Синтез речи

Лекция №3

Гриша Стерлинг, SberDevices

1. Tacotron 2
2. Global Style Tokens (GST)
3. Multispeaker TTS
4. Multilanguage TTS
5. Неавторегрессивный TTS

Данные для TTC

Single speaker:
- LJSpeech
- HiFiTTS

Multi speaker:
- VCTK
- LibriTTS

Дополнительная разметка:
- Montreal Forced Aligner (MFA)
- Reaper
- Audacity

Tacotron 2

seq2seq:

encoder + attention + decoder
+ postnet

+WaveNet vocoder

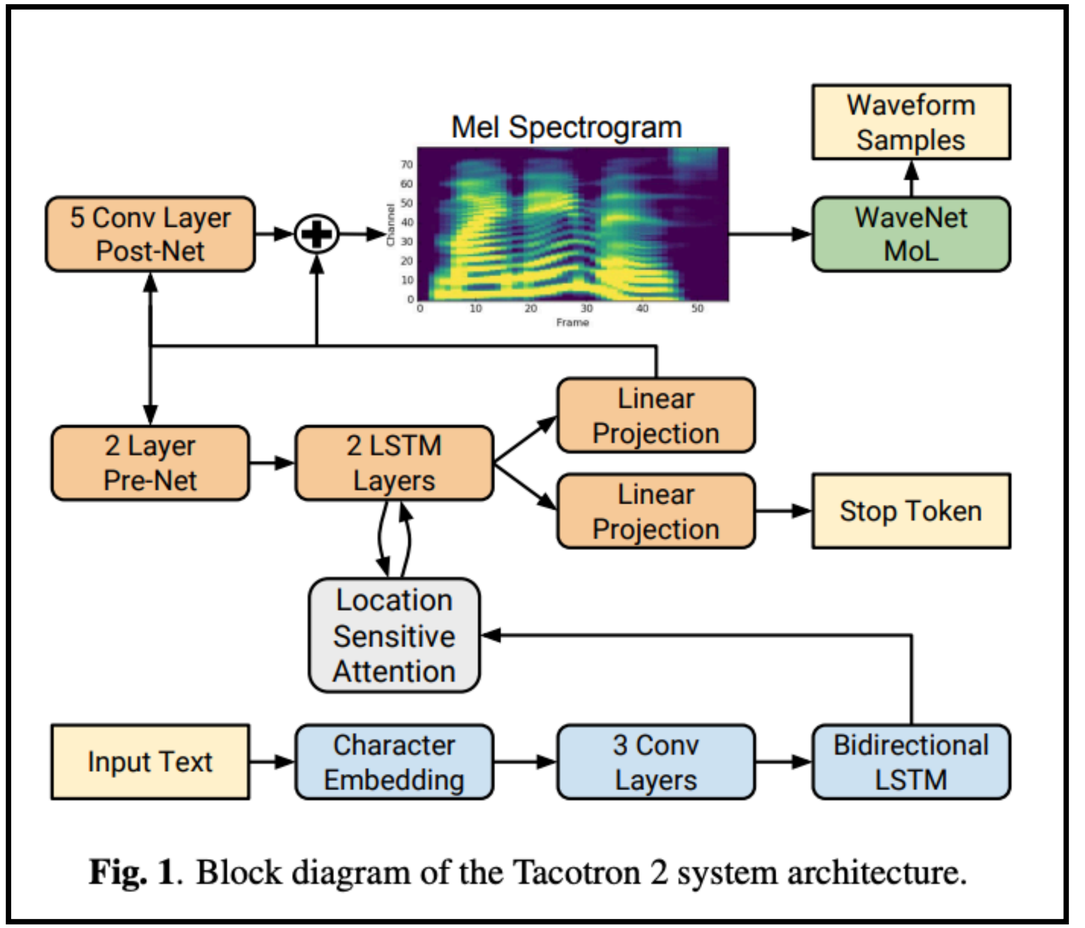
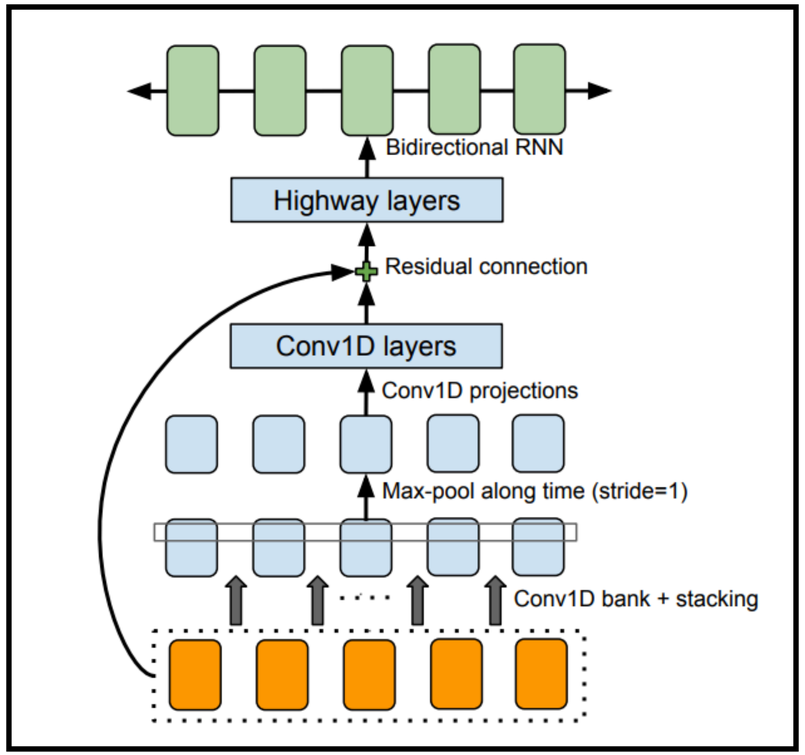


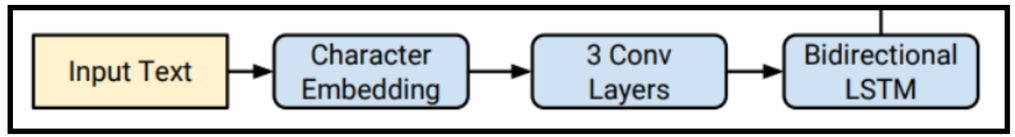
Fig. 1. Block diagram of the Tacotron 2 system architecture.

Tacotron 2 encoder

Tacotron encoder



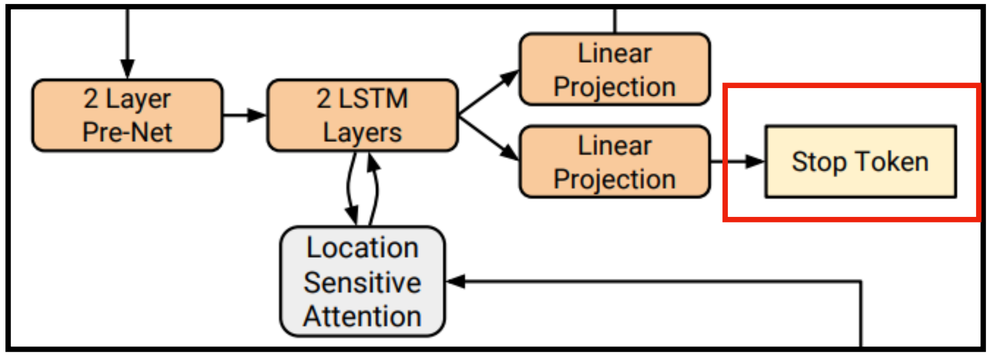
Tacotron 2 encoder



- нет FC после embedding слоя
- нет ResCon и Highway layers

Tacotron 2 decoder

LSTM input =
Concat(Prenet(last_frame), context)



encoder
out

Teacher forcing:

batch_size x num_letters x 512 -> batch_size x 512

Tacotron 2 attention

Location sensitive attention:

Bahdanau:

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

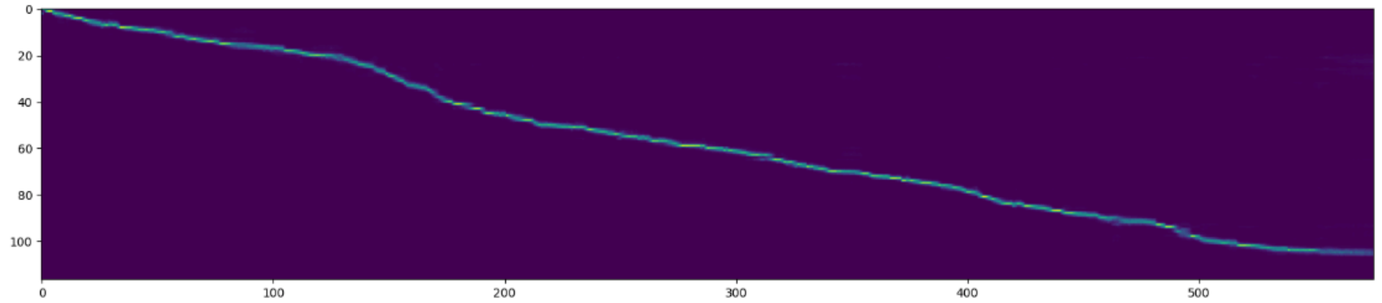
scores:

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

location features:

$$f_i = F * \alpha_{i-1}.$$

Всем привет



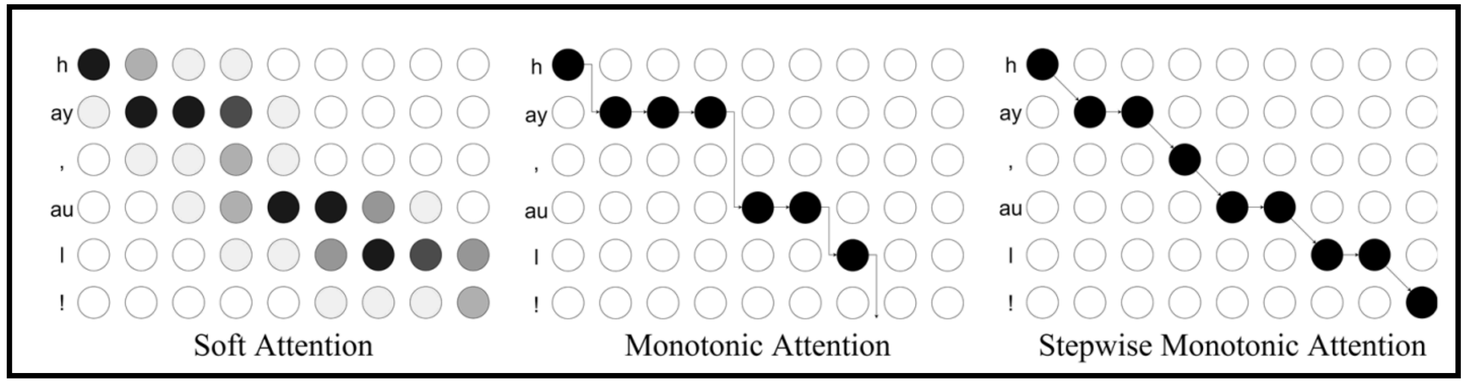
weights:

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T_x} \exp\left(e_{ik}\right)},$$

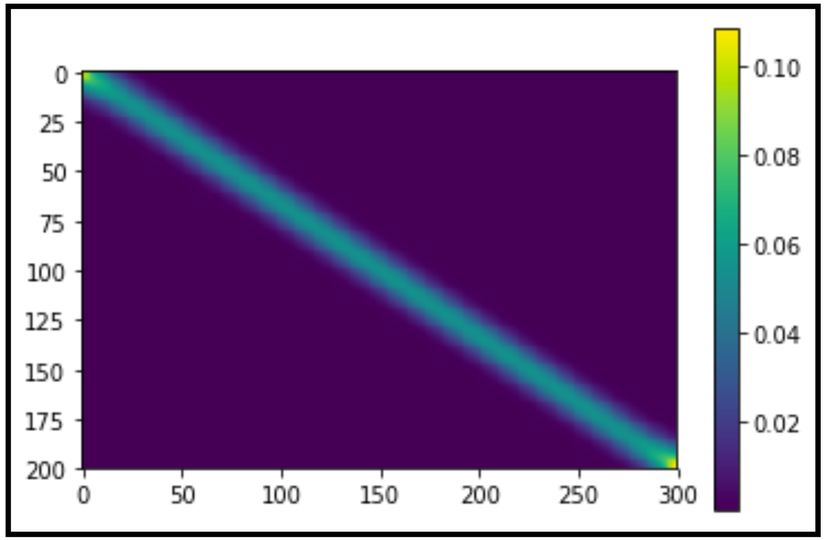
context = sum(encoder_out * alpha)

Tacotron 2 другие attention механизмы

Решаются проблемы:
- артефакты
- сходимость
- длинные предложения



Новые проблемы:
- энкодер учится хуже
- монотонность речи
- контекст вектор локальный



Tacotron 2 inference

Dropout:

«In order to introduce output variation at inference time, dropout with probability 0.5 is applied only to layers in the pre-net of the autoregressive decoder»

$$D\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i^2 D[X_i] + 2 \sum_{1 \leqslant i < j \leqslant n} c_i c_j \operatorname{cov}(X_i, X_j),$$

$$D[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

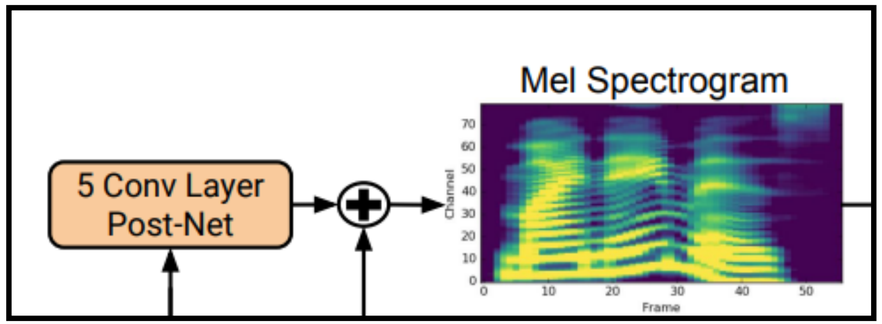
- без дропаута плохо говорит
- синтез каждый раз разный

Vocoder fine-tune:

1.

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

2.



Просодия

Prosody

Речь = кто + что + **как**

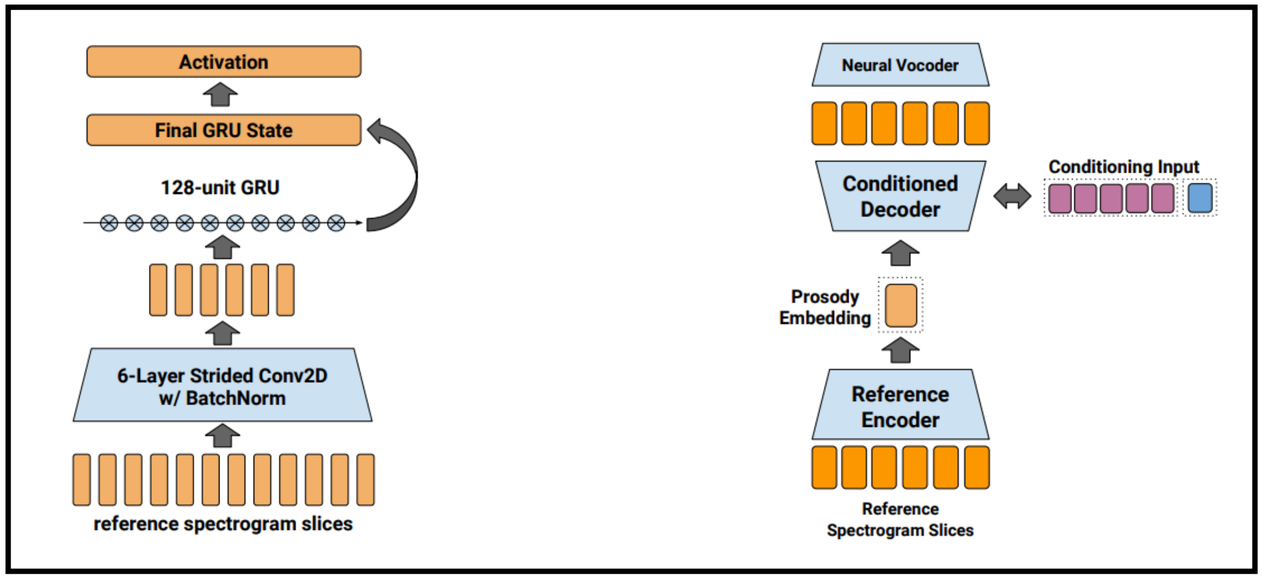
как = высокоуровневая просодия + низкоуровневая просодия

- эмоция
- громкость
- скорость
- тон

- эмфаза
- вопросы
- паузы

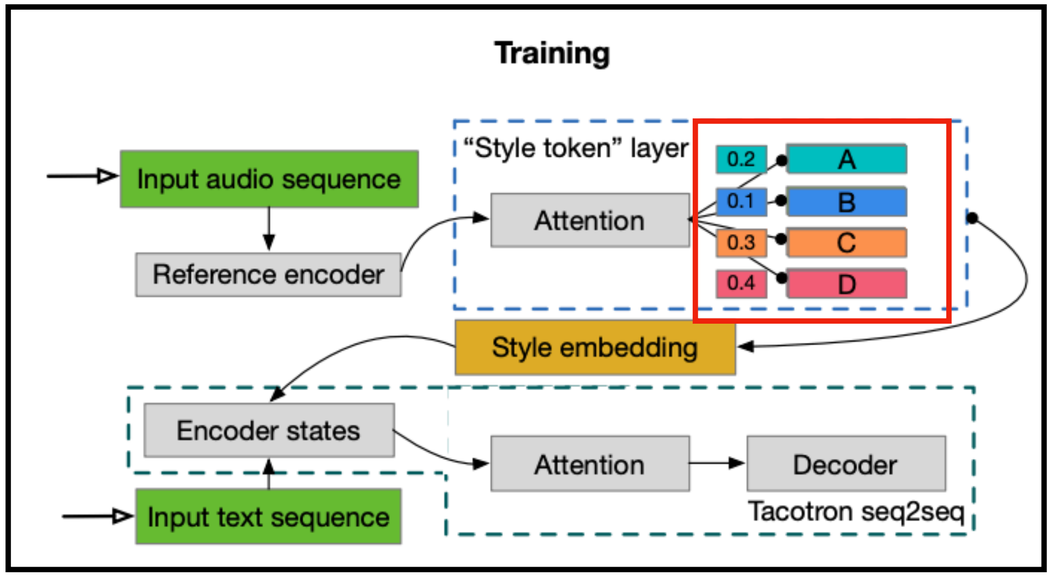
Global Style Tokens

Reference encoder:



ground truth spec -> style vector

GST:



ground truth spec -> mixture of styles

Global Style Tokens

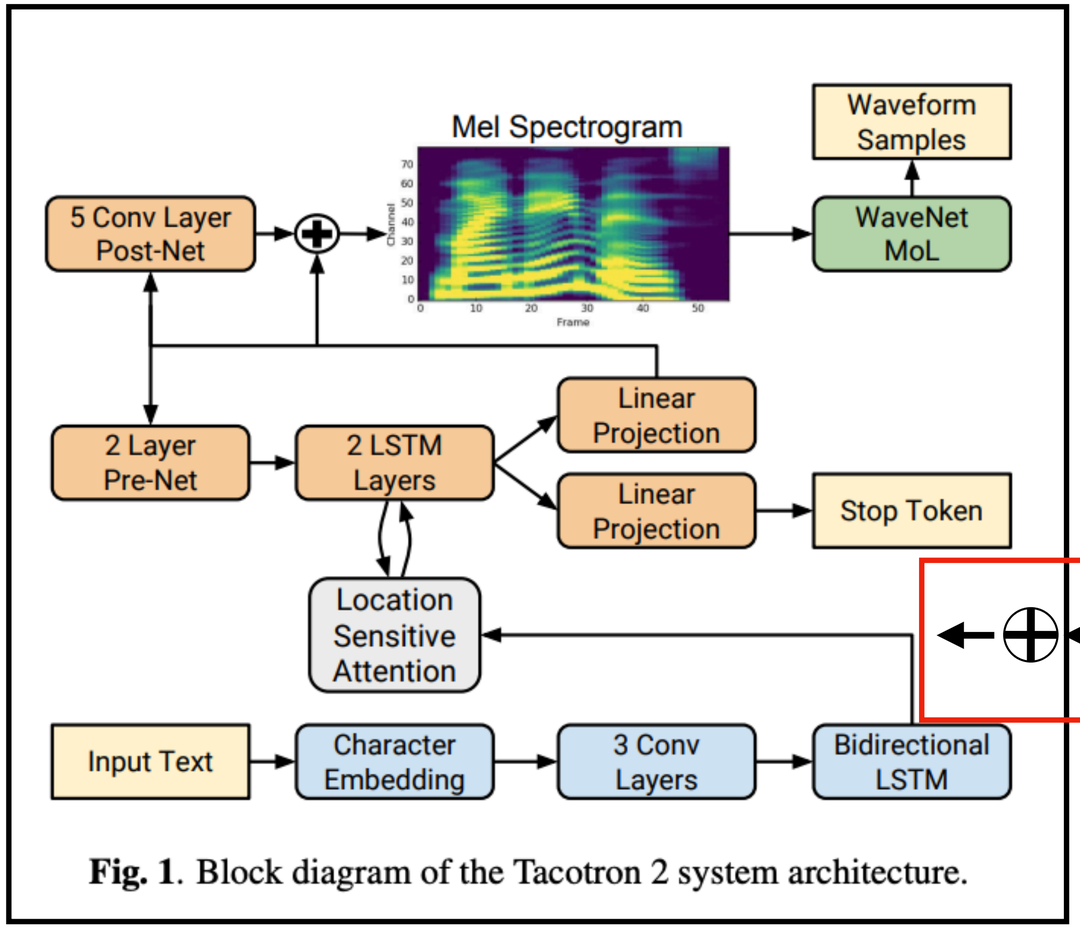


Fig. 1. Block diagram of the Tacotron 2 system architecture.

style embedding

Профит:

- style transfer
- лучше учится
- семплирование из «возможных способов озвучить текст»

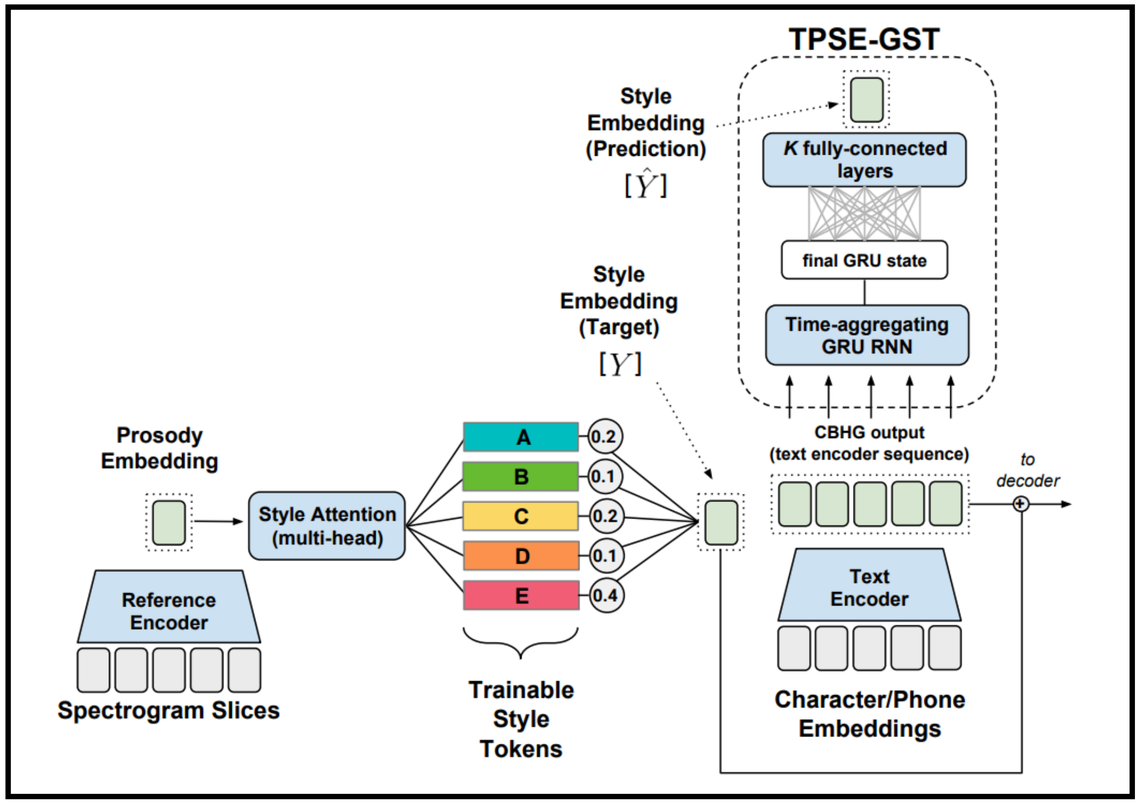
энкодер = текст энкодер + стиль

Global Style Tokens

Проблемы:

- не воспроизводится :)
- стиль выучивает длину, громкость и тон
- не интерпретируется
- неоткуда брать референс

Text predicted style embedding:



Multispeaker TTS

Задачи:

seen2seen - обычный ms tts

seen2unseen - thisvoicedoesnotexist

unseen2unseen - zero shot voice transfer

Схема решения:

конкат speaker embedding
к энкодеру

Multispeaker TTS

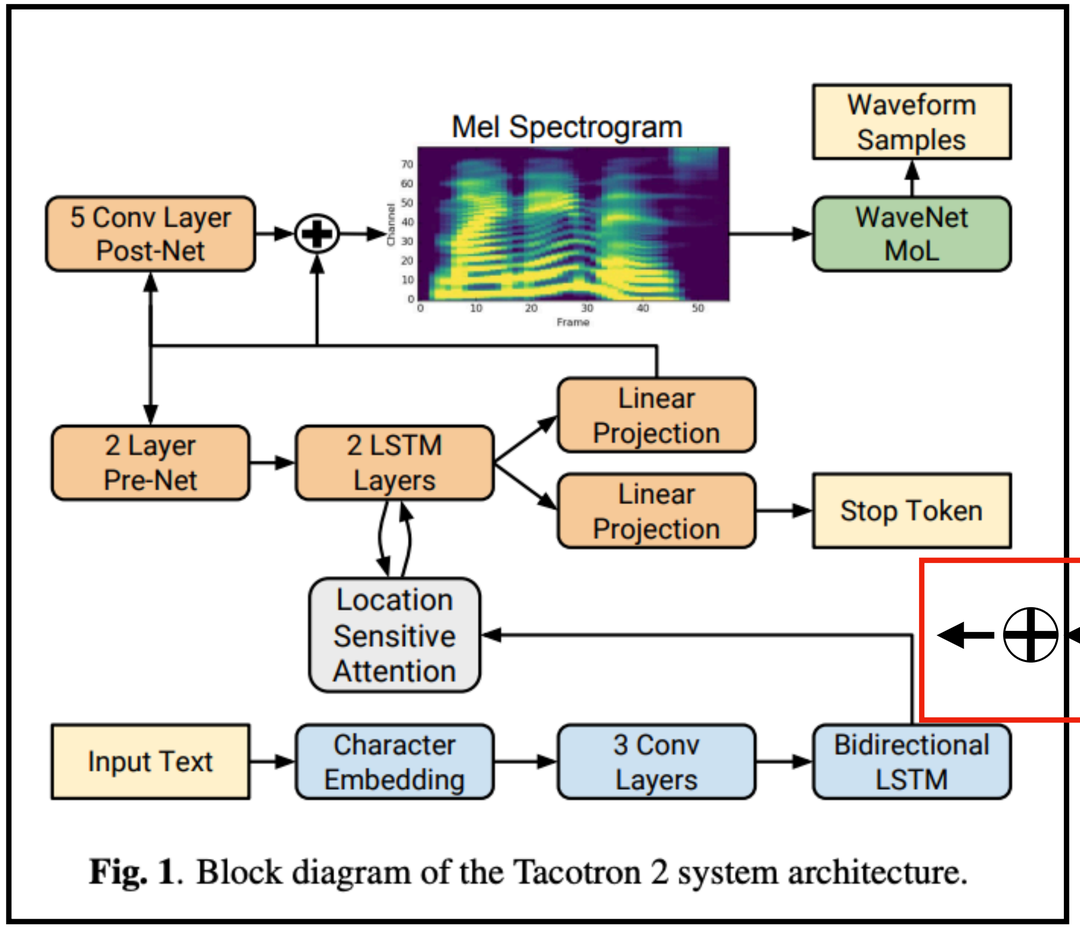


Fig. 1. Block diagram of the Tacotron 2 system architecture.

speaker embedding

- one-hot
- speaker verification
- learnable dictionary encoding (LDE)

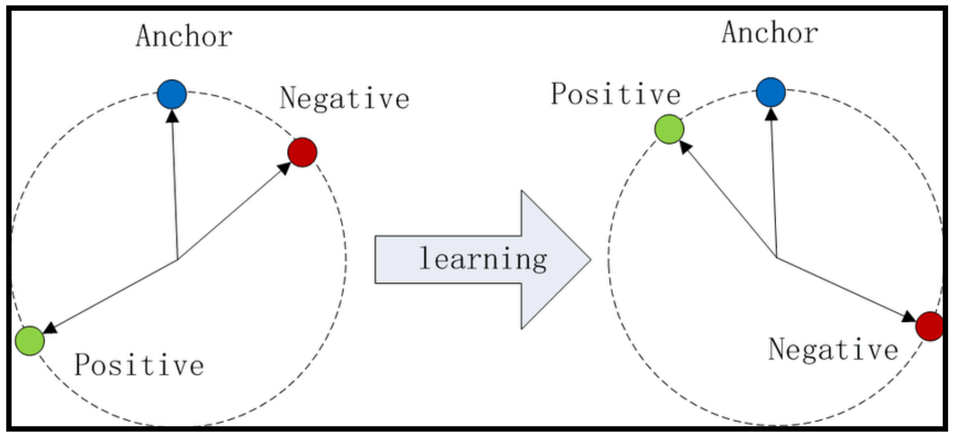
Multispeaker TTS

one-hot speaker embedding:

[num_speakers x embedding_dim]

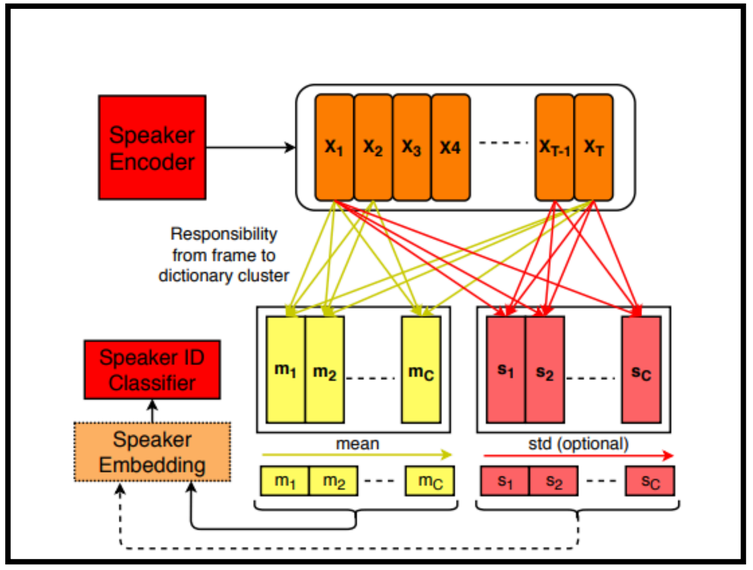
speaker verification:

- benchmark: VoxCeleb (1, 2)
- transfer learning
- metric learning



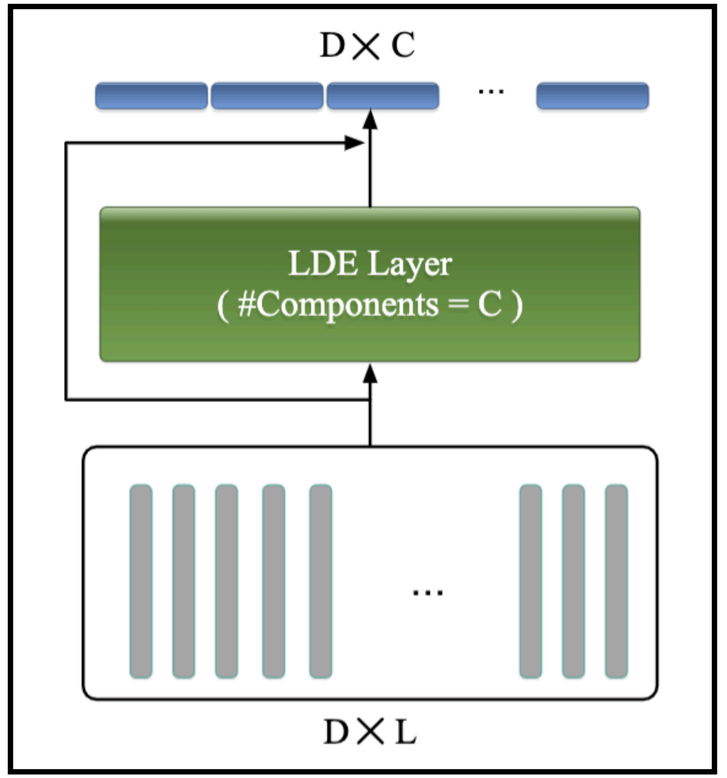
Multispeaker TTS

speaker embedding = mixture of encodings



k-means

Vanilla LDE:



Multilanguage TTS

Speaking fluently in foreign language

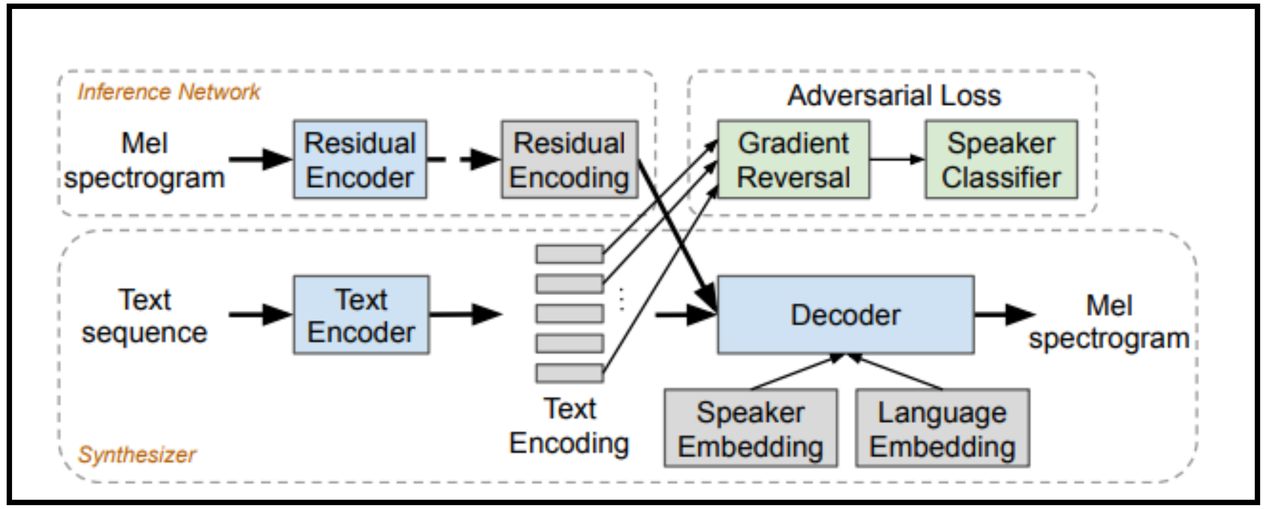
train:

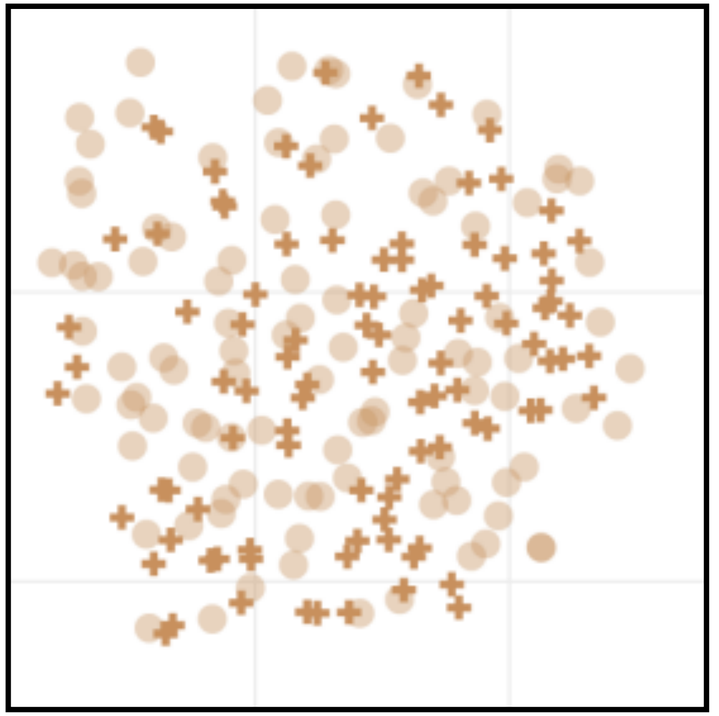
X, **en**
Y, ru

inference:

Y, en

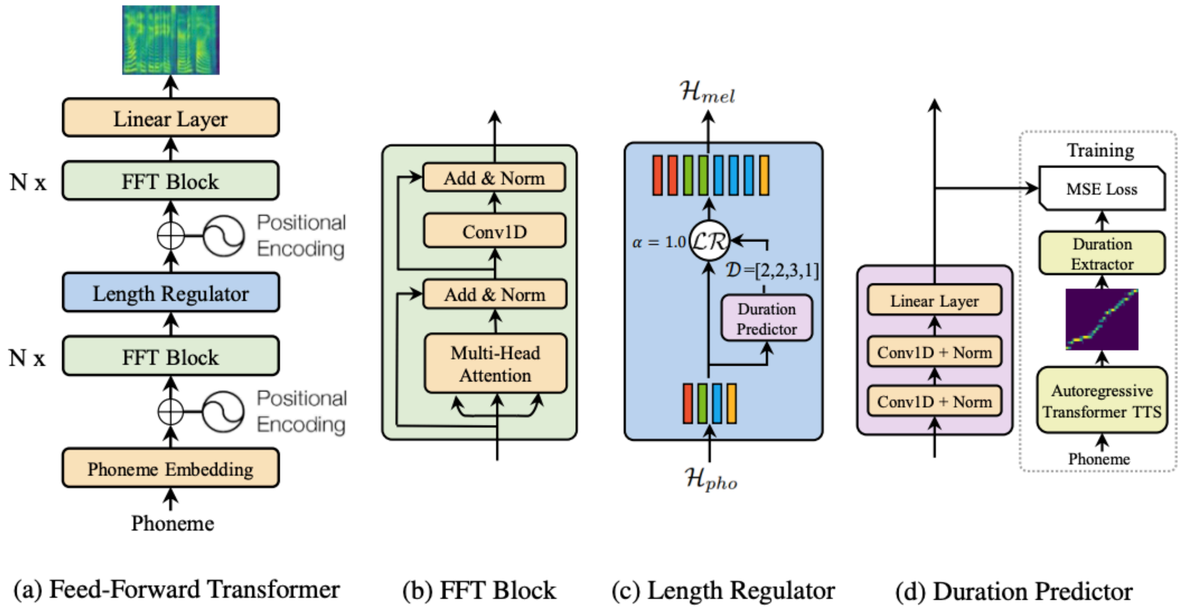
«прив**е**т мир»
«h**e**llo world»





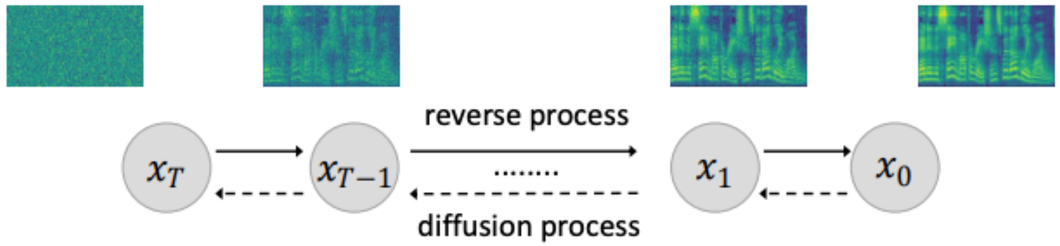
Non-autoregressive TTS

Fastspeech-family:
transformer + duration predictor



(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator (d) Duration Predictor

Diffusion-based:



Спасибо :)