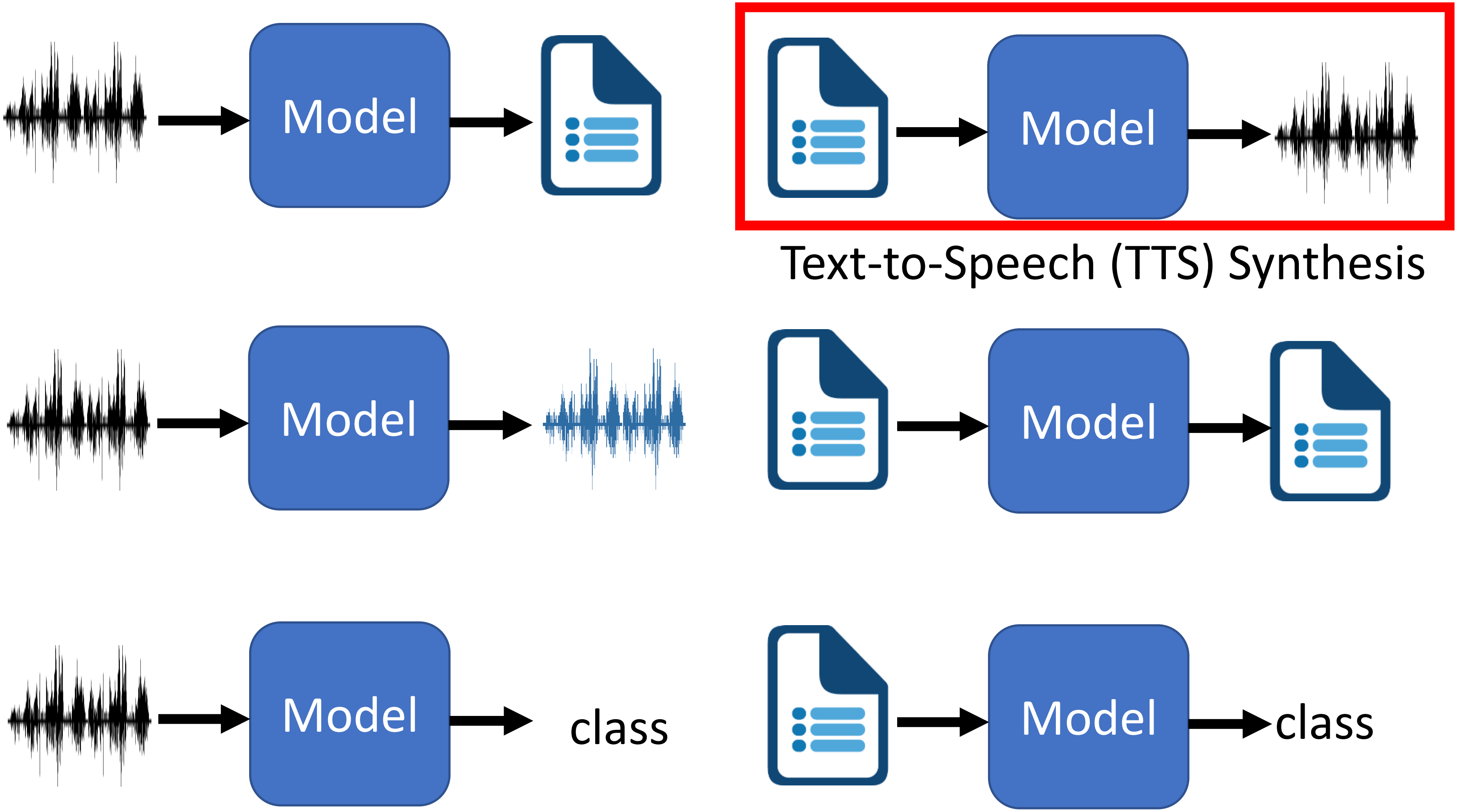
Speech Synthesis

李宏毅

Hung-yi Lee

One slide for this course



Model

Text-to-Speech (TTS) Synthesis

Model

Model

Model

Model

class

Model

class

Outline

- TTS before End-to-end
- Tacotron: End-to-end TTS
- Beyond Tacotron
- Controllable TTS

VODER (1939)

https://en.wikipedia.org/wiki/Voder

Source of video: https://www.youtube.com/watch?v=0rAyrmm7vv0

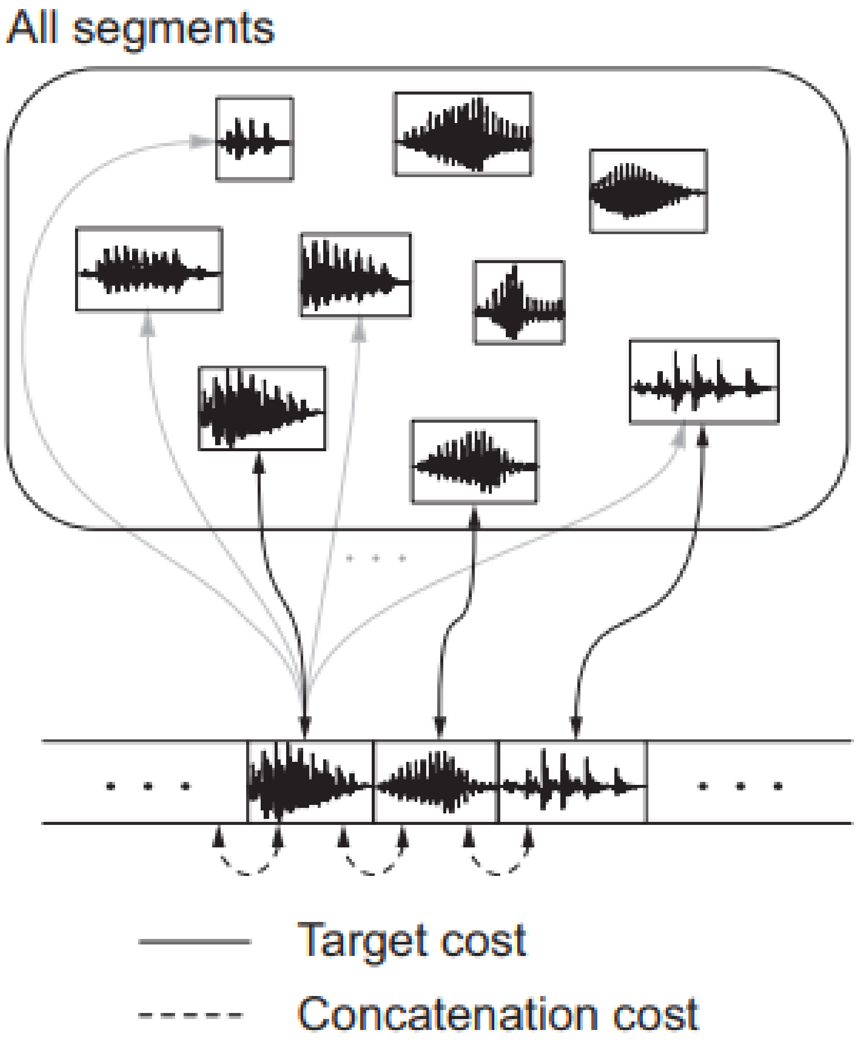
IBM computer (1960s)

- In 1961, John Larry Kelly Jr. using an IBM computer to synthesize speech at Bell lab.

Source of video and audio: https://youtu.be/UGsfwhb4-bQ
https://www.vintagecomputermusic.com/mp3/s2t9_Computer_Speech_Demonstration.mp3

Concatenative Approach

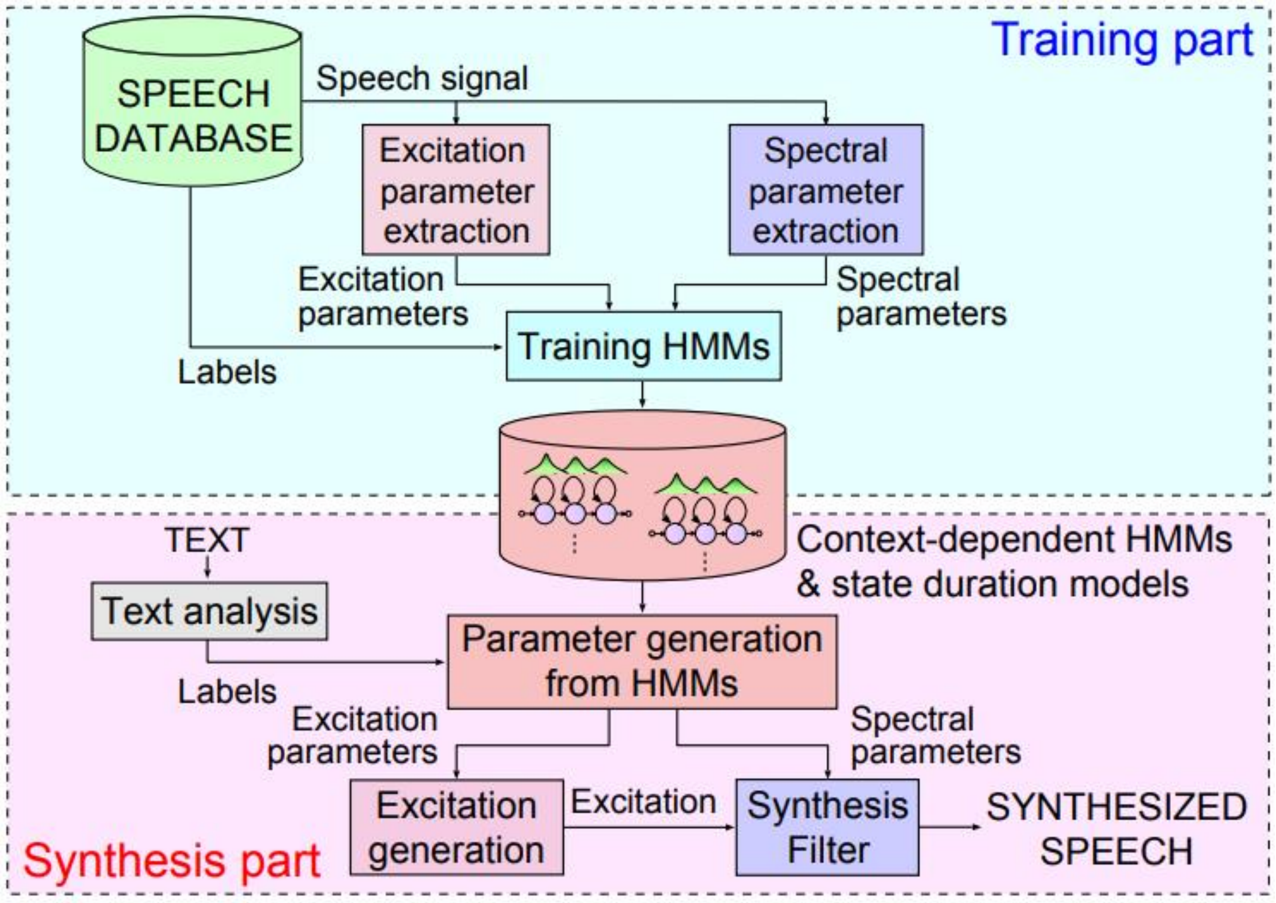
speeches from a large database



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

Parametric Approach

HMM/DNN-based Speech Synthesis System (HTS)



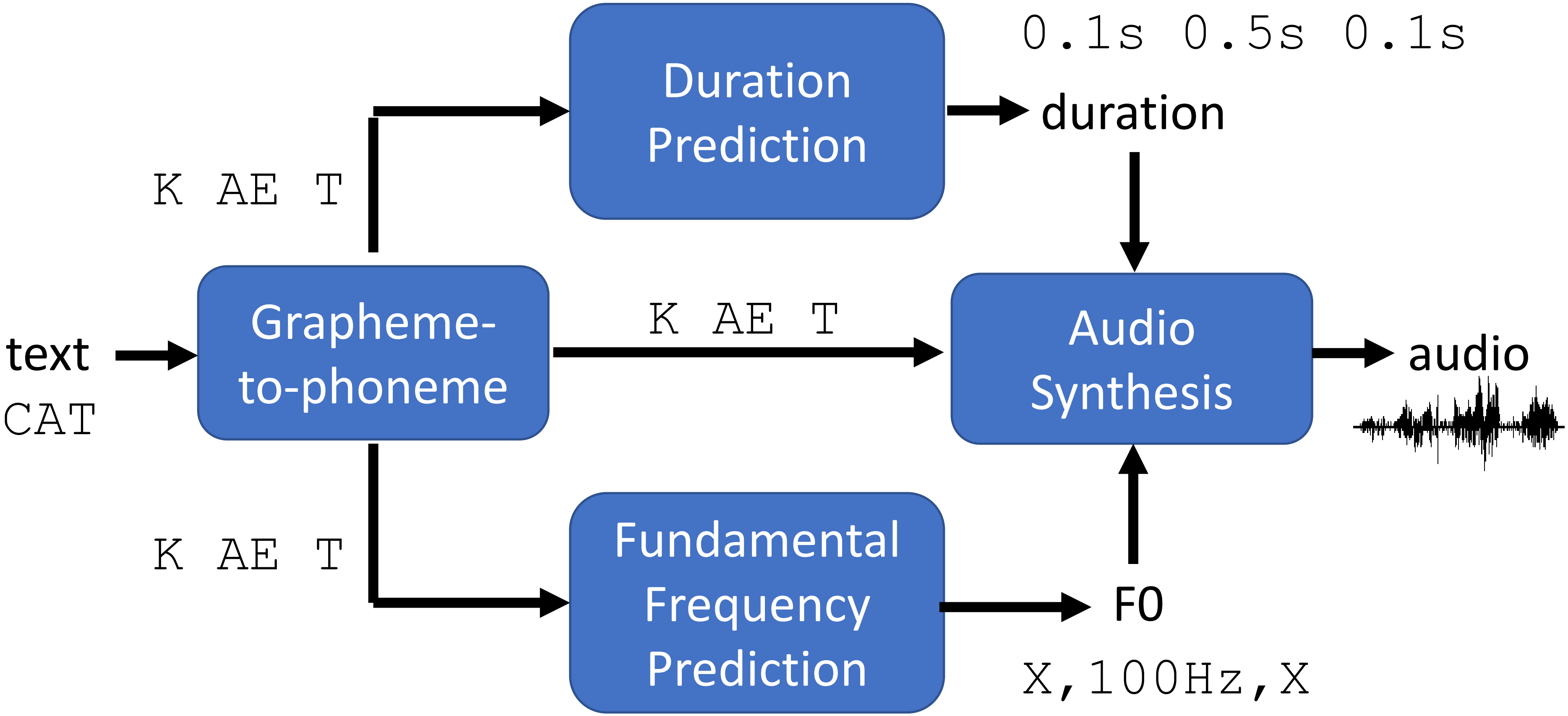
Source of image: http://hts.sp.nitech.ac.jp/?Tutorial

Deep Voice

[Arik, et al., ICML'17]

Deep Voice 3 is end-to-end.

[Ping, et al., ICLR'18]



text

CAT

Grapheme-to-phoneme

K AE T

K AE T

K AE T

Duration Prediction

duration

0.1s 0.5s 0.1s

Fundamental Frequency Prediction

F0

X,100Hz,X

Audio Synthesis

audio

All the components are deep learning based.

Outline

TTS before End-to-end

Tacotron: End-to-end TTS

Beyond Tacotron

Controllable TTS

Tacotron

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

TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

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, Rif A. Saurous[*]

Google, Inc.
{yxwang,rjryan,rif}@google.com

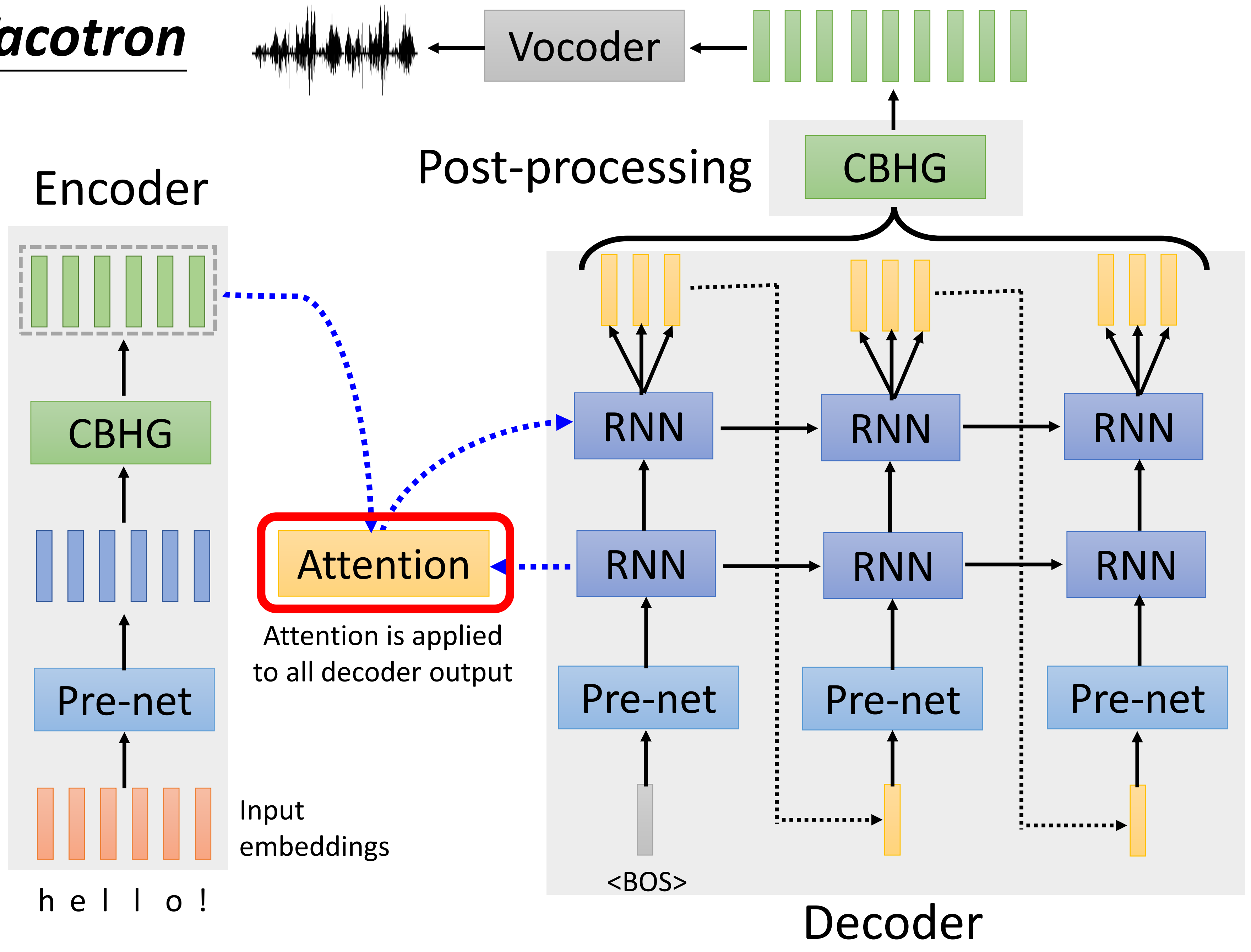
[*] These authors really like tacos.

[†] These authors would prefer sushi.

Before Tacotron …

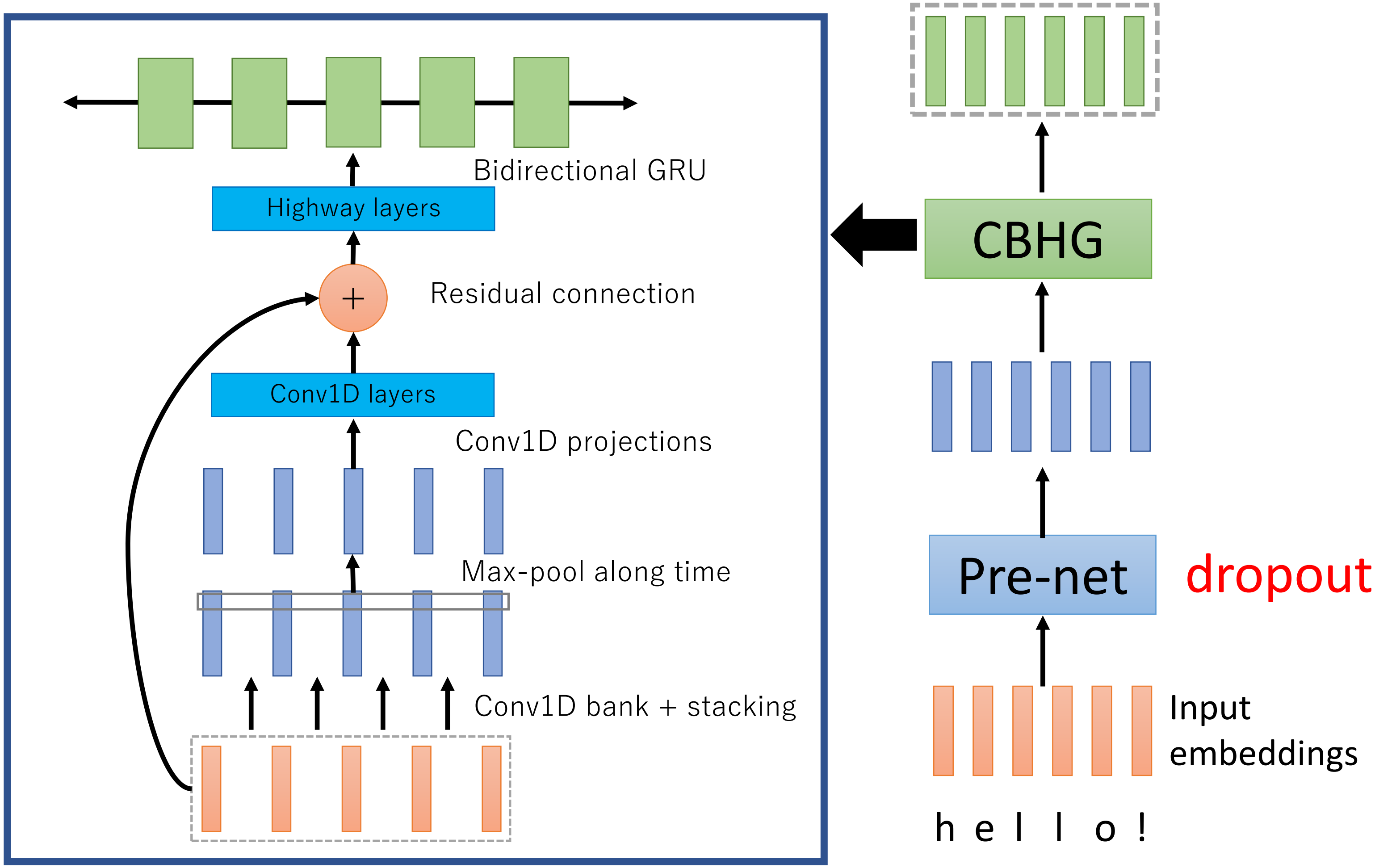
- Tacotron:
  - Input: character
  - Output: (linear) spectrogram
- First Step Towards End-to-end Parametric TTS Synthesis [Wang, et al., INTERSPEECH'16]
  - Input: phoneme
  - Output: acoustic features for STRAIGHT (vocoder)
- Char2wav [Sotelo, et al., ICLR workshop'17]
  - Input: character
  - Output: acoustic features for SampleRNN (vocoder)

Tacotron



Attention is applied to all decoder output

Encoder = Grapheme-to-phoneme?



Bidirectional GRU

Highway layers

Residual connection

Conv1D layers

Conv1D projections

Max-pool along time

Conv1D bank + stacking

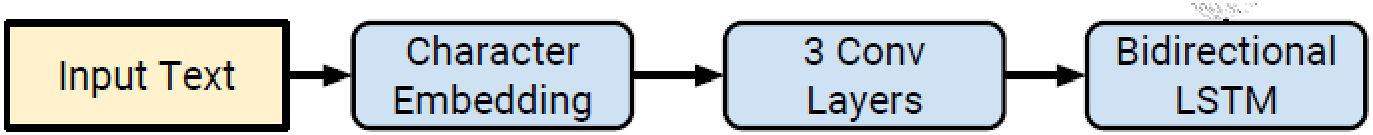
CBHG

Pre-net

dropout

Input embeddings

h e l l o !



Input Text

Character Embedding

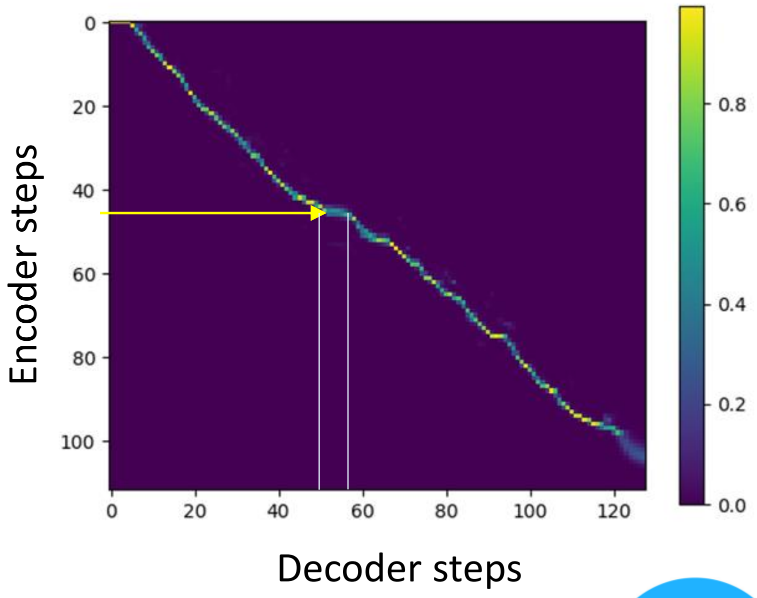
3 Conv Layers

Bidirectional LSTM

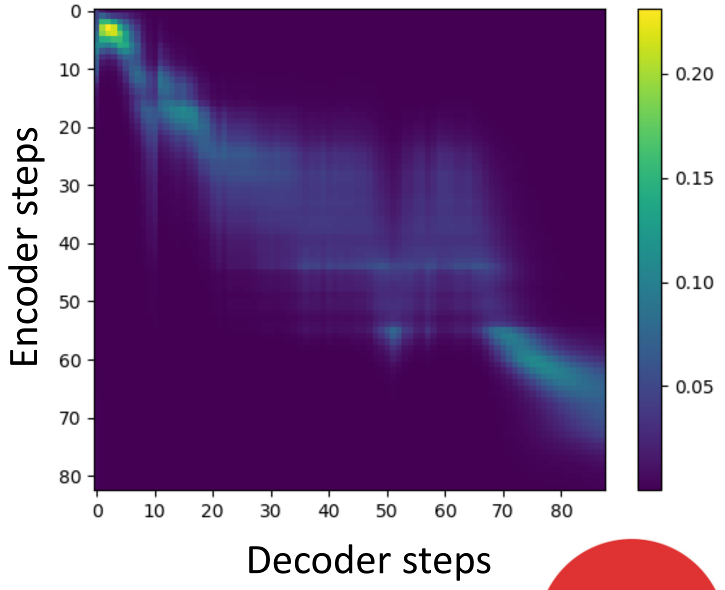
(v2)

Attention = Modeling Duration ?

- The output audio and input text much be monotonic aligned.



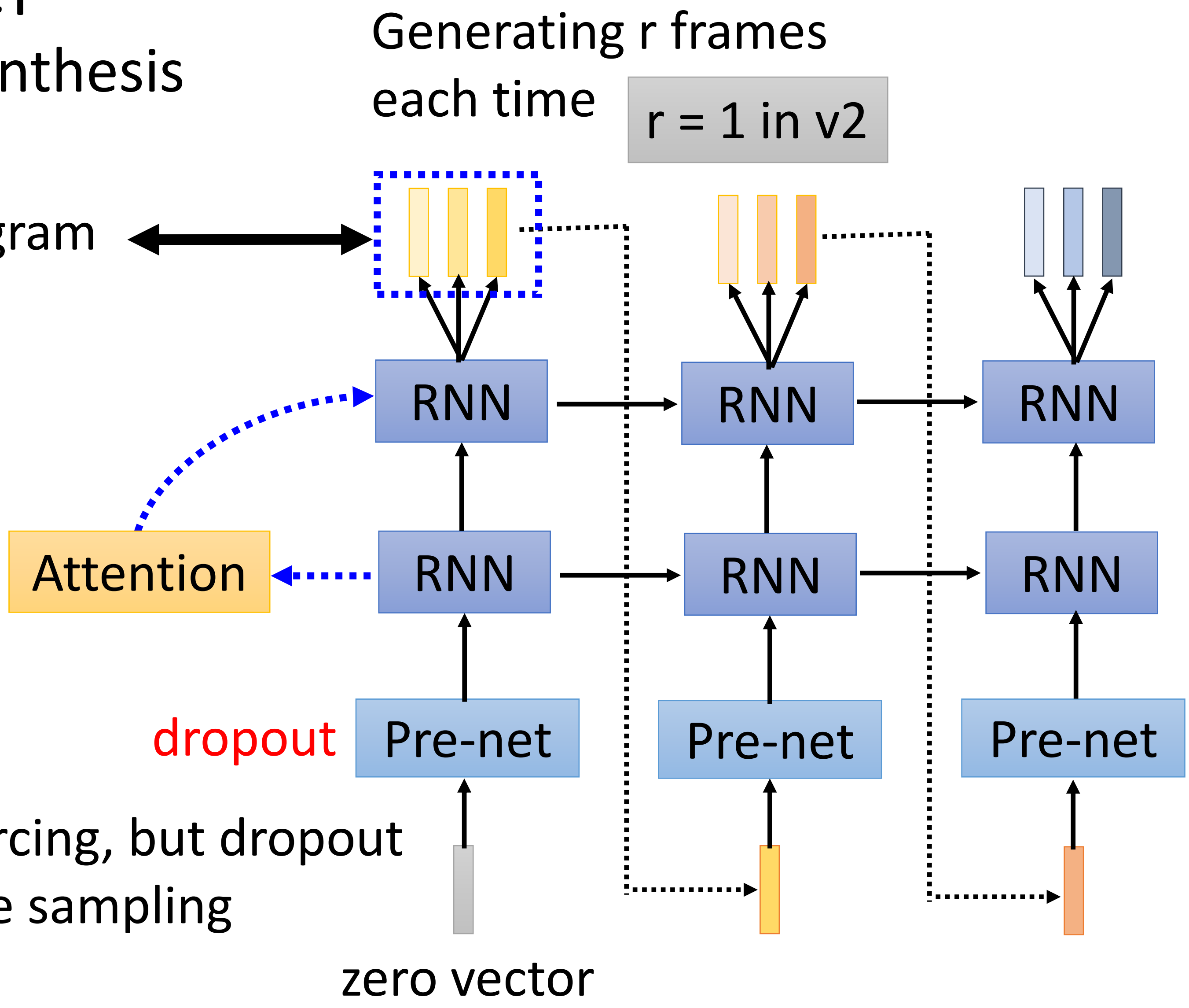
👍



👎

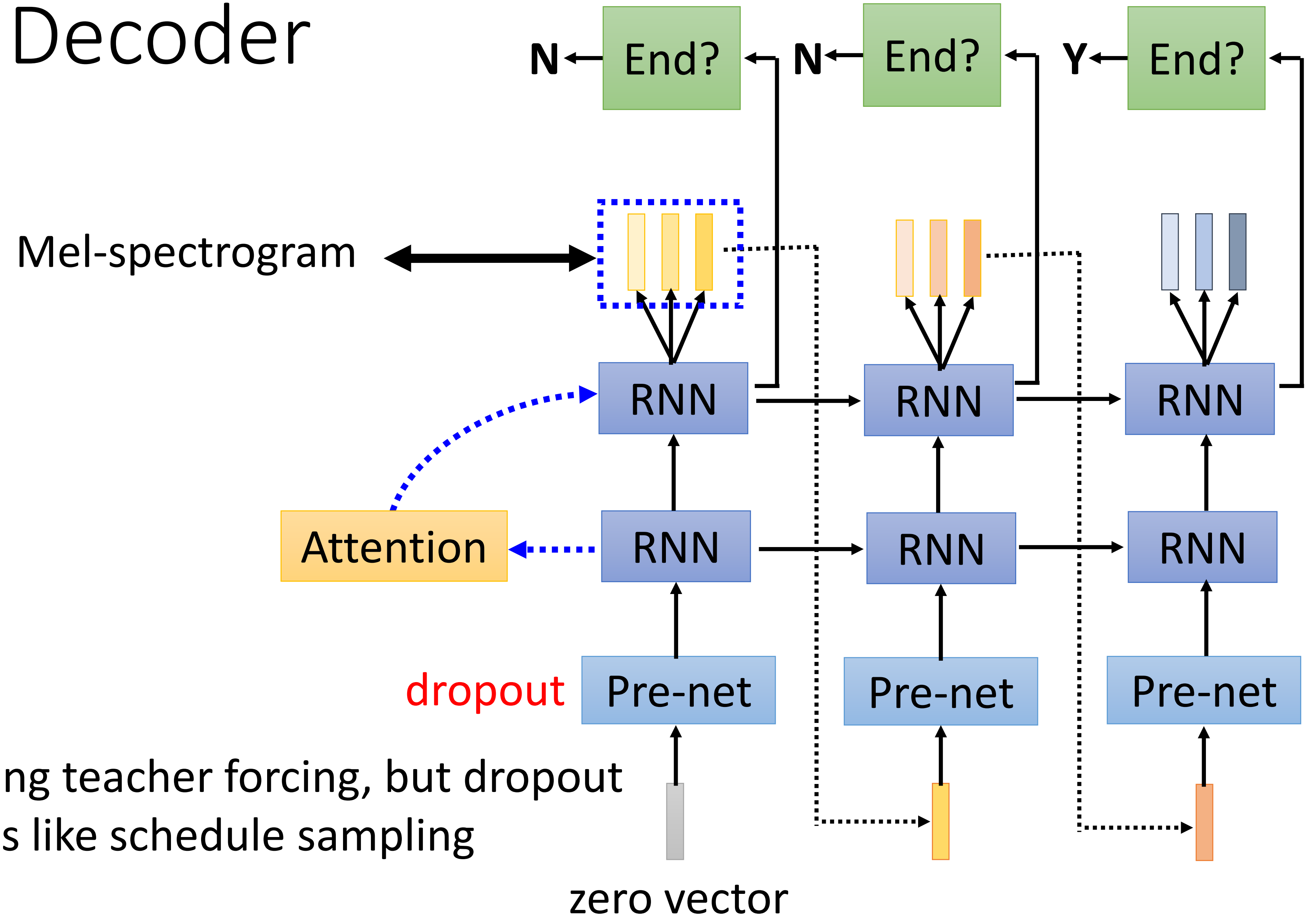
Decoder

→ Audio Synthesis



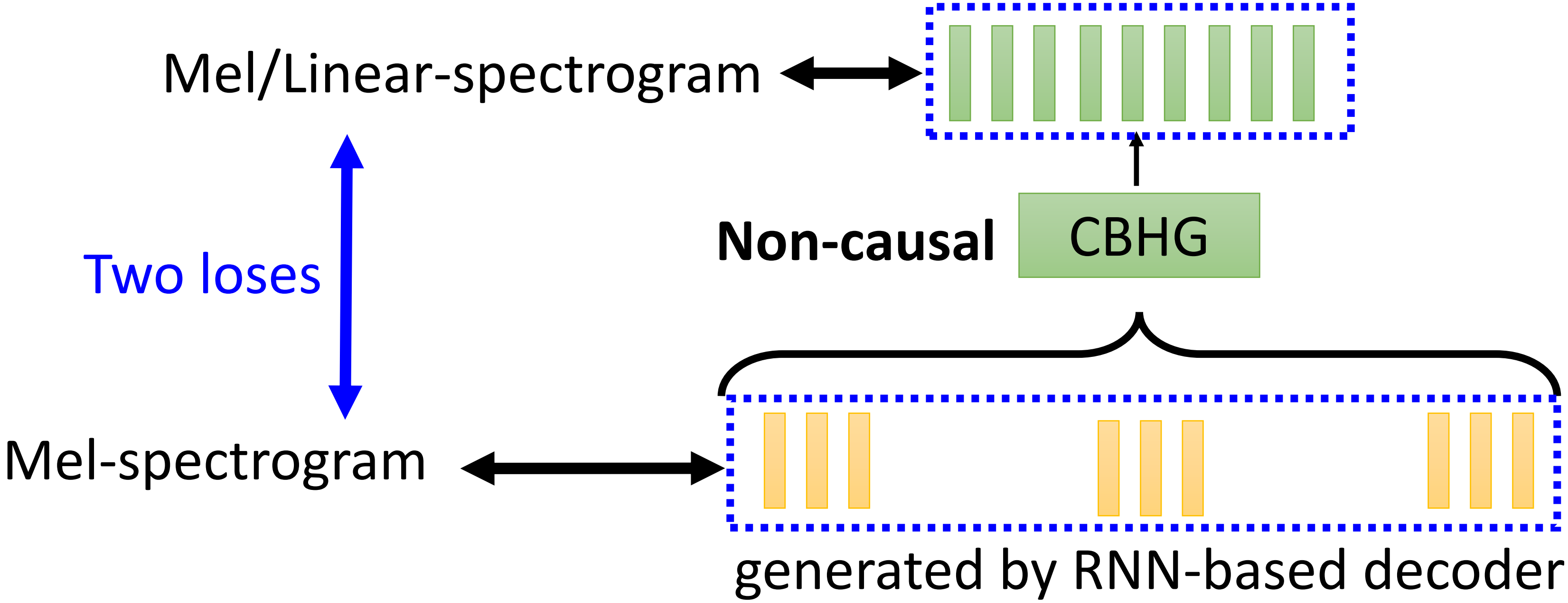
Using teacher forcing, but dropout acts like schedule sampling

Decoder



Using teacher forcing, but dropout acts like schedule sampling

Post processing



Mel/Linear-spectrogram

Two loses

Mel-spectrogram

Non-causal

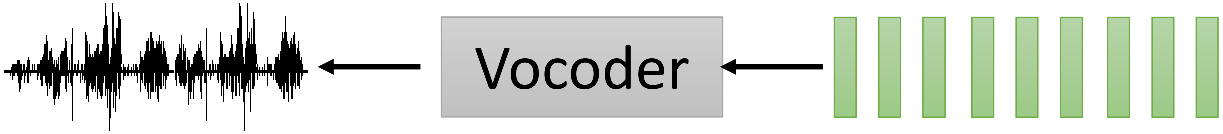
CBHG

generated by RNN-based decoder

Vocoder:

Griffin-Lim in v1

Wavnet in v2

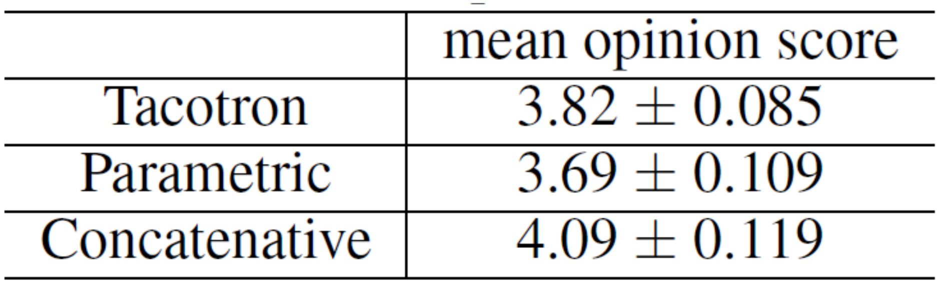


Vocoder

How good is Tacotron?

Version 1

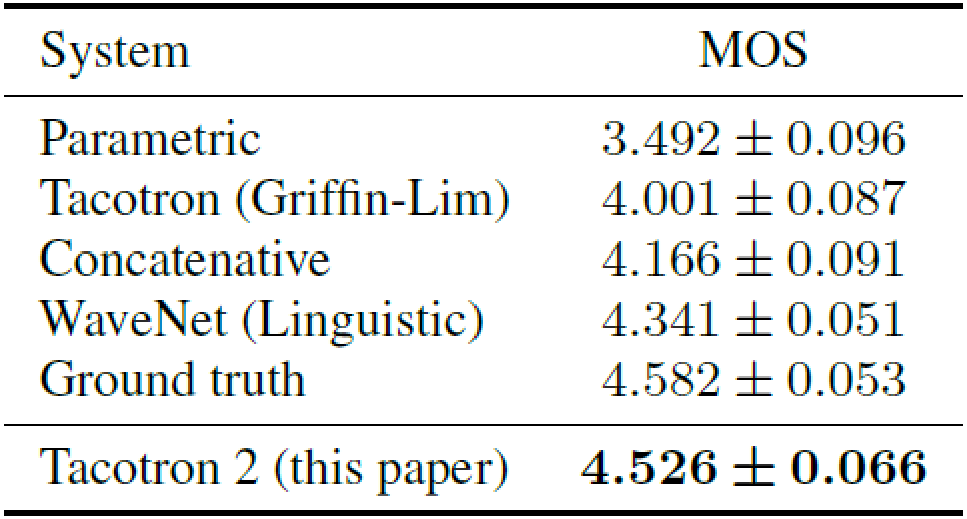
[Wang, et al., INTERSPEECH’17]



	mean opinion score
Tacotron	3.82 ± 0.085
Parametric	3.69 ± 0.109
Concatenative	4.09 ± 0.119

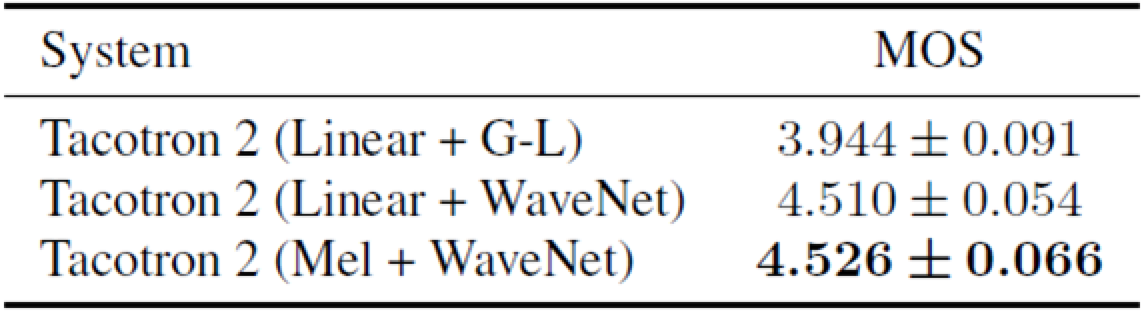
Version 2

[Shen, et al., ICASSP’18]



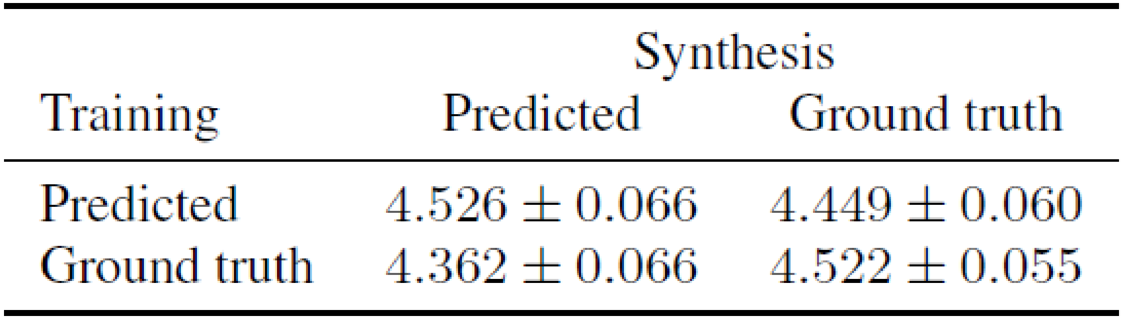
System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

How good is Tacotron?



System	MOS
Tacotron 2 (Linear + G-L)	3.944 ± 0.091
Tacotron 2 (Linear + WaveNet)	4.510 ± 0.054
Tacotron 2 (Mel + WaveNet)	**4.526 ± 0.066**

WaveNet is much better than Griffin-Lim



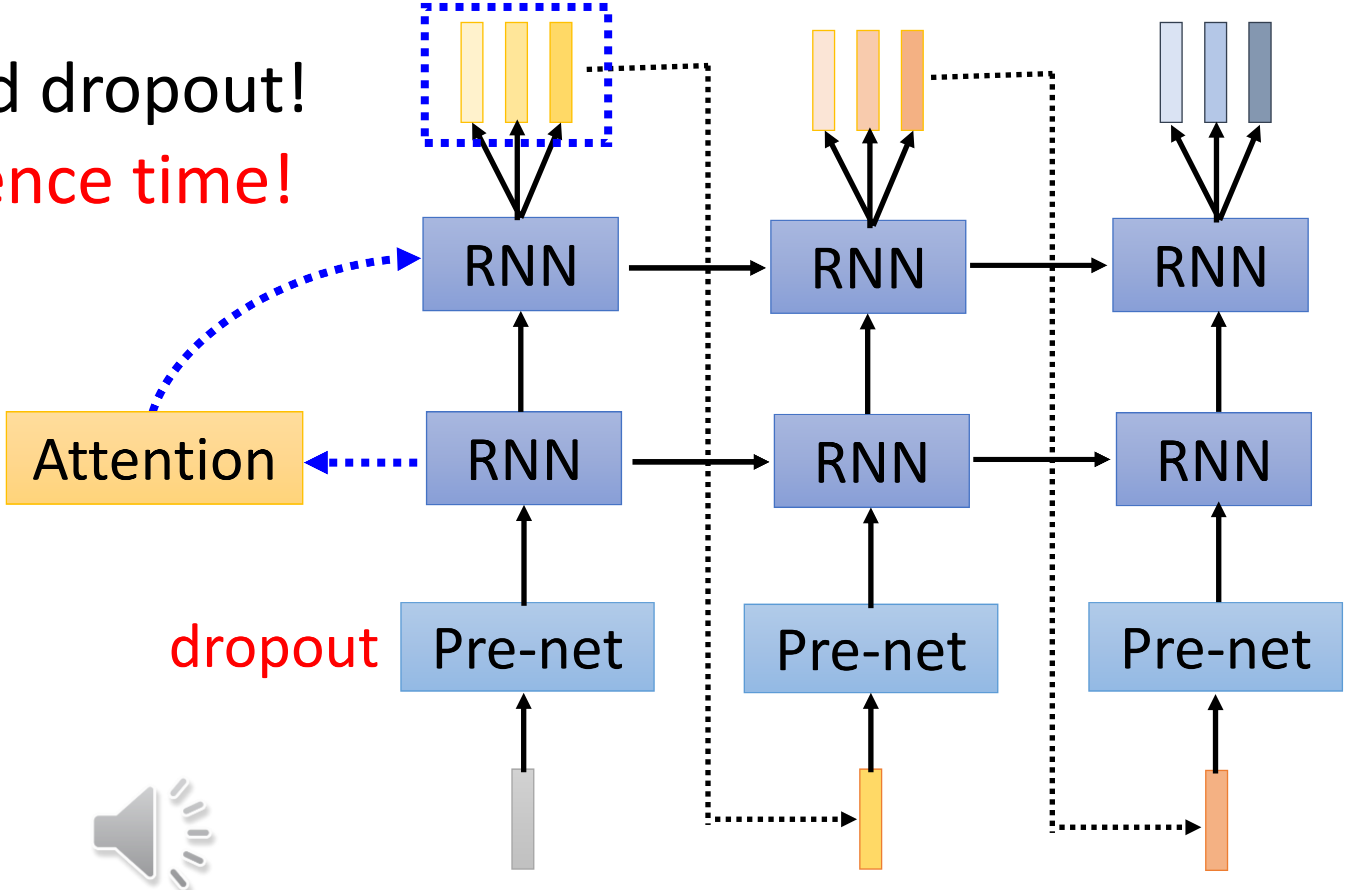
	Synthesis	
Training	Predicted	Ground truth
Predicted	4.526 ± 0.066	4.449 ± 0.060
Ground truth	4.362 ± 0.066	4.522 ± 0.055

WaveNet needs to be trained

Tip at Inference Phase

- You need dropout!

  At inference time!

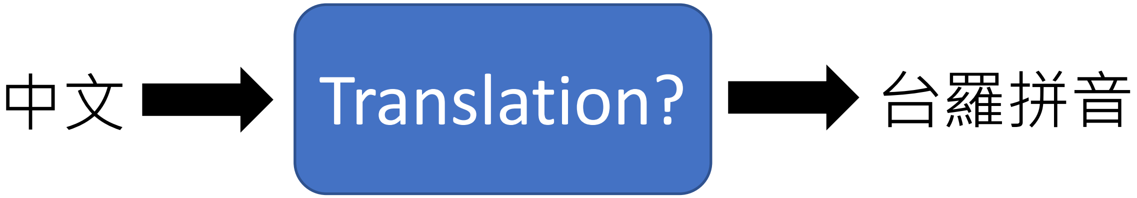


with dropout

without dropout

感謝杜濤同學提供實驗結果

用 Tacotron 做閩南語語音合成

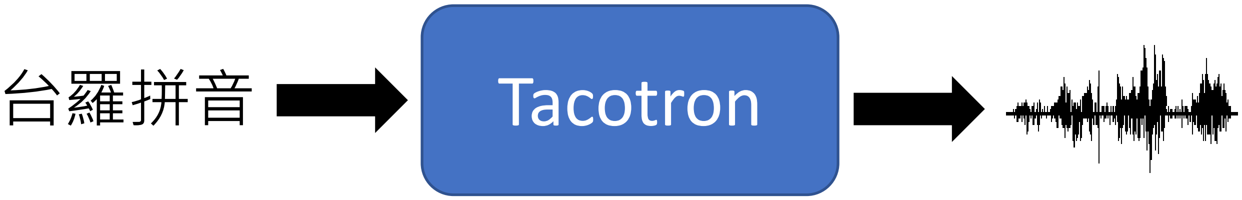

中文

Translation?

台羅拼音

https://i3thuan5.github.io/tai5-uan5_gian5-gi2_kang1-ku7/index.html

台灣語言工具


台羅拼音

Tacotron

Source of training data: https://suisiann-dataset.ithuan.tw/

台灣媠聲2.0

感謝張凱為同學提供實驗結果

Outline

TTS before End-to-end

Tacotron: End-to-end TTS

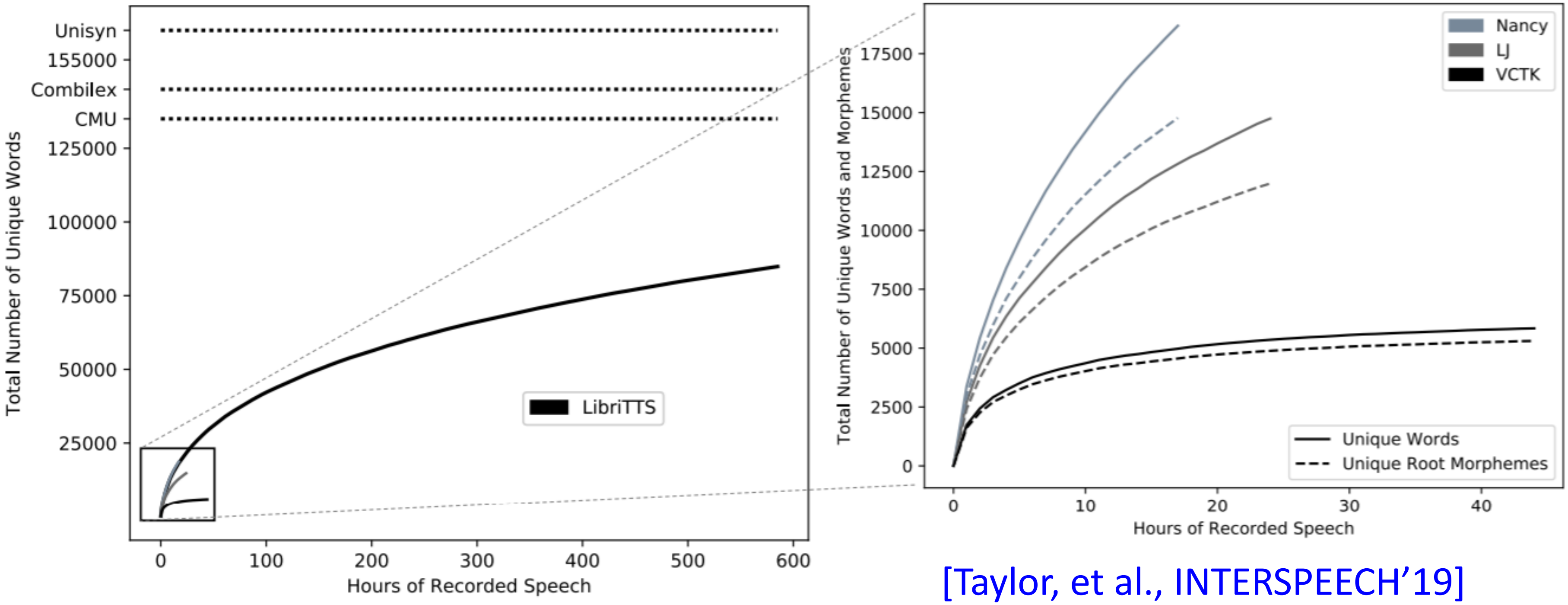
Beyond Tacotron

Controllable TTS

Mispronunciation

- The raters considered ground truth is better than Tacotron 2 because …
- *"… occasional mispronunciation by our system is the primary reason …"*

(LibriTTS dataset 585 hours)

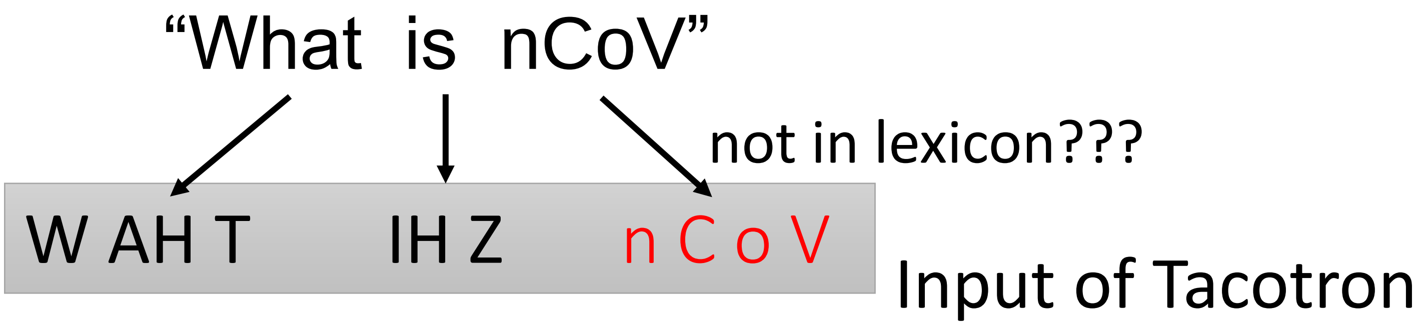


[Taylor, et al., INTERSPEECH'19]

Source of image: https://www.isca-speech.org/archive/Interspeech_2019/pdfs/2830.pdf

Mispronunciation

- Using a lexicon to transform word to phoneme, and using phoneme as Tacotron input
  - But lots of OOV words …



- Character and phoneme hybrid input [Ping, et al., ICLR’18]

If the pronunciation of machine is incorrect, one can add the word into the lexicon to fix the problem.

More information for Encoder

- Syntactic information

  [Guo, et al., INTERSPEECH’19]

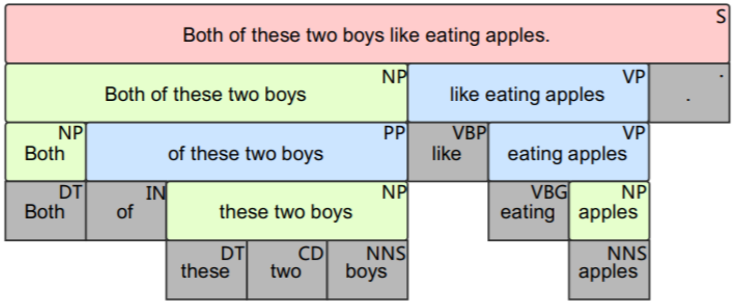


Figure 1: *An example of syntactically parsed tree*

小龍女對楊過說：
「我也想過過過過兒過過的生活」

Source of example:
https://youtu.be/kptTHjBi_ak

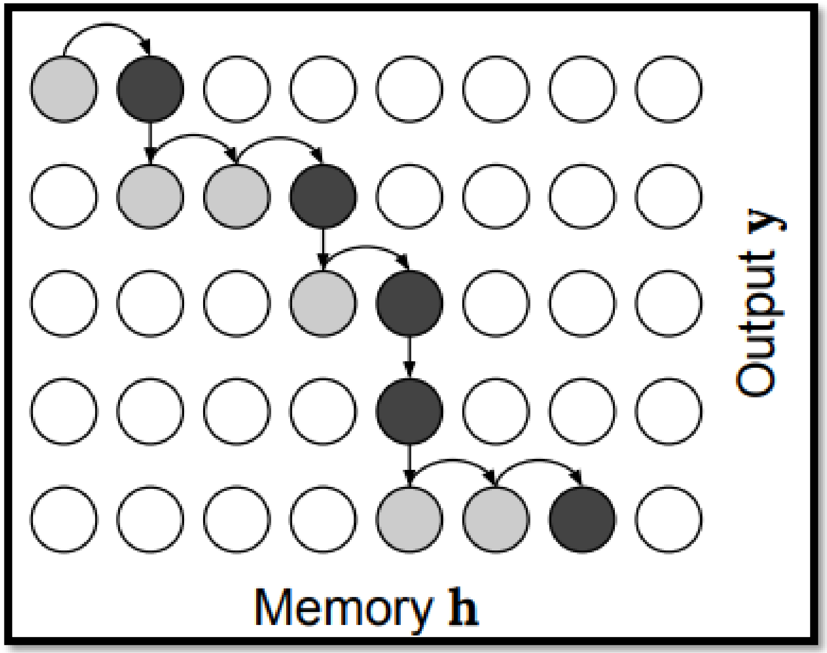
- BERT embedding as input

  [Hayashi, et al., INTERSPEECH’19]

Attention

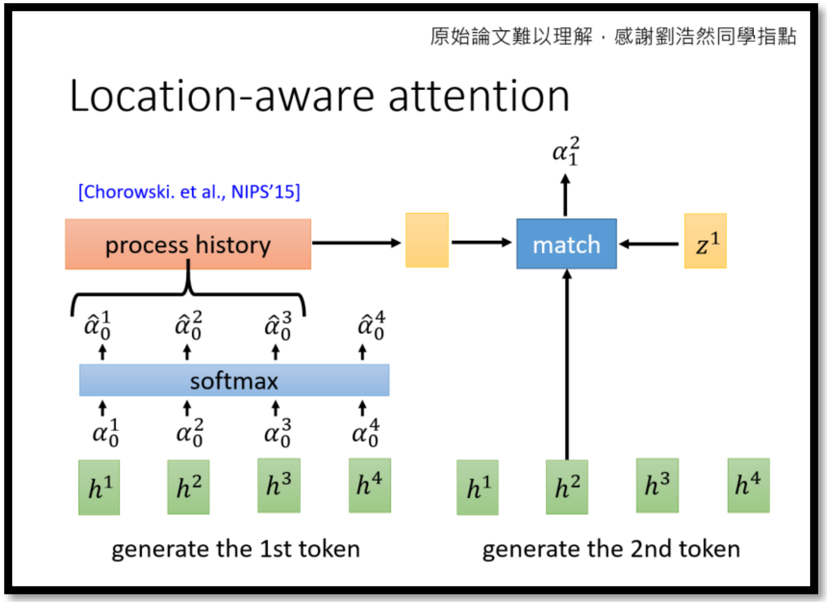
- Monotonic Attention

[Raffel,et al., ICML'17]



- Location-aware attention

(Have been mentioned when we talked about ASR)



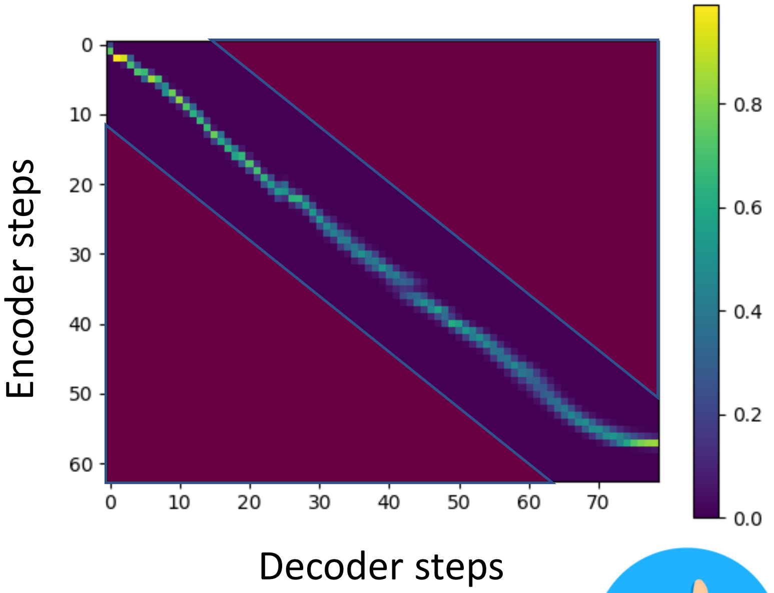
原始論文難以理解，感謝劉浩然同學指點

Location-aware attention

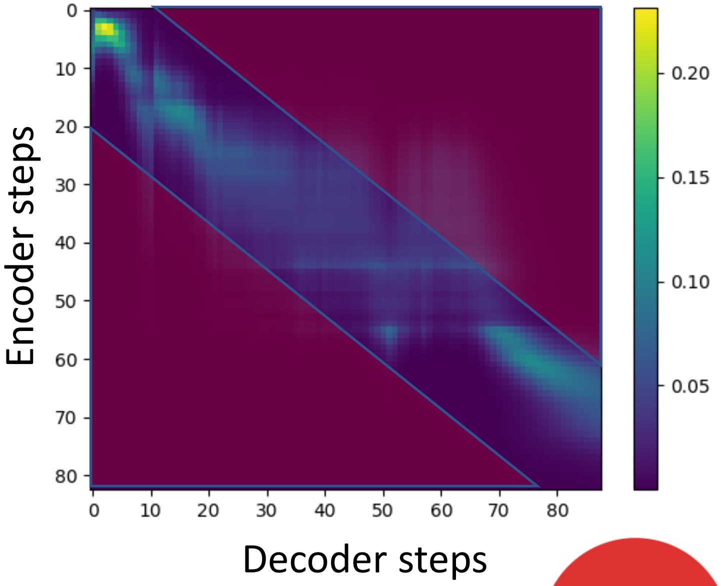
Attention

- Guided Attention [Tachibana, et al., ICASSP’18]

Penalizing the non-diagonal attention matrix during training



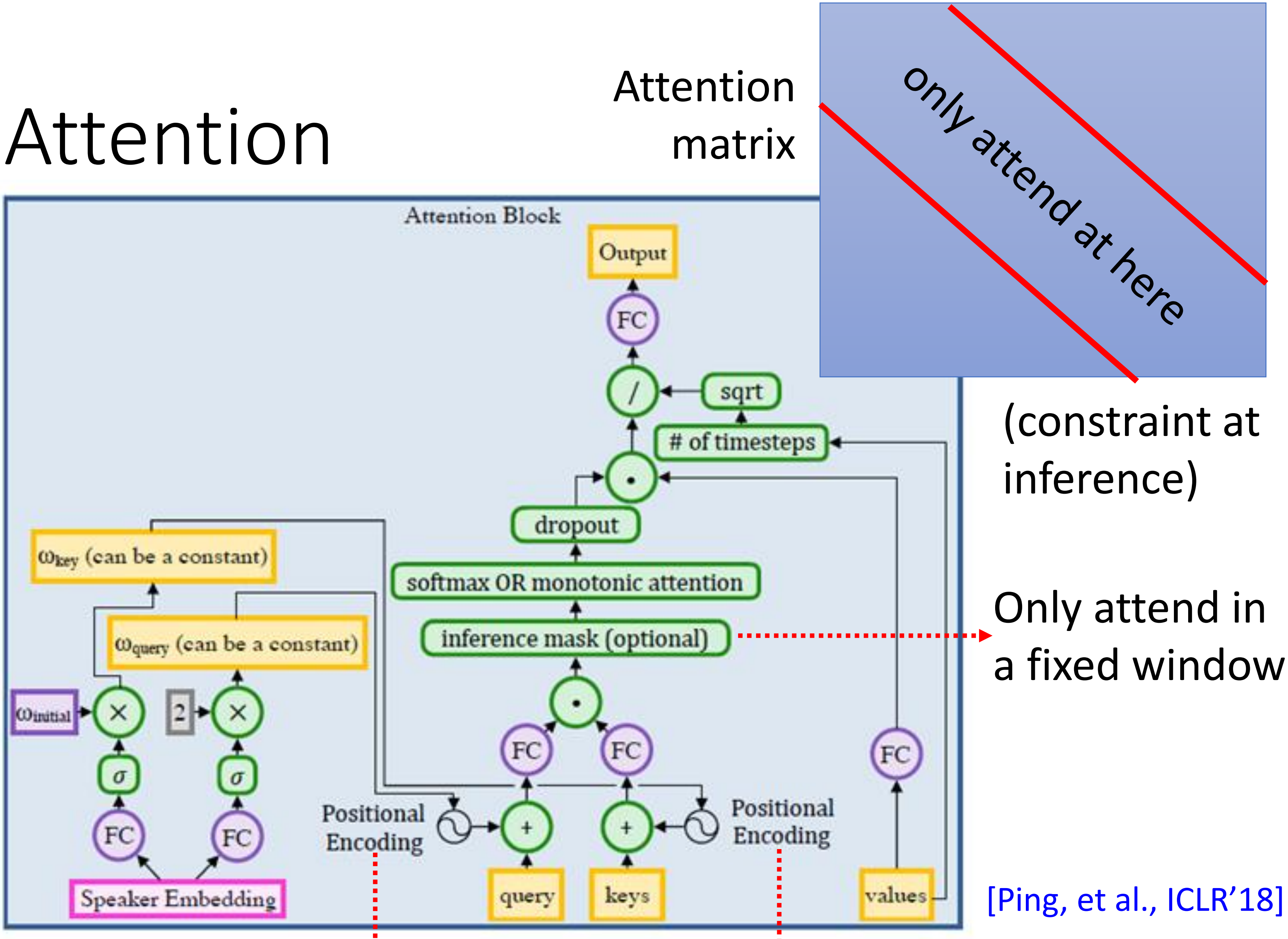
👍



👎

Attention

Attention matrix



(constraint at inference)

Only attend in a fixed window

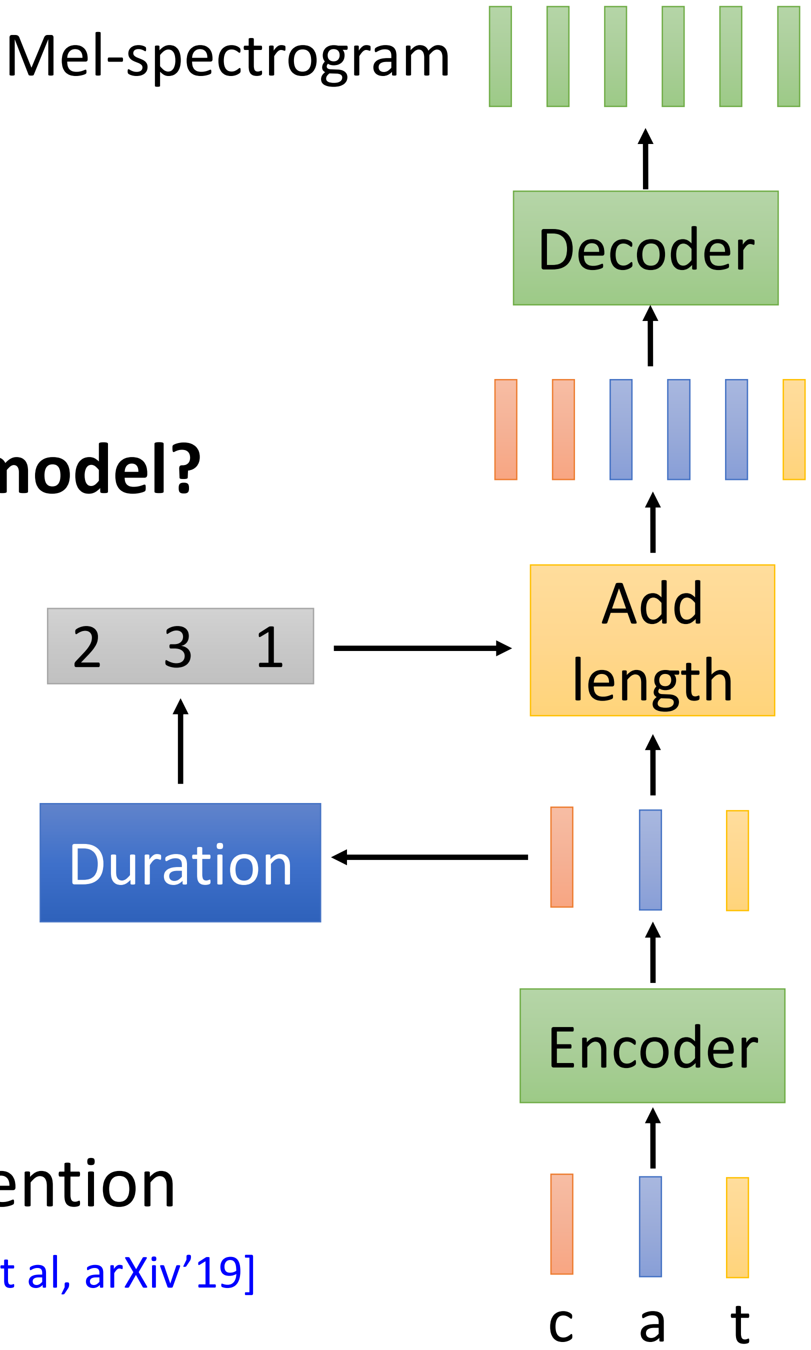
[Ping, et al., ICLR’18]

constraint attention by positional encoding

Fast Speech

[Ren, et al., NeurIPS’19]

How to train this model?



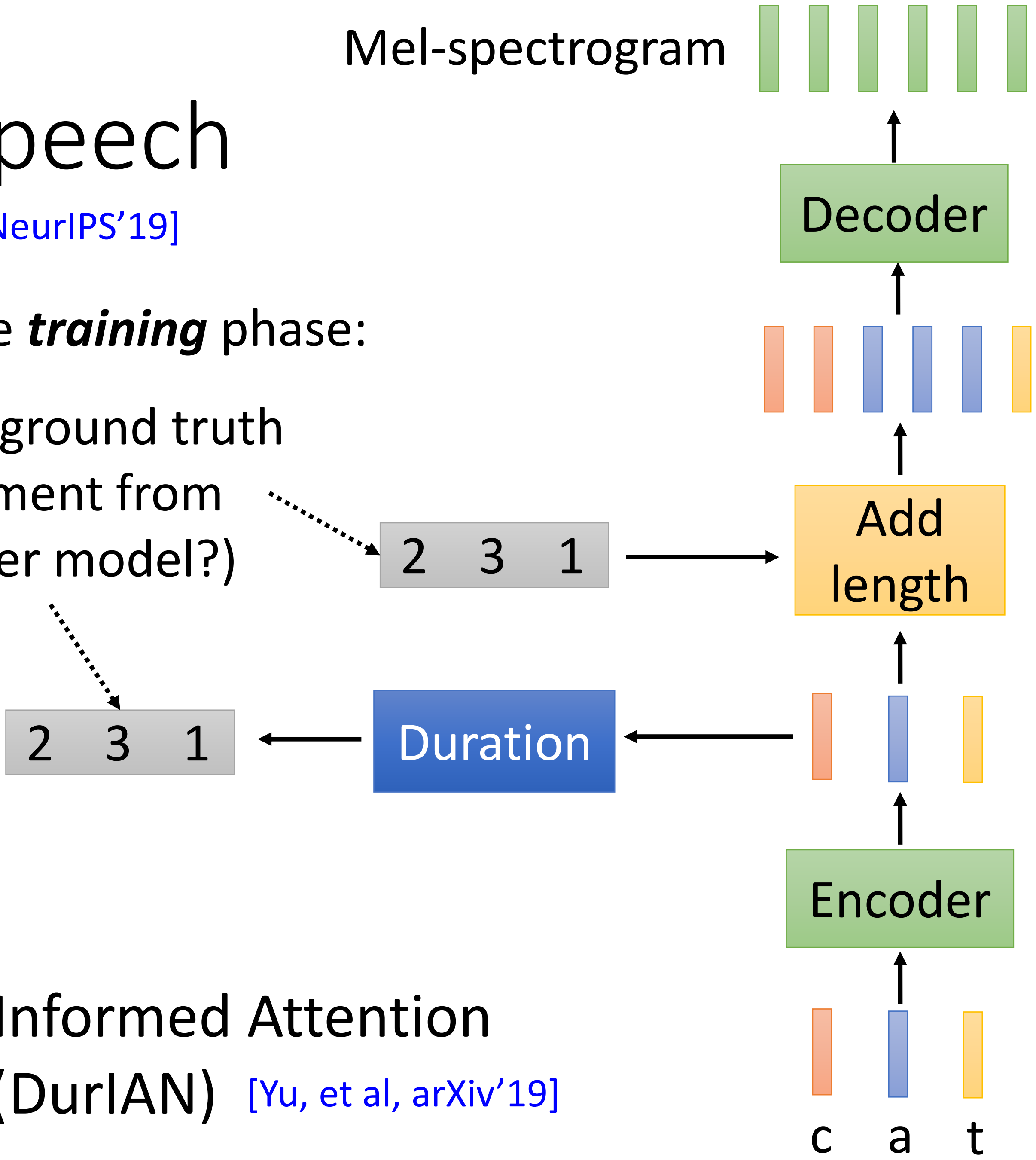
Duration Informed Attention Network (DurIAN) [Yu, et al, arXiv’19]

Fast Speech

[Ren, et al., NeurIPS’19]

During the ***training*** phase:

Using ground truth (alignment from another model?)

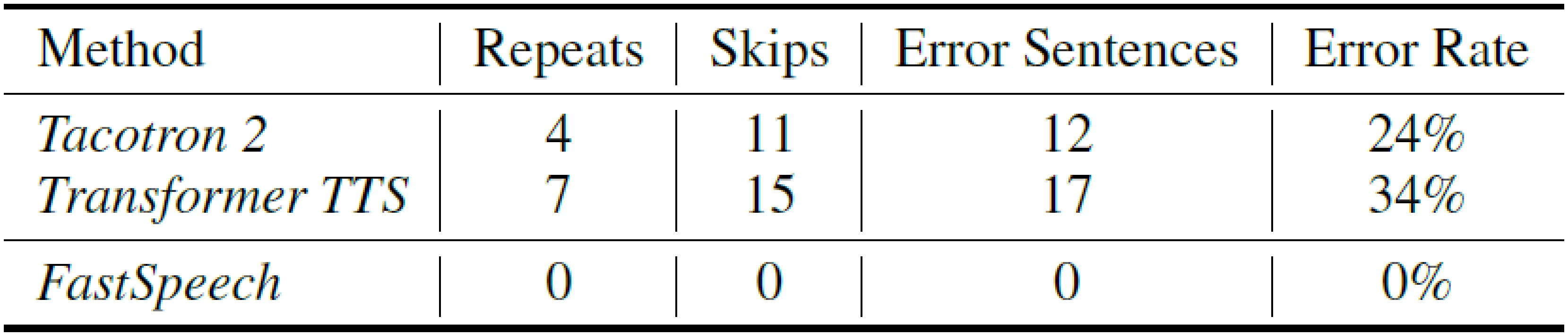


Duration Informed Attention Network (DurIAN) [Yu, et al, arXiv’19]

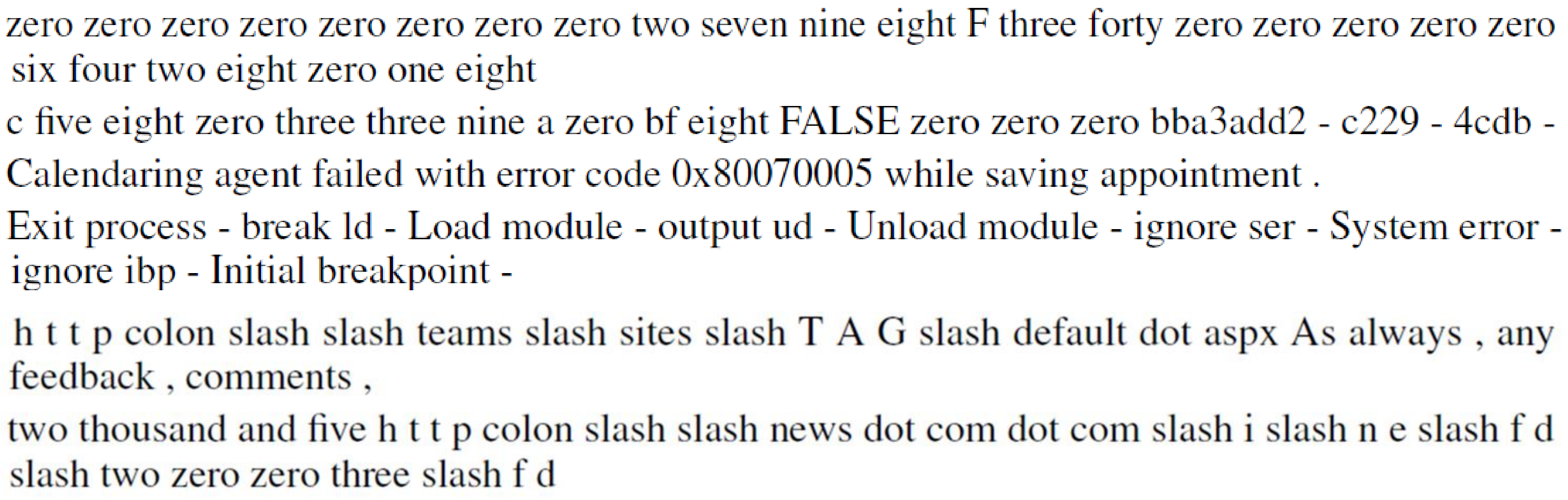
Fast Speech

Source of results:
https://arxiv.org/pdf/1905.09263.pdf

In 50 sentences:



Method	Repeats	Skips	Error Sentences	Error Rate
Tacotron 2	4	11	12	24%
Transformer TTS	7	15	17	34%
FastSpeech	0	0	0	0%



zero zero zero zero zero zero zero zero two seven nine eight F three forty zero zero zero zero zero six four two eight zero one eight

c five eight zero three three nine a zero bf eight FALSE zero zero zero bba3add2 - c229 - 4cdb -

Calendaring agent failed with error code 0x80070005 while saving appointment .

Exit process - break ld - Load module - output ud - Unload module - ignore ser - System error - ignore ibp - Initial breakpoint -

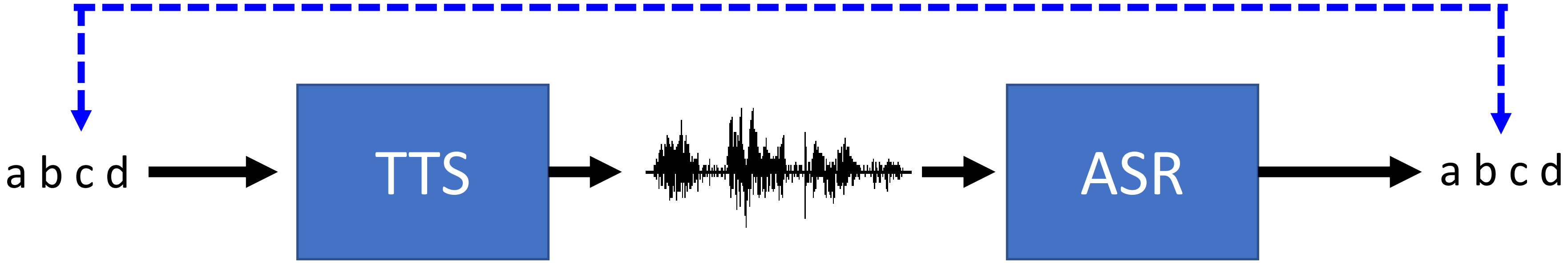
h t t p colon slash slash teams slash sites slash T A G slash default dot aspx As always , any feedback , comments ,

two thousand and five h t t p colon slash slash news dot com dot com slash i slash n e slash f d slash two zero zero three slash f d

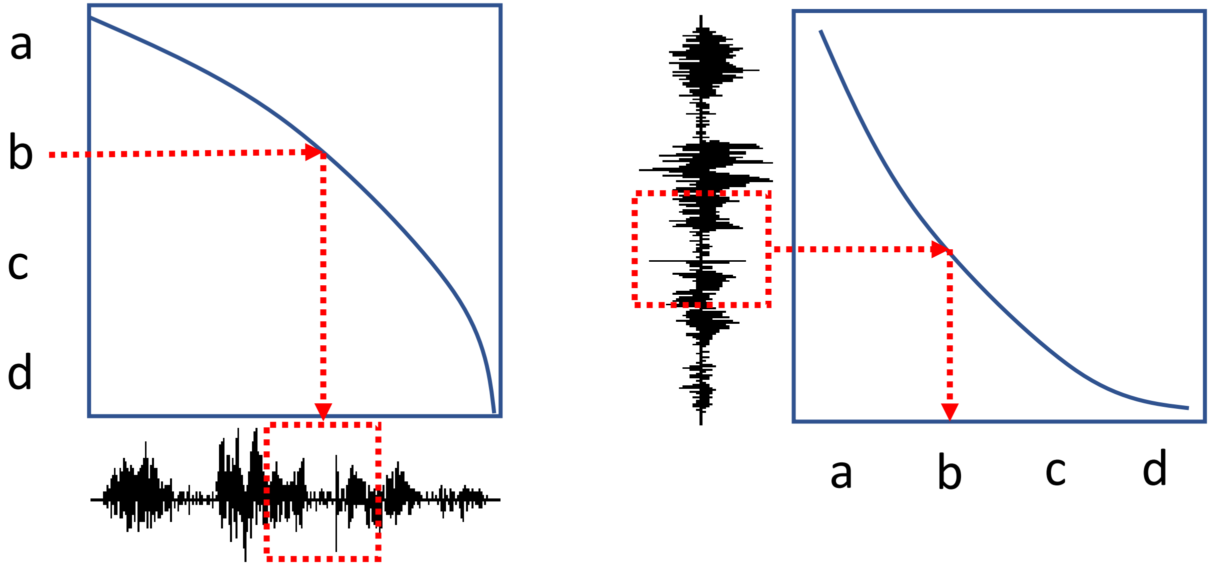
[Tjandra, et al., ASRU’17]

Using ASR to improve TTS

minimize recognition error



attention consistency

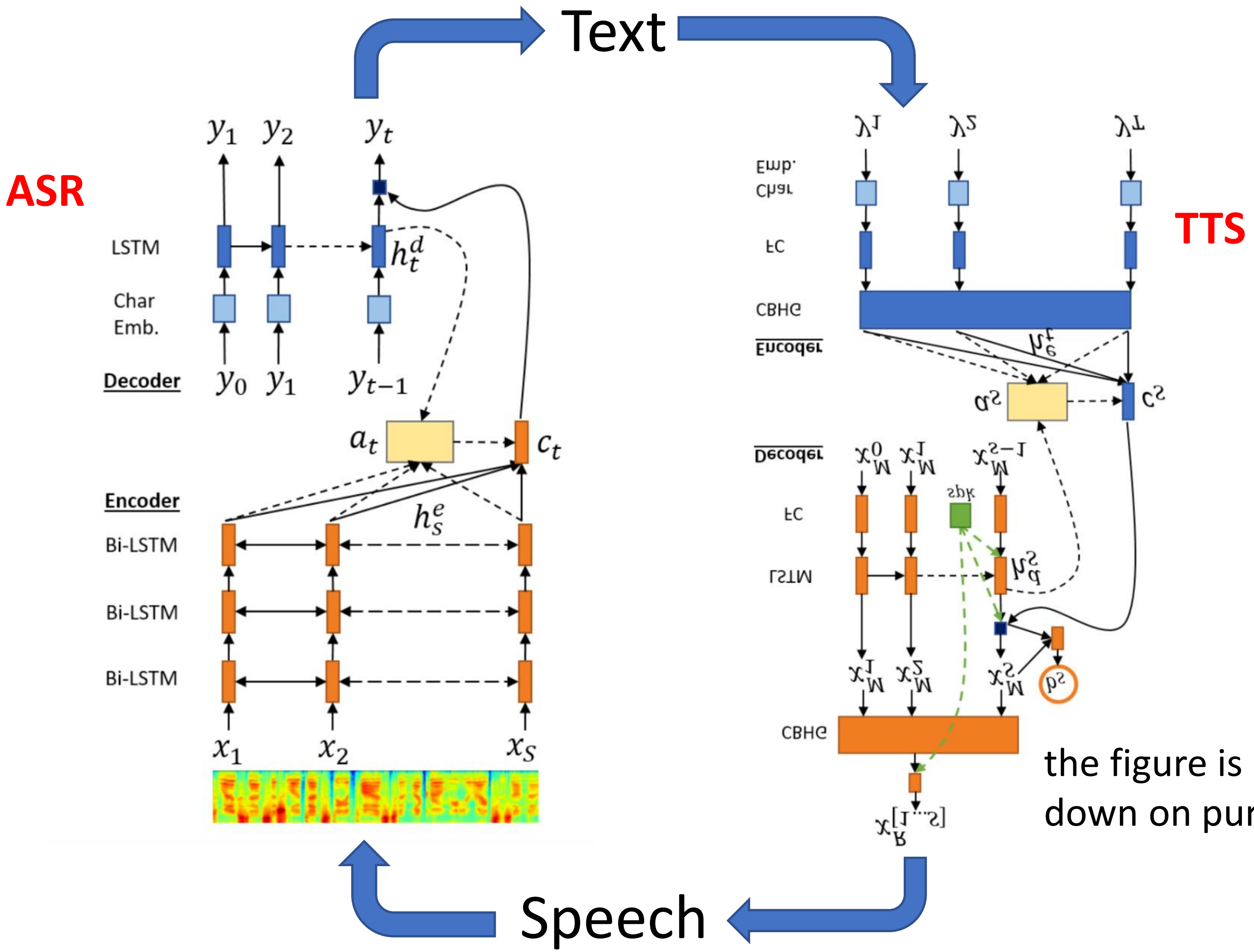


Dual Learning: ASR & TTS

ASR & TTS form a cycle.

Speech Chain

[Tjandra et al., ASRU 2017]



Text

Speech

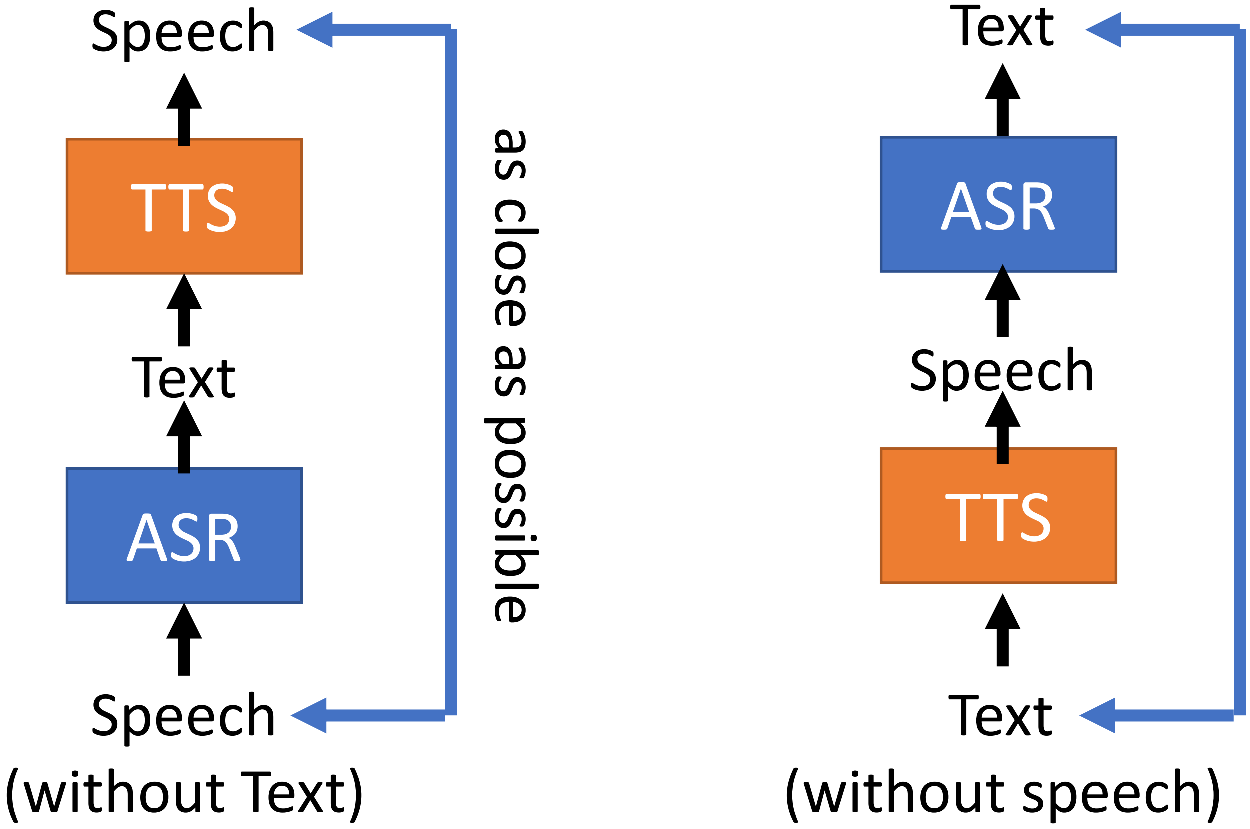
ASR

TTS

the figure is up upside down on purpose

Dual Learning: TTS v.s. ASR

- Given pretrained TTS and ASR system



Dual Learning: TTS v.s. ASR

- Experiments

Mel: mel-spectrogram
Raw: raw waveform

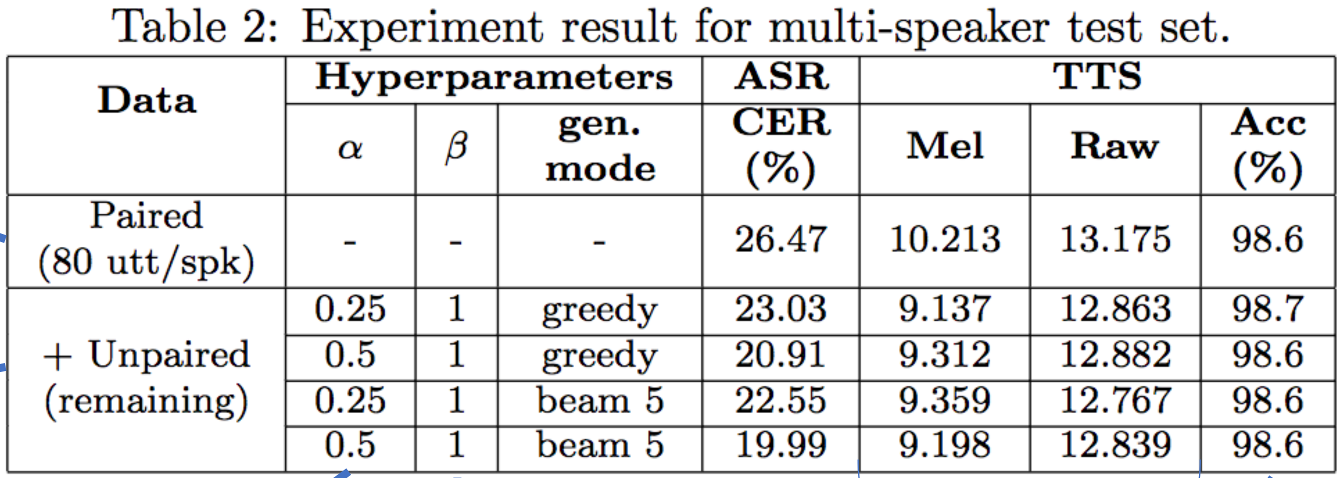


Table 2: Experiment result for multi-speaker test set.

Data	Hyperparameters			ASR	TTS		
	α	β	gen. mode	CER (%)	Mel	Raw	Acc (%)
Paired (80 utt/spk)	-	-	-	26.47	10.213	13.175	98.6
+ Unpaired (remaining)	0.25	1	greedy	23.03	9.137	12.863	98.7
	0.5	1	greedy	20.91	9.312	12.882	98.6
	0.25	1	beam 5	22.55	9.359	12.767	98.6
	0.5	1	beam 5	19.99	9.198	12.839	98.6

1600 utterance-sentence pairs

7200 unpaired utterances and sentences

supervised loss

unsupervised loss

mse

Prediction of the "end-of-utterance"

[Tjandra et al., ASRU 2017]

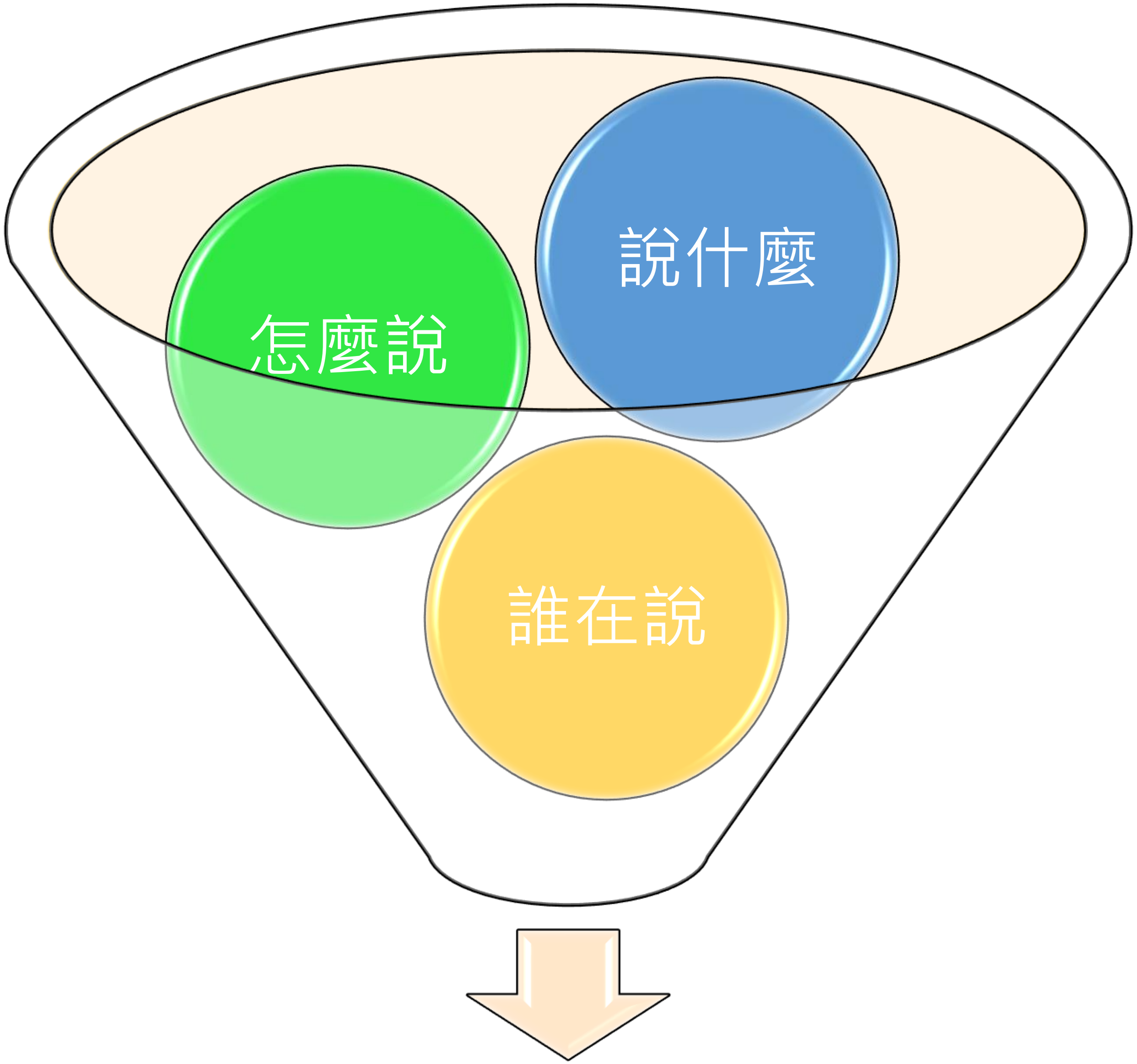
Outline

TTS before End-to-end

Tacotron: End-to-end TTS

Beyond Tacotron

Controllable TTS



說什麼

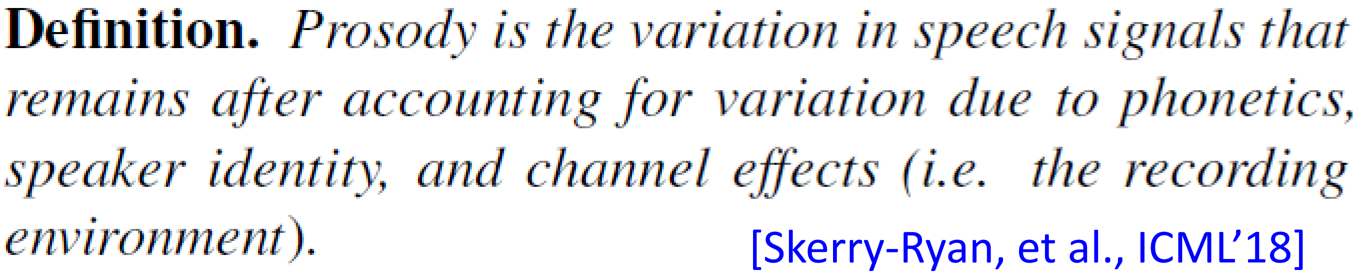
怎麼說

誰在說

Speech

Controllable TTS

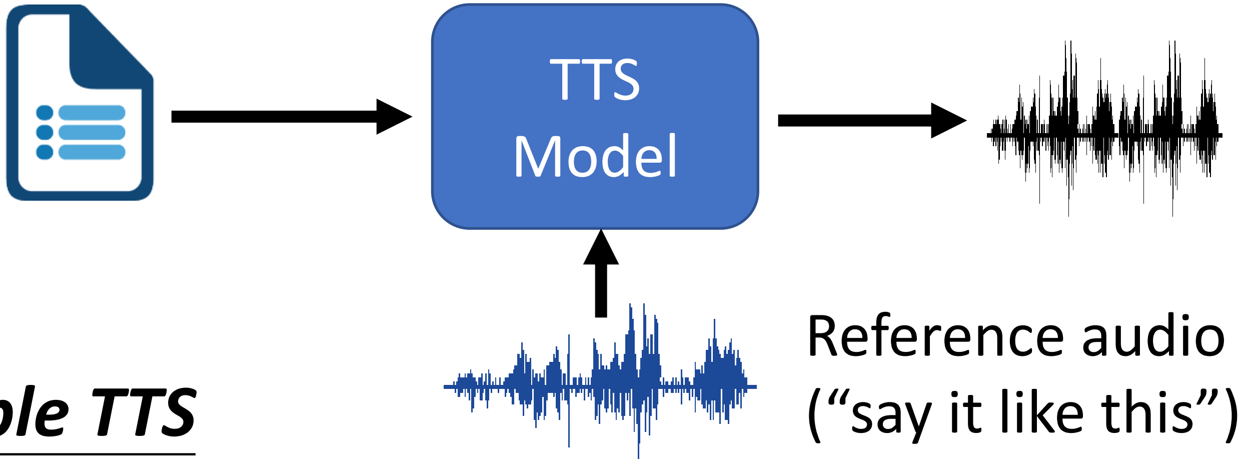
- 誰在說?
  - Voice Cloning
  - Lack of high quality single speaker data to train a speech synthesis system
- 怎麼說?
  - Intonation (語調), stress (重音), rhythm (韻律) …
  - Prosody (抑揚頓挫)



Definition. *Prosody is the variation in speech signals that remains after accounting for variation due to phonetics, speaker identity, and channel effects (i.e. the recording environment).*

[Skerry-Ryan, et al., ICML’18]

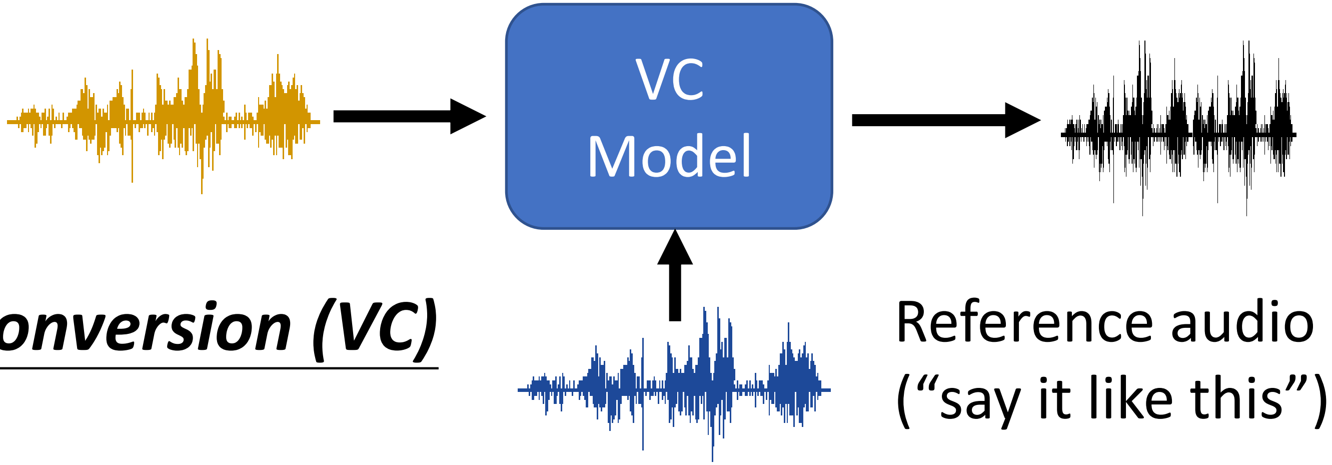
Controllable TTS v.s. VC



TTS Model

Reference audio (“say it like this”)

Controllable TTS

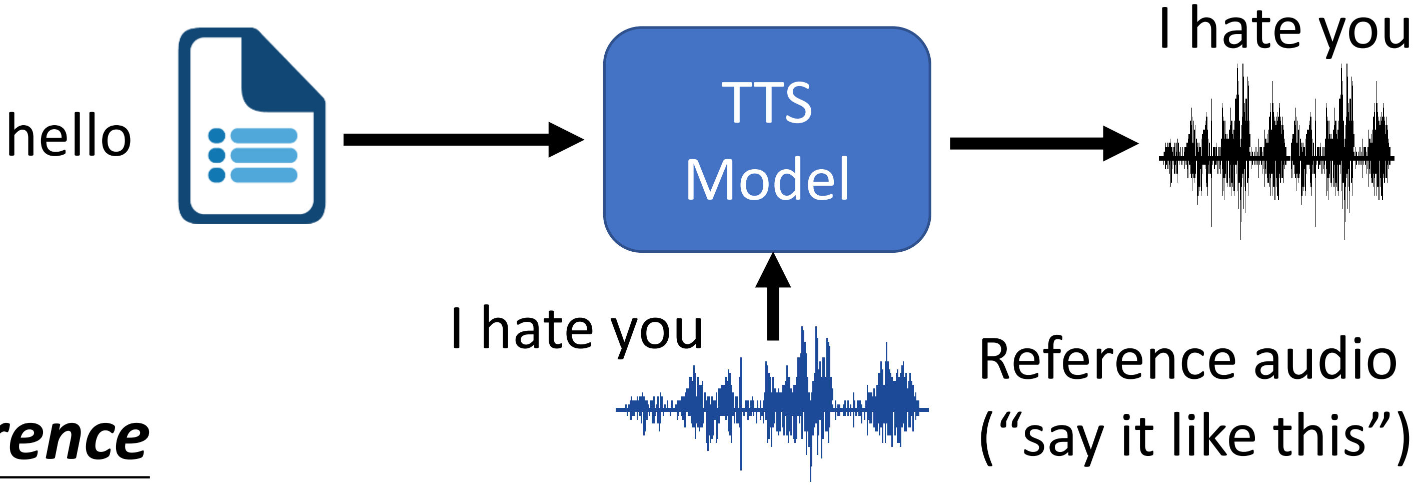


VC Model

Reference audio (“say it like this”)

Voice Conversion (VC)

Controllable TTS



hello

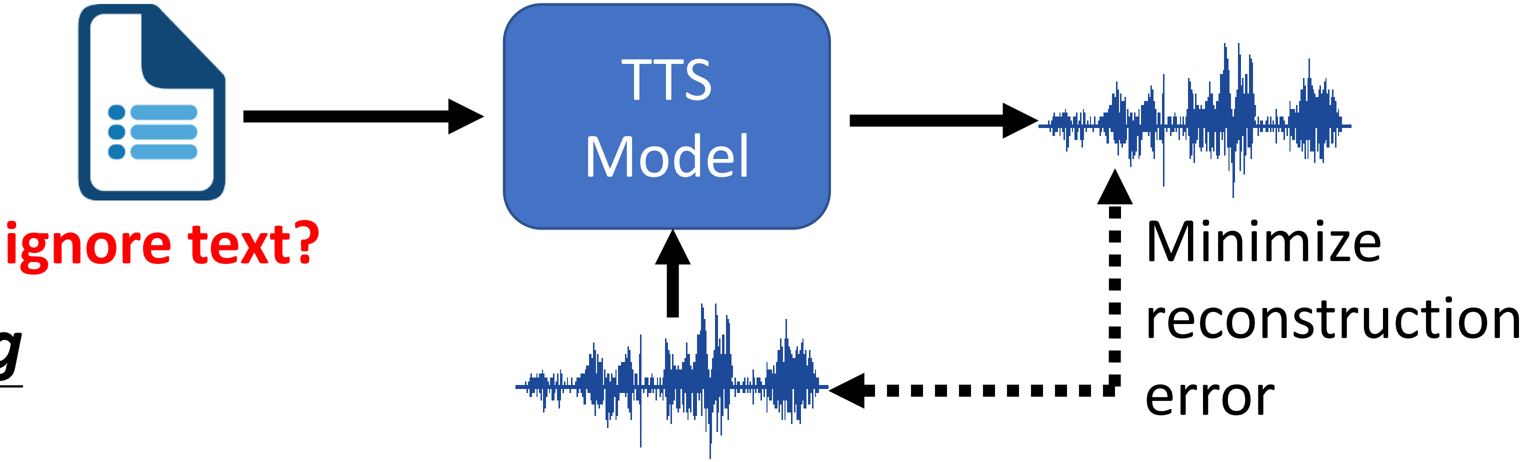
TTS Model

I hate you

I hate you

Reference audio
(“say it like this”)

Inference



TTS Model

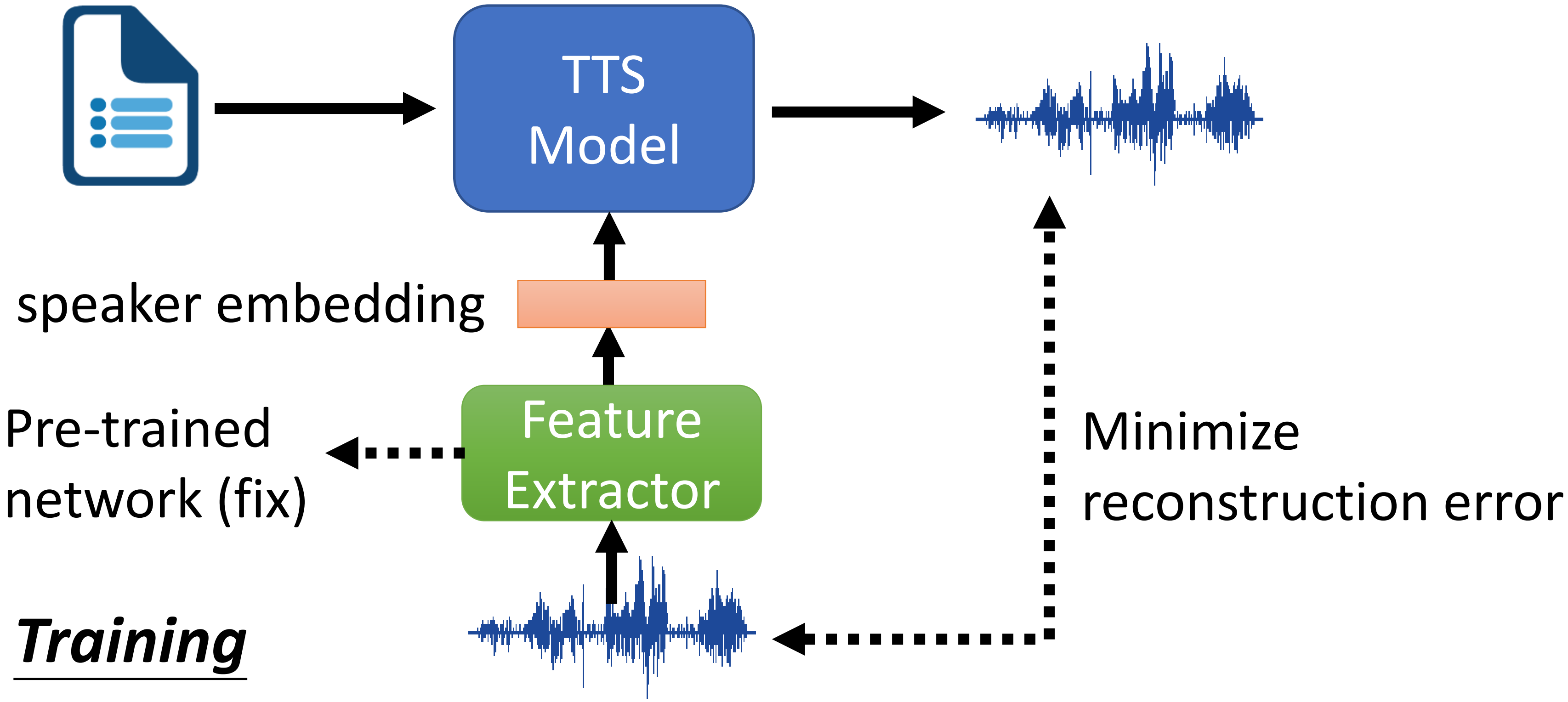
ignore text?

Minimize reconstruction error

Training

Voice Cloning

[Jia, et al., NeurIPS'18]



TTS Model

speaker embedding

Feature Extractor

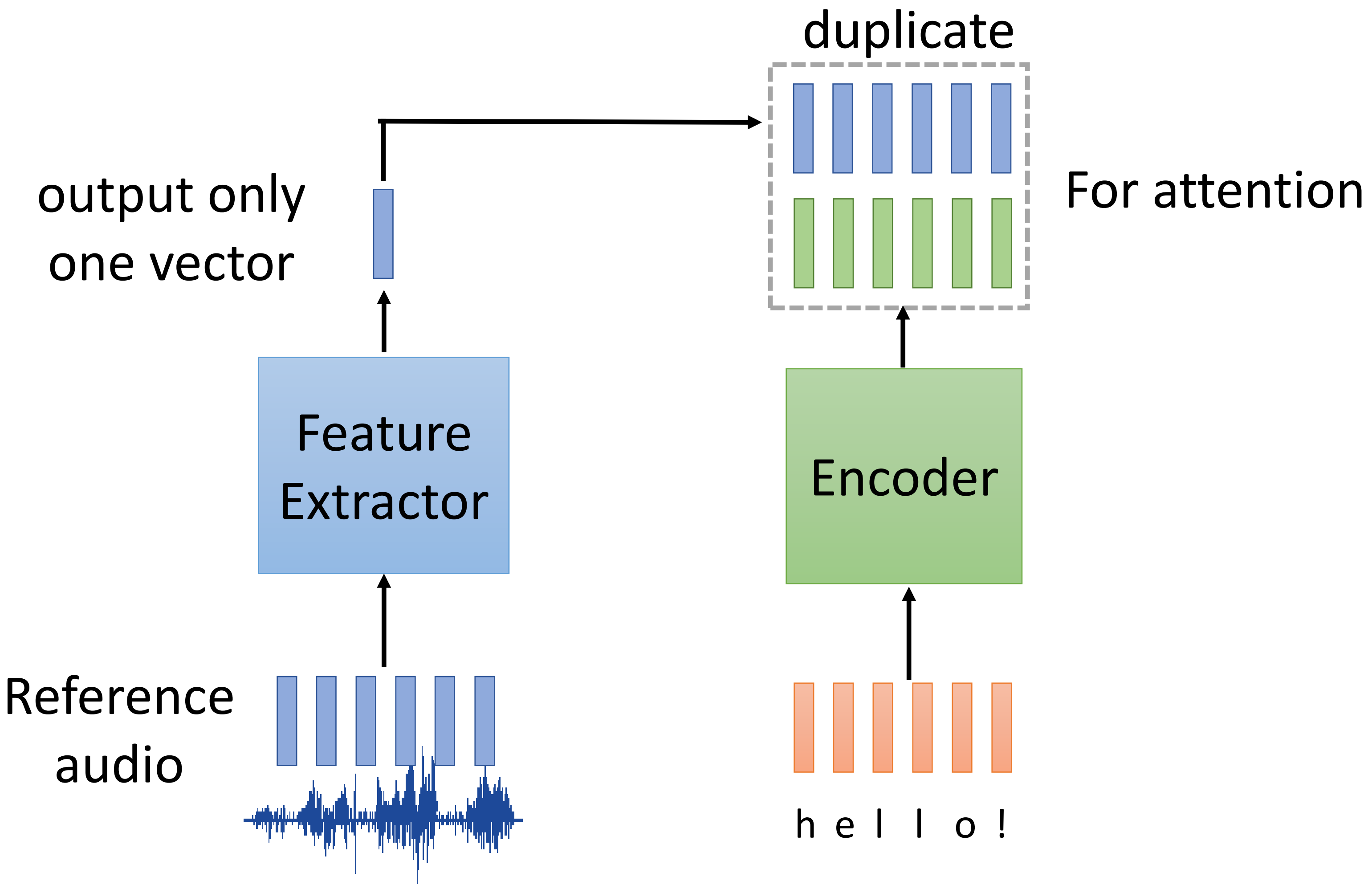
Pre-trained network (fix)

Minimize reconstruction error

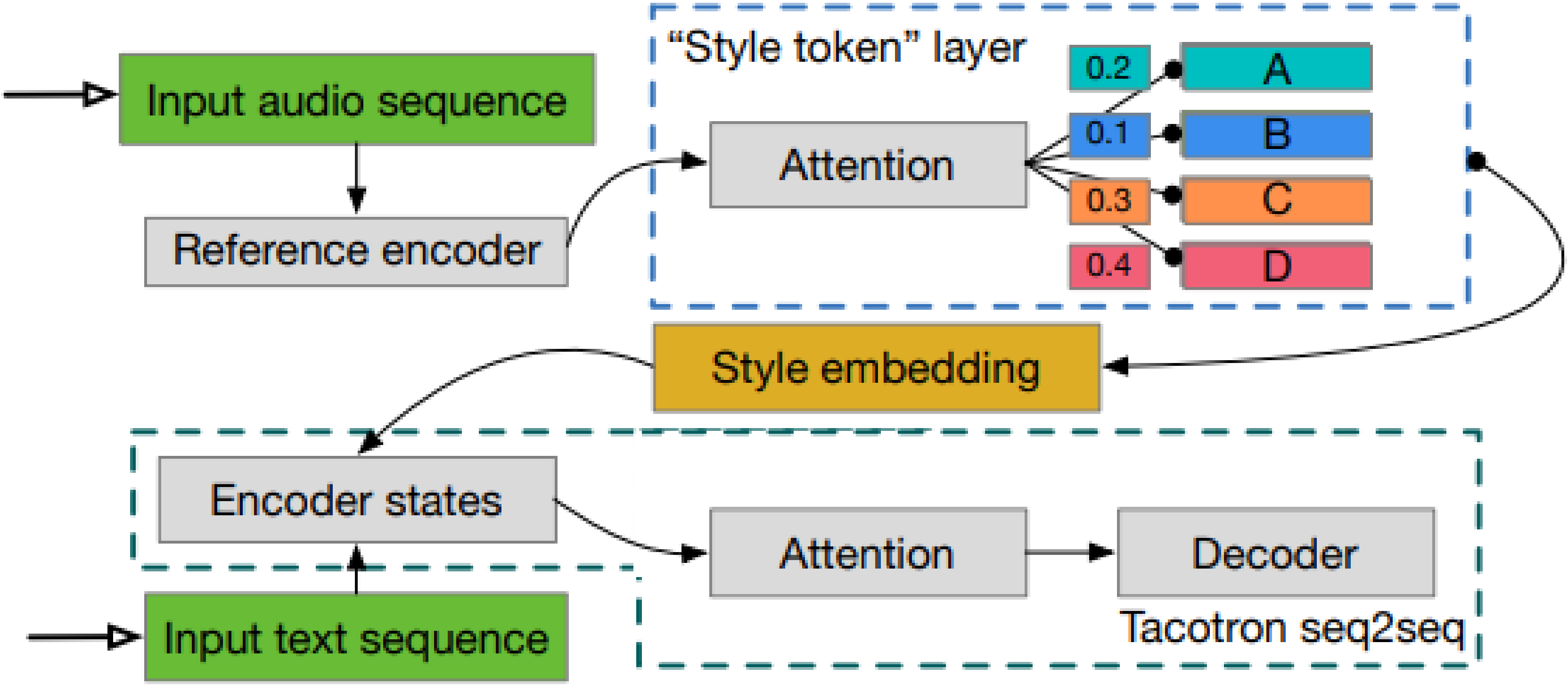
Training

GST-Tacotron

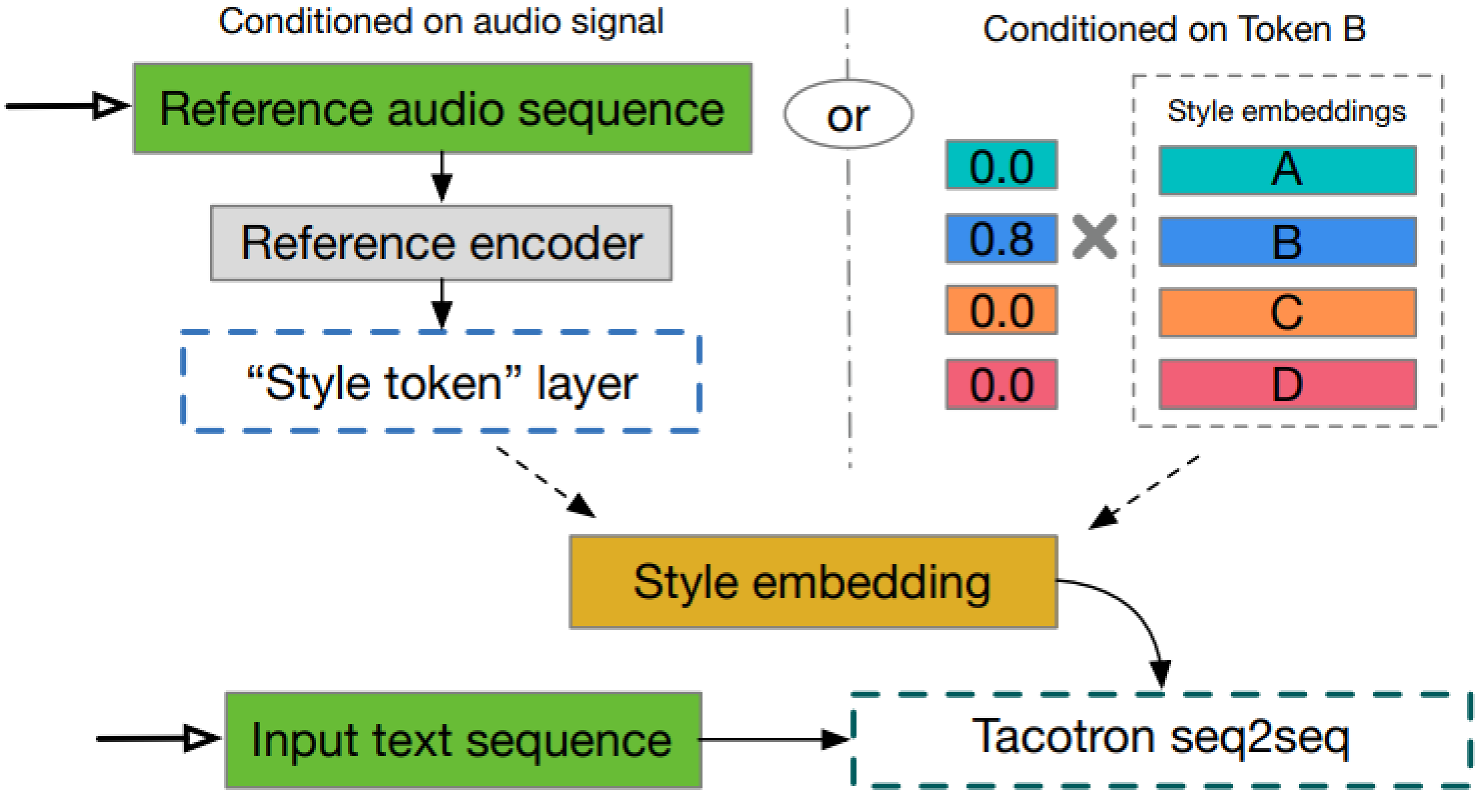
GST = global style tokens
[Wang, et al., ICML'18]



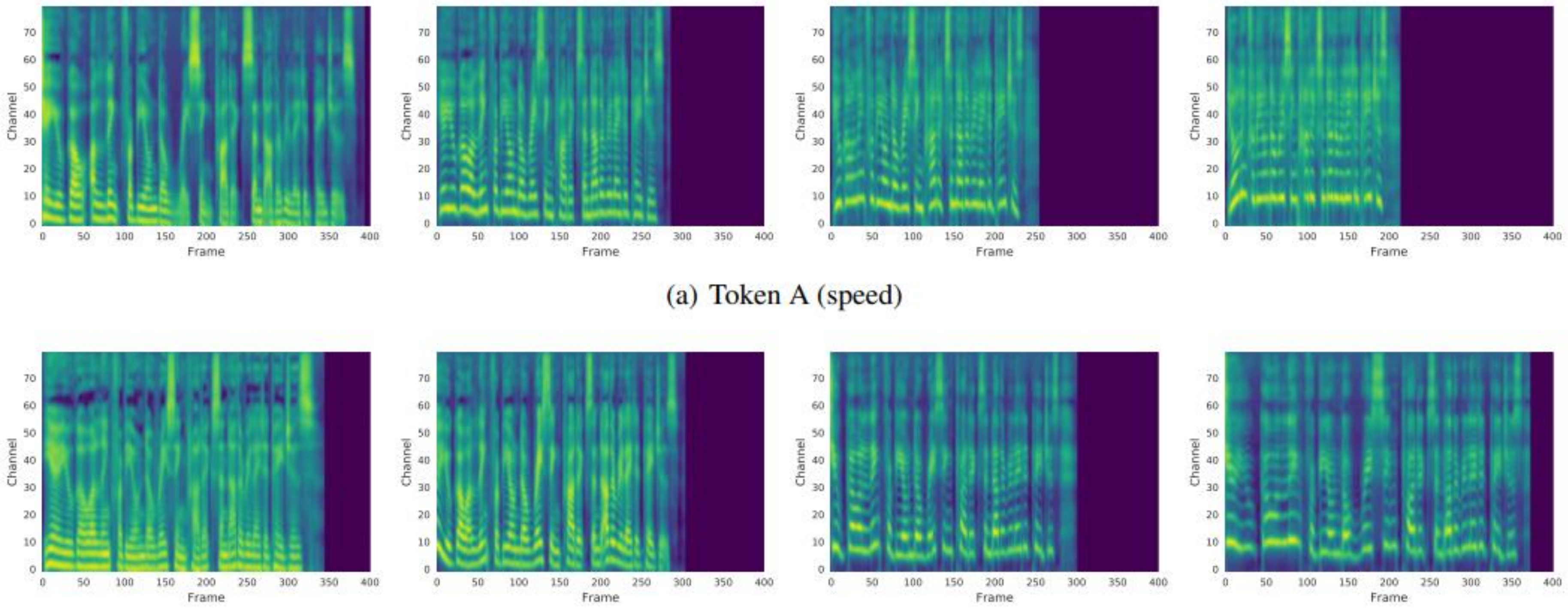
GST-Tacotron



GST-Tacotron



- What does the tokens effect?
  - One token corresponds to a lower pitch voice
  - One token for a decreasing pitch
  - One token for a faster speaking rate
  - ……



(a) Token A (speed)

(b) Token B (animated)

Source of image: https://arxiv.org/pdf/1803.09017.pdf

Concluding Remarks

- TTS before End-to-end
- Tacotron: End-to-end TTS
- Beyond Tacotron
- Controllable TTS

Reference

- [Wang, et al., INTERSPEECH’17] Yuxuan Wang, R.J. 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, 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, Yonghui Wu, Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, ICASSP, 2018
- [Wang, et al., INTERSPEECH’16] Wenfu Wang, Shuang Xu, Bo Xu, First Step Towards End-to-end Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention, INTERSPEECH, 2016
- [Sotelo, et al., ICLR workshop’17] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, Yoshua Bengio, Char2Wav: End-to-End Speech Synthesis, ICLR workshop, 2017
- [Taylor, et al., INTERSPEECH’19] Jason Taylor, Korin Richmond, Analysis of Pronunciation Learning in End-to-End Speech Synthesis, INTERSPEECH, 2019

Reference

- [Hayashi, et al., INTERSPEECH’19] Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Shubham Toshniwal, Karen Livescu, Pre-trained Text Embeddings for Enhanced Text-to-Speech Synthesis, INTERSPEECH, 2019
- [Zhang, et al., arXiv’19] Liqiang Zhang, Chengzhu Yu, Heng Lu, Chao Weng, Yusong Wu, Xiang Xie, Zijin Li, Dong Yu, Learning Singing From Speech, arXiv, 2019
- [Guo, et al., INTERSPEECH’19] Haohan Guo, Frank K. Soong, Lei He, Lei Xie, Exploiting Syntactic Features in a Parsed Tree to Improve End-to-End TTS, INTERSPEECH, 2019
- [Wang, et al., ICML’18] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, Rif A. Saurous, Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis, ICML, 2018
- [Skerry-Ryan, et al., ICML’18] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, Rif A. Saurous, Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron, ICML, 2018

Reference

- [Arik, et al., ICML’17] Sercan O. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi, Deep Voice: Real-time Neural Text-to-Speech, ICML, 2017
- [Ping, et al., ICLR’18] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, John Miller, Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning, ICLR, 2018
- [Ren, et al., NeurIPS’19] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu, FastSpeech: Fast, Robust and Controllable Text to Speech, NeurIPS, 2019
- [Yu, et al, arXiv’19] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, Dan Su, Dong Yu, DurIAN: Duration Informed Attention Network For Multimodal Synthesis, arXiv, 2019
- [Raffel,et al., ICML’17] Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, Douglas Eck, Online and Linear-Time Attention by Enforcing Monotonic Alignments, ICML, 2017

Reference

- [Tachibana, et al., ICASSP’18] Hideyuki Tachibana, Katsuya Uenoyama, Shunsuke Aihara, Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention, ICASSP, 2018
- [Jia, et al., NeurIPS’18] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis, NeurIPS, 2018
- [Arik, et al., NeurIPS’18] Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, Yanqi Zhou, Neural Voice Cloning with a Few Samples, NeurIPS, 2018
- [Tjandra, et al., ASRU’17] Andros Tjandra, Sakriani Sakti, Satoshi Nakamura, Listening while Speaking: Speech Chain by Deep Learning, ASRU, 2017
- [Liu, et al., SLT’18] Da-Rong Liu, Chi-Yu Yang, Szu-Lin Wu, Hung-Yi Lee, "Improving Unsupervised Style Transfer in End-to-End Speech Synthesis with End-to-End Speech Recognition", SLT, 2018