Improving Frame-Online Neural Speech Enhancement With Overlapped-Frame Prediction

Zhong-Qiu Wang, *Graduate Student Member, IEEE*, and Shinji Watanabe, *Senior Member, IEEE*

***Abstract*—Frame-online speech enhancement systems in the short-time Fourier transform (STFT) domain usually have an algorithmic latency equal to the window size due to the use of overlap-add in the inverse STFT (iSTFT). This algorithmic latency allows the enhancement models to leverage future contextual information up to a length equal to the window size. However, this information is only partially leveraged by current frame-online systems. To fully exploit it, we propose an overlapped-frame prediction technique for deep learning based frame-online speech enhancement, where at each frame our deep neural network (DNN) predicts the current and *several past frames* that are necessary for overlap-add, instead of only predicting the current frame. In addition, we propose a loss function to account for the scale difference between predicted and oracle target signals. Experiments on a noisy-reverberant speech enhancement task show the effectiveness of the proposed algorithms.**

***Index Terms*—Deep learning, online speech enhancement.**

I. Introduction

DEEP learning has elevated the performance of speech enhancement in the past decade [1]. Since the very first success of deep learning in offline enhancement [2], there have been growing interests in using DNNs for low-latency enhancement, as many applications require online real-time processing. For example, the recent deep noise suppression challenges [3] target at enhancement in a monaural teleconferencing setup, requiring a processing latency less than 40 ms on a specified processor. Similar latency requirements exist in other related challenges [4], [5]. The recent Clarity challenge [6] aims at multi-microphone enhancement in a hearing aid setup, requiring an algorithmic latency of at maximum 5 ms.

Numerous deep learning based approaches [7]–[31] have been proposed for frame-online speech enhancement by using frame-online (or causal) DNN modules such as uni-directional recurrent networks, causal normalization layers, causal convolutions and causal attention layers. To the best of our knowledge, almost all the current DNN models adopt a *single-frame prediction* strategy for online enhancement. That is, the DNN models predict a frame of target speech at the current frame based on the current and past frames, and the prediction at the current frame is overlap-added with the predictions at nearby frames for signal re-synthesis. This leads to an algorithmic latency equal to the window length. However, this strategy cannot fully leverage the future context afforded by the algorithmic latency. To explain this problem, we use an example STFT-based system shown in Fig. 1 for illustration, where the window size is 32 ms and hop size 8 ms. The frame-online DNN model operates in an online streaming fashion, processing one frame at a time when a new frame arrives and producing a predicted frame of signal for each input frame. The predicted signals at nearby frames are then overlapped and added together (see the red-frame rectangle) to get the final output signal for the sub-frame at frame t (marked in the top of the figure). As we can see, to get the prediction at sample index n (marked in the top), the system has to first fully observe the input signal at frame $t+3$ before the DNN can perform feed-forwarding. The algorithmic latency is hence equal to the 32 ms window length. In this system, to get the predicted signal at frame t, the model only takes in the input signals up to frame t. The insight of this letter is that we can actually use input signals up to frame $t+3$ for the DNN to get the predicted signal at frame t as well as at frame $t+1$, $t+2$ and $t+3$. The resulting predicted signals would very likely be better, because the DNN can leverage up to three frames of future context, and at the same time the algorithmic latency remains the same as the window size. This can be achieved by training our DNN model for *overlapped-frame prediction*, where the current and past frames necessary for overlap-add are predicted at each frame at once.

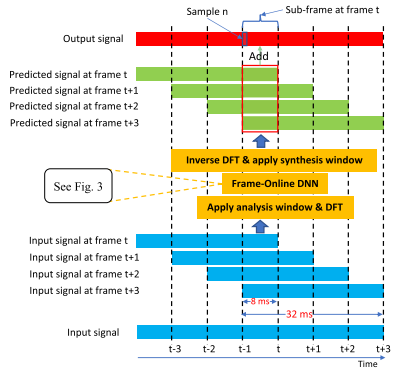

Fig. 1. Single-frame prediction. Best viewed in color.

Besides overlapped-frame prediction, we also propose a novel mechanism that equalizes the scales of predicted and oracle target signals before loss computation.

II. Proposed Algorithms

Let us denote the monaural mixture $y = s + v$ in the time domain as a summation of the target signal s and non-target signal

Manuscript received March 14, 2022; revised June 7, 2022; accepted June 13, 2022. Date of publication June 15, 2022; date of current version July 1, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rosangela Coelho. *(Corresponding author: Zhong-Qiu Wang.)*

The authors are with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: wang.zhongqiu41@gmail.com; shinjiw@ieee.org).

Digital Object Identifier 10.1109/LSP.2022.3183473

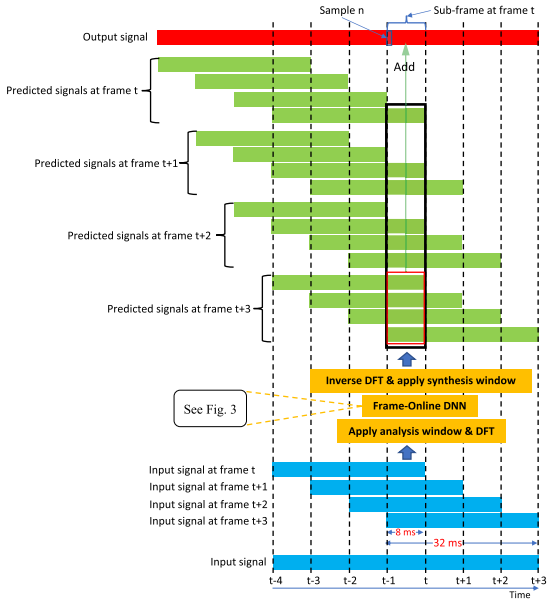

Fig. 2. Overlapped-frame prediction. Best viewed in color.

v. In the STFT domain, we denote the mixture as $Y = S + V$, where Y, S and V are the STFT spectra of y, s and v, respectively. Our DNNs predict S based on Y, if operating in the STFT domain; and predict s based on y, if in the time domain. This section describes the proposed overlapped-frame prediction, along with synthesis window design and DNN configurations, and its extensions to time-domain models.

A. Overlapped-Frame Prediction

Let us denote the the STFT window and hop sizes by W and H samples, assuming W is a multiple of H, and let $C = W/H$ be the number of overlapped frames in each window. The proposed system is illustrated in Fig. 2, which uses 32 ms window and 8 ms hop sizes as an example. Our DNN operates in a frame-online fashion and, at each frame, it predicts C frames consisting of the current and $C-1$ immediate past frames. To get the sub-frame output at frame t (marked in the top of the figure), we overlap the C frames of predicted signals produced at frame $t+C-1$, and add together the sub-frames marked by the red-frame rectangle (denoted as "partial sub-frame summation"). Clearly, the algorithmic latency is still equal to the window size, but we can leverage input signals up to frame $t+C-1$ to better predict each of the C frames that are overlap-added to get the sub-frame output at frame t. This algorithm requires our frame-online DNN model to have C outputs. Alternatively, we can overlap-add the sub-frames marked by the black-frame rectangle (denoted as "full sub-frame summation"). This alternative could lead to better performance as it summates more sub-frame predictions.

We point out that our approach is different from studies [14], [16] that look ahead extra frames and add extra latency to the window size. It is also different from studies [32] that predict a symmetric window of frames of time-frequency masks at each frame, where the motivation was about output ensembling.

B. Synthesis Window Design

The proposed partial and full sub-frame summation methods require proper synthesis windows that can achieve perfect reconstruction when used with an analysis window $g \in \mathbb{R}^W$.

For the partial sub-frame summation, following [33] we can use a regular synthesis window $l \in \mathbb{R}^W$ defined as:

$$l[n] = \frac{g[n]}{\sum_{e=0}^{W/H-1} g[eH + (n \bmod H)]^2}, \tag{1}$$

where $0 \le n < W$. By following the overlap-add procedure, we can verify that perfect reconstruction is satisfied, because

$$\sum_{e=0}^{W/H-1} g[eH+h] \times l[eH+h] = 1, \tag{2}$$

for any $0 \le h < H$.

For the full sub-frame summation, we design the synthesis window as follows:

$$l[n] = \frac{g[n]}{\sum_{e=0}^{W/H-1} \Big((e+1) \times g[eH + (n \bmod H)]^2\Big)}, \tag{3}$$

where the difference from (1) is that we have a weighting term for the W/H sub-frames in the denominator. An intuitive way to understand this is that among all the sub-frames marked by the black-frame rectangle in Fig. 2, four are the last sub-frame in various predicted frames, three are the second-last sub-frame, and so on. We therefore need the weighting term in the denominator to have a pair of analysis and synthesis windows that can achieve perfect reconstruction, which can be ensured, because, in the proposed overlap-add procedure, the condition below always holds for any $0 \le h < H$:

$$\sum_{e=0}^{W/H-1} g[eH+h] \times ((e+1) \times l[eH+h]) = 1. \tag{4}$$

C. DNN Configurations

Our STFT-domain DNN model is trained for complex spectral mapping [18], [34]–[37], where the real and imaginary (RI) components of the mixture Y are concatenated as input features for a DNN to predict the RI components of target speech S. The DNN architecture will be described in Section III-B.

When trained for overlapped-frame prediction, the DNN predicts K ($= C$) frames at each frame, essentially producing C estimated target spectrograms; and when trained for single-frame prediction, the DNN produces one estimated target spectrogram (i.e. $K = 1$). Let us denote the DNN-estimated RI components by $\hat{R}^{(k)}$ and $\hat{I}^{(k)}$, where $k \in \{1, 2, \ldots, K\}$ indexes the K spectrograms, and the enhanced speech by $\hat{S}^{(k)} = \hat{R}^{(k)} + j\hat{I}^{(k)}$, where j is the imaginary unit. We can define the loss function on the RI components and magnitudes of the estimated spectrograms:

$$\mathcal{L}_{\text{RI+Mag}} = \sum_k \Big(\|\hat{R}^{(k)} - \text{Real}(S)\|_1 + \|\hat{I}^{(k)} - \text{Imag}(S)\|_1 + \Big\| |\hat{S}^{(k)}| - |S| \Big\|_1 \Big), \tag{5}$$

where $\text{Real}(\cdot)$ and $\text{Imag}(\cdot)$ extract RI components, $|\cdot|$ computes magnitude, and $\|\cdot\|_1$ calculates the L_1 norm. Based on the described overlap-add mechanisms, an iSTFT is applied to resynthesize the estimated time-domain signal $\hat{s} = \text{iSTFT}(\hat{S}^{(:)})$, where $\hat{S}^{(:)}$ includes all the K estimated spectrograms. Following [38], we can train through the iSTFT and define the loss on

the re-synthesized signal and its magnitude:

$$\mathcal{L}_{\text{Wav+Mag}} = \|\hat{s} - s\|_1 + \Big\| |\text{STFT}(\hat{s})| - |\text{STFT}(s)| \Big\|_1, \quad (6)$$

where $\text{STFT}(\cdot)$ extracts a spectrogram from a signal[1]. When using this loss function with overlapped-frame prediction, at each frame we essentially use signals up to frame $t + C - 1$ to predict the sub-frame at frame t (marked in the top of Fig. 2).

When using mapping based approaches with the loss function in (6), the model needs to predict a signal that has the same gain as the target. Although this may not be a problem for offline processing, as the model can observe the entire input mixture to produce a reasonable gain, for frame-online processing this could be difficult. We propose to first compute a real-valued gain-equalization (GEQ) factor to balance the gain of the predicted signal with that of the target signal before loss computation:

$$\mathcal{L}_{\text{Wav+Mag,geq}} = \|\hat{\alpha}\hat{s} - s\|_1 + \Big\| |\text{STFT}(\hat{\alpha}\hat{s})| - |\text{STFT}(s)| \Big\|_1, \quad (7)$$

where $\hat{\alpha} = \text{argmin}_\alpha \|\alpha\hat{s} - s\|_2^2 = (\hat{s}^\mathsf{T} s)/(\hat{s}^\mathsf{T}\hat{s})$. Note that GEQ can facilitate model training by not restricting the same gain, but $\hat{\alpha}$ will not be one at run time. This would be fine, as automatic gain adjustment, a common module in modern speech communication, can adjust the gain to a desired level.

We always include in each loss function a magnitude loss, which can improve metrics that favor estimated signals with a good magnitude [38].

D. Extension to Time-Domain Models

Besides STFT-domain models, the proposed algorithms can also be used with time-domain models such as Conv-TasNet [13], by replacing the yellow blocks in Figs. 1 and 2 with time-domain models that use overlap-add for signal re-synthesis.

III. Experimental Setup

We validate our algorithms on a simulated monaural speech enhancement task. This section describes the simulated dataset, DNN architectures, and miscellaneous configurations.

A. Dataset

The WSJCAM0 corpus [39] contains 7,861, 742, and 1,088 clean speech signals in its training, validation, and test sets, respectively. Using the split of the clean signals in WSJCAM0, we simulate 39,245 (∼77.7 h), 2,965 (∼5.6 h) and 3,260 (∼8.5 h) noisy-reverberant mixtures as our training, validation and test sets, respectively. We sample the clips in the development set of FSD50k [40] to simulate the noises for training and validation, and those in the evaluation set for testing. For each simulated mixture, we sample up to 7 noise clips. We consider clips longer than 10 seconds as background noises and as foreground noises otherwise. Each simulated mixed noise file has one background noise and the rest are foreground noises. The energy level between the dry background noise and each dry foreground noise is sampled from the range $[-3, 9]$ dB. The directions of each noise source and the target speaker to the microphone are independently sampled from the range $[0, 2\pi)$. We treat each sampled clip as a point source, convolve each source with its corresponding room impulse response, and summate the convolved signals to create the mixture. After adding up all the spatialized noises, we scale the summated reverberant noise such that the signal-to-noise ratio between the target direct-path speech and the summated reverberant noise is equal to a value sampled from the range $[-8, 3]$ dB. The distance from each source to the microphone is drawn from the range $[0.75, 2.5]$ m. The reverberation time is drawn from the range $[0.2, 1.0]$ s. The sampling rate is 16 kHz.

[1]We tried to weight the two terms in the loss function, but did not observe gains, likely because they have similar scales due to the Parseval's theorem.

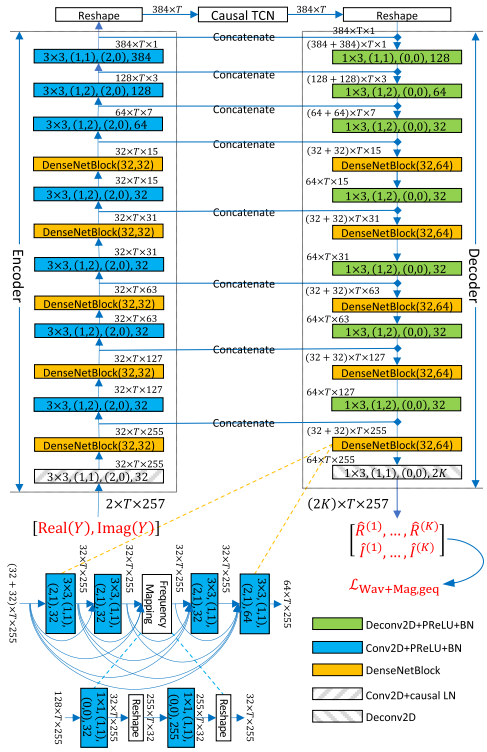

Fig. 3. Frame-online TCN-DenseUNet architecture adapted from [41]. During training, the tensor shape after each block is denoted in the format: *featureMaps* × *timeSteps* × *freqChannels*. Each one of Conv2D+PReLU+BN, Deconv2D+PReLU+BN, Conv2D, and Deconv2D blocks is shown in the format: *kernelTime* × *kernelFreq*, *(strideTime, strideFreq)*, *(padTime, padFreq)*, *featureMaps*. Each DenseNetBlock(g_1, g_2) has five Conv2D+PReLU+BN blocks with growth rate g_1 for the first four layers and g_2 for the last one.

B. DNN Architectures

Our STFT-domain frame-online DNN architecture is illustrated in Fig. 3. It is a modified version of the offline TCN-DenseUNet architecture, which has shown strong performance in our previous studies [36], [41]–[43] on speech separation. The major modifications include using causal layer normalization (LN) layers, batch normalization (BN) layers, and causal one- and two-dimensional convolution and deconvolution. The network is a causal temporal convolution network (TCN) sandwiched by a U-Net, where the encoder performs down-sampling along frequency and the decoder performs up-sampling along frequency to recover the original frequency resolution. DenseNet blocks, which concatenate the outputs of the previous convolutions as the input to the current convolution, are inserted at multiple frequency scales in the U-Net. The corresponding offline architecture is obtained by replacing causal LN and convolutions with non-causal ones. We choose this architecture as it shares many similarities with the architectures

TABLE I
RESULTS OF USING 32/8 MS WINDOW/HOP SIZES FOR TCN-DENSEUNET

ID	Systems	Frame-online?	Loss function	Sub-frame summation	SI-SDR	PESQ	eSTOI
0	Unprocessed	-	-	-	−6.2	1.44	0.411
1a	Single-frame pred.	yes	RI+Mag	-	3.4	2.11	0.708
1b	Overlapped-frame pred.	yes	RI+Mag	Partial	3.6	2.18	0.725
1c	Overlapped-frame pred.	yes	RI+Mag	Full	3.6	2.16	0.722
2a	Single-frame pred.	yes	Wav+Mag	-	3.5	2.17	0.719
2b	Overlapped-frame pred.	yes	Wav+Mag	Partial	4.0	2.23	0.733
2c	Overlapped-frame pred.	yes	Wav+Mag	Full	4.0	2.26	0.738
2d	Overlapped-frame pred.	yes	Wav+Mag,geq	Full	4.2	2.28	0.743
3a	Single-frame pred.	no	RI+Mag	-	4.8	2.44	0.776
3b	Single-frame pred.	no	Wav+Mag	-	4.6	2.46	0.776
3c	Single-frame pred.	no	Wav+Mag,geq	-	4.8	2.47	0.778

TABLE II
RESULTS OF USING 20/10 MS WINDOW/HOP SIZES FOR TCN-DENSEUNET

ID	Systems	Frame-online?	Loss function	Sub-frame summation	SI-SDR	PESQ	eSTOI
1a	Single-frame pred.	yes	RI+Mag	-	2.8	2.01	0.692
1b	Overlapped-frame pred.	yes	RI+Mag	Partial	3.0	2.05	0.699
1c	Overlapped-frame pred.	yes	RI+Mag	Full	3.0	2.05	0.698
2a	Single-frame pred.	yes	Wav+Mag	-	2.3	2.06	0.702
2b	Overlapped-frame pred.	yes	Wav+Mag	Partial	2.4	2.06	0.704
2c	Overlapped-frame pred.	yes	Wav+Mag	Full	3.0	2.16	0.721
2d	Overlapped-frame pred.	yes	Wav+Mag,geq	Full	3.4	2.16	0.724
3a	Single-frame pred.	no	RI+Mag	-	4.5	2.46	0.770
3b	Single-frame pred.	no	Wav+Mag	-	4.0	2.47	0.778
3c	Single-frame pred.	no	Wav+Mag,geq	-	4.5	2.49	0.784

TABLE III
RESULTS OF USING 4/2 MS WINDOW/HOP SIZES FOR CONV-TASNET

ID	Systems	Frame-online?	Loss function	Sub-frame summation	SI-SDR	PESQ	eSTOI
2a	Single-frame pred.	yes	Wav+Mag	-	2.2	1.79	0.658
2b	Overlapped-frame pred.	yes	Wav+Mag	Partial	2.1	1.77	0.652
2c	Overlapped-frame pred.	yes	Wav+Mag	Full	2.3	1.81	0.661
2d	Overlapped-frame pred.	yes	Wav+Mag,geq	Full	2.3	1.81	0.664
3b	Single-frame pred.	no	Wav+Mag	-	3.8	2.28	0.754
3c	Single-frame pred.	no	Wav+Mag,geq	-	3.9	2.26	0.754

in recent studies [17]–[20]. The RI components of different input (and output) signals are concatenated as feature maps in the network input (and output). The number of feature maps is 2 for the input tensor and is 2 K for the output tensor. Linear activation is used in the output layer to get the predicted RI components. Based on this architecture, overlapped-frame prediction uses only slightly more computation than single-frame prediction, since the increased computation, which only stems from the final 2D deconvolution and overlap-add, is negligible compared to the computation of the DNN.

We choose Conv-TasNet [13] as the time-domain model for overlapped-frame prediction, considering its popularity. Its default hyper-parameters reported in [13] are used.

C. Miscellaneous Configurations

The direct-path signal of the target speaker is used as the target for model training and as the reference for metric computation. SI-SDR in decibel (dB) [44], PESQ [45], and eSTOI [46] are used as the evaluation metrics. For TCN-DenseUNet, we experiment with the commonly used 32/8 and 20/10 ms window/hop sizes. Giving a 16 kHz sampling rate, we use a 512-point discrete Fourier transform to obtain 257-dimensional STFT features at each frame for both setups. This way, the same DNN architecture can be used. The square-root Hann window is used as the analysis window. For Conv-TasNet, we experiment with 4/2 ms window/hop sizes. It should be noted that a longer window looks ahead more future samples and has a higher algorithmic latency, and would likely have more improvement when used with overlapped-frame prediction.

IV. EVALUATION RESULTS

Table I, II and III respectively report the results of using 32/8, 20/10 and 4/2 ms window/hop sizes. In each table, we provide the offline results obtained by using the offline versions of the DNN architectures. These offline results can be viewed as the performance upper bound of the frame-online models.

Let us first look at Table I. Comparing 1a and 1b (or 2a and 2b), we observe that overlapped-frame prediction with partial sub-frame summation produces better performance than single-frame prediction. Using full sub-frame summation, 1c obtains worse performance than 1b, while 2c obtains similar or slightly better performance than 2b. This is likely because when we use the RI+Mag loss and do not train through iSTFT, the sub-frames summated by using full sub-frame summation contain sub-frame predictions produced at earlier frames (see the black-frame rectangle in Fig. 2), which are not as good as the ones produced at the current frame. When we train through iSTFT and use the Wav+Mag loss, the model may figure out how to best predict each sub-frame and best summate all the sub-frames (in the black-frame rectangle) to optimize the loss. Overall, our proposed systems in 2c show noticeable improvement over 1a and 2a, by better leveraging the 32 ms future context information.

In Table II, similar trend as in Table I is observed. The relative gains of overlapped-frame prediction over single-frame prediction are smaller than those in Table I. This is as expected as there is less future context (i.e., 20 ms) to exploit.

In Table III, the gains are very marginal as the future context (i.e., 4 ms) is even less. This Conv-TasNet experiment suggests that the window size needs to be reasonably large for overlapped-frame prediction to work well.

Although the gains brought by overlapped-frame prediction depend on the allowed future context and are small when the allowed context is limited, we always observe consistent and steady improvements in the considered setups.

In all the tables, the proposed gain equalization leads to slightly better results in the online cases (see 2c vs. 2 d).

V. CONCLUSION

We have proposed a novel overlapped-frame prediction technique for frame-online speech enhancement. It better leverages future context, without incurring extra algorithmic latency. The increased amount of computation only stems from the output layer and is negligible compared to that of the DNN backbone. The proposed technique can be easily modified for, or directly adopted by, numerous frame-online systems. It would likely yield better performance due to the better exploitation of the future context afforded by the algorithmic latency, no matter whether they are DNN- or non-DNN-based, whether they operate in the T-F domain or in the time domain, or whether they deal with speech enhancement or other related separation tasks, as long as they use overlap-add for signal re-synthesis.

REFERENCES

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[2] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[3] C. K. Reddy *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework," in *Proc. Interspeech*, 2020, pp. 2492–2496.

[4] R. Cutler *et al.*, "ICASSP 2022 acoustic echo cancellation challenge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 9107–9111.

[5] W. Rao *et al.*, "INTERSPEECH 2021 conferencingspeech challenge: Towards far-field multi-channel speech enhancement for video conferencing," 2021, *arXiv:2104.00960*.

[6] "Clarity challenge: Machine learning challenges for hearing devices," [Online]. Available: http://claritychallenge.org/

[7] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoust. Soc. Amer.*, vol. 141, pp. 4705–4714, 2017.

[8] G. Wichern and A. Lukin, "Low-latency approximation of bidirectional recurrent networks for speech denoising," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 66–70.

[9] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 696–700.

[10] K. Wilson *et al.*, "Exploring tradeoffs in models for low-latency speech enhancement," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 366–370.

[11] S. Wisdom *et al.*, "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 900–904.

[12] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 531–535.

[13] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[14] T. Yoshioka, Z. Chen, C. Liu, X. Xiao, H. Erdogan, and D. Dimitriadis, "Low-latency speaker-independent continuous speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6980–6984.

[15] S. Chakrabarty and E. A. P. Habets, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 787–799, Aug. 2019.

[16] S. Sonning, C. Schuldt, H. Erdogan, and S. Wisdom, "Performance study of a convolutional time-domain audio separation network for real-time speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 831–835.

[17] Y. Liu and D. Wang, "Causal deep CASA for monaural talker-independent speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1270–1279, 2020.

[18] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.

[19] Y. Hu *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[20] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6629–6633.

[21] C. Han, Y. Luo, and N. Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6404–6408.

[22] Y. Xia, S. Braun, C. K. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 871–875.

[23] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, 2020, pp. 3291–3295.

[24] S. Braun, H. Gamper, C. K. A. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 656–660.

[25] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6633–6637.

[26] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "ICASSP 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6628–6632.

[27] Z. Tu, J. Zhang, N. Ma, J. Barker, and C. Science, "A two-stage end-to-end system for speech-in-noise hearing aid processing," in *Proc. Clarity*, 2021, pp. 3–5.

[28] S. J. Yang, S. Wisdom, C. Gnegy, R. F. Lyon, and S. Savla, "Listening with Googlears : Low-latency neural multiframe beamforming and equalization for hearing aids," in *Proc. Clarity*, 2021.

[29] K. Zmolikova and J. H. Cernock, "BUT system for the first clarity enhancement challenge," in *Proc. Clarity*, 2021, pp. 1–3.

[30] X. Ren *et al.*, "A causal U-Net based neural beamforming network for real-time multi-channel speech enhancement," in *Proc. Interspeech*, 2021, pp. 1832–1836.

[31] C. Li, L. Yang, W. Wang, and Y. Qian, "SkiM: Skipping memory LSTM for low-latency real-time continuous speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 681–685.

[32] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, no. 5, pp. 2604–2612, 2016.

[33] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Audio, Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[34] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[35] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[36] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[37] Z.-Q. Wang and D. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 941–950, 2020.

[38] Z.-Q. Wang, G. Wichern, and J. Le Roux, "On the compensation between magnitude and phase in speech separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 2018–2022, 2021.

[39] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1995, vol. 1, pp. 81–84.

[40] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50 K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang. Process*., vol. 30, pp. 829–852, 2022.

[41] Z.-Q. Wang, G. Wichern, and J. Le Roux, "Leveraging low-distortion target estimates for improved speech enhancement," 2021, *arXiv:2110.00570*.

[42] Z.-Q. Wang, G. Wichern, and J. Le 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.

[43] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2001–2014, 2021.

[44] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - Half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[45] 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. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 749–752.

[46] 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, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.