Speech dereverberation constrained on room impulse response characteristics

Bahrman, Louis; Fontaine, Mathieu; Le Roux, Jonathan; Richard, Gaël

TR2024-121 September 04, 2024

Abstract

Single-channel speech dereverberation aims at extracting a dry speech signal from a recording affected by the acoustic reflections in a room. However, most current deep learning-based approaches for speech dereverberation are not interpretable for room acoustics, and can be considered as black-box systems in that regard. In this work, we address this problem by regularizing the training loss using a novel physical coherence loss which encourages the room impulse response (RIR) induced by the dereverberated output of the model to match the acoustic properties of the room in which the signal was recorded. Our investigation demonstrates the preservation of the original dere- verberated signal alongside the provision of a more physically coherent RIR.

Interspeech 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Speech dereverberation constrained on room impulse response characteristics

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

[1]LTCI, Telecom Paris, Institut polytechnique de Paris, France
[2]Mitsubishi Electric Research Laboratories (MERL), USA
[1]`firstname.lastname@telecom-paris.fr`, [2]`leroux@merl.com`

Abstract

Single-channel speech dereverberation aims at extracting a dry speech signal from a recording affected by the acoustic reflections in a room. However, most current deep learning-based approaches for speech dereverberation are not interpretable for room acoustics, and can be considered as black-box systems in that regard. In this work, we address this problem by regularizing the training loss using a novel physical coherence loss which encourages the room impulse response (RIR) induced by the dereverberated output of the model to match the acoustic properties of the room in which the signal was recorded. Our investigation demonstrates the preservation of the original dereverberated signal alongside the provision of a more physically coherent RIR.

Index Terms: Speech dereverberation, hybrid deep learning, room acoustics, acoustic matching, speech processing

1. Introduction

An acoustic signal captured in a closed room comprises several correlated components: a more so-called direct-path signal and a combination of early reflections plus late reverberation collectively coined as reverberation signal. The reverberation phenomenon may not be desirable in speech recording as it lowers its perceptual intelligibility [1]. This justifies the need to transform the reverberant signal to mitigate its effects in speech-related tasks such as speech enhancement or automatic speech recognition [2]. The process of speech dereverberation consists in removing the early reflections and late reverberation from a reverberant signal, thereby approximating the dry signal. This presents yet an ill-posed problem since it depends on deconvolution where the impulse response is unknown. In theory, the convolutive model used for dereverberation should represent the Room Impulse Response (RIR), which uniquely characterizes reverberation. As RIR is not minimum-phase [3] or lacks robustness to spatial variations [4], a wide range of models mitigate deconvolution errors using regularization of the known RIR [5, 6], deep generalization to a spatial neighbourhood [7], or by a posterior sampling of a diffusion process informed by the RIR [8].

A first approach is to directly model either the dry signal, the reverberant signal, or both for dereverberation purposes. Regarding the modelling of reverberation, it has been represented as a convolutive distortion, and approaches have been developed to concurrently represent the convolutive model and the dry signal [9]. One of the most notable methods is the Weighted Prediction Error (WPE) [10]. This method has widely benefited from further refinement, including hybrid approaches combining WPE with deep learning [11, 12]. While WPE estimates the time-frequency (T-F) filter used to synthesize a dry signal from a reverberant one, Forward Convolutive Prediction (FCP) [13] aims at estimating the filter mapping a dry signal estimated by a neural network to a reverberant mixture. It has been applied to tasks such as dereverberation in a monaural setting [14], unsupervised multichannel dereverberation [15] and source separation [16, 17].

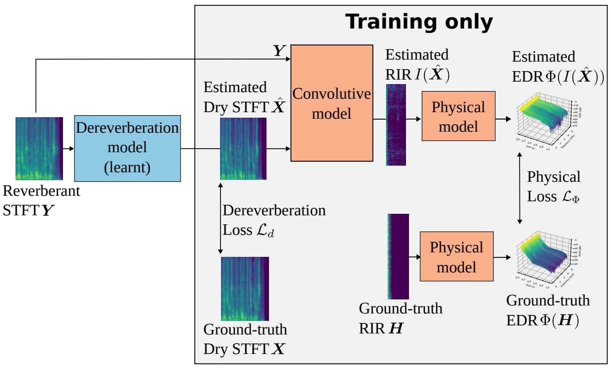

Figure 1: *Overview of the proposed method.*

The FCP is moreover closely related to the Convolutive Transfer Function (CTF) approximation, which considers reverberation as a subband filtering process. An observation model based on CTF has been used in conjunction with nonnegative matrix factorization (NMF) [18] and a diffusion model in [19]. However, neither the backward filter estimated by WPE nor the FCP or CTF have been constrained to be realistic with respect to room acoustics. Most state-of-the-art single-channel DNN-based dereverberation algorithms such as TF-GridNet [20] or UNet-based architectures [21] have shown good performance in various scenarios, yet are purely data-driven designed. The dereverberation task can on the other hand leverage not only the RIR itself but also the physical properties that constrain it leading to a physics-driven dereverberation paradigm. This has been made possible by the recent advances in blind room acoustic parameters estimation [22]. This approach has first been used to leverage the reverberation time RT_{60} in classical models [23], and refined using DNNs in [24, 25] for instance. At inference, they require a preliminary estimation of the RT_{60} but do not constrain the model output to match this property. Similarly, FullSubnet [26] has been used to target a signal with a shortened RT_{60} [27]. While this physically realistic approach simplifies the learning target of the DNN, the predicted signal is not necessarily dry.

This paper aims to bridge the gap between convolutive models and room acoustic properties estimation to constrain a deep dereverberation model. More precisely, for this preliminary study, we choose FullSubNet as a weakly physics-driven dereverberation algorithm and design physical losses inherited from a CTF model. Our contribution is two-fold: we show that 1) DNNs designed to only dereverberate speech are also able to implicitly model reverberation without increasing the number of parameters and 2) explicitly synthesize an RIR from a dereverberation model. More precisely, our proposed speech dereverberation constrained on RIR procedure demonstrates, through obtained objective scores, that we can maintain the overall quality of the original FullSubNet output while exhibiting a more physically consistent RIR. For reproducibility purposes and to help future research, we publicly distribute our code and pretrained models[1].

2. Reverberation in the T-F domain

Time-domain formulation: Assuming fixed source and microphone positions and no additive noise, a monaural reverberant (or wet) signal y can be represented as a convolution between a dry signal x and the room impulse response (RIR) h between the source and the microphone:

$$y_n = (h * x)_n, \tag{1}$$

where n denotes the time index and $*$ the convolution operator. **STFT filtering and Convolutive transfer function:** The time-invariant linear system of Eq. (1) can be formulated in the short-time Fourier transform (STFT) domain as interband and interframe convolution [28]:

$$Y_{f,t} = \sum_{f'=0}^{F-1} \sum_{t'=-\infty}^{\infty} \mathcal{H}_{f,f',t'} X_{f',t-t'}, \tag{2}$$

where $Y_{f,t}$ is the STFT coefficient of the reverberant signal at frequency $f = 0, \dots, F-1$ and time $t = 0, \dots, T_y - 1$, $\mathcal{H} \in \mathbb{C}^{F \times F \times T_h}$ is a tridimensional representation of the RIR and $X \in \mathbb{C}^{F \times T_x}$ is the STFT of the dry signal. As shown in [28], $\mathcal{H}$ can be obtained in closed form from the RIR h. Several approximations can be made from this model. Among them, the subband filtering operation, also named convolutive transfer function (CTF), considers the case where $\mathcal{H}_{f,f',t'}$ is nonzero only if $f = f'$ [9]. Crossband modelling, investigated in [28], considers an interband convolution kernel $\mathcal{C}_f$ of size $(2F' + 1)T_h$ for each frequency band f. The crossband filter can be estimated from the STFT coefficients of the dry signal (or an estimate of it) and the reverberant signal via a frequency-dependent least-squares optimization problem:

$$\mathcal{C}_f(\boldsymbol{X}) = \underset{\boldsymbol{C}_f}{\arg\min} \left\| \bar{\boldsymbol{X}}_f \boldsymbol{C}_f - \boldsymbol{Y}_f \right\|_2^2, \tag{3}$$

where

$$\boldsymbol{C}_f \triangleq \left[\boldsymbol{C}_{f,f'}^{\mathrm{T}} \right]_{f'=f-F'}^{f+F'} \in \mathbb{C}^{(2F'+1)T_h}, \tag{4}$$

$$\bar{\boldsymbol{X}}_f \triangleq \left[\boldsymbol{X}_{f'}^{(\mathrm{T})} \right]_{f'=f-F'}^{f+F'} \in \mathbb{C}^{T_y \times (2F'+1)T_h}. \tag{5}$$

$\mathcal{C}_f(\boldsymbol{X})$ is the concatenation of the crossband filters $\boldsymbol{C}_{f,f'}^{\mathrm{T}}$ mapping the frequencies $f' = f - F', \dots, f + F'$ of $\boldsymbol{X}$ to the frequency f of the reverberant STFT $\boldsymbol{Y}$. $\bar{\boldsymbol{X}}_f$ is the column-wise concatenation of the Toeplitz matrices $\boldsymbol{X}_{f'}^{(\mathrm{T})}$ of size $T_y \times T_h$ constructed from frequency bands $f' = f - F', \dots, f + F'$ of the dry STFT coefficients.

Room parameter estimation: Given an STFT representation $\boldsymbol{H}$ of an impulse response h, the energy decay relief (EDR) [29] is defined for each time-frequency bin (f, t) as:

$$\mathrm{EDR}(\boldsymbol{H})_{f,t} \triangleq \sum_{t'=t}^{+\infty} |H_{f,t'}|^2. \tag{6}$$

The EDR can be interpreted as a subband energy decay curve (EDC), representing a frequency-dependent energy decay. It has been used as a loss for RIR estimation [30].

3. Proposed Method

3.1. Overview

We propose to introduce a new loss term that imposes physical constraints on the RIR characteristics measured via the CTF approximation when training a dereverberation deep neural network (DNN). The general procedure to define our physical loss term is as follows. From the dereverberated output $\hat{\boldsymbol{X}}$ obtained by the DNN from a reverberant signal $\boldsymbol{Y}$, a convolutive model computes the CTF $\mathcal{C}(\hat{\boldsymbol{X}})$ mapping the output of the DNN to its reverberant input, following Eq. (3), and from it an estimate $I(\hat{\boldsymbol{X}})$ of the STFT of the corresponding RIR. A physical model is then used to compute an estimated physical property $\Phi(I(\hat{\boldsymbol{X}}))$ from the estimated CTF, and similarly a target physical property $\Phi(I(\boldsymbol{X}))$ from the oracle CTF obtained with the ground-truth anechoic signal. Their distance is finally used to define our physical loss function $\mathcal{L}_\phi$.

This new loss term $\mathcal{L}_\phi$ can be combined with a classical dereverberation loss $\mathcal{L}_d$ (e.g., assessing the reconstruction quality of the dry or direct-path signal) to train the DNN. A diagram of the training procedure is shown in Fig. 1.

Because the convolutive and physical models are not parametric, they do not need to be trained. At inference, for the dereverberation task, these blocks are discarded, and only the DNN is used. Hence, the number of parameters, as well as the computational complexity and memory footprint are the same as for the original model.

3.2. Corrected Convolutive Model

The number of crossbands is limited by the dimension of the least-squares system to solve at Eq. (3). For the system to have a unique solution, it is required that $\bar{\boldsymbol{X}}_f$ is full-rank, hence the relation $(2F' + 1)T_h < T_y$ must hold. Taking into account the length of the dry signals and RIRs in our training data, as well as the computational load, we limit ourselves to considering the subband ($F' = 0$) and 3-band ($F' = 1$) cases for the CTF. We solve Eq. (3) using QR decomposition.

It can be proven that the STFT $\boldsymbol{H}$ of the impulse response h can be computed from the convolutive interframe and interband filter $\mathcal{H}_{f,f',t}$ (if it were known):

$$H_{f,t} = \sum_{f'=0}^{F-1} (-1)^{f'} \mathcal{H}_{f,f',t} \tag{7}$$

where the multiplication by $(-1)^{f'}$ stems from the centering of the first STFT window. We can use this relationship to obtain an estimate of the STFT of the RIR from the CTF computed by either the clean speech $\boldsymbol{X}$ or its estimate $\hat{\boldsymbol{X}}$:

$$I(\boldsymbol{X})_{f,t} = \sum_{f'=f-F'}^{f+F'} (-1)^{f'} \mathcal{C}_{f,f',t}(\boldsymbol{X}), \tag{8}$$

[1] https://louis-bahrman.github.io/SD-cRIRc/

and similarly for $\hat{\boldsymbol{X}}$. Because our model only considers a few crossbands, this estimate will not yield the exact STFT $H_{f,t}$ of the RIR, but an approximation, even if it is computed on the CTF $\mathcal{C}(\boldsymbol{X})$ obtained from the clean speech $\boldsymbol{X}$. We define the modeling error at each T-F bin as $\mathcal{E}_{f,t} = I(\boldsymbol{X})_{f,t} - H_{f,t}$.

To make physical properties less dependent on this approximation, we attempt to compensate for the error via a spectral-subtraction-based correction. The spectral subtraction yields $I(\boldsymbol{X})^c_{f,t}$, an estimator of the RIR spectrum. The same error correction can be applied to the estimate $I(\hat{\boldsymbol{X}})$ of the RIR obtained from the estimate $\hat{\boldsymbol{X}}$ of the dry speech:

$$I(\boldsymbol{X})^c_{f,t} = \left(|I(\boldsymbol{X})_{f,t}|^2 - |\mathcal{E}_{f,t}|^2\right)^{1/2} e^{j\angle I(\boldsymbol{X})_{f,t}}, \tag{9}$$

$$I(\hat{\boldsymbol{X}})^c_{f,t} = \left(|I(\hat{\boldsymbol{X}})_{f,t}|^2 - |\mathcal{E}_{f,t}|^2\right)^{1/2} e^{j\angle I(\hat{\boldsymbol{X}})_{f,t}}. \tag{10}$$

Note that adjusting both target and estimated convolutive transfer functions by the same quantity will alter the nonlinear behaviour of the physical model employed.

If the spectrogram of the RIR that has been used for data generation is not available, one can still compare the physical properties estimated from $I(\hat{\boldsymbol{X}})$ and $I(\boldsymbol{X})$ directly without applying the correction.

3.3. Physical coherence loss

As an example of physical characteristic of interest to be used as a constraint on the RIR, we consider the dB-scaled EDR [29]. Given an STFT of an RIR or an approximation of it, $\boldsymbol{R}$, the dB-scaled EDR is obtained as:

$$\Phi_{f,t}(\boldsymbol{R}) \triangleq \mathrm{EDR}^s(\boldsymbol{R})_{f,t} = 10\log_{10}\frac{\mathrm{EDR}(\boldsymbol{R})_{f,t}}{\mathrm{EDR}(\boldsymbol{R})_{f,0}}. \tag{11}$$

The physical coherence loss $\mathcal{L}_\Phi$ can then be defined as a point-wise mean-squared error between the dB-scaled EDRs obtained from an estimate $\hat{\boldsymbol{R}}$ and a target $\boldsymbol{R}$. Since the tail of the EDR is very sensitive to CTF approximation errors and has high values on the log scale, both target and estimated EDRs are masked to exclude time-frequency bins where the target EDR is lower than -20 dB:

$$\mathcal{L}_\Phi(\hat{\boldsymbol{R}}, \boldsymbol{R}) = \sum_{f,t} \left|\Phi_{f,t}(\hat{\boldsymbol{R}}) - \Phi_{f,t}(\boldsymbol{R})\right|^2 \mathbb{1}_{\{\Phi_{f,t}(\boldsymbol{R}) > -20\}}. \tag{12}$$

We consider several variants for the selection of $\hat{\boldsymbol{R}}$ and $\boldsymbol{R}$, such as $I(\hat{\boldsymbol{X}})^c$ and $I(\boldsymbol{X})^c$, as described in Section 4.1.

3.4. Multi-objective training

To balance both physical coherence and reconstruction losses in a multi-task training setting, we use GradNorm [31]. GradNorm ensures that the gradients of both $\mathcal{L}_\Phi$ and $\mathcal{L}_d$ losses have equal norms across all weights. In our setting, $\mathcal{L}_\Phi$ is highly nonconvex with respect to the network parameters, so we prioritize the reconstruction loss over the physical coherence loss to stabilize training. After GradNorm has been applied, we further multiply the physical coherence loss by a constant weight w_Φ. Based on preliminary experiments, we set $w_\Phi = 0.1$.

4. Experiments

4.1. Model variants

We assess several variants of our method with FullSubNet (FSN) [26] as the baseline dereverberation model (see Fig. 1). The ability of FullSubNet to process spectrograms both in the full-band and subband directions is required to estimate a cross-band convolutive model. It has also been successfully used to solve the physically meaningful task of reverberation-time shortening [27]. We select its bidirectional version and keep the original training loss expressed as a mean square error on its complex ratio mask output [32] as the dereverberation loss $\mathcal{L}_d$.

The following variants are considered, representing different ways to compute the convolutive model. We define two kinds of approaches depending on whether subband or cross-band filters are considered to obtain the estimates of the RIR STFT, and which estimates and targets are compared:

- Subband approach (SB): $\mathcal{L}_\Phi(I(\hat{\boldsymbol{X}}), \boldsymbol{H})$, comparing the estimate from $\hat{\boldsymbol{X}}$ with ground-truth RIR STFT $\boldsymbol{H}$.
- Symmetric Subband approach (SSB): $\mathcal{L}_\Phi(I(\hat{\boldsymbol{X}}), I(\boldsymbol{X}))$, comparing the estimate from $\hat{\boldsymbol{X}}$ with the estimate from $\boldsymbol{X}$.
- Corrected Subband approach (CSB): $\mathcal{L}_\Phi(I(\hat{\boldsymbol{X}})^c, I(\boldsymbol{X})^c)$, comparing the corrected estimate from $\hat{\boldsymbol{X}}$ with the corrected estimate from $\boldsymbol{X}$.
- 3-band approach (3B): $\mathcal{L}_\Phi(I(\hat{\boldsymbol{X}}), \boldsymbol{H})$, similar to SB but computed using $F' = 1$ crossbands.

4.2. Miscellaneous configurations

As in the original FullSubNet, 49151 sample excerpts (around 3 s at 16 kHz) reverberant audios are processed in the STFT domain using a 512-sample Hann window with an overlap of 50 %. The network is trained for $330,000$ steps using the Adam optimizer with an initial learning rate (LR) of 10^{-4} and a One-cycle-LR with a maximum at 10^{-3}.

4.3. Training dataset

Similarly to [20], we simulated a training dataset by dynamically convolving dry speech signals with simulated RIRs. The dry speech signals are randomly sampled from the close-talking microphone recordings in the WSJ0 dataset [33]. The training set is composed of a total of 61 hours of recordings split into 31,350 audio excerpts. The simulated RIR dataset consists of 32,000 RIRs simulated using the pyroomacoustics library [34] with 2000 rooms whose dimensions and RT$_{60}$ are uniformly sampled in the respective ranges of $[5, 10] \times [5, 10] \times [2.5, 4]$ m^3, and $[0.2, 1.0]$ s. In each room, a source is randomly positioned and 16 microphones are sampled such that the source-microphone distance D is uniformly distributed in $[0.75, 2.5]$ m and both source and microphone are at least 50 cm from the walls. At training time, we use a dynamic mixing procedure consisting in randomly selecting a dry signal and RIR pair. In order to align the dry signal target and the direct-path, the samples before the direct path are discarded and it is normalised (so that the first impulse is of amplitude 1). This does not change the RIR distribution and compensates for the delay induced by the direct-path, both on the STFT $\boldsymbol{H}$ and on the oracle EDR $\Phi(\boldsymbol{H})$ so that they will start decreasing at the first frame.

We evaluate the proposed method on two different tasks: speech dereverberation and room impulse response characterization.

4.4. Metrics for evaluation and tasks

We evaluate the generalization performance of our metrics to both unseen sources and rooms. For dry sources, we consider the test set of WSJ0 [33], and Librispeech clean [35]. Two reverberation datasets are considered: one simulated using un-

Table 1: *Dereverberation scores ± standard deviation (std.) for FullSubNet (FSN) and its constraints versions.*

	Matched RIRs						Mismatched RIRs					
	WSJ0			LibriSpeech clean			WSJ0			LibriSpeech clean		
	STOI	SISDR	WB-PESQ	STOI	SISDR	WB-PESQ	STOI	SISDR	WB-PESQ	STOI	SISDR	WB-PESQ
FSN	0.93 ± 0.07	**5.1 ± 4.1**	2.23 ± 0.60	0.90 ± 0.11	**3.1 ± 4.3**	2.06 ± 0.55	0.87 ± 0.06	**0.9 ± 2.6**	1.60 ± 0.21	0.84 ± 0.10	**–0.8 ± 3.4**	1.53 ± 0.24
+ SB	0.92 ± 0.07	4.3 ± 4.2	2.10 ± 0.56	0.89 ± 0.11	2.5 ± 4.6	1.98 ± 0.51	0.86 ± 0.06	–0.3 ± 2.9	1.46 ± 0.19	0.82 ± 0.10	–1.9 ± 3.5	1.42 ± 0.21
+ CSB	0.92 ± 0.07	4.2 ± 4.6	2.11 ± 0.65	0.89 ± 0.11	2.2 ± 5.1	1.99 ± 0.59	0.86 ± 0.06	–0.7 ± 2.9	1.43 ± 0.18	0.82 ± 0.10	–2.4 ± 3.8	1.41 ± 0.21
+ SSB	0.93 ± 0.07	4.8 ± 4.1	2.19 ± 0.59	0.89 ± 0.11	2.6 ± 4.5	1.99 ± 0.52	0.87 ± 0.06	0.6 ± 2.7	1.57 ± 0.20	0.83 ± 0.10	–1.3 ± 3.8	1.49 ± 0.23
+ 3B	0.93 ± 0.07	4.9 ± 4.1	**2.24 ± 0.60**	0.90 ± 0.11	2.9 ± 4.6	**2.07 ± 0.57**	0.87 ± 0.06	0.7 ± 2.6	**1.61 ± 0.21**	0.84 ± 0.10	–1.0 ± 3.7	**1.54 ± 0.25**
input	0.86 ± 0.09	–0.2 ± 4.8	1.76 ± 0.67	0.85 ± 0.12	–1.0 ± 5.5	1.89 ± 0.76	0.75 ± 0.07	–4.5 ± 2.9	1.20 ± 0.11	0.74 ± 0.10	–5.2 ± 3.7	1.24 ± 0.16

seen rooms matching the same physical parameters as the training dataset described in Section 4.3 (”Matched RIRs”), and the other matching harder conditions (”Mismatched RIRs”): $\text{RT}_{60} \in [1.0, 1.5]$ s room size range in $[10, 15] \times [10, 15] \times [4, 6]$ m^3, $D \in [2.5, 4.0]$ m. The dereverberation performance between the baseline and the proposed approaches is evaluated using the Short-time-objective Intelligibility STOI, the Scale Invariant Signal-to-noise ratio (SISDR) [36], and the wide-band Perceptual Evaluation of Speech Quality WB-PESQ.

To demonstrate the acoustic matching capability acquired by the network constrained by RIR characteristics, we compare the energy decay curves (EDCs) predicted at the output of each version of the DNN using 3 convolutive models:

- EDC-Fourier: $\mathcal{L}_\Phi\left(\text{IDFT}\left[\frac{\text{DFT}(y_n)}{\text{DFT}(x_n)}\right], h_n\right)$, where (I)DFT is the (inverse) discrete Fourier transform, and $\Phi_n(r) = \sum_{n'=n}^{+\infty} |r(n')|^2$ [27].
- EDR-Subband: Corresponds to the SB loss, which we consider as a metric.
- EDR-Crossband: Corresponds to the 3B loss, which we consider as a metric.

5. Results and Discussion

5.1. Dereverberation

The results for the dereverberation task are presented in Table 1. Our proposed solution, FSN+3B, has a higher WB-PESQ on all datasets and acoustic conditions than the FSN baseline. All physically constrained variants exhibit similar performance in terms of STOI as the baseline. This means that the physical coherence loss and the dereverberation loss can be jointly optimized and that they both converge to equally performing optima in terms of STOI on the space of the DNN weights. The poorer results of our methods compared to the baseline in terms of SISDR can be explained by the DNN encountering difficulty in optimizing the phase of the complex mask mapping $\boldsymbol{Y}$ to $\boldsymbol{X}$ when it is constrained by a convolutive model. Considering this metric, the model trained on SSB performs similarly to the model trained on 3B. These losses are the ones that introduce the least constraints on the training and that are the least well-defined (SSB by introducing subband modelling errors, and 3B by being unstable). Because these two losses regularize the training in a physically realistic manner, they enable the model to perform better on unseen cases and to generalize to out-of-domain RIRs and source signals. Further experiments show that the dereverberation performance remains consistent when high SNR noise is added to the reverberant input of the model at test time. These results reflect FullSubNet's underlying design assumption that both Full- and Subband modelling are needed for the dereverberation task.

Table 2: *RIR estimation scores ± std. on the WSJ0 test set.*

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

5.2. RIR estimation

Table 2 compares the performance of all proposed approaches with respect to the energy decay of several convolutive models. The line denoted ”dry” shows $\mathcal{L}_\Phi(I(\boldsymbol{X}), \boldsymbol{H})$ for each convolutive model and energy decay EDC-Fourier, EDR-Subband, and EDR-Crossband. It represents the best theoretical performance each convolutive model can offer. The results of the 3B metric show a very high error and variance. This can be explained through Avargel's error analysis of the Crossband filtering [28], where it is shown that for a given SNR on the dry and reverberant signal, there exists only one single tuple (F', T_x) minimizing the mean-squared error. Further analysis shows that the length of the signal considered was too short for the Crossband method to perform well, hence its poor results. The results suggest that the RIR estimation task competes with the dereverberation task, as indicated by their differing performance rankings. The FSN+CSB variant is performing the best and is capable of modelling the subband model even better than the oracle subband model $I(\boldsymbol{X})$. This can be explained by the fact that forcing the model output to respect a subband model while maintaining its ability to process crossbands in its latent representation is very efficient to predict the STFT, but insufficient to perform dereverberation correctly. This assumption is indeed at the core of FullSubNet's design. Accordingly, a general guideline might be to resort to FSN+CSB for the RIR estimation task, and to FSN+3B for the dereverberation task.

6. Conclusion

We have proposed a novel approach for speech dereverberation which constrains the estimated room impulse response to well capture the acoustic properties of the room in which the signal was recorded. While the overall dereverberation performance remains comparable to the baseline model, having access to a realistic room impulse response characterizing the reverberated environment opens the path to a variety of controllable acoustic transformation applications (acoustic sound matching, realistic room shape modifications,...). Future work will be dedicated to the generalization of our approach to other DNN architectures.

7. Acknowledgements

This work was funded by the European Union (ERC, HI-Audio, 101052978). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work was performed using HPC resources from GENCI–IDRIS (Grant 2024-AD011014072R1). We would like to thank the thank the reviewers and meta reviewers for their insightful comments.

8. References

[1] T. Houtgast and H. J. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, 1985.

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Rooms: Robustness Against Reverberation for Automatic Speech Recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[3] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 165–169, Jul. 1979.

[4] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *Journal of Sound and Vibration*, vol. 102, no. 2, pp. 217–228, Sep. 1985.

[5] N. Cahill and R. Lawlor, "A novel approach to mixed phase room impulse response inversion for speech dereverberation," in *Proc. ICASSP*, Mar. 2008, pp. 4593–4596.

[6] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *Proc. ICASSP*, May 2014, pp. 5177–5181.

[7] R. Xu, G. Krishnan, C. Zheng, and S. K. Nayar, "Personalized Dereverberation of Speech," in *Proc. Interspeech*, Aug. 2023, pp. 3859–3863.

[8] J.-M. Lemercier, S. Welker, and T. Gerkmann, "Diffusion Posterior Sampling for Informed Single-Channel Dereverberation," *arXiv preprint arXiv:2306.12286*, Jun. 2023.

[9] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio source separation and speech enhancement*. Hoboken, NJ: John Wiley & Sons, 2018.

[10] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[11] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural Network-Based Spectrum Estimation for Online WPE Dereverberation," in *Proc. Interspeech*, Aug. 2017, pp. 384–388.

[12] K. Saito, N. Murata, T. Uesaka, C.-H. Lai, Y. Takida, T. Fukui, and Y. Mitsufuji, "Unsupervised Vocal Dereverberation with Diffusion-Based Generative Models," in *Proc. ICASSP*, Jun. 2023, pp. 1–5.

[13] Z.-Q. Wang, G. Wichern, and J. Le Roux, "Convolutive Prediction for Reverberant Speech Separation," in *Proc. WASPAA*, Oct. 2021, pp. 56–60, iSSN: 1947-1629.

[14] Z.-Q. Wang, G. Wichern, and J. L. Roux, "Convolutive Prediction for Monaural Speech Dereverberation and Noisy-Reverberant Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3476–3490, 2021.

[15] Z.-Q. Wang, "USDnet: Unsupervised Speech Dereverberation via Neural Forward Filtering," Feb. 2024, arXiv preprint arXiv:2402.00820.

[16] Z.-Q. Wang and S. Watanabe, "UNSSOR: Unsupervised Neural Speech Separation by Leveraging Over-determined Training Mixtures," Oct. 2023, arXiv preprint arXiv:2305.20054.

[17] R. Aralikatti, C. Boeddeker, G. Wichern, A. Subramanian, and J. Le Roux, "Reverberation as Supervision For Speech Separation," in *Proc. ICASSP*, Jun. 2023, pp. 1–5.

[18] D. Baby and H. Van hamme, "Supervised speech dereverberation in noisy environments using exemplar-based sparse representations," in *Proc. ICASSP*, Mar. 2016, pp. 156–160.

[19] P. Wang and X. Li, "RVAE-EM: Generative speech dereverberation based on recurrent variational auto-encoder and convolutive transfer function," Sep. 2023, arXiv preprint arXiv:2309.08157.

[20] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[21] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech Dereverberation Using Fully Convolutional Networks," in *Proc. EUSIPCO*, Sep. 2018, pp. 390–394.

[22] T. de M. Prego, A. A. de Lima, R. Zambrano-López, and S. L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," in *Proc. WASPAA*, 2015, pp. 1–5.

[23] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A New Method Based on Spectral Subtraction for Speech Dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, May 2001.

[24] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 102–111, Jan. 2017.

[25] Y. Li, Y. Liu, and D. S. Williamson, "A Composite T60 Regression and Classification Approach for Speech Dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 1–11, 2023.

[26] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement," in *Proc. ICASSP*, Jun. 2021, pp. 6633–6637.

[27] R. Zhou, W. Zhu, and X. Li, "Speech Dereverberation with A Reverberation Time Shortening Target," Nov. 2022, arXiv preprint arXiv:2204.08765.

[28] Y. Avargel and I. Cohen, "System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[29] J.-M. Jot, "An analysis/synthesis approach to real-time artificial reverberation," in *Proc. ICASSP*, Mar. 1992, pp. 221–224.

[30] A. Ratnarajah, I. Ananthabhotla, V. K. Ithapu, P. Hoffmann, D. Manocha, and P. Calamia, "Towards Improved Room Impulse Response Estimation for Speech Recognition," in *Proc. ICASSP*, Jun. 2023, pp. 1–5.

[31] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks," in *Proc. ICML*, Jul. 2018, pp. 794–803, iSSN: 2640-3498.

[32] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[33] J. S. Garofolo *et al.*, *CSR-I (WSJ0) Complete LDC93S6A*, Linguistic Data Consortium, Philadelphia, 1993, web Download.

[34] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms," in *Proc. ICASSP*, Apr. 2018, pp. 351–355.

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, Apr. 2015, pp. 5206–5210, iSSN: 2379-190X.

[36] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *Proc. ICASSP*, May 2019, pp. 626–630.