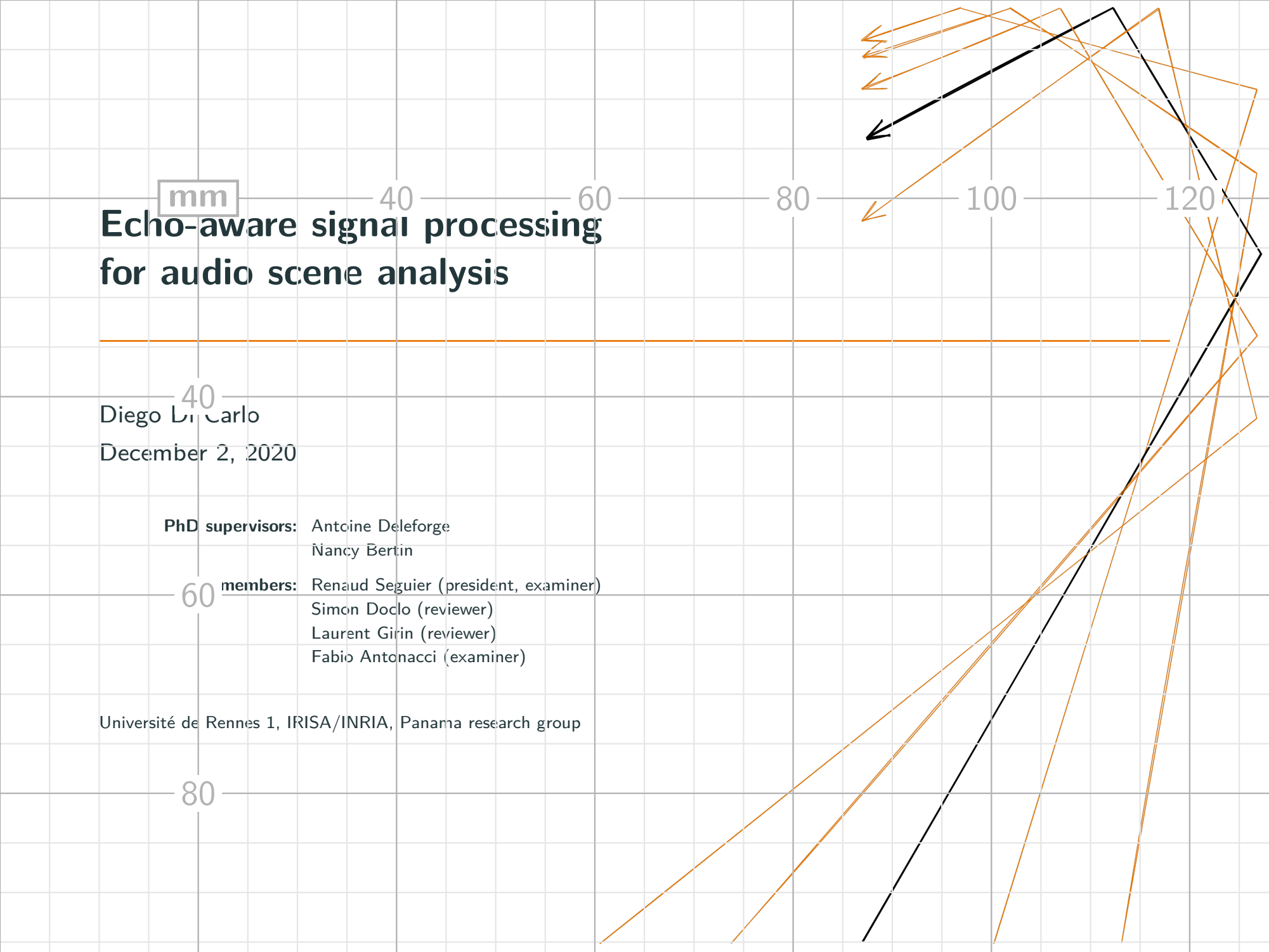


Echo-aware signal processing for audio scene analysis

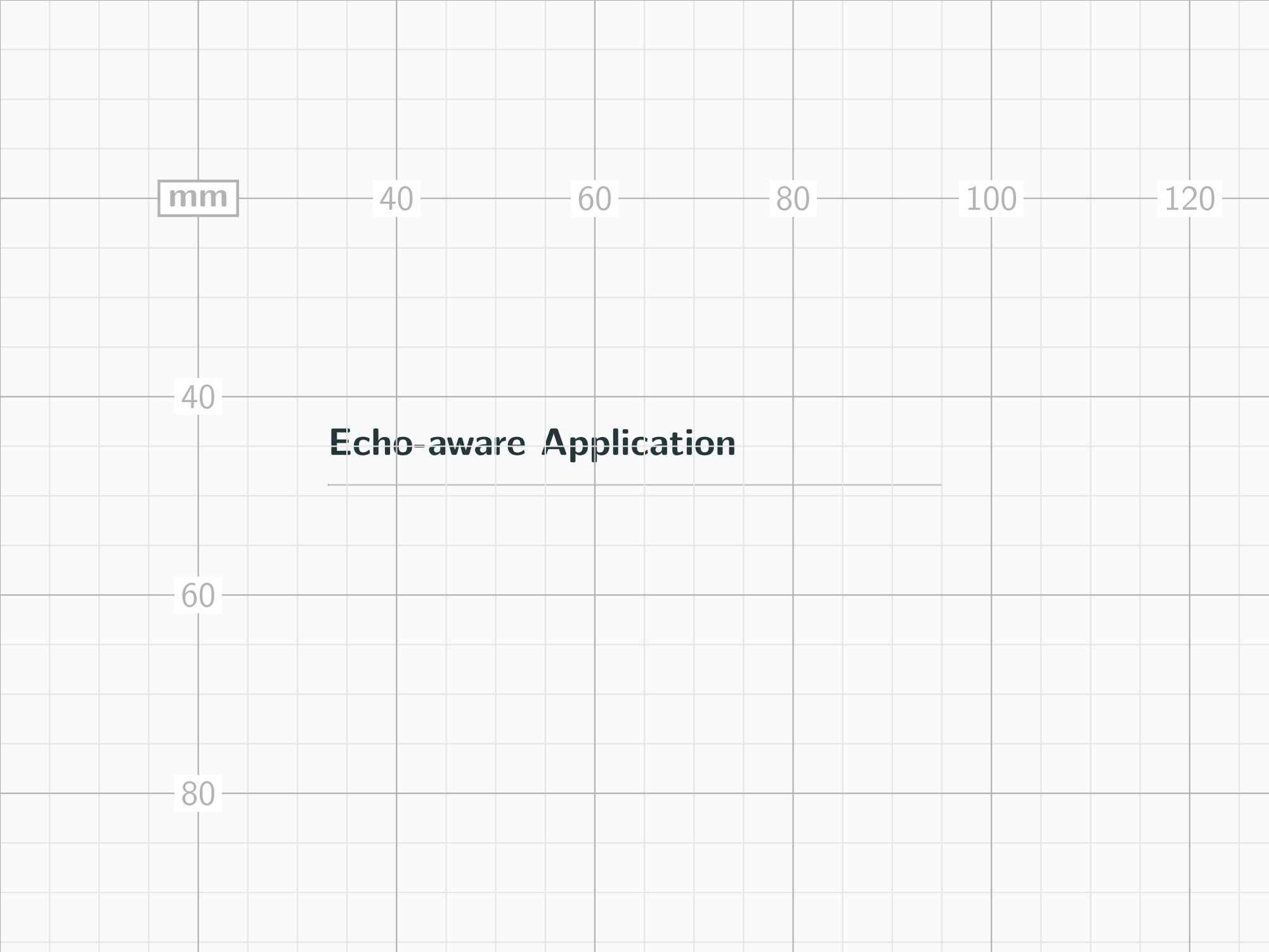
Diego Di Carlo

December 2, 2020

PhD supervisors: Antoine Deleforge
Nancy Bertin

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

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

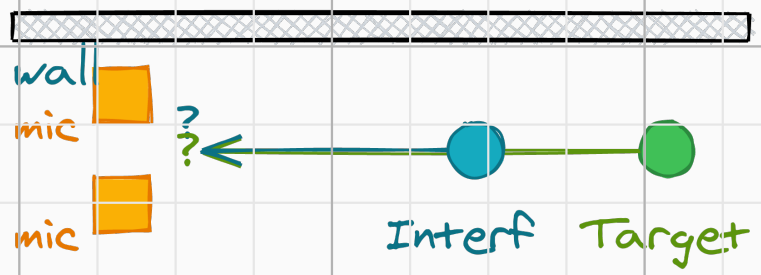

mm
40
60
80
100
120
40
60
80
Echo-aware Application

Echo-aware Application

Echo mm **me conten** 40 **ifferent time** 60 **rection** 80 100 120

Image Source Model
⇔
Image Microphone Model

40



What?
Echoes 60 opy

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

80

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 [? Crocco et al., 2017]

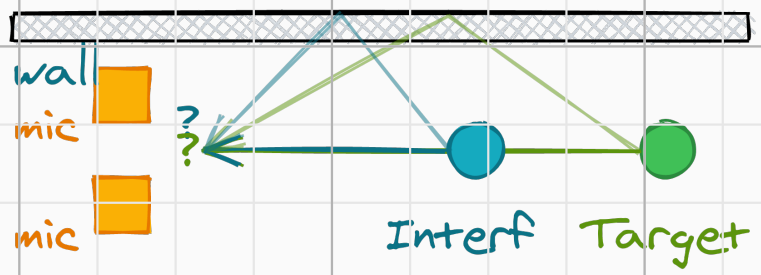
How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echo = mm me conten ifferent time ection

Image Source Model
⇔
Image Microphone Model



What?
Echoes opy

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

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 [? Crocco et al., 2017]

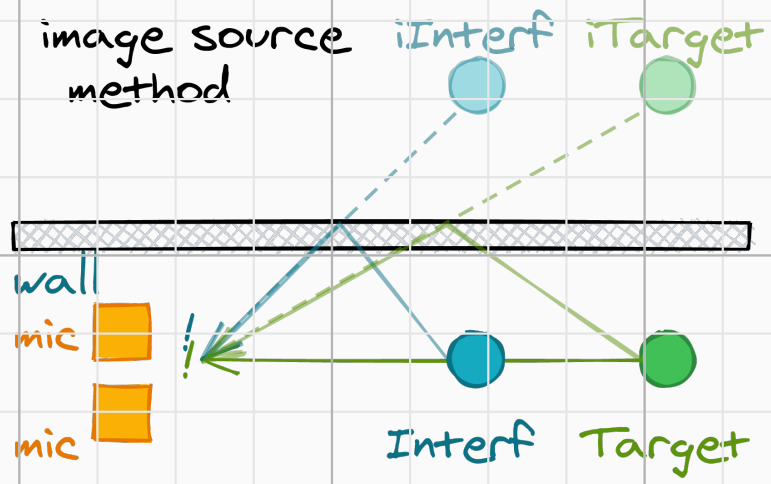
How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echo mm **me conten** 40 **ifferent time** 60 **rection** 80 100 120

Image Source Model
⇔
Image Microphone Model



40

What?
Echoes 60 opy

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

80

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 [? Crocco et al., 2017]

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echo mm **me conten** 40 **ifferent time** 60 **rection** 80 100 120

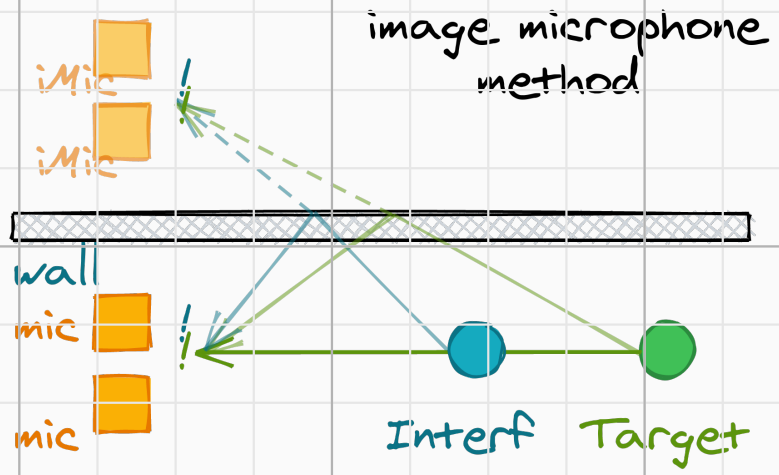


Image Source Model
⇔
Image Microphone Model

What?
Echoes 60 opy

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

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 [? Crocco et al., 2017]

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echo mm **me conten** 40 **ifferent time** 60 **rection** 80 100 120

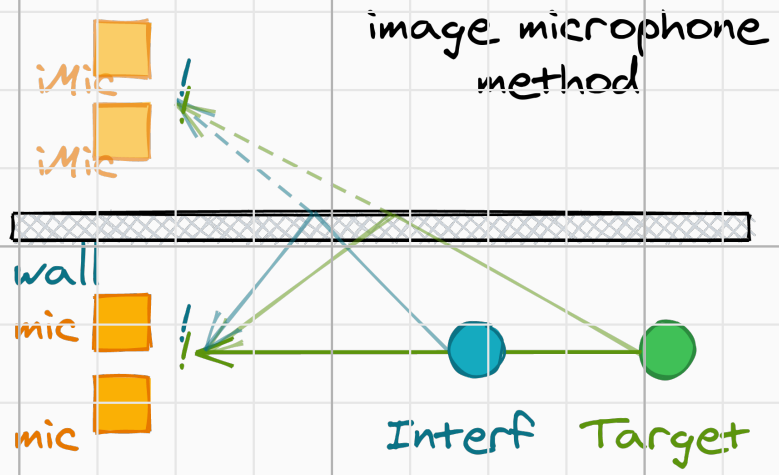


Image Source Model
⇔
Image Microphone Model

What?
Echoes 60 opy

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

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 [? Crocco et al., 2017]

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echo mm **me conten** 40 **ifferent time** 60 **rection** 80 100 120

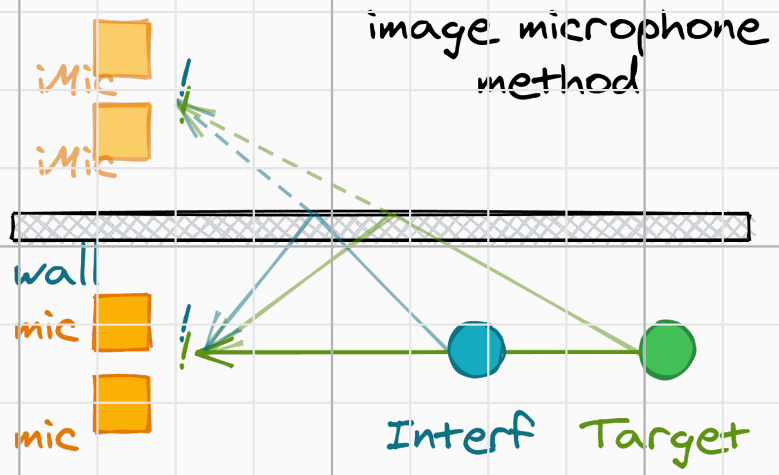


Image Source Model
⇔
Image Microphone Model

What?
Echoes 60 opy

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

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 [? Crocco et al., 2017]

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echo mm **me conten** 40 **ifferent time** 60 **rection** 80 100 120

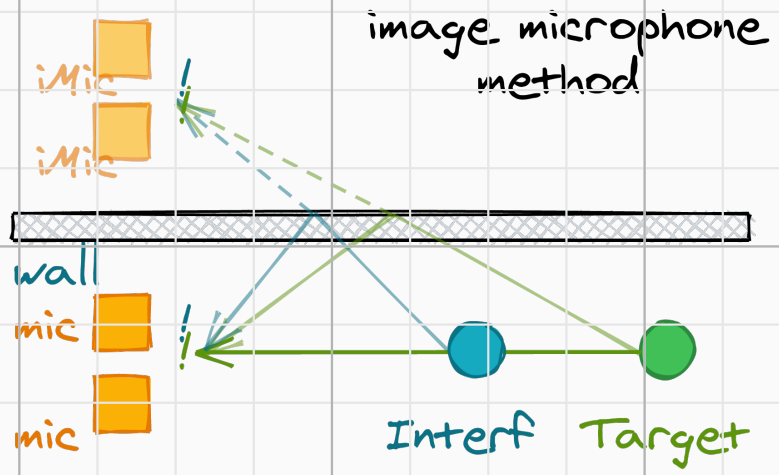


Image Source Model
⇔
Image Microphone Model

What?
Echoes 60 opy

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

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 [? Crocco et al., 2017]

How?
Echoes ∈ sound propagation

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

Echo-aware Application

Echo mm me conten 40 ifferent time 60 ection 80 100 120

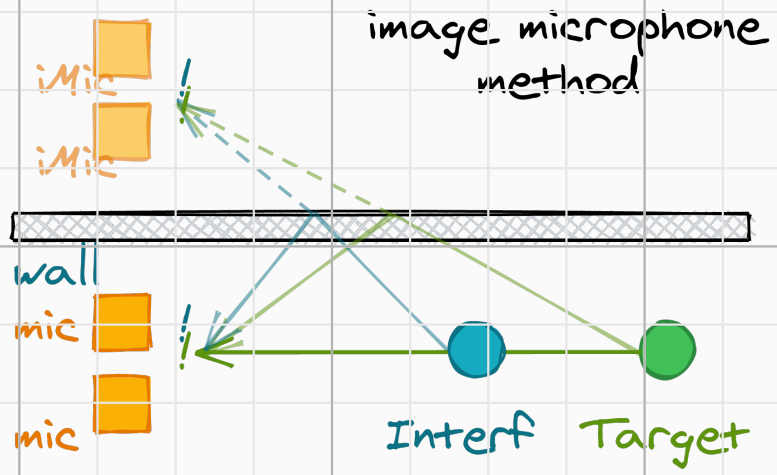


Image Source Model
⇔
Image Microphone Model

What?
Echoes 60 opy

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

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
  [? Crocco et al., 2017]

How?
Echoes ∈ sound propagation

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

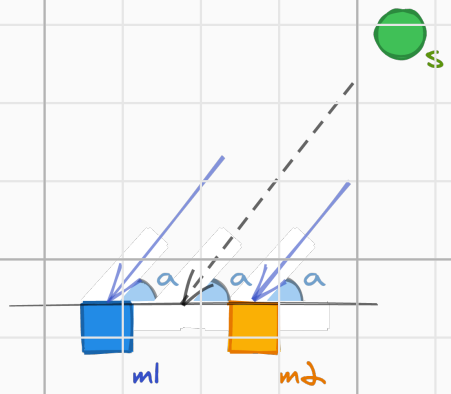
Sound Source Localization (SSL)

(common knowledge)

We consider here distance estimation.

SSL with 2 microphones

- Only one 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]

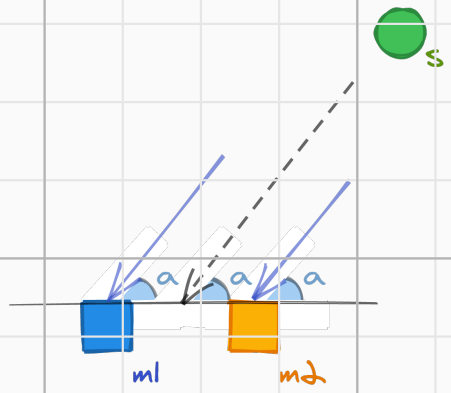
Sound Source Localization (SSL)

(common knowledge)

We consider here distance estimation.

SSL with 2 microphones

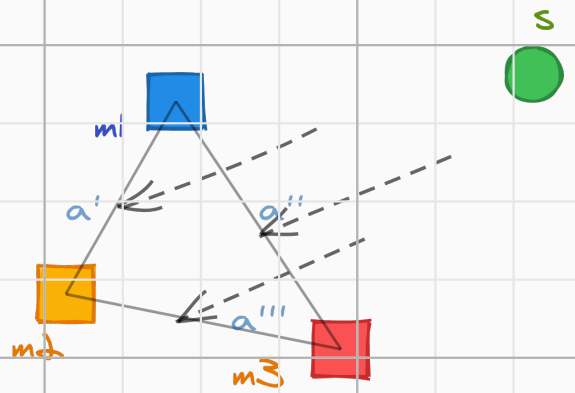
- Only one angle of arrival (AOA)
- can be approximated from TDOA using e.g. GCC-PHAT[1]
  (known limitation, but good in practice)



SSL with more microphones

- To get Direction of Arrival (DoA): azimuth (↔) and elevation (↕)

1. AOA for each pair can be "fused" 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]

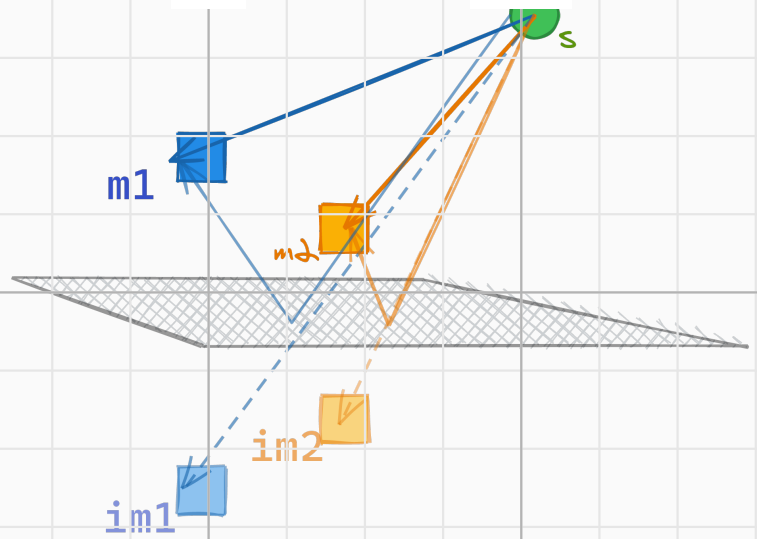
Sound Source Localization with Echoes

The Picnic Scenario:

- One source
- Two microphones

→ passive scenario

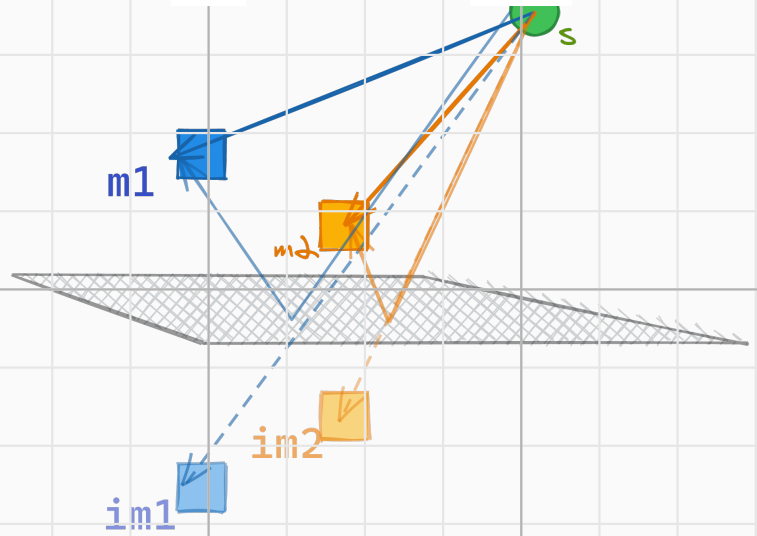
→ generalizable to any array geometry



Sound Source Localization with Echoes

The Picnic Scenario:

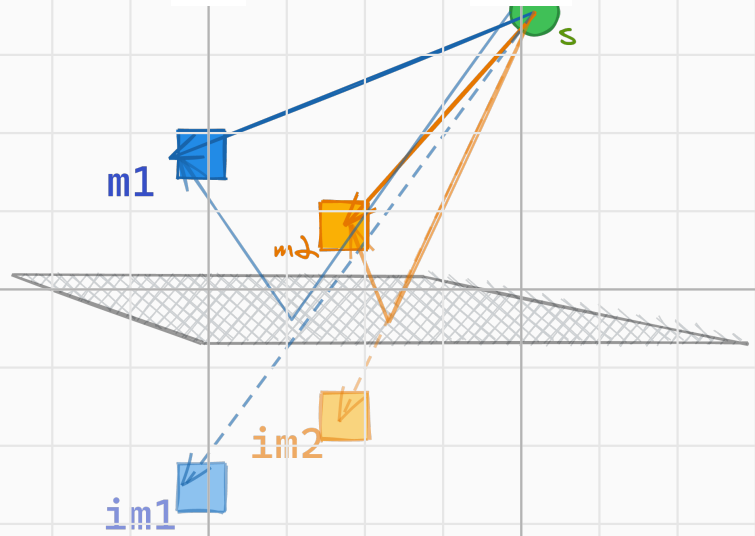
- One source
- Two microphones
  - → passive scenario
  - → generalizable to any array geometry
- Close to a very reflective surface
  - → First echo = Strongest echo
  - → α_{picnic} const. $\forall f$
  - → table-top device



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

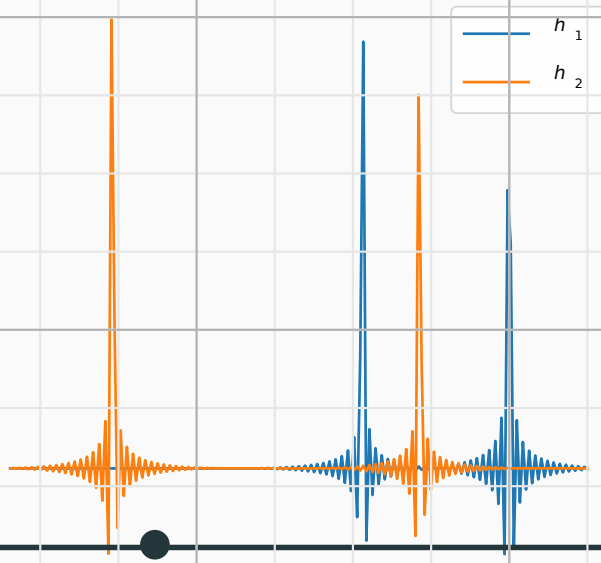


Each [illegible] is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

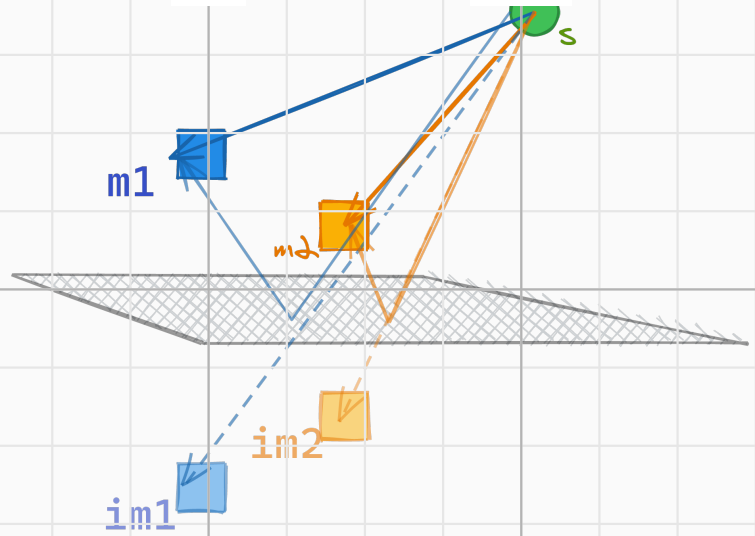
How t[illegible]ccess the *image* microphones?



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

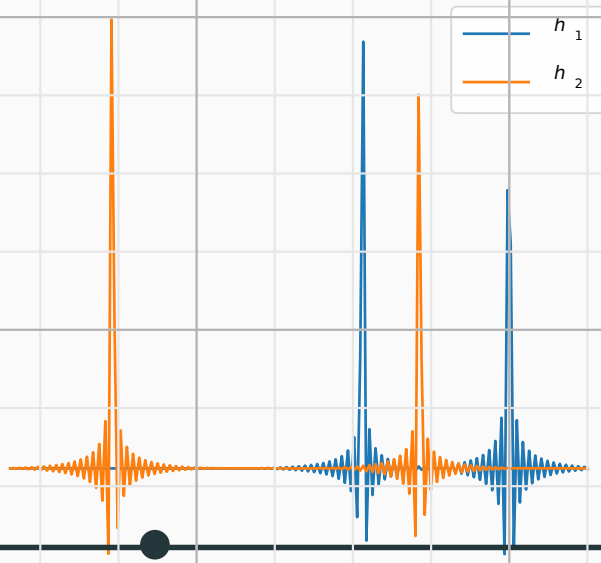


Each [illegible] is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

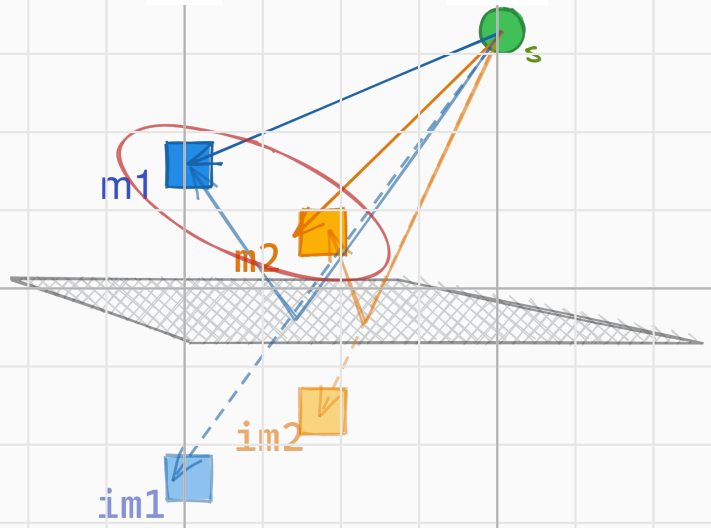
How t[illegible]ccess the *image* microphones?



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

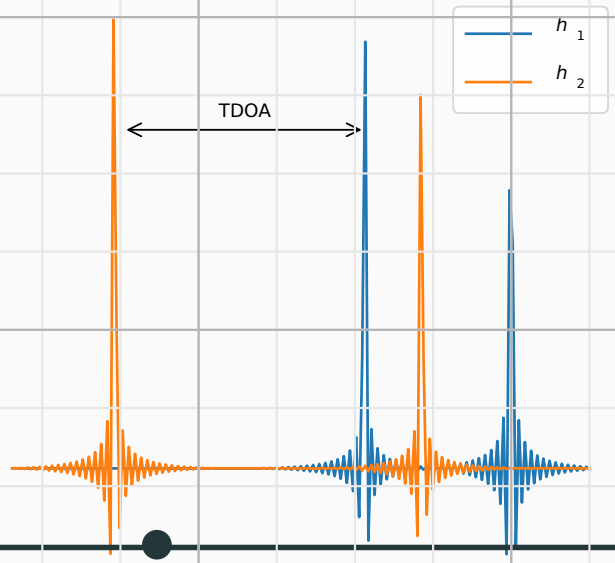


Each [illegible] is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

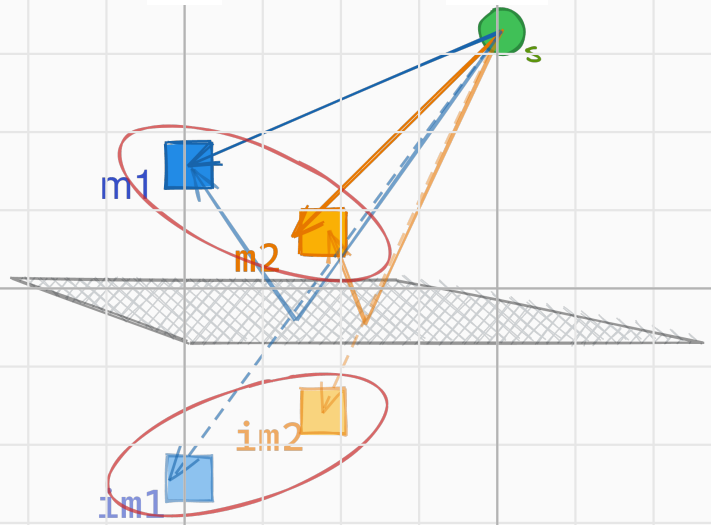
How t[illegible]ccess the *image* microphones?



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

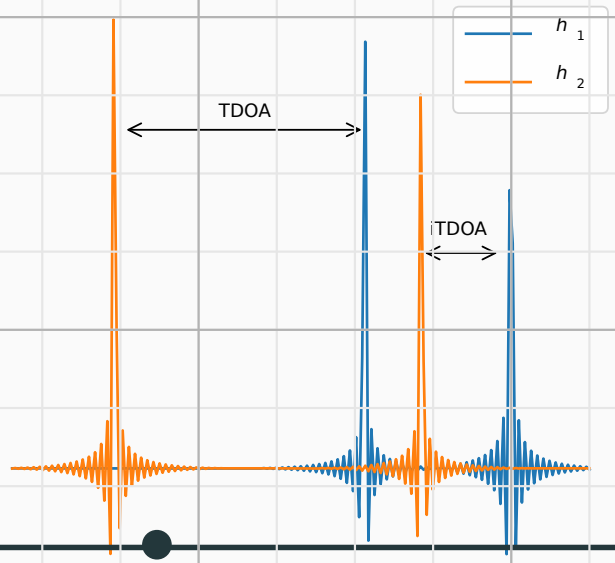


Each [illegible] is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

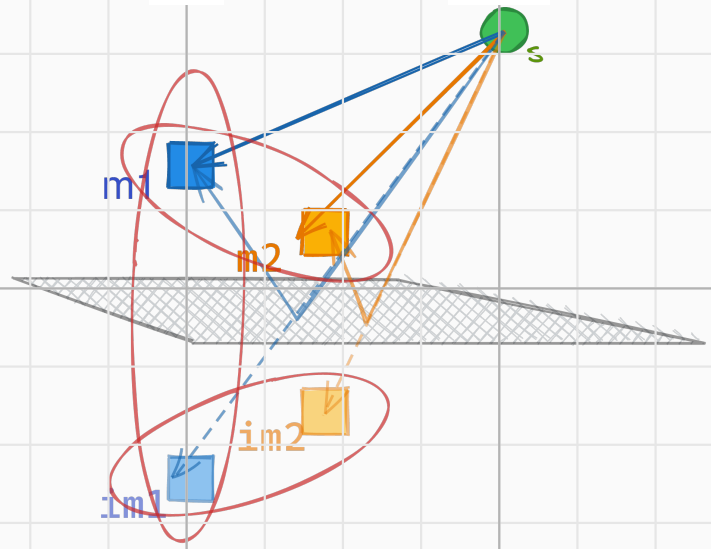
How t[illegible]ccess the *image* microphones?



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

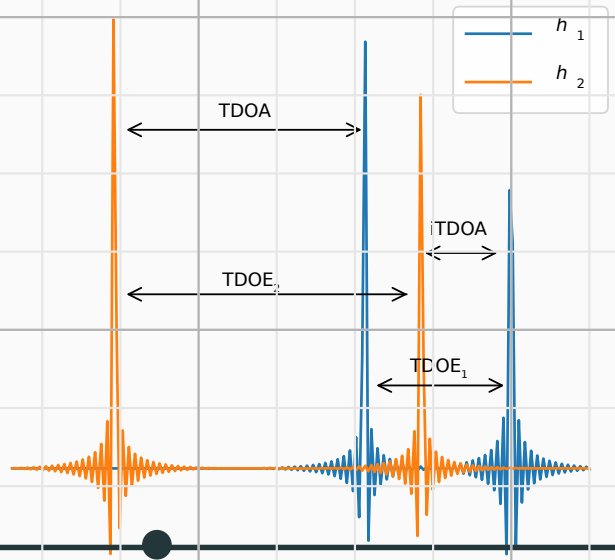


Each [illegible] is augmented with echoes

Mirage Array

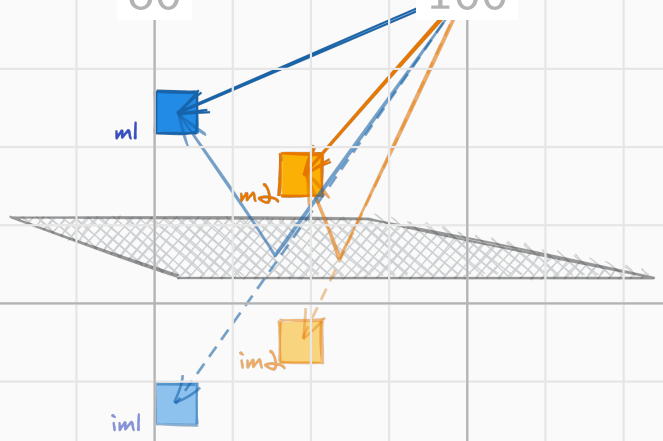
(Microphone Array Augmetation with Echoes)

How t[illegible]ccess 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 learning-based approach (`MLP`) for TDOAs estimation for each pair
2. fuse together the estimation ...
   - 2.1 use the error on a validation set as measure of uncertainty.
   - 2.2 DDDDDDOOOODOO DDOO BETTER HERE for each pair of the `Mirage` array (similar to SRP-PHAT[1]),
   - 2.3 knowing the position of the microphones,

Basel `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)

AO	Input	ACCURACY $\alpha < 10°$	ACCURACY $\alpha < 20°$
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 $\alpha < 10°$	ACCURACY $\alpha < 20°$
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)

AO	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
Mir	sp	63	82
GCC PHAT	sp	**82**	**97**

Observation

✓ co arable 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)

	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

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	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
- ✓ Takled "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
- ✓ Takled "impossible" localization
- ⚠ Performance depending on echo estimation methods (work in progress)

References i

mm o, M., Truc 40 A., and Del 60 , A. (2017). 80 100 120
Uncalibrated 3d room geometry estimation from sound impulse responses.
Journal of the Franklin Institute, 354(18):8678–8709.

DiBiase, J. H., Silverman, H. F., and Brandstein, M. S. (2001).
Robust localization in reverberant rooms.
In 40 *icrophone 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.

E n, J., Gaubitch, N. D., Moore, A. H., and Naylor, P. A. (2015).
60
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).
A stic slam.
80
IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(9):1484–1498.

References ii

Elko, G. W., Sondhi, M. M., and Jan, E.-E. (1993).
Spatially selective sound capture for speech and audio processing.
Speech Communication, 13(1-2):207–222.

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

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

Kowalczyk, K. (2019).
Raking early reflection signals for late reverberation and noise reduction.
The Journal of the Acoustical Society of America, 145(3):EL257–EL263.

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.

References iii

mm iff, H. (201

***Room acoustics*.**
CRC Press.

Leglaive, S., Badeau, R., and Richard, G. (2016).
Multichannel audio source separation with probabilistic reverberation priors.
I /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 *07 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 106–109. IEEE.

Ribeiro, F., Ba, D., Zhang, C., and Florêncio, D. (2010).
Turning enemies into friends: Using reflections to improve sound source localization.
In *010 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.