NTT

SPE-7.5

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

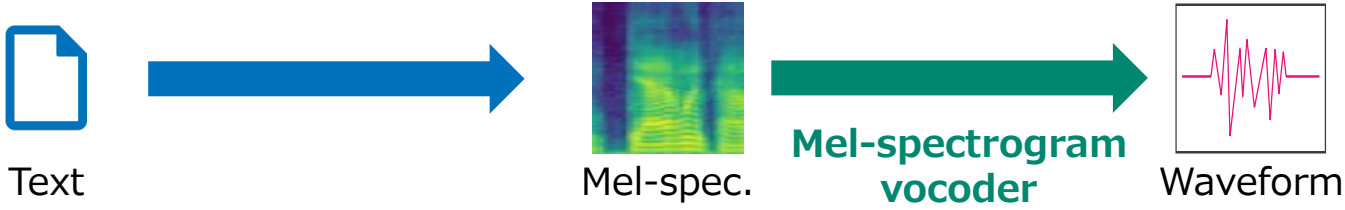
Shogo Seki

IEEE ICASSP 2022

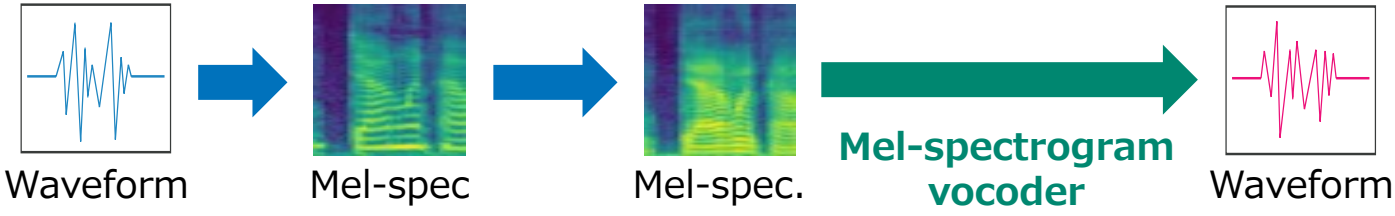
Background and Objective 1/5

Construction of fast and lightweight mel-spectrogram vocoder

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

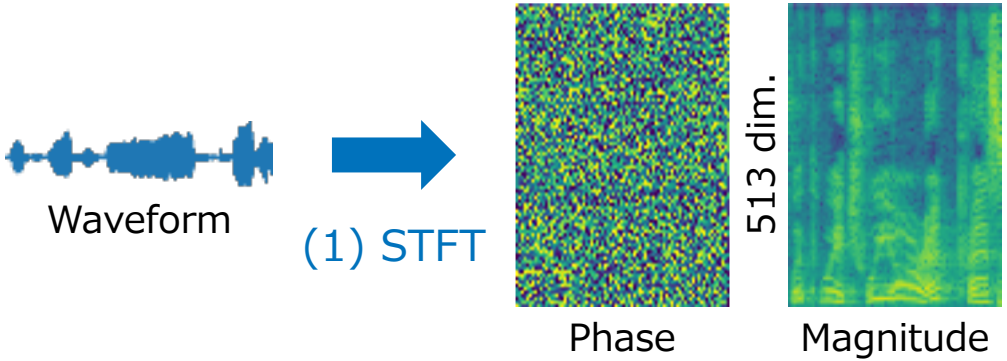


- Voice conversion (Waveform → Waveform)



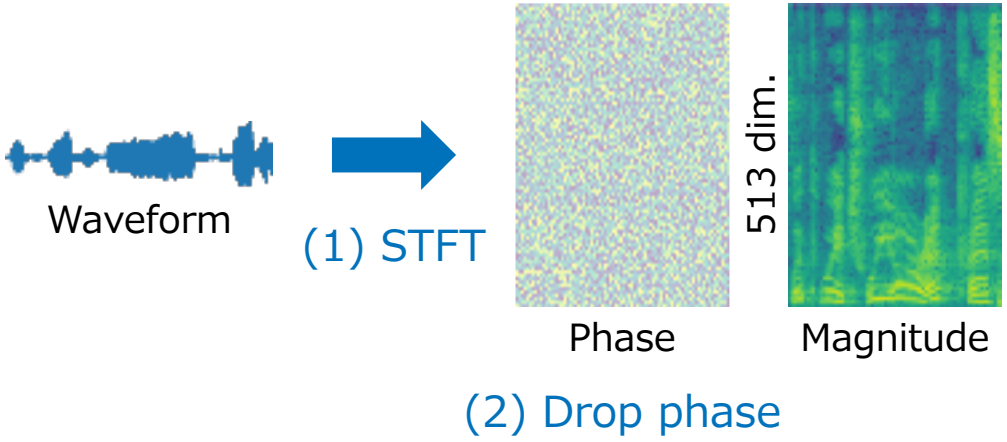
Background and Objective 2/5

Flow of mel-spectrogram extraction



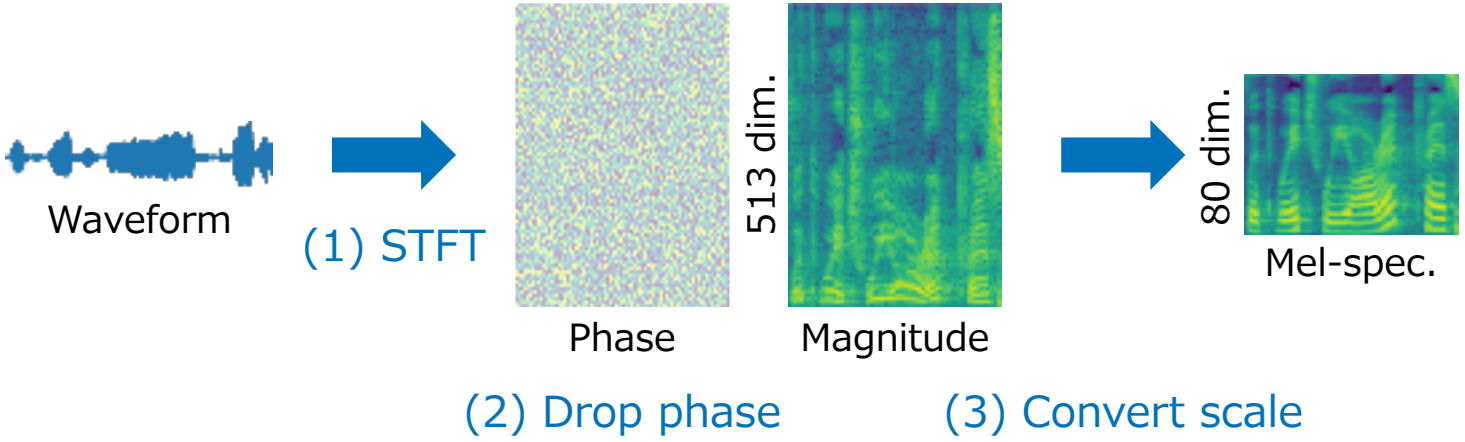
Background and Objective 2/5

Flow of mel-spectrogram extraction



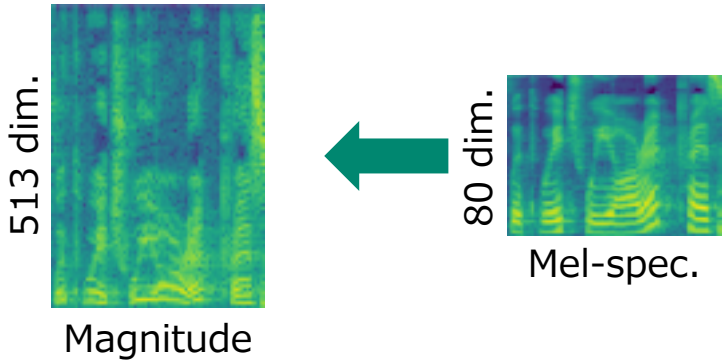
Background and Objective 2/5

Flow of mel-spectrogram extraction



Background and Objective 3/5

Flow of mel-spectrogram vocoder (signal processing solution)



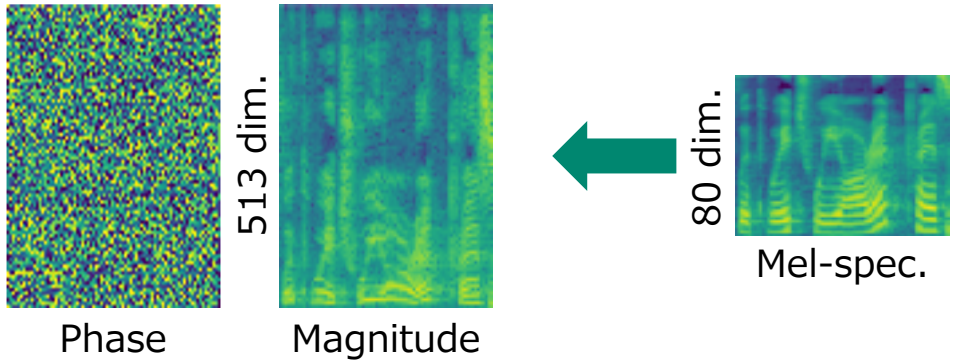
(3’) Recover scale

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)



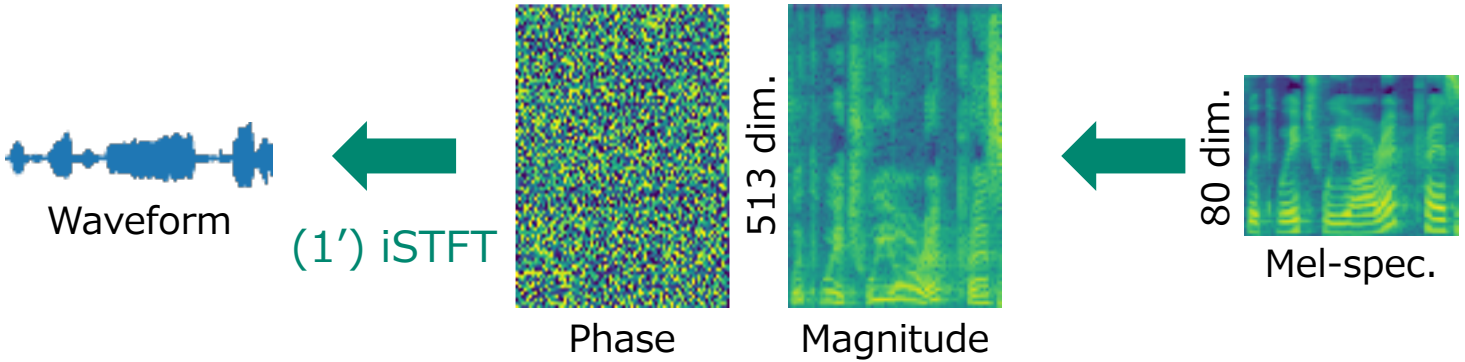
(2′) Reconstruct phase (3′) Recover scale

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)

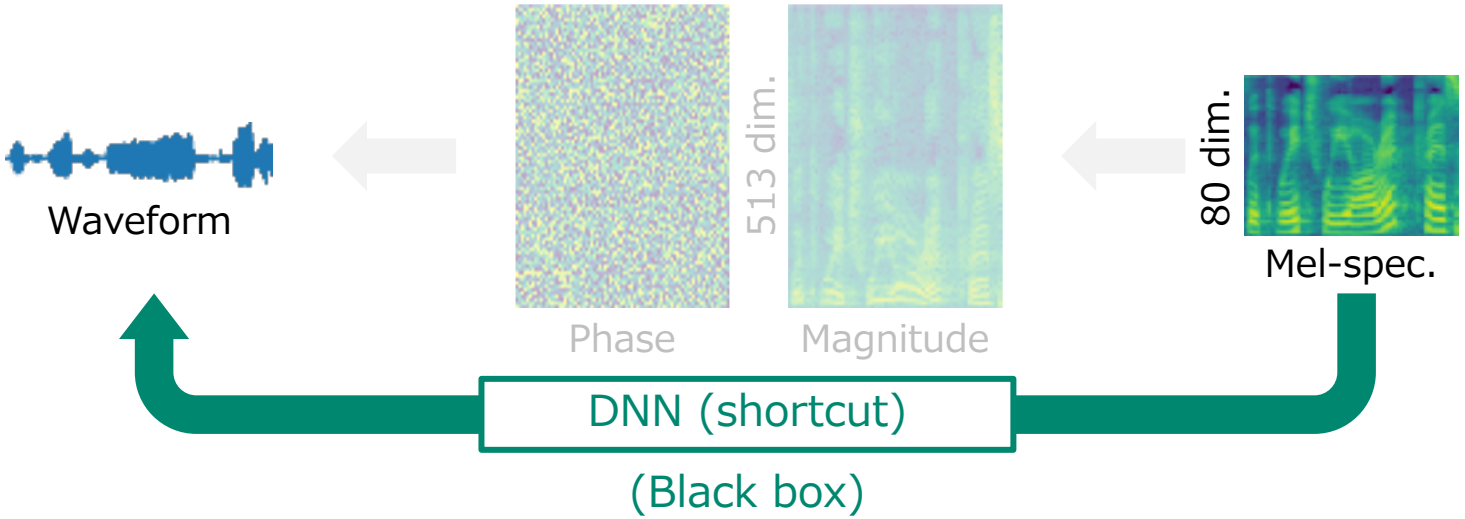


Pros: Exploits **time-frequency structure** explicitly

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

Background and Objective 4/5

Flow of mel-spectrogram vocoder (DNN shortcut solution)

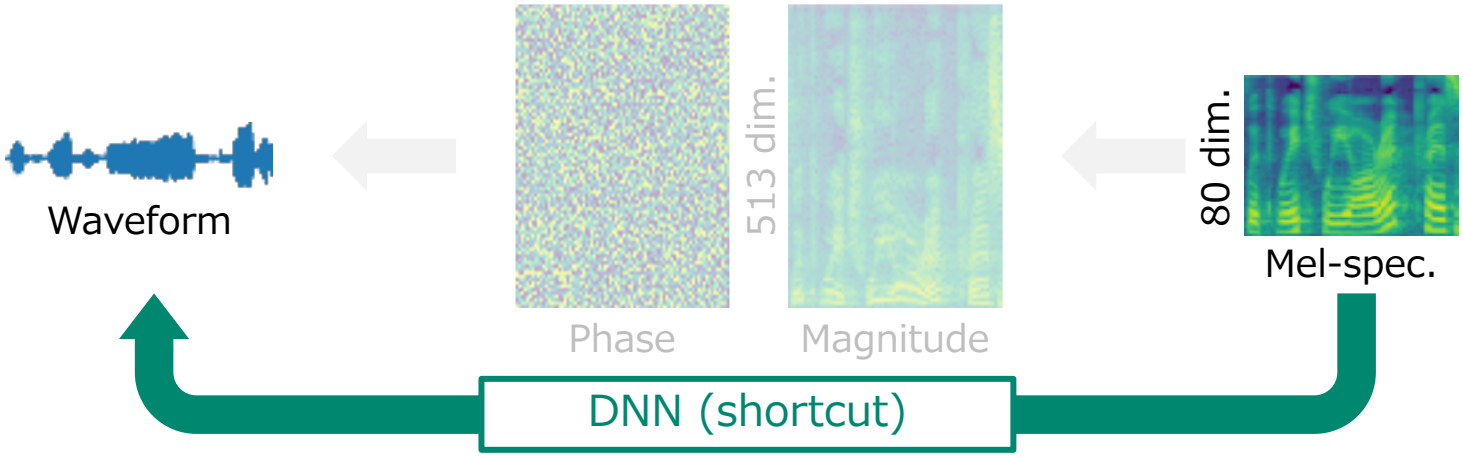


Pros: Does not require **redundant estimation** (reconstruction of high-dim. specs)

Cons: Cannot exploit **time-frequency structure** explicitly

Background and Objective 5/5

Flow of mel-spectrogram vocoder (DNN shortcut solution)



(Black box) ❶ Keep this advantage

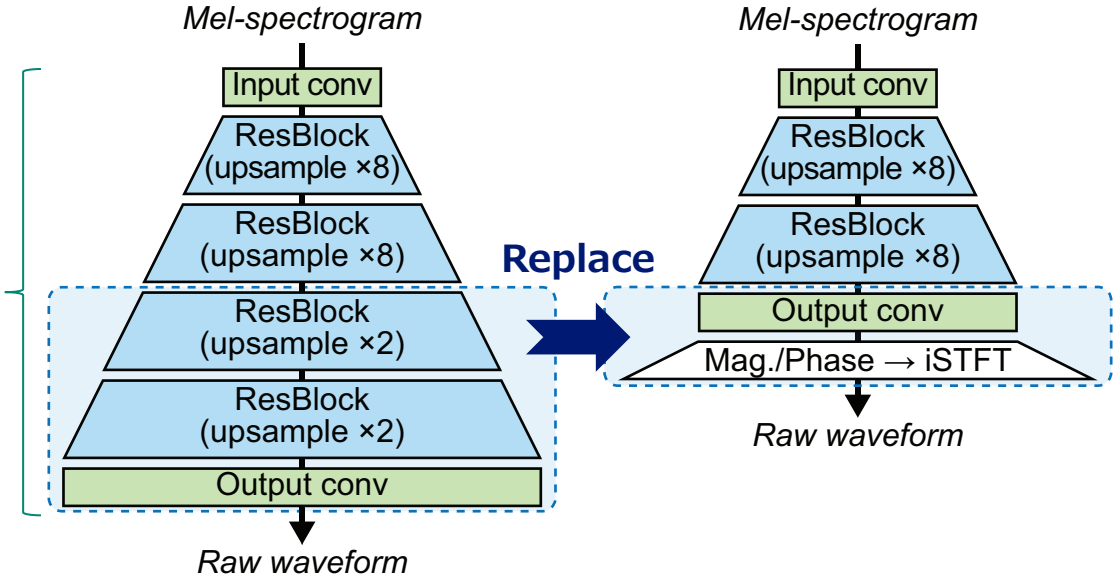
Pros: Does not require **redundant estimation** (reconstruction of high-dim. specs)

Cons: Cannot exploit **time-frequency structure** explicitly

❷ Can we exploit?

Proposal: *iSTFTNet*

Hybrid of DNN upsampling & iSTFT signal processing



Black box
→ Does not consider time-frequency structure

Replace

❶DNN upsampling
→ Simplifies frequency structure

❷iSTFT
→ Converts time-frequency explicitly

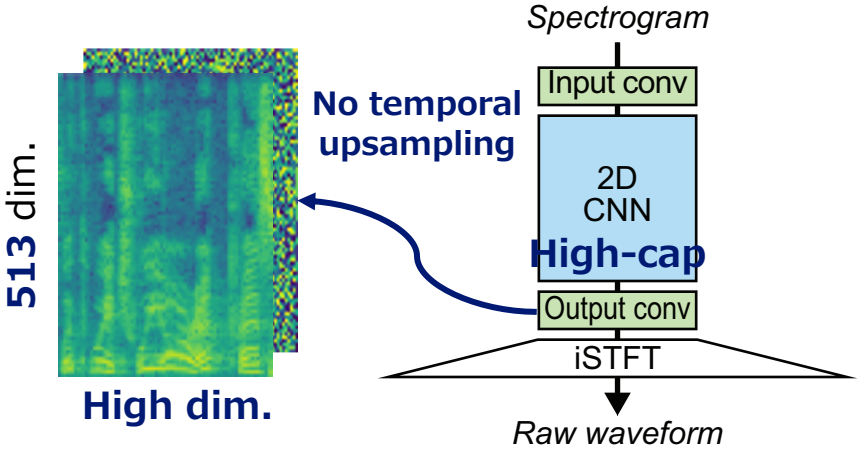
Mel-spectrogram vocoder (previous)

***iSTFTNet* (proposed)**

Fast & Lightweight

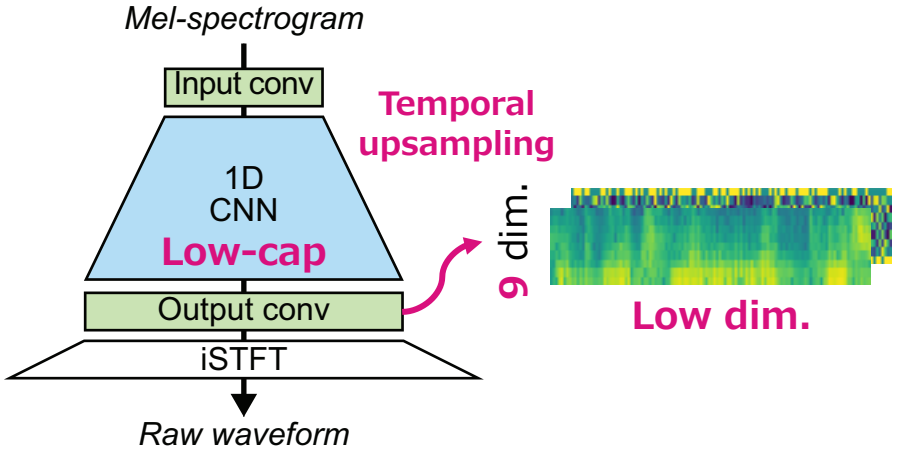
Related Work

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)

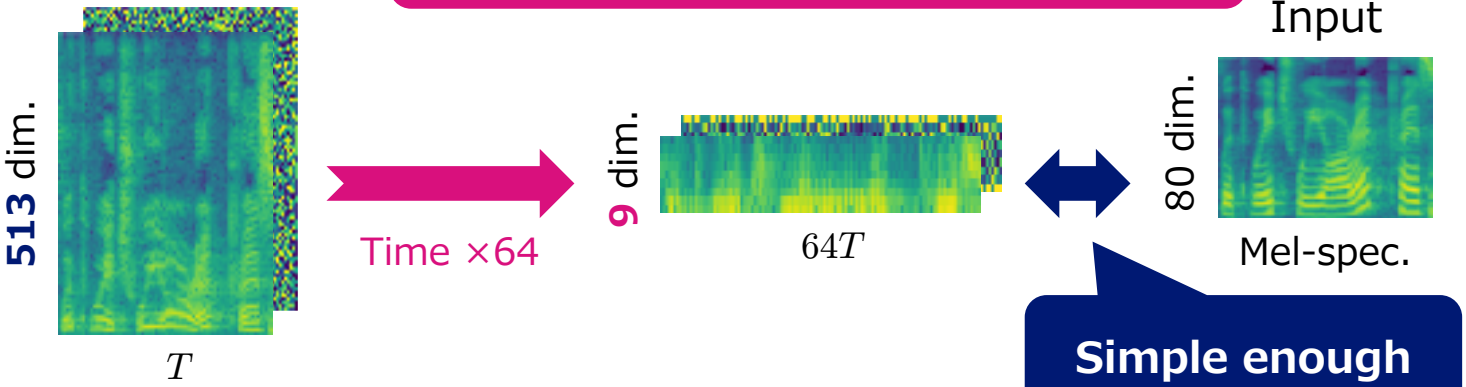
Theoretical Background

Time-frequency trade-off

$$f_1 \cdot 1 = f_s \cdot s = \text{constant}$$

FFT size: f_1 — Time scale: 1

We can simplify frequency structure by increasing time scale



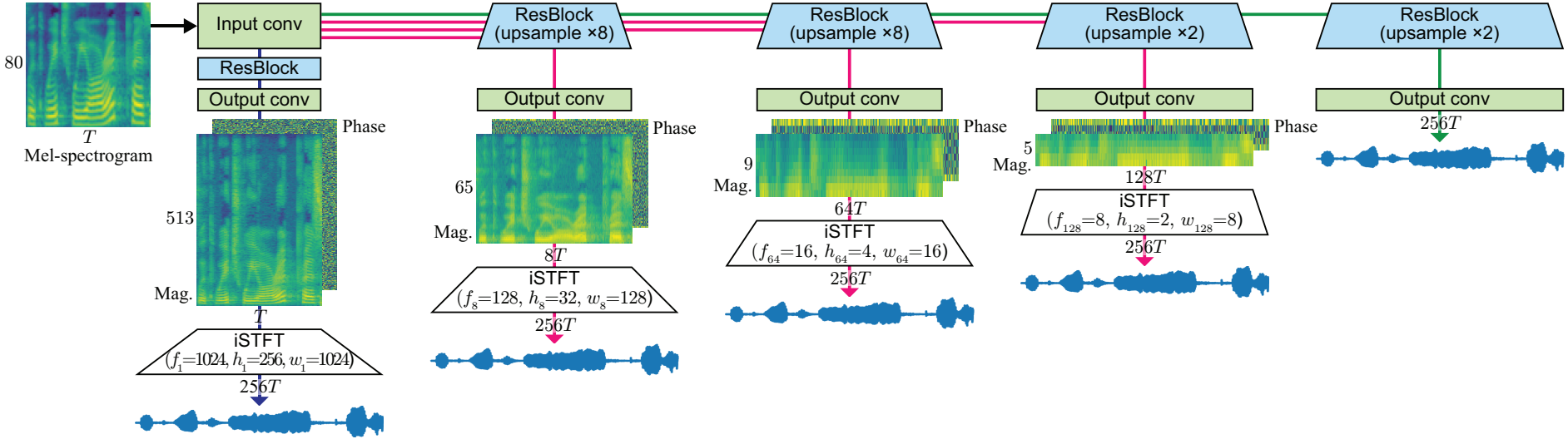
FFT: 1024
Hop: 256
Window: 1024

FFT: 16
Hop: 4
Window: 16

Simple enough

Architectures of *iSTFTNets*

NTT



(a) C1I
(No upsampling)

(b) *C8I* **(c) *C8C8I*** **(d) *C8C8C2I***

Hybrid: *iSTFTNet*

(e) C8C8C2C2
(Previous)

Fast
Lightweight
Low-quality?

Slow
Heavyweight
High-quality?

We examined effect on quality empirically

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**

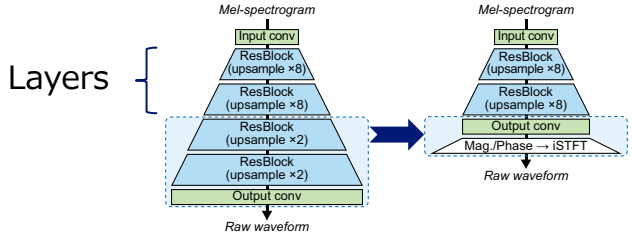
Experiment Setup 3/3

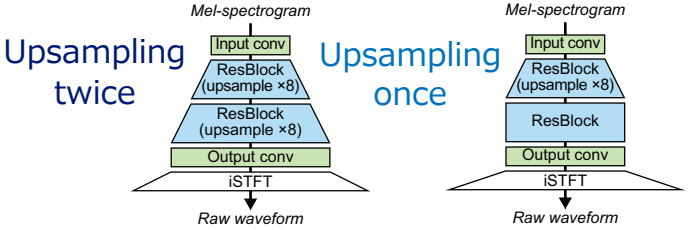
Validation items

1. **How many blocks should be retained?**

2. **Necessity of combining DNN upsampling and iSTFT**

3. **Comparison with benchmark models**

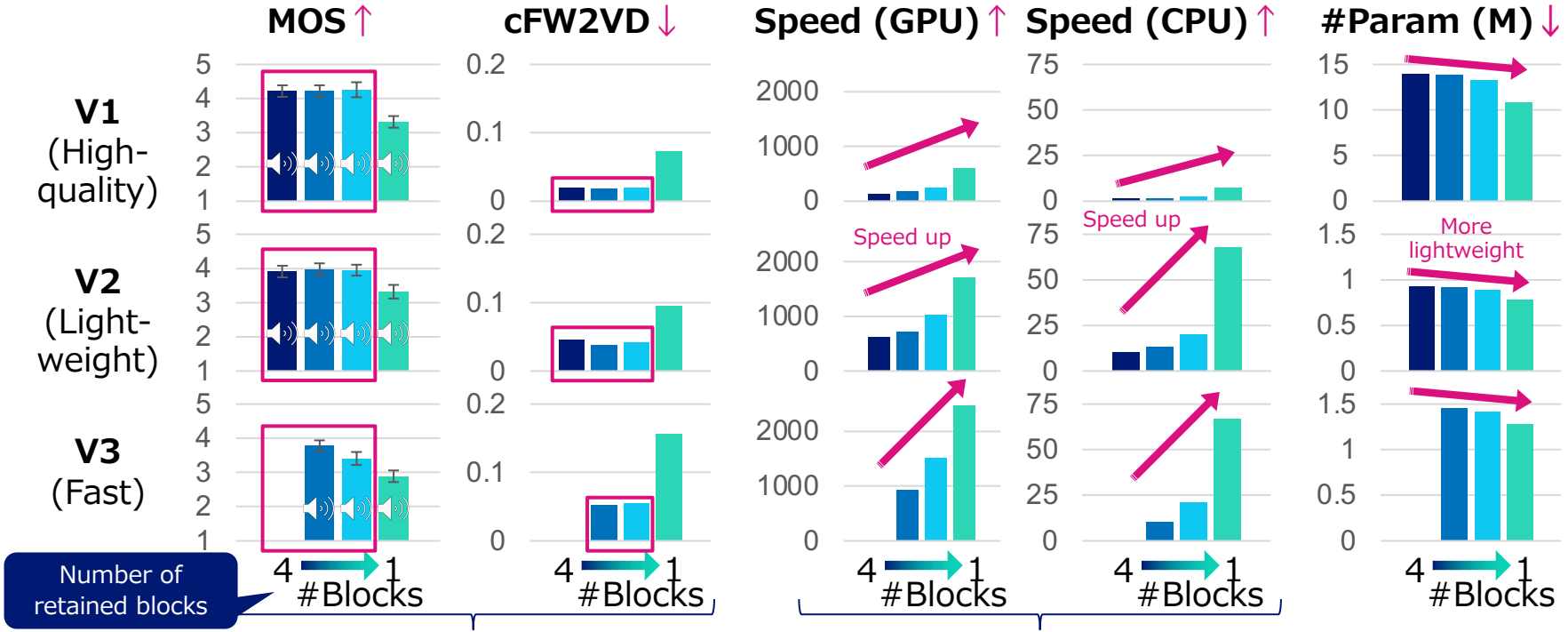




***iSTFTNet* (v2)** vs. **MB-MelGAN** vs. **PWG**

Results 1/3

1. How many blocks should be retained?

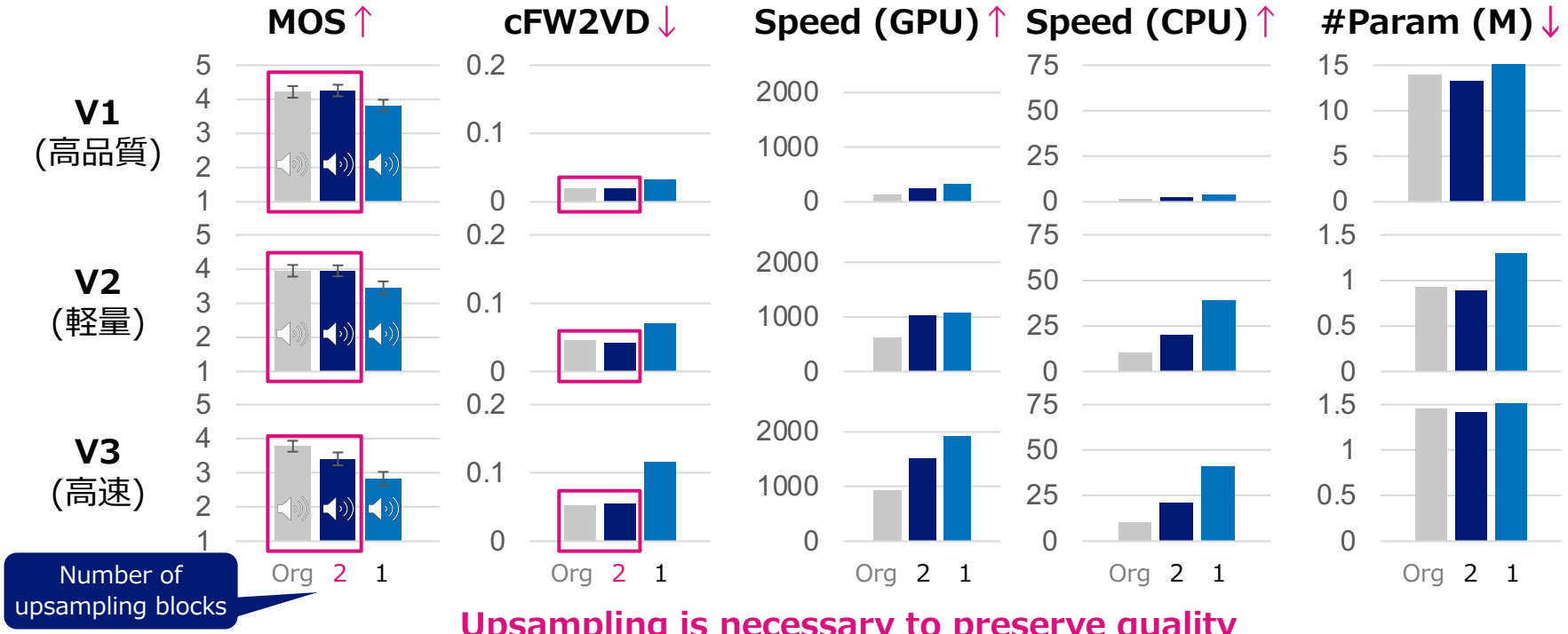


Quality is **preserved** when #blocks is 3/2 #Blocks↓ → **Speed**↑ #Blocks↓ → **Weight**↓

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

Results 2/3

2. Necessity of combining DNN upsampling and iSTFT



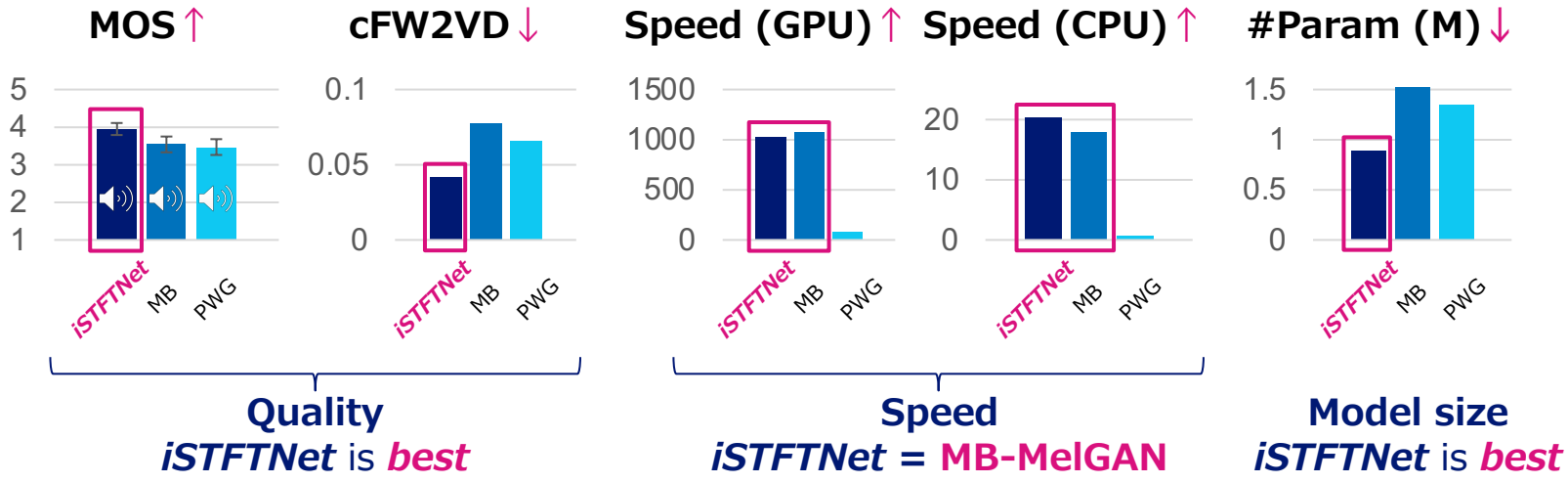
Upsampling is necessary to preserve quality

※ We also examined non-upsampling models and found that they suffer from training difficulties

Results 3/3

3. Comparison with benchmark models

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



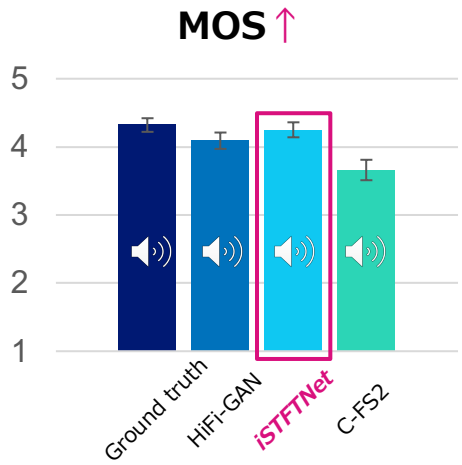
※ We also confirmed that **iSTFT-MelGAN** (iSTFTNet+MelGAN) outperforms **MB-MelGAN** in terms of quality

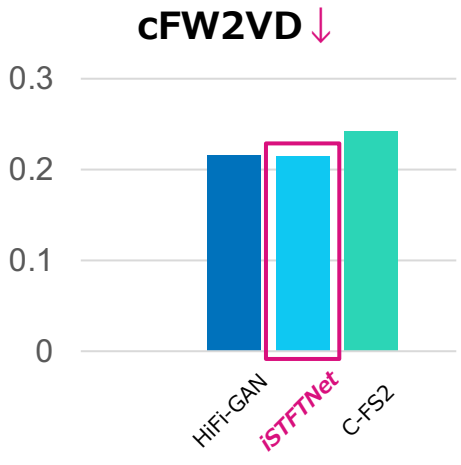
Application to TTS synthesis

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

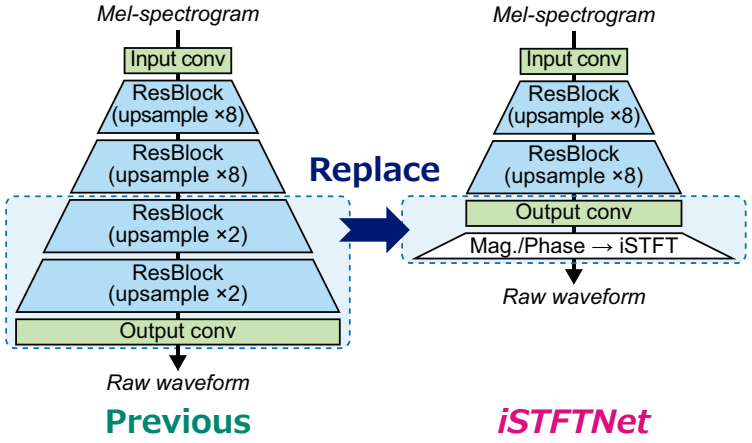
- ***iSTFTNet*: DNN** upsampling + **iSTFT**

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