iSTFTNet:
Fast and Lightweight Mel-Spectrogram Vocoder
Incorporating Inverse Short-Time Fourier Transform

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

Takuhiro Kaneko
Kou Tanaka
Hirokazu Kameoka
Shogo Seki
Background and Objective 1/5

Construction of fast and lightweight mel-spectrogram vocoder

- **Text-to-speech synthesis** (Text → Waveform)

  ![Text](image1)
  ![Mel-spec.](image2)
  ![Mel-spectrogram vocoder](image3)
  ![Waveform](image4)

- **Voice conversion** (Waveform → Waveform)

  ![Waveform](image5)
  ![Mel-spec](image6)
  ![Mel-spec.](image7)
  ![Mel-spectrogram vocoder](image8)
  ![Waveform](image9)

**Objective of this study:**
- **Compact & Expressive**
- **Speed-up & weight reduction**
Flow of mel-spectrogram extraction

1) STFT

Waveform

Phase

Magnitude
Background and Objective 2/5

Flow of mel-spectrogram extraction

1. STFT
   - Waveform
   - 513 dim. magnitude

2. Drop phase
   - Phase
   - Magnitude
Background and Objective 2/5

Flow of mel-spectrogram extraction

1. STFT
2. Drop phase
3. Convert scale
Background and Objective 3/5

Flow of mel-spectrogram vocoder (signal processing solution)

Pros: Exploits **time-frequency structure** explicitly

Cons: Requires **redundant estimation** (reconstruction of high-dim. specs)

(3’) Recover scale
Flow of mel-spectrogram vocoder (signal processing solution)

(2') Reconstruct phase  (3') Recover scale

Pros: Exploits time-frequency structure explicitly
Cons: Requires redundant estimation (reconstruction of high-dim. specs)
Flow of mel-spectrogram vocoder (signal processing solution)

Pros: Exploits **time-frequency structure** explicitly
Cons: Requires **redundant estimation** (reconstruction of high-dim. specs)
Flow of mel-spectrogram vocoder (DNN shortcut solution)

**Pros:** Does not require **redundant estimation** (reconstruction of high-dim. specs)

**Cons:** Cannot exploit **time-frequency structure** explicitly
Flow of mel-spectrogram vocoder (DNN shortcut solution)

Pros: Does not require redundant estimation (reconstruction of high-dim. specs)
Cons: Cannot exploit time-frequency structure explicitly

① Keep this advantage
② Can we exploit?
Proposal: iSTFTNet

Hybrid of DNN upsampling & iSTFT signal processing

Mel-spectrogram vocoder (previous)  
Replace  
iSTFTNet (proposed)

DNN upsampling  
→ Simplifies frequency structure

iSTFT  
→ Converts time-frequency explicitly

Fast & Lightweight

Black box  
→ Does not consider time-frequency structure
Related Work

Use of iSTFT for waveform synthesis

E.g., GAN Signal Reconstruction [Oyamada+2018 (ours)]

Synthesizes high-dim. spec. directly
→ Requires high-capacity model (e.g., 2D CNN)

iSTFTNet (Proposed)

Synthesizes low-dim. spec
→ Only requires low-capacity model (e.g., 1D CNN)
Theoretical Background

Time-frequency trade-off

\[ f_1 \cdot 1 = f_s \cdot s = \text{constant} \]

We can simplify frequency structure by increasing time scale

\text{FFT size} \times \text{Time scale}

Simple enough

Input

513 dim. \rightarrow 9 dim. \rightarrow 80 dim.

Time \times 64

\text{FFT: 1024}
\text{Hop: 256}
\text{Window: 1024}

\text{FFT: 16}
\text{Hop: 4}
\text{Window: 16}

\text{Mel-spec.}
Architectures of iSTFTNets

(a) C1I (No upsampling)
Fast
Lightweight
Low-quality?

(b) C8I

(c) C8C8I

(d) C8C8C2I
(Previous)

Hybrid: iSTFTNet

We examined effect on quality empirically

Copyright 2022 NTT CORPORATION
Experiment Setup 1/3

Data

- **Dataset**: LJSpeech dataset [Ito&Johnson2017]
  - **Speaker**: English female
  - **Audio clips**: 13,100 (24 h) (training: 12,600, validation: 250, evaluation: 250)
  - **Sampling rate**: 22.05 kHz

- **Audio feature**: 80-dimensional log-mel spectrograms
  - FFT size: 1024, hop length: 256, window length: 1024

Comparison model

- **Latest models**: 3 HiFi-GAN variants [Kong+2020]
  - V1 (High-quality), V2 (Lightweight), V3 (Fast)

- **Benchmark models**:
  - Multiband (MB)-MelGAN [Yang+2021], Parallel WaveGAN (PWG) [Yamamoto+2020]
  

Compare them with their iSTFTNet variants
Evaluation metrics

• Perceptual quality
  › **Subjective:** Mean opinion score (MOS) test
    » MOS↑ → Quality↑
  › **Objective:** Conditional Fréchet wav2vec distance (cFW2VD)
    » Calculates distance between real and generative distributions in wav2vec 2.0 [Baevski+2020]
    » High correlation with MOS (Spearman’s rank correlation: -0.93)
    » cFW2VD↓ → Quality↑

• Inference speed
  › **Relative speed** compared to real time on GPU/CPU
    » Relative speed↑ → Fast

• Model size
  › **Number of parameters**
    » #Param↓ → Lightweight
Validation items

1. How many blocks should be retained?

2. Necessity of combining DNN upsampling and iSTFT

3. Comparison with benchmark models

*iSTFTNet (v2) vs. MB-MelGAN vs. PWG*
1. How many blocks should be retained?

Quality is preserved when #blocks is 3/2

High-quality & fast & lightweight when #blocks is 3/2
Results 2/3

2. Necessity of combining DNN upsampling and iSTFT

- **MOS ↑**
- **cFW2VD ↓**
- **Speed (GPU) ↑**
- **Speed (CPU) ↑**
- **#Param (M) ↓**

### MOS

- **V1** (High quality)
- **V2** (Lightweight)
- **V3** (High speed)

### cFW2VD

- **0**
- **0.1**
- **0.2**

### Speed

- **Speed (GPU)**
- **Speed (CPU)**

### #Param

- **0**
- **1**
- **1.5**

---

**Upsampling is necessary to preserve quality**

※ We also examined non-upsampling models and found that they suffer from training difficulties
3. Comparison with benchmark models

- **iSTFTNet (v2)** vs. **MB-MelGAN** vs. **PWG**

- Quality: **iSTFTNet** is best
- Speed: **iSTFTNet** = **MB-MelGAN**
- Model size: **iSTFTNet** is best

※ We also confirmed that **iSTFT-MelGAN** (iSTFTNet+MelGAN) outperforms **MB-MelGAN** in terms of quality.
Application to TTS synthesis

Examination of applicability to text-to-speech synthesis

- **Original** vs. C-FS2+HiFi-GAN vs. C-FS2+iSTFTNet vs. C-FS2
  - C-FS2: Conformer+FastSpeech 2 [Guo+2021]

Text
made certain recommendations which it believes would, if adopted,
iSTFTNet
- Comparable with ground truth
- Better or comparable with HiFi-GAN/C-FS2

MOS ↑

<table>
<thead>
<tr>
<th></th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth</td>
<td>4.5</td>
</tr>
<tr>
<td>HiFi-GAN</td>
<td>4.0</td>
</tr>
<tr>
<td>iSTFTNet</td>
<td>4.5</td>
</tr>
<tr>
<td>C-FS2</td>
<td>3.5</td>
</tr>
</tbody>
</table>

cFW2VD ↓

<table>
<thead>
<tr>
<th></th>
<th>cFW2VD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiFi-GAN</td>
<td>0.2</td>
</tr>
<tr>
<td>iSTFTNet</td>
<td>0.1</td>
</tr>
<tr>
<td>C-FS2</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Summary and Future Work

Objective

• Construction of fast & lightweight mel-spectrogram vocoder

Proposal

• \textit{iSTFTNet}: DNN upsampling + iSTFT

Experiments

• \textit{iSTFTNet} is faster and more lightweight

Future work

• Applying our ideas to other neural vocoders

We also apply to multi-speaker & Japanese datasets

Audio samples

https://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/istftnet/