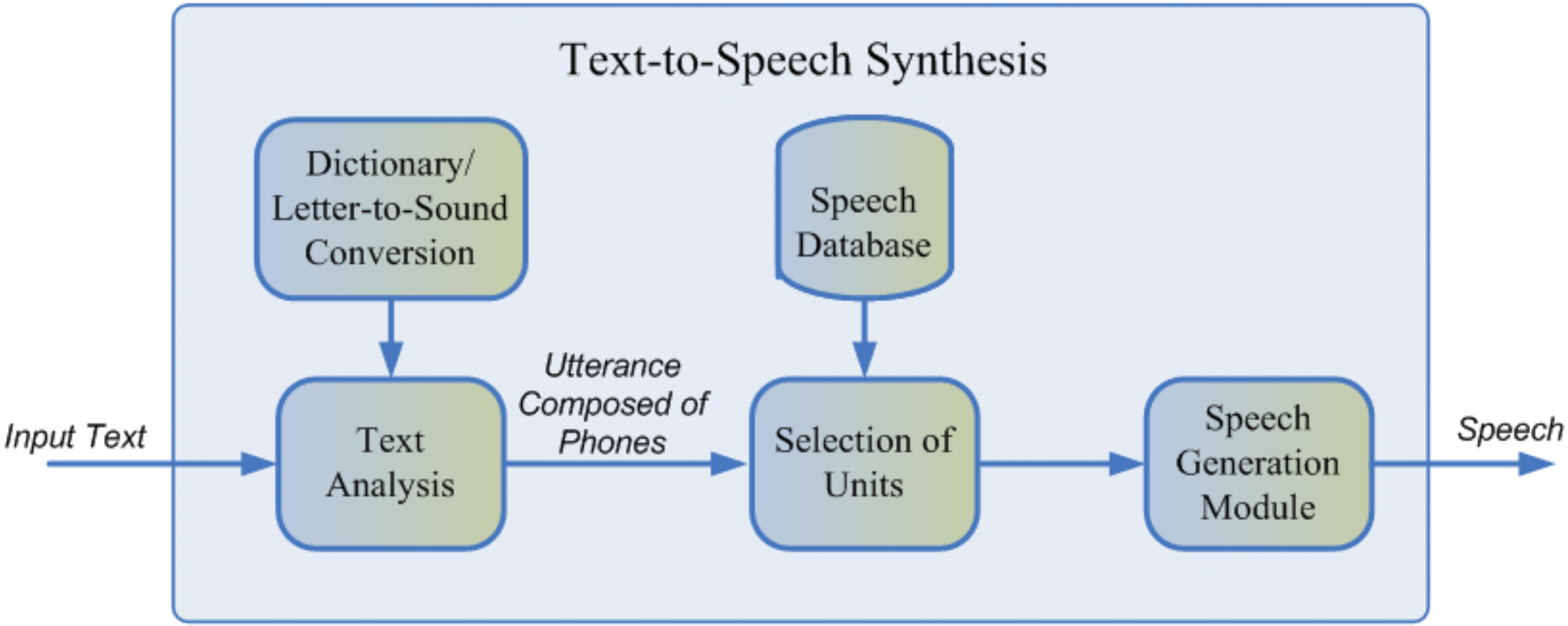
Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

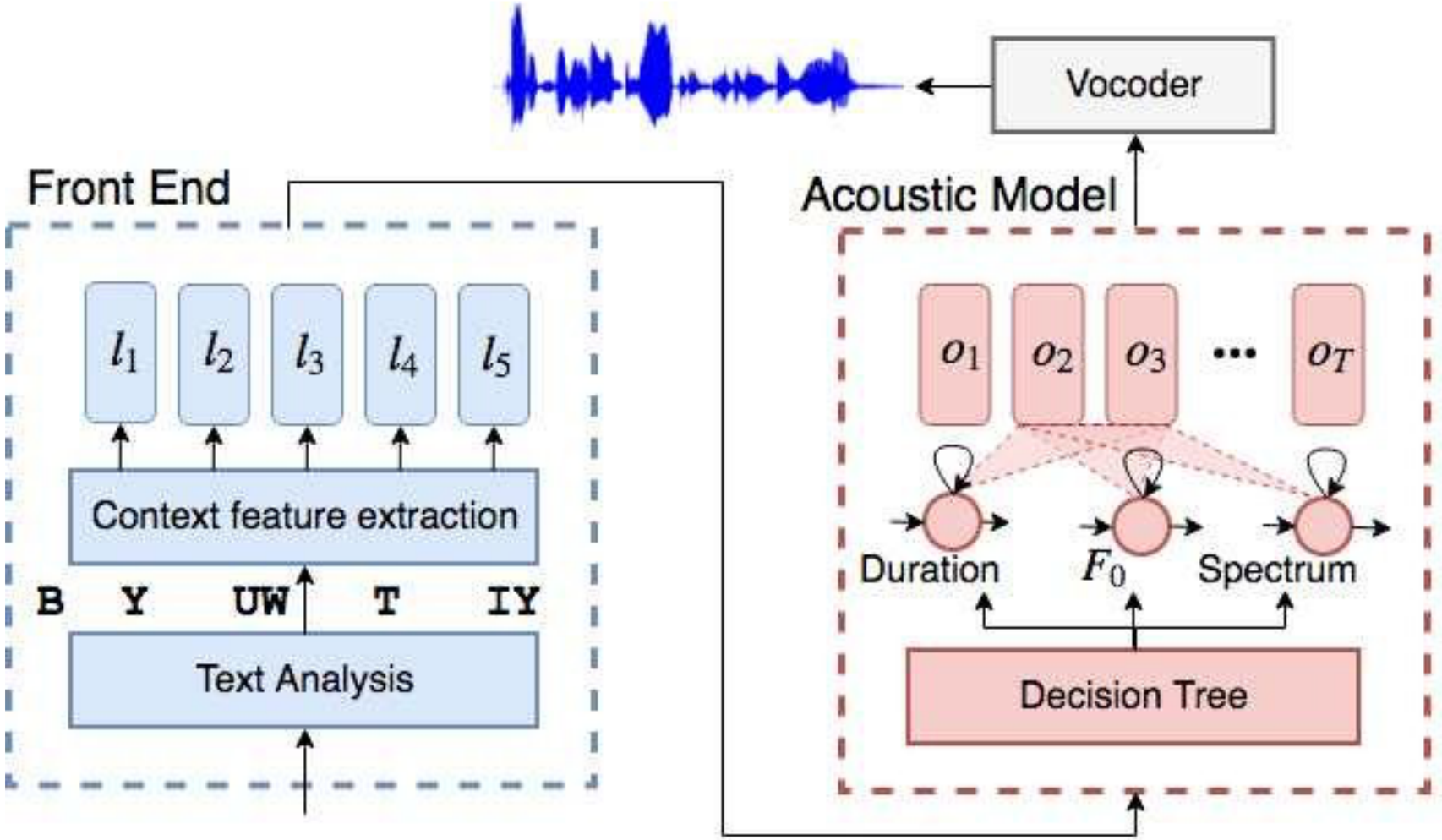
Plan

- First models of TTS
- WaveNet
- Spectrogram
- Tacotron
- Tacotron 2
- Experiments

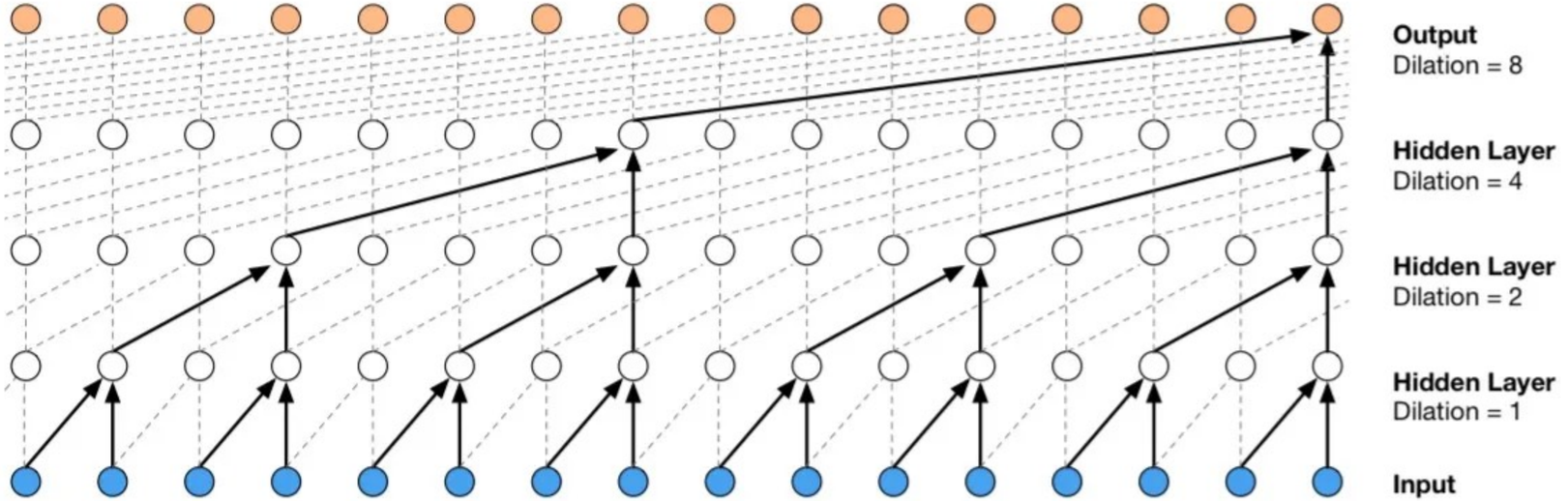
Concatenative model



Parametric model

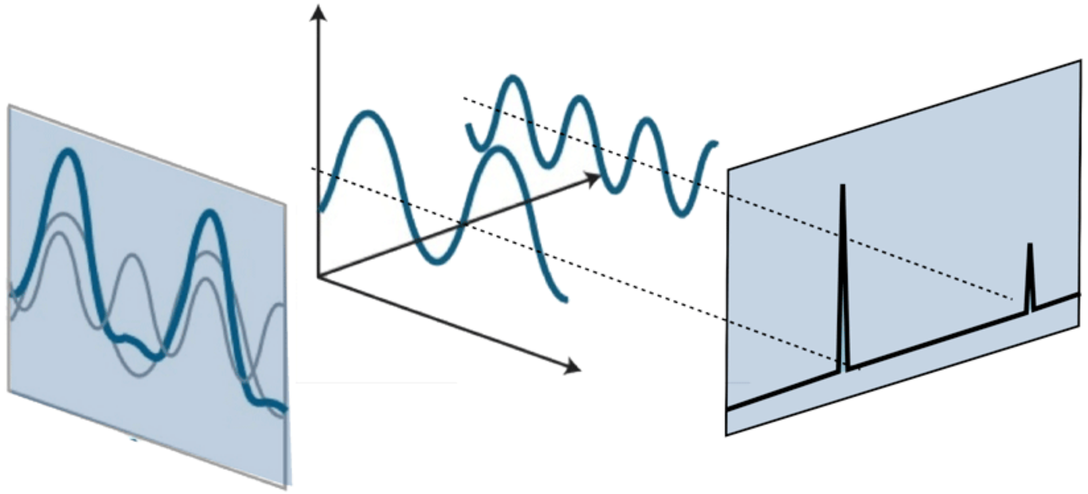


WaveNet



	Subjective 5-scale MOS in naturalness	
Speech samples	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

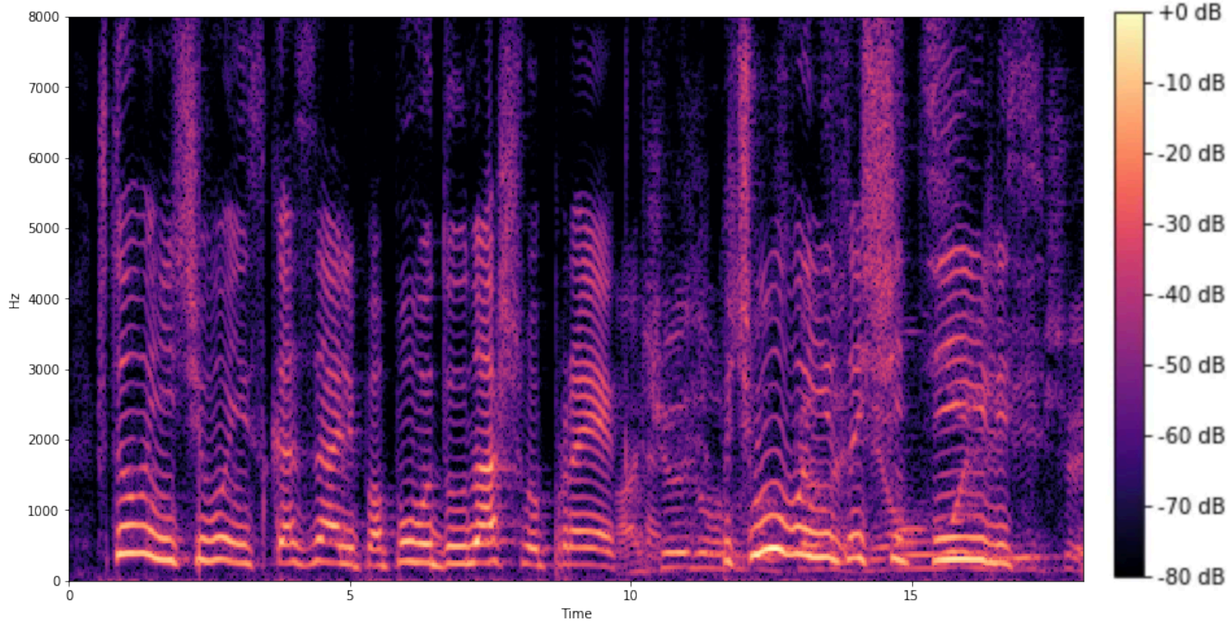
Spectrogram

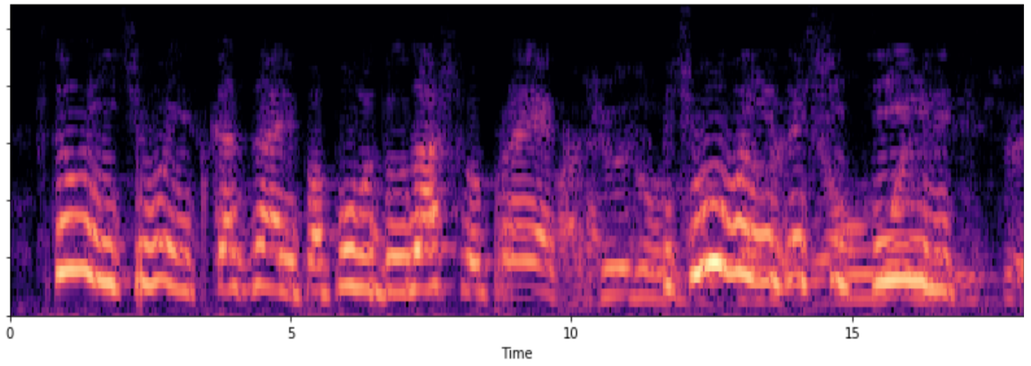


Time Domain
s(t)

FT

Frequency Domain
S(ω)

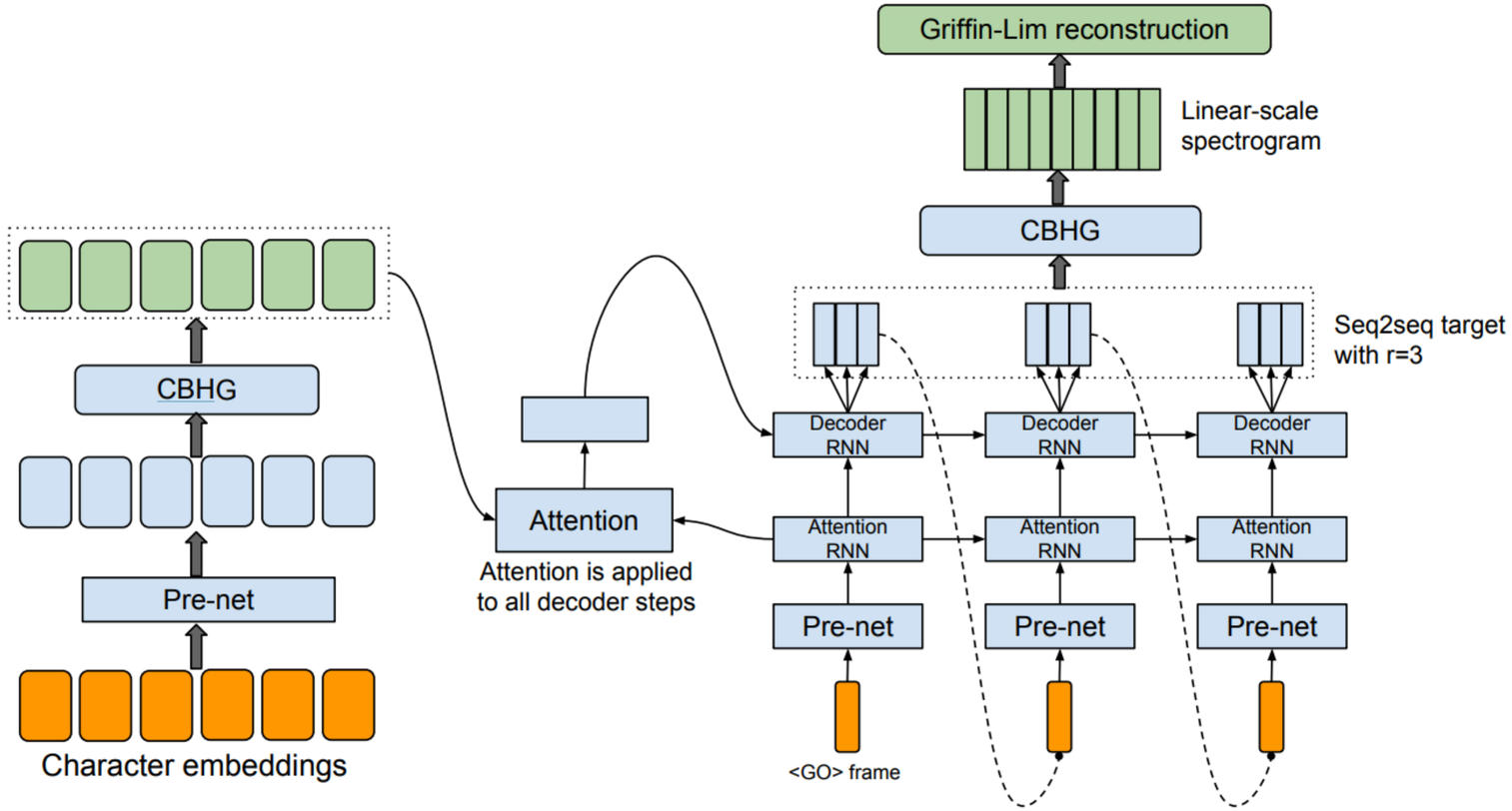




Tacotron

1. Get text
2. Generate Mel-Spectrogram
3. Applies Griffin-Lim algorithm

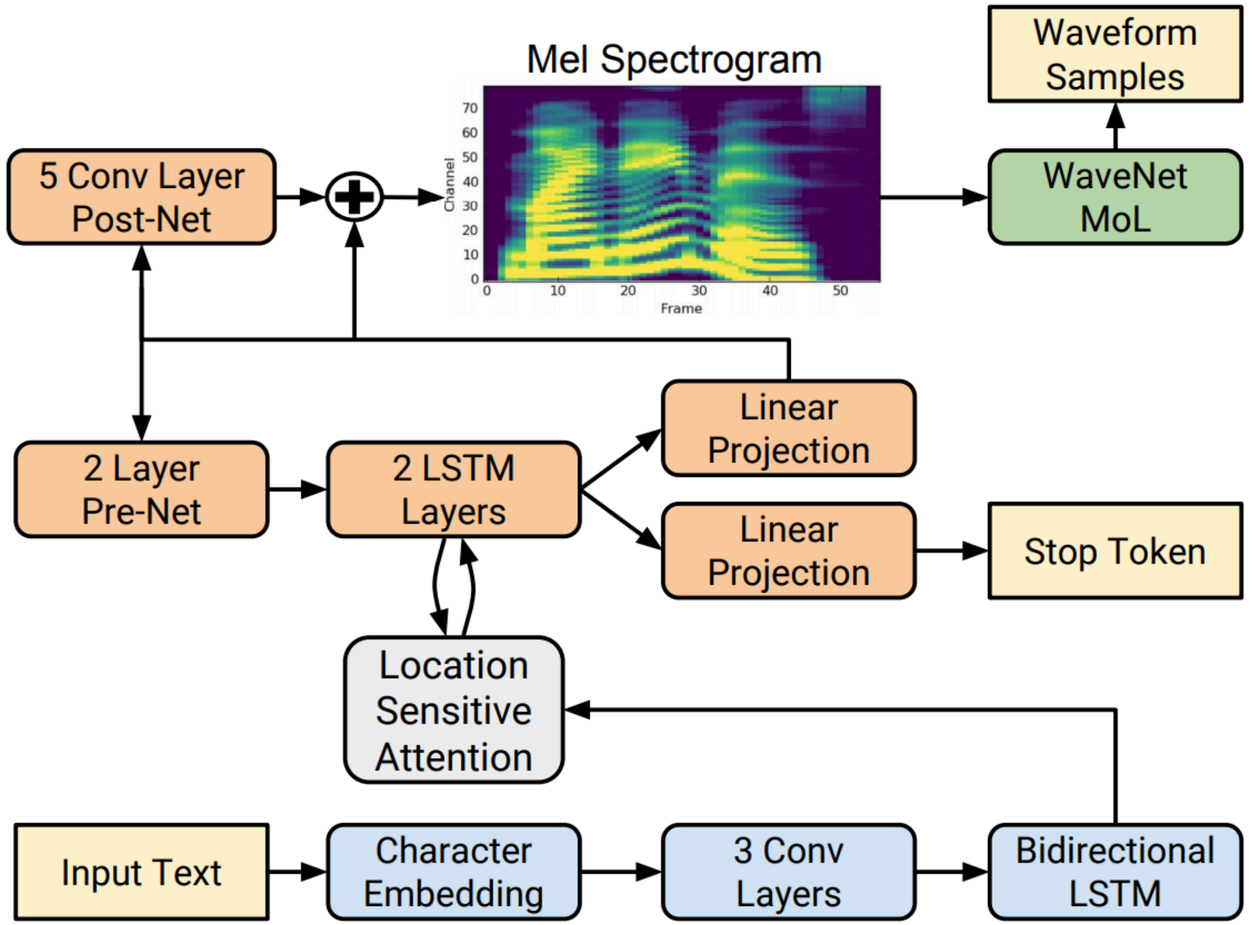
MOS test result: 3.82 ± 0.085



Tacotron 2

Difference from Tacotron

- Griffin-Lim replaced by modified Wavenet
- CBHG replaced by LSTM and convolutional layers
- Attention replaced by Location Sensitive Attention

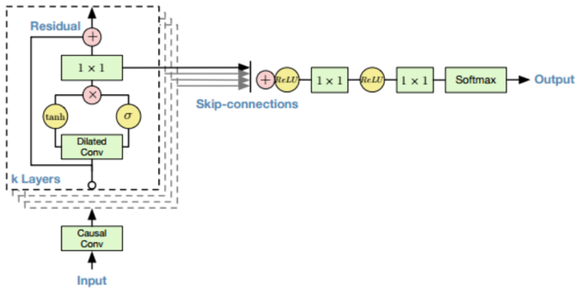


Wavenet modifications

1. Input is mel spectrogram

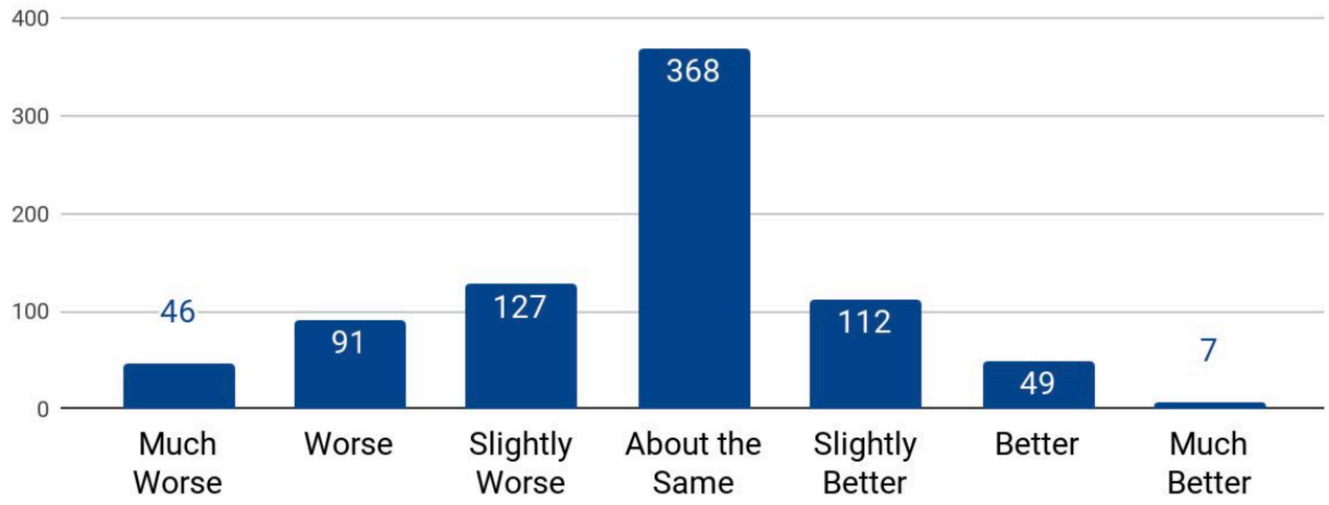
2. 30 dilated convolutional layers, grouped into 3 dilation cycles, i.e., the dilation rate of layer k (k=0,…, 29) is 2^(k mod 10)

3. To work with 12.5 ms frames, only 2 upsampling layers are used in the conditioning stack, instead of 3



MOS test

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)	$\mathbf{4.526 \pm 0.066}$



Modifications

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

Table 2. Comparison of evaluated MOS for our system when WaveNet trained on predicted/ground truth mel spectrograms are made to synthesize from predicted/ground truth mel spectrograms.

Modifications

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

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.

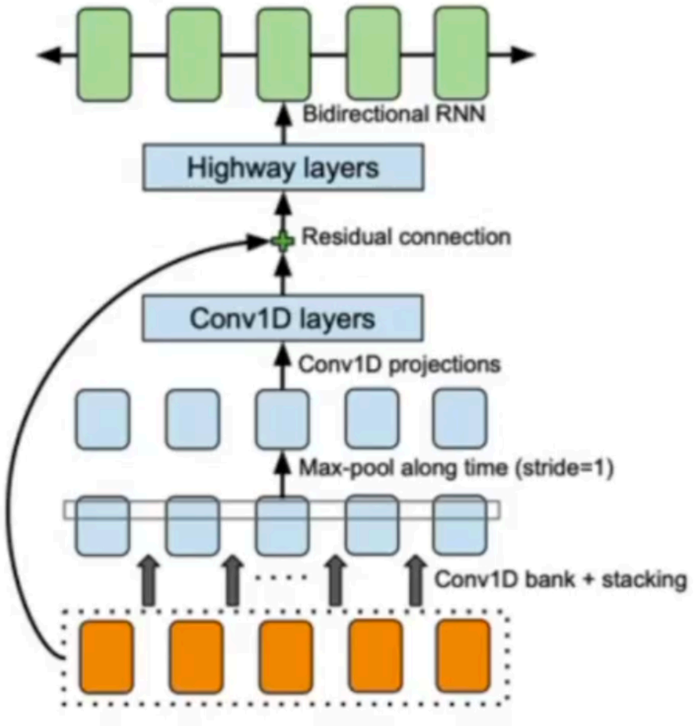
Modifications

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

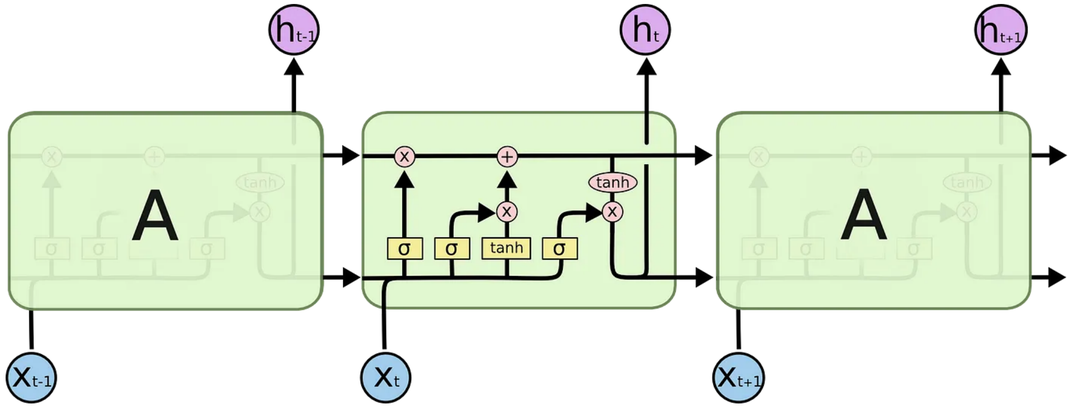
Table 4. WaveNet with various layer and receptive field sizes.

Additional materials

CBHG

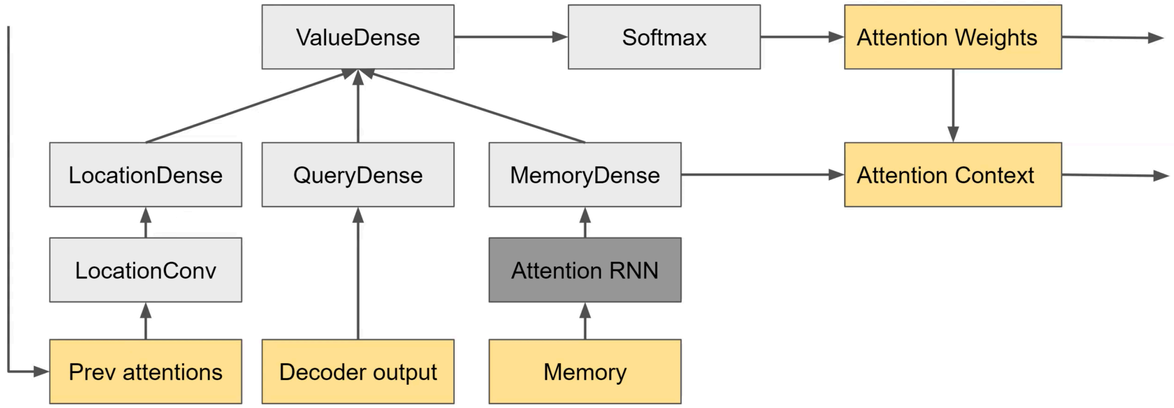


LSTM layer



Location-sensitive Attention

$$e_{i,j} = w^T \tanh(W s_{i-1} + V h_j + U f_{i,j} + b)$$



Links

- https://arxiv.org/abs/1712.05884
- https://arxiv.org/abs/1609.03499
- https://medium.com/swlh/a-simple-overview-of-rnn-lstm-and-attention-mechanism-9e844763d07b