Bandwidth Extension is All You Need

Paper ID: 4656

Jiaqi Su[1,2], Yunyun Wang[1], Adam Finkelstein[1], Zeyu Jin[2]
[1]Princeton University, [2]Adobe Research

Introduction

Many audio processing/synthesis applications target **16kHz**.

- Computationally easier target (relative to **48kHz**).
- Less immersive listening experience.

We propose:

- BWE as postprocessing for bridging from 8k/16kHz to 48kHz
- A waveform BWE method using GAN that achieves audio quality typically indistinguishable from real 48kHz audio.

Exisiting BWE methods: **limited extension up to 16kHz**

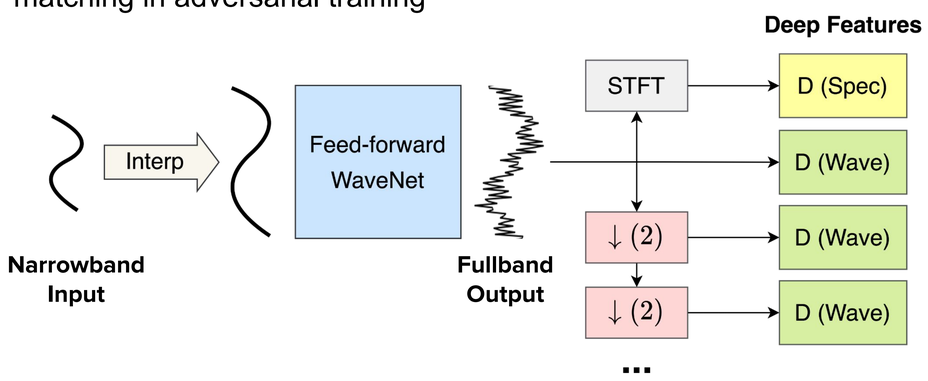
- Spectral methods compensate energy, but over-smooth spectrogram and introduce artifacts by phase approximation.
- Waveform methods still not close to real high-quality.

Evlauations show that our BWE method:

- Achieve close to real 48kHz audio quality for 16k-to-48k BWE; greatly improve over previous methods for 8k-to-48k BWE.
- Bring consistent quality boost to denoisers and vocoders.

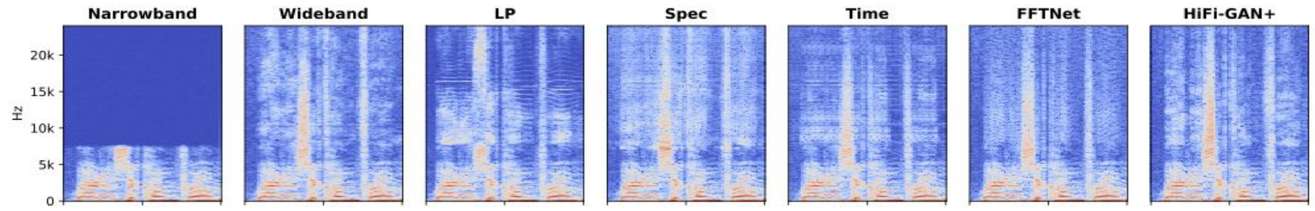
Method

Adapt from **HiFi-GAN [5]**: Feed-forward WaveNet with deep feature matching in adversarial training



- Discriminator on 128-coefficient log mel spectrogram
- Multi-scale discriminators on downsampled waveforms
- L1 loss, spectrogram losses, and feature matching loss with deep features of the discriminators
- Weight normalization to speed up convergence

Experiments



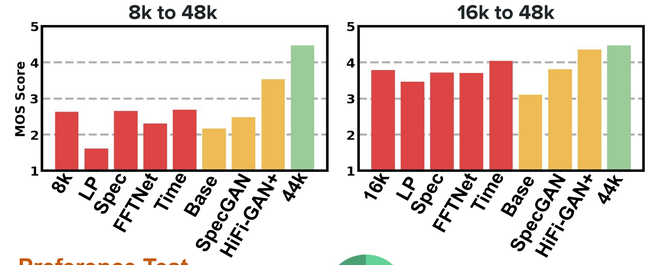
- ❖ **HiFi-GAN+**: our full approach
  - ➢ **Base:** feed-forward WaveNet
  - ➢ **SpecGAN:** spec discriminator only
- ❖ **Time:** time-domain EnvNet with GAN [28]
- ❖ **FFTNet:** FFTNet variant [6]
- ❖ **LP:** linear prediction based analysis synthesis [15]
- ❖ **Spec:** spectral 1D conv U-Net with GAN [20]

Experiment 1: BWE baseline comparison

Method	PSNR↑	LSD↓	PSNR	LSD	PSNR	LSD	PSNR	LSD
Input SR	8k		16k		8k		16k	
	VCTK Dataset				DAPS Dataset			
NB Input	**38.56**	15.81	**44.40**	14.84	35.95	12.87	**41.98**	11.50
LP	15.74	4.06	15.74	3.83	15.78	5.00	13.73	4.61
Spec	26.19	2.42	35.74	2.06	**36.26**	3.06	40.65	2.58
Time	22.99	2.03	29.90	1.92	31.60	2.82	31.07	3.10
FFTNet	36.33	**2.00**	40.59	**1.67**	35.38	2.80	39.62	2.44
Base	31.70	2.26	32.40	2.03	29.26	**2.67**	30.08	**2.34**
SpecGAN	12.75	2.15	31.78	1.95	10.57	2.85	26.56	2.45
HiFi-GAN+	33.53	2.13	32.16	1.83	30.60	2.80	29.28	2.35

➢ Numerical values of objective measures do not correlate well with perceptual quality.

Mean Opinion Score Test



Preference Test

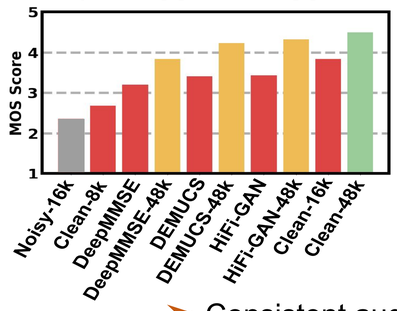
57.4% prefer 48kHz ground truth

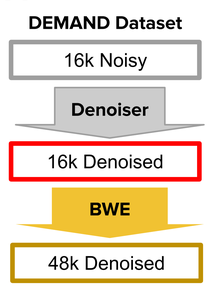
42.6% prefer HiFiGAN+
= 85.2% no preference

➢ 16k-to-48k BWE by HiFi-GAN+ is typically indistinguishable from real 48kHz.

Experiment 2: BWE for denoising

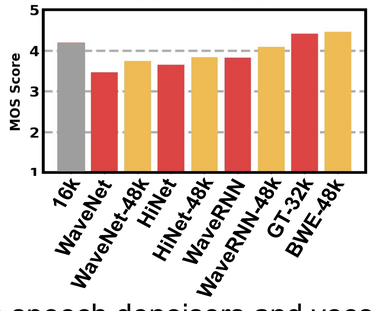
DeepMMSE [8], DEMUCS [7], HiFi-GAN [5]

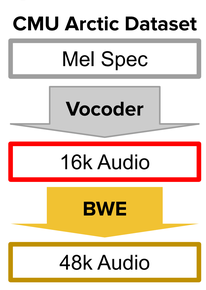




Experiment 3: BWE for vocoding

WaveNet [1], HiNet [13], WaveRNN [12]





➢ Consistent audio quality boost for various speech denoisers and vocoders

Conclusions

- Our waveform-to-waveform bandwidth extension method based on HiFi-GAN can achieve audio quality typically indistinguishable from real 48kHz audio.
- Applying BWE as post-processing can consistently boost quality in a wide variety of audio applications (denoisers, vocoders etc.).
- Exisiting objective measures do not correlate well with perceptual quality (for the 48kHz BWE task).