Tacotron: End-to-end speech synthesizer

Mahshid Alinoori

Table of Contents

Introduction

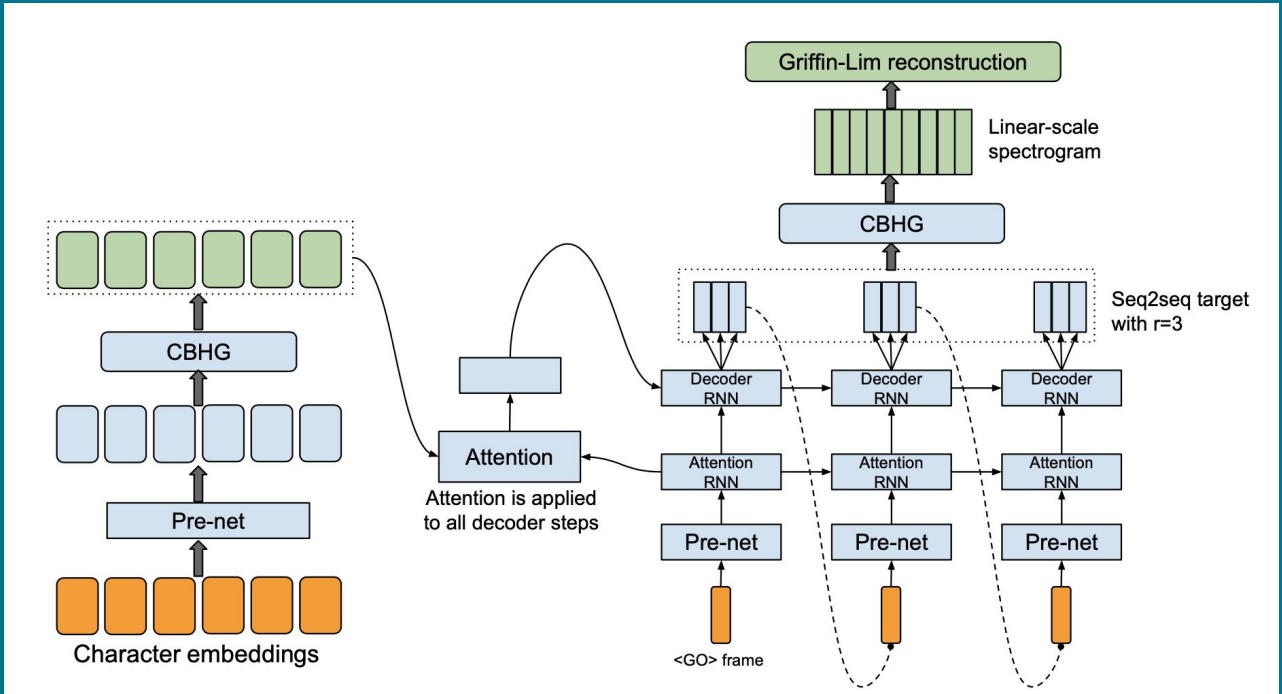
TTS pipelines consist of **various models** for capturing linguistic features, acoustic features, and synthesizing which are **trained independently**.

They suffer from drawbacks such as **compounded errors, feature engineering requirements,** and more difficult application of **conditioning.**

End-to-end systems makes all these steps much easier, they are more **robust** to the error and can be more **adaptive** to new data.

Tacotron is introduced as a fully end-to-end system to overcome the limitations of previous models and does not need any pre-training and is **trained from scratch**.

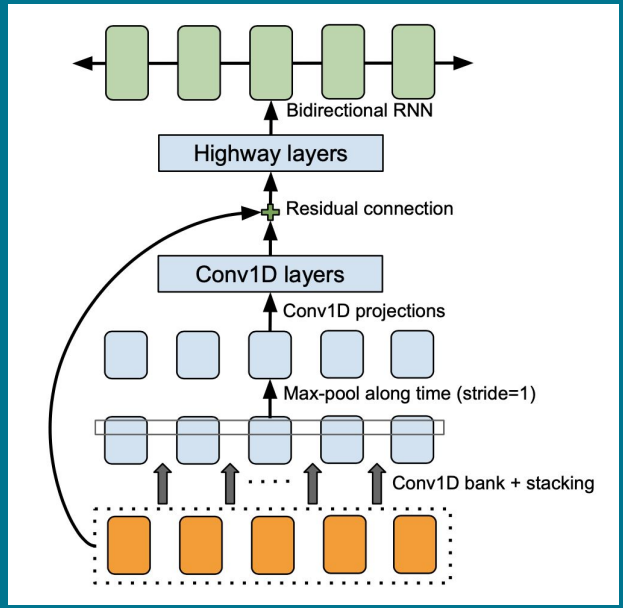
Tacotron Architecture



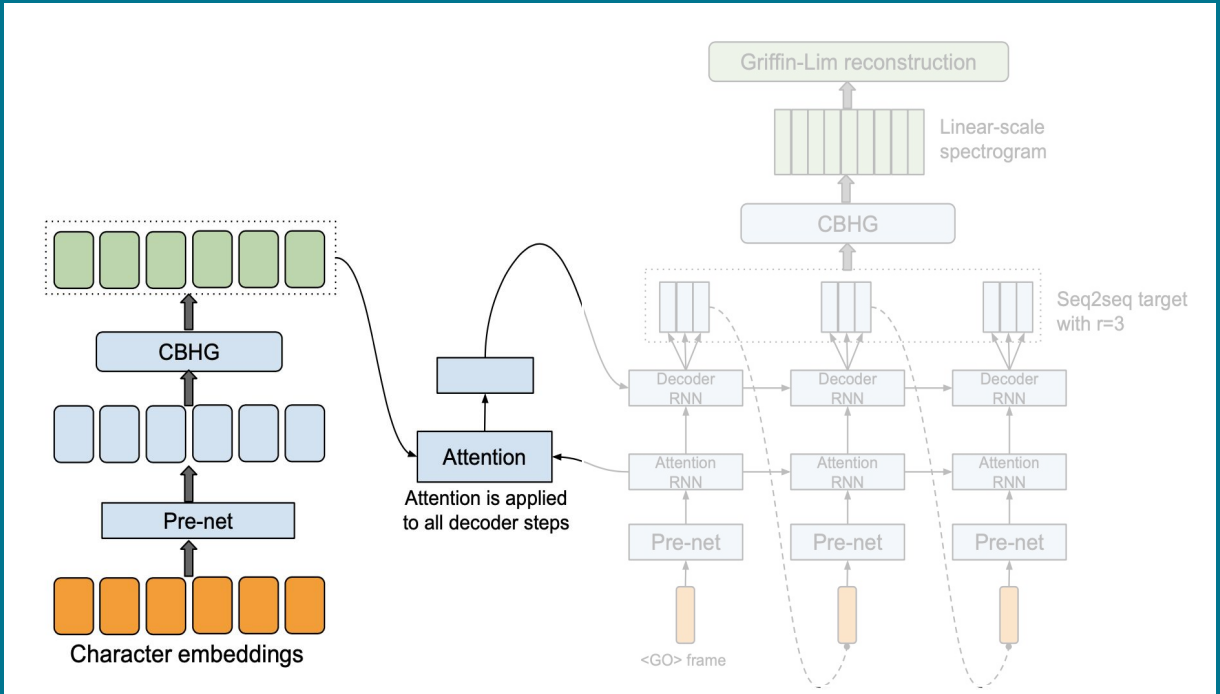
Model: seq2seq with attention
Consists of: an encoder, an attention-based decoder, and a post-processing net
Input: characters
Outputs: spectrogram frames

CBHG Module

Used for: extracting presentation from sequences.
1D convolutional banks: K sets of 1-D filters and model local and contextual information similar to k-grams
Highway layer: multi-layer highway networks to extract high-level features.
Bidirectional GRU RNN: for extracting sequential features

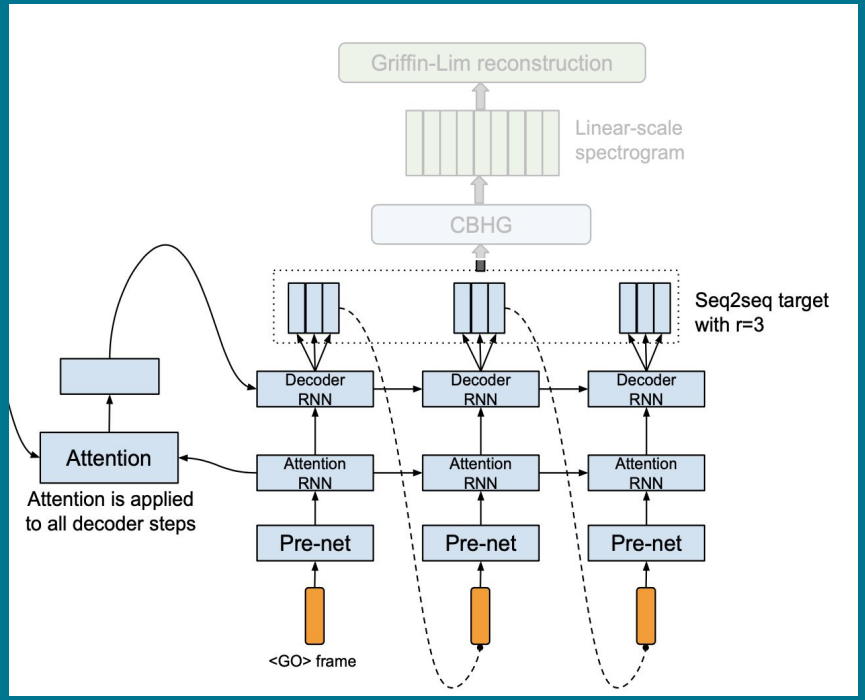


Encoder



Used for: extracting sequential representation of text
Input: character sequences represented as 1-hot vectors and embedded
Pre-net: used as non-linear transformation by a bottleneck layer with dropouts
CBHG: generates final encoder output

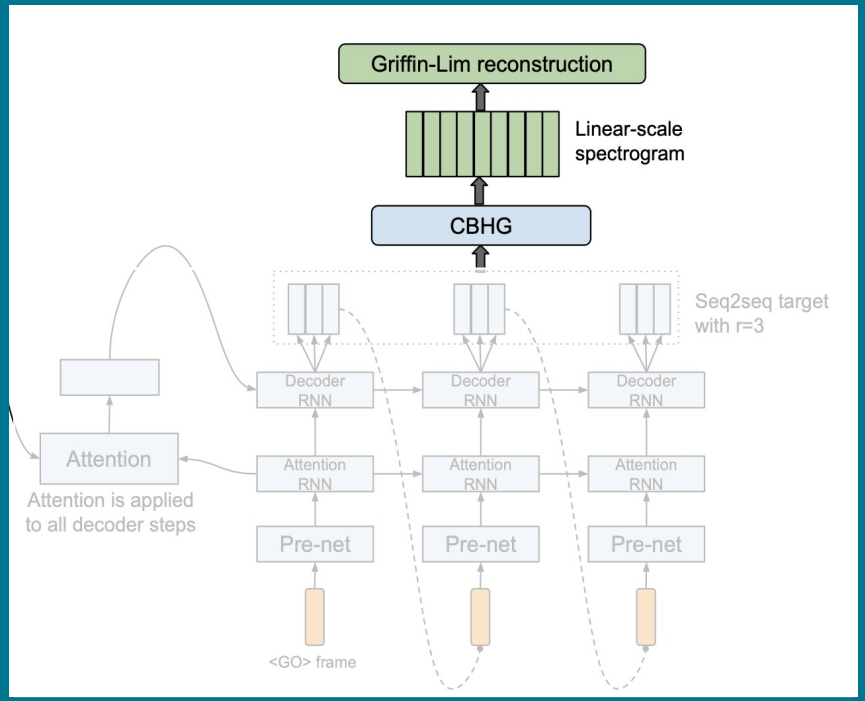
Decoder



Model: content-based tanh attention decoder
Implemented as: GRUs with residual connections

Input: concatenation of context vectors and attention RNN outputs
Output: 80-band mel-scale spectrogram
FC layer: The output layer of the decoder to predict multiple non-overlapping frames
Pre-net: non-linear transformation with drop-out

Post-processing



Used for: converting the seq2seq to a target that is finally synthesized into waveform and correcting prediction error for the frames

Implemented as: a simple version of CBHG
Output: spectral magnitude sampled on a linearly-scaled frequency
Griffin-Lim Algorithm: used for generating the waveform from the spectrogram

Experiments

Dataset

North American English dataset with 24.6 hours of speech

Ablation Studies

4 studies using vanilla seq2seq model, a GRU encoder, removing post-processing net, and the complete version

Results

Vanilla model: poor aligning
GRU encoder: noisier alignment
No post-processing: more synthesis artifacts

MOS Tests

Crowdsourced by native speakers and compared to state-of-the-art is placed in-between with a 3.82 MOS

Conclusion

Tacotron is an end-to-end TTS model.

Takes character and outputs raw spectrogram.

Consists of an encoder, an attention-based decoder, and a post-processing network.

Benefit of better performance and tricks for speeding up training and inference.

Thanks!