A Hybrid Model for Weakly-Supervised Speech Dereverberation

Louis Bahrman, Mathieu Fontaine, Gaël Richard

LTCI, Télécom Paris, Institut Polytechnique de Paris, France

Summary

Context

- DNN-based approaches for dereverberation often rely on dry/wet pairs
- Anechoïc data is rare and expensive
- Metrics-based weak supervision fails to generalize to other metrics

Main takeways

- Reverberation-based weak supervision for dereverberation
- Dereverberation supervised by only pairs of wet and RT_{60}

Code

Experiments

Data

- WSJ1 $\star$ Synthethic RIRs

Speech model

- BiLSTM (Weninger et al. 15)
  Simple model used in the baseline
- FullSubNet (FSN) (Hao et al. 21)
  Able to process phase, already used in reverberation-aware dereverberation.

Supervision variants

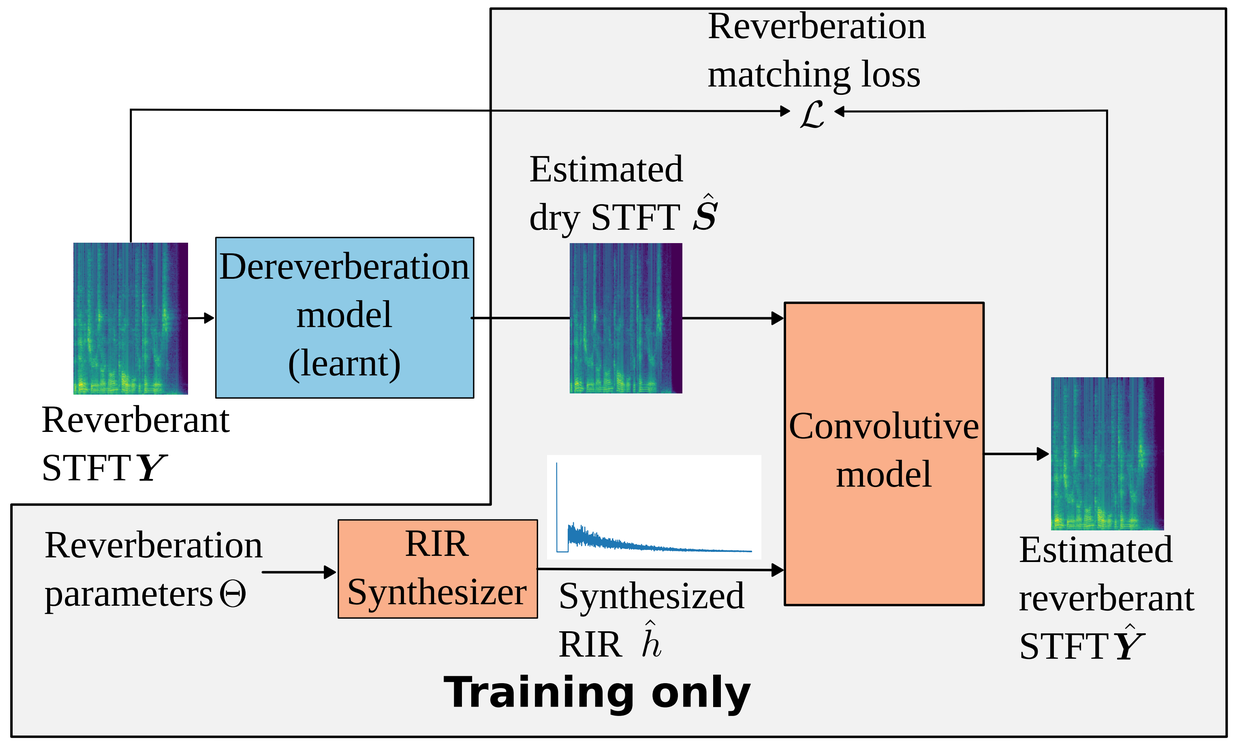
- Strong supervision
  - Using dry signal s:
    Speech model paired training loss
  - Using RIR h:
    Oracle RIR synthesis model
- Weak supervision (WS)

Baseline

Metrics-based weak supervision:
MetricGAN-U (Fu et al. 2022)

- Speech model (generator): BiLSTM
- Weak supervision (discriminator): SRMR metric learnt by a DNN

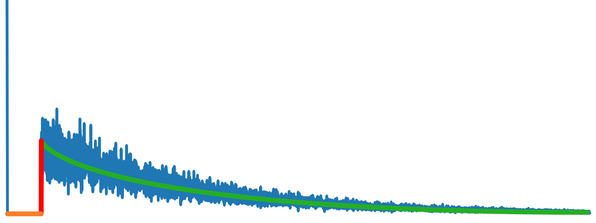
Method



RIR synthesizer

Inspired from Polack's model

$$\hat{h}_\Theta(n) = \begin{cases} |b(n)| e^{-\frac{3\ln(10)}{\mathrm{RT}_{60} f_s} n} & \text{if } n > 2n_m \\ 1 & \text{if } n = 0 \\ 0 & \text{otherwise} \end{cases}$$



Parameters

- RT_{60} : Reverberation time
- $b(n) \sim \mathcal{N}(0, \sigma^2)$
- $n_m = \frac{4Vf_s}{cA}$: Mixing time

Weak supervision variants

- $\Theta \triangleq \{\mathrm{RT}_{60}, \sigma, V, A\}$: all the parameters
- $\{\mathrm{RT}_{60}, \sigma\}$: fixed mixing time at 20 ms after the peak (mean over training dataset)
- $\{\mathrm{RT}_{60}\}$: fixed mixing time at 20 ms, fixed $\sigma = 0.02$ (median over the training dataset).

Convolutive model

Equivalent to $\hat{y} = \hat{s} \star \hat{h}_\Theta$ in time-frequency

$$\mathcal{C}(\hat{\boldsymbol{S}}, \hat{h}_\Theta) = \sum_{f'=f-F'}^{f+F'} \sum_{t'=0}^{\min(t;T_h)} \hat{\mathcal{H}}_{f,f',t'} \hat{S}_{f',t-t'},$$

with $\hat{\mathcal{H}}_{f,f',t'}$ constructed from $\hat{h}_\Theta(n)$. (Avargel and Cohen 07)

Reverberation matching loss

- Compare $\hat{\boldsymbol{Y}}_{f,t} \triangleq \mathcal{C}(\hat{\boldsymbol{S}}, \hat{h}_\Theta)$ and $\boldsymbol{Y}$
- Sum of two losses: $\mathcal{L} = \mathcal{L}_{\mathbb{C}} + \lambda \mathcal{L}_M$:

$$\mathcal{L}_{\mathbb{C}} = \sum_{f,t} \left[|\hat{Y}_{f,t} - Y_{f,t}|^2 \right]$$

$$\mathcal{L}_M = \sum_{f,t} \left[\left| \log\left(\frac{1+\gamma|\hat{Y}_{f,t}|}{1+\gamma|Y_{f,t}|} \right) \right|^2 \right]$$

With $\gamma = \lambda = 1$

Results

Model	WS?	Supervision	SISDR	ESTOI	WB-PESQ	SRMR
FSN	✗	cRM	5.6 ± 3.9	0.84 ± 0.09	2.55 ± 0.68	8.2 ± 3.5
		h	4.3 ± 4.0	0.77 ± 0.12	2.03 ± 0.69	7.8 ± 3.1
	✓	$\{\mathrm{RT}_{60}, \sigma, V, A\}$	1.0 ± 2.5	$\mathbf{0.71 \pm 0.14}$	$\mathbf{1.80 \pm 0.70}$	6.9 ± 2.8
		$\{\mathrm{RT}_{60}, \sigma\}$	1.1 ± 2.5	0.70 ± 0.14	1.78 ± 0.69	7.0 ± 2.8
		$\{\mathrm{RT}_{60}\}$	$\mathbf{2.9 \pm 3.4}$	0.71 ± 0.15	1.78 ± 0.71	6.9 ± 2.8
BiLSTM	✗	$\lvert S_{f,t}\rvert^2, \forall f,t$	1.3 ± 4.2	0.78 ± 0.12	2.25 ± 0.79	7.9 ± 3.0
		h	0.1 ± 4.1	0.70 ± 0.15	1.80 ± 0.70	7.2 ± 2.7
	✓	$\{\mathrm{RT}_{60}, \sigma, V, A\}$	0.8 ± 4.0	0.70 ± 0.15	1.81 ± 0.74	6.9 ± 2.7
		$\{\mathrm{RT}_{60}, \sigma\}$	0.7 ± 4.0	0.70 ± 0.15	1.78 ± 0.72	6.8 ± 2.7
		$\{\mathrm{RT}_{60}\}$	$\mathbf{1.6 \pm 3.7}$	$\mathbf{0.71 \pm 0.15}$	$\mathbf{1.84 \pm 0.75}$	6.9 ± 2.8
BiLSTM	✓	SRMR (baseline)	-1.5 ± 3.4	0.64 ± 0.18	1.78 ± 0.74	$\mathbf{10.9 \pm 4.2}$
Reverberant			-1.3 ± 3.4	0.68 ± 0.16	1.75 ± 0.74	6.9 ± 2.9