



# Single-channel Speech Enhancement Using Graph Fourier Transform

Chenhui Zhang, Xiang Pan

Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China

ch.zhang@zju.edu.cn, panxiang@zju.edu.cn

## Abstract

This paper presents combination of Graph Fourier Transform (GFT) and U-net, proposes a deep neural network (DNN) named G-UNet for single channel speech enhancement. GFT is carried out over speech data for creating inputs of U-net. The GFT outputs are combined with the mask estimated by U-net in time-graph (T-G) domain to reconstruct enhanced speech in time domain by Inverse GFT. The G-UNet outperforms the combination of Short time Fourier Transform (STFT) and magnitude estimation U-net in improving speech quality and de-reverberation, and outperforms the combination of STFT and complex U-net in improving speech quality in some cases, which is validated by testing on LibriSpeech and NOISEX92 dataset.

**Index Terms:** single-channel speech enhancement, graph fourier transform, deep learning

## 1. Introduction

Speech enhancement is increasingly becoming a technology worthy of attention recently, dedicated to improving both the accuracy of automatic speech recognition (ASR) modules and human perception. The goal of speech enhancement is to improve the intelligibility and perceptual quality of speech signals degraded by reverberation and noise. Traditional DNN based noisy speech enhancement frameworks are mainly done in time-frequency (T-F) domain or directly time [1, 2, 3] domain.

The T-F diagram estimation methods can be roughly divided into two categories. One only focuses on the magnitude spectrum of the signal while ignoring the phase information when performing T-F diagram estimation [4, 5, 6, 7]. The phase of the mixed speech is reused when recovering the original signal, which leads to a certain degree of performance degradation. The other introduces phase estimation on the basis of magnitude spectrum estimation, and uses the estimated magnitude and phase spectrum to restore the original speech signal [8, 9, 10, 11, 12]. However, the computation of complex DNN model is huge. As described in DCUNET[10] and DCCRN[12], keeping the DNN structure unchanged, extending convolution/recurrent layers from real operations to complex operations will result in twice as many parameters and four times as many calculations.

To reduce the complexity and further improve the performance, some other signal transformation, such as discrete cosine transform[13, 14], was introduced to replace traditional Fourier Transform. And some researchers attempted to introduce methods of graph signal processing into speech enhancement. Such as graph neural networks which exploits the spatial correlations in the multi-channel speech enhancement [15], and improved graph Wiener filtering for speech enhancement [16].

In this article, a signal front-end transform based on Graph Fourier Transform (GFT) is proposed. It treats the original sig-

nal as a segmented graph signal after framing, and uses graph theory to capture the characteristic information of the signal over a period of time. GFT can project the original signal to time-graph (T-G) domain, and generate a T-G diagram similar to T-F diagram. Then GFT is combined with existing mask-based estimation method to build G-UNet for single-channel speech enhancement based on U-net [17] deep learning network. In addition, M-UNet that pays attention to magnitude only and C-UNet [10] of the complex convolution are implied for comparison, proving the feasibility of GFT in the field of single-channel speech enhancement.

The contributions of this paper are as follows: (i) A signal front-end transform based on GFT is utilized to extract short-term signal features with graph structure. (ii) The mask estimation based on U-net is combined with GFT as G-UNet for single-channel speech enhancement, which can realize the estimation of clean speech under environment with noise and reverberation by using a real-valued DNN. It also proves that speech enhancement framework need not to be restricted in working in time or time-frequency domain traditionally (iii) This paper proves that estimation error in G-UNet resulting in phase error occurs while training G-UNet with SI-SNR as loss function, and speech quality is further improved when signs corrected. (iv) This paper aims to remind researchers to jump out of the limitations of the time/time-frequency domain, and explore whether using transformations other than STFT as front-end feasibility for DNN based speech enhancement.

## 2. Methods

### 2.1. Graph Fourier Transform

Graph Fourier Transform [18] uses the characteristic matrix of graph signal to transform the signal from time domain to graph domain. Consider that signal  $\mathbf{s}$  can be represented by graph structure, with the Adjacency matrix  $\mathbf{A}$  and Degree matrix  $\mathbf{D}$ .  $\mathbf{A}$  describes the distance between each node on the graph, and  $\mathbf{D}$  describes the number that each node connects with others on the graph. Eigenvalue decomposition of Laplacian matrix  $\mathbf{L} = \mathbf{D} - \mathbf{A}$  could be denoted as  $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ , where eigenvectors in  $\mathbf{U}$  are sorted in ascending order according to the size of the eigenvalues in  $\mathbf{\Lambda}$ . Transformed signal  $\mathbf{s}_g$  can be expressed as  $\mathbf{s}_g = \mathbf{U}^T \mathbf{s}$ . And Inverse Graph Fourier Transform (IGFT) reconstructs time domain signal  $\mathbf{s}$  from  $\mathbf{s}_g$ , expressed as  $\mathbf{s} = \mathbf{U} \mathbf{s}_g$ .

GFT takes  $\mathbf{u}_i$  as the base, and transforms the signal from the original time domain to the graph domain, where  $\mathbf{u}_i$  represents the eigenvector of the Laplacian matrix  $\mathbf{U} = [\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{n-1}]$  of the graph signal. The transformed signal is a linear combination of  $\mathbf{u}_i$ .

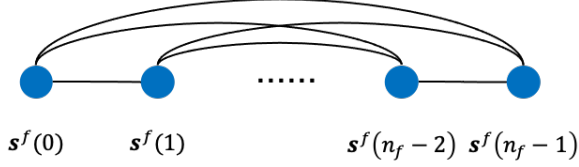


Figure 1: Structure of considered graph signal

## 2.2. GFT Used in Speech Signal

For a long sequence of speech signal, it can be expressed as  $\mathbf{s} = [s_0, s_1, \dots, s_{n-1}]_{n \times 1}^T$ , where  $n$  represents the total number of samples. Speech signal first needs to be divided into frames, and a sequence of speech signal could be represented as  $\tilde{\mathbf{s}} = [s_0^f, s_1^f, \dots, s_{m-1}^f]_{n_f \times m}$ , where  $s_i^f = [s_h, s_{h+1}, \dots, s_{h+n_f-1}]_{n_f \times 1}^T$  represents the  $i$ -th segment of the framed signal,  $n_f$  represents the number of samples of a single frame signal,  $h = i \times n_s$ ,  $n_s$  represents the number of points shifted by the frame window. Each single frame speech signal  $s^f$  could be considered stationary and regarded as a graph structure as shown in Fig. 1.

Each sampling point  $s^f(i)$  in  $s^f$  is regarded as a node on the graph, connected to other nodes. Adjacency matrix  $\mathbf{A}$  is used to represent the interval between sampling points in time domain and degree matrix  $\mathbf{D}$  is used to indicate the number of sampling points. Therefore,  $\mathbf{A}$  and  $\mathbf{D}$  of the graph could be represented as  $\mathbf{A} = [a_{ij}]_{n_f \times n_f}$ ,  $a_{ij} = |i - j|$ ,  $\mathbf{D} = \text{diag}(n_f - 1)_{n_f \times n_f}$ , and the Laplacian matrix  $\mathbf{L} = \mathbf{D} - \mathbf{A}$  is an orthogonal matrix. Laplacian matrix  $\mathbf{L}$  is subjected to eigenvalue decomposition as  $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ .

Let  $\mathcal{G} : \mathbb{R}^n \rightarrow \mathbb{R}^{n_f \times m}$  be the Short-time GFT proposed where  $n_f$  and  $m$  are the number of graph and time bins. T-G diagram of the original signal  $\mathbf{s}$  could be represented as  $\mathbf{S}$ .

$$\mathbf{S} = \mathcal{G}(\mathbf{s}) = \mathbf{U}^T \tilde{\mathbf{s}}. \quad (1)$$

And Inverse Short-time GFT is denoted as  $\mathcal{G}^\dagger : \mathbb{R}^{n_f \times m} \rightarrow \mathbb{R}^n$ , used to reconstruct  $\mathbf{S}$  into time domain signal  $\mathbf{s}$ .

$$\mathbf{s} = \mathcal{G}^\dagger(\mathbf{S}) = \text{Gri}f(\tilde{\mathbf{s}}) = \text{Gri}f(\mathbf{U}\mathbf{S}). \quad (2)$$

$\text{Gri}f(\cdot)$  denotes algorithm proposed in [19] used to reconstruct the framed signal into time domain signal, using rectangular window.

## 2.3. Speech Enhancement Based on GFT and Deep Learning

Environment considered in this paper contains reverberation and noise. Original speech to be enhanced can be expressed as  $\mathbf{x} = \mathbf{s} * \mathbf{r} + \mathbf{n}$ , where  $*$  stands for the convolution operator. Clean speech  $\mathbf{s}$  is affected by Room Impulse Response (RIR) function  $\mathbf{r}$  and additive noise  $\mathbf{n}$ . The goal of speech enhancement and separation is to recover  $\mathbf{s}$  from  $\mathbf{x}$ .

$\mathcal{G}$  is combined with deep learning to explore its feasibility in the field of speech enhancement. Speech enhancement network called G-Unet is proposed, which is shown in Fig. 2. G-Unet combines  $\mathcal{G}$  with the classic U-net [17] architecture. And inspired by DCCRN[12], a 2-layer Long Short-Term Memory (LSTM) layer is added to capture the sequential information

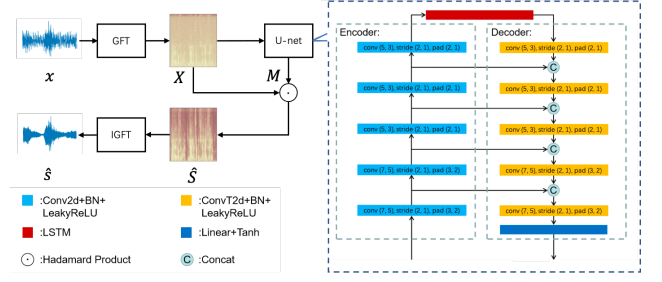


Figure 2: Illustration of proposed speech enhancement network G-Unet

among each speech frame between encoder output and decoder input. Note that parameters of DNN used here are all real-valued.

The original mixed speech signal  $\mathbf{x}$  is first transformed into T-G domain as  $\mathbf{X} = \mathcal{G}(\mathbf{x})$ , then the mask of  $\mathbf{X}$  is estimated by DNN based mask estimator, denoted as  $\mathbf{M}$ . And  $\hat{\mathbf{s}} = \mathcal{G}^\dagger(\mathbf{M} \odot \mathbf{X})$  is the enhanced speech signal.

Similar to ideal ratio mask (IRM)[20] in T-F domain, G-Unet uses the signal approximation method to estimate the real-valued IRM in T-G domain, denoted as IGRM. IGRM can be defined as:

$$\text{IGRM}_{t,g} = \frac{S_{t,g}}{X_{t,g}} \quad (3)$$

where  $S_{t,g}$  and  $X_{t,g}$  denotes transformed GFT diagram of clean and noisy signal, at  $t$ -th time frame and  $g$ -th graph index.

During training, Scale-Invariant Source-to-Noise Ratio (Si-SNR) [21] in time domain is chosen as the loss function [1] during training, which is defined as

$$\begin{cases} s_{\text{target}} = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle}{\|\mathbf{s}\|^2} \cdot \mathbf{s} \\ \text{Si} - \text{SNR} = 10 \times \log_{10} \left( \frac{\|s_{\text{target}}\|^2}{\|\hat{\mathbf{s}} - s_{\text{target}}\|^2} \right) \end{cases} \quad (4)$$

where  $\hat{\mathbf{s}}$  and  $\mathbf{s}$  denotes the estimated and clean speech,  $\|\cdot\|$  is  $\ell_2$  norm.

## 3. Experiments

### 3.1. Datasets

Clean speech dataset selected in this paper is LibriSpeech [22], which includes 251 speakers for training and 40 speakers for testing, and noise datasets used are NOISEX92 [23] and DEMAND [24]. Image source method [25] is utilized to generate the RIR functions for creating reverberant speech signals. Five rooms have different sizes except for a fixed height of 3 m, the RIR functions corresponding to room R1 R3 and R5 are utilized for training while the RIRs of room R2 and R5 respectively for validation and testing.

The reverberant speech signals are combined with noise to create the training set with SNR from -5 dB to 10 dB. All training data are sampled at 16 kHz. The number of samples in generated dataset is 25537 for training and 3002 for validation.

Different from the training set, the test set only contains noise from NOISEX92. Besides the speech from Librispeech, the test set with 1916 samples includes speech fragments of Chinese speakers. Note that speakers in the test set didn't appear in the training set.

### 3.2. Training Details

G-Net shown in Fig 2 is implemented in Pytorch and trained on training set for 30 epochs, with Adam optimizer using learning rate of 0.0008, and batch size of 8. For feature extraction,  $n_f$  was set to 512 and  $n_s$  was set to 256. channel size of each layer in encoder/decoder is kept same as DCUNET-10 in [10].

Besides, two existing speech enhancement approaches which use Short time Fourier Transform (STFT)  $\mathcal{F} : \mathbb{R}^n \rightarrow \mathbb{C}^{f \times m}$  as front-end transform are also implemented in this paper as baselines for comparison.

- M-Net: an U-Net structured model only estimate magnitude spectrum, and use noisy phase and estimated magnitude to reconstruct the signal. Similarly, channel size of each layer in encoder/decoder is kept same as DCUNET-10 in [10]. And a 2-layer real valued LSTM is also added between encoder and decoder.
- C-Net [10]: an U-Net structured model incorporating complex-valued building blocks to estimate complex-valued spectrogram, complex LSTM layer[12] is added between the encoder and decoder of U-Net. And C-Net channel of each layer in encoder/decoder is twice the channel size of G-Net/M-Net. A 2-layer complex valued LSTM, similar to DCCRN[12], is added between encoder and decoder. Compared to real-valued LSTM, the complex convolution and LSTM layer brings twice amount of parameters and four times amount of computation.

$\mathcal{F}$  is calculated using Hann window, the frame length is set to 64 ms with 32 ms overlap and FFT size is 1024. Loss function used during training of M-Net and C-Net is also SI-SNR represented in (4).

### 3.3. Evaluation Metrics

In this paper, SI-SNR in (4), PESQ, STOI [26], SRMR [27] are selected as performance indicators for evaluating the effect of speech enhancement.

- PESQ: Perceptual evaluation of speech quality, with a rating range of  $-0.5 \sim 4.5$ .
- STOI: Short-time objective intelligibility measure, with a rating range of  $0.0 \sim 1.0$ .
- SRMR: speech to reverberation modulation energy ratio.
- SI-SNR: Scale invariant source-to-noise ratio.

For all indicators, the higher the value corresponds the better the effect of speech enhancement.

### 3.4. Evaluations

G-Net is first evaluated on a single test corpus  $\mathbf{x}$  (with pink noise, SNR=0 dB),  $\mathbf{s}$  and  $\hat{\mathbf{s}}$  denotes the clean and estimated speech.  $\mathcal{G}(\mathbf{x}), \mathcal{G}(\mathbf{s}), \mathcal{G}(\hat{\mathbf{s}})$  are shown in Fig. 3. And T-F analysis is performed on  $\mathbf{x}, \mathbf{s}, \hat{\mathbf{s}}$ , shown in Fig 4.

It can be seen that G-Net could realize the estimation of clean T-G diagram(Fig. 3), meanwhile magnitude spectrum of clean speech can be restored through G-Net(Fig. 4). What's more interesting, shapes of T-G diagram and related spectrum, even positions of over suppression/noise residue (marked with red/white boxes) are extremely similar. In addition, the shortcomings of over-suppression and noise residue operated in T-G domain will be mapped to the T-F spectrum (Fig. 3.(c)/Fig. 4.(c)). Reduction and noise residue in T-F diagram are also labeled by red and white boxes in Fig 4.(c), which

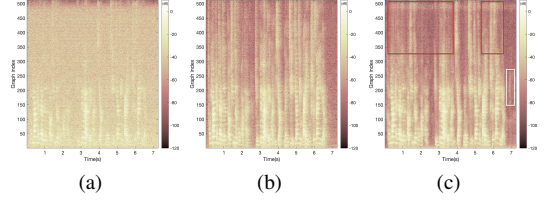


Figure 3: T-G diagrams of different speech. (a):mixed speech. (b):clean speech. (c):speech enhanced by G-Net.

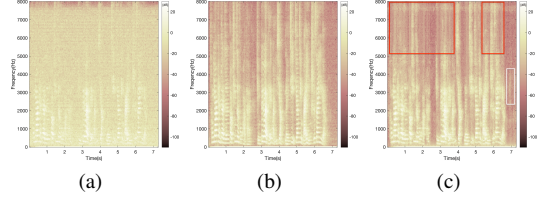


Figure 4: Spectrogram of different speech. (a):mixed speech. (b):clean speech. (c):speech enhanced by G-Net.

are almost at the same position of boxes in Fig. 3.(c). Proving that T-G bins and graph index in T-G domain are indeed related to T-F bins and frequency in T-F domain.

And the phase spectrum of  $\hat{\mathbf{s}}$  has changed from  $\mathbf{x}$ , but has a large deviation from  $\mathbf{s}$ , which might limits the performance of G-Net. The reason for deviation in the phase spectrum might be the deviation of  $sign(\mathbf{S})$  and  $sign(\hat{\mathbf{S}})$ .  $\mathbf{S} = \mathcal{G}(\mathbf{s}), \hat{\mathbf{S}} = \mathcal{G}(\hat{\mathbf{s}})$ , and  $sign(\mathbf{S})$  means the signs spectrum of  $\mathbf{S}$ , expressed by (5).

$$[sign(\mathbf{S})]_{ij} = [sgn(\mathbf{S}_{ij})], sgn(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \quad (5)$$

Without calculating loss in graph domain, G-Net pays more attention to the magnitude information that has a greater impact on hearing, while ignoring the phase information, that is, signs of the T-G diagram.  $sign(\mathbf{S})$  is introduced to get revised T-G diagram  $\hat{\mathbf{S}}_r$ . Note that the estimation of  $sign(\mathbf{S})$  is not implemented in this paper,  $sign(\mathbf{S})$  is only used to prove that estimation error of  $sign(\hat{\mathbf{S}})$  leads to phase error of  $\hat{\mathbf{s}}$ . The revising calculation could be equivalent to (6).  $abs(\cdot)$  computes the absolute value of each element of its input matrix.

$$\hat{\mathbf{S}}_r = abs(\hat{\mathbf{S}}) \odot sign(\mathbf{S}), \hat{\mathbf{s}}_r = \mathcal{G}^\dagger(\hat{\mathbf{S}}_r) \quad (6)$$

Phase error of  $\hat{\mathbf{s}}$  is denoted as  $|\Delta\Phi|$ ,  $|\Delta\Phi| \in [0, 2\pi]$ .  $|\Delta\Phi| = mean\{abs(\theta\{\mathcal{F}(\hat{\mathbf{s}})\} - \theta\{\mathcal{F}(\mathbf{s})\})\}$ ,  $\theta\{\cdot\}$  computes argument of each element in its input matrix,  $mean\{\cdot\}$  computes the mean of all elements in its input matrix.

Phase error and evaluation metrics of  $\mathbf{x}, \hat{\mathbf{s}}, \hat{\mathbf{s}}_r$  are shown in Table 1, G-Net-sr represents G-Net with signs of output T-G diagram revised. **Bold** in tables denotes the best scheme. It can be seen that phase of  $\mathbf{s}_r$  is corrected after signs were revised, which is manifested in the reduction of  $|\Delta\Phi|$  and the improvement of all metrics.

Then models are evaluated on the test set under different noise environments. In order to demonstrate the importance of

Table 1: Phase error and evaluation metrics of single test corpus and its enhanced results

	$ \Delta\Phi $	SI-SNR	SRMR	PESQ	STOI
Noisy	2.11	11.06	3.40	2.81	0.77
G-Unet	2.08	14.65	4.33	3.09	0.86
G-Unet-sr	<b>1.37</b>	<b>15.64</b>	<b>4.36</b>	<b>3.12</b>	<b>0.88</b>

the  $sign(\hat{\mathbf{S}})$ , G-Unet-sr was evaluated at the same time. Cuz the main purpose of our paper is to demonstrate the feasibility of using GFT as front-end transformation for DNN-based speech enhancement and out that speech enhancement could not be limited in time/T-F domain rather than state-of-the-art (SOTA) DNN model, it is necessary to control all same factors (including the structure of DNN models used) except for front-end transformation.

Table 2 shows the average performance of the three models in all noise environments (13 types in total) at different SNR. It can be seen that G-Unet always performs better than M-Unet on all metrics. And compared with C-Unet, G-Unet can approach or even exceed C-Unet in terms of SI-SNR, PESQ and STOI, which is more obvious for high SNR case. However, in terms of the performance of eliminating reverberation, G-Unet is weaker than C-Unet, showing a lower SRMR. The last row in Table 2 indicates that the application of  $sign(\mathbf{S})$  can improve all indicators under all SNRs, and SI-SNR, PESQ are always higher than C-Unet's.

Table 2: Average objective evaluation results

SNR	SI-SNR		SRMR		PESQ		STOI	
	0	5	0	5	0	5	0	5
Noisy	1.02	10.84	3.01	5.78	1.50	2.53	0.71	0.88
M-Unet	7.06	9.36	7.76	7.29	2.25	2.66	0.84	0.92
C-Unet	10.56	16.01	<b>8.51</b>	<b>7.96</b>	2.51	3.09	<b>0.86</b>	<b>0.95</b>
G-Unet	9.91	16.64	7.96	7.80	2.50	3.14	0.82	0.93
G-Unet-sr	<b>11.15</b>	<b>17.69</b>	8.14	7.87	<b>2.60</b>	<b>3.21</b>	0.84	0.94

In addition, G-Unet performs better under several noises, such as white, pink, volvo, machinegun, factory, which are more stationary than others. Table 3 and 4 show objective evaluation results of the three models under white noise and babble noise with different SNRs.

For the white Gaussina noise, G-Unet performs better than M-Unet on all metrics. Compared with metrics of C-Unet, SI-SNR and PESQ of G-Unet are always higher, and STOI is 0.02 lower at SNR=0 dB and equal at SNR=5 dB. Similarly, the application of  $sign(\mathbf{S})$  can improve all indicators under all SNRs.

Table 3: Objective evaluation results (White noise)

SNR	SI-SNR		SRMR		PESQ		STOI	
	0	5	0	5	0	5	0	5
Noisy	8.89	18.89	5.11	7.08	2.27	3.29	0.87	0.96
M-Unet	9.28	10.27	7.44	7.13	2.40	2.92	0.91	0.95
C-Unet	15.29	20.05	<b>8.08</b>	<b>7.76</b>	2.77	3.46	<b>0.94</b>	<b>0.98</b>
G-Unet	16.11	23.10	7.90	7.70	2.86	3.63	0.92	<b>0.98</b>
G-Unet-sr	<b>17.29</b>	<b>24.21</b>	7.99	7.73	<b>2.90</b>	<b>3.68</b>	0.93	<b>0.98</b>

For other types of noise, taking babble as background noise for example, the performance of G-Unet is weaker than C-Unet's. However, the gap between G-Unet and C-Unet decreases when SNR higher. Compared with M-Unet, G-Unet's SI-SNR, SRMR, and PESQ are always higher, while STOI is

0.06 lower at SNR=0 dB and 0.01 lower at SNR=5 dB. Similarly, the application of  $sign(\mathbf{S})$  can improve all indicators, making indicators higher than C-unet's in some cases.

Table 4: Objective evaluation results (Babble noise)

SNR	SI-SNR		SRMR		PESQ		STOI	
	0	5	0	5	0	5	0	5
Noisy	-1.88	8.12	2.24	5.38	0.99	2.03	0.59	0.82
M-Unet	5.73	8.87	7.86	7.34	2.06	2.51	0.80	0.91
C-Unet	<b>8.04</b>	14.17	<b>8.54</b>	<b>8.04</b>	2.22	2.88	<b>0.82</b>	<b>0.93</b>
G-Unet	6.12	13.40	7.94	7.86	2.10	2.80	0.74	0.90
G-Unet-sr	7.60	<b>14.68</b>	8.26	8.00	<b>2.24</b>	<b>2.91</b>	0.77	0.92

According to the evaluation results, we can make inferences as follow: (i) In terms of SI-SNR and PESQ, metrics about improving speech quality, G-Unet can always achieve higher values than M-Unet and higher than C-Unet under environments with several noises. (ii) G-Unet achieves a higher value than M-Unet but lower than C-Unet in terms of SRMR and STOI. (iii)  $sign(\hat{\mathbf{S}})$  has a deviation while taking SI-SNR in time domain as loss function, resulting in phase error between  $\hat{\mathbf{s}}$  and  $\mathbf{s}$ . Using  $sign(\mathbf{S})$  for revision can reduce the average phase error and improve all metrics of speech.

## 4. Conclusion

In this paper, a signal front-end transform based on GFT is proposed to combine with U-net to realize a mask-based single-channel speech enhancement network G-Unet. G-Unet outperforms M-Unet on improving speech quality and de-reverberation, and outperforms C-Unet on improving speech quality only when dealing with noise more stationary, due to extracting more features from the T-G domain over the T-F domain in these cases. When the sign estimation error in G-Unet resulting in the phase error has been corrected, the phase error can be reduced and speech quality can be further improved. The effectiveness of G-Unet for speech enhancement is verified on LibriSpeech in terms of SI-SNR, PESQ, SRMR and STOI.

## 5. Future Works

Finally, here list some points that are worth exploring:

- We didn't realize the estimation of  $sign(\mathbf{S})$ , but used clean speech's information to prove the effect of  $sign(\mathbf{S})$  in reducing phase error in Sec. 4. In future, we plan to design loss function to focus on signs of the T-G diagram, or use a dual-stream network to realize the estimation of signs of T-G diagram at the same time.
- This paper only focus on human perception-related objective metrics in Sec. 4, while not considering human perception itself. Post-filtering methods in T-F domain[28, 29] can be introduced to remove unnatural residual noise components.
- More SOTA models[11, 29, 30] working in T-F domain can be learnt and introduced in future.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 41776108 and No. 61571397).

## 6. References

- [1] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE*

- ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [2] A. Défossez, G. Synnaeve, and Y. Adi, “Real Time Speech Enhancement in the Waveform Domain,” in *Proc. Interspeech*, 2020, pp. 3291–3295.
  - [3] 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.
  - [4] K. Tan and D. Wang, “A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement,” in *Proc. Interspeech*, 2018, pp. 3229–3233.
  - [5] N. Takahashi, N. Goswami, and Y. Mitsufuji, “MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation,” *CoRR*, vol. abs/1805.02410, 2018.
  - [6] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, “Speech Enhancement Using Self-Adaptation and Multi-Head Self-Attention,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2020, pp. 181–185.
  - [7] A. A. Nair and K. Koishida, “Cascaded time + time-frequency unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2021, pp. 7153–7157.
  - [8] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2015, pp. 708–712.
  - [9] D. S. Williamson, Y. Wang, and D. Wang, “Complex Ratio Masking for Monaural Speech Separation,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 3, pp. 483–492, 2016.
  - [10] H. Choi, J. Kim, J. Huh, A. Kim, J. Ha, and K. Lee, “Phase-Aware Speech Enhancement with Deep Complex U-Net,” in *7th International Conference on Learning Representations, ICLR*, 2019.
  - [11] D. Yin, C. Luo, Z. Xiong, and W. Zeng, “PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network,” in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, 2020, pp. 9458–9465.
  - [12] 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,” in *Proc. Interspeech*, 2020, pp. 2472–2476.
  - [13] C. Geng and L. Wang, “End-to-end speech enhancement based on discrete cosine transform,” *CoRR*, vol. abs/1910.07840, 2019.
  - [14] Q. Li, F. Gao, H. Guan, and K. Ma, “Real-time monaural speech enhancement with short-time discrete cosine transform,” *CoRR*, vol. abs/2102.04629, 2021.
  - [15] P. Tzirakis, A. Kumar, and J. Donley, “Multi-channel speech enhancement using graph neural networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2021, pp. 3415–3419.
  - [16] T. Wang, H. Guo, X. Yan, and Z. Yang, “Speech signal processing on graphs: The graph frequency analysis and an improved graph Wiener filtering method,” *Speech Commun.*, vol. 127, pp. 82–91, 2021.
  - [17] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *CoRR*, vol. abs/1505.04597, 2015.
  - [18] A. Sandryhaila and J. M. F. Moura, “Discrete signal processing on graphs: Graph fourier transform,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2013, pp. 6167–6170.
  - [19] D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time Fourier transform,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 1983, pp. 804–807.
  - [20] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2013, pp. 7092–7096.
  - [21] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR - half-baked or well done?” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2019, pp. 626–630.
  - [22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2015, pp. 5206–5210.
  - [23] A. Varga and H. J. M. 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 Commun.*, vol. 12, no. 3, pp. 247–251, 1993.
  - [24] J. Thiemann, N. Ito, and E. Vincent, “The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings,” *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.
  - [25] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
  - [26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2010, pp. 4214–4217.
  - [27] T. H. Falk, C. Zheng, and W. Chan, “A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech,” *IEEE Trans. Speech Audio Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.
  - [28] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, “A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech,” in *Proc. Interspeech*, 2020, pp. 2482–2486.
  - [29] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, “ICASSP 2021 Deep Noise Suppression Challenge: Decoupling Magnitude and Phase Optimization with a Two-Stage Deep Network,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2021, pp. 6628–6632.
  - [30] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, “MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement,” in *Proc. Interspeech*, 2021, pp. 201–205.