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

"Sweet spot" for efficiency

8k, 16kHz

But lost sense of presence
Motivation

Audio Applications

- Vocoders
- Voice Conversion
- Source Separation
- Denoisers

Doubling computations at least

Limited dataset

Modelling Challenges

Expensive

48kHz
Motivation

Audio Applications

Vocoders
Voice Conversion
Source Separation
Denoisers

Bandwidth extension is all you need!
Previous Work: Bandwidth Extension

- **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]
  - FFTNet with perceptual loss [Feng 2019]

  - Only method reaching 44kHz

Most limited to 8-16kHz for wideband
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
  - **LP**: linear prediction based analysis synthesis \([\text{Bachhav 2018}]\)
  - **Spec**: a spectral-domain method using 1D conv U-Net with GAN \([\text{Eskimez 2019}]\)
  - **Time**: a time-domain method using EnvNet structure with GAN \([\text{Li 2019}]\)
  - **FFTNet**: FFTNet variant for BWE \([\text{Feng 2019}]\)

- Dataset
  - **Train**: VCTK dataset \([\text{Veaux 2016}]\)
  - **Test**: Device and Produced Speech (DAPS) clean set \([\text{Mysore 2015}]\)
Observation: Objective metrics do not correlate well with perceptual quality.
Experiments: BWE - Subjective Evaluations

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.

Amazon Mechanical Turk

<table>
<thead>
<tr>
<th>MOS Score</th>
<th>Baseline</th>
<th>Ours</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>8k</td>
<td>LP</td>
<td>Spec</td>
<td>FFTNet</td>
</tr>
<tr>
<td>16k</td>
<td>Base</td>
<td>SpecGAN</td>
<td>HiFi-GAN+</td>
</tr>
<tr>
<td>44k</td>
<td>44k</td>
<td>44k</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MOS Score</th>
<th>Baseline</th>
<th>Ours</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>16k</td>
<td>LP</td>
<td>Spec</td>
<td>FFTNet</td>
</tr>
<tr>
<td>8k</td>
<td>Base</td>
<td>SpecGAN</td>
<td>HiFi-GAN+</td>
</tr>
<tr>
<td>44k</td>
<td>44k</td>
<td>44k</td>
<td></td>
</tr>
</tbody>
</table>

- Prefer 48k ground truth: 57.4%
- Prefer HiFi-GAN+: 42.6%
- = 85.2% no preference
Experiments: BWE - Demo

16k to 48k

References
Ours
Baselines

16k
LP
Spec
FFTNet
Time
Base
Spec-GAN
HiFi-GAN+
44k
Experiments: BWE - Demo
Experiments: BWE for Denoising

Denoising algorithms

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

MOS Test

References
Ours
Baselines
Experiments: BWE for Denoising

Observation: Consistent quality boost to enhancement algorithms
Experiments: BWE for Denoising - Demo

MOS Score

Noisy-16k  Clean-16k  DeepMMSE  DeepMMSE-48k  DEMUCS  DEMUCS-48k  HiFi-GAN  HiFi-GAN-48k  Clean-48k

References
Ours
Baselines
Experiments: BWE for Waveform generation

CMU Arctic Dataset

Mel Spectrogram

Vocoder

16k Audio

BWE

48k Audio

Vocoder algorithms

MOS Test

<table>
<thead>
<tr>
<th>MOS Score</th>
<th>16k</th>
<th>WaveNet</th>
<th>WaveNet-48k</th>
<th>HiNet</th>
<th>HiNet-48k</th>
<th>WaveRNN</th>
<th>WaveRNN-48k</th>
<th>GT-32k</th>
<th>BWE-48k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Baselines</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>
Experiments: BWE for Waveform generation

Observation: Consistent quality boost to vocoder algorithms
Experiments: BWE for Waveform generation - Demo

![Diagram showing MOS scores for different models and conditions.](image-url)
More audio examples

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!