2021 IEEE International Conference on Acoustics, Speech and Signal Processing

Bandwidth Extension is All You Need

Jiaqi Su[1,2], Yunyun Wang[1], Adam Finkelstein[1], Zeyu Jin[2]

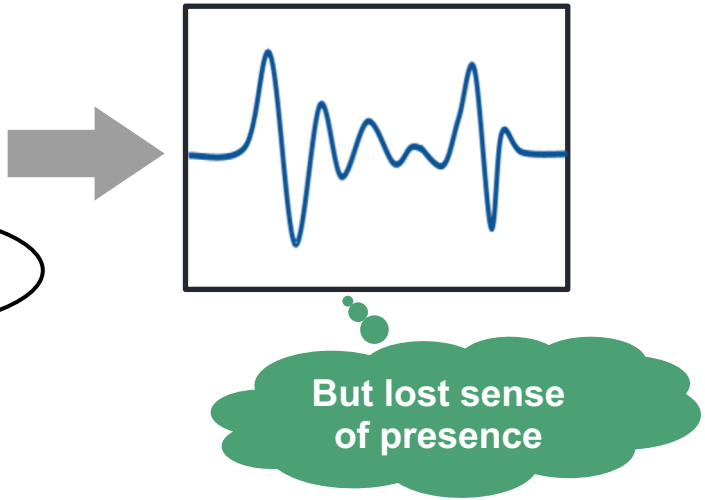
[1] Princeton University
[2] Adobe Research

Motivation

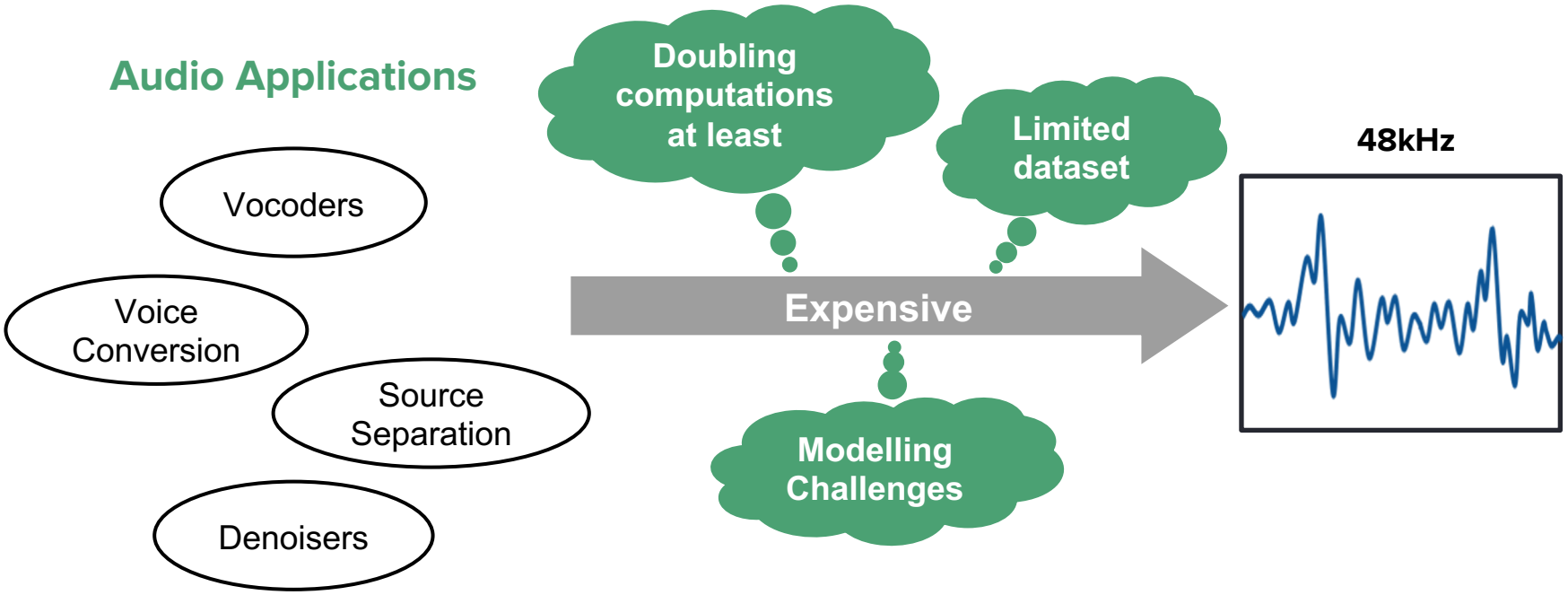
Audio Applications

- Vocoders
- Voice Conversion
- Source Separation
- Denoisers

8k, 16kHz



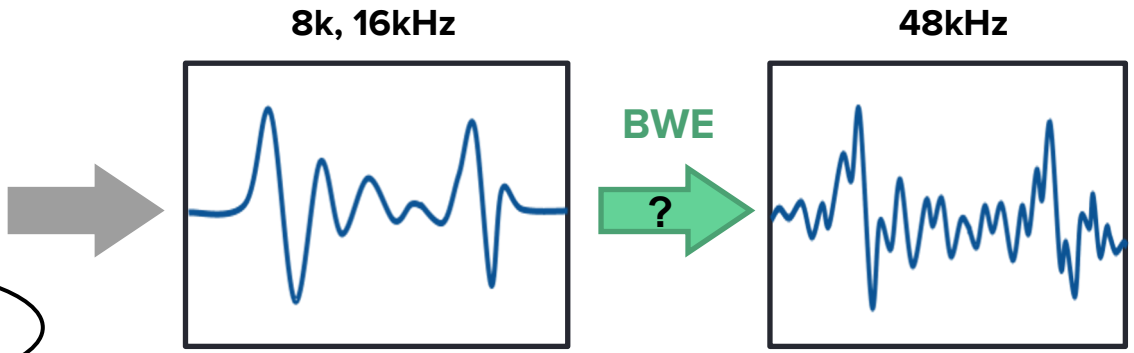
Motivation


Audio Applications
Vocoders
Voice Conversion
Source Separation
Denoisers
Doubling computations at least
Limited dataset
Expensive
Modelling Challenges
48kHz

Motivation

Audio Applications

- Vocoders
- Voice Conversion
- Source Separation
- Denoisers



Bandwidth extension is all you need!

Previous Work: Bandwidth Extension

Most limited to 8-16kHz for wideband

- **Traditional Signal Processing Methods**
  NMF *{Bansal 2005}*, LPC *[Bachhav 2018]*, HMMs *[Jax 2003]*, GMMs *[Seo 2014]*
- **Learning-based Spectral Methods**
  DNN *[Li 2015]*, Variational Auto-Encoders *[Bachhav 2020]*, U-Nets *[Eskimez 2019]*, RNN *[Schmidt 2018]*

➤ ***Over-smoothing details, phase approximation***

- **Learning-based Waveform Methods**
  - Audio super resolution *[Kuleshov 2017]*
  - WaveNet *[Wang 2018, Gupta 2019]*, Hierarchical RNN *[Ling 2018]*, EnvNet *[Li 2019]*, Time-Frequency Networks *[Lim 2018]*, time-frequency losses *[Wang 2020]*
  - FFTNet with perceptual loss *[Feng 2019]* ⟹ **Only method reaching 44kHz**

Previous Work: Generative Adversarial Networks

- **GAN in Bandwidth Extension**
  Simple discriminators on spectral features *[Li 2018, Eskimez 2019, Bachhav 2020]*

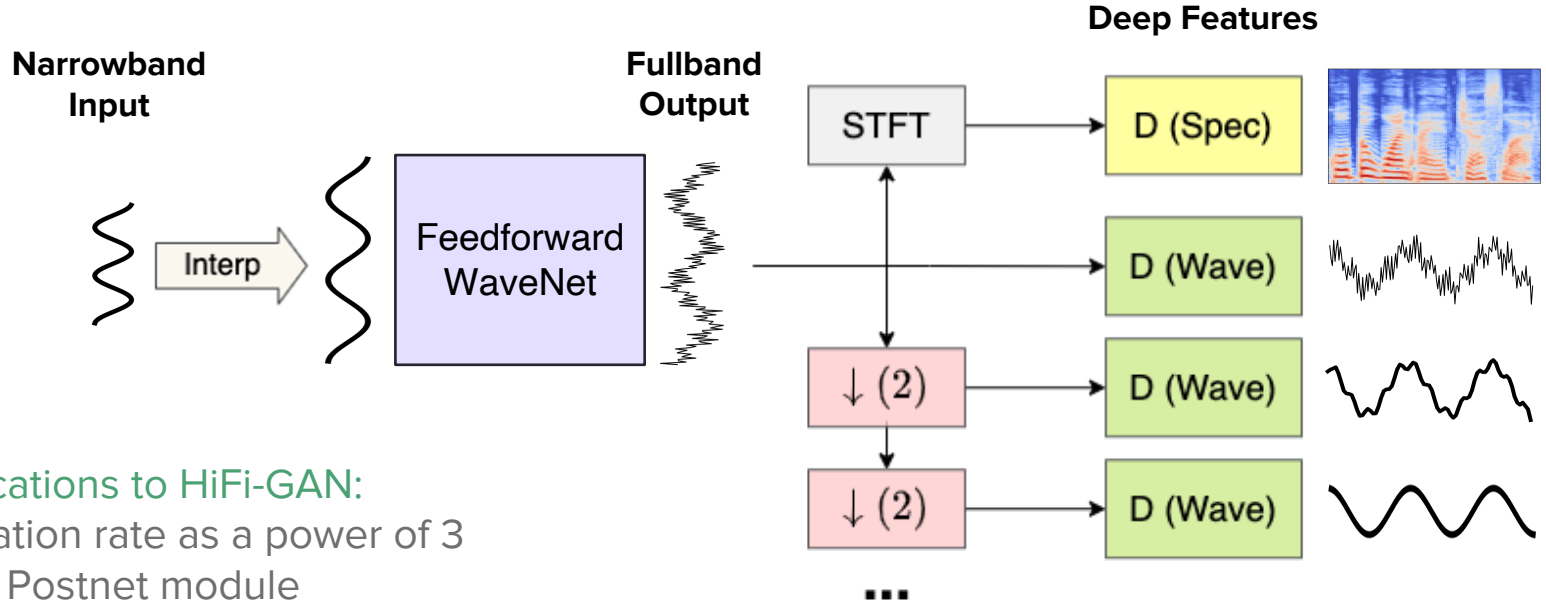
  ➤ ***Waveform discriminator rarely used***

- **GAN in other speech processing domains**
  - MelGAN *[Kumar 2019]*: feature matching loss of discriminators
  - HiFi-GAN *[Su 2020]:* multi-domain discriminators

  ➤ ***Similar ideas can apply to BWE problem***

Method

Adapt from **HiFi-GAN *[Su 2020]***: Feed-forward WaveNet with deep feature matching in adversarial training

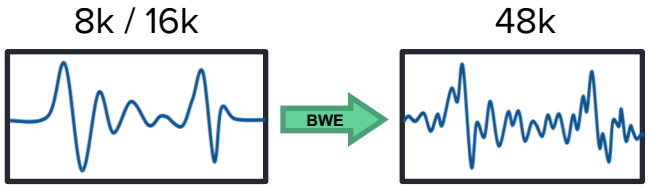


Modifications to HiFi-GAN:

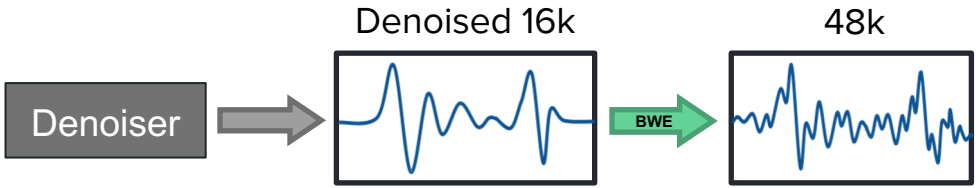
- Dilation rate as a power of 3
- No Postnet module
- Weight normalization

Experiments

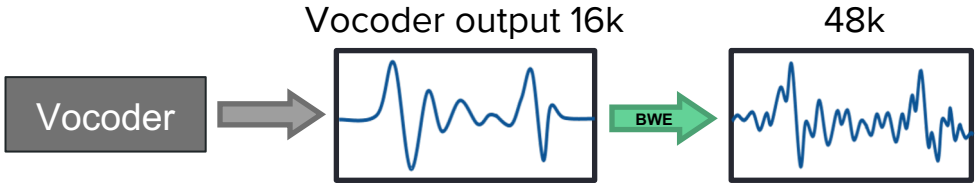
1. Clean speech bandwidth extension baseline comparison study



2. Bandwidth extension for speech denoising



3. Bandwidth extension for waveform generation



Experiments: BWE

- Baselines
  - **8k:** input (ground truth downsampled to 8k)
  - **16k:** input (ground truth downsampled to 16k)
  - **44k:** ground truth
  - **HiFi-GAN+**: our full approach
    - ❖ **Base:** feed-forward WaveNet
    - ❖ **SpecGAN:** use of the spectrogram discriminator only

  Ours

  - **LP:** linear prediction based analysis synthesis *[Bachhav 2018]*
  - **Spec:** a spectral-domain method using 1D conv U-Net with GAN *[Eskimez 2019]*
  - **Time:** a time-domain method using EnvNet structure with GAN *[Li 2019]*
  - **FFTNet:** FFTNet variant for BWE *[Feng 2019]*

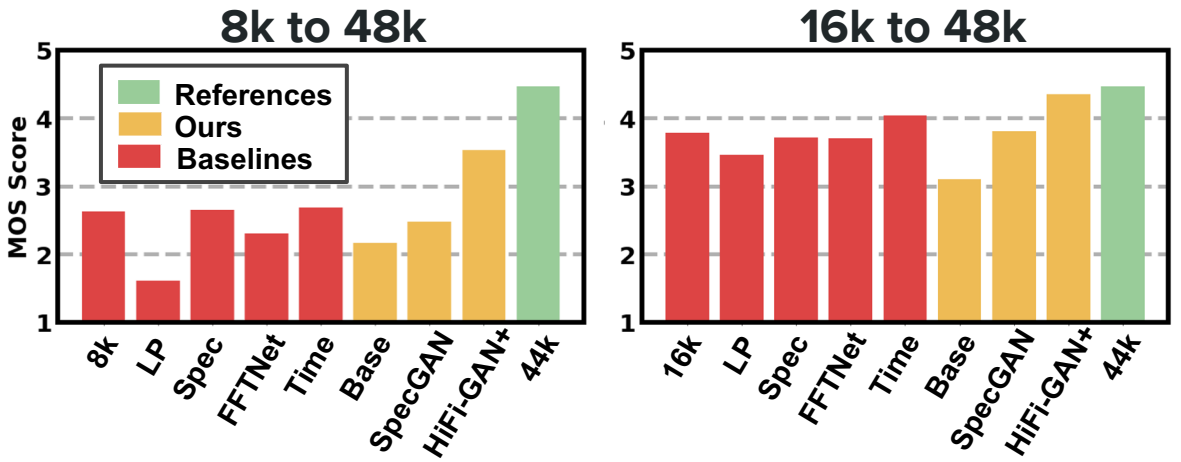
  Baselines
- Dataset
  - Train: VCTK dataset *[Veaux 2016]*
  - Test: Device and Produced Speech (DAPS) clean set *[Mysore 2015]*

Experiments: BWE - Objective Evaluations

Observation: Objective metrics do not correlate well with perceptual quality.

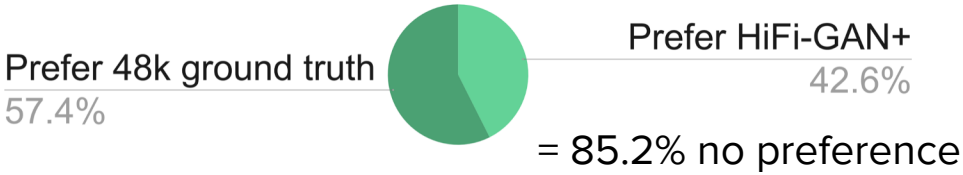
Experiments: BWE - Subjective Evaluations

Amazon Mechanical Turk



MOS Test

- 382 subjects
- 23,400 ratings

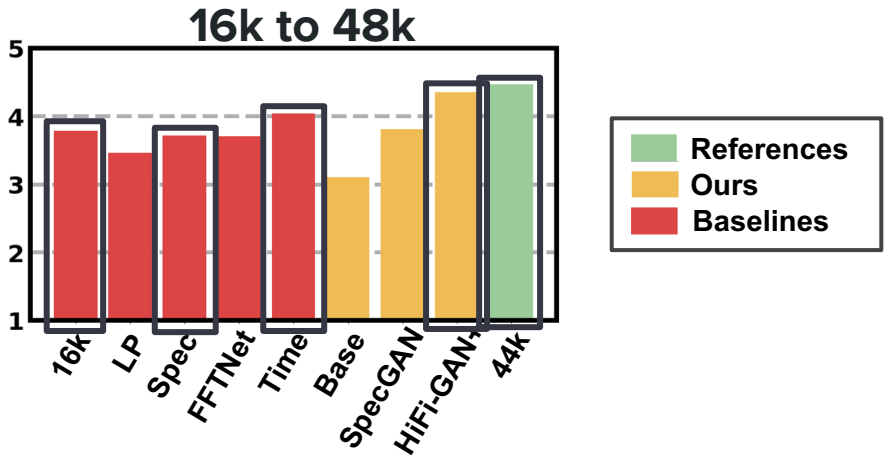


Preference test

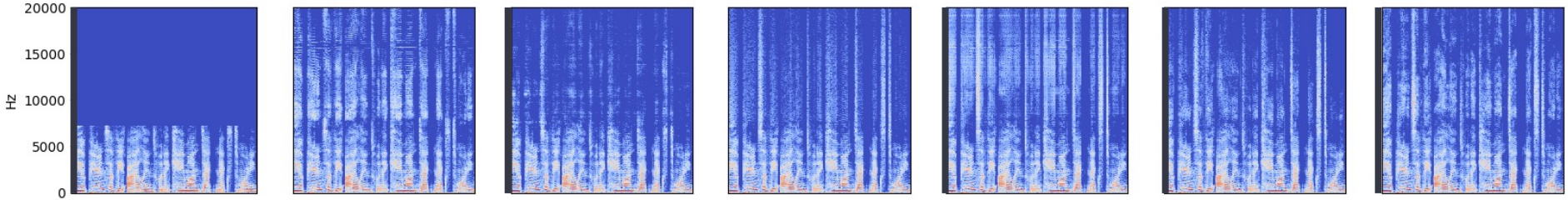
- 200 subjects
- 2,675 answers

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

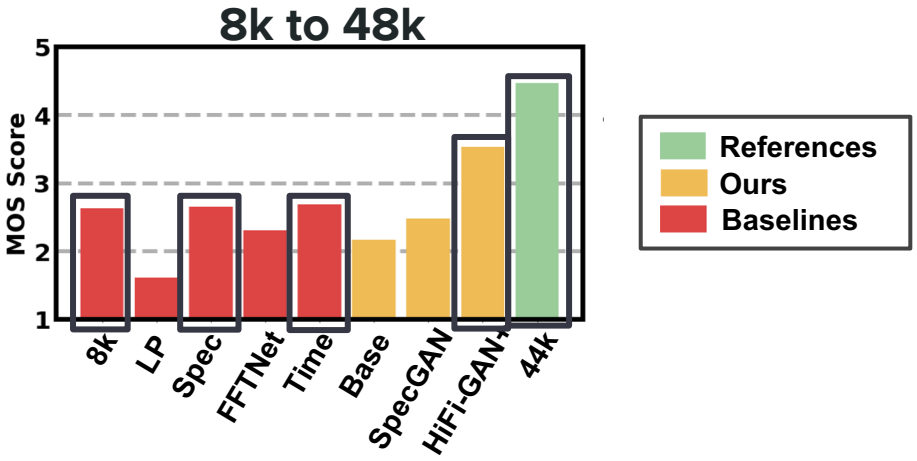
Experiments: BWE - Demo

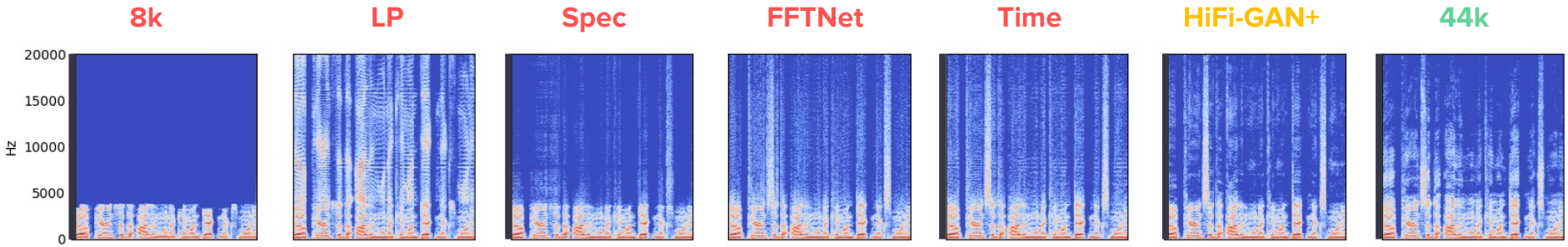


16k LP Spec FFTNet Time HiFi-GAN+ 44k



Experiments: BWE - Demo

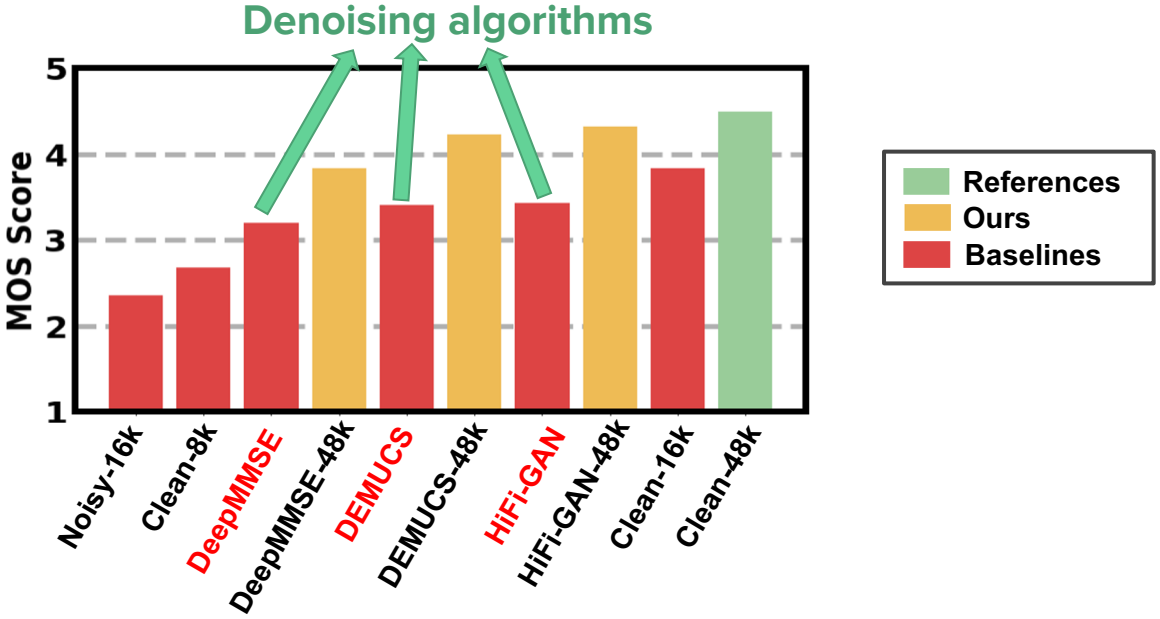




Experiments: BWE for Denoising

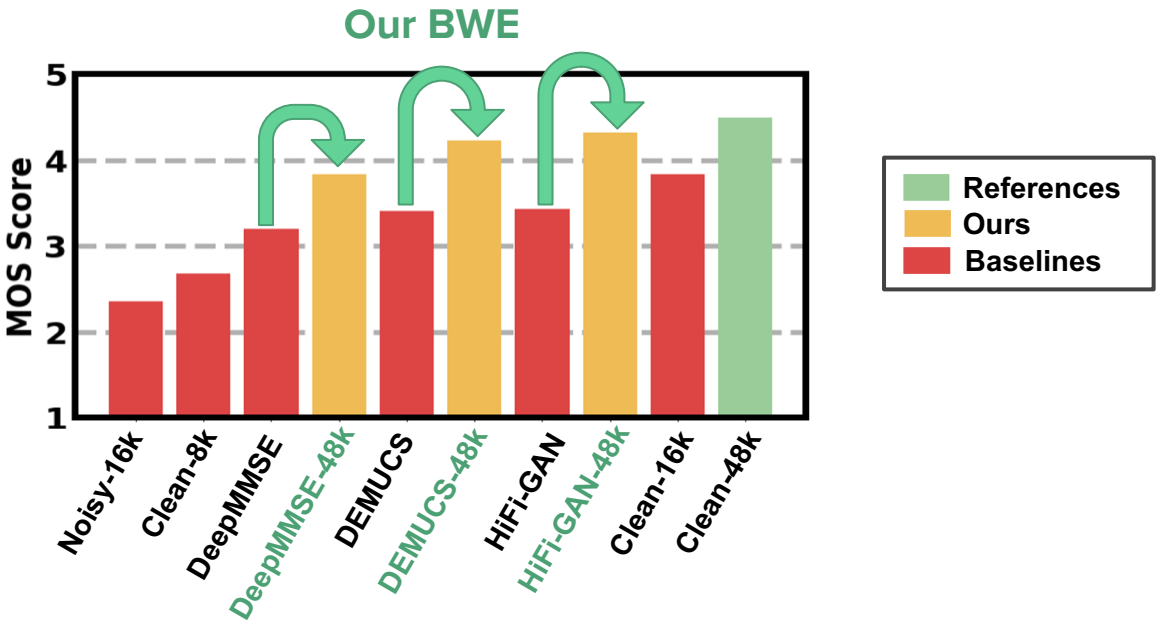
DEMAND Dataset → 16k Noisy → **Denoiser** → 16k Denoised → **BWE** → 48k Denoised

MOS Test



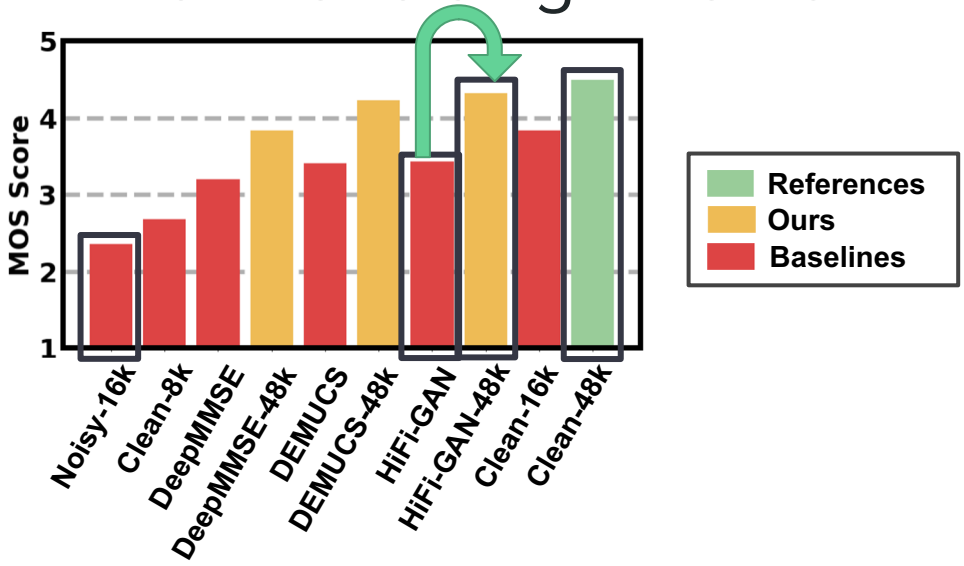
Experiments: BWE for Denoising

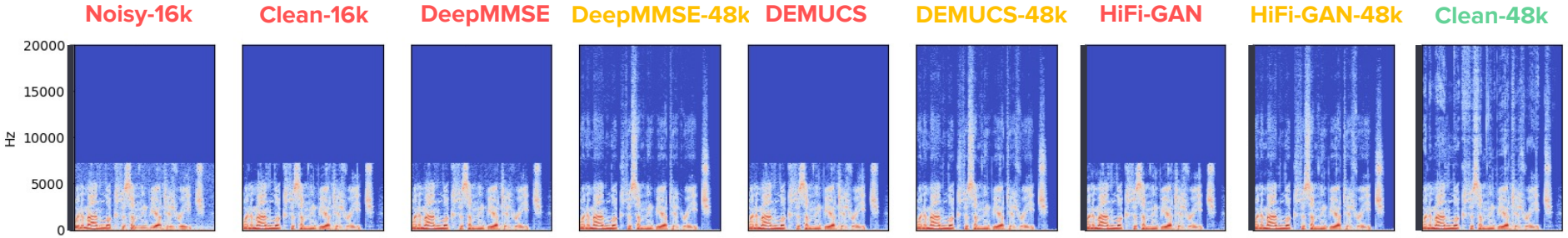
MOS Test



Observation: Consistent quality boost to enhancement algorithms

Experiments: BWE for Denoising - Demo



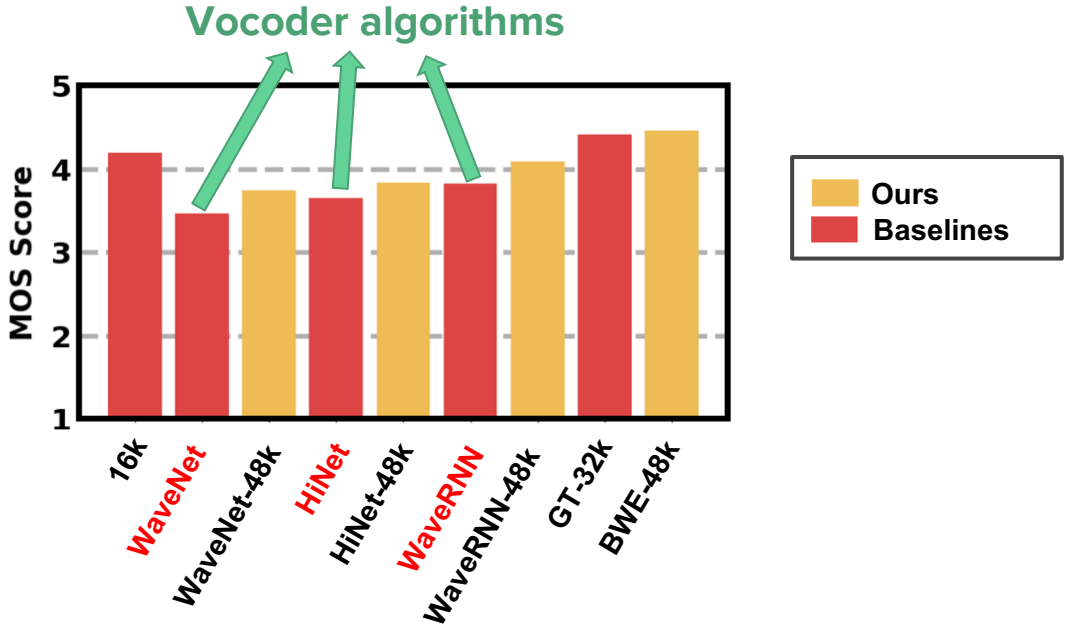


Experiments: BWE for Waveform generation

CMU Arctic Dataset

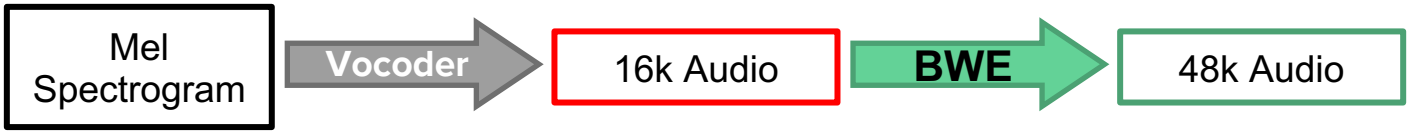
Mel Spectrogram → Vocoder → 16k Audio → BWE → 48k Audio

MOS Test

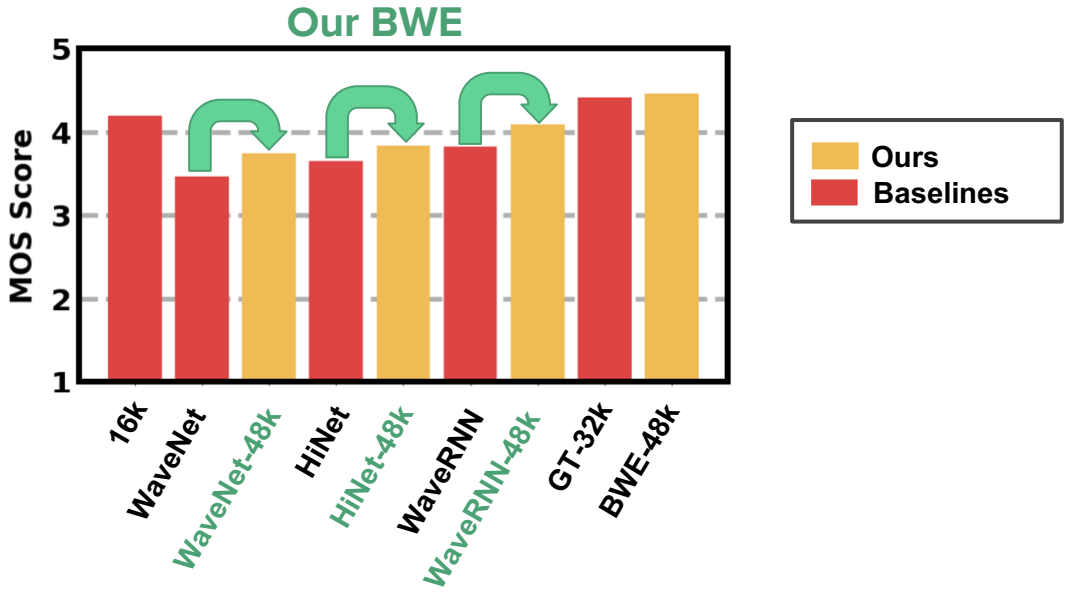


Experiments: BWE for Waveform generation

CMU Arctic Dataset

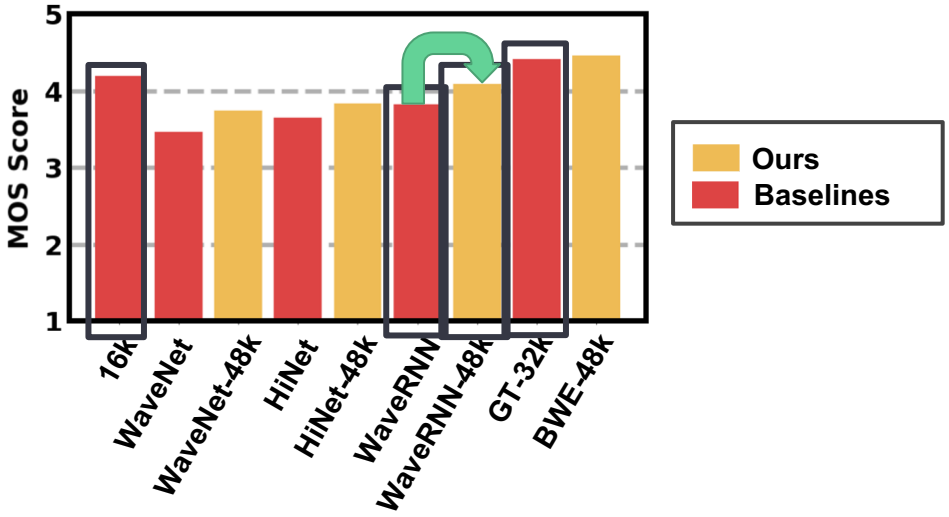


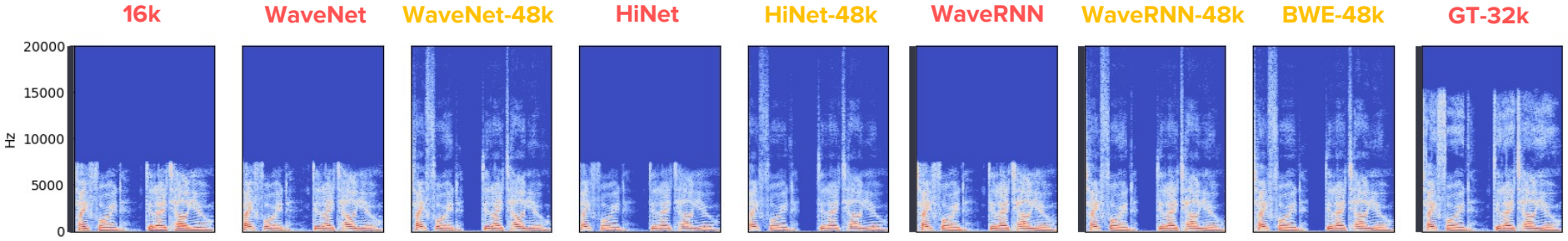
MOS Test



Observation: Consistent quality boost to vocoder algorithms

Experiments: BWE for Waveform generation - Demo





More audio examples

https://daps.cs.princeton.edu/projects/Su2020BWE/

Conclusions

- A bandwidth extension method based on HiFi-GAN targeting at up to **48kHz**, as a general tool to enhance other audio applications.
- Objective and subjective evaluations with STOA baselines on 8k-to-48kHz and 16k-to-48kHz BWE tasks.
- Evaluations on applying BWE to outputs of a variety of denoisers and vocoders.

Thanks for watching!