Interspeech 2022
18-22 September 2022, Incheon, Korea

Memory-Efficient Multi-Step Speech Enhancement with Neural ODE

Jen-Hung Huang and Chung-Hsien Wu

Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

P78101514@gs.ncku.edu.tw, chunghsienwu@gmail.com

Abstract

Although deep learning-based models proposed in the past years have achieved remarkable results on the speech enhancement tasks, the existing multi-step denoising methods require a memory size proportional to the number of steps during training, which makes it difficult to apply to large models. In this paper, we propose a memory-efficient multi-step speech enhancement method that requires only constant amount of memory for model training. This End-to-End method combines Neural Ordinary Differential Equations (Neural ODEs) with the Memory-efficient Asynchronous Leapfrog Integrator (MALI) for multi-step training. Experiments on the Voice Bank and DEMAND datasets showed that the multi-step method using MALI had better performance than the single-step method, with maximum improvements of 0.16 on PESQ and 0.5% on STOI. In addition to reducing the memory required for model training, this method is also quite competitive with the current state-of-the-art methods.

Index Terms: speech enhancement, deep learning, neural ODEs, multi step

1. Introduction

In recent years, more and more tasks that rely on speech as a medium for interaction and communication have been used in daily life, such as online meetings, hearing aids, speech recognition, etc. However, the environment is full of various background noises, which will seriously affect the performance of these tasks. Therefore, speech enhancement, which removes noises and improves the speech quality, is an important preprocessing task. Nevertheless, traditional speech enhancement methods based on statistical models, such as Wiener filtering [1], spectral subtraction [2] and minimum mean square error (MMSE) estimation [3, 4], are unlikely to effectively remove the non-stationary noise that permeates our daily life.

Recently, speech enhancement research used data-driven and larger model-based approaches [5, 6, 7, 8]. Among them, training the deep learning models such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) and Transformers for single-step denoising achieved significantly better results than traditional methods. Ideally, the performance can be improved by increasing the model size, but doing so will not only increase the training cost, but may also lead to overfitting, which makes the benefits less than expected. In contrast, as humans have the ability to understand content through repeatedly listening the speech signals disturbed by a certain degree of noise, it is reasonable to use iterative multi-step denoising method to achieve a better result. However, for existing multi-step speech denoising methods such as DSEGAN [9] and SNR-Progressive Learning [10], the memory consumption during training increases linearly with the number of steps. This makes it difficult to train larger models for multi-step denoising.

Algorithm 1 Forward Function $\psi_{f_\theta,\Delta t}$ of ALF

Input	z_{in}, v_{in}, t_{in}
Output	$z_{out}, v_{out}, t_{out}$
Forward	$t \leftarrow t_{in} + \Delta t/2$
	$z \leftarrow z_{in} + v_{in} \times \Delta t/2$
	$v \leftarrow f_\theta(z, t)$
	$v_{out} \leftarrow v_{in} + 2(v - v_{in})$
	$z_{out} \leftarrow z + v_{out} \times \Delta t/2$
	$t_{out} \leftarrow t + \Delta t/2$

Algorithm 2 Inverse Function $\psi^{-1}_{f_\theta,\Delta t}$ of ALF

Input	$z_{out}, v_{out}, t_{out}$
Output	z_{in}, v_{in}, t_{in}
Inverse	$t \leftarrow t_{out} - \Delta t/2$
	$z \leftarrow z_{out} - v_{out} \times \Delta t/2$
	$v \leftarrow f_\theta(z, t)$
	$v_{in} \leftarrow 2v - v_{out}$
	$z_{in} \leftarrow z - v_{in} \times \Delta t/2$
	$t_{in} \leftarrow t - \Delta t/2$

In this study, we propose a memory-efficient multi-step speech enhancement method that, different from previous methods, just requires a fixed memory size for multiple iterations. Our approach is based on the Neural Ordinary Differential Equations (Neural ODE) [11] and combined with the Memory-efficient Asynchronous Leapfrog Integrator (MALI) [12], which has the backward pass towards correct training direction. We apply this method to UNet [13] without special design and conduct experiments on the public Voice Bank + Demand dataset [14]. The results confirmed that this method not only reduces the memory size required to train a multi-step speech enhancement model, but also has quite competitive performance compared to the current state-of-the-art methods.

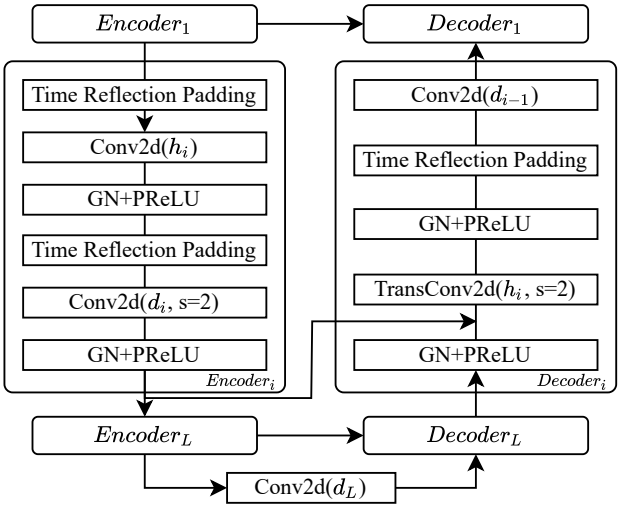
2. Methodology

2.1. Memory-efficient Asynchronous Leapfrog Integrator

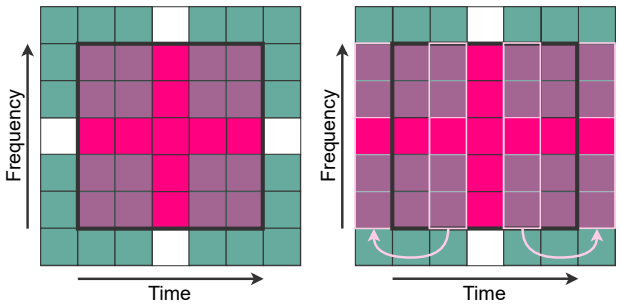
The original Neural Ordinary Differential Equations (Neural ODEs) [11], which use the Adjoint Method, computes gradients that are not accurate in the backward pass. As a result, deviations occur when the model parameters are corrected, and it is difficult to train stably when applied to large models. To solve this problem, Adaptive Checkpoint Adjoint (ACA) [15] which needs to record intermediate states and Memory-efficient Asynchronous Leapfrog Integrator (MALI) [12] based on Asynchronous Leapfrog (ALF) [16] is proposed.

$$z_T, v_T \leftarrow \psi^N_{f_\theta,\Delta t}(z_{t_0}, v_{t_0}),\ v_{t_0} = f_\theta(z_{t_0}) \tag{1}$$

10.21437/Interspeech.2022-868



(a)



(b) *left: 2d Zero Padding, right: Time Reflection Padding*

Figure 1: *(a) Overview of UNet architecture and (b) The Time Reflection Padding.*

Algorithm 3 Fixed Step Memory-efficient Asynchronous Leapfrog Integrator

Initialize

start time t_0, end time T, ODE function f_θ, step size Δt, and step number $N = \frac{T-t_0}{\Delta t}$

Forward

Input start state $(z_{t_0}\ v_{t_0})$, and $v_{t_0} = f_\theta(z_{t_0}, t_0)$
Output z_T
with no grad:
for i in $\{1, 2, \ldots, N\}$:
$z_{t_i}, v_{t_i}, t_i \leftarrow \psi_{f_\theta, \Delta t}(z_{t_{i-1}}, v_{t_{i-1}}, t_{i-1})$
Save final state (z_T, v_T) for backward

Backward

Input $a(T) = \begin{bmatrix} \frac{\partial L}{\partial z_T} & 0 \end{bmatrix}$
Output $a(t_0)$, $\frac{dL}{d\theta}$
Load final state (z_T, v_T)
Initialize $\frac{dL}{d\theta} \leftarrow 0$
for i in $\{N, \ldots, 2, 1\}$:
with no grad:
$z_{t_{i-1}}, v_{t_{i-1}}, t_{i-1} \leftarrow \psi^{-1}_{f_\theta, \Delta t}(z_{t_i}, v_{t_i}, t_i)$
with enable grad:
$z_{t_i}, v_{t_i}, t_i \leftarrow \psi_{f_\theta, \Delta t}(z_{t_{i-1}}, v_{t_{i-1}}, t_{i-1})$
$a(t_{i-1}) \leftarrow a(t_i) \begin{bmatrix} \frac{\partial z_{t_i}}{\partial z_{t_{i-1}}} & \frac{\partial z_{t_i}}{\partial v_{t_{i-1}}} \\ \frac{\partial v_{t_i}}{\partial z_{t_{i-1}}} & \frac{\partial v_{t_i}}{\partial v_{t_{i-1}}} \end{bmatrix}$
$\frac{dL}{d\theta} \leftarrow \frac{dL}{d\theta} + a(t_i) \begin{bmatrix} \frac{\partial z_{t_i}}{\partial \theta} \\ \frac{\partial v_{t_i}}{\partial \theta} \end{bmatrix}$

The forward-pass of MALI is represented in Eq. 1, which is composed of N ALF forward functions (Algo. 1) $\psi_{f_\theta, \Delta t}$. f_θ and Δt represent ordinary differential equation with trainable parameters θ and time step size, respectively. By relying on the inverse function $\psi^{-1}_{f_\theta, \Delta t}$ (Algo. 2) of $\psi_{f_\theta, \Delta t}$, MALI does not need to record the computation graph and intermediate states in the forward pass. As long as the final state z_T is saved, the model can be updated according to Algo. 3 by calculating the correct gradient through local backpropagation.

2.2. Model

2.2.1. UNet

This experiment uses UNet [13] as the ordinary differential equation f_θ used by MALI, and its structure is shown in Fig. 1a, which consists of pairs of encoder blocks and decoder blocks. Each block is composed of two layers of convolutions. When passing the second layer, downsample/upsample of 2×2 is performed, and GroupNorm [17] and PReLU [18] are added between all convolutions.

2.2.2. Time Reflection Padding

The use of zero padding in CNNs over multiple iterations may result in the gradual accumulation of boundary cut-off information, which affects non-boundary audio and causes distortion. In order to avoid this situation affecting the performance, the Time Reflection Padding shown in Fig. 1b is used as the padding method in UNet.

2.2.3. Dimension Augmentation

Dupont et al. [19] found that extending the dimensionality of the state causes Neural ODEs to substantially speed up learning and increase performance. In Eq. 2, the Complex Spectrogram is expanded from the original $\mathbb{R}^{B \times 2 \times F \times T}$ to $\mathbb{R}^{B \times D \times F \times T}$ as the initial state z_{t_0} of MALI. The proposed method performs a linear projection on the final state z_T in Eq. 3 to get the estimated spectrum $\tilde{S} \in \mathbb{R}^{B \times 2 \times F \times T}$.

$$z_{t_0} = \{X_{real}, X_{imag}, 0_1, \ldots, 0_{D-2}\} \tag{2}$$

$$\{\tilde{S}_{real}, \tilde{S}_{imag}\} = z_T W + b,\ W \in \mathbb{R}^{D \times 2} \tag{3}$$

Table 1: *Comparison with the state-of-the-art methods on the Voice Bank + DEMAND dataset.*

Methods	Year	Param.	PESQ	STOI (%)	CSIG	CBAK	COVL
Noisy	–	–	1.97	92.1	3.35	2.44	2.63
SOTA speech enhancement approaches							
SEGAN[20]	2017	43.2M	2.16	92.5	3.48	2.94	2.80
MMSEGAN[21]	2018	–	2.53	93.0	3.80	3.12	3.14
MetricGAN[22]	2019	1.86 M	2.86	–	3.99	3.18	3.42
CRGAN[23]	2020	–	2.92	94.0	4.16	3.24	3.54
DCCRN[24]	2020	3.7 M	2.68	93.7	3.88	3.18	3.27
RDL-Net[25]	2020	3.91 M	3.02	93.8	4.38	3.43	3.72
PHASEN[26]	2020	–	2.99	–	4.21	3.55	3.62
MHSA-SPK[27]	2020	–	2.99	–	4.15	3.42	3.53
T-GSA[28]	2020	–	3.06	93.7	4.18	3.59	3.62
TSTNN[29]	2021	0.92 M	2.96	95.0	4.17	3.53	3.49
DEMUCS[30]	2021	128 M	3.07	95.0	4.31	3.40	3.63
GaGNet[31]	2021	5.94 M	2.94	94.7	4.26	3.45	3.59
MetricGAN+[32]	2021	–	3.15	–	4.14	3.16	3.64
SE-Conformer[33]	2021	–	3.13	95.0	4.45	3.55	3.82
MMB-AIAT[34]	2021	0.90 M	3.11	94.9	4.45	3.60	3.79
CRB-AIAT[34]	2021	1.17 M	3.15	94.7	4.48	3.54	3.81
DB-AIAT[34]	2021	2.81 M	3.31	95.6	4.61	3.75	3.96
Proposed approaches							
Small-MALI-SE	2022	0.48 M	3.07	94.6	4.34	2.35	3.73
Medium-MALI-SE	2022	0.99 M	3.15	94.9	4.40	2.39	3.82
Large-MALI-SE	2022	2 M	3.21	95.0	4.46	2.46	3.88

Table 2: *Compare the performance of Small, Medium and Large models at different step sizes.*

Models	Step size	PESQ	STOI (%)	SISDR
Noisy	–	1.97	92.1	8.45
Small (0.48 M)	w/o MALI	2.97	94.2	19.29
	1	3.03	94.5	19.53
	0.5	3.07	94.6	19.51
	0.25	3.00	94.7	19.51
	0.125	3.04	94.6	19.75
Medium (0.99 M)	w/o MALI	2.99	94.4	19.09
	1	3.11	94.7	19.41
	0.5	3.15	94.8	19.62
	0.25	3.14	94.9	19.41
	0.125	3.07	94.9	19.75
Large (2 M)	w/o MALI	3.07	94.6	19.31
	1	3.11	94.8	19.53
	0.5	3.13	94.9	19.15
	0.25	3.20	94.9	19.70
	0.125	**3.21**	**95.0**	**19.97**

Table 3: *The effect of dimension augmentation (DA) on the performance of the Medium model.*

Models	Step size	PESQ	STOI (%)	SISDR
Noisy	–	1.97	92.1	8.45
Medium	1	3.11	94.7	19.41
	0.5	**3.15**	**94.8**	**19.62**
w/o DA	1	3.03	94.6	19.29
	0.5	3.02	94.7	19.32

3. Experiments

3.1. Datasets

The experiment in this study used the public data set proposed by [14], which contained 30 speakers in the Voice Bank corpus [35], in which 28 was used as the training data, and 2 was the test data. Also, this study referring to [36] selected the data from two speakers (p226 and p287) in the training data as the validation set. The selected data were further mixed with noise from DEMAND database [37] to produce noisy speech data.

3.2. Training Details

In this experiment, three sizes of UNets, Small (0.48 M), Medium (0.99 M) and Large (2 M), were constructed on the basis of different number of parameters, all of which are stacked by 4 encoders/decoders. All convolution kernel size and GroupNorm groups were set to 3 and 8. The channel setting of each layer was selected from $c = \{8, 16, 32, 48, 64, 96, 128, 196\}$. The hyperparameters d and h in Fig. 1a were $\{c_{i+1}, \dots c_{i+4}\}$, $\{c_{i+2}, \dots c_{i+5}\}$, and the dimension augmentation $D = c_i$. Small, Medium and Large models set the values of i to 0, 1 and 2, respectively. The start time t_0 and end time T of MALI were from 0 to 1. This experiment did not use the current time t as the conditional input to the model.

After down-sampling all speech to 16 kHz in preprocessing step, the complex spectrogram was obtained through STFT of 511-sample window size and 63-sample hop size. This study

used Hanning window as the window function when transforming. During training, the length of the speech segment of each data was 32194 samples (about 2 seconds), and reflection padding was performed when the audio file duration was less than the aforementioned length. We used RAdam [38] as the optimizer for this experiment, the learning rate for Small and Medium was 5e-3, for Large was 2e-3, and the batch size was set to 16. After training for 80 epochs, the model with the best performance on the validation set was selected as the final result.

3.3. Main Results

We used PSEQ [39], STOI [40], CSIG [41], CBAK [41] and COVL [41] as the metrics for speech enhancement performance evaluation. In Table. 1, MALI-UNet was compared with multiple state-of-the-art baselines using the above evaluation metrics, including SEGAN [20], MMSEGAN [21], MetricGAN [22], CRGAN [23], DCCRN [24], RDL-Net [25], PHASEN [26], MHSA-SPK [27], T-GSA [28], TSTNN [29], DEMUCS [30], GaGNet [31], MetricGAN+ [32], SE-Conformer [33]. UNet, which was not specially designed for speech enhancement tasks still had the ability to compete with these baselines after using MALI.

Table 2 shows the comparison of performance over Small, Medium and Large models at various step sizes. It can be observed that the multi-step denoising method using MALI had better performance than the single-step method without NODEs (w/o MALI) on the models of different sizes, with maximum improvements of 0.16 on PESQ and 0.5% on STOI. On the other hand, Table 3 shows that dimension augmentation had a large effect on the performance gain brought by multi-step training. Removing it will greatly reduce the information that can be reused between states, and even with MALI. There is no significant improvement in various evaluation metrics.

4. Conclusions

In this paper, we propose a memory-efficient multi-step speech enhancement method that combines Neural ODEs and MALI. The amount of memory required for this method to train is independent of the number of iterations. Therefore, it can also be used for models that consume a large amount of memory and does not require any structural changes when combined with other baselines. In addition, the study found that by expanding the dimension of the intermediate state, the performance of models of different sizes will be significantly improved after multiple iterations. On public datasets, using our method with an unspecified UNet had the ability to compete with other state-of-the-art methods.

5. References

[1] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[4] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[5] J.-C. Wang, Y.-S. Lee, C.-H. Lin, S.-F. Wang, C.-H. Shih, and C.-H. Wu, "Compressive sensing-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2122–2131, 2016.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[8] S.-C. Chu, C.-H. Wu, and Y.-W. Lin, "Speech enhancement based on masking approach considering speech quality and acoustic confidence for noisy speech recognition," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 536–540.

[9] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving gans for speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1700–1704, 2020.

[10] Y.-H. Tu, J. Du, T. Gao, and C.-H. Lee, "A multi-target snr-progressive learning approach to regression based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1608–1619, 2020.

[11] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *NeurIPS*, 2018.

[12] J. Zhuang, N. C. Dvornek, S. Tatikonda, and J. S. Duncan, "Mali: A memory efficient and reverse accurate integrator for neural odes," *International Conference on Learning Representations*, 2021.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[14] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks," in *Proc. Interspeech 2016*, 2016, pp. 352–356.

[15] J. Zhuang, N. Dvornek, X. Li, S. Tatikonda, X. Papademetris, and J. Duncan, "Adaptive checkpoint adjoint method for gradient estimation in neural ode," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 639–11 649.

[16] U. Mutze, "An asynchronous leapfrog method ii," *arXiv preprint arXiv:1311.6602*, 2013.

[17] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[19] E. Dupont, A. Doucet, and Y. W. Teh, "Augmented neural odes," in *NeurIPS*, 2019.

[20] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech 2017*, 2017, pp. 3642–3646.

[21] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5039–5043.

[22] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[23] Z. Zhang, C. Deng, Y. Shen, D. S. Williamson, Y. Sha, Y. Zhang, H. Song, and X. Li, "On Loss Functions and Recurrency Training for GAN-Based Speech Enhancement Systems," in *Proc. Interspeech 2020*, 2020, pp. 3266–3270.

[24] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," in *INTERSPEECH*, 2020.

[25] M. Nikzad, A. Nicolson, Y. Gao, J. Zhou, K. K. Paliwal, and F. Shang, "Deep residual-dense lattice network for speech enhancement," in *AAAI*, 2020.

[26] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9458–9465.

[27] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 181–185, 2020.

[28] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with gaussian-weighted self-attention for speech enhancement," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6649–6653, 2020.

[29] K. Wang, B. He, and W. Zhu, "Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7098–7102, 2021.

[30] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *INTERSPEECH*, 2020.

[31] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, p. 108499, 2022.

[32] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 201–205.

[33] E. Kim and H. Seo, "SE-Conformer: Time-Domain Speech Enhancement Using Conformer," in *Proc. Interspeech 2021*, 2021, pp. 2736–2740.

[34] G. Yu, A. Li, Y. Wang, Y. Guo, H. Wang, and C. Zheng, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," *arXiv preprint arXiv:2110.06467*, 2021.

[35] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–4.

[36] S. Fu, C. Yu, K. Hung, M. Ravanelli, and Y. Tsao, "Metricgan-u: Unsupervised speech enhancement/ dereverberation based only on noisy/ reverberated speech," *CoRR*, vol. abs/2110.05866, 2021. [Online]. Available: https://arxiv.org/abs/2110.05866

[37] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.

[38] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, April 2020.

[39] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.

[40] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[41] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.