

Supplementary Material

Title: **FastSpeech 2: Fast and High-Quality End-to-End Text to Speech**

Directory Structure

```
├─ audio_quality
|   ├─ ablation_fs2          audios for Table 6 (a) in the paper
|   |   ├─ fastspeech2
|   |   ├─ fastspeech2-energy
|   |   ├─ fastspeech2-pitch
|   |   └─ fastspeech2-CWT (Section 3.2.3, Predicting Pitch in Frequency Domain)
|   └─ ablation_fs2s        audios for Table 6 (b) in the paper
|       ├─ fastspeech2s
|       ├─ fastspeech2s-energy
|       ├─ fastspeech2s-pitch
|       ├─ fastspeech2s-CWT (Section 3.2.3, Predicting Pitch in Frequency Domain)
|       ├─ fastspeech2s-pitch-energy
|       └─ fastspeech2s-mel_decoder
|           (Section 3.2.3, Mel-Spectrogram Decoder in FastSpeech 2s)
|   └─ comparison_teacher_MFA audios for Table 5 (b) in the paper
|       ├─ fastspeech+MFA
|       └─ fastspeech+teacher
|   └─ main_MOS              audios for Table 1 in the paper
|       ├─ fastspeech        FastSpeech (Mel + PWG)
|       ├─ fastspeech 2      FastSpeech 2 (Mel + PWG)
|       ├─ fastspeech 2s     FastSpeech 2s
|       ├─ gt_pwg            GT (Mel + PWG)
|       ├─ gt_recording      GT, the ground-truth recordings
|       ├─ tacotron2         Tacotron 2 (Mel + PWG)
|       └─ transformer tts   Transformer TTS (Mel + PWG)
├─ variance_control          audio samples for variance control (Appendix E in the paper)
|   └─ details in variance_control/README.pdf
|   └─ fs2_energy
|       └─ control the energy (volume) in synthesized speech in FastSpeech 2
|           ├─ audio1        audio example 1
|           └─ audio2        audio example 2
|   └─ fs2_pitch            control the pitch in synthesized speech in FastSpeech 2
|       ├─ audio1          audio example 1
|       └─ audio2          audio example 2
|   └─ fs2s_energy
|       └─ control the energy (volume) in synthesized speech in FastSpeech 2s
|           ├─ audio1        audio example 1
|           └─ audio2        audio example 2
|   └─ fs2s_pitch          control the pitch in synthesized speech in FastSpeech 2s
|       ├─ audio1          audio example 1
|       └─ audio2          audio example 2
```