**Background**

Increased demand for efficient mel-spectrogram vocoder

Text-to-speech synthesis (Text → Waveform)

Voice conversion (Waveform → Waveform)

Objective of this study: Speed up & reduce weights

**Typical mel-spectrogram vocoders**

Signal processing-based solution

![Waveform](image1)

DNN-based shortcut solution

![Waveform](image2)

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

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

**Proposed: iSTFTNet**

Hybrid of DNN upsampling & iSTFT signal processing

**Theoretical Background**

Time-frequency trade-off

\[ f_1 \cdot 1 = f_2 \cdot s = \text{constant} \]

FFT size Time scale

We can simplify frequency structure by increasing time scale

**Architectures of iSTFTNets**

- **Model:** iSTFTNet
- **Ground truth:** 4.46 ± 0.042
- **iSTFTNet:** 4.22 ± 0.020
- **Conformer-FS2:** 4.09 ± 0.066

**Results (Synthesis from ground-truth mel-spectrogram)**

Q1. How many blocks should be retained?

- **iSTFTNet:**
  - Best quality: 4.22 ± 0.020
  - Reasonable quality: 4.09 ± 0.066

Q2. Necessity of combining DNN upsampling & iSTFT

- **iSTFTNet:**
  - Best quality: 4.22 ± 0.020
  - Reasonable quality: 4.09 ± 0.066

Q3. Comparison with benchmark models

- **Model:** iSTFTNet
  - Best quality: 4.22 ± 0.020
  - Reasonable quality: 4.09 ± 0.066

**Application to text-to-speech synthesis**

- **iSTFTNet** is better than or comparable with baselines
- **iSTFTNet** is comparable with ground truth