FastSpeech: Fast, Robust and Controllable Text to Speech

Yi Ren*, Yangjun Ruan*, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu

Motivation
Due to the long mel-spectrogram sequence and the autoregressive generation, end-to-end TTS models face several challenges:
- Slow inference speed for mel-spectrogram generation.
- Synthesized speech is not robust (word skipping and repeating).
- Synthesized speech is lack of controllability.

Our proposed FastSpeech can address the above-mentioned three challenges as follows:
- Greatly speeds up the mel-spectrogram generation (by 270x).
- Almost eliminate word skipping and repeating.
- Can adjust voice speed and control part of the prosody.

Our Method
Phoneme--->[FastSpeech]--->[Mel-spectrogram]--->[Vocoder]--->Voice

Feed-forward transformer: generate mel-spectrogram in parallel both in training and inference.
- FFT (Feed-Forward Transformer) block: basic block from Transformer, stack N layers.
- Replace dense connection with 1D convolution for speech.
- Share the same model structure between the phoneme and mel side.

Length Regulator: bridge the length mismatch between phoneme and mel sequence.

Duration Predictor is jointly trained with the FastSpeech model to predict the length of mel-spectrograms for each phoneme with the mean square error (MSE) loss. We extract the ground-truth phoneme duration from an autoregressive teacher TTS model as target.

Experiments
All experiments are conducted on LJSpeech dataset. We randomly split the dataset into 3 sets: 12500 samples for training, 300 samples for validation and 300 samples for testing.


Feed-forward transformer: generate mel-spectrogram in parallel both in training and inference.

Model Architecture
(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator (d) Duration Predictor

Experiments
<table>
<thead>
<tr>
<th>Method</th>
<th>Repeats</th>
<th>Skips</th>
<th>Error Sentences</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech 2</td>
<td>12</td>
<td>13</td>
<td>17</td>
<td>100%</td>
</tr>
<tr>
<td>Transformer TTS</td>
<td>12</td>
<td>13</td>
<td>17</td>
<td>100%</td>
</tr>
<tr>
<td>FastSpeech</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 3: The comparison of robustness between FastSpeech and other systems on the 50 particularly hard sentences. Each kind of word error is counted at most once per sentence.

Changing Voice Speed and Adding Breaks

Audio Samples: https://speechresearch.github.io/fastspeech/