Hunting Echoes
for
Auditory Scene Analysis

Diego Di Carlo • 18.02.2020

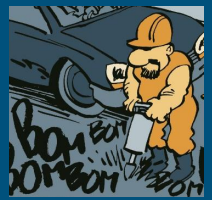
PhD Student in PANAMA at INIRA, Rennes (Fr)

Supervised by
Antoine DELEFORGE, Nancy BERTIN

Sound carries **information**...

Semantic information of sources

- Music
- Speech
- Noise



Temporal Information

- diarization/scheduling
- duration

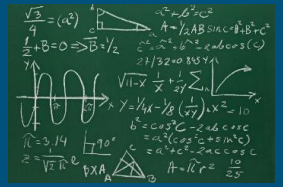
Spatial information due to *sound propagation*

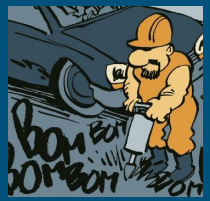
- Source/Sensors positions
- Type of environment

An auditory scene is all of this

Courtesy of [Deleforge17]

Analysis problems





Typical Audio **Inverse** (Analysis) **Problems**

- Automatic Speech Recognition
- Music Information Retrieval (MIR)
- Diarization
- Speech enhancement
- Denoising

- Sound Source Separation
- Sound Source Localization
- Acoustic Measurements
- Room Geometry Estimation
- Signal restoration and inpainting
- And many others

Courtesy of [Deleforge17]

Hunting Echoes
for
Auditory Scene Analysis

[OUTLINE]

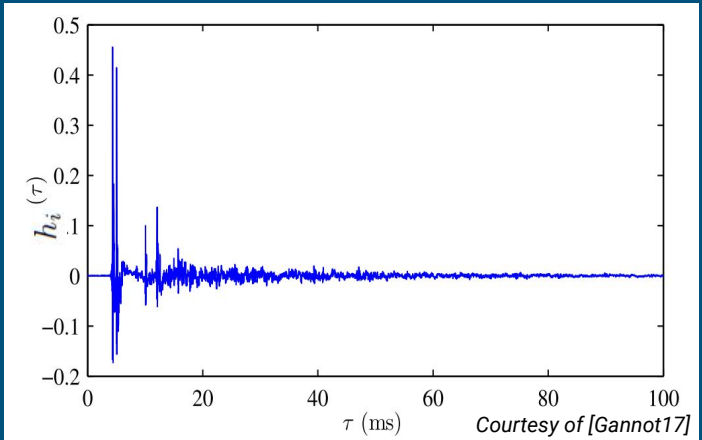
1. Definitions: What are the echoes?
2. Applications: What to do with them?
   a. *Sound Source Localization and* **MIRAGE**
      Presented at ICASSP19
   b. *Sound Source Separation and* **SEPARAKE**
      Presented at ICASSP18
3. Estimations: How to know them?
   a. *Learning-based* approach and **MIRAGE**
      Presented at ICASSP19
   b. *Continuous Dictionary* and **BLASTER**
      Accepted at ICASSP20
4. Real World data collection: How to play with them
   a. ***PicNic of the MUSIS*** *dataset*
   b. ***dEchoerate*** *dataset*

Acoustic Impulse Responses

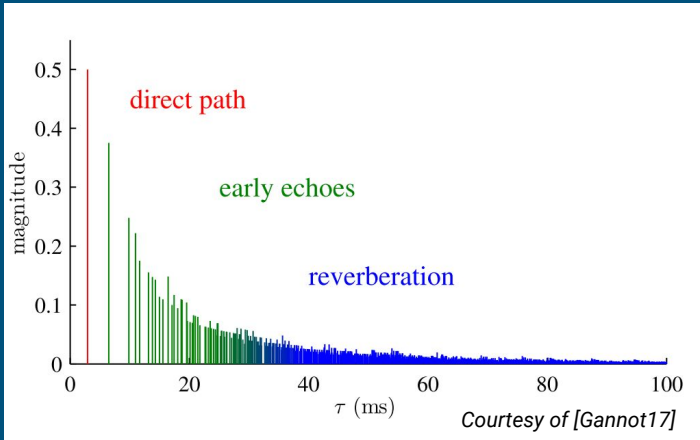
Acoustic Impulse Response (AIR):

the linear filtering effect due to the propagation of sound from a source to a microphone.

$$x_i(t) = (h_i * s)(t) + n_i(t), \quad i = 1, 2, \ldots, I$$



Courtesy of [Gannot17]



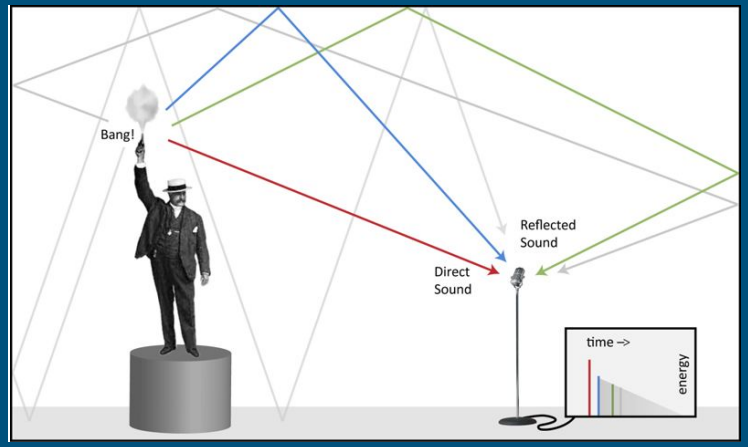
Courtesy of [Gannot17]

Acoustic Impulse Responses

Room Impulse Response (RIR),

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

$$x_i(t) = (h_i * s)(t) + n_i(t), \quad i = 1, 2, \ldots, I$$



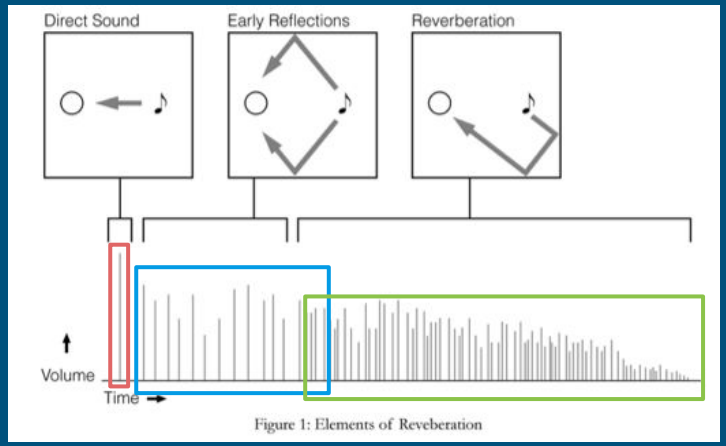


Figure 1: Elements of Reveberation

Acoustic Impulse Responses

AIR (or RIR) can be subdivided in:

- Direct (or anechoic) path
- Early reflections (**Echoes**)
- Late Reverberation

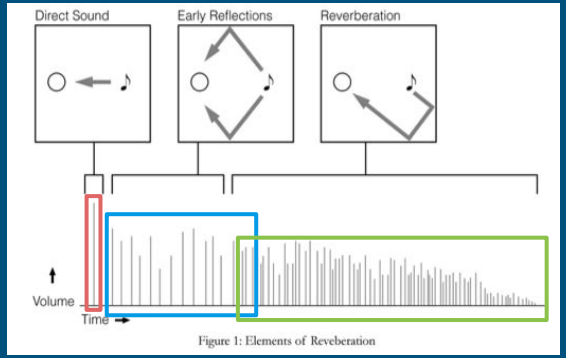


Figure 1: Elements of Reveberation

Acoustic Impulse Responses

AIR (or RIR) can be subdivided in:

- Direct (or anechoic) path
- Early reflections (**Echoes**)
- Late Reverberation

RIR can be modeled with the Image Method

$$h_i(t) = \sum_{r=0}^{R_i} \frac{\alpha_{i,r}}{4\pi c_{\text{air}} \tau_{i,r}} \delta(t - \tau_{i,r})$$

$$\tau_{i,r} = \|\mathbf{r}_i - \mathbf{s}_r\| / c_{\text{air}}$$

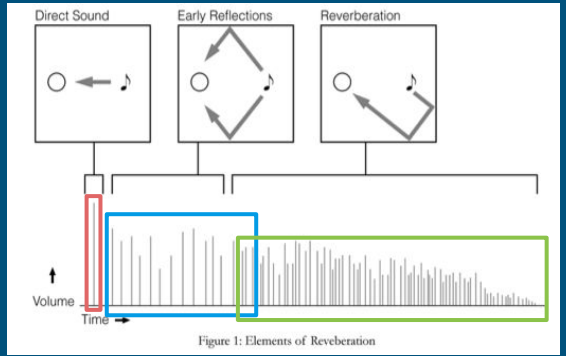
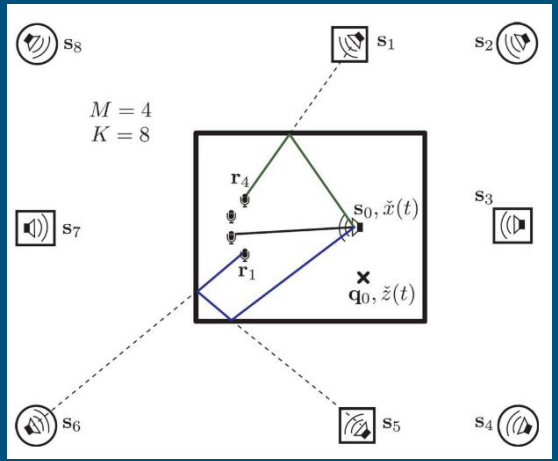


Figure 1: Elements of Reveberation



Courtesy of [Sheibler15]

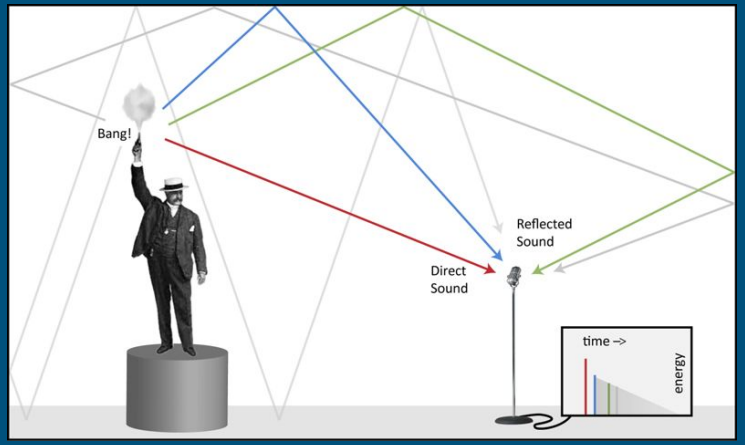
Acoustic Impulse Responses

Courtesy of https://www.cs.princeton.edu/~funk/acoustics.html

Acoustic Impulse Responses

RIRs account for ...

- … the geometry of the audio scene:
  - Room shape and size
  - Source position
  - Microphone position
  - … other objects (e.g. furnitures) sizes and shape.
- … the acoustic properties of the audio scene:
  - surfaces materials of walls or furnitures



Acoustic Impulse Responses

Estimating the full sound propagation is a **very difficult** task

For some DSP problems, the sound propagation is typically…

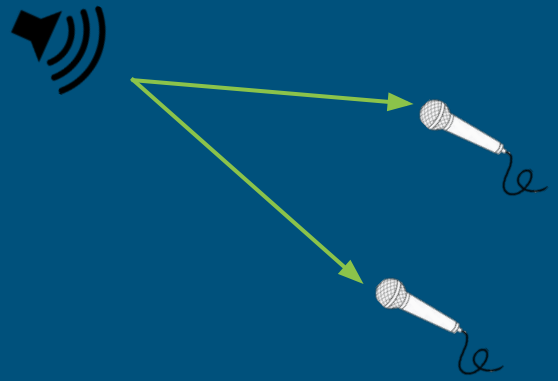
- … ignored *[Praetzlich et al. 2015, Le Roux et al. 2015]*

Acoustic Impulse Responses

Estimating the full sound propagation is a **very difficult** task.

For some DSP problems, the sound propagation is typically…

- … ignored *[Praetzlich et al. 2015, Le Roux et al. 2015]*;
- … assumed as a single anechoic path *[DiBiase 2010, Rickard 2007]*;

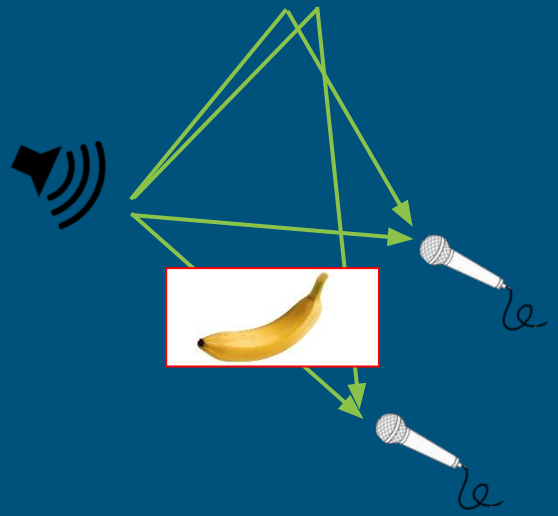


Acoustic Impulse Responses

Estimating the full sound propagation is a **very difficult** task.

For some DSP problems, the sound propagation is typically…

- … ignored *[Praetzlich et al. 2015, Le Roux et al. 2015]*;
- … assumed as a single anechoic path *[DiBiase 2010, Rickard 2007]*;
- … modelled entirely *[Ozerov et al. 2010, Nugraha et al. 2016]*;

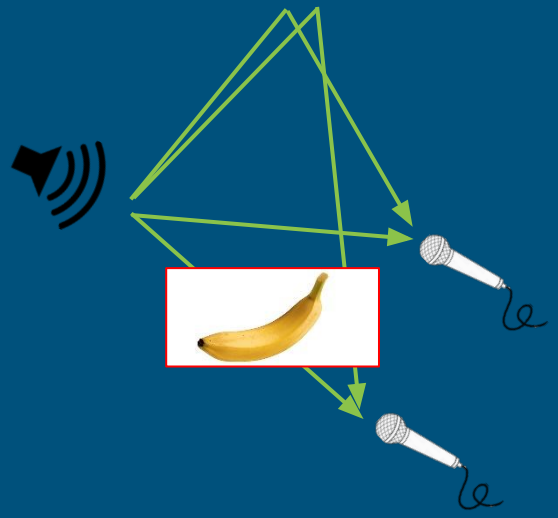


Acoustic Impulse Responses

Estimating the full sound propagation is a **very difficult** task.

For some DSP problems, the sound propagation is typically…

- … ignored *[Praetzlich et al. 2015, Le Roux et al. 2015]*;
- … assumed as a single anechoic path *[DiBiase 2010, Rickard 2007]*;
- … modelled entirely *[Ozerov et al. 2010, Nugraha et al. 2016]*;



Moreover, **strong reverberation** and **strong early reflections**

- detrimentally affect typical **Audio Inverse Problem** algorithm
- are usually modelled as **noise**, as something to reduce

If we know them?

Courtesy of [PinkFloyd70]

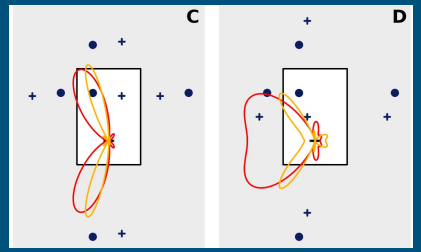
Can Echoes help?

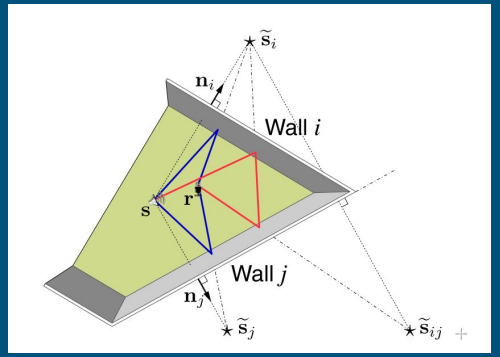
Recents **echo-aware** methods showed that the knowledge of early echoes increases their performances.

HP : Let us assume the echoes' coefficients and locations given

Echo-aware Inverse Problems

- Speech enhancement
  - Sound Source Separation
    - *[Scheibler et al. 18, Leglaive et al. 16]*
  - Beamforming
    - *[Dockmanic et al. 15, Flanagan et al. 93]*
- Source Localization
  - *[An et al. 18, Salvati et al. 16, Riberio et al. 10]*
- Microphone calibration
  - *[Salvati et al. 16, Dockmanic et al. 13]*
- Dereverberation and Room Equalization
  - *[Krishnan et al. 2018]*
- Room Geometry Estimation
  - *[Crocco et al. 16, Dockmanic et al. 13, ...]*

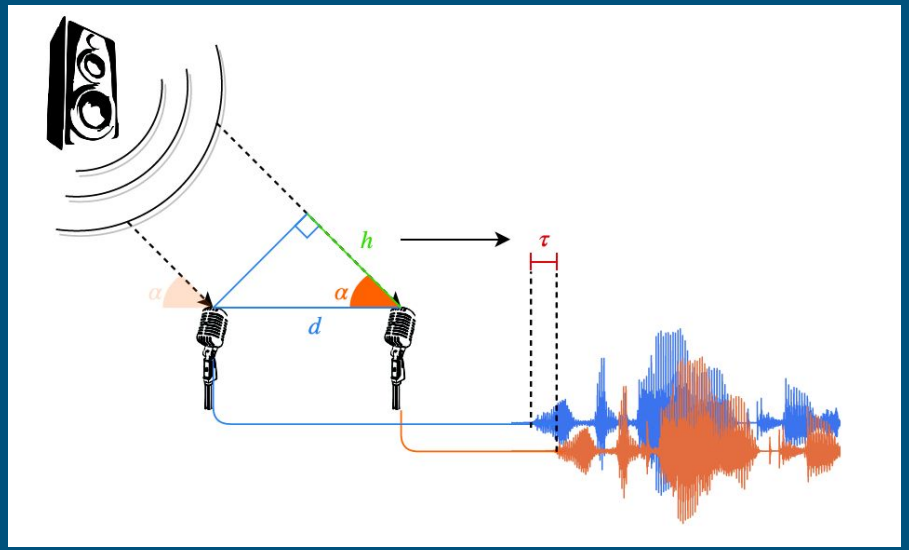




Courtesy of [Sheibler15, Dockmainc13]

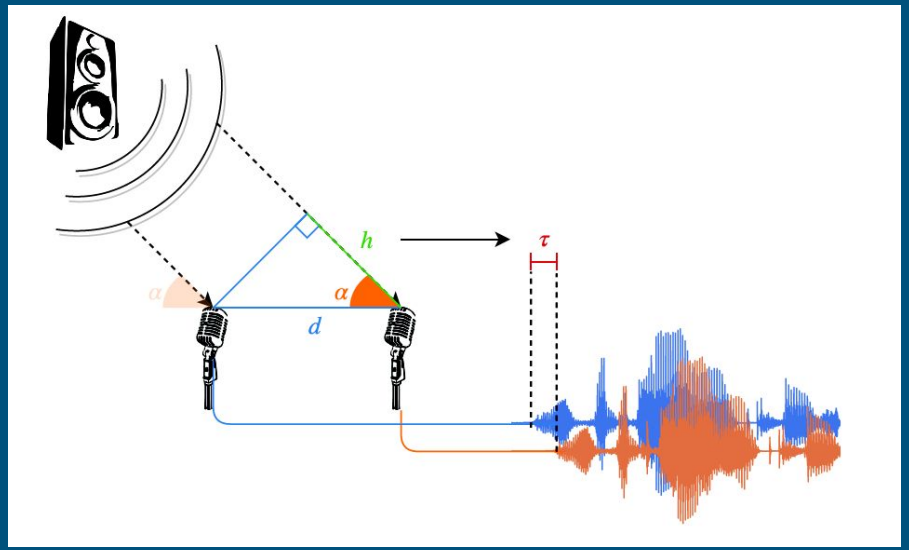
Background in Sound Source Localization

- SSL for single pair
- Direction of Arrival/TDOA/1D SSL estimation



Background in Sound Source Localization

- SSL for single pair
- Direction of Arrival/TDOA/1D SSL estimation with GCC-PHAT



$$\alpha = \arccos \frac{\tau\ c}{d}$$

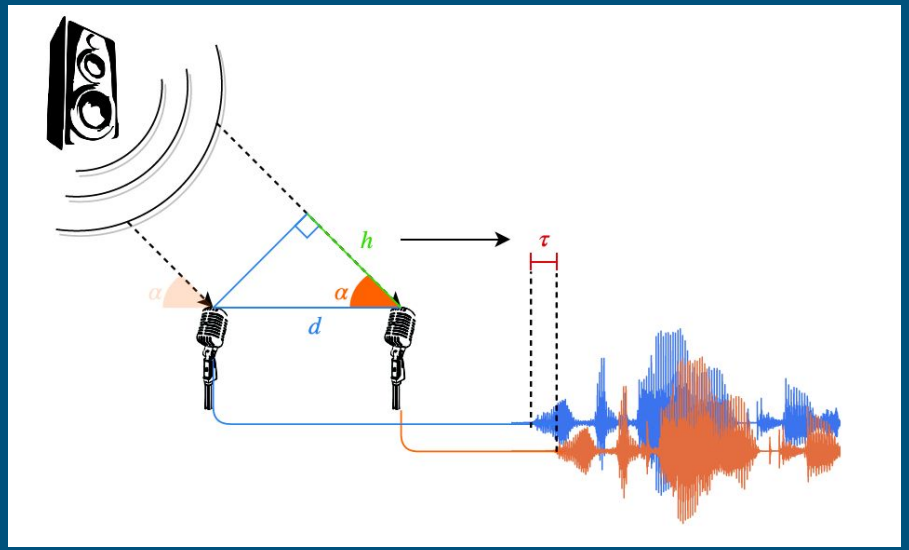
$$\tau = \arg\max_\tau\ (m_1 \text{✡} m_2)(\tau)$$

✡ is the *generalized cross correlation* function:

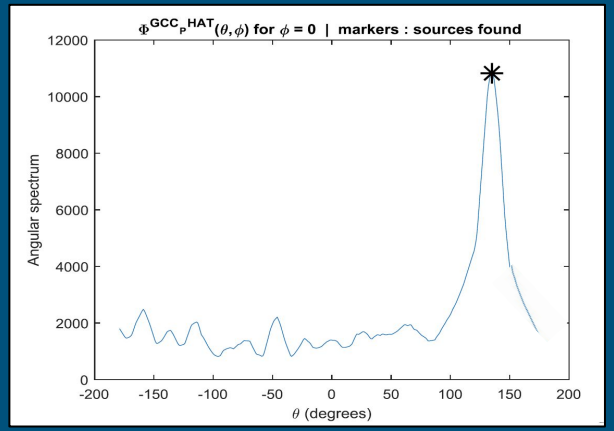
- CC + PHAT transform to remove source's autocorrelation

Background in Sound Source Localization

- SSL for single pair
- Direction of Arrival/TDOA/1D SSL estimation



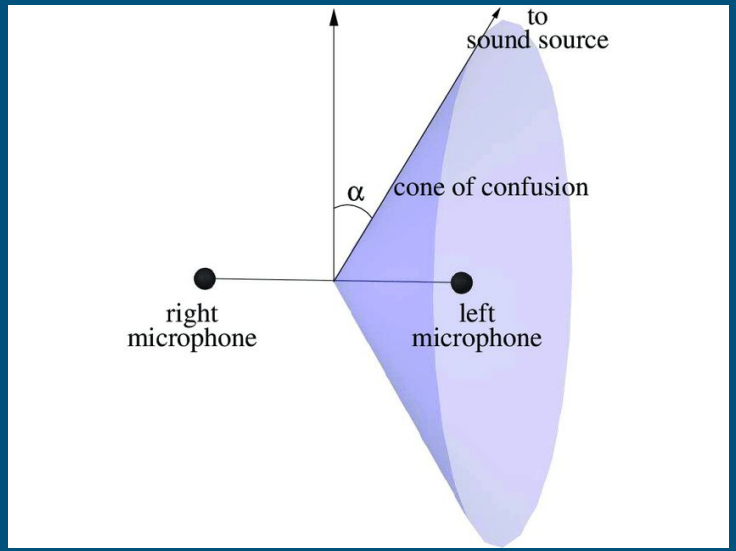
$$\tau = \arg\max_{\tau} \; (m_1 \star m_2)(\tau)$$

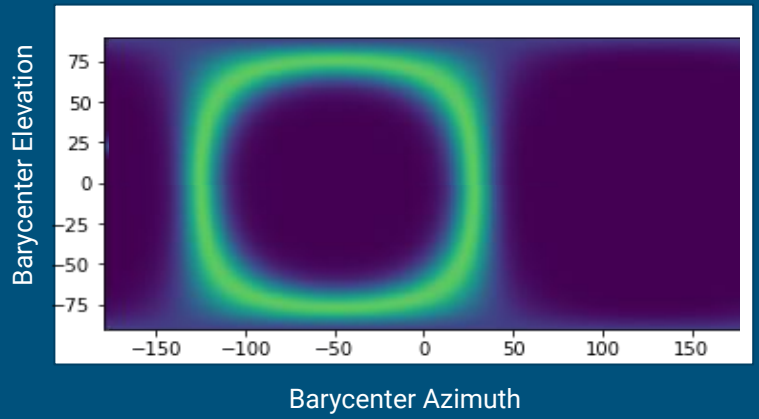


***Local** Angular Spectrum*

Background in Sound Source Localization

- SSL for single pair
- *If we know the geometry of the microphone array, we can define a **global coordinate system***

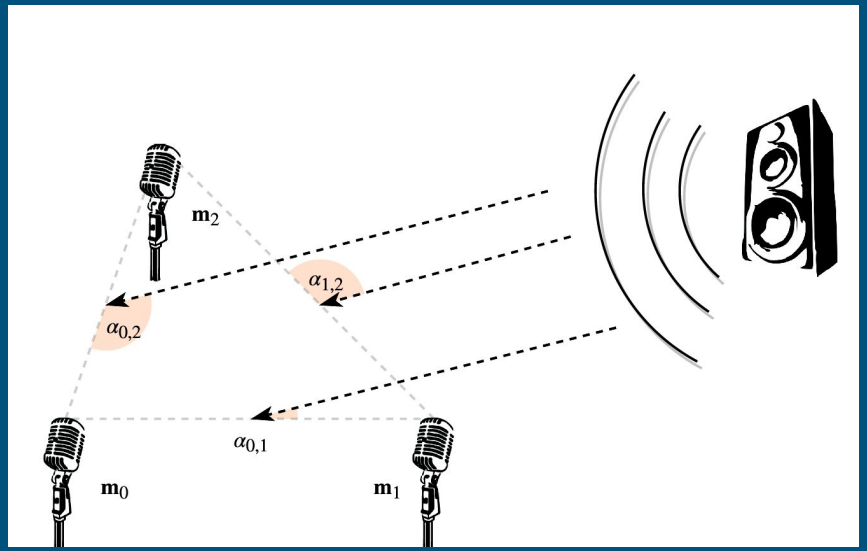




***Global** Angular Spectrum*

Background in Sound Source Localization

- SSL for Microphone Array - multiple pairs
- SRP-PHAT method as *Divide-et-Impera paradigm with GCC-PHAT at the leaves*

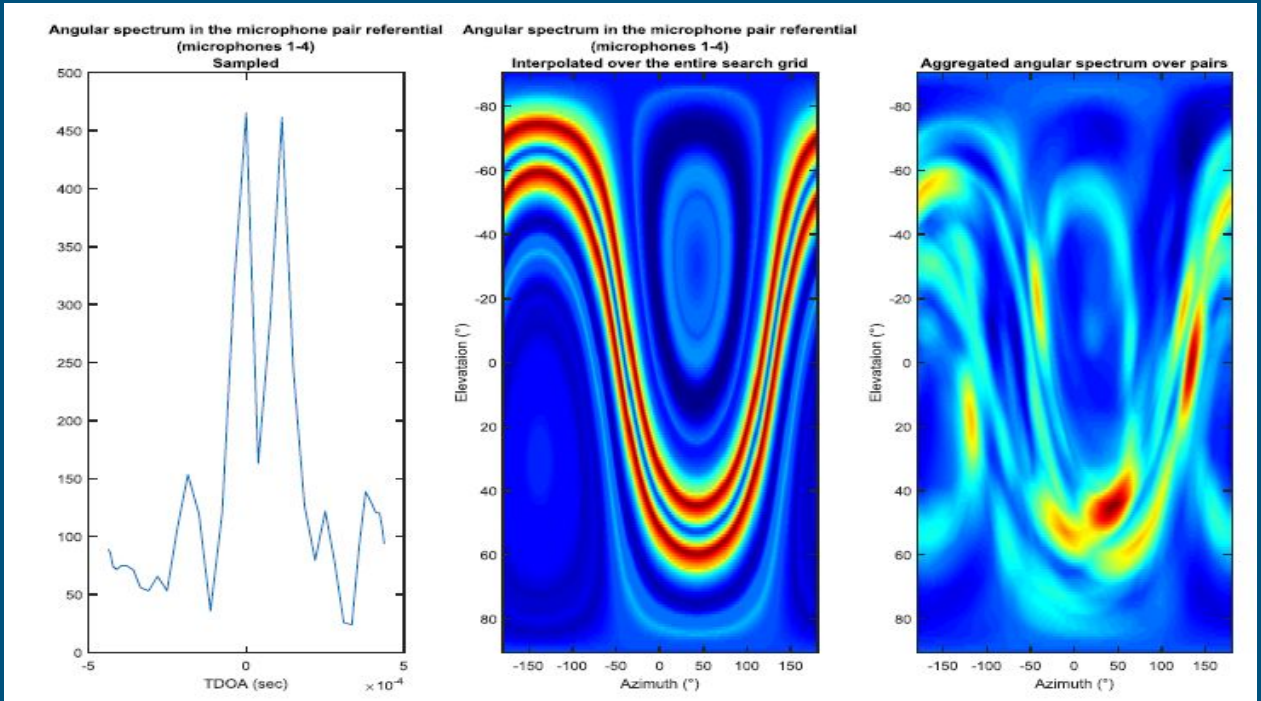


If we have more microphone?

1. Divide: we can define multiple pairs
   a. **local** coordinate system
   b. **global** coordinate system
2. Leaves: perform **1D SSL** for each pairs
3. Merging: aggregate all the local angular spectra in the global one
4. Impera: pull the **maximum** on the global 2D grid

[SRP-PHAT method, DiBiase et al, 2001]

Background in Sound Source Localization



SPR-PHAT method, *[DiBiase et al, 2001]*

Implemented in MBSS Locate, an open-source MATLAB implementation of SSL algorithm

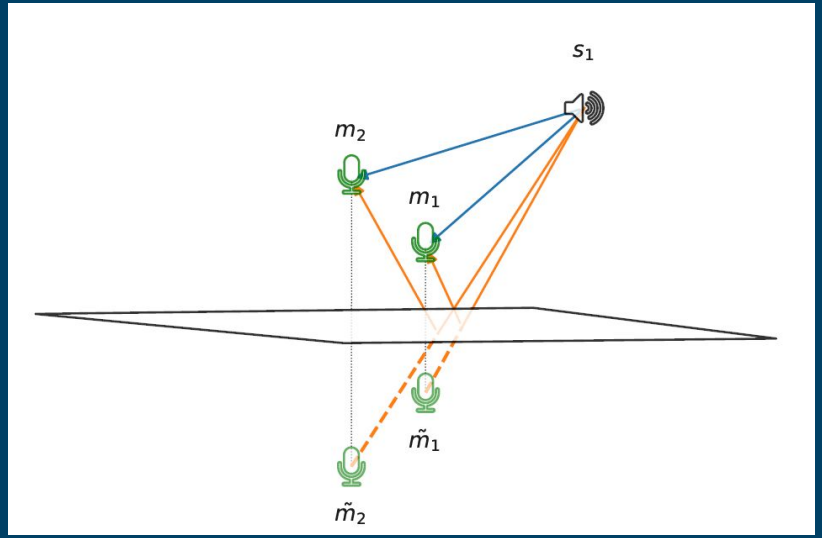
Available at:
http://bass-db.gforge.inria.fr/bss_locate/

MIRAGE

Microphones array augmentation with echoes

Presented at ICASSP 2019

[Diego Di Carlo, Antoine Deleforge, Nancy Bertin]

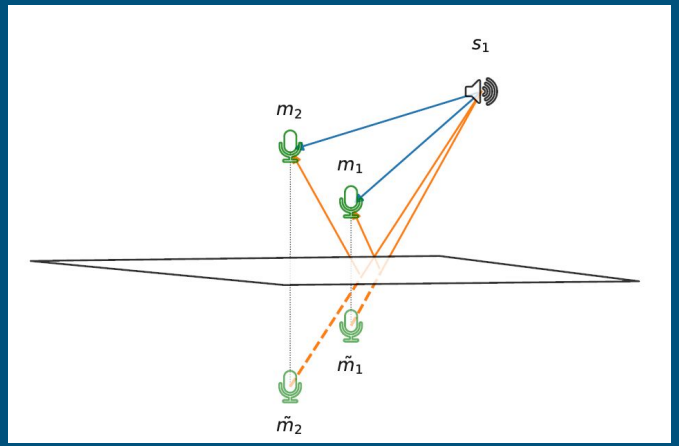


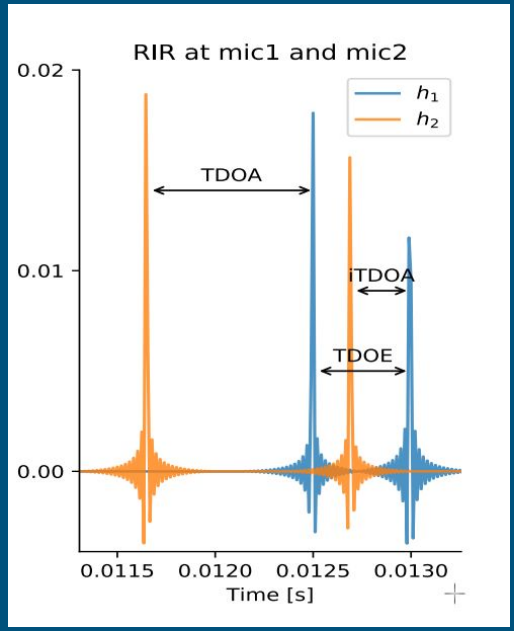
The signal received at m_1 *can be seen as the* ***sum*** *of anechoic signals received at* m_1 *and an* ***image microphone*** $\tilde{m}_1$

Image source -> Image microphones

Sound Source **Localization** *with a little help from echoes*

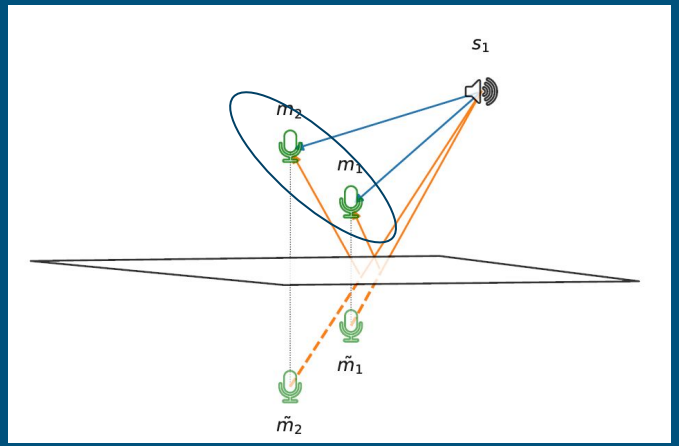
- More microphones… better audio signal processing!
- How to « access » image microphones?

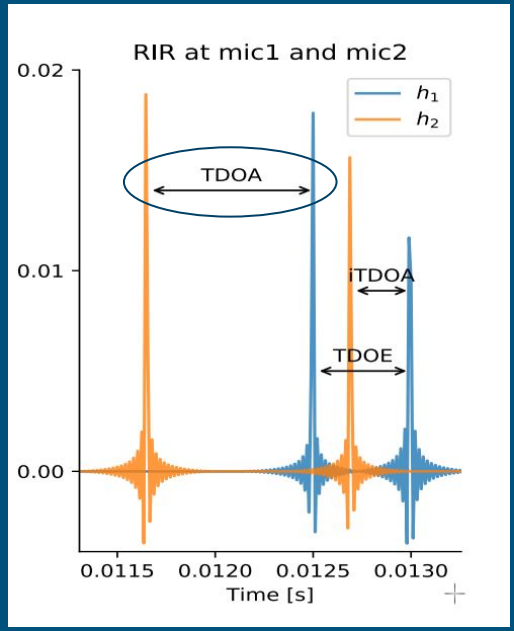




Sound Source **Localization** *with a little help from echoes*

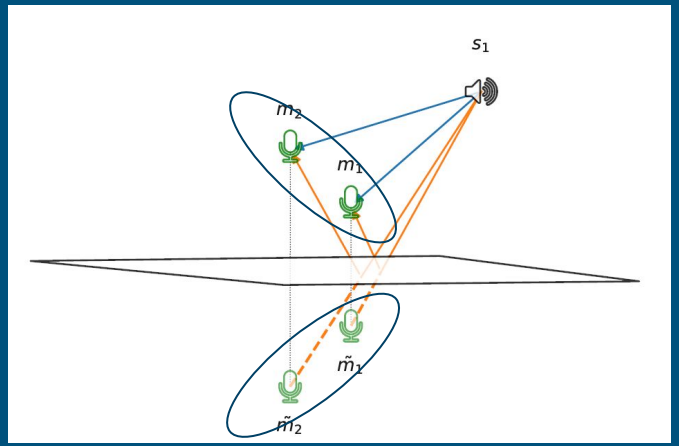
- More microphones… better audio signal processing!
- How to « access » image microphones?

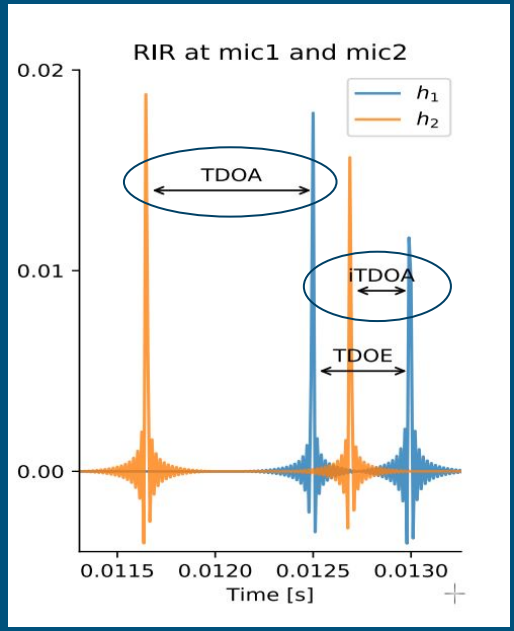




Sound Source **Localization** *with a little help from echoes*

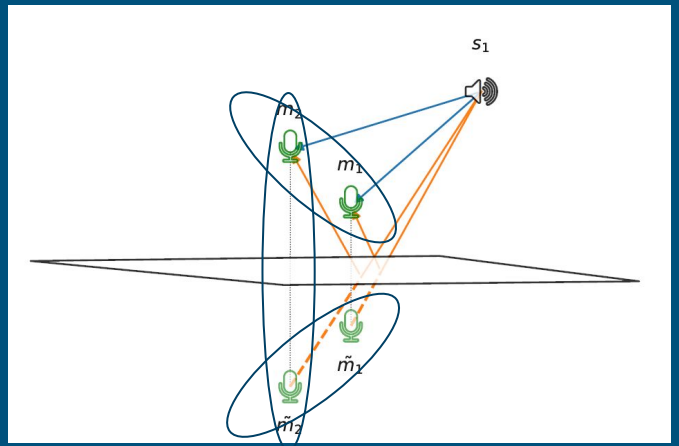
- More microphones… better audio signal processing!
- How to « access » image microphones?

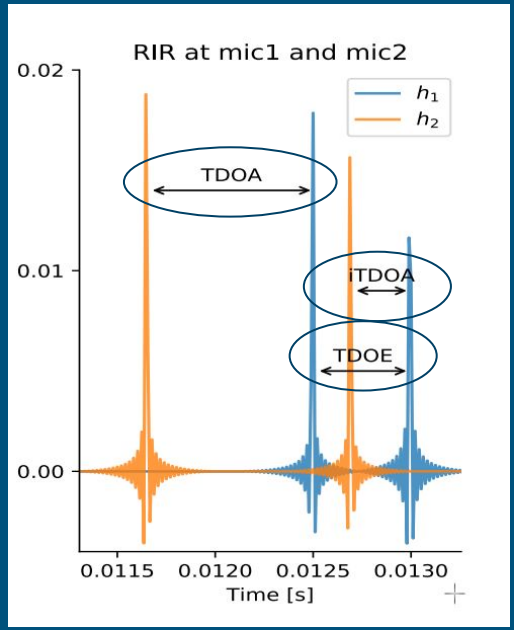




Sound Source **Localization** *with a little help from echoes*

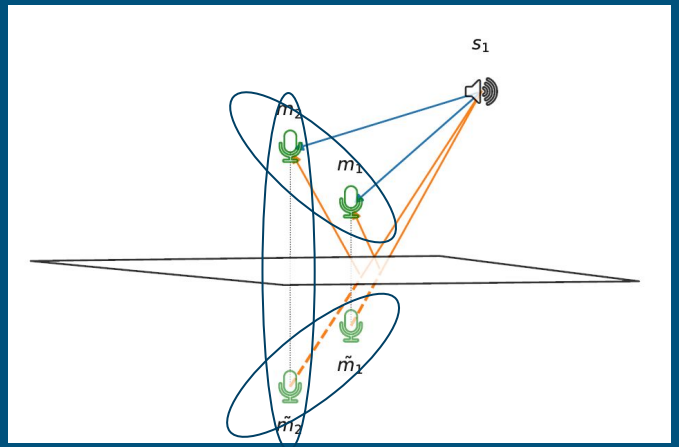
- More microphones… better audio signal processing!
- How to « access » image microphones?

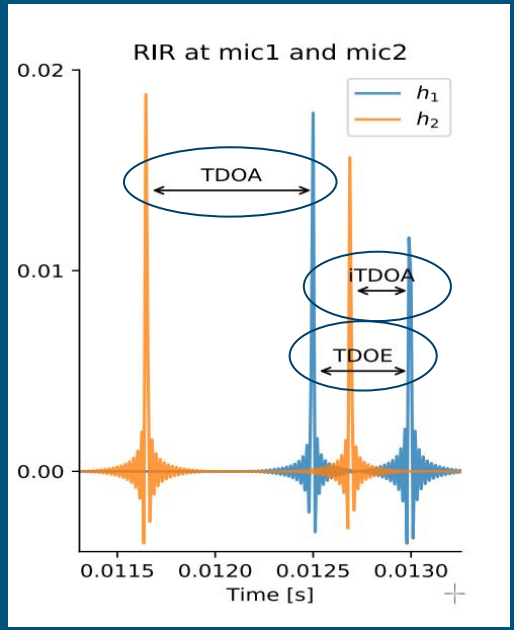




Sound Source **Localization** *with a little help from echoes*

- More microphones… better audio signal processing!
- How to « access » image microphones?

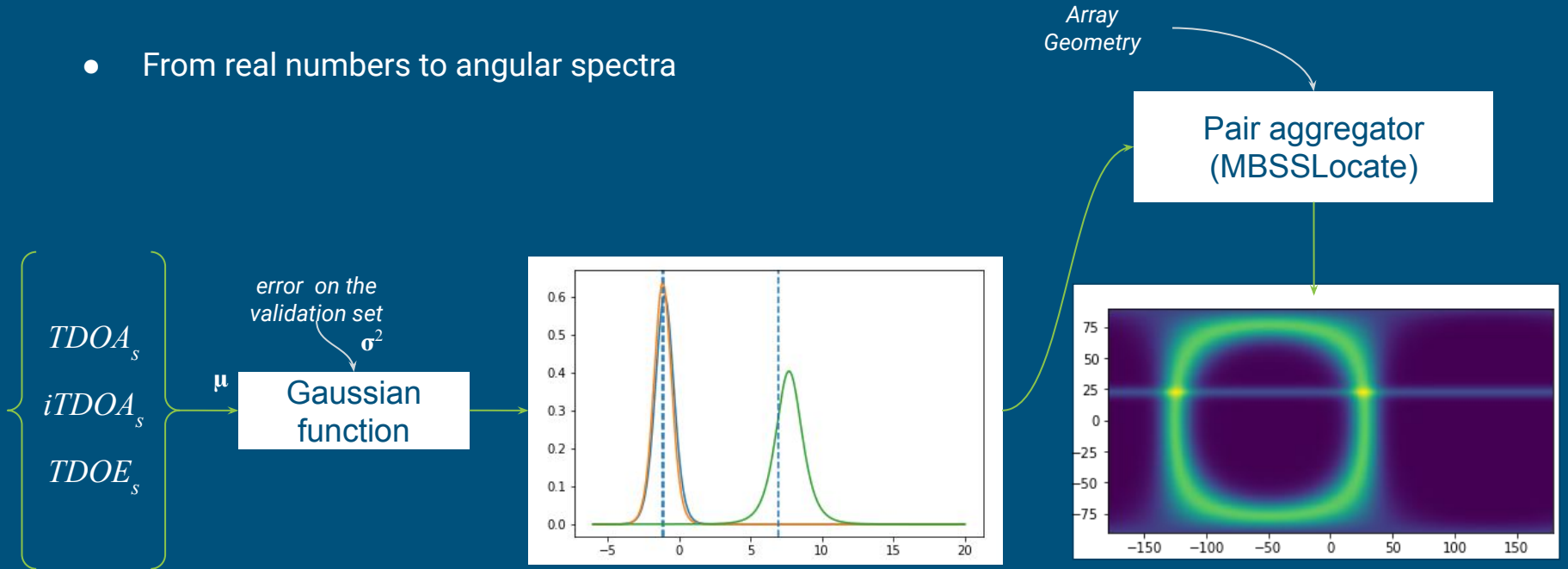




- Each pair in the **augmented array** is associated to impulse response characteristics

Sound Source **Localization** *with a little help from echoes*

- From real numbers to angular spectra

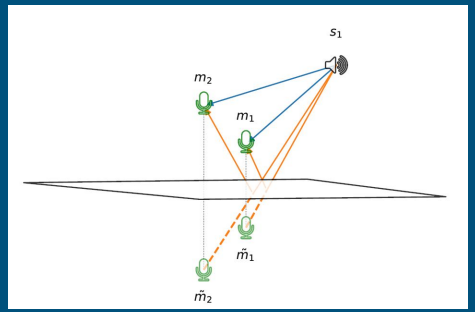


Sound Source **Localization** *with a little help from echoes*

- Aggregating (with MBSSLocate) time differences of arrival from multiple microphone pairs enables 2D sound source localization
- The microphone and surface positions are assumed known
- Promising «**impossible localization**» results using clean signals and white noise sources
- Future work:
  - ✓ Aggregating multiple pairs as in [*DiBiase 10*]
  - ✓ Define proper training dataset as in [*Google 18*]
  - ✓ Robust Regression Model
  - ✓ Test on real data
  - ❌ Perfect symmetries breaks the model

Results on test set
[ICASSP19]

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)

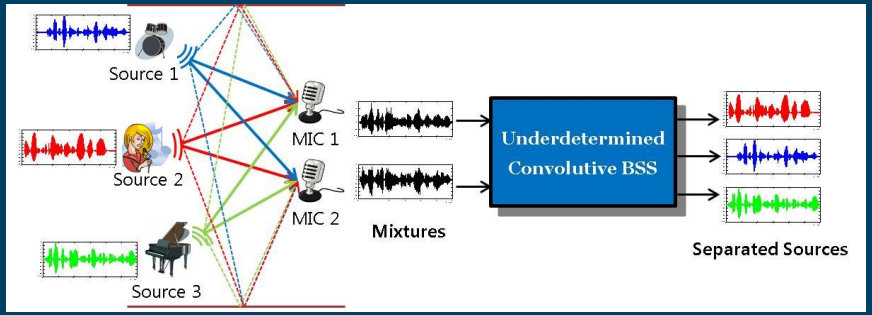


SEPARAKE

Sound source separation with a little help from echoes

Presented at ICASSP 2018

[Robin Scheibler, Diego Di Carlo, Antoine Deleforge, Ivan Dokmanic]

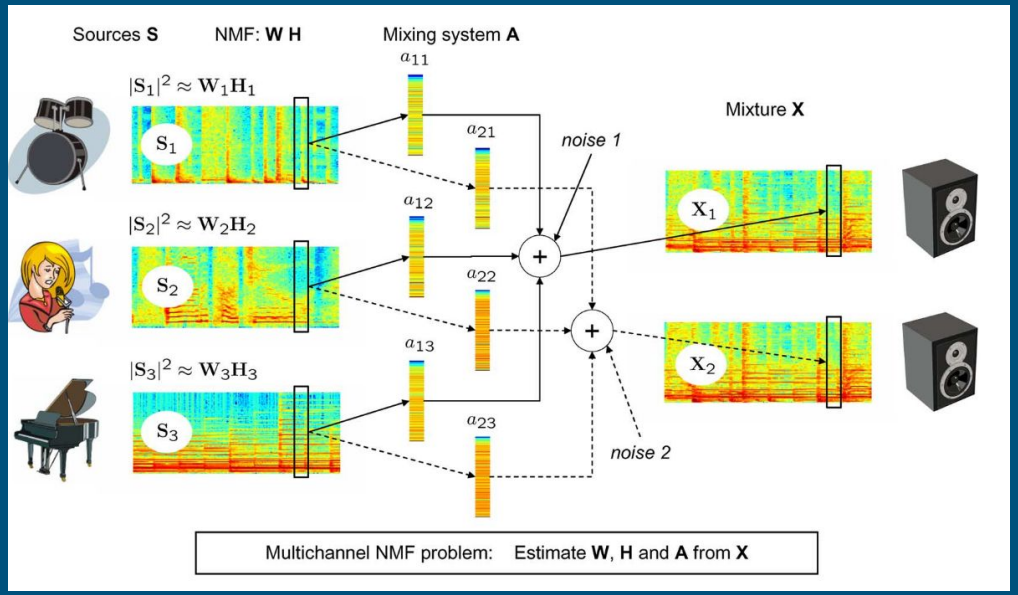


Courtesy of slsp.kaist.ac.kr

Model the source's acoustic propagations using known early echoes in NMF-based sound source separation algorithm

Sound Source **Separation** *with a little help from echoes*

Multichannel Audio Source Separation [*Ozerov et Fevotte, 2010*]:



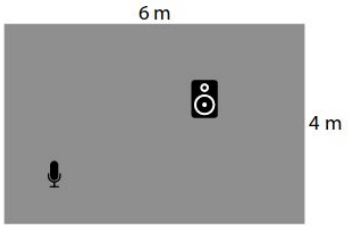
Courtesy of [Ozerov et al. 10]

Parameters can be estimated efficiently with the *Expectation-Minimization algorithm*

- *However it is a **non-convex optimization problem***
- *Good performances when a **good initialization point** is provided*
  - *E.g. pre-train a dictionary of template from a database of speech utterances*

Sound Source **Separation** *with a little help from echoes*

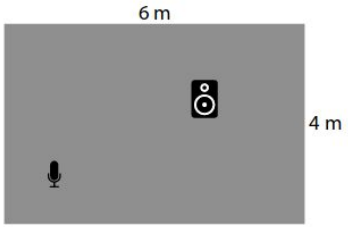
$$H_i(f) = \sum_{r=0}^{R_i} \frac{\alpha_{i,r}}{4\pi \|\mathbf{r}_i - \mathbf{s}\|} e^{-j2\pi f \|\mathbf{r}_i - \mathbf{s}\|/c_{\text{air}}} + \varepsilon_i(f)$$

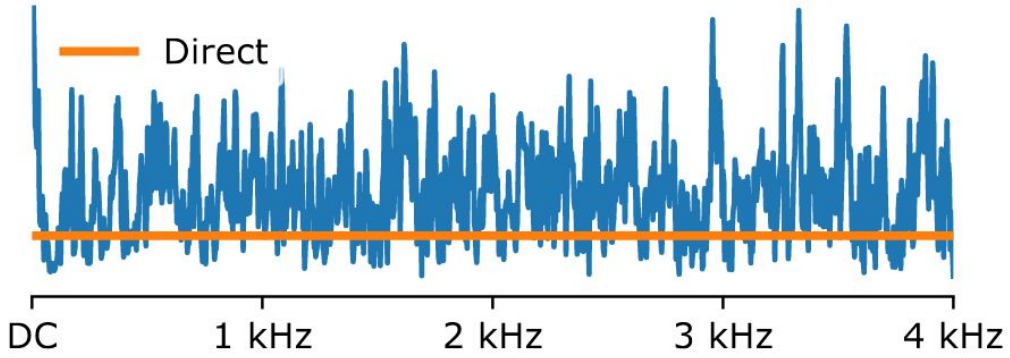


Courtesy of [Sheibler18]

Sound Source **Separation** *with a little help from echoes*

$$H_i(f) = \sum_{r=0}^{R_i} \frac{\alpha_{i,r}}{4\pi \|\mathbf{r}_i - \mathbf{s}\|} e^{-j2\pi f \|\mathbf{r}_i - \mathbf{s}\|/c_{\text{air}}} + \varepsilon_i(f)$$

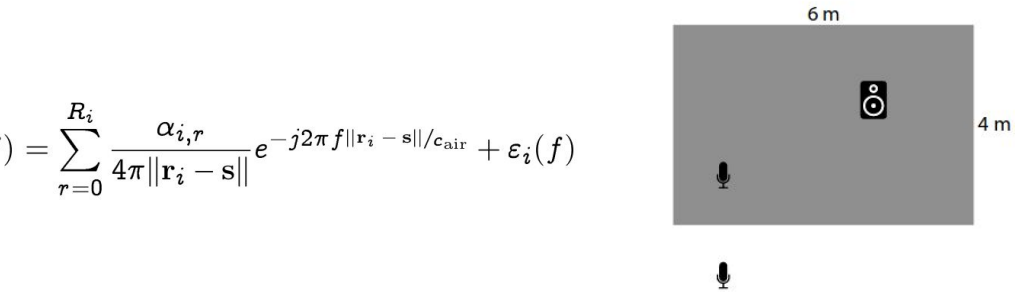


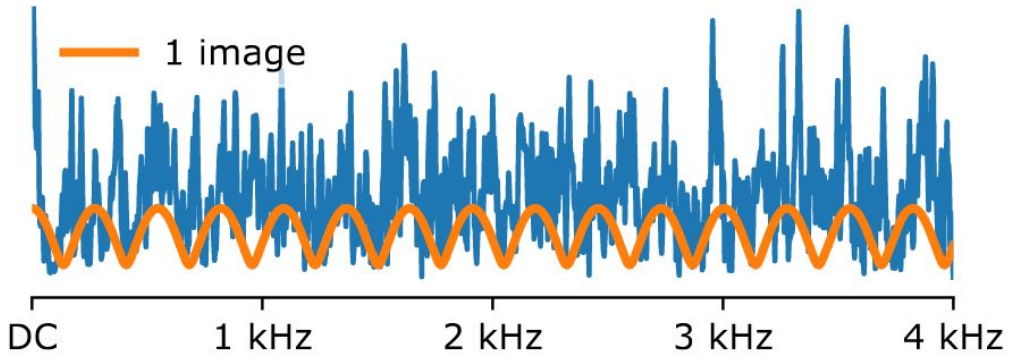


Courtesy of [Sheibler18]

Sound Source **Separation** *with a little help from echoes*

$$H_i(f) = \sum_{r=0}^{R_i} \frac{\alpha_{i,r}}{4\pi \|\mathbf{r}_i - \mathbf{s}\|} e^{-j2\pi f \|\mathbf{r}_i - \mathbf{s}\|/c_{\text{air}}} + \varepsilon_i(f)$$

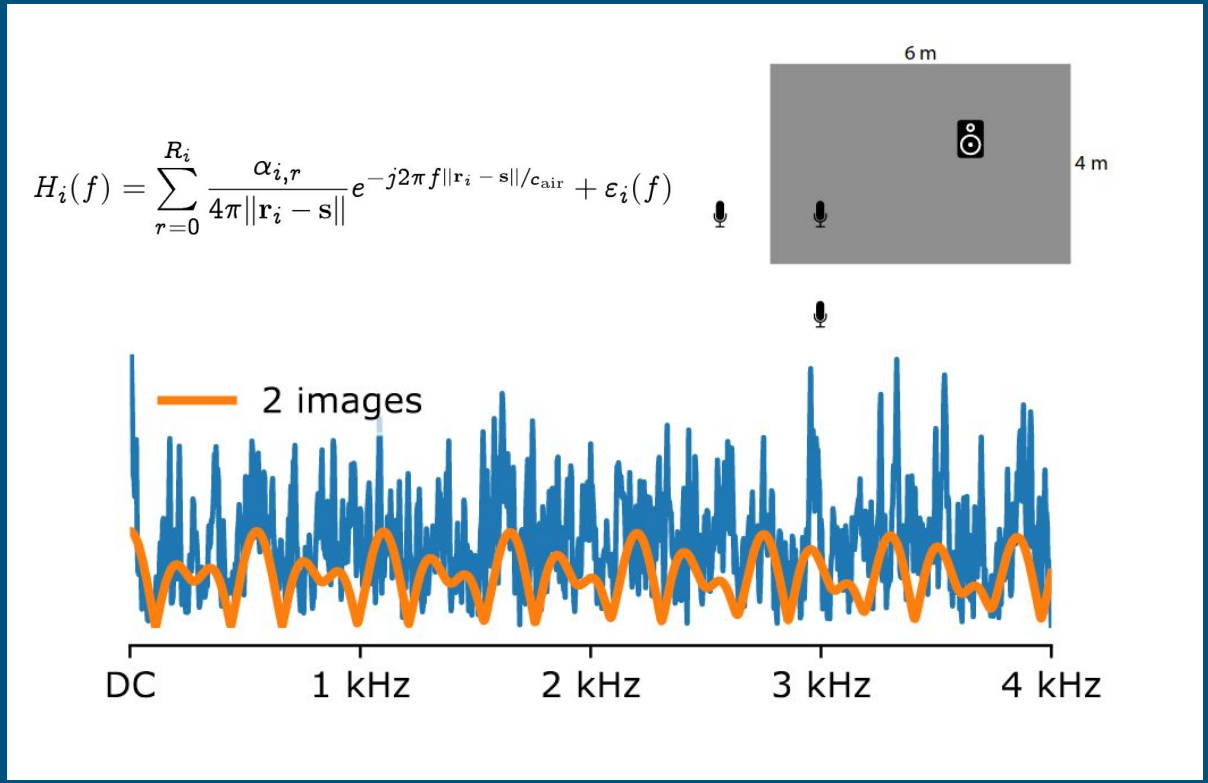




Courtesy of [Sheibler18]

Sound Source **Separation** *with a little help from echoes*

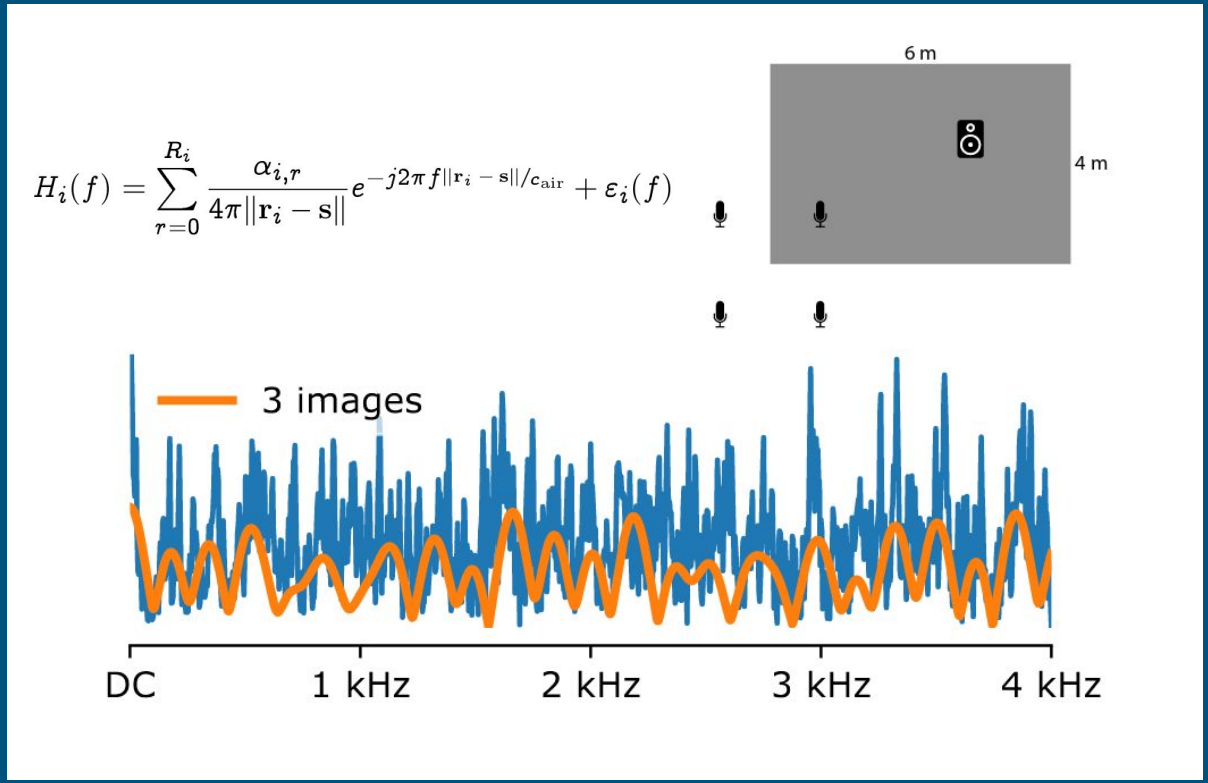
$$H_i(f) = \sum_{r=0}^{R_i} \frac{\alpha_{i,r}}{4\pi \|\mathbf{r}_i - \mathbf{s}\|} e^{-j2\pi f \|\mathbf{r}_i - \mathbf{s}\|/c_{\text{air}}} + \varepsilon_i(f)$$



Courtesy of [Sheibler18]

Sound Source **Separation** *with a little help from echoes*

$$H_i(f) = \sum_{r=0}^{R_i} \frac{\alpha_{i,r}}{4\pi \|\mathbf{r}_i - \mathbf{s}\|} e^{-j2\pi f \|\mathbf{r}_i - \mathbf{s}\|/c_{\text{air}}} + \varepsilon_i(f)$$

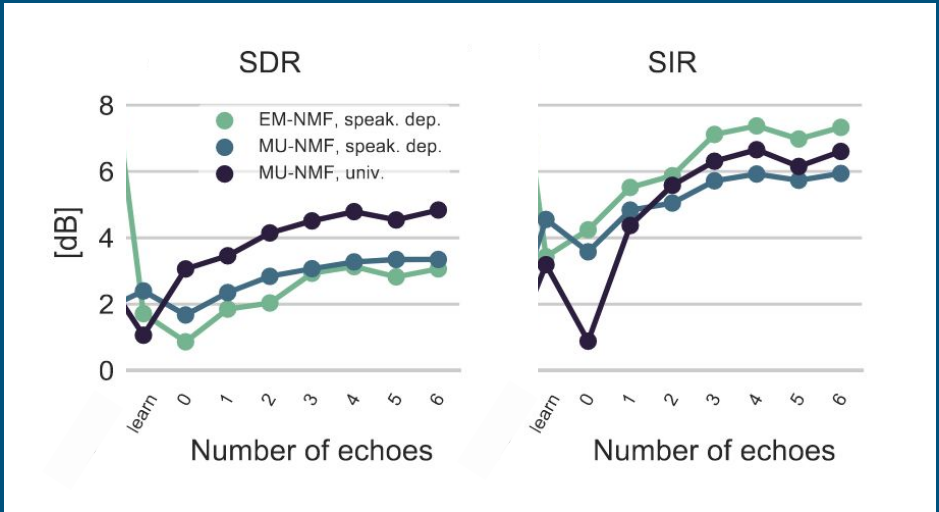


Courtesy of [Sheibler18]

Sound Source **Separation** *with a little help from echoes*

- **We used**
  - Learned dictionary for Speech templates
  - Knowledge of the audio scene
    => initialization point for the ATF
- **Experiments:**
  - 2 sources, 3 microphones
  - Simulation with *Pyroomacoustics*
  - *RT60 = 100 ms*
- ***Baselines***
  - *Learned ATF*
- ***Conclusion**:*
  - ✓ First echoes improve performances
  - ✓ The first are the most important
  - × Initial ATF/Echoes are known

Results on test set
[ICASSP18]



How to estimate them?

Courtesy of [Douglas54]

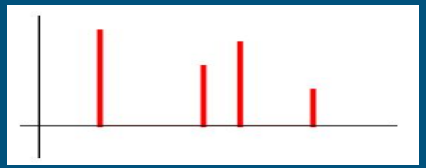
In all the presented works, the echoes are assumed known.

How to estimate them?

Echo Estimation problem

In the time domain, If we model only the specular reflection, the RIR from a fixed point source to the i-th microphone is modeled as **stream of Diracs:**

$$h_i(t) = \sum_{r=0}^{R_i} c_{i,r}\,\delta(t - \tau_{i,r}) + \varepsilon_i(t)$$



r-th Dirac's **coefficient**
(air and surface attenuation)

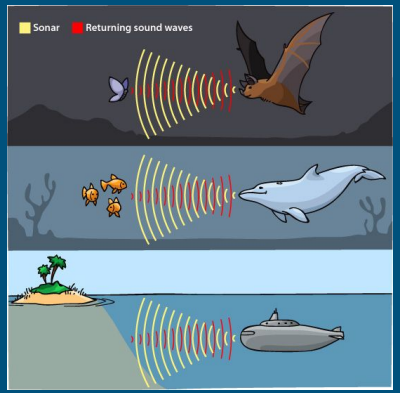
r-th Dirac's **position**
(sound propagation time)

Can we estimate them from sound?

Echo **estimation** problem

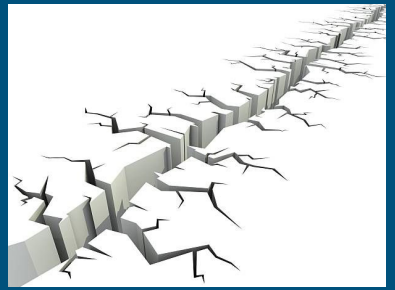
Active Echo estimation
Channel Identification

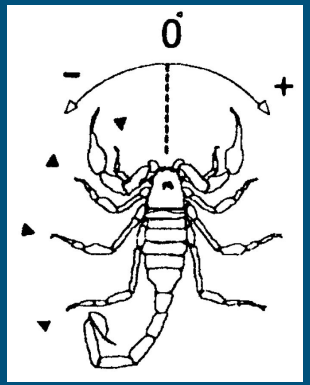
Given multichannel observations $x(t)$ and emitted signal $s(t)$, deduce the Diracs' coefficients $\{c_{m,k}\}$ and position $\{\tau_{m,k}\}$



Passive Echo estimation
Blind Channel Identification

As on the left, but given multichannel observations $x(t)$ only, and the source $s(t)$ is unknown.

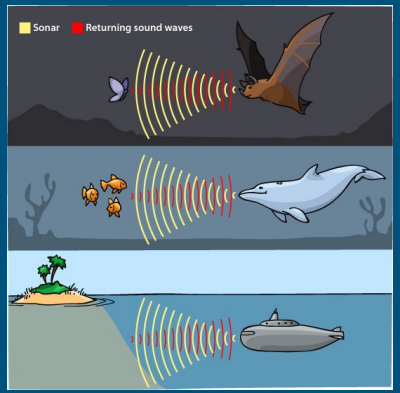




Echo **estimation** problem

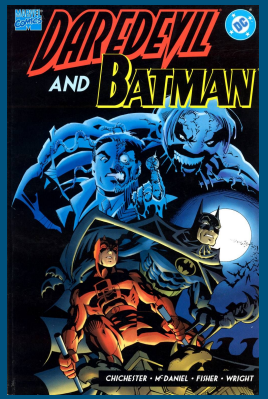
Active Echo estimation
Channel Identification

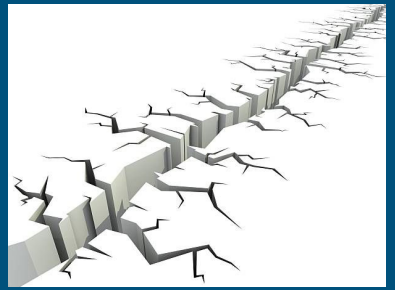
Given multichannel observations $x(t)$ and emitted signal $s(t)$, deduce the Diracs' coefficients $\{c_{m,k}\}$ and position $\{\tau_{m,k}\}$

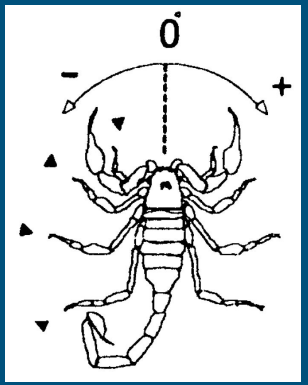


Passive Echo estimation
Blind Channel Identification

As on the left, but given multichannel observations $x(t)$ only, and the source $s(t)$ is unknown.



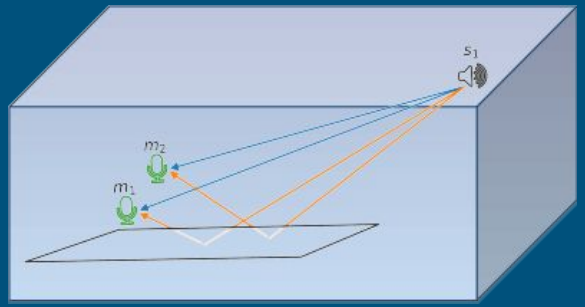




Learning-based Blind Echo Estimation

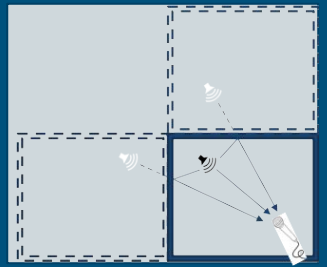
The **pic-nic scenario:**

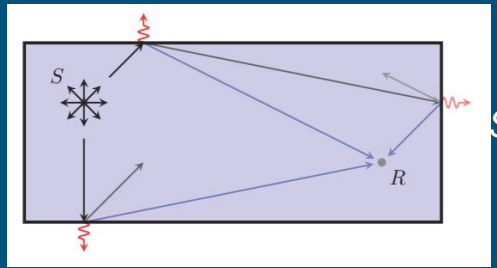
- Single Source
- **Two microphones**
- Random shoe-box rooms
- Nearest surface is the most reflective ~ table-top device



10’000 Auditory pic-nic scenes generated using *[Schimmel et al. 2009]* software

Specular reflection



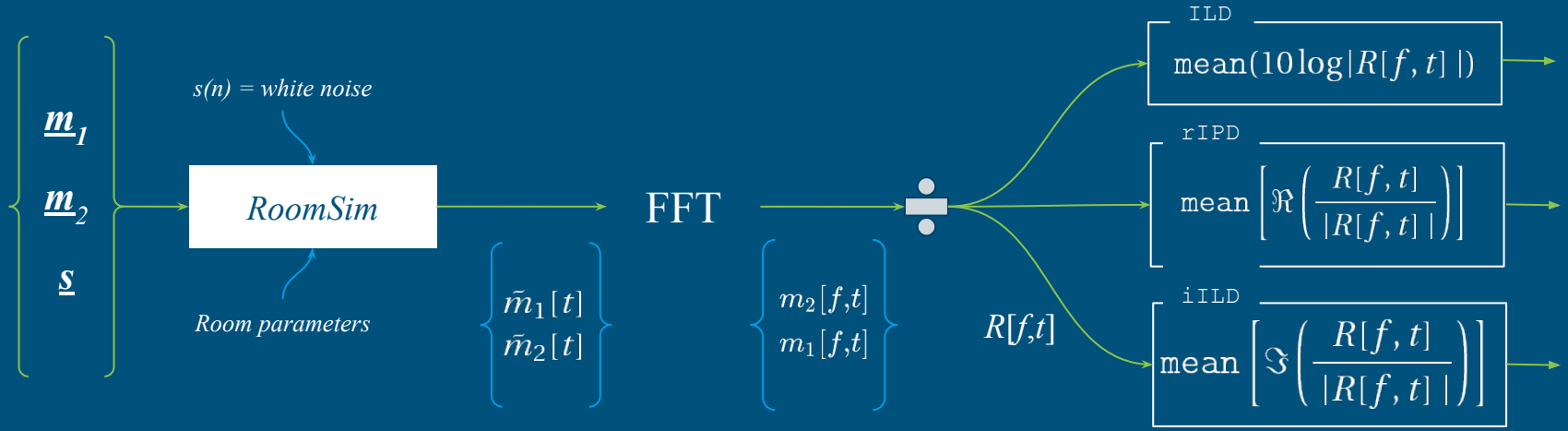


Scattering

Learning-based Blind Echo Estimation

Why only two microphones?

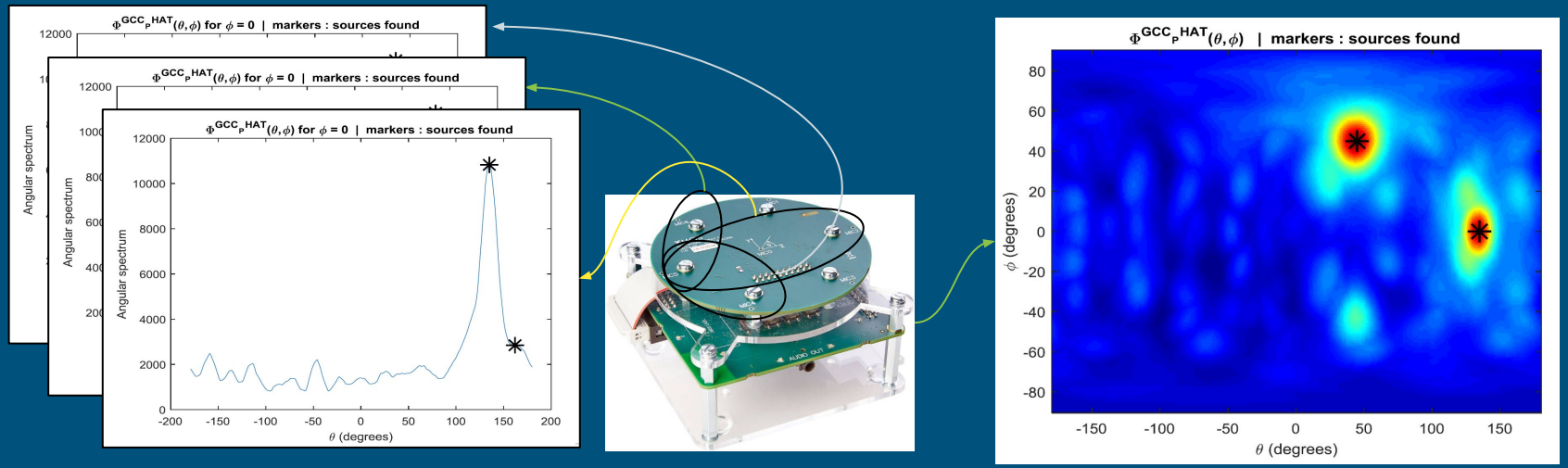
- The relative transfer function can be computed



Learning-based Blind Echo Estimation

Why only two microphones?

- The contribution of multiple microphone pairs can be aggregated together
  - If the geometry of the microphone array is known a priori [*MBSSLocate, DiBiase et al 2001*]



Learning-based Blind Echo Estimation

Why only two microphones?

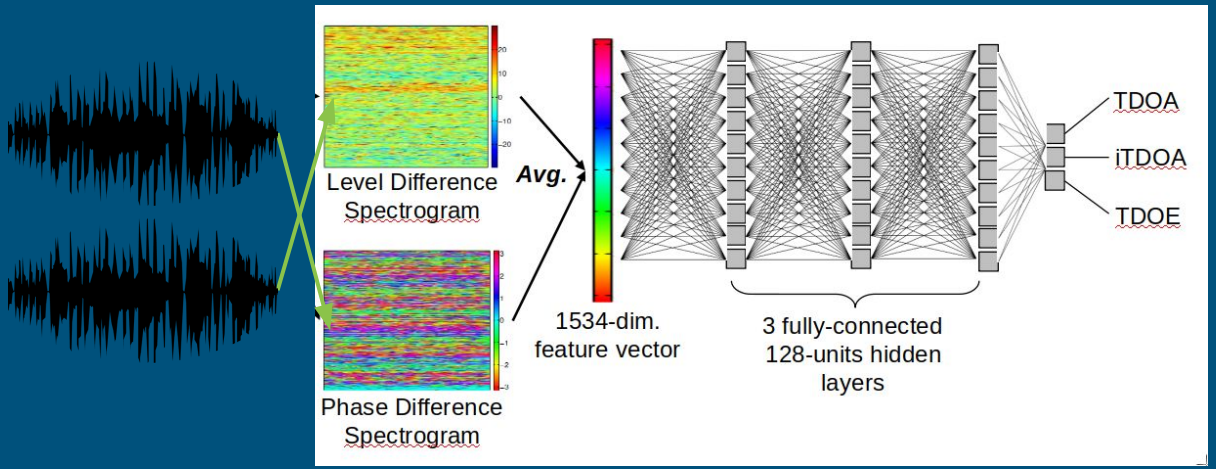
- The (instantaneous) relative transfer function can be computed

$$\begin{cases} \tilde{m}_1[t] = h_1[t] * \tilde{s}[t] \\ \tilde{m}_2[t] = h_2[t] * \tilde{s}[t] \end{cases} \Rightarrow R[f,t] = \frac{m_2[f,t]}{m_1[f,t]} = \frac{h_2[f]s[f,t]}{h_1[f]s[f,t]} = \frac{h_2[f]}{h_1[f]}$$

  - Ideally it removes the dependency from the source signal
  - Assumption:
    - **no noise**
    - **filter shorter** than the fft window
    - Source signals are **broadband signals** with **no-spectral holes**

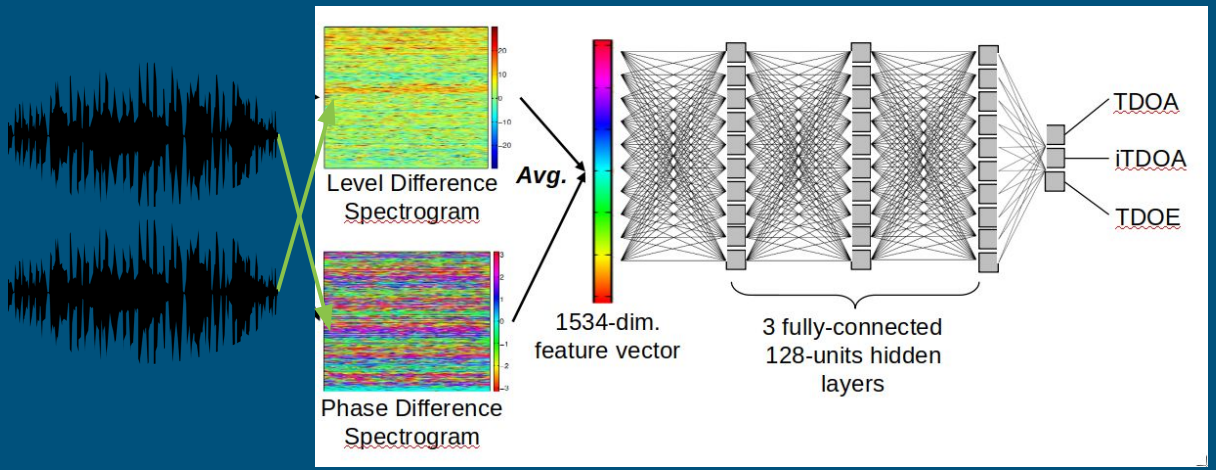
Learning-based Blind Echo Estimation

Simple Deep Neural Network Learning



Learning-based Blind Echo Estimation

Simple Deep Neural Network Learning



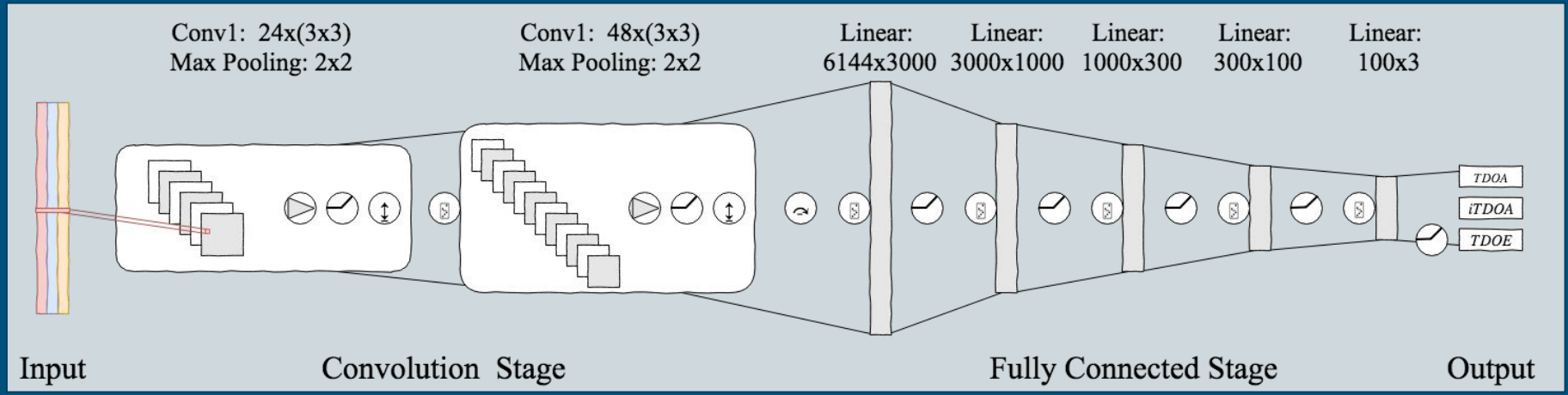
Results on test set
[Presented at ICASSP19]

	Input	nRMSE TDOA	nRMSE iTDOA	nRMSE TDOE
MIRAGE	wn	0.18	0.28	0.25
MIRAGE	wn+n	0.68	0.69	0.89
MIRAGE	sp	0.31	0.34	0.56
MIRAGE	sp+n	0.99	0.98	1.48
GCC-PHAT	wn	0.21	-	-
GCC-PHAT	wn+n	0.68	-	-
GCC-PHAT	sp	0.32	-	-
GCC-PHAT	sp+n	1.38	-	-

Also tried with a Gaussian Locally-Linear Mapping (GLLiM) ⇒ It failed

Learning Echo hunting continues...

- What's next? What's now?
  - State of The Art DNN architecture: CNN [*Chakrabarty et al 2017, Nguyen et al. 2018*]



Learning Echo hunting continues...

- What's next? What's now?
  - State of The Art DNN architecture: CNN [*Chakrabarty et al 2017, Nguyen et al. 2018*]
  - Gaussian and Student-T likelihood Loss Function for estimating both TDOA, iTDOA and TDOE and their uncertainties (~ Mixture Density Network *[Bishop94]*)

$$\mathcal{L}(\theta) = \frac{1}{3}\sum_{n=1}^{N} \left|\tau_{a,n} - \tilde{\tau_{a,n}}\right|^2 + \left|\tau_{i,n} - \tilde{\tau_{i,n}}\right|^2 + \left|\tau_{e,n} - \tilde{\tau_{e,n}}\right|^2$$

Learning Echo hunting continues...

- What's next? What's now?
  - State of The Art DNN architecture: CNN [*Chakrabarty et al 2017, Nguyen et al. 2018*]
  - Gaussian and Student-T likelihood Loss Function for estimating both TDOA, iTDOA and TDOE and their uncertainties (~ Mixture Density Network *[Bishop94]*)

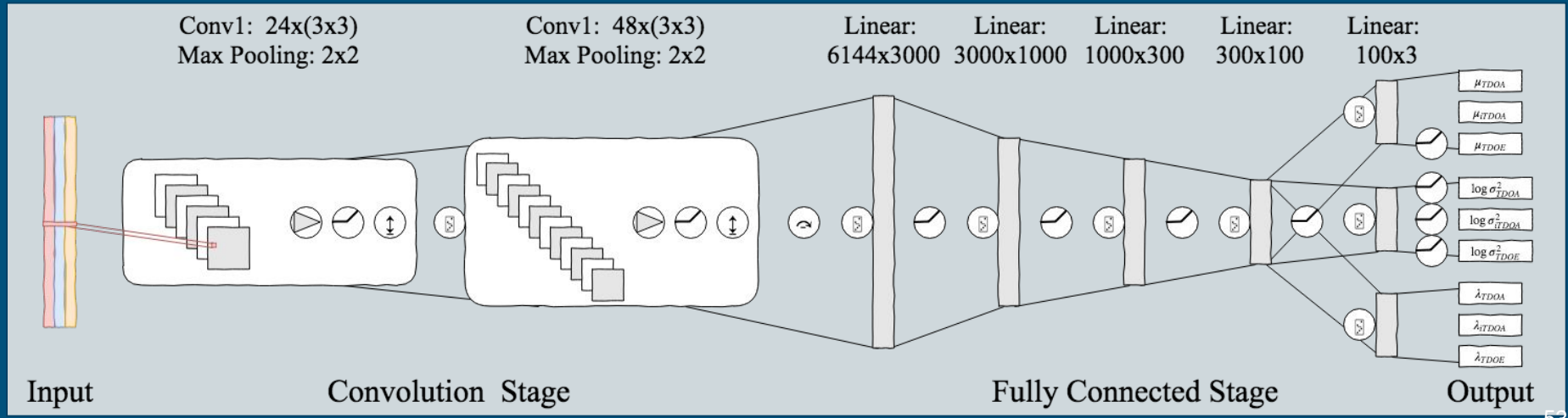
$$\mathcal{L}(\theta) = \frac{1}{3}\sum_{n=1}^{N} \left|\tau_{a,n} - \tilde{\tau_{a,n}}\right|^2 + \left|\tau_{i,n} - \tilde{\tau_{i,n}}\right|^2 + \left|\tau_{e,n} - \tilde{\tau_{e,n}}\right|^2$$

$$p(\tau_k|X;\theta) \sim \mathcal{N}(\mu_{\tau_k}(x_n;\theta), \sigma^2_{\tau_k}(x_n;\theta)) \quad k = a, i, e$$

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \log \sigma^2_{\tau_a}(x_n) + \frac{|\tau_a - \mu_{\tau_a}(x_n)|^2}{\sigma^2_{\tau_a}(x_n)} + \dots$$

Learning Echo hunting continues...

- What's next? What's now?
  - State of The Art DNN architecture: CNN [*Chakrabarty et al 2017, Nguyen et al. 2018*]
  - Gaussian and Student-T likelihood Loss Function for estimating both TDOA, iTDOA and TDOE and their uncertainties (~ Mixture Density Network *[Bishop94]*)



Learning Echo hunting continues...

- Results
  - State of The Art DNN architecture: CNN [*Chakrabarty et al 2017, Nguyen et al. 2018*]
  - Gaussian and Student-T likelihood Loss Function (~ Mixture Density Network *[Bishop94]*)

Results on test set
[Presented at ICASSP19]

	Input	nRMSE TDOA	iTDOA	TDOE
MIRAGE	wn	0.18	0.28	0.25
MIRAGE	wn+n	0.68	0.69	0.89
MIRAGE	sp	0.31	0.34	0.56
MIRAGE	sp+n	0.99	0.98	1.48
GCC-PHAT	wn	0.21	-	-
GCC-PHAT	wn+n	0.68	-	-
GCC-PHAT	sp	0.32	-	-
GCC-PHAT	sp+n	1.38	-	-

Current Results on test set
with noise

distr	snr	phase	test_signal	tdoa	itdoa	tdoe1
gaussian	0	Test	noise	0.103131	0.110806	0.248462
gaussian	15	Test	noise	0.102640	0.110342	0.280237
gaussian	30	Test	noise	0.101439	0.108265	0.323202
none	0	Test	noise	0.137354	0.145333	0.209920
none	15	Test	noise	0.192951	0.196020	0.284383
none	30	Test	noise	0.148980	0.151179	0.222592
student	0	Test	noise	0.099268	0.107615	0.237591
student	15	Test	noise	0.110567	0.111748	0.310297
student	30	Test	noise	0.106170	0.113793	0.294742

Learning Echo hunting continues...

- Results
  - State of The Art DNN architecture: CNN
  - Gaussian and Student-T likelihood Loss Function (~ Mixture Density Network *[Bishop94]*)
- Work in progress
  - *Estimate more echoes R>2*:
    - Multi-scale loss function
    - Physical based DNN
  - *Multichannel aggregation*
  - *Training with proper datasets*
    - One for picnic scenarios
    - One for distributed arrays
  - State-of-the-Art *RTF estimation*
  - *Test on real data*
  - *Performances in terms of beamforming*

Current Results on test set
with noise

distr	snr	phase	test_signal	tdoa	itdoa	tdoe1
gaussian	0	Test	noise	0.103131	0.110806	0.248462
gaussian	15	Test	noise	0.102640	0.110342	0.280237
gaussian	30	Test	noise	0.101439	0.108265	0.323202
none	0	Test	noise	0.137354	0.145333	0.209920
none	15	Test	noise	0.192951	0.196020	0.284383
none	30	Test	noise	0.148980	0.151179	0.222592
student	0	Test	noise	0.099268	0.107615	0.237591
student	15	Test	noise	0.110567	0.111748	0.310297
student	30	Test	noise	0.106170	0.113793	0.294742

BLASTER

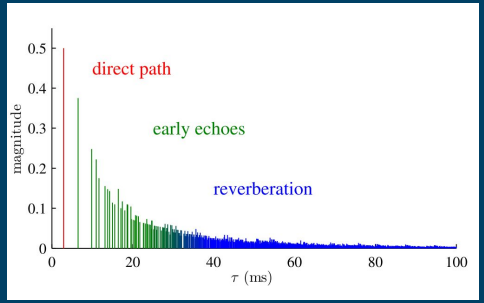
BLind and Sparse Technique
for
Acoustic Echo Retrieval

Accepted at ICASSP 2020

[Diego Di Carlo, Clement Elvira,
Antoine Deleforge, Nancy Bertin, Remi Gribonval]

***Observation**: the early part of RIR is*

- *Sparse*
- *"Coefficient are non-negative" (strong assumption)*
- *Location are off-grid*
- *Sum of Diracs functions ⇒ we know their closed-form*

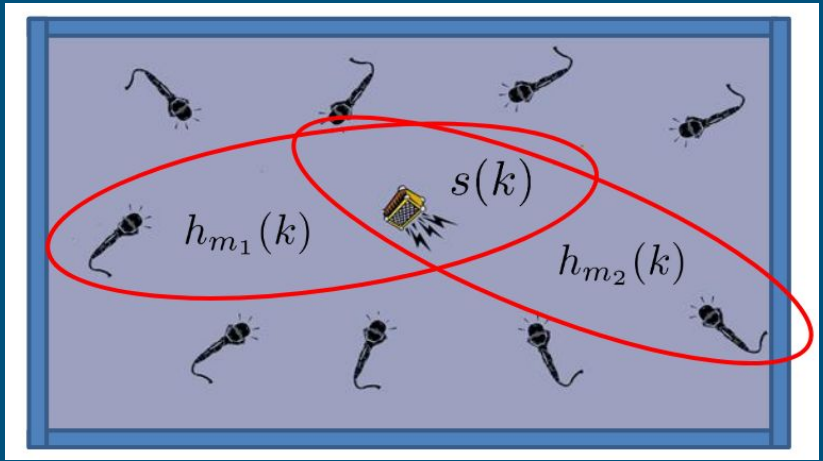


*⇒ we can formulated it as a **off-grid spike-retrieval problem** (Continuous Dictionary, Super-Resolution)*

Collaboration with Clement ELVIRA
Postdoc about Continuous Dictionary
at INRIA Rennes (Panama Team).

Blind Echo Estimation

For every couple of microphones, in noiseless case, the **cross-relation identity holds**



Courtesy of [Crocco15]

$$(h_r * h_l * s)(t) = (h_l * h_r * s)(t) \quad \forall l \neq r$$

Blind Echo Estimation

State of the Art baseline methods [*Lin et al. 2007, Crocco et al. 2015*]:

- (Toeplitz) **Cross relation** as cost-function

$$x_i(t) = (h_i * s)(t) + n_i(t), \quad i = 1, 2, \ldots, I$$

Cross Relation

$$(h_1 * h_2 * s)(t) = (h_2 * h_1 * s)(t)$$

Convolution as Toeplitz matrix multiplication

$$\mathbf{h}_1^*, \mathbf{h}_2^* = \underset{\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{R}^L}{\arg\min} \frac{1}{2} \|\mathbf{X}_1 \mathbf{h}_2 - \mathbf{X}_2 \mathbf{h}_1\|_2^2$$

Blind Echo Estimation

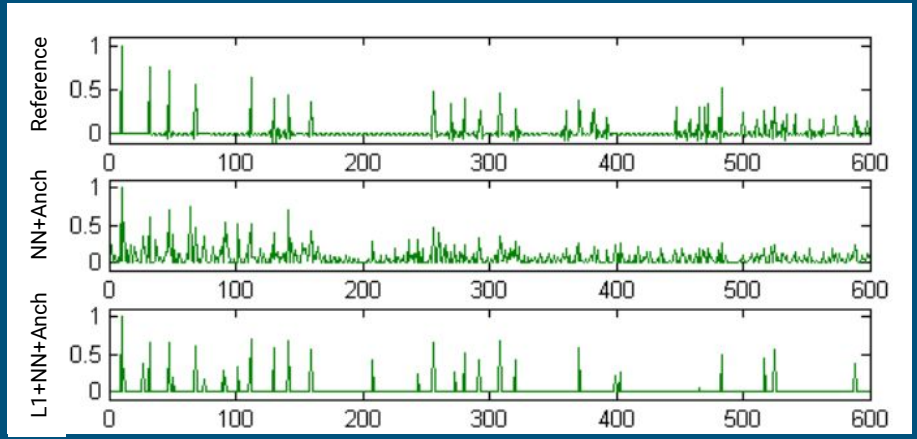
State of the Art baseline methods [*Lin et al. 2007, Crocco et al. 2015*]:

- (Toeplitz) **Cross relation** as cost function
- **Anchor constraint** for avoiding the trivial solution $\mathbf{h} = \mathbf{0}$
- **non-negativity constraints** and **sparse penalty** for robustness to noise

Blind and Sparse Non-negative (**BSN**) BCE *[Lin07]*

$$\mathbf{h}_1^*, \mathbf{h}_2^* = \underset{\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{R}^L}{\arg\min} \frac{1}{2} \|\mathbf{X}_1\mathbf{h}_2 - \mathbf{X}_2\mathbf{h}_1\|_2^2 + \lambda \|\mathbf{h}\|_1$$

$$s.t. \;\; \mathbf{h}_1[0] = 1, \mathbf{h} \geq 0$$



Courtesy of [Crocco15]

Blind Echo Estimation

The *discrete* **Blind Channel Estimation** problem:

$$\mathbf{h}_1^*, \mathbf{h}_2^* = \underset{\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{R}^L}{\arg\min} \frac{1}{2} ||\mathbf{X}_1 \mathbf{h}_2 - \mathbf{X}_2 \mathbf{h}_1||_2^2 + \lambda ||\mathbf{h}||_1 \quad s.t. \;\; \mathbf{h}_1[0] = 1, \mathbf{h} \geq 0$$

is equivalent to a **LASSO*** problem as follows:

$$\mathbf{x}^* = \underset{\mathbf{x}}{\arg\min} \frac{1}{2} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 + \lambda ||\mathbf{x}||_1 \quad s.t. \;\; \mathbf{x} \geq 0$$

where

$$\mathbf{b} = \mathbf{X}_2 \mathbf{e}_1, \quad \mathbf{x} = [\mathbf{h}_1[2:], \mathbf{h}_2]^{\mathrm{T}} \;\text{ and }\; \mathbf{A} = [-\mathbf{X}_2[:, 2:], \mathbf{X}_1]$$

Limitations:

- Location are off-grid
- Computational cost: algorithms scales linearly with the size of $\mathbf{A}$

Observation:

- $\sum_i \delta_i$ functions => we know their closed-form

LASSO: Least Absolute Shrinkage and Selection Operator

BLASTER - off-grid BCE

Methods from the theory of *Super Resolution/Sparse-Spikes Deconvolution/Continuous Dictionary:*

LASSO (discrete solution)

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1 \quad s.t.\ \ \mathbf{x} \geq 0$$

Dictionary $\mathbf{A}$:

- $\mathbf{A}\mathbf{e}_i$: select a column of $\mathbf{A}$
- $\mathbf{A}\mathbf{e}_i$: select an echo at locations $\boldsymbol{\tau}_i$

Solution $\mathbf{x} = [\mathbf{h}_1, \mathbf{h}_2]$:

- Looking for a Sparse vector $\mathbf{x}$

Sparsity enforced with **L1-norm**: $\|\cdot\|_1$

Solved with standard **LASSO solver**

closed-form from Fourier theory

$\mathbf{h}=\sum_i c_i\, \delta(t - \boldsymbol{\tau}_i)$ *is a spike measure*

BLASSO* (continuous solution)

$$\mu^* = \arg\min_{\mu} \frac{1}{2}\|\mathcal{A}\mu - \mathbf{b}\|_2^2 + \lambda\|\mu\|_{TV} \quad s.t.\ \ \mu \in \mathcal{M}^+$$

Operator $\mathbf{A}$:

- $\mathcal{A}(\delta\boldsymbol{\tau})$: select echo at location $\boldsymbol{\tau} \in [0, 1]$ *ms*

Solution $\mathbf{x}$:

- Spares positive measure $\mu = \sum_i c_i\, \delta(t - \boldsymbol{\tau}_i)$

Sparsity enforced with **Total-Variation Norm**: $\|\cdot\|_{TV}$

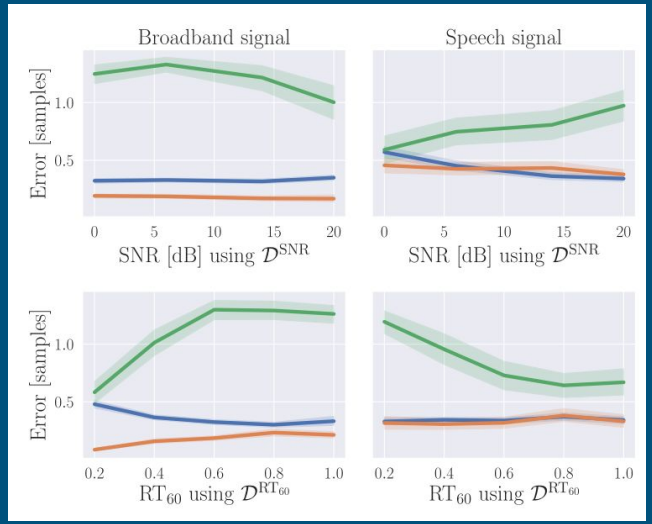
Solved with Gradient descent algorithm (**Frank-Wolfe algo**)

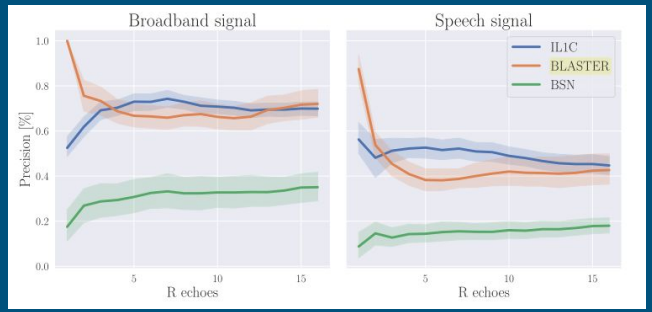
BLASTER - off-grid BCE

- **Experiments:**
  - 1 source and 2 microphones
  - Random geometry in random Shoeboxes
  - Simulation with *Pyroomacoustics*
  - ***Different SNR*** *and* ***Different RT60***
- ***Baselines***
  - ***BSN****: Blind, Sparse and Non-negative BCE [Lin et al. 07]*
  - ***IL1C****: Iterative weighted-L1 norm BCE [Crocco et al. 15]*
- ***Conclusion****:*
  - ✓ Perfect off-grid reconstruction for **noiseless** and **early case**
  - ✓ Similar or slightly worse performance wrt state of the art
  - ✗ Performances are source-dependent
  - ✗ Better performances for $R < 4$
- Future directions:
  - multichannel extensions
  - Sync model
  - Relative **Early** and **Denoised** Transfer Function

Results on test set

[ICASSP20]

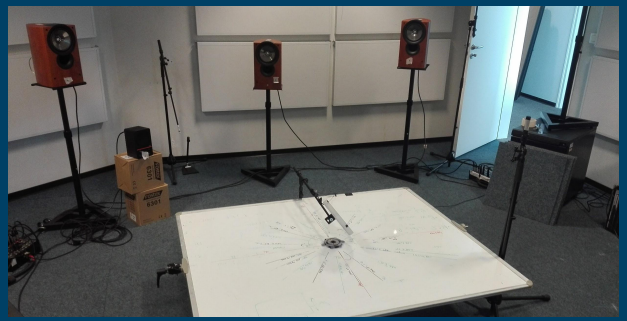




Picnic of the MUSIS dataset

A small dataset for Tabletop Device

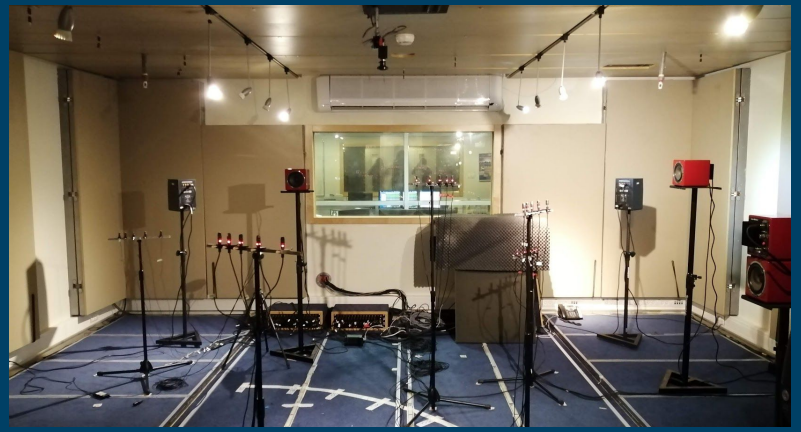
Research project funded by HONDA



אוניברסיטת בר-אילן
Bar-Ilan University

dEchoerate dB

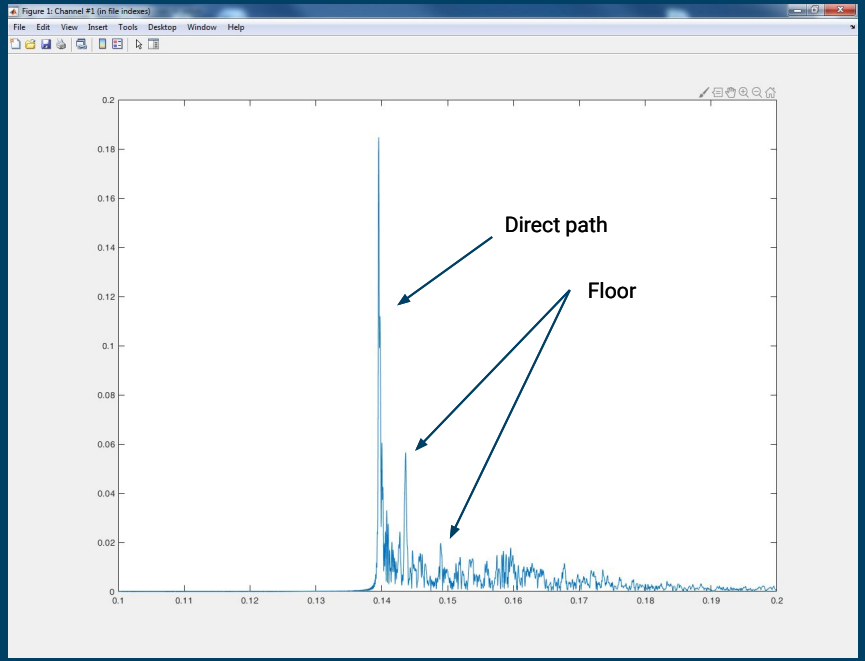
A dataset for Echoes Retrieval



אוניברסיטת בר-אילן
Bar-Ilan University

dEchoerate dB

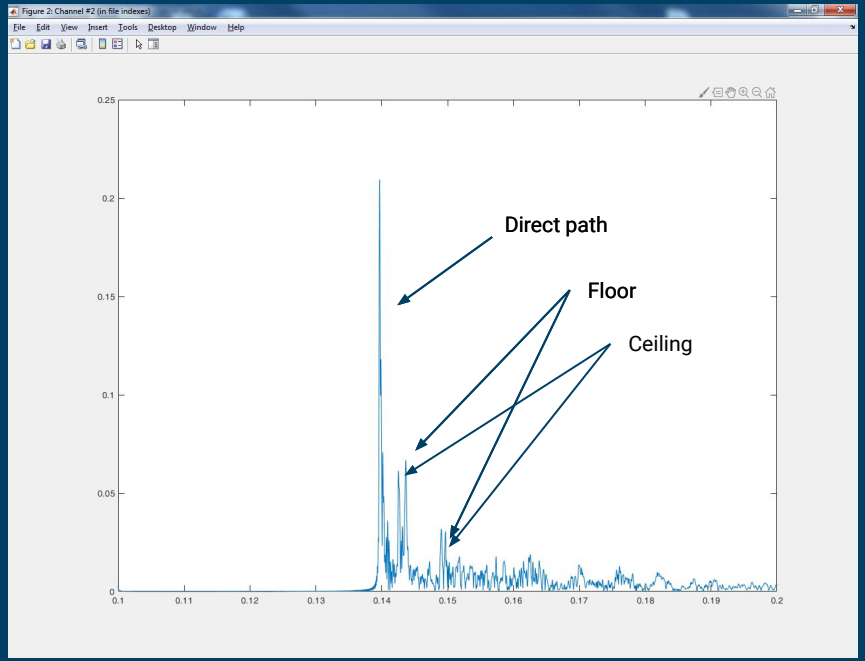
A database for Echoes Retrieval



אוניברסיטת בר-אילן
Bar-Ilan University

dEchoerate dB

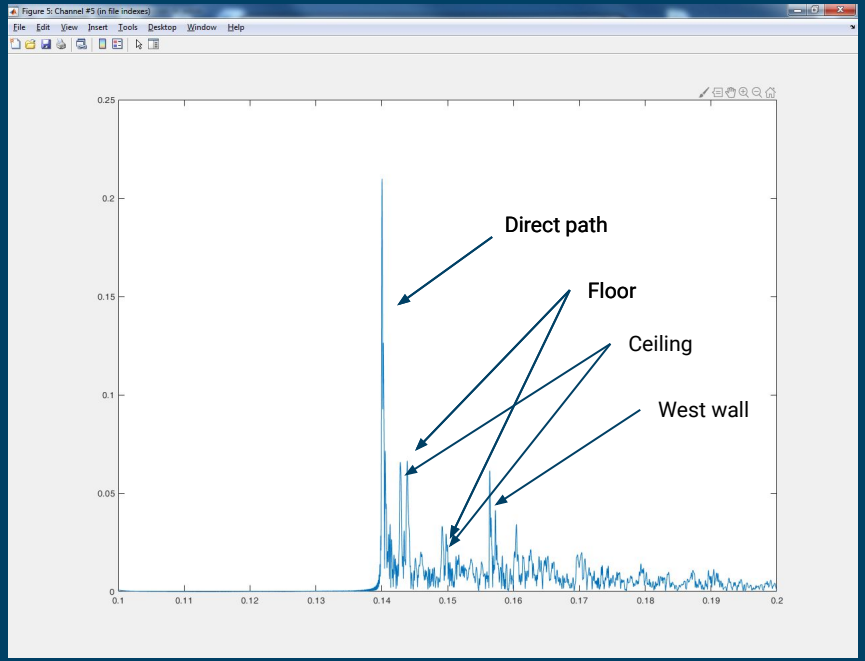
A database for Echoes Retrieval



אוניברסיטת בר-אילן
Bar-Ilan University

dEchoerate dB

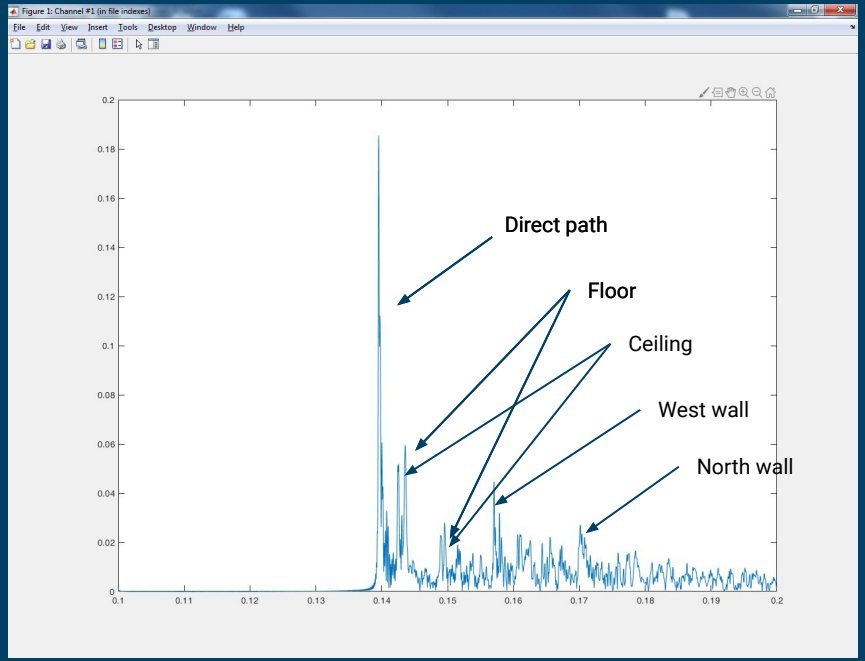
A database for Echoes Retrieval



אוניברסיטת בר-אילן
Bar-Ilan University

dEchoerate dB

A database for Echoes Retrieval



THANK YOU

Questions, suggestions
or re-explanations?

Diego DI CARLO
diego.di-carlo@inria.fr
chutlhu.github.io