INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Low-Delay Speech Enhancement Using Perceptually Motivated Target and Loss

Xu Zhang, Xinlei Ren, Xiguang Zheng, Lianwu Chen, Chen Zhang, Liang Guo, Bing Yu

Kuaishou Technology Co., Beijing, P.R. China
zhangxu07@kuaishou.com

Abstract

Speech enhancement approaches based on deep neural network have outperformed the traditional signal processing methods. This paper presents a low-delay speech enhancement method that employs a new perceptually motivated training target and loss function. The proposed approach can achieve similar speech enhancement performance compared to the state-of-the-art approaches, but with significantly less latency and computational complexities. Judged by the MOS tests conducted by the INTERSPEECH 2021 Deep Noise Suppression Challenge organizer, the proposed method is ranked the 2nd place for Background Noise MOS, and the 6th place for overall MOS.

Index Terms: Speech enhancement, time-frequency masking, deep neural network, single-channel

1. Introduction

Single channel speech enhancement is designed to separate the clean speech from the noisy speech mixtures. Traditional signal processing based methods aim to model the noise spectrum to perform spectrum subtraction [1] or the Wiener filtering [2]. In recent years, methods based on deep neural networks (DNNs) have outperformed the traditional approaches. These methods are usually trained in a supervised fashion and can be categorized into the time domain and the time-frequency domain approaches. The time domain methods proposed in [3], [4] directly work with the noisy speech waveforms to produce the estimated clean speech. While the time domain methods can achieve end-to-end processing, the trade-off is forfeiting the sparsity between the speech and the noise signals in the time-frequency domain, as described in [5]. The time-frequency domain methods proposed in [6], [7] and [8], [9] employ DNN to model the magnitude and complex spectrum of the clean speech, respectively. While setting the training target to the complex spectrum can achieve higher oracle upper bound than the magnitude spectrum, increased complexity is introduced, which may not be suitable for practical real-time applications. In addition, in last year's INTERSPEECH Deep Noise Suppression Challenge [10], methods using both of these training targets can achieve comparative perceptual quality [11]. In this work, we focus on low-delay low-complexity single-channel speech enhancement in the time-frequency domain with the perceptually optimal magnitude spectrum as the training target.

This paper proposed a preprocessing method to create a perceptually optimal magnitude spectrum as the training target. When the magnitude spectrum of the clean speech is combined with the phase of the input noisy signal, the upper bound of the oracle target speech signal is degraded especially for low signal to noise ratio (SNR) conditions. This is caused by the fact that for the low SNR time-frequency instants, the noisy phase is significantly different from the ideal speech phase. Before applying the proposed preprocessing method, the upper bound of the training target created by the commonly used ideal masks are studied. While similar study can be found in [12] where the Signal-to-Artifact ratio (SAR) of the oracle train targets created by different ideal mask are compared, this paper directly compares the PESQ [13] of the oracle training targets obtained by the classic Wiener mask[14], ideal ratio mask (IRM) [15] and the ideal amplitude mask (IAM) [12]. It has been found that the IAM can achieve the highest PESQ score under different SNR conditions. The IAM is further compressed using the proposed preprocessing method. The proposed preprocessing scheme can achieve on average of 0.11 PESQ improvement for the achievable training targets with SNR ranged from -10dB to 25dB. Specifically, for 5dB to 15dB conditions, the PESQ improvement is above 0.15.

In addition to the preprocessing scheme, a new loss function is proposed to incorporate the compressed IAM with the loss calculation from the magnitude spectrum. The proposed loss function introduces an IAM weighting factor to equalize the importance of the logarithm compressed magnitudes. The aim is to provide better balance between the time-frequency instants with low and high amplitudes.

The proposed method is evaluated using PESQ [13] and STOI [16]. As shown in Section 6, the proposed preprocessing scheme can achieve 0.04 PESQ and 0.11 STOI improvement compare to the identical configuration without using the preprocessing. The proposed IAM weighting factor in the loss function can achieve 0.13 PESQ and 0.02 STOI improvement. When using the preprocessing and loss function together, the proposed method can achieve similar PESQ performance compared with state-of-the-art methods with less system latency and complexity.

The proposed method is also entered the INTERSPEECH 2021 Deep Noise Suppression Challenge [17]. With 480-point FFT (corresponding to 30ms for 16kHz input signal) and 160-point stride (corresponding to 10ms for 16kHz input signal), the total system delay is 30ms + 10ms = 40ms, which satisfies the latency requirement of this challenge. The proposed model has 3.393M parameters, 756.868M FLOPs. The one-frame inference time on Intel Core i7 (2.6GHz) CPU is 0.386ms. The proposed method is ranked the 2nd place for Background Noise MOS, and the 6th place for overall MOS.

2. Signal Model

A noisy-reverberant mixture signal y is modeled by the following formula at the time domain:

$$y(t) = h(t) * x(t) + n(t) \tag{1}$$

where x(t) is the speech signal, h(t) is the transfer function from the talker to the microphone, * denotes the convolution, and n(t) is the noise. The purpose of the system is to estimate the x(t) from the y(t), including removing noise and reverberation from the captured signal.

http://dx.doi.org/10.21437/Interspeech.2021-1410

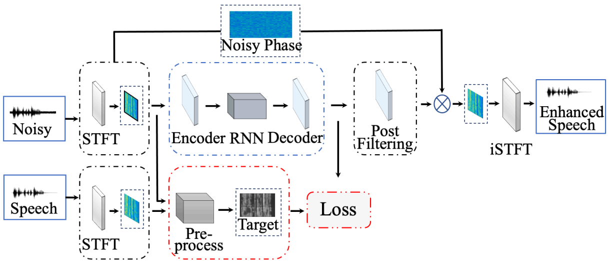


Figure 1: *Architecture of the proposed system*

The overview of the training process is shown in Figure 1. As shown, the short-time Fourier transform (STFT) is performed on the noisy speech and the clean speech to obtain the time-frequency domain magnitude and phase signals. The magnitude of the oracle clean speech is preprocessed to obtain the perceptually optimal oracle training target. The magnitude spectrum of the input noisy signal is fed into the encoder-decoder based DNN to estimate the magnitude of the clean speech. Signal processing based post-filtering is also employed to produce the final magnitude estimation, which will be combined with the noisy phase to form the time domain output using the Inverse short-time Fourier transform (iSTFT).

3. Training Target

While perfect reconstruction can be achieved by employing a complex clean speech spectrogram as the training target, approaches only predicting the magnitude spectrogram of the clean speech are widely adopted for real-time communication systems to significantly reduce the complexity [18]. In this work, we directly compare the PESQ scores of different training targets under a wide signal to noise ratio (SNR) range. In addition, a preprocessing method is proposed to further improve the perceptual quality of the oracle train targets.

The oracle training targets produced by three widely used masks are firstly derived:

$$x_{oracle_{irm}} = iSTFT\left(|Y|\cdot\frac{|X|}{|X|+|N|}\cdot\angle(Y)\right) \qquad (2)$$

$$x_{oracle_{Wiener}} = iSTFT\left(|Y|\cdot\frac{X^2}{X^2+N^2}\cdot\angle(Y)\right) \qquad (3)$$

$$x_{oracle_{iam}} = iSTFT\left(|Y| * \frac{|X|}{|Y|} * \angle(Y)\right) \qquad (4)$$

where $|\cdot|$ and $\angle(\cdot)$ denote magnitude and phase calculation from the complex time-frequency spectrum.

2000 oracle training targets are produced for each of the SNR (ranged from -10 dB to 25dB) and ideal mask condition using the VCTK speech dataset [19] and steady noise including while, pink, babble, street, etc. The PESQ results using the original clean speech signal as the reference is shown in Figure 2 where the error bar indicates 95% confidence interval. As shown, the IAM (green) achieved the highest PESQ scores compared to the IRM (blue) and the Wiener-like mask (grey).

In addition, a mask compression scheme is proposed to compress the IAM from the standard setting to significantly improve the perceptual quality of the IAM oracle train target:

$$x_{oracle_{iam-\gamma}} = iSTFT\left(|Y| * \left(\frac{|X|}{|Y|}\right)^{\gamma} * \angle(Y)\right) \qquad (5)$$

where γ represents the compression ratio and $\gamma = 1$ is the standard setting for the IAM. Experiment results for different choice of γ from 0.2 to 1.4 are shown in Figure 3.

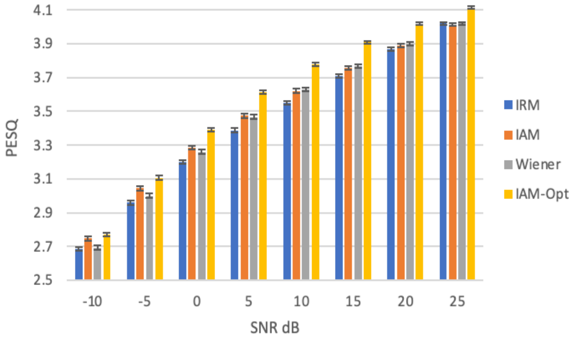


Figure 2: *Comparison among different oracle training targets*

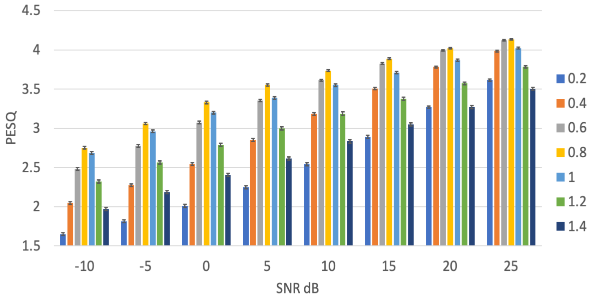


Figure 3: *Comparison among different compression ratios (γ)*

As shown in Figure 3, 2000 oracle training targets using (5) for each of the γ option are produced to compare with the clean reference using PESQ. The error bar indicates 95% confidence interval. $\gamma = 0.8$ achieved 0.116 (average) and more than 0.15 (for 5dB, 10dB, 15dB and 20dB SNR conditions) PESQ improvement compared to the original ($\gamma = 1.0$) IAM oracle training targets.

4. DNN Model

As shown in Figure 1, the DNN model in this work is based on the convolution recurrent network (CRN) [20] that nested a Recurrent Neural Network (RNN) module inside the Convolutional Neural Network (CNN) based encoder-decoder structure. After taking a 480-point (30ms for 16kHz input) STFT with 160-point stride (10ms for 16kHz input), the STFT domain magnitude representation $Y \in \mathbb{R}^{F\times T}$ (a context window of $F = 241$ STFT magnitude coefficients with $T = 2$ frames) of the input audio signal is fed to $L_c \in \mathbb{N}$ ($L_c = 2$) convolutional encoder to model the time-frequency domain local interconnections. The feature maps of the last convolutional encoder layer [batch, frame, frequency, channel] are reshaped over the frequency axis to form the sequential representations to the $L_r \in \mathbb{N}$ ($L_r = 2$) recurrent layers [batch, frame, frequency × channel]. The output of the recurrent layer is reshaped back to the 3-dimensional tensor to form the input to the $L_d \in \mathbb{N}$ ($L_d = 2$) convolutional decoder. The final output tensor is produced using the output of the last convolutional decoder layer by a feed-forward layer with the sigmoid activation function.

In detail, for each of the convolutional layer, rectified linear unit (ReLU) activation function and batch normalization is employed. Dropout mechanism is also employed at the end of the convolutional layer to further prevent overfitting. The dropout parameter is set to 0.3. After L_c convolutional layers,

Table 1: *Design of the convolutional recurrent neural network*

Layer	Encoder (Decoder)			RNN	FC
	Channels	Size	Stride	Units	Units
Conv2D	90	[1,9]	[1,3]		
Conv2D	90	[2, 3]	[1,2]		
GRU				256	
GRU				256	
Conv2D Transpose	8	[1,5]	[1,2]		
Conv2D Transpose	1	[1, 3]	[1,1]		
Dense					241

the output is a tensor $\mathcal{L} \in \mathbb{R}^{M \times F' \times T'}$, where M is the number of output feature maps of the last encoder layer, F' and $T'$are the remaining number of frequency and time indexes. The recurrent network is implemented by the Gate Recurrent Unit (GRU) to model the sequential nature of the input signal. After stacking the output feature maps $\mathcal{L}$ over the frequency axis, the input to the GRU is a tensor $\mathcal{H} \in \mathbb{R}^{(M \cdot F') \times T'}$. The output of recurrent layer is reshaped back to $\Gamma \in \mathbb{R}^{M \times F'' \times T''}$ to form the input to the convolutional decoder. Details of the hyperparameters is presented in Table 1.

A new IAM-weighted loss is designed to better suppress the noise for the low-SNR and magnitude time-frequency regions. For each time-frequency instant, the loss can be calculated by:

$$Loss = W_{IAM} * |ln(X'_{mag} + 1) - ln(X_{mag} + 1)| \quad (6)$$

where

$$W_{IAM} = e^{(\frac{a}{b+IAM})} \quad (7)$$

$$IAM = Y_{mag} * \left(\frac{X_{mag}}{Y_{mag}}\right)^{\gamma} * \angle(Y) \quad (8)$$

In this work, γ is set to 1, a is set to 2 and b is set to 1. The proposed IAM weight is aimed to balance the importance of the speech-dominant and noise-dominant time-frequency instances, where the IAMs with smaller value can have higher weights in the total loss. The logarithm compression part intents to reduce the level difference between the low magnitude and high magnitude time-frequency instants. As presented in the evaluation section, improved noise suppression performance can be achieved using the proposed loss. The X'_{mag} is the predicted clean amplitude spectrum, the X_{mag} is the target amplitude spectrum, Y_{mag} is the noisy amplitude spectrum. $\angle(\cdot)$ denotes magnitude and phase calculation from the complex time-frequency spectrum.

5. Postfiltering

To further improve the subjective quality, envelope postfiltering [18] is employed to refined the IAM estimated by the DNN (IAM_{dnn}) to produce the final IAM (IAM_{pf}):

$$IAM_{pf} = \frac{(1+\tau) \cdot IAM_{dnn}}{(1+\frac{\tau \cdot IAM_{dnn}{}^2}{IAM_{sin}{}^2})} \quad (9)$$

where

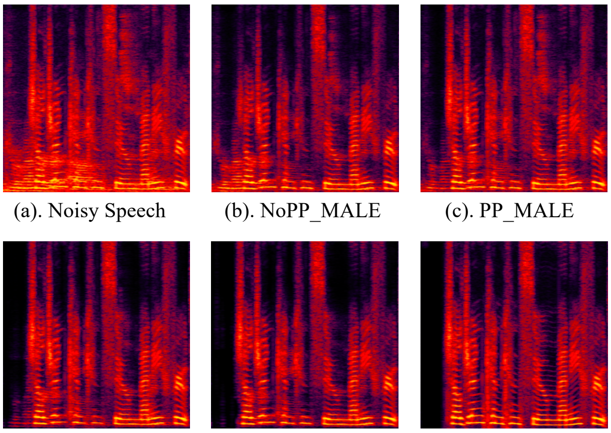
$$IAM_{sin} = IAM_{dnn} \cdot \sin(\frac{\pi \cdot IAM_{dnn}}{2}) \quad (10)$$

Table 2: *PESQ and STOI results on the INTERSPEECH DNS challenge synthesis test set*

Data hour	Algorithm	PESQ_NB	STOI
500h	NoPP_MALE	3.03	94.83
	PP_MALE	3.07	94.94
	NoPP_WO-MALE	3.16	94.85
	PP_WO-MALE	3.18	94.96
2000h	PP_WO-MALE	3.25	95.36

In this work, τ is set to 0.02. The final estimated magnitude is

$$X'_{pf} = Y \cdot IAM_{pf} \quad (11)$$


(a). Noisy Speech (b). NoPP_MALE (c). PP_MALE

(d). NoPP_WO-MALE (e). PP_WO-MALE (f). Clean Speech

Figure. 4: *Results comparison of the spectrum*

6. Experiments and Results

6.1. Datasets

The clean speech dataset of the experiments is the DNS Challenge (INTERSPEECH 2021) clean speech dataset which contains multiple languages, such as English, Chinese, French, etc. The noise data is the DNS Challenge (INTERSPEECH 2021) noise set. We synthesize the training data with SNR that is randomly chosen from {-5, 0, 5, 10, 15, 20, 25, 30} dB. Silence, clean speech and pure noise samples are also included. To further improve the robustness in reverberated environments, the noisy and target signals are convolved with measured or simulated room impulse responses. Varies EQ filters are also applied to simulate the frequency response curve of various microphones. To avoid speech distortion introduced by dereverberation, speech with 75ms early reflection is used as training target. The sampling rate of the audio signal is 16kHz.

6.2. Results

First, to evaluate the proposed preprocessing method and loss function, models with different configurations are compared as following.

- **NoPP_MALE**: The model without preprocessing and with Mean Absolute Logarithmic Error (**MALE**) loss function.
- **PP_MALE**: The model with preprocessing and **MALE** loss function.

Table 3: *PESQ and STOI results on the DNS challenge (INTERSPEECH 2020) synthesis test dataset*

Algorithm	Para(M)	Latency(ms)	PESQ_NB	STOI
Noisy	---	---	2.45	91.52
NSNet [10]	5.1M	40	2.87	94.47
DTLN [21]	1.0M	40	3.04	94.76
DCCRN[22]	3.7M	62.5	3.27	---
FullSubNet[23]	5.6M	64	3.31	96.11
Proposed	3.4M	40	3.25	95.36

- **NoPP_WO-MALE**: The model without preprocessing and with the IAM Weight Optimized MALE (**WO-MALE**) loss function.
- **PP_WO-MALE**: The model preprocessing and **WO-MALE** loss function.

All of the models are trained with 500 hours of data, and the PP_WO-MALE model with 2000h data also be trained to form the model submitted to the DNS Challenge (INTERSPEECH 2021). The Short-Term Objective Intelligibility (STOI) and the perceptual evaluation of speech quality (PESQ) are used as evaluation metric. Table 2 presents the results on the synthesis test set of the DNS Challenge (INTERSPEECH 2020).

From Table 2, it can be observed that the model with preprocessing can achieve higher PESQ and STOI scores than the model without preprocessing. It should be noted that the PESQ improvement is less than the oracle target speech as discussed in Section 3. This is mainly caused by two reasons. First, the network may not be above to fully exploit the difference between the oracle target with and without preprocessing. Second, the SNR distribution of the synthesis test set is not identical to the experiment conducted in Section 3. It can also be observed that the WO-MALE model can achieve higher PESQ and STOI scores than the MALE conditions. Finally, the proposed PP_WO-MALE model achieves the highest PESQ and STOI scores. The PP_WO-MALE trained with 2000h dataset further improves the PESQ and STOI scores.

An illustrative example is also presented in Figure 4 where the clean speech is in the presence of a background competing speaker at the beginning of the sentence and also with constant background noise. As shown, the proposed PP_WO-MALE Condition can achieve suppress more steady background noise while remove the competing speaker.

This 2000h PP_WO-MALE model is compared with the state-of-the-art methods including the top-ranked methods in last year's INTERSPEECH DNS Challenge using the same test set. NSNet [10] is the official baseline method of the INTERSPEECH DNS 2020 Challenge. DTLN [21] is a low-complexity model with single-frame-in, single-frame-out manner and DCCRN [22] is the top-ranked method in the subjective listening test of the INTERSPEECH DNS 2020 Challenge. FullSubNet [23] is the top-ranked method of the ICASSP DNS 2021 Challenge.

As listed in Table 3, the proposed method can achieve more than 0.20 PESQ improvement compared to the methods with the identical (40ms) latency. And comparing the DCCRN and the FullSubNet which both have about 50% higher latency and larger model size, the proposed low-delay and low-complexity method still achieves competitive PESQ performance.

Table 4: *DNS challenge (INTERSPEECH2021) background noise MOS*

Team	Sta	Emo	NET	NE	Mus	Eng	Ovr
Ours	4.68	4.56	4.52	4.53	4.21	4.54	4.51
NSNet2	4.21	3.61	4.23	4.03	3.11	3.94	3.88
Noisy	2.86	1.93	2.91	3.11	2.14	2.3	2.6

Table 5 : *DNS challenge (INTERSPEECH2021) overall MOS*

Team	Sta	Emo	NET	NE	Mus	Eng	Ovr
Ours	3.62	3.30	3.40	3.59	3.12	3.27	3.41
NSNet2	3.28	2.75	3.31	3.25	2.78	2.93	3.07
Noisy	3.03	2.28	3.00	3.04	2.57	2.52	2.77

The DNS challenge (INTERSPEECH2021) organizers has provided the official subjective evaluation results listed in Table 4 and Table 5. It summarizes the final P.808 subjective evaluation results for real-time deep noise suppression track. The abbreviations of Table 4 (Background Noise MOS) and V (Overall MOS) are listed as follows: Stationary (Sta), Emotional Speech (Emo), Non-English-tonal (NET), Non-English (NE), Musial Instruments (Mus), English (Eng), Overall (Ovr). While significant outperformed the baseline system, the proposed method ranked the 2nd place in the Background Noise MOS result, and the 6th place in the Overall MOS result (the rank of Microsoft is ignored for the contest).

7. Complexity

With 480-point FFT (corresponding to 30ms for 16kHz input signal) and 160-point stride (corresponding to 10ms for 16kHz input signal), the total system delay is 30ms + 10ms = 40ms, which satisfies the latency requirement of this challenge. The proposed model has 3.393M parameters, 756.868M FLOPs. The one-frame inference time on Intel Core i7 (2.6GHz) CPU is 0.386ms.

8. Conclusions

In this study, a low-delay low-complexity speech enhancement system is proposed. We demonstrated that the model with preprocessing and newly designed WO_MALE loss function can achieve the best performance. We also compared the performance with some state-of-the-art methods in the DNS2020, the proposed system achieved similar PESQ performance with less system latency and complexity. In the DNS Challenges (INTERSPEECH 2021) results provided by the challenge organizer, the proposed system got the 2nd place in the Background Noise MOS result, and 6th place in the Overall MOS result (the rank of Microsoft is ignored for the contest).

9. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. Wouters, "The Potential for Speech Intelligibility Improvement Using the Ideal Binary Mask and the Ideal Wiener Filter in Single Channel Noise Reduction Systems: Application to Auditory Prostheses," *IEEE*

Trans. Audio Speech Lang. Process., vol. 21, no. 1, pp. 63–72, Jan.

[3] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation,” *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019

[4] R. Giri, U. Isik, and A. Krishnaswamy, “Attention Wave-U-Net for Speech Enhancement,” in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2019, pp. 249–253

[5] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004

[6] Y. Wang, A. Narayanan, and D. Wang, “On Training Targets for Supervised Speech Separation,” *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014

[7] T. Grzywalski and S. Drgas, “Using Recurrences in Time and Frequency within U-net Architecture for Speech Enhancement,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6970–6974.

[8] D. S. Williamson and D. Wang, “Time-Frequency Masking in the Complex Domain for Speech Dereverberation and Denoising,” *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017,

[9] D. Yin, C. Luo, Z. Xiong, and W. Zeng, “PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network,” presented at the AAAI, Nov. 2019.

[10] C. K. A. Reddy *et al.*, “The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results,” *ArXiv200513981 Cs Eess*, Oct. 2020, Accessed: Apr. 03, 2021.

[11] “Deep Noise Suppression Challenge - INTERSPEECH 2020 - Microsoft Research.” https://www.microsoft.com/en-us/research/academic-program/deep-noise-suppression-challenge-interspeech-2020/#!results (accessed Mar. 26, 2021).

[12] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 708–712

[13] ITU, “P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.” 2001.

[14] M. Parviainen, P. Pertilä, T. Virtanen, and P. Grosche, “Time-Frequency Masking Strategies for Single-Channel Low-Latency Speech Enhancement Using Neural Networks,” in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2018, pp. 51–55

[15] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7092–7096

[16] 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

[17] C. K. A. Reddy *et al.*, *Interspeech 2021 Deep Noise Suppression Challenge*. 2021.

[18] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, “A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech,” in *Proc. Interspeech 2020*, 2020, pp. 2482–2486

[19] J. Yamagishi, C. Veaux, and K. MacDonald, “CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92),” Nov. 2019,

[20] K. Tan and D. Wang, “A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement,” in *Proc. Interspeech 2018*, 2018, pp. 3229–3233

[21] N. L. Westhausen and B. T. Meyer, “Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression,” in *Proc. Interspeech 2020*, 2020

[22] Y. Hu *et al.*, “DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement,” in *Proc. Interspeech 2020*, 2020

[23] X. Hao, X. Su, R. Horaud, and X. Li, “FullSubNet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement,” in *Proc. ICASSP 2021*, 2021