Tacotron1, 2

Pattern Recognition & Machine Learning Laboratory
Hyeon-Woo Bae

Aug. 11, 2021

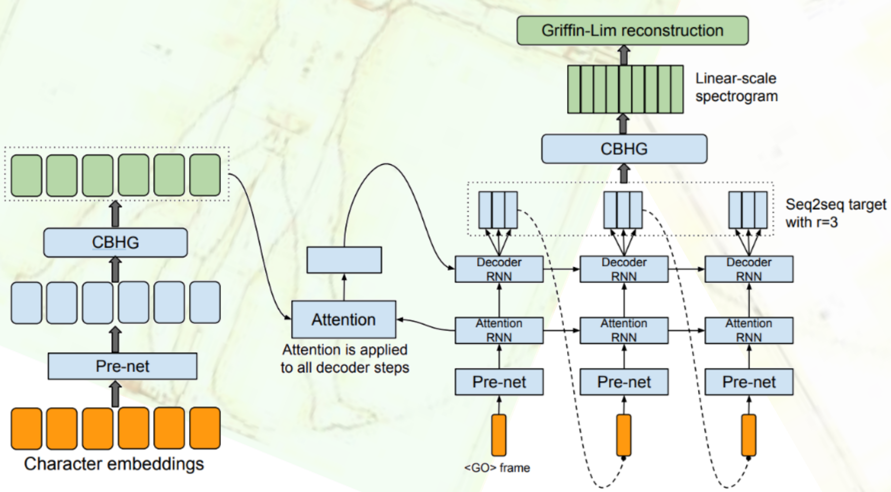
Tacotron: Towards End-to-End Speech Synthesis [*Y. Wang et al.*, 2017] (1/5)

- **Goal**
  - ➢ Presenting end-to-end text-to-speech (TTS) system that can be trained on <text, audio> pairs
- **Motivation**
  - ➢ Complexity of modern text-to-speech (TTS) designs
    - Previous system consists of multiple stages, such as a text analysis frontend, an acoustic model and an audio synthesis module
- **Contribution**
  - ➢ Alleviates the need for laborious feature engineering
  - ➢ Easily allows for rich conditioning on various attributes, such as speaker or language, or high-level features like sentiment
    - Conditioning can occur at beginning of the model
  - ➢ Single model is likely to be more robust than a multi-stage model
    - Error of multi-stage model can compound

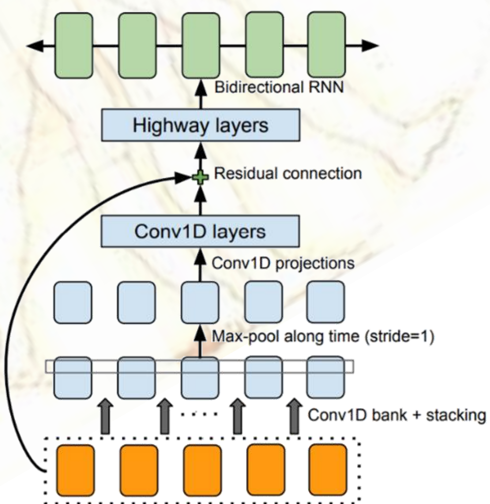
Y. Wang et al., "Tacotron: Towards End-to-End Speech Synthesis," in Interspeech, 2017

Tacotron: Towards End-to-End Speech Synthesis [Y. Wang et al., 2017] (2/5)

- **Model Architecture**
  - CBHG (1-D Convolutional Bank + Highway network + bidirectional GRU) Module
    - Bank of 1-D convolutional filters
      - Convolving input sequence
    - Highway networks
      - Extract high-level features
    - Bidirectional gated recurrent unit (GRU)
      - Extract sequential features from both forward and backward context

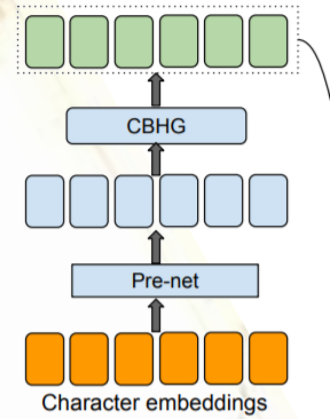


Architecture of Tacotron

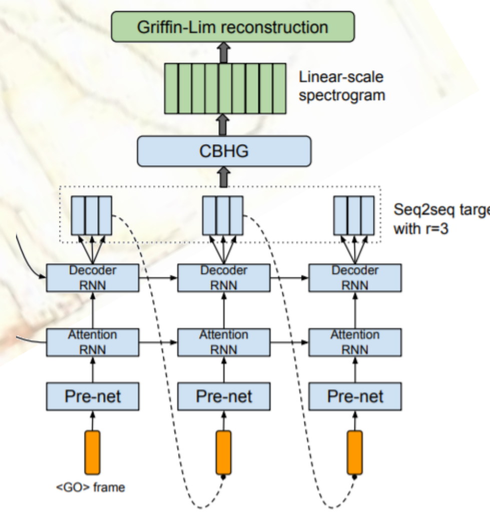


Architecture of CBHG Module

- **Model Architecture (Cont.)**
  - **Encoder**
    - **Extract robust sequential representations of text**
    - **Use bottleneck layer with dropout as the pre-net**
      - **Help convergence and improve generalization**
    - **CBHG reduce overfitting, make fewer mispronunciations**
  - **Decoder**
    - **Concatenate context vector and attention RNN output**
    - **Predicting r frames at once**
      - **Reduce model size, training time**
      - **Increase convergence speed**
      - **Much faster and stable alignment**
    - **Use seq2seq target as mel spectrogram**
  - **Post-processing net**
    - **Convert mel spectrogram to linear spectrogram**
    - **Can see full decoded sequence**
    - **Use a CBHG module**
  - **Griffin-Lim algorithm**
    - **Synthesize linear spectrogram to waveform**
    - **Fast and Simple**



Architecture of Encoder

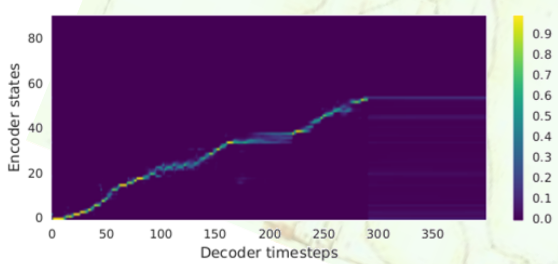


Architecture of Decoder

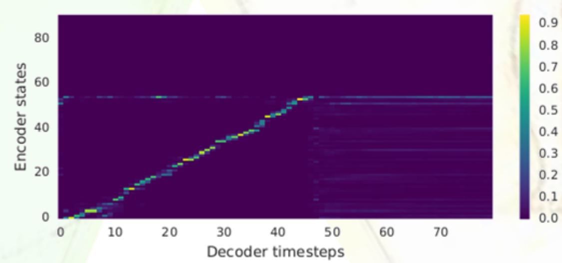
Y. Wang et al., "Tacotron: Towards End-to-End Speech Synthesis," in Interspeech, 2017

Tacotron: Towards End-to-End Speech Synthesis [*Y. Wang et al.*, 2017] (4/5)

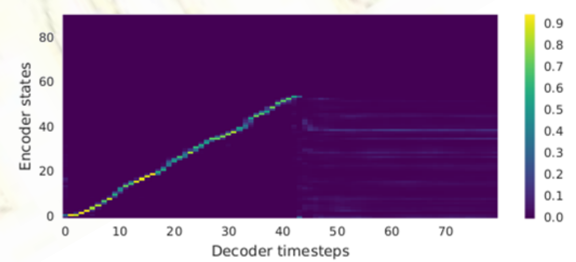
- **Experiment**
  - **Training data set**
    - **Internal North American English dataset, contains about 24.6 hours of speech data spoken by a professional female speaker**
  - **Ablation analysis**
    - **No pre-net and post-processing net**
      - **Attention tends to get stuck for many frames before moving forward**
      - **Bad speech intelligibility**
    - **CBHG encoder replaced by GRU encoder**
      - **GRU encoder is noisier**
      - **Noisy alignment leads to mispronunciations**
      - **CBHG encoder reduces overfitting and generalizes well**



Attention alignment of Vanilla seq2seq+scheduled sampling



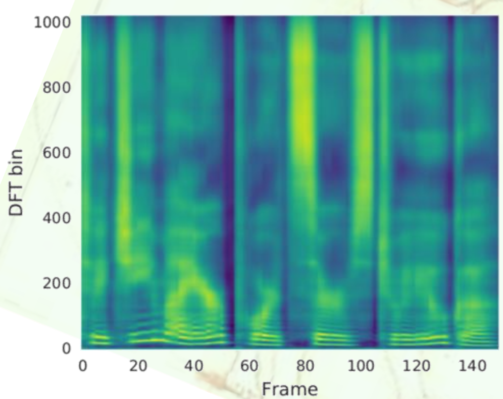
Attention alignment of GRU encoder



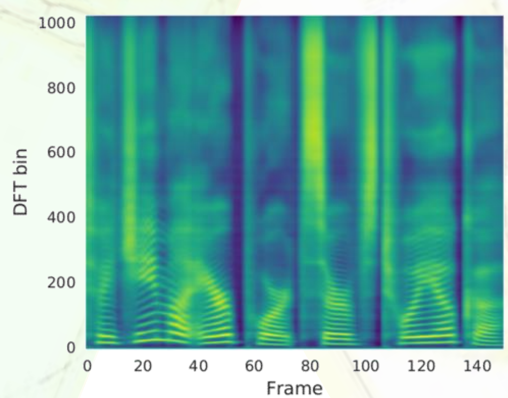
Attention alignment of Tacotron

Tacotron: Towards End-to-End Speech Synthesis [*Y. Wang et al.*, 2017] (5/5)

- **Result**
  - **Benefit of using post-processing net**
    - **Prediction from the post-processing net contains better resolved harmonics and high frequency formant structure**
  - **Mean opinion score test**
    - **Asked to rate the naturalness of the stimuli in a 5-point Likert scale score**
    - **Tacotron achieves an MOS of 3.82**
    - **Tacotron outperforms the parametric system**



Spectrogram without post-processing net



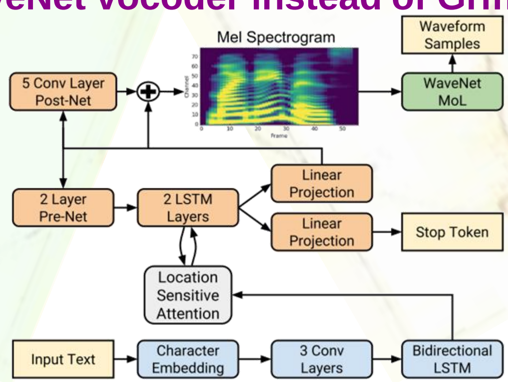
Spectrogram with post-processing net

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

5-scale mean opinion score evaluation

Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions [*J. Shen et al.*, 2018] (1/4)

- **Goal**
  - Presenting end-to-end text-to-speech (TTS) system that can be trained on <text, audio> pairs
- **Motivation**
  - Complexity of modern text-to-speech (TTS) designs
    - Previous system consists of multiple stages, such as a text analysis frontend, an acoustic model and an audio synthesis module
- **Contribution**
  - Modify architecture of Tacotron 1
    - Using LSTM instead of CBHG module
    - Using location-sensitive attention instead of additive attention mechanism
    - Using modified WaveNet vocoder instead of Griffin-Lim algorithm



Architecture of Tacotron 2

Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions [*J. Shen et al.*, 2018] (2/4)

- Model Architecture
  - Encoder
    - Convert character sequence into a hidden feature representation
    - Input characters are represented using character embedding
    - Output of final convolutional layer is passed into a single bidirectional LSTM
  - Attention
    - Use location-sensitive attention
      - Add attention weights from previous decoder time steps to additive attention mechanism
      - Reduce potential failure mode where some subsequences are repeated or ignored by decoder
  - Decoder
    - Predict a mel spectrogram one frame at a time from the encoded input sequence
    - Concatenation of LSTM output and the context vector is projected through a linear transform to predict the target spectrogram frame
    - Stop token prediction is used to determine when to terminate generation
  - WaveNet vocoder
    - Modified version of the WaveNet architecture
      - Can invert a mel spectrogram instead of linguistic condition

Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions [*J. Shen et al.*, 2018] (3/4)

- **Experiment**
  - **Dataset**
    - Internal US English dataset, which contains 24.6 hours of speech from a single professional female speaker
  - **Evaluation**
    - Tacotron 2 significantly outperforms all other TTS systems and results in a Mean Opinion Score (MOS) comparable to that of the ground truth audio
  - **Ablation studies**
    - Predicted features versus Ground truth
    - Linear spectrograms
    - Post processing network
      - Without post-net : 4.429 ± 0.071
      - With post-net : 4.526 ± 0.066
    - Simplifying WaveNet

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

Comparison of evaluated MOS for WaveNet

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

Comparison of evaluated MOS for Griffin-Lim vs. WaveNet

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)	$\mathbf{4.526 \pm 0.066}$

MOS evaluations of TTS systems

Total layers	Num cycles	Dilation cycle size	Receptive field (samples / ms)	MOS
30	3	10	6,139 / 255.8	4.526 ± 0.066
24	4	6	505 / 21.0	4.547 ± 0.056
12	2	6	253 / 10.5	4.481 ± 0.059
30	30	1	61 / 2.5	3.930 ± 0.076

WaveNet with various layer and receptive field sizes