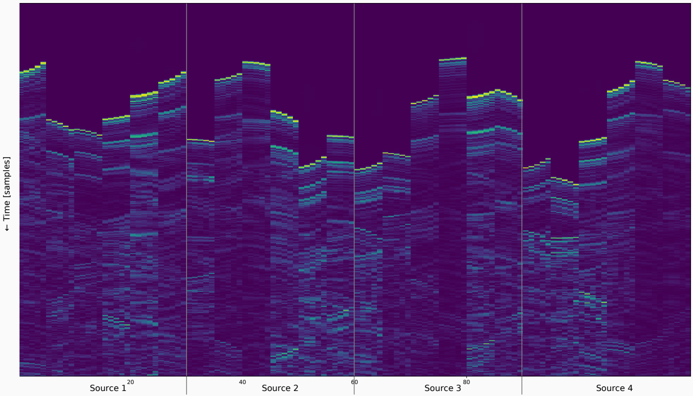
- each column correspond to the absolute values of one RIR
- every 5 columns corresponds to one array
- every 30 column corresponds to one sound source

dEchorate: the skyline view



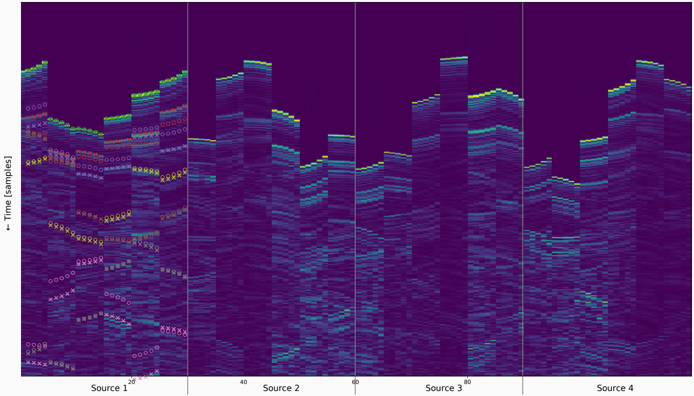
- each column correspond to the absolute values of one RIR
- every 5 columns corresponds to one array
- every 30 column corresponds to one sound source

dEchorate: the skyline view



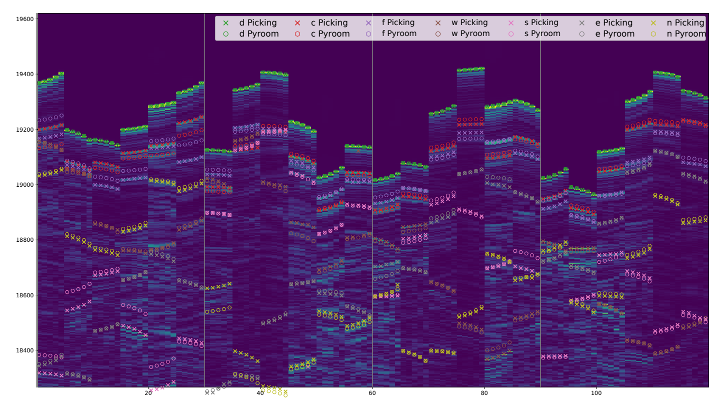
- each column correspond to the absolute values of one RIR
- every 5 columns corresponds to one array
- every 30 column corresponds to one sound source

dEchorate: the skyline view



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

dEchorate: the skyline view



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

Echo-aware Speech Enhancement with dEchorate

Speech Enhancement (SE)

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

interfering sources	+	noise	+	reverberation
$\downarrow$		$\downarrow$		$\downarrow$
source separation		denoising		dereverberation room equalization

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 room equalization

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

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

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 room equalization

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

$$\mathbf{x}[f,t] = \mathbf{h}[f]\mathbf{s}[f,t] + \mathbf{n}[f,t] \in \mathbb{C}^I \longrightarrow \mathbf{w}^{\mathsf{H}}[f] \in \mathbb{C}^I \longrightarrow \mathbf{w}^{\mathsf{H}}[f]\mathbf{x}[f,t] \approx \mathbf{s}[f,t]$$

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

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 room equalization

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

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

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

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

Reducing output energy + distortionless ⇔ reduce any uncorrelated noise

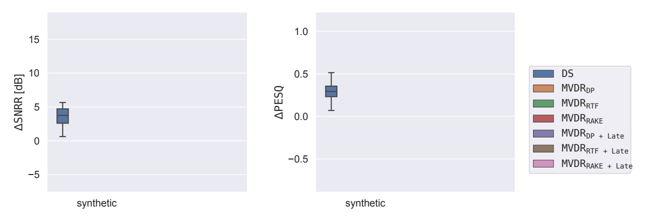
Echo-aware Speech Enhancement

Closed-form solution, but it requires:

	Noise covariance matrix	RIRs
DS	-	Direct Path (AOA)

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

Data: dEchorate dataset, RT60 = 600 ms)



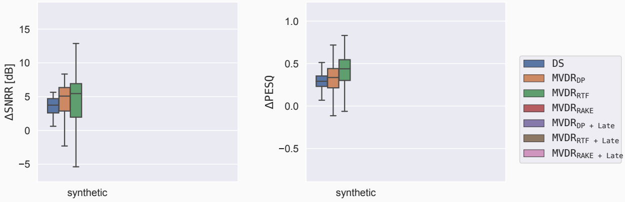
Echo-aware Speech Enhancement

Closed-form solution, but it requires:

	Noise covariance matrix	RIRs
DS	-	Direct Path (AOA)
$\mathrm{MVDR}_{\mathrm{DP}}$	Noise	Direct Path (AOA)
$\mathrm{MVDR}_{\mathrm{ReTF}}$	Noise	Relative Transfer Function

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

Data: dEchorate dataset, RT60 $= 600$ ms)



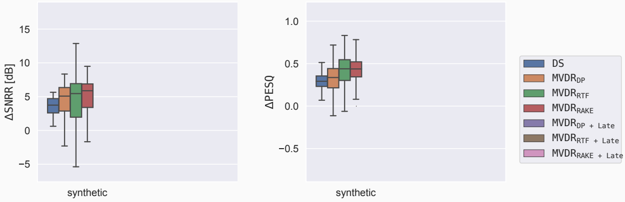
Echo-aware Speech Enhancement

Closed-form solution, but it requires:

	Noise covariance matrix	RIRs
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	4 Early Echoes

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

Data: `dEchorate` dataset, RT60 $= 600$ ms)



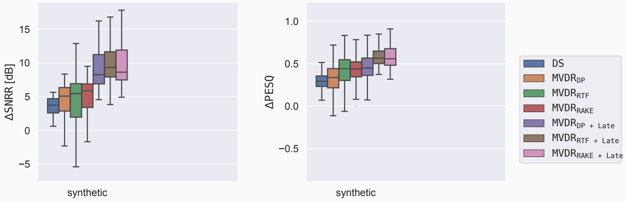
Echo-aware Speech Enhancement

Closed-form solution, but it requires:

	Noise covariance matrix	RIRs
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	4 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	4 Early Echoes

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

Data: dEchorate dataset, RT60 = 600 ms)



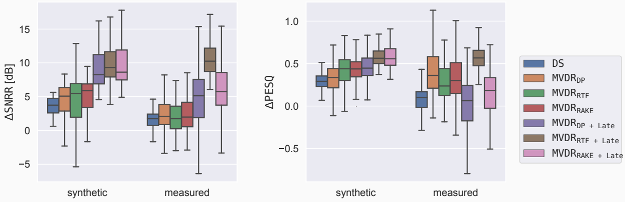
Echo-aware Speech Enhancement

Closed-form solution, but it requires:

	Noise covariance matrix	RIRs
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	4 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	4 Early Echoes

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

Data: dEchorate dataset, RT60 = 600 ms)



Conclusion

Summary of contributions

How to estimate them?

In passive stereo scenario:

- Learning-based method
  - off grid estimation
  - depends on source and # echoes
- Analytical method
  - estimation on first echo' TDOAs
  - only on synthetic data

How to use them?

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

Where to find them?

- `dEchorate`
  Echo-aware database for both estimation and application
  - echo annotation $\Leftrightarrow$ geometry annotation
  - synthetic $\Leftrightarrow$ real RIRs

Echo-aware perspective

Directions for future work:

- on **estimation**
  - develop theoretical guaranties for off-grid acoustic echo retrieval
  - for DNN: extended physic-based learning or other learning paradigm (i.e., unfolding or curriculum learning)
- on **application**
  - other field of echoes: (Seismology, Underwater acoustic, Volcanology, Sniper Detection, etc.)
- on `dEchorate`
  - Synthetic to Real RIRs (style transfer, new types to acoustic simulators)
  - Benchmark for echo-aware algorithms
- **"close the loop":** audio analysis $\Leftrightarrow$ echo estimation
  in the thesis only the $\Rightarrow$ direction.

List of publications and artifacts

Publications

- Estimation
  - deep learning method in [Di Carlo et al., 2019]
  - `Blaster`— analytical method in [Di Carlo et al., 2020]
- Application
  - `Mirage`— sound source localization in [Di Carlo et al., 2019]
  - `Separake`— sound source separation in [Scheibler et al., 2018]
- 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 on multichannel `Mirage`

Code

- `dEchorate`: GUI and code for `dEchorate`
- `Risotto`: library for ReTF estimation
- `Brioche`: library for 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.

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.

References ii

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.

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.

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

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.

Kataria, S., Gaultier, C., and Deleforge, A. (2017).
Hearing in a shoe-box: binaural source position and wall absorption estimation using virtually supervised learning.
In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 226–230. 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.

Perotin, L., Défossez, A., Vincent, E., Serizel, R., and Guérin, A. (2019).
Regression versus classification for neural network based audio source localization.
In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 343–347. IEEE.

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.

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., Deleforge, A., and Gribonval, R. (2018).
Mulan: a blind and off-grid method for multichannel echo retrieval.
In *Advances in Neural Information Processing Systems*, pages 2182–2192.