A Hybrid Model for Weakly-Supervised Speech Dereverberation

Louis Bahrman, Mathieu Fontaine, Gael Richard

A Hybrid Model for Weakly-Supervised Speech Dereverberation

Louis Bahrman, Mathieu Fontaine*, Gaël Richard**

*LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France

***Abstract*—This paper introduces a new training strategy to improve speech dereverberation systems using minimal acoustic information and reverberant (wet) speech. Most existing algorithms rely on paired dry/wet data, which is difficult to obtain, or on target metrics that may not adequately capture reverberation characteristics and can lead to poor results on non-target metrics. Our approach uses limited acoustic information, like the reverberation time (RT60), to train a dereverberation system. The system's output is resynthesized using a generated room impulse response and compared with the original reverberant speech, providing a novel reverberation matching loss replacing the standard target metrics. During inference, only the trained dereverberation model is used. Experimental results demonstrate that our method achieves more consistent performance across various objective metrics used in speech dereverberation than the state-of-the-art.**

***Index Terms*—Speech dereverberation, hybrid deep learning, reverberation modeling, speech processing.**

I. Introduction

Acoustic signals captured in closed rooms are affected by reflections from room walls and diffraction from obstacles on its path, in a process coined as reverberation. These effects may not be desirable in speech recordings as they lower speech intelligibility [1]. This justifies the need for dereverberation methods to mitigate the reverberation phenomenon in speech-related tasks such as speech enhancement and automatic speech recognition [2]. Dereverberation task has been historically solved by using statistical signal-processing methods [3]–[5]. The nonlinearity of the task naturally calls for deep neural networks (DNNs) extensions requiring in practice a large amount of annotated data and learning strategies. These learning-based approaches can be supervised in various ways.

Best-performing discriminative approaches as TFGridNet [6] learn to predict a dry signal from a reverberant one decomposing the time-frequency signal into time and spectral subband modules and hence require paired dry/wet data. FullSubNet [7] aims to estimate a Complex Ratio Mask (cRM) for retrieving the dry signal. However, these techniques require generating a large amount of paired data and may lack robustness if the test data significantly differs from the training dataset. This lack of paired data has motivated the development of new approaches which can dereverberate using unpaired signals only as in Cycle-consistent Generative Adversarial Networks (GANs) [8], [9].

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).

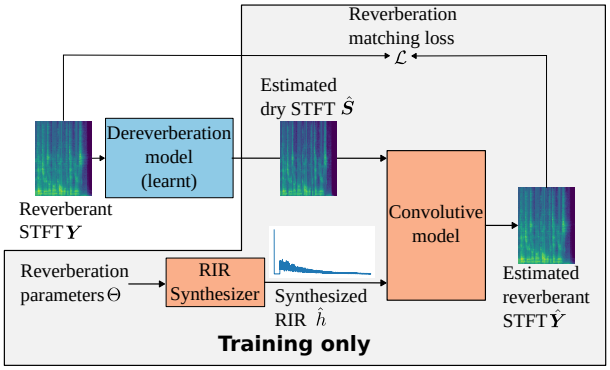

Fig. 1. Overview of the proposed method

On the other hand, generative models like variational autoencoders [10], [11] or diffusion models [12] are used to learn the prior dry signals without having access to any reverberant signal at training. Although these models require less supervision, they do not solve the problem of data scarcity, since dry speech data is harder to obtain than reverberant speech data.

Few approaches are designed to require only reverberant signals at training time. The current best-performing model for dereverberation supervised only by reverberant signals is MetricGAN-U [13]. Its training framework is based on a GAN, where the discriminator is trained to mimic the behaviour of a target metric, and the generator to optimize its performance with respect to this evaluation metric. It has been successfully applied on the dereverberation task, using Speech to Reverberant Modulation energy Ratio (SRMR) [14] as a target metric to be optimized.

Besides, both supervised and unsupervised approaches for dereverberation have been improved by leveraging reverberation models. Such approaches can be considered as hybrid deep learning, in the sense that they combine DNN priors with statistical models or signal-based representations of the reverberation. Indeed, reverberation has been classically represented as a convolutive distortion and approaches have been developed to concurrently estimate the convolutive model and the dry signal [3], [4], [15]. The reverberation model can be implicitly modelled, or explicitly used. A popular choice to implicitly model reverberation is the Convolutive Transfer

Function (CTF) approximation, which considers reverberation as a subband filtering process. It has been used in the weighted prediction error (WPE) method [3] and its neural enhancements [16] or with diffusion models [17] and variational approaches [18]. An observation model based on CTF has been used in conjunction with discriminative approaches under the name Forward Convolutive Prediction (FCP) [19]. FCP has even been used for unsupervised learnt dereverberation in USDNet [20]. USDNet is powerful in a multichannel setting but shows only subpar performance in a monaural setting.

The convolutive model can also be explicitly modelled. Recent work shows that models designed to only dereverberate are also able to explicitly model reverberation [21] and that it is possible to constrain a diffusion posterior sampling to match acoustic properties [22]. Other methods even assume that some reverberation properties are available at inference [23], [24], or even the full room impulse response (RIR) which uniquely characterizes the reverberation process [25]. This has been made possible by the recent advances in blind room acoustic parameters estimation [26], [27].

So far, MetricGAN-U remains the best approach for dereverberation supervised by reverberant signals, outperforming unsupervised approaches such as WPE. We qualify this approach as a *metrics-based weak supervision* since it requires a nonintrusive metric to compute its training target. However, metrics-based supervision is known to potentially be detrimental to performance regarding other criteria [28].

In this article, we propose to alleviate this issue by introducing a novel hybrid weak supervision framework for dereverberation, called *reverberation-based weak supervision*. We train a deep neural network to predict a dry estimate from a reverberant signal, such that a reverberation model applied to the dry estimate matches its reverberant input. We show that reverberation-based weak supervision performs better than metrics-based weak supervision on various objective measures. For reproducibility purposes and to help future research, we publicly distribute examples, code and pretrained models [1].

II. Reverberation model

A. Late reverberation and mixing time

Assuming fixed source and microphone positions and no additive noise, a monaural reverberant signal y can be represented as a convolution between a dry signal s and the room impulse response (RIR) h between the source and the microphone:

$$y(n) = (s \star h)(n), \tag{1}$$

where n denotes the time index and $\star$ the convolution operator. The RIR h can be divided into three parts: the direct path corresponds to its first peak h_d followed by early echoes h_e and, after the mixing time, late reverberation h_l:

$$h = h_d + h_e + h_l, \tag{2}$$

The support of h_d, h_e and h_l are disjoint, and several definitions for the mixing time n_m have been proposed. In [29], the mixing time in samples is defined using statistical properties of ergodic rooms, as a multiple of the mean free path [30]:

$$n_m = \frac{4Vf_s}{cA}, \tag{3}$$

where V, f_s, c and A are respectively the room volume, sampling frequency, speed of sound and area of the walls.

B. Polack's late reverberation model

A simple yet powerful model for late reflections is Polack's model [31]. This model states that the late reverberation h_l is a realization of an exponentially decaying white noise:

$$h_l(n) = b(n)e^{-(n+3n_m)/\tau}, \tag{4}$$

where $b(n) \sim \mathcal{N}(0, \sigma^2)$ is a centered Gaussian distribution, and τ depends on the reverberation time RT_{60} and the sampling frequency f_s as:

$$\tau = \frac{\mathrm{RT}_{60} f_s}{3\ln(10)}. \tag{5}$$

C. Convolution in Time-Frequency domain

The time-invariant linear system of Eq. (1) can be formulated in the short-time Fourier transform (STFT) domain as interband and interframe convolution [32]:

$$Y_{f,t} = \sum_{f'=0}^{F-1} \sum_{t'=0}^{\min(t;T_h)} \mathcal{H}_{f,f',t'} S_{f',t-t'}, \tag{6}$$

where $\boldsymbol{Y} \triangleq \{Y_{f,t}\}_{f,t=0}^{F-1,T_y-1} \in \mathbb{C}^{F\times T_y}$ are the STFT coefficients of the reverberant signal at frequency $f = 0, \ldots, F-1$ and time $t = 0, \ldots, T_y - 1$, $\mathcal{H} \triangleq \{\mathcal{H}_{f,f',t}\}_{f,f',t=0}^{F-1,F-1,T_h-1} \in \mathbb{C}^{F\times F\times T_h}$ is a tridimensional representation of the RIR and $\boldsymbol{S} \triangleq \{S_{f,t}\}_{f,t=0}^{F-1,T_s-1} \in \mathbb{C}^{F\times T_s}$ is the STFT of the dry signal. As shown in [32], $\mathcal{H}$ can be obtained in closed form from the RIR $h \in \mathbb{R}^{N_h}$ as:

$$\mathcal{H}_{f,f',t'} = \sum_{m=-N+1}^{N-1} h(t'L - m) W_{f,f'}(m), \tag{7}$$

where N is the STFT window length, L the hop-size and

$$W_{f,f'}(m) = \frac{1}{F} \sum_{n=0}^{N-1} w_s(n+m) w_a(n) e^{\frac{j2\pi(f'(n+m)-fn)}{F}} \tag{8}$$

with w_s, w_a the synthesis and analysis windows respectively.

III. Proposed method

A. Overview

We propose to supervise the training of a dereverberation deep neural network (DNN) using a conventional reverberation model. The general training procedure is as follows. Given a reverberant signal $\boldsymbol{Y}$ as in the previous section, the DNN outputs a dry signal estimate $\hat{\boldsymbol{S}} \triangleq \{\hat{S}_{f,t}\}_{f,t=0}^{F-1,T_s-1} \in \mathbb{C}^{F\times T_s}$. In parallel, a reverberation model $\mathcal{R}$ synthesizes an approximated RIR $\hat{h}$ from a few reverberation model parameters $\Theta \triangleq \{\mathrm{RT}_{60}, \sigma, V, A\}$. Both the estimated dry STFT $\hat{\boldsymbol{S}}$ and the synthesized RIR $\hat{h} \in \mathbb{R}^{N_h}$ are convolved in a cross-band convolutive model $\mathcal{C}$ (see Eq. (10)), to compute an

[1] https://louis-bahrman.github.io/Hybrid-WSSD/

estimate of the reverberant spectrogram $\hat{\boldsymbol{Y}}$. The standard dereverberation loss requiring pairs of dry and wet signals is replaced by a reverberation matching loss $\mathcal{L}$, computing the distance between the estimated and ground-truth reverberant spectrograms $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$. A diagram of the training procedure is shown in Fig. 1. Because the RIR synthesis and convolutive model are not parametric, they do not need to be trained. These blocks are discarded at inference, 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.

B. RIR synthesizer

The RIR synthesizer aims at synthesizing an RIR for which the late reverberation h_l matches Polack's model, and the direct path h_d is a peak of amplitude 1. To better match Polack's model with our data distribution without changing its energy distribution, and based on preliminary experiments, we decided to synthesize an RIR using the absolute value of the Gaussian distribution used in Polack's model. According to the mean free path property, the direct path peak should be on average positioned at the sample corresponding to the mean free path of the room n_m. As stated in [29], the mixing time, corresponding to the beginning of the late reverberation h_l, is set at 3 times the mean free path. However, to better align the dry and reverberant signals, we discard the RIR samples before the first peak. Hence, the synthetic RIR becomes:

$$\mathcal{R}(\Theta)(n) = \begin{cases} |b(n)|e^{-\frac{3\ln(10)}{\mathrm{RT}_{60} f_s} n} & \text{if } n > 2n_m \\ 1 & \text{if } n = 0 \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where $b(n)$ is drawn from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$, and n_m corresponds to Eq. (3). In this model, at fixed RT_{60}, σ^2 is proportional to the inverse of the direct-to-reverberant ratio (DRR), which has been proven to have great influence on dereverberation performance [33].

C. Convolutive model and reverberation matching loss

To better backpropagate the training gradient to the dereverberation model whose output might be in the time-frequency plane, we consider a time-frequency cross-band convolutive model and reverberation matching loss. Given $\hat{h} = \mathcal{R}(\Theta)$, and $\hat{\boldsymbol{S}}$ the dry speech estimate outputted by the DNN, we define the time-frequency convolutive model as:

$$\hat{\boldsymbol{Y}}_{f,t} \triangleq \mathcal{C}(\hat{\boldsymbol{S}}, \hat{h}) = \sum_{f'=f-F'}^{f+F'} \sum_{t'=0}^{\min(t;T_h)} \hat{\mathcal{H}}_{f,f',t'} \hat{S}_{f',t-t'}, \tag{10}$$

with $\hat{\mathcal{H}}_{f,f',t'} \triangleq \sum_{m=-N+1}^{N-1} \hat{h}(t'L - m) W_{f,f'}(m)$ and the notations in Eq. (10) coinciding to those of Eq. (6-8). Based on [32], we set the number of crossbands F' to 4.

Our reverberation-matching loss corresponds to the commonly used mean-squared error estimator for the deconvolution problem. Since this problem can be ill-posed for low-amplitude signals, a regularization term matching the log-magnitudes of the reverberant estimate and ground truth is added, and the model training loss is, with $\lambda = \gamma = 1$ as in [34]:

$$\mathcal{L} = \sum_{f,t} \left[|\hat{Y}_{f,t} - Y_{f,t}|^2 + \lambda \left| \log\left(\frac{1 + \gamma|\hat{Y}_{f,t}|}{1 + \gamma|Y_{f,t}|} \right) \right|^2 \right] \tag{11}$$

IV. Experiments

We compare our proposed "reverberation-based weak supervision" with a baseline "metrics-based weak supervision" as implemented in MetricGAN-U.

A. DNN variants

We assess several variants of our method with FullSubNet (FSN) [7]. The ability of FullSubNet to process complex STFT representations both in the full-band and subband directions is required to be paired with our proposed cross-band convolutive model and reverberation matching loss. It has also been proven to be able to jointly model physical properties of a room and dry speech [21], and to be paired with reverberation-informed training strategies [35]. We also consider the baseline BiLSTM model [36] used as a generator in MetricGAN-U. This model is much simpler as it only allows to processing STFT magnitude masks, and will serve as an indicator for the behaviour of our proposed loss with a less expressive model.

B. Supervision variants

We considered several supervision variants classified as weak supervision and strong supervision.

Weak supervision (WS): WS variants include using Polack's model with either

- $\Theta \triangleq \{\mathrm{RT}_{60}, \sigma, V, A\}$: all the parameters, including those used to estimate the mixing time.
- $\{\mathrm{RT}_{60}, \sigma\}$: a fixed mixing time set as 20 ms after the peak, corresponding to the mean of all mixing times in the training dataset.
- $\{\mathrm{RT}_{60}\}$: a fixed mixing time at 20 ms and a median value of Polack's variance $\sigma = 0.02$ over the training dataset. This is the least-supervised model and is motivated by realistic scenarios where only the reverberation time can be computed from a reverberant signal [26].

Strong Supervision: Those variants include using more oracle information such as

- the exact RIR h as an oracle RIR synthesis model. This variant should be considered as an upper bound for our proposed reverberation-based weak supervision's performance, as it is equivalent to having access to pairs of dry and reverberant signals as supervision.
- each model's original paired training loss as supervision. BiLSTM is trained using the mean squared error between dry and dereverberated magnitude spectra. FSN is trained to minimize the euclidean distance between its estimated and the ground-truth ideal complex ratio mask (cRM).

We also consider MetricGAN-U's metrics-based weak supervision as a baseline. It corresponds to the BiLSTM model trained with the weak supervision of the SRMR metric.

TABLE I
DEREVERBERATION SCORES ± STANDARD DEVIATION
(FOR EACH METRIC, THE HIGHER THE BETTER)

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
	✓	Θ	1.0 ± 2.5	$\mathbf{0.71 \pm 0.14}$	$\mathbf{1.80 \pm 0.70}$	6.9 ± 2.8
		$\{\text{RT}_{60}, \sigma\}$	1.1 ± 2.5	0.70 ± 0.14	1.78 ± 0.69	7.0 ± 2.8
		$\{\text{RT}_{60}\}$	$\mathbf{2.9 \pm 3.4}$	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
	✓	Θ	0.8 ± 4.0	0.70 ± 0.15	1.81 ± 0.74	6.9 ± 2.7
		$\{\text{RT}_{60}, \sigma\}$	0.7 ± 4.0	0.70 ± 0.15	1.78 ± 0.72	6.8 ± 2.7
		$\{\text{RT}_{60}\}$	$\mathbf{1.6 \pm 3.7}$	$\mathbf{0.71 \pm 0.15}$	$\mathbf{1.84 \pm 0.75}$	6.9 ± 2.8
BiLSTM	✓	SRMR [13]	-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

C. 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 %. We use a learning rate decay based on the training loss on a validation set, and early stopping based on the SISDR metric on a validation set.

D. Dataset

Similarly to [6], [21], 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 WSJ1 dataset [37]. The training set is composed of a total of 73 hours of recordings split into 60307 audio excerpts. The simulated RIR dataset consists of 32,000 RIRs drawn from 2000 rooms simulated using the image source method implemented in the pyroomacoustics library [38]. Room dimensions and RT60 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. The source-microphone distance 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 of 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 direct-path is of amplitude 1). This does not change the RIR distribution and compensates for the delay induced by the direct-path to match the RIR synthesis procedure.

V. RESULTS AND DISCUSSION

We evaluate the performance of our proposed reverberation-based weak supervision for the dereverberation task on unseen speakers from WSJ1 (Hub and Spokes S1 to S4) and rooms. The performance is evaluated using the Scale-Invariant Signal-to-Noise ratio (SISDR) [39], Extended Short-Time Objective Intelligibility (ESTOI) [40], Wide-Band Perceptual Evaluation of Speech Quality (WB-PESQ) [41] and SRMR [14] metrics. To assert the statistical significance of our result analysis despite the high measured variances, we opted for a non-parametric Wilcoxon test with a significance level of 0.001 for the null hypothesis to be rejected. The results are presented in Table I. The line denoted "Reverberant" corresponds to unprocessed signals and the best significant weak supervision variant for each metric and dereverberation model is highlighted. All of the proposed methods show an improvement of the SISDR, ESTOI and WB-PESQ metrics, meaning that they can successfully dereverberate speech. The baseline (BiLSTM + SRMR) excels in terms of SRMR, but this performance comes at the cost of its SISDR and ESTOI results, which are degraded compared to the reverberant input. This result confirms the main drawback of metrics-based weak supervision, in the sense that it tends to solely optimize the metric it is trained on, disregarding the others. Indeed, all our proposed methods perform better than the baseline on all other metrics than SRMR. This demonstrates the superiority of reverberation-based weak supervision over metrics-based weak supervision. The best-performing method FullSubNet benefits from strong supervision, both when trained on its original complex masking loss or using the oracle RIR. On the other hand, the less-complex BiLSTM widely benefits from weak supervision, and performs better in terms of SISDR when weakly supervised by Polack's model than when it has access to the ground-truth RIR h. For this model, the weakest supervision by RT_{60} yields not only superior results to other weak-supervision variants for all metrics except SRMR, but even improves the model's SISDR performance above its original supervision based on magnitude spectra. This is due to the BiLSTM's design, which is meant only for a spectral magnitude masking loss, without alleviating the STFT phase. Hence, when the estimated dry signal is reverberated using a ground-truth RIR h, the estimated reverberant STFT phase is perturbed to a large extent, whereas reverberation by Polack's model yields a phase that is closer to the complex circular Gaussian model at the core of BiLSTM's design. Another noticeable result occurs for both models in the reverberation-based weakly supervision by Polack's model. Comparing reverberation-weak supervision approaches, we remark that they perform better in terms of SISDR when having no access to the acoustic parameters used to estimate the mixing time and Polack's model σ. Hence, fixing σ, V and A is equivalent to making the DRR only dependant on the RT_{60}, which can be easily computed from reverberant speech [26], and seems to regularize our proposed training procedure for dereverberation when evaluated with synthetic RIRs.

VI. CONCLUSION

We have proposed a novel approach for weakly-supervised speech dereverberation, by training a deep neural network to predict a dry estimate from a reverberant signal, such that a reverberation model applied on the dry estimate matches its reverberant input. Results demonstrate the superiority of our reverberation-based weak supervision over metrics-based weak supervision. This method opens the path to a variety of dereverberation techniques for data-scarce scenarios and various signals such as music. Future work will be dedicated to leveraging a more powerful RIR synthesis model that can estimate the RT_{60} from reverberant signals only, and to extending this work to weakly-supervised generative approaches for dereveberation to better model the probabilistic RIR model.

REFERENCES

[1] T. Houtgast and H. J. M. 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, Mar. 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] 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.

[4] 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.

[5] N. D. Gaubitch, M. R. P. Thomas, and P. A. Naylor, "Dereverberation using LPC-based approaches," in *Speech dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. London: Springer London, 2010, pp. 95–128.

[6] 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 Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3221–3236, 2023.

[7] 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.

[8] H. Muckenhirn, A. Safin, H. Erdogan, F. De Chaumont Quitry, M. Tagliasacchi, S. Wisdom, and J. R. Hershey, "CycleGAN-based Unpaired Speech Dereverberation," in *Proc. Interspeech*. ISCA, Sep. 2022, pp. 196–200.

[9] G. Yu, Y. Wang, C. Zheng, H. Wang, and Q. Zhang, "CycleGAN-based Non-parallel Speech Enhancement with an Adaptive Attention-in-attention Mechanism," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec. 2021, pp. 523–529.

[10] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational Autoencoder for Speech Enhancement with a Noise-Aware Encoder," in *Proc. ICASSP*, Jun. 2021, pp. 676–680.

[11] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A Recurrent Variational Autoencoder for Speech Enhancement," in *Proc. ICASSP*, May 2020, pp. 371–375.

[12] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *Proc. ICASSP*, 2023, pp. 1–5.

[13] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "MetricGAN-U: Unsupervised Speech Enhancement/ Dereverberation Based Only on Noisy/ Reverberated Speech," in *Proc. ICASSP*, May 2022, pp. 7412–7416.

[14] T. H. Falk, C. Zheng, and W.-Y. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

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

[16] 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.

[17] 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.

[18] P. Wang and X. Li, "Rvae-em: Generative speech dereverberation based on recurrent variational auto-encoder and convolutive transfer function," in *Proc. ICASSP*, 2024, pp. 496–500.

[19] 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.

[20] Z.-Q. Wang, "USDnet: Unsupervised Speech Dereverberation via Neural Forward Filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 3882–3895, 2024.

[21] L. Bahrman, M. Fontaine, J. Le Roux, and G. Richard, "Speech dereverberation constrained on room impulse response characteristics," in *Interspeech 2024*, 2024, pp. 622–626.

[22] E. Moliner, J.-M. Lemercier, S. Welker, T. Gerkmann, and V. Välimäki, "BUDDy: Single-Channel Blind Unsupervised Dereverberation with Diffusion Models," May 2024.

[23] 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.

[24] 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.

[25] J.-M. Lemercier, S. Welker, and T. Gerkmann, "Diffusion posterior sampling for informed single-channel dereverberation," in *Proc. WASPAA*, 2023, pp. 1–5.

[26] 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.

[27] W. Mack, S. Deng, and E. A. Habets, "Single-Channel Blind Direct-to-Reverberation Ratio Estimation Using Masking," in *Proc. Interspeech*. ISCA, Oct. 2020, pp. 5066–5070.

[28] D. de Oliveira, S. Welker, J. Richter, and T. Gerkmann, "The pesqetarian: On the relevance of goodhart's law for speech enhancement," in *Proc. Interspeech*, 2024, pp. 3854–3858.

[29] B. A. Blesser, "An interdisciplinary synthesis of reverberation viewpoints," *journal of the audio engineering society*, vol. 49, pp. 867–903, october 2001.

[30] W. B. Joyce, "Sabine's reverberation time and ergodic auditoriums," *J. Acoust. Soc. Am.*, vol. 58, no. 3, pp. 643–655, 09 1975.

[31] J.-D. Polack, "La transmission de l'energie sonore dans les salles," PhD Thesis, 1988.

[32] 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.

[33] E. A. P. Habets, S. Gannot, and I. Cohen, "Late Reverberant Spectral Variance Estimation Based on a Statistical Model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, Sep. 2009.

[34] S. Schwär and M. Müller, "Multi-Scale Spectral Loss Revisited," *IEEE Signal Process. Lett.*, vol. 30, pp. 1712–1716, 2023.

[35] R. Zhou, W. Zhu, and X. Li, "Speech dereverberation with a reverberation time shortening target," in *Proc. ICASSP*, 2023, pp. 1–5.

[36] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR," in *Latent Variable Analysis and Signal Separation*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds. Cham: Springer International Publishing, 2015, pp. 91–99.

[37] J. S. Garofolo *et al.*, *CSR-II (WSJ1) Complete LDC94S13A*, Linguistic Data Consortium, Philadelphia, 1994.

[38] 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.

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

[40] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[41] I. Rec, "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, CH–Geneva*, vol. 41, pp. 48–60, 2005.