Deep Internal Learning for Single-Channel Speech Dereverberation

Yuxi Zhang, Emilie d'Olne, Vikas Tokala, Patrick A. Naylor
Department of Electrical and Electronic Engineering
Imperial College London, United Kingdom
{yuxi.zhang22, emilie.dolne16, v.tokala, p.naylor}@imperial.ac.uk

***Abstract*—Deep Internal Learning (DIL) is a paradigm with high potential in deep learning, as it reduces reliance on external training data and uses a lightweight model compared to traditional deep learning methods. Since its introduction, DIL has been explored in the field of image processing, including applications such as image super-resolution, denoising, and deblurring. However, its potential for other signal-processing tasks such as speech enhancement remains relatively underexplored. In this study, we investigate the feasibility of utilizing DIL for single-channel dereverberation. Specifically, we develop a small speech-specific Convolutional Neural Network (CNN) that is trained exclusively on example pairs derived from the observed reverberant speech itself. We evaluate our approach in the context of a dereverberation task in oracle experiments and more practical scenarios, considering rooms with a range of reverberation times. Experimental results demonstrate that the DIL approach can achieve higher speech enhancement and dereverberation scores compared to the traditional Weighted Prediction Error (WPE) algorithm. This paper showcases the potential of DIL for speech enhancement and formulates several open issues for future research on this topic.**

***Index Terms*—Deep Internal Learning, Speech Enhancement, Speech Dereverberation**

I. Introduction

Reverberation is a common type of distortion found in real-world speech signals. It can have a detrimental effect on perceived speech quality, human intelligibility, and Automatic Speech Recognition (ASR) performance. Communication technologies such as hearing aids [1], [2] and teleconferencing systems [3] require robust dereverberation techniques to improve speech intelligibility and thereby facilitate more effective communication. In ASR-integrated applications, such as voice-controlled assistants [4], Speaker Identification (SID) [5], and other speech processing tasks [6], [7], reverberation caused by complex acoustic environments must be mitigated to achieve higher recognition accuracy.

Room reverberation is typically modeled as a convolution of the source signal with an impulse response representing the acoustic propagation channel [8]. In recent years, various approaches to dereverberation have been proposed. Traditional approaches such as Weighted Prediction Error (WPE) algorithm [9]–[11] utilize linear prediction to reduce reverberation components in speech signals. Recent dereverberation methods leverage Machine Learning (ML), particularly diffusion models [6], [12]. The performance of these ML approaches is particularly effective when they generalize well to unseen test data. This generalization ability is usually achieved through extensive training, both in terms of the quantity and diversity of the training data. However, curating sufficiently diverse and large-scale training datasets remains a challenge, and training such models incurs high computational costs, leading to practical difficulties.

In this paper, we explore a contrasting ML approach based on Deep Internal Learning (DIL) that requires no (external) training data. Instead, DIL leverages the structural patterns within a single input to produce an enhanced signal. While such zero-shot methods like DIL are generally expected to underperform compared to generalized models, our objective is to investigate how close their performance could be to fully-fledged ML approaches, with the compensating advantage of reducing training and inference complexity. As a first step toward applying DIL in speech enhancement, we focus on single-channel speech dereverberation. The key contributions of this work are summarized as follows:

1. **Novel Application**: We introduce DIL to the domain of speech enhancement and demonstrate its suitability for scenarios where developing a generalized model with external training data may be expensive, impractical or infeasible.
2. **Efficiency and Practicality**: We highlight the potential of DIL to remove reliance on external training data and provide a lightweight, efficient model for speech dereverberation.
3. **Empirical Validation**: We develop a novel formulation of DIL for speech enhancement and provide simulation results to indicate the level of performance obtained in oracle and more realistic scenarios for the chosen task of dereverberation.

The rest of the paper is structured as follows: First, we very briefly summarize the existing research for speech dereverberation. Next, we introduce the DIL concept, originally applied in image processing, before presenting our novel development of DIL for speech dereverberation. We then evaluate the feasibility of DIL in an oracle experiment. Finally, we extend our investigation to more practical scenarios.

II. Early Approaches for Speech Dereverberation

The noiseless reverberant signal, $y[n]$, captured by a single microphone is commonly modeled as

$$y[n] = \sum_{\tau=0}^{T} h[\tau]x[n-\tau] = h[n] * x[n], \tag{1}$$

Demo results are available at https://yuxiz0826.github.io/demos/

where n denotes the discrete time index, $h[n]$ is the Room Impulse Response (RIR) modeled as a finite impulse response (FIR) filter of order T, $x[n]$ is the anechoic speech signal of interest, and $*$ denotes the convolution operator. Specifically, $h[n]$ can be decomposed into early reflections $h_e[n]$, which may be beneficial, and late reflections $h_\ell[n]$, which are detrimental to speech intelligibility and ASR performance, as

$$y[n] = h_e[n] * x[n] + h_\ell[n] * x[n]. \tag{2}$$

The goal of dereverberation is to reduce the effect of late reflections represented by $h_\ell[n]$. By applying the Short-time Fourier Transform (STFT) on (1), we obtain

$$Y(l,k) = H(l,k)X(l,k), \tag{3}$$

where $Y(l,k)$, $X(l,k)$ and $H(l,k)$ represent the STFT of $y[n]$, $x[n]$, and $h[n]$, respectively, with l denoting the time frame index and k representing the frequency bin index. This formulation forms the basis for many dereverberation approaches, including both model-based and data-driven techniques.

A widely used model-based approach for speech dereverberation is the WPE algorithm [9]–[11]. It employs linear prediction and optimization to estimate and subtract the reverberant component, thereby estimating $h_e[n] * x[n]$. Suitable for both single- and multi-channel settings, WPE serves as the baseline model in this work. Although it offers advantages including relatively low computational complexity and robustness across various reverberation conditions, WPE's performance may be constrained in single-channel scenarios and when the observed speech duration is short, as both conditions can lead to accumulated estimation errors [13].

Deep Neural Network (DNN)-based methods, both supervised and unsupervised, discriminative and generative, have been extensively explored [12], [14]–[19]. These approaches enhance adaptability by modeling speech characteristics and acoustic variations during training. However, the training process typically relies heavily on large datasets, which introduces substantial computational costs and challenges in generalizing to unseen conditions. Recent work has shown that self-supervised speech enhancement techniques can reduce dependence on reverberant-clean pairs by utilizing unpaired or solely reverberant data [20], [21]. However, these approaches still demand considerable training data, increasing the training cost, particularly when data acquisition is challenging. Consequently, this has spurred our interest in alternative methods that minimize data requirements while maintaining adaptability across diverse conditions.

III. DIL in Image Processing

DIL is a deep learning paradigm initially developed for image processing tasks such as image super-resolution, denoising, and deblurring [22]–[25]. Unlike traditional DNNs that rely on large external datasets for training, DIL enables networks to be trained exclusively on examples extracted from a single image input I at test time by leveraging its internal self-similarity [23], [26]. This inherent recurrence or similarity suggests that even when an image is degraded, key features such as edges, textures, or patterns remain sufficiently preserved to allow reconstruction. This fundamental insight allows DIL to infer and restore the missing details of the original image by analyzing the repetitive information across different scales. More specifically, for the image super-resolution, an image-specific model is trained to learn the mapping from a downscaled version $I \downarrow q$ (where q is the super-resolution scale factor) to its corresponding observed test image I. Once trained, the model can be applied to reconstruct the desired high-resolution version $I \uparrow q$ of the input image. This self-contained (internal) learning eliminates the dependence on extensive external training datasets, positioning DIL as a powerful alternative to traditional deep learning methods.

Despite its success in image processing, there has been limited exploration of DIL in other domains, such as speech processing. Speech spectrograms, like images, are two-dimensional representations that exhibit structural recurrence, including harmonics, formants, and rhythmic patterns. This parallel between images and speech spectrograms suggests that DIL could potentially be extended to various speech processing tasks, such as speech denoising, dereverberation, and bandwidth extension. In this work, we specifically explore the application of DIL to speech dereverberation due to its conceptual similarity to image deblurring.

IV. DIL for Speech Dereverberation

A. Proposed Framework

Unlike conventional DNN speech dereverberation approaches that require training on a large number of reverberant-anechoic speech pairs, DIL exploits example pairs derived solely from the reverberant spectrogram input itself, eliminating the need for external clean references.

We consider a clean speech log-magnitude spectrogram $X(l,k)$ and the corresponding observed reverberant spectrogram $Y(l,k)$, given by

$$Y(l,k) = H_1(l,k)X(l,k), \tag{4}$$

where $H_1(l,k)$ represents the room reverberation in the time-frequency domain. The first step in DIL is to create a degraded version $Y'(l,k)$, which is a more reverberant counterpart of $Y(l,k)$, such that

$$Y'(l,k) = H_2(l,k)Y(l,k), \tag{5}$$

where $H_2(l,k)$ introduces additional reverberation to $Y(l,k)$. In general, $H_2(l,k)$ is unknown and represents a degradation model. We will consider both oracle $H_2(l,k) = H_1(l,k)$, and approximations $H_2(l,k) \neq H_1(l,k)$, which distinguish $H_2(l,k)$ from the original RIR $H_1(l,k)$.

Fig. 1 presents our DIL framework. In practice, the only available information is $Y(l,k)$. By training on paired examples $V(k)$ and $U(k)$ extracted from $Y'(l,k)$ and $Y(l,k)$ respectively, a speech-specific model $M(Y',Y)$ learns the mapping between $Y'(l,k)$ and $Y(l,k)$. During inference, this

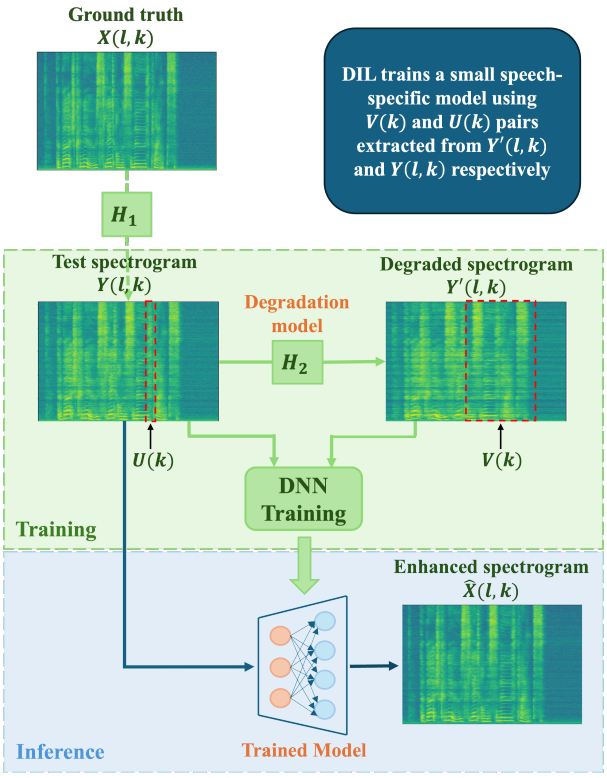

Fig. 1. DIL Framework for Speech Dereverberation

model is applied to $Y(l,k)$ to give an estimate, $\hat{X}(l,k)$, of $X(l,k)$. This DIL framework uses $M(Y',Y)$ as a proxy for the ideal model $M(Y,X)$, which would directly estimate $X(l,k)$ from $Y(l,k)$. The method's success depends on its sensitivity to the accuracy of approximating $M(Y,X)$ with $M(Y',Y)$. This point will be addressed in Section *V*.

B. Data Pre-processing & Model Architecture

The STFT is computed using Hann-windowed frames with a length $N = 1024$, hop-size of 128 samples and sampling frequency 16 kHz. This DIL model operates only on the spectrogram magnitude, assuming the phase remains unchanged during reconstruction. However, this assumption may introduce artifacts, as discussed later.

Since the diversity of reverberant-clean relationships within a single reverberant speech example is significantly lower than that in an entire training set, the DIL approach enables a lightweight model architecture to be employed, compared to conventional DNNs, which typically require deeper structures to learn this mapping. We therefore adopt a simple fully convolutional network with 10 convolutional layers, each using 3×3 kernels (stride 1), ReLU activation, and a 0.2 dropout rate, except the last layer where both are omitted. Fig. 2 illustrates the model architecture. The network processes a single-channel spectrogram (1×21×513) through 16 feature maps per hidden layer while preserving spatial dimensions. A fully connected layer maps features back to a single-channel output (1x513). A global residual connection adds the input spectrogram to the output, preserving essential details while leveraging learned features. For optimization, we use the MSEloss with the Adam optimizer, initializing the learning rate at 10^{-5}. A MultistepLR scheduler adjusts the learning rate at epochs 100 and 150 (decay factor 0.1). Early stopping is applied with a patience of 5 epochs, terminating training when no significant improvement in loss ($\Delta < 10^{-5}$) is observed.

The temporal context and windowing strategy construct the training input $V(k)$ by extracting 21 consecutive frames centered at frame l from $Y'(l,k)$. The corresponding training target $U(k)$ contains the full range frequency of $Y(l,k)$ at frame l. These example pairs are highlighted in Fig. 2 with a dashed red outline. During inference, the same windowing strategy is applied to $Y(l,k)$ before feeding it into the trained model. The model outputs are assembled to reconstruct the complete spectrogram $\hat{X}(l,k)$. Since no information is available for the first and last 10 frames of $\hat{X}(l,k)$, they are temporarily filled with values from $Y(l,k)$ during reconstruction. However, these boundary frames are excluded from the evaluation to prevent edge effects. Notably, while $Y'(l,k)$ serves as a pseudo-label to guide the training, all samples originate from the test input itself without incorporating any external data. This exclusive reliance on the information within the signal defines DIL as a zero-shot learning paradigm, enabling adaptation without pre-trained models.

V. Experiments & Results

A. Experiment Set-up

We randomly selected 100 clean speech signals from the IEEE Sentences dataset, including both male and female voices. To generate the RIRs, we used the Image-Source Method (ISM) [27] with three different reverberation times (T60): [204, 513, 972] ms, as summarized in Table I.

TABLE I
Configurations of Experimental Room Parameters

	T60=204 ms	**T60=513 ms**	**T60=972 ms**
Room Size	[5,6,4]	[8, 9, 5]	[8, 9, 5]
Source Position	[3,4,2]	[7,8,3]	[7,8,3]
Microphone Position	[3,3,2]	[7,5,3]	[7,5,3]
Reflection Factor	0.6	0.8	0.9

Performance evaluation for speech dereverberation lacks a single standardized metrics. Thus, a combination of metrics were used. We used Perceptual Evaluation of Speech Quality (PESQ) [28] for quality assessment, Short-Time Objective Intelligibility (STOI) [29] for intelligibility, and Normalized Signal-to-Reverberation Ratio (NSRR) [30] to quantify the power ratio between the direct path and reverberant components of the received signal.

B. Oracle Case: $H_2(l,k) = H_1(l,k)$

To evaluate the feasibility of the DIL approach for speech dereverberation, we first conducted an oracle experiment where the ground truth $H_1(l,k)$ was assumed to be known, and the oracle choice $H_2(l,k) = H_1(l,k)$ was used. The results

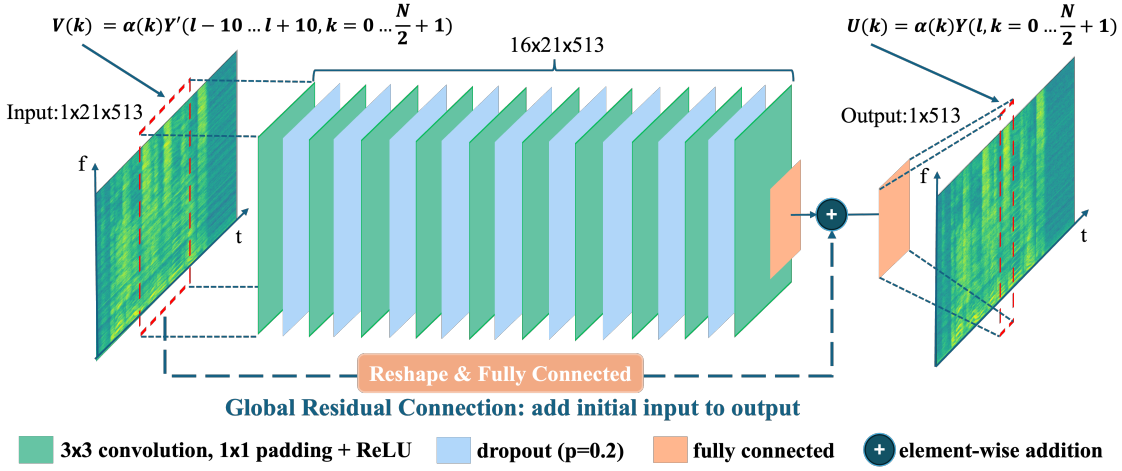

Fig. 2. Model Architecture with the windowing strategy for data augmentation. The red window $V(k)$ represents a single training input example, while the red window $U(k)$ denotes its corresponding training target. By shifting both windows one time frame to the right, the next training example pair is obtained. $a(k)$ denotes the frequency weight, which was 1 in our experiments.

TABLE II
PERFORMANCE COMPARISON. MEAN (STD. DEV.) DIL_O REPRESENTS THE ORACLE CASE.

	T60 = 204 ms			T60 = 513 ms			T60 = 972 ms		
	PESQ	**STOI**	**NSRR (dB)**	**PESQ**	**STOI**	**NSRR (dB)**	**PESQ**	**STOI**	**NSRR (dB)**
Observed	1.99 (0.19)	0.72 (0.03)	-7.22 (1.08)	1.82 (0.16)	0.69 (0.03)	-8.83 (1.96)	1.75 (0.15)	0.67 (0.03)	-9.42 (2.07)
WPE	2.19 (0.23)	0.77 (0.02)	-6.39 (1.05)	1.96 (0.20)	0.75 (0.03)	-8.09 (1.95)	1.87 (0.18)	0.72 (0.03)	-8.64 (2.02)
DIL_o	**2.70 (0.17)**	**0.89 (0.02)**	**2.11 (1.93)**	**2.46 (0.14)**	**0.86 (0.02)**	**2.30 (1.75)**	**2.38 (0.15)**	**0.85 (0.02)**	**2.16 (1.68)**
DIL	**2.49 (0.19)**	**0.85 (0.02)**	**-1.1 (1.66)**	**2.31 (0.14)**	**0.83 (0.02)**	**0.77 (1.54)**	**2.28 (0.14)**	**0.82 (0.02)**	**0.99 (1.68)**

of this experiment, (**DIL_o**), were compared with the baseline WPE, which was configured according to the recommendations in [10], [11]. Table II shows the mean performance and standard deviations over 100 test utterances. **Observed** refers to the unprocessed reverberant speech input. Audio demonstrations are available at https://yuxiz0826.github.io/demos/.

The oracle experiment results demonstrate the upper bound performance of this DIL approach across various reverberation conditions. While DIL experiences a slight performance degradation as T60 increases, it consistently outperforms WPE, achieving superior scores across all metrics. Notably, DIL maintains positive NSRR values under all conditions, indicating significant dereverberation improvements, whereas WPE's NSRR scores remain negative.

Although oracle results provide an upper bound for the DIL approach, PESQ and STOI do not reach their theoretical maximums. One potential reason is that the DIL model operates solely on the spectrogram magnitude without phase information. As a result, phase distortions in the reconstructed waveform may negatively impact perceptual quality and intelligibility, leading to a potential reduction in scores. This limitation also manifests as certain artifacts in the enhanced audio, raising an open question related to their nature. Incorporating phase estimation is therefore a key direction for future work.

C. Realistic Case: $H_2(l,k) \neq H_1(l,k)$

This experiment evaluates whether the DIL approach remains effective under more realistic assumptions when $H_1(l,k)$ is typically unknown. Specifically, only the T60 of $H_1(l,k)$ is assumed to be known, which can be estimated using existing techniques [31]. Here, $H_2(l,k)$ is generated by multiplying uniformly distributed random noise samples with an exponentially decaying envelope corresponding to the assumed T60 of $H_1(l,k)$. The experimental results are again compared with WPE, as presented in Table II (**DIL**).

We observe that while the performance of DIL slightly decreases compared to the oracle case, it remains robust in reducing reverberation and still consistently outperforms WPE across all T60 conditions. This demonstrates DIL's effectiveness even without full oracle knowledge of $H_1(l,k)$.

However, this experiment also highlights several key limitations. First, an accurate estimation of T60 was assumed, which may not hold in practical applications, as T60 estimation algorithms often introduce errors and variability. Therefore, assessing the model's sensitivity to potential inaccuracies is crucial in T60 estimation. Second, the randomized reverberation filter used for $H_2(l,k)$ does not fully capture real RIR charateristics, such as early reflections and late reverberation. This highlights the model's sensitivity to deviations between $H_1(l,k)$ and its approximation $H_2(l,k)$, suggesting the future work should evaluate DIL using measured RIRs

from actual acoustic environments. Additional improvements include multi-channel processing, which could leverage spatial information to enhance dereverberation further.

VI. Discussion & Conclusion

This paper proposes the DIL paradigm for speech enhancement and presents its formulation and evaluation for speech dereverberation, eliminating reliance on external training data. We demonstrate its effectiveness in single-channel speech dereverberation in both oracle and more realistic scenarios. Our experimental results show that DIL outperforms the WPE baseline by 0.35 in PESQ, 0.08 in STOI and 8.86 in NSRR (dB) for T60 around 0.5 s, delivering robust performance across varying T60 conditions. Compared to the state-of-the-art ML algorithm [12], which reports an improvement over WPE by 0.7 in PESQ and 0.16 in STOI, DIL maintains strong performance while avoiding the need for a pre-trained clean speech model. Incorporating phase information during training could help mitigate artifacts, while using measured RIRs may improve robustness in real-world applications. Future work will examine the model's sensitivity to T60 estimation errors and explore multi-channel setups to further enhance the reverberation adaptability in diverse acoustic environments.

References

[1] J.-M. Lemercier, J. Thiemann, R. Koning, and T. Gerkmann, "A neural network-supported two-stage algorithm for lightweight dereverberation on hearing devices," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2023, no. 1, p. 18, May 2023.

[2] T. Lei, Z. Hou, Y. Hu, W. Yang, T. Sun, X. Rong, D. Wang, K. Chen, and J. Lu, "A Low-Latency Hybrid Multi-Channel Speech Enhancement System For Hearing Aids," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Jun. 2023, pp. 1–2.

[3] M. Togami, Y. Kawaguchi, and N. Nukaga, "Online speech dereverberation with time-varying assumption of acoustic transfer functions for teleconferencing systems," in *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*, Aug. 2012, pp. 136–141.

[4] R. Gomez, T. Kawahara, and K. Nakadai, "Robust hands-free Automatic Speech Recognition for human-machine interaction," in *2010 10th IEEE-RAS International Conference on Humanoid Robots*, Dec. 2010, pp. 138–143.

[5] W. Zhang, C. Boeddeker, S. Watanabe, T. Nakatani, M. Delcroix, K. Kinoshita, T. Ochiai, N. Kamo, R. Haeb-Umbach, and Y. Qian, "End-to-End Dereverberation, Beamforming, and Speech Recognition with Improved Numerical Stability and Advanced Frontend," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Jun. 2021, pp. 6898–6902.

[6] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech Enhancement and Dereverberation With Diffusion-Based Generative Models," *IEEE Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2351–2364, 2023.

[7] G. Li, J. Deng, M. Geng, Z. Jin, T. Wang, S. Hu, M. Cui, H. Meng, and X. Liu, "Audio-Visual End-to-End Multi-Channel Speech Separation, Dereverberation and Recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2707–2723, 2023.

[8] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer Science & Business Media, Jul. 2010.

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2008, pp. 85–88.

[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, Language Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[11] T. Yoshioka and T. Nakatani, "Generalization of Multi-Channel Linear Prediction Methods for Blind MIMO Impulse Response Shortening," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[12] E. Moliner, J.-M. Lemercier, S. Welker, T. Gerkmann, and V. Välimäki, "BUDDy: Single-Channel Blind Unsupervised Dereverberation with Diffusion Models," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2024, pp. 120–124.

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

[14] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning Spectral Mapping for Speech Dereverberation and Denoising," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[15] Y. Zhao, Z.-Q. Wang, and D. Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2017, pp. 5580–5584.

[16] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 4506–4510.

[17] 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. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2022, pp. 7412–7416.

[18] C. Li, T. Wang, S. Xu, and B. Xu, "Single-channel Speech Dereverberation via Generative Adversarial Training," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 1309–1313.

[19] V. Kothapally and J. H. L. Hansen, "SkipConvGAN: Monaural Speech Dereverberation Using Generative Adversarial Networks via Complex Time-Frequency Masking," *IEEE Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1600–1613, 2022.

[20] Y.-C. Wang, S. Venkataramani, and P. Smaragdis, "Self-supervised Learning for Speech Enhancement," in *Proceedings of the 37 Th International Conference on Machine Learning*, Vienna, Austria, 2020.

[21] T. Hussain, R. E. Zezario, Y. Tsao, and A. Hussain, "Speech Dereverberation Based on Self-supervised Residual Denoising Autoencoder with Linear Decoder," in *Proceedings of ELM 2022*, K.-M. Björk, Ed. Cham: Springer Nature Switzerland, 2024, pp. 46–57.

[22] T. Tirer, R. Giryes, S. Y. Chun, and Y. C. Eldar, "Deep Internal Learning: Deep learning from a single input," *IEEE Signal Processing Magazine*, vol. 41, no. 4, pp. 40–57, Jul. 2024.

[23] A. Shocher, N. Cohen, and M. Irani, ""Zero-Shot" Super-Resolution Using Deep Internal Learning," in *Proceedings of the IEEE Conference on Computer Vision and Patrn Recognition*, 2018, pp. 3118–3126.

[24] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep Image Prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[25] J. Zukerman, T. Tirer, and R. Giryes, "BP-DIP: A Backprojection based Deep Image Prior," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 675–679.

[26] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 349–356.

[27] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[28] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)-a New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 2, Salt Lake City, UT, USA, May 2001, pp. 749–752.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[30] P. A. Naylor, E. A. P. Habets, J. Y.-C. Wen, and N. D. Gaubitch, "Models, Measurement and Evaluation," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. London: Springer, 2010, pp. 21–56.

[31] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of Room Acoustic Parameters: The ACE Challenge," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.