INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

DeFT-AN RT: Real-time Multichannel Speech Enhancement using Dense Frequency-Time Attentive Network and Non-overlapping Synthesis Window

Dongheon Lee[1], *Dayun Choi*[1], *Jung-Woo Choi*[1]

[1]Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea

donghen0115@kaist.ac.kr, cdy3773@kaist.ac.kr, jwoo@kaist.ac.kr

Abstract

In real-time speech enhancement models based on the short-time Fourier transform (STFT), algorithmic latency induced by the STFT window size can induce perceptible delays, leading to reduced immersion in real-time applications. This study proposes an efficient real-time enhancement model based on dense frequency-time attentive network (DeFT-AN). The vanilla DeFT-AN consists of cascaded dense blocks and time-frequency transformers, which allow for a smooth transition between time frames through a temporal attention mechanism. To inherit this advantage and reduce algorithmic latency, we develop the lightweight and causal version of DeFT-AN with dual-window size processing that utilizes synthesis windows shorter than analysis windows. The benefit of DeFT-AN in identifying temporal context enables the use of non-overlapping synthesis windows, and experimental results show that the model can achieve the highest performance with the lowest algorithmic latency among STFT-based models.

Index Terms: real-time speech enhancement, dual-window size approach, non-overlapping synthesis window

1. Introduction

In recent years, there has been a surge in demand for teleconferencing due to the pandemic situation and a growing interest in augmented reality and the metaverse. In both these scenarios, it is essential to suppress noises and reverberations from various room conditions in real-time and only transmit the user's voice. Real-time multichannel speech enhancement is a task to achieve this goal and aims at recovering clean speech from multichannel signals captured in a noisy reverberant environment.

Real-time speech enhancement, however, is a complex task that necessitates meeting three extra criteria in addition to those required for general speech enhancement. The challenge lies in achieving causality, low algorithmic latency, and computational complexity. The recent 5th deep noise suppression (DNS) challenge [1] specifies that the model must be causal, which can be accomplished by uni-directional RNNs (RNN [2], LSTM [3], GRU [4], etc.), causal convolution, causal attention, and causal normalization. However, the enhancement performance of the causal system is typically lower than the non-causal system due to missing future information. Next, the algorithmic latency should be less than 20 ms [1]. Algorithmic latency includes the delay caused by algorithm structures such as the STFT window size or encoder kernel size in deep neural networks, which differs from computational latency induced by the computation of the algorithm. Algorithmic latency exceeding the real-time requirement can be perceptible and make the listener less immersive in the communication. Additionally, the real-time factor (RTF) involved with the computational complexity should be less than 0.5 [1]. The RTF is the ratio of the time taken to execute one processing step (computing time) to the corresponding signal duration. Although an RTF lower than 1 is required for real-time operation, values less than 0.5 is recommended for practical applications to accommodate the possible variation in computing time [1].

Real-time speech enhancement can be realized in either the time domain [5, 6] or STFT domain [7, 8]. STFT-domain approaches generally have shown better performance through the utilization of frequency information and spatiotemporal loss functions that can reflect human perception [9, 10]. In STFT-domain approaches, an analysis window is applied to each time frame to convert them to STFT-domain, which is then fed into the deep learning model. The output STFT generated from the model is then transformed back to the temporal waveform by taking inverse STFT (iSTFT), windowing individual frames using a synthesis window, and applying overlap-add operations. STFT-domain approaches typically use a 32 ms window size and 8 ms hop size [11], which results in algorithmic latency of 32 ms that does not meet the real-time requirements of the 5th DNS challenge. To reduce algorithmic latency, STFT-domain approaches like Embedding and Beamforming Network (EaBNet) [12] use a window size of 20 ms and a hop size of 10 ms, but this results in lowered frequency resolution. To address the algorithmic latency problem, time-domain approaches such as the Time-domain audio separation Network (TasNet) [13] and Skipping Memory (SkiM) [14] have been proposed. However, the receptive field of the convolutional encoder in these approaches is smaller than those of STFT-domain approaches, making it difficult to enhance speech in a highly reverberant environment. Thus, the challenge is to handle reverberations from various environments while meeting real-time requirements.

Recently, a study [11] has considered a dual-window size processing [15] for achieving low algorithmic latency with the STFT-based model. Although the proposed model is a non-causal system and has not been designed for real-time processing, the low algorithmic latency of the dual-window size processing can be helpful for real-time speech enhancement [16]. The conventional dual window size processing involves using different sizes for the analysis and synthesis window. The size of the synthesis window utilized for the waveform generation can range from the hop size to the analysis window size. The synthesis window as small as the hop size can reduce the algorithmic latency to the hop size, as shown in Figure 1. But in this case, time frames do not overlap each other and discontinuity can occur at frame boundaries. The discontinuities often reduce speech enhancement performance and cause audible artifacts. If non-overlapping synthesis windows can be implemented with minimal speech degradation, it will be highly beneficial for reducing algorithmic latency in real-time processing.

10.21437/Interspeech.2023-2437

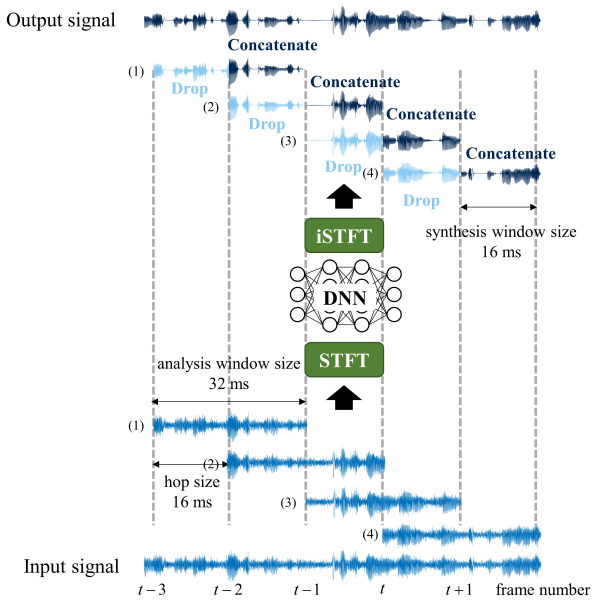

Figure 1: *Illustration of non-overlapping synthesis window*

In this study, we propose a real-time multichannel speech enhancement model, DeFT-AN RT, that enables the use of non-overlapping synthesis windows with negligible audible artifacts. The proposed model is based on Dense Frequency-Time Attentive Network (DeFT-AN) [9] that can handle multichannel noisy reverberant speech with long reverberation times and has shown the highest enhancement performance for various datasets [10, 17]. By leveraging its ability to generate an optimal time-frequency mask through temporal and spectral attention, we demonstrate that the dual-window size processing combined with the causal and lightweight version of DeFT-AN can surpass the performance of state-of-the-art real-time multichannel speech enhancement models reported to date. The proposed model is a transformer-based model, where the causal attention of the T-conformer refers to only the present and previous frames to determine temporal context.

We demonstrate that this advantage of causal attention makes it possible to use short synthesis windows without mutual overlap in time. The discontinuities in the temporal waveform synthesized by concatenating non-overlapping iSTFTs can be minimized, and thus, the algorithmic latency can be reduced to the size of the non-overlapping short synthesis window. We also attempt to reduce the high computational complexity of

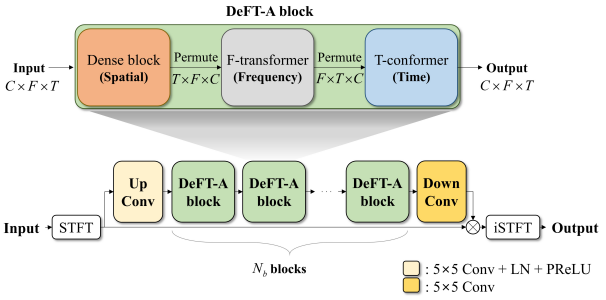

Figure 2: *Overall structure of DeFT-AN*

the transformer-based model by utilizing grouped convolution [18, 19] and lightweight attention. These modifications allow us to reduce the computational complexity to approximately one-seventh of the vanilla DeFT-AN.

The rest of the paper is organized as follows. Section 2 introduces the new architecture and explains modifications from DeFT-AN made for real-time processing. In section 3, we present speech enhancement results in noisy reverberant environments and a comparison to conventional time-domain and STFT-domain speech enhancement models. We demonstrate that the proposed model achieves better performance than the baseline real-time models and lower algorithmic latency and RTF than conventional STFT-based models.

2. Real-time architecture design

2.1. Proposed model

The proposed model is based on the Dense Frequency-Time Attentive Network (DeFT-AN) [9] proposed for multichannel speech enhancement in various reverberant environments, which utilizes the complex spectral masking of noisy speech in the STFT domain. DeFT-AN has achieved high performance by using sub-blocks that sequentially aggregate spatial, spectral, and temporal information. The model's masking network (Figure 2) comprises an up-convolution (Up-Conv) layer for increasing the size of the channel dimension, serial DeFT-A blocks for capturing information along different dimensions, and a down-convolution for reconstructing real and imaginary components of the target clean speech in the STFT domain. Each DeFT-A block includes a dense block, F-transformer, and T-conformer for aggregating information in spatial, spectral, and temporal information, respectively. Despite the remarkable performance of DeFT-AN in highly reverberant and unseen environments, its high computational complexity and algorithmic latency are the major obstacles to implementing a real-time speech enhancement model.

To reduce the computational complexity and memory usage of DeFT-AN, we propose the following modifications. First, we incorporate grouped convolution in dense blocks to reduce the number of parameters and computational complexity of the original 2D convolution. Second, we utilize lightweight attention in F-transformer and T-conformer, which was originally proposed for the computer vision task [20]. The lightweight attention decreases the length of key and value features of the attention layer by a factor of k using a strided convolution with equal kernel size and stride (k). This reduces the size of the attention map by a factor of k. In this study, we implement lightweight attention by using a strided 1D convolution before applying linear projection to extract the key and value features. Figure 3 illustrates the example of lightweight attention implemented for the attention layer of the F-transformer. Lastly, unlike the vanilla DeFT-AN, we remove the layer normalizations after 1×1 convolution in the feedforward layer. Figure 4 presents the schematic of the proposed DeFT-A block, which consists of a dense block, F-transformer, and T-conformer.

2.2. Non-overlapping synthesis window

The next challenge is to reduce algorithmic latency originating from the window size of the STFT-based model. As described in Section 1, using a synthesis window as small as the hop size can reduce algorithmic latency. Nevertheless, frames generated without consideration of continuity with neighboring frames can cause abrupt changes at both ends.

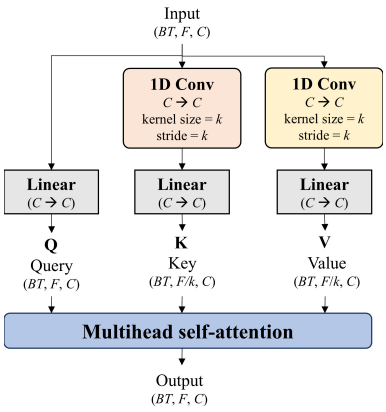

Figure 3: *Lightweight attention layer of F-transformer for low computational complexity*

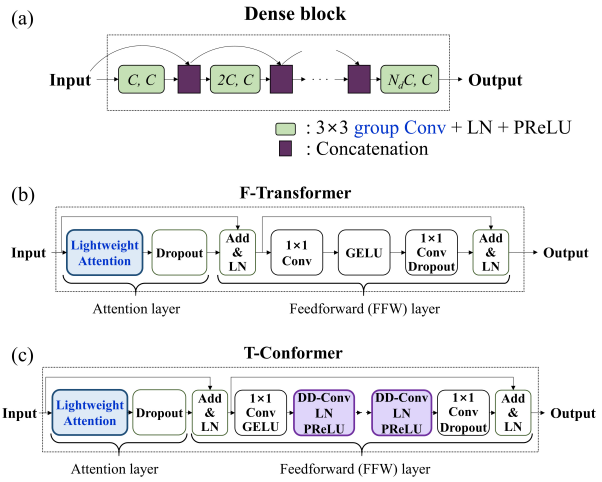

Figure 4: *Modified DeFT-A block for low computational complexity. (a) Dense block with grouped convolution, (b) F-transformer and (c) T-conformer using lightweight attention*

In this study, we investigate the ability of the transformer-based model in handling temporal context to mitigate the discontinuity problem. The T-conformer structure in DeFT-AN RT utilizes causal lightweight attention, for which the attention map is structured as an upper triangular matrix. The size of the attention map is $(T \times T)$, where T is equal to the number of time frames, which allows the model to synthesize STFT data considering all the present and past samples. Unlike convolutional neural networks (CNNs) that only analyze the temporal relationship within the receptive field of convolution kernels, transformers designed for STFT data can exploit the entire temporal relationship in a single layer. Therefore, the proposed model has the potential to use non-overlapping synthesis windows with a little quality degradation. In Section 3, it is demonstrated that DeFT-AN RT can provide a high signal-to-distortion ratio and minimize discontinuities even when non-overlapping synthesis windows are used to minimize algorithmic latency.

3. Experiments

3.1. Experimental setup

The experiments were conducted with two datasets: spatialized WSJCAM0 dataset and spatialized DNS challenge dataset. Both datasets were simulated from four microphones arranged on a circle of a 10-cm radius with equal inter-element angles. The spatialized WSJCAM0 dataset was constructed by spatializing speeches from the WSJCAM0 corpus [21] using the simulated room impulse responses (RIRs) and noises from the channels 0, 2, 4, and 6 of the REVERB challenge dataset [22]. The reverberation times (RT60) of RIRs were within the [0.2, 1.3] seconds range, and signal-to-noise ratios (SNRs) of noisy speech were in the range of [5, 25] dB. The details on creating this dataset can be found in [10]. The spatialized DNS challenge dataset consists of spatialized speech and noise from the DNS challenge 2020 corpus [23], with SNRs and RT60s ranging between [-10, 10] dB and [0.2, 1.2] s, respectively. The algorithm for creating the spatialized DNS challenge dataset is described in [17]. The RIRs were simulated using *Pyroomacoustics* [24], which uses the image source method. Different rooms were used for training, validation, and testing, so the tests were done for the rooms unseen during the training. All utterances were resampled at 16 kHz, and 4-s long utterances were randomly selected for training.

The final model of DeFT-AN RT has the following parameter settings: 4 DeFT-A blocks, 4 dense blocks, 4 groups of dense blocks, 3 sequential dilated convolutions in a T-conformer, and a channel dimension length of 64 in the Up-Conv. The complex spectra of the four microphone signals were extracted using STFT with a rectangular analysis window of 32 ms length and 50% overlap. The RI components of complex spectra with 257 frequency bins were utilized. The model was trained for 70 epochs using ADAM optimizer [25] with an initial learning rate of 0.0004 and a learning rate scheduler reducing the learning rate by half if the validation loss does not decrease after 3 epochs. The proposed network was trained by the phase-constrained magnitude (PCM) loss [26], which is defined as the sum of the real and imaginary (RI) magnitude losses for speech and noise. The analysis window size, hop size, and synthesis window size were set to 32 ms, 16 ms, and 16 ms, respectively. Both the hop size and synthesis window size were configured as 16 ms, meaning no overlap between the synthesis windows.

3.2. Experimental results

We conducted the parameter study with four variants of the proposed model, as well as with the vanilla DeFT-AN, to validate the effectiveness of modifications introduced in the proposed model. The parameter study was conducted using only the spatialized WSJCAM0 dataset. The first two results of Table

Table 1: *Comparison of DeFT-AN variants for parameter study*

Parameter	SI-SDR	PESQ	STOI	Model size	MAC/s	latency
Vanilla DeFT-AN (non-causal)	**15.7**	**3.63**	**98.1**	2.7 M	95.6 G	32 ms
DeFT-AN RT (proposed)	12.1	3.42	96.6	**1.15 M**	**13.4 G**	16 ms
Single window size	12.6	3.43	97.0	1.15 M	13.4 G	32 ms
DeFT-AN RT (hop 8 ms)	12.8	3.48	96.9	1.15 M	26.8 G	**8 ms**
Without grouped convolution	12.7	3.45	97.2	2.25 M	31.6 G	16 ms
Vanilla attention	12.9	3.47	97.0	0.82 M	17.7 G	16 ms

Table 2: *Performance comparison with baseline real-time multichannel speech enhancement models*

	WSJCAM0			DNS challenge			Model size	MAC/s	RTF	latency
	SI-SDR	**PESQ**	**STOI**	**SI-SDR**	**PESQ**	**STOI**				
MC SkiM	4.8	2.10	83.4	2.5	1.50	62.7	5.3 M	**1.9 G**	**0.32**	**1.25 ms**
EaBNet	7.3	2.97	93.2	3.5	2.24	80.5	2.8 M	7.4 G	0.77	20 ms
causal DRCNet (non-real-time)	10.0	3.34	95.7	5.9	2.56	86.3	1.4 M	7.5 G	2.02	32 ms
DeFT-AN RT	**12.1**	**3.42**	**96.6**	**6.8**	**2.77**	**89.4**	**1.2 M**	13.4 G	0.48	16 ms

1 present the speech enhancement performance of the vanilla DeFT-AN and the proposed real-time model (DeFT-AN RT), evaluated in terms of SI-SDR, PESQ, and STOI. The real-time model exhibits more than a 3 dB reduction in SI-SDR compared to the non-real-time model, but it is natural to have such performance degradation with the real-time model at the expense of causal convolution, shorter algorithmic latency (32 ms $\rightarrow$ 16 ms), low computational complexity (95.6G $\rightarrow$ 13.4G in MAC/s), and small model size (2.7M $\rightarrow$ 1.15M).

The first variant (single window size) shown in the next row of Table 1 is identical to DeFT-AN RT except for the same window size and hop size (50% overlap) used for the synthesis window as in the analysis window. As shown in Table 1, the performance difference between the single- and dual-window models was not substantial, with an SI-SDR difference of around 0.5 dB and almost the same PESQ and STOI. This result indicates that DeFT-AN RT exhibiting only half algorithmic latency can generate clean speech well from non-overlapping synthesis windows without the need for overlapping between time frames.

The next variant still uses the dual window size approach but its hop size was reduced to 8 ms to secure shorter algorithmic latency. The performance is slightly higher or comparable to the single window size and proposed models, but its computational complexity is doubled. Thus, when enough computing power is available, the dual-window approach combined with the DeFT-AN can reduce algorithmic latency to 8 ms.

We also investigated the effectiveness of grouped convolution and lightweight attention in reducing computational complexity. In the corresponding two variants (without grouped convolution, original attention), we used the conventional 2D convolution instead of grouped convolution in dense blocks and full linear projection of the original DeFT-AN in the F-transformer and T-conformer, respectively. From the results presented in Table 1, we can see that grouped convolution and lightweight attention contribute to a small change in speech enhancement performances. However, the grouped convolution significantly reduces the computational complexity (from 31.6 $\rightarrow$ 13.4 GMAC/s). Using lightweight attention slightly increases the model size, while computational complexity is reduced by 24.3%, and GPU memory usage is reduced by 32.3% by using a small attention map compared to vanilla attention.

Finally, the proposed model was compared to baseline methods. The first method is multichannel (MC) SkiM [14], which is the time domain approach originally proposed for speech separation but can also be used for multichannel speech enhancement. The second model, EaBNet [12], is an STFT-based real-time speech enhancement model with a small STFT window size. The last model is the causal DRCNet [27]. DRCNet is a complex spectral mapping model based on U-Net with a dense stack of BLSTM and 2D convolution. We included DRCNet as one of the baselines because the single-stage model of DRCNet is the second best-performing model after DeFT-AN on the spatialized WSJCAM0 dataset. For a fair comparison with real-time models, we used the causal version of DRCNet. However, note that the causal DRCNet has RTF higher than 1.0 (2.02) and is not a real-time processing model. The performance comparison with baseline models was conducted using both datasets, and the results are presented in Table 2. RTF was evaluated on AMD Ryzen 7 3700X CPU clocked at 3.60GHz. The performance of MC SkiM was the lowest among baseline models, indicating that its short encoder kernel size cannot cover the long reverberation time of datasets dealt with in this study. EaBNet utilizes reduced window and hop size to meet real-time requirements and shows enhanced performance across all evaluation measures. Causal DRCNet is unsuitable as a real-time model due to its algorithmic latency of 32 ms and high RTF, but its performance was significantly better than other real-time models. The proposed DeFT-AN has lower algorithmic latency and RTF than those of EaBNet, but it outperforms all baseline models including causal DRCNet. The only downside of the proposed model is its high computational complexity (13.4 GMAC/s), but its RTF is the smallest among the STFT-based models because transformers in DeFT-AN use convolution layers supporting parallel processing.

4. Conclusion

We presented a real-time multichannel speech enhancement model, DeFT-AN RT, which inherits the high enhancement performance and parallel processing ability of DeFT-AN for real-time processing. The proposed model utilizes grouped convolution and lightweight attention to reduce computational complexity and adopts a dual-window size approach such that target speech waveforms can be synthesized without overlap of iSTFT frames. The lightweight attention of the T-conformer is capable of analyzing all temporal relations up to the present, which suppresses possible discontinuities at frame boundaries. This non-overlapping synthesis window shortened the algorithmic latency to a level suitable for real-time processing. Through the training and testing over reverberant and noisy datasets, the proposed model demonstrated its remarkable speech enhancement performance surpassing all baseline real-time speech enhancement models and the smallest real-time factor among STFT-based real-time models.

5. Acknowledgement

This work was supported by the BK21 Four program through the National Research Foundation (NRF) funded by the Ministry of Education of Korea, the National Research Council of Science and Technology (NST) granted by the Korean government (MSIT)(No. CRC21011), and the Center for Applied Research in Artificial Intelligence (CARAI) funded by DAPA and ADD (UD190031RD).

6. References

[1] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler *et al.*, "5th deep noise suppression challenge at IEEE ICASSP," Nov. 2022, [Online]. Available:https://www.microsoft.com/en-us/research/academic-program/deep-noise-suppression-challenge-icassp-2023/.

[2] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst. for Cognitive Science, CA, USA, Tech. Rep., 1985.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[4] K. Cho, B. V. Merrinboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1724–1734.

[5] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May. 2019, pp. 6875–6879.

[6] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020, pp. 3291–3295.

[7] X. Ren, X. Zhang, L. Chen, X. Zheng, C. Zhang, L. Guo *et al.*, "A causal U-net based neural beamforming network for real-time multi-channel speech enhancement," in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug. 2021, pp. 1832–1836.

[8] C. Xue, W. Huang, W. Chen, and J. Feng, "Real-time multi-channel speech enhancement based on neural network masking with attention model," in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, Aug. 2021, pp. 1862–1866.

[9] D. Lee and J.-W. Choi, "DeFT-AN: Dense frequency-time attentive network for multichannel speech enhancement," *IEEE Signal Processing Letters*, vol. 30, pp. 155–159, Mar. 2023.

[10] Z.-Q. Wang and D. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May. 2020, pp. 486–490.

[11] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, "STFT-domain neural speech enhancement with very low algorithmic latency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 397–410, Nov. 2022.

[12] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May. 2022, pp. 6487–6491.

[13] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, April. 2018, pp. 696–700.

[14] C. Li, L. Yang, W. Wang, and Y. Qian, "SkiM: Skipping memory LSTM for low-latency real-time continuous speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May. 2022, pp. 681–685.

[15] D. Mauler and R. Martin, "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement," in *Proc. 15th European Signal Processing Conference*, Poznan, Poland, Sep. 2007, pp. 222–226.

[16] S. U. Wood and J. Rouat, "Unsupervised low latency speech enhancement with RT-GCC-NMF," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 332–346, Apr. 2019.

[17] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "TPARN: Triple-path attentive recurrent network for time-domain multichannel speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May. 2022, pp. 6497–6501.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2017.

[19] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1492–1500.

[20] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang *et al.*, "Cmt: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 12 175–12 185.

[21] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAMO: a british english speech corpus for large vocabulary continuous speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Detroit, MI, USA, May. 1995, pp. 81–84.

[22] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, pp. 1–19, Jan. 2016.

[23] C. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020, pp. 2492–2496.

[24] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 351–355.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, May. 2015.

[26] A. Pandey and D.-L. Wang, "Dense CNN with self-attention for time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1270–1279, Mar. 2021.

[27] J. Liu and X. Zhang, "DRC-NET: Densely connected recurrent convolutional neural network for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May. 2022, pp. 166–170.