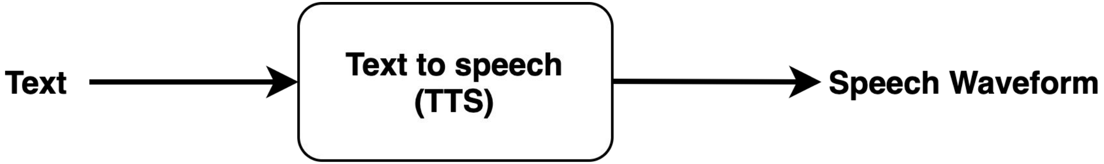
Neural Text To Speech Synthesis

Abdellah EL MEKKI

Text To Speech Synthesis

- The artificial production of human speech from text.

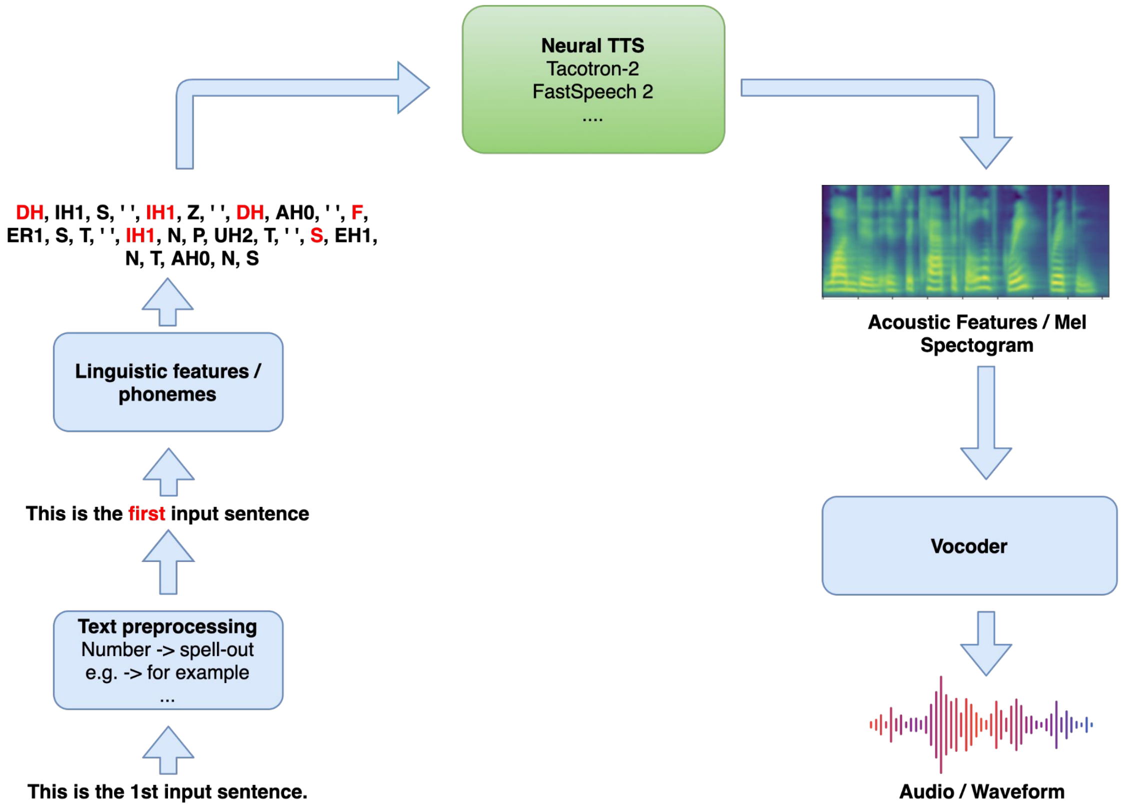


- TTS technologies
  - Concatenative speech synthesis
  - Statistical parametric speech synthesis
  - Neural network based end-to-end speech synthesis
- **Disciplines:** Acoustics, linguistics, digital signal processing, statistics and deep learning.

Text To Speech Synthesis

- Neural based end-to-end TTS
  - **Text Analysis:** text → phoneme
    - Text normalization, grapheme-to-phoneme conversion
  - **Acoustic Model:** phoneme → mel-spectrogram
    - Tacotron 2, DeepVoice 3, TransformerTTS, FastSpeech 1/2
  - **Vocoder:** mel-spectrogram → waveform
    - WaveNet, WaveRNN, LPCNET, WaveGlow, MelGAN, PWG (Parallel WaveGAN)

The Neural Text-To-Speech Framework



TTS vs ASR

	ASR	TTS
Dataset	Can be multi-speaker	One speaker
Text	No need for phonemes level annotation.	Phonemes level is mandatory.
Mapping	One-to-one (Every audio have one writing possibility).	One-to-many (Every text can be spoken using different styles. E.g. duration, pitch, sound volume, speaker, style, emotion, etc)

FastSpeech 2: Fast and High-Quality End-to-End Text to Speech

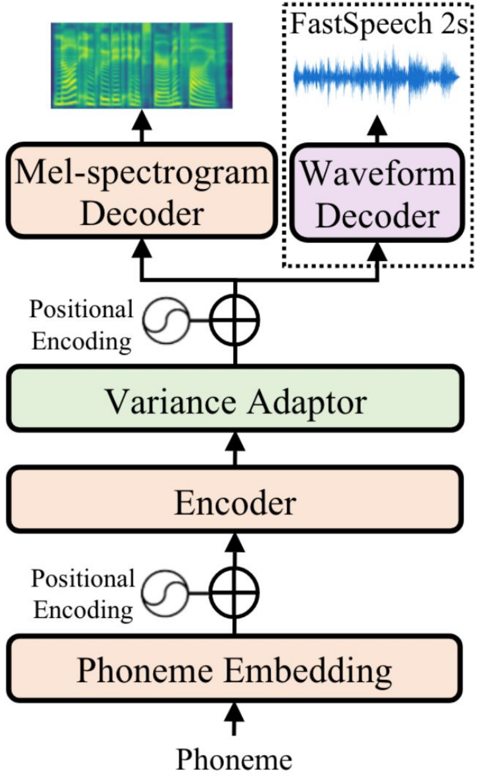
Yi Ren[1,*], **Chenxu Hu**[1,*], **Xu Tan**[2], **Tao Qin**[2], **Sheng Zhao**[3], **Zhou Zhao**[1,†], **Tie-Yan Liu**[2]

[1]Zhejiang University
{rayeren,chenxuhu,zhaozhou}@zju.edu.cn

[2]Microsoft Research Asia
{xuta,taoqin,tyliu}@microsoft.com

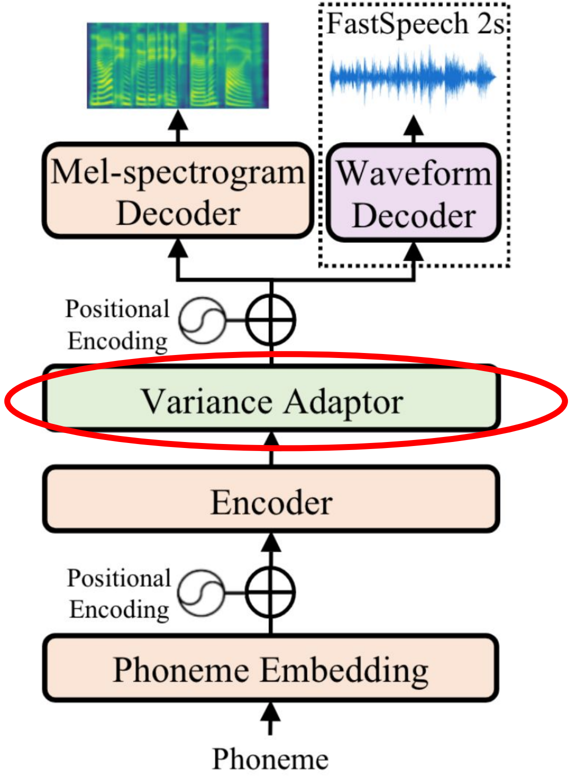
[3]Microsoft Azure Speech
Sheng.Zhao@microsoft.com

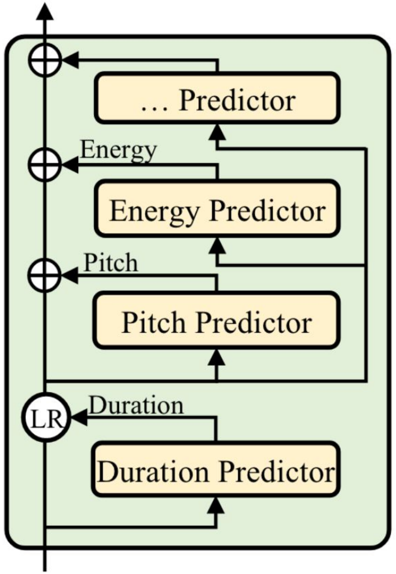
FastSpeech 2



- End-to-end model.
- Input:
  - Phonemes
- Encoder:
  - Transformer encoder
- Output:
  - Mel-spectogram
  - Waveform (FastSpeech 2s)
- Variance Adaptor:
  - Duration
  - Pitch
  - Energy
  - …

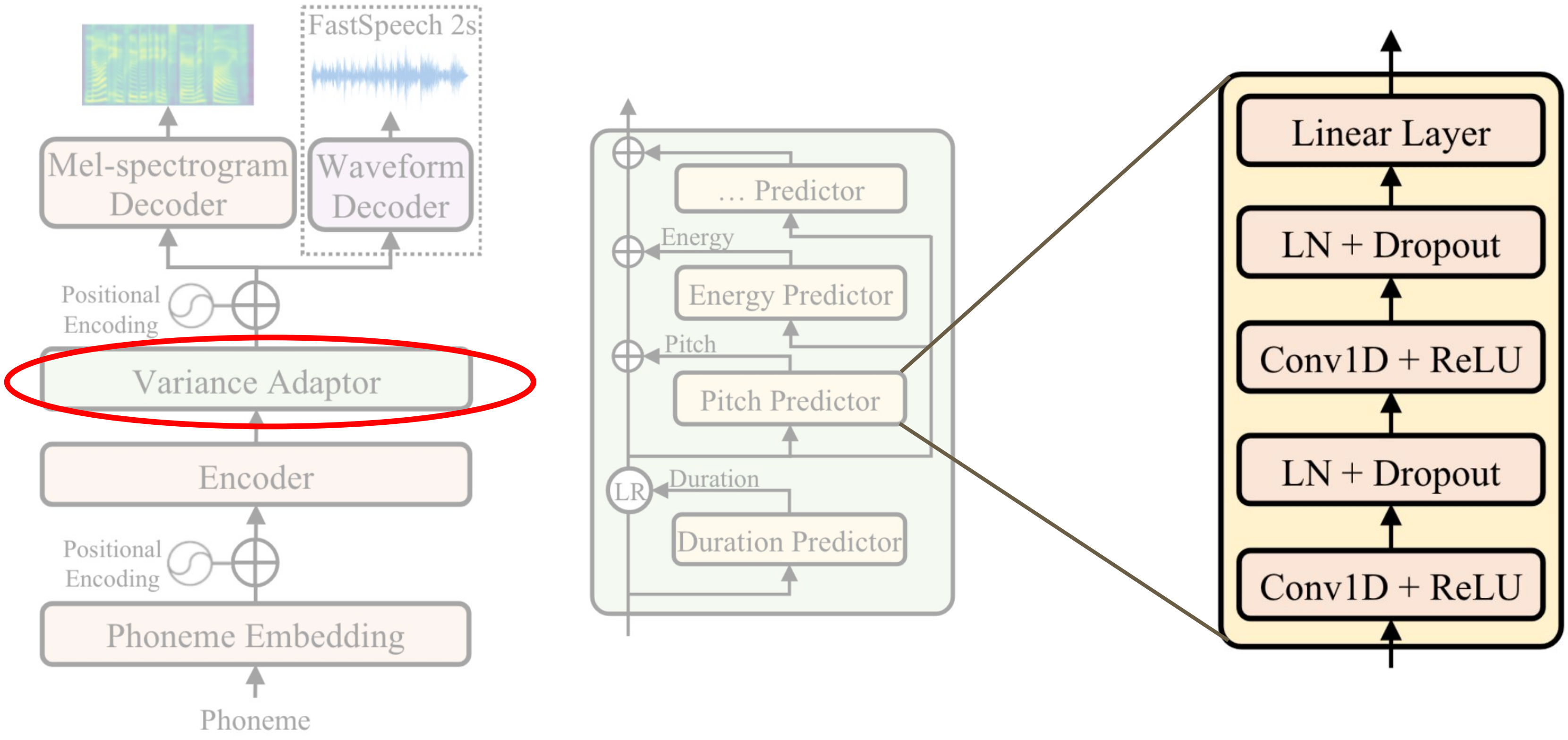
Variance Adaptor



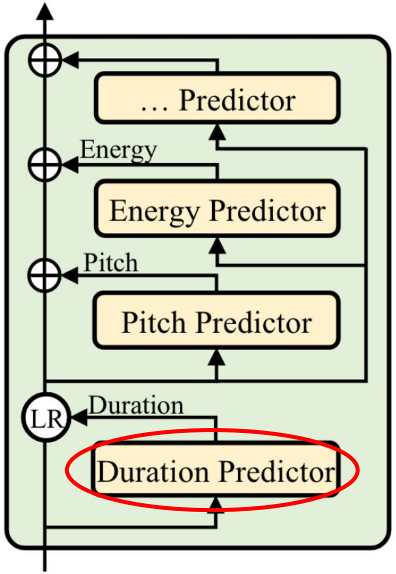


- **Phoneme duration:** how long the speech voice sounds.
- **Pitch:** a key feature to convey emotions and greatly affects the speech prosody.
- **Energy:** frame-level magnitude of mel-spectrograms and directly affects the volume and prosody of speech.

Variance Predictor

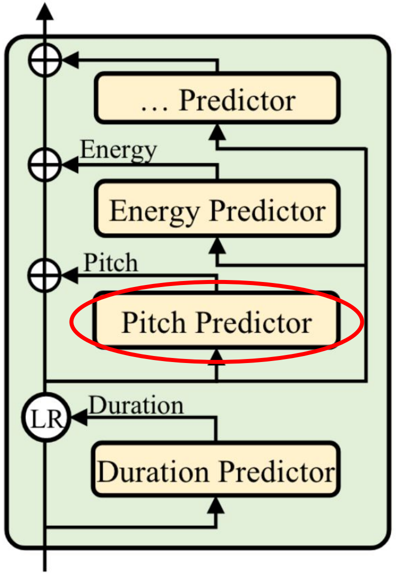


Duration Predictor



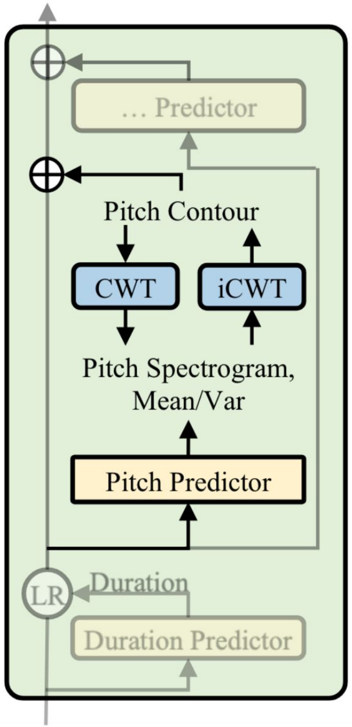
- **Input**
  - Phoneme hidden sequence
- **Output**
  - Duration of phoneme (How many mel frames correspond to this phoneme)
- **Optimization**
  - Mean square error (MSE) loss
- **Training data**
  - Durations are extracted using Montreal forced alignment (MFA).
    - Forced alignment is a technique to take an orthographic transcription of an audio file and generate a time-aligned version using a pronunciation dictionary to look up phones for words.

Pitch Predictor



- **Issue:**
  - High variations of ground-truth pitch
- **Input**
  - Phoneme hidden sequence
- **Output**
  - Pitch spectrogram
- **Optimization**
  - Mean square error (MSE) loss

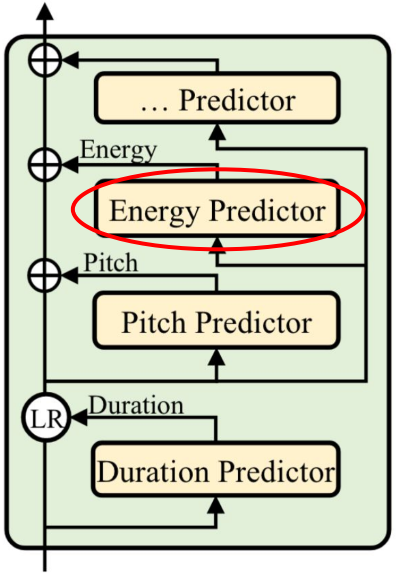
CWT Pitch Prediction



- **Motivation:** To better predict the variations in pitch contour
- **Idea:** Use continuous wavelet transform (CWT) to decompose the continuous pitch contour to pitch spectrogram.

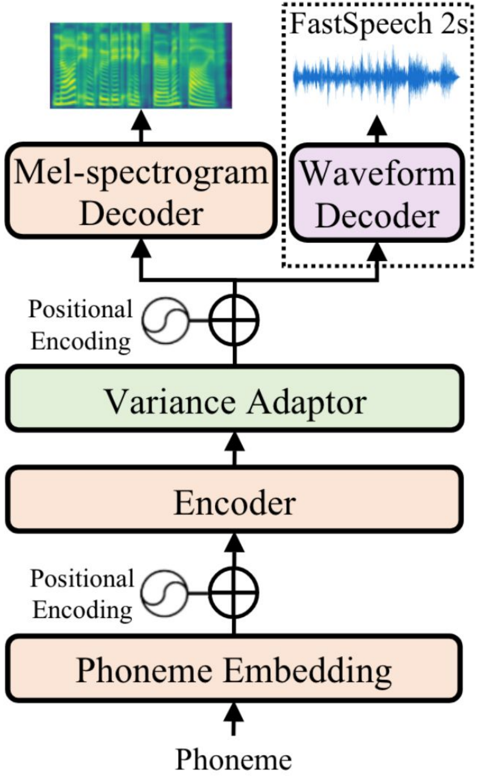
- During training
  - **Text input** –fit-> **pitch spec** <-CWT– **pitch contour**
- During inference
  - **Text input** –predict -> **pitch spec** –iCWT-> **pitch contour**.

Energy Predictor



- **Input**
  - Spectrogram frame
- **Output**
  - L2-norm of the amplitude of the frame
  - Phoneme-level average
- **Optimization**
  - Mean square error (MSE) loss

FastSpeech 2



- End-to-end model.
- Input:
  - Phonemes
- Output:
  - Mel-spectrogram
  - Waveform (FastSpeech 2s)
- Variance Adaptor:
  - Duration
  - Pitch
  - Energy
  - …

Experimental Setup

- Dataset: LJSpeech
- Language: English
- Dataset size: 24 hours - 13,100 audio clips.
  - Train: 12,228 samples
  - Validation: 349 samples
  - Test: 523 samples
- Grapheme-to-phoneme: https://github.com/Kyubyong/g2p
- Raw waveform to mel-spectrograms:
  - Frame size: 1024
  - Hop size: 256
  - Sample rate: 22050

Results

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

Thanks!