

Синтез речи

Лекция №2

Гриша Стерлинг, SberDevices

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

Данные для ТТС

Single speaker:

- LJSpeech
- HiFiTTS

Multi speaker:

- VCTK
- LibriTTS

Дополнительная разметка:

- Montreal Forced Aligner (MFA) hmm-gmm
- Reaper, pyreaper, CREPE
- Audacity
- Yin algorithm (плохой)

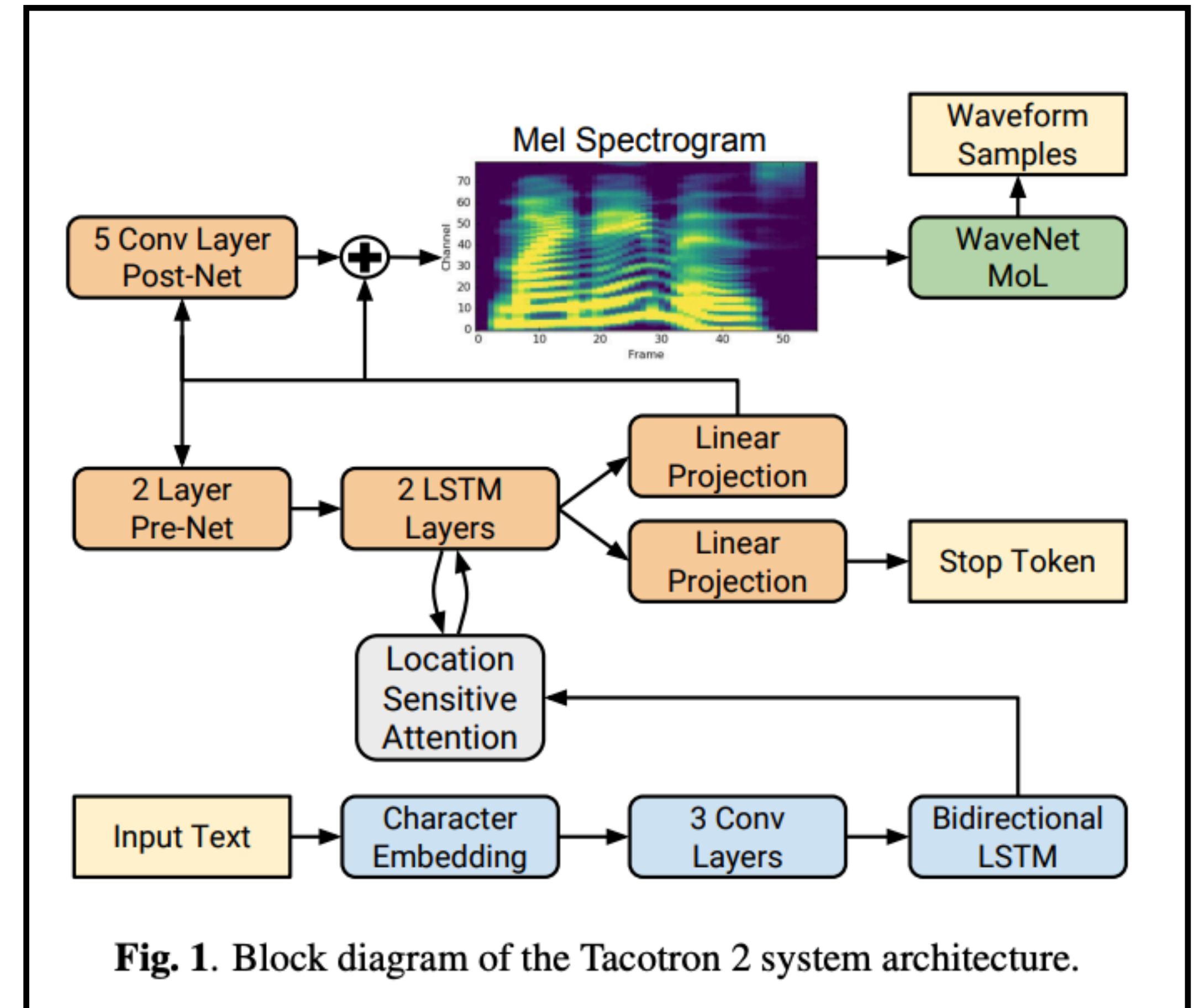
Tacotron 2

seq2seq:

encoder + attention + decoder
+ postnet

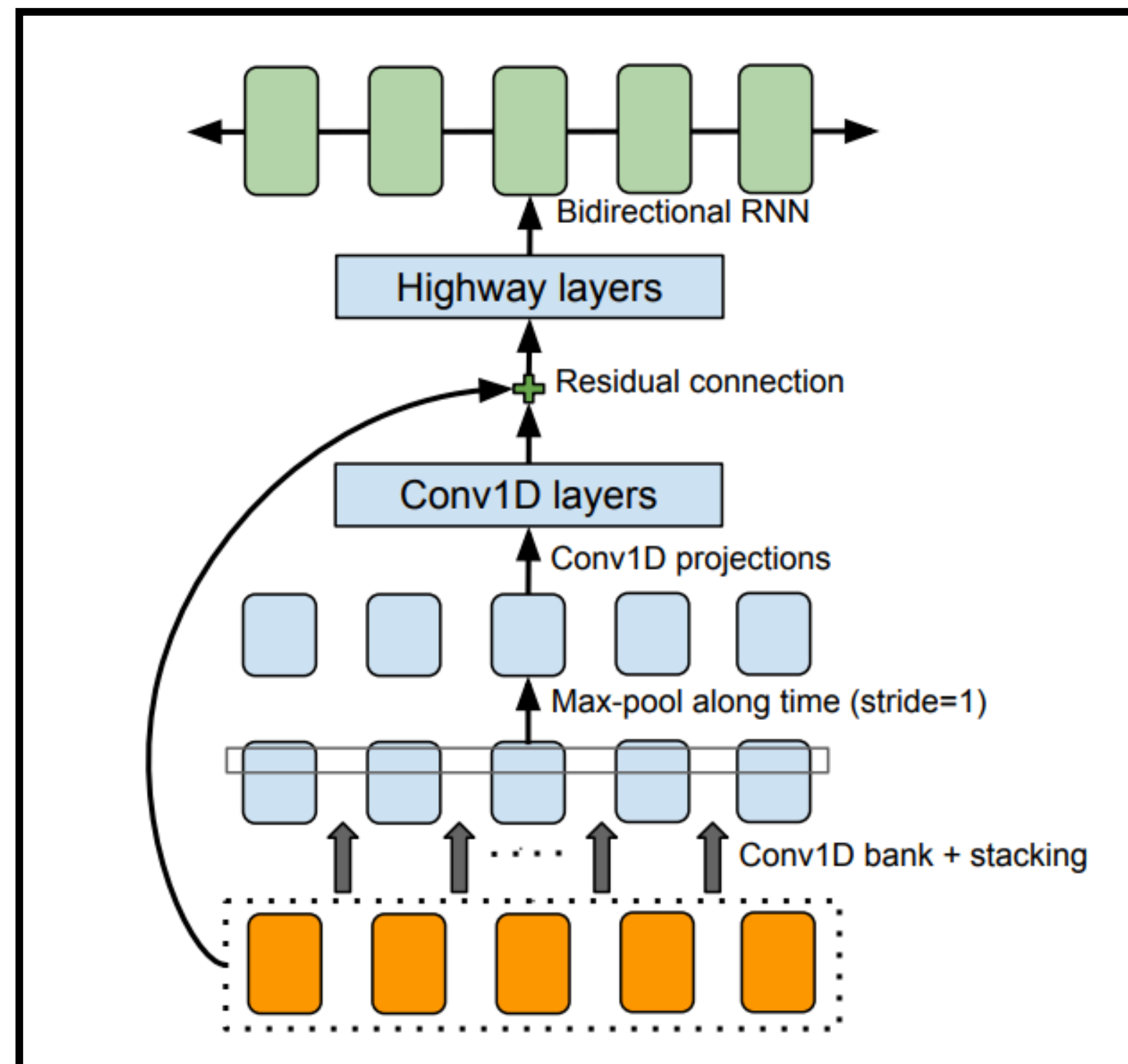
$B \times T \times 512$

+WaveNet vocoder

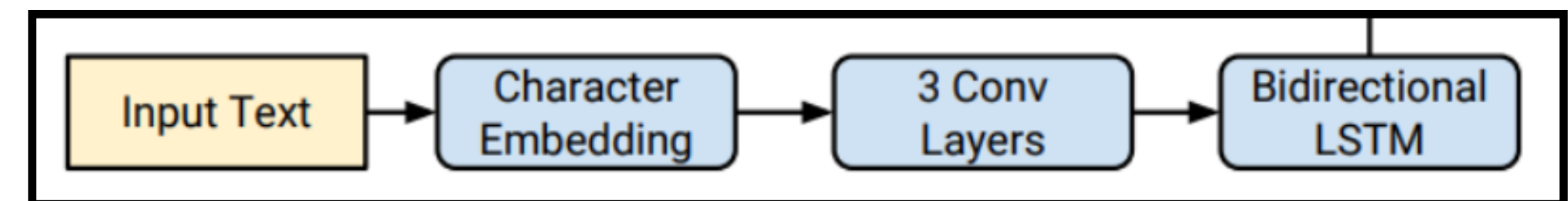


Tacotron 2 encoder

Tacotron encoder



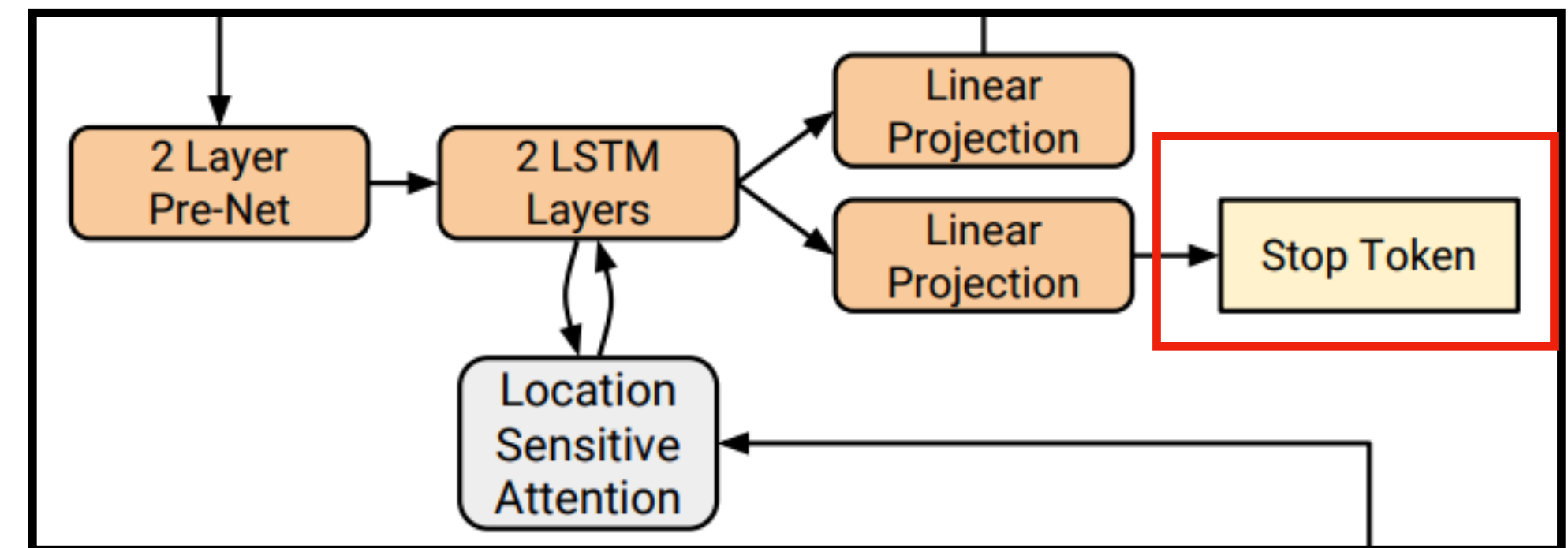
Tacotron 2 encoder



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

Tacotron 2 decoder

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



Teacher forcing:

batch_size x num_letters x 512 -> batch_size x 512

Tacotron 2 attention

Location sensitive attention:

scores:

$$e_{i,j} = w^T \tanh(Ws_{i-1} + Vh_j + Uf_{i,j} + b)$$

location features:

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

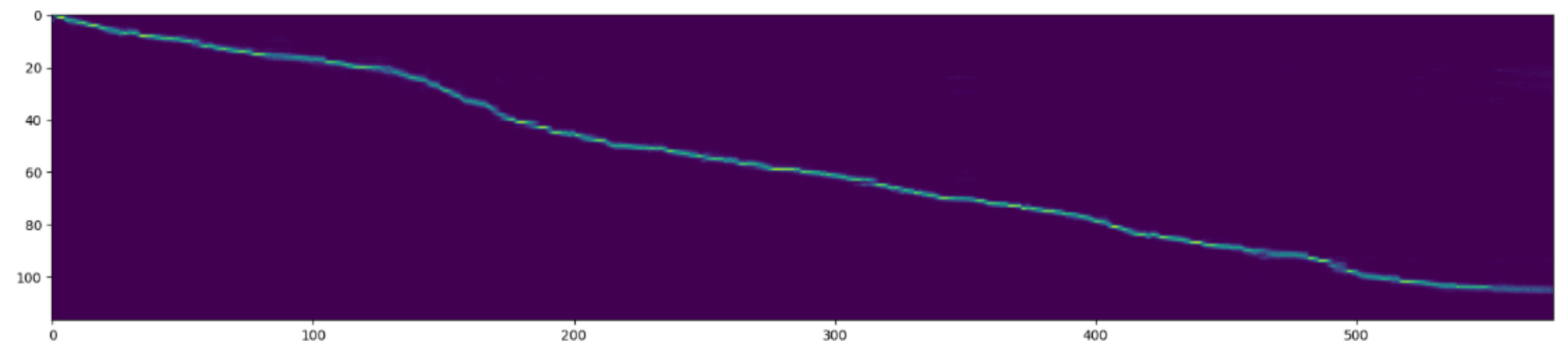
weights:

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

Bahdanau:

$$e_{i,j} = w^T \tanh(Ws_{i-1} + Vh_j + b)$$

Всем привет

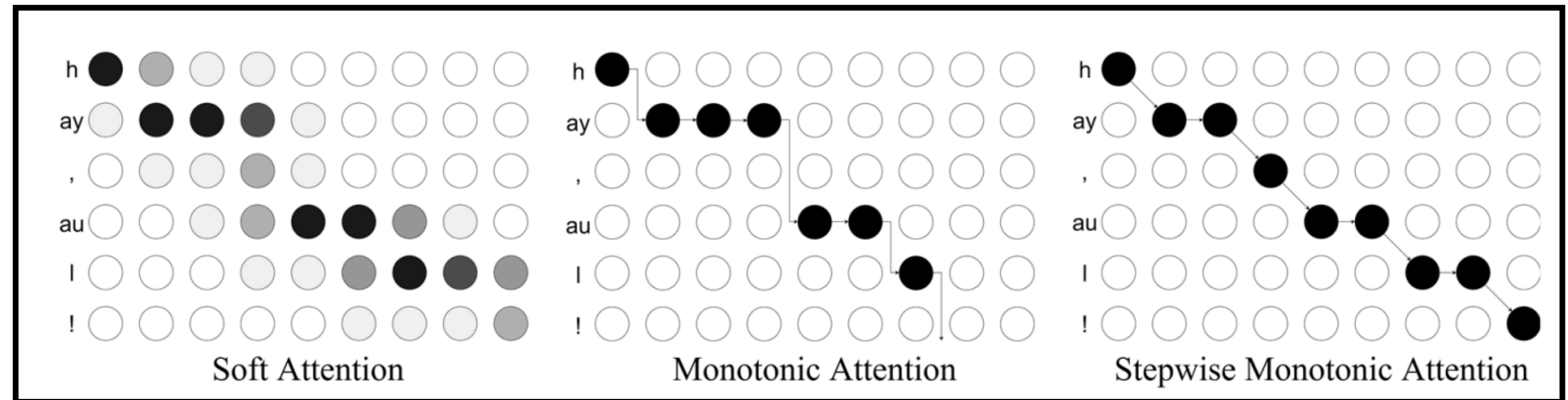


context = sum(encoder_out * alpha)

Tacotron 2 другие attention механизмы

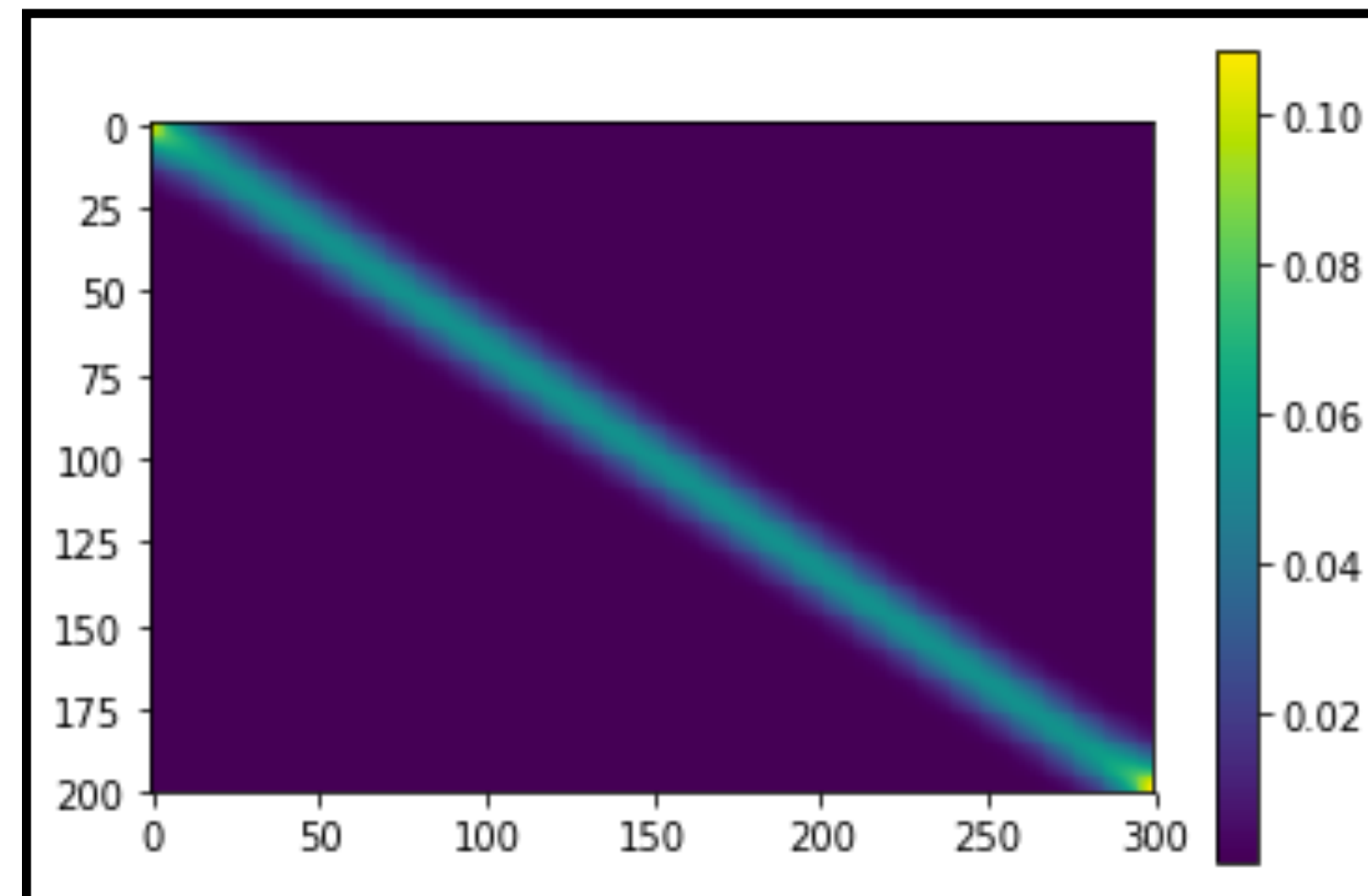
Решаются проблемы:

- артефакты
- сходимость
- длинные предложения



Новые проблемы:

- энкодер учится хуже
- монотонность речи
- контекст вектор локальный



Guided attention

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»

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

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

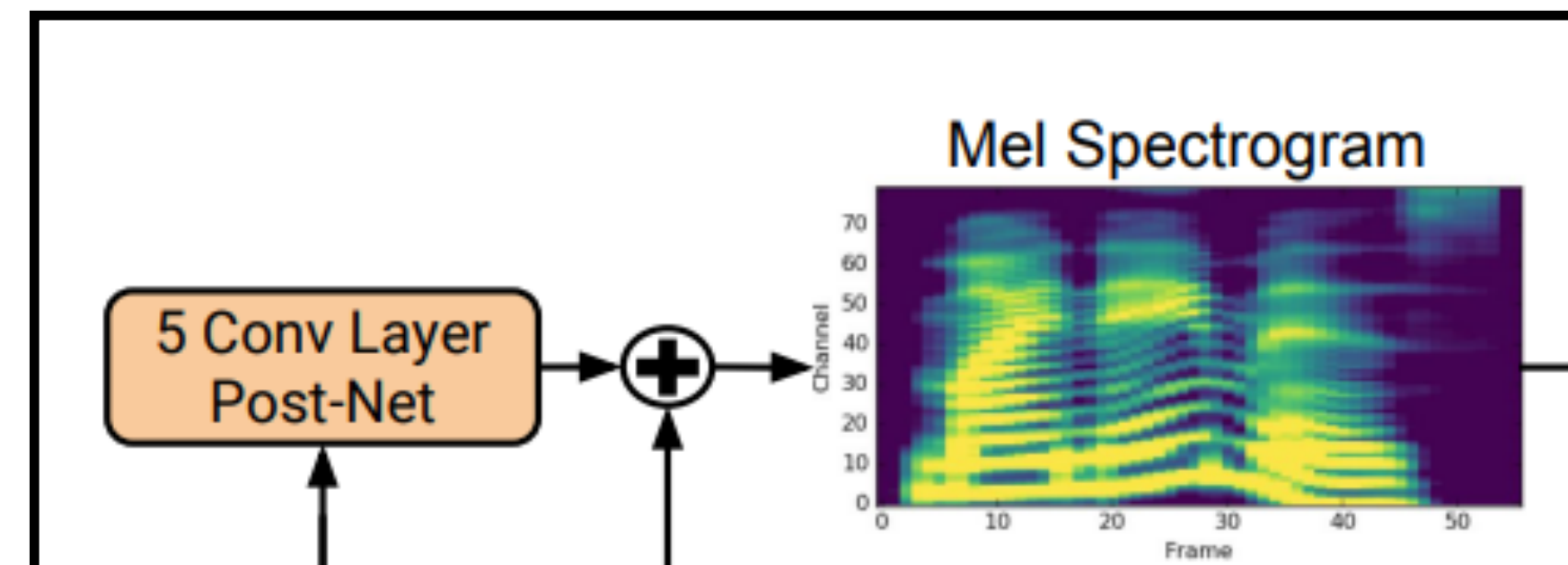
- без дропаута плохо говорит
- синтез каждый раз разный

Vocoder fine-tune:

1.

Training	Synthesis	
	Predicted	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**

q-Taco

как = высокоуровневая просодия + низкоуровневая просодия

- ЭМОЦИЯ
- громкость
- скорость
- ТОН

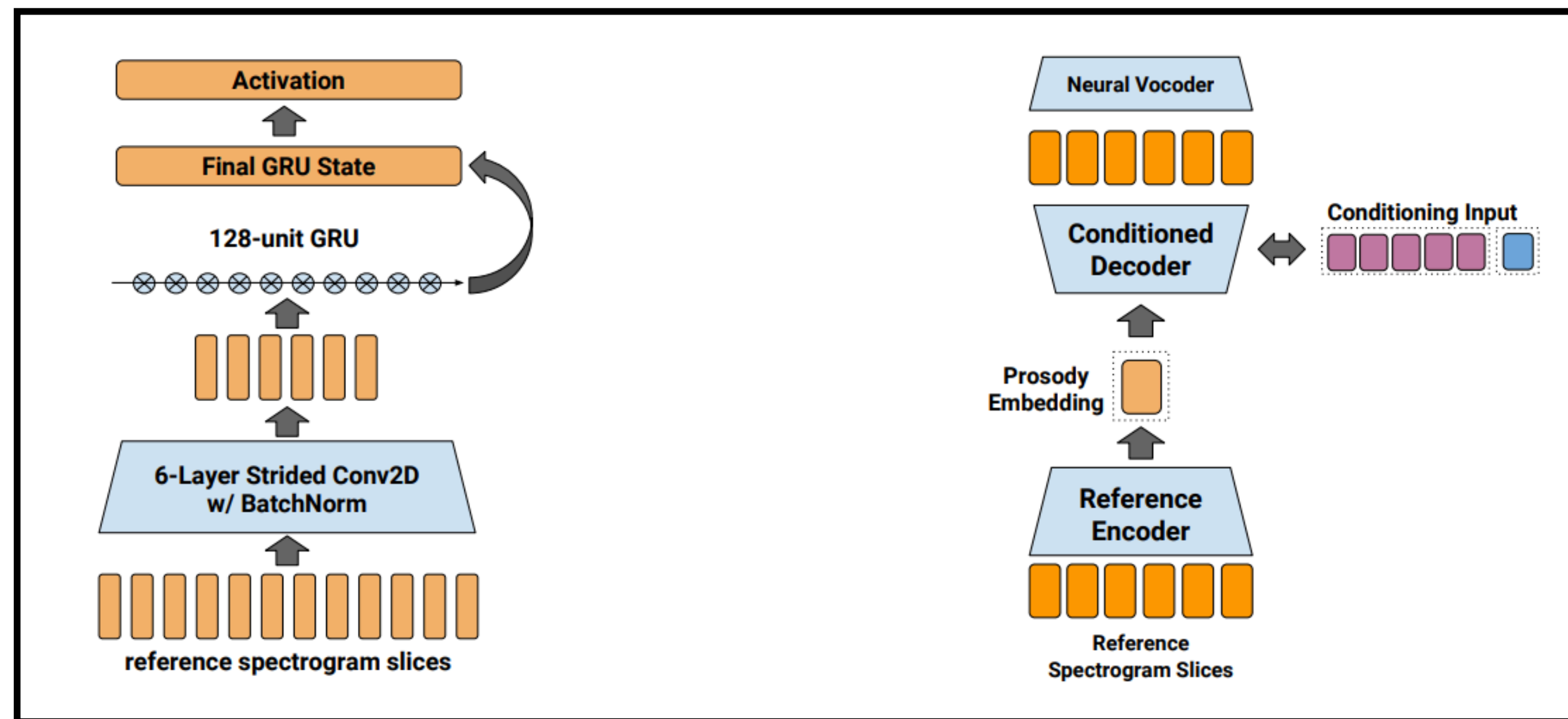
- эмфаза
- вопросы
- паузы

Будешь чай? Будешь чай?

Global Style Tokens

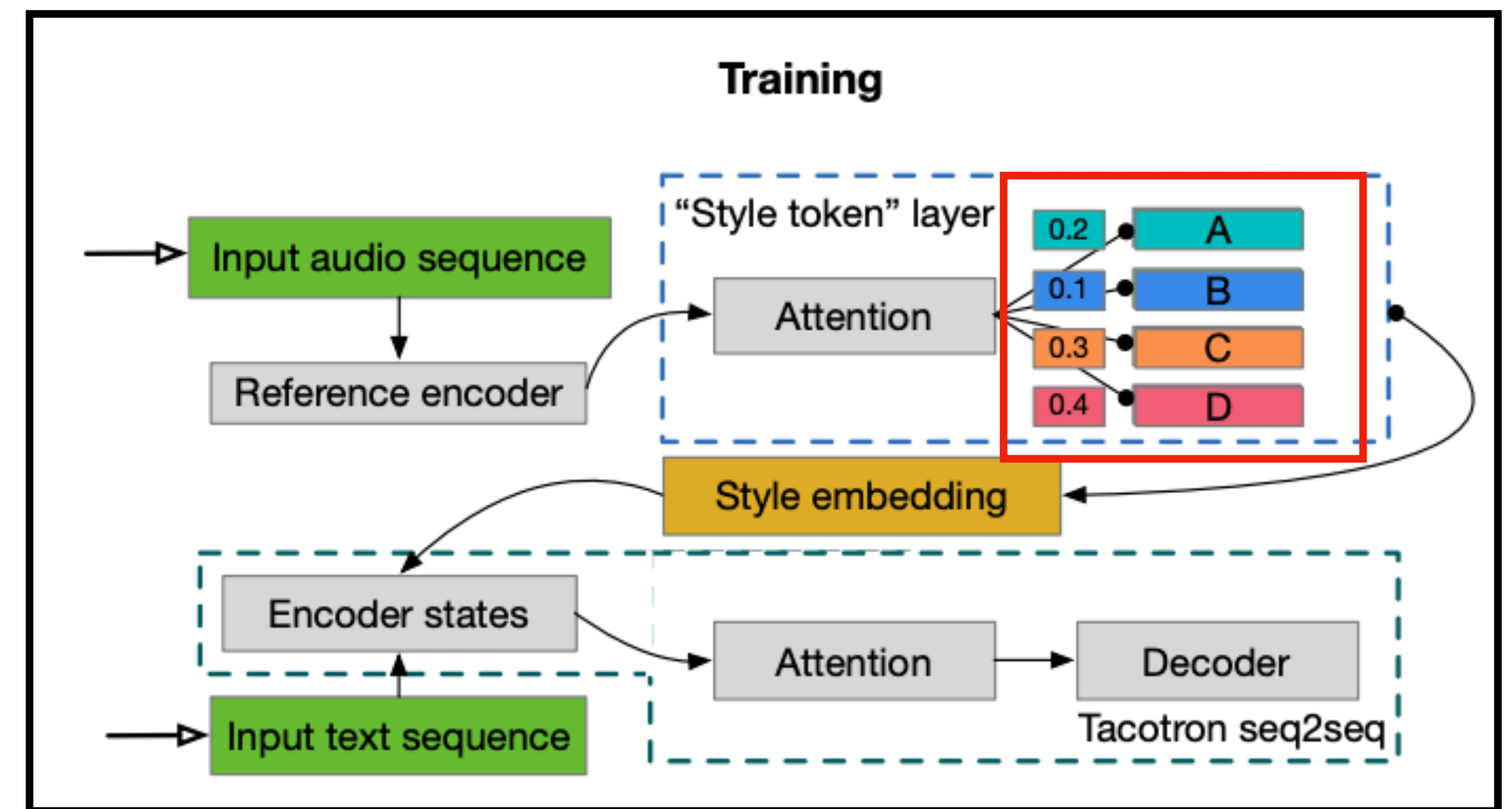
Blizzard challenge 2013

Reference encoder:



ground truth spec -> style vector

GST:



ground truth spec -> mixture of styles

Global Style Tokens

Профит:

- style transfer
- лучше учится
- семплирование из «ВОЗМОЖНЫХ СПОСОБОВ ОЗВУЧИТЬ ТЕКСТ»

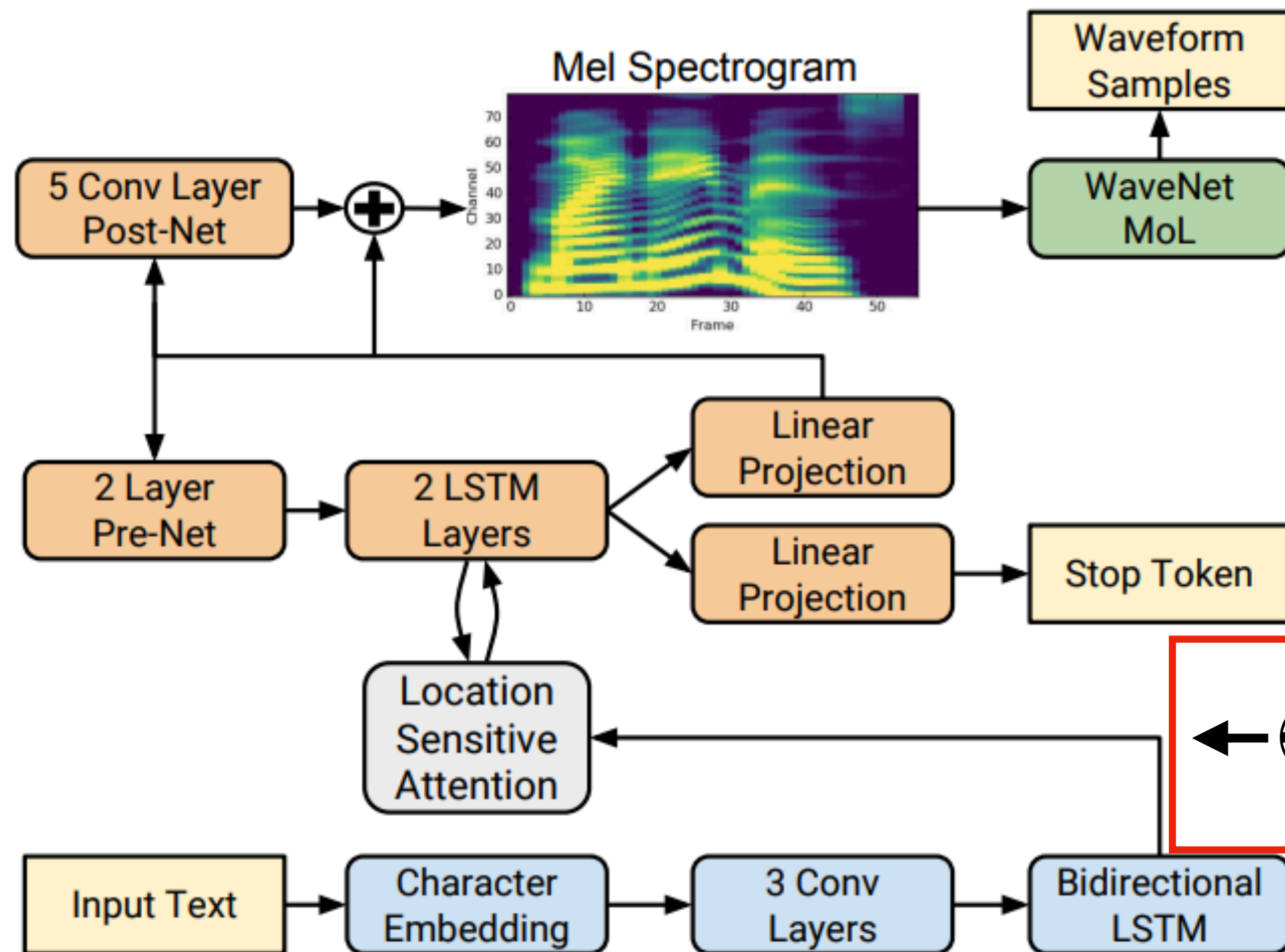


Fig. 1. Block diagram of the Tacotron 2 system architecture.

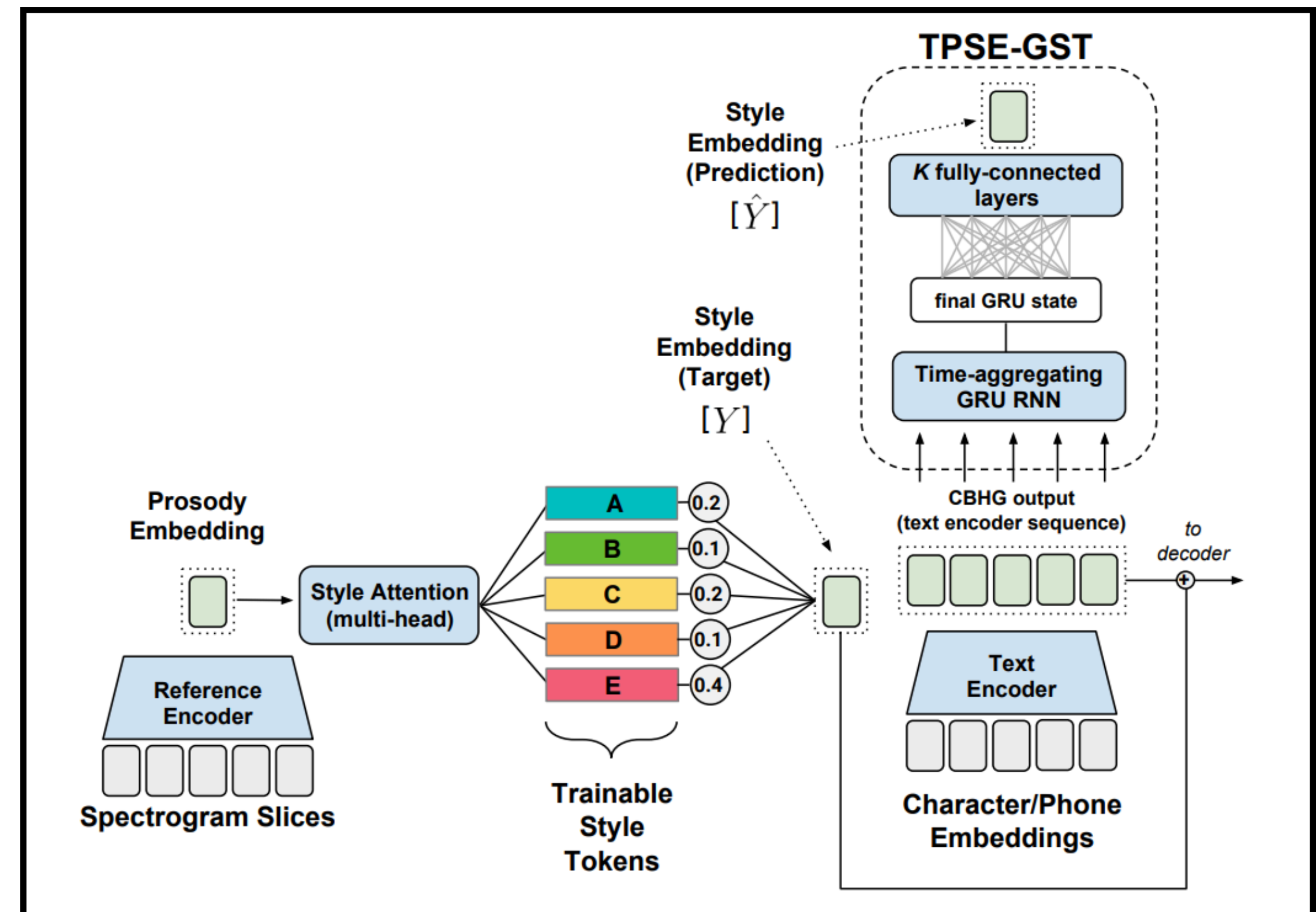
ЭНКОДЕР = ТЕКСТ ЭНКОДЕР + СТИЛЬ

Global Style Tokens

Text predicted style embedding:

Проблемы:

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



Multispeaker TTS

Задачи:

seen2seen - обычный ms tts

seen2unseen - thisvoicedoesnotexist

unseen2unseen - zero shot voice transfer

Схема решения:

конкат speaker embedding
к энкодеру

Multispeaker TTS

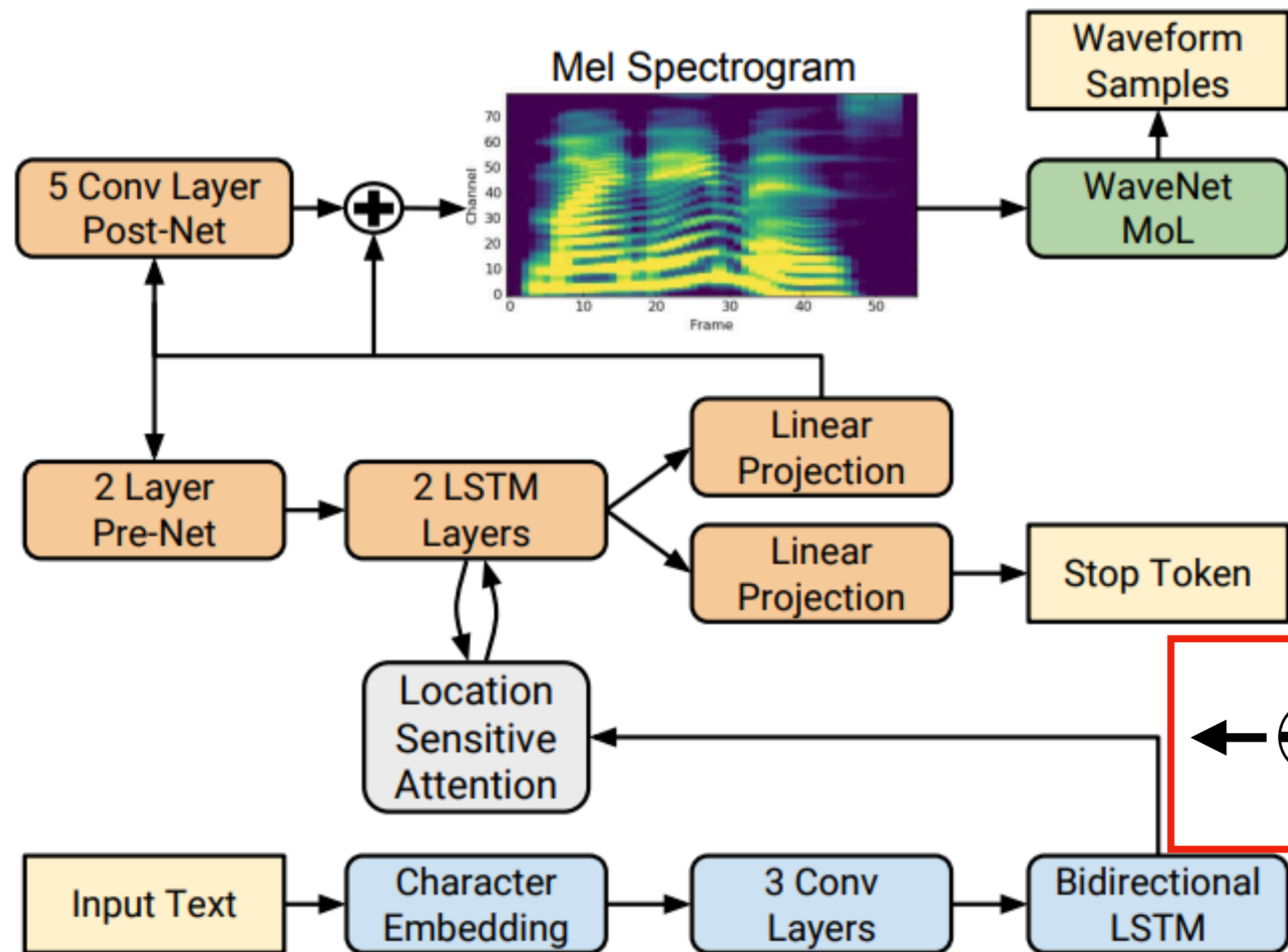


Fig. 1. Block diagram of the Tacotron 2 system architecture.

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

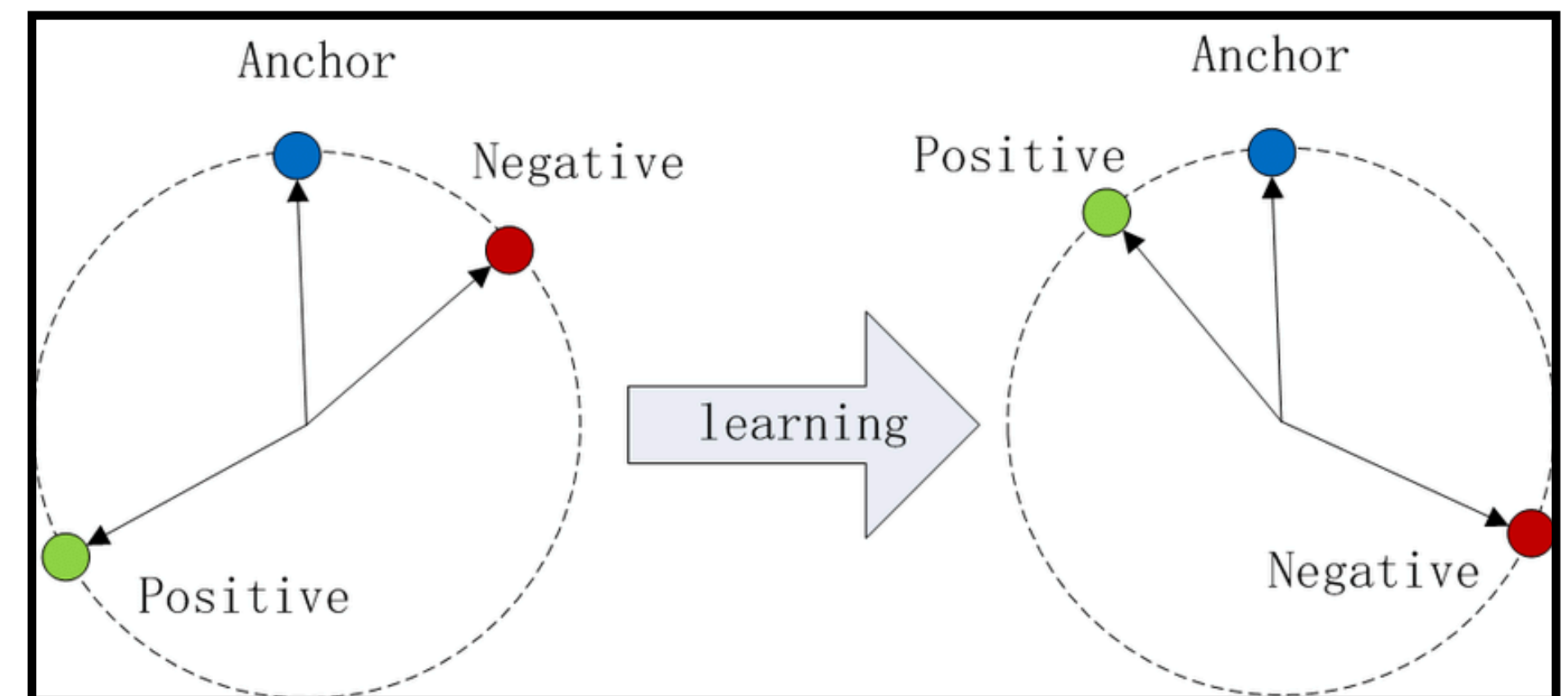
Multispeaker TTS

one-hot speaker embedding:

[num_speakers x embedding_dim]

speaker verification:

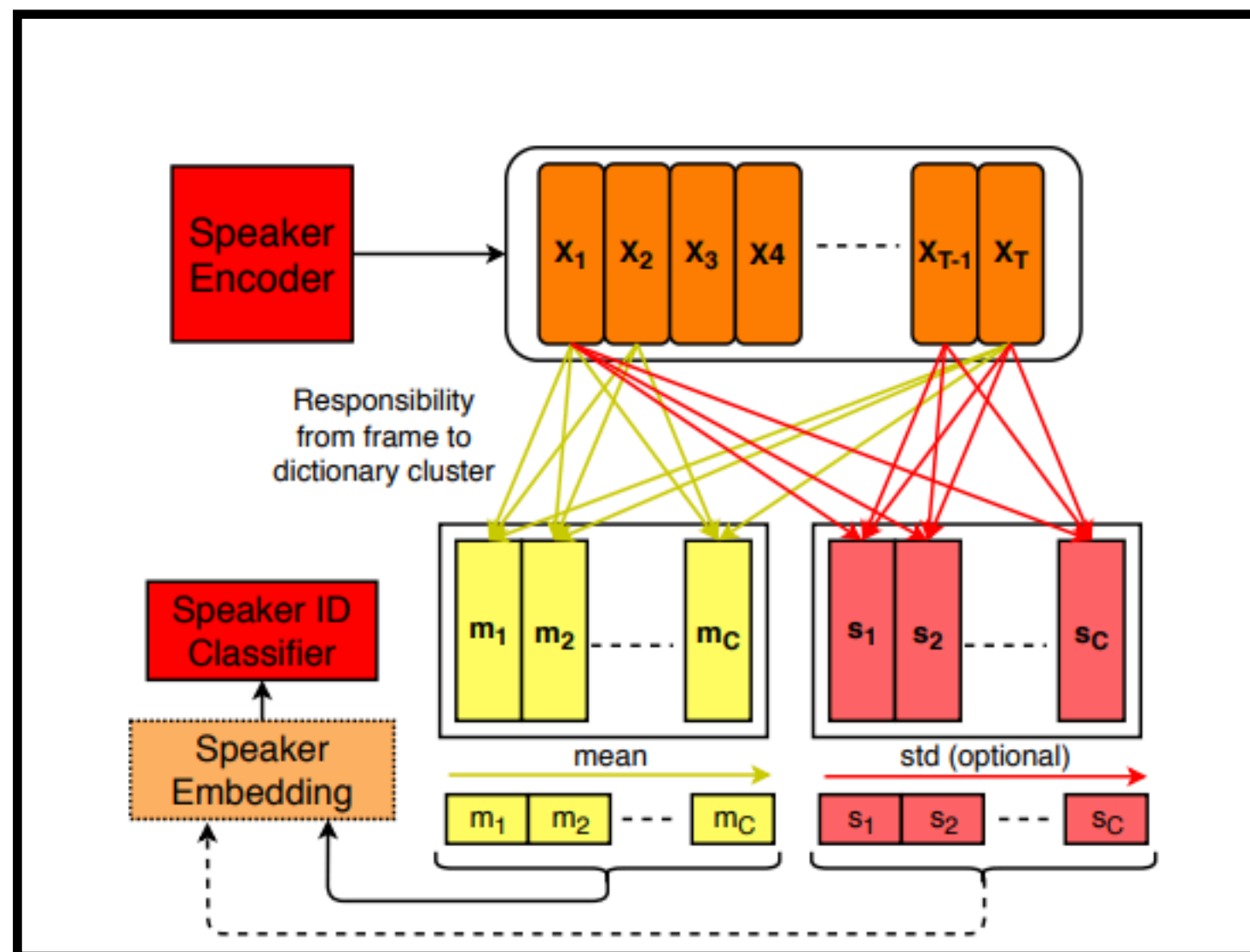
- benchmark: VoxCeleb (1, 2)
- transfer learning
- metric learning



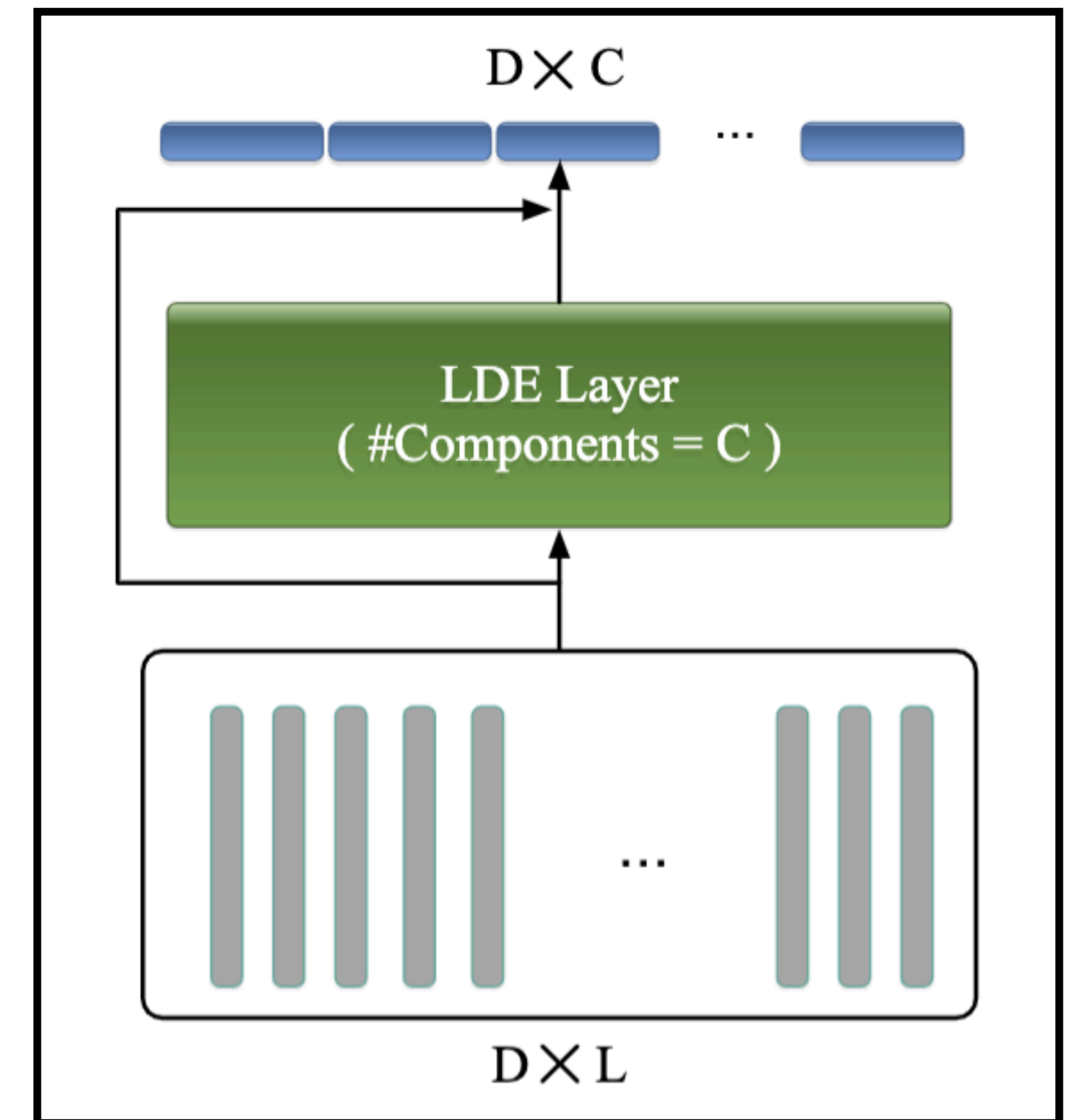
Multispeaker TTS

speaker embedding = mixture of encodings

k-means



Vanilla LDE:



Multilanguage TTS

Speaking fluently in foreign language

train:

X, en

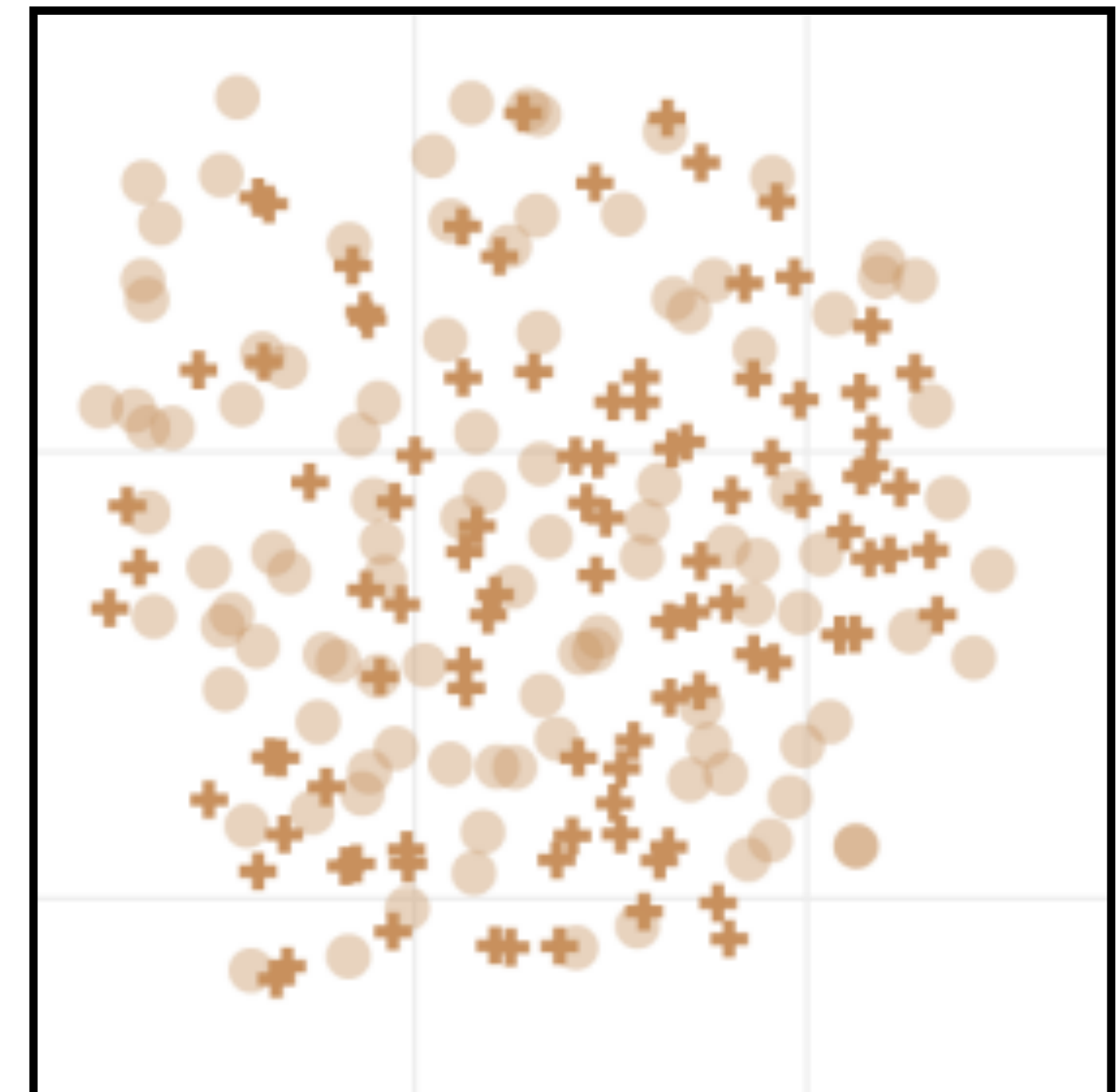
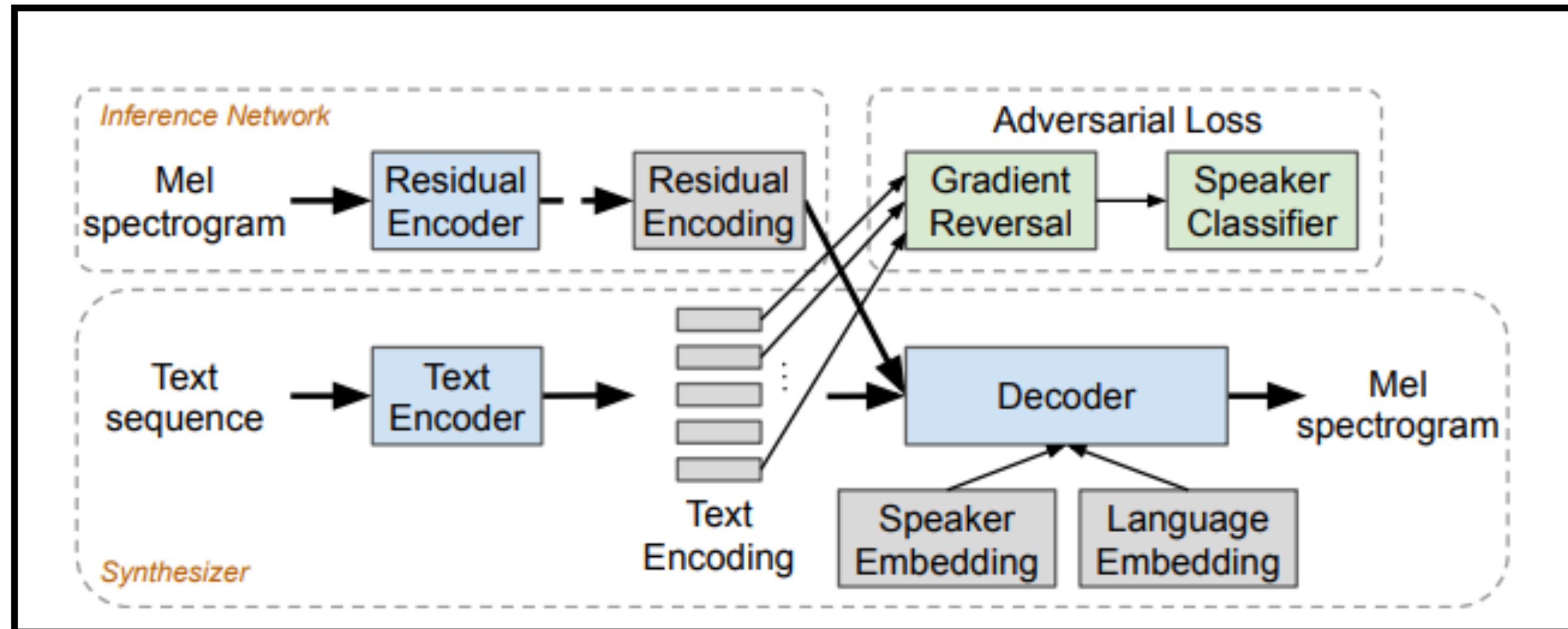
Y, ru

inference:

Y, en

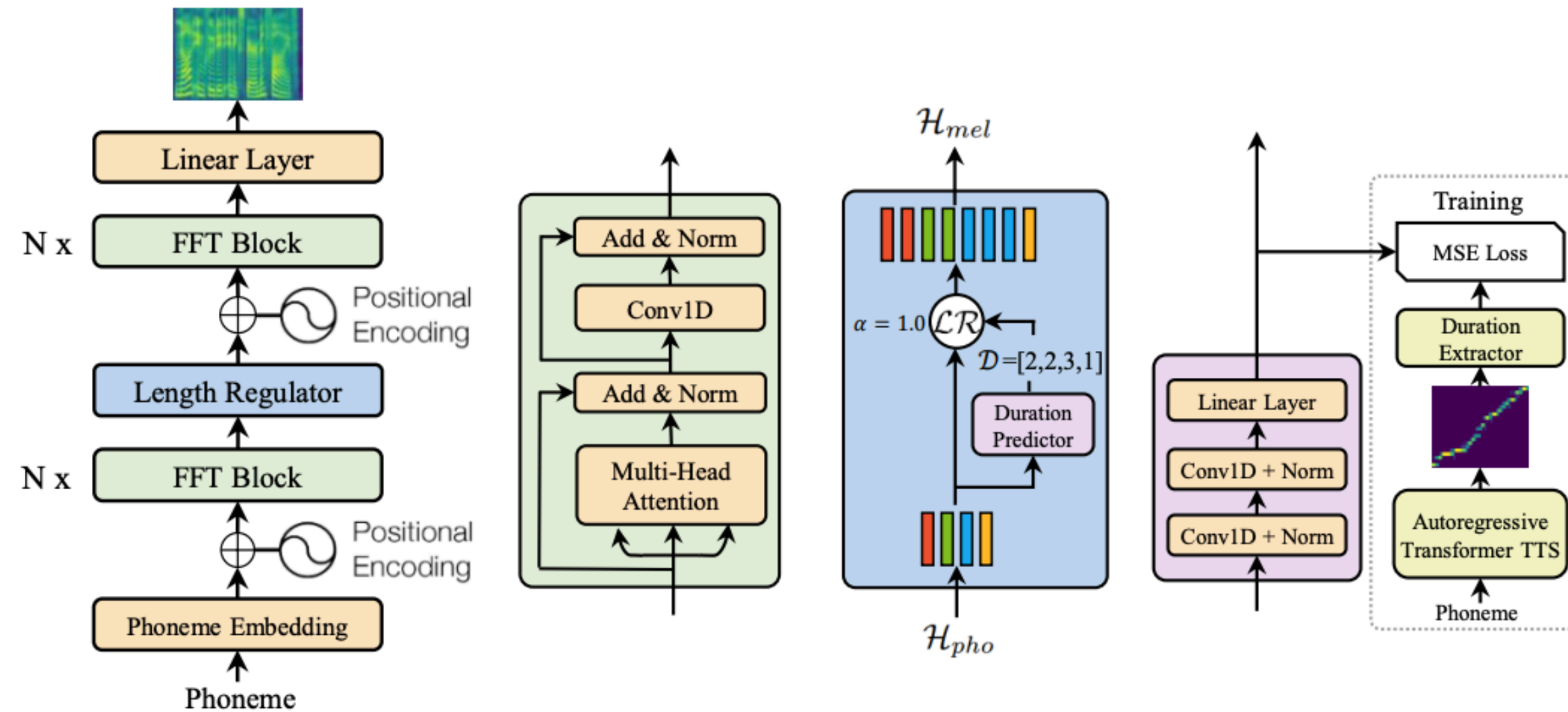
«привет мир»

«hello world»



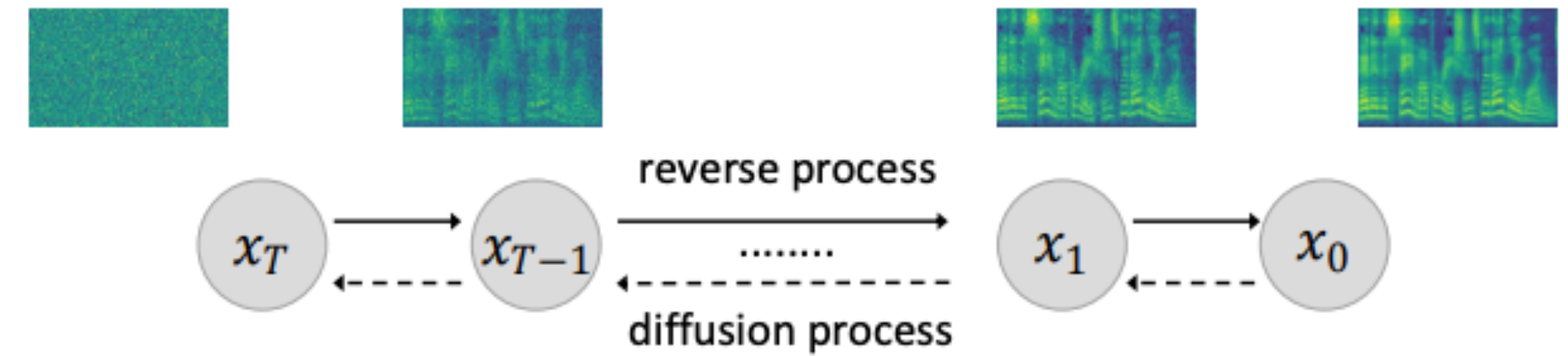
Non-autoregressive TTS

Fastspeech-family:
fastpitch fastspeech-2
transformer + duration predictor



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

Diffusion-based:



Спасибо :)