Echo-aware signal processing for audio scene analysis

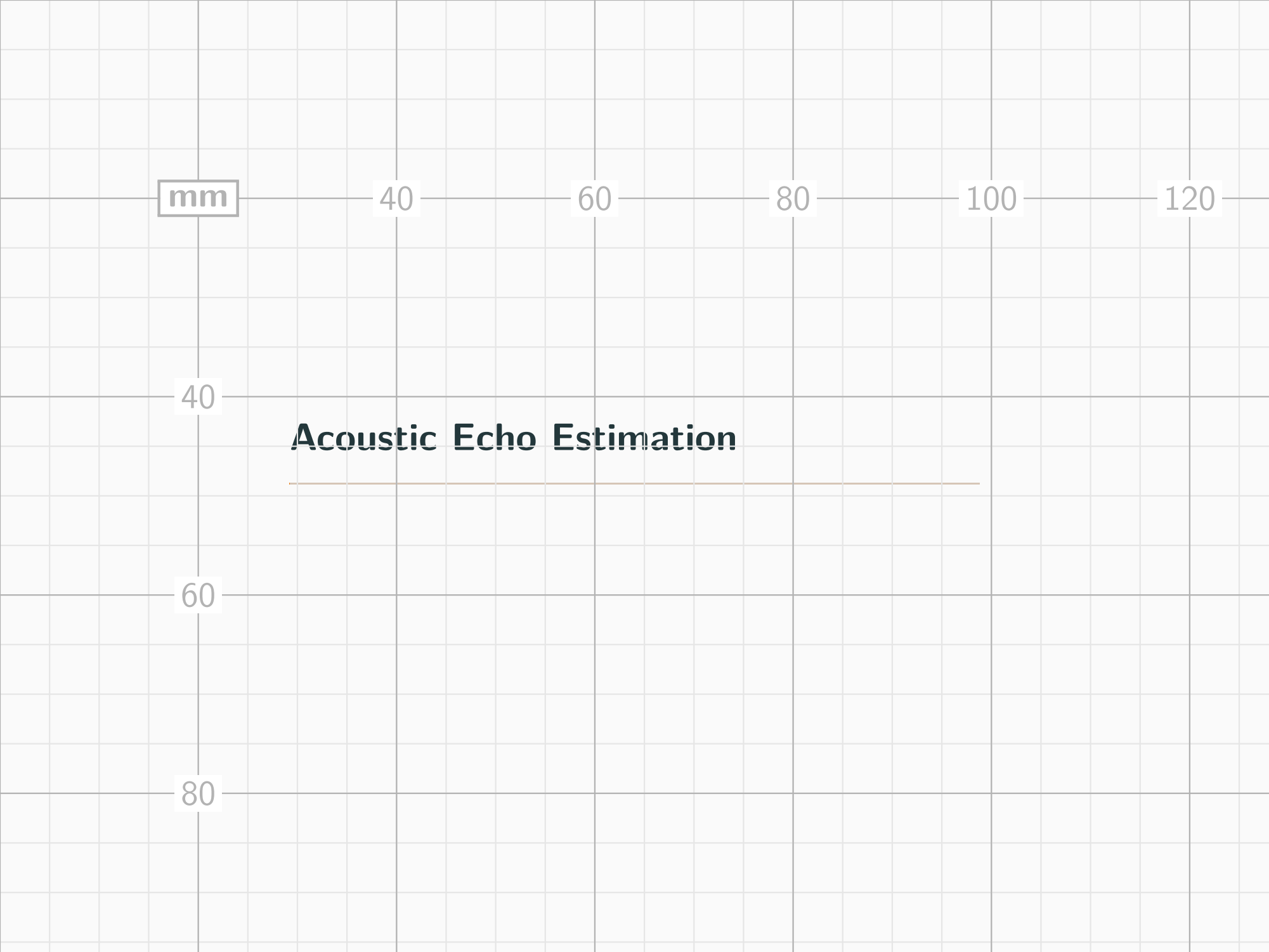
Diego Carlo

November 27, 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



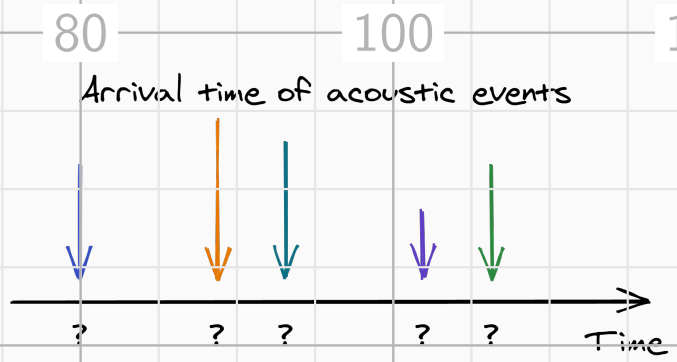
Acoustic Echo Estimation

Acoustic Echo Retrieval (AER)

Estimating early (strong) reflection from microphones recordings

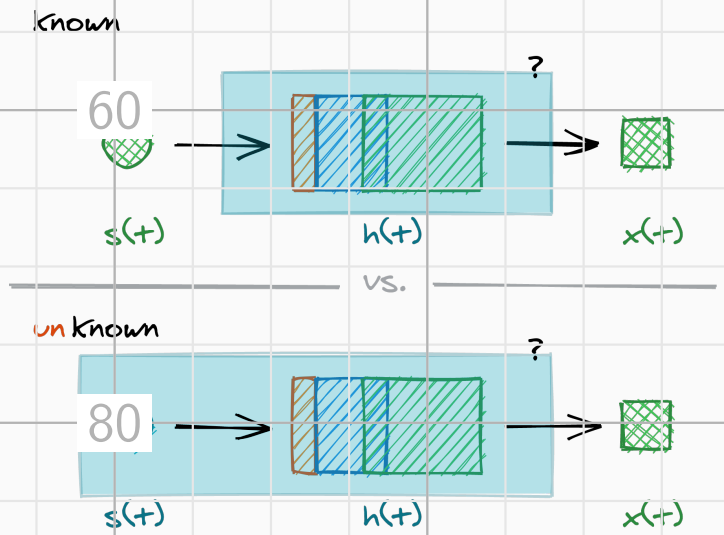
$$\{\tilde{x}_i\}_i \longrightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$

estimate $\tau_i^{(r)}$ $\longrightarrow$ TOAs estimation



Scenarios:

Source signal is



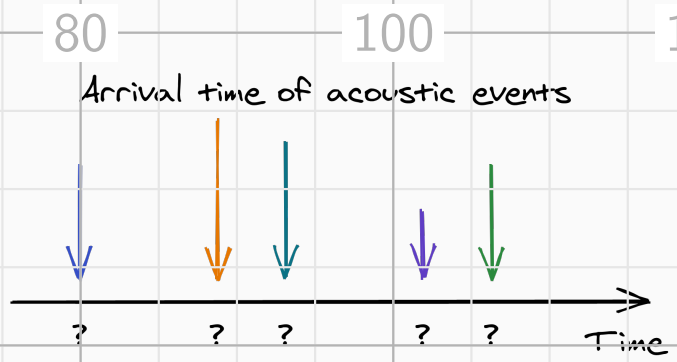
Our case: signal source and passive system of (I microphones)

Acoustic Echo Retrieval (AER)

Estimating early (strong) reflection from microphones recordings

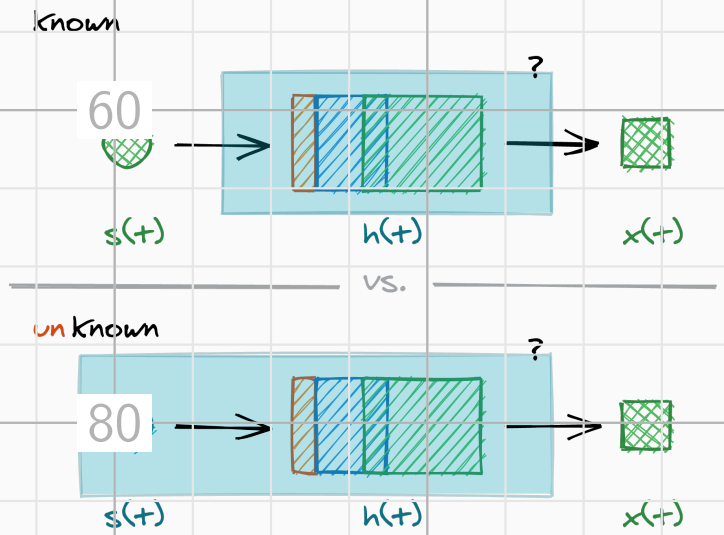
$$\{\tilde{x}_i\}_i \longrightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$

estimate $\tau_i^{(r)}$ $\longrightarrow$ TOAs estimation



Scenarios:

Source signal is



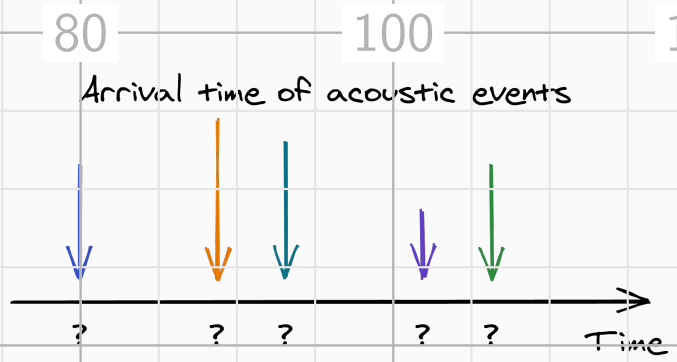
Our case: signal source and passive system of (I microphones)

Acoustic Echo Retrieval (AER)

Estimating early (strong) reflection from microphones recordings

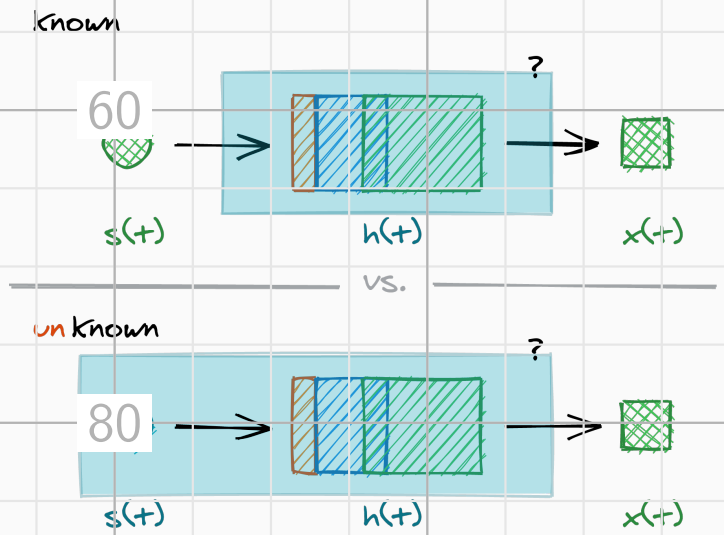
$$\{\tilde{x}_i\}_i \longrightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$

estimate $\tau_i^{(r)}$ $\longrightarrow$ TOAs estimation



Scenarios:

Source signal is



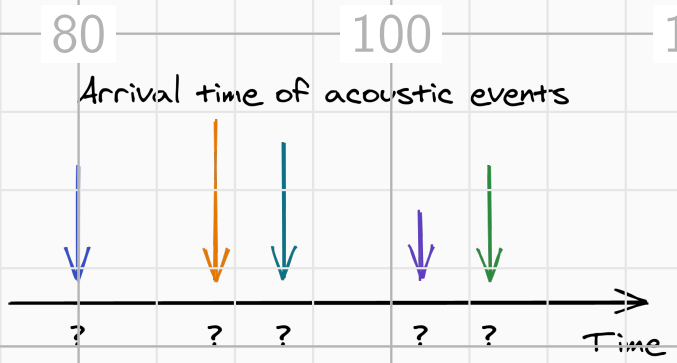
Our case: signal source and passive system of (I microphones)

Acoustic Echo Retrieval (AER)

Estimating early (strong) reflection from microphones recordings

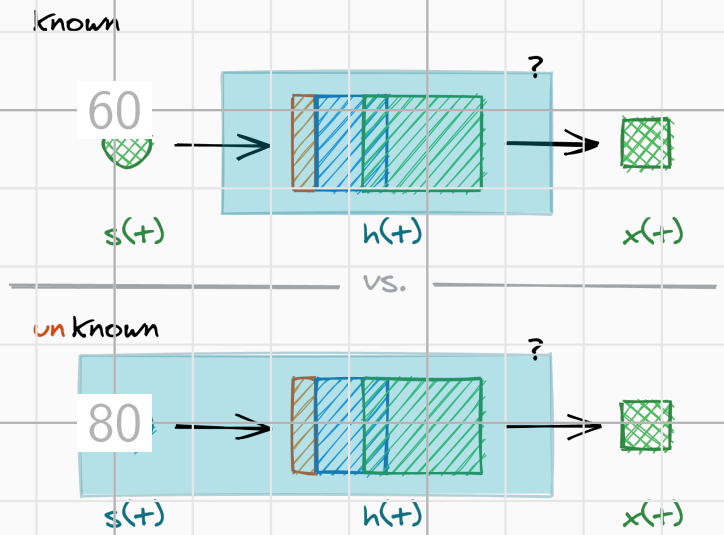
$$\{\tilde{x}_i\}_i \longrightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$

estimate $\tau_i^{(r)}$ $\longrightarrow$ TOAs estimation



Scenarios:

Source signal is



Active

- **intrusive** or specific setup
- **non-blind** problem

(**Application:** sonar, calibration, measurements, etc.)

Passive

- **passive** and more common setups
- **blind inverse** problem (harder)

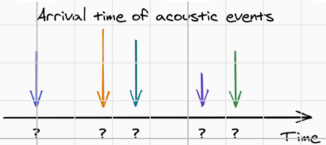
(**Applications:** passive listening, smart speakers, etc.)

Acoustic Echo Retrieval (AER)

Estimating early (strong) reflection for microphones recordings

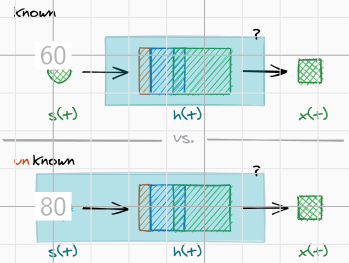
$$\{\tilde{x}_i\}_i \longrightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$

estimating $\tau_i^{(r)}$ $\rightarrow$ TOAs estimation

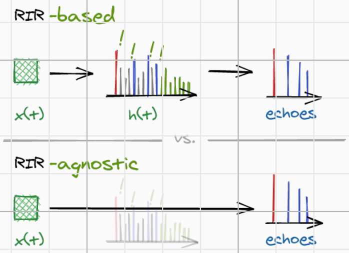


Scenarios:

Source signal is



Estimation is

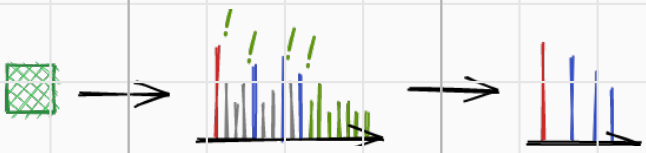


1

Our case: signal source and passive system of (I microphones)

Passive Acoustic Echo Estimation

RIR-based approaches



1. BCE problem $\Longrightarrow$ RIRs
2. Peak picking $\Longrightarrow$ Echoes

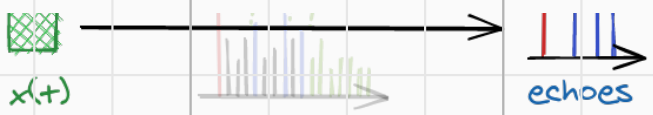
Pros

- BCE is well studied (eg. LASSO)
  (many frameworks, solver, >20 years of literature)
- Currently state of the art good in some application

Cons

- Full RIR is estimated
- Peak picking
- On-grid estimation

RIR-agnostic approaches



1. Estimation in $\Theta = \{\tau_i^{(r)}\}_{i,r}$

- No full RIRs + no peak picking
  - Lower complexity
  - Sub-sampling accuracy
- Echo model is respected (bypassing sampling)
  (Sparsity, Non-negativity)

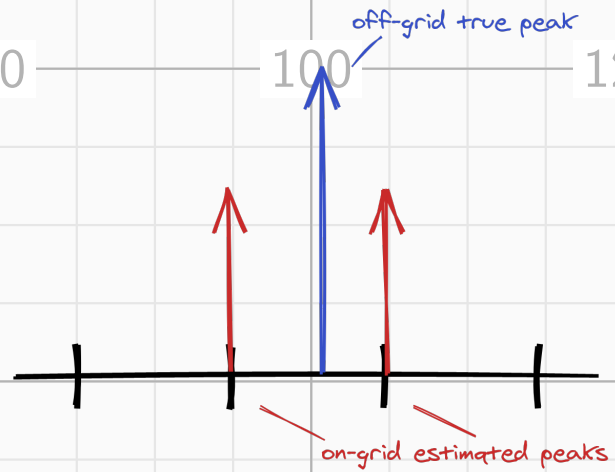
exploratory (no solver)

no direct translation of BCE

- easy ill-conditioned

Limitations & bottleneck

- **Estimation is on grid**
  - Sparsity and non-negativity not true "on grid"
  - *Body guard* effect [Duval and Peyré, 2017]
    - → low recall ⟹ low accuracy
    - → slow convergence
  - Peak Picking
    - → manually tuned / corrected
    - → echo labeling (NP-hard problem)



How about higher F_s?

→ Increase Precision

Sorry, computational bottleneck!

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

How to solve this?

RIR-agnostic + off-grid

1. Learning-based
2. Analytical

Learning-based off-grid AER

Recall: AER $\Leftrightarrow \{\tilde{x}_i\} \overset{?}{\longrightarrow} \{\tau_i^{(r)}, \alpha_i^{(r)}\}_r$

Observations:

- This *direct* mapping is difficult, the *inverse* "is not"
  $\rightarrow$ acoustic simulators: mic/src/room $\longrightarrow$ $\tau_i^{(r)}, \alpha_i^{(r)}, \quad \tilde{h}_i, \quad \tilde{x}_i$
- Acoustic simulator are "simple", versatile and fast
  $\rightarrow$ many data
- This approach is successfully in *Sound Source Localization*
  $\rightarrow$ position is related to echoes
  [Nguyen et al., 2018, Perotin et al., 2019] ⚠ Not only DNN

Idea: (Deep) Learning-based AER

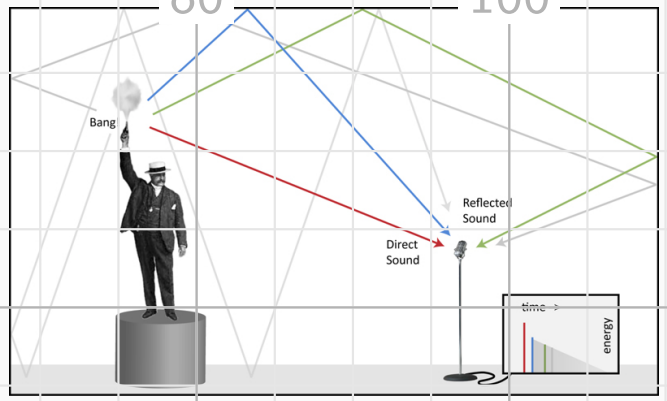
1. Extend virtually learning-based SSL to AER
2. Estimate first echo estimation (simple but important)
3. Only 2 microphones

Proposed Models

Output: mm target

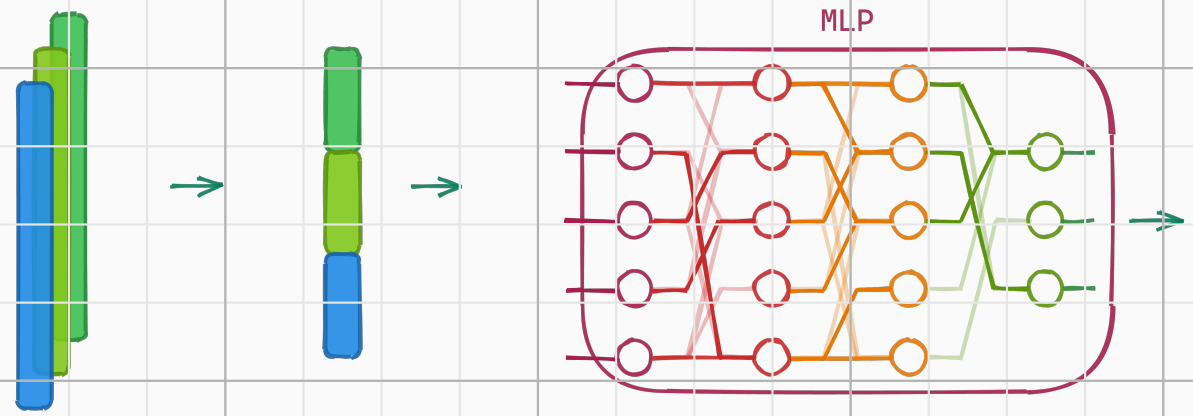
First order echoes $\tau_i^{(r)}$ for 2 microphones

Without loss of generality: TDOAs



Input: features

$\frac{X_2[f,t]}{X_1[f,t]} \in \mathbb{C} \rightarrow \underset{t}{\text{avg.}}$

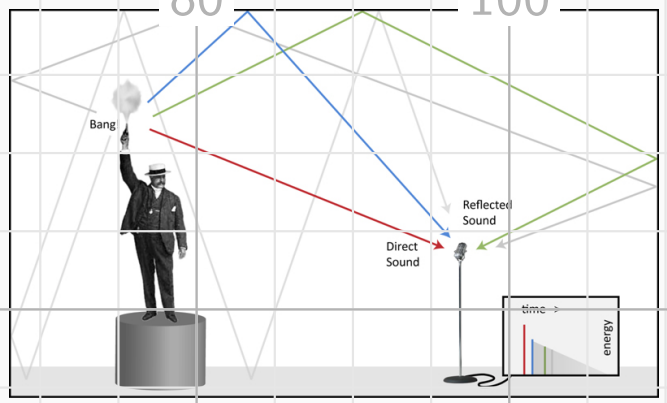


Proposed Models

Output: mm target

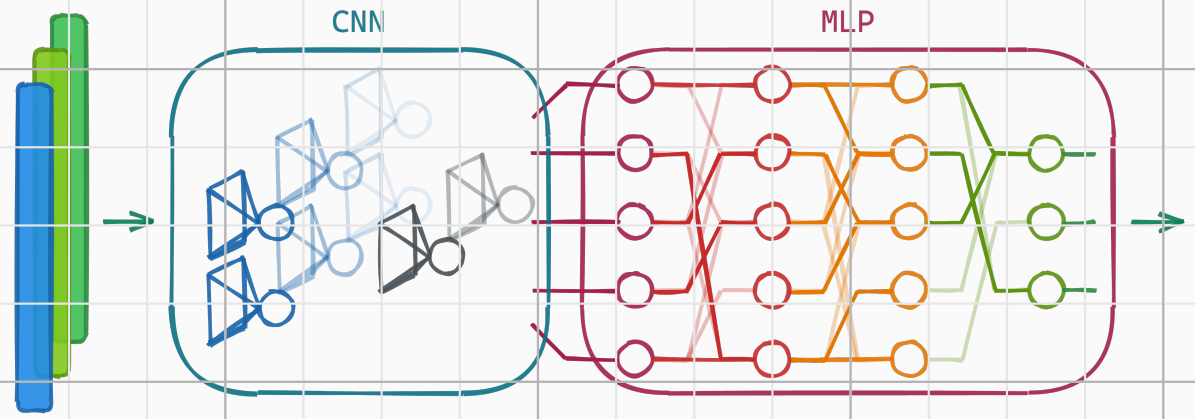
First order echoes $\tau_i^{(r)}$ for 2 microphones

Without loss of generality: TDOAs



Input: features

$$\frac{X_2[f,t]}{X_1[f,t]} \in \mathbb{C} \rightarrow \underset{t}{\text{avg.}}$$



Data and Model

mm 40 60 80 100 120

40

1: better than a baseline?

Can we estimate the echos?

Is is robust to noise?

60

80

Experimental resutls

mm 40 60 80 100 120

Proposed Method: `MLP`, `CNN`, $\texttt{CNN}_{\mathcal{N}}$, $\texttt{CNN}_{\mathcal{T}}$

1: better than a baseline?

40

Baseline `GCC PHAT` (Only real TDOA)[Knapp and Carter, 1976]

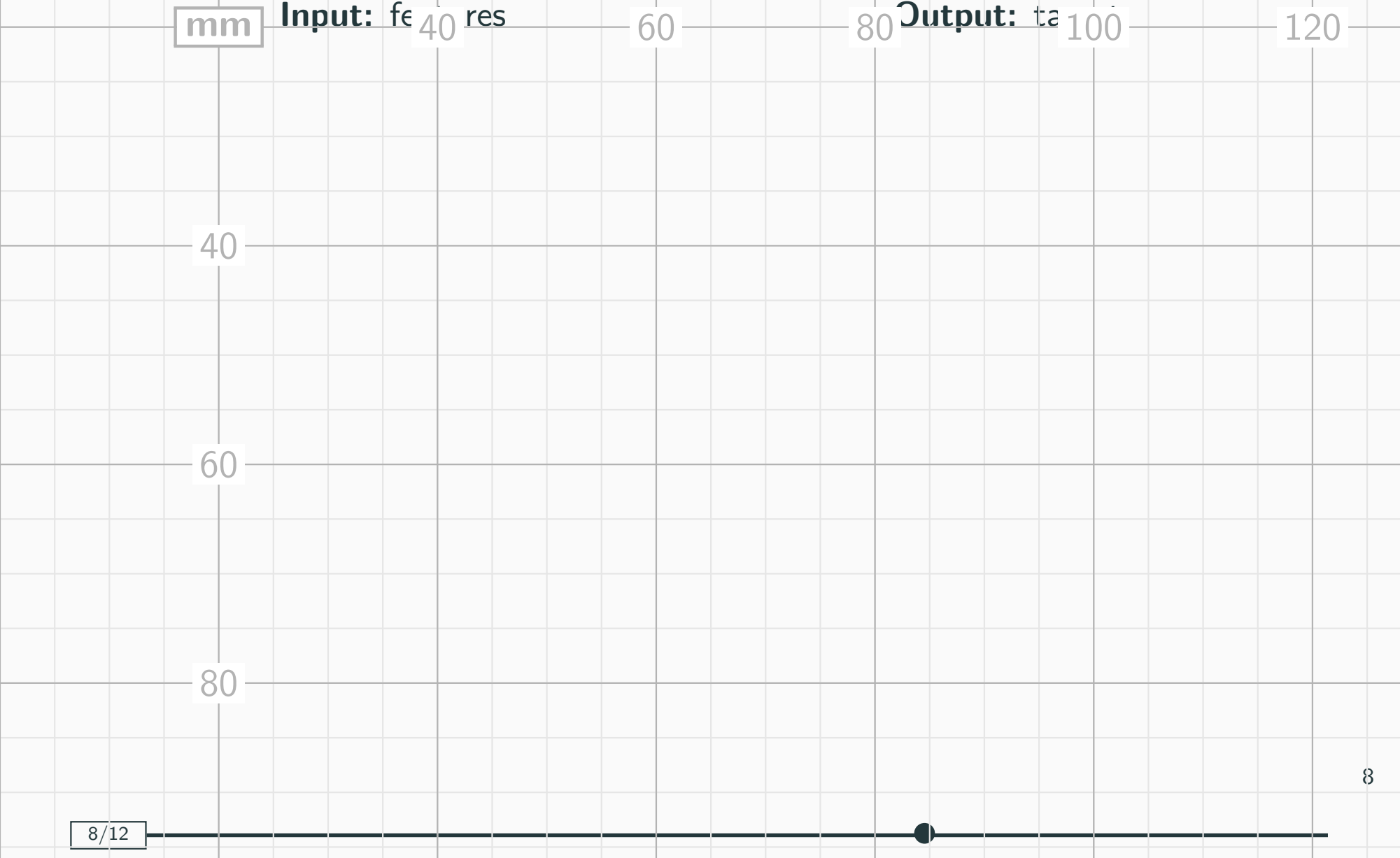
2: CNN better than MLP?

Can we estimate the echos? 60

Is is robust to noise?

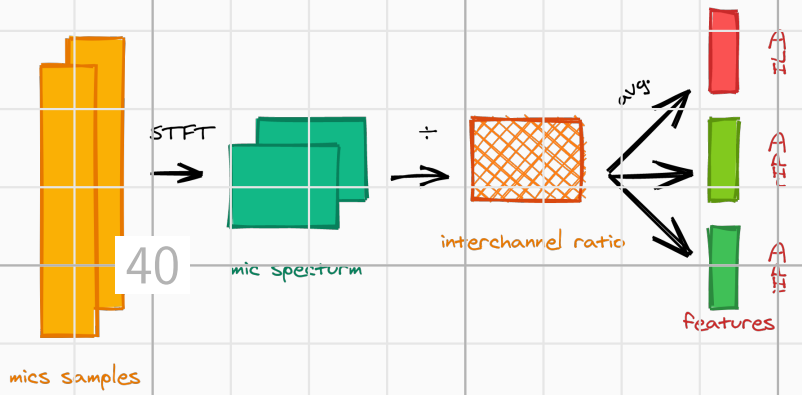
80

Lantern- Data & Models



Lantern- Data & Models

Input: features

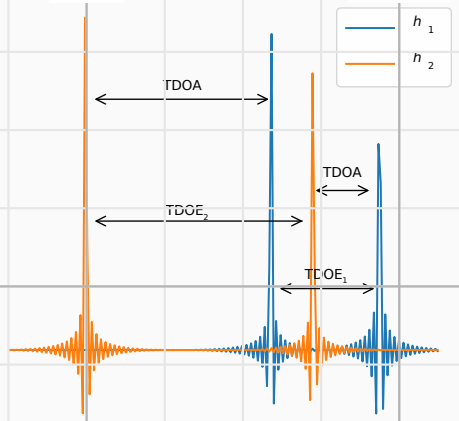


Relative Transfer Function

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

This is the **instantaneous** ReTF

Output: targets



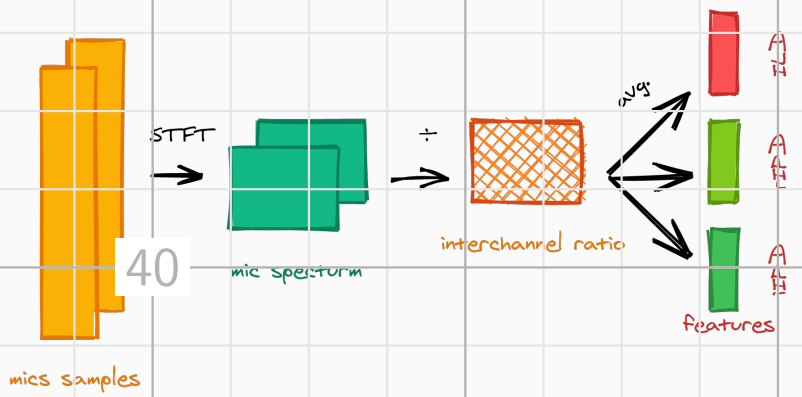
TDOAs inter and intra reflection

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

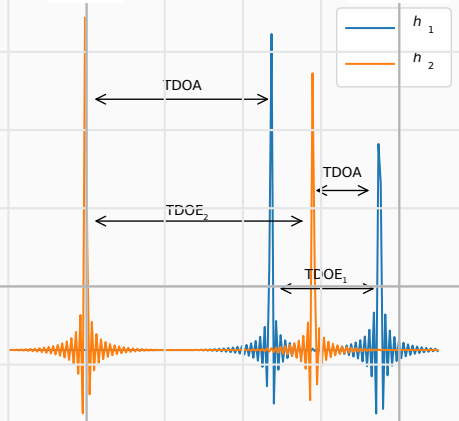
First strongest echo $\Leftrightarrow$ close surface

Lantern- Data & Models

Input: features



Output: target



Relative Transfer Function

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

This is the **instantaneous** ReTF

TDOAs inter and intra reflection

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

First strongest echo $\Leftrightarrow$ close surface

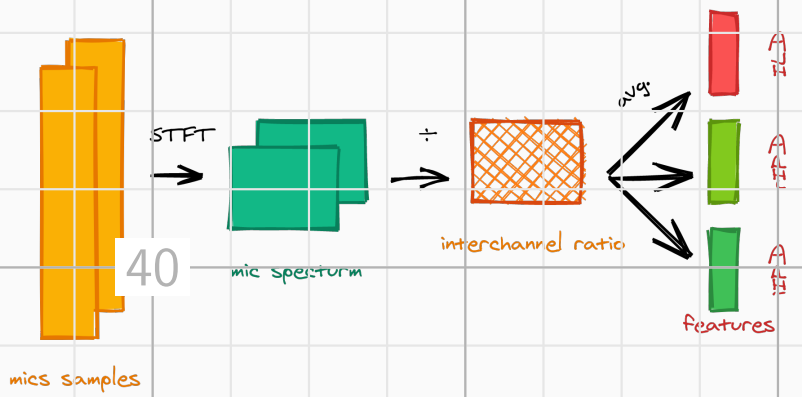
Model

- Architecture: **CNN** [Chakrabarty and Habets, 2017, Nguyen et al., 2018]

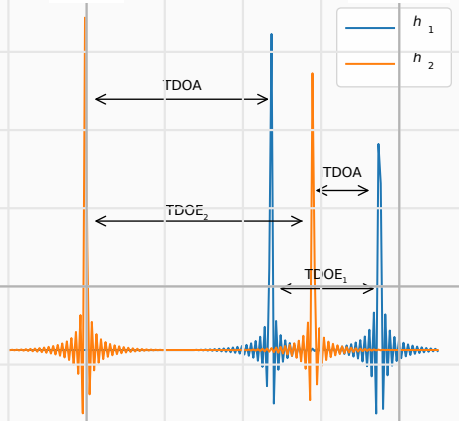
8

Lantern- Data & Models

Input: features



Output: target



Relative Transfer Function

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

This is the **instantaneous** ReTF

TDOAs inter and intra reflection

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

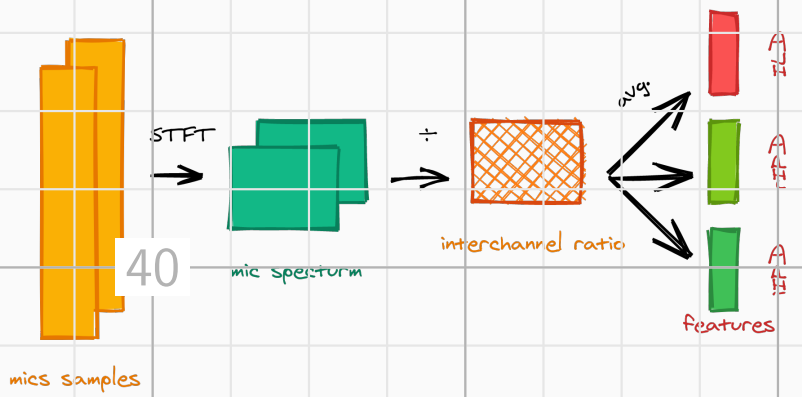
First strongest echo $\Leftrightarrow$ close surface

Model

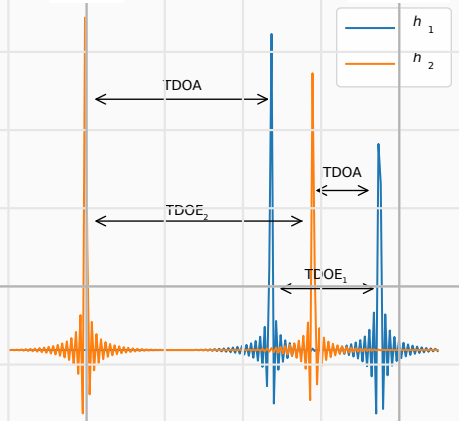
- Architecture: **CNN** [Chakrabarty and Habets, 2017, Nguyen et al., 2018]
- Loss Function:
  1. RMSE (Multi-label regression) on $\mathcal{V}$
  2. Gaussian log-likelihood $\rightarrow \{\mu, \sigma^2\}$
  3. Student log-likelihood $\rightarrow \{\mu, \lambda, \nu\}$

Lantern- Data & Models

Input: features



Output: target



Relative Transfer Function

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

This is the **instantaneous** ReTF

TDOAs inter and intra reflection

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

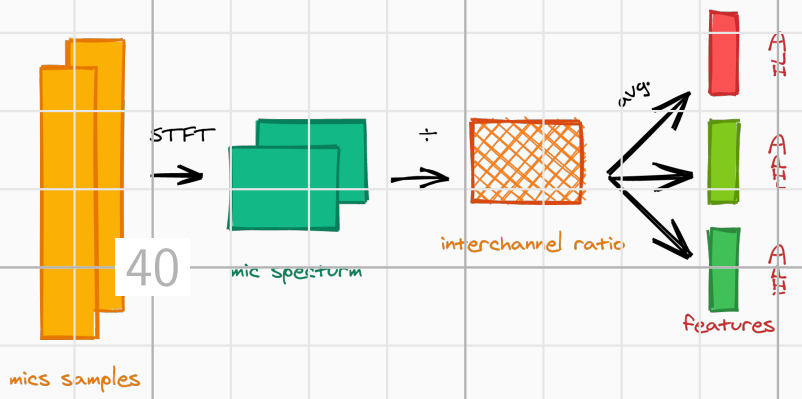
First strongest echo $\Leftrightarrow$ close surface

Model

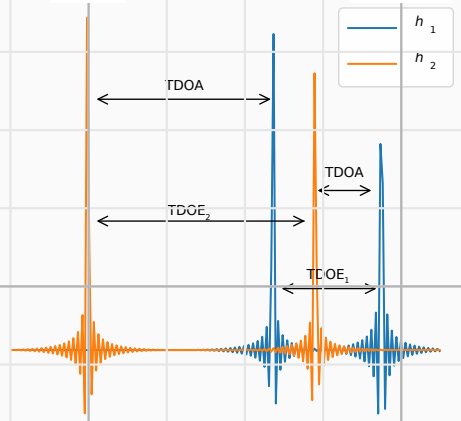
- Architecture: **CNN** [Chakrabarty and Habets, 2017, Nguyen et al., 2018]
- Loss Function:
  1. RMSE (Multi-label regression) on $\mathcal{V}$
  2. Gaussian log-likelihood $\rightarrow \{\mu, \sigma^2\}$
  3. Student log-likelihood $\rightarrow \{\mu, \lambda, \nu\}$

Lantern- Data & Models

Input: features



Output: target



Relative Transfer Function

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

This is the **instantaneous** ReTF

TDOAs inter and intra reflection

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

First strongest echo $\Leftrightarrow$ close surface

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 $\rightarrow \{\mu, \lambda, \nu\}$

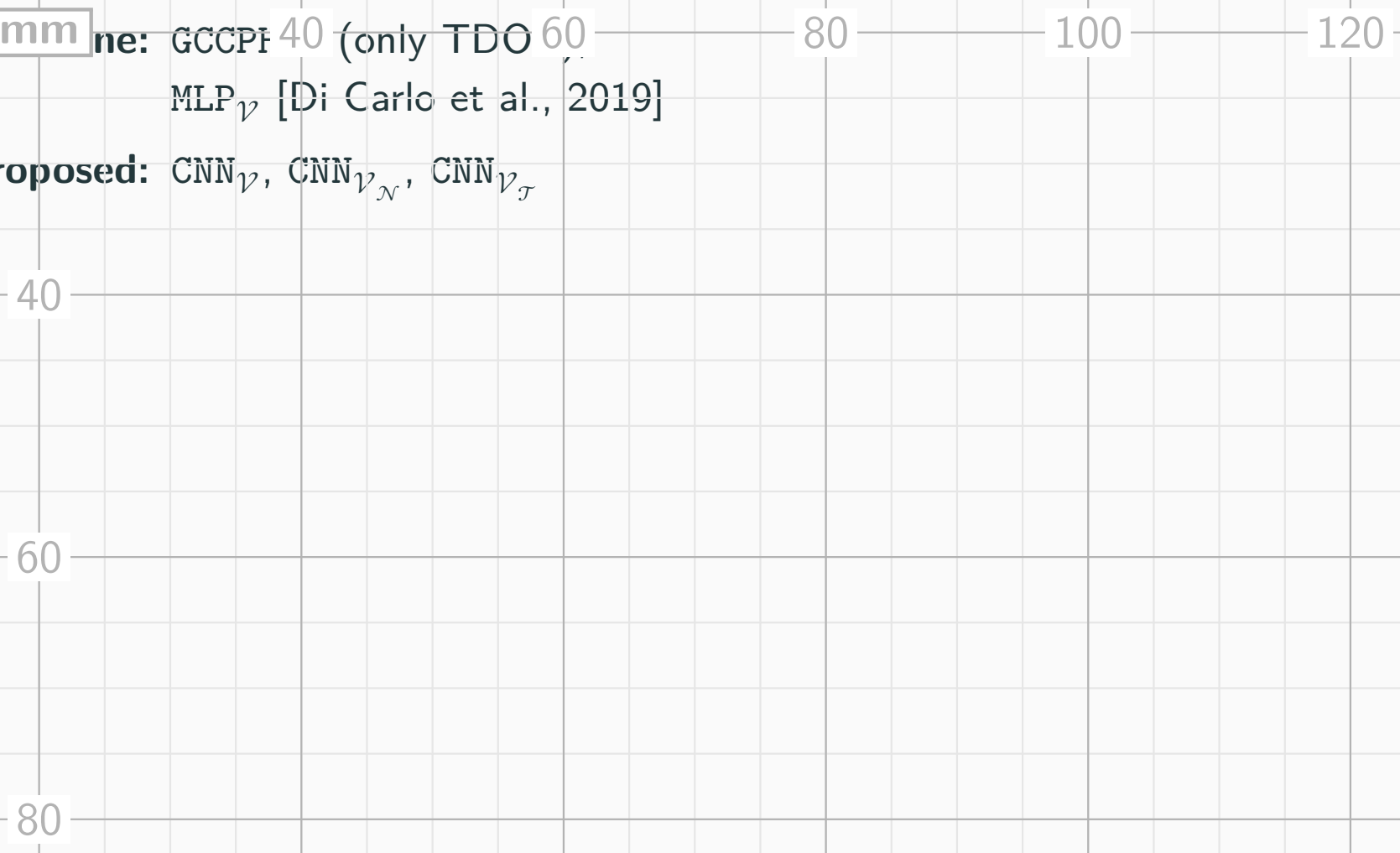
} Generative models $\leftarrow$ similar for data fusion

MDN [Bishop, 1994]

Lantern- Experiments & Resuls

mm ne: GCCPH (only TDO
$\mathrm{MLP}_{\mathcal{V}}$ [Di Carlo et al., 2019]

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



Lantern- Experiments & Resuls

mm ne: GCCPI (only TDO
$\mathrm{MLP}_{\mathcal{V}}$ [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)

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:

- random RT60, random SNR
- broadband source (wn)
- instantaneous RTF

Test: similar to train

Lantern- Experiments & Resuls

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

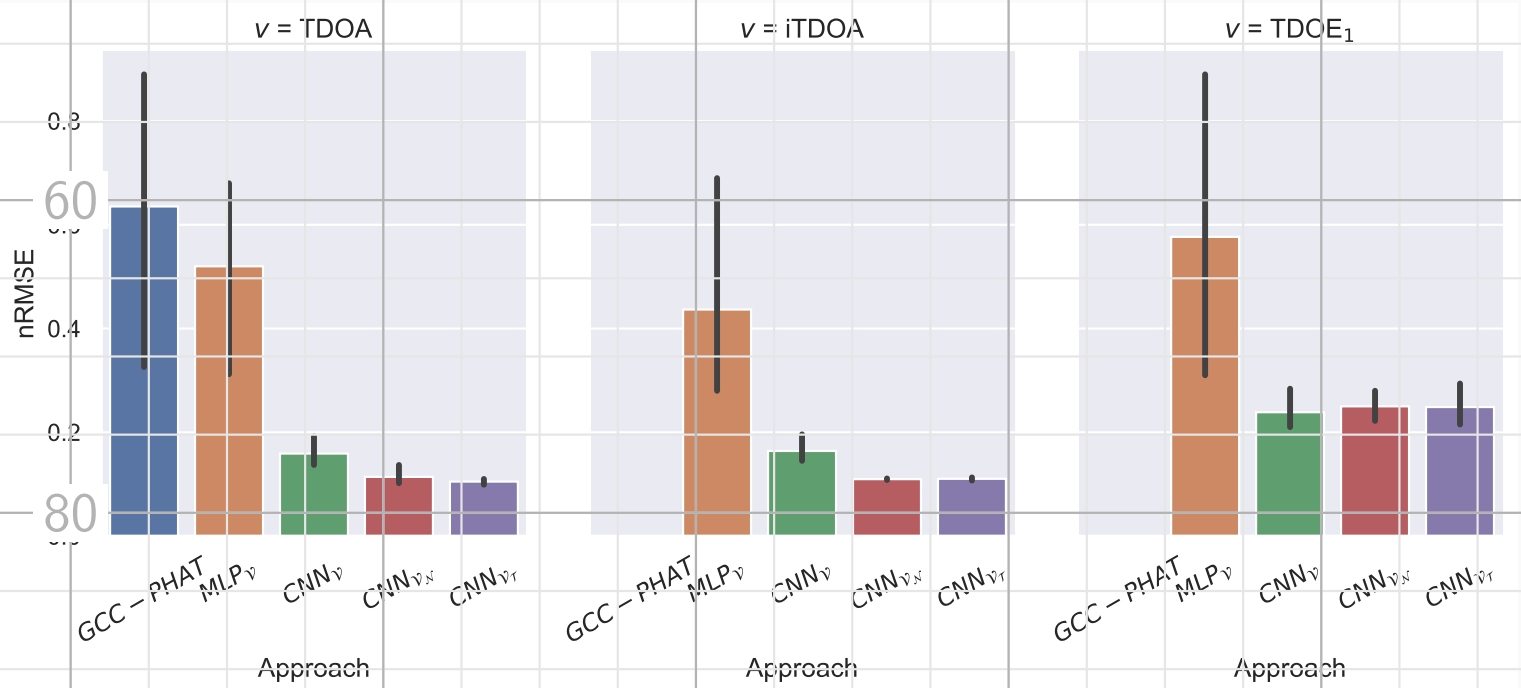
Proposed: $\text{CNN}_{\mathcal{V}}$, $\text{CNN}_{\mathcal{V}_{\mathcal{N}}}$, $\text{CNN}_{\mathcal{V}_{\mathcal{T}}}$

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

Train:

- random RT60, random SNR
- broadband source (wn)
- instantaneous RTF

Test: similar to train



Lantern- Experiments & Resuls

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

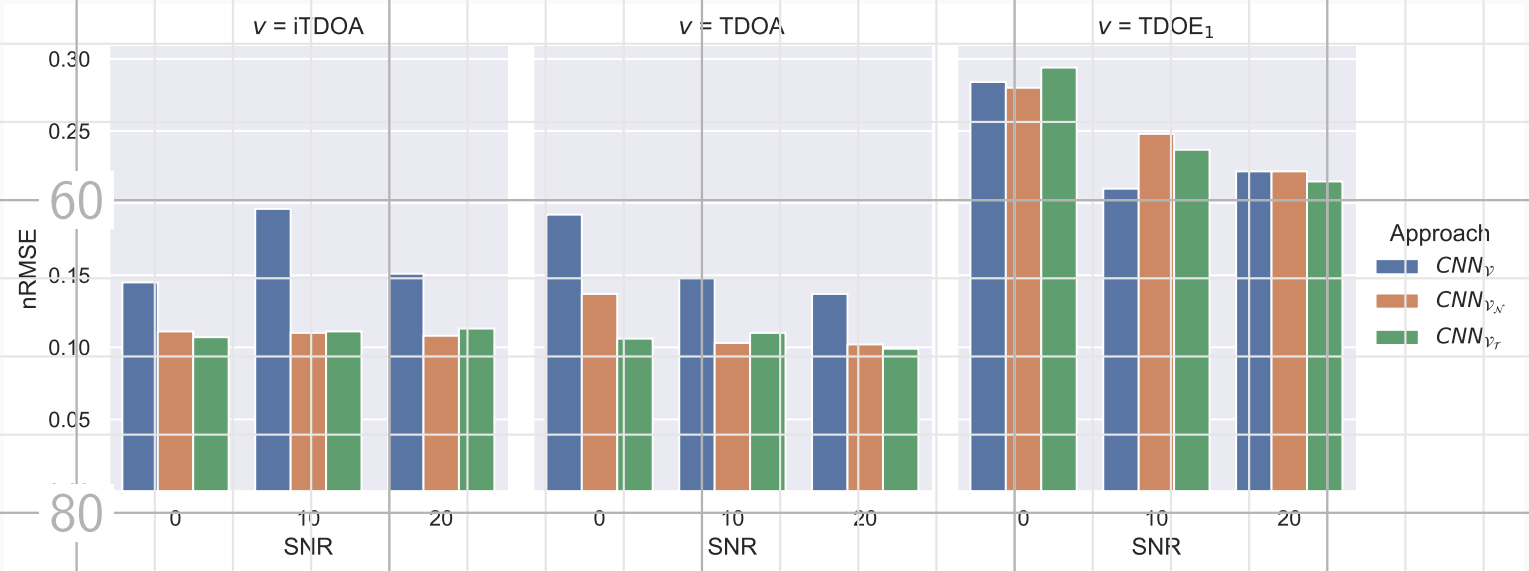
Proposed: $\text{CNN}_{\mathcal{V}}$, $\text{CNN}_{\mathcal{V}_{\mathcal{N}}}$, $\text{CNN}_{\mathcal{V}_{\mathcal{T}}}$

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

Train:

- random RT60, random SNR
- broadband source (wn)
- instantaneous RTF

Test: similar to train



✓ Generative better than Gaussian $\approx$ Student-T

✗ Bigger error on TDOE

References i

Bishop, C. M. (1994).

Mixture density networks.

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.

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.

References ii

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

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.

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

References iii

Perotin, L., Défossez, A., Vincent, E., Serizel, R., and Guérin, A. (2019). **Regression versus classification for neural network based audio source localization.** In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 343–347. IEEE.