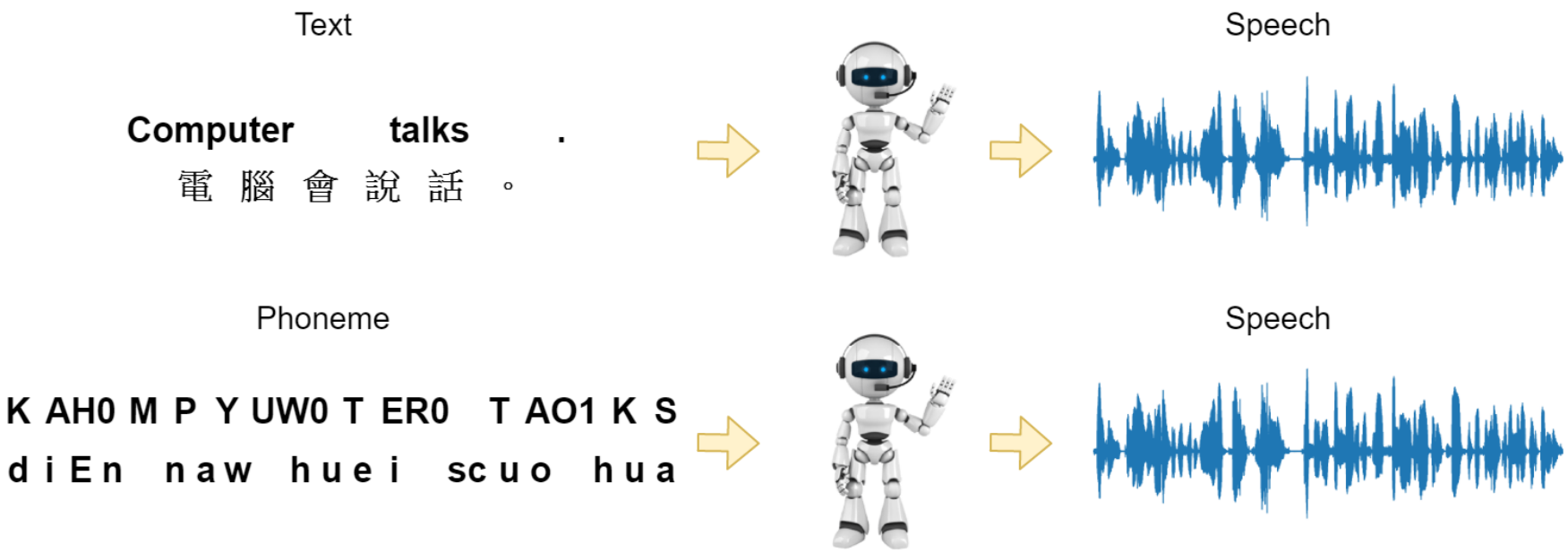
An Introduction to Text-to-Speech

Chung-Ming Chien 簡仲明
<r08922080@ntu.edu.tw>

What is Text-to-Speech?



- More than simply recording audio samples for each phoneme and combine them!

Outline

- History
- End-to-end TTS: Tacotron
- Non-autoregressive TTS: FastSpeech
- Hands-on Lab

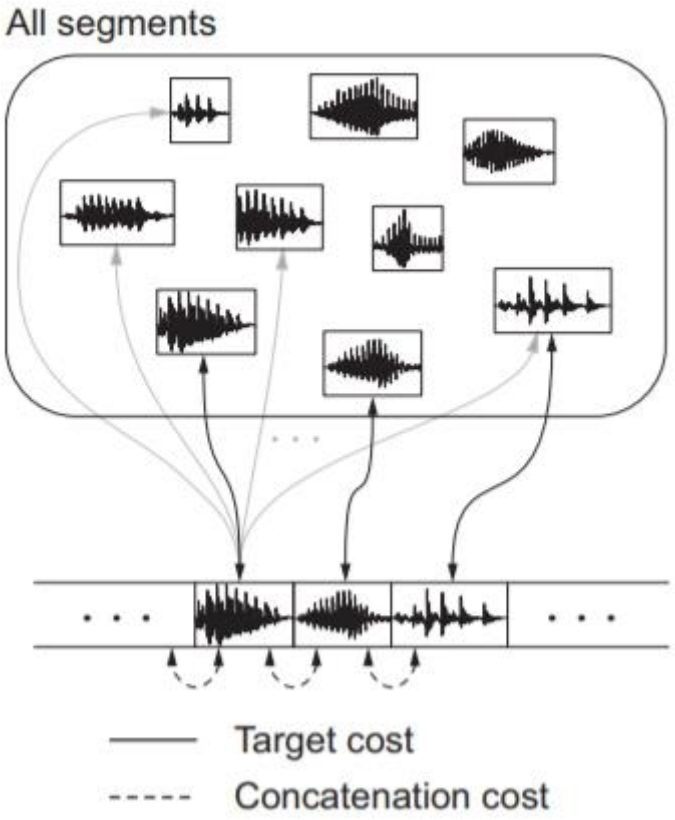
Outline

- History
- End-to-end TTS: Tacotron
- Non-autoregressive TTS: FastSpeech
- Hands-on Lab

First Trial: Voder (1939)

https://en.wikipedia.org/wiki/Voder
https://www.youtube.com/watch?v=0rAyrmm7vv0

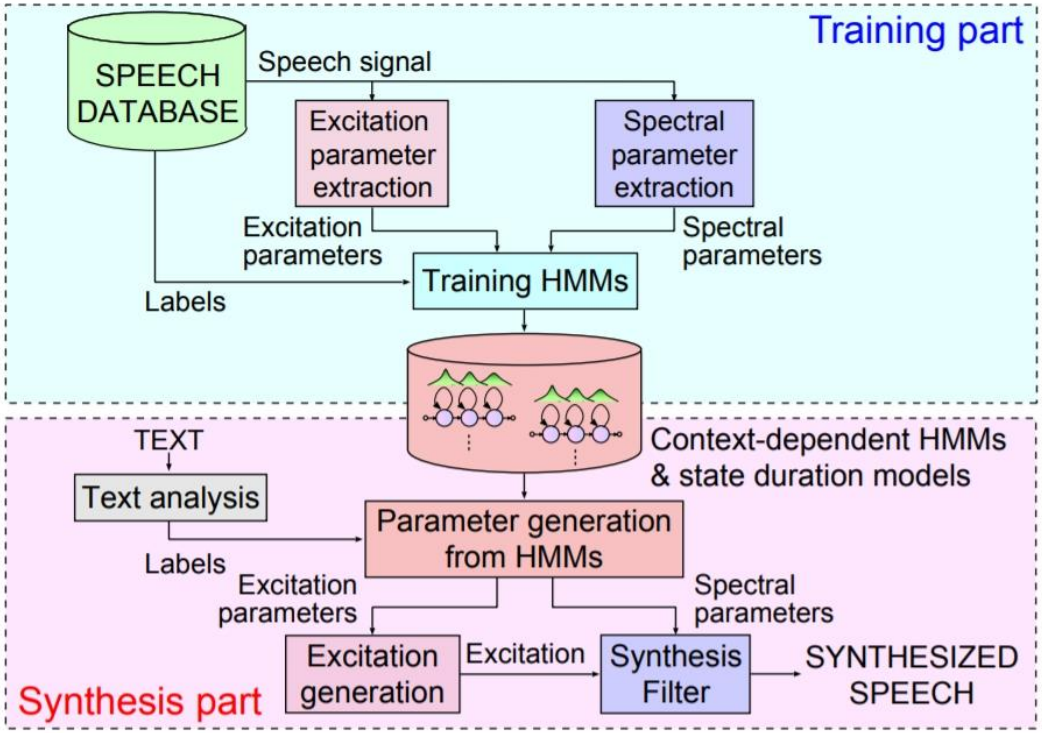
Concatenative Approach



- How to select clips?
- How to concatenate them smoothly?

https://www.cs.cmu.edu/~pmuthuku/mlsp_page/lectures/spss_specom.pdf

Parametric Approach



- HMM generates acoustic features (f0, energy…)
- Vocoder generates waveform based on acoustic features

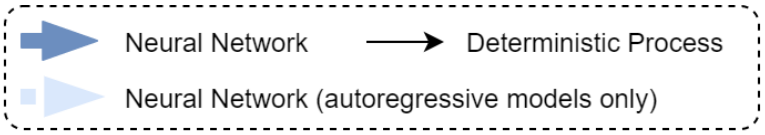
https://www.sp.nitech.ac.jp/~tokuda/tokuda_interspeech09_tutorial.pdf

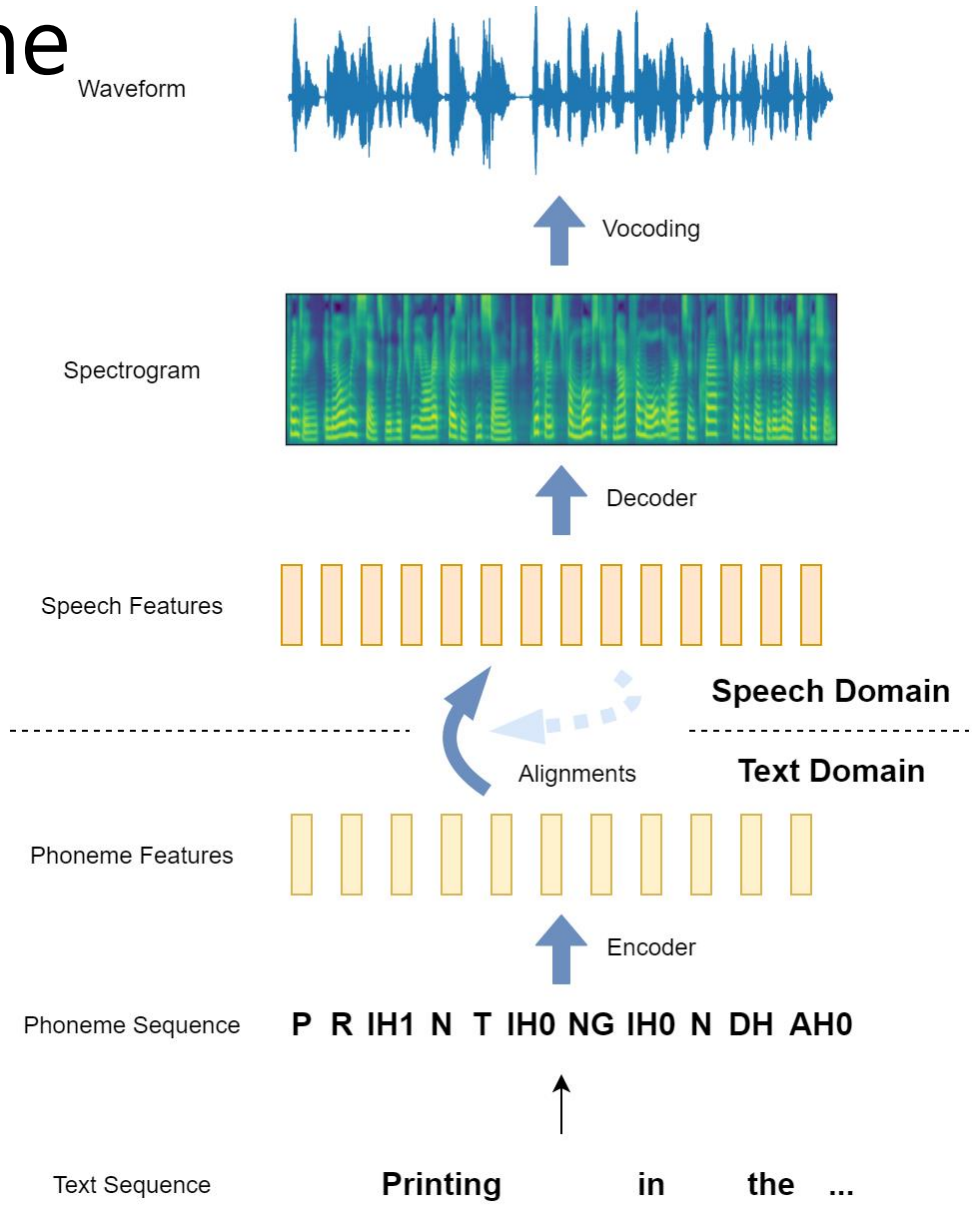
Outline

- History
- End-to-end TTS: Tacotron
- Non-autoregressive TTS: FastSpeech
- Hands-on Lab

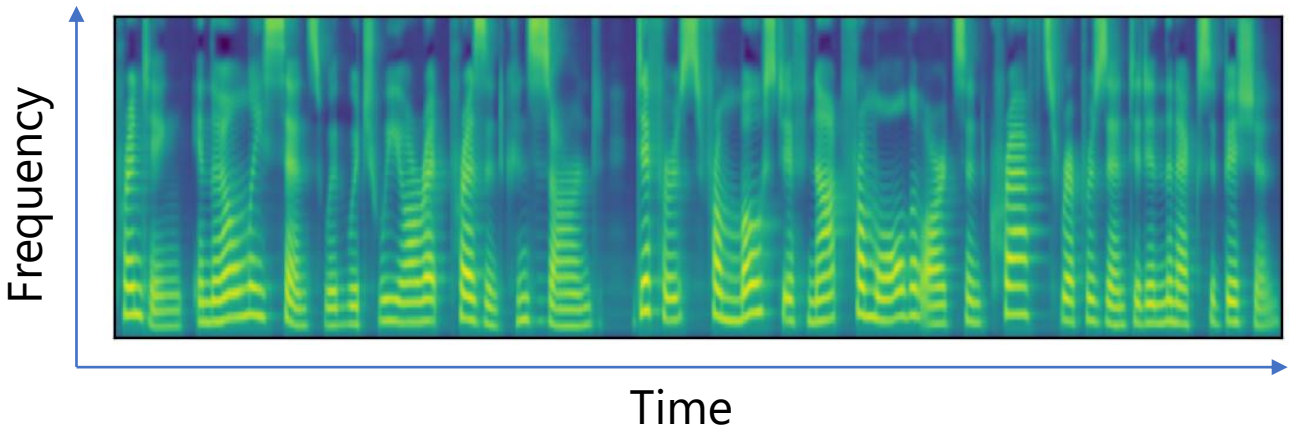
General TTS Pipeline

- Input: Text
- Output: Waveform





Spectrogram & Vocoder



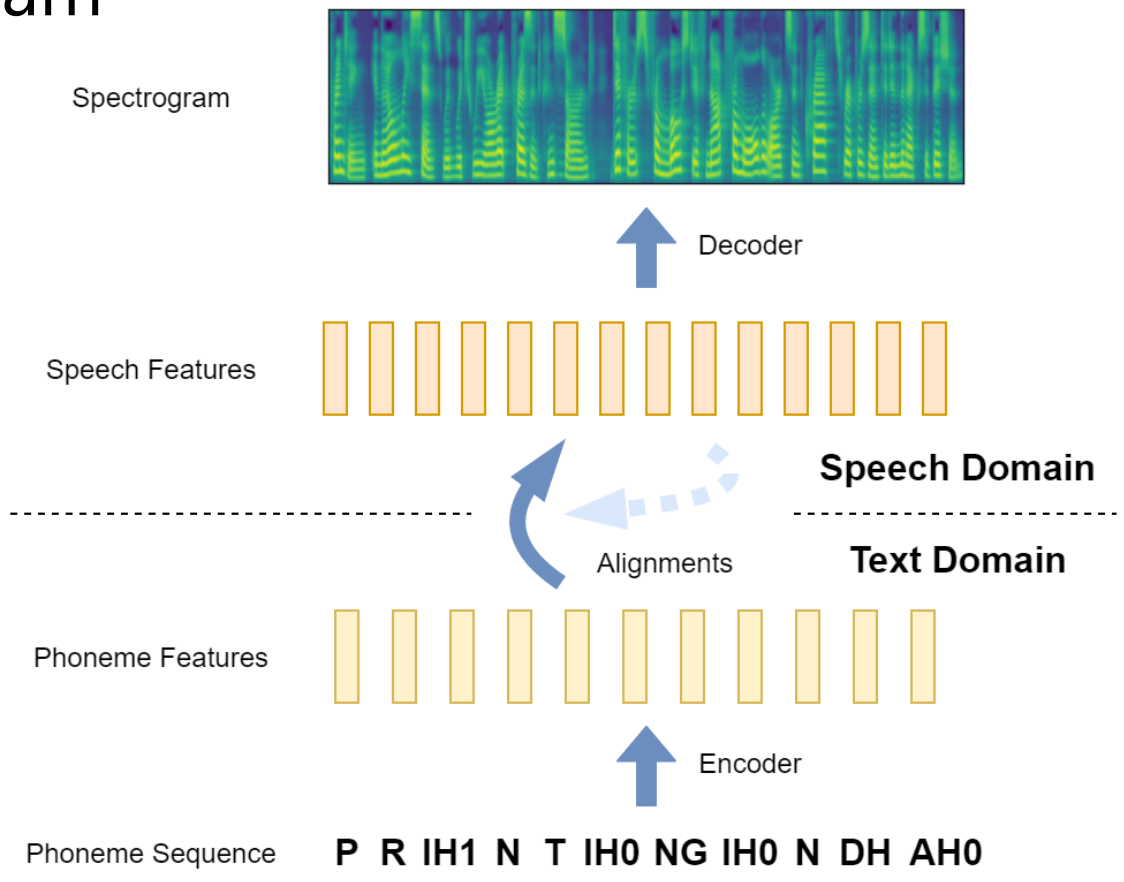
- Spectrogram: the spectrum of frequencies of a signal as it varies with time

- Vocoder: convert spectrogram back to time-domain waveform

General TTS Pipeline

- Input: phoneme
- Output: spectrogram

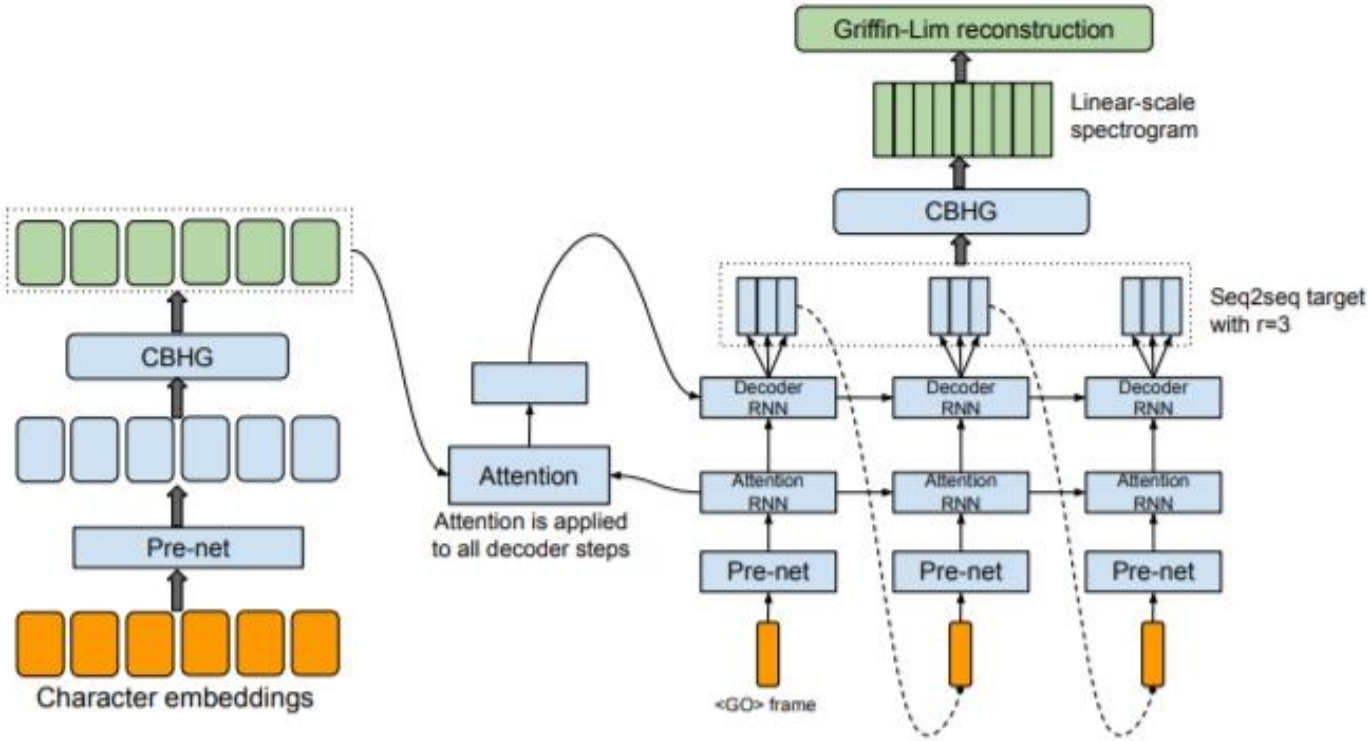
- Components
  - encoder
  - alignment
  - decoder



Tacotron & Tacotron 2

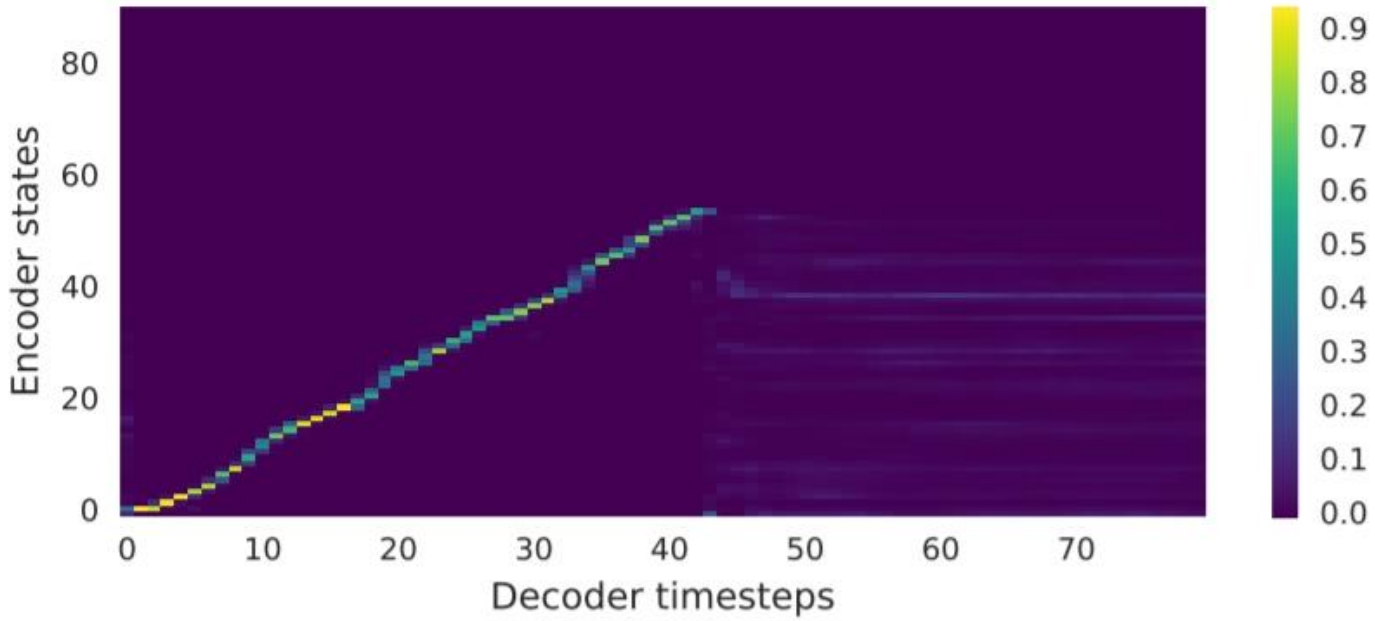
[Wang, et al., INTERSPEECH'17]
[Shen, et al., ICASSP'18]

- Sequence-to-sequence model with attention



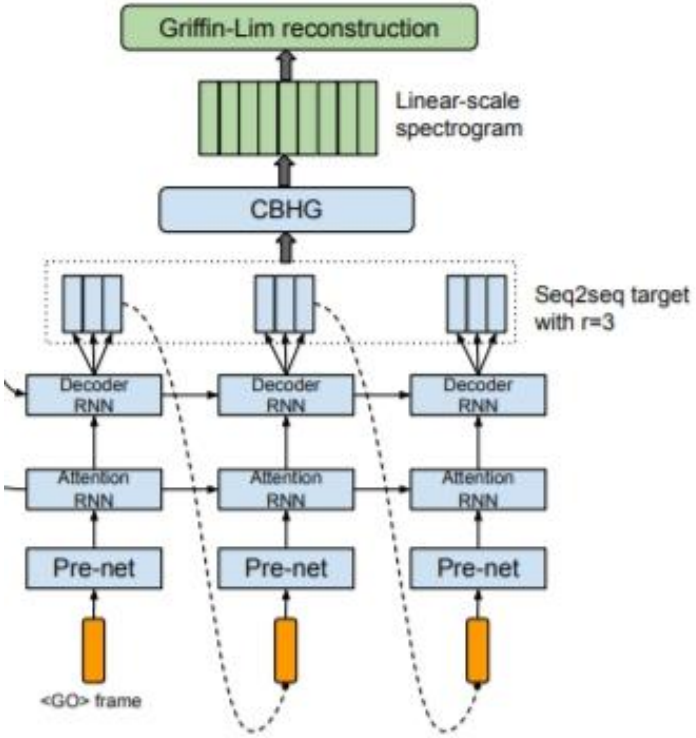
Attention

- Which encoder state the decoder is focusing on at every decoder step.
- The output audio and input text should be monotonically aligned



Inefficiency of Autoregressive Models

- Training time: unparallelizable
- Inference time: causal inference



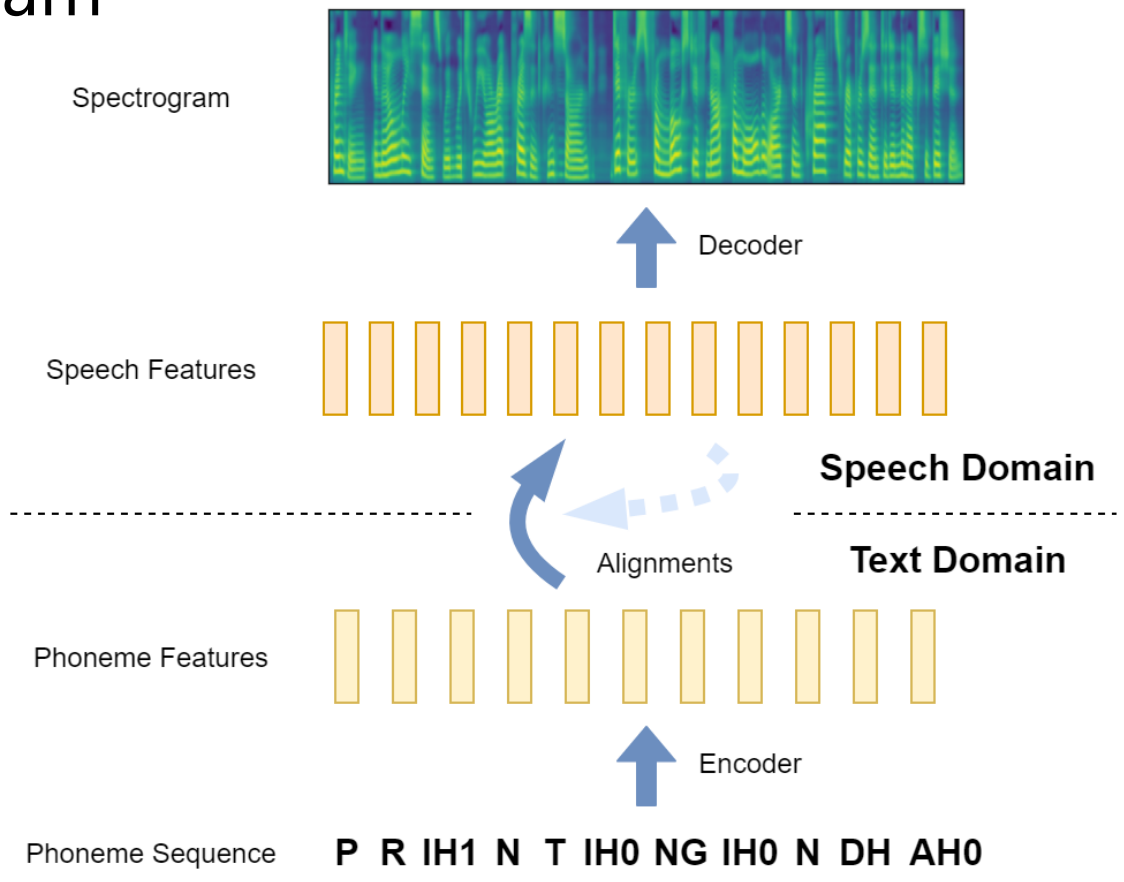
Outline

- History
- End-to-end TTS: Tacotron
- Non-autoregressive TTS: FastSpeech
- Hands-on Lab

General TTS Pipeline

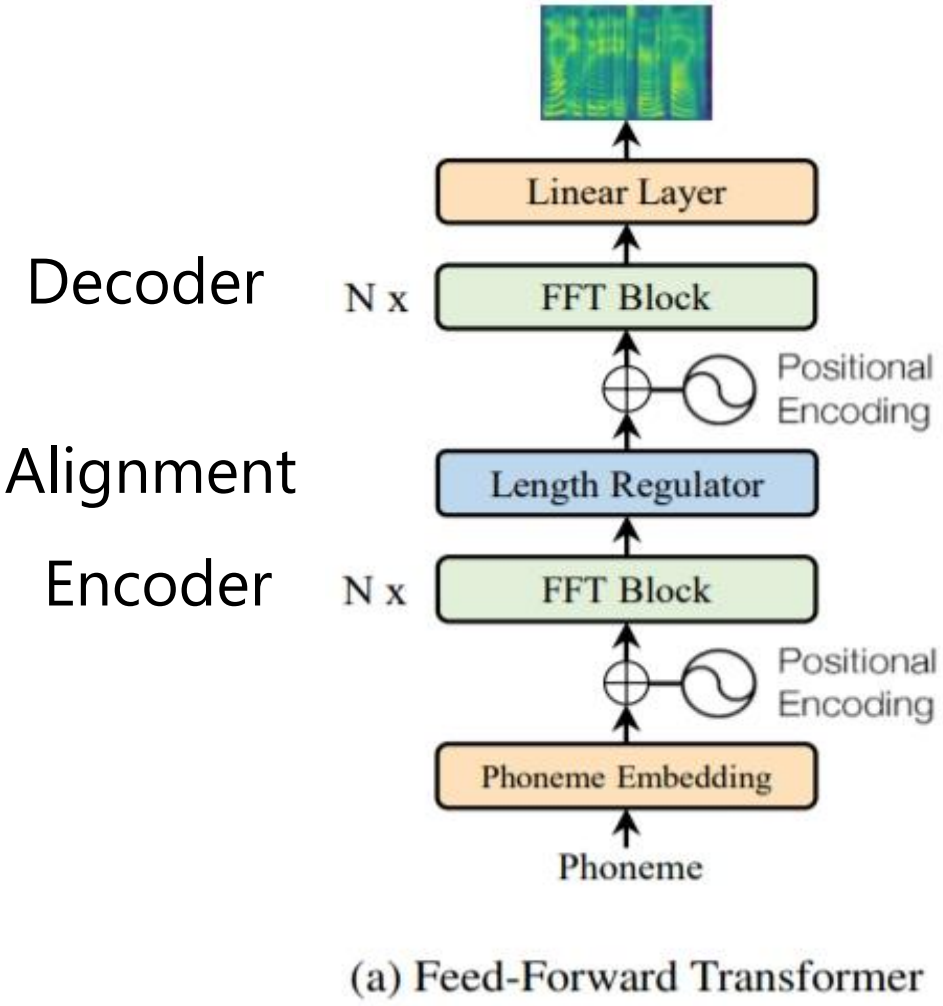
- Input: phoneme
- Output: spectrogram

- Components
  - encoder
  - alignment
  - decoder



FastSpeech & FastSpeech 2

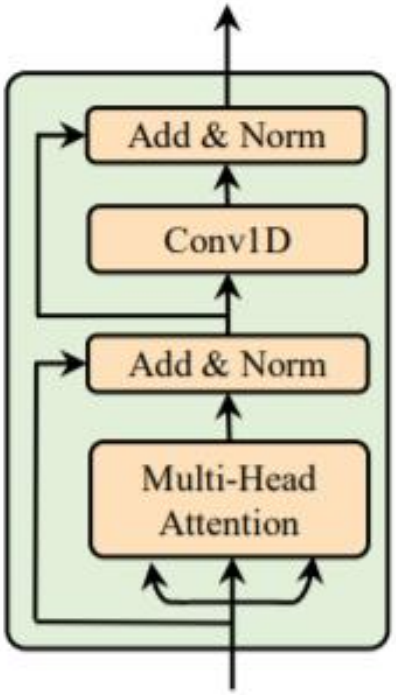
[Ren, et al., NeurIPS'19]
[Ren, et al., arXiv'20]



(a) Feed-Forward Transformer

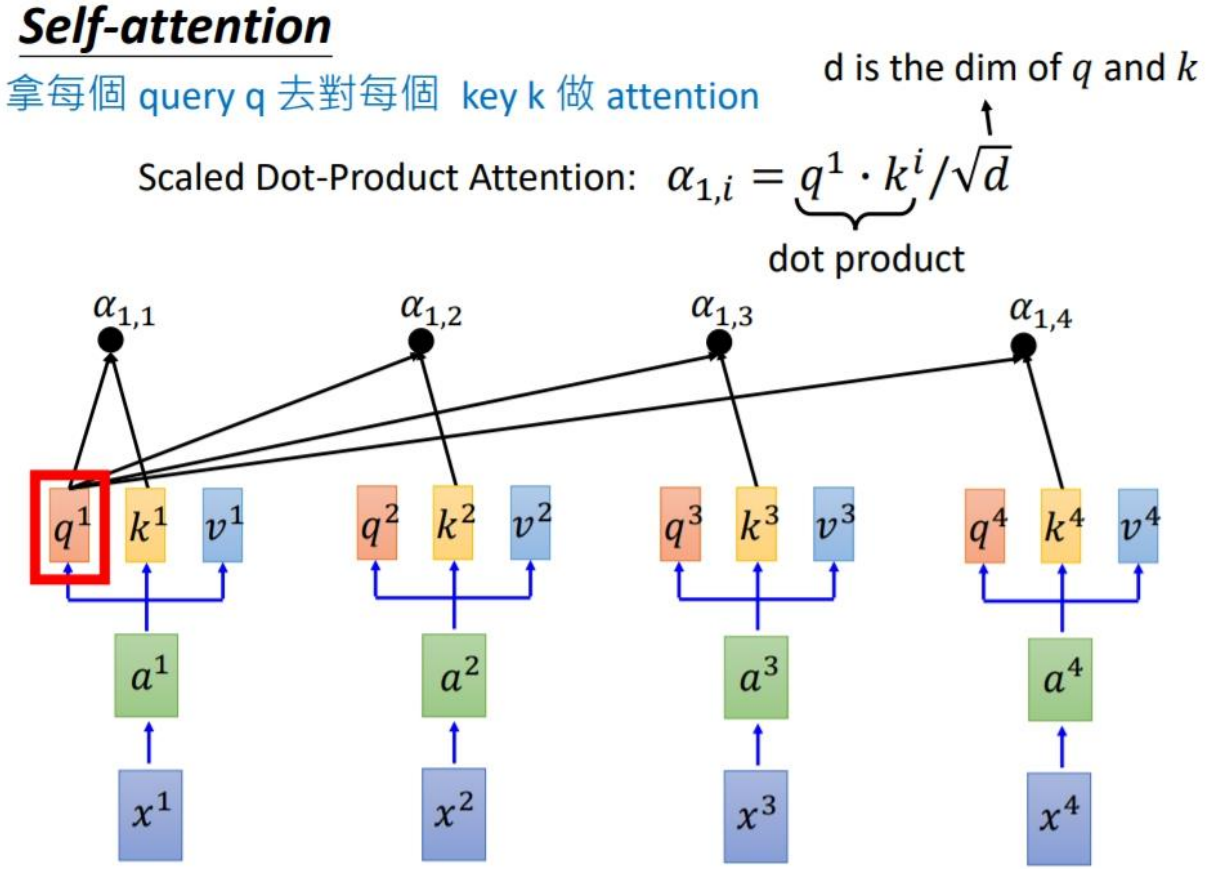
Encoder & Decoder

- Transformer block
  - self-attention
  - multi-head attention
  - non-autoregressive



Encoder & Decoder

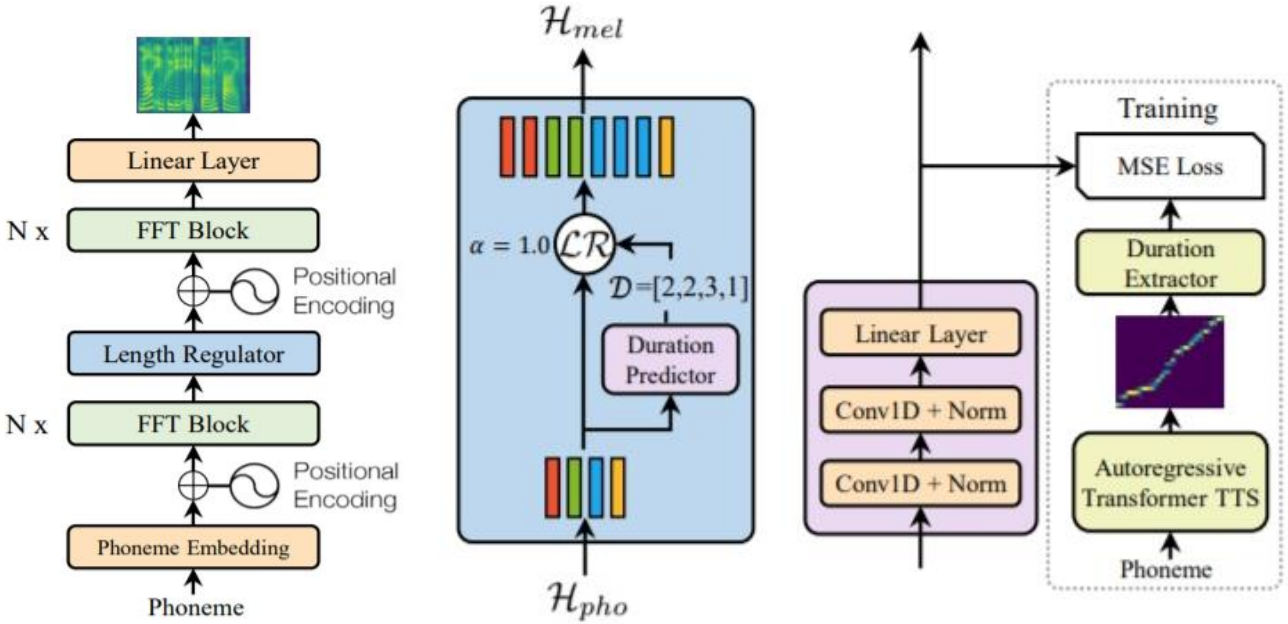
- Transformer block



https://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2019/Lecture/Transformer%20(v5).pdf

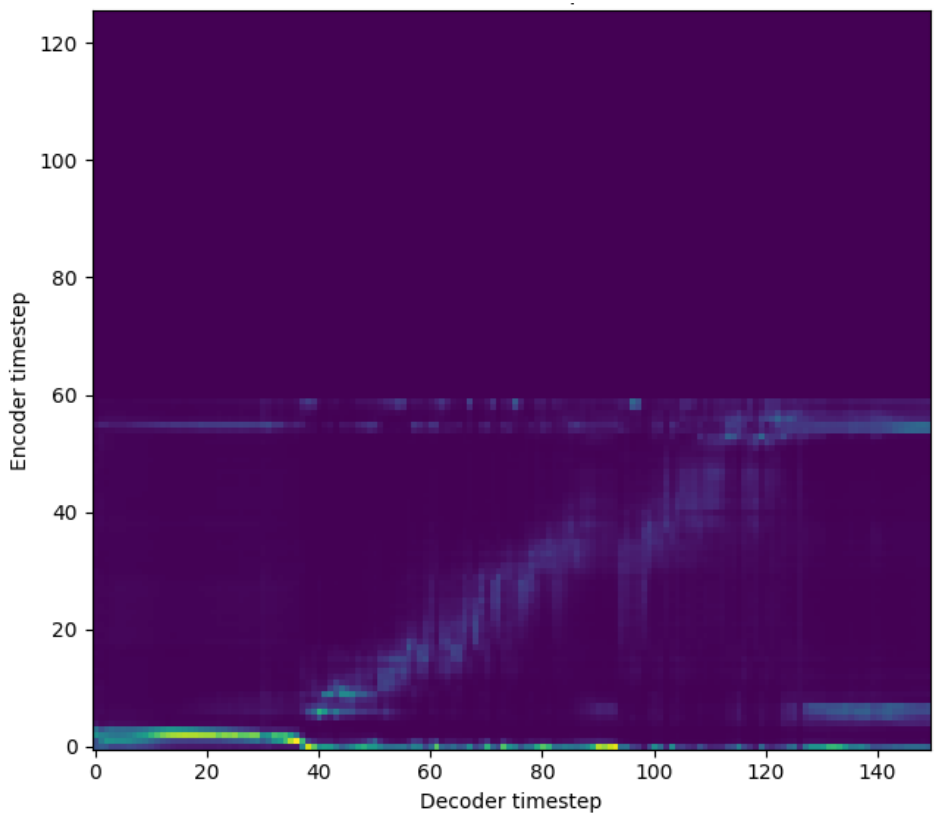
Length Regulator

- Predict duration of each phoneme
- Need ground-truth duration (from another TTS/ASR model or off-the-shelf packages)



FastSpeech & FastSpeech 2

- Fast training/inference
- Avoid the error caused by false alignment



Outline

- History
- End-to-end TTS: Tacotron
- Non-autoregressive TTS: FastSpeech
- Hands-on Lab

Hands-on Lab

- Github repository: https://github.com/ming024/FastSpeech2
- Colab: https://colab.research.google.com/drive/1q553nFsrYYSDX-xE1rZn0WGHp780Me8j?usp=sharing

Reference

- [Wang, et al., INTERSPEECH'17] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark and Rif A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis", INTERSPEECH, 2017
- [Shen, et al., ICASSP'18] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis and Yonghui Wu "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions", ICASSP, 2018
- [Ren, et al., NeurIPS'19] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao and Tie-Yan Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech", NeurIPS, 2019

Reference

- [Ren, et al., arXiv'20] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao and Tie-Yan Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech", arXiv, 2020
- Heiga Zen, Keiichi Tokuda and Alan W. Black, "Statistical parametric speech synthesis", Speech Communication, 2009
- Keiichi Tokuda and Heiga Zen, "Fundamentals and recent advances in HMM-based speech synthesis", INTERSPEECH tutorial, 2009
- Prof. Hung-yi Lee's Lecture Slides: Speech Synthesis, Transformer
- Unofficial PyTorch Tacotron 2 implementation by NVIDIA
- Unofficial PyTorch FastSpeech 2 implementation by ming024