CS11-737 Multilingual NLP

Text-to-Speech Synthesis

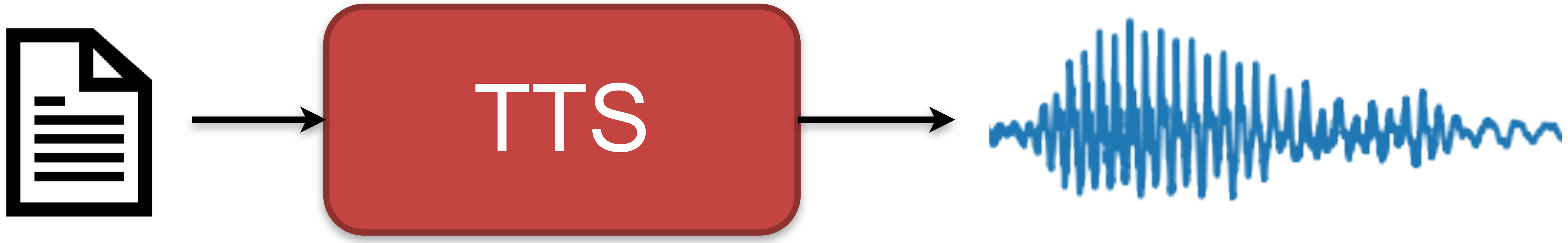
Lei Li

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

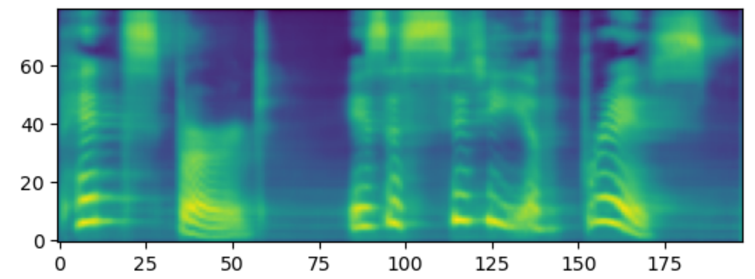
TTS Pipeline

Text
Text Analysis
sequence of phonemes
Acoustic Model
sequence of acoustic features (MFCC)
Vocoder
speech waveform

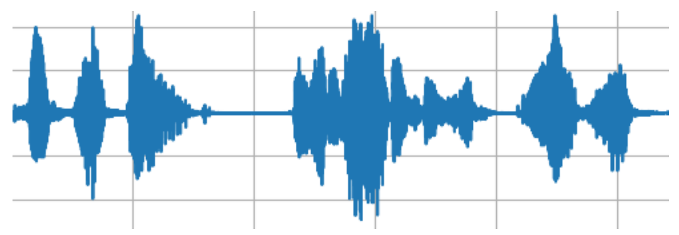
Pittsburgh is a city of bridge. →

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

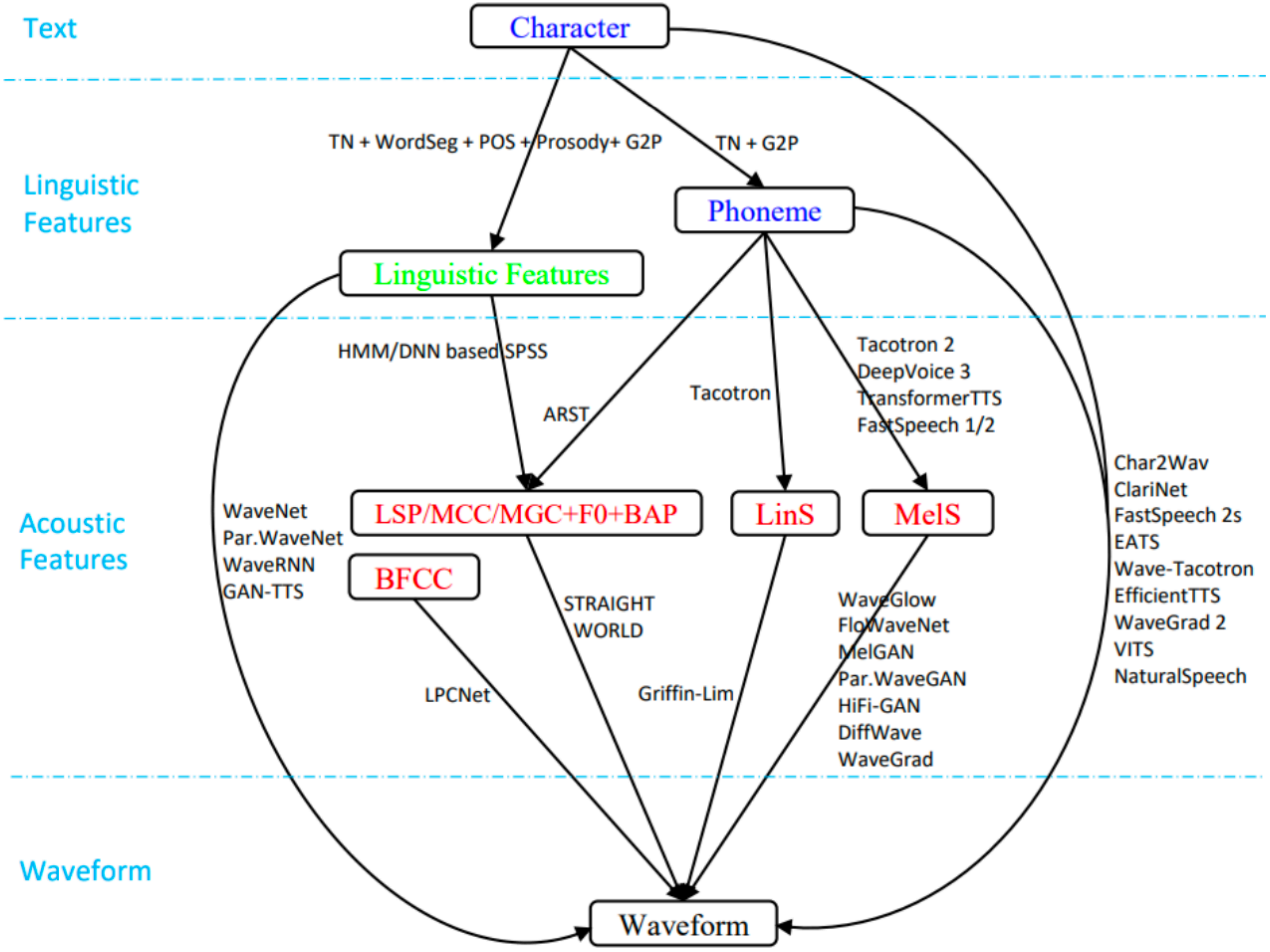
→



→



TTS technologies



Text

Character

TN + WordSeg + POS + Prosody+ G2P

TN + G2P

Linguistic Features

Phoneme

Linguistic Features

HMM/DNN based SPSS

ARST

Tacotron

Tacotron 2
DeepVoice 3
TransformerTTS
FastSpeech 1/2

Acoustic Features

WaveNet
Par.WaveNet
WaveRNN
GAN-TTS

LSP/MCC/MGC+F0+BAP

LinS

MelS

BFCC

Char2Wav
ClariNet
FastSpeech 2s
EATS
Wave-Tacotron
EfficientTTS
WaveGrad 2
VITS
NaturalSpeech

STRAIGHT
WORLD

LPCNet

Griffin-Lim

WaveGlow
FloWaveNet
MelGAN
Par.WaveGAN
HiFi-GAN
DiffWave
WaveGrad

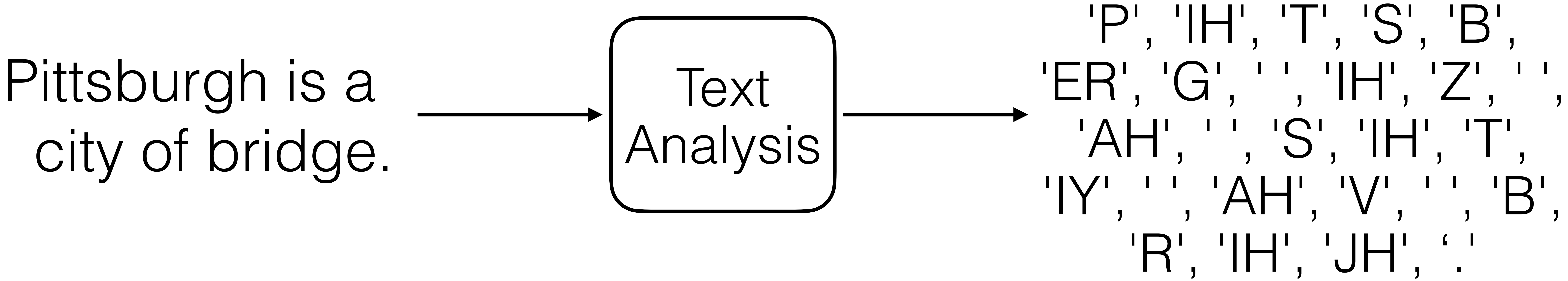
Waveform

Waveform

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.

Text Analysis

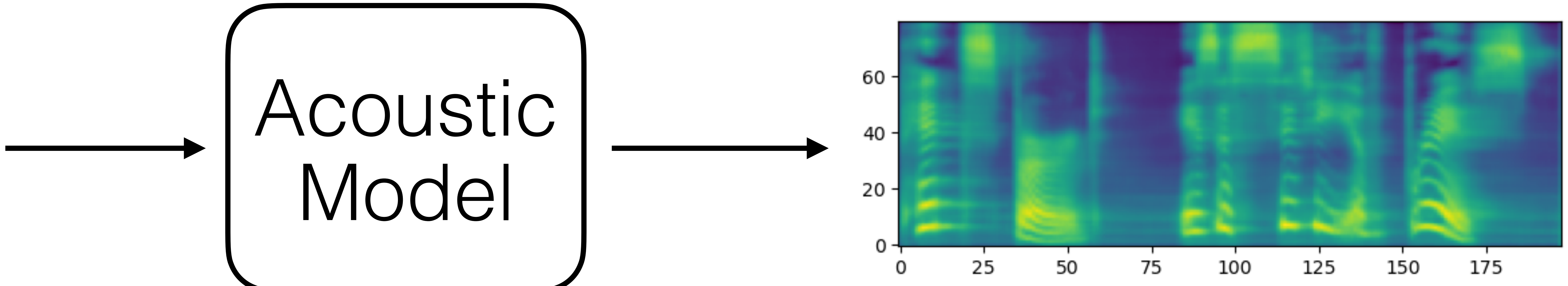
'P', 'IH', 'T', 'S', 'B', 'ER', 'G', ' ', 'IH', 'Z', ' ', 'AH', ' ', 'S', 'IH', 'T', 'IY', ' ', 'AH', 'V', ' ', 'B', 'R', 'IH', 'JH', ‘.'

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', ‘.'

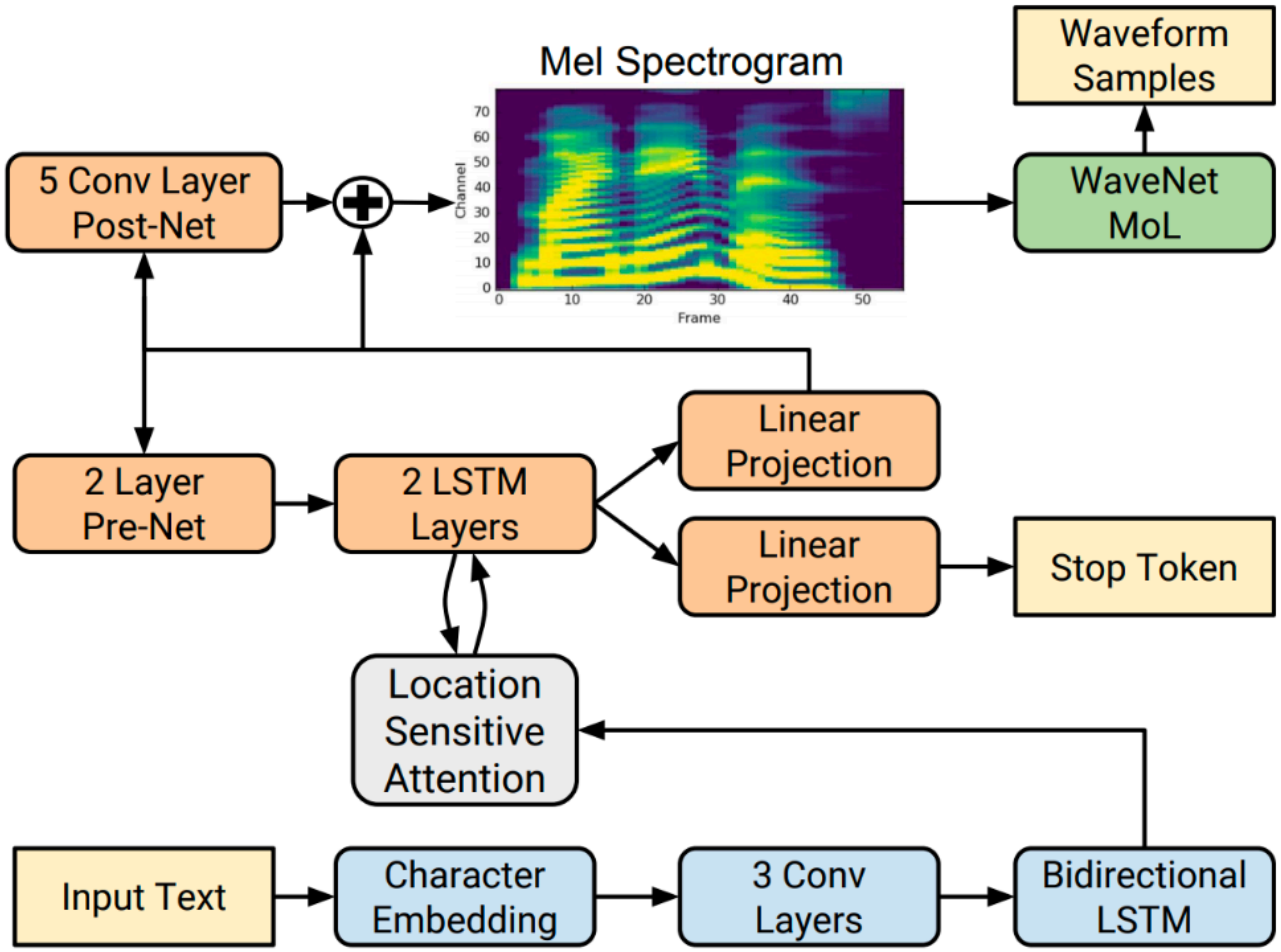


Acoustic Model

Mel-spectrogram

Tacotron2

- RNN based approach



Vocoder

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

60
40
20
0
0
25
50
75
100
125
150
175

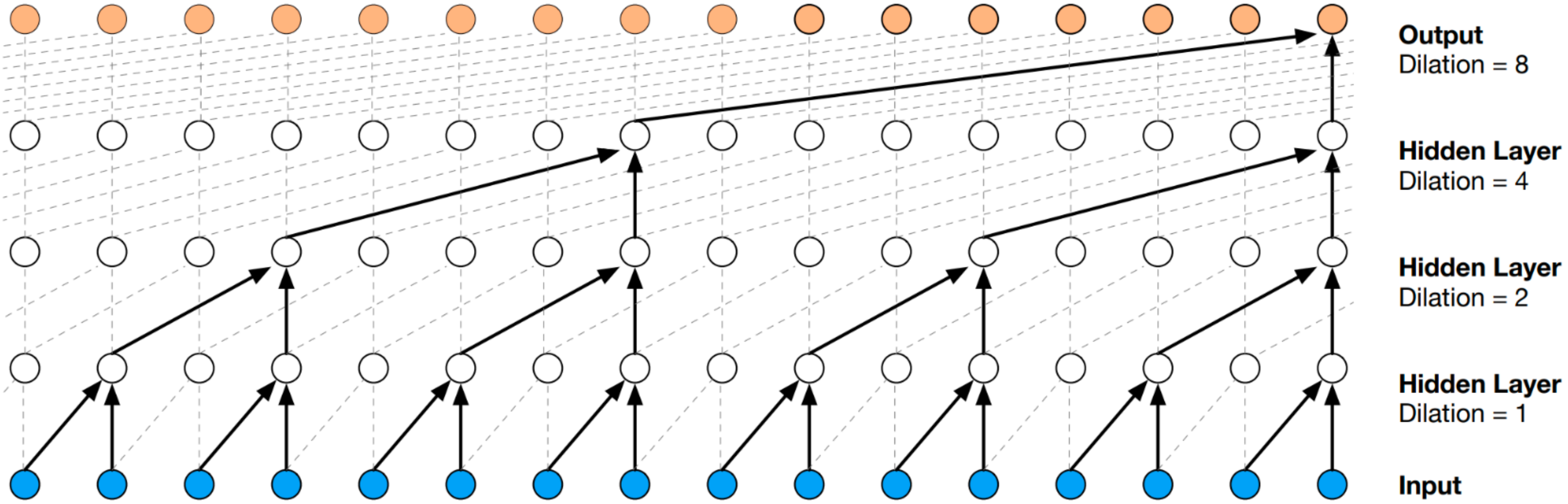
Vocoder

Mel-spectrogram

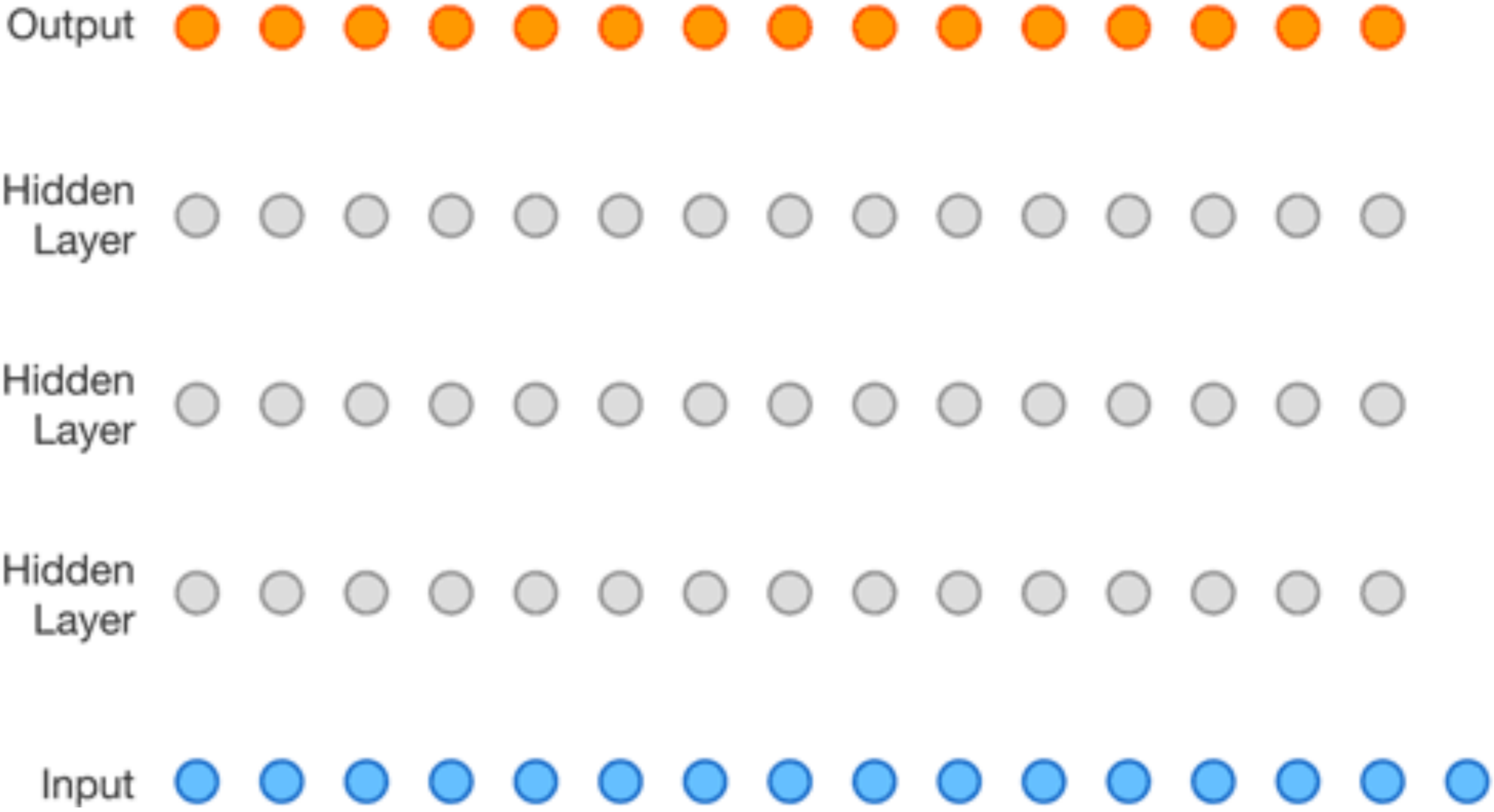
“Pittsburgh is a city of bridge”

Vocoder — WaveNet

- autoregressive model with dilated causal convolution



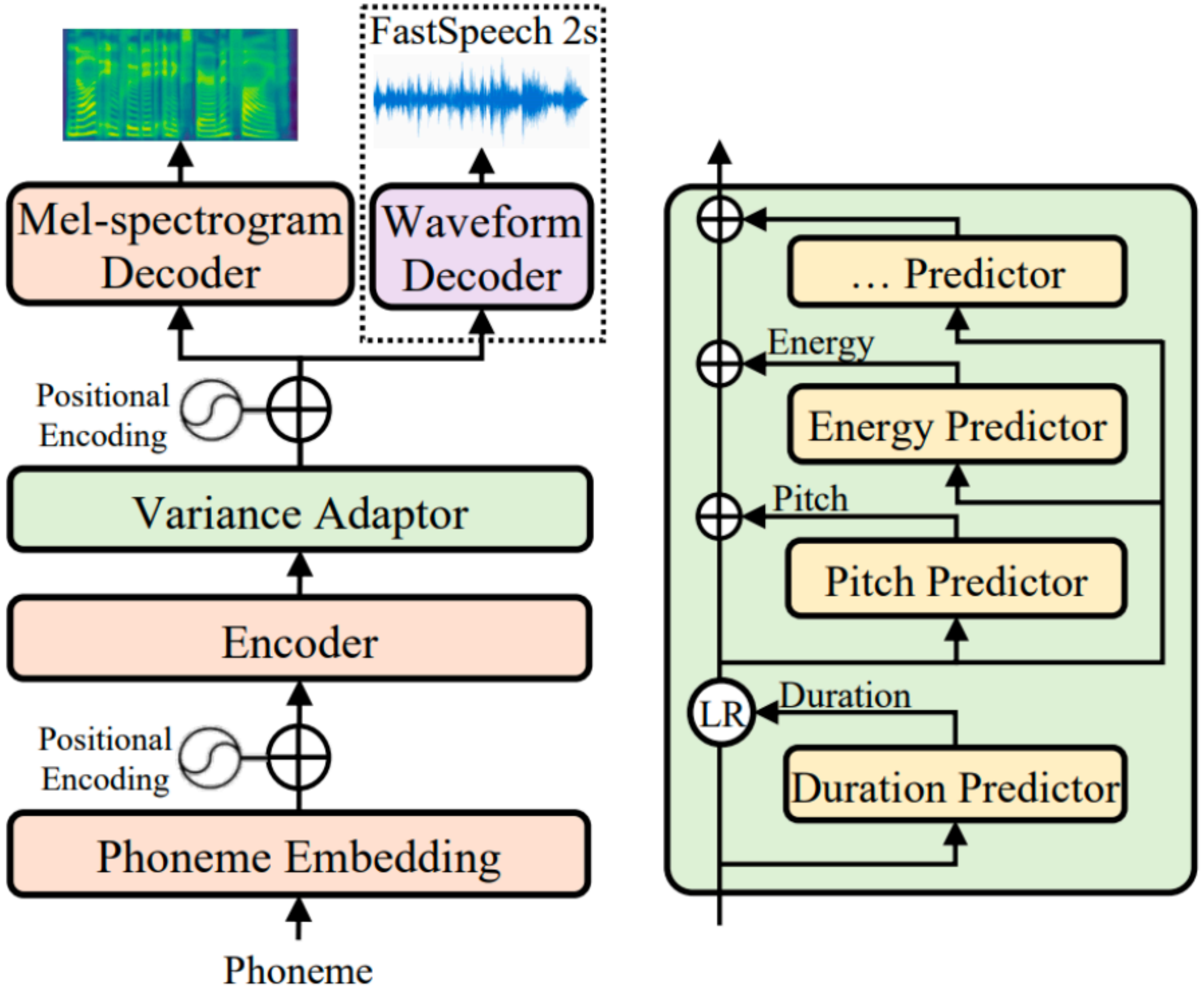
WaveNet



End-to-end TTS

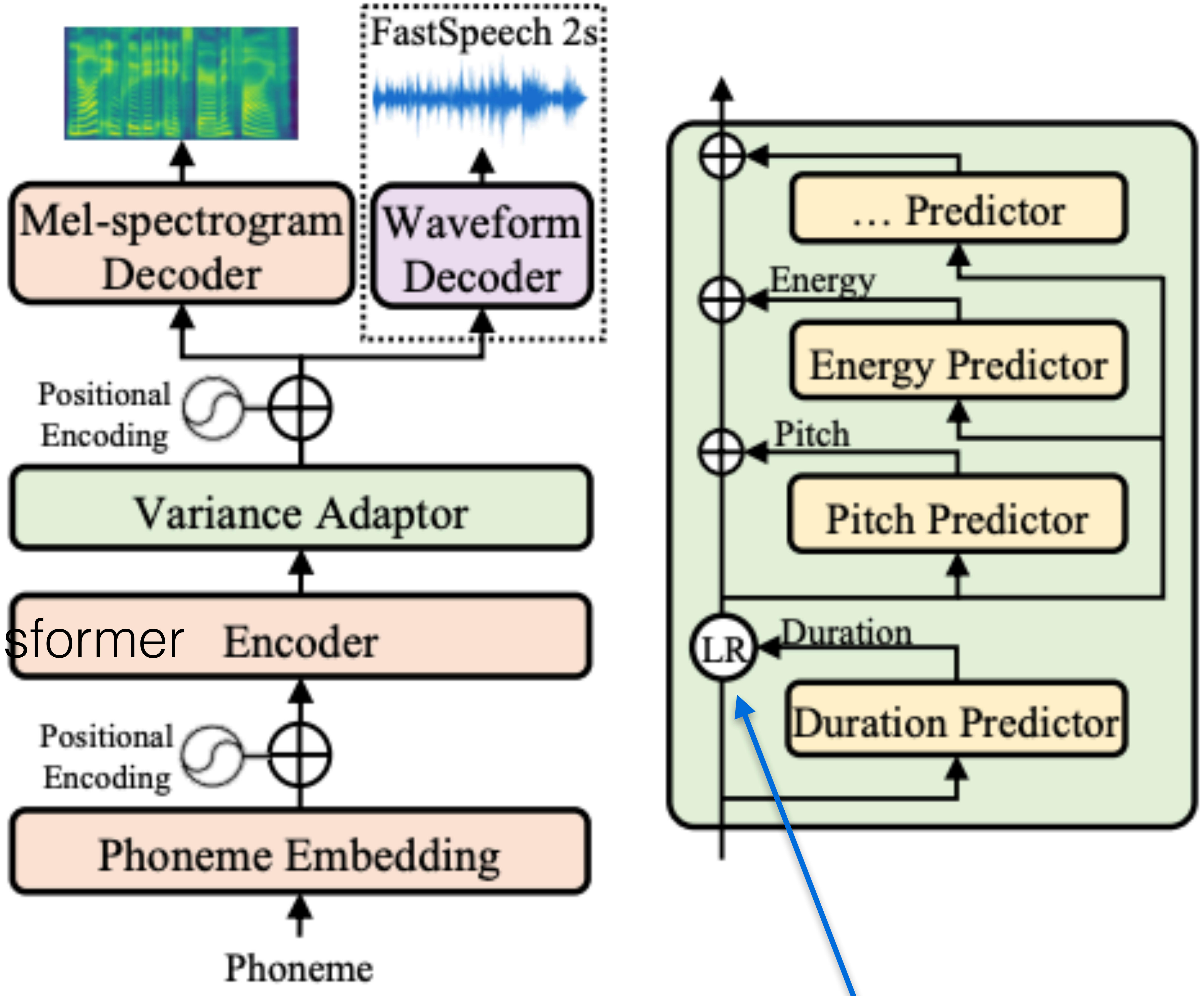
FastSpeech/FastSpeech2/2s

- 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



Transformer+ Conv1D

transformer

(a) FastSpeech 2

(b) Variance adaptor

length reg

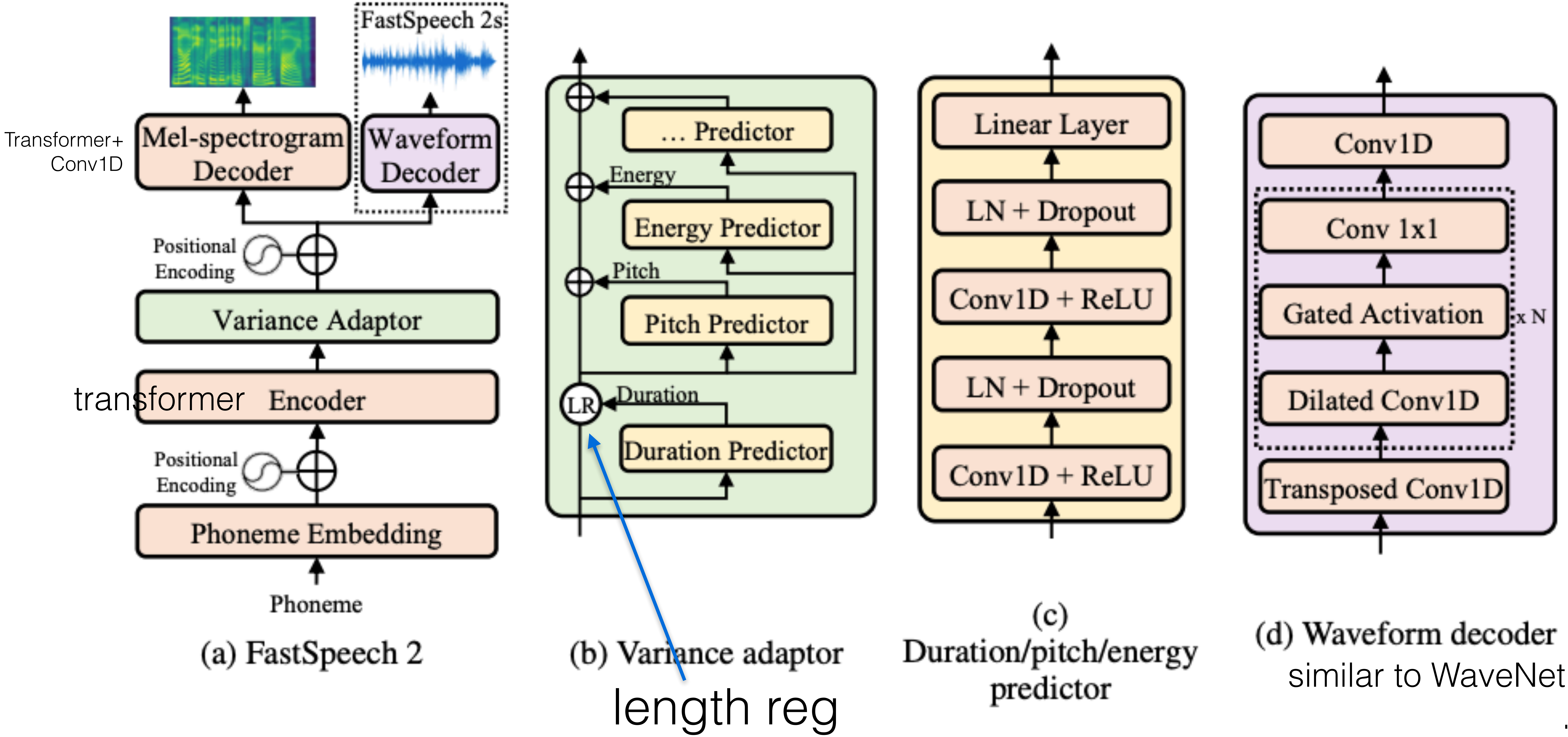
the amplitude of STFT for each frame, discrete to 256 and map to embedding

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

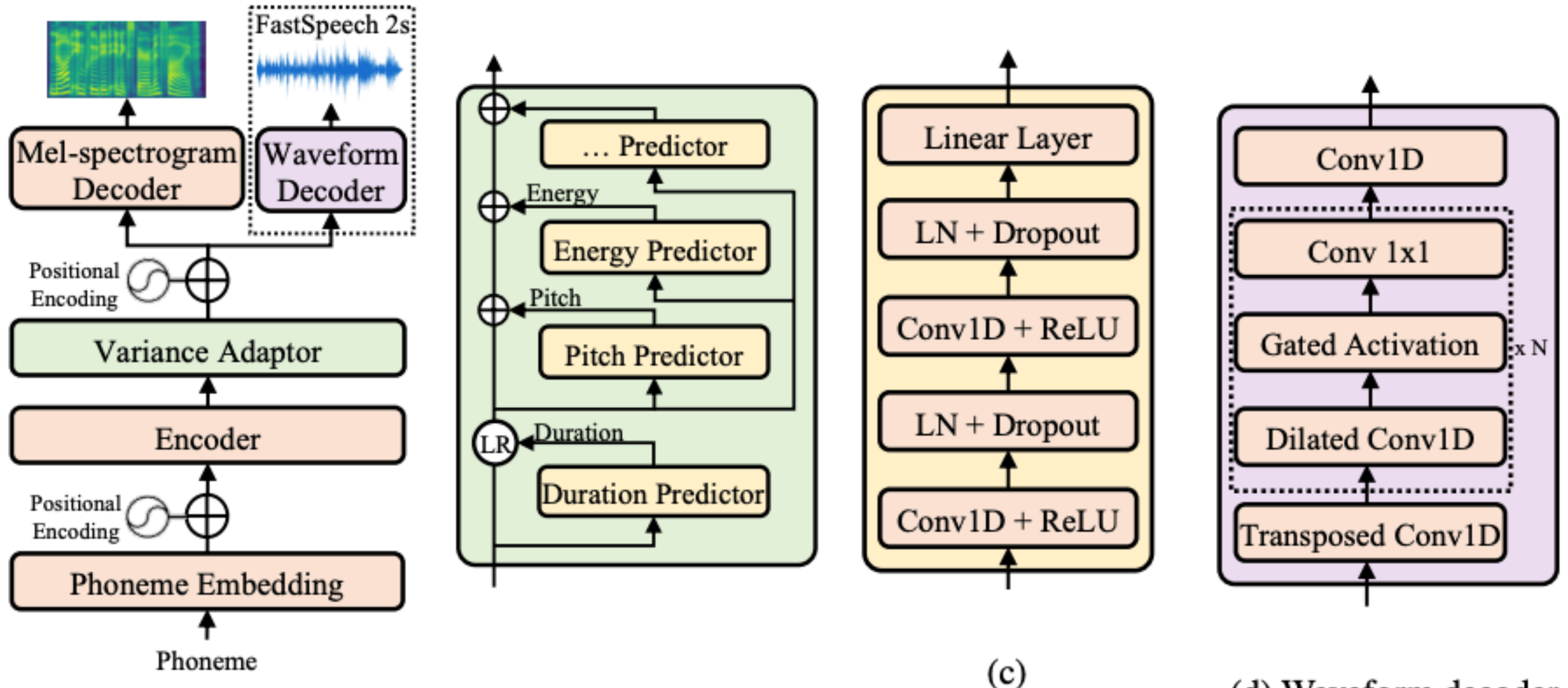
(b) Variance adaptor

(c) Duration/pitch/energy predictor

(d) Waveform decoder

similar to WaveNet

Model Setup



(a) FastSpeech 2

(b) Variance adaptor

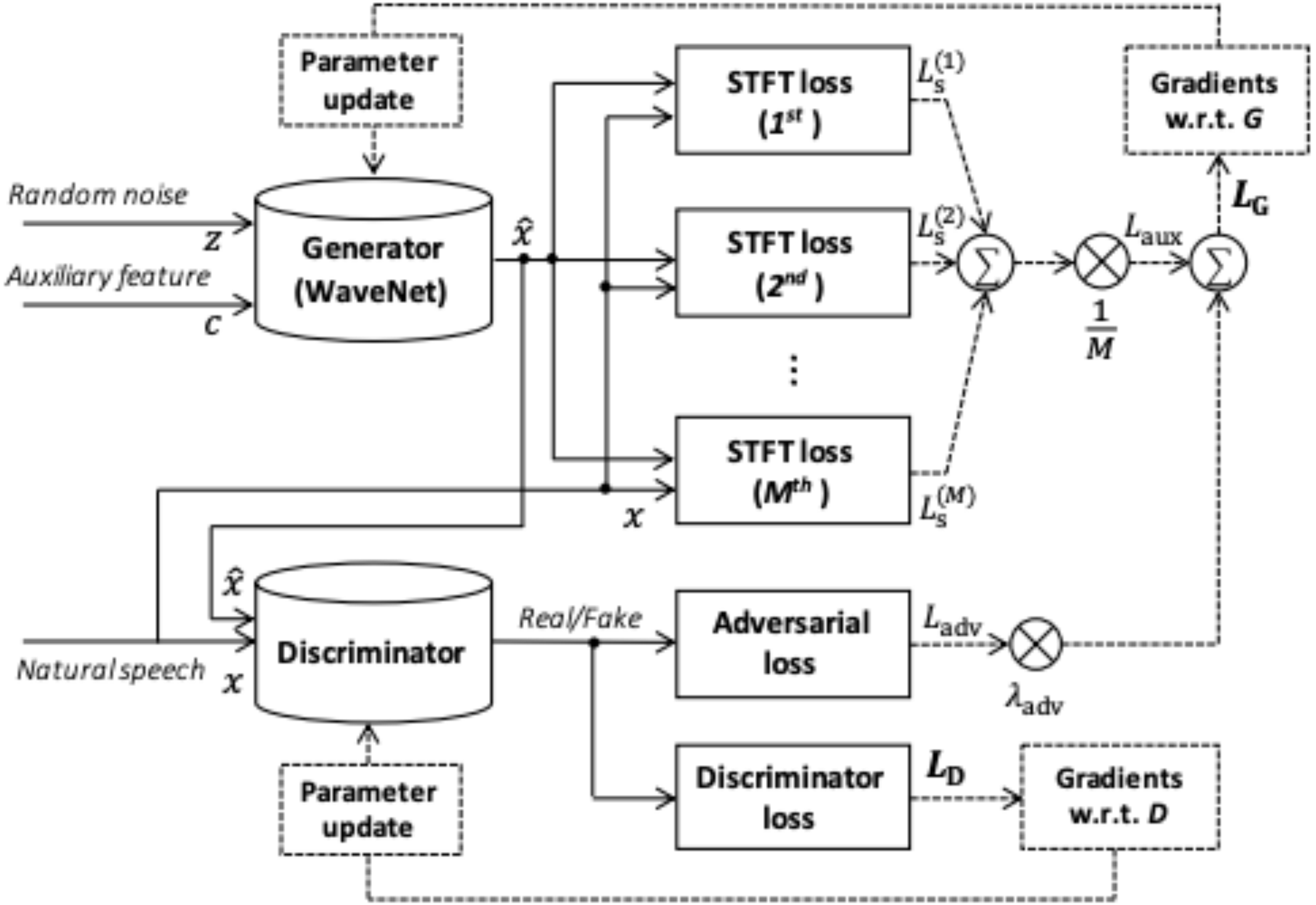
(c) Duration/pitch/energy predictor

(d) Waveform decoder

Hyperparameter	**FastSpeech/FastSpeech 2/2s**
Phoneme Embedding Dimension	256
Pre-net Layers	/
Pre-net Hidden	/
Encoder Layers	4
Encoder Hidden	256
Encoder Conv1D Kernel	9
Encoder Conv1D Filter Size	1024
Encoder Attention Heads	2
Mel-Spectrogram Decoder Layers	4
Mel-Spectrogram Decoder Hidden	256
Mel-Spectrogram Decoder Conv1D Kernel	9
Mel-Spectrogram Decoder Conv1D Filter Size	1024
Mel-Spectrogram Decoder Attention Headers	2
Encoder/Decoder Dropout	0.1
Variance Predictor Conv1D Kernel	3
Variance Predictor Conv1D Filter Size	256
Variance Predictor Dropout	0.5
Waveform Decoder Convolution Blocks	30
Waveform Decoder Dilated Conv1D Kernel size	3
Waveform Decoder Transposed Conv1D Filter Size	64
Waveform Decoder Skip Channlel Size	64
Batch Size	48/48/12
Total Number of Parameters	23M/27M/28M

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