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 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)

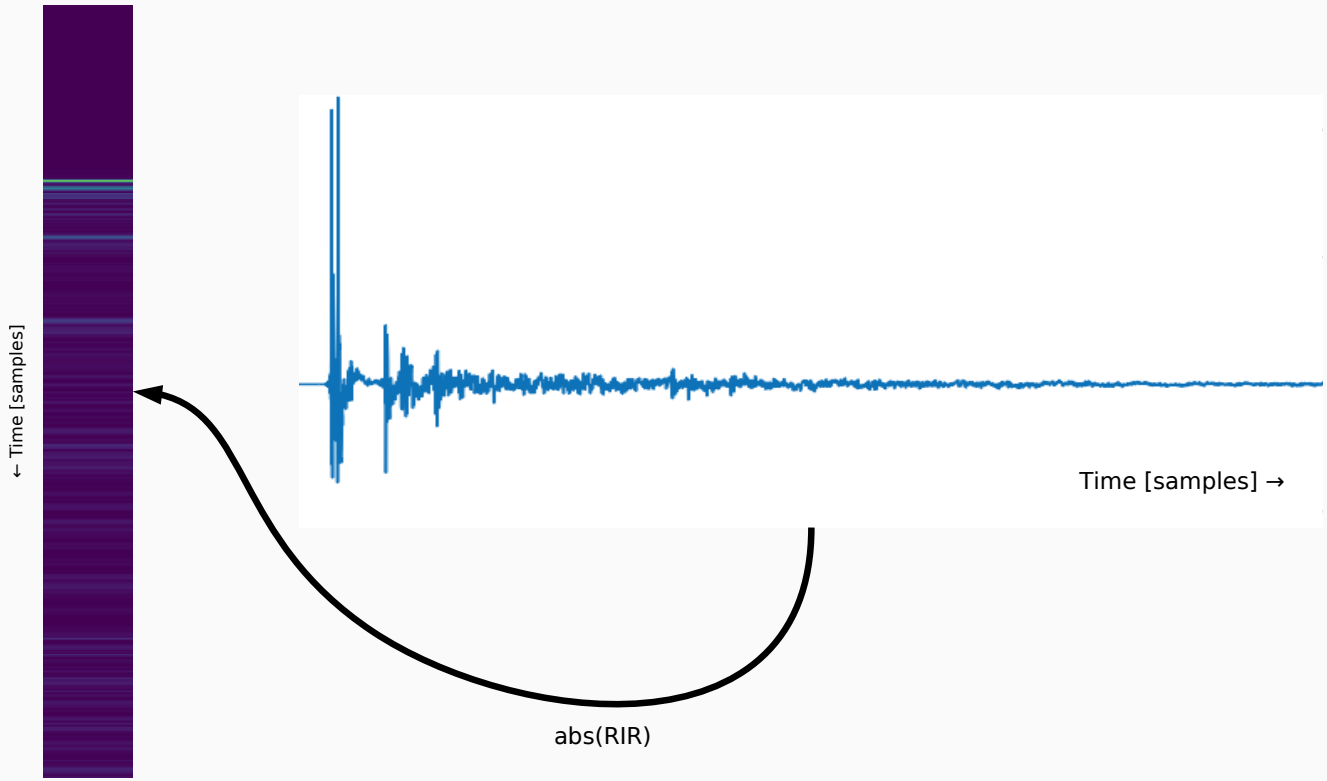
dEchorate

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

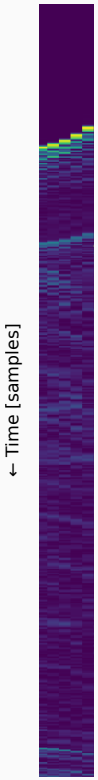


- each column correspond to the absolute values of one RIR

dEchorate: the skyline view

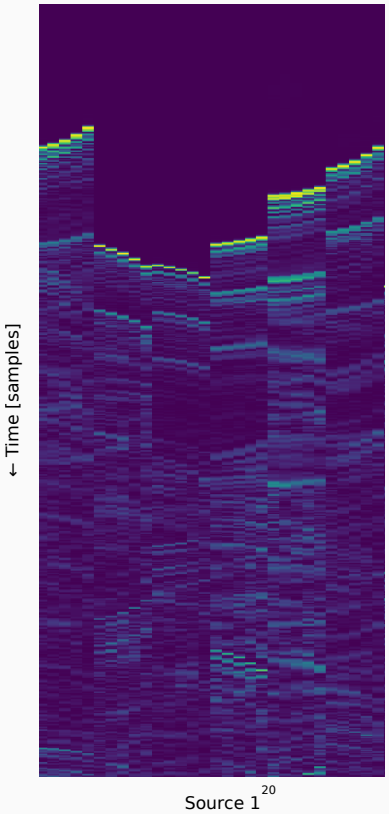
- each column correspond to the absolute values of one RIR

dEchorate: the skyline view



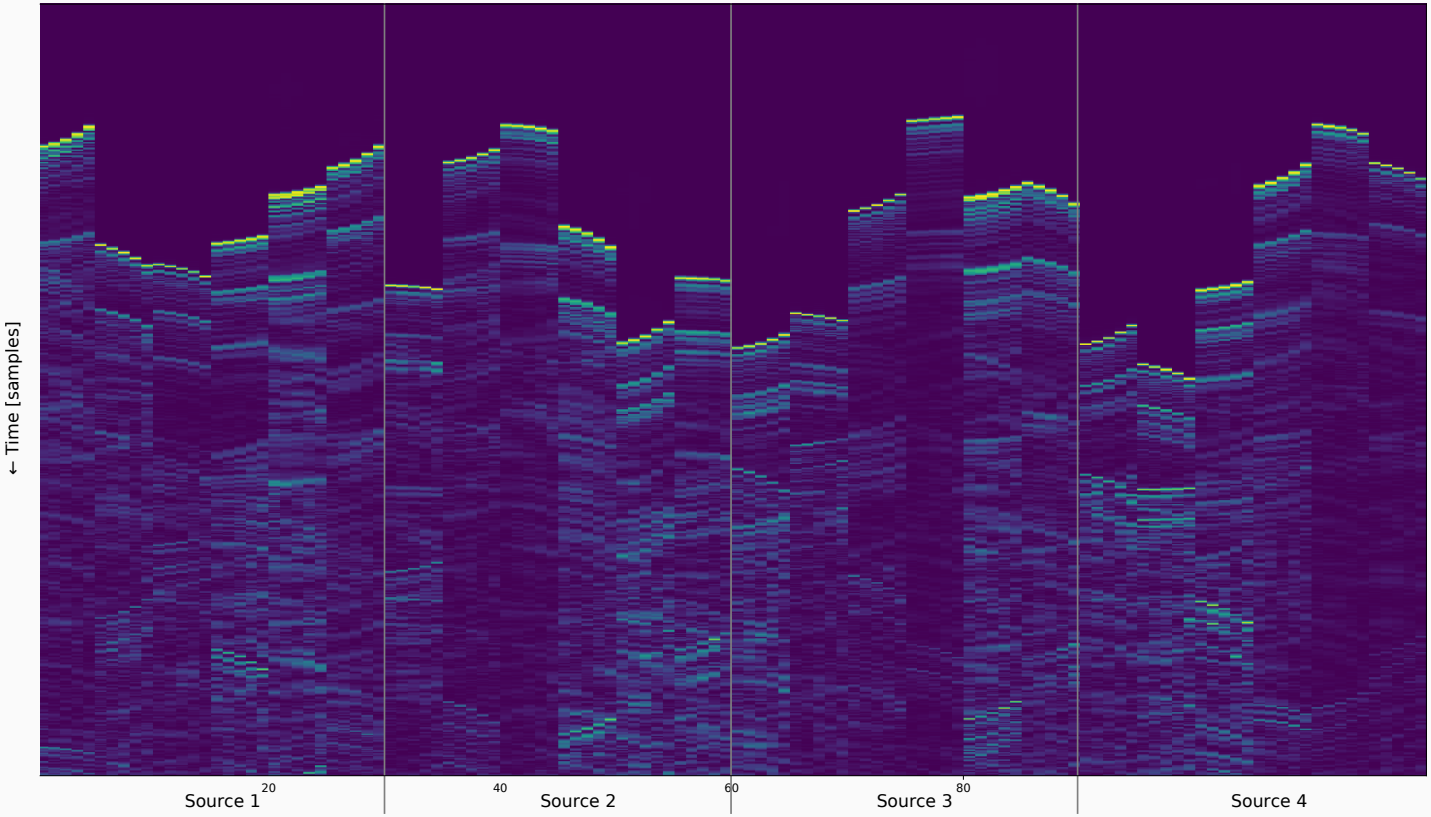
- each column correspond to the absolute values of one RIR
- a block of 5 columns corresponds to one array

dEchorate: the skyline view



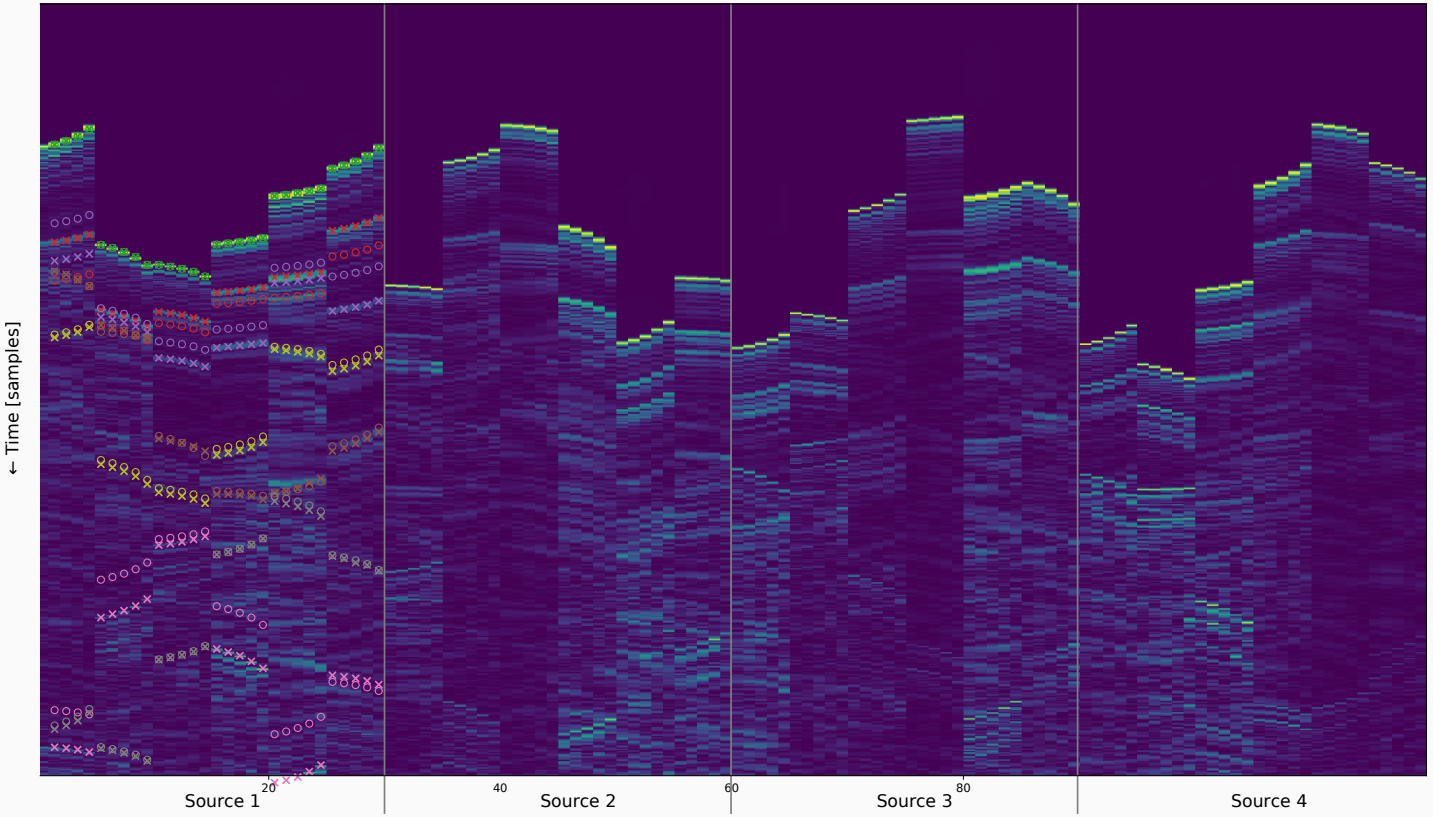
- each column correspond to the absolute values of one RIR
- a block of 5 columns corresponds to one array
- a block of 30 columns corresponds to 6 array for 1 sound source

dEchorate: the skyline view



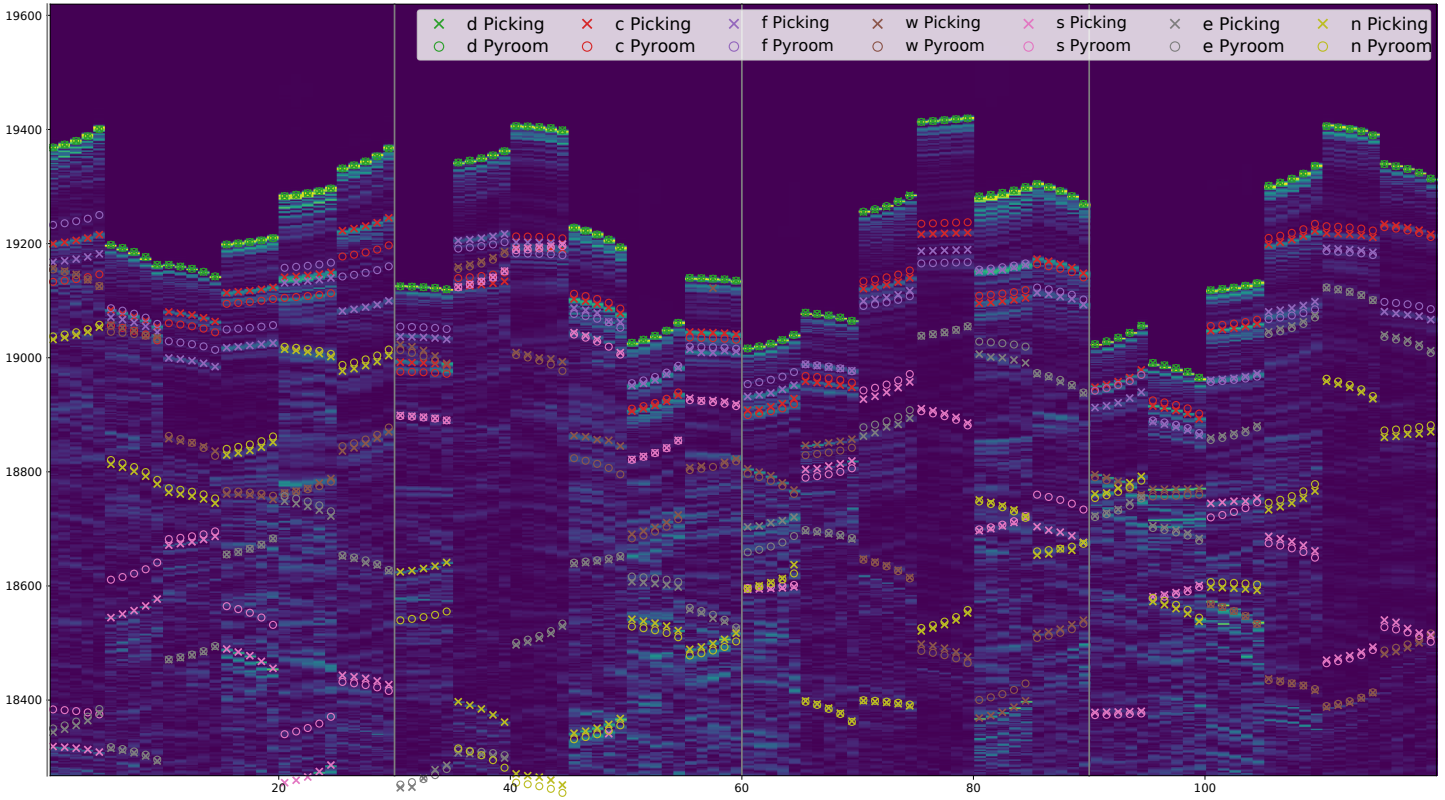
- each column correspond to the absolute values of one RIR
- a block of 5 columns corresponds to one array
- a block of 30 columns corresponds to 6 array for 1 sound source
- $\times$ corresponds to manual echo location, $\circ$ to geometric annotation

dEchorate: the skyline view



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

interfering sources	+	noise	+	reverberation
↓		↓		↓
source separation		denoising		dereverberation room equalization

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

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

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

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

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

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

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

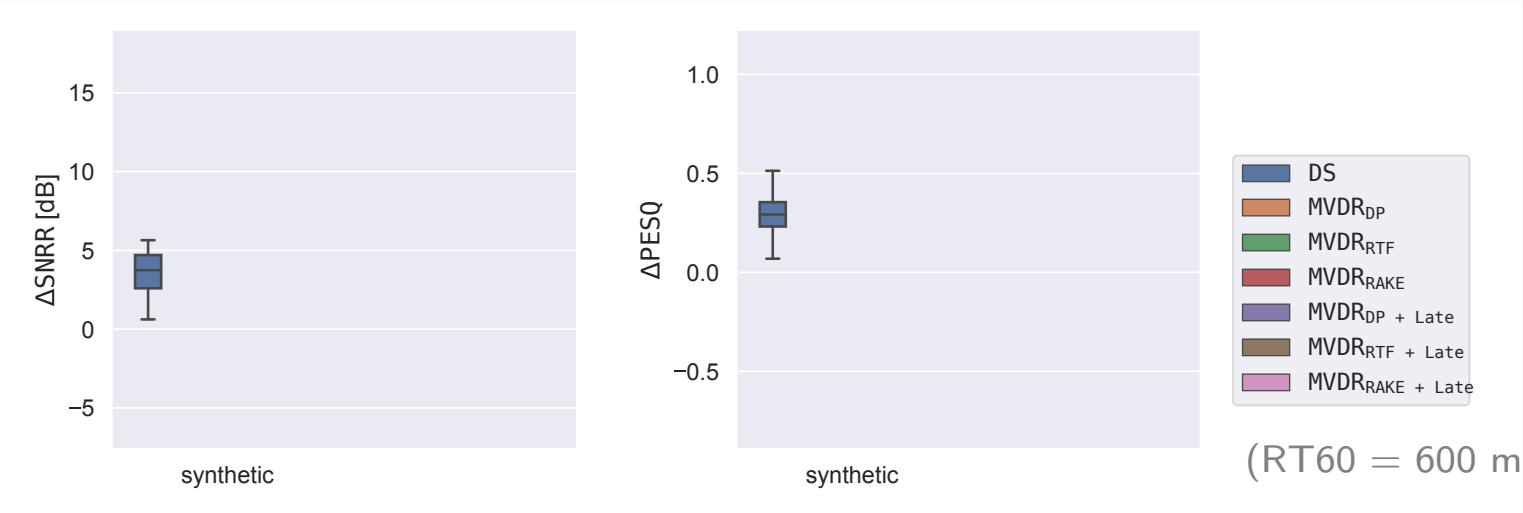
$$\widehat{\mathbf{W}} = \arg\min_{\mathbf{W}} \mathbb{E}\left\{ \left\| \mathbf{W}^{\mathsf{H}}\mathbf{X} \right\|_2^2 \right\} \quad \text{s.t.} \quad \mathbf{W}^{\mathsf{H}}\mathbf{H} = 1$$

Reducing output energy + distortionless ⇔ reduce any 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)

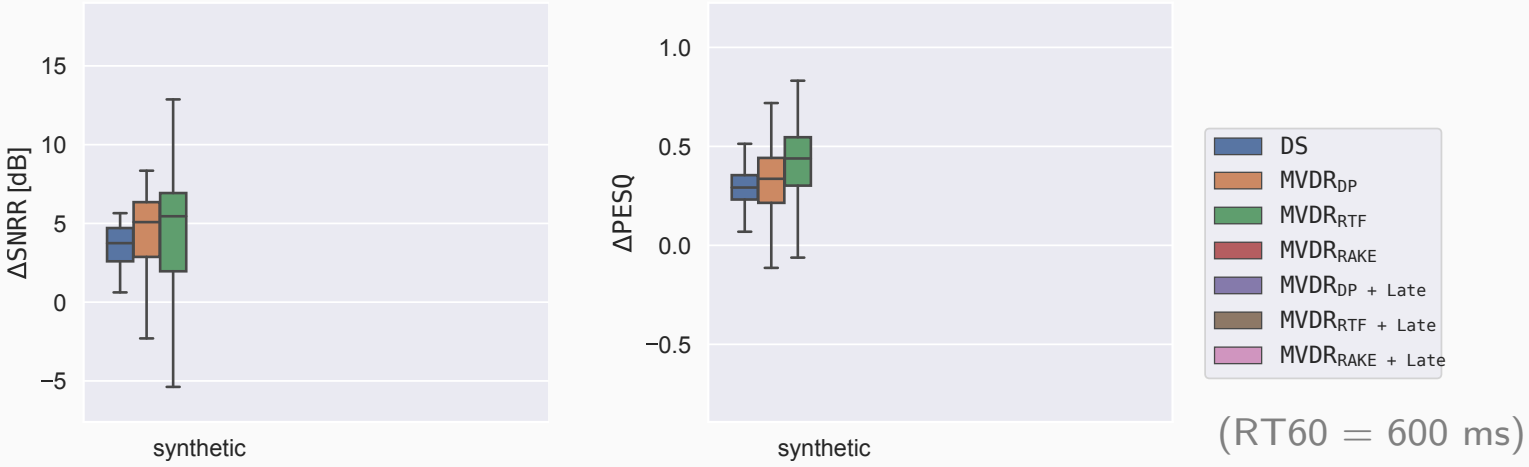


(RT60 = 600 ms)

Speech Enhancement with dEchorate

Methods	Noise covariance matrix	RIRs
DS	-	Direct path (AOA)
MVDR_{DP}	Noise	Direct path (AOA)
$\text{MVDR}_{\text{ReTF}}$ [1]	Noise	Relative Transfer Function

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



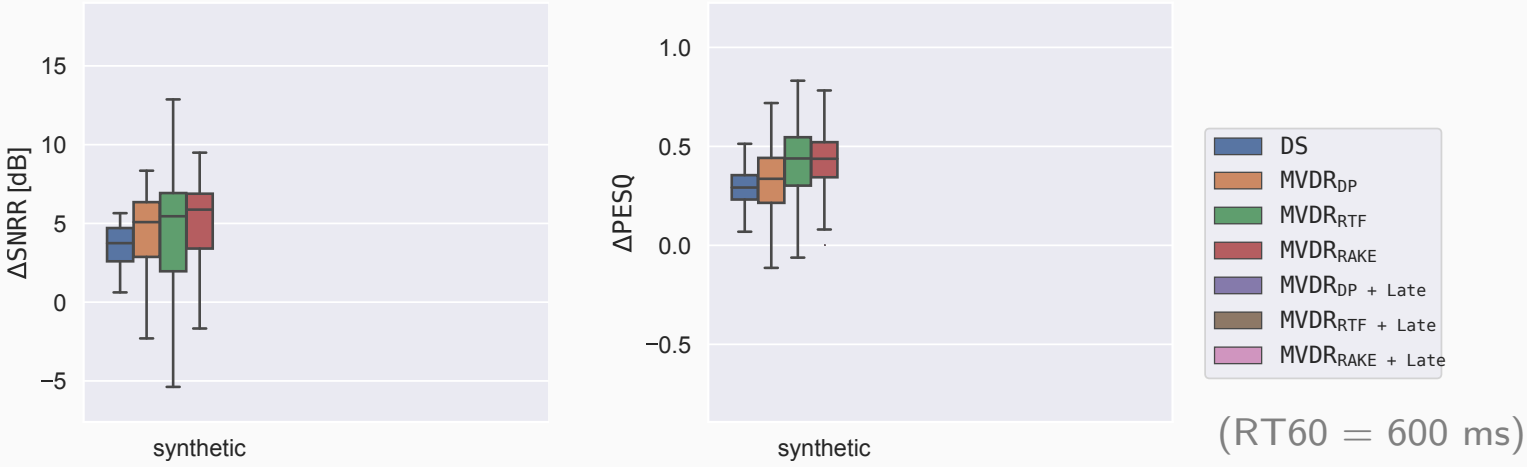
(RT60 = 600 ms)

[1] Using [Markovich-Golan et al., 2018],

Speech Enhancement with dEchorate

Methods	Noise covariance matrix	RIRs
DS	-	Direct path (AOA)
MVDR_{DP}	Noise	Direct path (AOA)
$\text{MVDR}_{\text{ReTF}}$ [1]	Noise	Relative Transfer Function
$\text{MVDR}_{\text{Rake}}$ [2]	Noise	4 strongest echoes per channel

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



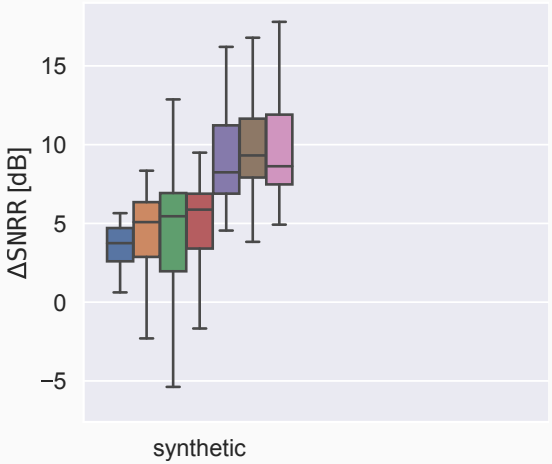
(RT60 = 600 ms)

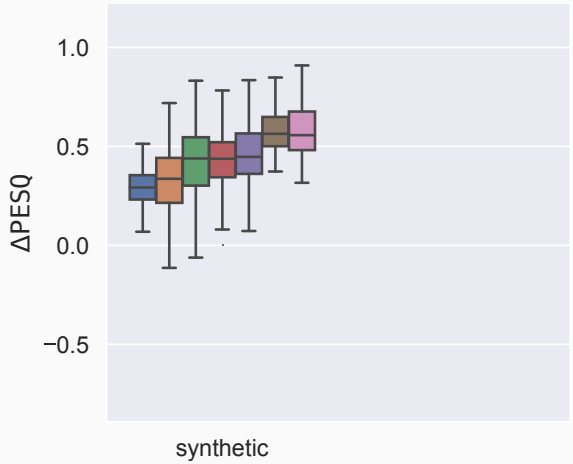
[1] Using [Markovich-Golan et al., 2018], [2] Using [Kowalczyk, 2019],

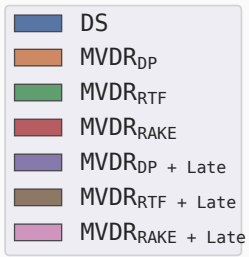
Speech Enhancement with dEchorate

Methods	Noise covariance matrix	RIRs
DS	-	Direct path (AOA)
MVDR_{DP}	Noise	Direct path (AOA)
$\text{MVDR}_{\text{ReTF}}$ [1]	Noise	Relative Transfer Function
$\text{MVDR}_{\text{Rake}}$ [2]	Noise	4 strongest echoes per channel
$\text{MVDR}_{\text{DP+Late}}$	Noise + Late Diffusion [3]	Direct path (AOA)
$\text{MVDR}_{\text{ReTF+Late}}$ [1]	Noise + Late Diffusion [3]	Relative Transfer Function
$\text{MVDR}_{\text{Rake+Late}}$ [2]	Noise + Late Diffusion [3]	4 strongest echoes per channel

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







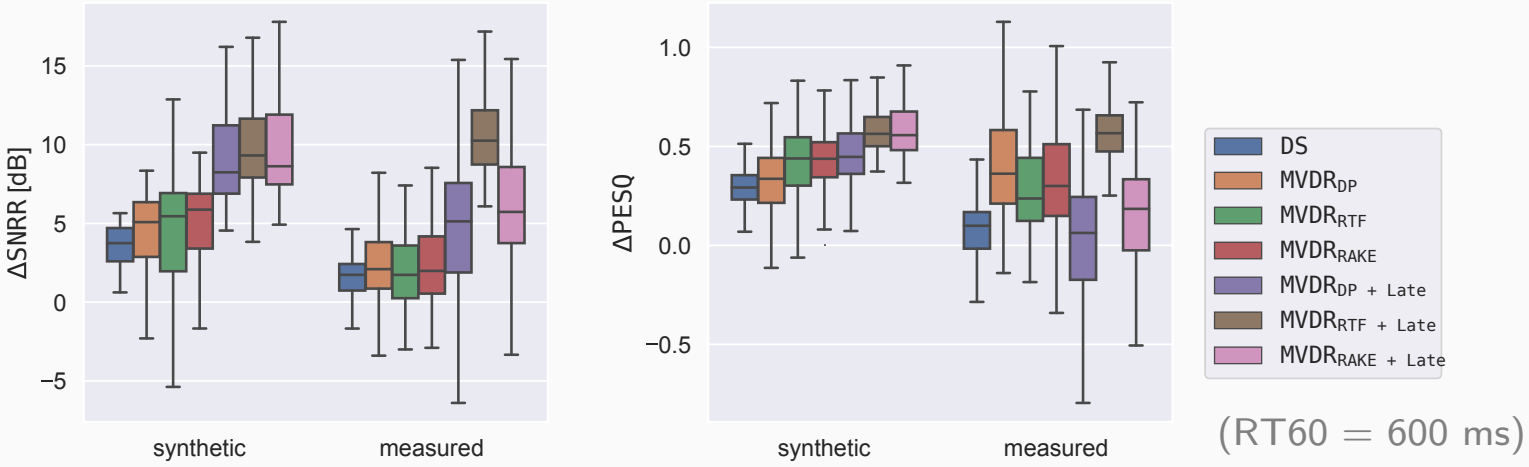
(RT60 = 600 ms)

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

Speech Enhancement with dEchorate

Methods	Noise covariance matrix	RIRs
DS	-	Direct path (AOA)
MVDR_{DP}	Noise	Direct path (AOA)
$\text{MVDR}_{\text{ReTF}}$ [1]	Noise	Relative Transfer Function
$\text{MVDR}_{\text{Rake}}$ [2]	Noise	4 strongest echoes per channel
$\text{MVDR}_{\text{DP+Late}}$	Noise + Late Diffusion [3]	Direct path (AOA)
$\text{MVDR}_{\text{ReTF+Late}}$ [1]	Noise + Late Diffusion [3]	Relative Transfer Function
$\text{MVDR}_{\text{Rake+Late}}$ [2]	Noise + Late Diffusion [3]	4 strongest echoes per channel

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



(RT60 = 600 ms)

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

Conclusion

Summary of contributions

How to estimate them?

In passive stereo scenario:

- Analytical method
  - ✓ direct estimation
  - ✗ depends on source and # echoes
- Learning-based method
  - ✓ estimation on 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 ↰

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:

- on **estimation**
  - develop theoretical guaranties 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`
  - Synthetic to Real RIRs (style transfer, new type of acoustic simulator)
  - Benchmark data for echo-aware algorithms
- **"close the loop":** echo estimation $\Leftrightarrow$ audio analysis
  - in the thesis only the $\Rightarrow$

List of publications and artifacts

- 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 artifacts

- 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


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.


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.

References ii


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.


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.


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.

References iii


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.


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.