Interspeech 2022
18-22 September 2022, Incheon, Korea

FFC-SE: Fast Fourier Convolution for Speech Enhancement

Ivan Shchekotov[*12], *Pavel Andreev*[*123], *Oleg Ivanov*[1], *Aibek Alanov*[124], *Dmitry Vetrov*[24]

[*] Equal contribution
[1] Samsung AI Center, Moscow
[2] Higher School of Economics, Moscow
[3] Skolkovo Institute of Science and Technology, Moscow
[4] Artificial Intelligence Research Institute, Moscow

i.shchekotov@partner.samsung.com, p.andreev@samsung.com

Abstract

Fast Fourier convolution (FFC) is the recently proposed neural operator showing promising performance in several computer vision problems. The FFC operator allows employing large receptive field operations within early layers of the neural network. It was shown to be especially helpful for inpainting of periodic structures which are common in audio processing. In this work, we design neural network architectures which adapt FFC for speech enhancement. We hypothesize that a large receptive field allows these networks to produce more coherent phases than vanilla convolutional models, and validate this hypothesis experimentally. We found that neural networks based on Fast Fourier convolution outperform analogous convolutional models and show better or comparable results with other speech enhancement baselines.

Index Terms: speech enhancement, fast fourier convolution

1. Introduction

Speech enhancement is of major interest in the audio processing community, as it has a fundamental importance in telecommunication. There are a lot of solutions for this problem in traditional signal processing, but each such solution relies on some assumptions on the underlying noise model. Due to the recent advances in deep learning, data-driven approaches have dominated the area of modern speech enhancement.

One popular line of deep learning techniques tackling speech enhancement is based on the time domain signal retrieval. These approaches often utilize a convolutional encoder-decoder (CED) structure. For example, [1] and [2] follow an adversarial training pipeline and use a CED network as a generator employing a fully-convolutional discriminator for training. Some of these approaches additionally use neural modules that can capture long-range temporal sequence information, such as long short-term memory cells [3] and transformers [4]. However, since these techniques directly map a noisy waveform to the clean one, they typically leave aside any information about signal spectrum, causing potential inefficiencies. One recent attempt to explicitly take into account spectral information during generation is [5]. The authors propose a universal model for vocoding, speech enhancement and bandwidth extension that takes as inputs both waveform and spectrogram and achieves state-of-art results. We show that the quality of speech enhancement can be further improved by our models.

Another line of research is built upon the estimation of short-time Fourier transform (STFT) representations. Approaches of these lines aim to predict STFT coefficients of clean signal directly [6] or correct spectrum of the noisy signal by estimating various masks for modification of magnitudes or both magnitudes and phases [7, 8, 9]. For instance, MetricGAN [10] and MetricGAN+ [11] papers use Bidirectional LSTM to predict binary masks for spectrogram optimizing common speech quality objective metrics directly and report state-of-the-art results for these metrics. The direct estimation of phases is challenging. Different tricks are proposed to simplify this task. These techniques include decoupling magnitude and phase estimation [12] and the usage of separate vocoder networks for waveform synthesis [13]. However, these methods tend to use large neural networks, requiring substantial computational resources. We found that one of the limiting factors for phase prediction is local receptive field of these networks, preventing effective use of models parameters. We observed that phase estimation can be significantly facilitated by non-local neural operators, leading to much smaller model sizes while achieving better quality.

We propose new neural architectures based on fast Fourier convolution (FFC) operator [14] which we adapt for speech enhancement problems. The FFC layers were originally proposed for computer vision tasks as a non-local operator replacing vanilla convolutional layers within existing neural networks. Fast Fourier convolution has the global receptive field and was shown to be helpful for the restoration of periodic backgrounds in inpainting problems [15]. These properties of FFC are especially helpful for the complex spectrum prediction. Indeed, the harmonics of spectrogram are known to form periodic structures which can be naturally handled by fast Fourier convolution (see Figure 1). Besides, we experimentally observe that a large receptive field of FFC is useful for producing coherent phases. Based on these insights, we design new neural architectures for direct complex-valued spectrogram estimation in speech enhancement problems. The proposed models achieve state-of-art performance on VoiceBank-DEMAND [16] and Deep Noise Suppression [17] datasets with much fewer parameters than the baselines. The implementation will become publicly available.

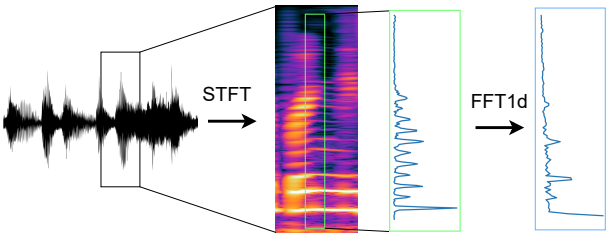

Figure 1: *Harmonics of short-time Fourier transform constitute periodic structures which can be naturally processed in Fourier domain by global branch of fast Fourier convolution.*

10.21437/Interspeech.2022-603

2. Proposed method

We consider the standard single-channel speech denoising problem. In other words, our goal is to learn a mapping from noisy waveform $y = x + n$ with additive noise n to the clean one x. We tackle this problem by neural architectures equipped with a non-local neural operator named fast Fourier convolution [14]. We adapt this operator for complex spectrum processing and propose two neural architectures which use this operator as a basic block.

2.1. Fast fourier convolution

Fast Fourier convolution (FFC) [14] is a neural operator that allows performing non-local reasoning and generation within a neural network. FFC uses channel-wise fast Fourier transform [18] followed by a point-wise convolution and inverse Fourier transform, thus it globally affects input tensor across dimensions involved in Fourier transform. FFC splits channels into local and global branches. The local branch uses conventional convolutions for local updates of feature maps. Global branch performs Fourier transform of the feature map and updates it in spectral domain affecting global context.

In this work, we perform Fourier transform across frequency dimensions of feature maps (corresponding to STFT representations) only (see Figure 1), whereas in computer vision the Fourier transform is applied across both image dimensions [14, 15]. Specifically, we implement the global branch of the FFC layer in three steps:

1. Apply real fast Fourier transform across frequency dimension of the input feature map and concatenate real and imaginary parts of spectrum across channel dimension:

$$\mathbb{R}^{C\times F\times T} \xrightarrow{\text{fft1d}} \mathbb{C}^{C\times F/2\times T} \xrightarrow{\text{concat}} \mathbb{R}^{2C\times F/2\times T}.$$

2. Apply convolutional block (with 1×1 kernel) in the frequency domain:

$$\mathbb{R}^{2C\times F/2\times T} \xrightarrow{\text{conv}-\text{bn}-\text{relu}} \mathbb{R}^{2C\times F/2\times T}.$$

3. Apply inverse Fourier transfrom:

$$\mathbb{R}^{2C\times F/2\times T} \xrightarrow{\text{concat}} \mathbb{C}^{C\times F/2\times T} \xrightarrow{\text{ifft1d}} \mathbb{R}^{C\times F\times T}.$$

where C, F, T are the number of channels, dimension corresponding to frequency and dimension corresponding to time, respectively. Global and local branches interact with each other through summation of activations, as illustrated in Figure 2. We use the same variation of FFC that was explored in [15] for image inpainting, except we utilize one-dimensional Fourier transform across the frequency dimension.

2.2. FFC-AE

We implement two neural network architectures for speech enhancement. The first one (FFC-AE) is inspired by [15]. This architecture consists of the convolutional encoder (strided convolution) which downsamples the input STFT representation across time and frequency dimensions by a factor of two. The encoder is followed by a series of residual blocks, each consisting of two sequential fast Fourier convolution modules. The output of residual blocks is then upsampled by transposed convolution and used to predict real and imaginary parts of the denoised complex-valued spectrogram. The architecture is depicted on Figure 3 (left). We call this model fast Fourier convolutional autoencoder (FFC-AE).

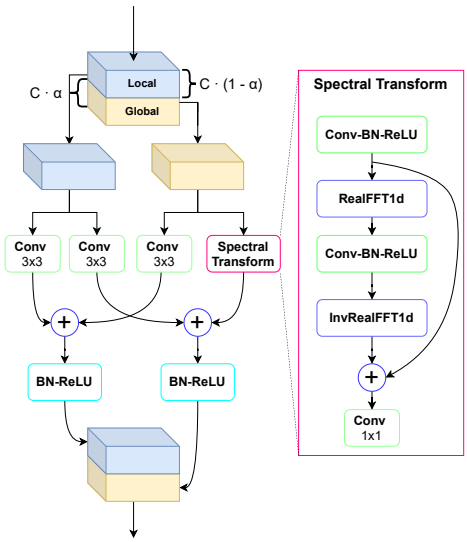

Figure 2: *Fast Fourier Convolution neural module for speech enhancement. Parameter $\alpha \in [0, 1]$ controls the ratio of channels used in the global branch of the module.*

Although bigger downsampling factors lead to a further reduction in the number of operations during inference, we found that it also leads to significant performance degradation, while factor 2 provides a good trade-off between performance and complexity for STFT with window size of 1024 and hop length of 256.

2.3. FFC-UNet

The second architecture is inspired by the classic work [19]. We incorporate FFC layers into U-Net architecture as shown in Figure 3 (right). At each level of the U-Net structure, we utilize several residual FFC blocks with convolutional upsampling or downsampling. We find it beneficial to make the parameter α (ratio of channels going to a global branch of fast Fourier convolution) dependent on the U-Net level at which FFC is used. Higher levels of U-Net structure work with higher resolutions of data at which periodic structures are present, while lower levels work at a coarse scale that lacks periodic structure. More generally, as noted in [14] the deeper layers of neural networks are mainly supposed to exploit local patterns, while the topmost layers highly demand contextual inference. Thus, the global branch of FFC layers is less useful at the coarse scales and we decrease the parameter α starting from 0.75 at the topmost level to 0 at the bottom layer with step 0.25.

2.4. Training

The predicted STFT representation is converted into waveform by inverse short-time Fourier transform. We use the multi-discriminator adversarial training framework proposed in [5] for time-domain models' training. It consists of three losses, namely LS-GAN loss $\mathcal{L}_{GAN}$ [20], feature matching loss $\mathcal{L}_{FM}$ [21, 22], and mel-spectrogram loss $\mathcal{L}_{Mel}$ [23]:

$$\mathcal{L}(\theta) = \mathcal{L}_{GAN}(\theta) + \lambda_{fm}\mathcal{L}_{FM}(\theta) + \lambda_{mel}\mathcal{L}_{Mel}(\theta) \quad (1)$$

$$\mathcal{L}(\varphi_i) = \mathcal{L}_{GAN}(\varphi_i), \quad i = 1, \ldots, k. \quad (2)$$

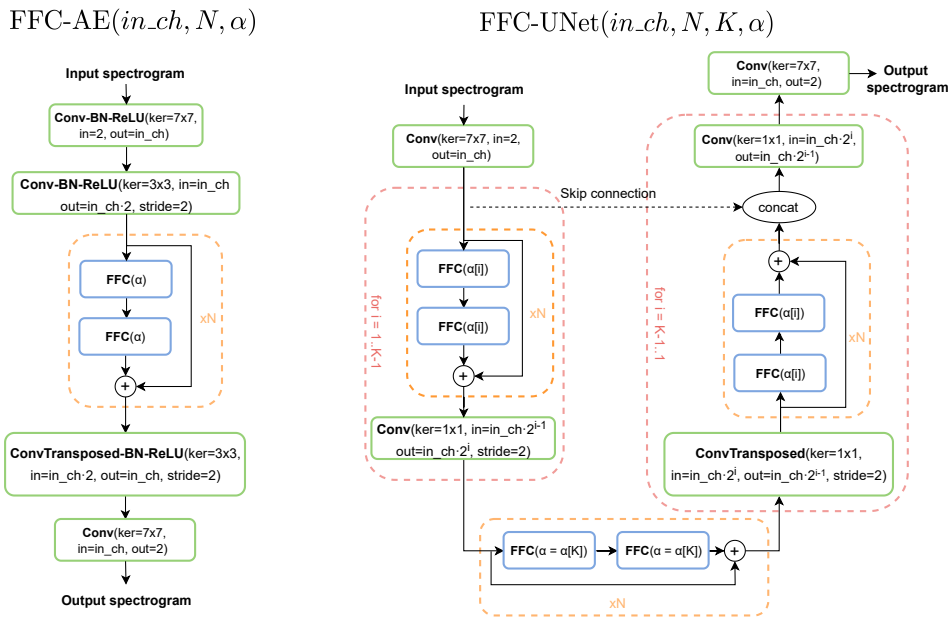

Figure 3: *Proposed architectures for speech enhancement. Left: fast Fourier convolutional autoencoder which adopts architecture introduced in [15] for speech enhancement task. Right: fast Fourier convolutional U-Net. Parameter in_ch controls the overall width of the networks, N defines the number of FFC residual blocks, K is the depth of the FFC-UNet architecture, α (real number $\in [0, 1]$ in case of FFC-AE, K numbers $\in [0, 1]$ in case of FFC-Unet) controls the proportion of channels going to the global branch.*

where $\mathcal{L}(\theta)$ denotes loss for generator with parameters θ, $\mathcal{L}(\varphi_i)$ denotes loss for i-th discriminator with parameters φ_i (all discriminators are identical, except initialized differently). In all experiments we set $\lambda_{fm} = 2$, $\lambda_{mel} = 45$, $k = 3$.

3. Experiments and Results

3.1. Datasets

We use two benchmarks for the evaluation of the effectiveness of the proposed speech denoising models. All audio recordings were sampled at 16 kHz.

The first one is VoiceBank-DEMAND dataset [16] which is a standard benchmark for speech denoising systems. The train set consists of 28 speakers with 4 signal-to-noise ratios (SNR) (15, 10, 5, and 0 dB) and contains 11572 utterances. The test set (824 utterances) consists of 2 speakers unseen by the model during training with 4 SNR (17.5, 12.5, 7.5, and 2.5 dB).

The second benchmark is the Deep Noise Suppression (DNS) challenge [17]. We synthesize 100 hours of training data using provided codes and default configuration. The only modification is that we do not utilize artificial reverberation during synthesis. The models are tested on two kinds of test sets. The first one (DNS-INDOMAIN) is a hold-out data randomly selected and excluded from synthesized 100 hours of training data. The second one (DNS-BLIND) is a standard blind test set from the DNS repository. This data is recorded in the presence of noise in real-world scenarios.

3.2. Metrics

Objective metrics We use conventional metrics WB-PESQ [25], extended STOI [26], scale-invariant signal-to-distortion ratio (SI-SDR) [27], COVL, CBAK, CSIG [28] for objective evaluation of samples in the concerned tasks. Metrics for all baselines and our models were calculated using the publicly available implementations and were not reused from original papers. In addition to conventional speech quality metrics, we considered absolute objective speech quality measure based on direct MOS score prediction by a fine-tuned wav2vec2.0 [29] model (WV-MOS), which was found to have better system-level correlation with subjective quality measures than the other objective metrics [5].

Subjective metrics We use 5-scale MOS tests for subjective quality evaluation following procedure described in [5]. All audio clips were normalized to prevent the influence of audio volume differences on the raters. The referees were restricted to be English speakers with proper listening equipment.

3.3. Experimental Setup

In all our experiments, signals are transformed to the spectral domain using STFT with Hann window of size 1024 and hop size 256. For FFC-AE model we set $\alpha = 0.75$, $N = 9$, $in\_ch = 32$ and $in\_ch = 64$ for V0 and V1 versions, respectively. For FFC-UNet $K = 4, N = 4$, $in\_ch = 32$ and α is gradually decreased with depth as described in Section 2.3. Models are trained for 800 000 iterations with batch size being equal to 8. Adam optimizer is used with learning rate 0.0002. ResUNet-Decouple+ [12] was trained with the same loss function as reported in the original paper, the number iterations was set 800 000 and learning rate to 0.0002.

3.4. Experimental Results

In addition to baselines from the literature [11, 12, 13, 5, 3, 6], we compare against fully convolutional U-Net model (vanilla U-Net) and a model which is the same as FFC-AE except for all Fourier units at the global branch are replaced with vanilla convolutions (FFC-AE (abl.)). These models follow exactly the same training setup as the proposed models for a clear illustration of the FFC importance.

Phase estimation We test the ability of the FFC-AE model to

Table 1: *Speech denoising results on Voicebank-DEMAND dataset. Best three results are highlighted in bold.*

Model	MOS	WV-MOS	SI-SDR	STOI	PESQ	CSIG	CBAK	COVL	# Params (M)
Ground Truth	4.46 ± 0.06	4.50	-	1.00	4.64	5.0	5.0	5.0	-
Input	3.44 ± 0.06	2.99	8.4	0.79	1.97	3.34	2.82	2.74	-
MetricGAN+ [11]	3.82 ± 0.06	3.90	8.5	0.83	**3.13**	4.12	3.16	3.62	2.7
ResUNet-Decouple+ [12]	3.94 ± 0.04	4.13	**18.4**	0.84	2.45	3.38	3.15	2.89	102.6
DEMUCS (non-caus.) [3]	4.06 ± 0.03	**4.37**	**18.5**	**0.87**	**3.03**	**4.36**	**3.51**	**3.72**	60.8
VoiceFixer [13]	4.10 ± 0.03	4.14	-18.5	0.75	2.38	3.6	2.37	2.96	122.1
HiFi++ [5]	4.15 ± 0.07	4.27	**18.4**	0.86	2.76	4.09	3.35	3.43	**1.7**
FFC-AE-V0 (ours)	$\mathbf{4.24 \pm 0.09}$	4.34	17.9	0.86	2.88	4.25	3.40	3.57	**0.42**
FFC-AE-V1 (ours)	$\mathbf{4.33 \pm 0.03}$	**4.37**	17.5	**0.87**	2.96	**4.34**	**3.42**	**3.66**	**1.7**
FFC-UNet (ours)	$\mathbf{4.28 \pm 0.03}$	**4.38**	18.1	**0.87**	**2.99**	**4.35**	**3.47**	**3.69**	7.7
FFC-AE-V1 (abl.)	3.98 ± 0.07	4.05	16.7	0.84	2.68	3.94	3.23	3.31	2.9
vanilla UNet	4.10 ± 0.07	4.11	17.2	0.85	2.73	3.94	3.28	3.34	20.7

Table 2: *Speech denoising results on DNS dataset. * indicates results on DNS-BLIND. Best three results are highlighted in bold.*

Model	MOS	MOS*	WV-MOS	WV-MOS*	SI-SDR	STOI	PESQ	CSIG	CBAK	COVL	# Params (M)
Ground Truth	4.40 ± 0.08	-	3.845	-	-	1.00	4.64	5.0	5.0	5.0	-
Input	2.75 ± 0.07	2.43 ± 0.08	1.195	0.80	-	0.69	1.49	2.59	2.32	1.99	-
DEMUCS [3]	3.52 ± 0.15	2.94 ± 0.08	**3.32**	**2.83**	**15.56**	0.82	2.20	3.44	3.21	2.81	33.5
HiFi++ [5]	3.54 ± 0.08	2.75 ± 0.06	2.91	2.32	11.69	0.82	2.20	3.65	3.00	2.92	**1.7**
ResUNet-Dec+ [12]	3.63 ± 0.04	2.51 ± 0.08	2.94	1.86	14.78	0.81	2.09	2.82	3.06	2.43	102.6
FullSubNet [24]	3.73 ± 0.02	$\mathbf{3.08 \pm 0.09}$	2.90	2.41	**14.96**	0.82	2.43	3.59	**3.27**	3.0	5.6
FFC-AE-V0 (ours)	$\mathbf{3.92 \pm 0.09}$	2.88 ± 0.09	3.20	2.58	12.86	**0.83**	**2.44**	**3.84**	3.17	**3.15**	**0.42**
FFC-AE-V1 (ours)	$\mathbf{4.02 \pm 0.05}$	$\mathbf{3.10 \pm 0.07}$	**3.33**	**2.76**	14.12	**0.85**	**2.61**	**3.98**	**3.31**	**3.31**	**1.7**
FFC-UNet (ours)	$\mathbf{4.00 \pm 0.06}$	$\mathbf{3.11 \pm 0.08}$	**3.35**	**2.70**	**15.48**	**0.86**	**2.69**	**4.08**	**3.44**	**3.41**	7.7

estimate phases given spectrograms on the LJ-Speech dataset [30] and compare against analogous architectures with vanilla convolutions which in contrast to FFC do not have global receptive fields. The models were trained to predict phases (sine and cosine) by guidance of losses described in 2.4 and were provided with magnitude spectrograms. The results are shown in Table 3. FFC-AE significantly outperforms FFC-AE (abl.) and vanilla UNet models while having fewer parameters.

Table 3: *Phase estimation on LJ-Speech dataset*

Model	MOS	WV-MOS	# Params (M)
Ground Truth	4.51 ± 0.05	4.23	-
FFC-AE-V0 (ours)	$\mathbf{4.47 \pm 0.04}$	**4.11**	0.4
vanilla UNet	4.31 ± 0.04	3.97	20.7
FFC-AE-V0 (abl.)	3.96 ± 0.08	3.81	0.7

Speech enhancement We compare the quality of the proposed models with strong baselines on both benchmarks. On Voicebank-DEMAND, as it can be seen from Table 1, our models significantly outperform all the baselines by MOS and give competitive results on objective metrics. On DNS benchmark (Table 2) our models have better quality than all the competitors considering DNS-INDOMAIN test set and perform competitively with FullSubNet [24] (one of the top-ranked models in DNS Challenge 2021) in terms of MOS on DNS-BLIND test set.

Noteworthy, our models performed better or comparably with the closest baselines FullSubNet and DEMUCS on DNS-BLIND set without employing dynamic data synthesis, reverberation simulation and augmentation techniques. Thus, DEMUCS and FullSubNet models which employ these techniques were in an advantageous position from this point of view. We believe that the generalization of our models to the blind test set can be further improved, considering more advanced data generation pipelines.

4. Conclusions

In this paper, we adapted the fast Fourier convolution operator for speech enhancement problems. We observe that neural architectures built upon fast Fourier convolution significantly outperform vanilla convolution-based architectures in terms of quality of speech enhancement, phase estimation and parameter efficiency. In general, the proposed architectures deliver state-of-art results on speech denoising benchmarks, being significantly smaller than the baselines. Future work should consider extending the results to real-time streaming scenarios. Importantly, we believe that the success of fast Fourier convolution can be translated to other speech processing tasks, such as voice conversion and neural vocoding.

5. Acknowledgements

This work was supported by Samsung Research. Dmitry Vetrov was supported by the grant provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement No. 000000D730321P5Q0002 and the agreement with HSE University No. 70-2021-00139.

6. References

[1] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "Seanet: A multi-modal speech enhancement network," *arXiv preprint arXiv:2009.02095*, 2020.

[2] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[3] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.

[4] E. Kim and H. Seo, "SE-Conformer: Time-Domain Speech Enhancement Using Conformer," in *Proc. Interspeech 2021*, 2021, pp. 2736–2740.

[5] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for neural vocoding, bandwidth extension and speech enhancement," *arXiv preprint arXiv:2203.13086*, 2022.

[6] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6633–6637.

[7] 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," in *International Conference on Learning Representations*, 2018.

[8] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[9] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, 2017, pp. 136–140.

[10] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[11] 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," *arXiv preprint arXiv:2104.03538*, 2021.

[12] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep resunet for music source separation," *arXiv preprint arXiv:2109.05418*, 2021.

[13] H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: Toward general speech restoration with neural vocoder," *arXiv preprint arXiv:2109.13731*, 2021.

[14] L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4479–4488, 2020.

[15] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2149–2159.

[16] C. Valentini-Botinhao *et al.*, "Noisy speech database for training speech enhancement algorithms and tts models," 2017.

[17] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper *et al.*, "Icassp 2022 deep noise suppression challenge," *arXiv preprint arXiv:2202.13288*, 2022.

[18] H. J. Nussbaumer, "The fast fourier transform," in *Fast Fourier Transform and Convolution Algorithms*. Springer, 1981, pp. 80–111.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[20] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[21] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*. PMLR, 2016, pp. 1558–1566.

[22] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *arXiv preprint arXiv:1910.06711*, 2019.

[23] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.

[24] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: a full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6633–6637.

[25] 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.

[26] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[27] 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.

[28] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[29] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[30] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.