# SPEECH DEREVERBERATION CONSTRAINED ON ROOM IMPULSE RESPONSE CHARACTERISTICS

**Louis Bahrman[1], Mathieu Fontaine[1], Jonathan Le Roux[2], Gaël Richard[1]**

[1]LTCI, Télécom Paris, IP-Paris, France; [2]MERL, Cambridge, MA, USA

## SUMMARY

### Context

- DNN-based approaches for dereverberation are often not consistent with room acoustics
- We introduce a physical coherence loss which regularizes the training in a physically consistent manner

### Main takeways

- Dereverberation performance comparable with the baseline, at no additional cost
- Extraction of an RIR from a dereverberation model

### Code

## EXPERIMENTS

### Tasks

- Dereverberation
- RIR estimation

### Data

- Training data: WSJ0, synthethic RIRs.
- Test data: WSJ0 or Librispeech clean, synthethic RIRs drawn from 2 settings:
  - Matched: Similar room parameters
  - Mismatched: Harder room parameters

**Baseline** FullSubNet (Hao et al. 21) (FSN)

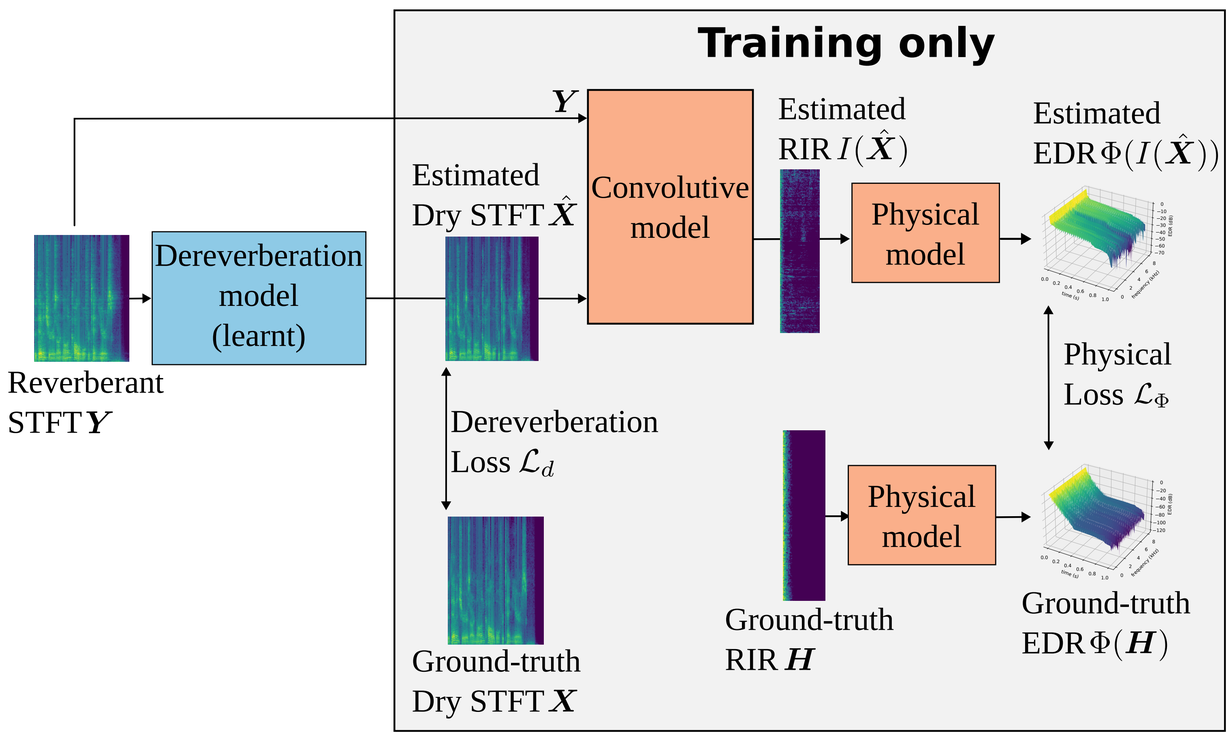
### RIR estimation metrics

- EDC-Fourier: Energy Decay Curve distance, deconvolution as $\mathrm{IDFT}\left[\frac{\mathrm{DFT}(y_n)}{\mathrm{DFT}(x_n)}\right]$
- EDR: Energy Decay Relief distance

## RESULTS: RIR ESTIMATION

| WSJ0 | | EDC (↓) | EDR (↓) | |
|---|---|---|---|---|
| | | Fourier | Subband | Crossband |
| Matched | FSN | 66.2 ± 28 | 39.0 ± 12 | 99.6 ± 24 |
| | +SB | 60.5 ± 21 | **32.7 ± 7** | 100.7 ± 22 |
| | +CSB | **52.6 ± 24** | 34.1 ± 13 | **97.8 ± 24** |
| | +SSB | 76.4 ± 23 | 39.9 ± 10 | 102.9 ± 23 |
| | +3B | 67.1 ± 27 | 38.7 ± 11 | 100.0 ± 24 |
| | dry | 0.0 ± 0 | 36.7 ± 10 | 75.0 ± 19 |
| Mismatched | FSN | 86.4 ± 15 | 37.8 ± 7 | 116.7 ± 6 |
| | +SB | 66.3 ± 16 | 27.6 ± 6 | 114.9 ± 7 |
| | +CSB | **63.1 ± 16** | **25.6 ± 4** | **113.6 ± 7** |
| | +SSB | 86.2 ± 14 | 40.4 ± 8 | 117.9 ± 6 |
| | +3B | 86.8 ± 15 | 37.5 ± 7 | 117.2 ± 6 |
| | dry | 0.0 ± 0 | 38.4 ± 8 | 84.4 ± 12 |

## METHOD




### Convolutive model

Reverberation in the T-F domain

$$Y_{f,t} = \sum_{f'=0}^{F-1} \sum_{t'=-\infty}^{\infty} \mathcal{H}_{f,f',t'} X_{f',t-t'}$$

$\mathcal{H}_{f,f',t'}$ is a tridimensional representation of an RIR $\boldsymbol{H}$ (Avargel and Cohen 07).

RIR extraction steps

1. Compute the crossband filter for each $f$:

$$\mathcal{C}_f(\hat{\boldsymbol{X}}) = \arg\min_{\boldsymbol{C}_f} \left\| \bar{\hat{\boldsymbol{X}}}_f \boldsymbol{C}_f - \boldsymbol{Y}_f \right\|_2^2$$

   where $\boldsymbol{Y}_f$, $\bar{\hat{\boldsymbol{X}}}_f$ are constructed by concatenating respectively $Y_{f',t}$ and Toeplitz matrices of $\hat{X}_{f',t}$, at different crossbands $f' = f - F', \ldots, f + F'$.

2. Extract a representation of the estimated RIR STFT:

$$I(\hat{\boldsymbol{X}})_{f,t} = \sum_{f'=f-F'}^{f+F'} (-1)^{f'} \mathcal{C}_{f,f',t}(\hat{\boldsymbol{X}})$$

3. (Optional) Correct the Subband modeling error $\mathcal{E}_{f,t} = I(\boldsymbol{X})_{f,t} - H_{f,t}$, by spectral subtraction:

$$I(\hat{\boldsymbol{X}})^c_{f,t} = \left( |I(\hat{\boldsymbol{X}})_{f,t}|^2 - |\mathcal{E}_{f,t}|^2 \right)^{1/2} e^{j\angle I(\hat{\boldsymbol{X}})_{f,t}}$$

### Physical model

Given a ground-truth or estimated RIR STFT $\boldsymbol{R}$, compute the bandwise dB-scaled Energy Decay Relief $\Phi_{f,t}(\boldsymbol{R}) = 10 \log_{10} \frac{\mathrm{EDR}(\boldsymbol{R})_{f,t}}{\mathrm{EDR}(\boldsymbol{R})_{f,0}}$, where:

$$\mathrm{EDR}(\boldsymbol{R})_{f,t} \triangleq \sum_{t'=t}^{+\infty} |R_{f,t'}|^2$$

### Losses

- DNN Loss $\mathcal{L}_d$: Complex Ideal Ratio Mask between dereverberated and dry speech
- Physical coherence loss $\mathcal{L}_\Phi(\hat{\boldsymbol{R}}, \boldsymbol{R})$

$$= \sum_{f,t} \left| \Phi_{f,t}(\hat{\boldsymbol{R}}) - \Phi_{f,t}(\boldsymbol{R}) \right|^2 \mathbb{1}_{\{\Phi_{f,t}(\boldsymbol{R}) > -20\}}$$

Both losses are weighted using GradNorm.

### Variants

- Subband (SB)
- Symmetric Subband (SSB)
- Corrected Subband (CSB)
- 3-band (3B)

| Variant | Crossbands | Loss |
|---|---|---|
| SB | $F' = 0$ | $\mathcal{L}_\Phi(I(\hat{\boldsymbol{X}}), \boldsymbol{H})$ |
| SSB | $F' = 0$ | $\mathcal{L}_\Phi(I(\hat{\boldsymbol{X}}), I(\boldsymbol{X}))$ |
| CSB | $F' = 0$ | $\mathcal{L}_\Phi(I(\hat{\boldsymbol{X}})^c, I(\boldsymbol{X})^c)$ |
| 3B | $F' = 1$ | $\mathcal{L}_\Phi(I(\hat{\boldsymbol{X}}), \boldsymbol{H})$ |

## RESULTS: DEREVERBERATION

| | Matched RIRs | | | | Mismatched RIRs | | | |
|---|---|---|---|---|---|---|---|---|
| | WSJ0 | | LibriSpeech clean | | WSJ0 | | LibriSpeech clean | |
| | SISDR | WB-PESQ | SISDR | WB-PESQ | SISDR | WB-PESQ | SISDR | WB-PESQ |
| FSN | **5.1 ± 4.1** | 2.23 ± 0.60 | **3.1 ± 4.3** | 2.06 ± 0.55 | **0.9 ± 2.6** | 1.60 ± 0.21 | **–0.8 ± 3.4** | 1.53 ± 0.24 |
| + SB | 4.3 ± 4.2 | 2.10 ± 0.56 | 2.5 ± 4.6 | 1.98 ± 0.51 | –0.3 ± 2.9 | 1.46 ± 0.19 | –1.9 ± 3.5 | 1.42 ± 0.21 |
| + CSB | 4.2 ± 4.6 | 2.11 ± 0.65 | 2.2 ± 5.1 | 1.99 ± 0.59 | –0.7 ± 2.9 | 1.43 ± 0.18 | –2.4 ± 3.8 | 1.41 ± 0.21 |
| + SSB | 4.8 ± 4.1 | 2.19 ± 0.59 | 2.6 ± 4.5 | 1.99 ± 0.52 | 0.6 ± 2.7 | 1.57 ± 0.20 | –1.3 ± 3.8 | 1.49 ± 0.23 |
| + 3B | 4.9 ± 4.1 | **2.24 ± 0.60** | 2.9 ± 4.6 | **2.07 ± 0.57** | 0.7 ± 2.6 | **1.61 ± 0.21** | –1.0 ± 3.7 | **1.54 ± 0.25** |
| input | –0.2 ± 4.8 | 1.76 ± 0.67 | –1.0 ± 5.5 | 1.89 ± 0.76 | –4.5 ± 2.9 | 1.20 ± 0.11 | –5.2 ± 3.7 | 1.24 ± 0.16 |

STOI remains the same as the baseline across all variants