FastSpeech 2: A Text-to-Speech AI Model

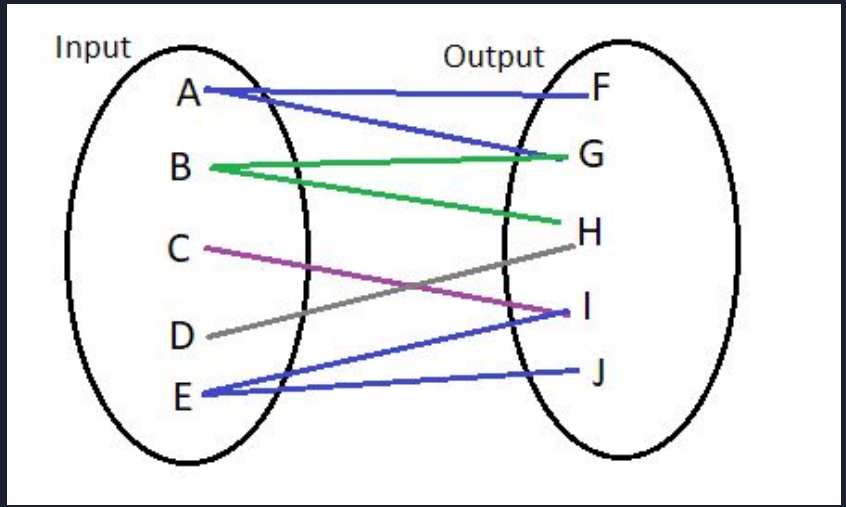
Luke Buchanan, Seth Wertz

Text-to-Speech Significance

- Increased Time Spent Online
- Accessibility
  - Visual
  - Cognitive
  - Age
- Efficiency

Challenges

- One-to-Many Mapping Problem
  - Example: Enough, through, cough, bought, drought, dough
- Human Speech Variance



FastSpeech 1

- Microsoft and Zhejiang University 2019
- Phoneme Encoding
- The Student-Teacher Model
- Autoregression

FastSpeech 2 Model Overview

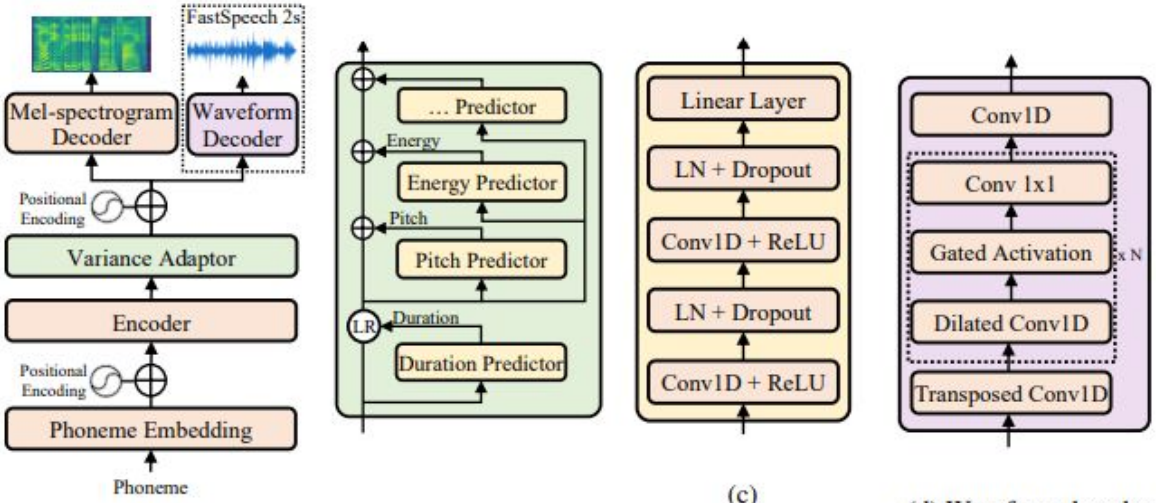


Figure 1: The overall architecture for FastSpeech 2 and 2s. LR in subfigure (b) denotes the length regulator proposed in FastSpeech. LN in subfigure (c) denotes layer normalization.

FastSpeech 2 Model Overview

- Encoder based on self-attention transformer layers
  - Uses phoneme embeddings
- Variance adaptor
  - Predicts duration, pitch, and energy of phonemes
  - Each predictor uses CNN structure
- Convolutional layer decoder

FastSpeech 2s

- Fully end-to-end TTS module
  - Doesn’t generate intermediate spectrogram
- Generates speech waveform directly from text
- Higher quality speech than Fastspeech 1
- Faster voice synthesis than Fastspeech 2

Results

- Trained on 13,000 English voice samples
- Mean Opinion Score (MOS)
- Ground Truth (GT)

Method	MOS
GT	4.30 ± 0.07
GT (Mel + PWG)	3.92 ± 0.08
Tacotron 2 (Shen et al., 2018) (Mel + PWG)	3.70 ± 0.08
Transformer TTS (Li et al., 2019) (Mel + PWG)	3.72 ± 0.07
FastSpeech (Ren et al., 2019) (Mel + PWG)	3.68 ± 0.09
FastSpeech 2 (Mel + PWG)	3.83 ± 0.08
FastSpeech 2s	3.71 ± 0.09

(a) The MOS with 95% confidence intervals.

Method	CMOS
FastSpeech 2	0.000
FastSpeech	-0.885
Transformer TTS	-0.235

(b) CMOS comparison.

Table 1: Audio quality comparison.

Model Speedup

- Faster to Train, Slower Inference

Method	Training Time (h)	Inference Speed (RTF)	Inference Speedup
Transformer TTS (Li et al., 2019)	38.64	9.32×10^{-1}	/
FastSpeech (Ren et al., 2019)	53.12	1.92×10^{-2}	48.5×
FastSpeech 2	**17.02**	1.95×10^{-2}	47.8×
FastSpeech 2s	92.18	$\mathbf{1.80 \times 10^{-2}}$	**51.8**×

Demo

https://cmchien.ttic.edu/FastSpeech2/

https://github.com/ming024/FastSpeech2