

Синтез речи

Лекция №1

Гриша Стерлинг

Синтез речи

Задача:

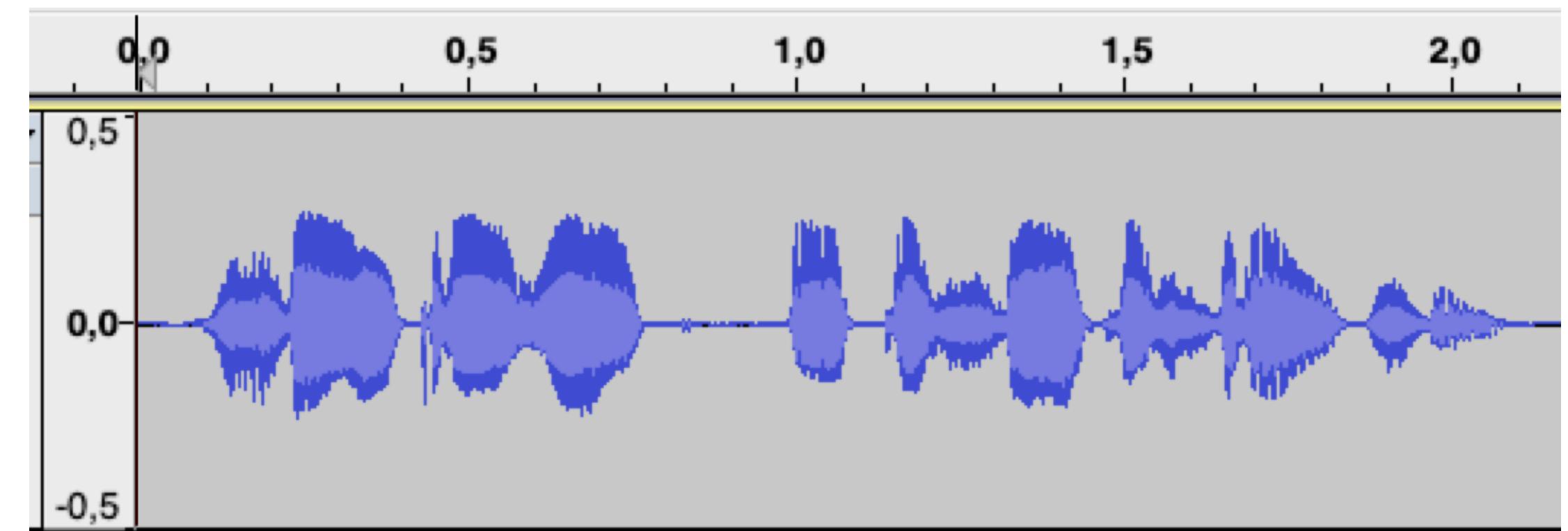
- озвучить заданный текст голосом

Задачи со звездочкой:

- style transfer
- несуществующим голосом
- controllable speech synthesis
- на другом языке
- эмоции
- шепот
- субвокализации, смех

«Всем привет, это синтез речи»

->

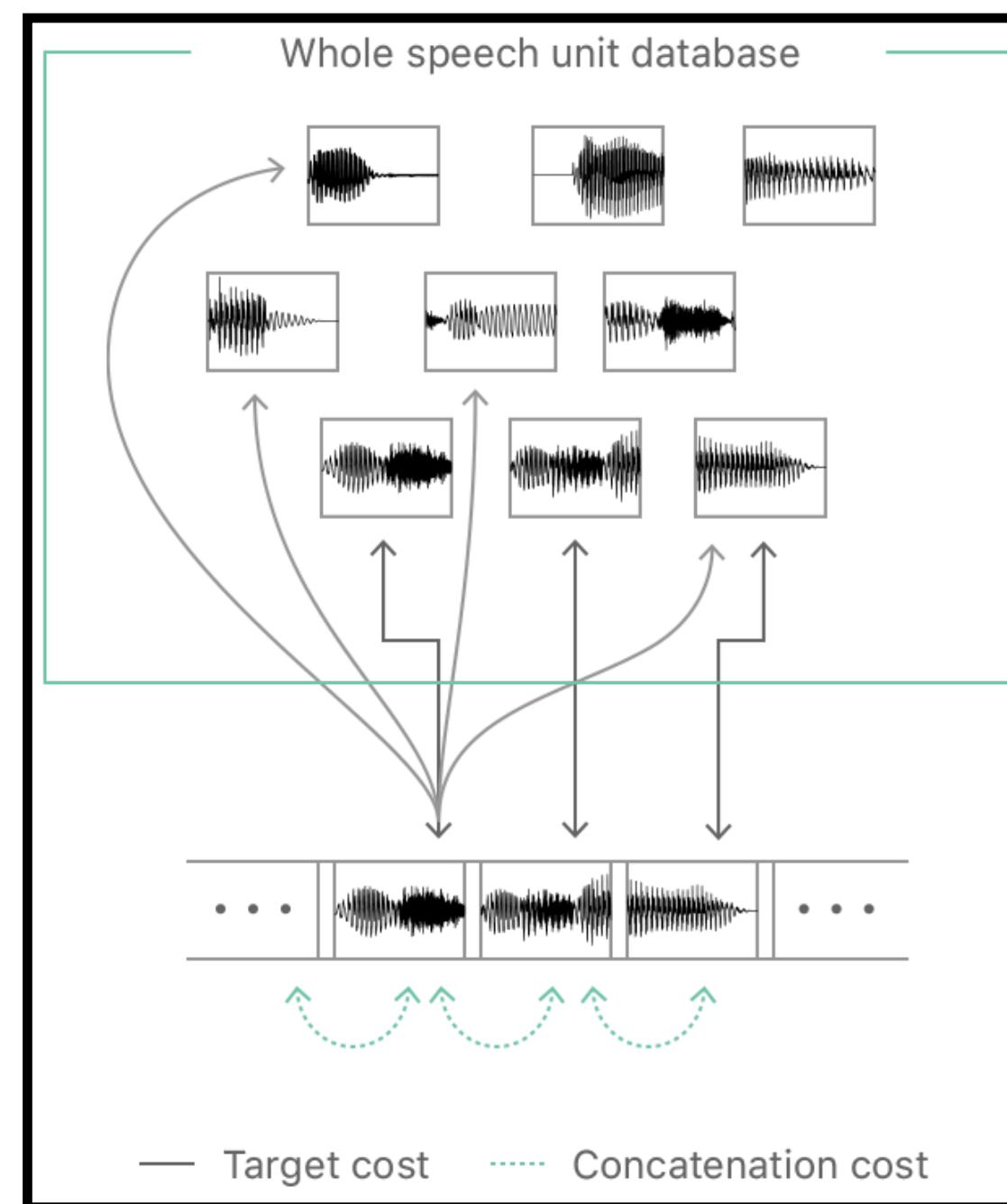


Синтез речи

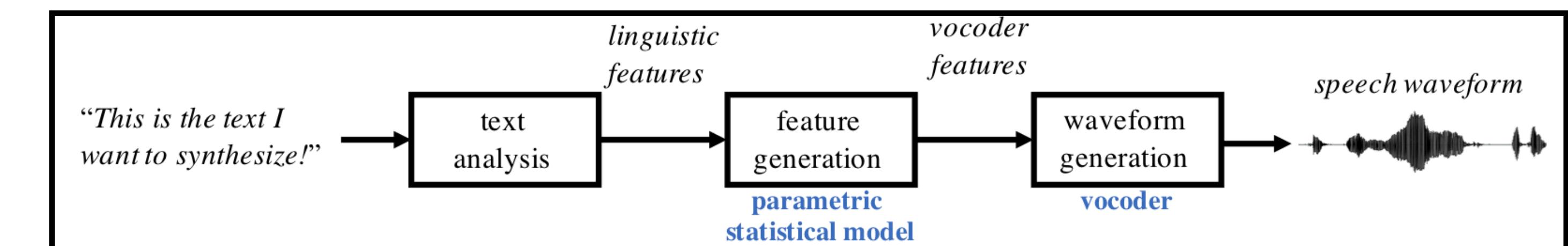


Concatenative
(unit selection)

Parametric

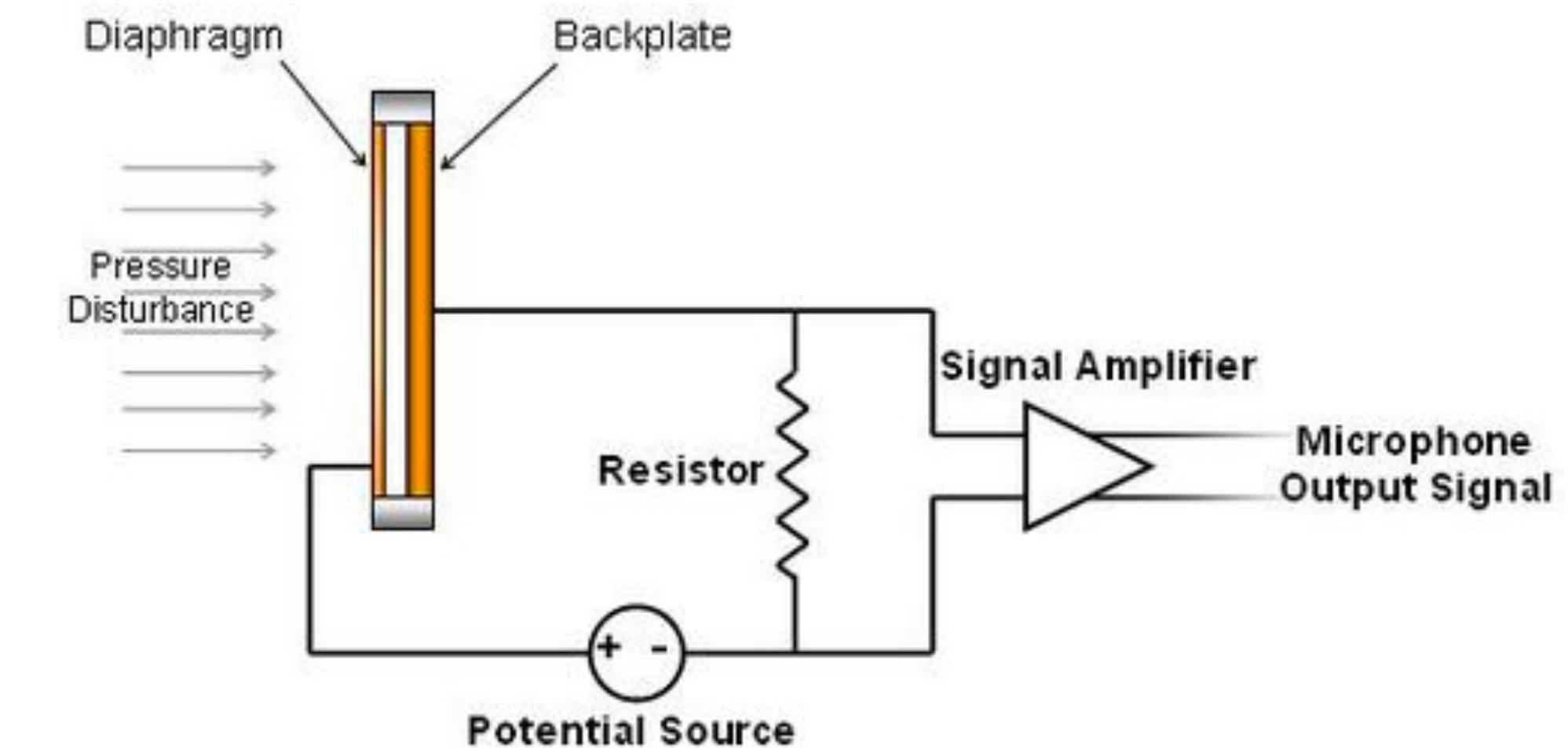
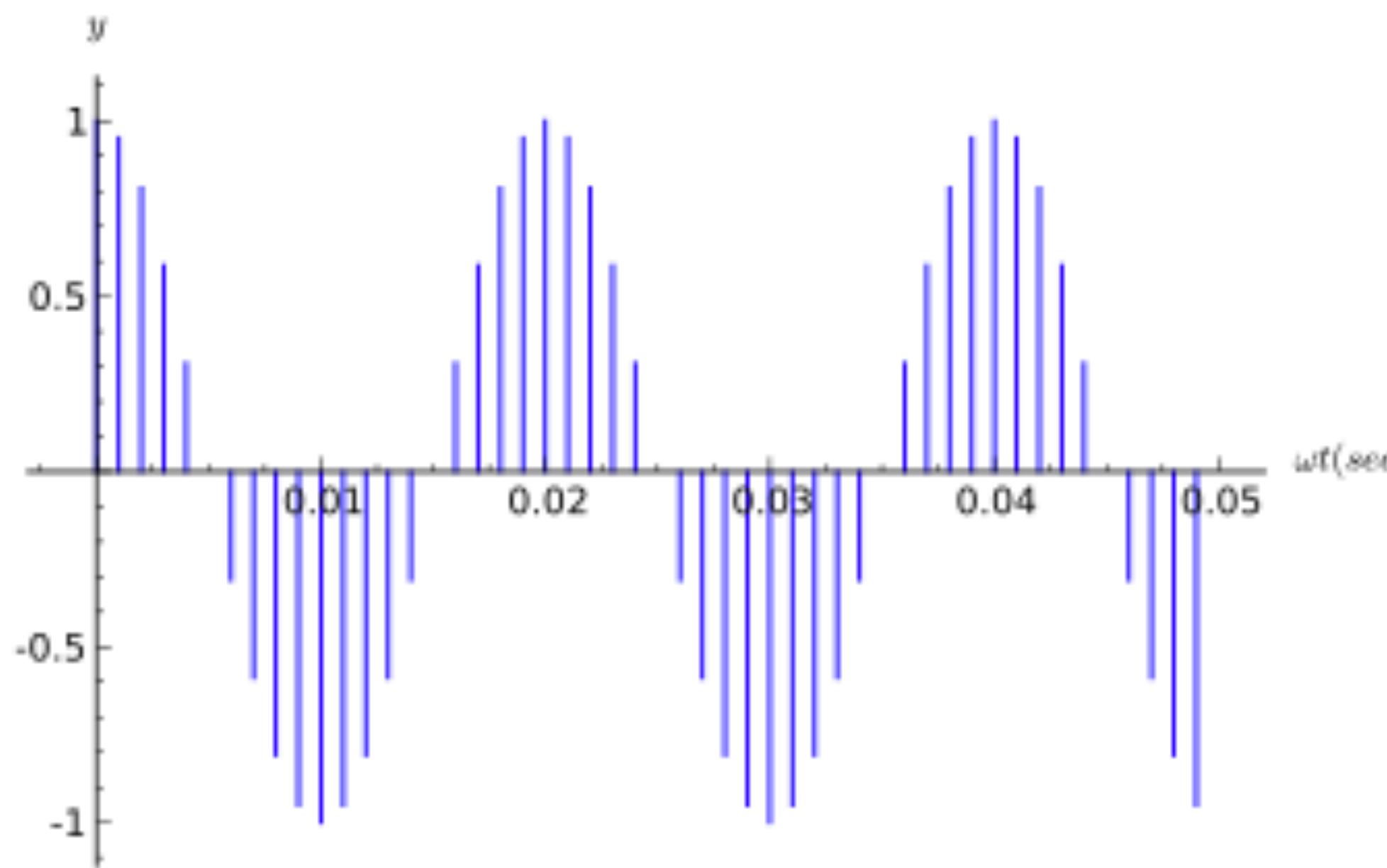


- 2 стадии:
- акустическая модель
 - вокодер (**voice coder**)



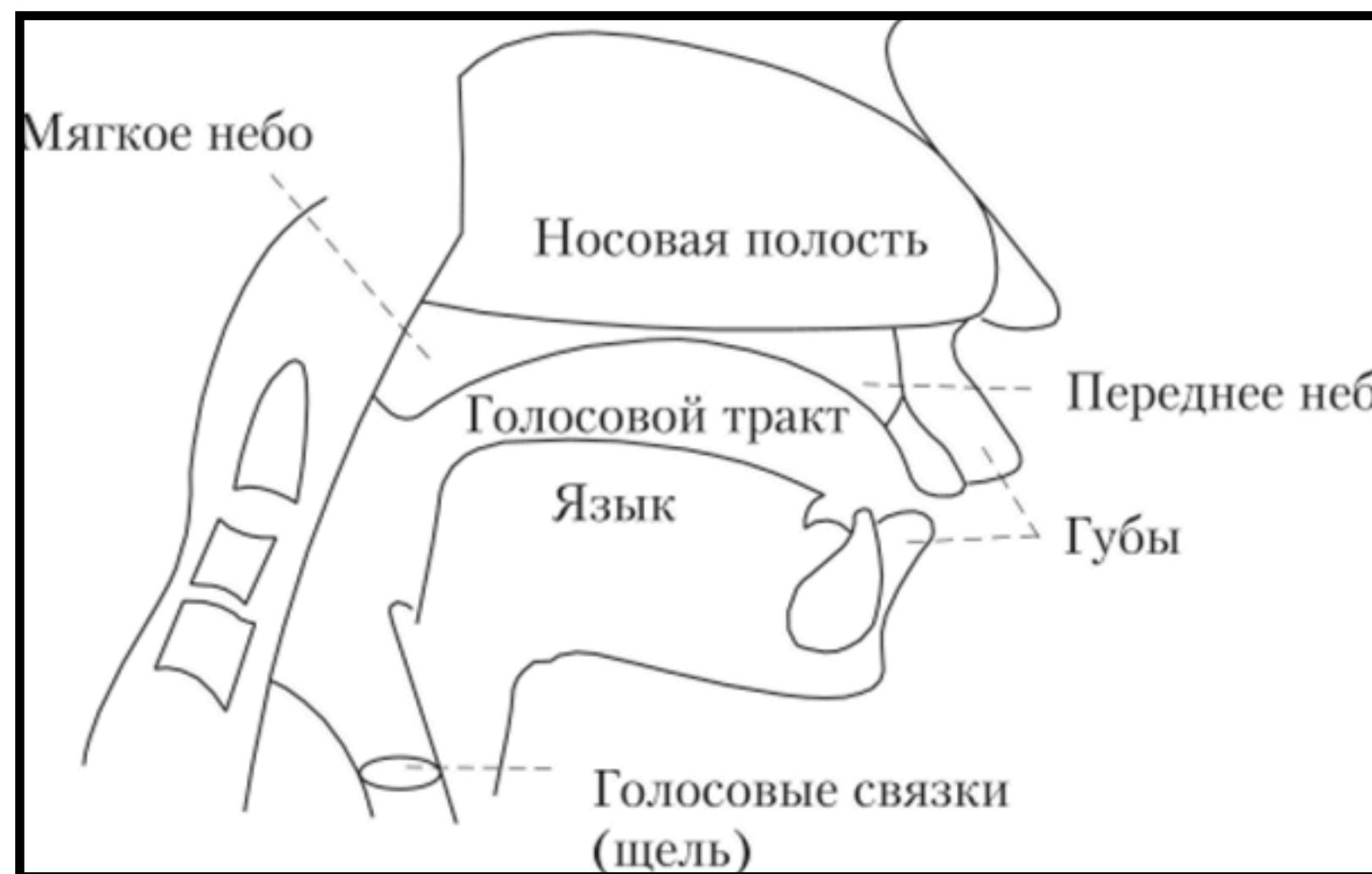
Цифровое представление звука

- Квантизация:
 - по времени
 - по амплитуде
- Устройство микрофона:

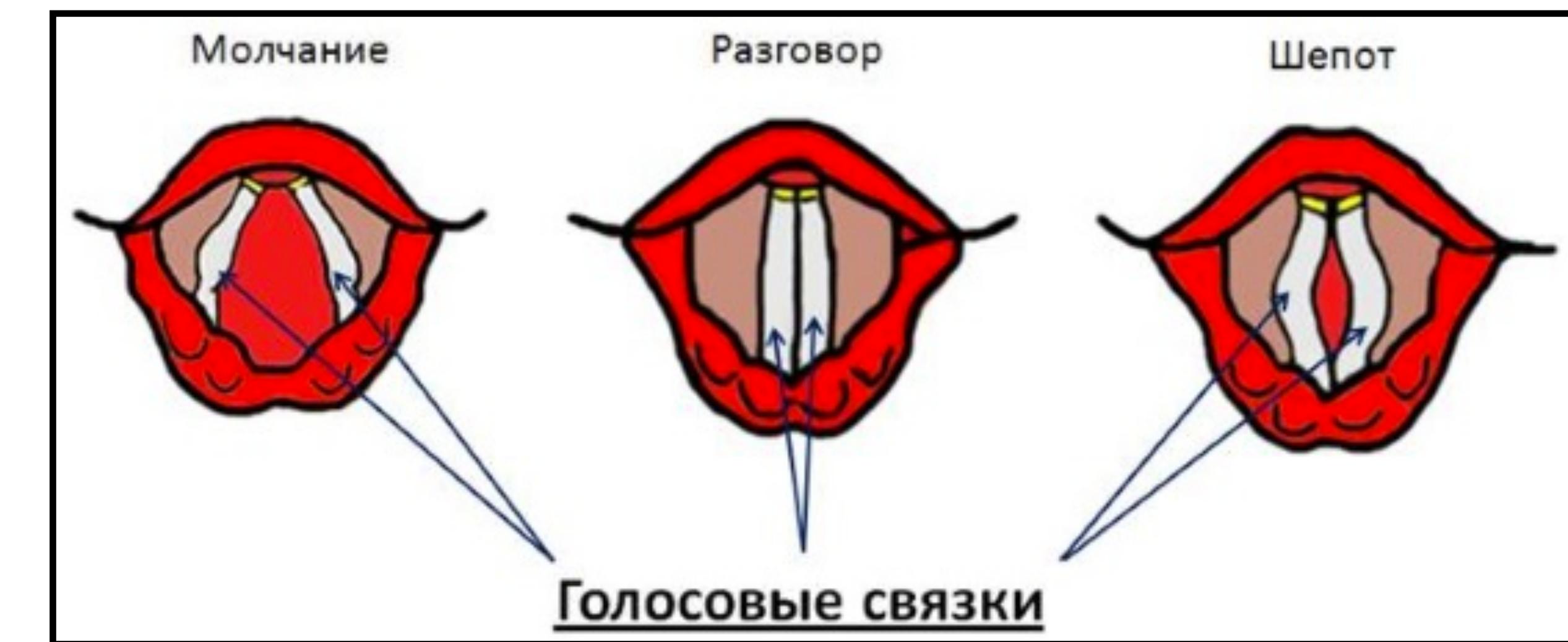


Что такое речь

Речевой тракт человека

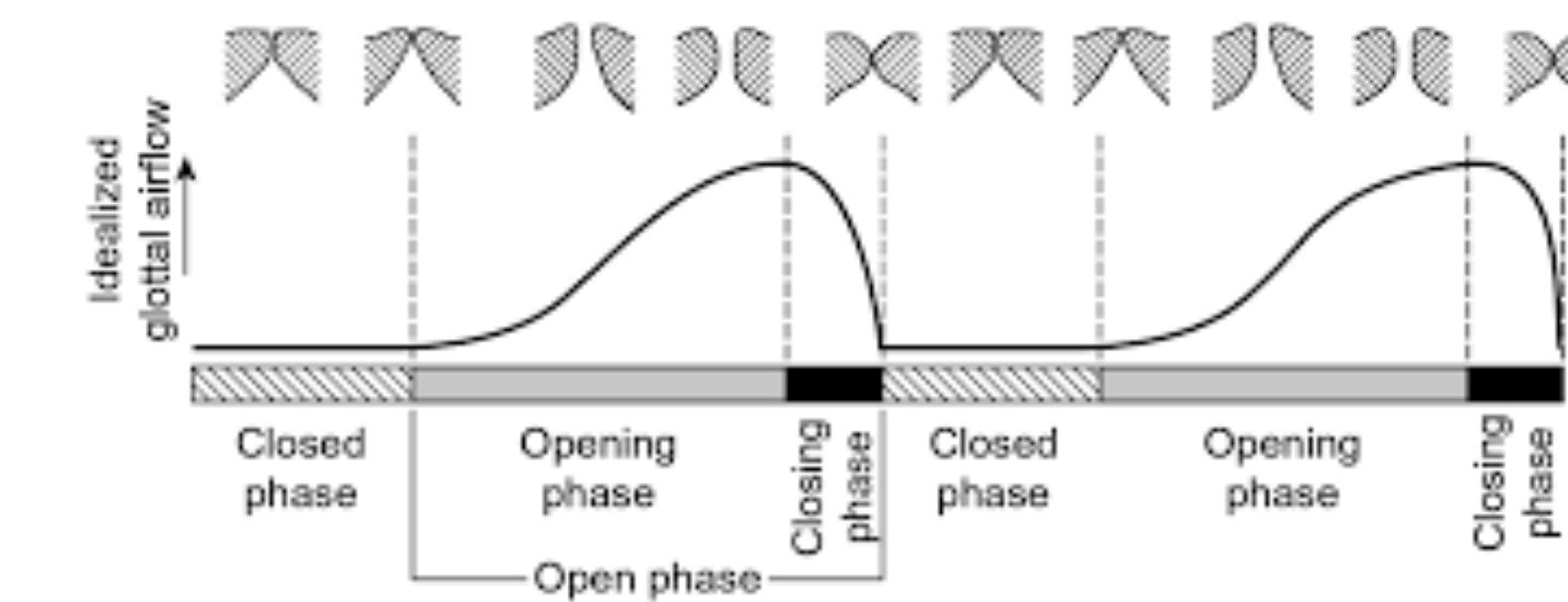


Голосовые связки

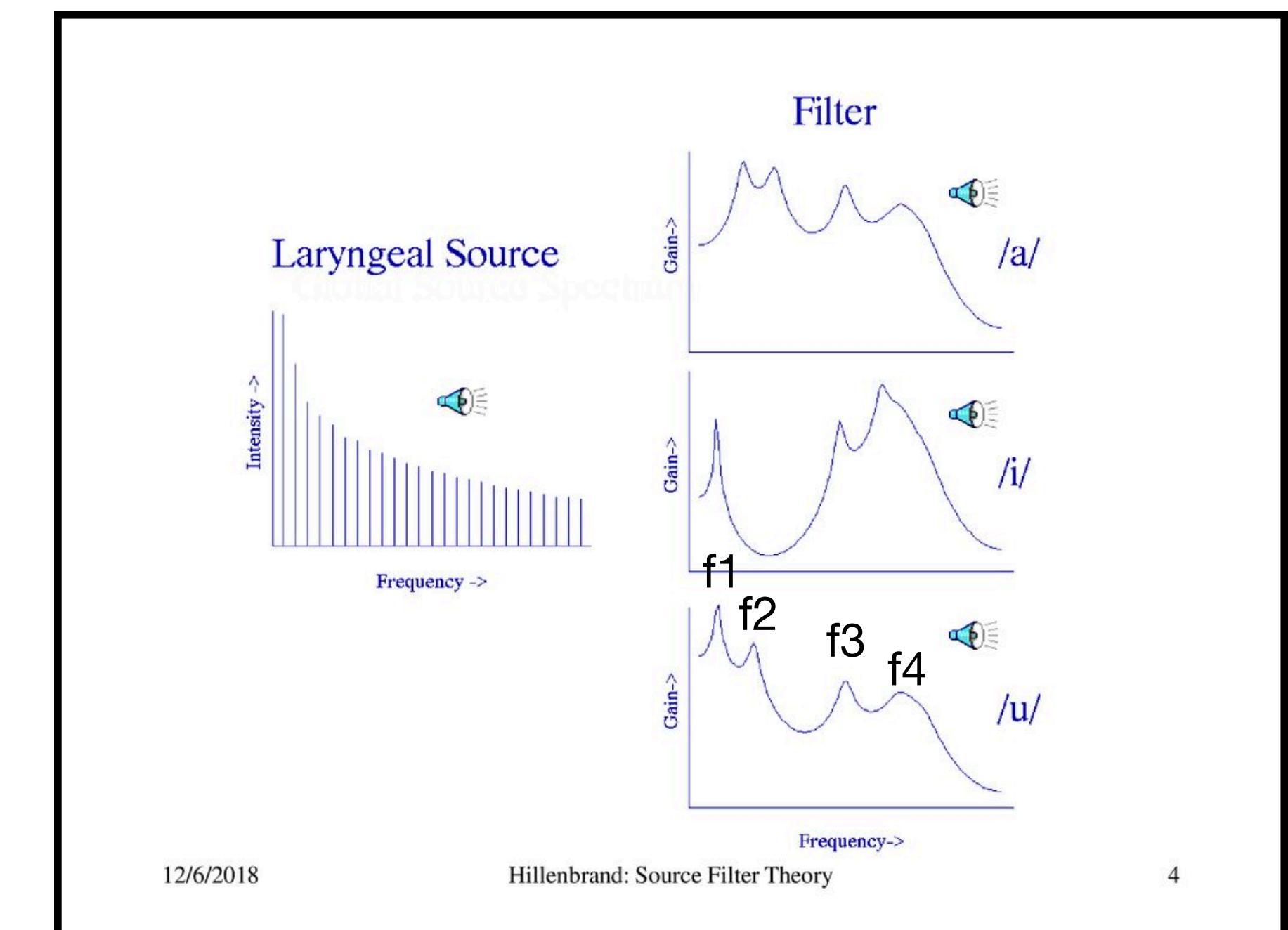
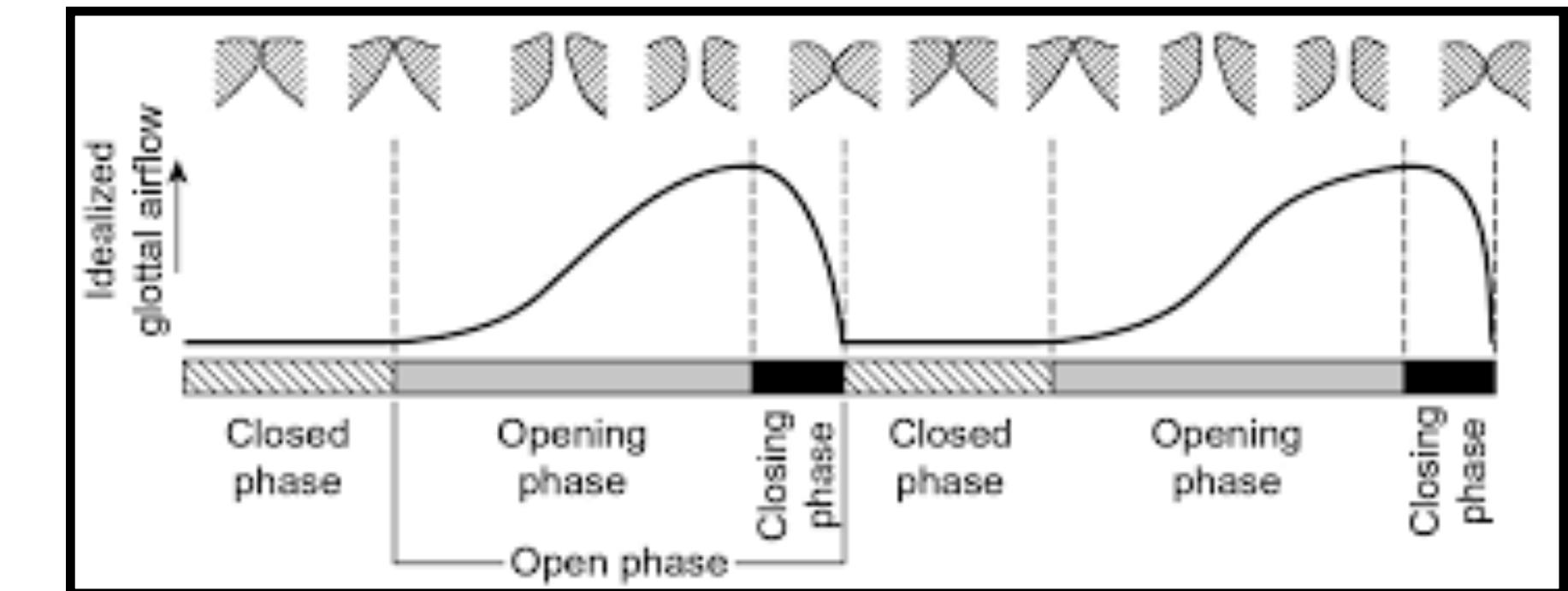
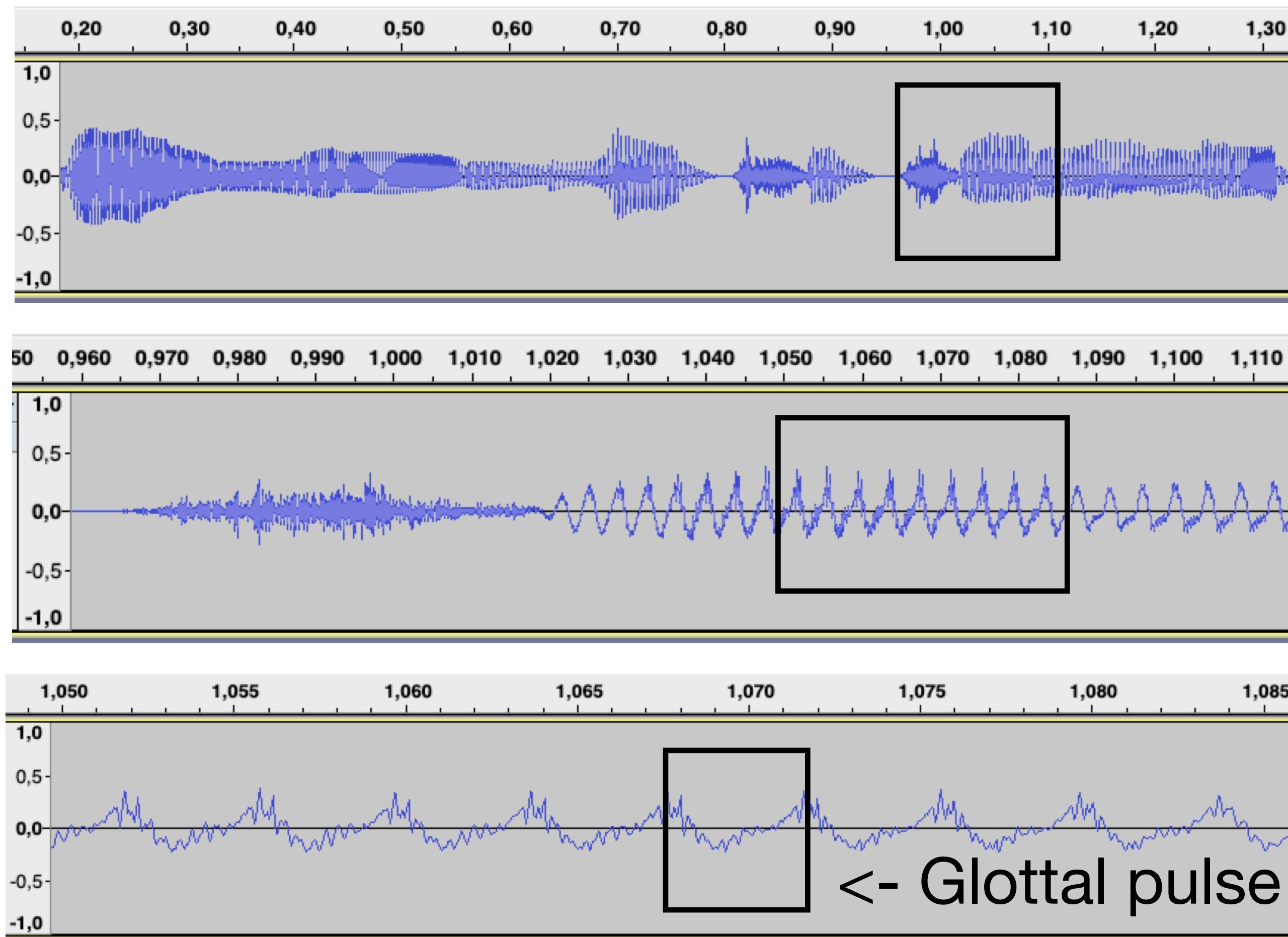


$$y(t) = h(t) * x(t),$$

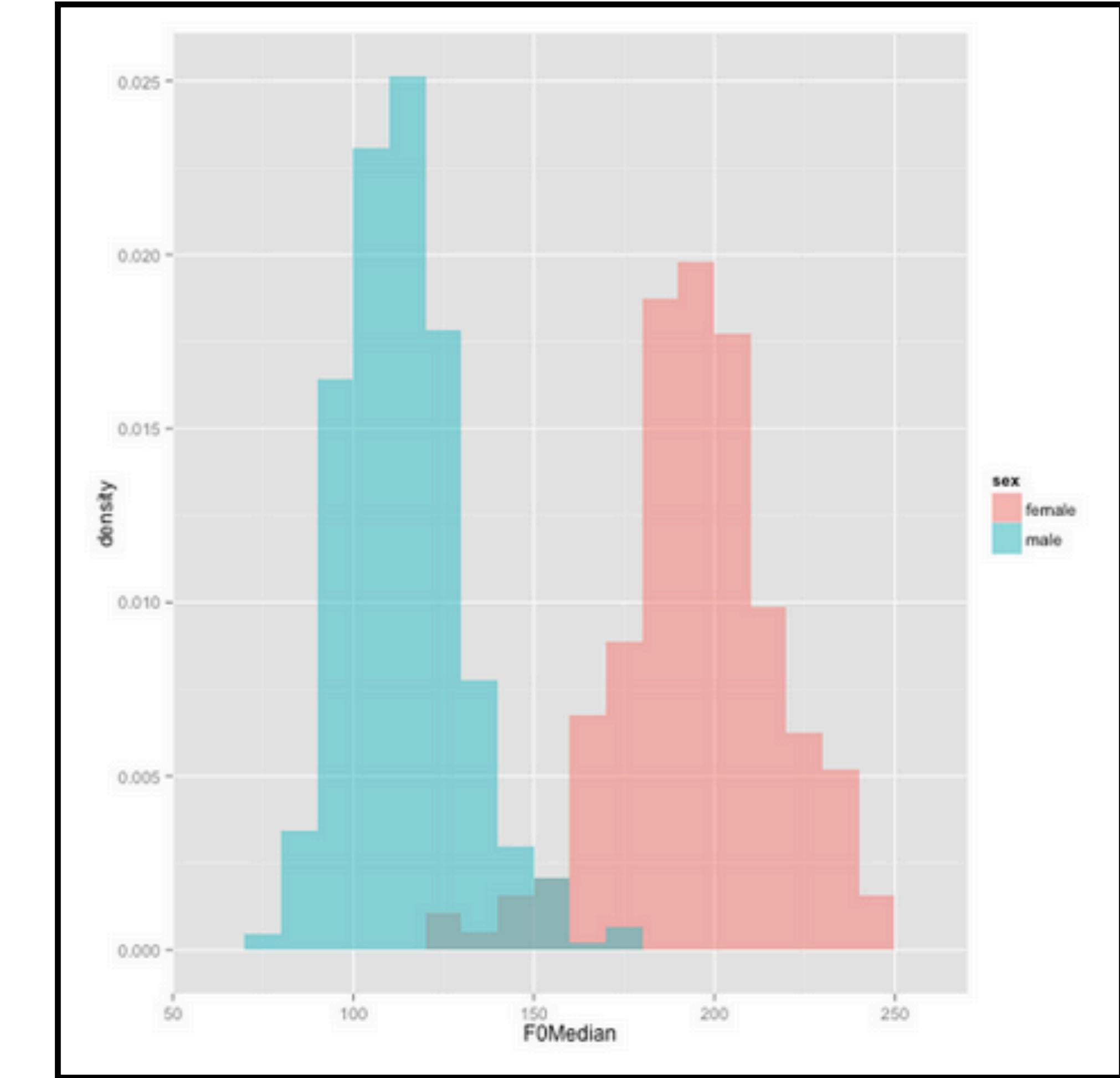
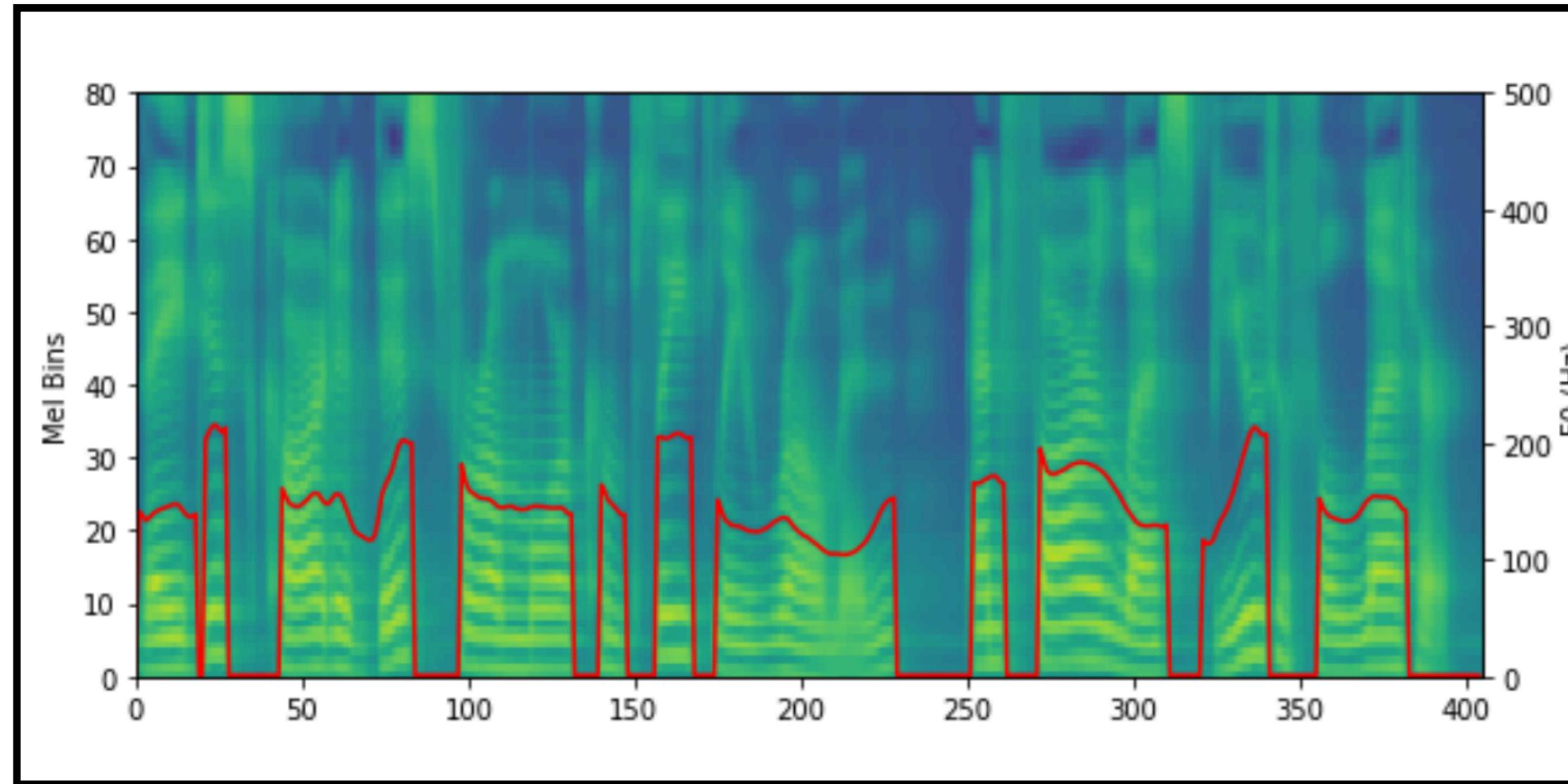
Source-filter theory



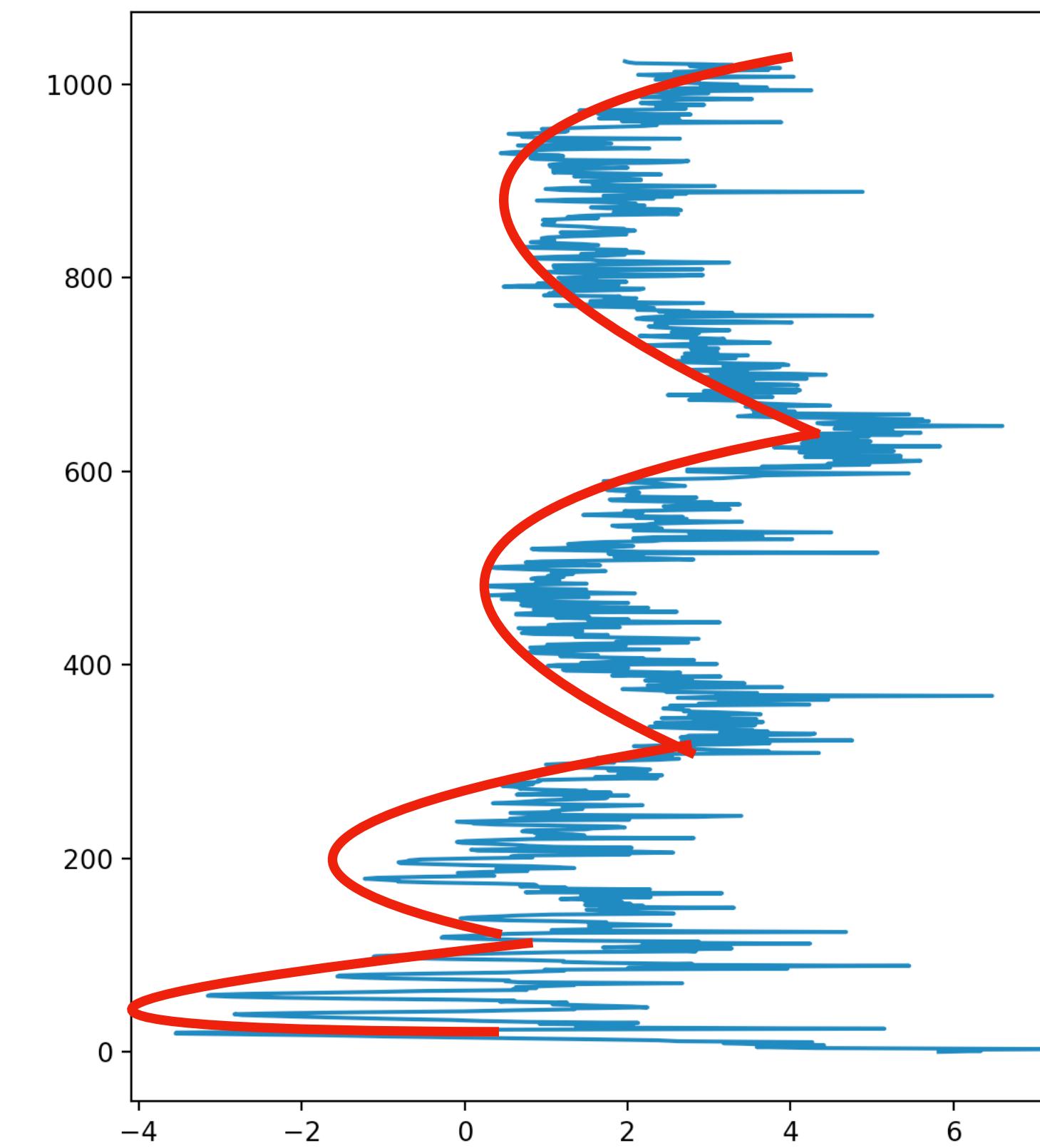
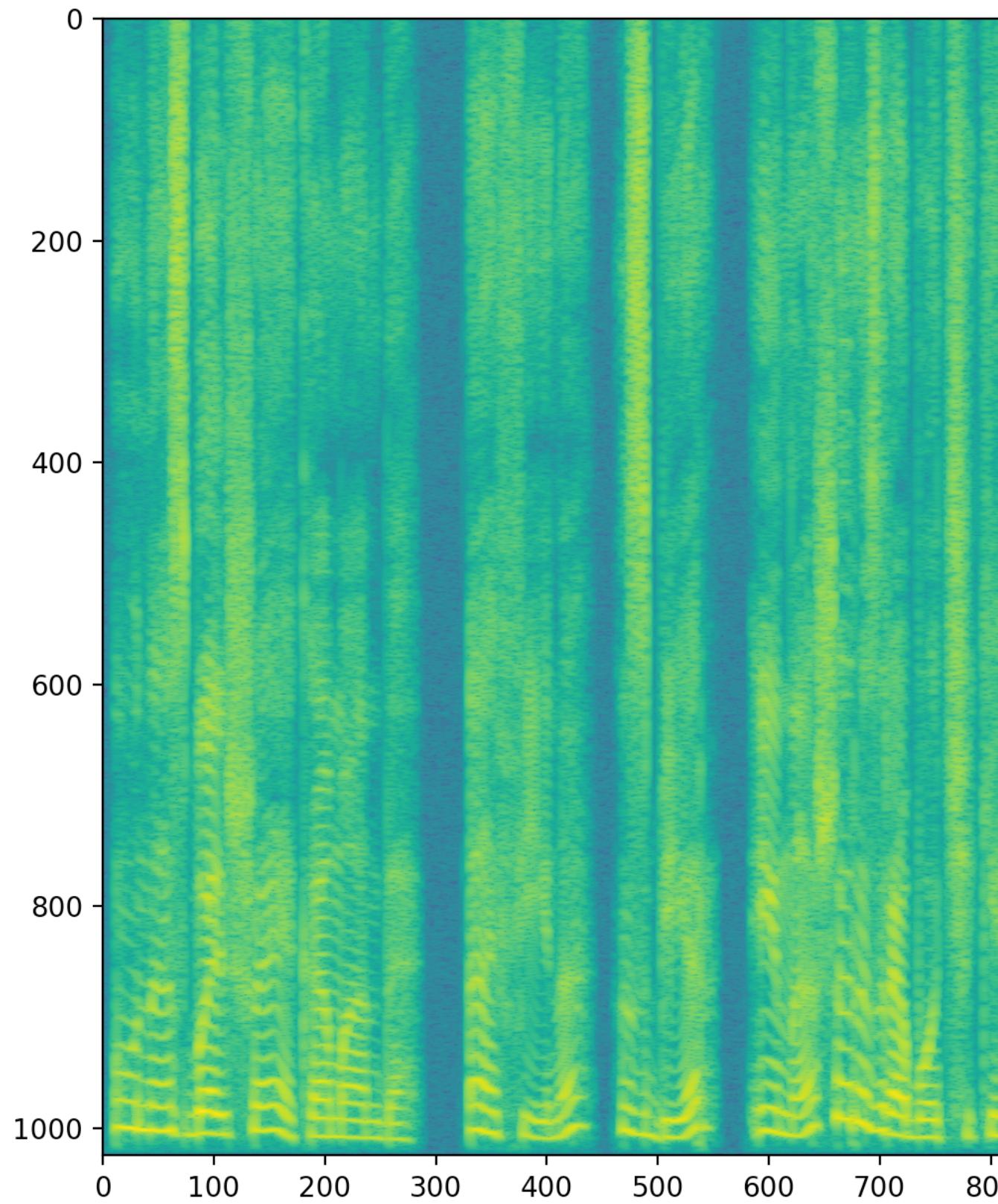
Что такое речь



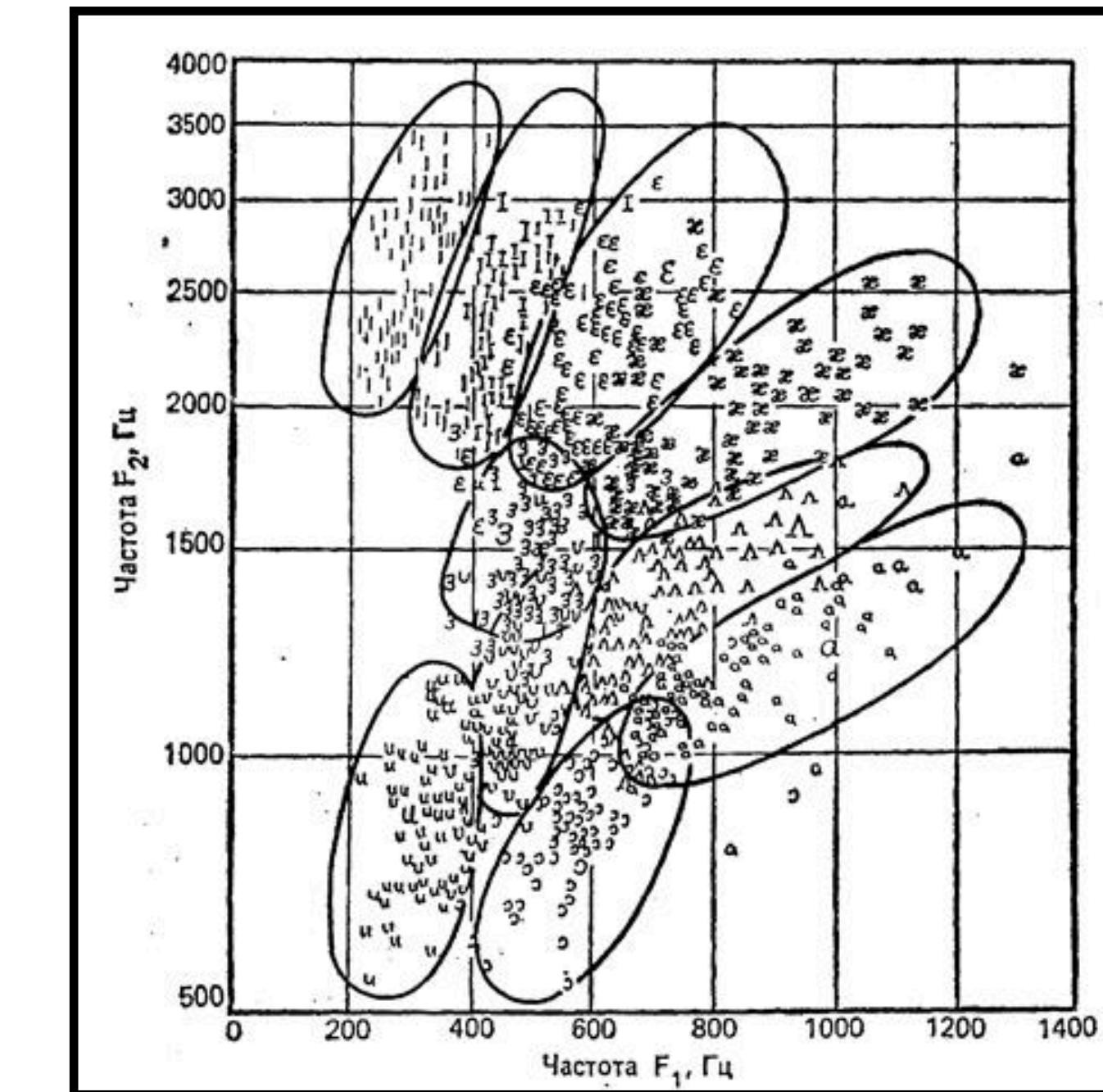
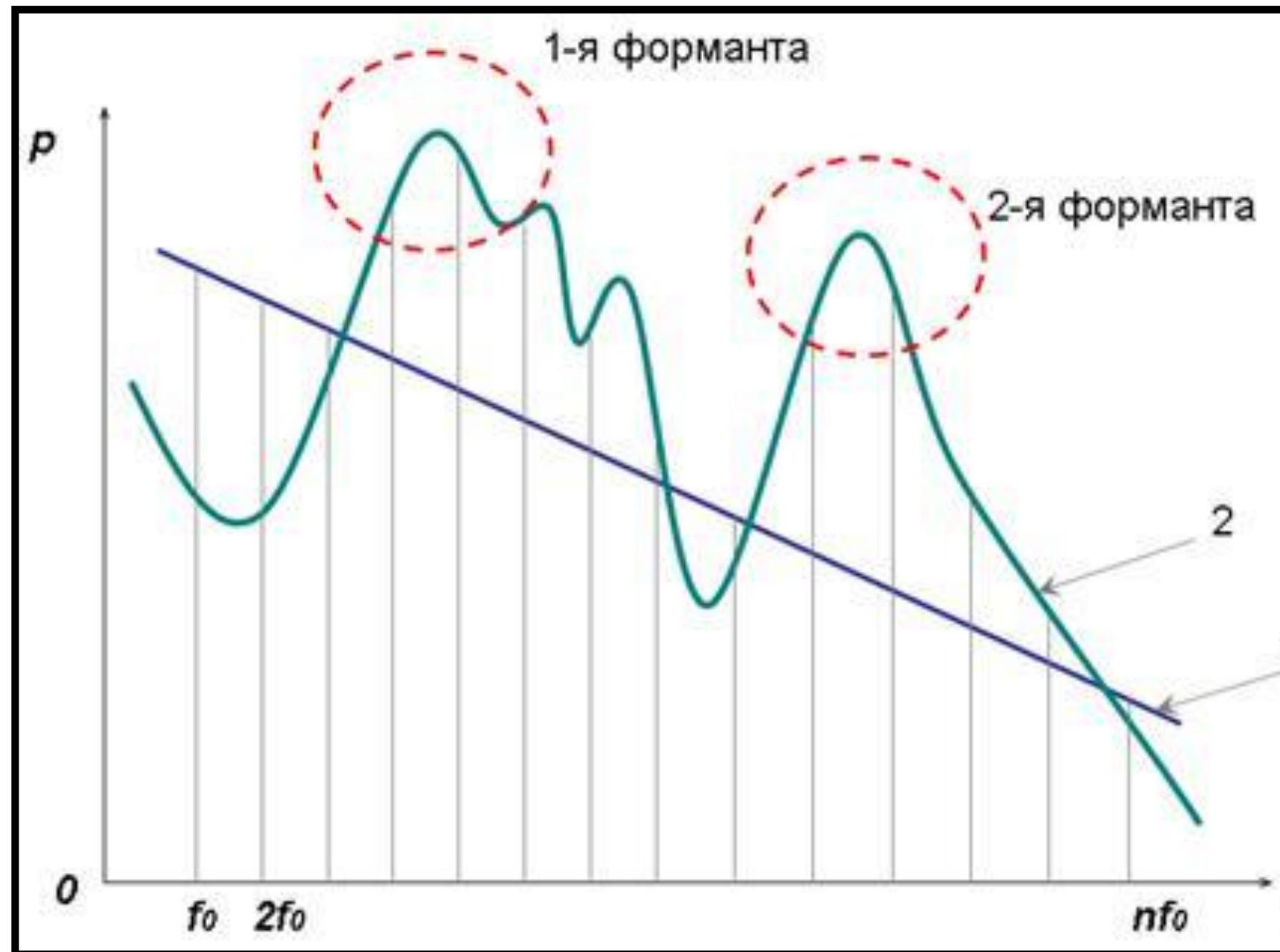
pitch (f0, частота основного тона)



Форманты

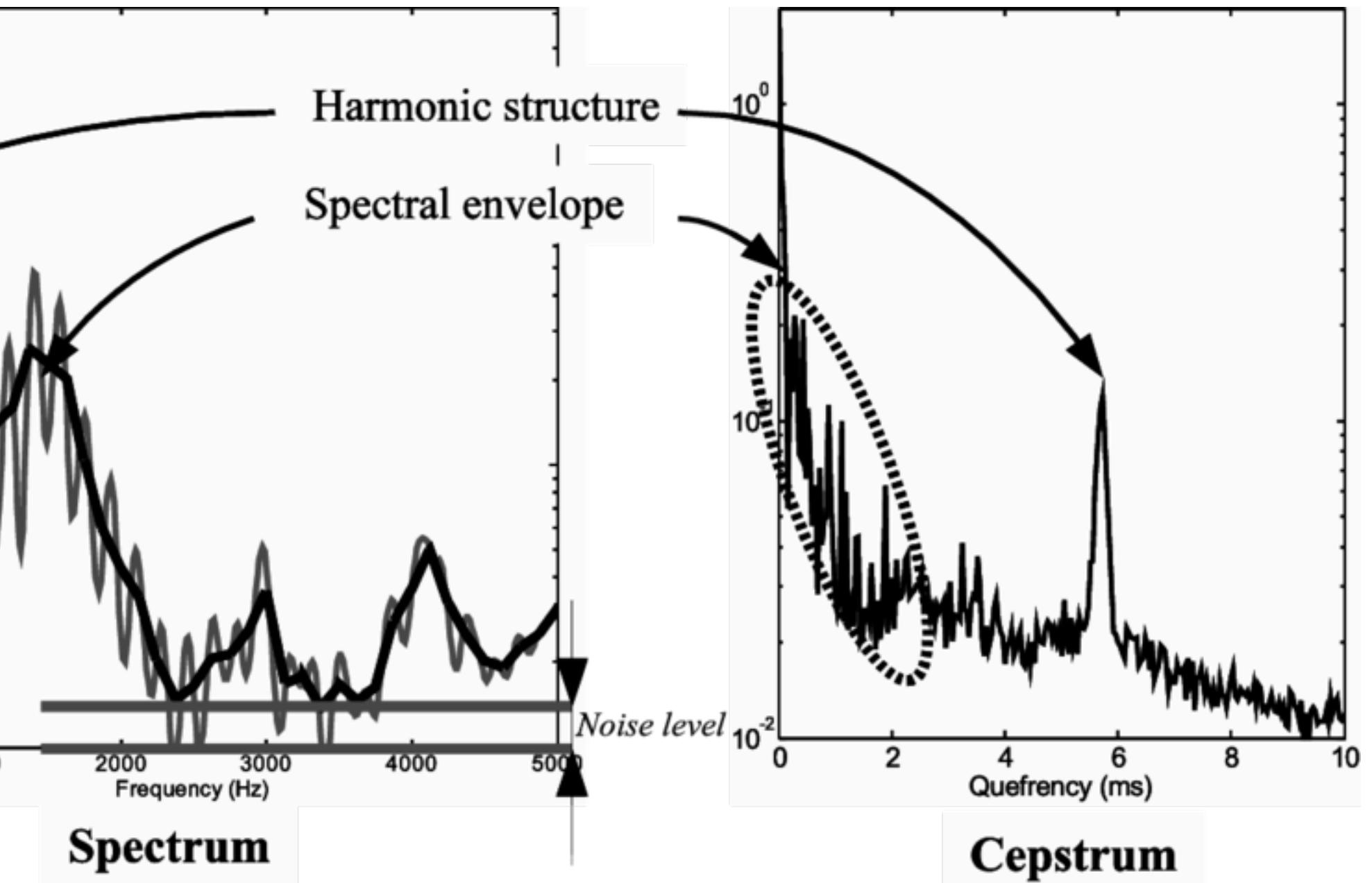
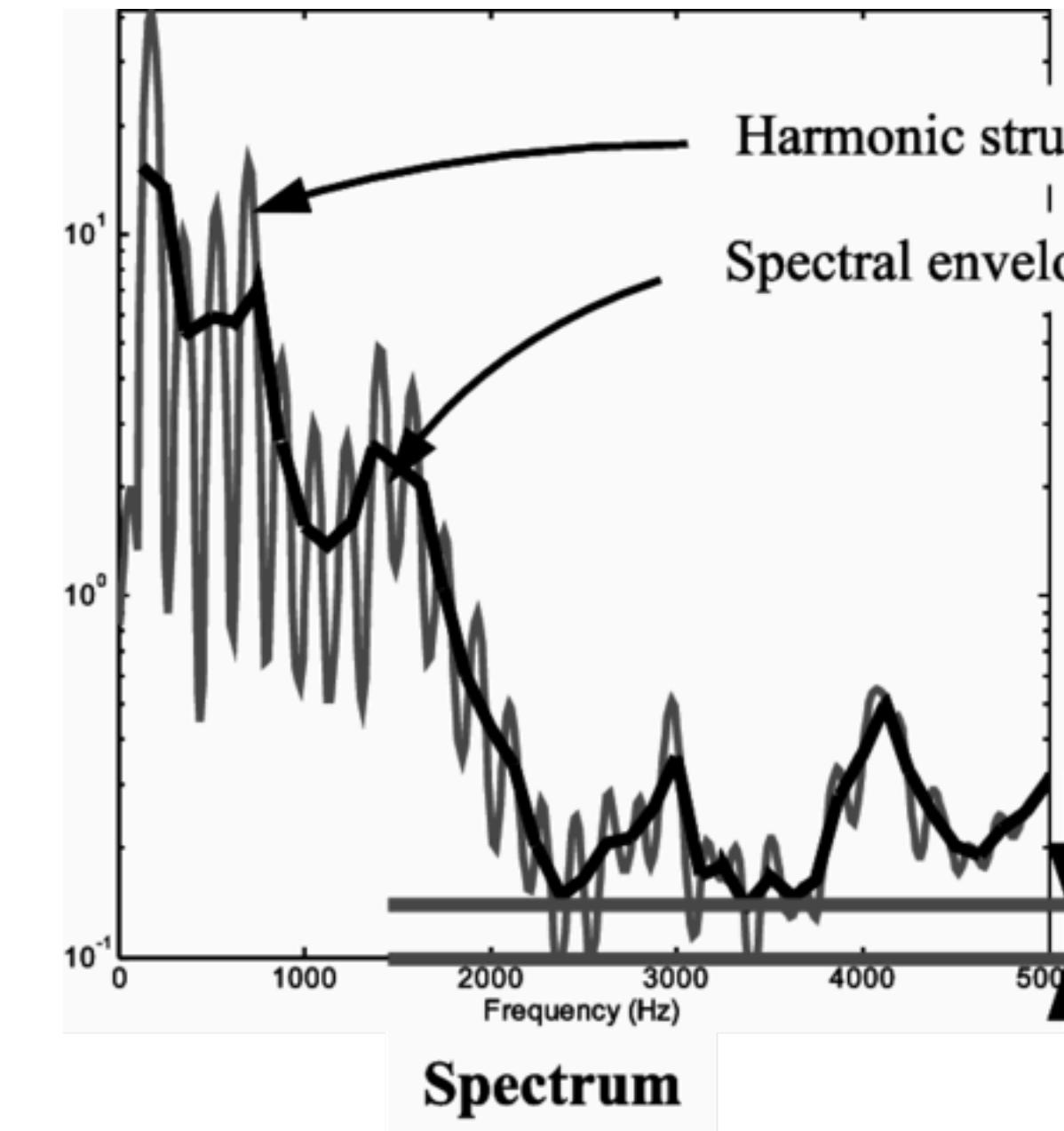
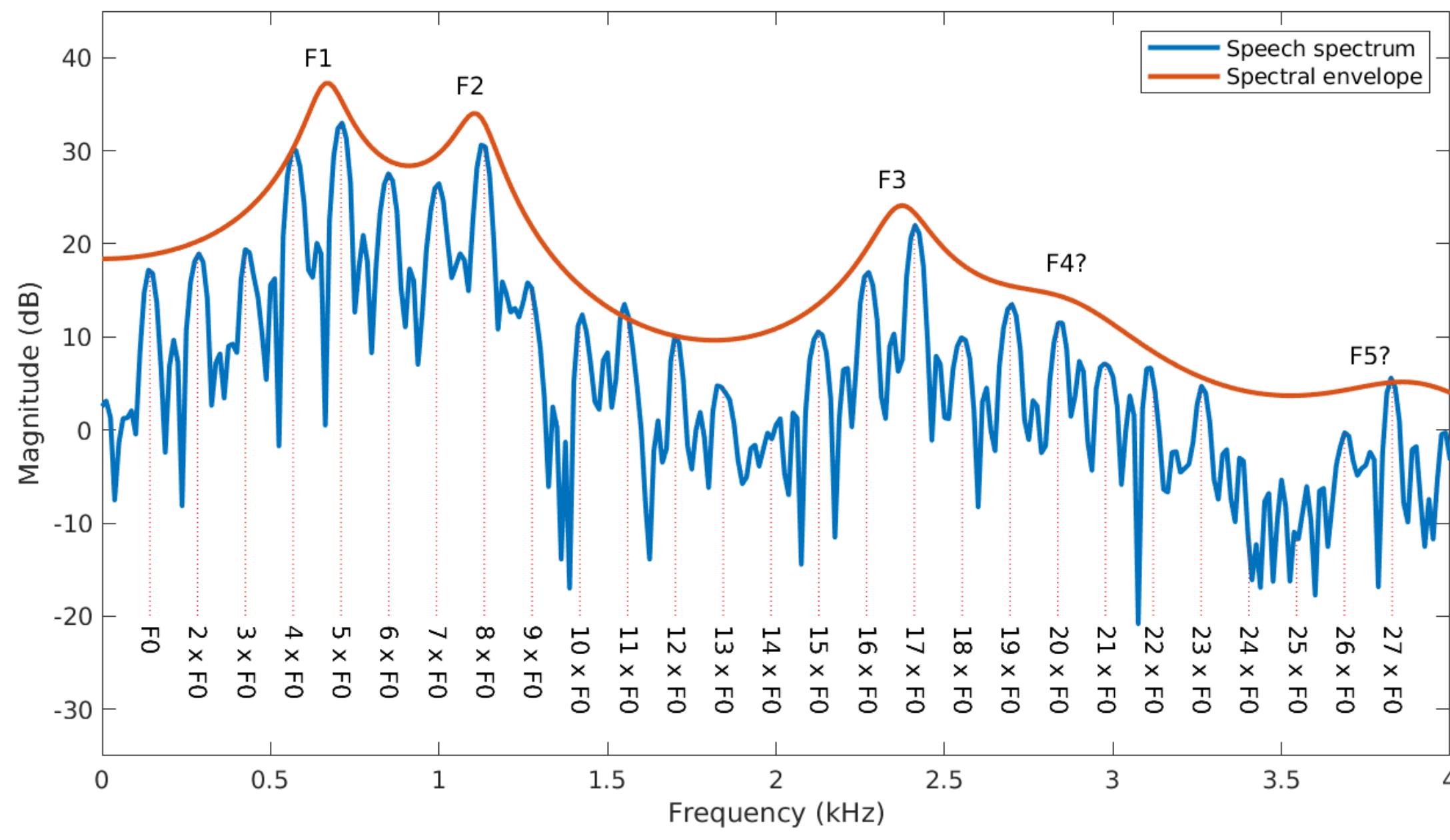


Форманты



Кепстр

- Спектр выглядит как периодический сигнал -> FFT

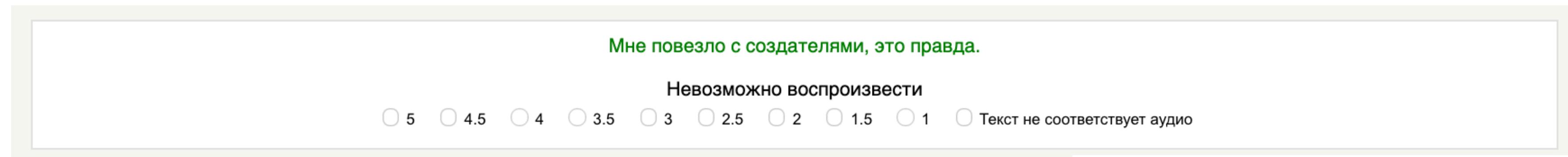


F0, F1, F2, F3, ...

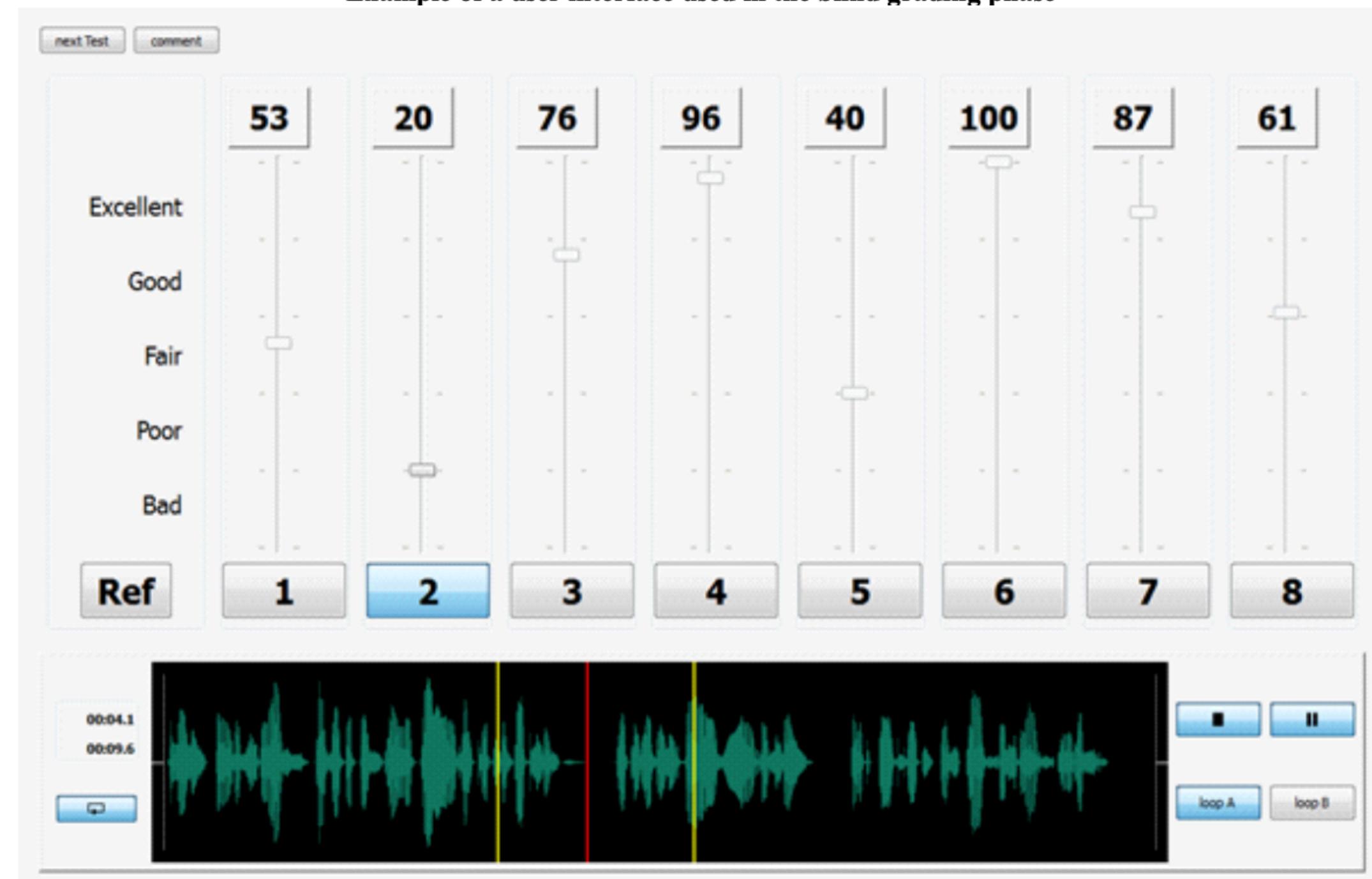
- F0 - частота основного тона (связано с голосовыми связками)
- F1, F2, ... - форманты (связано с остальным речевым трактом)
- Pitch - субъективное восприятие высоты звука (в расчетах берут F0 и не парятся)
- F0 - отвечает за эмоции, интонацию, экспрессию
- Форманты отвечают за звуки (какой звук говорится)

Метрики качества

MOS:



MUSHRA:



$$MOS = \frac{\sum_{n=1}^N R_n}{N}$$

Where R are the individual ratings for a given stimulus by N subjects.

- + reference
- + hidden reference
- + few samples
- + 1-2 anchors

Метрики качества

SBS:
+ CMOS

Я не смог подтвердить перевод. давайте попробуем позже

▶ 0:04 / 0:04 ————— Вариант А ————— 🔍 ⋮

▶ 0:04 / 0:04 ————— Вариант В ————— 🔍 ⋮

1 ⚡ Точно А 2 ⚡ Не могу выбрать 3 ⚡ Точно В

PSER:

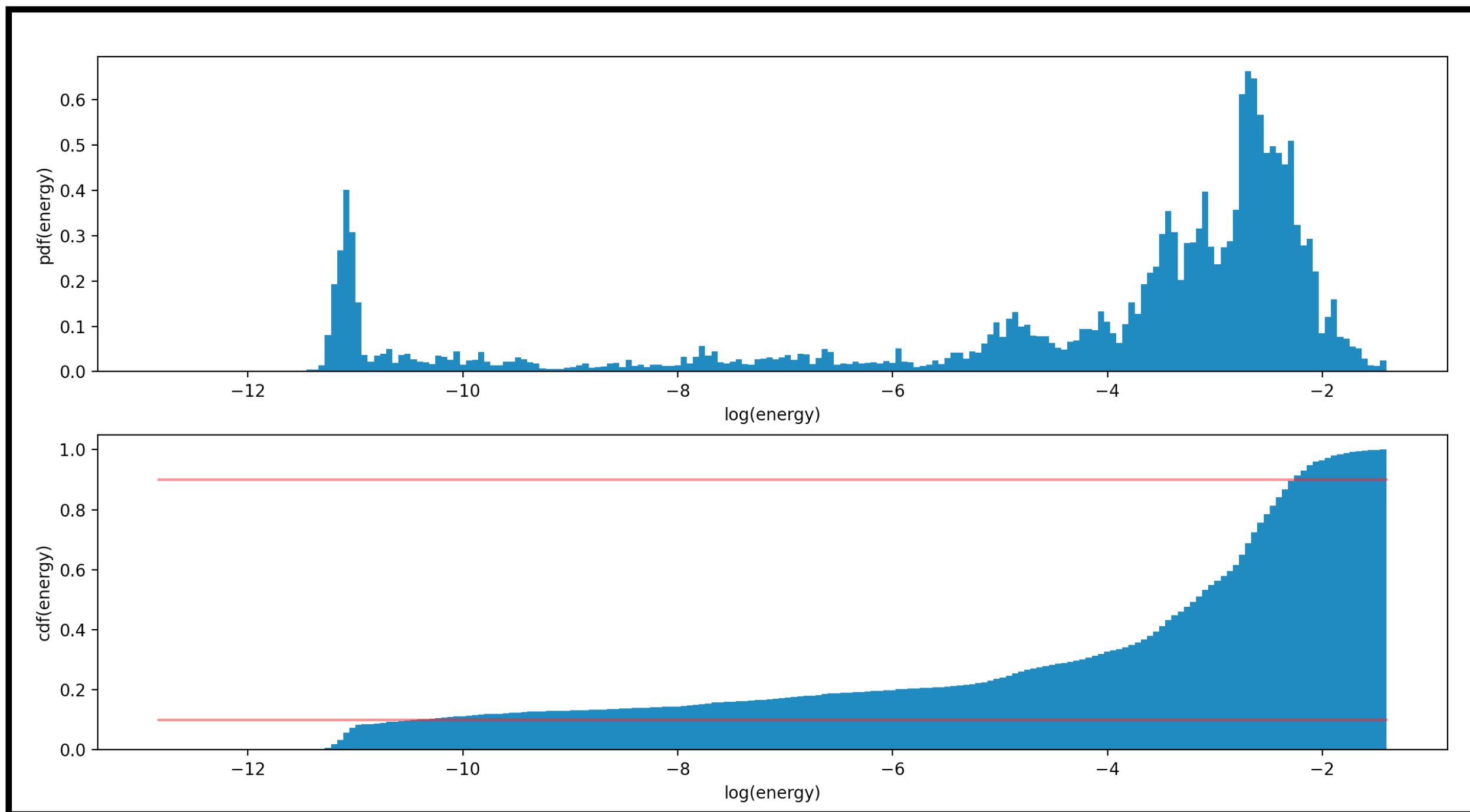
благодарю! мои разработчики будут рады!

▶ 0:00 / 0:03 ————— 🔍 ⋮

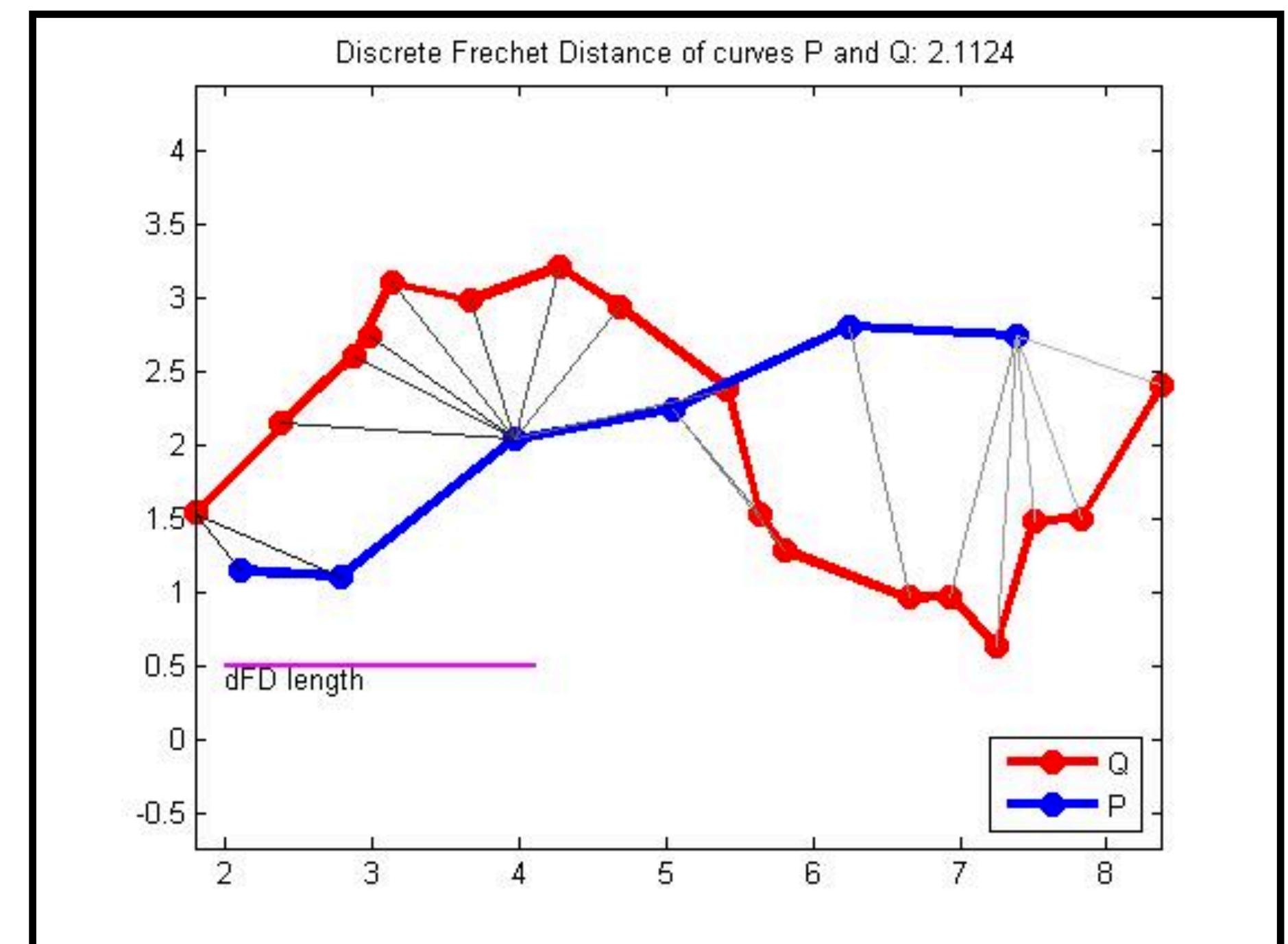
- Лишние паузы
- Не хватает пауз
- Неверное произношение
- Пропущены нужные / вставлены лишние звуки
- Некачественное аудио
- Неверная интонация
- Наличие заиканий
- Всё верно

Метрики качества

SNR:



Frechet distance:



Метрики качества

MCD:

$$\frac{10\sqrt{2}}{\ln 10} \frac{1}{T} \sum_{t=1}^T \sqrt{\sum_i (C_{ti} - \hat{C}_{ti})^2}.$$

Mel-cepstral distortion

nn-based:

MBNet: MOS Prediction for Synthesized Speech with Mean-Bias Network

Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, Tao Qin

Neural MOS Prediction for Synthesized Speech Using Multi-Task Learning With Spoofing Detection and Spoofing Type Classification

Yeunju Choi, Youngmoon Jung, Hoirin Kim

MOSNet: Deep Learning based Objective Assessment for Voice Conversion

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, Hsin-Min Wang

+utmos

Данные для ТТС

Single speaker:

- LJSpeech
- HiFiTTS

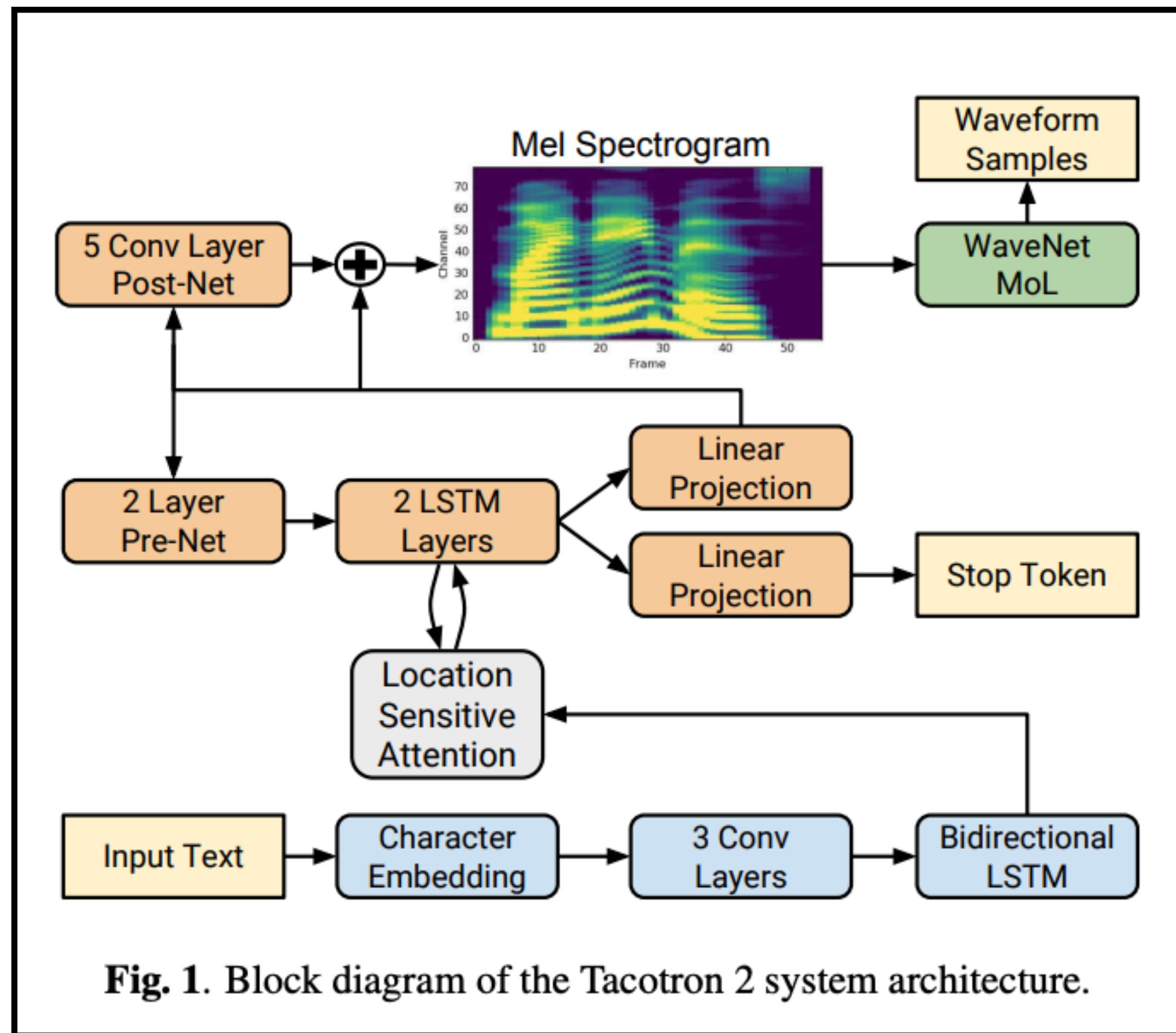
Multi speaker:

- VCTK
- LibriTTS

Дополнительная разметка:

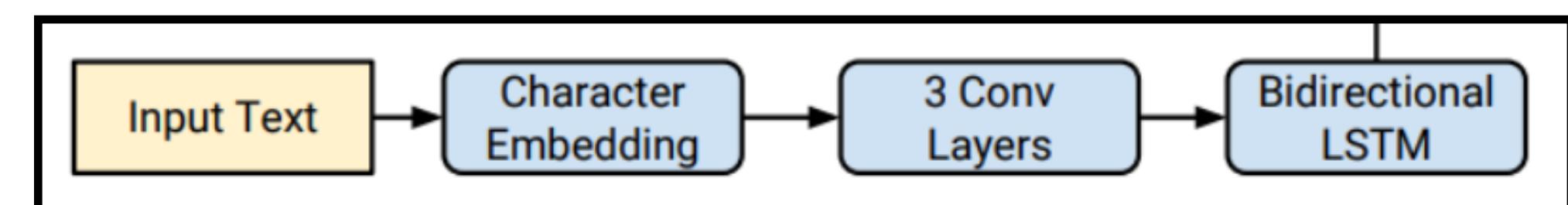
- Montreal Forced Aligner (MFA) hmm-gmm
- Reaper, pyreaper, CREPE
- Audacity
- Yin algorithm (плохой)

Tacotron 2



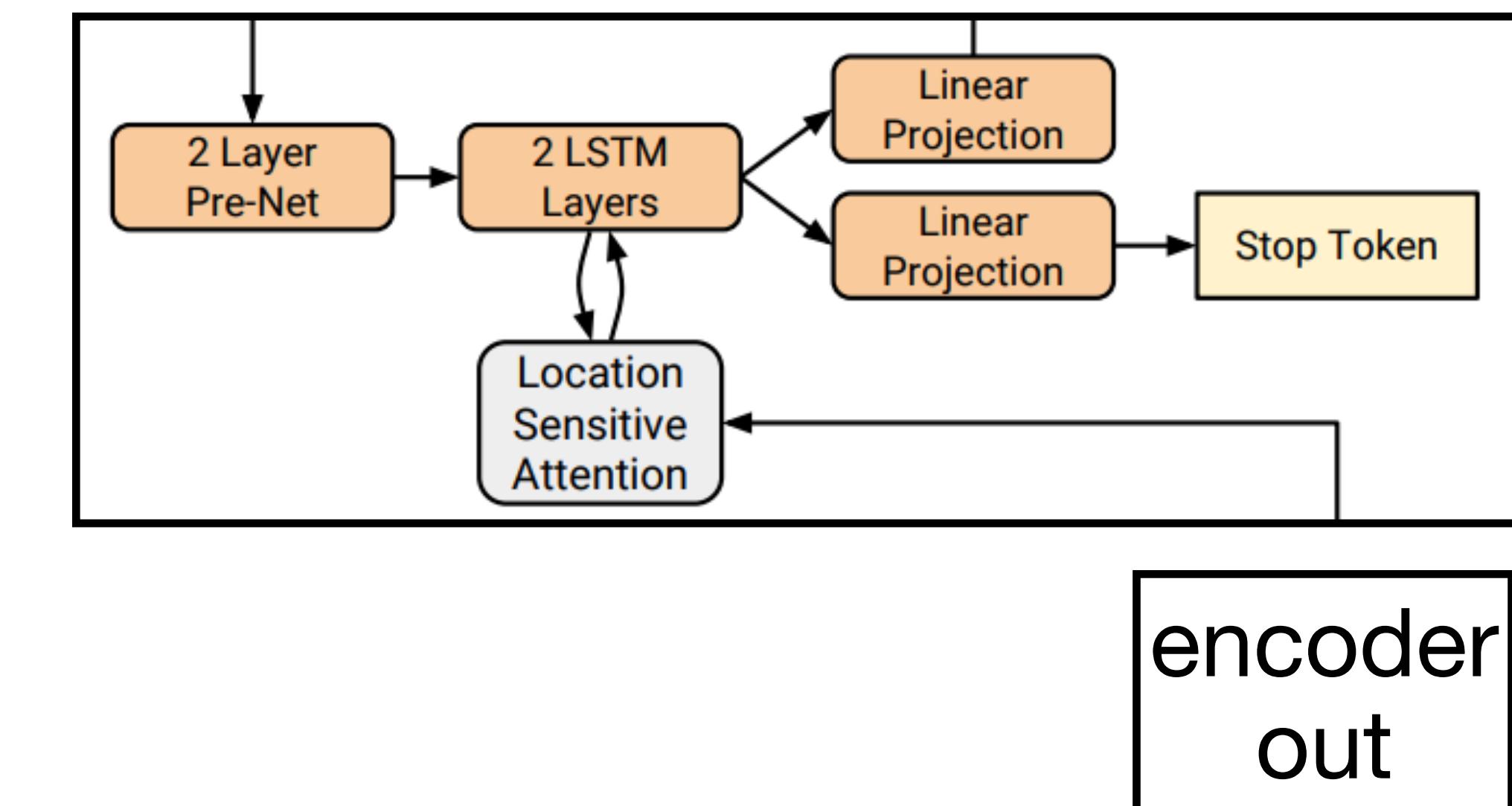
seq2seq:

encoder + attention + decoder
+ postnet



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:

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

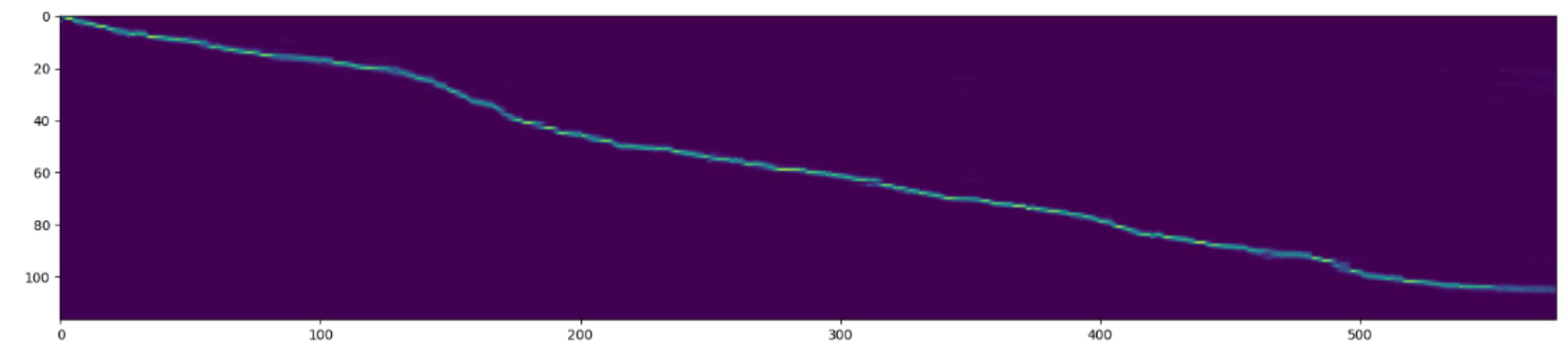
weights:

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

Bahdanau:

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

Всем привет

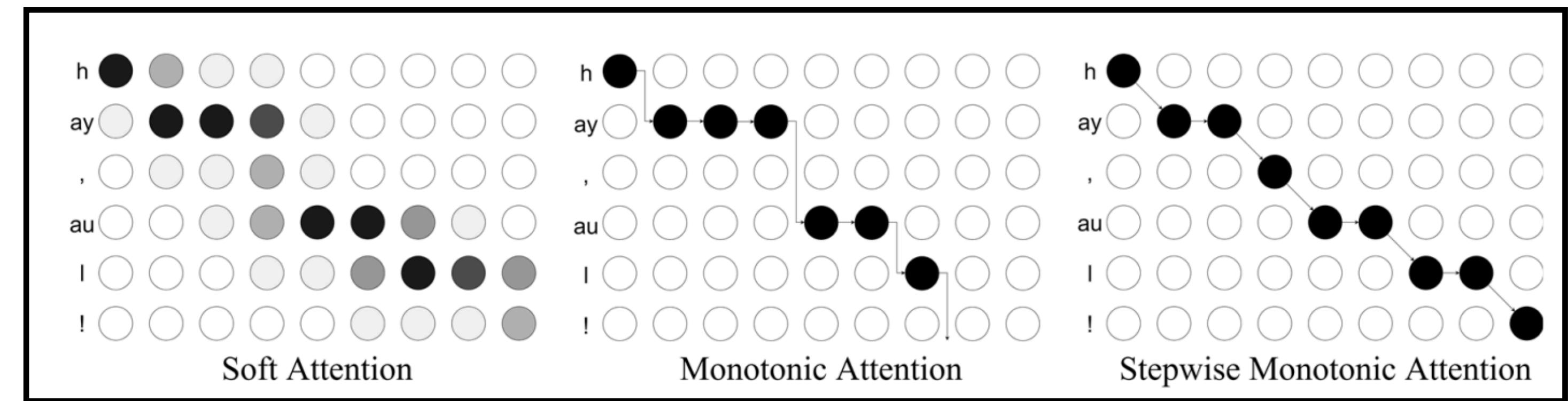


context = sum(encoder_out * alpha)

Tacotron 2 другие attention механизмы

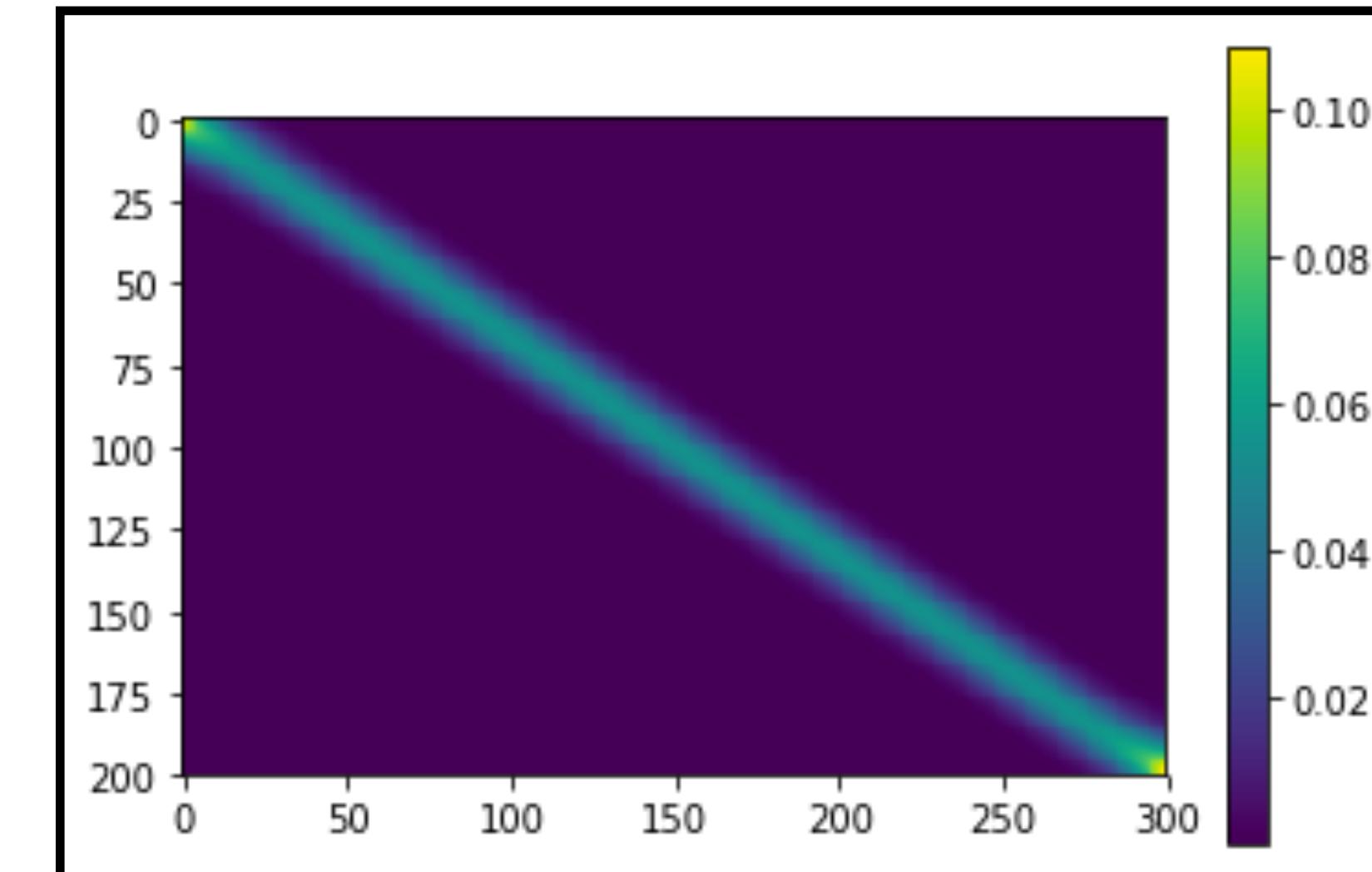
Решаются проблемы:

- 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»

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

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

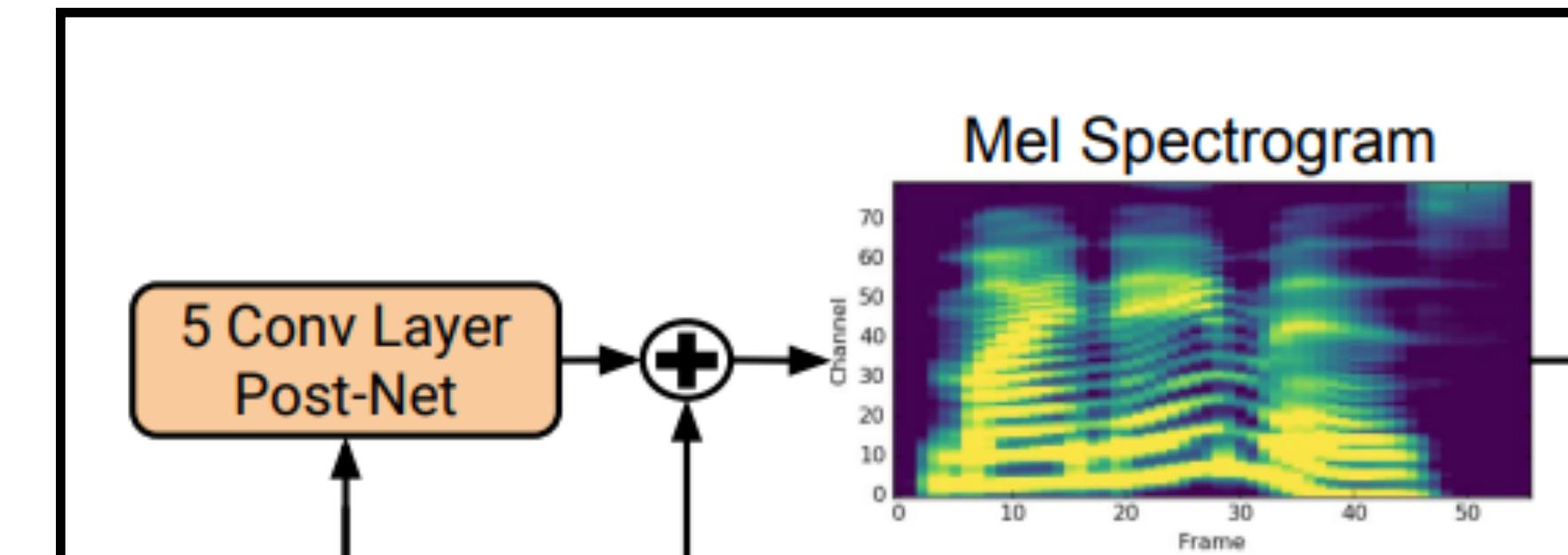
- без дропаута плохо говорит
- синтез каждый раз разный

Vocoder fine-tune:

1.

| Training | Predicted | 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

как = высокоуровневая просодия + низкоуровневая просодия

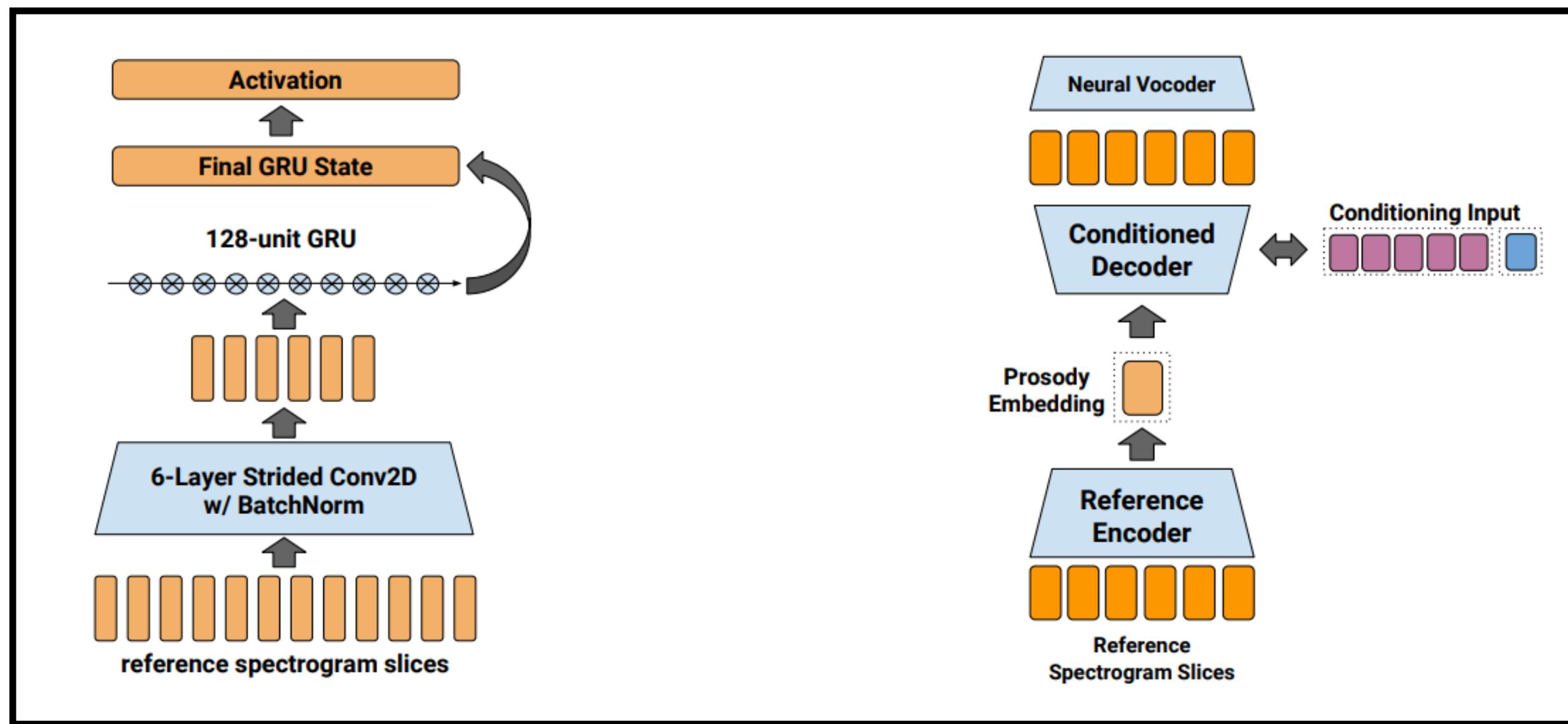
- ЭМОЦИЯ
 - громкость
 - скорость
 - ТОН
- эмфаза
 - вопросы
 - паузы

Будешь чай? Будешь чай?

Global Style Tokens (GST)

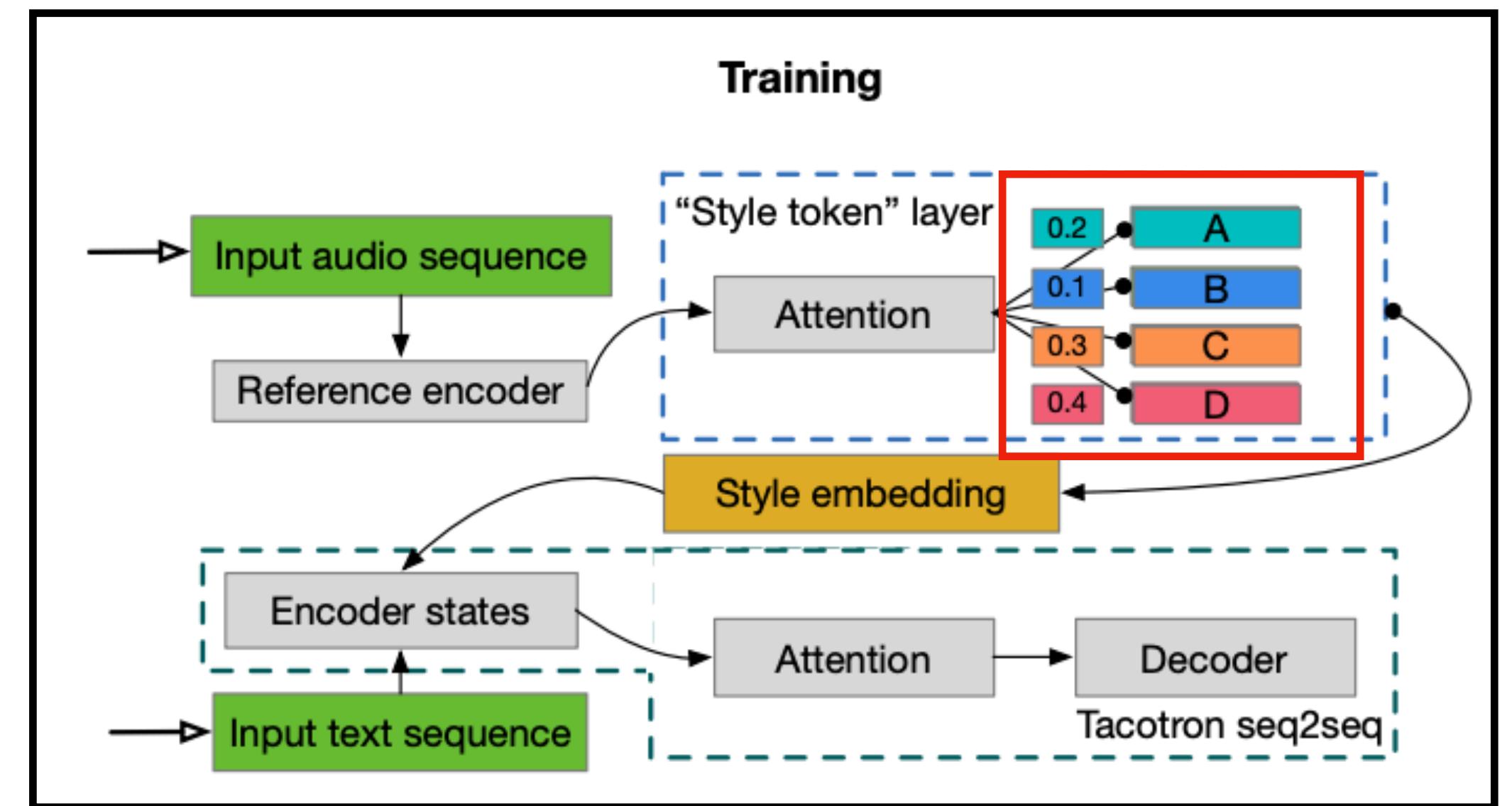
Blizzard challenge 2013

Reference encoder:



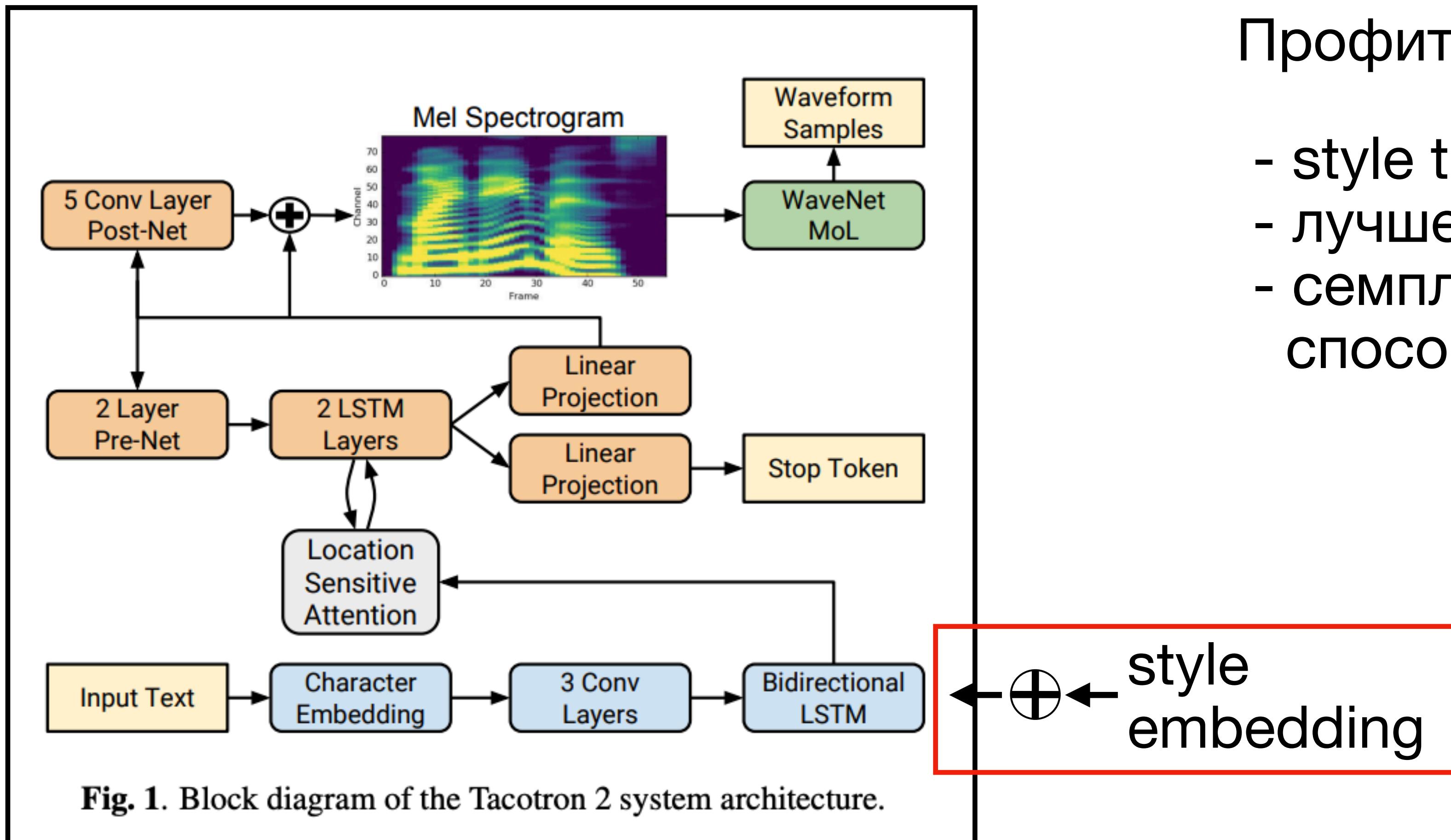
ground truth spec -> style vector

GST:



ground truth spec -> mixture of styles

Global Style Tokens



Профит:

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

\oplus ← style
embedding

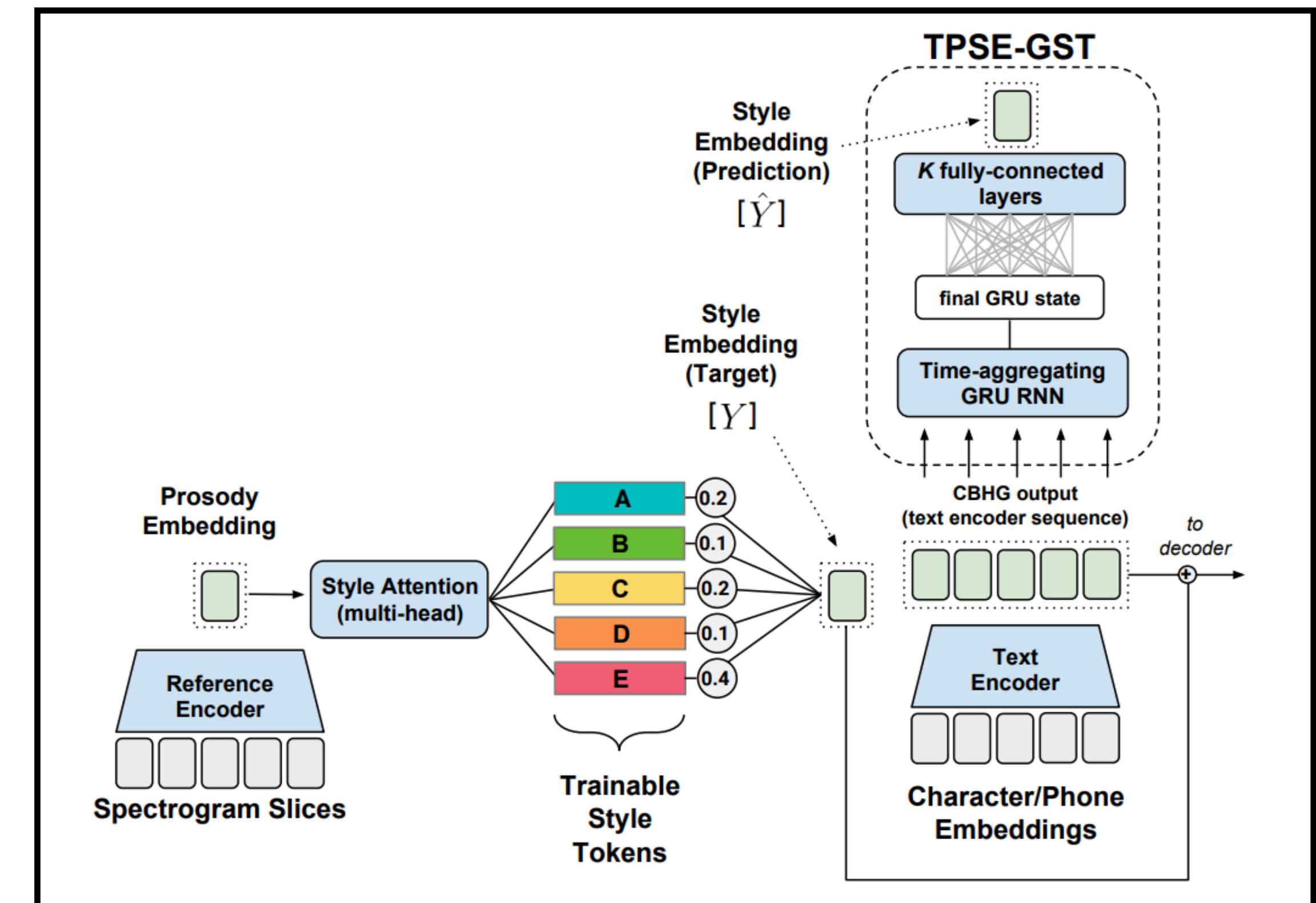
Энкодер = текст энкодер + стиль

Global Style Tokens

Text predicted style embedding:

Проблемы:

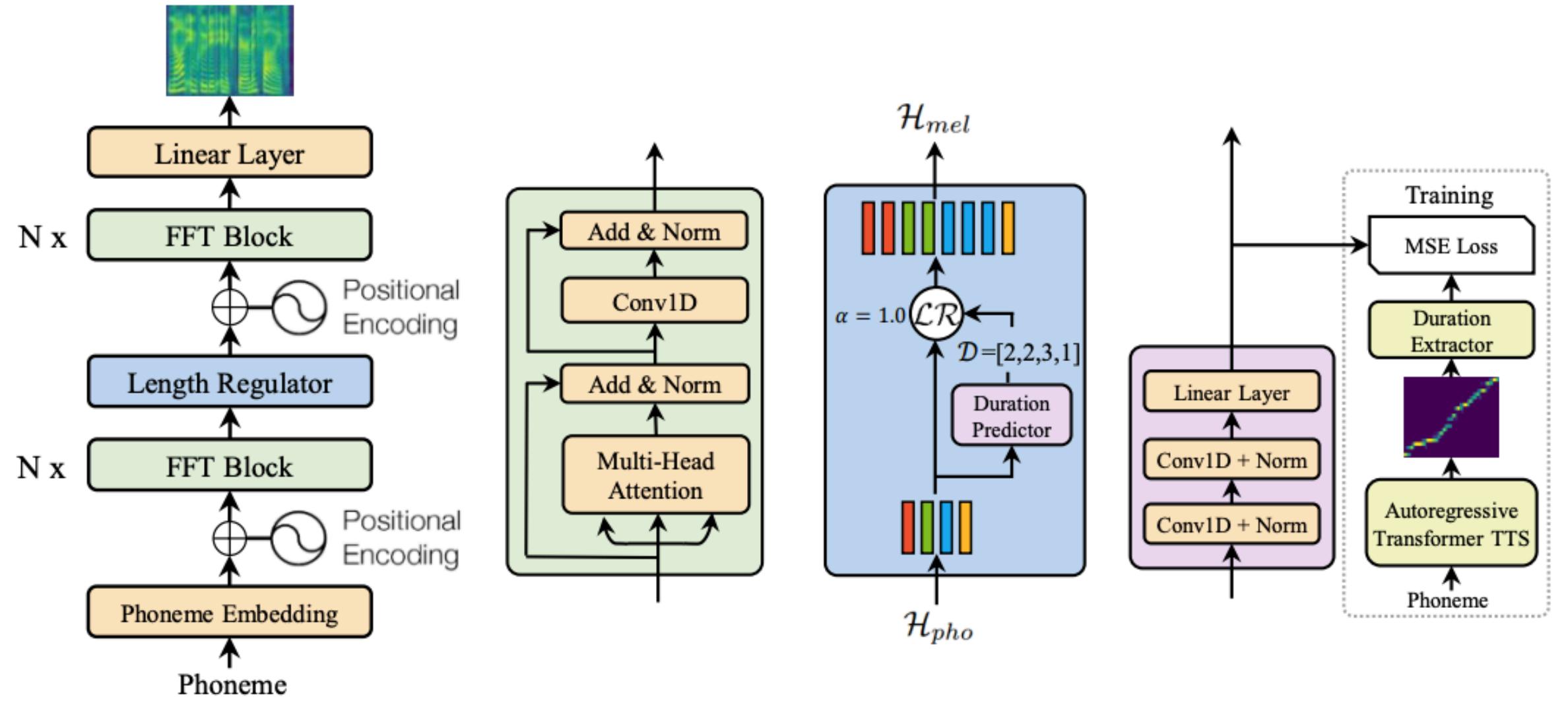
- не воспроизводится :)
- стиль выучивает длину, громкость и тон
- не интерпретируется
- неоткуда брать референс



Non-autoregressive TTS

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

Diffusion-based:



(a) Feed-Forward Transformer

(b) FFT Block

(c) Length Regulator

(d) Duration Predictor

