CS11-737 Multilingual NLP

Text-to-Speech Synthesis

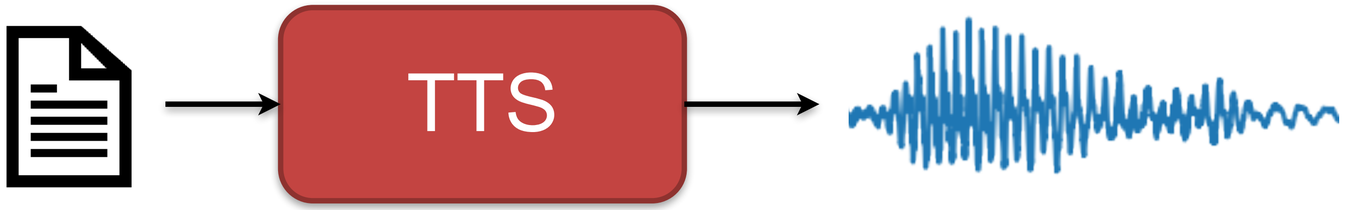
Lei Li

https://lileicc.github.io/course/11737mnlp23fa/

Carnegie Mellon University
Language Technologies Institute

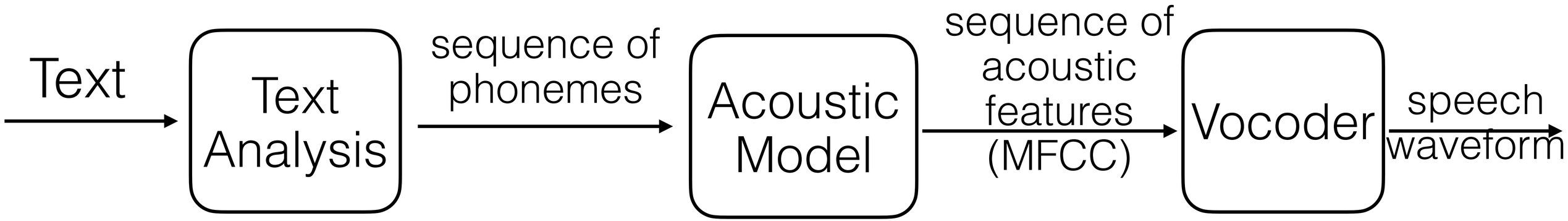
Text-to-Speech Synthesis (TTS)

- produce speech waveform from text input



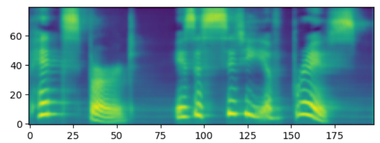
Inverse problem of ASR

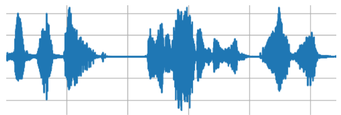
TTS Pipeline



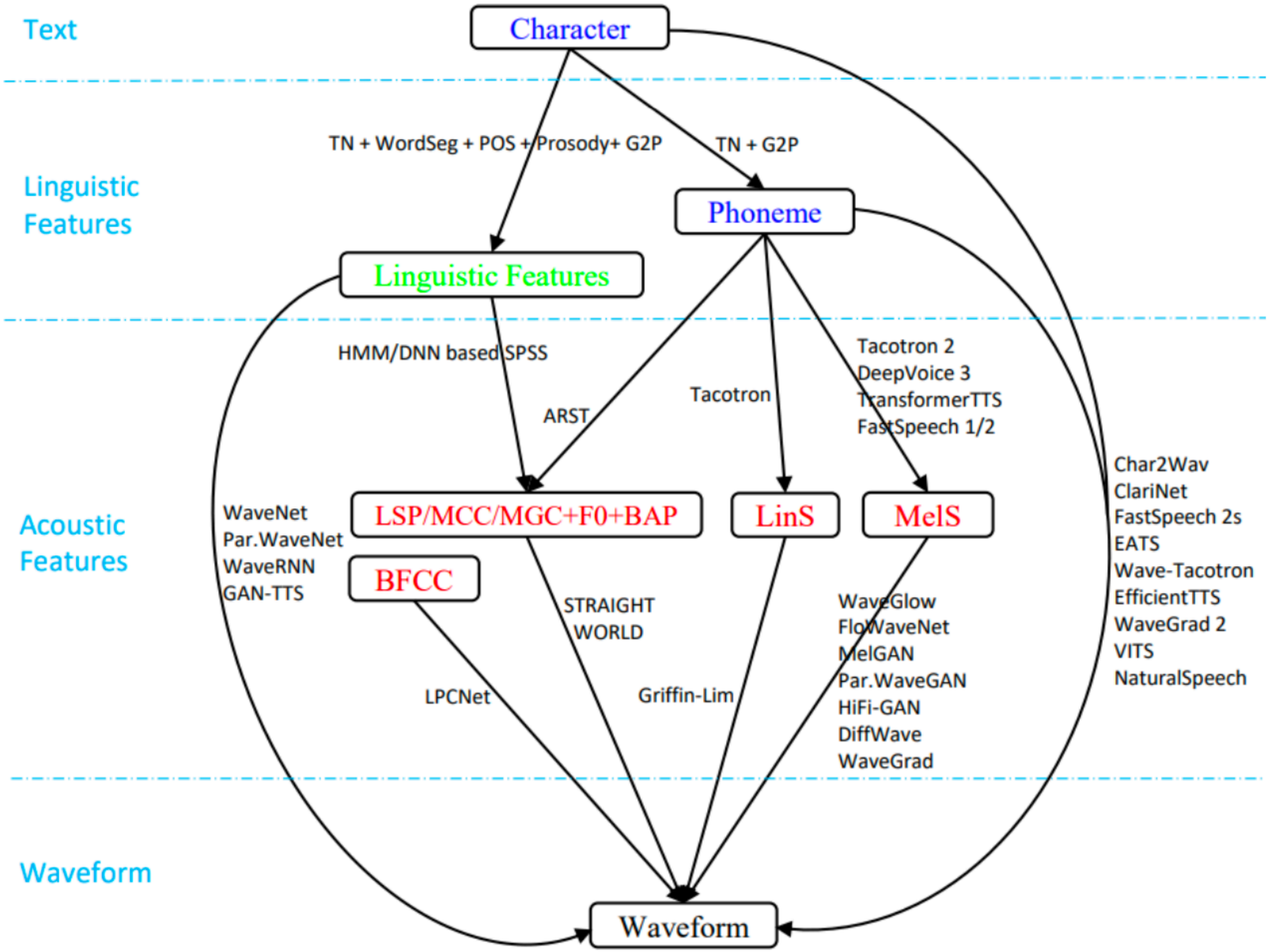
Pittsburgh is a city of bridge.

'P', 'IH', 'T', 'S', 'B',
'ER', 'G', ' ', 'IH', 'Z', ' ',
'AH', ' ', 'S', 'IH', 'T',
'IY', ' ', 'AH', 'V', ' ', 'B',
'R', 'IH', 'JH', '.'





TTS technologies

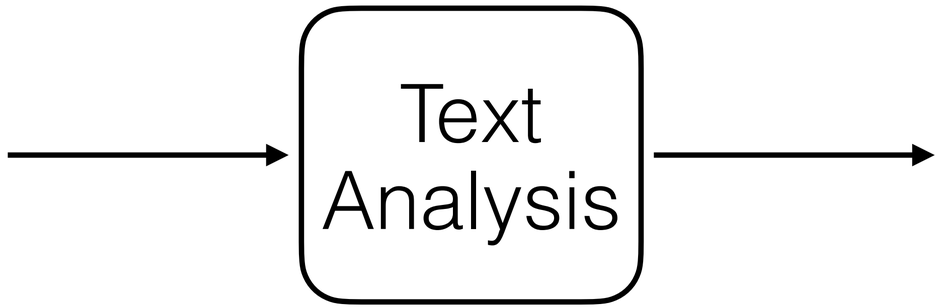


TTS Pipeline — Text Analysis

- Transform text into linguistic features:
  - text normalization:
    - 1989 -> nineteen eighty nine
    - Jan. 24 -> January twenty-fourth
  - homograph disambiguation:
    - do you live (/l ih v/) near a zoo with live (/l ay v/) animals?
  - Grapheme-to-phoneme conversion
    - speech -> s p iy ch
  - ToBI (Tones and Break Indices)
  - Phrase/word/syllable segmentation
  - Part-of-speech tagging

Text to Phoneme

Pittsburgh is a city of bridge.



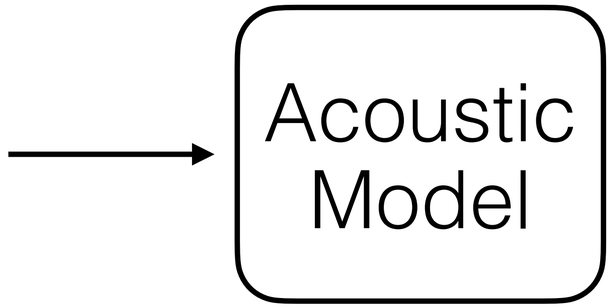
'P', 'IH', 'T', 'S', 'B', 'ER', 'G', ' ', 'IH', 'Z', ' ', 'AH', ' ', 'S', 'IH', 'T', 'IY', ' ', 'AH', 'V', ' ', 'B', 'R', 'IH', 'JH', '.'

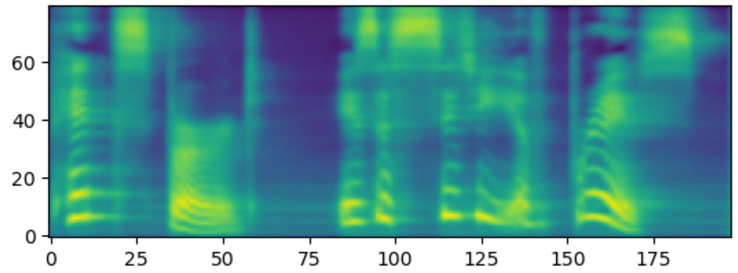
Acoustic Model

- Transform a sequence of phonemes intro audio features

- Mel-scale Frequency Cepstral Coefficients (MFCC)
  - Tactron uses 80 channel MFCC, 50ms per frame, 12.5ms frame shifting.

'P', 'IH', 'T', 'S', 'B',
'ER', 'G', ' ', 'IH', 'Z', ' ',
'AH', ' ', 'S', 'IH', 'T',
'IY', ' ', 'AH', 'V', ' ', 'B',
'R', 'IH', 'JH', '.'

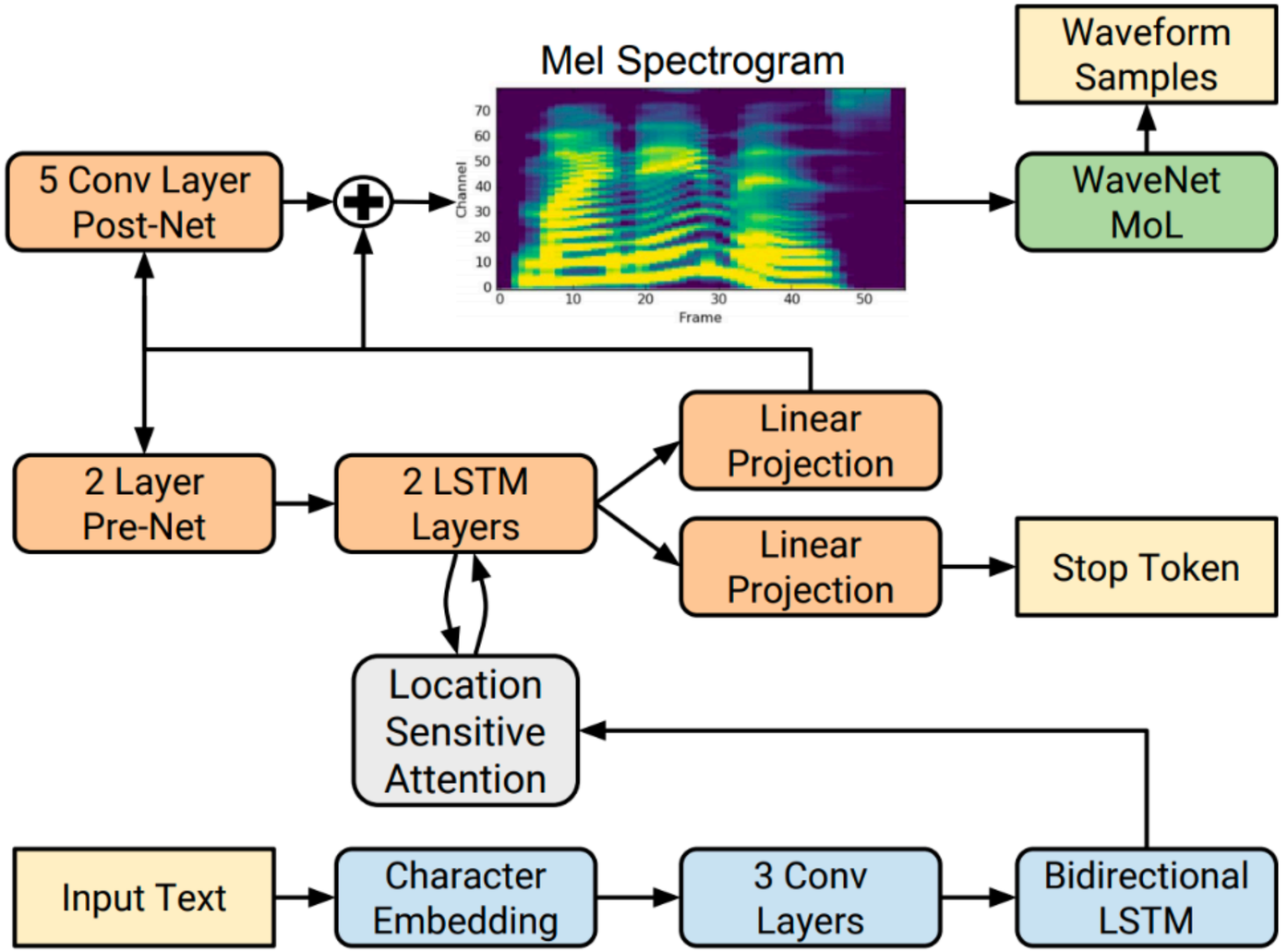




Mel-spectrogram

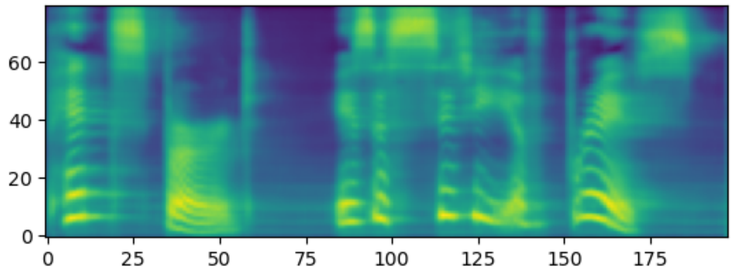
Tacotron2

- RNN based approach



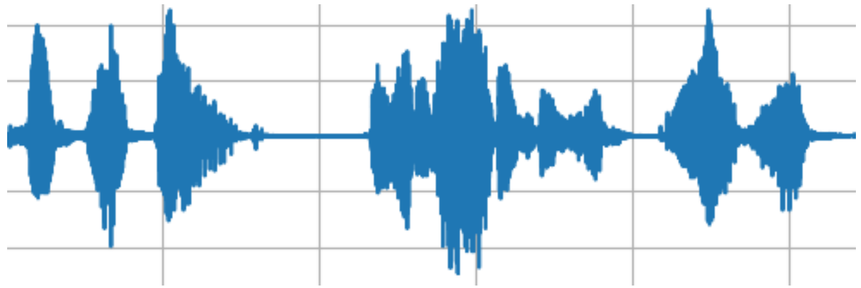
Vocoder

- Transform acoustic features (mel-spectrogram) to speech waveform signals



Mel-spectrogram

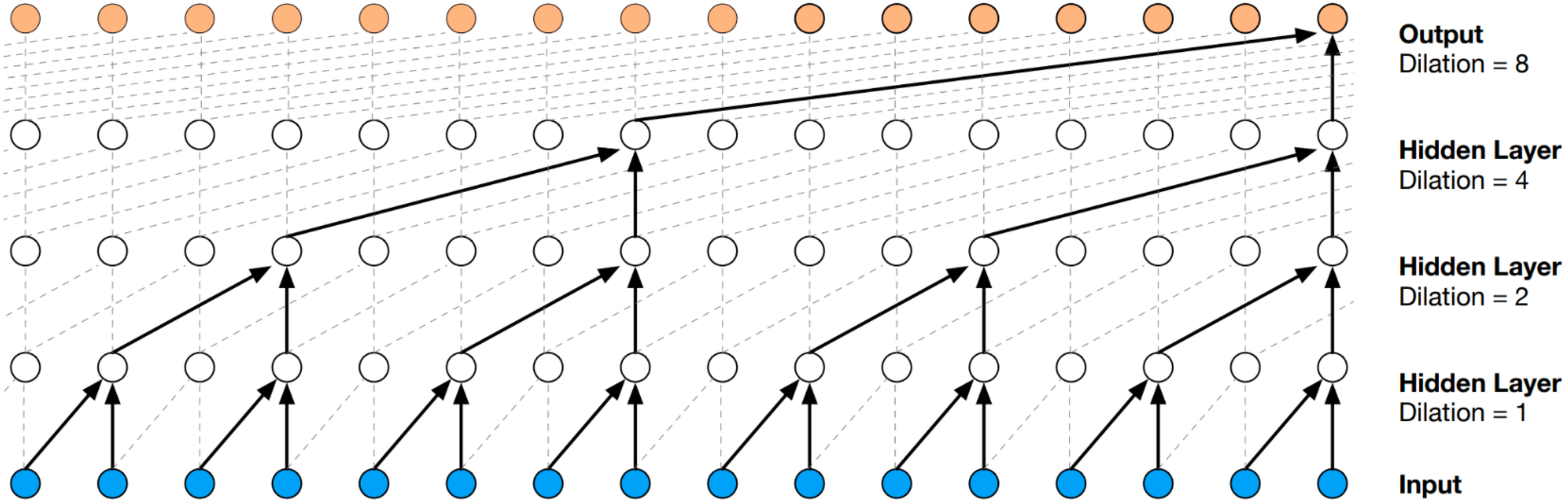
Vocoder



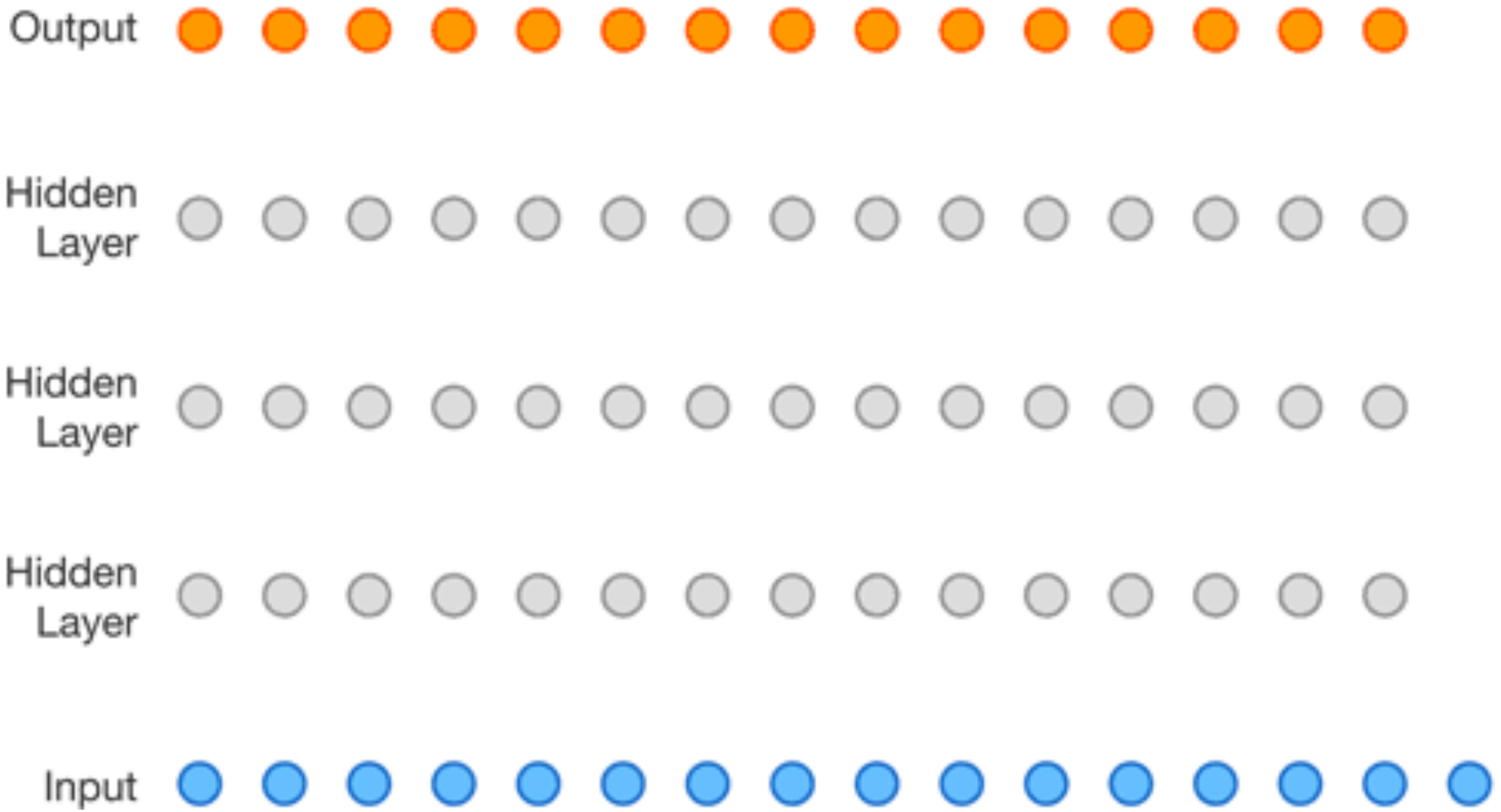
“Pittsburgh is a city of bridge”

Vocoder — WaveNet

- autoregressive model with dilated causal convolution



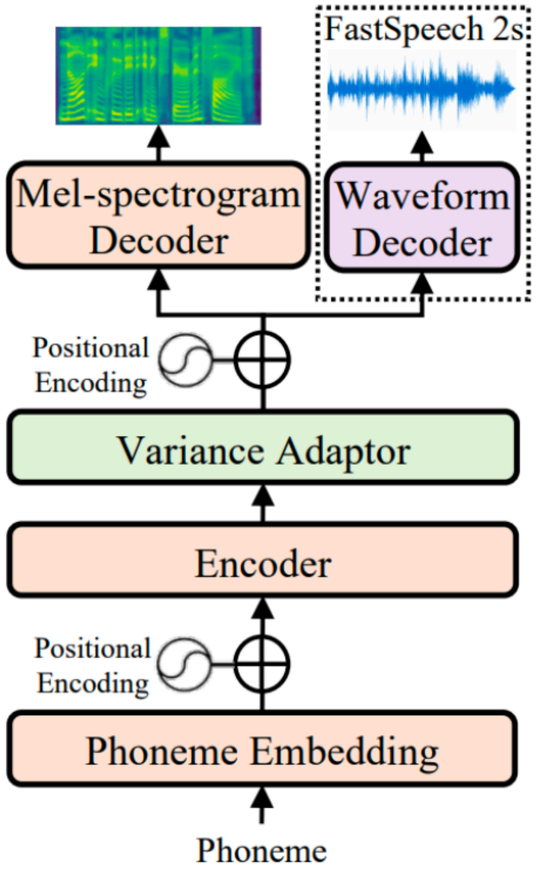
WaveNet



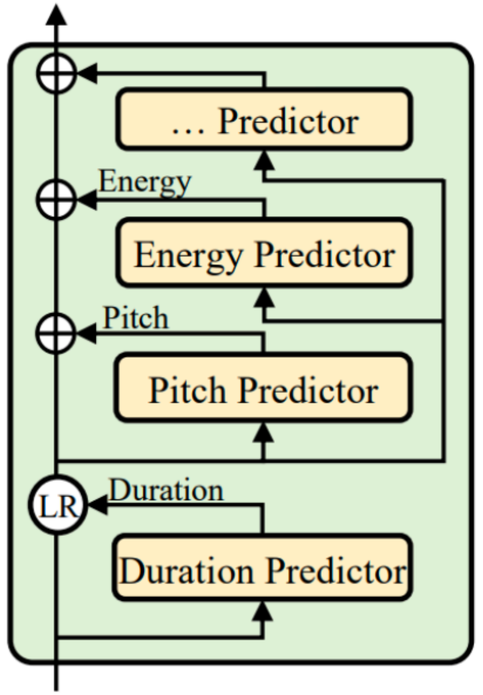
End-to-end TTS

FastSpeech/FastSpeech2/2s

- Generate mel-spectrogram in parallel
- use variance adaptor to predict duration, pitch, energy
- FastSpeech2s: generating wave directly

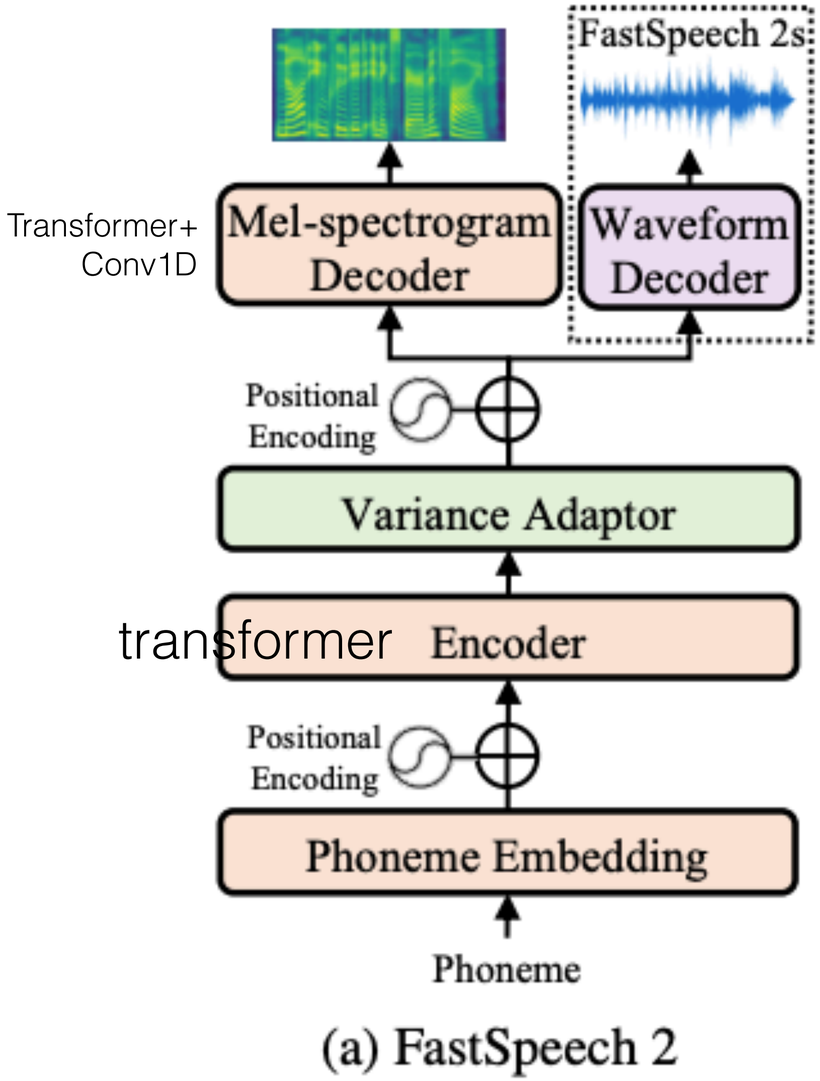


(a) FastSpeech 2

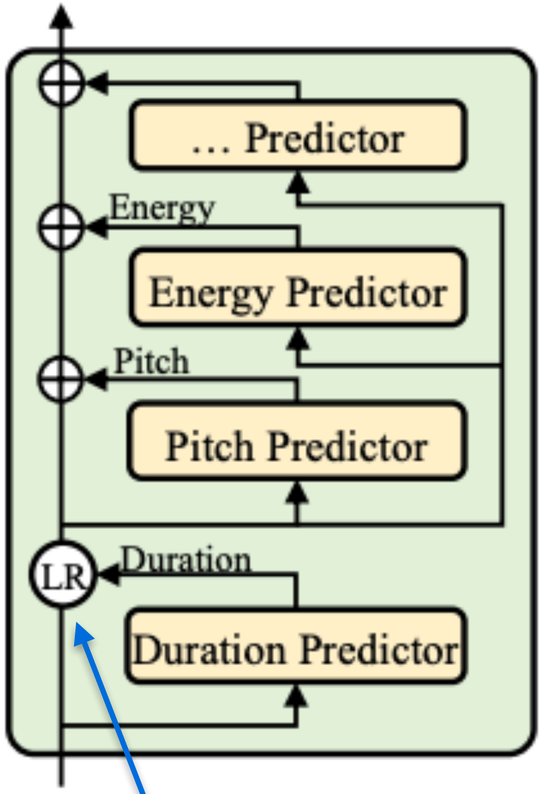


(b) Variance adaptor

FastSpeech2/2s



(a) FastSpeech 2



(b) Variance adaptor

length reg

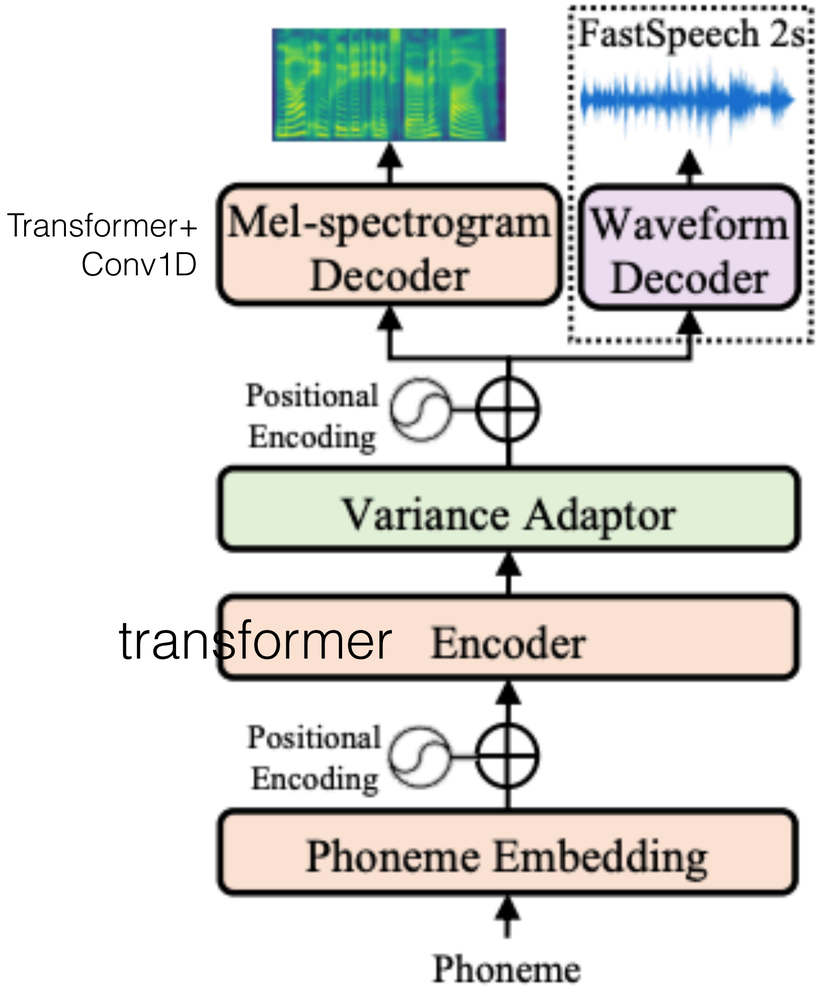
the amplitude of STFT for each frame, discrete to 256 and map to embedding

predicts F_0 of each phoneme, map to 256 values in log-scale and embedding vector

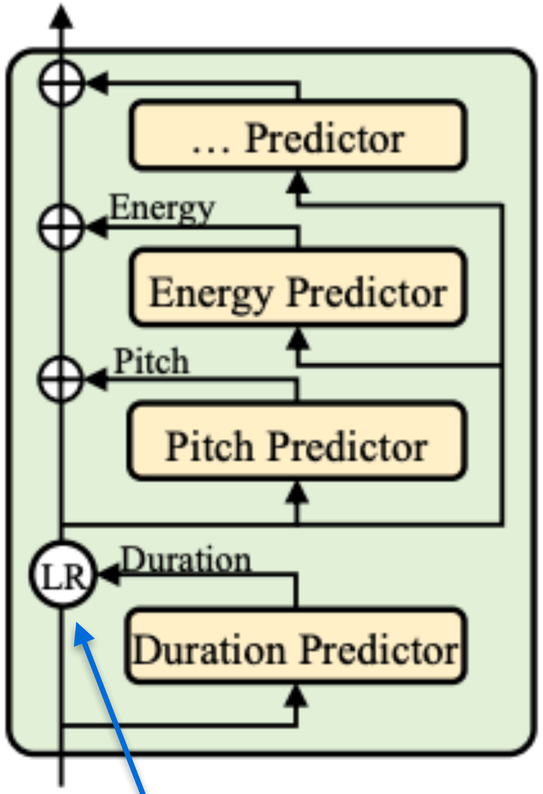
predicts num. of mel frames of each phoneme

Montreal forced alignment (MFA) tool to construct groudtruth

FastSpeech2s

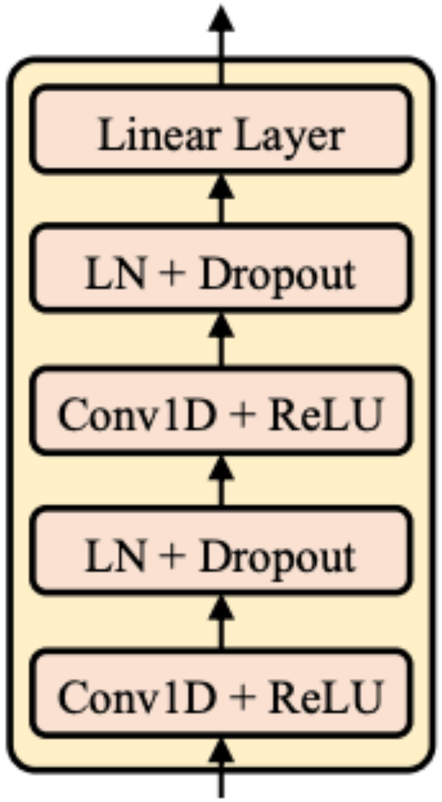


(a) FastSpeech 2

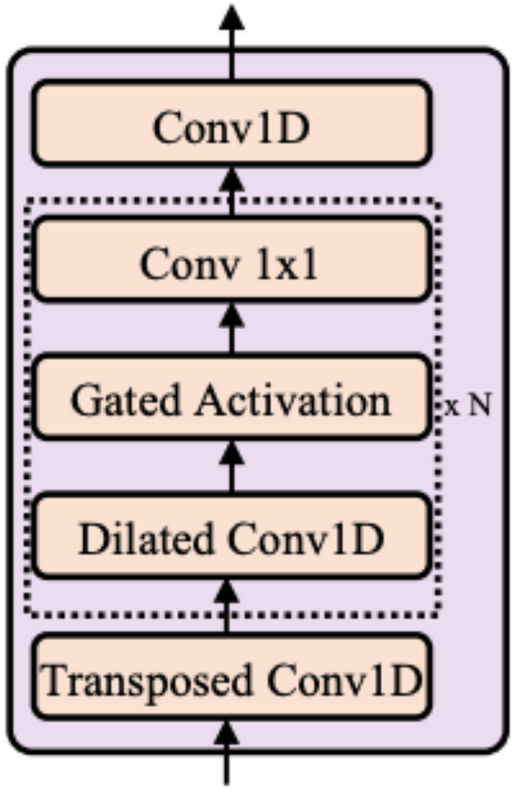


(b) Variance adaptor

length reg



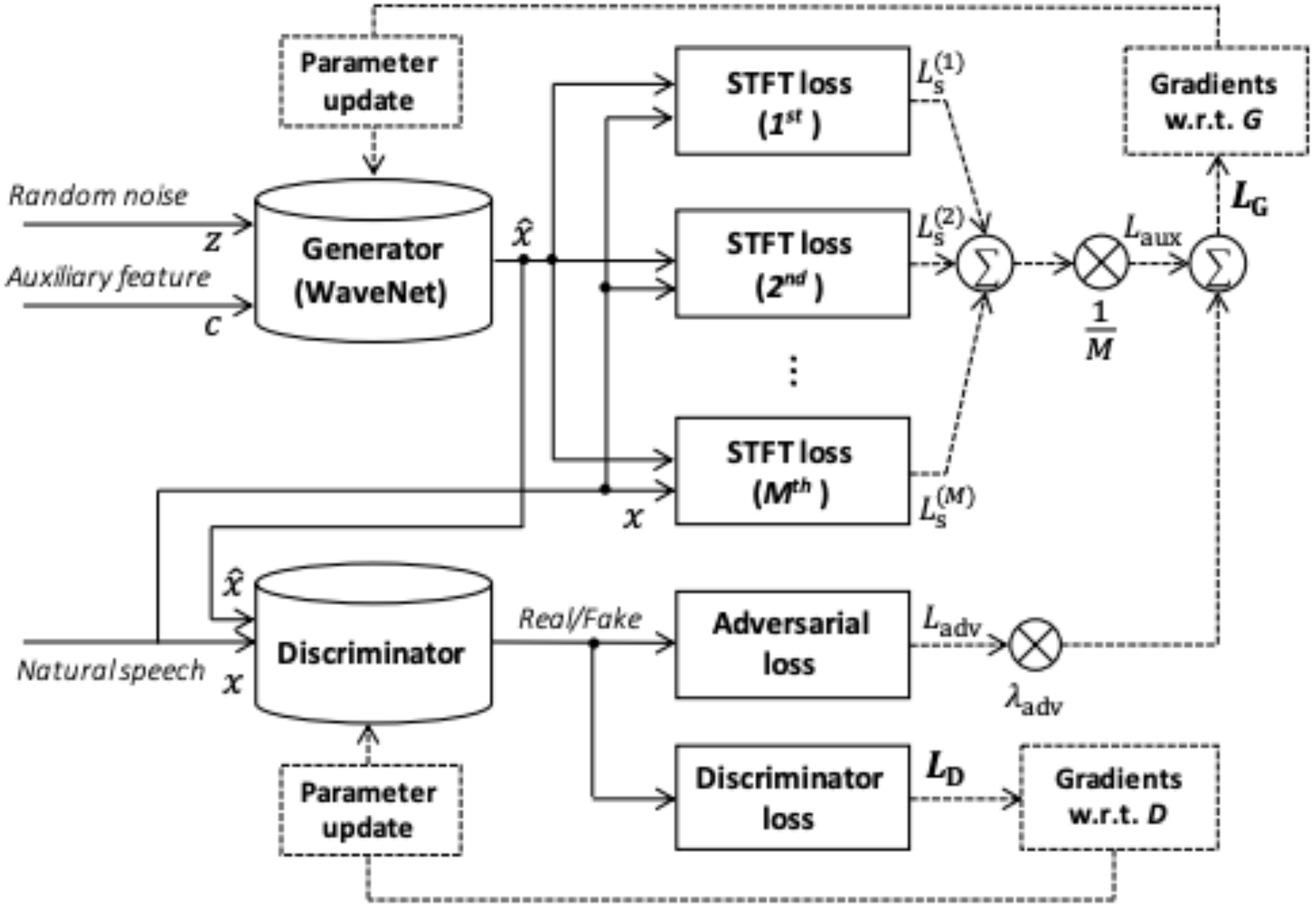
(c)
Duration/pitch/energy predictor



(d) Waveform decoder
similar to WaveNet

Training FastSpeech2s

use loss from Parallel WaveGAN



Code Example

- see python notebook

Summary

- Text preprocessing for TTS
- Acoustic model to generate acoustic features for each frame
- Vocoder to generate waveform
- FastSpeech2s: end-to-end tts

Language in 10

Code Walkthrough

- https://github.com/ming024/FastSpeech2