Tacotron: Towards End-to-End Speech Synthesis

Speech Synthesis

- Задача синтеза – перевод текста в речь.
- Речь должна:
  - Передовать смысл сообщения.
  - Не допускать ошибок.
  - Звучать натурально.
- Критерий качества – Mean Opinion Score (MOS)

Подходы к синтезу

- Concatenative synthesis
- Parametric synthesis

Подходы к синтезу

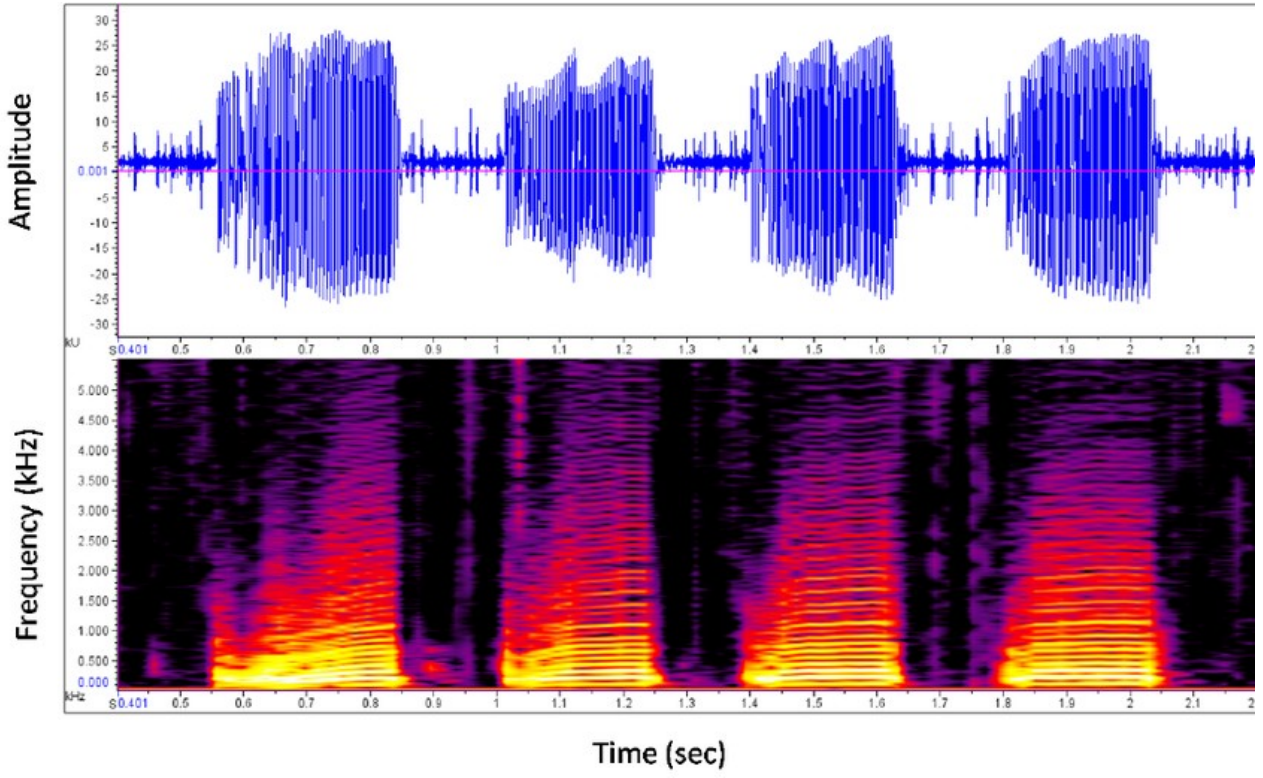
- Concatenative synthesis
- Parametric synthesis
  - Extract linguistic features
  - Duration model
  - Acoustic feature prediction model
  - Signal-processing-based vocoder

Данные

- На вход можно: текст, фонемы, спец. символы.
- Аудио.

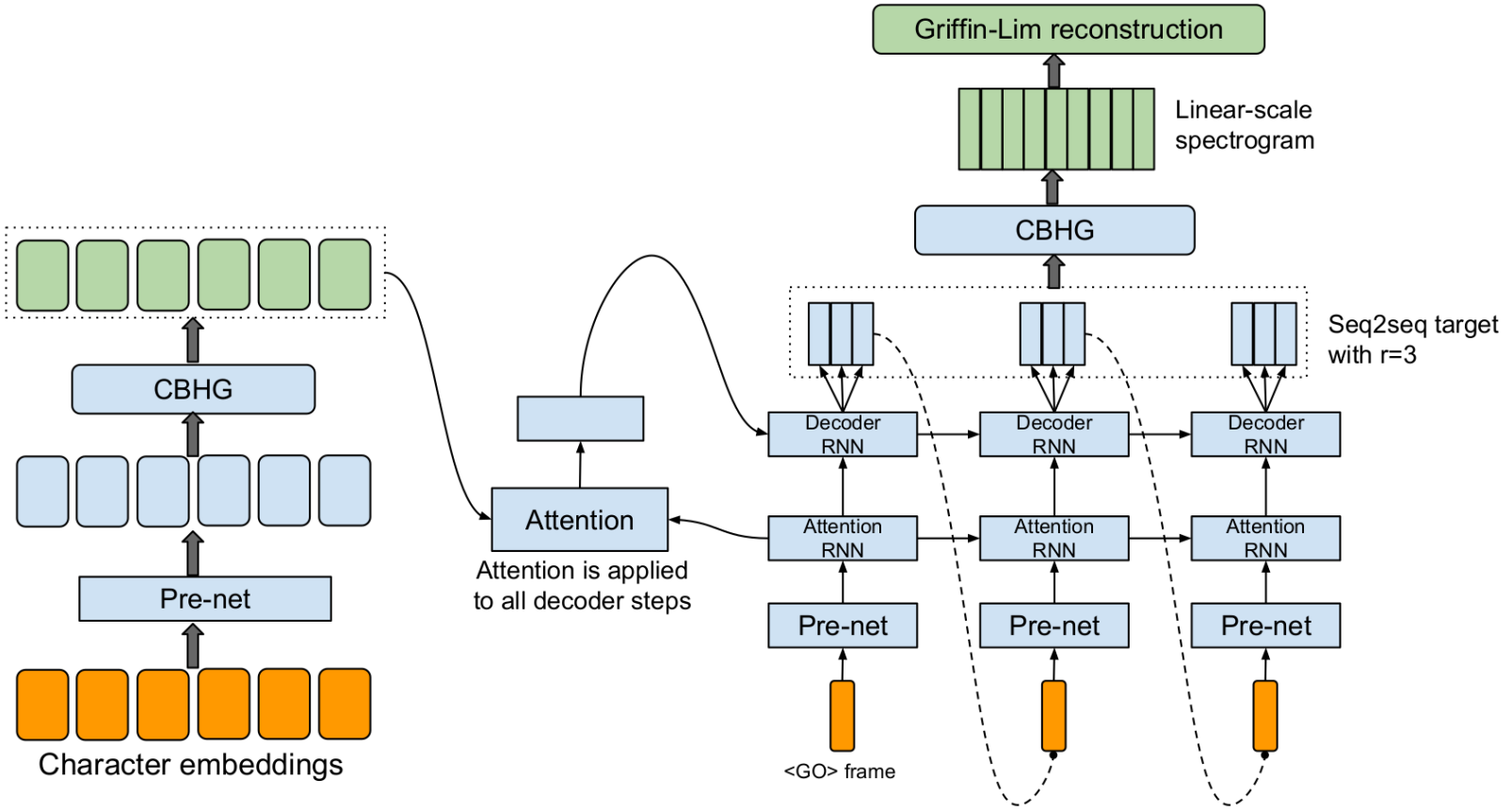
Spectrogram

- Спектрограмма – представление спектра частот звука.

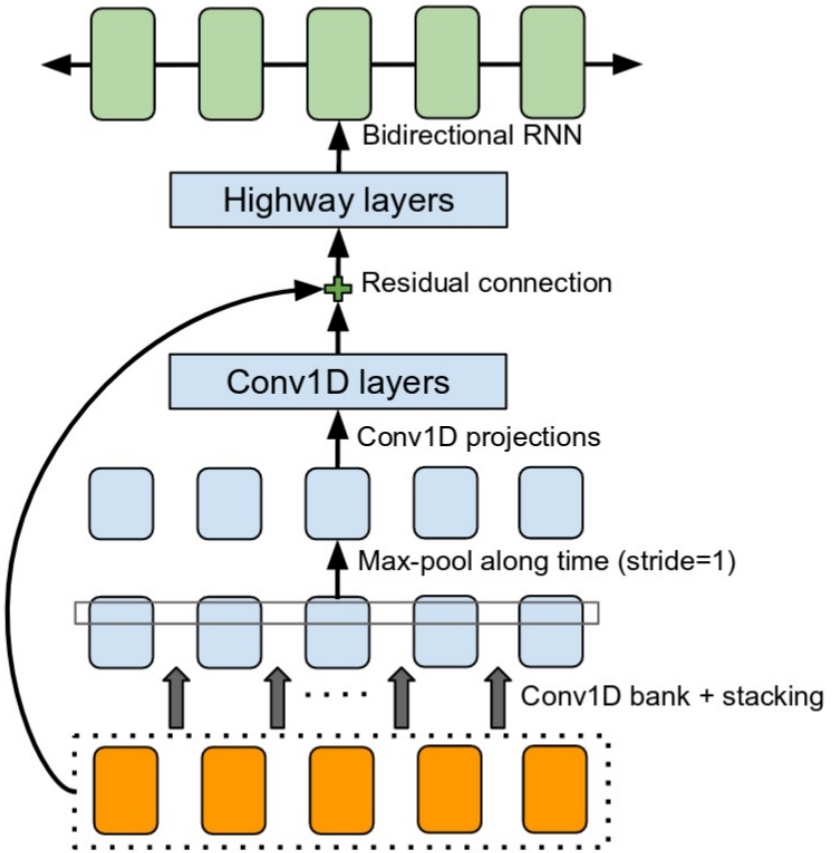


Модель Такотрона

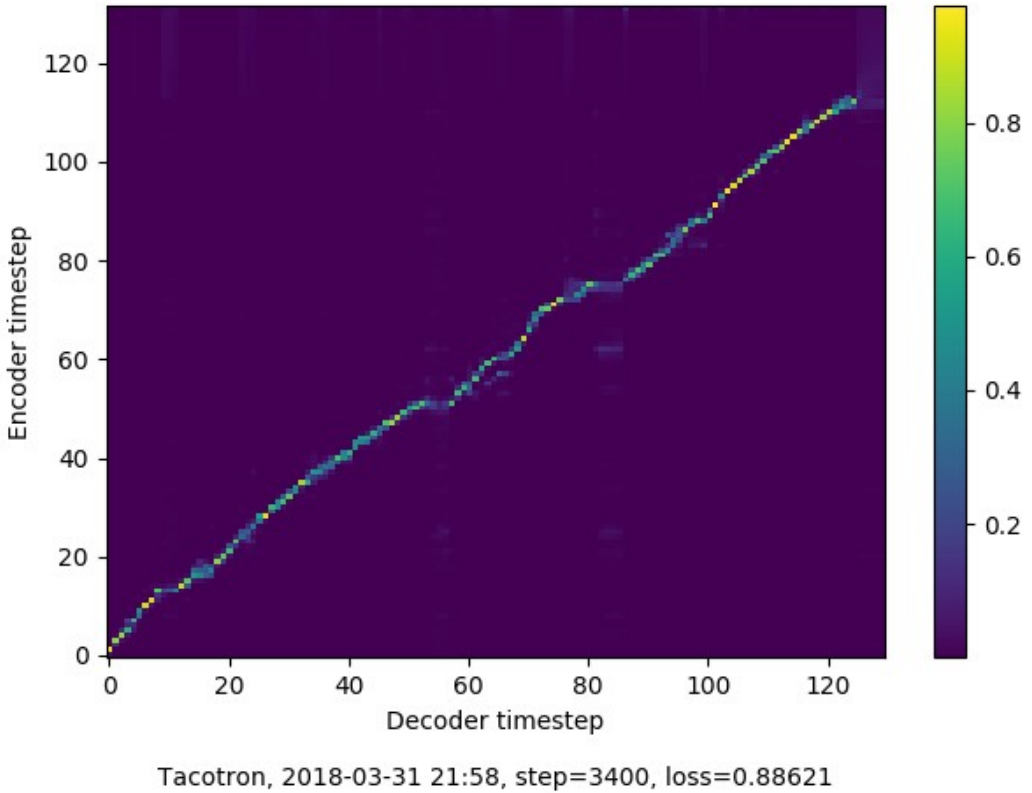
- Encoder
- Attention-based decoder
- Post-processing net


Griffin-Lim reconstruction
Linear-scale spectrogram
CBHG
Seq2seq target with r=3
Decoder RNN
Decoder RNN
Decoder RNN
Attention
Attention RNN
Attention RNN
Attention RNN
Attention is applied to all decoder steps
Pre-net
Pre-net
Pre-net
<GO> frame
CBHG
Pre-net
Character embeddings

CBHG block


Bidirectional RNN
Highway layers
Residual connection
Conv1D layers
Conv1D projections
Max-pool along time (stride=1)
Conv1D bank + stacking

Attention


Encoder timestep
0
20
40
60
80
100
120
Decoder timestep
0
20
40
60
80
100
120
0.8
0.6
0.4
0.2
Tacotron, 2018-03-31 21:58, step=3400, loss=0.88621

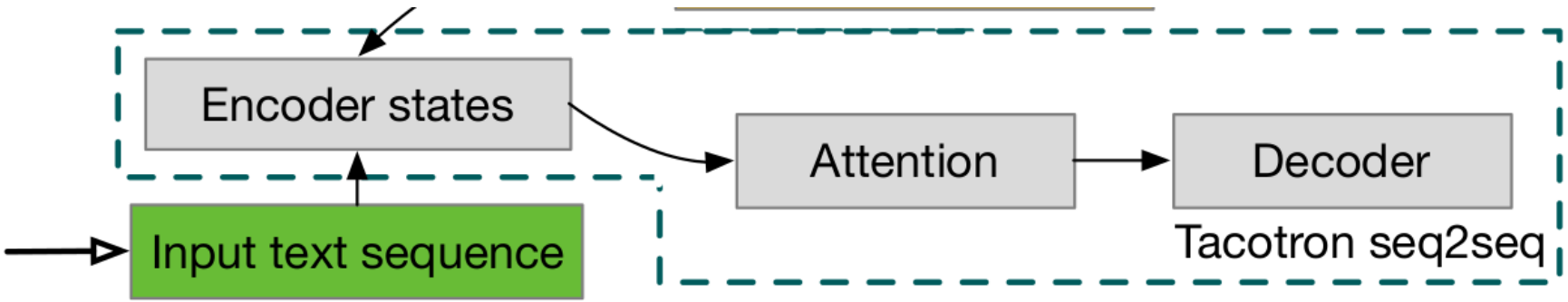
Вокодеры

- Griffin-Lim
- WaveNet
- Parallel WaveNet

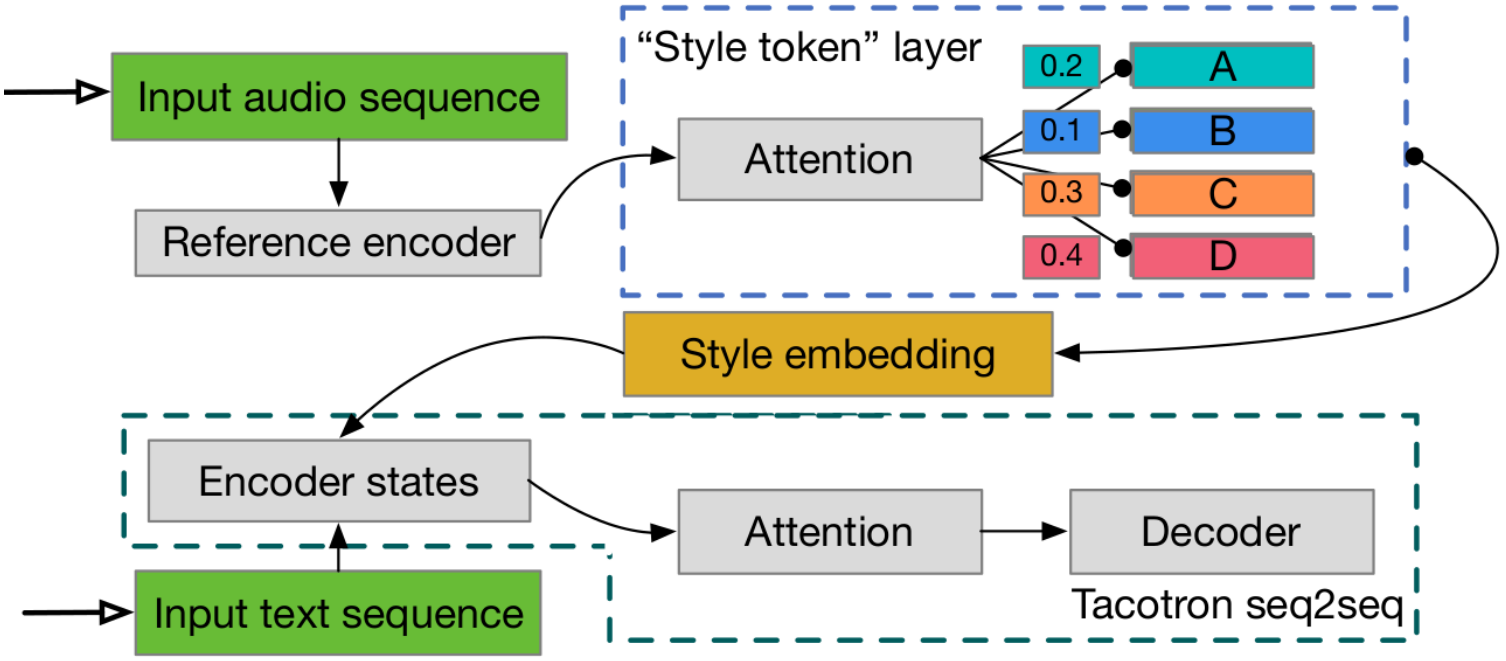
System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

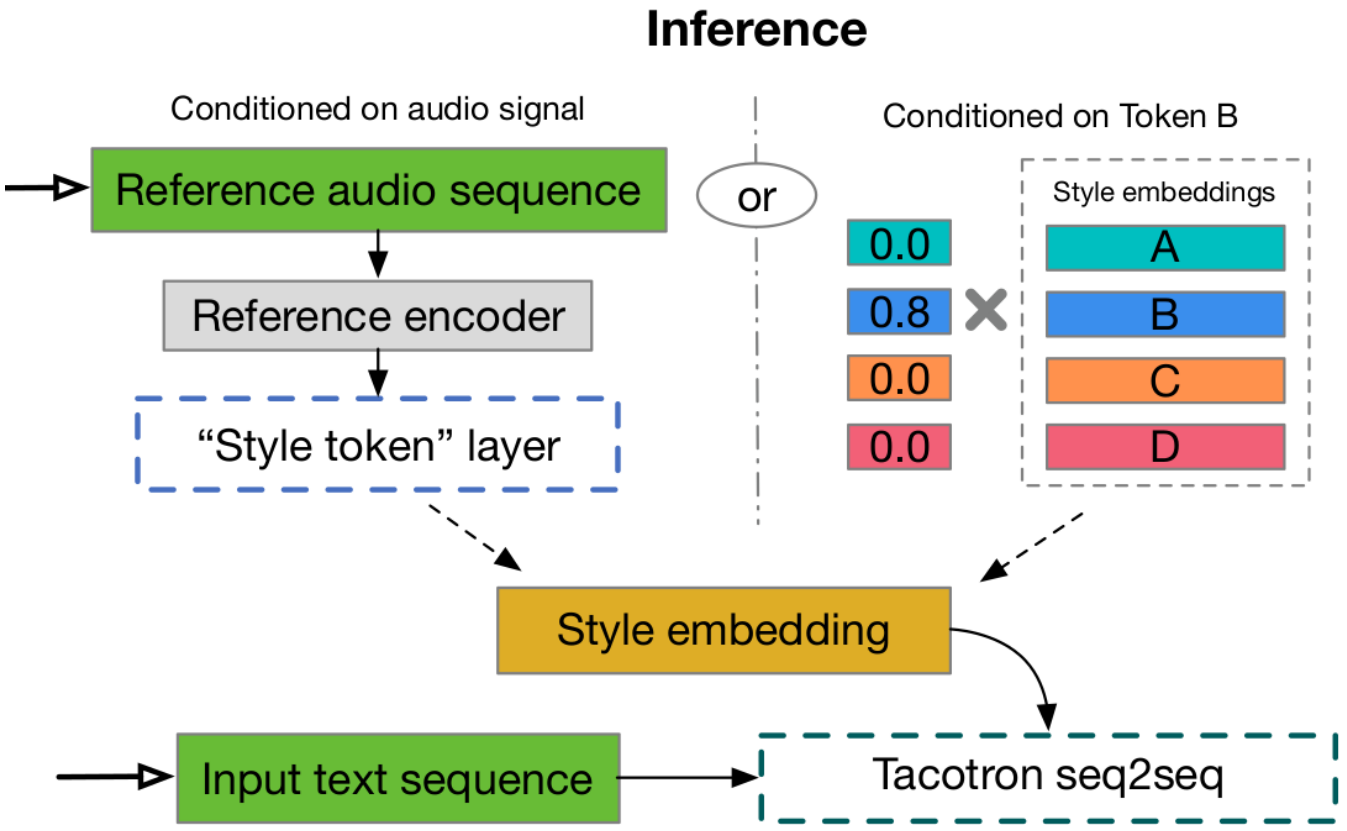
Prosody modeling

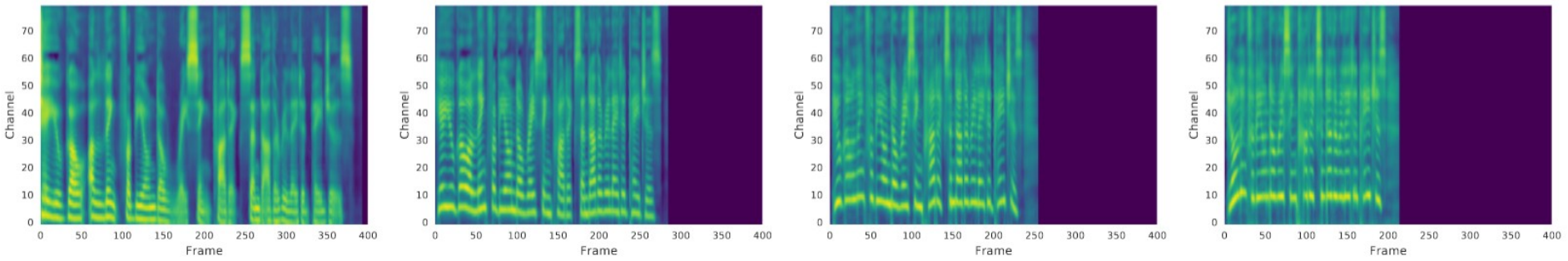
- В тексте нет информации о:
  - Интонациях
  - Особенностях голоса
  - Эффектов окружения


Encoder states
Attention
Decoder
Input text sequence
Tacotron seq2seq

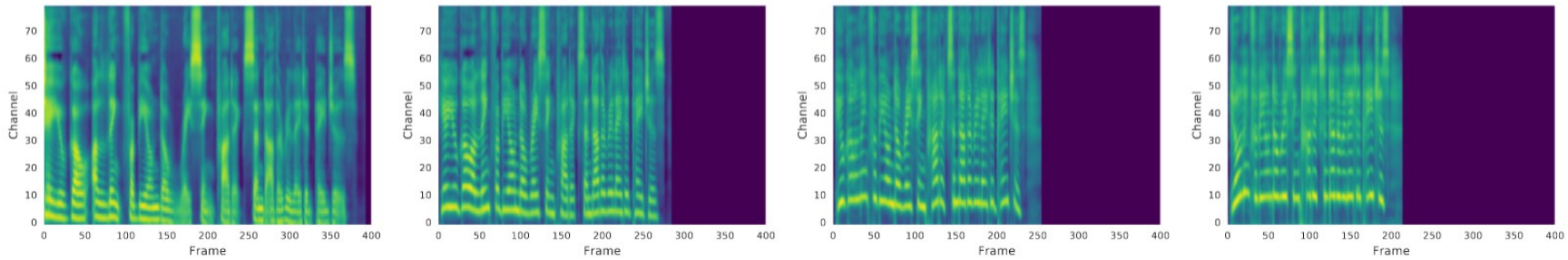
Training


Input audio sequence
Reference encoder
“Style token” layer
Attention
0.2
A
0.1
B
0.3
C
0.4
D
Style embedding
Encoder states
Attention
Decoder
Input text sequence
Tacotron seq2seq

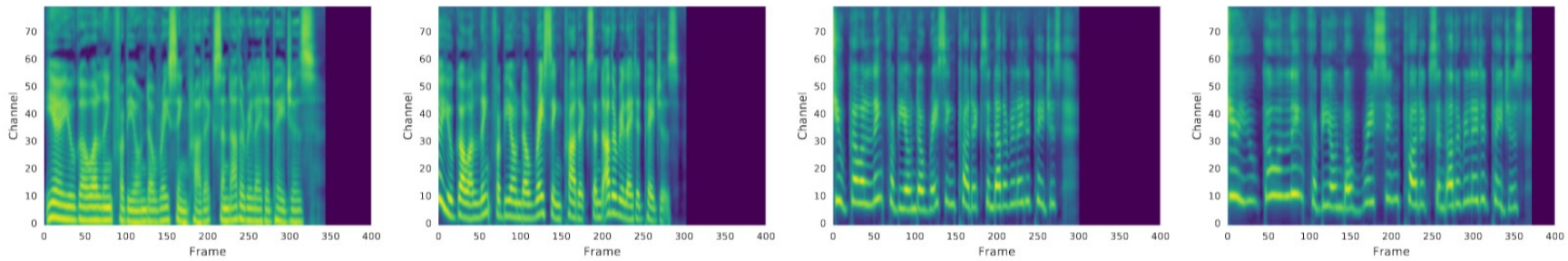

Inference
Conditioned on audio signal
Reference audio sequence
Reference encoder
“Style token” layer
or
Conditioned on Token B
Style embeddings
0.0
0.8
0.0
0.0
A
B
C
D
Style embedding
Input text sequence
Tacotron seq2seq



(a) Token A (speed)



(a) Token A (speed)



(b) Token B (animated)

NOISE %	BASELINE TACOTRON	GST
50%	2.819 ± 0.269	4.080 ± 0.075
75%	1.819 ± 0.227	3.993 ± 0.074
90%	1.609 ± 0.131	4.031 ± 0.082
95%	1.353 ± 0.090	3.997 ± 0.066

- https://google.github.io/tacotron/ - статьи по такотрону.