

# HiFi++: a Unified Framework for Neural Vocoding, Bandwidth Extension and Speech Enhancement

Pavel Andreev<sup>\*123</sup> Aibek Alanov<sup>\*12</sup> Oleg Ivanov<sup>\*1</sup> Dmitry Vetrov<sup>24</sup>

## 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 neural vocoding, bandwidth extension, and speech enhancement. We show that with the improved generator architecture and simplified multi-discriminator training, HiFi++ performs on par with the state-of-the-art in these tasks while spending significantly less memory and computational resources. The effectiveness of our approach is validated through a series of extensive experiments.

## 1. Introduction

The problem of conditional speech generation has great practical importance. The applications of conditional speech generation include neural vocoding, bandwidth extension, speech denoising (also referred to as speech enhancement), and many others. One recent success in the field of conditional speech generation is related to the application of generative adversarial networks (Kumar et al., 2019; Kong et al., 2020a). 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 further improve the HiFi model (Kong et al., 2020a) by designing more efficient generator and discriminator neural networks.

The key contribution of this work is a novel HiFi++ generator architecture that allows to outperform the HiFi model

with significantly less parameters. 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 made substantially more lightweight than the original HiFi model while achieving better performance than the latter in terms sample quality.

Kong et al. (2020a); You et al. (2021) argue that the HiFi model success can be largely attributed to the multi-resolution discrimination framework. In this paper, we show that this framework is unnecessary and can be simplified to several absolutely identical discriminators operating on the same resolution. Thus, we claim that the success of multi-resolution discriminators is mainly related to the effect of generative multi-adversarial networks (Durugkar et al., 2016), i.e. usage of several discriminators during adversarial training. In addition to the conceptual simplification of the discrimination framework, we reduce the number of discriminators’ parameters and their computational complexity facilitating faster training.

Crucially, we show that the resulting framework 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 neural vocoding, bandwidth extension and speech enhancement tasks. The model is significantly more lightweight than the examined counterparts while having better or comparable quality.

To sum up, the contributions of this paper are as follows:

1. We propose the novel HiFi++ generator architecture by introducing three additional modules: SpectralUnet, WaveUNet, and SpectralMaskNet subnetworks. The new architecture allows obtaining better quality with much fewer parameters.
2. We demystify the importance of a multi-resolution discrimination framework for conditional waveform generation and propose new discriminators which are light, simple, and fast while being able to provide the

<sup>\*</sup>Equal contribution <sup>1</sup>Samsung AI Center, Moscow, Russia <sup>2</sup>Higher School of Economics, Moscow Russia <sup>3</sup>Skolkovo Institute of Science and Technology, Moscow Russia <sup>4</sup>Artificial Intelligence Research Institute, Moscow Russia. Correspondence to: Pavel Andreev <p.andreev@samsung.com, pavel.andreev@skoltech.ru>.

same quality as original HiFi discriminators.

3. We propose a unified framework for neural vocoding, bandwidth extension, and speech enhancement delivering results on par with state-of-the-art in these domains.

## 2. Background

### 2.1. Neural vocoding

The majority of modern speech synthesis systems decompose this task into two stages. In the first stage, low-resolution intermediate representations (e.g., linguistic features, mel-spectrograms) are predicted from text data (Li et al., 2020; Shen et al., 2018; 2020). In the second stage, these intermediate representations are transformed to raw waveform (Kong et al., 2020a; Oord et al., 2016; Prenger et al., 2019). Neural vocoders relate to the techniques used in the second stage of the speech synthesis process.

More formally, we consider a neural vocoder as a learnable mapping from the mel-spectrogram  $x = \text{Mel}(y)$  to the raw waveform  $y$ .

### 2.2. Bandwidth extension

Frequency bandwidth extension (Kuleshov et al., 2017; Lin et al., 2021) (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  $s$ :

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

where  $\text{lowpass}(\cdot, s/2)$  means applying a low-pass filter with the cutoff frequency  $s/2$  (Nyquist frequency at the sampling rate  $s$ ),  $\text{Resample}(\cdot, S, s)$  denotes downsampling the signal from the sampling frequency  $S$  to the frequency  $s$ . Following recent works (Wang & Wang, 2021; Sulun & Davies, 2020; Liu et al., 2021), we randomize low-pass filter type and order during training for model robustness.

### 2.3. Speech enhancement

Audio denoising (Fu et al., 2019; Tagliasacchi et al., 2020) 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  $n$ .

## 3. HiFi++

### 3.1. Improved HiFi-GAN Generator Architecture

The HiFi generator (Kong et al., 2020a) was recently proposed as a highly computationally efficient fully convolutional network that solves the neural vocoding with speech quality comparable to autoregressive counterpart while being several orders of magnitude faster. The key part of this architecture is a multi-receptive field fusion (MRF) module which allows to model diverse receptive field patterns. By adjusting parameters of the HiFi architecture one can obtain a good trade-off between computational efficiency and sample quality of the model.

In this paper, we propose a novel HiFi++ architecture that outperforms the largest HiFi model (HiFi V1) with 8 times smaller size (see Table 1). To obtain such high performance with very limited capacity we significantly improve the HiFi generator by introducing new modules: SpectralUNet, WaveUNet and SpectralMaskNet (see Figure 1). The HiFi++ generator is based on the HiFi part 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. These modules allow to achieve a decent sample quality with the reduced by an order of magnitude HiFi part and to make the overall HiFi++ architecture light and efficient compared to HiFi V1 model. 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-spectrogram (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 and takes the 1d sequence as an input (in the case of

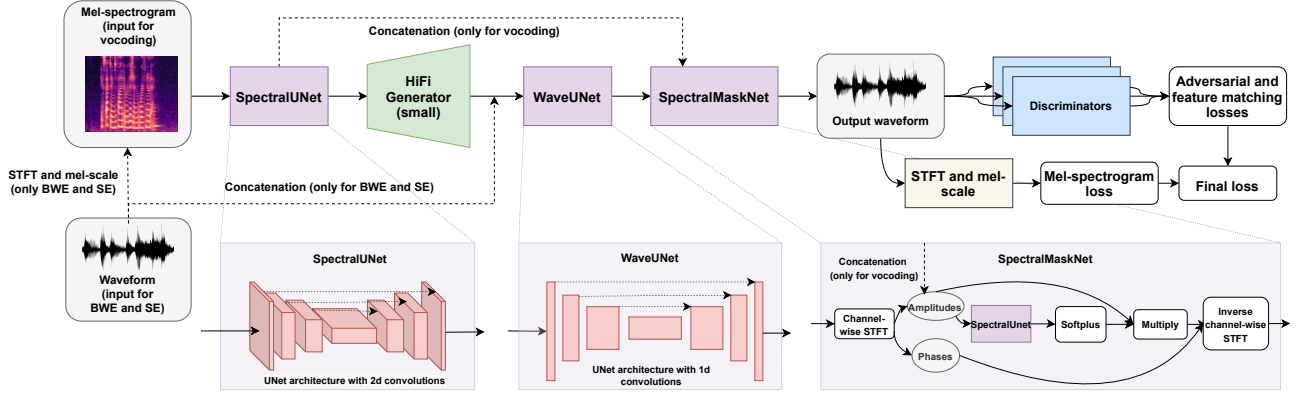


Figure 1. HiFi++ architecture and training pipeline. The HiFi++ generator consists of the small instance of the original HiFi generator and three introduced modules SpectralUNet, WaveUNet and SpectralMaskNet. The generator’s architecture is identical for all three concerned tasks (vocoding, bandwidth extension and speech enhancement) except WaveUNet uses additional raw waveform input in bandwidth extension and speech enhancement tasks, while SpectralMaskNet uses the output of SpectralUNet in vocoding.

BWE and SE problems it also takes the input waveform concatenated, see Figure 1). This module operates directly on time domain and it can be considered as a time domain postprocessing mechanism that improves the output of the HiFi part. The WaveUNet module is an instance of the well-known architecture Wave-U-Net (Stoller et al., 2018) 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 (Wisdom et al., 2019; Jansson et al., 2017).

Overall, the introduced modules allows to design such HiFi++ generator that has model complexity smaller by almost an order of magnitude than HiFi V1 and achieves the comparable and even better sample quality (see Table 1). We demonstrate the similar performance gain and high efficiency for bandwidth extension and speech enhancement (see Tables 2 and 3). These results show the effectiveness and importance of the introduced modules for speech-related tasks.

Table 1. Vocoding results on LJSpeech.  $G$  size (M) means the number of parameters of the generator in millions.  $G$  MACs (G) denotes for the number of the multiply-accumulate operations in billions.  $D$  size (M) and  $D$  MACs (G) are similar but relates to discriminators in total. Training time (D) is computed on a single V100 GPU in days. WV-MOS is an introduced objective metric (see Section 5.2).

Model	Model quality				Model complexity				
	MOS	WV-MOS	STOI	PESQ	$G$ size (M)	$G$ MACs (G)	$D$ size (M)	$D$ MACs (G)	Training time (D)
Ground Truth	$4.70 \pm 0.04$	4.23	1.00	4.64					
WaveGlow	$3.79 \pm 0.05$	3.95	0.97	3.14	87.73	N/A	-	-	N/A
MelGAN	$3.53 \pm 0.05$	3.62	0.93	2.01	4.26	5.94	7.41	1.98	N/A
HiFi V3	$4.20 \pm 0.03$	3.92	0.91	2.61	1.46	2.36	48.5	11.42	14.8
HiFi V2	$4.46 \pm 0.06$	4.01	0.91	2.83	<b>0.92</b>	<b>1.42</b>	48.5	11.42	13.8
HiFi V1	$4.51 \pm 0.05$	4.02	0.92	3.42	13.92	22.70	48.5	11.42	20.1
HiFi++ (ours)	<b><math>4.56 \pm 0.06</math></b>	<b>4.14</b>	<b>0.98</b>	<b>3.61</b>	1.72	2.78	<b>1.86</b>	<b>0.86</b>	<b>9.2</b>

We believe that it can be further improved or extended to obtain even better trade-off between the model performance and its complexity.

### 3.2. Light, Simple and Fast Discriminators Reduce Training Complexity

The HiFi generator is trained in an adversarial manner against two types of discriminators: the multi-period discriminator (MPD) and the multi-scale discriminator (MSD). MPD consists of several sub-discriminators each processing different periodic sub-signals of input audio. The aim of MPD discriminators is to identify various periodic patterns of the speech. MSD also consists of several sub-discriminators that evaluate input waveforms at different temporal resolutions. It was proposed in MelGAN (Kumar et al., 2019) to process consecutive patterns and long-term dependencies. The HiFi training consists of 5 MPD discriminators and 3 MSD discriminators which in total have capacity almost 5 times larger than HiFi V1 generator and significantly slow down the training process. Kong et al. (2020a) argue that such complicated and expensive multi-resolution discrimination framework is one of the key factors of high quality performance of HiFi model which is supported by ablation study.

In this paper, we show that this framework is over-complicated and redundant and can be simplified to 3 identical discriminators that have capacity 25 times smaller than HiFi discriminators and notably reduce the training time (see Table 1). Firstly, we illustrate that the ablation study from HiFi paper (Kong et al., 2020a) is misleading because it shows that without MPD discriminators the model performance is drastically degraded. However, we observe that it is a consequence of a poor hyperparameter selection and with a more accurate one the model can achieve the same quality as without MPD (see Table 4). Further, we demonstrate that we can substitute MSD discriminators that operate on different input resolutions for identical much smaller discriminators that process the waveform on the initial resolution. So, from our findings it follows that the benefit of the HiFi multi-resolution discriminators can be mainly explained by the well-known effect in the GAN literature of generative multi-adversarial networks (Durugkar et al., 2016).

The main point of this effect of generative multi-adversarial networks (Durugkar et al., 2016) is that the performance of the generative model can be easily improved by training against multiple discriminators with the same architectures but different initialization. The more discriminators the better sample quality the model can achieve, however this effect saturates very fast with the number of discriminators. We empirically observe that for speech tasks 3 discriminators are a good trade-off between quality and training time.

Overall, we propose to utilize only 3 identical light, simple and fast discriminators instead of 8 different heavy and slow discriminators of the original HiFi model. We show that our discrimination framework drastically reduce training complexity both in terms of memory consumption and computations while achieving the better performance than the HiFi model (see Table 1).

### 3.3. Unified Framework for Neural Vocoding, Bandwidth Extension and Speech Enhancement

The proposed HiFi++ generator and efficient multi-discrimination training can be considered as a unified framework for conditional speech generation. So, almost without modifications it can be easily applied to neural vocoding, bandwidth extension (BWE) and speech enhancement (SE) problems as the most important tasks in this field. Moreover, the HiFi++ model outperforms existing baselines in each task with significantly less model complexity (see Section 5). These results empirically confirm that the HiFi++ architecture is robust and suitable for speech-related tasks.

This framework consists of the lightweight HiFi++ generator, discriminators and training loss terms (see Figure 1). Across all three problems these parts are the same except the generator architecture that has minor changes depending on each task. The only modifications are: i) an additional skip-connection from the input waveform to the WaveUNet module for BWE and SE for the model to take into account the input raw waveform information; ii) an additional skip-connection from the output of the SpectralUNet to the SpectralMaskNet for vocoding facilitating usage of the clean input spectrogram in SpectralMaskNet module.

Further we will describe training loss terms to fully introduce our framework.

#### 3.3.1. TRAINING LOSS

**GAN Loss** As we use multi-discrimination training we have  $k$  identical discriminators  $D_1, \dots, D_k$  (in all experiments we use  $k = 3$ ). As an adversarial training objective we use LS-GAN (Mao et al., 2017) that provides non-vanishing gradient flows compared to the original GAN loss (Goodfellow et al., 2014). LS-GAN losses for the generator  $G_\theta$  with parameters  $\theta$  and the discriminators  $D_{\varphi_1}, \dots, D_{\varphi_k}$  with parameters  $\varphi_1, \dots, \varphi_k$  are defined as

$$\mathcal{L}_{GAN}(\varphi_i) = \mathbb{E}_{(x,y)} [(D_{\varphi_i}(y) - 1)^2 + D_{\varphi_i}(G_\theta(x))^2], \quad i = 1, \dots, k, \quad (2)$$

$$\mathcal{L}_{GAN}(\theta) = \sum_{i=1}^k \mathbb{E}_x [(D_{\varphi_i}(G_\theta(x)) - 1)^2], \quad (3)$$



where  $y$  denotes the ground truth audio and  $x = f(y)$  denotes the input condition and the transform  $f$  can be mel-spectrogram, low-pass filter or adding noise.

**Feature Matching Loss** The feature matching loss is computed as  $L_1$  distance between intermediate discriminator feature maps computed for ground-truth sample and conditionally generated one (Larsen et al., 2016; Kumar et al., 2019). It was successfully employed to speech synthesis (Kumar et al., 2019) to stabilize the adversarial training process. The feature matching loss is computed as

$$\mathcal{L}_{FM}(\theta) = \sum_{i=1}^k \mathbb{E}_{(x,y)} \left[ \sum_{j=1}^T \frac{1}{N_j} \|D_{\varphi_i}^j(y) - D_{\varphi_i}^j(G_{\theta}(x))\|_1 \right], \quad (4)$$

where  $T$  denotes the number of layers in the discriminator;  $D_{\varphi_i}^j$  and  $N_j$  denote the activations and the size of activations in the  $j$ -th layer of the  $i$ -th discriminator, respectively.

**Mel-Spectrogram Loss** The mel-spectrogram loss is the  $L_1$  distance between the mel-spectrogram of a waveform synthesized by the generator and that of a ground truth waveform. It is defined as

$$\mathcal{L}_{Mel}(\theta) = \mathbb{E}_{(x,y)} [\|\phi(y) - \phi(G_{\theta}(x))\|_1], \quad (5)$$

where  $\phi$  is the function that transforms a waveform into the corresponding mel-spectrogram.

**Final Loss** Our final losses for the generator and discriminator are as

$$\mathcal{L}(\theta) = \mathcal{L}_{GAN}(\theta) + \lambda_{fm}\mathcal{L}_{FM}(\theta) + \lambda_{mel}\mathcal{L}_{Mel}(\theta) \quad (6)$$

$$\mathcal{L}(\varphi_i) = \mathcal{L}_{GAN}(\varphi_i), \quad i = 1, \dots, k. \quad (7)$$

In all experiments we set  $\lambda_{fm} = 2$  and  $\lambda_{mel} = 45$ .

## 4. Related work

### 4.1. Neural vocoding

Neural vocoders can be categorized into several families according to types of generative models. Autoregressive models (Kalchbrenner et al., 2018; Oord et al., 2016) have demonstrated their proficiency in generating high fidelity speech; however, due to the sequential nature of generation, these models are inherently slow. Flow-based models (Oord et al., 2018; Prenger et al., 2019) do not need to generate samples sequentially. Therefore they can be successfully parallelized while maintaining the quality of autoregressive counterparts. However, these models are typically too large to be used in real-time applications. Another line of research is devoted to diffusion probabilistic models (Kong et al., 2020b; Chen et al., 2020). These models were shown to provide sample quality matching the strong autoregressive

baselines using a fewer sequential operations. Still, iterative nature of deep diffusion probabilistic models makes them inferior to generative adversarial networks in terms of generation speed.

Vocoders based on generative adversarial networks (GANs) (Kumar et al., 2019; Kong et al., 2020a) have recently been shown to be a superior type of generative models in both quality of generated speech and inference speed. Specifically, Kong et al. (2020a) showed that GAN-based vocoders are able to achieve superior quality over all publicly available neural vocoders while requiring less resources for inference. In our work, we further improve their results and design more efficient architectures for discriminators and generator.

### 4.2. Bandwidth extension

Several prior works (Birnbaum et al., 2019; Lin et al., 2021; Wang & Wang, 2020) tackle the bandwidth extension problem by waveform-to-waveform or joint time-frequency neural architectures equipped with different supervised reconstruction losses. Birnbaum et al. (2019) (TFiLM) proposed a temporal feature-wise linear modulation layer that uses a recurrent neural network to alter the activations of a convolutional model. The authors applied this layer to convolutional encoder-decoder neural architecture operating in waveform domain (Kuleshov et al., 2017) and observe significant benefits of these layers for bandwidth extension quality. Lin et al. (2021) (2S-BWE) considered a two-stage approach to frequency bandwidth extension. At the first stage signal spectrum is predicted by either temporal convolutional network (TCN) (Bai et al., 2018) or convolutional recurrent network (CRN) (Tan & Wang, 2018) while at the second stage raw waveform is refined by WaveUNet model. The authors demonstrated the benefits of the two-stage generation procedure and the superiority of the concerned approach over baselines. Nevertheless, as we will show in the next section the perceptual sample quality achieved by these models tends to be poor especially for low input frequency bandwidths.

Other works (Kim & Sathe, 2019; Li et al., 2021) addressed audio super-resolution problem with generative adversarial networks. Li et al. (2021) considered waveform-to-waveform fully convolutional encoder-decoder SEANet architecture introduced by Tagliasacchi et al. (2020) and multi-scale discriminators (Kumar et al., 2019) operating in waveform domain. In contrast, our method explicitly takes into account information about signal spectrum and uses a more efficient discrimination framework. In Section 5 we will show that allows it to achieve superior quality compared to SEANet models while requiring less trainable parameters.

Concurrent work (VoiceFixer) (Liu et al., 2021) proposed to solve bandwidth extension problem as a part of general

speech restoration problem (i.e., different types of signal distortions are considered at the same time). The authors propose a two-stage approach for speech restoration. At the first stage ResUNet (Kong et al., 2021) is used to reconstruct signal mel-spectrogram. At the second stage, TFGAN vocoder (Tian et al., 2020) is employed to synthesize waveform given reconstructed mel-spectrogram. Our approach is similar in the sense that we also built our model upon neural vocoder, however, we train our model in an end-to-end manner and achieve more efficient performance (see Section 5).

### 4.3. Speech denoising

The recent deep learning papers on the topic form two lines of research.

The first one operates at the waveforms level, or in the time domain. Stoller et al. (2018) proposed to adapt UNet (Ronneberger et al., 2015) model to unidimensional time-domain signal processing for solving the problem of audio sources separation, which is a general case of speech denoising problem. The proposed convolutional encoder-decoder (CED) architecture became common for speech enhancement neural network models. For example, Pascual et al. (2017) follow an adversarial training pipeline and use a CED network as a generator employing a fully-convolutional discriminator for training. The above-mentioned SEANet model (Tagliasacchi et al., 2020) also solves the speech denoising problem and uses a fully-convolutional architectures of generators and discriminators. Defossez et al. (2020) proposes the DEMUCS architecture for the speech denoising problem. DEMUCS (Défossez et al., 2019) is a CED network with gated convolutions and long short-term memory modules in the bottleneck part. The model is trained using a joint time and frequency domain reconstruction loss. Recently, Gulati et al. (2020) efficiently combined convolution neural networks and transformers for speech recognition in the time domain. The resulted model is called Conformer and showed state-of-the-art performance in different audio processing problems. In particular, Kim & Seo (2021) (SE-Conformer) successfully adapted Conformer architecture for time-domain speech denoising.

The second line of papers doesn't rely on time-domain information and uses a high-level spectrogram representation of the audio instead. Many approaches constituting this line utilize a spectral masking technique, i.e., for each point of the spectrogram they predict a real-valued multiplicative factor lying in  $[0, 1]$ . For instance, MetricGAN (Fu et al., 2019) and MetricGAN+ (Fu et al., 2021) papers use Bidirectional LSTM combined with spectral masking to optimize common speech quality objective metrics directly, and report state-of-the-art results for these metrics.

## 5. Experiments

All training hyper-parameters and implementation details are released with source code as a part of supplementary material.

### 5.1. Data

**Neural vocoding** We use public LJ-Speech dataset (Ito & Johnson, 2017) which is standard in the speech synthesis field. LJ-Speech is a single speaker dataset that consists of 13,100 audio clips with a total length of approximately 24 hours. We use train-validation split from HiFi paper (Kong et al., 2020a) with sizes of 12950 train clips and 150 validation clips. Audio samples have a sampling rate of 22 kHz; it was used as-is.

**Bandwidth extension** We use publicly available dataset VCTK (Yamagishi et al., 2019) 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 (Valentini-Botinhao et al., 2017) 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 about the data can be found in the original paper.

### 5.2. Evaluation

**Objective evaluation** We use conventional metrics WB-PESQ (Rix et al., 2001), STOI (Taal et al., 2011), scale-invariant signal-to-distortion ratio (SI-SDR) (Le Roux et al., 2019) for objective evaluation of samples in the concerned tasks.

**MOSNet** In addition to conventional speech quality metrics, we considered absolute objective speech quality measure based on direct MOS score prediction (MOSNet) (Lo et al., 2019). We found that the original MOSNet model has poor speech quality assessment performance and is not able to detect obvious distortions (see Appendix C) and hypothesize that this is due to the outdated neural network architecture used for this model. Thus, we trained a modern neural network model wav2vec2.0 (Baevski et al., 2020) on the same data as MOSNet, i.e. large listening evaluation results released by the Voice Conversion Challenge 2018 (Lorenzo-Trueba et al., 2018). We call this model WV-MOS. Please

Table 2. Bandwidth extension results on VCTK dataset. \* indicates re-implementation.

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

refer to Appendix C for more details. We found this new model to be a more proficient predictor of subjective speech quality than PESQ, STOI, SI-SDR, and MOSNet (at system-level) and therefore report this metric for all our models.

**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. Please refer to Appendix D for further details.

### 5.3. Neural Vocoding

To demonstrate the effectiveness of the proposed HiFi++ framework for the neural vocoding both in terms of quality and computational efficiency we performed an extensive evaluation of the generated samples, and the complexity measurement. To assess the model quality we use the subjective MOS test and objective WV-MOS, STOI and PESQ metrics. To measure the model complexity we consider the size of the generator ( $G$  size) and its complexity in terms of the multiply-accumulate operations ( $G$  MACs). Also in the same manner we calculate the size and complexity of discriminators in total ( $D$  size and  $D$  MACs). To compare the empirical complexity of the training we measure the overall training time on single V100 GPU.

The results and comparison with existing baselines can be found in Table 1. Notably, HiFi++ model outperforms HiFi V1 significantly in terms of objective metrics WV-MOS, STOI and PESQ and achieves comparable or even better MOS. To explicitly compare subjectively HiFi++ and HiFi V1 we perform a pair-wise test and obtain statistically significant superiority of HiFi++ over HiFi V1 (p-value equals  $3 \cdot 10^{-3}$  for the binomial test, please refer to Appendix D for details). At the same time the HiFi++ generator has 1.72M parameters which is 8 times smaller than the HiFi generator size (13.92M). In terms of computational efficiency our

model also have 8 times less MACs than HiFi (2.78 GMACs vs 22.70 GMACs) and it finishes training more than 2 times faster (9.2 days vs 20.1 days). If we compare our HiFi++ with the smallest HiFi V2 model we observe that the former significantly surpass the latter in terms of quality metrics and still is more efficient in terms of the training time. These results show that HiFi++ architecture is more suitable for this task and allows much better trade-off between model quality and its computational complexity.

### 5.4. 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. Before subsampling signal to the desired sampling rate (2 kHz, 4 kHz, or 8 kHz) we apply a low-pass filter which is randomly selected among butterworth, chebyshev,essel, and elliptic filters of different orders to avoid aliasing and encourage model robustness (Liu et al., 2021). The sub-sampled signal is then resampled back to a 16 kHz sampling rate using polyphase filtering.

The results and comparison with other techniques are outlined in Table 2. Our model HiFi++ provides a better trade-off between model size and quality of bandwidth extension than other techniques. Specifically, our model is 5 times smaller than the closest baseline SEANet (Li et al., 2021) 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, please refer to Appendix D for details).

Importantly, these results highlight the importance of adversarial objectives for speech frequency bandwidth exten-

Table 3. Speech denoising results on Voicebank-DEMAND dataset. \* indicates re-implementation.

Model	MOS	WV-MOS	SI-SDR	STOI	PESQ	# Param (M)
Ground truth	4.60 $\pm$ 0.03	4.50	-	1.00	4.64	-
HiFi++ (ours)	<b>4.33 <math>\pm</math> 0.06</b>	4.14	18.4	<b>0.95</b>	2.70	<b>1.7</b>
VoiceFixer	<b>4.32 <math>\pm</math> 0.05</b>	4.14	-18.5	0.89	2.38	122.1
DEMUCS	4.22 $\pm$ 0.05	<b>4.37</b>	<b>18.5</b>	<b>0.95</b>	3.03	60.8
MetricGAN+	4.01 $\pm$ 0.09	3.90	8.5	0.93	<b>3.13</b>	2.7
*SEANet	3.99 $\pm$ 0.09	4.19	13.5	0.92	2.36	9.2
*SE-Conformer	3.39 $\pm$ 0.09	3.88	15.8	0.91	2.16	1.8
Input	3.36 $\pm$ 0.06	2.99	8.4	0.92	1.97	-

sion models. Surprisingly, the SEANet model (Li et al., 2021) appeared to be the strongest baseline among examined counterparts leaving the others far behind. This model uses adversarial objective similar to ours. The TFilm (Birnbaum et al., 2019) and 2S-BWE (Lin et al., 2021) models use supervised reconstruction objectives and achieve very poor performance, especially for low input frequency bandwidths.

### 5.5. Speech Enhancement

The comparison of the HiFi++ with baselines is demonstrated in the Table 3. Our model achieves comparable performance with VoiceFixer (Liu et al., 2021) and DEMUCS (Defossez et al., 2020) counterparts while being drastically smaller. 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 (Tagliasacchi et al., 2020) 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 MetricGAN+ model (Fu et al., 2021). While this model is explicitly trained to optimize PESQ and achieves superior values of this metric, this success does not spread on other objective and subjective metrics.

### 5.6. 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 slightly increased capacity to match the size of the initial HiFi++ architecture.

The results of the ablation study are shown in Table 4 (see Appendix A), which reveal how each module contributes to the HiFi++ performance. We can see that SpectralUNet and SpectralMaskNet is essential for neural vocoding because their absence notably degrades the model performance in terms of all three objective metrics. It can be explained by the fact that these modules operates mostly in frequency domain which is important for vocoding task. In contrast, the WaveUNet module shows a minor effect on the HiFi++ performance in the case of vocoding because in this task there are no any external distortions of the input. This module shows its main potential in the case of BWE or SE problems. Therefore, we perform an additional ablation study of the WaveUNet for the BWE problem and we obtain that it is essential for this task.

We also assess how HiFi model can be extended to other task besides the neural vocoding and implement the HiFi generator for the BWE task in a straightforward way by reducing the frequencies of the input mel-spectrogram (which we call "vanilla HiFi generator"). We see that such naive approach works poorly compared to HiFi++ model while it has two times bigger capacity.

## 6. Conclusion

In this work, we introduce the universal HiFi++ framework for neural vocoding, 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 all three tasks. Remarkably, our model obtains such results being much smaller (in some



cases by two orders of magnitude) and computationally more efficient than existing counterparts. Such strong performance across all three problems is based on the proposed HiFi++ generator architecture and the simplified multi-discrimination training.

## References

- Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.
- Bai, S., Kolter, J. Z., and Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.
- Birnbaum, S., Kuleshov, V., Enam, Z., Koh, P. W., and Ermon, S. Temporal film: Capturing long-range sequence dependencies with feature-wise modulations. *arXiv preprint arXiv:1909.06628*, 2019.
- Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.
- Défossez, A., Usunier, N., Bottou, L., and Bach, F. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*, 2019.
- Defossez, A., Synnaeve, G., and Adi, Y. Real time speech enhancement in the waveform domain. In *Interspeech*, 2020.
- Durugkar, I., Gemp, I., and Mahadevan, S. Generative multi-adversarial networks. *arXiv preprint arXiv:1611.01673*, 2016.
- Fu, S.-W., Liao, C.-F., Tsao, Y., and Lin, S.-D. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*, pp. 2031–2041. PMLR, 2019.
- Fu, S.-W., Yu, C., Hsieh, T.-A., Plantinga, P., Ravanelli, M., Lu, X., and Tsao, Y. Metricgan+: An improved version of metricgan for speech enhancement. *arXiv preprint arXiv:2104.03538*, 2021.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pp. 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015.
- Ito, K. and Johnson, L. The lj speech dataset. <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., and Weyde, T. Singing voice separation with deep u-net convolutional networks. 2017.
- Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A., Dieleman, S., and Kavukcuoglu, K. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pp. 2410–2419. PMLR, 2018.
- Kim, E. and Seo, H. SE-Conformer: Time-Domain Speech Enhancement Using Conformer. In *Proc. Interspeech 2021*, pp. 2736–2740, 2021. doi: 10.21437/Interspeech.2021-2207.
- Kim, S. and Sathe, V. Bandwidth extension on raw audio via generative adversarial networks. *arXiv preprint arXiv:1903.09027*, 2019.
- Kong, J., Kim, J., and Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *arXiv preprint arXiv:2010.05646*, 2020a.
- Kong, Q., Cao, Y., Liu, H., Choi, K., and Wang, Y. Decoupling magnitude and phase estimation with deep resunet for music source separation. *arXiv preprint arXiv:2109.05418*, 2021.
- Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020b.
- Kuleshov, V., Enam, S. Z., and Ermon, S. Audio super resolution using neural networks. *arXiv preprint arXiv:1708.00853*, 2017.
- Kumar, K., Kumar, R., de Boissiere, T., Geste, L., Teoh, W. Z., Sotelo, J., de Brébisson, A., Bengio, Y., and Courville, A. Melgan: Generative adversarial networks for conditional waveform synthesis. *arXiv preprint arXiv:1910.06711*, 2019.
- Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pp. 1558–1566. PMLR, 2016.
- Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. Sdr-half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630. IEEE, 2019.
- Li, N., Liu, Y., Wu, Y., Liu, S., Zhao, S., and Liu, M. Robusttrans: A robust transformer-based text-to-speech model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8228–8235, 2020.

- Li, Y., Tagliasacchi, M., Rybakov, O., Ungureanu, V., and Roblek, D. Real-time speech frequency bandwidth extension. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 691–695. IEEE, 2021.
- Lin, J., Wang, Y., Kalgaonkar, K., Keren, G., Zhang, D., and Fuegen, C. A two-stage approach to speech bandwidth extension. *Proc. Interspeech 2021*, pp. 1689–1693, 2021.
- Liu, H., Kong, Q., Tian, Q., Zhao, Y., Wang, D., Huang, C., and Wang, Y. Voicefixer: Toward general speech restoration with neural vocoder. *arXiv preprint arXiv:2109.13731*, 2021.
- Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., and Wang, H.-M. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*, 2019.
- Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., and Ling, Z. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*, 2018.
- Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017.
- Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pp. 3918–3926. PMLR, 2018.
- Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
- Pascual, S., Bonafonte, A., and Serra, J. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.
- Prenger, R., Valle, R., and Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3617–3621. IEEE, 2019.
- Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. 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)*, volume 2, pp. 749–752. IEEE, 2001.
- Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.
- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.
- Shen, J., Jia, Y., Chrzanowski, M., Zhang, Y., Elias, I., Zen, H., and Wu, Y. Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling. *arXiv preprint arXiv:2010.04301*, 2020.
- Stoller, D., Ewert, S., and Dixon, S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.
- Sulun, S. and Davies, M. E. On filter generalization for music bandwidth extension using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 15 (1):132–142, 2020.
- Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.
- Tagliasacchi, M., Li, Y., Misiunas, K., and Roblek, D. Seanet: A multi-modal speech enhancement network. *arXiv preprint arXiv:2009.02095*, 2020.
- Tan, K. and Wang, D. A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech*, pp. 3229–3233, 2018.
- Tian, Q., Chen, Y., Zhang, Z., Lu, H., Chen, L., Xie, L., and Liu, S. Tfgan: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis. *arXiv preprint arXiv:2011.12206*, 2020.
- Valentini-Botinhao, C. et al. Noisy speech database for training speech enhancement algorithms and tts models. 2017.
- Wang, H. and Wang, D. Time-frequency loss for cnn based speech super-resolution. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 861–865. IEEE, 2020.
- Wang, H. and Wang, D. Towards robust speech super-resolution. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

- Wisdom, S., Hershey, J. R., Wilson, K., Thorpe, J., Chinen, M., Patton, B., and Saurous, R. A. Differentiable consistency constraints for improved deep speech enhancement. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 900–904. IEEE, 2019.
- Yamagishi, J., Veaux, C., MacDonald, K., et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). 2019.
- You, J., Kim, D., Nam, G., Hwang, G., and Chae, G. Gan vocoder: Multi-resolution discriminator is all you need. *arXiv preprint arXiv:2103.05236*, 2021.
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

## A. Ablation Study

In the introduction it was mentioned that the ablation study of MPD discriminators from (Kong et al., 2020a) is confusing because hyperparameters were selected poorly. To support our statement empirically we perform such ablation study and show that HiFi V3 can be trained without MPD discriminators while achieving the same model quality.

Table 4. Ablation study

Model	WV-MOS	STOI	PESQ	$G$ size (M)	$D$ size (M)
<b>Vocoding on LJSpeech</b>					
Ground Truth	4.23	1.00	4.64		
Baseline (HiFi++)	4.09	0.98	3.44	1.72	1.86
w/o SpectralUNet	4.04	0.97	3.38	1.74	1.86
w/o WaveUNet	4.08	0.98	3.60	1.77	1.86
w/o SpectralMaskNet	4.01	0.96	2.91	1.76	1.86
HiFi V3	4.02	0.95	2.68	1.46	48.5
HiFi V3 w/o MPD	4.01	0.96	2.72	1.46	7.41
<b>BWE (1kHz) on VCTK</b>					
Ground Truth	4.17	1.00	4.64		
Baseline (HiFi++)	3.70	0.86	1.74	1.72	1.86
w/o WaveUNet	3.55	0.79	1.40	1.75	1.86
Vanilla HiFi generator	3.43	0.44	1.41	3.43	1.86

## B. Architecture details

### B.1. HiFi-GAN (small)

We use  $V2$  configuration of the original HiFi-GAN generator for the mel-spectrogram upsampling part of our model (HiFi-GAN small). We outline its architecture on the Figures 2 and 3.

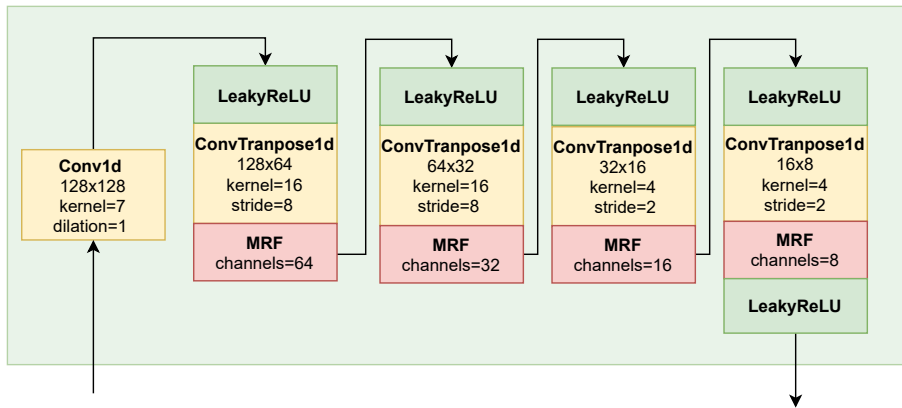


Figure 2. HiFi-GAN (small) architecture.

### B.2. WaveUNet and SpectralUNet

We use standard fully-convolutional multiscale encoder-decoder architecture for WaveUNet and SpectralUNet networks. The architectures of these networks are depicted on Figures 4 and 5. Each downsampling block of WaveUNet model



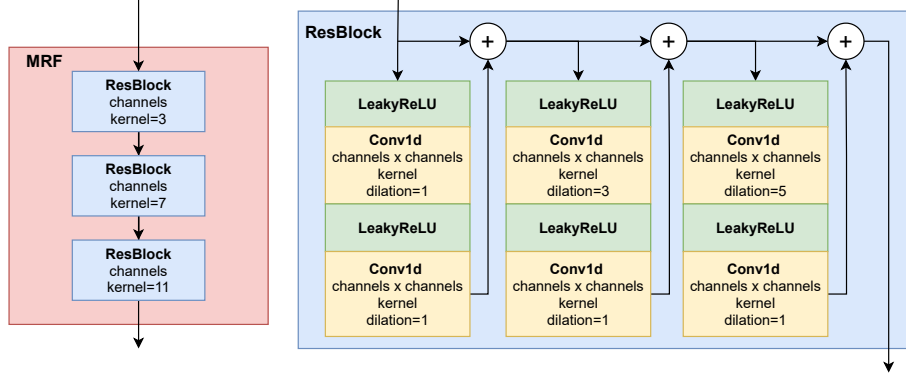
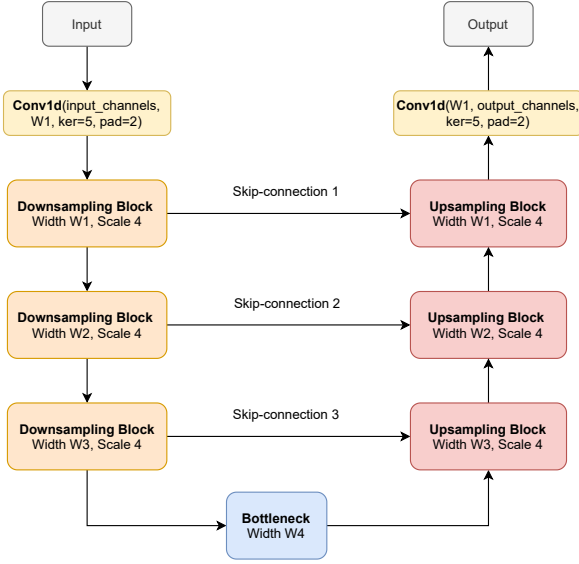
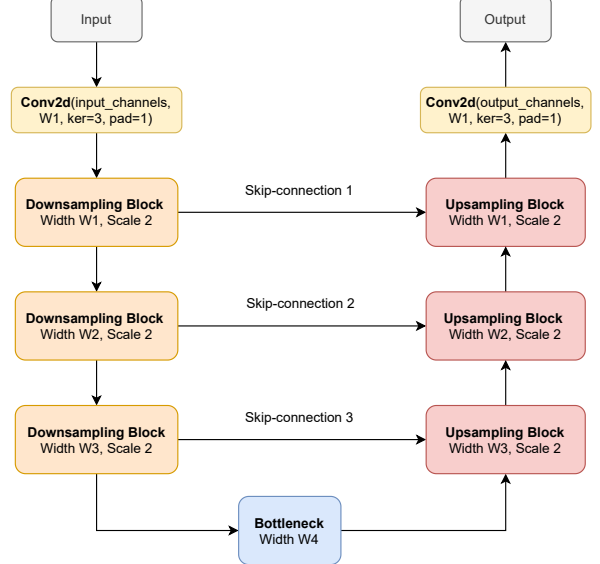


Figure 3. HiFi-GAN generator blocks.

performs  $\times 4$  downsampling of the signal across time dimension. Analogously, each downsampling block in SpectralUNet downscales signal  $\times 2$  across time and frequency dimensions. The width  $[W1, W2, W3, W4]$  and block depth parameters control number of parameters and computational complexity of resulting networks. The structure of WaveUNet blocks is outlined on the Figure 6.


 Figure 4. The architecture of WaveUNet model. Block widths  $[W1, W2, W3, W4]$  are equal to  $[10, 20, 40, 80]$ .

 Figure 5. The architecture of SpectralUNet model. Block widths  $[W1, W2, W3, W4]$  are equal to  $[8, 12, 24, 32]$ .

## C. WV-MOS

### C.1. MOSNet

The quality assessment of generated speech is an important problem in the speech processing domain. The popular objective metrics such as PESQ, STOI, and SI-SDR are known to be poorly correlated with subjective quality assessment. Meanwhile, mean opinion score (MOS) — an audio quality metric that is measured from human annotators' feedback is a de-facto standard for speech synthesis systems benchmarking (Kong et al., 2020a) where no objective metrics are usually reported. Not only is obtaining MOS expensive, but also it leads to incomparable (from paper to paper) results due to different sets of annotators participating in evaluations.

One recent attempt to address this problem was performed by Lo et al. (2019). The paper proposes a deep learning-based

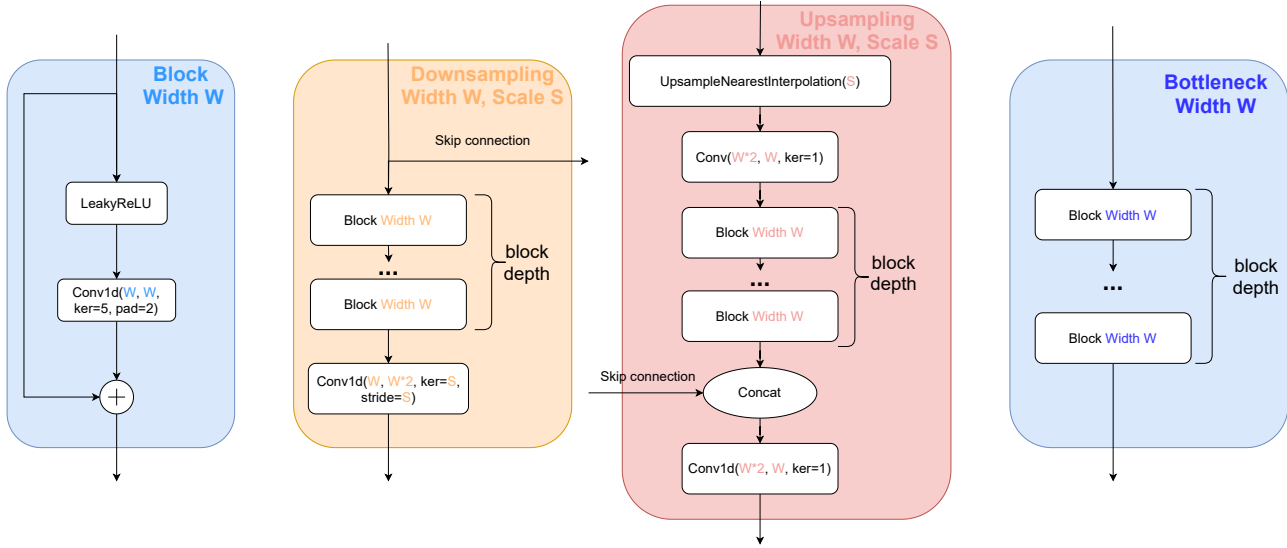


Figure 6. WaveUNet blocks. WaveUNet and SpectralUNet blocks share the same architectural structure except SpectralUNet uses 2d convolutions with kernel size  $3 \times 3$  instead of 1d convolutions with kernel size being equal to 5. Block depth is equal to 4.

objective assessment to model human perception in terms of MOS, referred to as MOSNet. MOSNet used raw magnitude spectrogram as the input feature and three neural network-based models, namely CNN, BLSTM, and CNN-BLSTM are used to extract valuable features from the input and fully connected (FC) layers and pooling mechanisms to generate predicted MOS. For training, the authors use MSE Loss with MOS evaluations of VCC 2018 as the targets. Among considered architectures of feature extractors, a combination of CNN with BLSTM turned out to be the most preferable on most of the test metrics. The advantages of MOSNet over conventional objective speech quality metrics (e.g. PESQ, STOI) are twofold. Firstly, it doesn't require a reference signal for quality assessment and predicts MOS score directly from the generated audio. Second, deep learning models have great potential in the prediction of subjective quality measurements as it was earnestly shown in the computer vision domain (Zhang et al., 2018).

### C.2. MOSNet is not able to identify obvious corruptions

Nevertheless, we found MOSNet model to be a low proficiency speech quality predictor. A simple yet demonstrative example is that MOSNet fails to identify records corrupted by low-pass filters. We applied low pass filters of different cutoff frequencies (1 kHz, 2 kHz, and 4 kHz) to 50 samples from the VCTK dataset and measured the ratio of cases where MOSNet assigned a higher score to the corrupted sample than to the reference one. We also examined additive noise as corruption and measured the analogous failure rate for the Voicebank-Demand test part. The results can be found in Table 5.

### C.3. WV-MOS: a new absolute objective speech quality measure

We hypothesized that the MOSNet failure can be to a significant extent attributed to the outdated neural network architecture and lack of pretraining. For this reason, we propose to utilize for MOS prediction a modern neural network architecture wav2vec2.0 (Baevski et al., 2020). Wav2vec2.0 model is pretrained in a contrastive self-supervised manner, thus, its representations are task-agnostic and can be useful for a variety of downstream tasks.

Following Lo et al. (2019), we use listening evaluation results released by Voice Conversion Challenge 2018 (Lorenzo-Trueba et al., 2018) for model training. We augment pretrained wav2vec2.0 model with a 2-layer MLP head and train the resulting neural network to predict assigned MOS scores. We use mean squared error as training objective, the batch size is equal to 64, the learning rate was set to 0.0001. As a result, we observe higher Spearman's rank correlation factors on VCC 2018 test set (0.62/0.93 for our model versus 0.59/0.90 for MOSNet on utterance/system levels, respectively). More importantly, we observe our model to better generalize to the audio samples beyond VCC 2018 (see Table 5 and Table 6).

We use crowd-sourced studies conducted during this work for validation of the proposed model. Based on crowd-sourced

Table 5. MOS prediction assessment results (toy example). The MOSNet fails to identify such obvious corruption as the application of 4kHz low-pass filters.

Data	Error rate (%) (MOSNet)	Error rate (%) (WV-MOS)
Low-pass filter (1 kHz)	0	0
Low-pass filter (2 kHz)	2.1	0
Low-pass filter (4 kHz)	34.0	0
Voicebank-demand	12.3	3.6

MOS scores we computed utterance-level (i.e., assigned MOSes are averaged for each sample) and system-level (i.e., assigned MOSes are averaged for each model) correlations of the predicted MOS scores with the ones assigned by referees. The results are demonstrated in the Table 6. Interestingly, at the utterance level, the WV-MOS score is correlated much better with the assigned MOS-es for bandwidth extension task while being inferior in this sense to PESQ for vocoding and speech enhancement tasks. At the system level, the situation is more unambiguous as WV-MOS outperforms all examined metrics in terms of Spearman’s correlation with assigned MOS scores.

Table 6. MOS prediction assessment results (crowd-sourced studies). The table shows Spearman’s rank correlations with crowd-sourced studies. The WV-MOS scores tend to have a much better correlation with crowd-sourced MOSes than the original MOSNet and better system-level correlations than all other examined metrics.

Data	Utterance-level					System-level				
	WV-MOS	MOSNet	PESQ	STOI	SI-SDR	WV-MOS	MOSNet	PESQ	STOI	SI-SDR
Vocoding	0.47	0.03	<b>0.73</b>	0.41	0.29	<b>0.96</b>	0.11	0.89	0.45	0.57
BWE (1 kHz)	<b>0.86</b>	0.52	-0.07	0.38	-0.02	<b>0.97</b>	0.67	-0.17	0.55	-0.07
BWE (2 kHz)	<b>0.85</b>	0.36	0.11	0.50	0.01	<b>0.90</b>	0.71	-0.02	0.57	-0.12
BWE (4 kHz)	<b>0.69</b>	0.25	0.08	0.24	-0.16	<b>0.86</b>	0.40	0.38	0.29	-0.38
SE	0.64	0.44	<b>0.67</b>	0.42	0.45	<b>0.68</b>	0.61	0.5	0.32	0.43

## D. Subjective evaluation

We measure mean opinion score (MOS) of the model using a crowd-sourcing adaptation of the standard absolute category rating procedure. Our MOS computing procedure is as follows.

1. Select a subset of 40 random samples from the test set (once per problem, i. e. for vocoding, bandwidth extension, speech enhancement).
2. Select a set of models to be evaluated; inference their predictions on the selected subset.
3. Randomly mix the predictions and split them into the pages of size 20 almost uniformly. Almost uniformly means that on each page there are at least  $\lfloor \frac{20}{\text{num.models}} \rfloor$  samples from each model.
4. Insert additional 4 trapping samples into random locations on each page: 2 samples from groundtruth, and 2 samples of a noise without any speech.
5. Upload the pages to the crowd-sourcing platform, set the number of assessors for each page to at least 30. Assessors are asked to work in headphones in a quiet environment; they must listen the audio until the end before assess it.
6. Filter out the results where the groundtruth samples are assessed with anything except 4 (good) and 5 (excellent), or the samples without voice are assessed with anything except 1 (bad).
7. Split randomly the remaining ratings for each model into 5 almost-equal-size groups, compute their mean and std.

Since the models are distributed uniformly among the pages, assessor’s biases affect all models in the same way, so the relative order of the models remains. On the other hand, assessor will have access to all variety of the models on one page and thus can scale his ratings better. The other side is that the models rating are not independent from each other in this setting, because assessors tend to estimate the sample quality relatively to the average sample of the page, i. e. the more bad models are in comparison – the bigger MOSes are assigned to the good ones. 4 trapping samples per page is also a reasonable choice, because one cannot just random guess the correct answers for these questions.

The drawback of MOS is that sometimes it requires too much assessors per sample to determine confidently which model is better. The possible solution is to use a simplified version of comparison category rating, i. e. preference test. This test compares two models, assessor is asked to chose which model produces the best output for the same input. If assessor doesn’t hear the difference, the option “equal” must be selected.

1. Select a subset of 40 random samples from the test set.
2. Randomly shuffle this set split it into the pages of size 20.
3. Select randomly 10 positions on each page where Model1’s prediction will be first.
4. Insert additional 4 trapping samples into random locations on each page: each trapping sample is a pair of a clean speech from groundtruth and its noticeable distorted version. The order of models in trapping sample is random, but on each page there are 2 samples with one order and 2 samples with another.
5. Upload the pages to the crowd-sourcing platform, set the number of assessors for each page to at least 30. Assessors are asked to work in headphones in a quiet environment; they must listen the audio until the end before assess it.
6. Filter out the results where the trapping samples are classified incorrectly.
7. Use sign test to reject the hypothesis that the models generate the speech of the same [median] perceptual quality.

## E. Implementation details

### E.1. Results reproduction

As a part of this submission supplementary material, we provide all source codes that are needed to train and infer our models. We also attach configuration files that contain all the necessary information regarding the model’s specification and hyperparameters.

### E.2. Baselines

We re-implement the 2S-BWE model (Lin et al., 2021) closely following the description provided in the paper. We adopt the method for 1 kHz and 2kHz input bandwidths following the same logic as in 4 kHz. The first minor difference for these cases is that we mirror phase spectrogram 2 times for 2 kHz (from 2 kHz to 4 Khz, then the resulting spectrogram is mirrored again to obtain 8 kHz bandwidth) and analogously 3 times for 1 kHz. The second difference is that the TCN network needs to output more channels for 1 kHz and 2 kHz as the larger spectrogram needs to be predicted. Thus, we change the number of the output channels of the last TCN layer. All 2S-BWE models are trained with the same set of hyperparameters described in the paper.

We implement the SEANet model following the original paper (Tagliasacchi et al., 2020). For a fair comparison with our work, we did not restrict the model to be streaming and did not reduce the number of channels as described by Li et al. (2021). We run the SEANet model for the same number of iterations as the HiFi++, besides that we use the same hyperparameters as described in the original paper.

We also tried to reproduce results of Kim & Seo (2021) (SE-Conformer). Noteworthy, we didn’t obtain metrics that are close to the reported in the article.

For comparison with HiFi-GAN (Kong et al., 2020a), MelGAN (Kumar et al., 2019), Waveglow (Prenger et al., 2019), TFilm (Birnbaum et al., 2019), VoiceFixer (Liu et al., 2021), MetricGAN+ (Fu et al., 2021), DEMUCS (Défossez et al., 2019) we employ the official implementations provided by the authors.