Wave-U-Net Discriminator:
Fast and Lightweight Discriminator for Generative Adversarial Network-Based Speech Synthesis

ICASSP 2023

Audio samples
https://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/waveunetd/
Background

Advancement of speech synthesis

• Two-stage approach

  ![
  ![Text](image1)
  ![
  ![Acoustic model](image2)
  ![Intermediate representation](image3)
  ![Neural vocoder](image4)
  ![Waveform](image5)
  
  Common objective: High-quality speech synthesis

• End-to-end approach

  ![
  ![Text](image1)
  ![
  ![End-to-End](image6)
  ![Waveform](image5)
  
  Common objective: High-quality speech synthesis

Copyright NTT CORPORATION
Background

GAN [Goodfellow+2014]-based speech synthesis

- Two-stage approach (e.g., HiFi-GAN [Kong+2020])

- End-to-end approach (e.g., VITS [Kim+2021])

Challenge: How to design an adequate discriminator?
Challenge of GAN-based approach

Speech has multilevel (e.g., multiscale) structures

Short relationship

Long relationship

Discriminator must capture multilevel structures
Previous study

An ensemble of discriminators was used

Multiscale discriminator
[Kumar+2019]

Multiperiod discriminator
[Kong+2020]

Multilevel structures can be captured
Model size & computation time increase according to #discriminators
Previous study

An ensemble of discriminators was used

Research question:
Can we replace an ensemble of discriminators with a single but expressive discriminator?

Multilevel structures can be captured
Model size & computation time increase according to #discriminators
Our solution

Wave-U-Net Discriminator

Multilevel structures can be captured via an encoder and decoder

#discriminators is one → Fast and lightweight
Method
Previous discriminator

Encoder architecture (e.g., MelGAN [Kumar+2019], HiFi-GAN [Kong+2020])

Real/fake is determined using the abstracted features
Previous discriminator

Encoder architecture (e.g., MelGAN [Kumar+2019], HiFi-GAN [Kong+2020])

Real/fake is determined using the abstracted features

→ Multiple discriminators are required to capture detailed structures
Wave-U-Net discriminator

Encoder-decoder architecture

Real/fake is determined in a sample-wise manner
Wave-U-Net discriminator

Encoder-decoder architecture

Real/fake is determined in a sample-wise manner
→ One discriminator is sufficient to capture detailed structures
Challenge in training

Unstable training of Wave-U-Net discriminator

Wave-U-Net discriminator is deeper than typical discriminator → Causes unstable training (saturate adversarial losses)
Techniques for stable training 1

Careful normalization

- Global normalization

\[
b = \frac{a}{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (a_i^i)^2 + \epsilon}}
\]

Prevents Wave-U-Net Discriminator from restricting itself to specific features

Copyright NTT CORPORATION
Techniques for stable training 2

Introduction of residual connections \([\text{He+2016}]\)

Real or fake (sample-wise)

Prevents the gradient vanishing problem
Validation items

1. Evaluation on neural vocoders

*Dataset dependency* was investigated

- **Datasets:**
  - *LJSpeech* [Ito&Johnson2017]: Single English female speaker
  - *VCTK* [Yamagishi+2016]: Multiple English speakers
  - *JSUT* [Sonobe+2017]: Single Japanese female speaker

- **Baseline:** *HiFi-GAN* [Kong+2020]

2. Evaluation on *end-to-end TTS*

*Task dependency* was investigated

- **Datasets:** *LJSpeech* [Ito&Johnson2017]
- **Baseline:** *VITS* [Kim+2021]

---

**Performance was examined when the original ensemble of discriminators was replaced with a Wave-U-Net discriminator**
Evaluation metrics

Speech quality

• **Subjective metric: MOS ↑**
  › Mean opinion score on naturalness

• **Objective metric: cFW2VD ↓ [Kaneko+2022]**
  › Distance between real and synthesized speech in wav2vec 2.0 [Baevski+2020]

Training speed

• **Time (s/batch) ↓**
  › Time required for a discriminator to process real and synthesized speech in a batch

Model size

• **# Param (M) ↓**
  › Number of parameters of a discriminator
Evaluation on neural vocoders
Results 1/3

Evaluation on neural vocoder in LJSpeech

- **HiFi-GAN** (baseline)
- **Wave-U-Net D** (proposed)

**MOS ↑**
- Comparable

**cFW2VD ↓**
- Comparable

**Time (s/batch) ↓**
- Fast

**# Param (M) ↓**
- Lightweight

Wave-U-Net discriminator reduces computation time & model size while retaining speech quality
Results 2/3

Evaluation on neural vocoder in VCTK

Wave-U-Net discriminator reduces computation time & model size while retaining speech quality
Results 3/3

Evaluation on neural vocoder in JSUT

- HiFi-GAN (baseline)
- Wave-U-Net D (proposed)

MOS $\uparrow$

<table>
<thead>
<tr>
<th>Wave-U-Net D</th>
<th>Comparable</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiFi-GAN</td>
<td>Comparable</td>
</tr>
</tbody>
</table>

cFW2VD $\downarrow$

<table>
<thead>
<tr>
<th>Wave-U-Net D</th>
<th>Comparable</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiFi-GAN</td>
<td>Comparable</td>
</tr>
</tbody>
</table>

Time (s/batch) $\downarrow$

<table>
<thead>
<tr>
<th>Wave-U-Net D</th>
<th>Fast $\downarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiFi-GAN</td>
<td>$\downarrow$</td>
</tr>
</tbody>
</table>

# Param (M) $\downarrow$

<table>
<thead>
<tr>
<th>Wave-U-Net D</th>
<th>Lightweight $\downarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiFi-GAN</td>
<td></td>
</tr>
</tbody>
</table>

Wave-U-Net discriminator reduces computation time & model size while retaining speech quality
Evaluation on end-to-end TTS
Results

Evaluation on end-to-end TTS

- VITS (baseline)
- Wave-U-Net D (proposed)

Wave-U-Net discriminator reduces computation time & model size while retaining speech quality
Conclusion
Conclusion

Objective

• Make a discriminator faster & more lightweight

Proposal

• Wave-U-Net Discriminator

Experiments

• Make a discriminator faster & more lightweight while retaining speech quality

Future work

• Application to other tasks
  ➔ Singing speech synthesis, emotional speech synthesis, …