Echo-aware signal processing for audio scene analysis

Diego Di Carlo

November 30, 2020

PhD supervisors: Antoine Deleforge
Nancy Bertin

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

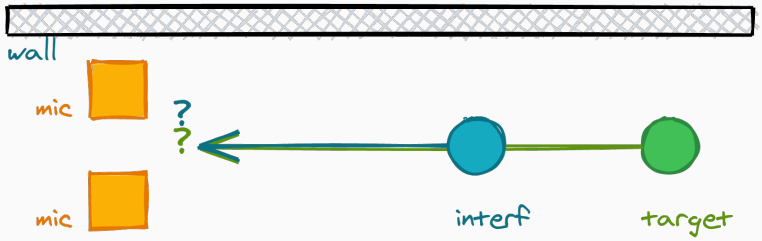
Université de Rennes 1, IRISA/INRIA, Panama research group

Echo-aware Application

Echo-aware Application

Echoes = same content, different time/direction

Image Source Model
⇔
Image Microphone Mode



Recent 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

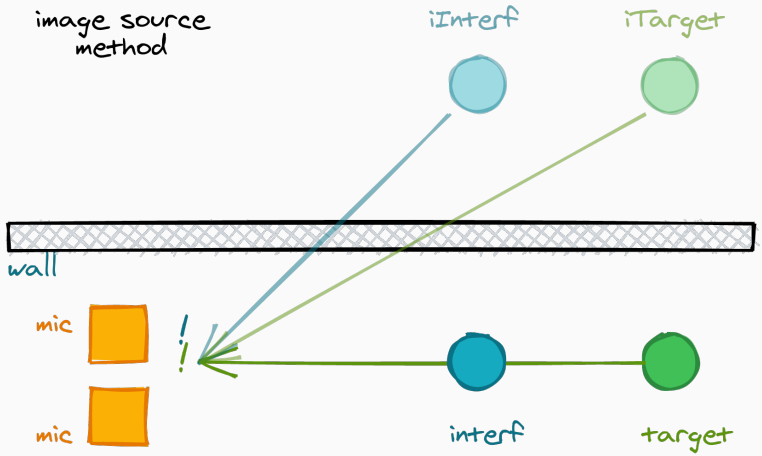


Image Source Model
⇔
Image Microphone Mode

Recent 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

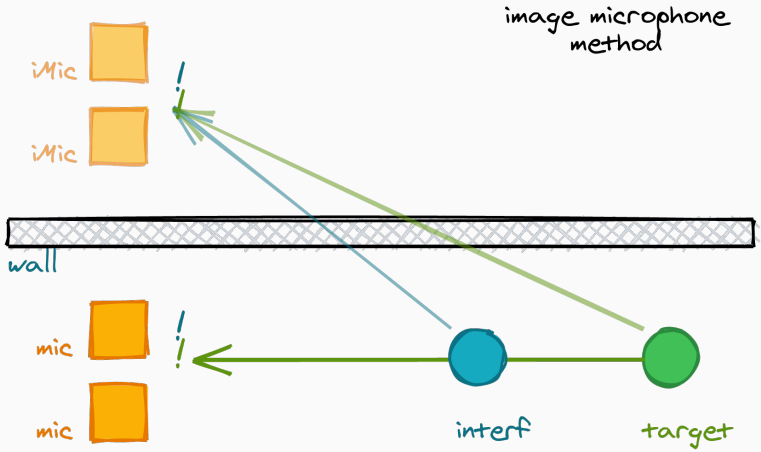


Image Source Model
⇔
Image Microphone Mode

Recent 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

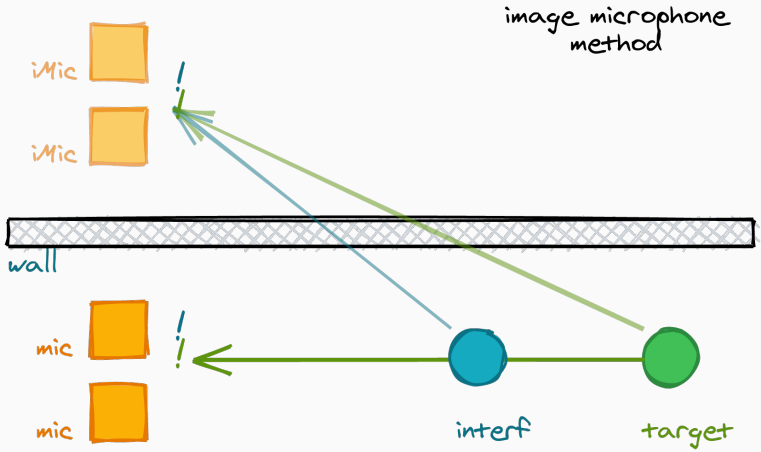


Image Source Model
⇔
Image Microphone Mode

Recent 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

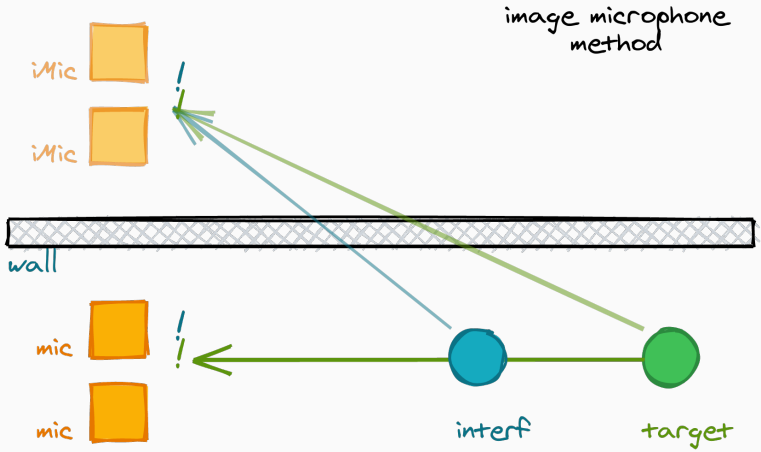


Image Source Model
⇔
Image Microphone Mode

Recent 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

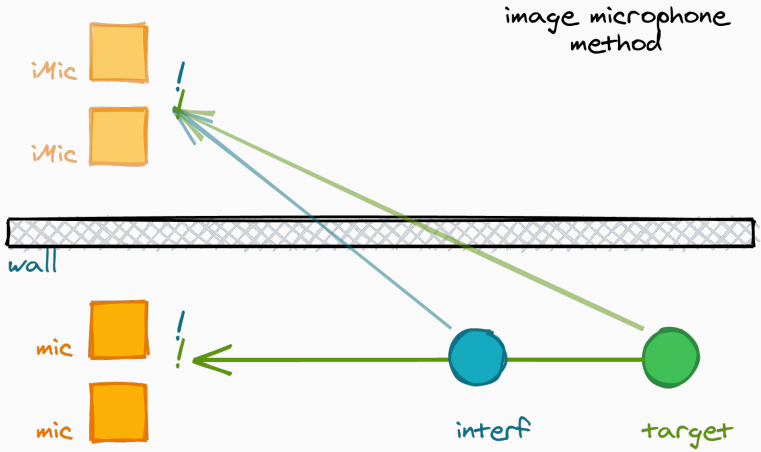


Image Source Model
⇔
Image Microphone Mode

Recent 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

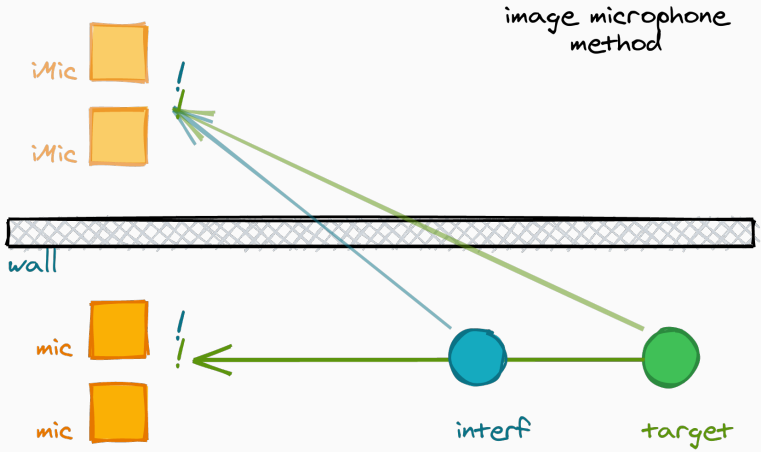


Image Source Model
⇔
Image Microphone Mode

Recent 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

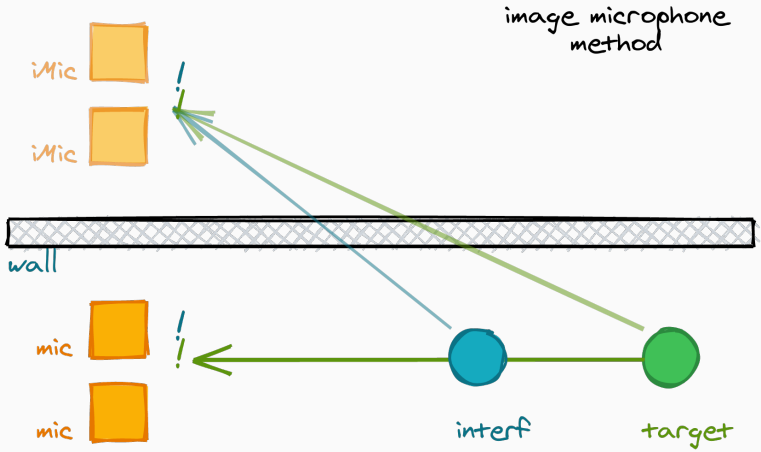


Image Source Model
⇔
Image Microphone Mode

Recent 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]

1

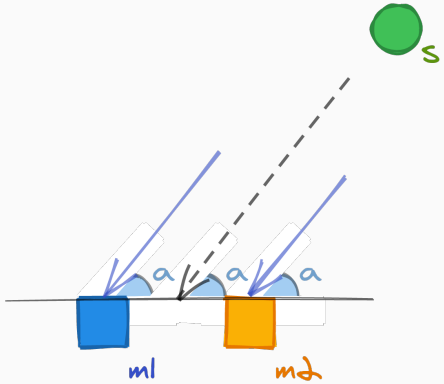
Sound Source Localization (SSL)

(common knowledge)

SSL $\rightarrow$ 3D position of sound source

SSL with 2 microphones

- Only angle of arrival (AOA)
- can be approximated from TDOAusing e.g.GCC-PHAT
  [Knapp and Carter, 1976]
  (known limitation, but good in practice)



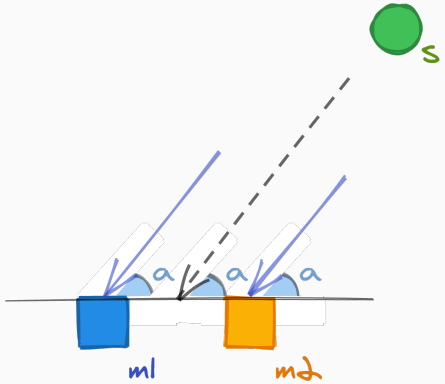
Sound Source Localization (SSL)

(common knowledge)

SSL $\rightarrow$ 3D position of sound source

SSL with 2 microphones

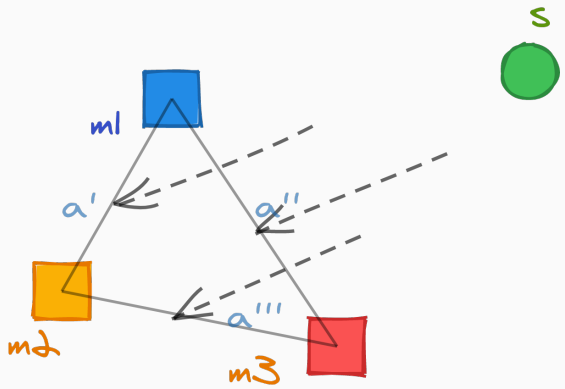
- Only angle of arrival (AOA) ⬈
- can be approximated from TDOAusing e.g.GCC-PHAT [Knapp and Carter, 1976]
  (known limitation, but good in practice)



SSL with more microphones

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

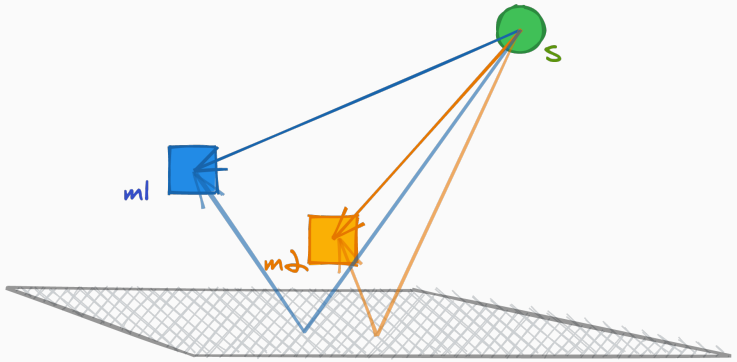
1. AOA for each pair can be "fuse" together (e.g. angular spectra in SRP-PHAT [DiBiase et al., 2001])
   (known limitation, but good in practice)



Sound Source Localization with Echoes

The Picnic Scenario:

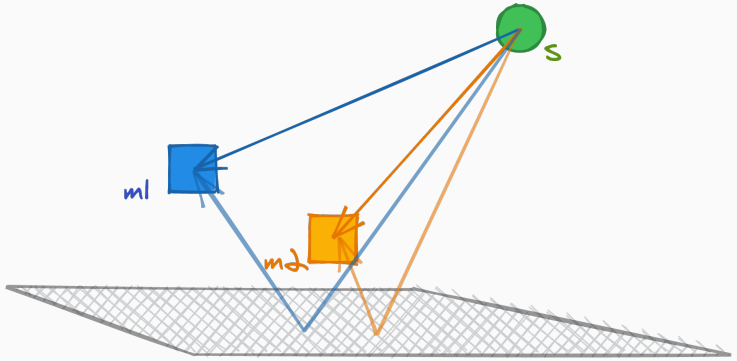
- One source
- Two microphones
  - → passive scenario
  - → generalizable to any array geometry



Sound Source Localization with Echoes

The Picnic Scenario:

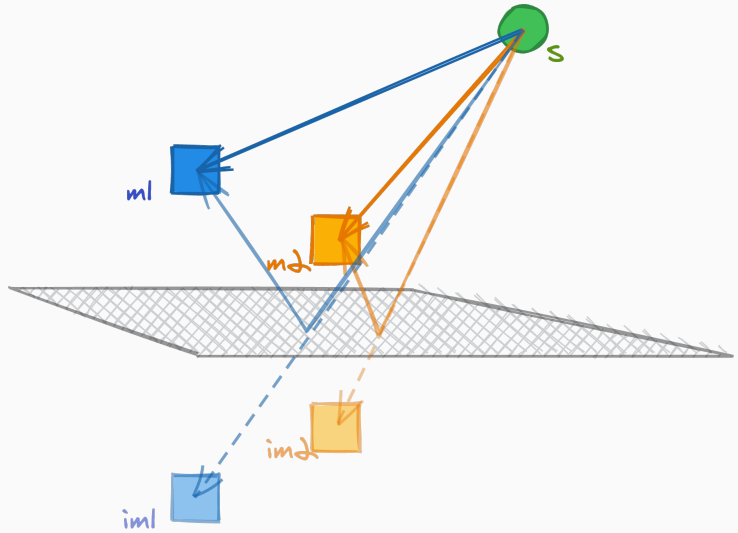
- One source
- Two microphones
  - → passive scenario
  - → generalizable to any array geometry
- Close to a very reflective surface
  - → First echo = Strongest echo
  - → α_{picnic} const. $\forall f$
  - → table-top device



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

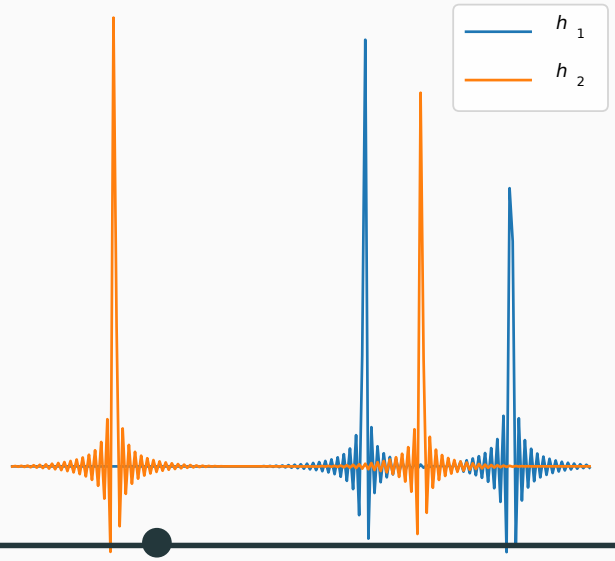


Each pair is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

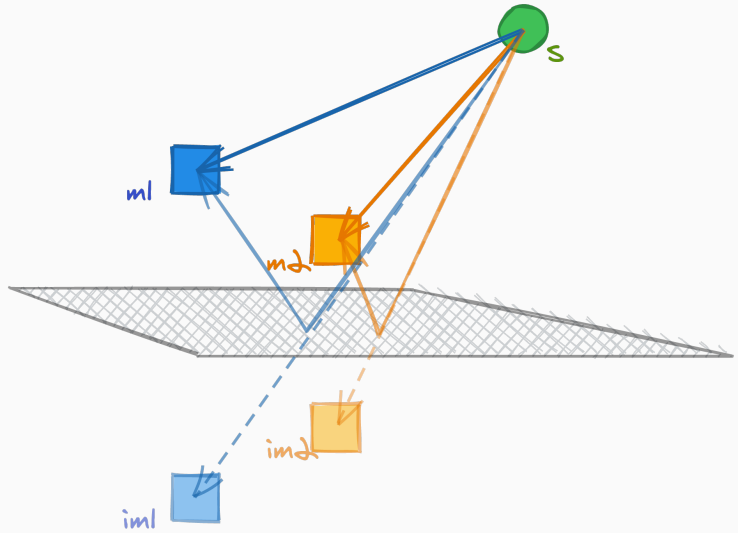
How to access the *image* microphones?



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

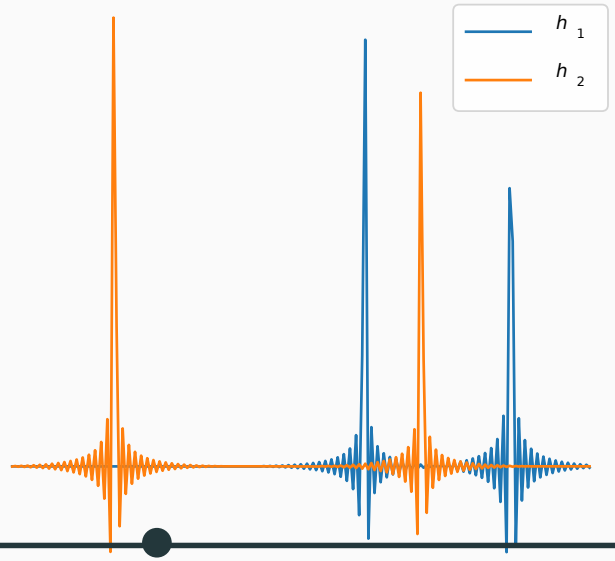


Each pair is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

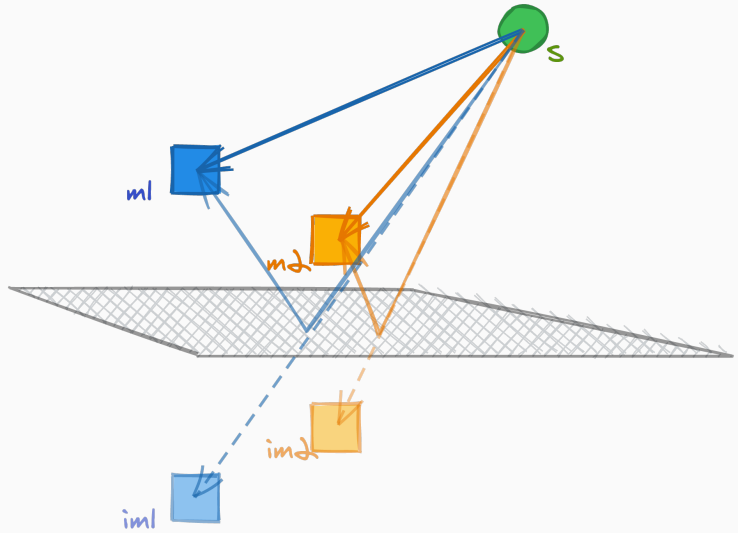
How to access the *image* microphones?



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

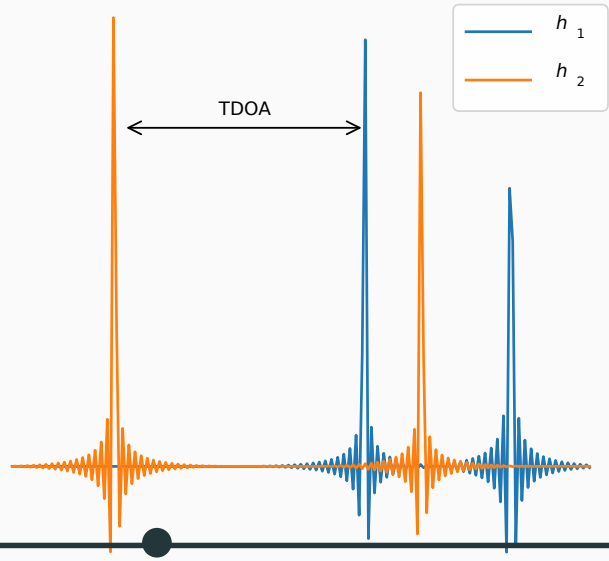


Each pair is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

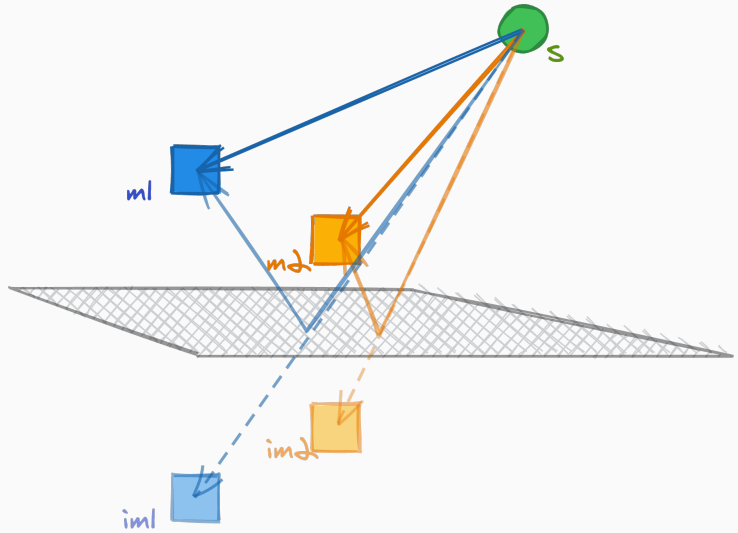
How to access the *image* microphones?



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

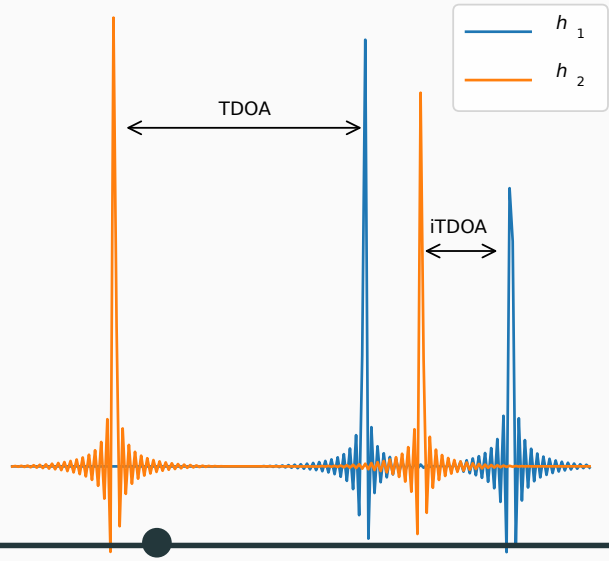


Each pair is augmented with echoes

Mirage Array

(Microphone Array Augmetation with Echoes)

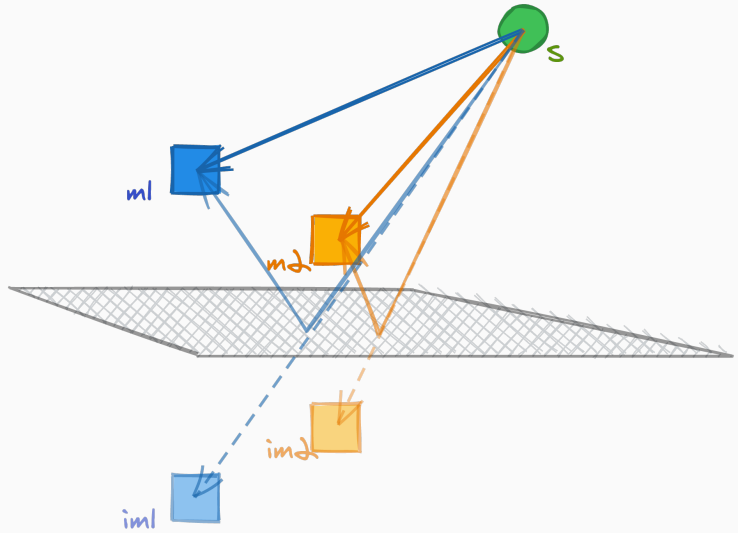
How to access the *image* microphones?



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

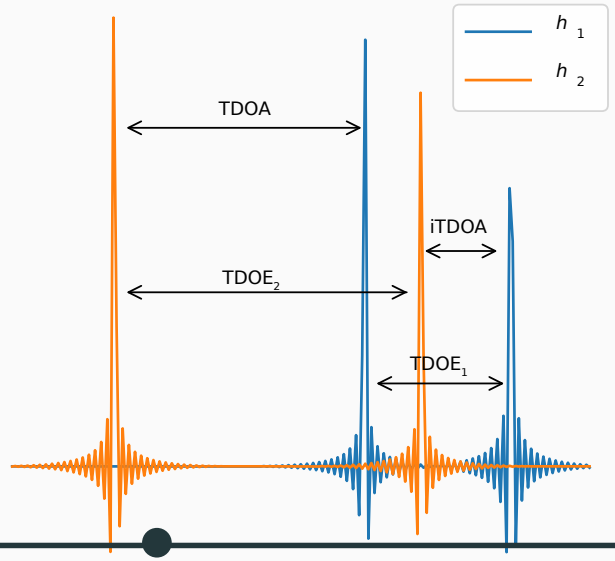


Each pair is augmented with echoes

Mirage Array

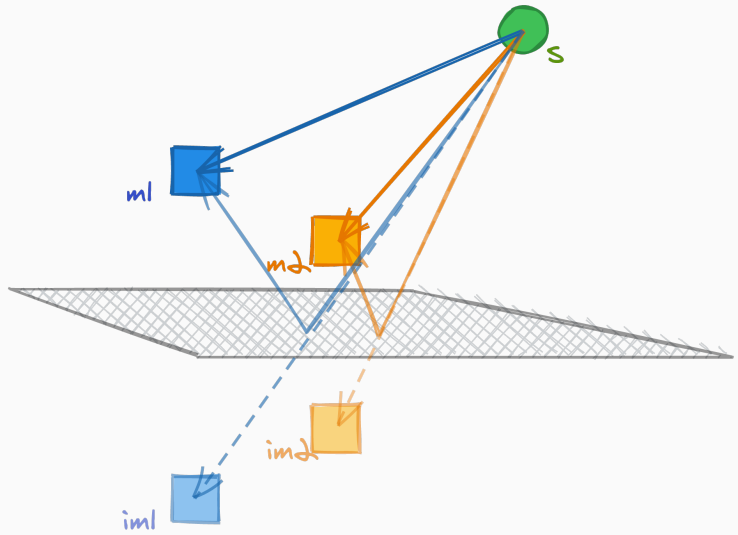
(Microphone Array Augmetation with Echoes)

How to access the *image* microphones?



Sound Source Localization with Echoes

Idea: perform DoA estimate with this.
Recall: these TDOAs are the same of the DNN-based method



Proposed Approach:

- use proposed DNN-based TDOA estimation
- fuse together the estimation ...
  - knowing the position of the microphones;
  - use the error on a validation set as measure of uncertainty.

Baseline: GCC-PHAT on true microphones

Experimental results

Proposed: `MLP` with `Mirage` (mod of SRP-PHAT[1])

Baseline: `GCC PHAT`[2]

Data: virtually generated close dataset as for `Lantern`

Metric: Accuracy (<ang10, <20°) (↰ also error in the manuscript)

AOA estimation

- ✓ Similar when wn
- ✗ Huge drop when noise
- ✗ Huge drop when speech and noise

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

2D SSL estimation
(both Az. and El.)

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

References i


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.


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.


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.

References ii


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.


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.


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.

References iii


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.


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.

References iv


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.