ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10097255

HIFI++: A UNIFIED FRAMEWORK FOR BANDWIDTH EXTENSION AND SPEECH ENHANCEMENT

Pavel Andreev[*123], *Aibek Alanov*[*124], *Oleg Ivanov*[*1], *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

ABSTRACT

Generative adversarial networks have recently demonstrated outstanding performance in neural vocoding outperforming best autoregressive and flow-based models. In this paper, we show that this success can be extended to other tasks of conditional audio generation. In particular, building upon HiFi vocoders, we propose a novel HiFi++ general framework for bandwidth extension and speech enhancement. We show that with the improved generator architecture, HiFi++ performs better or comparably with the state-of-the-art in these tasks while spending significantly less computational resources. The effectiveness of our approach is validated through a series of extensive experiments.

Index Terms— speech enhancement, bandwidth extension

1. INTRODUCTION

The problem of conditional speech generation has great practical importance. The applications of conditional speech generation include neural vocoding, bandwidth extension (BWE), speech enhancement (SE, also referred to as speech denoising), and many others. One recent success in the field of conditional speech generation is related to the application of generative adversarial networks [1, 2]. Particularly, it was demonstrated that GAN-based vocoders could drastically outperform all publicly available neural vocoders in both quality of generated speech and inference speed. In this work, we adapt the HiFi model [2] to the bandwidth extension and speech enhancement tasks by designing new generator.

The key contribution of this work is a novel HiFi++ generator architecture that allows to efficiently adapt the HiFi-GAN framework to the BWE and SE problems. The proposed architecture is based on the HiFi generator with new modules. Namely, we introduce spectral preprocessing (SpectralUnet), convolutional encoder-decoder network (WaveUNet) and learnable spectral masking (SpectralMaskNet) to the generator's architecture. Equipped with these modifications, our generator can be successfully applied to the bandwidth extension and speech enhancement problems. As we demonstrate through a series of extensive experiments, our model performs on par with state-of-the-art in bandwidth extension and speech enhancement tasks. The model is significantly more lightweight than the examined counterparts while having better or comparable quality.

2. BACKGROUND

Bandwidth extension Frequency bandwidth extension [3, 4] (also known as audio super-resolution) can be viewed as a realistic increase of signal sampling frequency. Speech bandwidth or sampling rate may be truncated due to poor recording devices or transmission channels. Therefore super-resolution models are of significant practical relevance for telecommunication.

For the given audio $x = \{x_i\}_{i=1}^{N}$ with the low sampling rate s, a bandwidth extension model aims at restoring the recording in high resolution $y = \{x_i\}_{i=1}^{N \cdot S/s}$ with the sampling rate S (i.e., expand the effective frequency bandwidth). We generate training and evaluation data by applying low-pass filters to a high sample rate signal and then downsampling the signal to the sampling rate s:

$$x = \mathrm{Resample}(\mathrm{lowpass}(y, s/2), s, S), \tag{1}$$

where $\mathrm{lowpass}(\cdot, s/2)$ means applying a low-pass filter with the cutoff frequency $s/2$ (Nyquist frequency at the sampling rate s), $\mathrm{Resample}(\cdot, S, s)$ denotes downsampling the signal from the sampling frequency S to the frequency s. Following recent works [5, 3], we randomize low-pass filter type and order during training for model robustness.

Speech enhancement Audio denoising [6, 7] is always a major interest in audio processing community because of its importance and difficulty. In this task, it is required to clean the original signal (most often speech) from extraneous distortions. We use additive external noise as distortion. Formally speaking, given the noisy signal $x = y + n$ the denoising algorithm predicts the clean signal y, i.e. suppresses the noise n.

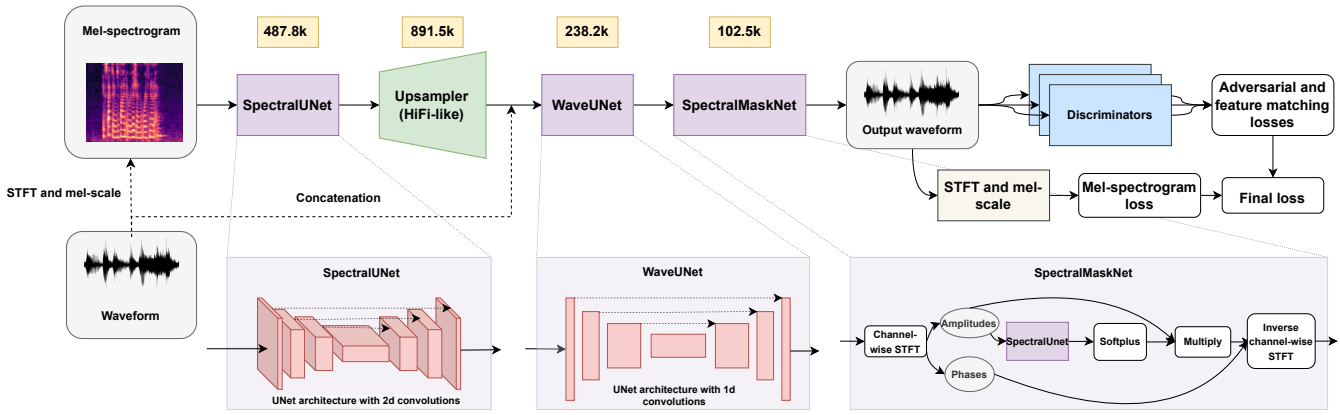

Fig. 1. HiFi++ architecture and training pipeline. The HiFi++ generator consists of the HiFi-like Upsampler and three introduced modules SpectralUNet, WaveUNet and SpectralMaskNet (their sizes are in yellow boxes). The generator's architecture is identical for BWE and SE.

3. HIFI++

3.1. Adapting HiFi-GAN Generator For Bandwidth Extension and Speech Enhancement

In this paper, we propose a novel HiFi++ architecture that adapts HiFi generator [2] to the SE and BWE problems by introducing new modules: SpectralUNet, WaveUNet and SpectralMaskNet (see Figure 1). The HiFi++ generator is based on the HiFi part ($V2$ version of HiFi-GAN generator) that takes as an input the enriched mel-spectrogram representation by the SpectralUNet and its output goes through postprocessing modules: WaveUNet corrects the output waveform in time domain while SpectralMaskNet cleans up it in frequency domain. We also tried to change the order of WaveUNet and SpectralMaskNet modules and did not observe significant improvements. We describe the introduced modules in details in the next paragraphs.

SpectralUNet We introduce the SpectralUNet module as the initial part of the HiFi++ generator that takes the input mel-spectogram (see Figure 1). The mel-spectrogram has a two-dimensional structure and the two-dimensional convolutional blocks of the SpectralUnet model are designed to facilitate the work with this structure at the initial stage of converting the mel-spectrogram into a waveform. The idea is to simplify the task for the remaining part of the HiFi++ generator that should transform this 2d representation to the 1d sequence. We design the SpectralUNet module as UNet-like architecture with 2d convolutions. This module also can be considered as the preprocess part that prepares the input mel-spectrogram by correcting and extracting from it the essential information that is required for the desired task.

WaveUNet The WaveUNet module is placed after the HiFi part (Upsampler) and takes several 1d sequences concatenated with the input waveform as an input. This module operates directly on time domain and it can be considered as a time domain postprocessing mechanism that improves the output of the Upsampler and merges the predicted waveform with the source one. The WaveUNet module is an instance of the well-known architecture Wave-U-Net [8] which is a fully convolutional 1D-UNet-like neural network. This module outputs the 2d tensor which consists of m 1d sequences that will be processed and merged to the output waveform by the next SpectralMaskNet module.

SpectralMaskNet We introduce the SpectralMaskNet as the final part of the generator which is a learnable spectral masking. It takes as an input the 2d tensor of m 1d sequences and applies channel-wise short-time Fourier transform (STFT) to this 2d tensor. Further, the SpectralUNet-like network takes the amplitudes of the STFT output (in the case of vocoding it takes also the output of SpectralUNet module concatenated) to predict multiplicative factors for these amplitudes. The concluding part consists of the inverse STFT of the modified spectrum (see Figure 1). Importantly, this process does not change phases. The purpose of this module is to perform frequency-domain postprocessing of the signal. We hypothesize that it is an efficient mechanism to remove artifacts and noise in frequency domain from the output waveform in a learnable way. Note that similar techniques have been used in speech enhancement literature as a standalone solution [9].

3.2. Training objective

We use the multi-discriminator adversarial training framework that is based on [2] for time-domain models' training. However, instead of multi-period and multi-scale discriminators we use several identical discriminators that are based on multi-scale discriminators but operate on the same resolutions and have smaller number of weights (we reduce channel number in each convolutional layer by the factor of 4). We

employ three losses, namely LS-GAN loss $\mathcal{L}_{GAN}$ [10], feature matching loss $\mathcal{L}_{FM}$ [1], and mel-spectrogram loss $\mathcal{L}_{Mel}$ [2]:

$$\mathcal{L}(\theta) = \mathcal{L}_{GAN}(\theta) + \lambda_{fm}\mathcal{L}_{FM}(\theta) + \lambda_{mel}\mathcal{L}_{Mel}(\theta) \quad (2)$$

$$\mathcal{L}(\varphi_i) = \mathcal{L}_{GAN}(\varphi_i), \quad i = 1, \dots, k. \quad (3)$$

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 $\lambda_{fm} = 2$, $\lambda_{mel} = 45$, $k = 3$.

4. EXPERIMENTS

All training hyper-parameters and implementation details will be released with source codes.

4.1. Data

Bandwidth extension We use publicly available dataset VCTK [11] (CC BY 4.0 license) which includes 44200 speech recordings belonging to 110 speakers. We exclude 6 speakers from the training set and 8 recordings from the utterances corresponding to each speaker to avoid text level and speaker-level data leakage to the training set. For evaluation, we use 48 utterances corresponding to 6 speakers excluded from the training data. Importantly, the text corresponding to evaluation utterances is not read in any recordings constituting training data.

Speech denoising We use VCTK-DEMAND dataset [12] (CC BY 4.0 license) for our denoising experiments. The train sets (11572 utterances) consists of 28 speakers with 4 signal-to-noise ratio (SNR) (15, 10, 5, and 0 dB). The test set (824 utterances) consists of 2 speakers with 4 SNR (17.5, 12.5, 7.5, and 2.5 dB). Further details can be found in the original paper.

4.2. Evaluation

Objective evaluation We use conventional metrics WB-PESQ [13], STOI [14], scale-invariant signal-to-distortion ratio (SI-SDR) [15], DNSMOS [16] for objective evaluation of samples in the SE task. In addition to conventional speech quality metrics, we considered absolute objective speech quality measure based on direct MOS score prediction by a fine-tuned wave2vec2.0 model (WV-MOS), which was found to have better system-level correlation with subjective quality measures than the other objective metrics[1].

Subjective evaluation We employ 5-scale MOS tests for subjective quality assessment. 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.

[1]https://github.com/AndreevP/wvmos

4.3. Bandwidth Extension

In our bandwidth extension experiments, we use recordings with a sampling rate of 16 kHz as targets and consider three frequency bandwidths for input data: 1 kHz, 2kHz, and 4 kHz. The models are trained independently for each bandwidth. The results and comparison with other techniques are outlined in Table 1. Our model HiFi++ provides a better tradeoff between model size and quality of bandwidth extension than other techniques. Specifically, our model is 5 times smaller than the closest baseline SEANet [17] while outperforming it for all input frequency bandwidths. In order to validate the superiority of HiFi++ over SEANet in addition to MOS tests we conducted pair-wise comparisons between these two models and observe statistically significant dominance of our model (p-values are equal to $2.8 \cdot 10^{-22}$ for 1 kHz bandwidth, 0.003 for 2 kHz, and 0.02 for 4 kHz for the binomial test).

Importantly, these results highlight the importance of adversarial objectives for speech frequency bandwidth extension models. Surprisingly, the SEANet model [17] appeared to be the strongest baseline among examined counterparts leaving the others far behind. This model uses adversarial objective similar to ours. The TFilm [18] and 2S-BWE [4] models use supervised reconstruction objectives and achieve very poor performance, especially for low input frequency bandwidths.

4.4. Speech Enhancement

The comparison of the HiFi++ with baselines is demonstrated in the Table 2. Our model achieves comparable performance with state-of-the-art models VoiceFixer [3] and DB-AIAT [19] counterparts while being dramatically more computationally efficient. Interestingly, VoiceFixer achieves high subjective quality while being inferior to other models according to objective metrics, especially to SI-SDR and STOI. Indeed, VoiceFixer doesn't use waveform information directly and takes as input only mel-spectrogram, thus, it misses parts of the input signal and is not aiming at reconstructing the original signal precisely leading to poor performance in terms of classic relative metrics such as SI-SDR, STOI, and PESQ. Our model provides decent relative quality metrics as it explicitly uses raw signal waveform as model inputs. At the same time, our model takes into account signal spectrum, which is very informative in speech enhancement as was illustrated by the success of classical spectral-based methods. It is noteworthy that we significantly outperform the SEANet [7] model, which is trained in a similar adversarial manner and has a larger number of parameters, but does not take into account spectral information.

An interesting observation is the performance of the MetriGAN+ model [6]. While this model is explicitly trained to optimize PESQ and achieves high values of this metric, this success does not spread on other objective and subjective metrics.

Table 1. Bandwidth extension results on VCTK dataset. * indicates re-implementation.

	BWE (1kHz)		**BWE (2kHz)**		**BWE (4kHz)**		
Model	MOS	WV-MOS	MOS	WV-MOS	MOS	WV-MOS	# Param (M)
Ground truth	4.62 ± 0.06	4.17	4.63 ± 0.03	4.17	4.50 ± 0.04	4.17	-
HiFi++ (ours)	$\mathbf{4.10 \pm 0.05}$	**3.71**	$\mathbf{4.44 \pm 0.02}$	**3.95**	$\mathbf{4.51 \pm 0.02}$	4.16	**1.7**
*SEANet [17]	3.94 ± 0.09	3.66	$\mathbf{4.43 \pm 0.05}$	**3.95**	4.45 ± 0.04	**4.17**	9.2
VoiceFixer [3]	3.04 ± 0.08	3.21	3.82 ± 0.06	3.50	4.34 ± 0.03	3.77	122.1
*2S-BWE (TCN) [4]	2.01 ± 0.06	2.34	2.98 ± 0.08	3.07	4.10 ± 0.04	3.96	2.7
*2S-BWE (CRN) [4]	1.97 ± 0.06	2.17	2.85 ± 0.04	3.16	4.27 ± 0.05	4.05	9.2
TFiLM [18]	1.98 ± 0.02	1.65	2.67 ± 0.04	2.27	3.54 ± 0.04	3.49	68.2
input	1.87 ± 0.08	0.39	2.46 ± 0.04	1.74	3.36 ± 0.06	3.17	-

Table 2. Speech denoising results on Voicebank-DEMAND dataset. * indicates re-implementation.

Model	MOS	WV-MOS	SI-SDR	STOI	PESQ	DNSMOS	# Par (M)	# MACs (G)
Ground truth	4.46 ± 0.05	4.50	-	1.00	4.64	3.15	-	-
DB-AIAT [19]	$\mathbf{4.40 \pm 0.05}$	**4.38**	**19.4**	**0.96**	**3.27**	**3.18**	2.8	41.8
HiFi++ (ours)	4.31 ± 0.05	4.36	17.9	0.95	2.90	3.10	**1.7**	**2.8**
VoiceFixer [3]	4.21 ± 0.06	4.14	-18.5	0.89	2.38	3.13	122.1	34.4
DEMUCS [20]	4.17 ± 0.06	4.37	18.5	0.95	3.03	3.14	60.8	38.1
*SEANet [17]	4.00 ± 0.06	4.19	13.5	0.92	2.36	3.05	9.2	4.50
MetricGAN+ [6]	3.98 ± 0.06	3.90	8.5	0.93	3.13	2.95	2.7	28.5
Input	3.45 ± 0.07	2.99	8.4	0.92	1.97	2.53	-	-

4.5. Ablation Study

To validate the effectiveness of the proposed modifications, we performed the ablation study of the introduced modules SpectralUNet, WaveUNet and SpectralMaskNet. For each module, we consider the architecture without this module *with increased capacity* of HiFi generator part to match the size of the initial HiFi++ architecture.

The results of the ablation study are shown in Table 3, which reveal how each module contributes to the HiFi++ performance. We also compare against vanilla HiFi generator model which takes mel-spectrogram as the only input. The structure of the vanilla HiFi generator is the same as in $V1$ and $V2$ versions from HiFi-GAN paper, except the parameter "upsample initial channel" is set to 256 (it is 128 for $V2$ and 512 for $V1$). We can see that WaveUNet and SpectralMaskNet are essential components of the architecture, as their absence notably degrades the model performance. SpectralUNet has no effect on quality of SE and minor positive effect on BWE (statistical significance of improvement is ensured by pairwise test). However, since we match the number of parameters for ablation models with HiFi++, this positive effect comes at no cost, thus, it is useful to include SpectralUNet into generator architecture.

Table 3. Ablation study results.

	BWE (1kHz)	**SE**
Model	MOS	MOS
Ground truth	4.50 ± 0.06	4.48 ± 0.05
Baseline (HiFi++)	$\mathbf{3.92 \pm 0.04}$	$\mathbf{4.27 \pm 0.04}$
w/o SpectralUNet	3.83 ± 0.06	$\mathbf{4.26 \pm 0.05}$
w/o WaveUNet	3.46 ± 0.06	4.19 ± 0.03
w/o SpectralMaskNet	3.51 ± 0.06	4.17 ± 0.05
vanilla HiFi	3.42 ± 0.05	4.17 ± 0.04
input	1.69 ± 0.05	3.51 ± 0.06

5. CONCLUSION

In this work, we introduce the universal HiFi++ framework for bandwidth extension and speech enhancement. We show through a series of extensive experiments that our model achieves results on par with the state-of-the-art baselines on BWE and SE tasks. Remarkably, our model obtains such results being much more effiecient (in some cases by two orders of magnitude) than existing counterparts.

6. REFERENCES

[1] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *arXiv preprint arXiv:1910.06711*, 2019.

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

[3] Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, and DeLiang Wang, "Voicefixer: A unified framework for high-fidelity speech restoration," *arXiv preprint arXiv:2204.05841*, 2022.

[4] Ju Lin, Yun Wang, Kaustubh Kalgaonkar, Gil Keren, Didi Zhang, and Christian Fuegen, "A two-stage approach to speech bandwidth extension," *Proc. Interspeech 2021*, pp. 1689–1693, 2021.

[5] Heming Wang and Deliang Wang, "Towards robust speech super-resolution," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[6] Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," *arXiv preprint arXiv:2104.03538*, 2021.

[7] Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek, "Seanet: A multi-modal speech enhancement network," *arXiv preprint arXiv:2009.02095*, 2020.

[8] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[9] Scott Wisdom, John R Hershey, Kevin Wilson, Jeremy Thorpe, Michael Chinen, Brian Patton, and Rif A Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 900–904.

[10] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[11] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

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

[13] Antony W Rix, John G Beerends, Michael P Hollier, and Andries 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)*. IEEE, 2001, vol. 2, pp. 749–752.

[14] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper 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.

[15] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John 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.

[16] Chandan KA Reddy, Vishak Gopal, and Ross Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 886–890.

[17] Yunpeng Li, Marco Tagliasacchi, Oleg Rybakov, Victor Ungureanu, and Dominik Roblek, "Real-time speech frequency bandwidth extension," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 691–695.

[18] Sawyer Birnbaum, Volodymyr Kuleshov, Zayd Enam, Pang Wei Koh, and Stefano Ermon, "Temporal film: Capturing long-range sequence dependencies with feature-wise modulations," *arXiv preprint arXiv:1909.06628*, 2019.

[19] Guochen Yu, Andong Li, Chengshi Zheng, Yinuo Guo, Yutian Wang, and Hui Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7847–7851.

[20] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020.