Echo-aware signal processing for audio scene analysis

Diego Di Carlo

December 3, 2020

PhD supervisors: Antoine Deleforge
Nancy Bertin

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

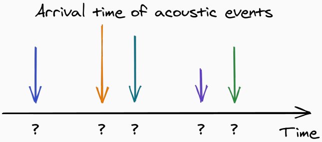
Université de Rennes 1, IRISA/INRIA, Panama research group

Acoustic Echo Estimation

Acoustic Echo Retrieval

Estimating early (strong) reflections for microphones recordings, i.e.,

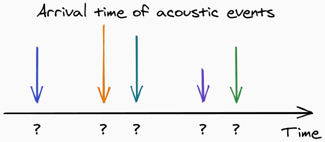
$$\{\tilde{x}_i\}_i \longrightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$



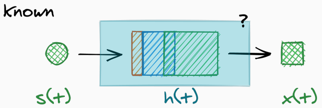
Acoustic Echo Retrieval

Estimating early (strong) reflections for microphones recordings, i.e.,

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



Two scenarios:

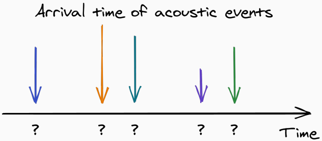


- **intrusive** or specific setups
- **non-blind** problem
  (**Applications:** sonar, measurements, etc.)

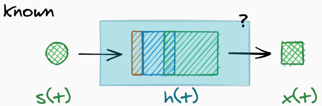
Acoustic Echo Retrieval

Estimating early (strong) reflections for microphones recordings, i.e.,

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



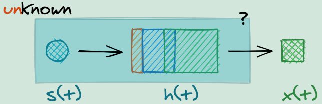
Two scenarios:



intrusive or specific setups

non-blind problem

(**Applications:** sonar, measurements, etc.)



passive and more common setups

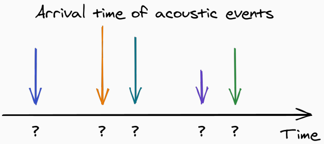
blind inverse problem (harder)

(**Applications:** recording on smart speakers, etc.)

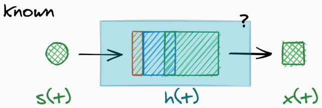
Acoustic Echo Retrieval

Estimating early (strong) reflections for microphones recordings, i.e.,

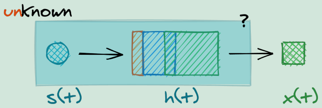
$$\{\tilde{x}_i\}_i \longrightarrow \{\tau_i^{(r)}, \alpha_i^{(r)}\}_{i,r}$$



Two scenarios:



- **intrusive** or specific setups
- **non-blind** problem
  (**Applications:** sonar, measurements, etc.)

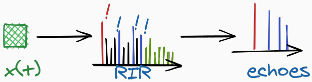


- **passive** and more common setups
- **blind inverse** problem (harder)
  (**Applications:** recording on smart speakers, etc.)

Our case: signal source and passive microphone array

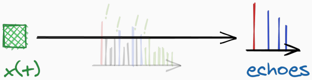
Passive Acoustic Echo Retrieval

RIR-based approaches



1. Discrete optimization $\Longrightarrow$ RIRs
2. Peak picking $\Longrightarrow$ Echoes

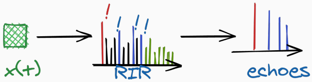
RIR-agnostic approaches



1. Direct estimation of $\left\{\tau_i^{(r)}, \alpha_i^{(r)}\right\}$ e.g., with maximum-likelihood

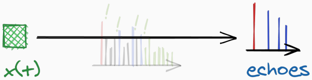
Passive Acoustic Echo Retrieval

RIR-based approaches



1. Discrete optimization $\Longrightarrow$ RIRs
2. Peak picking $\Longrightarrow$ Echoes

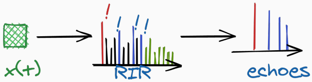
RIR-agnostic approaches



1. Direct estimation of $\left\{\tau_i^{(r)}, \alpha_i^{(r)}\right\}$ e.g., with maximum-likelihood

Passive Acoustic Echo Retrieval

RIR-based approaches

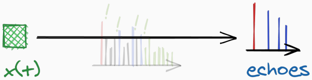


1. Discrete optimization $\implies$ RIRs
2. Peak picking $\implies$ Echoes

✓ BCE is well and known studied

✓ reasonably good for some application
[Crocco and Del Bue, 2016]

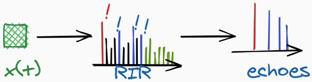
RIR-agnostic approaches



1. Direct estimation of $\left\{\tau_i^{(r)}, \alpha_i^{(r)}\right\}$ e.g., with maximum-likelihood

Passive Acoustic Echo Retrieval

RIR-based approaches



1. Discrete optimization $\implies$ RIRs
2. Peak picking $\implies$ Echoes

✓ BCE is well and known studied

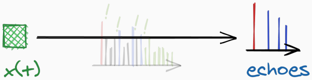
✓ reasonably good for some application
[Crocco and Del Bue, 2016]

✗ Full RIRs need to be estimated

✗ Peak picking has hyperparameters

✗ Issues due to *discrete* estimation

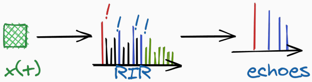
RIR-agnostic approaches



1. Direct estimation of $\left\{\tau_i^{(r)}, \alpha_i^{(r)}\right\}$ e.g., with maximum-likelihood

Passive Acoustic Echo Retrieval

RIR-based approaches



1. Discrete optimization $\Longrightarrow$ RIRs
2. Peak picking $\Longrightarrow$ Echoes

✓ BCE is well and known studied

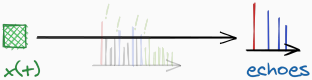
✓ reasonably good for some application
[Crocco and Del Bue, 2016]

✗ Full RIRs need to be estimated

✗ Peak picking has hyperparameters

✗ Issues due to *discrete* estimation

RIR-agnostic approaches



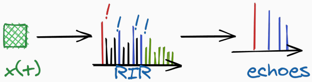
1. Direct estimation of $\left\{\tau_i^{(r)}, \alpha_i^{(r)}\right\}$ e.g., with maximum-likelihood

✓ No full RIRs & no peak picking
- → lower complexity
- → less hyperparameters

✗ exploratory
(few works on audio)

Passive Acoustic Echo Retrieval

RIR-based approaches



1. Discrete optimization $\implies$ RIRs
2. Peak picking $\implies$ Echoes

✓ BCE is well and known studied

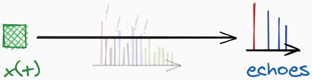
✓ reasonably good for some application
[Crocco and Del Bue, 2016]

✗ Full RIRs need to be estimated

✗ Peak picking has hyperparameters

✗ Issues due to *discrete* estimation

RIR-agnostic approaches



1. Direct estimation of $\left\{\tau_i^{(r)}, \alpha_i^{(r)}\right\}$ e.g., with maximum-likelihood

✓ No full RIRs & no peak picking

→ lower complexity
→ less hyperparameters

✗ exploratory
(few works on audio)

Proposed appoach RIR-agnostic & continuous:

1. Analytical approach
2. Learning-based approach

(Discrete) RIR-based methods: the State of the Art

Key ingredient – *Cross relation identity*

Signal model

$$x_1 = h_1 \star x$$
$$x_2 = h_2 \star x$$

(Discrete) RIR-based methods: the State of the Art

Key ingredient – *Cross relation identity*

Convolving with filters:

$$h_2 \star x_1 = h_2 \star h_1 \star x$$
$$h_1 \star x_2 = h_1 \star h_2 \star x$$

(Discrete) RIR-based methods: the State of the Art

Key ingredient – *Cross relation identity*

Commutativity of convolution:

$$h_2 \star x_1 = h_2 \star h_1 \star x$$
$$h_1 \star x_2 = h_2 \star h_1 \star x$$

(Discrete) RIR-based methods: the State of the Art

Key ingredient – *Cross relation identity*

Subtraction

$$\left.\begin{aligned} h_2 \star x_1 &= h_2 \star h_1 \star x \\ h_1 \star x_2 &= h_2 \star h_1 \star x \end{aligned}\right\} \rightarrow x_1 \star h_2 - x_2 \star h_1 = 0$$

(Discrete) RIR-based methods: the State of the Art

Key ingredient – *Cross relation identity*

Subtraction

$$\left.\begin{aligned} h_2 \star x_1 &= h_2 \star h_1 \star x \\ h_1 \star x_2 &= h_2 \star h_1 \star x \end{aligned}\right\} \rightarrow x_1 \star h_2 - x_2 \star h_1 = 0$$

Ideas:

1. Echo TOAs $\propto$ sampling frequency
2. Find echoes $\rightarrow$ **find sparse non-negative vectors** h_1, h_2 of length L
3. Modeled as **Lasso**-like problem

(Discrete) RIR-based methods: the State of the Art

Key ingredient – *Cross relation identity*

Subtraction

$$\left.\begin{aligned} h_2 \star x_1 &= h_2 \star h_1 \star x \\ h_1 \star x_2 &= h_2 \star h_1 \star x \end{aligned}\right\} \rightarrow x_1 \star h_2 - x_2 \star h_1 = 0$$

Ideas:

1. Echo TOAs $\propto$ sampling frequency
2. Find echoes $\rightarrow$ **find sparse non-negative vectors** h_1, h_2 of length L
3. Modeled as **Lasso**-like problem

$$= \texttt{Toeplitz}(x_i) h_j \in \mathcal{O}(L^2)$$

$$\hat{h}_1, \hat{h}_2 \in \underset{h_1, h_2 \in \mathbf{R}^n}{\arg\min} \; \|x_1 \star h_2 - x_2 \star h_1\|_2^2 + \lambda \mathcal{P}(h_1, h_2) \quad \text{s.t.} \quad \mathcal{C}(h_1, h_2)$$

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

$\mathcal{C}(h_1, h_2) \longrightarrow$ constraints e.g. nonnegativity, anchor

(Discrete) RIR-based methods: the State of the Art

Key ingredient – *Cross relation identity*

Subtraction

$$\left.\begin{aligned} h_2 \star x_1 &= h_2 \star h_1 \star x \\ h_1 \star x_2 &= h_2 \star h_1 \star x \end{aligned}\right\} \rightarrow x_1 \star h_2 - x_2 \star h_1 = 0$$

Ideas:

1. Echo TOAs $\propto$ sampling frequency
2. Find echoes $\rightarrow$ **find sparse non-negative vectors** h_1, h_2 of length L
3. Modeled as **Lasso**-like problem

$$= \texttt{Toeplitz}(x_i)h_j \in \mathcal{O}(L^2)$$

$$\hat{h}_1, \hat{h}_2 \in \underset{h_1, h_2 \in \mathbf{R}^n}{\arg\min} \; \|x_1 \star h_2 - x_2 \star h_1\|_2^2 + \lambda \mathcal{P}(h_1, h_2) \quad \text{s.t.} \quad \mathcal{C}(h_1, h_2)$$

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

$\mathcal{C}(h_1, h_2) \longrightarrow$ constraints e.g. nonnegativity, anchor

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

Proposed approach: analytical & continuous

C. Elvira.

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

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

Proposed approach: analytical & continuous

C. Elvira.

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

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

Observation 2: $\mathscr{F}\delta_{\mathrm{echo}}$ **is known in closed-form**

Proposed approach: analytical & continuous

C. Elvira.

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

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

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

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

$$\mathbf{X}_i = \mathtt{DFT}(x_i) \simeq \mathcal{F}\tilde{x}_i(nF_s) \qquad n = 0 \ldots N-1$$

Proposed approach: analytical & continuous

C. Elvira.

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

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

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

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

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

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

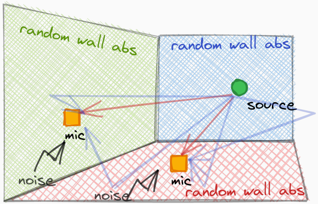
$\sim$ **Lasso**, but $\mathcal{F}h_i(f)$ is a continuous function $\rightarrow$ **BLasso** [Azais et al., 2015]

- ✓No huge matrix
- ✓ Solutions is a train of Dirac
- ✓ No peak picking
- ✓ Perfect in noiseless & synthetic case

Experimental results

Syntetic Dataset at 16 kHz

- 2 microphones, 1 sound source (noise and speech)
- shoebox with random geometry
- $\mathcal{D}^{\mathrm{SNR}}$: $\mathtt{SNR} \in [0, 20]$ dB, $\mathtt{RT}_{60} = 400$ ms
- $\mathcal{D}^{\mathrm{RT}60}$: $\mathtt{RT}_{60} = [100, 1000]$ ms, $\mathtt{SNR} = 20$ dB



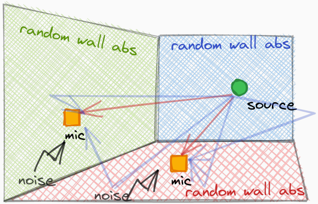
[1] [Lin et al., 2007]

[2] [Crocco and Del Bue, 2015]

Experimental results

Syntetic Dataset at 16 kHz

- 2 microphones, 1 sound source (noise and speech)
- shoebox with random geometry
- $\mathcal{D}^{\text{SNR}}$: $\texttt{SNR} \in [0, 20]$ dB, $\texttt{RT}_{60} = 400$ ms
- $\mathcal{D}^{\text{RT}60}$: $\texttt{RT}_{60} = [100, 1000]$ ms, $\texttt{SNR} = 20$ dB



Baselines: discrete RIR-based methods based on LASSO

- BSN: Blind, Sparse and Non-negative[1]
- IL1C: iteratively-weighted ℓ_1 constraint[2] $\rightarrow$ State of the Art

hyperparameters and peak-picking tuned via cross-validation

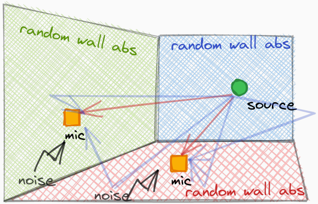
[1] [Lin et al., 2007]

[2] [Crocco and Del Bue, 2015]

Experimental results

Syntetic Dataset at 16 kHz

- 2 microphones, 1 sound source (noise and speech)
- shoebox with random geometry
- $\mathcal{D}^{\mathrm{SNR}}$: $\mathtt{SNR} \in [0, 20]$ dB, $\mathtt{RT}_{60} = 400$ ms
- $\mathcal{D}^{\mathrm{RT}60}$: $\mathtt{RT}_{60} = [100, 1000]$ ms, $\mathtt{SNR} = 20$ dB



Baselines: discrete RIR-based methods based on LASSO

- BSN: Blind, Sparse and Non-negative[1]
- IL1C: iteratively-weighted ℓ_1 constraint[2] $\rightarrow$ State of the Art

hyperparameters and peak-picking tuned via cross-validation

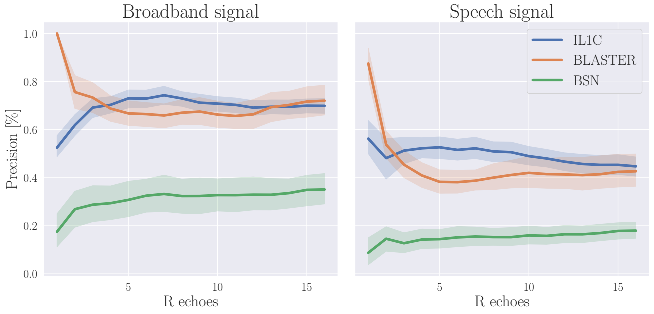
Proposed method: Blind and Sparse Technique for Echo Retrieval (Blaster)

[1] [Lin et al., 2007]

[2] [Crocco and Del Bue, 2015]

Precision per # of echoes

Metric: Precision = how many estimated echoes are correct (within 2 samples)



(RT_{60} = 400 ms and SNR = 20 dB.)

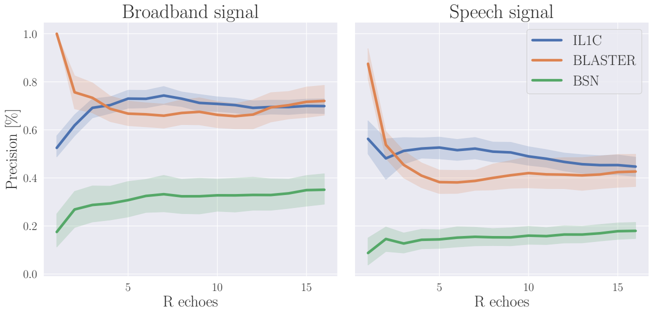
✗ Sensitive kto # echoes

✗ Sensitive source signal

✓ Good for 2 echoes [Scheibler et al., 2018, Di Carlo et al., 2019]

Precision per # of echoes

Metric: Precision = how many estimated echoes are correct (within 2 samples)



($RT_{60} = 400$ ms and SNR $= 20$ dB.)

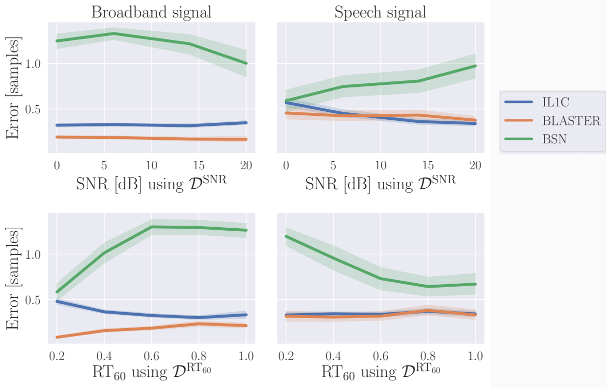
✗ Sensitive kto # echoes

✗ Sensitive source signal

✓ Good for 2 echoes [Scheibler et al., 2018, Di Carlo et al., 2019]

Error per Dataset/Signal while recovering 7 echoes

Metric: RMSE on matched echoes = error on the correct guess



(Fs = 16 kHz)

✓Lower RMSE ✓ Robustness to SNR and RT_{60} ✗ Source signal dependent

Learning-based approach

Idea:

1. Use virtually supervised deep learning models
2. Estimate first echo (simple but important) (↰ used in the next section)
3. Only 2 microphones attending 1 sound source

Motivations:

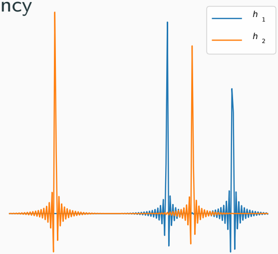
- $x_i \rightarrow \tau_i^{(r)}$ is difficult, while $\tau_i^{(r)} \rightarrow x_i$ "is not"
  $\rightarrow$ acoustic simulators: mic/src/room geometry $\longrightarrow \left\{\tau_i^{(r)}, \alpha_i^{(r)}\right\}, \ \tilde{h}_i, \quad \tilde{x}_i$
- Acoustic simulator are "simple", versatile and fast
  $\rightarrow$ allow to create large dataset

Learning-based approach

Inputs:
Interchannel level and phase difference features from

$$R[f] = \underset{t}{\text{avg.}} \frac{X_2[f,t]}{X_1[f,t]} \approx \underset{t}{\text{avg.}} \frac{H_2[f]S[f,t]}{H_1[f]S[f,t]}$$

$\approx$ the relative transfer function $\rightarrow$ remove source dependency



Outputs:
Inter and intra Time Difference of Arrivals (TDOAs)

HP: close-surface scenario: first $\Leftrightarrow$ strongest echo

Loss Function

1. RMSE (Multi-label regression) $\rightarrow$ TDOAs
2. Gaussian log-likelihood $\rightarrow \{\mu_\tau, \sigma_\tau^2\}\ \forall \tau \in$ TDOAs
3. Student log-likelihood $\rightarrow \{\mu_\tau, \lambda_\tau, \nu_\tau\}\ \forall \tau \in$ TDOAs

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

Learning-based approach

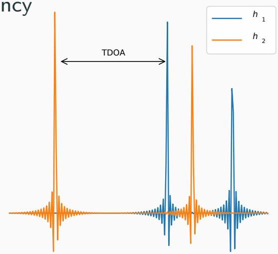
Inputs:
Interchannel level and phase difference features from

$$R[f] = \underset{t}{\text{avg.}} \frac{X_2[f,t]}{X_1[f,t]} \approx \underset{t}{\text{avg.}} \frac{H_2[f]S[f,t]}{H_1[f]S[f,t]}$$

$\approx$ the relative transfer function $\rightarrow$ remove source dependency

Outputs:
Inter and intra Time Difference of Arrivals (TDOAs)

HP: close-surface scenario: first $\Leftrightarrow$ strongest echo



Loss Function

1. RMSE (Multi-label regression) $\rightarrow$ TDOAs
2. Gaussian log-likelihood $\rightarrow \{\mu_\tau, \sigma^2_\tau\}\ \forall \tau \in$ TDOAs
3. Student log-likelihood $\rightarrow \{\mu_\tau, \lambda_\tau, \nu_\tau\}\ \forall \tau \in$ TDOAs

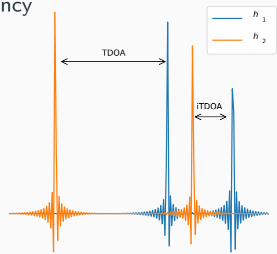
Architecture: MLP, CNN [Chakrabarty and Habets, 2017, Nguyen et al., 2018]

Learning-based approach

Inputs:
Interchannel level and phase difference features from

$$R[f] = \underset{t}{\text{avg.}} \frac{X_2[f,t]}{X_1[f,t]} \approx \underset{t}{\text{avg.}} \frac{H_2[f]S[f,t]}{H_1[f]S[f,t]}$$

$\approx$ the relative transfer function $\rightarrow$ remove source dependency



Outputs:
Inter and intra Time Difference of Arrivals (TDOAs)

HP: close-surface scenario: first $\Leftrightarrow$ strongest echo

Loss Function

1. RMSE (Multi-label regression) $\rightarrow$ TDOAs
2. Gaussian log-likelihood $\rightarrow \{\mu_\tau, \sigma^2_\tau\}\ \forall\tau \in$ TDOAs
3. Student log-likelihood $\rightarrow \{\mu_\tau, \lambda_\tau, \nu_\tau\}\ \forall\tau \in$ TDOAs

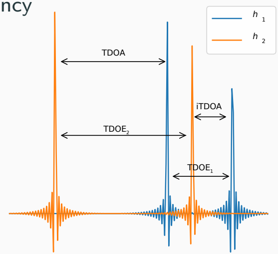
Architecture: MLP, CNN [Chakrabarty and Habets, 2017, Nguyen et al., 2018]

Learning-based approach

Inputs:
Interchannel level and phase difference features from

$$R[f] = \underset{t}{\text{avg.}} \frac{X_2[f,t]}{X_1[f,t]} \approx \underset{t}{\text{avg.}} \frac{H_2[f]S[f,t]}{H_1[f]S[f,t]}$$

$\approx$ the relative transfer function $\rightarrow$ remove source dependency



Outputs:
Inter and intra Time Difference of Arrivals (TDOAs)

HP: close-surface scenario: first $\Leftrightarrow$ strongest echo

Loss Function

1. RMSE (Multi-label regression) $\rightarrow$ TDOAs
2. Gaussian log-likelihood $\rightarrow \{\mu_\tau, \sigma^2_\tau\}\ \forall\tau \in$ TDOAs
3. Student log-likelihood $\rightarrow \{\mu_\tau, \lambda_\tau, \nu_\tau\}\ \forall\tau \in$ TDOAs

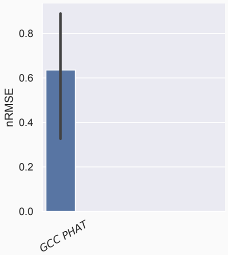
Architecture: MLP, CNN [Chakrabarty and Habets, 2017, Nguyen et al., 2018]

Experimental results

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

Baseline: GCC PHAT [Knapp and Carter, 1976]

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



More echoes

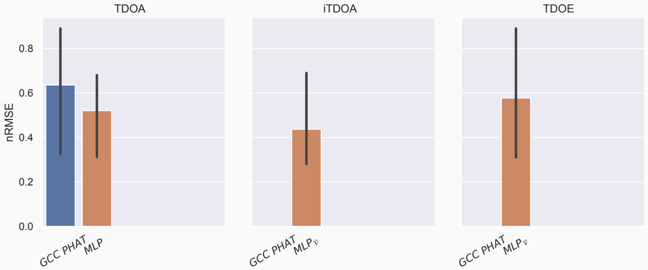
Real data

Experimental results

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

Baseline: GCC PHAT [Knapp and Carter, 1976]

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



Observation:

✓ MLP outperforms GCC PHAT on TDOA estimation

More echoes

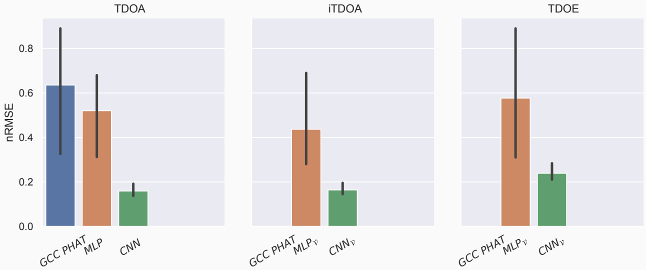
Real data

Experimental results

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

Baseline: GCC PHAT [Knapp and Carter, 1976]

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



Observation:

- ✓ MLP outperforms GCC PHAT on TDOA estimation
- ✓ CNN outperforms MLP (lower error and smaller variance)

More echoes

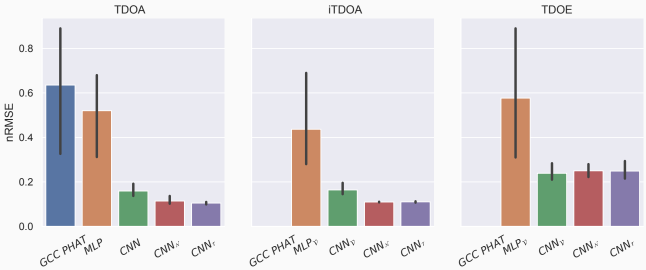
Real data

Experimental results

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

Baseline: `GCC PHAT` [Knapp and Carter, 1976]

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



Observation:

- ✓ `MLP` outperforms `GCC PHAT` on TDOA estimation
- ✓ `CNN` outperforms `MLP` (lower error and smaller variance)
- ✓ $\texttt{CNN}_{\mathcal{N}}$ and $\texttt{CNN}_{\mathcal{T}}$ outperform `CNN` (lower error and smaller variance)
- ✗ TDOA between DP and 1st echo more difficult

More echoes

Real data

Aissa-El-Bey, A. and Abed-Meraim, K. (2008).
Blind simo channel identification using a sparsity criterion.
In *2008 IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*, pages 271–275. IEEE.

Azais, J.-M., De Castro, Y., and Gamboa, F. (2015).
Spike detection from inaccurate samplings.
Applied and Computational Harmonic Analysis, 38(2):177–195.

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

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

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

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.

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., Habets, E. A., Kellermann, W., and Naylor, P. A. (2013).
Blind system identification using sparse learning for tdoa estimation of room reflections.
IEEE Signal Processing Letters, 20(7):653–656.

References iii

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

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

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

Scheibler, R., Di Carlo, D., Deleforge, A., and Dokmanić, I. (2018).
Separake: Source separation with a little help from echoes.
In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6897–6901. IEEE.

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