Tutorial on end-to-end text-to-speech synthesis

Part 2 – Tactron and related end-to-end systems

安田裕介 (Yusuke YASUDA)
National Institute of Informatics

エンドツーエンド音声合成に向けたNIIにおけるソフトウェア群 Part 2

〜 TacotronとWaveNetのチュートリアル 〜

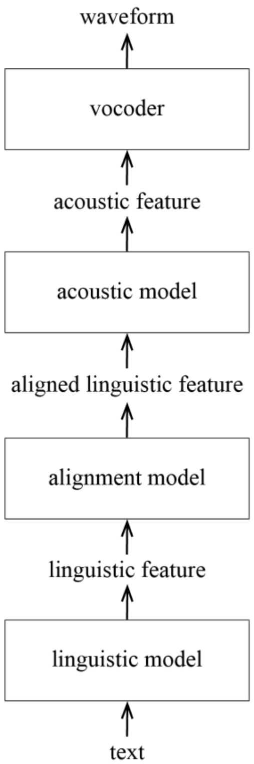
安田裕介
National Institute of Informatics

About me

- Name: 安田裕介
- Status:
  - A PhD student at Sokendai / NII
  - A programmer at work
- Former geologist
  - bachelor, master
- Github account: TanUkkii007

Table of contents

TTS architecture: traditional pipeline



- Typical pipeline architecture for statistical parametric speech synthesis

- Consists of task-specific models

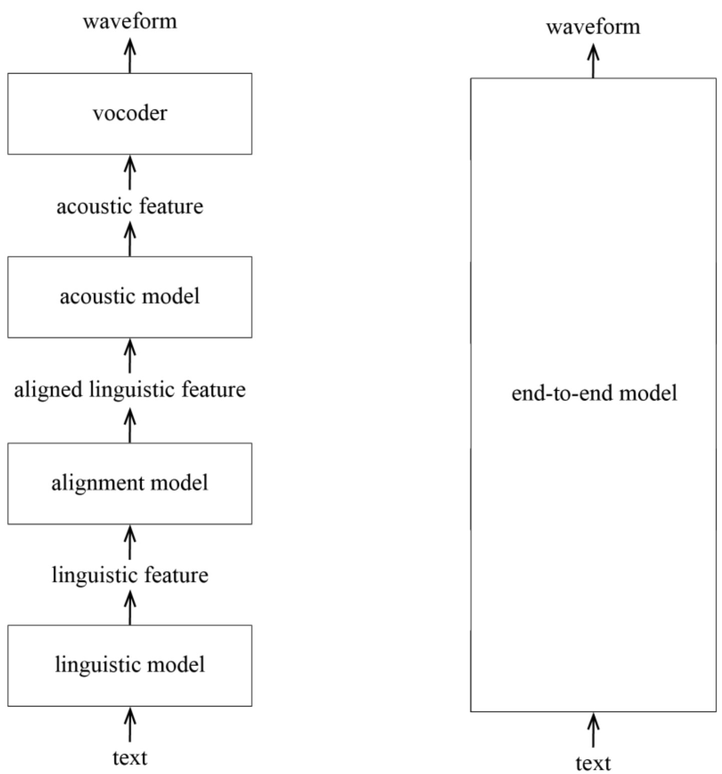
  - linguistic model

  - alignment (duration) model

  - acoustic model

  - vocoder

TTS architecture: End-to-end model



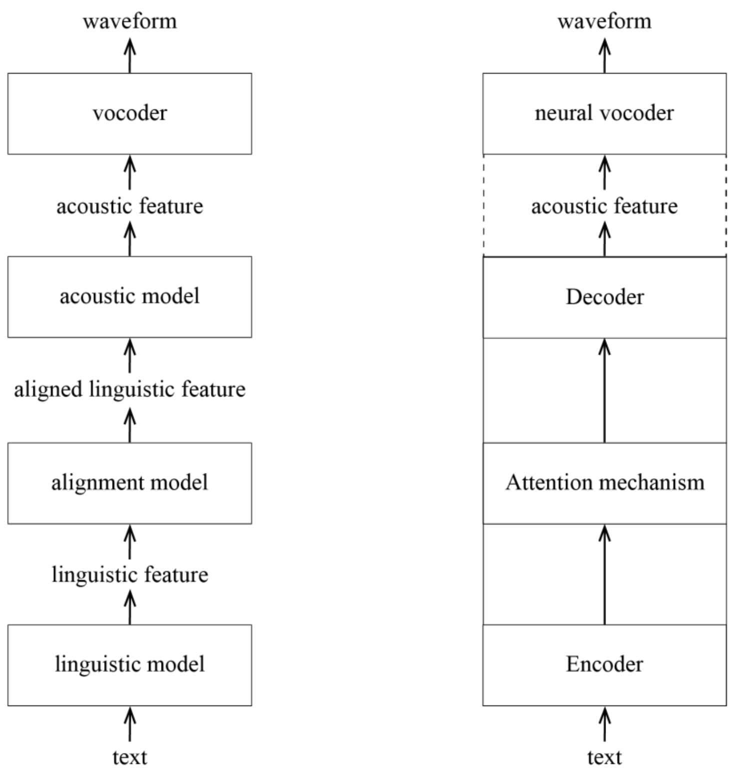
- End-to-end model directly converts text to waveform

- End-to-end model does not require intermediate feature extraction

- Pipeline models accumulate errors across predicted features

- End-to-end model's internal blocks are jointly optimized

End-to-end model: Encoder-Decoder with Attention



- Building blocks
  - Encoder
  - Attention mechanism
  - Decoder
  - Neural vocoder

Conventional end-to-end models may not include a waveform generator, but some recent full end-to-end models contain a neural waveform generator, e.g. ClariNet[1].

[1] Wei Ping, Kainan Peng, Jitong Chen: ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech. CoRR abs/1807.07281 (2018)

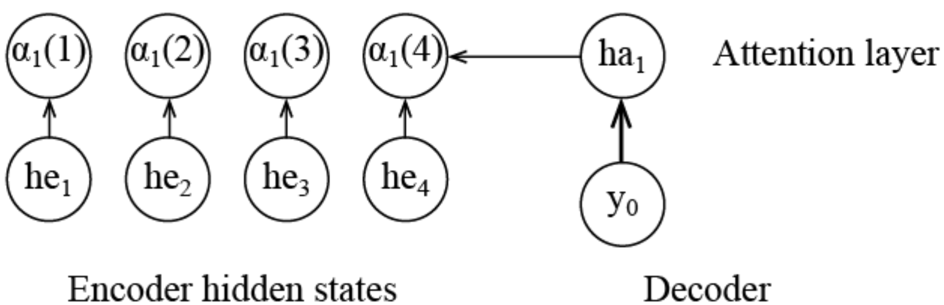
End-to-end model: Decoding with attention mechanism

alignment score

$e_i(j) = \text{score}(he_j, ha_i)$

alignment

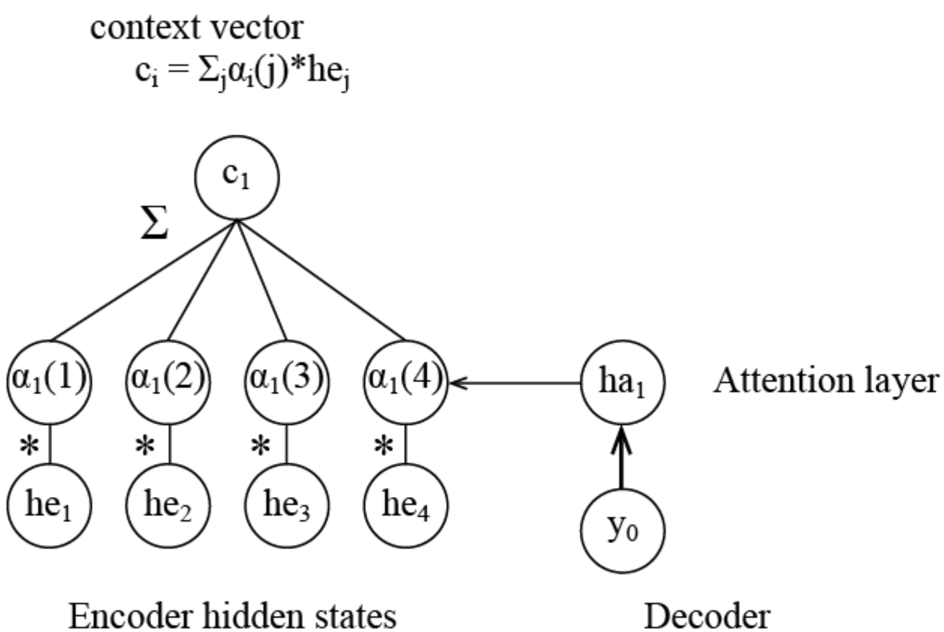
$\alpha_i(j) = \text{softmax}(e_i(j))$



Time step 1.

- Assign alignment probabilities to encoded inputs
- Alignment scores can be derived from attention layer's output and encoded inputs

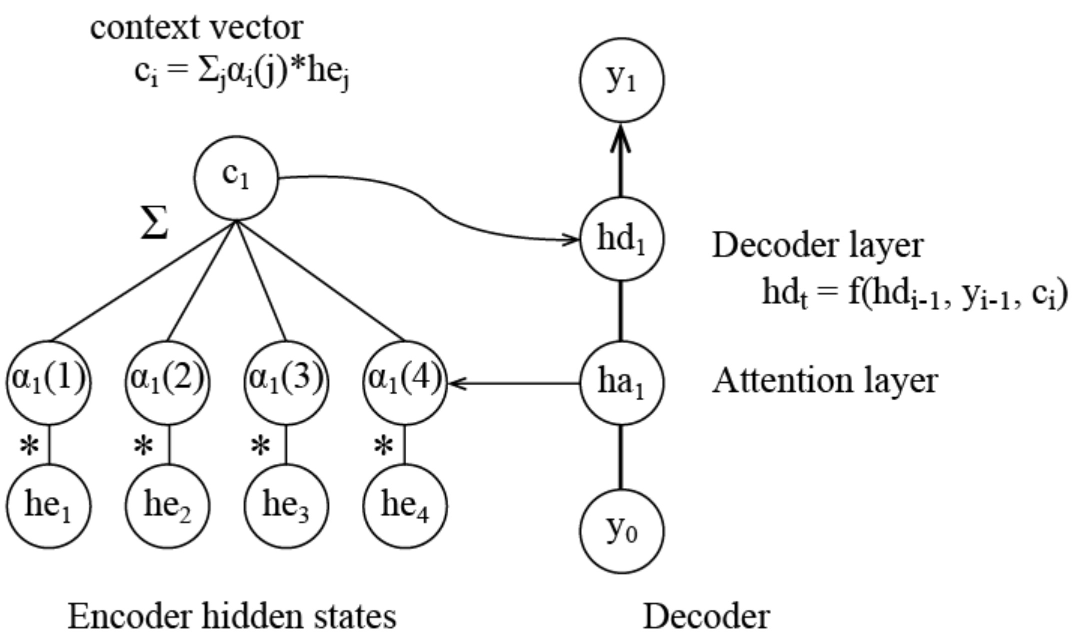
End-to-end model: Decoding with attention mechanism



Time step 1.

- Calculate context vector
- Context vector is the sum of encoded inputs weighted by alignment probabilities

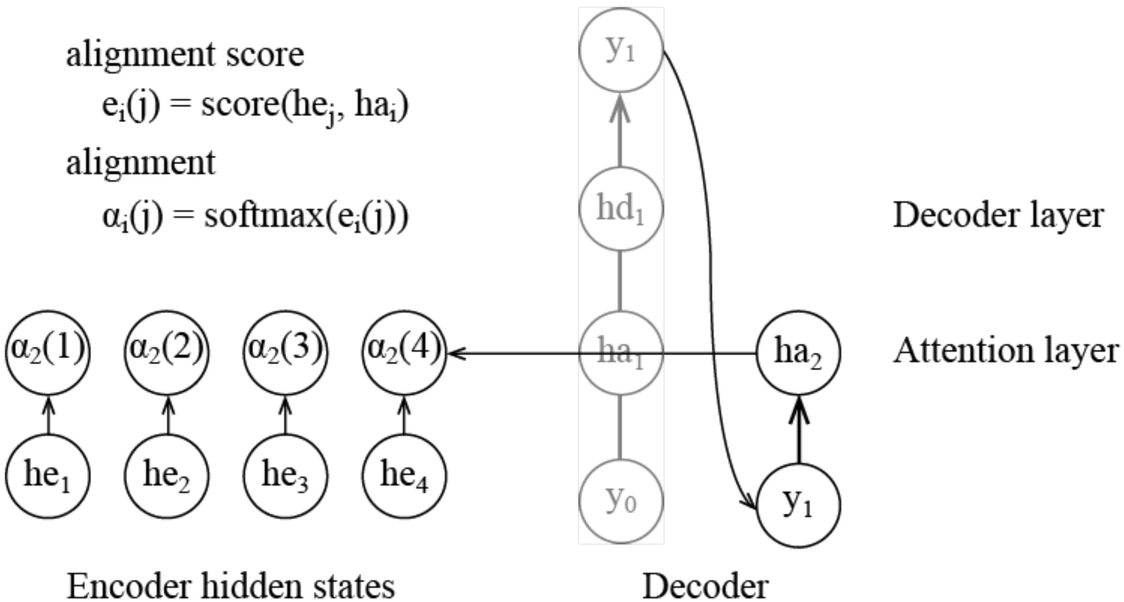
End-to-end model: Decoding with attention mechanism



Time step 1.

- Decoder layer predicts an output from the context vector

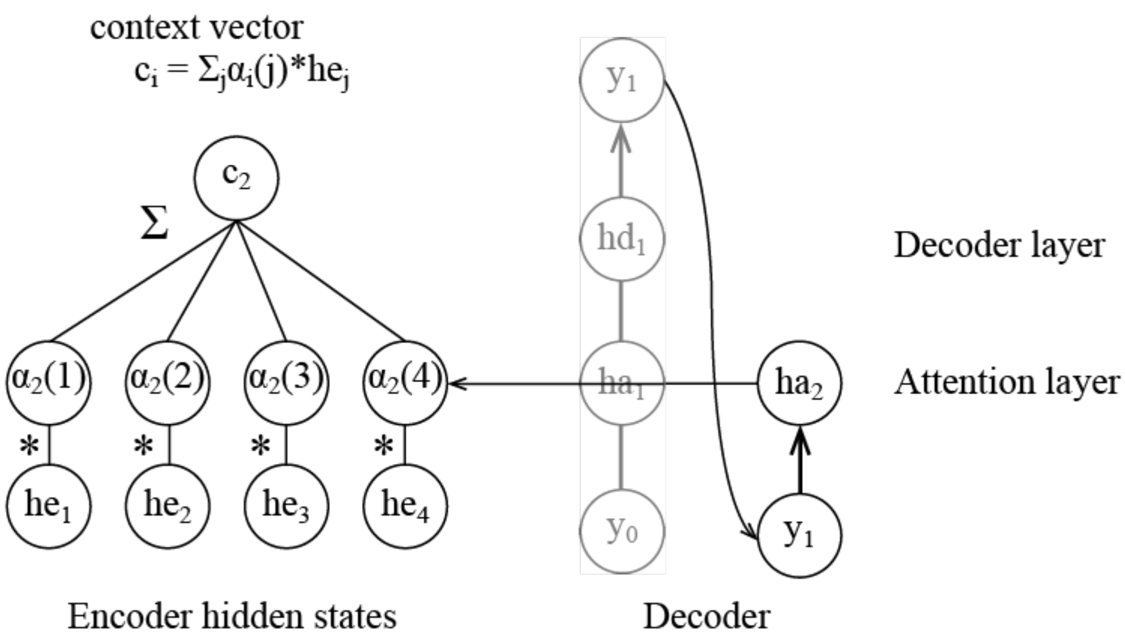
End-to-end model: Decoding with attention mechanism



Time step 2.

- The previous output is fed back to next time step
- Assign alignment probabilities to encoded inputs

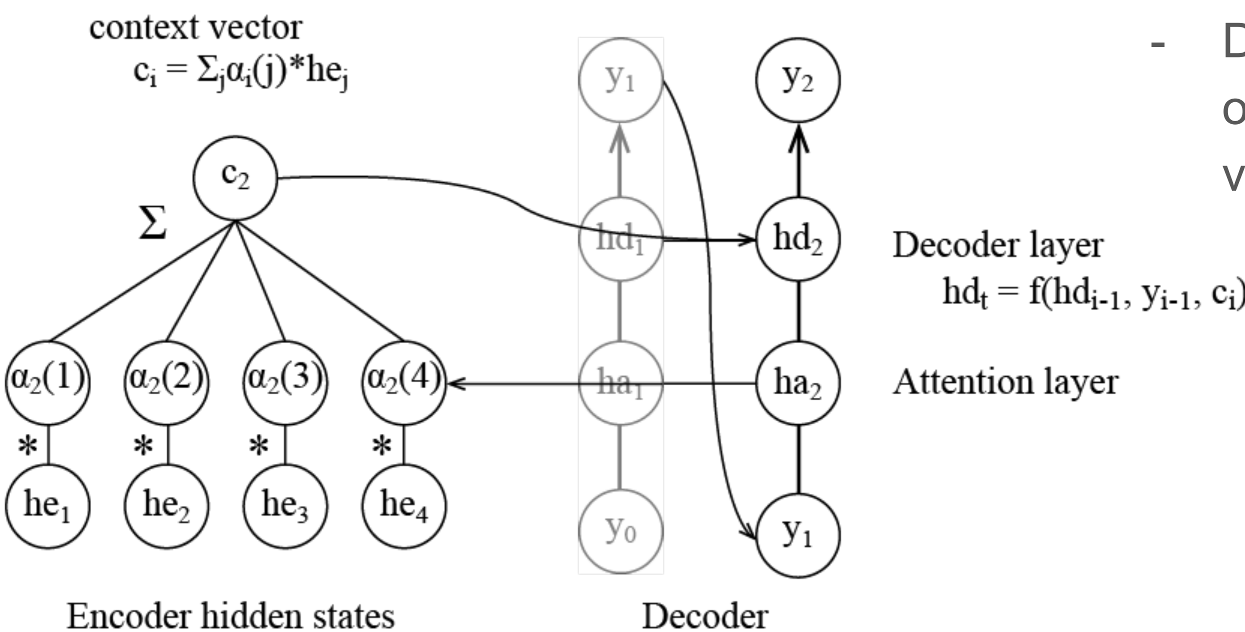
End-to-end model: Decoding with attention mechanism



Time step 2.

- Calculate context vector

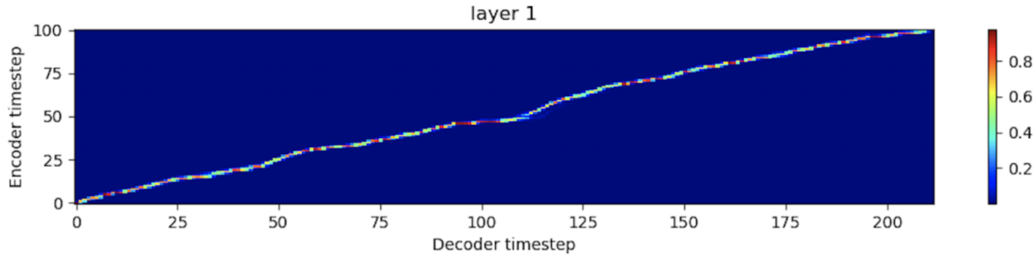
End-to-end model: Decoding with attention mechanism



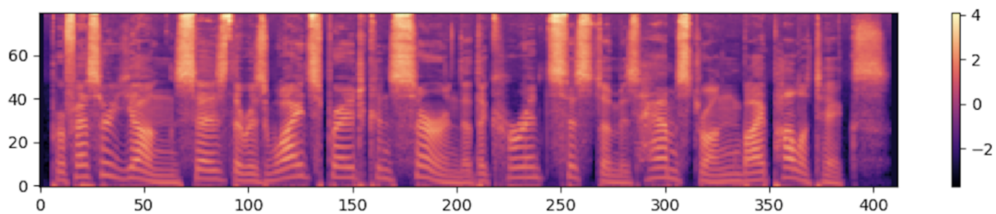
Time step 2.

- Decoder layer predict an output from the context vector

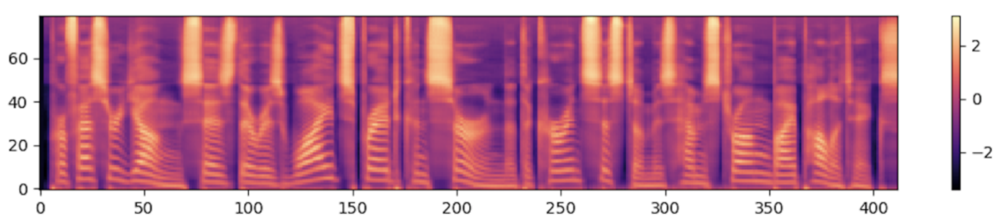
End-to-end model: Alignment visualization



alignment



ground truth mel spectrogram



predicted mel spectrogram

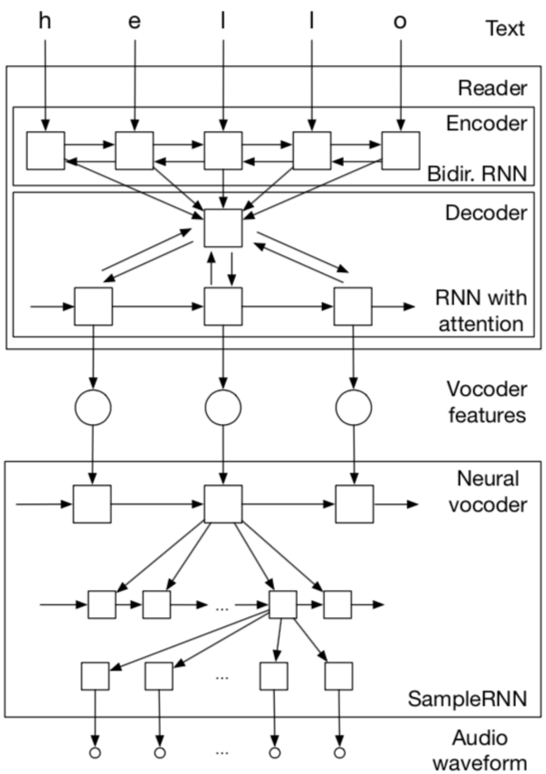
Example: Char2Wav

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, Yoshua Bengio:
Char2Wav: End-to-End Speech Synthesis. ICLR 2017

https://openreview.net/forum?id=B1VWyySKx

Char2Wav is one of the earliest work on end-to-end TTS. Its simple sequence-to-sequence architecture with attention gives a good proof-of-concept for end-to-end TTS as a start line. Char2Wav also uses advanced blocks like a neural vocoder and the multi-speaker embedding.

Char2Wav: Architecture



Encoder: bidirectional GRU RNN

Decoder: GRU RNN

Attention: GMM attention

Neural vocoder: SampleRNN

Figure is from Sotelo et al. (2017)

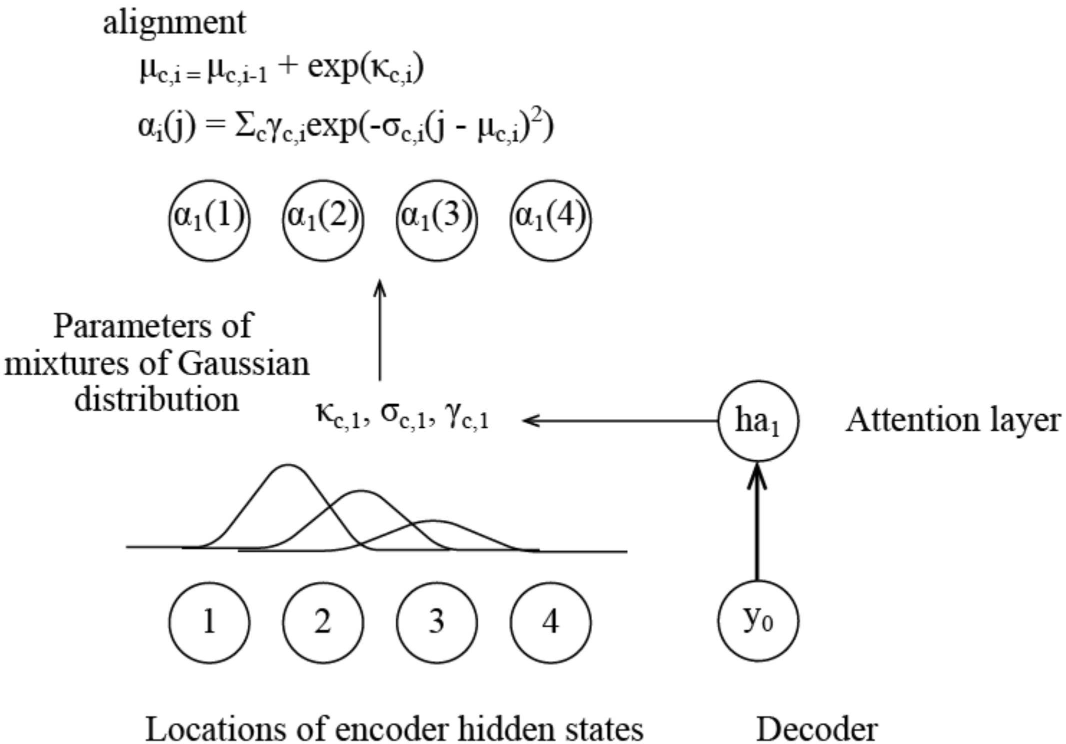
Char2Wav: Source and target choice

- Input: character or phoneme sequence

- Output: vocoder parameters

- Objective: mean square error or negative GMM log likelihood loss

Char2Wav: GMM attention



- Proposed by Graves (2013) [3]

- Location based attention

- Alignment is based on input location

- Alignment does not depend on input content

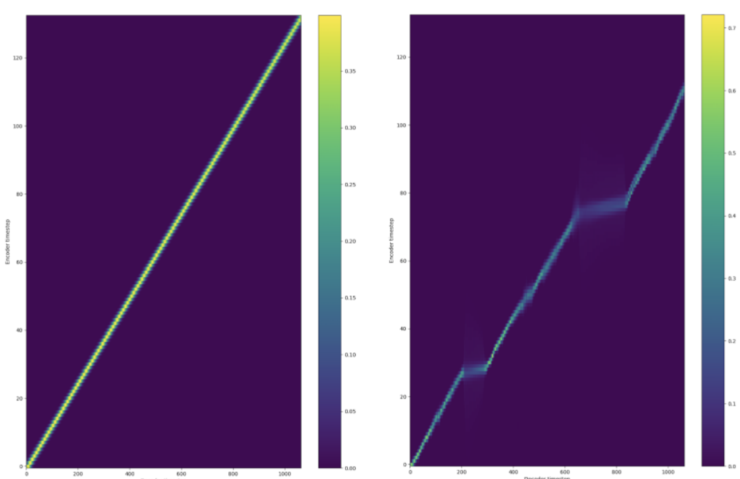
[3] Alex Graves: Generating Sequences With Recurrent Neural Networks. CoRR abs/1308.0850 (2013)

Char2Wav: GMM attention

$$\mu_{c,i} = \mu_{c,i-1} + \exp(\kappa_{c,i})$$

- Monotonic progress
  - The mean value μ always increases as time progresses
- Robust to long sequence prediction

epoch 0 epoch 130



Visualization of GMM attention alignment from Tacotron predicting vocoder parameters. r=2.

Char2Wav: Limitations

- Target features: Char2Wav uses vocoder parameters as a target. Vocoder parameters are challenging target for sequence-to-sequence system.
  - e.g. vocoder parameters for 10s speech requires 2000 iteration to predict
- Architecture: As Tacotron paper demonstrated, a vanila sequence-to-sequence architecture gives poor alignment.

These limitations were solved by Tacotron.

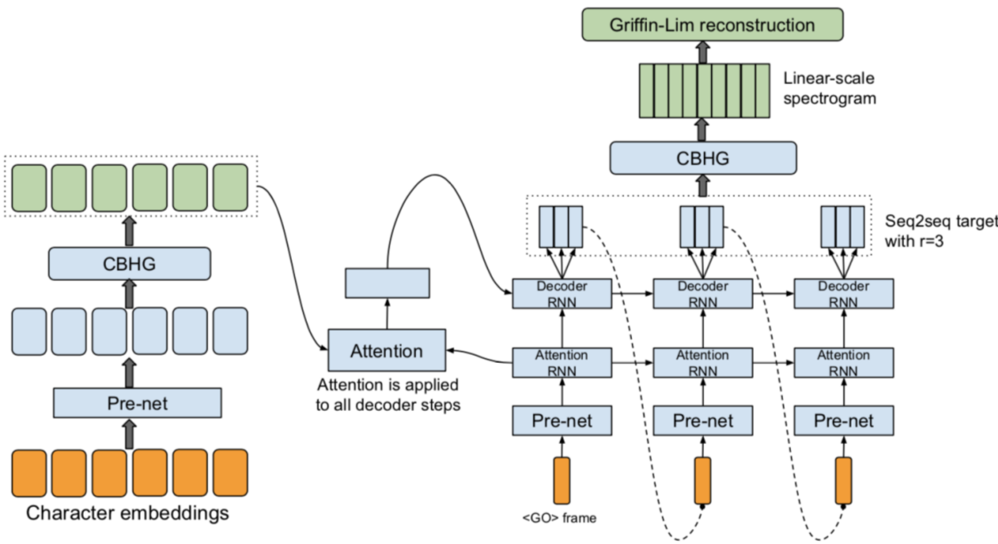
Example: Tacotron

Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous:
Tacotron: Towards End-to-End Speech Synthesis. INTERSPEECH 2017: 4006-4010

https://arxiv.org/abs/1703.10135

Tacotron is the most influential method among modern architectures because of several cutting-edge techniques.

Tacotron: Architecture



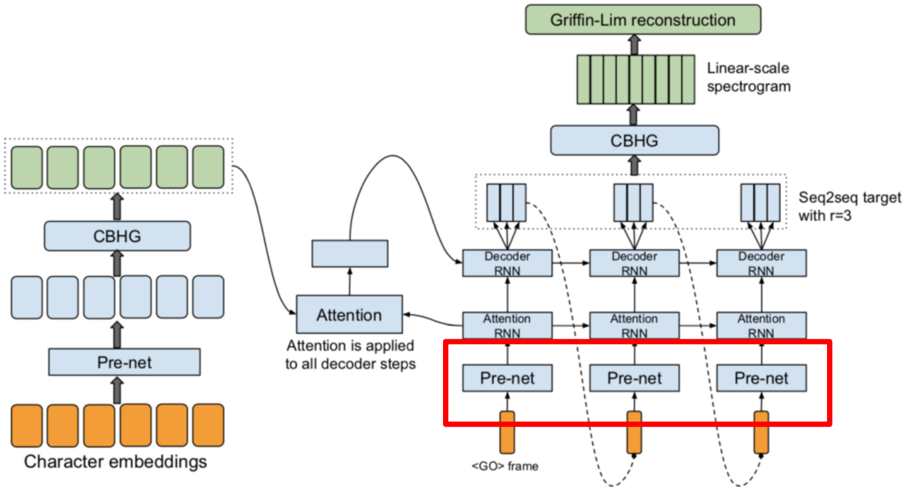
- CBHG encoder
- Encoder & decoder pre-net
- Reduction factor
- Post-net
- Additive attention

Figure is from Wang et al. (2017)

Tacotron: Source and target choice

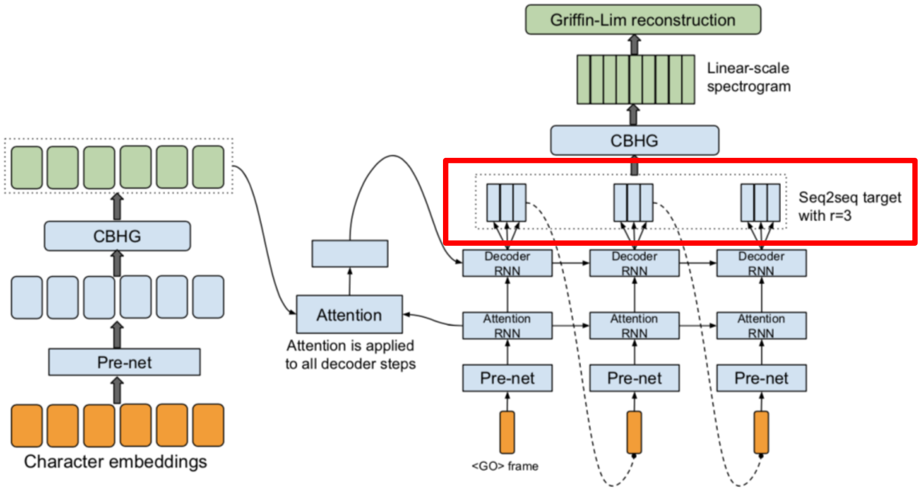
- Input: character sequence
- Output: mel and linear spectrogram (50ms frame length, 12.5ms frame shift)
  - **x2.5 shorter total sequence than that of vocoder parameters**
- Objective: L1 loss

Tacotron: Decoder pre-net



- The most important architecture advancement from vanila seq2seq
- 2 FFN + ReLU + dropout
- Crucial for alignment learning

Tacotron: Reduction factor



- Predicts multiple frames at a time
- Reduction factor r: the number of frames to predict at one step
- Large r →
  - Small number of iteration
  - Easy alignment learning
  - Less training time
  - Less inference time
  - Poor quality

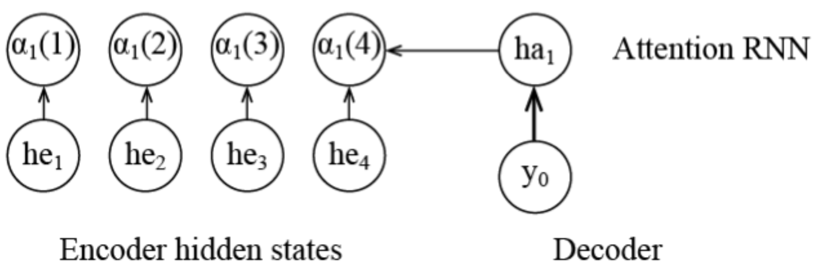
Tacotron: Additive attention

alignment score

$$e_i(j) = w^T \tanh(W\, ha_i + V\, he_j + b)$$

alignment

$$\alpha_i(j) = \mathrm{softmax}(e_i(j))$$

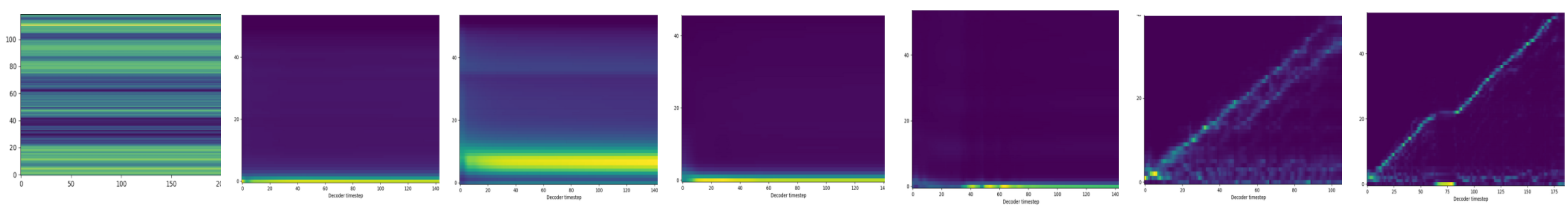


- Proposed by Bahdanau et al. (2014) [5]

- Content based attention

- Distance between source and target is learned by FFN

- No structure constraint



epoch 0 epoch 1 epoch 3 epoch 7 epoch 10 epoch 50 epoch 100

[5] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio: Neural Machine Translation by Jointly Learning to Align and Translate. CoRR abs/1409.0473 (2014)

Tacotron: Limitations

- Training time is too long: > 2M steps (mini batches) to converge
- Waveform generation: Griffin-Lim is handy but hurts the waveform quality.
- Stop condition: Tacotron predicts fixed-length spectrogram, which is inefficient both at training and inference time.

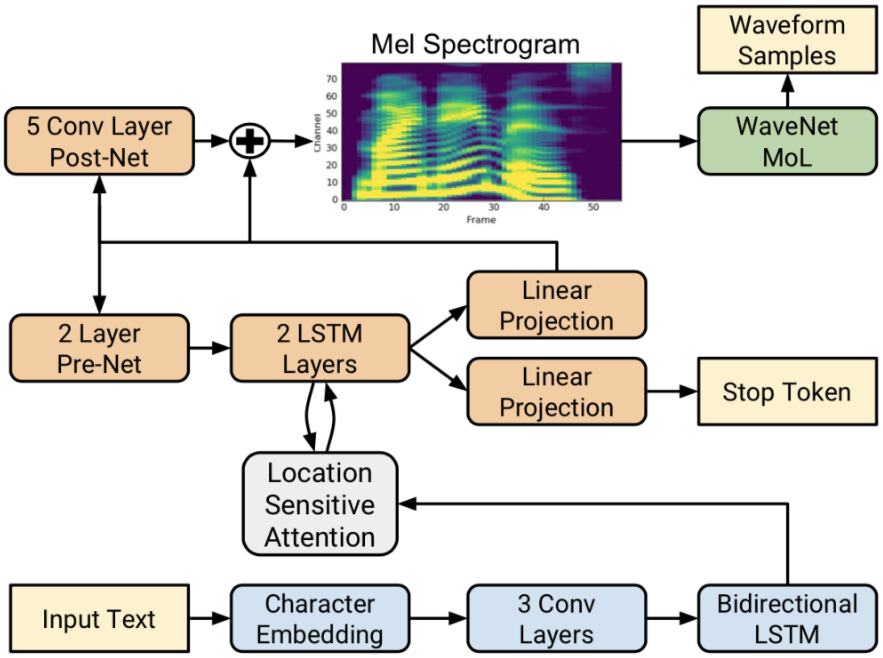
Example: Tacotron2

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu: Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. ICASSP 2018: 4779-4783

https://arxiv.org/abs/1712.05884

Tacotron2 is a surprising method that achieved human level quality of synthesized speech. Its architecture is an extension of Tacotron. Tacotron2 uses WaveNet for high-quality waveform generation.

Tacotron2: Architecture



- CNN layers instead of CBHG at encoder and post-net

- Stop token prediction

- Post-net improves predicted mel spectrogram

- Location sensitive attention

- No reduction factor

Figure is from Shen et al. (2018)

Tacotron2: Trends of architecture changes

1. Larger model size
- character embedding: 256 → 512
- encoder: 256 → 512
- decoder pre-net:
  (256, 128) → (256, 256)
- decoder: 256 → 1024
- attention: 256 → 128
- post-net: 256 → 512

2. More regularizations:
- Encoder applies dropout at every layer. The original Tacotron applies dropout in pre-net only at encoder.
- Zoneout regularization for encoder bidirectional[7] LSTM and decoder LSTM
- Post-net applies dropout at every layer
- L2 regularization

Large model needs heavy regularization to prevent overfitting

CBHG module is replaced by CNN layers at encoder and postnet probably due to limited regularization applicability.

[7] David Krueger, Tegan Maharaj, János Kramár, Mohammad Pezeshki, Nicolas Ballas, Nan Rosemary Ke, Anirudh Goyal, Yoshua Bengio, Hugo Larochelle, Aaron C. Courville, Chris Pal: Zoneout: Regularizing RNNs by Randomly Preserving Hidden Activations. CoRR abs/1606.01305 (2016)

Tacotron2: Source and target choice

- Input: character sequence
- Output: mel spectrogram
- Objective: mean square error

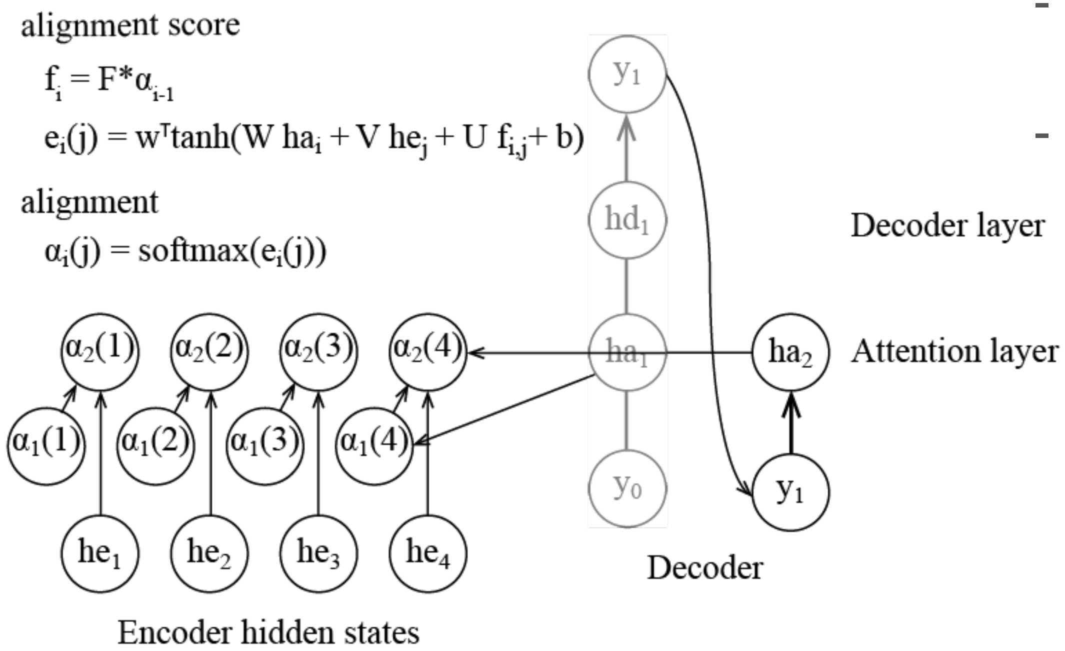
Tacotron2: Waveform synthesis with WaveNet

- Upsampling rate: 200
- Mel vs Linear: no difference
- Human-level quality was achieved by WaveNet trained with ground truth alignmed predicted spectrogram

System	MOS
Tacotron 2 (Linear + G-L)	3.944 ± 0.091
Tacotron 2 (Linear + WaveNet)	4.510 ± 0.054
Tacotron 2 (Mel + WaveNet)	$\mathbf{4.526 \pm 0.066}$

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

Tacotron2: Location sensitive attention



- Proposed by Chorowski et al., (2015)[8]
- Utilizes both input content and input location

[8] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, Yoshua Bengio: Attention-Based Models for Speech Recognition. NIPS 2015: 577-585

Tacotron2: Limitations

- WaveNet training in Tacotron2:
  - WaveNet is trained with ground truth-aligned **predicted** mel spectrogram.
  - Tacotron2's errors are corrected by WaveNet.
  - However, this is unproductive: one WaveNet for each different Tacotron2

- WaveNet training in normal condition
  - WaveNet is trained with ground truth spectrogram
  - However, its MOS score still is still inferior to the human speech.

Our work: Tacotron for Japanese language

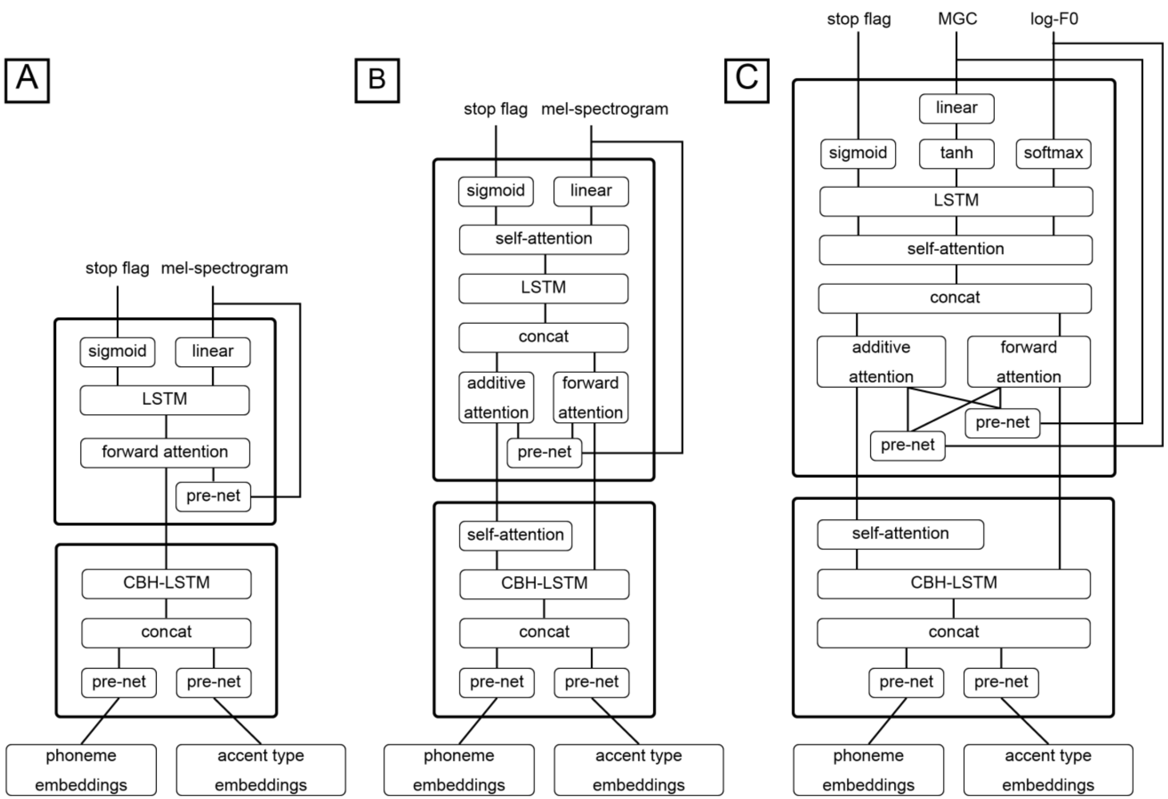
Yusuke Yasuda, Xin Wang, Shinji Takaki, Junichi Yamagishi:
Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language. CoRR abs/1810.11960 (2018)

https://arxiv.org/abs/1810.11960

We extended Tacotron for Japanese language.

- To model Japanese pitch accents, we introduced accentual type embedding as well as phoneme embedding
- To handle long term dependencies in accents: we use additional components

Japanese Tacotron: Architecture

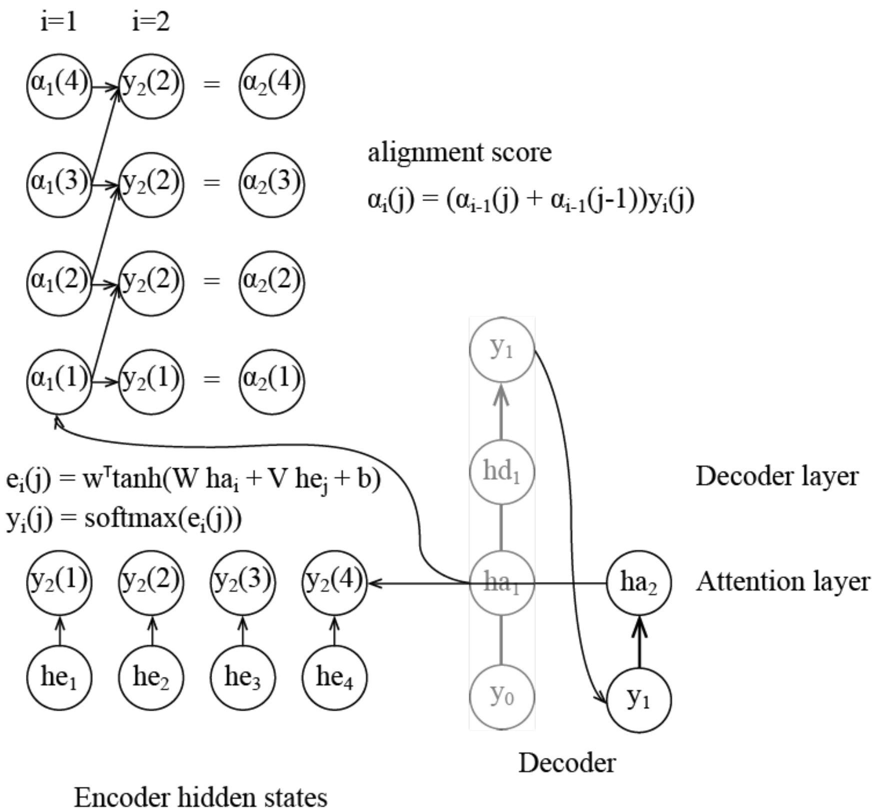


- Accentual type embedding
- Forward attention
- Self-attention
- Dual source attention
- Vocoder parameter target

Japanese Tacotron: Source and target choice

- Source: phoneme and accentual type sequence
- Target:
  - mel spectrogram
  - vocoder parameters
- Objective:
  - L1 loss (mel spectrogram, mel generalized cepstrum)
  - softmax cross entropy loss (discretized F0)

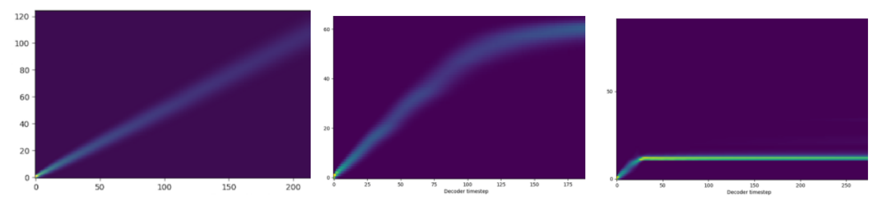
Japanese Tacotron: Forward attention



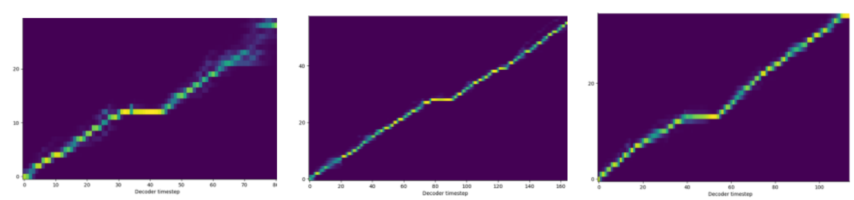
- Proposed by Zhang et al. (2018) [10]

- Precluding left-to-right progress

- Fast alignment learning



epoch 0 epoch 1 epoch 3

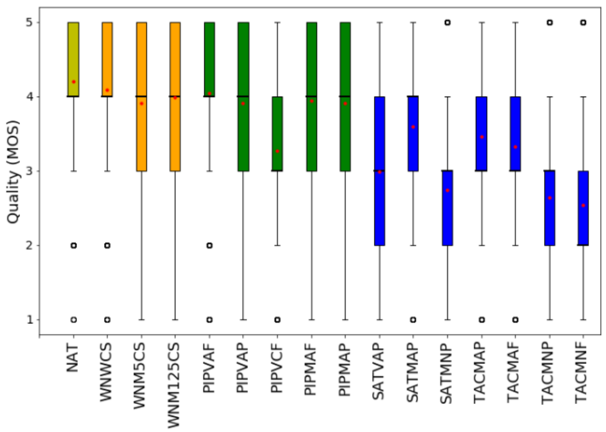


epoch 7 epoch 10 epoch 50

[10] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Rong Dai: Forward Attention in Sequence-To-Sequence Acoustic Modeling for Speech Synthesis. ICASSP 2018: 4789-4793

Japanese Tacotron: Limitations

- Speech quality is still worse than that of traditional pipeline system
  - Input feature limitation
  - Configuration is based on Tacotron, not Tacotron2



Summary

	Char2Wav	**Tacotron**	**VoiceLoop [11]**	**DeepVoice3 [12]**	**Tacotron 2**	**Transformer [13]**	**Japanese Tacotron**
network type	RNN	RNN	memory buffer	CNN	RNN	self-attention	RNN
input	character/ phoneme	character	phoneme	character/ phoneme	character	phoneme	phoneme
seq2seq output	vocoder	mel	vocoder	mel	mel	mel	mel/ vocoder
post-net output	-	linear	-	linear/ vocoder	mel	mel	-
attention mechanism	GMM	additive	GMM	dot-product	location sensitive	dot-product	forward
waveform synthesis	SampleRNN	Griffin-Lim	WORLD	Griffin-Lim/ WORLD/ WaveNet	WaveNet	WaveNet	WaveNet

Japanese Tacotron: Implementation

URL: https://github.com/nii-yamagishilab/self-attention-tacotron

Framework: Tensorflow, Python

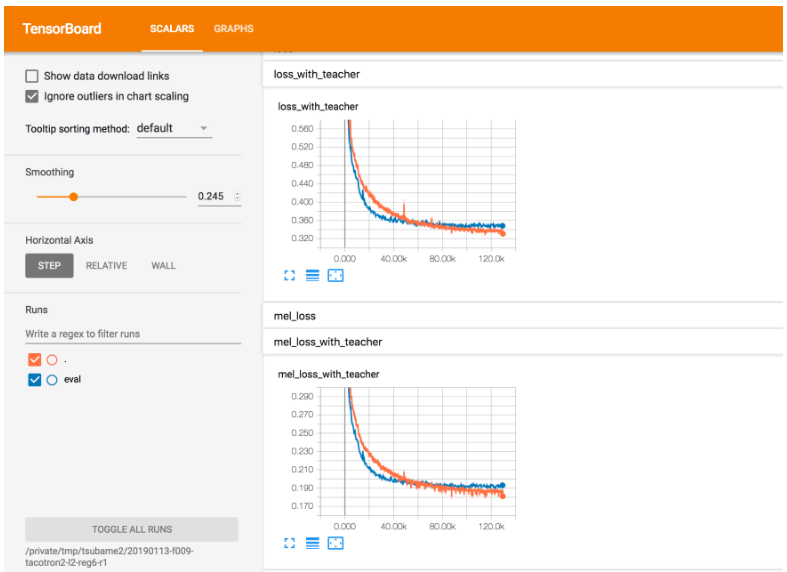
Supported datasets: LJSpeech, VCTK, (...coming more in the future)

Features:

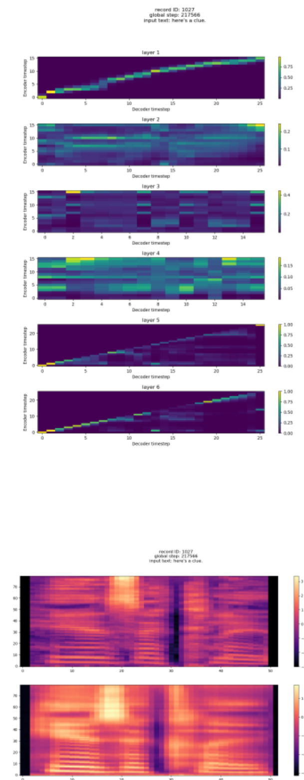
- Tacotron, Tacotron2, Japanese Tacotron model
- Combination choices for encoder, decoder, attention mechanism
- Force alignment
- Mel spectrogram, vocoder parameter for target
- Compatible with our WaveNet implementation

Japanese Tacotron: Experimental workflow

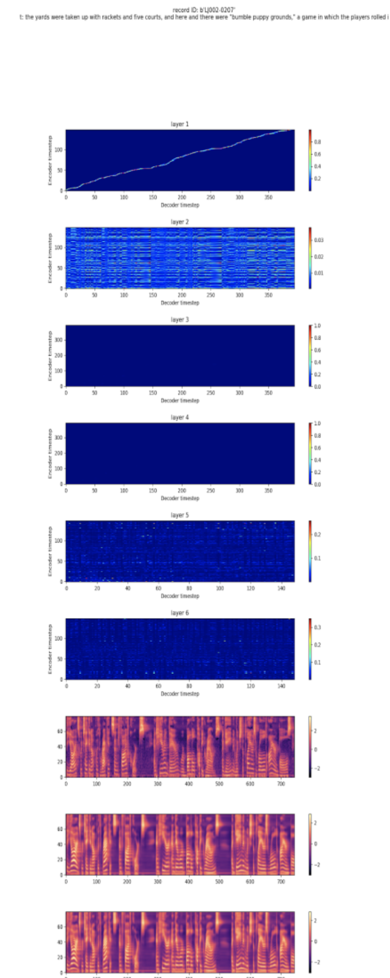
training & validation metrics



validation result



test result



Japanese Tacotron: Audio samples

NAT TACMAP TACMNP SATMAP SATMNP SATWAP

Bibliography

- [1] Wei Ping, Kainan Peng, Jitong Chen: ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech. CoRR abs/1807.07281 (2018)
- [2] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, Yoshua Bengio: Char2Wav: End-to-End Speech Synthesis. ICLR 2017
- [3] Alex Graves: Generating Sequences With Recurrent Neural Networks. CoRR abs/1308.0850 (2013)
- [4] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous: Tacotron: Towards End-to-End Speech Synthesis. INTERSPEECH 2017: 4006-4010
- [5] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio: Neural Machine Translation by Jointly Learning to Align and Translate. CoRR abs/1409.0473 (2014)

Bibliography

- [6] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yonghui Wu: Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. ICASSP 2018: 4779-4783
- [7] David Krueger, Tegan Maharaj, János Kramár, Mohammad Pezeshki, Nicolas Ballas, Nan Rosemary Ke, Anirudh Goyal, Yoshua Bengio, Hugo Larochelle, Aaron C. Courville, Chris Pal: Zoneout: Regularizing RNNs by Randomly Preserving Hidden Activations. CoRR abs/1606.01305 (2016)
- [8] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, Yoshua Bengio: Attention-Based Models for Speech Recognition. NIPS 2015: 577-585
- [9] Yusuke Yasuda, Xin Wang, Shinji Takaki, Junichi Yamagishi: Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language. CoRR abs/1810.11960 (2018)

Bibliography

- [10] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Rong Dai: Forward Attention in Sequence-To-Sequence Acoustic Modeling for Speech Synthesis. ICASSP 2018: 4789-4793
- [11] Yaniv Taigman, Lior Wolf, Adam Polyak, Eliya Nachmani: VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop. ICLR 2018
- [12] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, John Miller: Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. ICLR 2018
- [13] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, Ming Zhou: Close to Human Quality TTS with Transformer. CoRR abs/1809.08895 (2018)

Acknowledgement

This work was partially supported by JST CREST Grant Number JPMJCR18A6, Japan and by MEXT KAKENHI Grant Numbers (16H06302, 17H04687, 18H04120, 18H04112, 18KT0051), Japan.