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
Construction of fast and lightweight mel-spectrogram vocoder

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

- Voice conversion (Waveform → Waveform)

Objective of this study:

Speed-up & weight reduction

Compact & Expressive
Flow of mel-spectrogram extraction

1. STFT

Waveform

Phase

Magnitude

513 dim.
Flow of mel-spectrogram extraction

(1) STFT

Waveform

(2) Drop phase

Phase

Magnitude

513 dim.
Background and Objective 2/5

Flow of mel-spectrogram extraction

1. STFT
2. Drop phase
3. Convert scale

Waveform → STFT → Phase → Drop phase → Mel-spec. → Convert scale
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)
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)
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

Keep this advantage
Can we exploit?
Proposal: **iSTFTNet**

Hybrid of DNN upsampling & iSTFT signal processing

Replace

1. DNN upsampling
   - Simplifies frequency structure

2. iSTFT
   - Converts time-frequency explicitly

Fast & Lightweight

Mel-spectrogram vocoder (previous)

Mel-spectrogram

Input conv

ResBlock (upsample $\times 8$)

ResBlock (upsample $\times 8$)

ResBlock (upsample $\times 2$)

ResBlock (upsample $\times 2$)

Output conv

Raw waveform

Output conv

ResBlock (upsample $\times 8$)

ResBlock (upsample $\times 8$)

Mag./Phase $\rightarrow$ iSTFT

Raw waveform

iSTFTNet (proposed)

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

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

FFT size \times \text{Time scale}

We can simplify frequency structure by increasing time scale

- 513 dim.
  - Time \times 64
  - FFT: 1024
  - Hop: 256
  - Window: 1024

- 9 dim.
  - 64T

- 80 dim.
  - Mel-spec.

Simple enough

Copyright 2022 NTT CORPORATION
Architectures of iSTFTNets

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

(b) C8I

(c) C8C8I

(d) C8C8C2I

(e) C8C8C2C2 (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
Experiment Setup 2/3

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
**Results 1/3**

1. How many blocks should be retained?

**Quality is preserved when \#blocks is 3/2**

- **V1** (High-quality)
  - Number of retained blocks: 4
  - Quality preserved when \#blocks is 3/2 for **V1** (High-quality)

- **V2** (Lightweight)
  - Number of retained blocks: 1

- **V3** (Fast)
  - Number of retained blocks: 4

**#Blocks ↓ → Speed ↑**

**#Blocks ↓ → Weight ↓**

**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) ↓ |
|--------|-------|----------|---------------|---------------|-------------|
| **V1** |       |          |               |               |             |
| (高品質)|       |          |               |               |             |
| **V2** |       |          |               |               |             |
| (軽量) |       |          |               |               |             |
| **V3** |       |          |               |               |             |
| (高速)  |       |          |               |               |             |

Number of upsampling blocks

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

- \textit{iSTFTNet (v2)} vs. MB-MelGAN vs. PWG

\begin{itemize}
  \item \textbf{Quality}
    \begin{itemize}
      \item \textit{iSTFTNet is best}
    \end{itemize}

  \item \textbf{Speed}
    \begin{itemize}
      \item \textit{iSTFTNet = MB-MelGAN}
    \end{itemize}

  \item \textbf{Model size}
    \begin{itemize}
      \item \textit{iSTFTNet is best}
    \end{itemize}
\end{itemize}

\begin{figure}
\centering
\begin{tabular}{|c|c|c|}
\hline
 & MOS & cFW2VD \\
\hline
\textit{iSTFTNet} & 5 & 0.1 \\
\textit{MB} & 4 & 0.05 \\
\textit{PWG} & 3 & 0 \\
\hline
\textbf{Quality} & \textit{iSTFTNet is best} & \\
\hline
\textit{iSTFTNet} & 1500 & 1000 \\
\textit{MB} & 20 & 10 \\
\textit{PWG} & 1.5 & 0.5 \\
\hline
\textbf{Speed} & \textit{iSTFTNet = MB-MelGAN} & \\
\hline
\textit{iSTFTNet} & 1.5 & 0.5 \\
\textit{MB} & 1 & 1 \\
\textit{PWG} & 0 & 0 \\
\hline
\textbf{Model size} & \textit{iSTFTNet is best} & \\
\hline
\end{tabular}
\end{figure}

\textit{※ 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
Summary and Future Work

Objective
• Construction of fast & lightweight mel-spectrogram vocoder

Proposal
• iSTFTNet: DNN upsampling + iSTFT

Experiments
• iSTFTNet is faster and more lightweight

Future work
• Applying our ideas to other neural vocoders

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

We also apply to multi-speaker & Japanese datasets