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

DBNet: A Dual-branch Network Architecture Processing on Spectrum and Waveform for Single-channel Speech Enhancement

Kanghao Zhang, Shulin He, Hao Li, Xueliang Zhang

College of Computer Science, Inner Mongolia University, China
{imu.koer, heshulin, lihao}@mail.imu.edu.cn, cszxl@imu.edu.cn

Abstract

In real acoustic environment, speech enhancement is an arduous task to improve the quality and intelligibility of speech interfered by background noise and reverberation. Over the past years, deep learning has shown great potential on speech enhancement. In this paper, we propose a novel real-time framework called DBNet which is a dual-branch structure with alternate interconnection. Each branch incorporates an encoder-decoder architecture with skip connections. The two branches are responsible for spectrum and waveform modeling, respectively. A bridge layer is adopted to exchange information between the two branches. Systematic evaluation and comparison show that the proposed system substantially outperforms related algorithms under very challenging environments. And in INTERSPEECH 2021 Deep Noise Suppression (DNS) challenge, the proposed system ranks the top 8 in real-time track 1 in terms of the Mean Opinion Score (MOS) of the ITU-T P.835 framework.

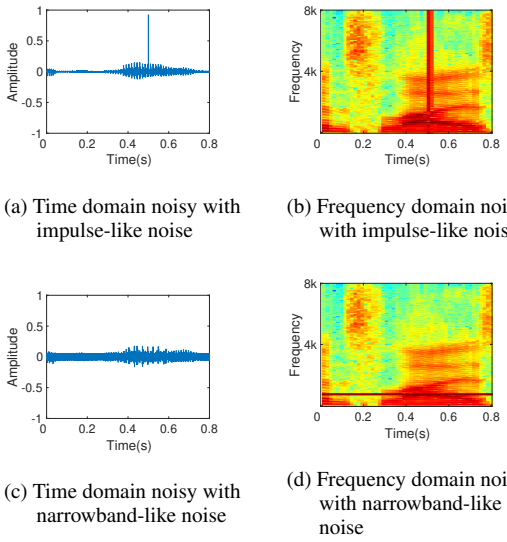
Index Terms: Speech enhancement, time domain, frequency domain, real-time

1. Introduction

With the COVID-19 pandemic, lots people are working online, and the demand for reliable real-time speech enhancement algorithms has increased sharply. During these times, we need to ensure clear call quality and effective communication and cooperation with others without delay. Our communication are usually disturbed by a lot of background noise, such as washing machines, passing trucks, and the strong reverberation. These will affect the efficiency of our work and communication. Recently, many researchers from academia and industry have made significant contributions to monaural speech enhancement. However, due to the diversity of noise types in reality, even the the state-of-the-art algorithms cannot handle challenging environments well.

With the development of deep learning, many research regard speech enhancement as a supervised learning problem [1] [2] [3], and obtained excellent performance. Usually the neural network input is time domain signal or STFT domain signal. In [2] [4], the authors study the noise reduction problem in the short-time Fourier transform (STFT) domain. In [1] [5], the time-domain signal after framing is directly feed to the neural network. Both frequency domain and time domain has its own advantages. The STFT method is more in line with human hearing perception, and the characteristics of speech is more explicit. Time domain method does not damage the signal by STFT, avoids the well-known invalid STFT problem.

Lim et al. [6] introduced a time-frequency network to jointly optimize the time and frequency domains of a signal for the task of audio super resolution. They show that combine these two domains can boost the audio super resolution performance and obtain the state-of-the-art in both quantitatively and qualitatively.



(a) Time domain noisy with impulse-like noise

(b) Frequency domain noisy with impulse-like noise

(c) Time domain noisy with narrowband-like noise

(d) Frequency domain noisy with narrowband-like noise

Figure 1: *Time- and frequency-domain noisy signal with two kinds of noises.*

As we know, for an impulse-like noise as shown in Fig. 1(a) and 1(b), it is easy to eliminate in the time domain. Only a few samples need to be removed. In the frequency domain, the impulse-like nose pollutes the entire frequency band, and it is difficult to be eliminated by a speech enhancement method based on the frequency domain. In contrast, for a narrowband-like noise as shown in Figure 1(c) and 1(d), the noise is distributed on the narrowband and the frequency domain-based method covers well. In the time-domain, the noise and the speech are coupled together, and it is hard to decouple by an enhancement method based on time-domain. In this paper, in order to reduce noise better, we proposed a novel speech enhancement algorithm called DBNet, which combines the time domain and frequency domain together. The DBNet is a dual-branch structure with alternate interconnection. Each branch incorporates an encoder decoder architecture with skip connections. The two branches are responsible for spectrum and waveform modeling, respectively. A bridge layer is adopted to exchange information between the two branches. Experiments show that the proposed method has excellent results on the WSJ0 SI-84 [7] and DNS Challenge [8].

The rest of the paper is organized as follows. The structure of the proposed DBNet is described in Section 2. Section 3 describes the experimental setup. The experimental results are revealed in Section 4. We conclude this paper in Section 5.

http://dx.doi.org/10.21437/Interspeech.2021-1042

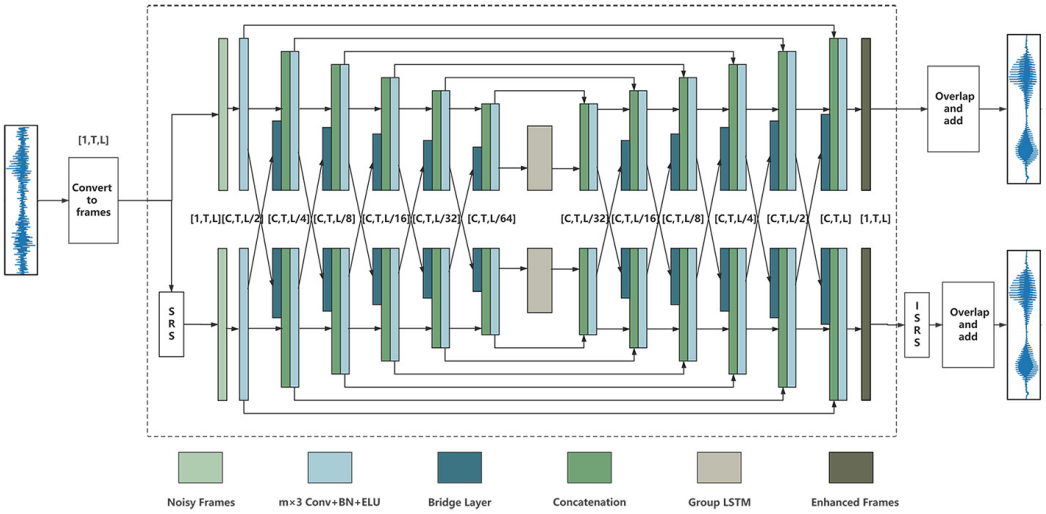


Figure 2: *Diagram of the proposed DBNet model*

2. Model Description

The overall structure of the proposed method is shown in Fig. 2. In the following sections, three important parts: SRS module, GLSTM, bridge layer, will be introduced one by one.

2.1. Shift Real Spectral (SRS) Module

Recently, Liu et al. [9] proposed a new separation target based on the time-frequency representation of SRS in their work, which proved the superiority of SRS over STFT. First, SRS takes the phase into consideration, which improves the speech intelligibility and quality. Second, SRS is a spectral representation method in real field instead of the complex field, all of the elements of input are real numbers. So it reduces the modeling difficulty and provide convenience for the information interaction module of our model. Therefore in this paper, SRS is adopted as our frequency domain input based on the two advantages above.

2.2. Gated Convolution and Group LSTM

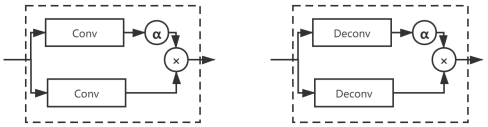
Dauphin et al. [10] improved the masked convolution for image convolution modeling and proposed gated convolution (GCNN) which is described as:

$$\begin{aligned} y &= (x * W_1 + b_1) \odot \sigma(x * W_2 + b_2) \\ &= v_1 \odot \sigma(v_2) \end{aligned} \tag{1}$$

where W and b represent kernel and bias, respectively. $*$ and $\odot$ denote operation of convolution and element-wise multiplication, respectively. σ represents a nonlinear activation function. The GCNN can reduce the vanishing gradient problem for deep architectures by providing a linear path for the gradients, so we replaced the convolution in the original crn with it. A diagram of gated convolution is shown in Fig. 3.

Model efficiency is pretty important, and many application scenarios have higher requirements for processing speed and memory usage. However, due to the introduction of the dual-branch structure, the amount of calculation and memory occupied by the model are much higher than that of the standard convolutional recurrent network. Gao et al. [11] proposed a grouped recursive neural network (RNN) strategy, which reduces the complexity of the model while ensuring the performance of the model. The process of a group RNN is shown in Fig. 4.

In this paper, the group LSTM contains two layers of RNN, and each layer has two LSTMs to learn the features within each group. Between the two layers, a frame-level rearrangement is used to establish the inter-group relationship of features, which guarantees the utilization of inter-group correlations to a certain extent.



(a) Gated Convolution (b) Gated Deconvolution

Figure 3: *Gated Convolution and Gated Deconvolution, where α denotes a sigmoid funtion*

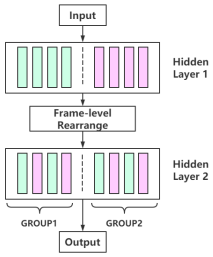


Figure 4: *The process of group LSTM*

2.3. Bridge Layer

Bridge Layer is a linear unit, which is responsible for converting information from one branch to another. Actually, the bridge

layer consists of two independent vectors with the same length as the frame length. The two vectors are respectively responsible for the information conversion from the time domain to the frequency domain and its inverse process. We take the real part of the fast Fourier transform (FFT) variable as the initialization parameter of these trainable vectors to fit the situation that SRS is used as the frequency domain representation.

2.4. Parameters Setting

The encoder has six layers, each consist of a bridge layer, a gated convolution followed by batch normalization and ELU nolinearity. Note that a module is added before the input of frequency branch to calculate the frequency-domain representation of the real number field. The input size of the model is [batch_size, 1, seq_len, features]. The first layer of the encoder expands the input channel from 1 to 64. After that,the rest layer of the encoder do the following operations. The bridge layer first translates the features from the other branch and then concatenate it with the output of the previous layer along the channel axis, finally pass the features to a gated convolution. The kernel and the stride of convolution are set to (1, 3) and (1, 2), respectively. The final output size of the encoder is [batch_size, 64, seq_len, features / 64]. The output is fed into group LSTM layers to extract the long-term relationship of the features in time domain and frequency domain, respectively.

Similarly, the decoder also has six layers. In addition to the features from another domain, each layer also gets the skip connection of the corresponding layer of the encoder, and concatenate them with the output of the previous layer along the channel axis. The decoder uses gated deconvolution to double the feature dimensions layer by layer to reconstruct the signal to the original size. The last layer of decoder enhance the signal to one channel. Finally, the output is converted into speech through the overlap-and-add operation. Note that the speech from frequency branch is regarded as final result.

2.5. Loss Function

In the early experiments, we used a loss based on STFT magnitude which was proposed in [1] and can be described as:

$$L_{Mag}(s,\hat{s}) = \frac{1}{T\cdot F}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1}[(|S_r(t,f)|+|S_i(t,f)|) - (|\hat{S}_r(t,f)|+|\hat{S}_i(t,f)|)] \quad (2)$$

where T and F represent the number of time frames and frequency dimensions, S and $\hat{S}$ denote STFTs of s and $\hat{s}$, respectively. S_r and $\hat{S}_i$ represent the real and imaginary parts of S, respectively. Note that the output of the network contains two enhanced utterances, one of which is from the time branch and the other from the frequency branch and they are optimized independently. So the total loss is defined as:

$$L = L_{Mag}(s,\hat{s}_t) + L_{Mag}(s,\hat{s}_f) \quad (3)$$

However, we found that magnitude loss introduced a large number of unknown artifacts. Although it does not affect the objective evaluation score, it would bring terrible auditory perception. Therefore, in DNS Challenge, the magnitude loss is replaced with the phase constrained magnitude loss proposed in [12] and achieved good subjective evaluation scores in the competition.

3. Experiment

3.1. Datasets

In this study, we evaluated the performance of our proposed model on the WSJ0 SI-84 dataset [7] which includes 7138 utterances from 83 speakers (42 males and 41 females). We used the utterances of 77 speakers for training and the rest for test. We used 10000 non-speech sounds from a sound effect library (available at `www.sound-ideas.com`) [13] and generated 320000 and 3000 utterances at the SNRs uniformly sampled from {-5dB, -4dB, -3dB, -2dB, -1dB, -0dB} for training and validation, respectively. For the test set, two noises (babble and cafeteria) from Auditec CD (available at `http://www.auditec.com`) are used to generate 300 mixtures at each SNR of -5dB, 0dB, and 5dB.

3.2. Baselines

In this study, we compared the proposed dual-branch network with another 3 baselines, namely CRN [14], GCRN [2] and AECNN [1], which are given as follows:

- CRN: it is a casual convolutional recurrent network in T-F domain. The network uses 5 convolution layers as the encoder and 5 deconvolution layers as the decoder. Two LSTM layers are used for sequence modeling. This network receiving magnitude as input. The number of channels is decreased and the number of parameters is 4.5M.
- GCRN: it is a causal gated convolutional recurrent network for complex spectral mapping. The structure is similar to CRN except that GCRN has two decoders to model real and imaginary, respectively. The input of network is complex spectral. We kept the best configuration in [2] and the number of parameters is 9.76M.
- AECNN: it is a autoencoder based fully convolutional neural network in the time domain. Raw waveform is chunked into frames with a large tiem frame size (1.024s). We kept the best configuration in [1]. The number of parameters is 18M.
- DBNet: the structure of two branches are same. 6 (de) convolution block are set for encoder and decoder. The number of channels is 64 for each layer. The kernel size of (1, 3) and stride of (1, 2) are set for time and frequency axis. The input is time frames and SRS for time branch and frequency branch, respectively. The numebr of parameters is 2.9M.

3.3. System Settings

All utterances are sample at 16kHz. The frames are extracted using a rectangular window and a hamming window of size 20ms for time domain and frequency domain, respectively. The overlap is 10ms. The models are trained using the Adam optimizer [15] with a learning rate of 0.001. And the batch size is set to 32 at the utterance level. Note that a random segment of 7 seconds is intercepted from an utterance if it is larger than 7 seconds. The smaller utterances are zero-padded to match the size of the largest utterance in the batch.

3.4. Evaluate Metrics

The performance is evaluated with two objective metrics: short-time objective intelligibility (STOI) [16] and perceptual evaluation of speech quality (PESQ) [17]. STOI values typically range

Table 1: *STOI and PESQ comparisons between DBNet and the baseline models*

Metrix	STOI								PESQ								Casual?
Test Noise	Babble				Cafeteria				Babble				Cafeteria				
Test SNR	-5db	0db	5db	AVG	-5db	0db	5db	AVG	-5db	0db	5db	AVG	-5db	0db	5db	AVG	
Mixture	59.1	71.6	83.3	71.3	57.5	70.1	82.4	70.0	1.52	1.76	2.07	1.78	1.40	1.69	2.07	1.72	
CRN	77.9	88.0	93.2	86.4	75.7	86.6	92.7	85.0	1.99	2.50	2.91	2.47	2.01	2.47	2.89	2.46	✓
GCRN	81.1	90.7	94.7	88.8	78.0	88.8	94.1	87.0	2.03	2.65	3.07	2.58	1.97	2.56	3.03	2.52	✓
AECNN	80.5	90.6	94.2	88.4	**80.0**	89.4	94.0	87.8	2.00	2.57	2.88	2.48	2.03	2.55	2.93	2.50	✓
DBNet	**83.1**	**91.8**	**95.3**	**90.0**	79.6	**89.9**	**94.8**	**88.1**	**2.15**	**2.78**	**3.16**	**2.70**	**2.08**	**2.69**	**3.10**	**2.62**	✓
GCRN-NC	84.1	92.1	95.5	90.5	81.3	90.5	95.0	89.0	2.22	2.81	3.17	2.73	2.08	2.72	3.07	2.62	×
DBNet-NC	**87.3**	**93.5**	**96.2**	**92.3**	**84.3**	**91.9**	**95.7**	**90.6**	**2.56**	**3.06**	**3.34**	**2.99**	**2.43**	**2.95**	**3.31**	**2.90**	×

from 0 to 1, which can be roughly interpreted as percent correct. PESQ values range from -0.5 to 4.5. For both of the STOI and PESQ metrics, the higher number indicates better performance.

4. Result and Analysis

4.1. Objective Comparisons

First, we compare our model with baselines in terms of both STOI and PESQ. The results are given in Table 1 and the best results are marked in bold. For STOI, DBNet is better than all baselines at all SNRs and noises except AECNN. However, AECNN is a model based on large frame size so it is not suitble for real-time scenes. Compared with GCRN, an average improvement of 1.20% and 1.10% is obtained for babble and cafeteria, respectively. For PESQ, GCRN is the best baseline and an average improvement of 0.12 and 0.10 is obtained for babble and cefeteria, respectively. As for non causal system, DBNet-NC outperforms GCRN-NC and obtained an average improvement of 1.7 for STOI and 0.27 for PESQ.

In conclusion, the proposed dual-branch model outperforms both AECNN which is a time-domain based model and GCRN which is a frequency-domain based model for complex spectrogram mapping, indicating that alternate interconnection of the information of two domains can significantly improve the performance of the model and improve parameter utilization.

4.2. Deep Noise Suppression Challenge

The Deep Noise Suppression (DNS) challenge is designed to foster innovation in the area of noise suppression to achieve superior perceptual speech quality. Detailed information about the challenge can be found at the following link:
`https://www.interspeech2021.org`

To evaluate the performance of the proposed method on more complicated and real acoustic scenarios, the proposed model was trained with the DNS-Challenge wide band dataset which contains more complex acoustic scenarios including reverberation, singing, emotions, and non-English speech. The settings of generating training set is described as follow. The SNRs of training mixtures vary from -5dB to 25dB. Around 30% of utterances are convolved with provided synthetic and real room impulse responses (RIRs) before mixed with different noise signals, and we process speech, noise and reverberation by using the spectral augmentation filters proposed in [18]. Moreover, there is a 5% chance that there may be multiple compound noises in one utterance. To meet the requirement of DNS-Challenge, the number of channels is appropriately decreased. In addition, kernel size of convolution blocks is set to (2, 3).

We use DNSMOS [19] which is a reliable non-intrusive objective speech quality metric as our evaluation metrics at training stag and take Mean Opinion Score (MOS) of the ITU-T P.835 framework as result. The results of the evaluation using the ITU-T P.835 criterion [20] which is provided by the organizer are shown in Table 2. It is obvious that the proposed model outperforms the baseline (NSnet2) [21] by overall 0.12 DMOS score. Then we calculated the processing latency of our algorithm according to the competition requirements. In this model, the frame size T = 20ms, and the overlap between consecutive frames Ts = 10ms, so the latency T =30ms, which meets the requirements. We also evaluated the memory access cost (MAC), and the result is 2.847G per second.

Table 2: *Subjective evaluation with P.835 criterion on the DNS Challenge*

	Stationary	Emotional	Tonal	Non-English	Musical	English	Overall
Noisy	3.03	2.28	3	3.04	2.57	2.52	2.77
NSnet2	3.28	2.75	3.31	3.25	2.78	2.93	3.07
DBNet	3.44	3.07	3.37	3.43	2.78	2.93	3.19

5. Conclusions

In this study, we propose a novel single-channel speech enhancement system, which consists of two denoising branches on time domain and frequency domain. The results turns out that the proposed model outperforms other advanced models in terms of objective intelligibility and quality scores.

We explain our work has excellent performance because the two network branches have different learning focuses, and the features learned from different domains can complement each other. According to the principle of STFT, convolution in the time domain is equivalent to direct product in the frequency domain. Operations in the time domain tend to focus more on local information, and operations in the frequency domain focus more on the relationship between frames. A reasonable combination of the two can achieve better performance. And the proposed model has fewer parameters which indicate the two branch structure improves parameter utilization. Subjective results showed that the proposed system ranks the top 8 in the Mean Opinion Score (MOS) of the ITU-T P.835 for real-time track 1 of the INTERSPEECH 2021 Deep Noise Suppression (DNS) challenge.

6. Acknowledgements

The author would like to thank Yongjie Yan, Tailong Zhang and Pengjie Shen for their valuable comments. This research was partly supported by the China National Nature Science Foundation (No. 61876214).

7. References

[1] A. Pandey and D. Wang, "A new framework for cnn based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 7, pp. 1179–1188, 2019.

[2] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 380–390, 2019.

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

[4] Q. Li, F. Gao, H. Guan, and K. Ma, "Real-time monaural speech enhancement with short-time discrete cosine transform," *arXiv preprint arXiv:2102.04629*, 2021.

[5] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6629–6633.

[6] T. Y. Lim, R. A. Yeh, Y. Xu, M. N. Do, and M. Hasegawa-Johnson, "Time-frequency networks for audio super-resolution," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 646–650.

[7] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*, 1992, pp. 357–362.

[8] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," *arXiv preprint arXiv:2101.01902*, 2021.

[9] Y. Liu, H. Zhang, and X. Zhang, "Using shifted real spectrum mask as training target for supervised speech separation," in *Interspeech*, 2018, pp. 1151–1155.

[10] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*, 2017, pp. 933–941.

[11] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, and T.-Y. Liu, "Efficient sequence learning with group recurrent networks," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 799–808.

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

[13] 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," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. 2604–2612, 2016.

[14] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Interspeech*, 2018, pp. 3229–3233.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio Speech and Language Processing (TASLP)*, vol. 19, no. 7, pp. 2125–2136, 2011.

[17] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment: Part i: Time-delay compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, 2002.

[18] J.-M. Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *IEEE 20th international workshop on multimedia signal processing*, 2018, pp. 1–5.

[19] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1270–1279.

[20] B. Naderi and R. Cutler, "A crowdsourcing extension of the itu-t recommendation p. 835 with validation," *arXiv preprint arXiv:2010.13200*, 2020.

[21] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 871–875.