



Funnel Deep Complex U-net for Phase-Aware Speech Enhancement

Yuhang Sun¹, Linju Yang¹, Huifeng Zhu¹, Jie Hao¹

¹Hertz Lab, OPPO Research Institute, Beijing, China

{sunyuhang, yanglinju, zhuhuifeng, haojie}@oppo.com

Abstract

The emergence of deep neural networks has made speech enhancement well developed. Most of the early models focused on estimating the magnitude of spectrum while ignoring the phase, this gives the evaluation result a certain upper limit. Some recent researches proposed deep complex network, which can handle complex inputs, and realize joint estimation of magnitude spectrum and phase spectrum by outputting real and imaginary parts respectively. The encoder-decoder structure in Deep Complex U-net (DCU) has been proven to be effective for complex-valued data. To further improve the performance, in this paper, we design a new network called Funnel Deep Complex U-net (FDCU), which could process magnitude information and phase information separately through one-encoder-two-decoders structure. Moreover, in order to achieve better training effect, we define negative stretched-SI-SNR as the loss function to avoid errors caused by the negative vector angle. Experimental results show that our FDCU model outperforms state-of-the-art approaches in all evaluation metrics.

Index Terms: speech enhancement, complex network, phase-aware, deep learning

1. Introduction

High quality speech signals will significantly increase the performance on voice communication, hearing aids and automatic speech recognition (ASR) [1]. In realistic environments, the speech signal quality will become poor due to effects of the noise. Speech enhancement is the task of separating the clean speech from noise given a noisy speech as an input, it thus plays an important role in speech signal processing and has been extensively studied in recent years. The performance of conventional methods like spectral subtraction [2], Wiener filter [3], Minimum Mean Square Error (MMSE) [4], Non-negative matrix factorization [5] is limited by scenarios. Then in computational auditory scene analysis (CASA) [6], the concept of time-frequency (T-F) mask was proposed, which inspired speech enhancement being formulated as a supervised learning problem. Recently, deep neural networks have greatly improved the speech quality and intelligibility, especially when non-stationary noise exists [7]. Most of the early DNN-based models used T-F representations of speech signals as the input features and focused on estimating the magnitude spectrogram regardless of estimation for phase spectrogram, then used the noisy phase to reconstruct the waveform. The reason is previous researches found that magnitude occupies a dominant position in resynthesized speech especially at high SNR [8][9]. Besides, the phase cannot be represented intuitively, so directly estimating the phase spectrum is not a easy work [10]. However, as the SNR decreases, the influence of phase becomes more and

more obvious [11]. Therefore, the accurate estimation of phase has become an important task in the field of speech enhancement [12][13].

DNN based speech enhancement researches are mainly divided into two directions: mapping-based method and masking-based method. Mapping-based method uses the DNN model to directly map the spectrum of the noisy speech to the spectrum of the clean speech, while masking-based method predicts masks by applying the noisy speech signals [14]. Ideal binary mask (IBM) [15] and ideal ratio mask (IRM) [16] and spectral magnitude mask (SMM) [17] only contributes to estimate the magnitude, then the emergence of phase-sensitive mask (PSM) [18] made it possible to incorporate phase information. Soon afterwards, complex ratio mask (CRM) [8] was developed to jointly estimate real and imaginary components. The complex-valued mask needs a structure that could process complex domain operations, and recently some studies have made significant progress in complex networks [19][20]. Deep Complex U-net (DCU) [21], applying the advantages of U-net [22] to complex model, is the more prominent one of these architectures and achieves remarkable performance in phase-aware speech enhancement.

On the other hand, in order to obtain more accurate phase information, researches on speech separation has proposed end-to-end methods [24][24][25], which can also be applied in speech enhancement. These approaches take a encoder-separator-decoder [26][27] structure while the encoder and decoder are similar to but different from short-time Fourier transform (STFT) and inverse short-time Fourier transform (ISTFT) [28]. The separator is between the encoder and decoder, usually adopts a structure that could capture the long-range temporal sequence information. These methods work directly on time domain signals, avoiding the degradation of speech quality caused by inaccurate phase estimation in the spectral domain.

Motivated by the previous works, we proposed a method of using complex neural networks for speech enhancement in the T-F domain. We designed a new architecture called Funnel Deep Complex U-net (FDCU), which jointly applies the masking-based method and the mapping-based method for the first time. In our model, we get estimated magnitude spectrogram by IRM and directly map the noisy complex spectrum to estimated phase spectrogram. Besides, inspired by Scale-Invariant-SNR (SISNR) [26], we proposed a new loss function called Stretched-Scale-Invariant-SNR (S-SISNR), which performs better than SISNR at low signal-to-noise-ratio (SNR). In our experiments, our proposed model FDCU outperforms the approaches for comparison significantly.

The rest of this paper is organized as follows. Section 2 introduces our model in detail. Section 3 presents the experimental setup and analyses the result. Section 4 concludes this paper.

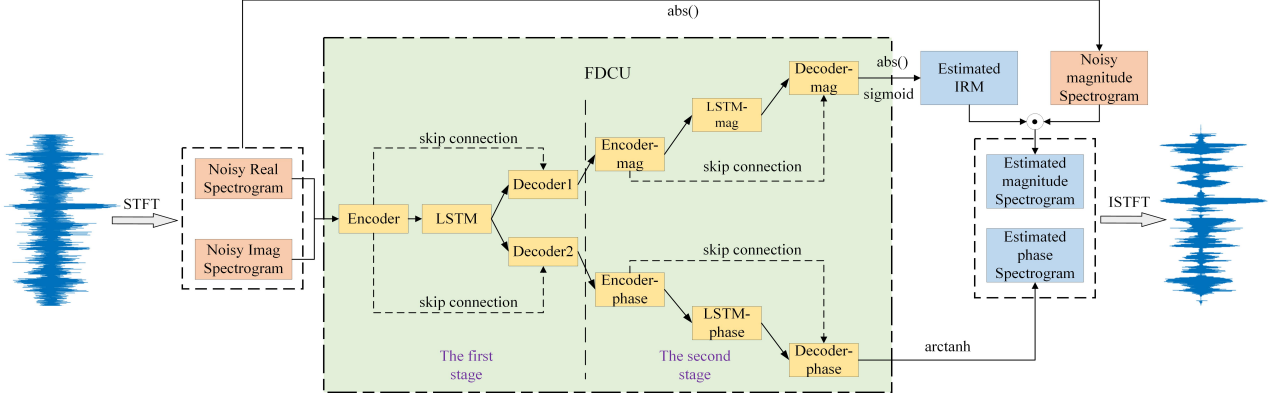


Figure 1: The specific structure of FDCU

2. The FDCU Model

Speech enhancement is to obtain as pure a speech signal as possible in the noisy speech signal. Given a clean speech signal x and a noise signal n , the noisy speech signal y can be expressed as follows:

$$y = x + n \quad (1)$$

where $\{y, x, n\} \in M^{N \times 1}$, N is the number of samples in the signal. Our goal is to estimate \hat{x} as close to x as possible when y is given.

2.1. Network architecture

DCU, first proposed in [21], is a simple complex encoder-decoder structure with skip connection. Then in [29], a complex LSTM layer is inserted between the encoder and the decoder, which is called DCCRN, could better model the temporal dependencies and achieve better performances. Unlike the method proposed in [14] that model the real and imaginary part separately, our model adopts a two-stage funnel structure, which includes three encoders, three LSTM parts and four decoders, and is shown in figure 1.

The input of our model is the complex STFT spectrogram. The first stage of the network is a one-encoder-two-decoders structure, and the outputs of the two decoders are both complex values, respectively model the magnitude spectrogram and the phase spectrogram. Previous studies have shown that DCU has the ability of extracting the target features[26], naturally, we use this structure to get more accurate magnitude features and phase features we need. Therefore in the second stage, we connect an encoder-decoder structure after each decoder of the first stage. As is illustrated, there are two paths in the second stage, the magnitude-path and the phase-path.

In magnitude-path, we use masking-based method to calculate the magnitude from the complex output, and get IRM matrix through the *sigmoid* function. Given the real part and the imaginary part in the output of magnitude-path are Mag_{real} and Mag_{imag} , the estimated IRM is as follows:

$$\hat{IRM} = \text{sigmoid}(\sqrt{Mag_{real}^2 + Mag_{imag}^2}) \quad (2)$$

Then the clean magnitude spectrogram can be estimated by element-wise production of \hat{IRM} and the noisy magnitude spectrogram.

In the phase-path, we directly map the complex spectrum of the noisy speech to the phase spectrum of the clean speech. If the complex output of phase-path are $Phase_{real}$ and $Phase_{imag}$, We can get the estimated phase matrix as:

$$\hat{P} = \text{arctanh}\left(\frac{Phase_{imag}}{Phase_{real}}\right) \quad (3)$$

Finally, we reconstruct the estimated speech signal through complex ISTFT. In order to make better use of temporal dependencies, complex LSTM is adopted in our architecture. Besides, skip-connection is also applied to achieve an efficient model.

2.2. Complex blocks of the network

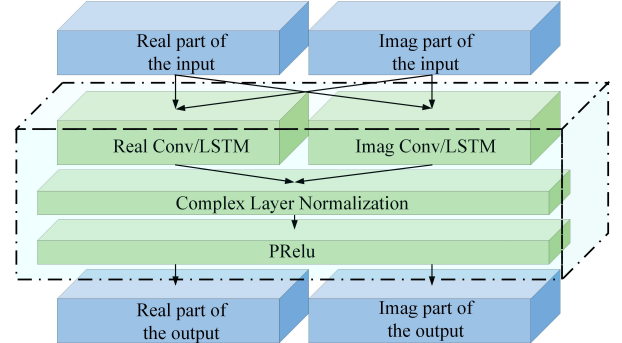


Figure 2: A complex block module. The kernel can be complex Convolution or LSTM layer

The structure of a complex block is shown in figure 2. A complex convolution block, used in each layer of the encoder and decoder, aims at processing deep features of the complex STFT spectrogram. Each block consists of complex Cov2d, complex layer normalization and PRelu[29]. Analogously, we also encapsulated LSTM layer in a block by substituting the core, which is the network structure within the dashed box in the figure 2, from complex Cov2d to complex LSTM.

All the encoders and decoders in our model contain multiple complex convolution blocks, while the three LSTM parts in figure 1 contain only one complex LSTM block.

2.3. Training target

The target of our network is to estimate the IRM and the phase spectrum. The estimated magnitude spectrum is calculated by

IRM, and then combined with the estimated phase spectrum to resynthesize the enhanced speech signal.

In typical complex networks, the estimator is trained to calculate the CRM and then minimize the difference between the estimated time-domain waveform and the target time-domain waveform. Alternatively, we adopt IRM and phase spectrum, so the loss in our model is calculated as:

$$loss = loss_f(\text{ISTFT}(Y \cdot \hat{IRM}, \hat{P}), s) \quad (4)$$

where Y denotes the noisy magnitude spectrogram, \hat{IRM} and \hat{P} denote the estimated IRM and phase spectrum, \cdot means element-wise production, ISTFT is inverse Short-time Fourier transform, and s represents the target speech signal. The loss function $loss_f$ is the proposed S-SISNR.

2.4. Loss function

SISNR is a commonly used time-domain loss function in recent researches by making itself negative [25]. The SISNR is defined as:

$$s_{target} = \langle \hat{s}, s \rangle / \|s\|^2 \quad (5)$$

$$e_{noise} = \hat{s} - s_{target} \quad (6)$$

$$\text{SISNR} = 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \quad (7)$$

where \hat{s} and s represent the estimated and target clean sources respectively.

Now we further derive the expression of SISNR according to the algebra operation rules:

$$\begin{aligned} \text{SISNR} &= 10 \log_{10} \frac{(\frac{|\hat{s}| |s| \cos(\theta)}{|s|})^2 |s|^2}{|\hat{s}|^2 + |\hat{s}|^2 \cos^2(\theta) - 2 \frac{|\hat{s}|}{|s|} \cos(\theta) |\hat{s}| |s| \cos(\theta)}}{\cos^2(\theta) |\hat{s}|^2} \\ &= 10 \log_{10} \frac{\cos^2(\theta) |\hat{s}|^2}{|\hat{s}|^2 - \cos^2(\theta) |\hat{s}|^2} \\ &= 10 \log_{10} \frac{\cos^2(\theta)}{1 - \cos^2(\theta)} = 10 \log_{10} \cot^2(\theta) \end{aligned} \quad (8)$$

where θ represent the angle between two vectors s and \hat{s} . Obviously, SISNR depends only on θ .

Figure 4-(a) shows the functional relationship between SISNR and θ . We found that SISNR has multiple extreme points in the range of $-\pi$ to π , and as θ gets closer to $-\pi$ or π , the SISNR-value becomes larger, so it is possible for the model to regard $-\pi$ or π as the optimal value in the optimization process, but the estimated-vector \hat{s} is actually very different from the target vector s . This is not conducive for us to restore more accurate phase information.

Therefore, we doubled the period of SISNR, and definite it as another loss function: Stretched SISNR (S-SISNR), which can be expressed as:

$$\text{S-SISNR} = 10 \log_{10} \cot^2(\frac{\theta}{2}) = 10 \log_{10} \frac{1 + \cos(\theta)}{1 - \cos(\theta)} \quad (9)$$

Since it is complicate to calculate the half angle, after the derivation of the trigonometric function, we find that it can be represented by $\cos(\theta)$, which is more easy to compute. As is shown in Figure 3-(b), S-SISNR has only one extreme point in the range of $-\pi$ to π , so the negative loss will be small only if θ is close to 0.

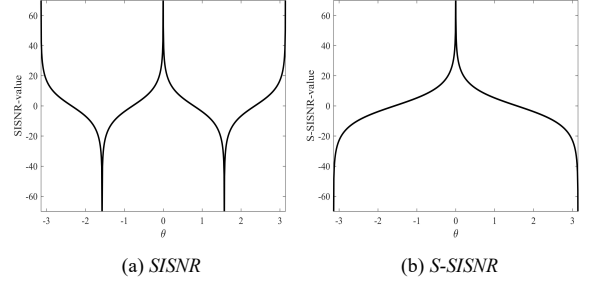


Figure 3: Extreme points of (a) SISNR and (b) S-SISNR

3. Experiments

3.1. Datasets

We evaluate our model on speech enhancement problem using TIMIT dataset [30], which contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. We choose NOISEX-92 [31] as the noise dataset. The 462 speakers in TIMIT training set are randomly divided into two parts, 400 speakers for the training set and the remaining 62 speakers for the validation set. The evaluation set contains all 168 speakers in TIMIT complete test set. The speech-noise mixtures are generated by randomly selecting utterances from the speech set and noise set at random SNR at -5 dB, 0dB and 5 dB. The total duration of the training set, the validation set and the evaluation set are about 28 hours, 8 hours and 6 hours respectively. The training set, validation set and evaluation set are mutually exclusive.

3.2. Training setup

All mixed speech signals used in the experiment are resampled to 16Khz, the input feature of our neural network is the complex spectrum obtained from STFT, while the Hanning window length and hop size is 1024 and 256 in STFT. All models are built based on pytorch[32] and are trained by using Adam[33] optimizer. Different learning rates are applied during training, starting at 0.0001 and decay 0.05 after each epoch. If the loss does not decrease for five epochs, the training stops. All the encoders in our model shared the same structure, the output channel of each layer in encoder is set to {32, 32, 64, 64, 64, 64, 64, 64, 64, 64}. The kernel sizes are {(7,1), (7,1), (7,5), (7,5), (7,5), (5,3), (5,3), (5,3), (5,3), (5,3)} and the stride is set to {(1,1), (1,1), (2,2), (2,1), (2,2), (2,1), (2,2), (2,1), (2,2), (2,1)}. The structures of all decoders are also the same. The input channels of decoders are {64, 128, 128, 128, 128, 128, 128, 64, 64} because of the skip connection, the hidden channels in all LSTM layers are 128.

3.3. Evaluation metrics

We evaluate our model with three commonly used objective metrics in speech enhancement, including SISNR, Perceptual Evaluation of Speech Quality (PESQ) [34] and Short time Objective Speech Intelligibility (STOI) [35].

PESQ: the most common metric to evaluate the speech quality. PESQ returns a score from -0.5 to 4.5 by comparing the enhanced speech and the clean speech.

STOI: a metric to predict intelligibility of speech signal. The value ranges from 0 to 1. The higher STOI means the cleaner enhanced speech signal.

Table 1: *SISNR, PESQ and STOI comparison for the different models. Higher score means better performance where bold text indicates the best performance for each metrics.*

	SISNR			PESQ			STOI			
	SNR	-5dB	0dB	5dB	-5dB	0dB	5dB	-5dB	0dB	5dB
noisy		-1.574	2.368	7.139	1.626	1.934	2.208	0.582	0.716	0.805
DCUnet		5.339	9.410	12.881	2.412	2.633	3.011	0.699	0.802	0.875
DCCRN		6.712	10.753	14.414	2.427	2.759	3.112	0.730	0.828	0.903
FDCUnet (SISNR)		8.056	11.874	15.536	2.616	3.000	3.336	0.743	0.851	0.917
FDCUnet (S-SISNR)		8.507	12.546	16.124	2.644	3.001	3.316	0.772	0.859	0.912

3.4. Experimental results and analysis

Table 1 shows the performance of our model as well as DCUnet and DCCRN. The specific structures and hyperparameters of these two models are shown in [21] and [29]. All of these models use the same dataset and metrics that we trained and evaluated in our model. In this result, our model increased the SISNR, PESQ and STOI significantly, respect to the noisy speech; What's more, we could find that our model achieves better performance than the other models. According to our analysis, this is because we applied the FDCU structure by combining masking-based method and mapping-based method, which can make our model more effective in speech enhancement tasks. The proposed loss function S-SISNR is also compared with the original loss function SISNR in our experiment. The S-SISNR outperforms to SISNR with respect to most metric scores, especially in the case of low SNR. As is shown in table 1, when SNR is at -5 dB, three metrics are increased by 5.6%, 1.1% and 3.9% respectively. However, this advantage gradually decreases as the SNR increases, we found that SISNR provides better PESQ and STOI when SNR is at 5dB. This proves that S-SISNR is better for the model processing low SNR noisy speech signals. Figure 3 can explain this result well, the higher SNR means θ is more likely to be distributed around 0, when $|\theta| < \pi/2$, the model is easier to converge when SISNR is applied because of the larger slope. When SNR gets lower, $|\theta|$ has more probability of being distributed between $\pi/2$ to π , under this circumstance, S-SISNR can optimize the model better than SISNR.

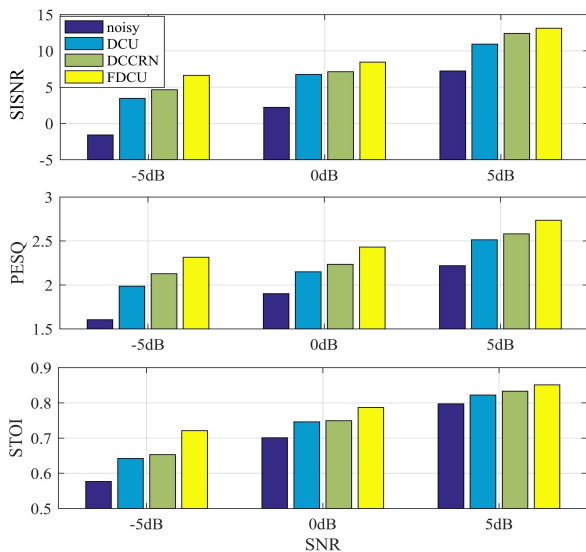


Figure 4: *The results of (a) SISNR, (b) PESQ and (c) STOI on the hu_nonspeech evaluation set.*

To verify the generalization performance of our model on different noise datasets, we replaced the noise set NOISEX-92 with hu_nonspeech [36], and made it a new evaluation set. Figure 4 shows the metric scores for our model (optimizing with S-SISNR) on different SNR, with respect to the others, it's shown that our model gives the best result. In particular, the advantages of our model can be better reflected at low SNR. This result indicates that FDCU architecture has the modeling capabilities for estimating clean speech and could generalize well to unseen noises.

Figure 5 shows the Magnitude spectrum of an example utterance processed by different models. The mixture is at -5dB SNR and the noise is Fighter Jets-f16. As is illustrated, the spectrogram processed by FDCU has less residual noises. Particularly, FDCU suppressed high frequency noise better. Besides, it can be seen from the difference between figure 5-(e) and figure 5-(f) that S-SISNR has a better effect on suppressing mid-frequency noise.

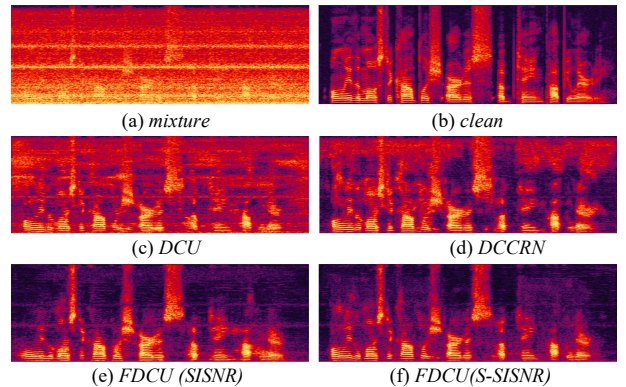


Figure 5: *Magnitude spectrum of an example utterance processed by different models. (a) the mixture signal, (b) the clean signal, (c) processed by DCU, (d) processed by DCCRN, (e) processed by FDCU, the loss function is basing on SISNR, (f) processed by FDCU, the loss function is basing on S-SISNR.*

4. Conclusions

In this work, we proposed a Funnel Deep Complex U-net architecture for speech enhancement. Our model incorporated the masking-based method and the mapping-based method to estimate clean waves from noisy. Besides, we also designed a new loss function S-SISNR, which can further improve the performance of the model especially at low SNR. By combining these strategies, our model got excellent scores in the evaluation metrics.

In the future, we try applying the FDCU to the front end of Automatic Speech Recognition (ASR), in order to improve the effects of ASR in noisy scenarios. And the real time capability, the overall size and hardware efficiency of the model will be evaluated on other larger datasets.

5. References

- [1] Abdulbaqi, Jalal, et al. "Residual Recurrent Neural Network for Speech Enhancement," *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6659-6663, 2020.
- [2] Berouti, M., Schwartz, R., & Makhoul, J. "Enhancement of speech corrupted by acoustic noise," *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, vol. 4, pp. 208-211, 1979.
- [3] Bingyin, et al. "Wiener filtering based speech enhancement with Weighted Denoising Auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13-29, 2014.
- [4] Hu, Yi , and P. C. Loizou . "Speech enhancement based on wavelet thresholding the multitaper spectrum," *Transactions on Speech and Audio Processing*. IEEE, vol. 12, no. 1, pp. 59-67, 2004.
- [5] Wilson, Kevin W., et al. "Speech denoising using nonnegative matrix factorization with priors," *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 4029-4032, 2008.
- [6] Wang, DeLiang, and Guy J. Brown. "Computational auditory scene analysis: Principles, algorithms, and applications," *Wiley-IEEE press*, 2006.
- [7] Sun, Lei, et al. "Multiple-target deep learning for LSTM-RNN based speech enhancement," *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, pp. 136-140, 2017.
- [8] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 483-492, 2016.
- [9] D. L. 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.
- [10] Pejman Mowlae, Josef Kulmer, Johannes Stahl, Florian Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*. John Wiley and Sons, Dec. 2016
- [11] K. Paliwal, K. W'ojcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465-494, 2011.
- [12] Timo Gerkmann, Martin Krawczyk-Becker, Jonathan Le Roux, "Phase Processing for Single Channel Speech Enhancement: History and Recent Advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55-66, Mar. 2015.
- [13] Mowlae P, Saeidi R, Stylianou Y. "Advances in phase-aware signal processing in speech communication," *speech communication*, vol. 81, pp. 1-29, 2016.
- [14] Tan, Ke, and DeLiang Wang. "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6865-6869, 2019.
- [15] Wang, DeLiang. "On ideal binary mask as the computational goal of auditory scene analysis." *Speech separation by humans and machines*. Springer, Boston, MA, pp. 181-197, 2005.
- [16] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 7092-7096, 2013.
- [17] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849-1858, 2014.
- [18] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phasesensitive and recognition-boosted speech separation using deep recurrent neural networks," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708-712.
- [19] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," arXiv preprint arXiv:1705.09792, 2017.
- [20] Pandey, Ashutosh, and DeLiang Wang. "Exploring deep complex networks for complex spectrogram enhancement." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6885-6889, 2019.
- [21] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," arXiv preprint arXiv:1903.03107, 2019.
- [22] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, pp. 234-241, 2015.
- [23] Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499, 2016.
- [24] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M.Hasegawa-Johnson, "Speech enhancement using Bayesian wavenet," in *Proc. Interspeech, Stockholm, Sweden*, 2017, pp. 2013-2017.
- [25] Rethage, Dario, Jordi Pons, and Xavier Serra. "A wavenet for speech denoising." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5069-5073, 2018.
- [26] Luo, Yi , and N. Mesgarani . "Conv-TasNet: Surpassing Ideal Time - Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 1256-1266, 2019.
- [27] Luo, Yi, Zhuo Chen, and Takuya Yoshioka. "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 46-50, 2020.
- [28] Kishore, Vinith, Nitya Tiwari, and Periyasamy Paramasivam. "Improved speech enhancement using TCN with multiple encoder-decoder layers." in *Proc. Interspeech, Shanghai, China*, 2020 , pp. 4531-4535.
- [29] Hu, Yanxin, et al. "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement." arXiv preprint arXiv:2008.00264, 2020.
- [30] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S.Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical reportn*, vol. 93, 1993.
- [31] Varga, Andrew, and Herman JM Steeneken. "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems." *Speech communication*, vol. 12, no. 3, pp. 247-251, 1993.
- [32] Paszke A, Gross S, Massa F, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". *Advances in Neural Information Processing Systems*, vol. 32, 8026-8037, 2019.
- [33] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.
- [34] Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, Rec. ITU-T P.862, International Telecommunications Union, Geneva, Switzerland, 2001.
- [35] 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. Processing*, vol. 19, no. 7, pp. 2125-2136, 2011.
- [36] Hu G. and Wang D.L, "A tandem algorithm for pitch estimation and voiced speech segregation". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067-2079, 2010.