Just AI

Современные методы управления просодией в задаче синтезе речи

Мылзенова Дарима, Just AI

Just AI

План доклада

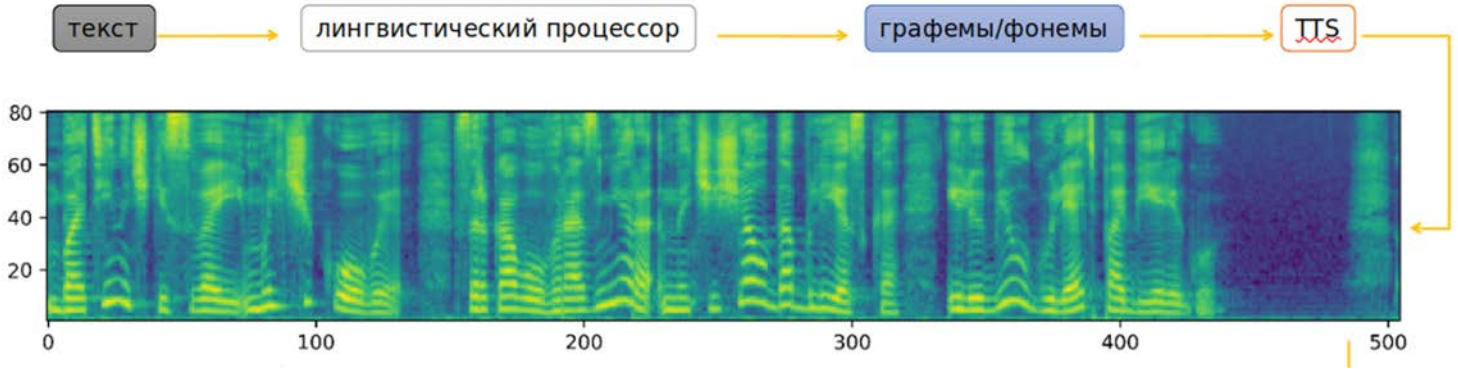
1. Что такое просодия и почему это важно для TTS
2. Контроль просодии через дополнительные признаки (с разметкой)
3. Нейросетевые архитектуры для контроля просодии

Just AI

Задача синтеза речи

Just AI

Задача синтеза речи



Проблема: задача one-to-many!

Just AI

Задача синтеза речи

Речевой сигнал

Лингвистическая информация

Паралингвистическая информация

Спикер

Эффекты записи

Просодия

темп
интонация
энергия, F0,
интонация

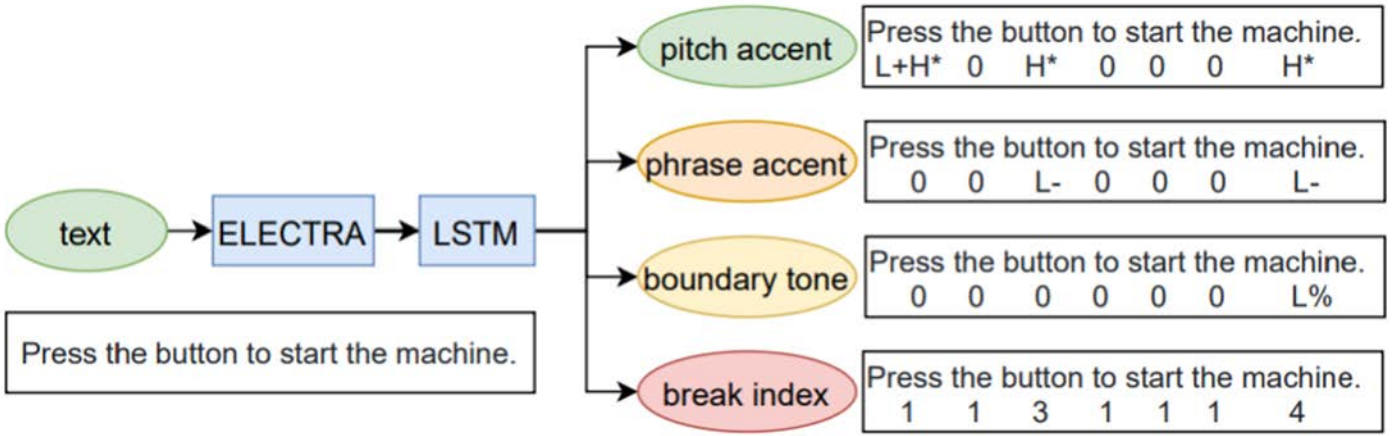
Just AI

Как повлиять на просодию?

1. Разметить просодические фичи и подавать на вход модели
2. Без разметки - построить латентное представление просодии.

Just AI

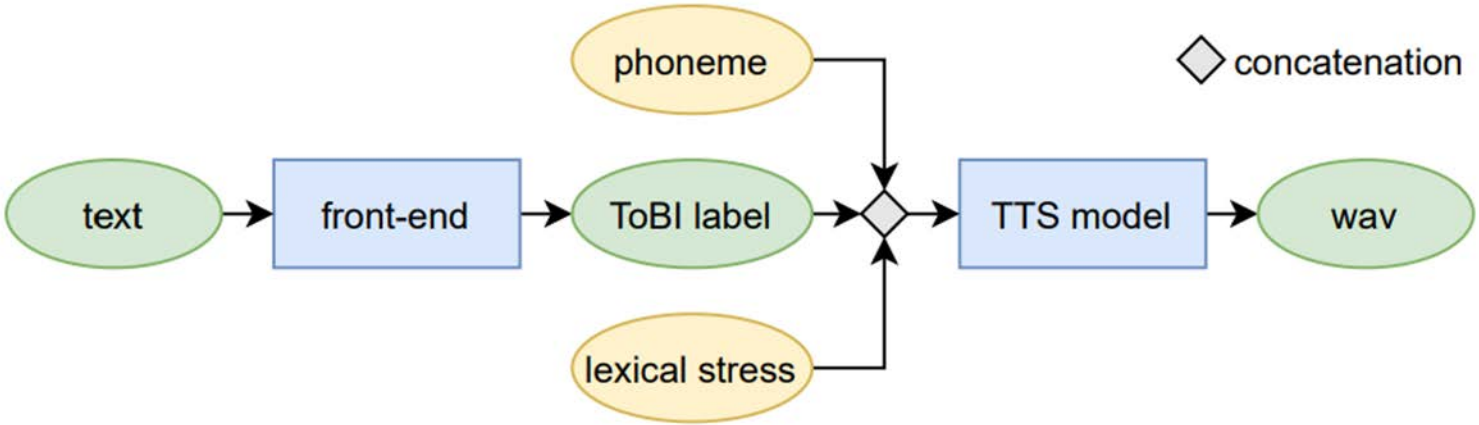
Supervised control: Tones and Breaks indices



https://www.isca-speech.org/archive/pdfs/interspeech_2021/zou21_interspeech.pdf

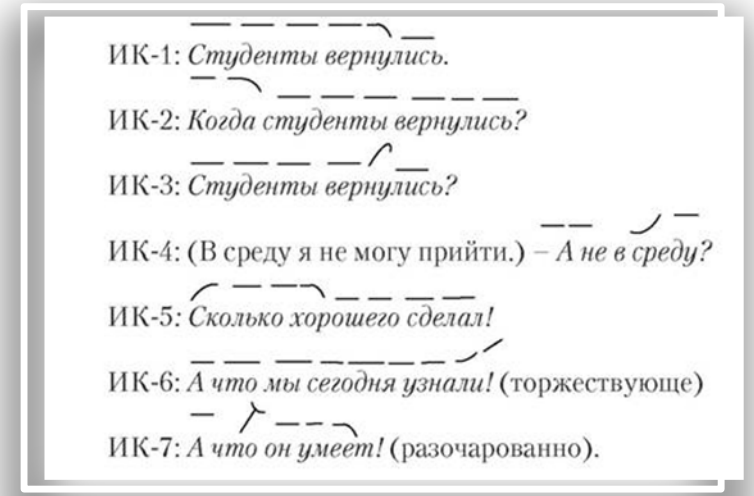
Just AI

Supervised control: Tones and Breaks indices



https://www.isca-speech.org/archive/pdfs/interspeech_2021/zou21_interspeech.pdf

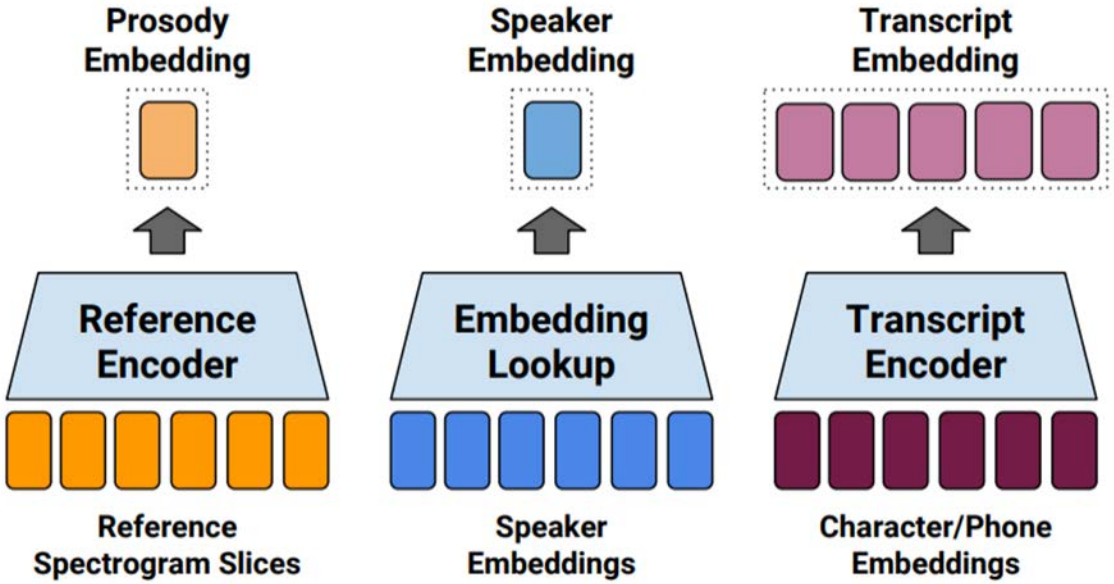
Supervised control: интонационные контуры



Брызгунова Е. А. Интонация // Русская грамматика / Н. Ю. Шведова (гл. ред.). — М.: Наука, 1980.

Just AI

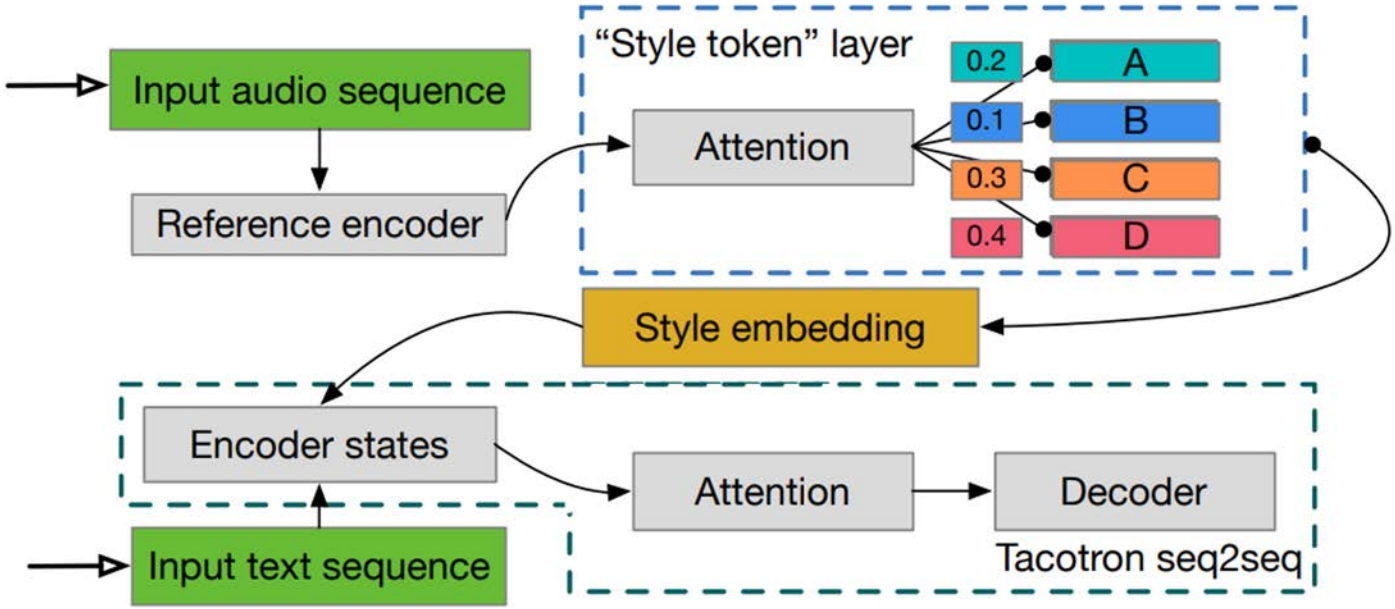
Unsupervised control: Tacotron 2



https://arxiv.org/pdf/1803.09047.pdf

Just AI

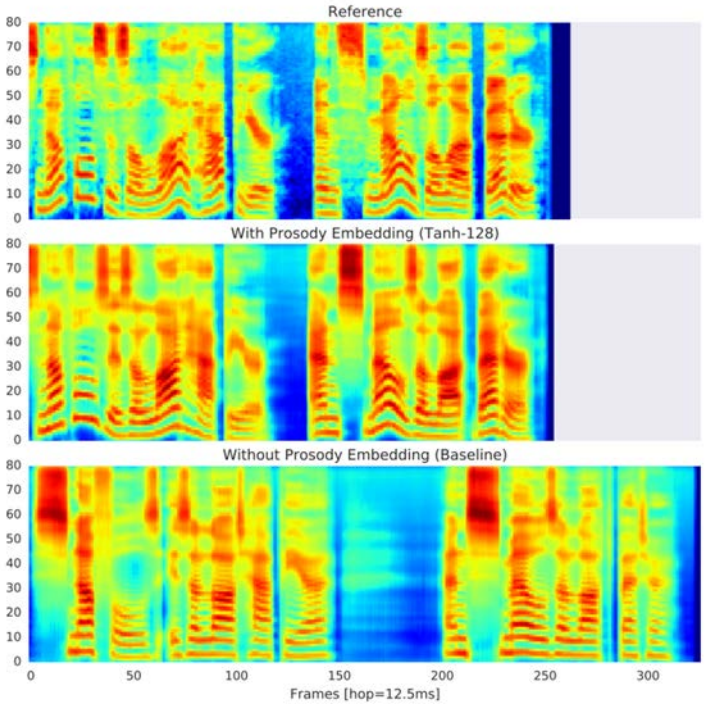
Unsupervised control with GST



https://arxiv.org/pdf/1803.09017.pdf

Just AI

Unsupervised control: Tacotron 2



референс

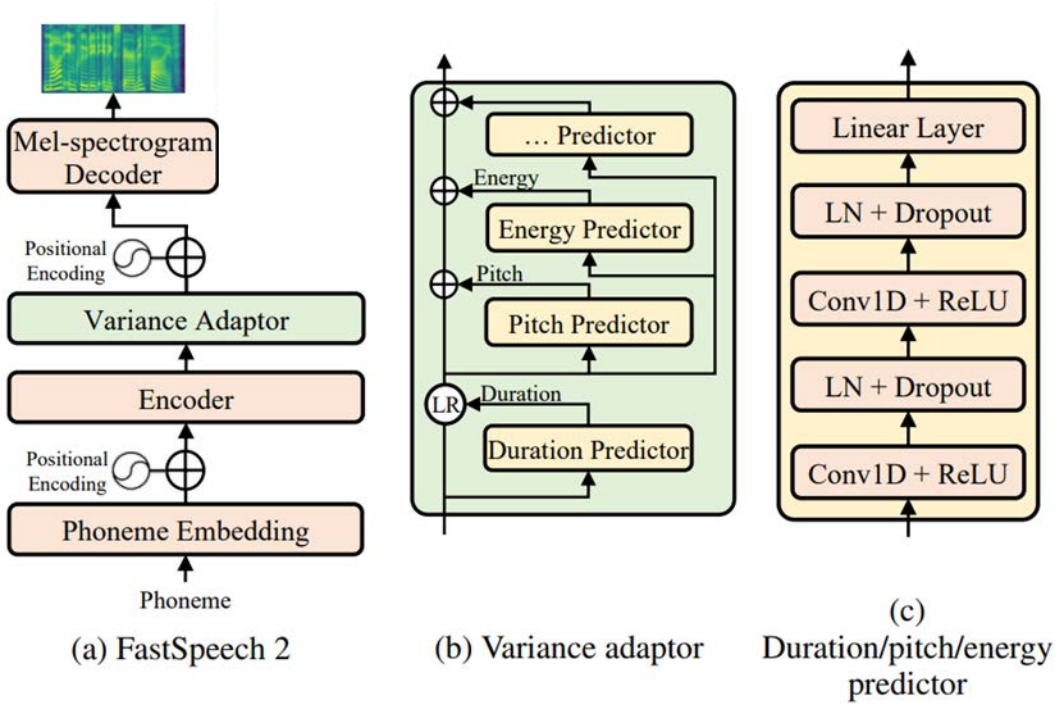
синтез с референсом

синтез безлайн модели

https://arxiv.org/pdf/1803.09047.pdf

Just AI

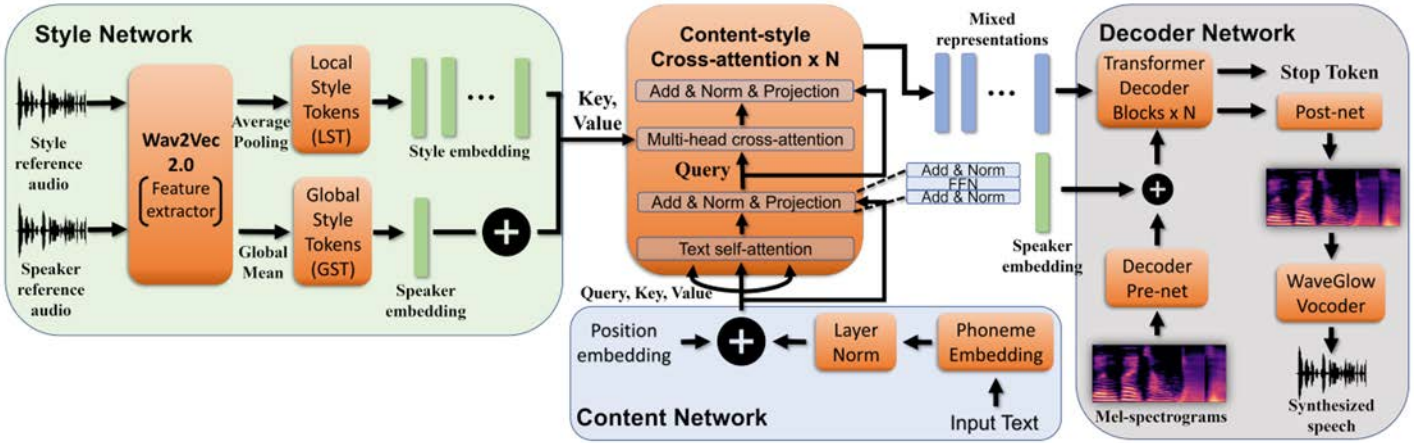
Unsupervised control: FastSpeech 2



https://arxiv.org/pdf/2006.04558.pdf

Just AI

Fine-grained control with LST



https://arxiv.org/pdf/2110.06306.pdf

Just AI

Flowtron

- Normalizing Flows
  p(x) -> p(z)

$$z = f_0 \circ f_1 \circ \cdots \circ f_k(x)$$

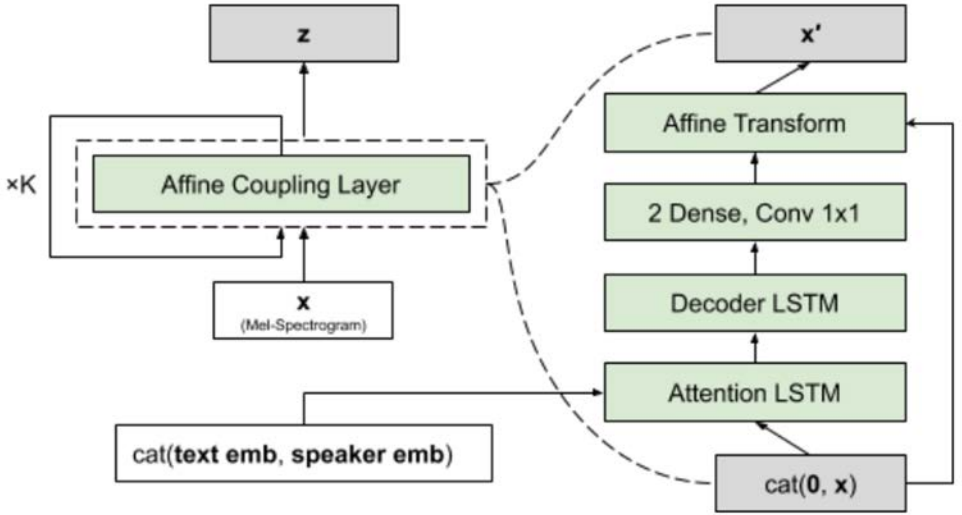
$$z \sim N(z; 0, I)$$

  x - фрейм спектра,
  f_i - обратимые

- Во время инференса сэмплируем случайный z, обращаем f_i и получаем x

  x = f_1

$$x = f_k^{-1} \circ f_{k-1}^{-1} \circ \cdots \circ f_0^{-1}(z)$$



https://arxiv.org/pdf/2005.05957.pdf

Just AI

Flowtron

Сэмплируя с большей дисперсией можно добиться более экспрессивной и живой речи.

Обуславливая параметры распределения можно сделать качественный перенос просодии.

https://developer.nvidia.com/blog/training-your-own-voice-font-using-flowtron/

Just AI

Summary

- Управление просодией с разметкой
- Предсказание латентного вектора просодии и стиля через Reference Encoder и GST
- Предсказание питча, энергии и длительности с FastSpeech2
- Более детальное предсказание стиля - LST
- Экспрессивный синтез с FlowTron