TACOTRON

Николаева Софья
НИУ ВШЭ
24 апреля 2020

INTRODUCTION

TACOTRON, sequence-to-sequence generative text-to-speech (TTS) model that synthesizes speech directly from characters.

TACOTRON achieves a 3.82 subjective 5-scale mean opinion score on US English.

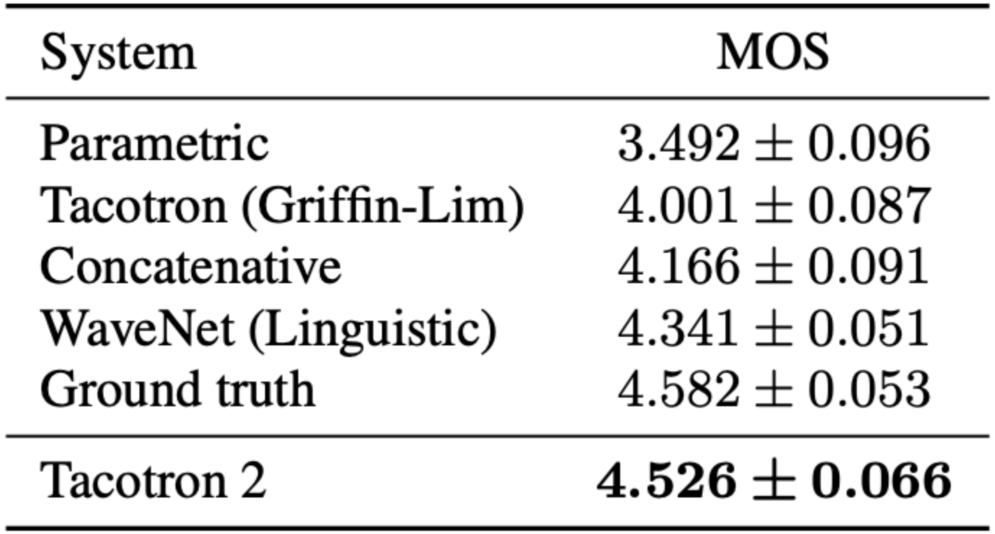
TACOTRON2 achieves a mean opinion score (MOS) of 4.53 comparable to a MOS of 4.58 for professionally recorded speech.

BEFORE TACOTRON

Concatenative synthesis with unit selection – the process of stitching small units of pre-recorded waveforms together.

Parametric speech synthesis generates smooth trajectories of speech features to be synthesized by a vocoder.

WaveNet – a generative model of time domain waveforms, produces audio quality that begins to rival that of real human speech.

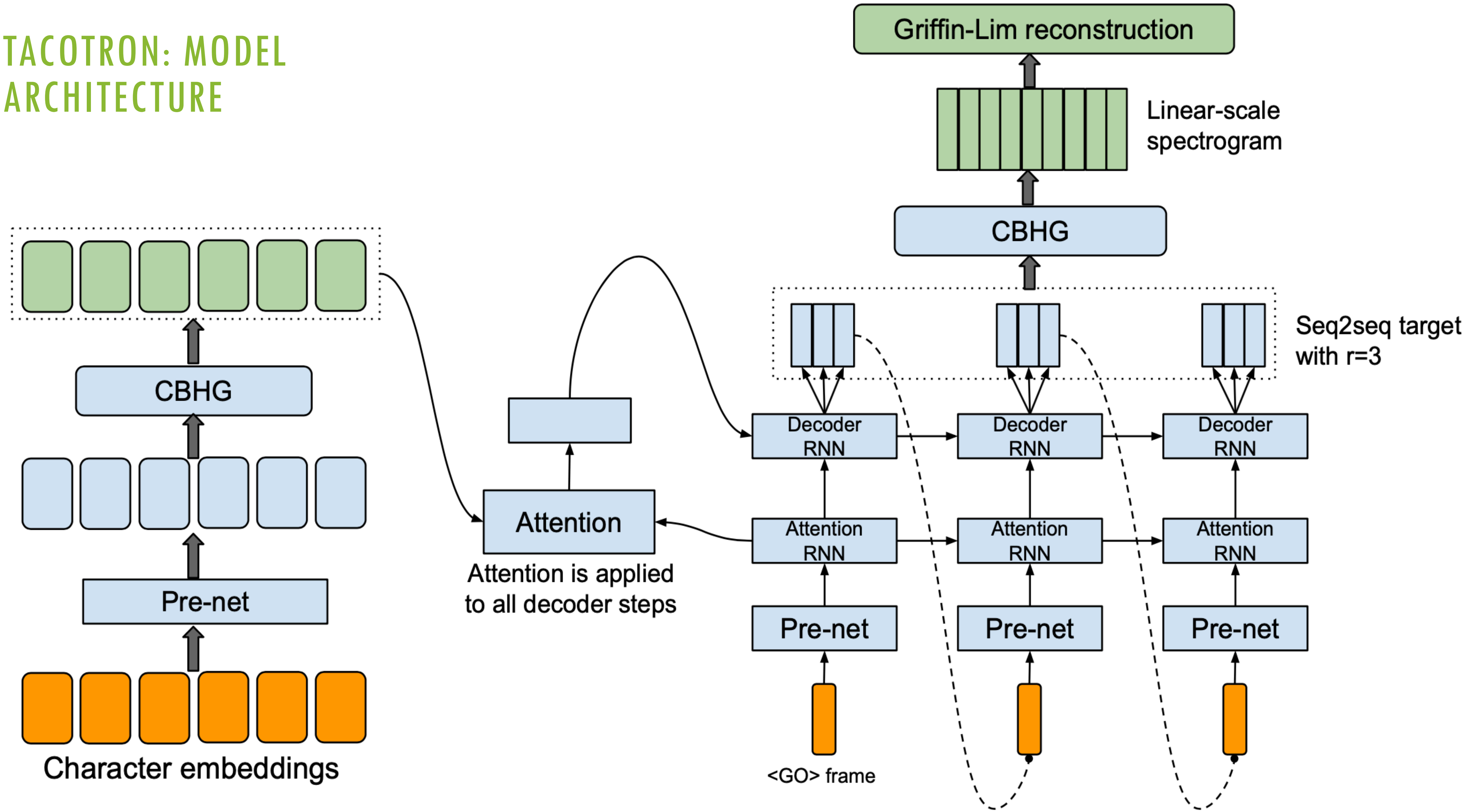


System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2	**4.526 ± 0.066**

TACOTRON (2017)

Tacotron, a sequence-to-sequence architecture for producing magnitude spectrograms from a sequence of characters, simplifies the traditional speech synthesis pipeline by replacing the production of these linguistic and acoustic features with a single neural network trained from data alone.

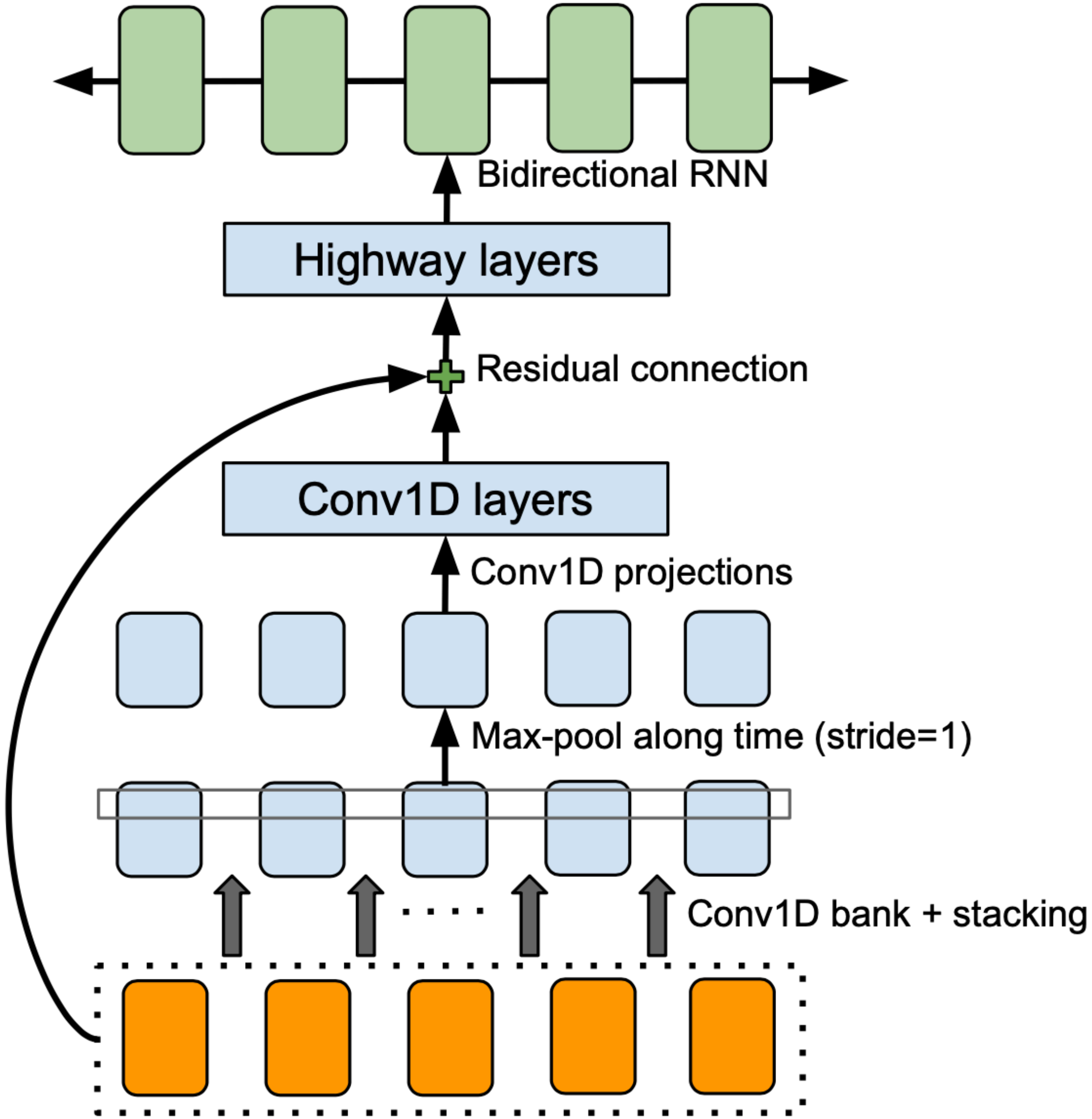
TACOTRON: MODEL ARCHITECTURE



TACOTRON: MODEL ARCHITECTURE CBHG MODULE

The CBHG (1-D convolution bank + highway network + bidirectional GRU) building module.

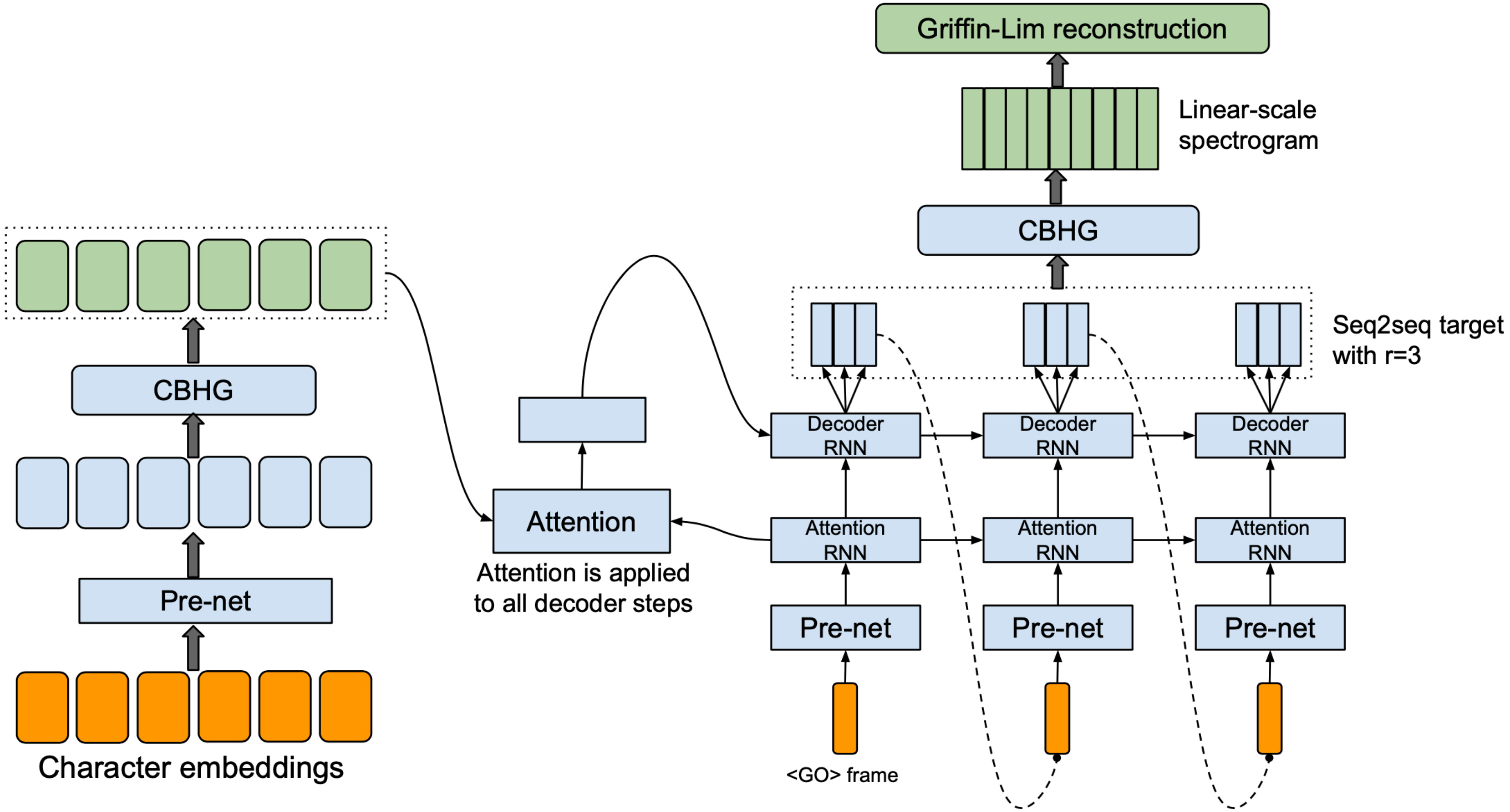
1) The input sequence is first convolved with K sets of 1-D convolutional filters. They explicitly model local and contextual information.
2) The convolution outputs are stacked together and further max pooled along time.
3) Then the processed sequence go to a few 1-D convolutions, whose outputs are added with the original input sequence via residual connections.
4) The outputs are fed into a multi-layer highway network to extract high-level features.
5) Finally, we stack a bidirectional GRU RNN on top to extract sequential features from both forward and backward context.



TACOTRON: MODEL ARCHITECTURE
ENCODER

The goal of the encoder is to extract robust sequential representations of text.

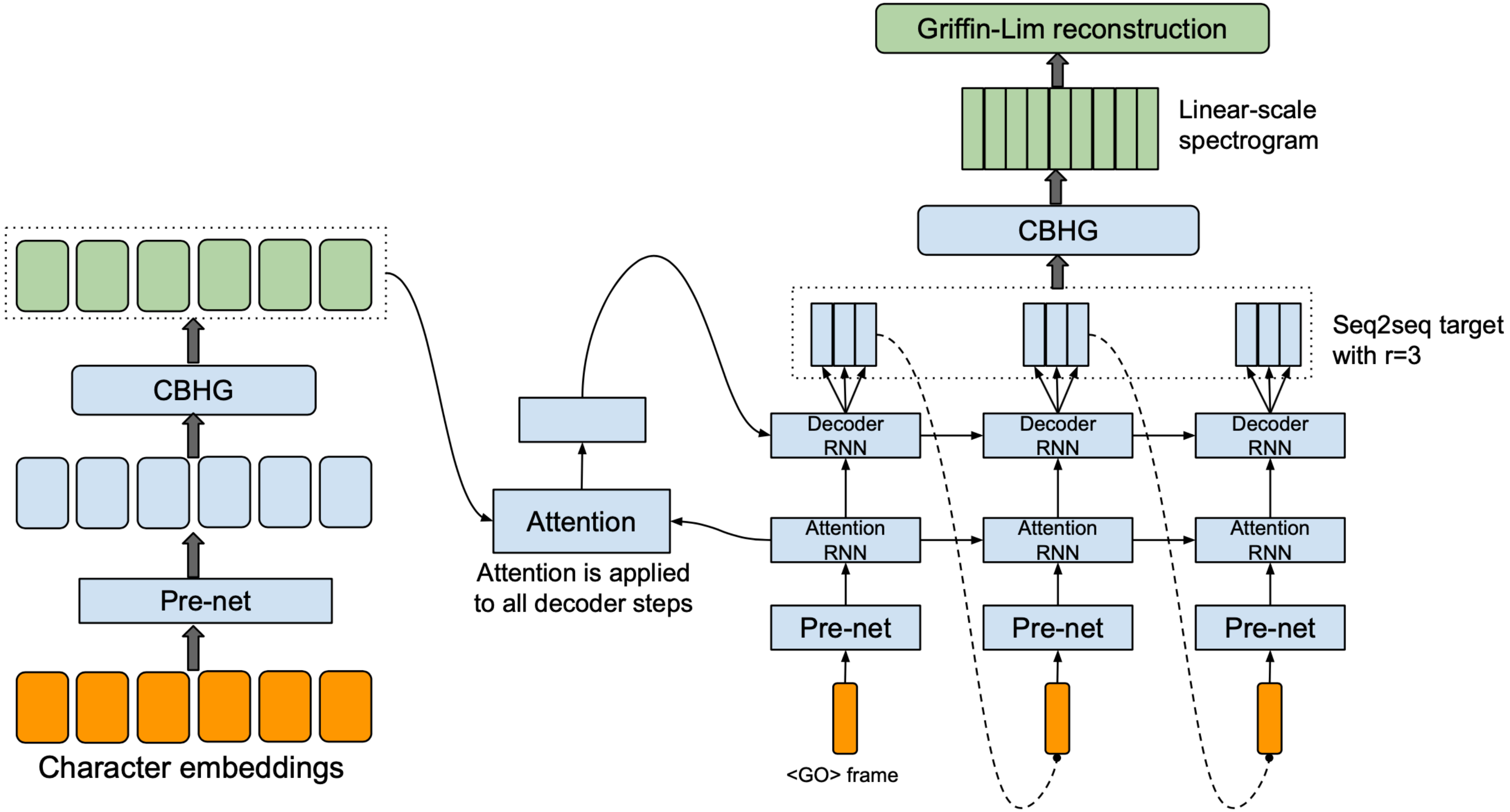
1. Applying a set of non-linear transformations (pre-net, in this work a bottleneck layer with dropout) , to each embedding.
2. A CBHG module transforms the pre-net outputs into the final encoder representation used by the attention module.



TACOTRON: MODEL ARCHITECTURE
DECODER

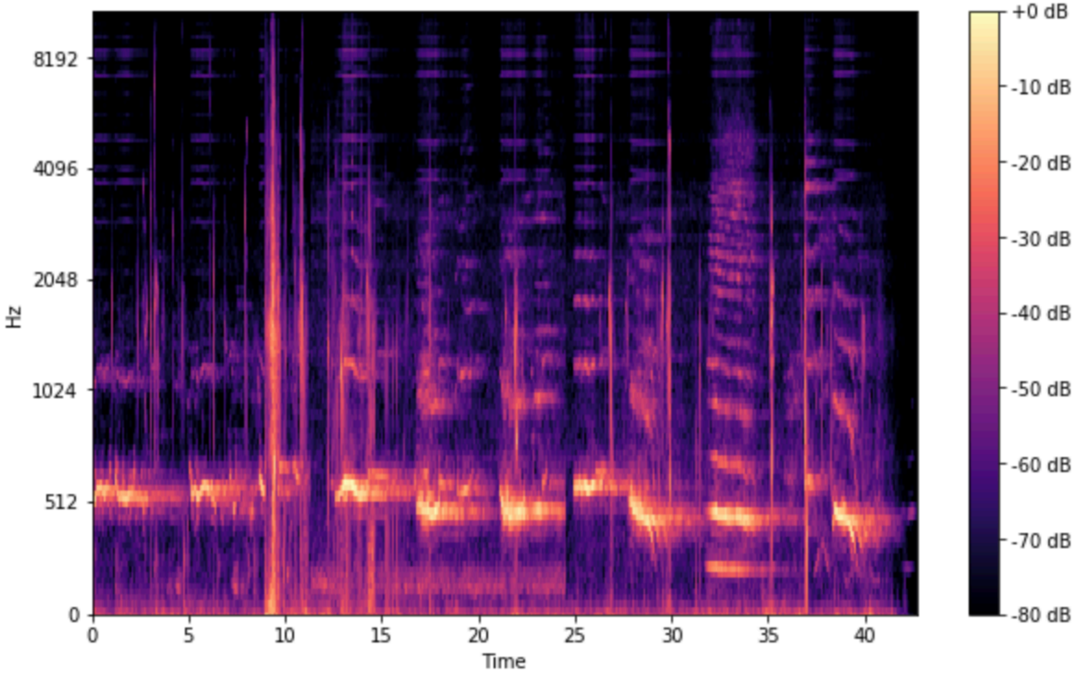
They concatenate the context vector and the attention RNN cell output to form the input to the decoder RNNs. They use a stack of GRUs with vertical residual connections for the decoder.

They use 80-band mel-scale spectrogram as the target.



MEL-SPECTOGRAM

A mel-frequency spectrogram is related to the linear-frequency spectrogram, i.e., the short-time Fourier transform (STFT) magnitude. It is obtained by applying a nonlinear transform to the frequency axis of the STFT, inspired by measured responses from the human auditory system, and summarizes the frequency content with fewer dimensions.



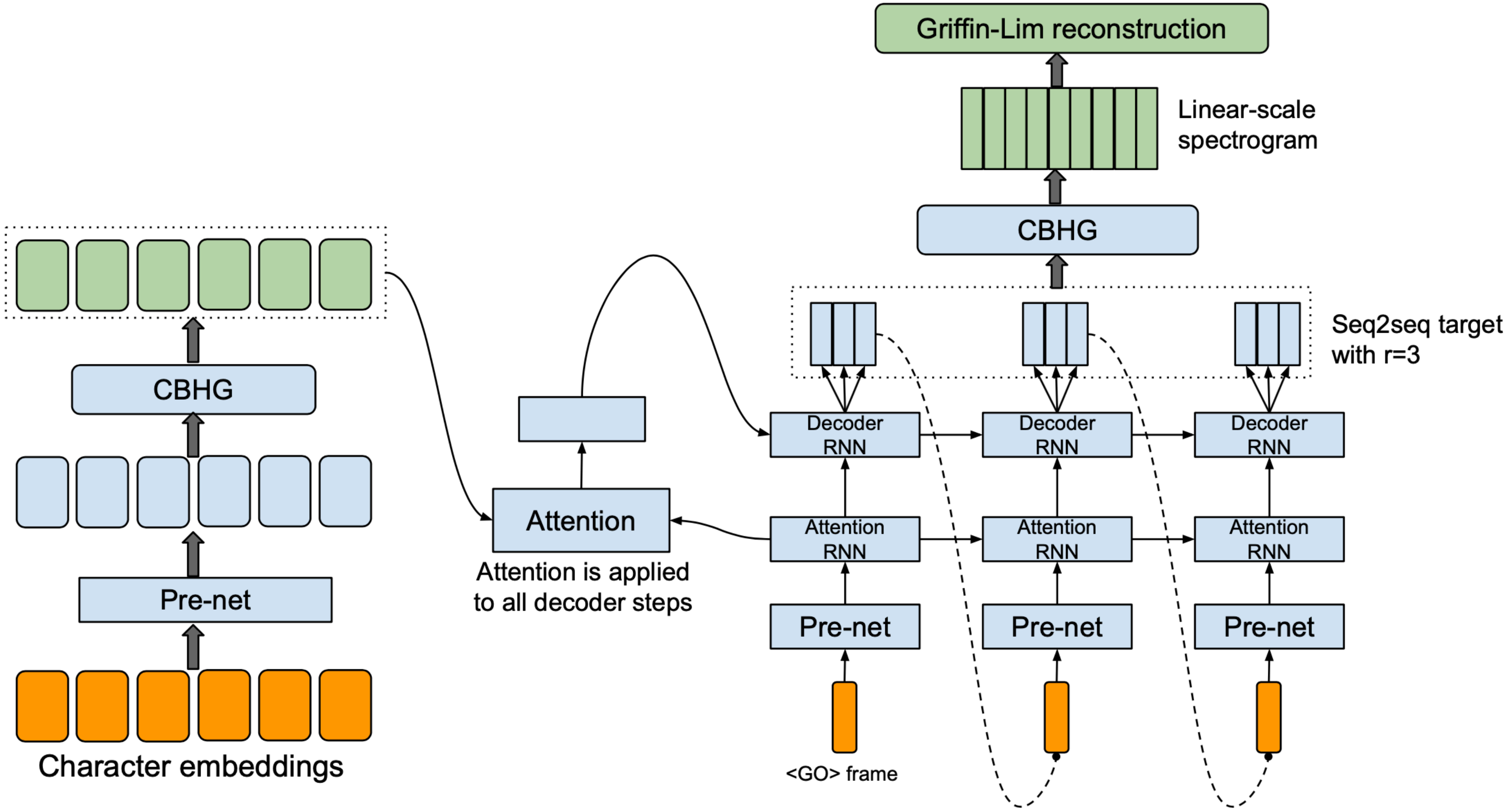
TACOTRON: MODEL ARCHITECTURE
DECODER

They use a post-processing network to convert from the seq2seq target to waveform.

They use a simple fully-connected output layer to predict the decoder targets.

Trick: predicting multiple, non-overlapping output frames at each decoder step.

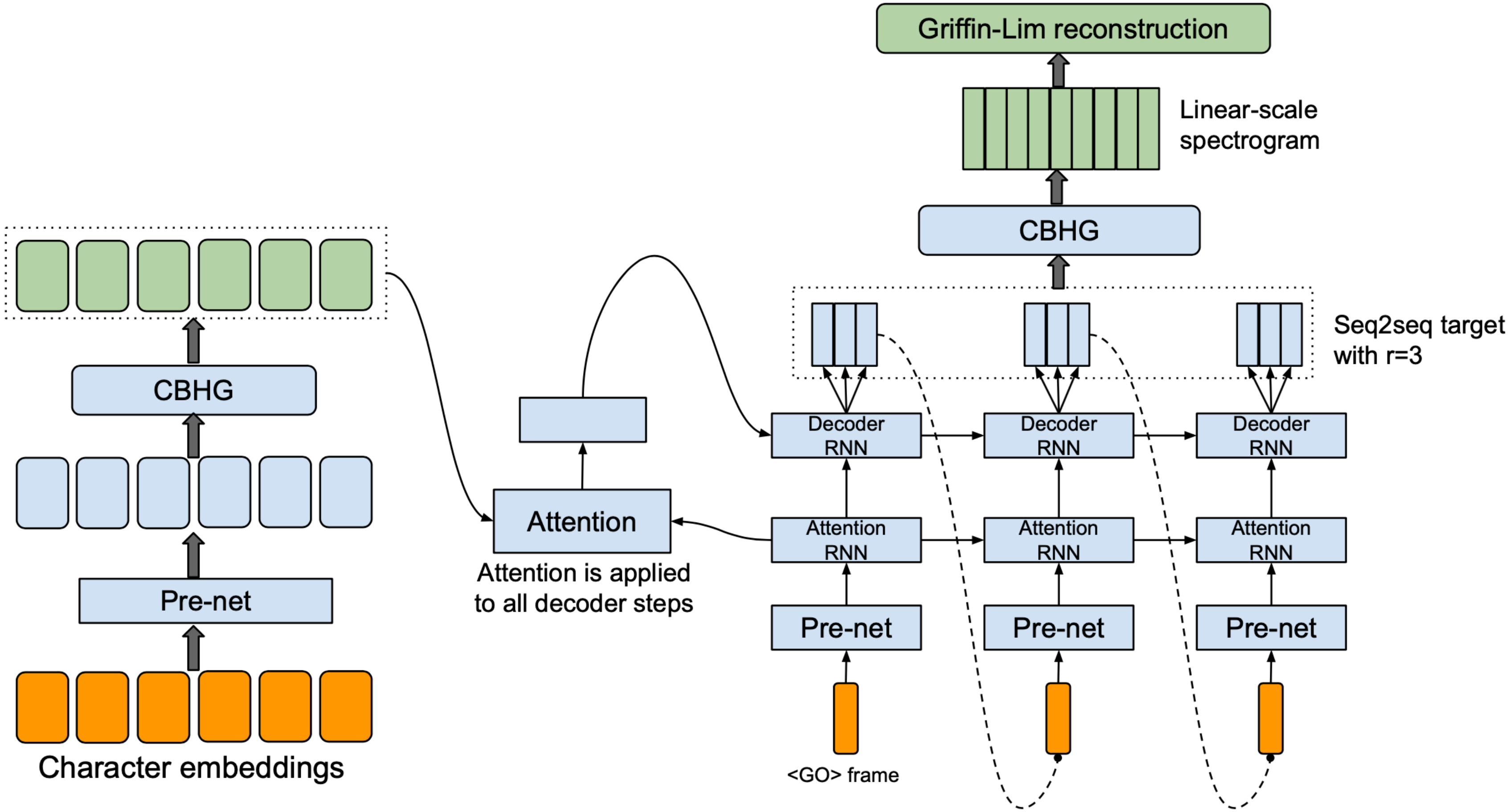
Predicting r frames at once divides the total number of decoder steps by r, which reduces model size, training time, and inference time.



TACOTRON: MODEL ARCHITECTURE
DECODER

In inference, at decoder step t, the last frame of the r predictions is fed as input to the decoder at step t + 1.

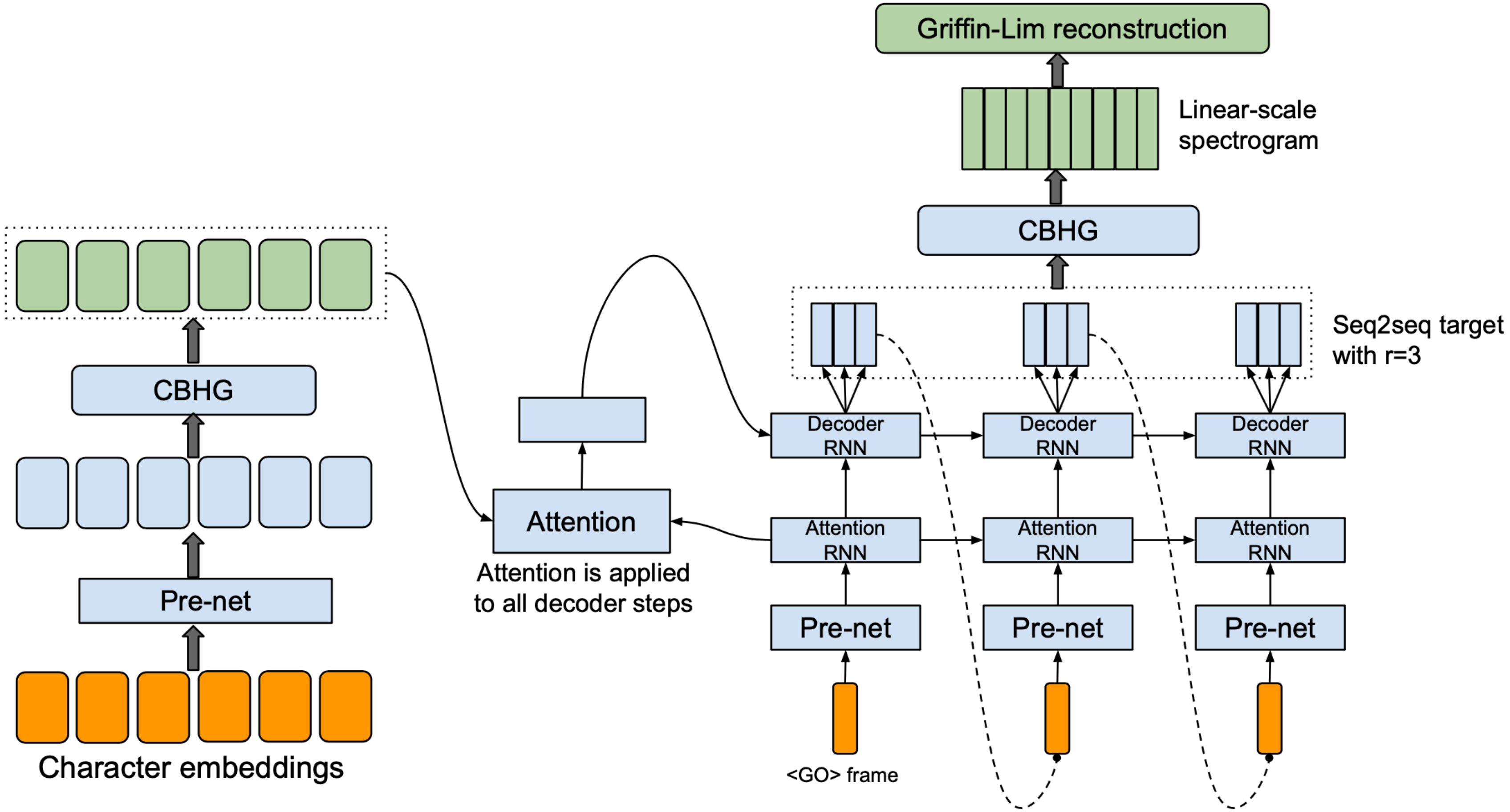
During training, they always feed every r-th ground truth frame to the decoder.



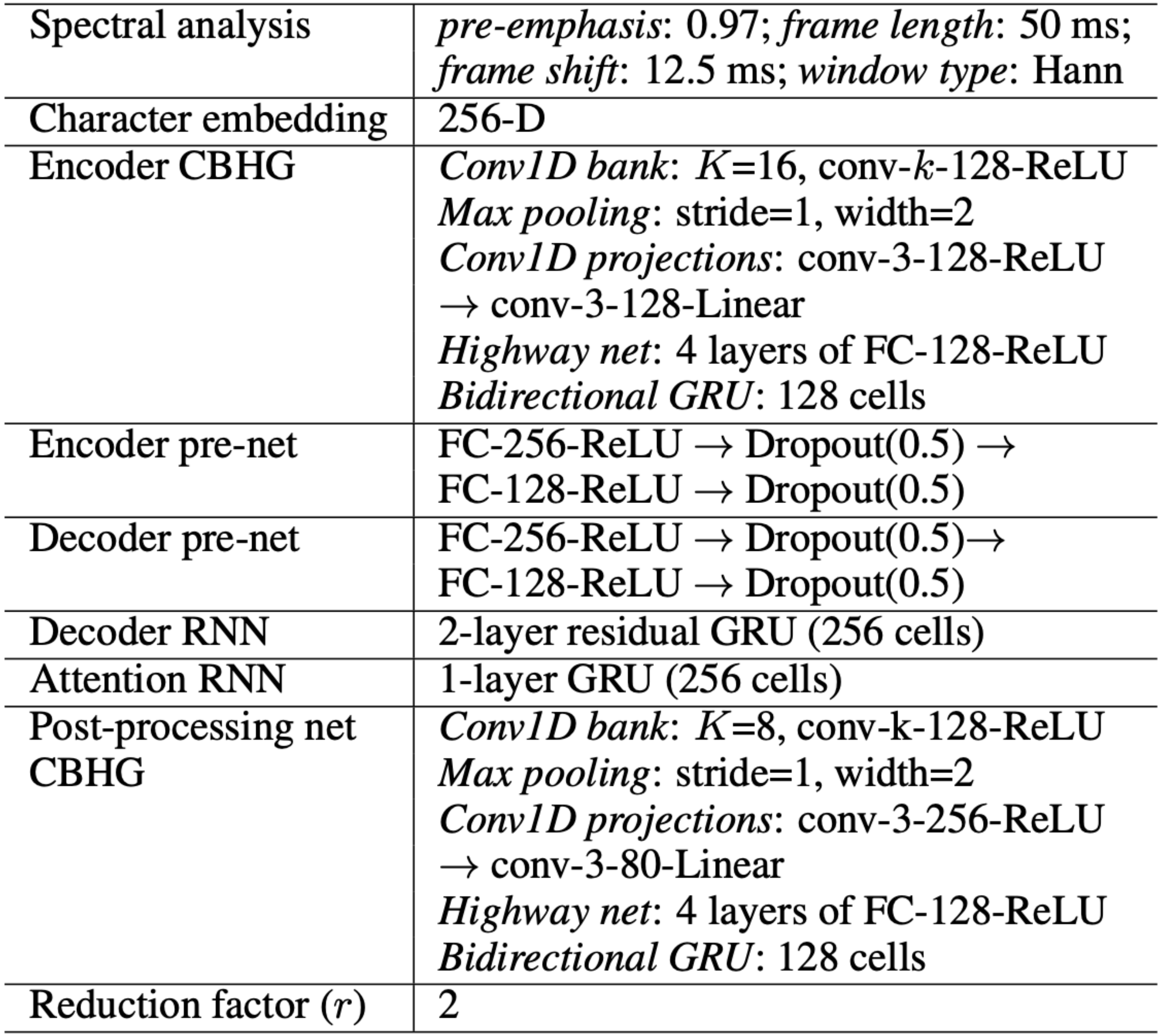
TACOTRON: MODEL ARCHITECTURE
POST-PROCESSING NET

The task is to convert the seq2seq target to a target that can be synthesized into waveforms.

Motivation: it can see the full decoded sequence, bc it has both forward and backward information to correct the prediction error for each individual frame.

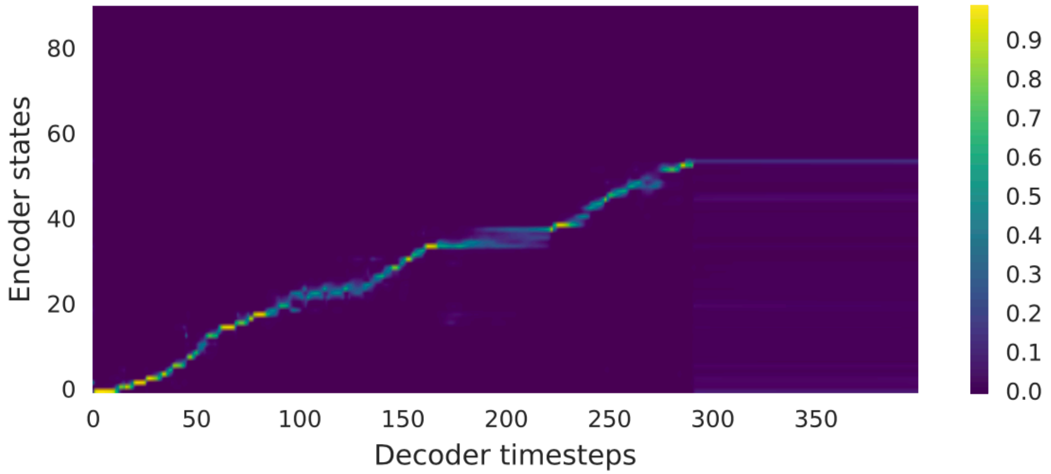


MODEL DETAILS

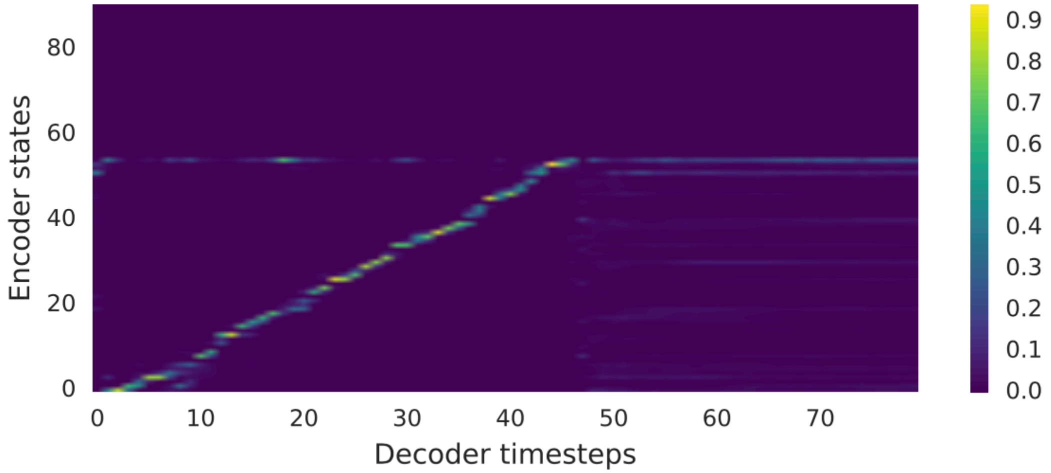


Spectral analysis	*pre-emphasis*: 0.97; *frame length*: 50 ms; *frame shift*: 12.5 ms; *window type*: Hann
Character embedding	256-D
Encoder CBHG	*Conv1D bank*: K=16, conv-k-128-ReLU *Max pooling*: stride=1, width=2 *Conv1D projections*: conv-3-128-ReLU → conv-3-128-Linear *Highway net*: 4 layers of FC-128-ReLU *Bidirectional GRU*: 128 cells
Encoder pre-net	FC-256-ReLU → Dropout(0.5) → FC-128-ReLU → Dropout(0.5)
Decoder pre-net	FC-256-ReLU → Dropout(0.5)→ FC-128-ReLU → Dropout(0.5)
Decoder RNN	2-layer residual GRU (256 cells)
Attention RNN	1-layer GRU (256 cells)
Post-processing net CBHG	*Conv1D bank*: K=8, conv-k-128-ReLU *Max pooling*: stride=1, width=2 *Conv1D projections*: conv-3-256-ReLU → conv-3-80-Linear *Highway net*: 4 layers of FC-128-ReLU *Bidirectional GRU*: 128 cells
Reduction factor (r)	2

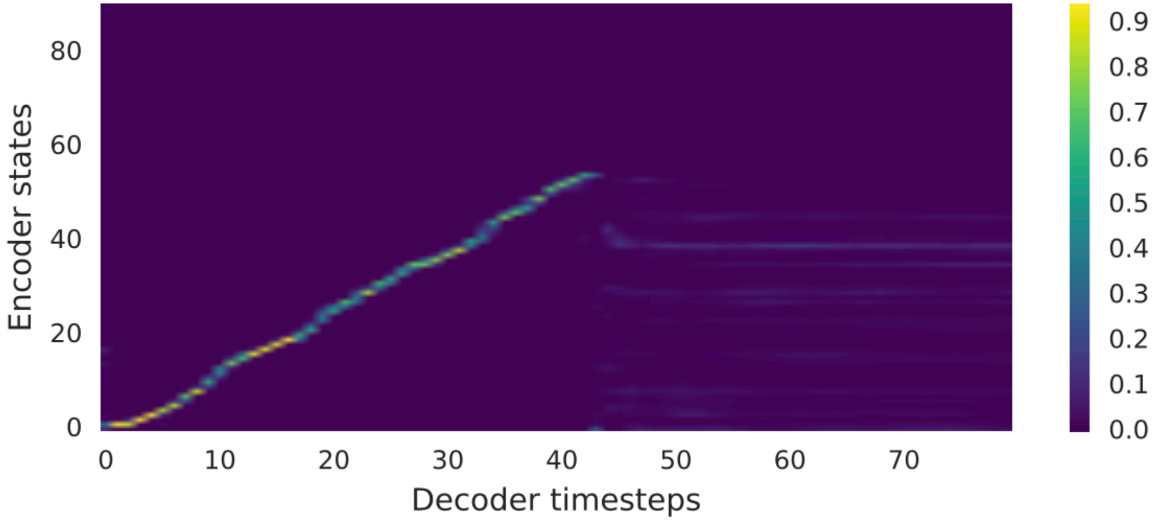
EXPERIMENTS



(a) Vanilla seq2seq + scheduled sampling



(b) GRU encoder



(c) Tacotron (proposed)

TACOTRON2 (2018)

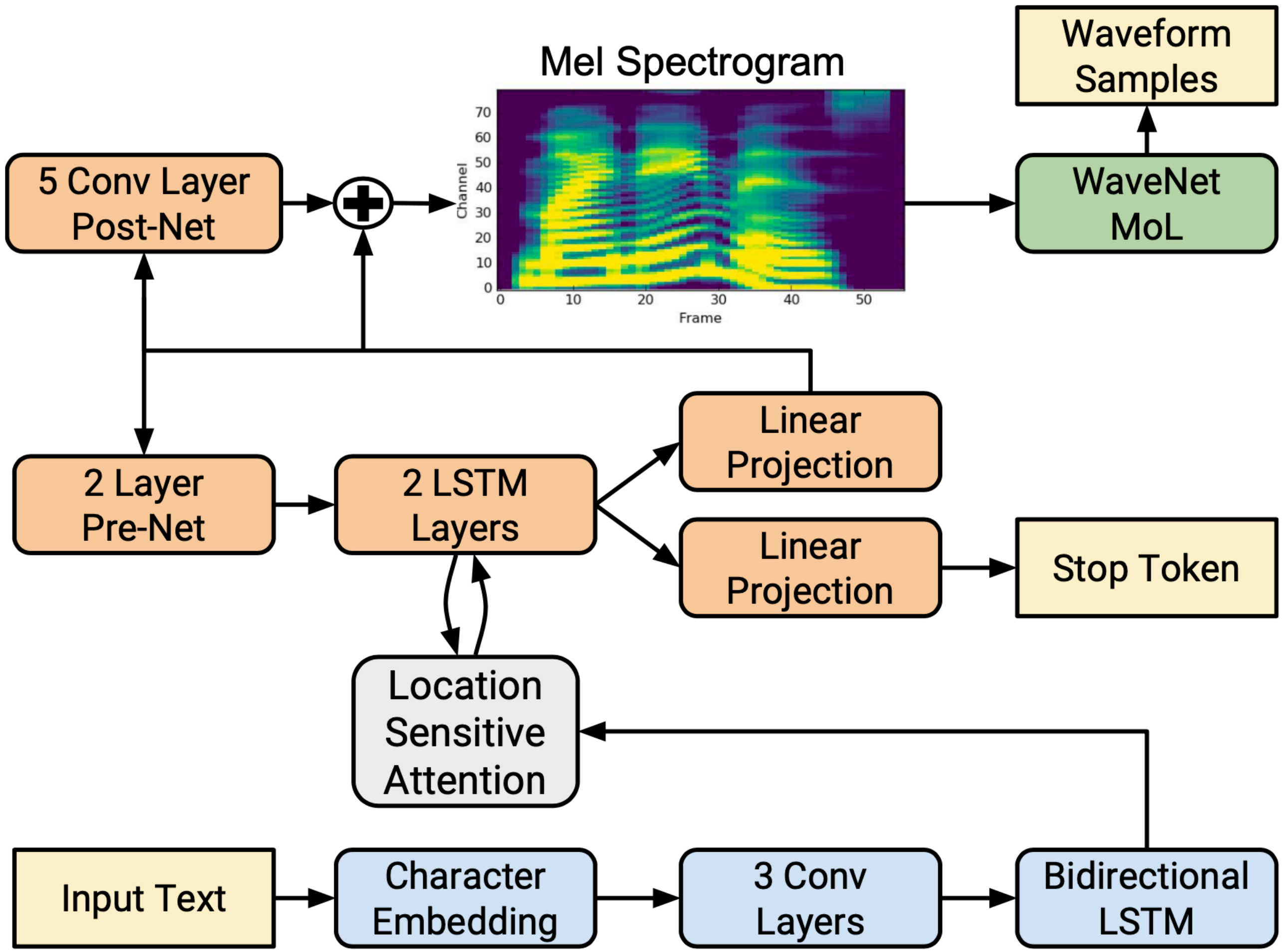
Tacotron 2, a neural network architecture for speech synthesis directly from text. The system consists of two components:

1. a recurrent seq2seq feature prediction network with attention which predicts a sequence of mel spectrogram frames from an input character sequence
2. a modified version of WaveNet which generates time-domain waveform samples conditioned on the predicted mel spectrogram frames.

TACOTRON2: MODEL ARCHITECTURE
ENCODER

Input characters are represented using a learned character embedding, which are passed through a stack of 3 convolutional layers.

The output of the final convolutional layer is passed into a single bi-directional LSTM to generate the encoded features.

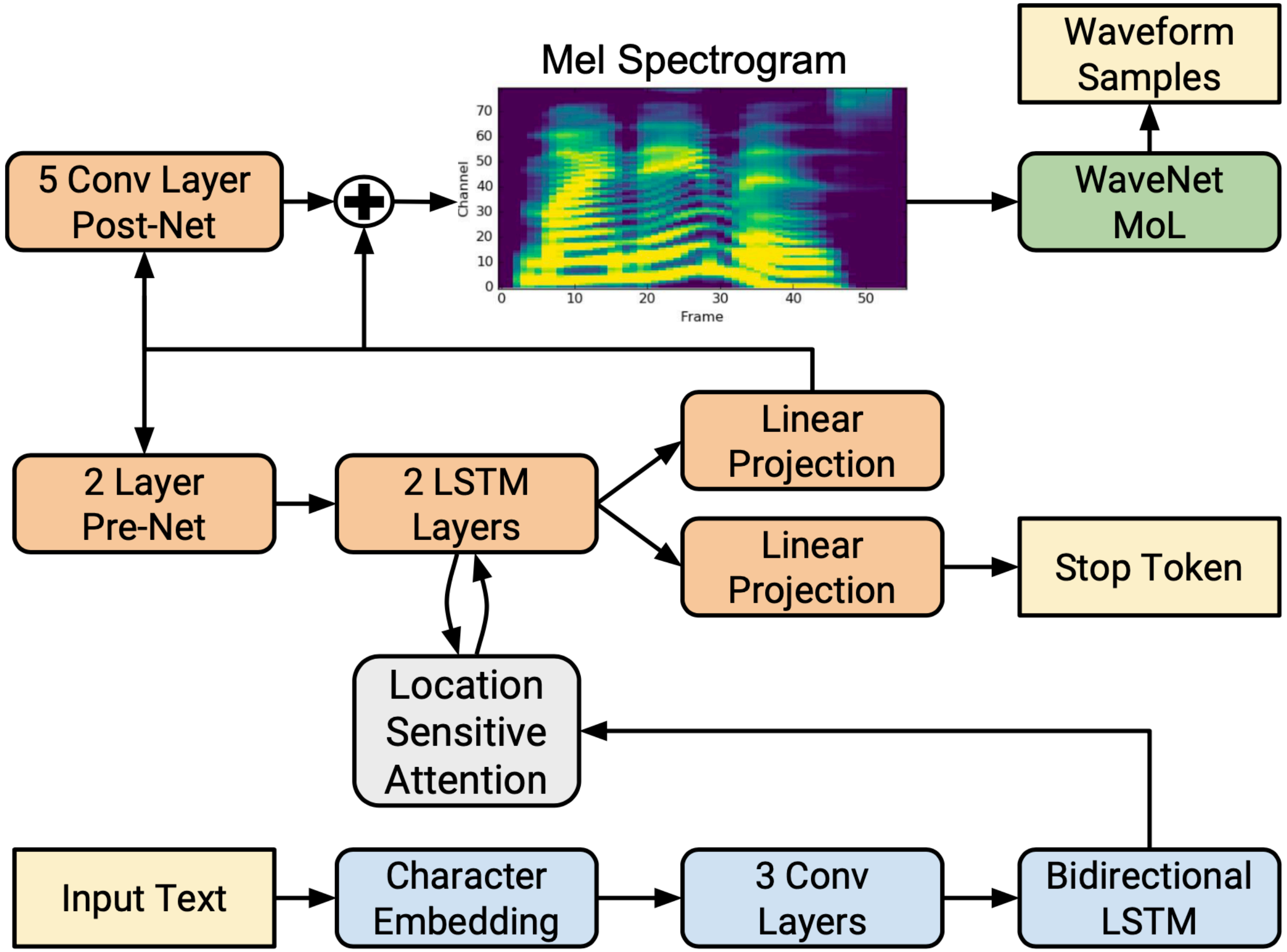


TACOTRON2: MODEL ARCHITECTURE
ATTENTION

The encoder output is consumed by an attention network which summarizes the full encoded sequence as a fixed-length context vector for each decoder output step.

They use the location-sensitive attention.

Attention probabilities are computed after projecting inputs and location features to 128-dimensional hidden representations.



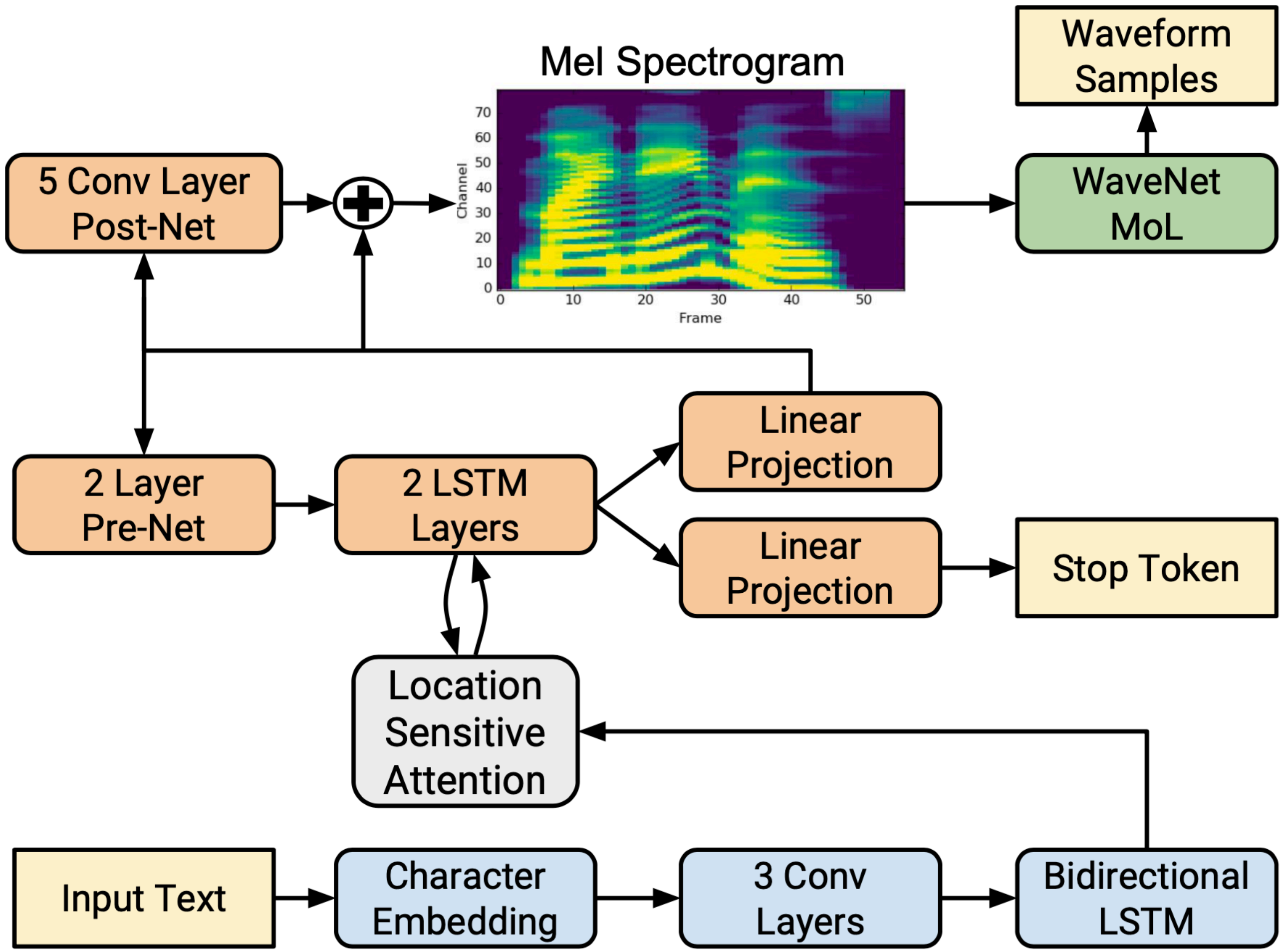
TACOTRON2: MODEL ARCHITECTURE
DECODER

The prediction from the previous time step is first passed through a small *pre-net*.

The pre-net output and attention context vector are concatenated and passed through a stack of 2 uni-directional LSTM layers.

The concatenation of the LSTM output and the attention context vector is projected through a linear transform.

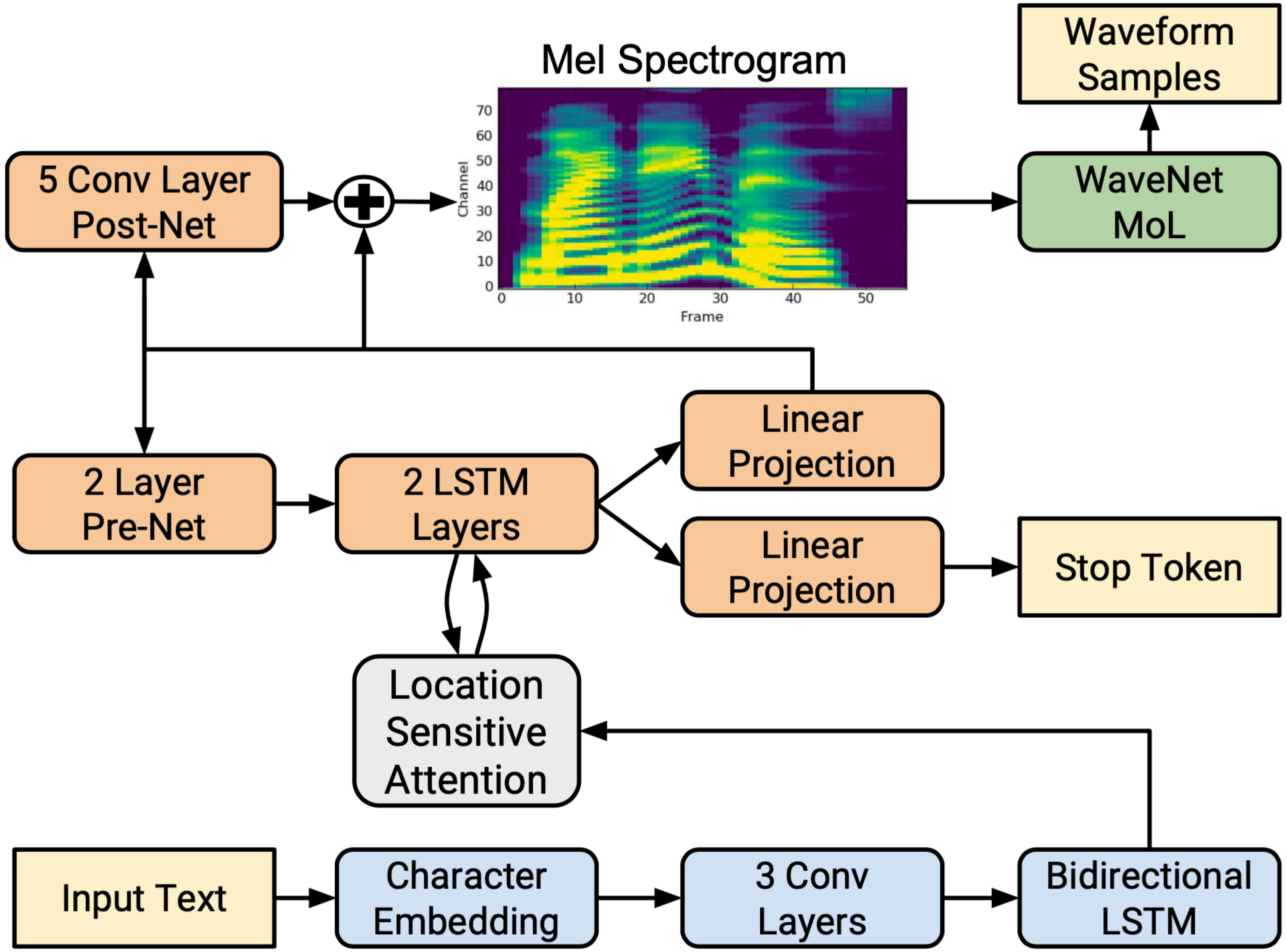
Finally, the predicted mel spectrogram is passed through a 5-layer convolutional *post-net* which predicts a residual.



TACOTRON2: MODEL ARCHITECTURE
STOP TOKEN

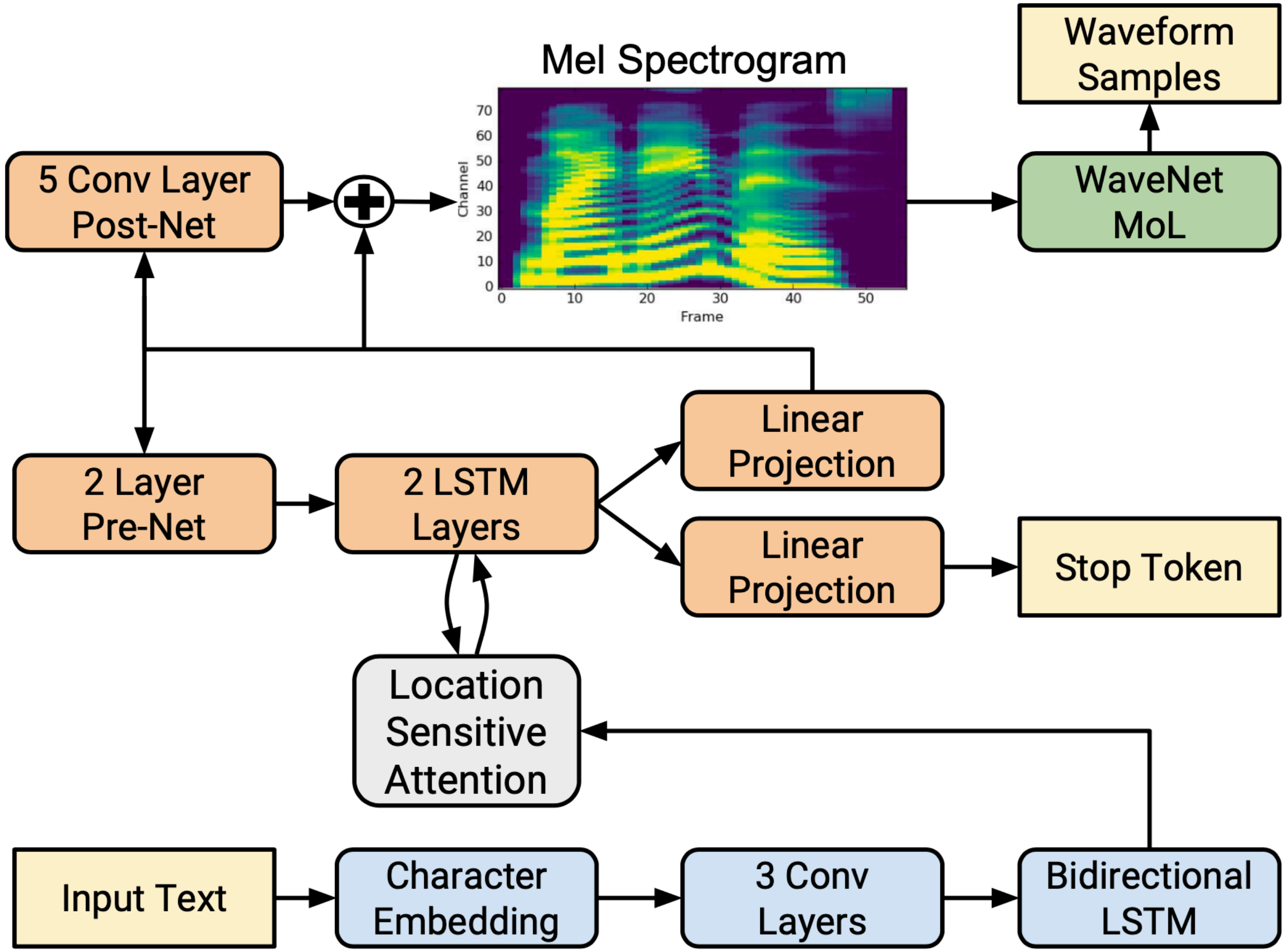
The concatenation of decoder LSTM output and the attention context is projected down to a scalar for predict the probability that the output sequence has completed.

This “stop token” prediction is used during inference to allow the model to dynamically determine when to terminate generation instead of always generating for a fixed duration.



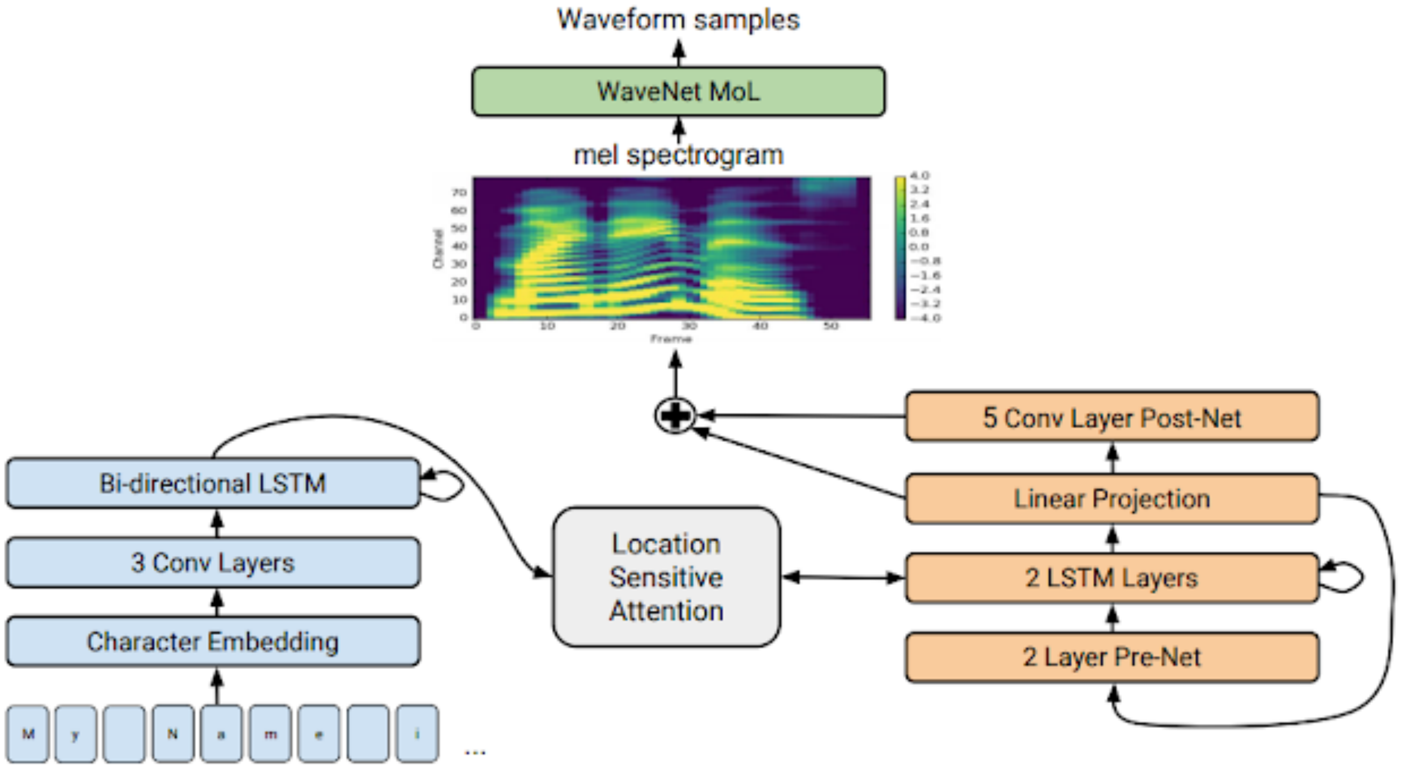
TACOTRON2: MODEL ARCHITECTURE
WAVE NET

Instead of predicting discretized buckets with a softmax layer, they use a 10- component mixture of logistic distributions (MoL) to generate 16-bit samples at 24 kHz. To compute the logistic mixture distribution, the WaveNet stack output is passed through a ReLU activation followed by a linear projection to predict parameters (mean, log scale, mixture weight) for each mixture component.

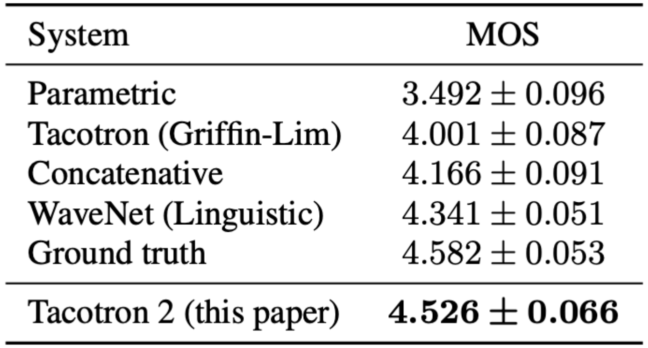


TACOTRON 2: MODEL ARCHITECTURE

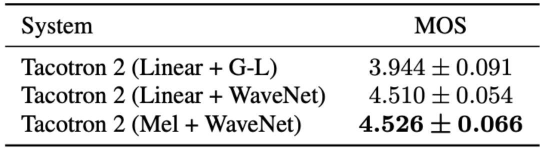
Using vanilla LSTM and convolutional layers in the encoder and decoder instead of “CBHG” stacks and GRU recurrent layers.



TACOTRON 2 RESULTS



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)	**4.526 ± 0.066**



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**

Table 3. Comparison of evaluated MOS for Griffin-Lim vs. WaveNet as a vocoder, and using 1,025-dimensional linear spectrograms vs. 80-dimensional mel spectrograms as conditioning inputs to WaveNet.

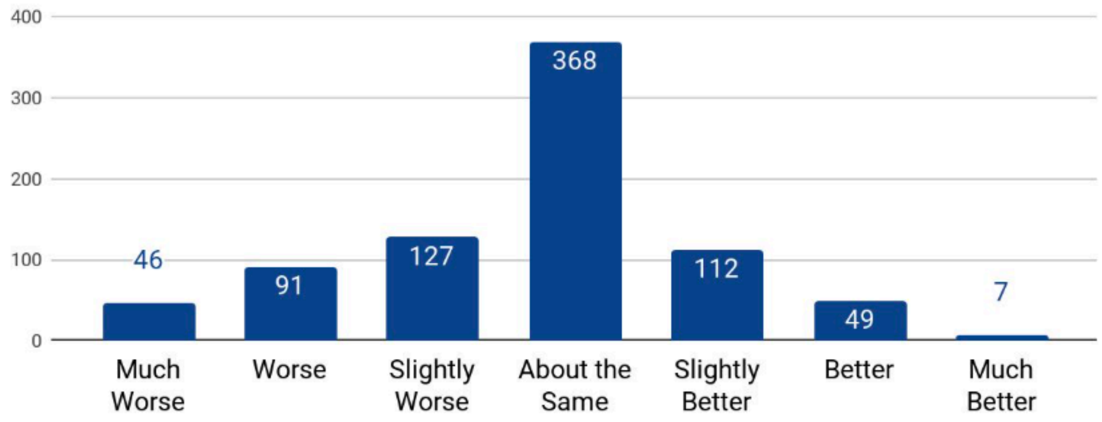


Fig. 2. Synthesized vs. ground truth: 800 ratings on 100 items.

BIBLIOGRAPHY

Tacotron https://arxiv.org/abs/1703.10135

Tacotron 2 https://arxiv.org/abs/1712.05884

Audio samples and other papers on Tacotron
https://google.github.io/tacotron/index.html