CFTNet: Complex-valued Frequency Transformation Network for Speech Enhancement

Conference Paper · August 2023

DOI: 10.21437/Interspeech.2023-280

2 authors:

Nursadul Mamun
University of Texas at Dallas
46 PUBLICATIONS **240** CITATIONS
SEE PROFILE

John H. L. Hansen
University of Texas at Dallas
704 PUBLICATIONS **15,322** CITATIONS
SEE PROFILE

INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

CFTNet: Complex-valued Frequency Transformation Network for Speech Enhancement

Nursadul Mamun, John H.L. Hansen

Cochlear Implant Processing Laboratory, Center for Robust Speech Systems (CRSS-CILab),
Department of Electrical & Computer Engineering, The University of Texas at Dallas
(nursadul.mamun,john.hansen)@utdallas.edu

Abstract

It is widely known that the presence of multi-speaker babble noise greatly degrades speech intelligibility. However, suppressing noise without creating artifacts in human speech is challenging in environments with a low signal-to-noise ratio (SNR), and even more so if noise is speech-like such as babble noise. Deep learning-based systems either enhance the magnitude response and reuse distorted phases or enhance the complex spectrogram. Frequency transformation block (FTB) has emerged as a useful architecture to implicitly capture harmonic correlation which is especially important for people with hearing loss (hearing aid/ cochlear implant users). This study proposes a complex-valued frequency transformation network (CFTNet) for speech enhancement, which leverages both a complex-valued U-Net and FTB to capture sufficient low-level contextual information. The proposed system learns a complex transformation matrix to accurately recover speech in the time-frequency domain from a noisy spectrogram. Experimental results demonstrate that the proposed system can achieve significant improvements in both seen and unseen noise over state-of-art networks. Furthermore, the proposed CFTNet can suppress highly nonstationary noise without creating musical artifacts commonly observed in conventional enhancement methods.

Index Terms: Speech Enhancement, Complex-value Network, Frequency Transformation Block, Deep Neural network, U-Net

1. Introduction

Cochlear implants (CI) allow CI recipients to achieve near-to-normal speech intelligibility in quiet acoustic conditions. However, speech understanding in the presence of background sounds or competing talkers is one of the main challenges for CI users in daily life [1, 2]. Speech enhancement (SE) techniques have been utilized to eliminate background noise from captured speech signals and are beneficial [3, 4, 5]. More recently, deep neural network-based approaches have shown considerable improvements in performance by reducing non-stationary noise [4, 6, 7, 8]. Unlike most signal processing methods, deep neural networks learn patterns for speech enhancement and generalize them to larger unseen scenarios with the help of non-linear optimization. Convolutional neural networks (CNN) can efficiently address local temporal-spectral structures of speech and can effectively separate the speech and noise components in noisy signals [9]. However, CNN models cannot preserve global information and spatial arrangement of the previous features [10]. More specifically, DNN- and CNN-based models have limited capability to restore high-frequency components of speech, thus leading to a lower speech-to-distortion ratio of enhanced speech. In addition, correlation in harmonics refers to the presence of similar patterns in the distribution of frequencies present in a signal or image. In natural images, these harmonics are mostly local and can take the form of repeating patterns in texture while those in speech frequency spectrograms are non-local and can be seen as repeating peaks in the frequency spectrum. Therefore, conventional CNN kernels cannot capture the harmonics. Alternatively, fully convolutional networks (FCN) can model high and low-frequency components of raw waveforms simultaneously [10]. To further increase the performance and ability of speech-denoising techniques, researchers have used various architectures such as U-Net [11, 12, 13, 14], ResNet [15], DenseNet [16], Convolutional Recurrent Network (CRN) [17] and R-CED [18]. Among them, U-Net-like models have been successfully used in several speech applications such as speech denoising [14], speech dereverberation [11], speech to language technology [19] etc.

Typically, U-Net compresses features along the encoder and then reconstructs along the decoder. To localize, high-resolution features from the contracting (encoder) path are combined with the up-sampled (decoder) path to increase the resolution of the reconstructed speech. Several studies showed that common noise reduction algorithms suppress some of the harmonics that exist in the original signal, which directly influences speech quality [20, 21]. This suggests that the regeneration of such harmonics can restore distorted frequencies and improve the quality of the enhanced signal. Therefore, the existing U-Net structure-based SE algorithms cannot efficiently exploit harmonics and thus produce musical noise in the enhanced signal [22]. In addition, most of these SE networks focus only on processing the magnitude spectrogram and use the original noisy phase to reconstruct the signal. Recent studies have revealed that phase plays a crucial role in perceptual quality in SE [19] and this motivated the researcher to design different phase-aware SE networks such as deep complex convolution recurrent neural (DCCRN) networks [23].

In this study, we introduce a complex-valued frequency transformation network (CFTNet) for speech enhancement. CFTNet uses U-Net style CNN as a backbone [12] and incorporates frequency transformation layers (FTL) to exploit correlation among all frequency harmonics, which have been proven to be useful to capture global correlations over frequency for T-F representations [24]. This allows the network to use limited frequency information to reconstruct missing frequency components in the distorted signal. The proposed CFTNet employs complex-valued convolution in the encoder/decoder layers and complex-valued GRU at the bottleneck based on its effectiveness to reconstruct enhanced phase along with enhanced magnitude.

This paper is organized as follows: Sec. 2 briefly introduces the proposed CFTNet for single-channel speech enhancement.

10.21437/Interspeech.2023-280

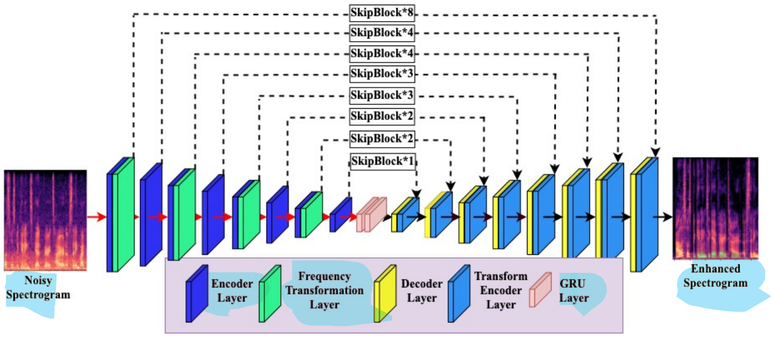

Figure 1: *Fig. 1. Basic block diagram of CFTNet with the complex frequency transformed module.*

Details on the experimental setup are discussed in Sec. 3, followed by results in Sec. 4. Finally, we conclude this study in Sec. 5.

2. Methodology

Here, we briefly discuss the standard U-Net and then propose modifications that construct CFTNet and the proposed SE network.

2.1. General Network Architecture

The goal of CFTNet is to parse degraded speech and recover high-quality signal content. The overall architecture of the proposed CFTNet is depicted in Fig. 1. The network consists of three main components: (i) a fully convolutional complex-valued encoder-decoder network (Cplx-UNet), (ii) complex-valued SkipBlocks (SB) within the skip connections between encoder and decoder, and (iii) complex-valued frequency transformation modules. The encoder network is designed using a series of encoder blocks followed by an FTL module in the alternate layers until the encoder downsamples to a single pixel. This ensures the decoder uses all spectral and temporal features learned by the encoder. Each encoder/decoder block is built upon complex-valued convolution layers to ensure successive enhancement of both magnitude and phase. A complex convolutional block in the skip connection reduces the semantic gap between the encoder and decoder blocks and thus guides the decoder to reconstruct the enhanced output. Although U-Net can capture contextual correlation for prediction, it yields less attention toward harmonic correlation. The proposed CFTNet can capture long-range dependencies and correlations among harmonics using FTL layers that CNN does not possess due to its localized receptive fields. Therefore, CFTNet employs an FTL block along with the encoder layer to exploit harmonic structures in the frequency components.

2.2. Complex-Valued Encoder/Decoder Layer

Complex convolution is the key difference between a complex-valued network and a real-valued network. The use of complex convolution in the U-Net architecture is to perform consistent improvement in both magnitude and phase in a T-F representation of noisy speech toward reconstructing a clean speech signal. Each complex-valued encoder layer in the proposed network consists of a complex convolution layer followed by complex batch normalization and complex nonlinear activation function. The complex-valued decoder layer is like a complex-valued encoder layer except complex convolution is substituted for complex-transpose convolution. Next, an algorithmic formulation of the complex convolution is presented. Let $X = X_r + jX_i$ represent the complex input such that $W = W_r + jW_i$ represents the complex kernel of the network. The resulting output of the convolution can be represented as,

$$Z = W * X = (W_r + jW_i) * (X_r + jX_i)$$

$$Z = (W_r * X_r + W_i * X_i) + j(W_r * X_i - W_i * X_r)$$

2.3. Complex-valued Frequency Transformation (FTL)

A frequency transformation block is a technique for relating correlation among harmonics along the frequency axis in a T-F representation [24]. A complex FTL is inserted after encoder layers so that output features have a full-frequency receptive field. In the speech, existing networks employ FTL frameworks with real-valued networks that operate on the magnitude response [22]. Here, we extend the FTL framework to address the complex domain by proposing a complex-valued FTL module that attends to features in frequency while maintaining interdependence between real and imaginary components of the complex-valued feature map. The FTL module consists of three stacked CNN layers in the attention module: (1) one fully connected layer, (2) one CNN layer used in a frequency transformation matrix (FTM), and (3) one CNN layer used for concatenation. Consider a set of complex-valued feature maps,

$$U_0(t) \varepsilon T \times F \times C$$

that is extracted from stacked CNN layers in the encoder consisting of a sequence of 1-D frequency vectors with a total of T frames. The trainable FTM can then be represented as,

$$W_{FTM} \varepsilon R^{F \times E},$$

where C, T, and F denote the channels, time, and frequency axis, respectively. We first apply the attention module to the incoming feature maps, which are then point-wise multiplied with the input features to exploit the inter-channel relationship of the features and output U_a. Next, a trainable FTM is applied to the feature maps at the time step, t_0, and ensures the global frequency correlation along the frequency axis. Finally, the output features of the FTM module are concatenated with the input features, $U_0(t)$ using a CNN layer to ensure both global and local frequency correlation among harmonics.

2.4. Complex-valued SkipBlocks

A skip connection in a U-Net architecture passes high-dimensional features from the encoder layer to the appropriate decoder layer. This enables the network to preserve spatial features lost during the down-sampling operation and guides the network to propagate from encoder to decoder. Although skip connections have been shown to significantly impact the development of robust networks, a recent study [25] identified a possible semantic gap while sharing features between the encoder and decoder and strengthening feature representation. This could be due to incompatible feature sets shared between the encoder and decoder that cause an adversarial impact on speech synthesis. Inspired by the success of image segmentation [25] and speech dereverberation [11], this study uses complex convolution blocks in the skip connection. This ensures a network that shares similar spectral information to improve learning capabilities.

3. Evaluation Methods

The speech stimuli used in this study were sentences from TIMIT [26]. The dataset consists of 6,300 phonetic transcribed speech utterances (approx. 3.5 hours) of American English speakers. The length of each sentence varies from 3 to 5 seconds. A subset of 800 sentences was used to train the model. These sentences were distorted with eight different noises from the AURORA dataset at 5 different SNRs: -10, -5, 0, 5, and 10 dB. These environmental noise conditions included samples from the following: airport, babble, car, exhibition, station, street, speech-shaped, and white Gaussian noise. A second subset of 50 samples was used to test the model distorted in three seen (babble, car, and speech-shaped noise) and two unseen (restaurant and train) noise types at 7 different SNR levels (-7.5, -5, -2.5, 0, 2.5, 5, and 10 dB). Seen noise refers to the noise type which is seen by the model during training whereas unseen noise is completely unknown by the model. All speech samples and noises were resampled at 16 kHz.

3.1. Network Architecture

Complex-valued frequency transformation network uses a frequency transformation block, complex-valued convolution, and complex-valued GRU layers to estimate a generalized nonlinear mapping from a noisy speech T-F spectrum to a corresponding clean speech spectrum. First, the short-time Fourier transform (STFT) of the speech signal with a frame size of 16 ms and an overlap of 8 ms is computed. Next, the network architecture with eight layers of encoder-decoder pairs, four FTB layers, two GRU layers as bottleneck layers, and convolution layers in the skip connections are employed as shown in Fig. 1. Each encoder layer uses convolution layers with a kernel size of 3×3 and stride of 2×1. Similarly, decoder layers use the same parameters apart from the transposed convolution. To ensure harmonic correlation in the frequency axis, we use an FTB layer in each alternate layer of the encoder. Parameters of the FTB layer are selected based on the parameters in the corresponding encoder layer. The proposed system is based on the complex spectrogram using complex-valued convolutions and complex-valued GRU layers to ensure consistent advancement in both magnitude and phase. The network is trained for 50 epochs with an Adam optimizer, an initial learning rate of 0.0003, and a batch size of 16. Lastly, a combination of scale-invariant signal-to-distortion ratio (SI-SDR) loss and frequency loss (STFT loss) was used as an objective function to minimize mean square error (MSE) between the network prediction and the corresponding clean spectrogram. This STFT loss calculates the spectral convergence and spectral magnitude losses in the STFT domain where SI-SDR is responsible for channel variations, interference, and artifacts in the time domain signal.

3.2. Evaluation metrics

A total of 5 objective metrics is used to evaluate the intelligibility and quality of reconstructed speech. Short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) are frequently used intrusive metrics for speech intelligibility and quality measurement [27, 28]. STOI is a function of T-F representation of the signal that compares temporal envelopes of estimated and reference signals in short-time regions and maps them to a range between [0, 1]. PESQ for the narrow band is a perceptual evaluation related to subjective opinion and varies from -0.5 to 4.5. Spectrogram orthogonal polynomial measure (SOPM) predicts speech intelligibility using orthogonal polynomial features and varies from 0 to 1 [29]. To investigate distortion in the frequency domain, the log-spectral distance (LSD) and scale-invariant signal-to-distortion ratio (SI-SDR) metrics are used as objective metrics [30]. The higher the objective score, the better the quality, except for LSD.

4. Results

This section presents simulated results of the proposed CFTNet algorithm in terms of objective scores. Performance is evaluated using a speech intelligibility metric, a quality metric, and two speech distortion metrics. The estimated scores for the CFTNet are compared with scores from three existing algorithms for different seen and unseen noises and SNRs. In addition, the effect of different loss functions on the training of the proposed algorithm is also presented.

Table 1: *Mean objective scores for CFTNet model trained using five different loss functions. Significance is denoted in bold at the level.*

Loss Function	Objective Metric				
	STOI	PESQ	SISDR	LSD	SOPM
Unprocessed	0.69	1.30	0.35	6.72	0.76
SISDR	0.82	1.58	4.79	9.24	0.84
STFTLoss	0.81	1.76	-15.94	3.96	0.76
SNRLoss	0.83	1.54	-8.52	4.72	0.87
SISDR+ α*FreqLoss	**0.86**	**2.14**	**8.05**	**3.25**	**0.88**

4.1. Effect of loss function on CFTNet

Speech intelligibility and quality of a speech signal improve when the loss between the reconstructed and target signal decreases. A proper selection of loss functions guides the network to the global minima. Different objective metrics are used as a loss function to evaluate the network performance with seen and unseen objective metrics. Table 1 represents the mean objective scores for the proposed network for 5 different loss functions. The objective score represents the mean score for 50 speech samples distorted at five different noisy conditions and seven different SNRs. In general, the objective score for the unprocessed signal is lower than the CFTNet, irrespective of the objective metric and loss function. The combination of time-domain and frequency-domain, metrics demonstrated the best performance with respect to the other four objective metrics. Therefore, the combination of SI-SDR and STFTLoss (with $\alpha = 25$) metric is used as a loss function for the proposed network and for further evaluation.

4.2. Effect of seen and unseen noise on CFTNet

To analyze the performance of the proposed network in seen and unseen noises and SNRs, objective scores are computed for different noises and presented in Table 2. Each score represents the average objective intelligibility or quality score of 150 and 100 speech samples for seen (50×3) and unseen (50×2) noise, respectively. In general, objective scores for enhanced speech are higher than unprocessed speech. Improvement in STOI is higher for lower SNRs and improvement in PESQ is higher for higher SNRs. Performance is found to increase as seen noise and SNRs were incorporated into the training set.

Table 2: *Mean objective scores for CFTNet model trained using five different loss functions. Significance is denoted in bold at the level.*

	STOI				**PESQ**			
SNR (dB)	**Seen Noise**		**Unseen Noise**		**Seen Noise**		**Unseen Noise**	
	Noisy	Enh.	Noisy	Enh.	Noisy	Enh.	Noisy	Enh.
-7.5	0.50	0.80	0.51	0.64	1.14	1.67	1.15	1.22
-5	0.56	0.83	0.56	0.71	1.14	1.86	1.17	1.32
-2.5	0.63	0.87	0.64	0.78	1.17	2.07	1.20	1.50
0	0.69	0.90	0.71	0.84	1.21	2.32	1.25	1.73
2.5	0.77	0.92	0.77	0.88	1.30	2.55	1.99	1.95
5	0.81	0.94	0.83	0.91	1.39	2.76	1.41	2.28
10	0.91	0.97	0.91	0.96	1.68	3.17	1.68	2.88
Mean	**0.70**	**0.89**	**0.70**	**0.82**	**1.29**	**2.34**	**1.41**	**1.84**

4.3. Ablation study

To analyze the performance of the proposed network, predicted scores for the CTFNet are compared with scores from three different existing networks, CRN [17] and UNet-SCB [11], DCCRN [23]. Scores are predicted for 50 samples at five different noises and seven different SNRs and averaged scores are reported in Table 3. Relative enhancement achieved by different algorithms is measured using STOI, PESQ, SI-SDR, and LSD metrics. Results indicate that the proposed CFTNet provides benefits in speech enhancement in terms of all four-objective metrics over unprocessed signals and the enhanced signal from CRN, UNet-SCB, and DCCRN networks.

Table 3: *Ablation study of the proposed network. Average improvement across all noise types and SNRs are presented in terms of STOI, PESQ, SI-SDR, and LSD.*

Models	STOI	PESQ	SI-SDR	SOPM	LSD
Noisy	0.69	1.29	0.35	0.76	6.72
CRN	0.71	1.30	1.60	0.81	8.80
UNet-SCB	0.72	1.35	2.14	0.78	6.50
DCCRN	0.80	1.62	4.90	0.87	**3.14**
CFTNet	**0.86**	**2.14**	**8.05**	**0.88**	3.25

5. Conclusion

To analyze the performance of the proposed network, predicted scores for the CTFNet are compared with scores from three different existing networks, CRN and UNet-SCB, DCCRN. Scores are predicted for 50 samples at five different noises and seven different SNRs and averaged scores are reported. Relative enhancement achieved by different algorithms is measured using STOI, PESQ, SI-SDR, and LSD metrics. Up-and-downward arrows represent improvement using different objective metrics. Results indicate that the proposed CFTNet provides benefits in speech enhancement in terms of all five objective metrics over unprocessed signals and the enhanced signal from CRN, UNet-SCB, and DCCRN networks.

6. Acknowledgment

This work was supported by Grant No. R01 DC016839-02 from the National Institute on Deafness and Other Communication Disorders (NIDCD), National Institutes of Health (NIH).

7. References

[1] F.-G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, “Cochlear implants: system design, integration, and evaluation,” *IEEE reviews in biomedical engineering*, vol. 1, pp. 115–142, 2008.

[2] J. H. Hansen, H. Ali, J. N. Saba, M. R. Charan, N. Mamun, R. Ghosh, and A. Brueggeman, “Cci-mobile: Design and evaluation of a cochlear implant and hearing aid research platform for speech scientists and engineers,” in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019, pp. 1–4.

[3] F. Bolner, T. Goehring, J. Monaghan, B. Van Dijk, J. Wouters, and S. Bleeck, “Speech enhancement based on neural networks applied to cochlear implant coding strategies,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6520–6524.

[4] C. Lee, H. Hasegawa, and S. Gao, “Complex-valued neural networks: A comprehensive survey,” *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 8, pp. 1406–1426, 2022.

[5] C. K. A. Reddy, N. Shankar, G. S. Bhat, R. Charan, and I. Panahi, “An individualized super-gaussian single microphone speech enhancement for hearing aid users with smartphone as an assistive device,” *IEEE signal processing letters*, vol. 24, no. 11, pp. 1601–1605, 2017.

[6] N. Mamun, S. Majumder, and K. Akter, “A self-supervised convolutional neural network approach for speech enhancement,” in *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*. IEEE, 2021, pp. 1–5.

[7] A. Pandey and D. Wang, “Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.

[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[9] N. Mamun, S. Khorram, and J. H. Hansen, “Convolutional neural network-based speech enhancement for cochlear implant recipients,” in *in Interspeech*, vol. 2019. NIH Public Access, 2019, p. 4265.

[10] 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.

[11] V. Kothapally, W. Xia, S. Ghorbani, J. H. Hansen, W. Xue, and J. Huang, “Skipconvnet: Skip convolutional neural network for speech dereverberation using optimally smoothed spectral mapping,” *in Interspeech*, 2020.

[12] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[13] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[14] K. Akter, N. Mamun, and M. A. Hossain, "A tf masking based monaural speech enhancement using u-net architecture," in *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2023, pp. 1–5.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

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

[18] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *in Interspeech*, 2016.

[19] A. Joglekar, S. O. Sadjadi, M. Chandra-Shekar, C. Cieri, and J. H. Hansen, "Fearless steps challenge phase-3 (fsc p3): Advancing slt for unseen channel and mission data across nasa apollo audio," *in Interspeech-2021*, 2021.

[20] C. Plapous, C. Marro, and P. Scalart, "Speech enhancement using harmonic regeneration," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–157.

[21] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura, and Y. Yamashita, "Single-channel speech enhancement with phase reconstruction based on phase distortion averaging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1559–1569, 2018.

[22] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *18th International Society for Music Information Retrieval Conference*, 2017, pp. 23–27.

[23] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *Interspeech*, 2020.

[24] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9458–9465.

[25] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural networks*, vol. 121, pp. 74–87, 2020.

[26] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.

[27] 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.

[28] 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*, vol. 19, no. 7, pp. 2125–2136, 2011.

[29] N. Mamun, M. S. Zilany, J. H. Hansen, and E. E. Davies-Venn, "An intrusive method for estimating speech intelligibility from noisy and distorted signals," *The Journal of the Acoustical Society of America*, vol. 150, no. 3, pp. 1762–1778, 2021.

[30] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.