Echo-aware signal processing for audio scene analysis

Diego Di Carlo

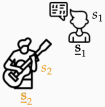
November 24, 2020

suprevisors: Antione DELEFORGE, Nancy BERTIN
collaborators: Clément Elvira, Robin Scheibler, Ivan Dokmanić, Sharon Gannot, Pini A

INRIA IRISA

Introduction

Scenario

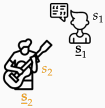


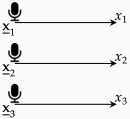
Sound

- produced by sources

Attention: artificial sound vs (natural) microphone recordings

Scenario



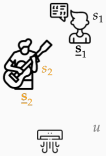


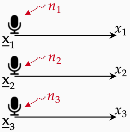
Sound

- produced by sources
- recorded by microphones

Attention: artificial sound vs (natural) microphone recordings

Scenario



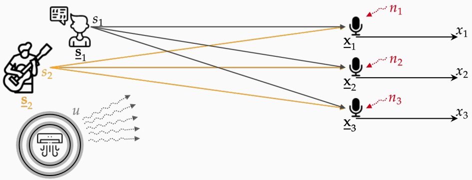


Sound

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

Attention: artificial sound vs (natural) microphone recordings

Scenario

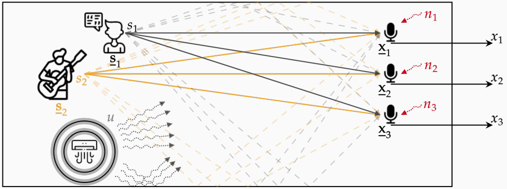


Sound

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

Attention: artificial sound vs (natural) microphone recordings

Scenario

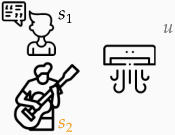


Sound

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

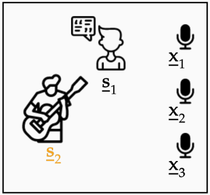
Attention: artificial sound vs (natural) microphone recordings

Semantic information



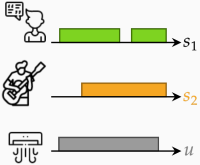
about source nature and semantic content

Spatial information



about source position and room geometry

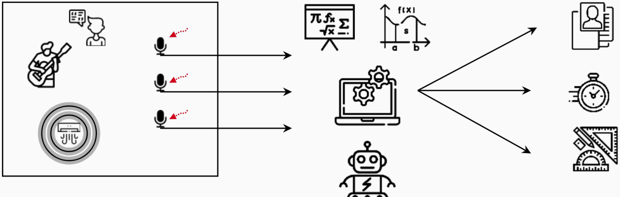
Temporal information



about events activity

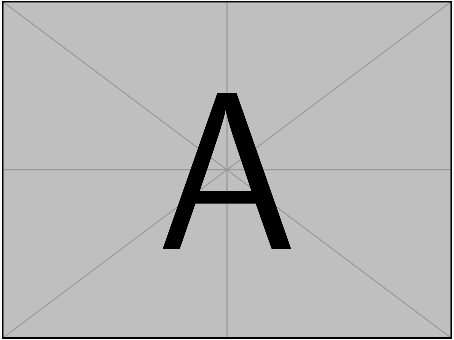
Audio Scene Analysis

Extraction and organization of all the information in the sound



Signal Processing

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



Some problems

• Speaker Verification	Who?
• Sound Source Separation	What?
• Speech Enhancement	
• Automatic Speech Recognition	
• Sound Source Localization	Where?
• Room Geometry Estimation	
• Voice Activity Detection	When?
• Diarization	
• RT_{60} estimation	How?
• Wall Absorption Estimation	
• *and many many other*	

Everything is connected
HOW → WHERE → WHEN → WHAT

Acoustic Echoes

- Elements of the sound propagation
- Sound reflection standing out for time and strength w.r.t. to the total reverberation
- repetition of the source sound but later
- both outdoor and indoor

Audio signal processing w.r.t. sound propagation [?]

- ignore it
- assume it free-field
- model it entirely
- model as few reflection $\rightarrow$ echo-aware processing

Thesis goal and contribution

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

Thesis objective

1. provide new methodologies and data to process and estimate acoustic echoes
2. Turning echoes into friends [?]
   extend previous classical methods for audio scene analysis

contribution
Estimation

- Knowledge-based echo estimation, aka `Blaster`
- Learning-based echo estimation, aka `Lantern`

Application

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

Echo-aware signal processing
for audio scene analysis

Modeling

Acoustic Impulse Response

Sound propagates

Sound source $\rightarrow$? $\rightarrow$ microphone

Sound propagates

$$s \rightarrow h \rightarrow x$$

Acoustic Impulse Response

Sound propagates

Sound source $\rightarrow$ environment $\rightarrow$ microphone

Sound propagates

$$s \rightarrow h \rightarrow x$$

Sound interacts with environment

- it is reflected (specularly and diffusely)
- it is diffracted
- it is absorbers and transmitted

Acoustic Impulse Response (AIR)
the linear filtering effect due to the propagation of sound from a source to a microphone.

$$x_i(t) = (a * s)(t)$$

Acoustic Impulse Response

Sound propagates

Sound source $\rightarrow$ room $\rightarrow$ microphone

Sound propagates

$$s \rightarrow h \rightarrow x$$

Sound interacts with environment

- it is reflected (specularly and diffusely)
- it is diffracted
- it is absorbers and transmitted

Echoes and Room Impulse Response

RIRs can be modeled with the Image Methods

- specular reflection only
- “playing billiard in a concert hall”
- for shoebox room it is is the solution for physics
- in frequency domain it writes as

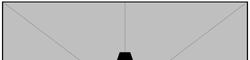
RIRs accounts for
the geometry of the room

- Room shape and size
- Mic and Source position
- presence of objects

the acoustic properties of the audio scene

- surface materials
- objects materials

examples



Room Impulse Response

$$\tilde{x}_i = (\tilde{h}_i * \tilde{s})(t) \longrightarrow \tilde{X}_i(f) = \tilde{H}_{ij}(f)\tilde{S}(f)$$

the linear filtering effect due to the propagation of sound from a source to a microphone in a indoor space

Observation
Our vision is limited both in time (finite and discrete) and in frequency (finite and discrete)

$$x_i[n] = \ldots \tag{1}$$

Signal model in the frequency domain

$$x_i = (h_i * s)(t) \;\longrightarrow\; X(f) = H_i(f)S(f)$$

Approximations

- Narrowband Approximation
- DTFT echo model in the DFT

Interim Conclusion I

Approximations

- Echoes are well described by specular reflection
- Echoes are off-grid by nature
- Sampling and quantization make them hard
- Processing in the discrete frequency domain, but with continuous time echo model

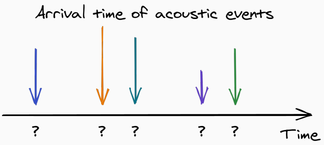
Acoustic Echo Estimation

Acoustic Echo Retrieval

The acoustic echoes retrieval (AER) problem

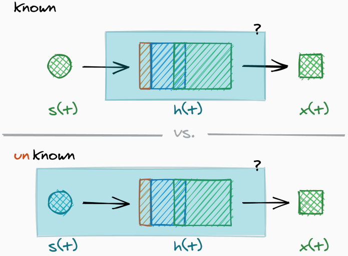
Estimating early (strong) acoustic reflections:

- their time of arrivals → TOAs Estimation
  ↪ sufficient sometimes
- their amplitude
  ↪ closed-form form TOA



Approaches

Source signal is



Active approaches

- solve and easier problem
- Intrusive or specific setup
- Single microphones

Application Sonar, Calibration, Measurements

Passive approaches

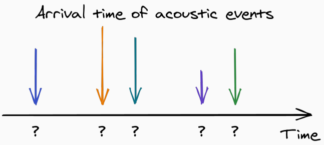
- more difficult problem — blind inverse problem

Acoustic Echo Retrieval

The acoustic echoes retrieval (AER) problem

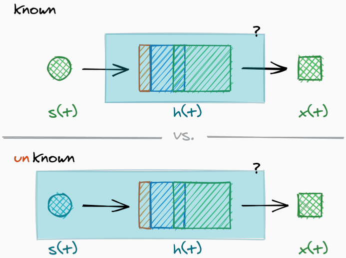
Estimating early (strong) acoustic reflections:

- their time of arrivals → TOAs Estimation
  ↪ sufficient sometimes
- their amplitude
  ↪ closed-form form TOA

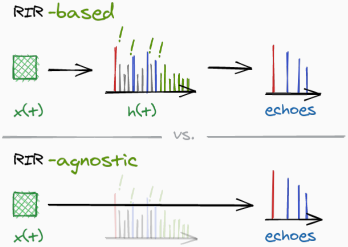


Approaches

Source signal is



Estimation is

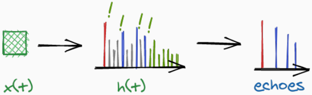


Scenario: signal source, only TOAs and passive system

Passive Acoustic Echo Estimation

Passive Acoustic Echo Estimation:

RIR-based approaches



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

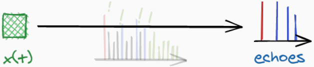
Pros

- SIMO BCE is well studied (elegant framework)
- It works well is some scenarios and in practice
  $\hookrightarrow$ if not limitation

Cons

- Full RIR
- dependent of manually tuned peak picking
- Pathological issue (sampling and body-guard
- Complexity

RIR-agnostic approaches



1. Estimation directly in the echoes parameters space $\{\tau, \alpha\}$
   and direction of arrivals can be used instead

Performed with

- Cross-correlation
  on-grid, eg. EM, Acoustic Cameras
- Cross-relation with super-resolution
  off-grid, [?, ?]

Pro

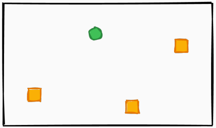
- No need for full RIRs
- Sub-sampling accuracy
- Low complexity
- Sparsity and Non-negativity are respected

Cons

- Exploratory

AER as discrete SIMO BCE

Key ingredient – *Cross relation identity*



$$x_i = h_i * s$$

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

Ideas:

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

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

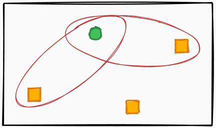
$\mathcal{P}(\mathbf{h_1}, \mathbf{h_2}) \longrightarrow$ sparse promoting regularizer

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

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

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

Key ingredient – *Cross relation identity*



$$x_i = h_i * s$$

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

Ideas:

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

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

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

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

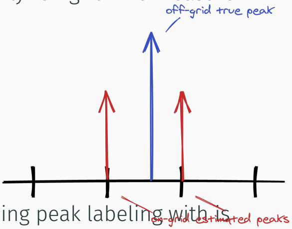
$\mathbf{x_i} * \mathbf{h}_j$ computed as $\mathcal{T}(\mathbf{x}_i)\mathbf{h_j} \in \mathcal{O}(L^2)$

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

AER as SIMO BCE: Limitations & bottleneck

1. Estimation is on-grid

- Sparsity and non-negativity Echoes are not necessarily “on grid” We measure filters, not diracs
- *Body guard* effect [Duval and Peyré, 2017]
  - $\rightarrow$ low recall $\Longrightarrow$ low accuracy
  - $\rightarrow$ slow convergence



... and Pick Picking

$\rightarrow$ Manually tuned peaking or peak disambiguation (using peak labeling with is NP-hard and need other prior knowledge)

Increase the sampling frequency, F_s

$\rightarrow$ Increase Precision

Computational bottleneck

- Bigger vectors and matrices
  - $\longrightarrow$ memory usage
- Computational complexity: at best $\mathcal{O}(F_s^2)$ per iteration
- the higher the sampling frequency, the more ill-conditioned

AER: Challenges and Contribution

State Of The Art

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

AER: Challenges and Contribution

State Of The Art

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

$\Longrightarrow$ however

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

AER: Challenges and Contribution

State Of The Art

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

$\Longrightarrow$ however

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

$\Longrightarrow$ we propose

`Blaster`

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

[Di Carlo et al., 2020] Collaboration

`Lantern`

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

[Di Carlo et al., 2019]

Blaster- Knowledge-based Off-grid AER

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

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

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

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

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

Idea: Recover echoes by matching a finite number of frequencies

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

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

✓no Toeplitz matrix

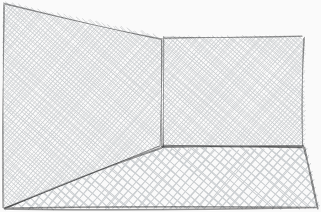
✓ Solutions is a train of Dirac

✓ anchor prevents trivial solution

In the manuscript:

Methods

- BSN: Blind Sparse and Nonnegative SIMO BCE [Lin et al., 2007]
- IL1C: Iteratively-weighted ℓ_1 Constraint SIME BCE [Crocco and Del Bue, 2015]
- Blaster: Proposed off-grid approach



Baseline method are xvalidated on other dataset

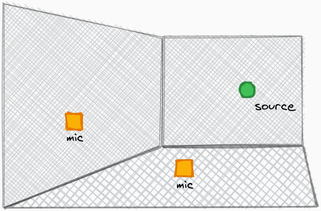
Metrics

- Precision (how many estimated echoes are correct)
- RMSE (error on the correct guess)

Blaster- Experimental Results

Methods

- BSN: Blind Sparse and Nonnegative SIMO BCE [Lin et al., 2007]
- IL1C: Iteratively-weighted ℓ_1 Constraint SIME BCE [Crocco and Del Bue, 2015]
- Blaster: Proposed off-grid approach



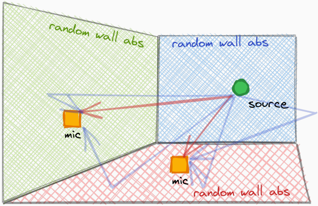
Baseline method are xvalidated on other dataset

Metrics

- Precision (how many estimated echoes are correct)
- RMSE (error on the correct guess)

Methods

- BSN: Blind Sparse and Nonnegative SIMO BCE [Lin et al., 2007]
- IL1C: Iteratively-weighted ℓ_1 Constraint SIME BCE [Crocco and Del Bue, 2015]
- Blaster: Proposed off-grid approach



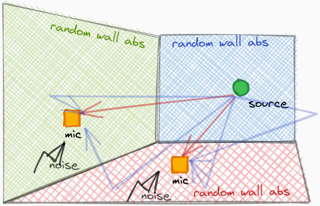
Baseline method are xvalidated on other dataset

Metrics

- Precision (how many estimated echoes are correct)
- RMSE (error on the correct guess)

Methods

- BSN: Blind Sparse and Nonnegative SIMO BCE [Lin et al., 2007]
- IL1C: Iteratively-weighted ℓ_1 Constraint SIME BCE [Crocco and Del Bue, 2015]
- Blaster: Proposed off-grid approach



Baseline method are xvalidated on other dataset

Metrics

- Precision (how many estimated echoes are correct)
- RMSE (error on the correct guess)

Blaster- Observation

Observation

- ✓ Smaller RMSE
- ✗ Sensitive to source signal

Future work:

- Real data
- Extention to Multichannel
- Relative Transfer Function

Lantern- Data-driven off-grid AER

Observation 1: Mapping from observation to echo is extremely difficult
Later echoes are not considered, may help

Observation 2: We have acoustic simulators
Acoustic simulators based on ISM

source position, room $\leftarrow$ reverberation elements $\leftarrow$

annotation for free

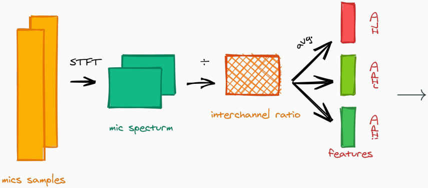
Observation 3: (Deep) Learning-based methods successful for localization
Echoes are strongly related to the source position

Idea: Use Deep Learning for AER

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

Which mapping?

Input: features

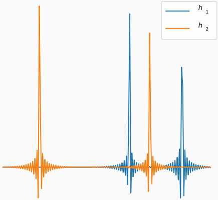


Relative Transfer Function

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

This is the **instantaneous ReTF**

Output: target



TDOAs inter and intra reflection

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

First strongest echo ⇔ close surface

$\texttt{TDOE}_2$ is a combination of $\mathcal{V}$

Virtually Supervised Learning (= data from acoustic simulator)

Which Model?

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

Generative models ← for data fusion, similar to **MDN** [Bishop, 1994]

Lantern- Experiments & Resuls

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

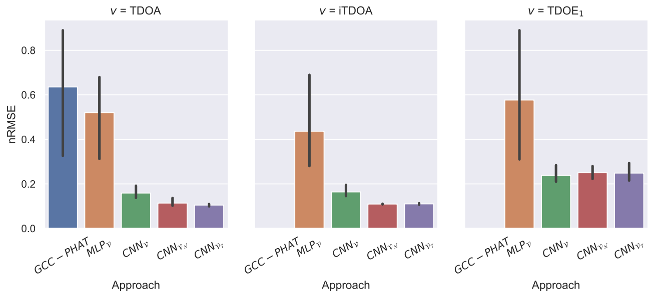
Proposed: $\mathrm{CNN}_{\mathcal{V}}$, $\mathrm{CNN}_{\mathcal{V}_{\mathcal{N}}}$, $\mathrm{CNN}_{\mathcal{V}_{\mathcal{T}}}$

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

Train:

- RT60, SNR
- white noise
- instantaneous RTF

Test: similar to train



✓ CNNs outperform GCC-PHAT and MLP

✓ CNNs more robust to noise

✗ Gaussian ∼ Student-T

Lantern- Experiments & Resuls

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

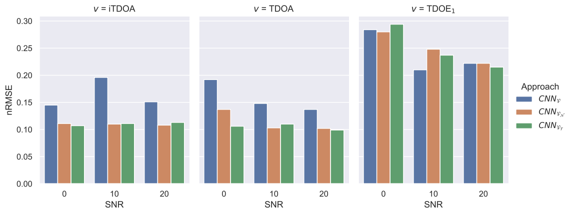
Proposed: $\text{CNN}_{\mathcal{V}}$, $\text{CNN}_{\mathcal{V}_{\mathcal{N}}}$, $\text{CNN}_{\mathcal{V}_{\mathcal{T}}}$

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

Train:

- RT60, SNR
- white noise
- instantaneous RTF

Test: similar to train



✓ Generative better than Normal

✗ Gaussian Student-T ∼

Echo-aware Application

Sound propagation is

- completely ignored
- assumed free-field case (*anechoic*)
- model it full (*reverberant*)
- *learned* it full (*reverberant*)
- model few early echoes (*multipath*)

$$x_i(t) = (h_i * s)(t)$$

$$h_i(t) = h_i^d(t) + h_i^e(t) + h_i^r(t)$$

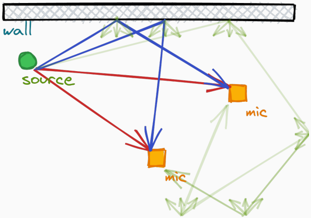
Recall

- anechoic case: easy mapping, but incoherence or wrong processing
- reverberant case: difficult mapping and estimation, but coherent processing

What can we do with echoes?

Sound propagation is

- completely ignored
- assumed free-field case (*anechoic*)
- model it full (*reverberant*)
- *learned* it full (*reverberant*)
- model few early echoes (*multipath*)



$$x_i(t) = (h_i * s)(t)$$

$$h_i(t) = h_i^d(t) + h_i^e(t) + h_i^r(t)$$

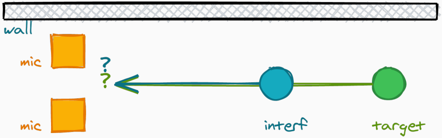
Recall

- anechoic case: easy mapping, but incoherence or wrong processing
- reverberant case: difficult mapping and estimation, but coherent processing

What can we do with echoes?

Echo-aware Application

Echoes = same content, different time/direction



Echoes helps indoor processing:

What?
Echoes = repetitions

- Sound Source Separation
- Speech Enhancement (Dereverberation, Denoising, Room Equalization)

Where?
Echoes $\in$ indoor propagation

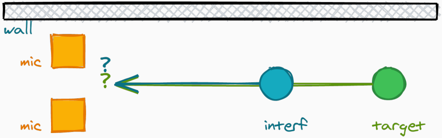
- Sound Source Localization
- Microphone Calibration
- Room Geometry Reconstruction

How?
Echoes $\in$ sound propagation

- Blind Channel Estimation
- Acoustic Measurements

Echo-aware Application

Echoes = same content, different time/direction



Echoes helps indoor processing:

What?
Echoes = repetitions

- Sound Source Separation
- Speech Enhancement (Dereverberation, Denoising, Room Equalization)

Where?
Echoes ∈ indoor propagation

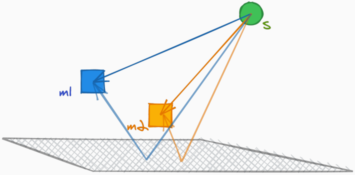
- Sound Source Localization
- Microphone Calibration
- Room Geometry Reconstruction

How?
Echoes ∈ sound propagation

- Blind Channel Estimation
- Acoustic Measurements

The Picnic Scenario:

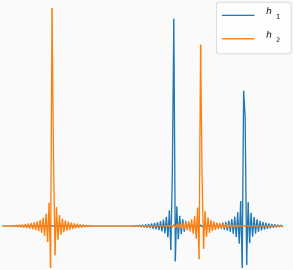
- One source
- Two microphones
  - $\rightarrow$ passive scenario
  - $\rightarrow$ generalizable later
- Close to a very reflective surface
  - $\rightarrow$ First echo = Strongest echo
  - $\rightarrow$ α_{picnic} const. $\forall f$
  - $\rightarrow$ table-top device



How to access the *image* microphones?
$\implies$ each pair is augmented with echoes

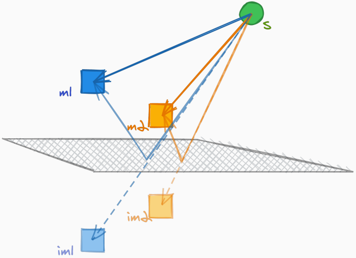
Mirage Array

idea: use SSL algorithm on this augmented array
recall: echoes are known



The Picnic Scenario:

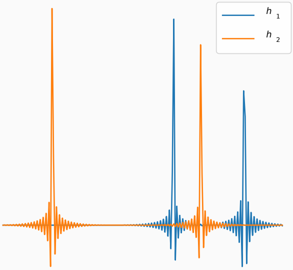
- One source
- Two microphones
  - → passive scenario
  - → generalizable later
- Close to a very reflective surface
  - → First echo = Strongest echo
  - → α_{picnic} const. $\forall f$
  - → table-top device



How to access the *image* microphones?
$\implies$ each pair is augmented with echoes

Mirage Array

idea: use SSL algorithm on this augmented array
recall: echoes are known

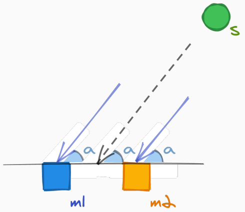


Mirage- Sound Source Locatization with Echoes

SSL with 2 microphones

1. only Angle of Arrival (AOA) w.r.t. the frame of the pair

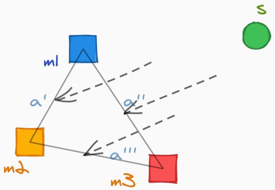
- e.g. GCC-PHAT for TDOA estimation
  (known limitation, but good in practice)
- TDOA to AoA known frame distance



SSL with more microphones

1. For each pair m:
   $\text{AOA}_m \leftarrow$ TDOA-based 2-mic-SSL
2. “Aggregate” together all the observation
   (Angular spectra, Probability distributions)

eg. SRP-PHAT



Baseline:
GCC-PHAT on true microphones

Proposed Approach:
Using Lantern (DNN-based TDOA estimation)
problem: real value not estimation $\rightarrow$ Generative Model given the TDOA axis

Mirage- Results

Data
Virtually generated dataset as for Lantern

AOA estimation
normalized nRMSE
Angular Error —
mean and accuracy

	Input	nRMSE TDOA	nRMSE iTDOA	nRMSE TDOE	ACCURACY $\theta < 10$	ACCURACY $\theta < 20$
MIRAGE	wn	0.18	0.28	0.25	4.10 (77)	5.97 (97)
MIRAGE	wn+n	0.68	0.69	0.89	5.00 (26)	9.89 (54)
MIRAGE	sp	0.31	0.34	0.56	4.83 (63)	7.26 (82)
MIRAGE	sp+n	0.99	0.98	1.48	4.60 (16)	9.88 (35)
GCC-PHAT	wn	0.21	-	-	4.22 (81)	6.19 (97)
GCC-PHAT	wn+n	0.68	-	-	4.03 (65)	5.34 (83)
GCC-PHAT	sp	0.32	-	-	4.08 (82)	5.34 (97)
GCC-PHAT	sp+n	1.38	-	-	4.70 (19)	8.38 (32)

Azimuth Elevation Estimation
Angular Error —
mean and accuracy

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

✓ Solved “impossible” localization

✗ Performance depending on echo estimation

Echo-aware Dataset

Echo-aware Datasets

Data in audio signal processing

1. are necessary for validating (and learning) models
2. collecting real data is a not always possible
   annotation and recording require expertise, equipment and time
3. dataset of real data cannot be easily shared
   they do not generalize to different use-cases and scenarios (array, recording scenario)
4. simulated data are used instead: quantity, versatility, annotation easiness and "quality"

Echo-aware Data in audio signal processing

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

For RooGE: good geo. annotation, but no variety of acoustic scenarios
[Dokmanić et al., 2013, Crocco et al., 2017, Remaggi et al., 2019]

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

dEchorate realization

dEchorate: echo-aware dataset
Recorded Acoustic lab of Bar'Ilan (Shoebox)
Annotated during confinement COVID-2020
Collaboration with prof. Sharon Gannot and ing. Pinchas Tandeitnik

Key features:

- many acoustic environments (revolving panels)
- 6 nULA with 5 mics and 4 sound sources
- geometry annotated
- echo annotated
- measured RIRs and *matching* simulated RIRs

Echo Annotation

1. RIR estimation with ESS [Farina, 2007]
2. IPS with beacon
3. GUI for echo annotation
   Skyline, Matched Filter, Assisted Peak Picking
4. Refined position with Least Square optimization
5. iterate including ceiling (perfectly flat)

dEchorate realization

dEchorate: echo-aware dataset
Recorded Acoustic lab of Bar'Ilan (Shoebox)
Annotated during confinement COVID-2020
Collaboration with prof. Sharon Gannot and ing. Pinchas Tandeitnik

Key features:

- many acoustic environments (revolving panels)
- 6 nULA with 5 mics and 4 sound sources
- geometry annotated
- echo annotated
- measured RIRs and *matching* simulated RIRs

Echo Annotation

1. RIR estimation with ESS [Farina, 2007]
2. IPS with beacon
3. GUI for echo annotation
   Skyline, Matched Filter, Assisted Peak Picking
4. Refined position with Least Square optimization
5. iterate including ceiling (perfectly flat)

dEchorate: echo-aware dataset
Recorded Acoustic lab of Bar'Ilan (Shoebox)
Annotated during confinement COVID-2020
Collaboration with prof. Sharon Gannot and ing. Pinchas Tandeitnik

Key features:

- many acoustic environments (revolving panels)
- 6 nULA with 5 mics and 4 sound sources
- geometry annotated
- echo annotated
- measured RIRs and *matching* simulated RIRs

Echo Annotation

1. RIR estimation with ESS [Farina, 2007]
2. IPS with beacon
3. GUI for echo annotation
   Skyline, Matched Filter, Assisted Peak Picking
4. Refined position with Least Square optimization
5. iterate including ceiling (perfectly flat)

Room Geometry Estimation

Estimating the room geometry: shape, volume or reflector position)
from signal or form TOAs and labels

If TOAs annotation (label and value) are available, RooGE as Image Source Inversion:
For each wall/label:

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

Other methods differs for prior knowledge and setup [?, ?, ?]

Room Geometry Estimation

Estimating the room geometry: shape, volume or reflector position)
from signal or form TOAs and labels

If TOAs annotation (label and value) are available, RooGE as Image Source Inversion:
For each wall/label:

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

Other methods differs for prior knowledge and setup [?, ?, ?]

IMAGE EXAMPLE HERE

Room Geometry Estimation

Estimating the room geometry: shape, volume or reflector position)
from signal or form TOAs and labels

If TOAs annotation (label and value) are available, RooGE as Image Source Inversion:
For each wall/label:

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

Other methods differs for prior knowledge and setup [?, ?, ?]

TABLES RESULTS HERE

Speech Enhancement
Improve the quality of a *target* sound source with respect:

- interferences, i.e. form other sources $\rightsquigarrow$ sound source separation
- background noise $\rightsquigarrow$ denoising
- reverberation $\rightsquigarrow$ dereverberation, room equalization

Spatial filtering via Beamformers

- Is a speech enhancement techniques for multichannel
- vs. Wiener Filtering, the target is distortionless
- in anechoic case, it correspond to delay-and-sum beamformer
- physical interpretation with steering vector based on DOA
- both in time and frequency domain

Speech Enhancement
Improve the quality of a *target* sound source with respect:

- interferences, i.e. form other sources ⇝ sound source separation
- background noise ⇝ denoising
- reverberation ⇝ dereverberation, room equalization

Spatial filtering via Beamformers

- Is a speech enhancement techniques for multichannel
- vs. Wiener Filtering, the target is distortionless
- in anechoic case, it correspond to delay-and-sum beamformer
- physical interpretation with steering vector based on DOA
- both in time and frequency domain

Beamforming: Delay and Sum

$$\mathbf{y}[l,k] = \mathbf{W}^{\mathsf{H}}\mathbf{x}[l,k]$$

STFT Signal Model

$$\mathbf{x}[l,k] = \mathbf{H}[k]\mathbf{s}[l,k] + \mathbf{n}[l,k]$$

Beamforming: Filter and Sum

$$\mathbf{y}[l,k] = \mathbf{W}^{\mathsf{H}}\mathbf{x}[l,k]$$

Beamforming in the STFT domain: apply filter and sum independently at each frequency bin

The PSD of various components
asd

Different Criteria and Solution

- DS
- MVDR - DP
- MVDR ReTF

IMAGE RESULTS

Interim conclusion (3/4)

`dEchorate` dataset for echo-aware signal processing

- designed for AER, SE and RooGE
- Geometrical annotation $\longleftrightarrow$ image source annotation $\longleftrightarrow$ Signal Annotation
- Measured Real RIRs and equivalent synt RIR
- also speech, noise, babble noise and different room conf (+fornitures)
- GUI, tools and code

Application
Echo Estimation

- Huge difference between real and simulated data

Room Geometry Reconstruction

- some annotation inconsistencies are noticed (but manually corrected)

Echo-aware Speech Enhancement

- a
- b

Conclusion

2D Outline

Thesis outline with projects

Backup slides

Sometimes, it is useful to add slides at the end of your presentation to refer to during audience questions.

The best way to do this is to include the `appendixnumberbeamer` package in your preamble and call `\appendix` before your backup slides.

will automatically turn off slide numbering and progress bars for slides in the appendix.

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.

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

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

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

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

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

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

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

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

Di Carlo, D., 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.

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

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

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

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.

Lin, Y., Chen, J., Kim, Y., and Lee, D. D. (2007).
Blind sparse-nonnegative (bsn) channel identification for acoustic time-difference-of-arrival estimation.
In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 106–109. IEEE.

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

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

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

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

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

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