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

Fusion-Net: Time-Frequency Information Fusion Y-Network for Speech Enhancement

Santhan Kumar Reddy Nareddula, Subrahmanyam Gorthi, Rama Krishna Sai S. Gorthi

Department of Electrical Engineering, Indian Institute of Technology, Tirupati, India
santhanreddy6@gmail.com,s.gorthi@iittp.ac.in,rkg@iittp.ac.in

Abstract

This paper proposes a deep learning-based densely connected Y-Net as an effective network architecture for the fusion of time and frequency domain loss functions for speech enhancement. The proposed architecture performs speech enhancement in the time domain while fusing information from the frequency domain. Y-network consists of an encoder branch followed by two decoder branches, where the first and second decoder loss functions enforce speech enhancement in time and frequency domains respectively. Each layer of the proposed network is formed with densely connected blocks comprising dilated and causal convolutions for significant feature collection and error backpropagation. The proposed model is trained on a publicly available data set of 28 speakers with 40 different noise conditions. The evaluations are performed on an independent, unseen test set of 2 speakers and 20 different noise conditions. The results from the proposed method are compared with five state-of-the-art methods using various metrics. The proposed method has resulted in an overall perceptual evaluation of speech quality of 3.4. It has outperformed the existing methods by a significant margin in terms of all the evaluation metrics.

Index Terms: Speech Enhancement, Dense Network, Fusion Y-Net, Fully Convolutional Layer, Charbonnier Loss Function

1. Introduction

Speech enhancement is a key step for many speech related tasks like automatic speech recognition, improving the quality of audio in mobile communications, hearing aids and cochlera implants. Conventional approaches for speech enhancement include methods like Wiener filtering [1], spectral subtraction, statistical based approaches [2] and subspace algorithms. These methods can be categorized under unsupervised methods as they do not require any annotated training data. Machine learning-based speech enhancement methods are found to significantly outperform the conventional unsupervised algorithms by posing it as a supervised problem.

In the past few years, Deep Neural Networks (DNNs) based speech enhancement methods have provided state-of-the-art results. Initially proposed DNNs-based enhancement methods use spectral masking or time-frequency filtering [3]. In these methods, the input speech signal is usually converted into time-frequency representations like Short Time Fourier Transform (STFT) and only magnitude of the spectrum is enhanced. The phase of noisy signal is unaltered because of the difficulty to learn its structure and it was believed that phase of signal was not of great importance for speech enhancement [4].

Recent studies reveal that both phase and magnitude are indeed important for better enhancement of the speech [5]. Hence, later researchers started focusing on both magnitude and phase information for speech enhancement. The first approach under this category [6] performed complex spectrogram enhancement in which the STFT of the noisy signal is given as an input to a DNN, and the resulting spectrogram is compared with the clean STFT. Another approach under this category is a time domain approach in which the time domain waveform is directly given to a network with no time-frequency transformation, and the enhanced signal is compared with clean time-domain signal. Some networks that are used this approach include U-Net [7] and SEGAN [8].

The level of pre-processing required in the time domain-based speech enhancement methods is usually less compared to the frequency domain approaches. For better extraction of features and learning even with minimal pre-processing of the input signal, a network with deep architecture is required. The network proposed in [9] contains such an encoder-decoder structure with dense blocks and it provides a deep network for speech enhancement in the time domain.

It can be noted from the aforementioned literature that most of the existing speech enhancement methods are based on features extracted from either the time domain or the frequency domain. However, the features extracted from both these domains can be potentially complementary to each other. Motivated by this observation, this paper proposes an elegant architecture that has natural ability to fuse the time and frequency domain inferences while performing direct time domain speech enhancement. For better learning of the network, we introduce Charbonnier loss function[10], which is smooth and differentiable with L_2 loss function near origin and robust to reject outliers with L_1 loss function else where.

The rest of the paper is organized as follows. The next section presents the details of the proposed Y-Net architecture. Section 3 presents the experimental setup. Section 4 presents evaluations results. Finally conclusions are presented in section 5.

2. Proposed Approach

As discussed in the introduction, the existing deep learning approaches so far have been relying on speech enhancement either in time or in frequency domain only. The primary reason for this is that most of the well known deep learning architectures rely on auto-encoder and have an encoder-decoder structure. Popular networks for speech enhancement in time domain also employ attention as in wave-U-Net [11] or GAN framework as in SEGAN [8]. These methods aim to enhance speech quality considering only time-domain information, without consideration of any frequency details. While other techniques in the literature enhance the speech considering only the frequency domain information [6].

In recent years, there are a few attempts to come up with end-to-end deep learning architectures that relay on both time and frequency domain inferences for speech enhancement. For example, the speech enhancement approach in [9], having

http://dx.doi.org/10.21437/Interspeech.2021-1184

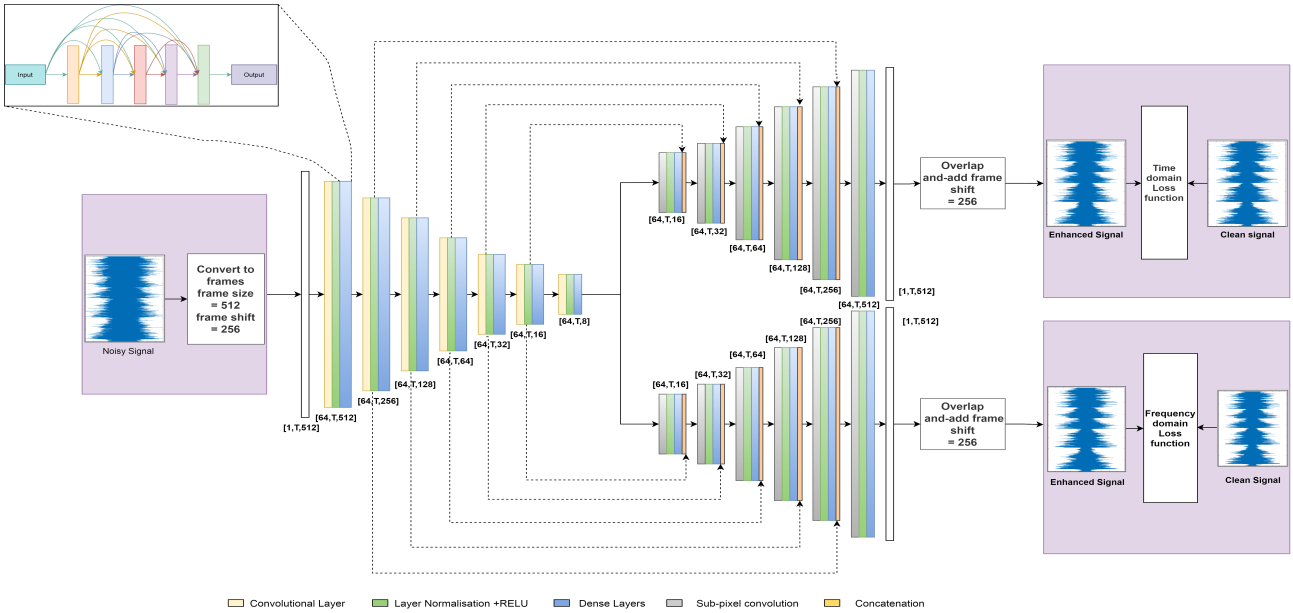

Figure 1: ***Fusion Y-Net*** *: Proposed Fusion NET with the time and frequency domain loss functions.*

encoder-decoder architecture, consists of both time and frequency domain loss functions at the decoder itself. Since, the information from both domains are employed to a single decoder branch, this might lead to unsuccessful blending of information. The proposed architecture in this paper combines information coming from both the domains through different decoder branch and has the provision for end-to-end learning. The developed architecture has an encoder and two independent decoder branches for back-propagation of time and frequency domain losses, respectively. Through the proposed architecture, the enhancement of speech signal happen in time-domain itself, while frequency domain decoder branch helps to modify the weights of time-domain encoder and acting as a catalyst for a better enhancement of the speech signal.

As against the the well known L_2 loss function or widely employed L_1 loss in the existing DNNs for speech enhancement, we propose Charbonnier loss function for speech enhancement, that effectively blends the salient aspects of both of these and has demonstrated performance in image super-resolution/image enhancement applications.

2.1. Proposed Architecture :

We propose a Fusion-Net architecture as shown in figure1 that has time domain loss at one decoder and the frequency domain loss at the other decoder block. The time-domain noisy signal is given as input to the encoder. In contrast to the base Y-Net in [12], the proposed Y-Network (i) comprises of dense blocks in both encoder and decoders, (ii) uses dilated and causal convolutions in dense blocks for effective speech processing, (iii) incorporates sub-pixel convolutions for up-sampling in the decoders layers, while strided convolutions are used for down-sampling in encoder layer, and (iv) proposes loss functions that are appropriate for time and frequency domain similarity assessment for speech enhancement.

2.2. Signal Flow through the Proposed Architecture:

The input speech signal frames are fed to the encoder. (1,1) convolutions are employed on the input signal to increase the number of channels to 64, and then these feature maps are fed to the dense block that consists of 5 layers of 2D convolutions. These dense blocks have dilated and causal convolutions to increase the receptive field over layers. Causal convolutions are performed across the frames ensure no information leakage from the future frames. Convolution in all dense blocks uses a kernel(filter) of size (2,3) with output 64 channels. In encoder, down-sample is performed through convolution with kernel of size (1,3) and a stride of (1,2), from layer to layer. Each convolution is followed by layer normalization [13] and PReLU [14] activation.

The output from the encoder is given to two decoder blocks, whose architecture is similar to that of encoder branch. Here up-sampling is done by sub-pixel convolution instead of transposed convolution to avoid checkerboard artifacts [18] as suggested in image super resolution [13] literature. The input to the every decoder layer is obtained by concatenating the output from the previous layer and output from the corresponding layer from the encoder. We use a kernel size of (1,3) for performing the up-sampling with sub-pixel convolution by a factor of two. To convert output channels form 64 to 1 again (1,1) convolutions are employed.

2.3. Proposed Loss Functions:

Once the speech signal frames are obtained by overlap-and-add method, these frames are fed as input to the encoder branch of Y-Net and enhanced signal output is obtained at both the decoders. First decoder block is employed with the time domain loss function and the second decoder block with the frequency domain loss. As shown in the Table 3, we experimented two loss functions in each of these domains. In this work, we propose the Charbonnier loss function, a differentiable variant of L_1 norm, for effective speech enhancement and demonstrate that it is quite effective in back-propagating the

Table 1: *Quantitative evaluation and comparison of the proposed Fusion Y-Net, with 5 state-of-the-art methods.*

Network Name	**PESQ**	**CSIG**	**CBAK**	**COVL**	**SSNR**
Noisy	1.97	3.35	2.44	2.63	1.68
Wiener [15]	2.22	3.23	2.68	2.67	5.07
SEGAN [8]	2.16	3.48	2.94	2.80	7.73
Wave-U-Net [16]	2.40	3.52	3.24	2.96	9.97
Wave-U-Net+ Attention [11]	2.62	3.91	3.35	3.27	10.05
Multi head attention [17]	2.99	4.15	3.42	3.57	-
Our model (Fusion Y-Net$_3$(TDB))	**3.40**	**4.57**	**3.80**	**4.0**	**10.67**

error information even in DNNs. This is one of the robust loss function, that can handle outliers that may arise from the noise in speech enhancement problems. It is profoundly used in deep learning based super resolution networks and optical flow estimation, however have not been payed much attention in the speech enhancement. In this work, we introduce Charbonnier loss in speech enhancement, for the first time, for smooth learning and and robust denoising.

The time domain L_2 and $Charbonnier$ loss functions are given by

$$L_2(x, \hat{x}) = \frac{1}{M} \sum_{n=0}^{M-1} [\hat{x}(n) - x(n)]^2 \quad (1)$$

$$Charbonnier(x, \hat{x}) = \frac{1}{M} \sum_{n=0}^{M-1} \sqrt{[\hat{x}(n) - x(n)]^2 + \alpha} \quad (2)$$

where $\hat{x}(n)$ denotes the enhanced signal, $x(n)$ represents clean signal and α is a constant to stabilise the training process. The form of equations for frequency domain loss functions remain same, except the replacement of $x(n)$ by complex spectrum $X(k)$.

The enhanced signal from second decoder of Y-Net is compared with clean signal in frequency domain. Here, we apply STFT to both the signals for representing the frequency information. Here, again, we consider two possibilities for the loss functions, one by applying L1 norm [19], other by applying Charbonnier loss on the STFT coefficients.

The time and frequency domain losses are combined as mentioned below:

$$L(x, \hat{x}) = \gamma * L_t(x, \hat{x}) + (1 - \gamma) * L_f(x, \hat{x}) \quad (3)$$

γ is hyper parameter set by experimentation and cross validation.

3. Data and Experimental Setup

The proposed method is evaluated on the publicly available and widely used CSTR VCTK corpus data set [20]. The training data consists of 28 speakers (14 male and 14 female) with 40 different noise conditions (10 noises $\times$ 4 signal-to-noise(SNRs)). The SNR values used for training are: 15 dB, 10 dB, 5 dB and 0 dB. Similarly for test set consists of 2 speakers with 20 different noise conditions (5 noises $\times$ 4 SNRs). There are around 400 sentences available from each speaker. Total training set consists of 11,572 audio files and test set consists of 824 audio files. Test set is made totally unseen from training set with different noise conditions and speakers.

Given utterances are down sampled from 48 kHz to 16 kHz similar to the experimental setup presented in [8]. In training phase the given speech signal is sliced into chunks of waveforms with a sliding window of approximately 1 second of speech (16384 samples) with an overlap of 50%. Testing set has been generated without an overlap through the whole duration of signal. The obtained waveforms are then converted to frames which are extracted by a rectangular window of size 512 samples with an overlap of 256 (50%) samples. For the Stochastic Gradient Descent(SGD) based optimisation, we used Adam optimiser. The training of the model is done for 60 epochs with step learning rate as shown in the given Table.2 with a batch size of 24. Energy normalisation is done for both train and test set. The value of γ in equation 3 is set to 0.85. The GPU processors used for training are NVIDIA-2080TI with a graphic card of 11GB.

Table 2: *Step wise learning rate for training the network*

Epochs	9	18	9	24
Learning Rate	$2e-4$	$1e-4$	$5e-5$	$1e-5$

4. Results and Discussion

In order to evaluate the quality of enhanced speech, we compute the following measures, all these metrics compare enhanced signal with its label.

- PESQ [21]: Perceptual evaluation of speech quality, predicts the Mean Opinion Score(MOS) of degraded audio sample. range(-0.5 to 4.5)
- CSIG [22] : It predicts the MOS of signal distortion concentrating only the speech signal.(ranges 1-5)
- CBAK [22] : It predicts the MOS of background noise intensiveness.(ranges 1-5)
- COVL [22] : It predicts MOS of the overall effect (background and the speech, ranges 1-5)
- SSNR [23] : Segmental Signal-to-Noise ratio. (ranges 0 to ∞)

All these composite score metrics are correlated with human listen ratings, and more details can be found in [22]. Notice that higher metrics values indicate better signal enhancement.

Table 1 shows the comparison of our network with state-of-the art networks. Noisy in the table represents scores of given noisy data set. Wiener filter is one of the popular conventional filter that estimates the enhanced signal, experimental results shows that CSIG metric score of this filter is indeed less than noisy signal and rest of metric scores are also

Table 3: *Results of the ablation study for the proposed Fusion Y-Net and the existing encoder-decoder networks.*

Network Name	Loss Functions	Evaluation	PESQ	CSIG	CBAK	COVL	SSNR
DCED-L_2	L_2-T	Normal evaluation	3.31	4.43	3.66	3.88	9.85
DCED-($char$)	$char$-T	Normal evaluation	3.36	4.52	3.7	3.96	10.25
Fusion Y-Net$_1$	L_2-T, L_1-F	Both decoders	2.98	3.46	3.35	3.23	7.21
Fusion Y-Net$_1$(TDB)	L_2-T, L_1-F	TDB	3.15	3.93	3.58	3.55	9.44
Fusion Y-Net$_2$	L_2-T, $char$-F	Both decoders	3.08	3.69	3.41	3.39	7.25
Fusion Y-Net$_2$(TDB)	L_2-T, $char$-F	TDB	3.34	4.50	3.74	3.94	10.06
Fusion Y-Net$_3$	$char$-T, $char$-F	Both decoders	3.10	3.80	3.46	3.46	7.6
Fusion Y-Net$_3$(TDB)	**$char$-T, $char$-F**	**TDB**	**3.40**	**4.57**	**3.80**	**4.0**	**10.67**

less. SEGAN framework tries to achieve speech enhancement in deep learning with adversarial training, the network showed some decent results. But PESQ score of SEGAN is less than wiener filter. Later encoder- decoder architectures like wave-U-Net and Wave-U-Net + Attention has boosted up speech enhancing capabilities in time-domain networks. In recent times, time-frequency masking methods has better performance when compared to encoder-decoder architectures. Taking these intuitions into account, we proposed a network that should learn frequency and time domain information and deep enough to extract features better. We also identify that robust loss function is necessary for learning through the deep network architectures for speech enhancement, and thus employ Charbonnier loss in the proposed Fusion Y-Net. Our Fusion Y-Net surpassed all the current existing algorithms in terms of subjective metrics and approximate improvement of 5% to 10% can seen on every metric.

4.1. Ablation Study

Following experiments are done by varying the loss functions during training. All the results are tabulated in the Table: 3 Encoder-decoder architecture is one of the best network for enhancing speech signal in time domain. We first experimented with loss functions of both domains at single decoder in encoder-decoder architecture, and the results are not satisfactory, perhaps due to the complexity of learning the encoder from the common decoder.

Fusion Y-Net$_1$: The experiments with Y-Net architecture with an encoder and two decoder branches corresponding to time and frequency domain loss functions is carried out. The training of the network was done with L_2 time domain and L_1 frequency domain loss function. The evaluation of this model was done by inferring the enhanced signal (i) as weighted sum of both the decoder block outputs referred to as Fusion Y-Net$_1$ and (ii) only from first decoder block referred as Fusion Y-Net$_1$ Time Domain Branch (TDB). These results show that taking inference directly from time domain alone is better than the combined inference. However, employing the frequency domain branch, while training has significantly improved the performance than a single decoder architecture.

Fusion Y-Net$_2$: It is noticed from the recent literature on deep learning based image enhancement [24] that Charbonnier loss function performs well in signal enhancement applications. Hence, in this network, we experiment replacing frequency domain L_1 loss with Charbonnier loss function while keeping time domain loss function unaltered. The training and evaluation procedures are same as that in Fusion Y-Net$_1$ and the evaluation results with Fusion Y-Net$_2$(TDB) provided better results than Fusion Y-Net$_1$(TDB), indicating the advantage of Charbonnier loss, even for speech enhancement.

Fusion Y-Net$_3$: Motivated by the improvements observed in Fusion Y-Net$_2$ by the introduction of Charbonnier loss, in this network, both the time and the frequency domain loss functions are replaced with Charbonnier loss function. The evaluation results with time domain decoder branch Fusion Y-Net$_3$(TDB) have outperformed all the variants discussed above.

As an another aspect of ablation study, we have compared the proposed Fusion Y-Net$_3$ with Densely Connected Encoder-single Decoder blocks (DCED) with L_1 loss and Charbonnier loss functions, respectively and we found that Fusion Y-Net has outperformed these two network architectures as well, indicating the need for two decoder architecture.

5. Conclusions

This paper proposes a new fully convolutional Y-Net architecture learnt by robust Charbonnier loss function for speech enhancement. The proposed method's main contribution is its ability to fuse information from both the time and the frequency domains through loss functions of respective decoder branches of the Y-Network. Further, the use of dense building blocks in Y-Net trained by smooth and robust Charbonnier loss enabled better feature extraction and the proposition of deeper architecture for speech enhancement. To our knowledge, this is the first speech enhancement framework that has independent decoder branches driven by time-domain and frequency-domain loss functions. Fusion of these inferences is conveniently addressed in the proposed architecture and that lead to a better enhancement of the speech signal. Ablation studies show that training the network using Charbonnier loss function for both time-frequency domains, while considering the speech enhancement in time-domain is effective for better enhancement of input time domain noisy speech signal. The comparative results have demonstrated that the proposed Fusion Y-Net architecture has significantly improved the accuracies of speech enhancement compared to the recent state-of-art algorithms, resulting in around 10% improvement in PESQ, and an overall average improvement of around 8% across all composite scores.

6. Acknowledgements

Santhan kumar Reddy thanks Analog Devices India(ADI) for sponsoring his master's program at IIT Tirupati.

7. References

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

[2] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

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

[4] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[5] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[6] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2017, pp. 1–6.

[7] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[8] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proc. Interspeech 2017*, 2017, pp. 3642–3646. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1428

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

[10] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 2432–2439.

[11] 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)*. IEEE, 2019, pp. 249–253.

[12] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro, "Y-net: joint segmentation and classification for diagnosis of breast biopsy images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 893–901.

[13] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

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

[15] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.

[16] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

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

[18] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," *arXiv preprint arXiv:1707.02937*, 2017.

[19] 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*, vol. 27, no. 7, pp. 1179–1188, 2019.

[20] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

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

[23] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall, 1988.

[24] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.