FastSpeech: Fast, Robust and Controllable Text to Speech

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

Microsoft

浙江大学 ZHEJIANG UNIVERSITY

Motivation

Due to the long sequence of the mel-spectrogram and the autoregressive nature, end-to-end TTS systems face several challenges:

- Slow inference speed for mel-spectrogram generation.
- Synthesized speech is usually not robust.
- Synthesized speech is lack of controllability.

Our proposed FastSpeech can address the above-mentioned three challenges as follows:

- Greatly speeds up the synthesis process.
- Reduce the ratio of the skipped words and repeated words.
- Easily adjust voice speed and control part of the prosody

Our Method

Phoneme –[Fastspeech]--> Mel-spectrogram ----[waveglow]----> Voice

Feed-forward transformer: generate mel-spectrogram in parallel both in training and inference (speedup)

- FFT (Feed-Forward Transformer) block: basic block from Transformer, stack N layers.
- Replace dense connection with 1D convolution in speech problem.
- Share the same model structure between the phoneme side and mel side.

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.

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

speech

1.0x speed s p iy ch ← 2 2 3 1 (α =1)

s s p p iy iy iy ch

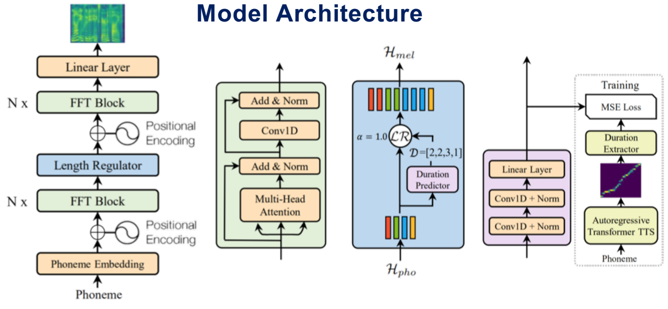
2.0x speed s p iy ch ← 2 2 3 1 (α = 0.5 (1 1 1.5 0.5 → 1 1 2 1))

s p iy iy ch

0.5x speed s p iy ch ← 2 2 3 1 (α = 2 (4 4 6 2))

s s s s p p p p iy iy iy iy iy iy ch ch

Model Architecture



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

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.

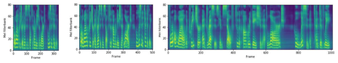
Method	MOS
GT	4.41 ± 0.08
GT (Mel + WaveGlow)	4.00 ± 0.09
Tacotron 2 [22] (Mel + WaveGlow)	3.86 ± 0.09
Merlin [28] (WORLD)	2.40 ± 0.13
Transformer TTS [14] (Mel + WaveGlow)	3.88 ± 0.09
FastSpeech (Mel + WaveGlow)	3.84 ± 0.08

Table 1: The MOS with 95% confidence intervals.

Method	Repeats	Skips	Error Sentences	Error Rate
Tacotron 2	4	11	12	24%
Transformer TTS	7	15	17	34%
FastSpeech	0	0	0	0%

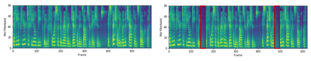
Table 3: The comparison of robustness between FastSpeech and other systems on the 50 particularly hard sentences. Each kind of word error is counted at most once per sentence.

Changing speed and adding breaks



(a) 1.5x Voice Speed (b) 1.0x Voice Speed (c) 0.5x Voice Speed

Figure 3: The mel-spectrograms of the voice with 1.5x, 1.0x and 0.5x speed respectively. The input text is "*For a while the preacher addresses himself to the congregation at large, who listen attentively*".



(a) Original Mel-spectrograms (b) Mel-spectrograms after adding breaks

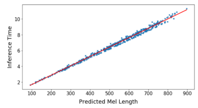
Figure 4: The mel-spectrograms before and after adding breaks between words. The corresponding text is "*that he appeared to feel* ***deeply*** *the force of the reverend gentleman's observations,* ***especially*** *when the chaplain spoke of*". We add breaks after the words "*deeply*" and "*especially*" to improve the prosody. The red boxes in Figure 4b correspond to the added breaks.

Experiments

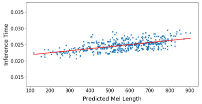
Inference Latency

Method	Latency (s)	Speedup
Transformer TTS [14] (Mel)	6.735 ± 3.969	/
FastSpeech (Mel)	0.025 ± 0.005	269.40×
Transformer TTS [14] (Mel + WaveGlow)	6.895 ± 3.969	/
FastSpeech (Mel + WaveGlow)	0.180 ± 0.078	38.30×

Table 2: The comparison of inference latency with 95% confidence intervals. The evaluation is conducted on a server with 12 Intel Xeon CPU, 256GB memory, 1 NVIDIA V100 GPU and batch size of 1. The average length of the generated mel-spectrograms for the two systems are both about 560.



(b) Transformer TTS



(a) FastSpeech

Ablation Studies

System	CMOS
FastSpeech	0
FastSpeech without 1D convolution in FFT block	-0.113
FastSpeech without sequence-level knowledge distillation	-0.325

Table 4: CMOS comparison in the ablation studies.

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

taoqin@microsoft.com
rayeren@zju.edu.cn
xuta@microsoft.com