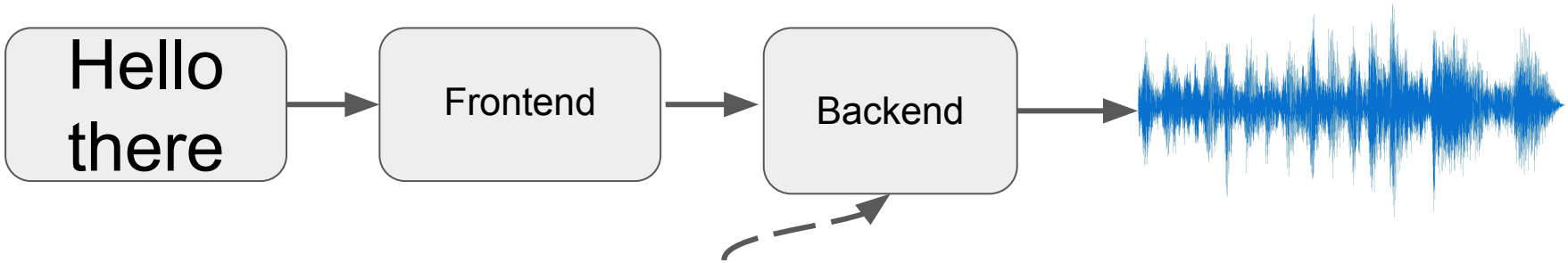
Text-to-Speech

Lecture plan:

- overview
- metrics
- datasets
- approaches

Overview



Text Analysis aka frontend:

- text normalization
- grapheme to phoneme (G2P)

Additional info:

- emphasis
- emotion
- speaker id

Backend:

- Acoustic model (AM) + Vocoder
- End2End (E2E)

Errors

Hard:

- wrong stress
- wrong pronunciation

Soft:

- naturalness
- noisiness

Metrics

Objective:  *

* surrogate metrics:

- WER/CER
- SR
- SER
- Neural MOS

Subjective:

- MOS
- MUSHRA
- SBS
- Robotness

Datasets:

- LJ Speech - EN, single speaker, ~24 hours
- Libri-TTS - EN, multi-speaker, ~585 hours
- RUSLAN - RU, single speaker, ~29 hours
- NATASHA - RU, single speaker, ~13 hours
- M-AILABS - multi language, ~1000 hours, 47 hours of Russian

Acoustic models

Tacotron family:

- Tacotron2
- GST-Tacotron
- Tacotron + Style reconstruction loss

Attentions:

- Location Sensitive Attention
- Guided Attention
- Monotonic Attention

Fast family:

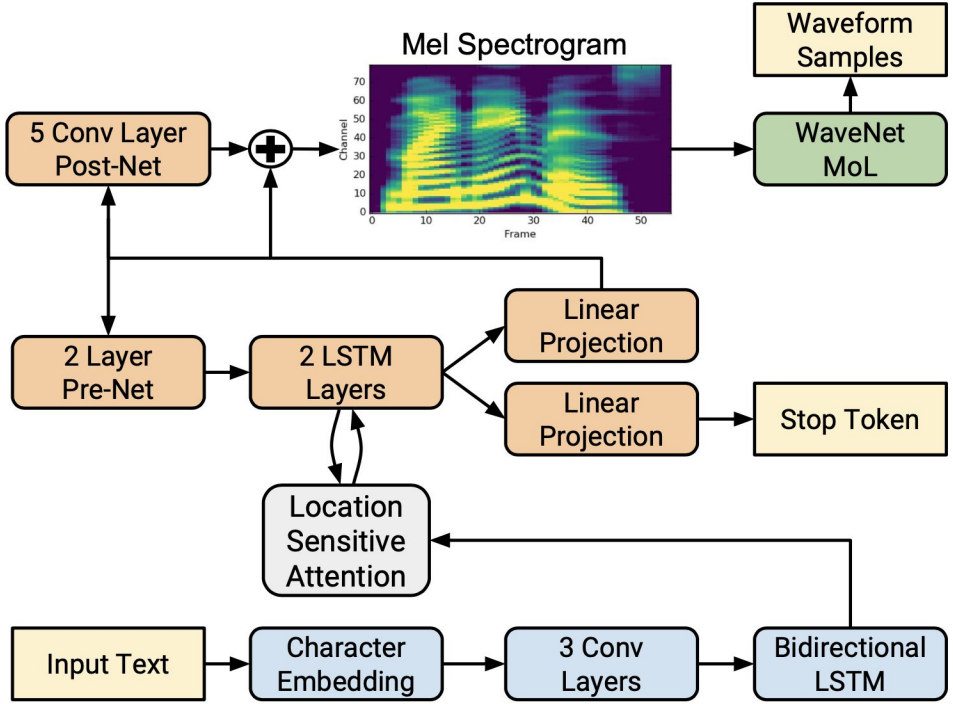
- FastSpeech2
- FastPitch
- AdaSpeech

Tacotron2

h_j - hidden state'ы lstm'ки encoder'a
s_i - i-ый hidden state decoder'a

$e_{i,j} = Score(s_{i-1}, \alpha_{i-1,j}, h)$ - energy
$\alpha_{i,j} = \exp(e_{i,j}) / \sum_{j=1}^{L} \exp(e_{i,j})$ - weights
$g_i = \sum_{j=1}^{L} \alpha_{i,j} h_j$
$y_i \sim Decoder(s_{i-1}, g_i)$

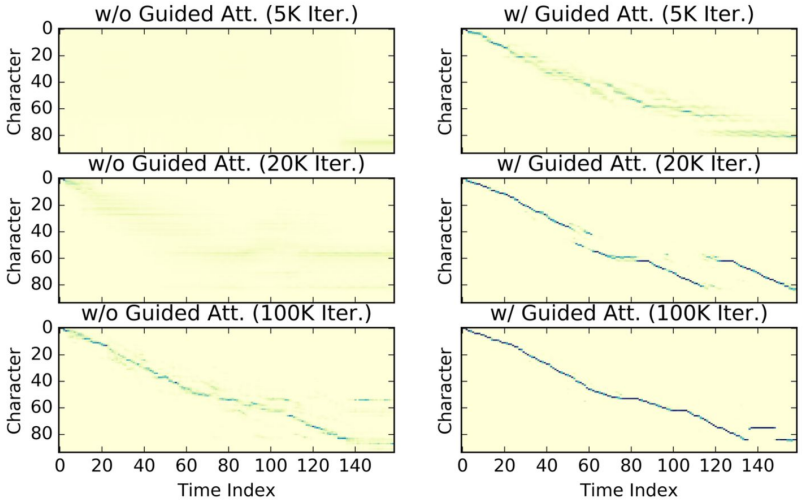
LSA energy:
$f_i = F * \alpha_{i-1}$
$e_{i,j} = w^T \tanh(W s_{i-1} + V h_j + U f_{i,j} + b)$
F, W, V, U, b - trainable parameters



Guided Attention

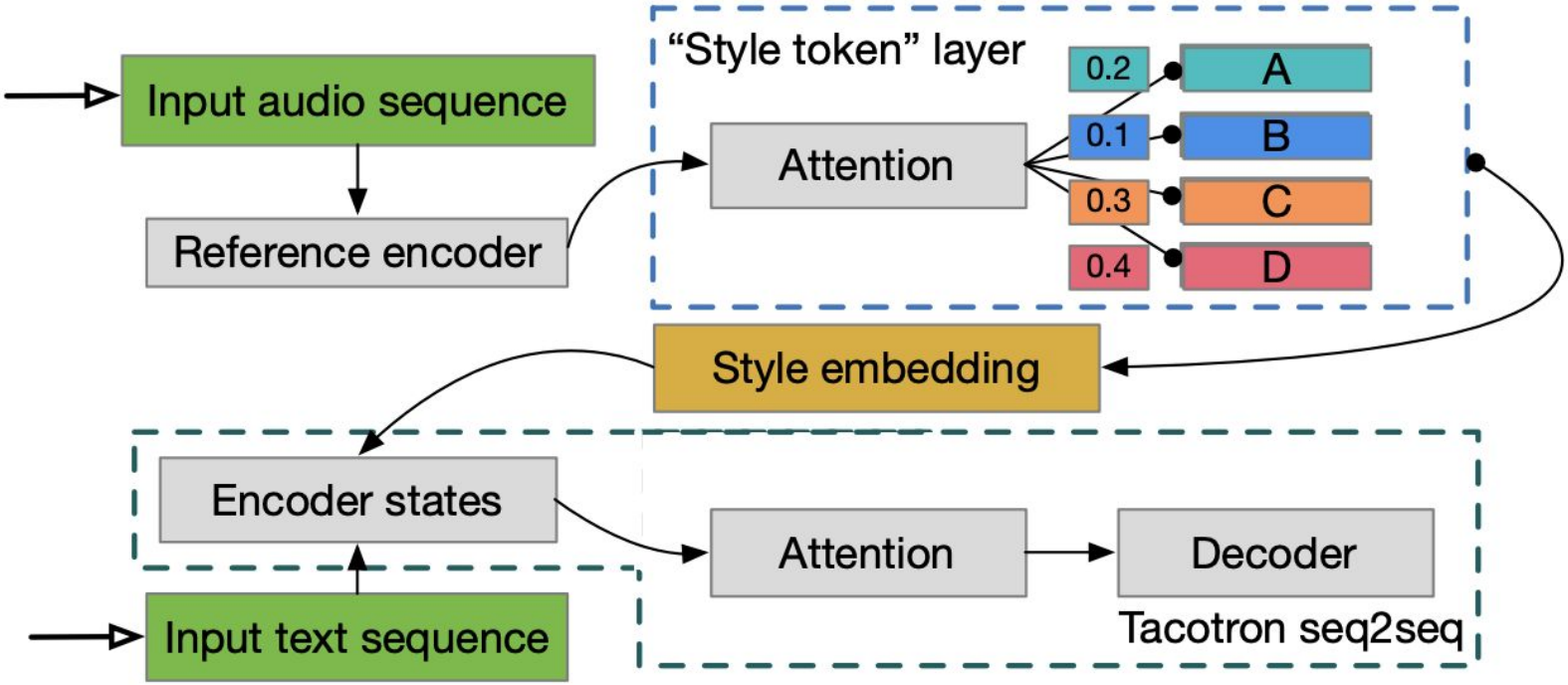
$n \sim at$, where $a \sim N/T$

N - length of text sequence, T - length of mel sequence



$$\mathcal{L}_{\text{att}}\left(A\right) = \mathbb{E}_{nt}[A_{nt}W_{nt}], \text{ where } W_{nt} = 1 - \exp\left\{-(n/N - t/T)^2/2g^2\right\}$$

Global Style Token (GST) Tacotron



Tacotron + Style reconstruction loss

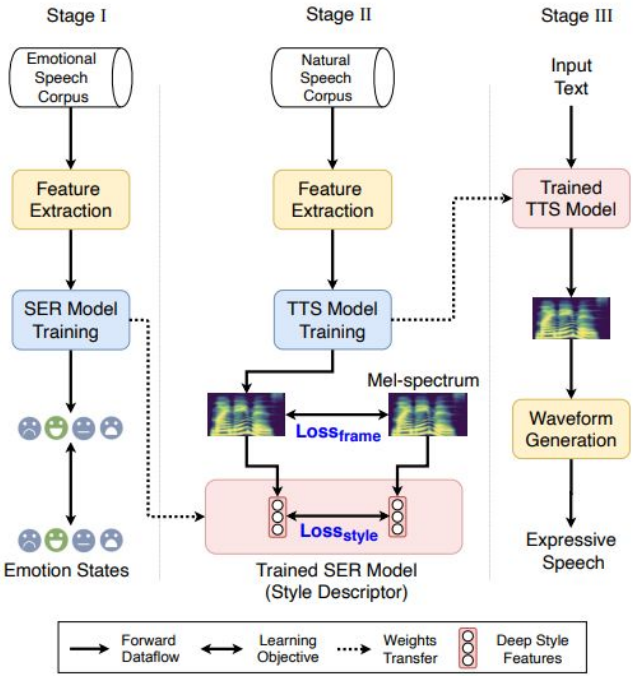
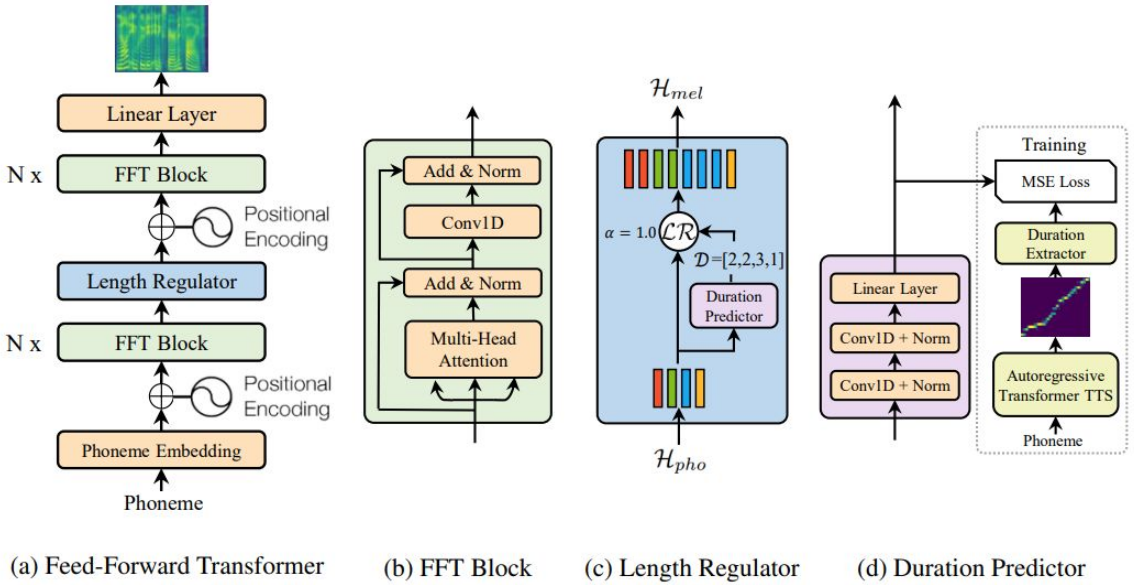


Fig. 2: Overall framework of a *Tacotron-PL* system in three stages: Stage I for training of style descriptor; Stage II for training of *Tacotron-PL*; Stage III for run-time inference.

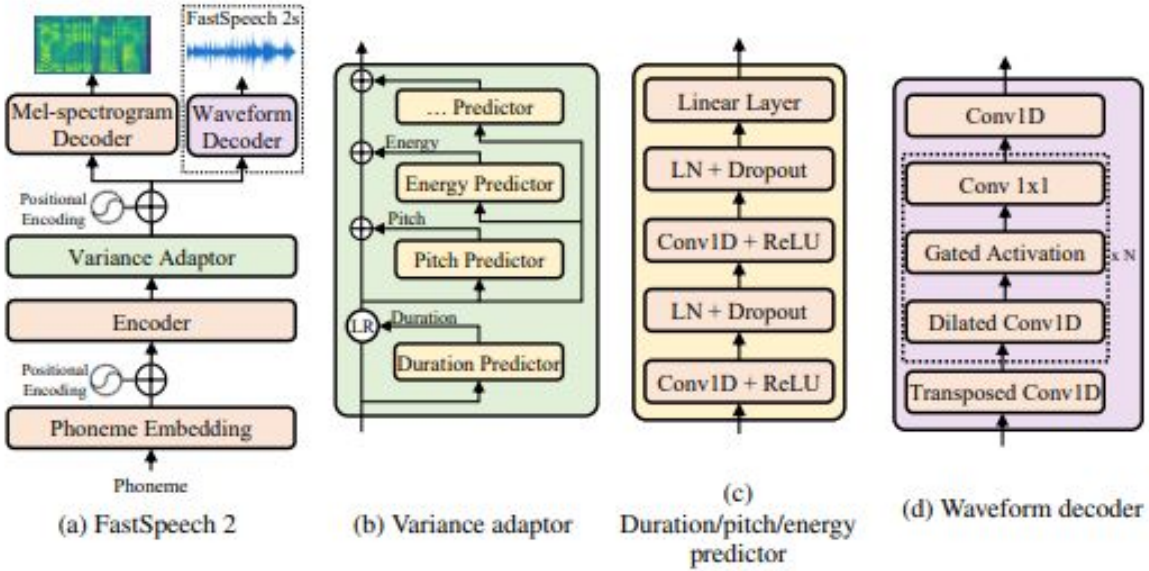
- pretrained speaker emotion recognition
- additional loss for style
- simpler and more expressive inference

Fast Family

- FastSpeech
- FastPitch
- FastSpeech2



FastSpeech2 & FastPitch



AdaSpeech

Main goals:

- to handle different acoustic conditions
- to finetune for new speakers with small number of parameters and without quality degradation

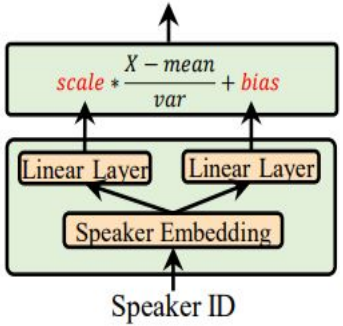


Figure 3: Conditional LayerNorm.

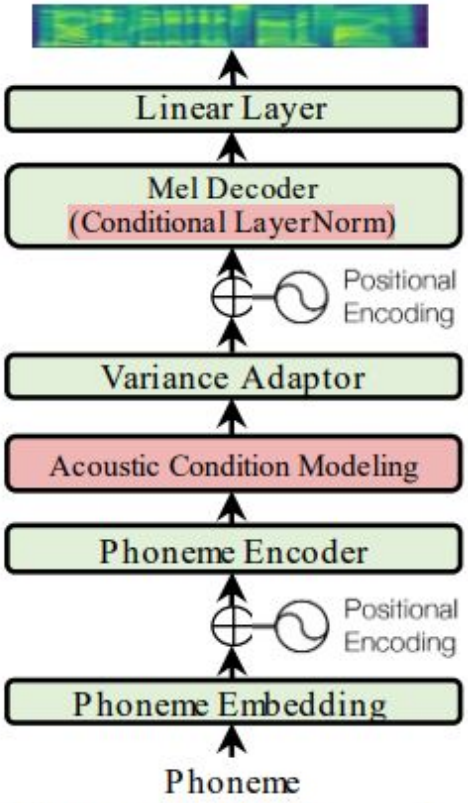
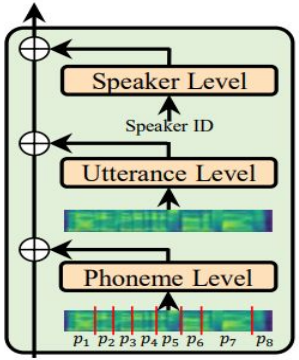
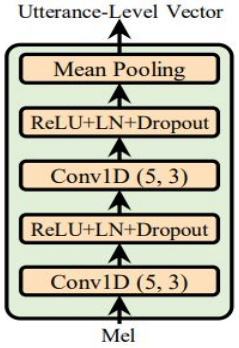


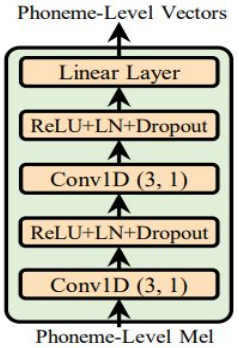
Figure 1: AdaSpeech.



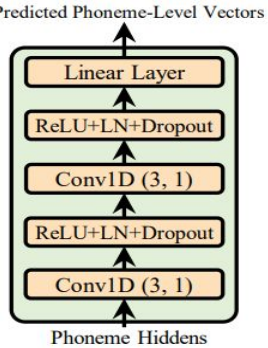
(a) Overall.



(b) Utterance level.



(c) Phoneme level.



(d) Phoneme level.

Ссылки:

- neural mos - NN for mos prediction
- g2p - russian g2p
- mfa - text-speech aligner on HMMs
- unnamed dataset - russian, single speaker, bad quality
- Best ml memes in the multiverse.