

Синтез речи

Лекция №1

Гриша Стерлинг, SberDevices

Синтез речи

Задача:

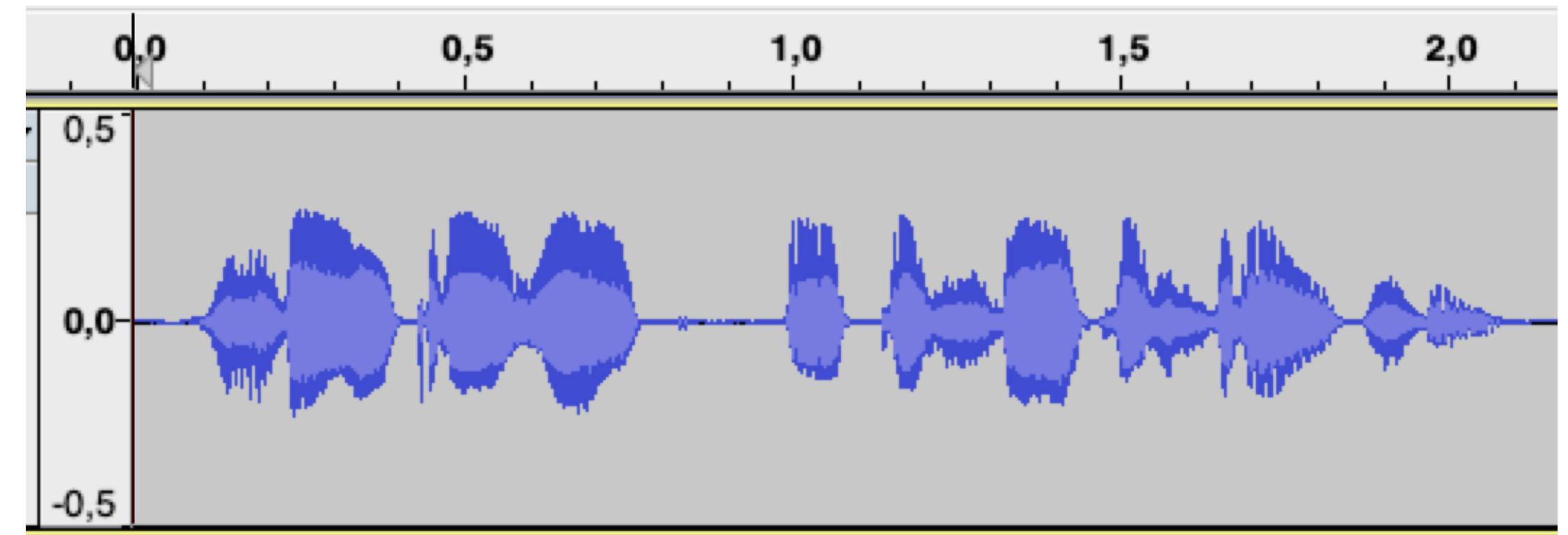
- озвучить заданный текст голосом

Задачи со звездочкой:

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

«Всем привет, это синтез речи»

->

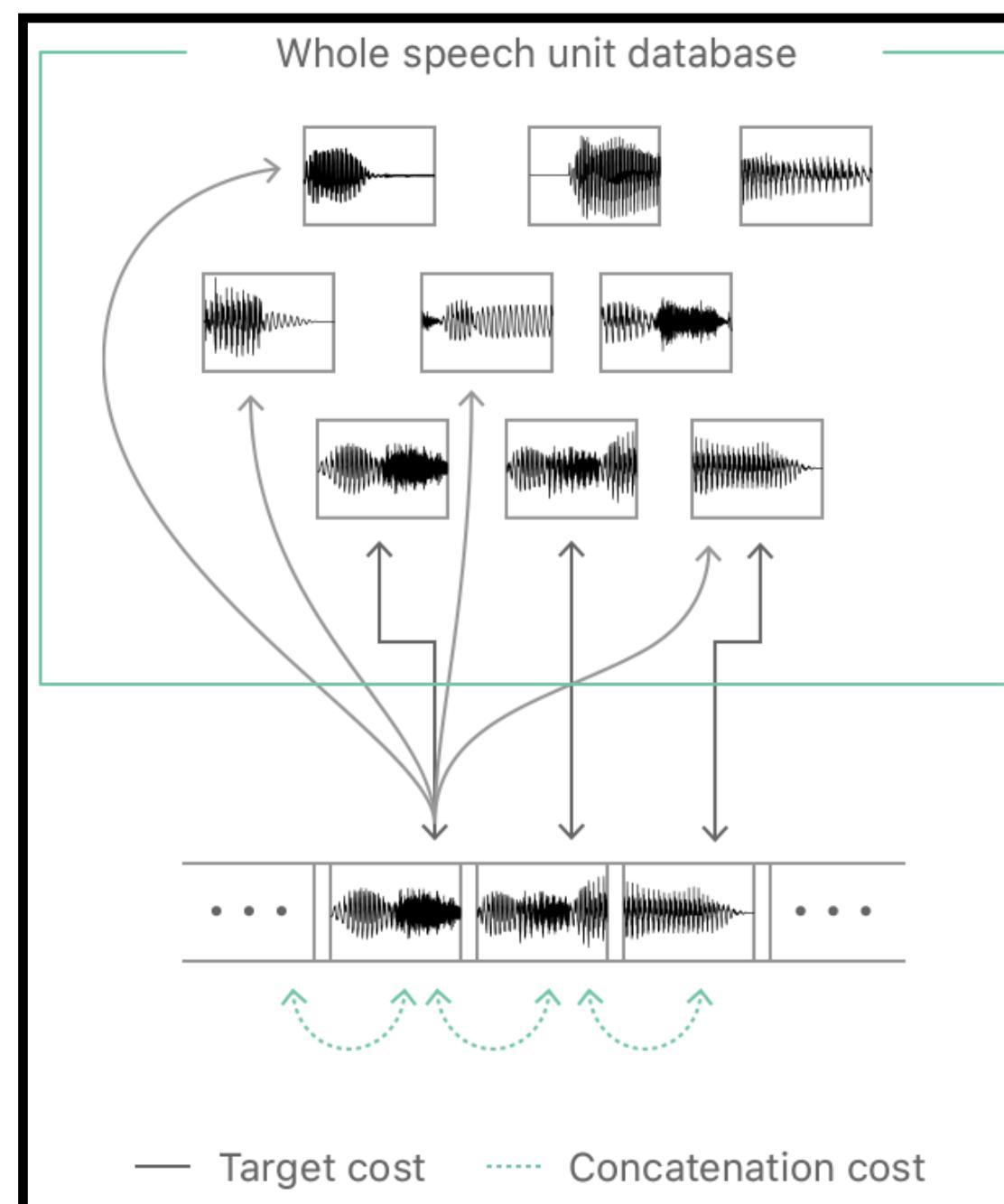


Синтез речи

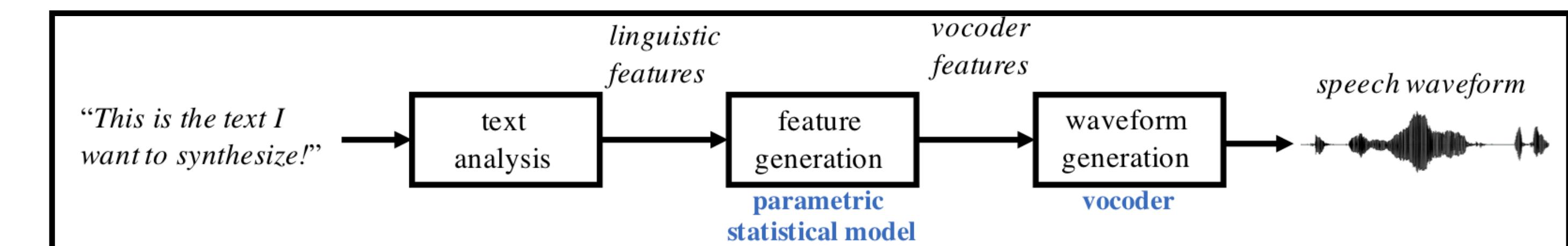


Concatenative
(unit selection)

Parametric

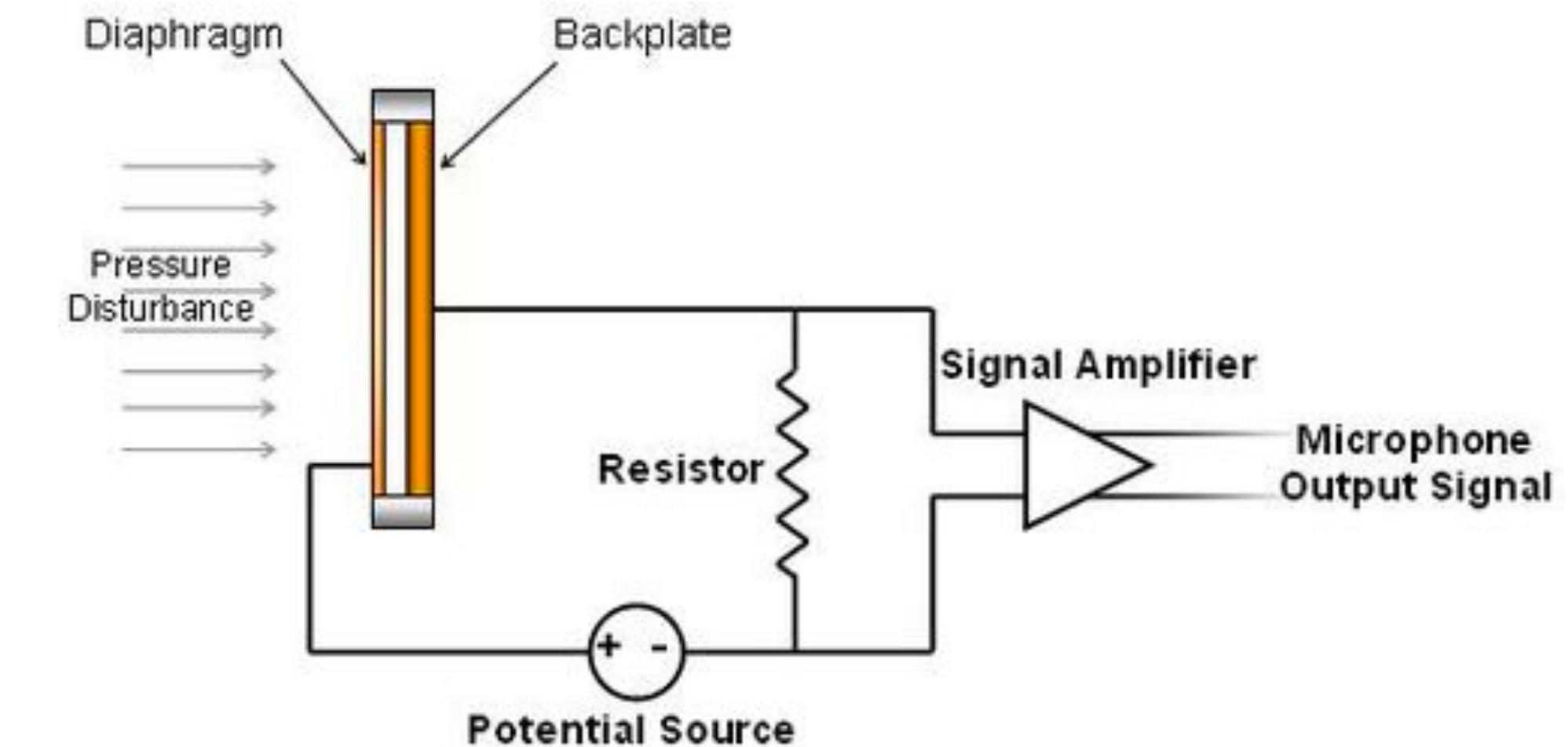
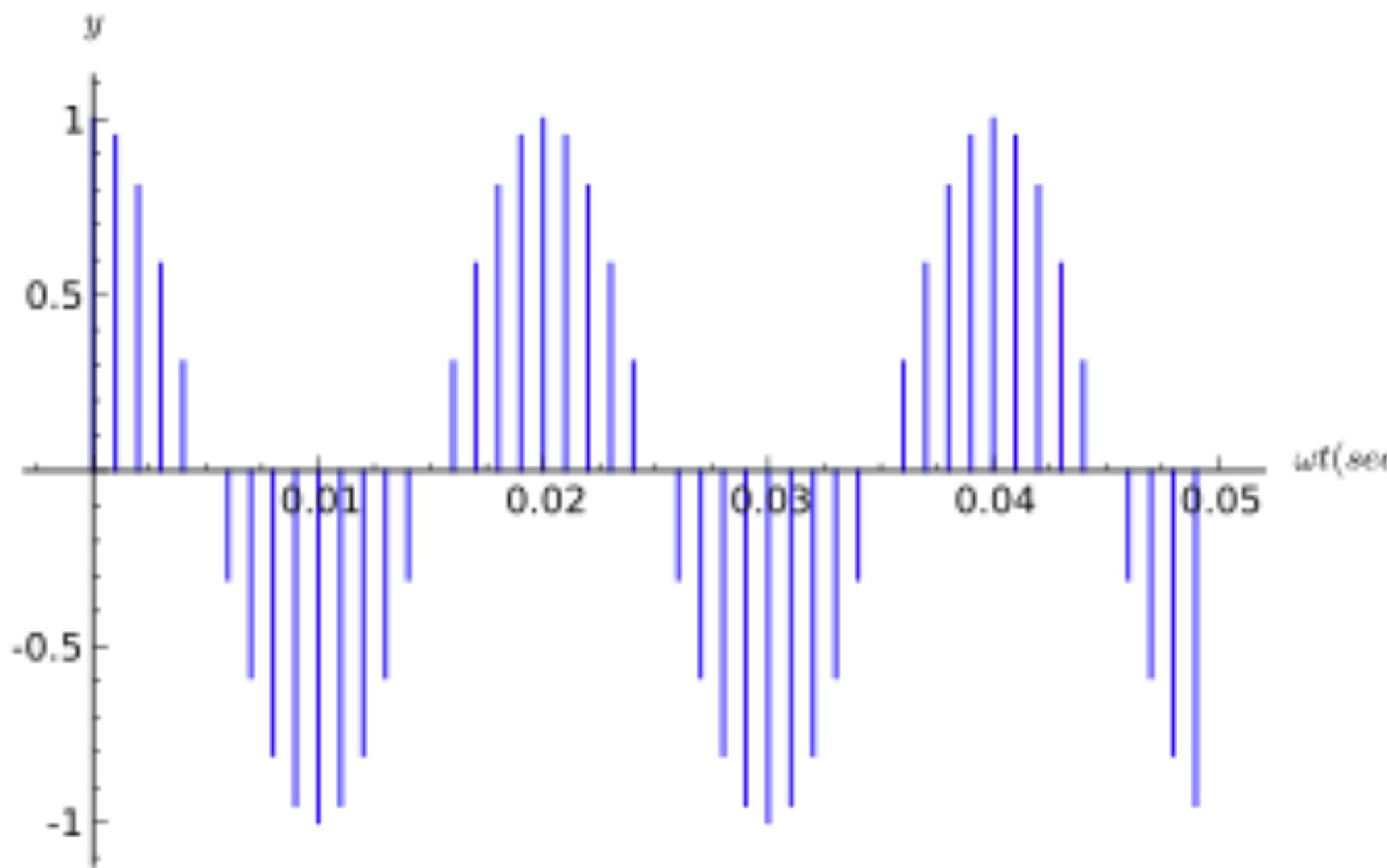


- 2 стадии:
- акустическая модель
 - вокодер (**voice coder**)



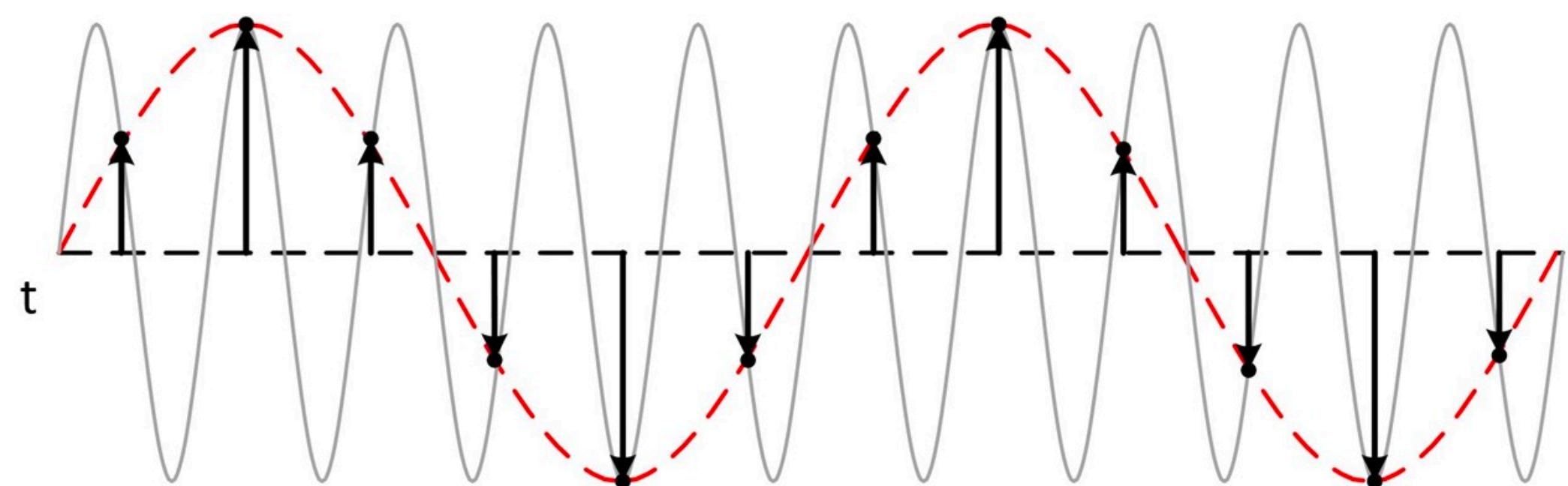
Цифровое представление звука

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



Цифровой сигнал:

- Теорема Котельникова

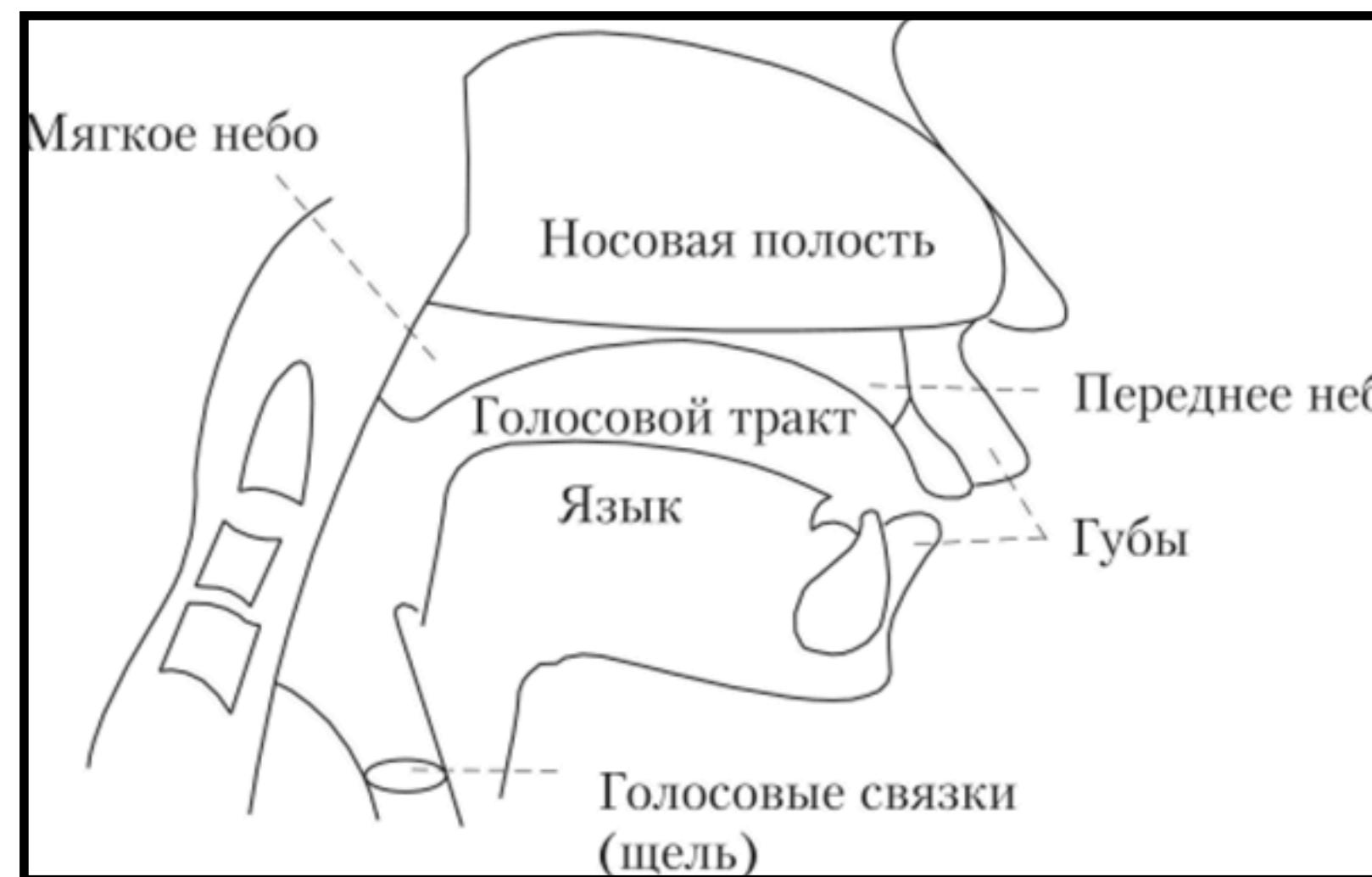


[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

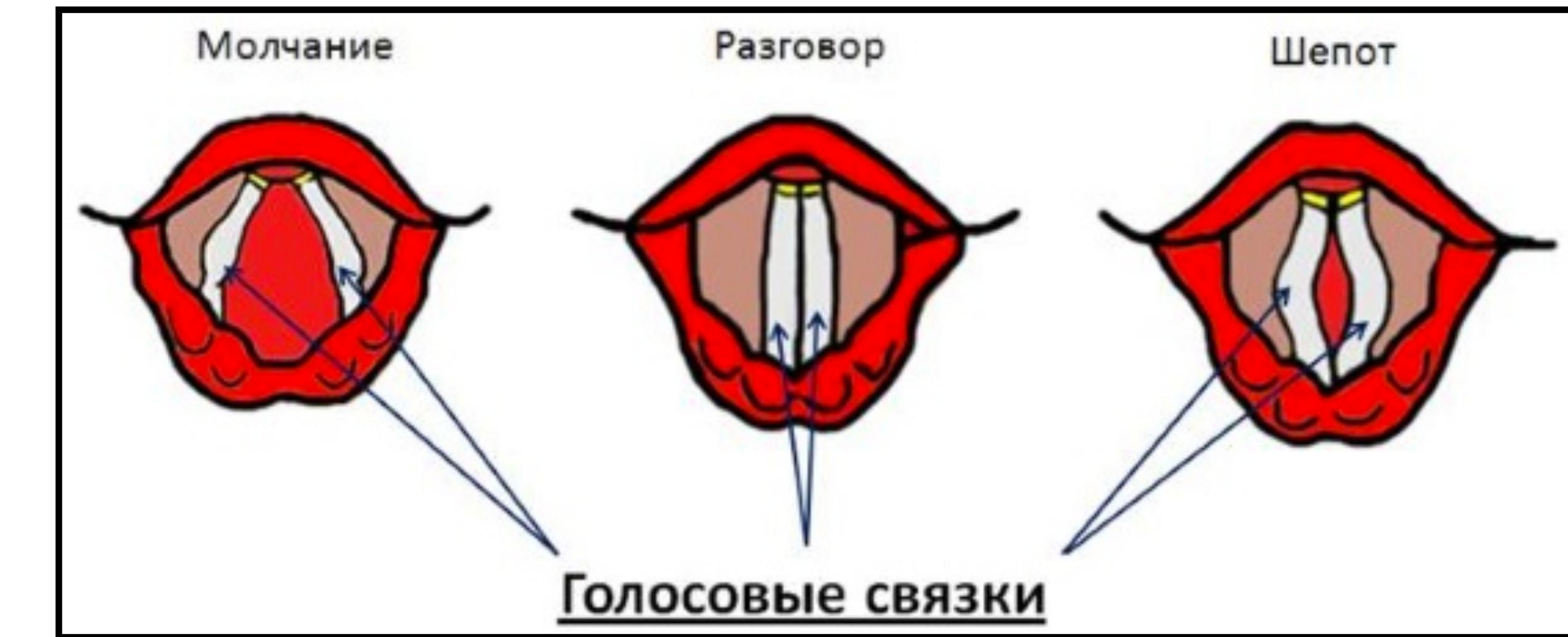
- Параметры:
 - sample rate (частота дискретизации):
8 кГц .. 48 кГц
 - encoding (bits per sample)
8, 16, 24, 32, 64
 - mono, stereo

Что такое речь

Речевой тракт человека

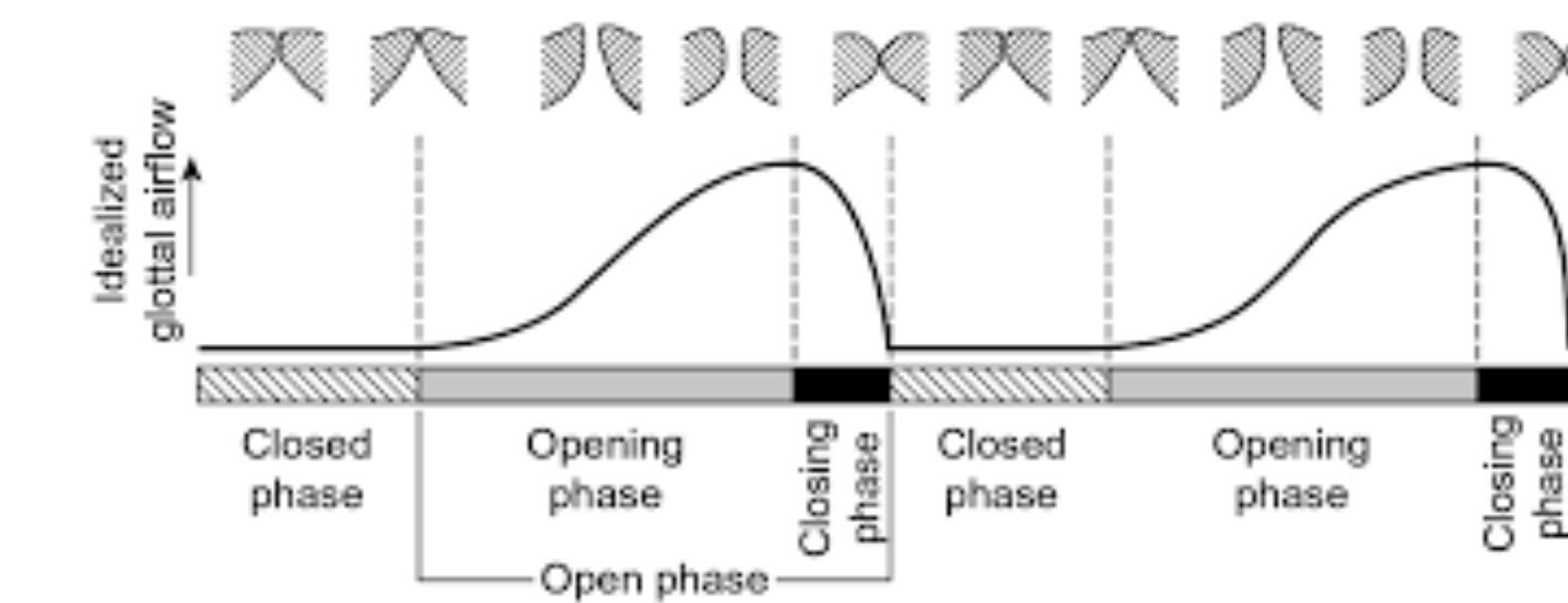


Голосовые связки

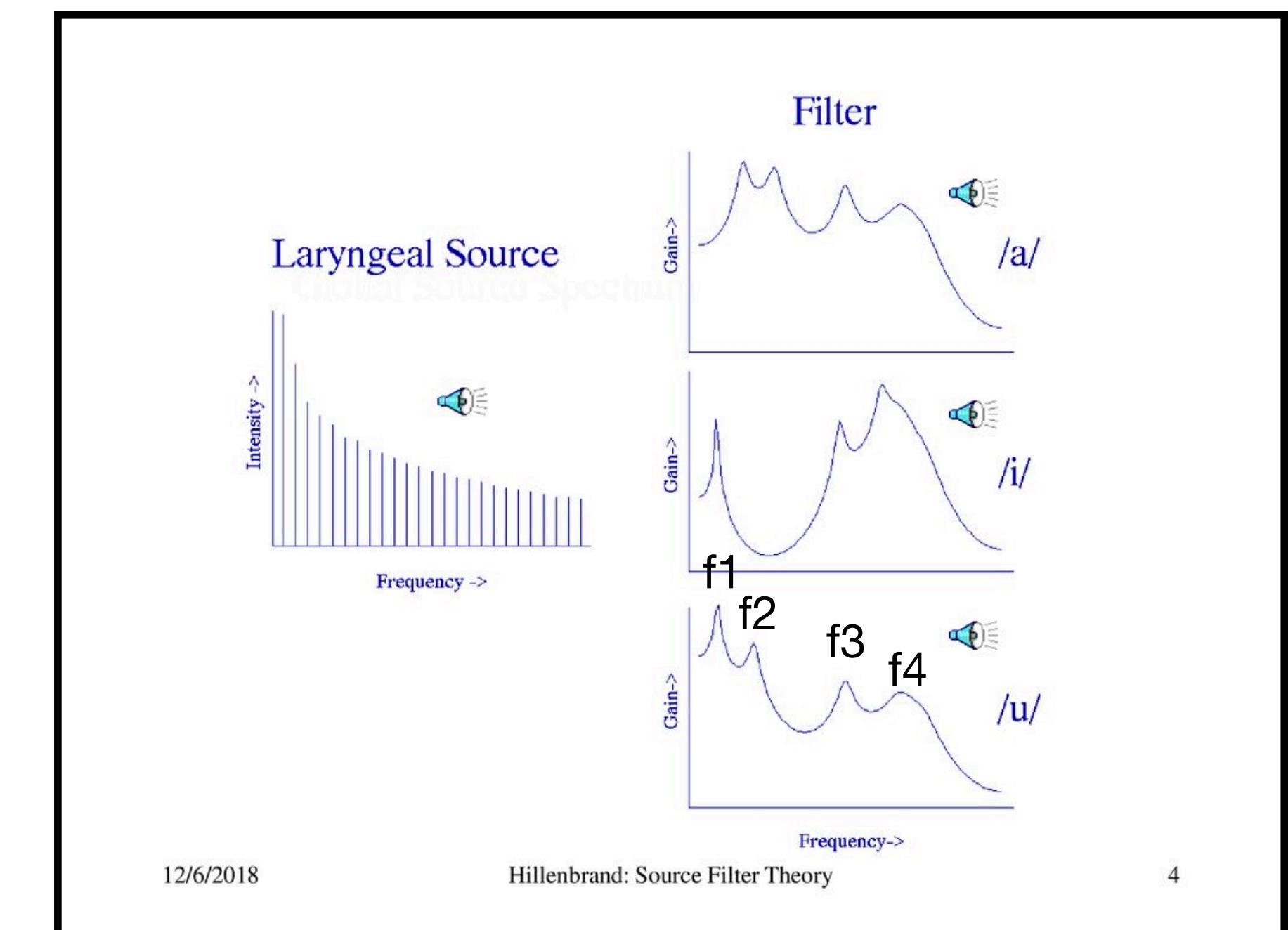
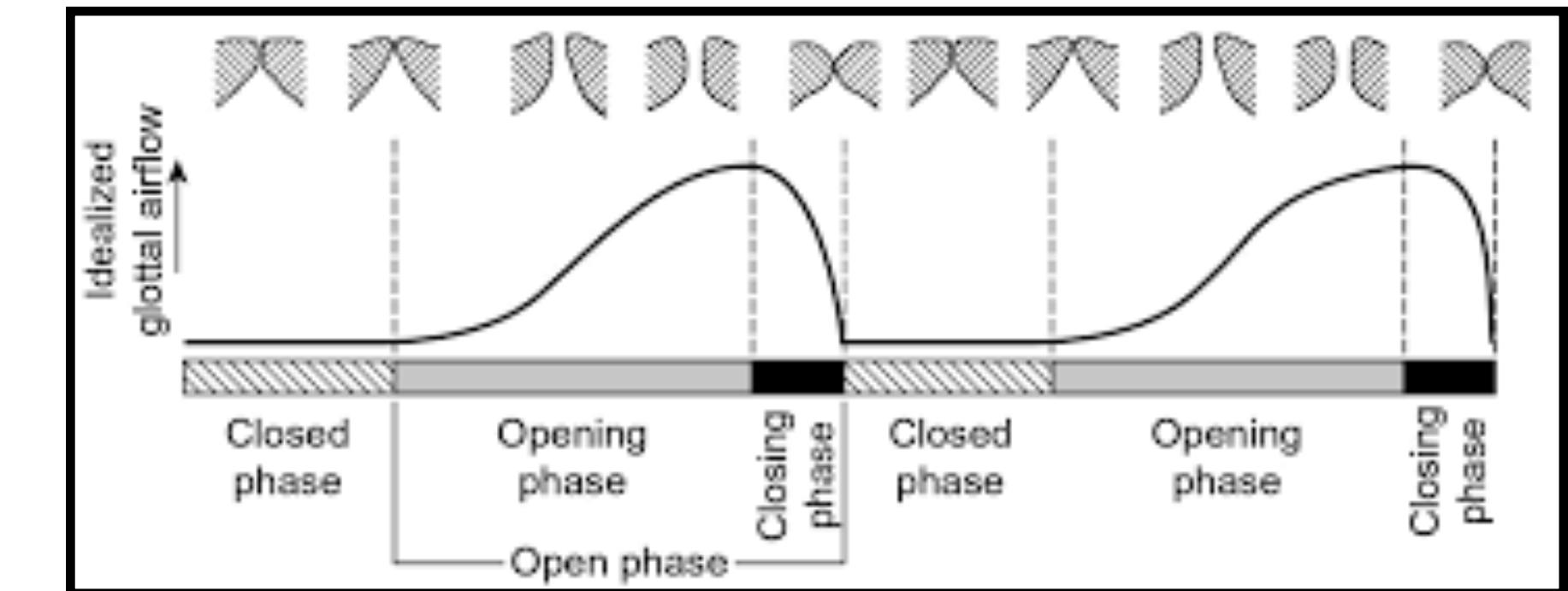
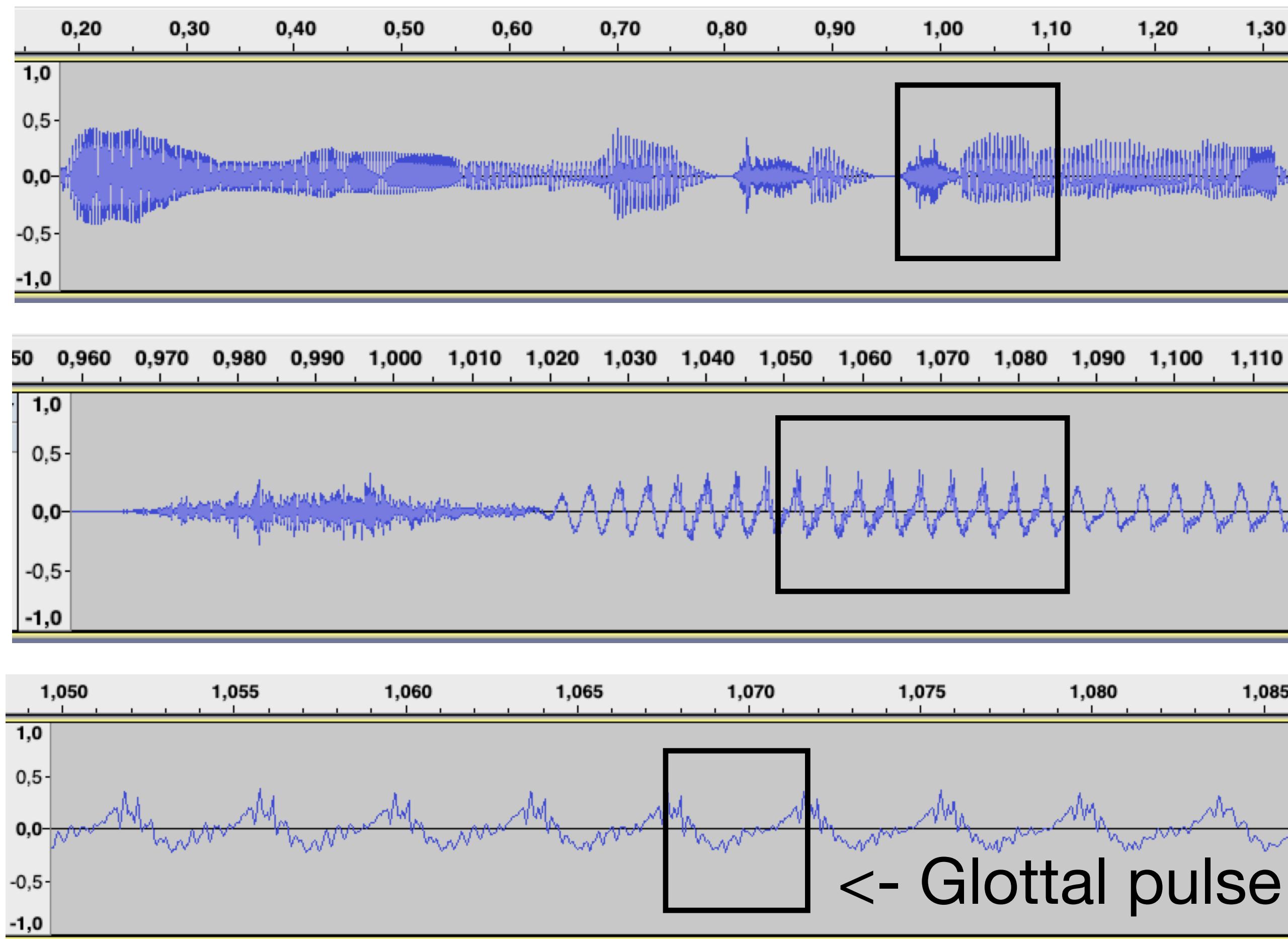


$$y(t) = h(t) * x(t),$$

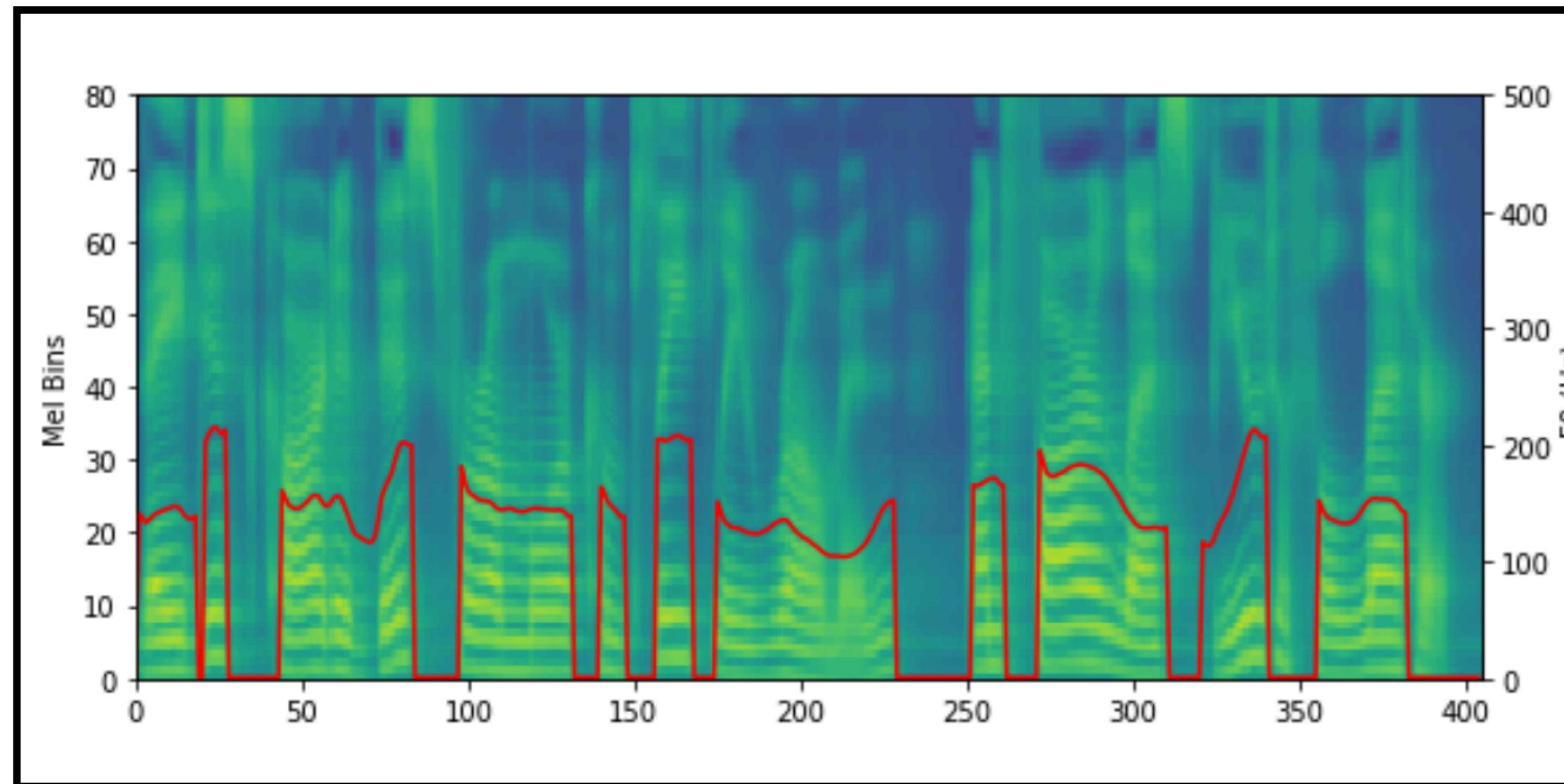
Source-filter theory



Что такое речь



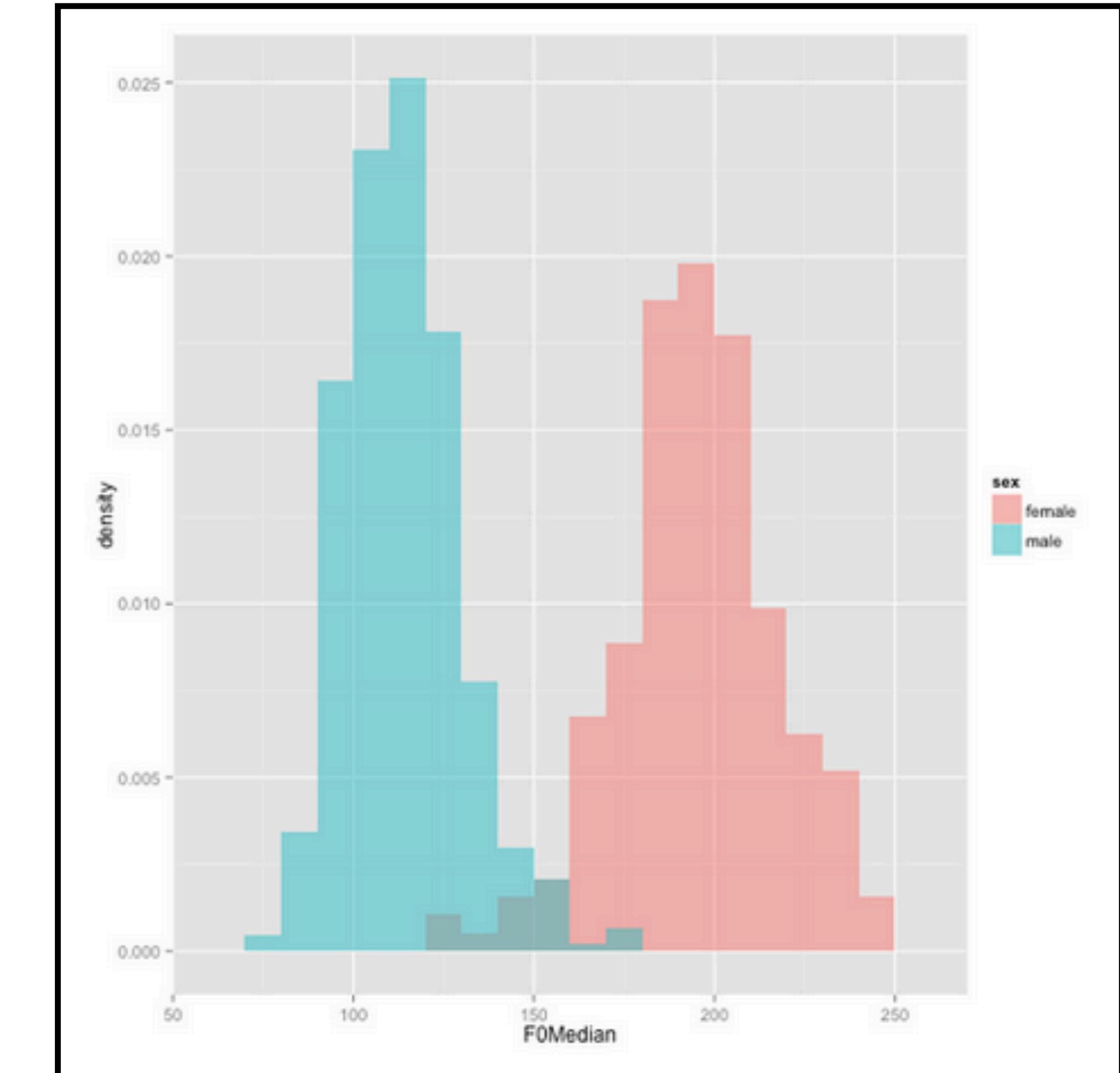
pitch (f0, частота основного тона)



voiced/unvoiced

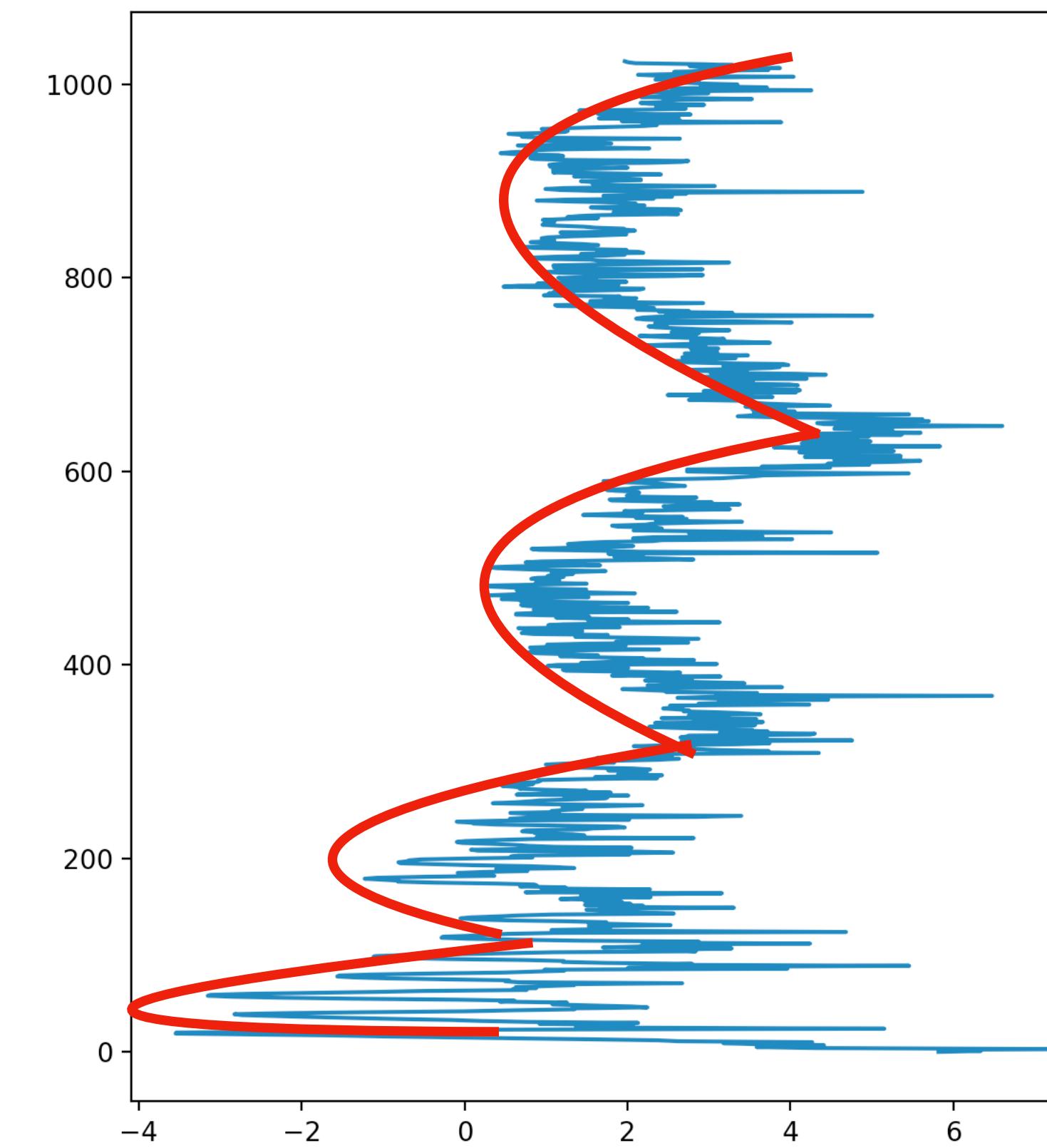
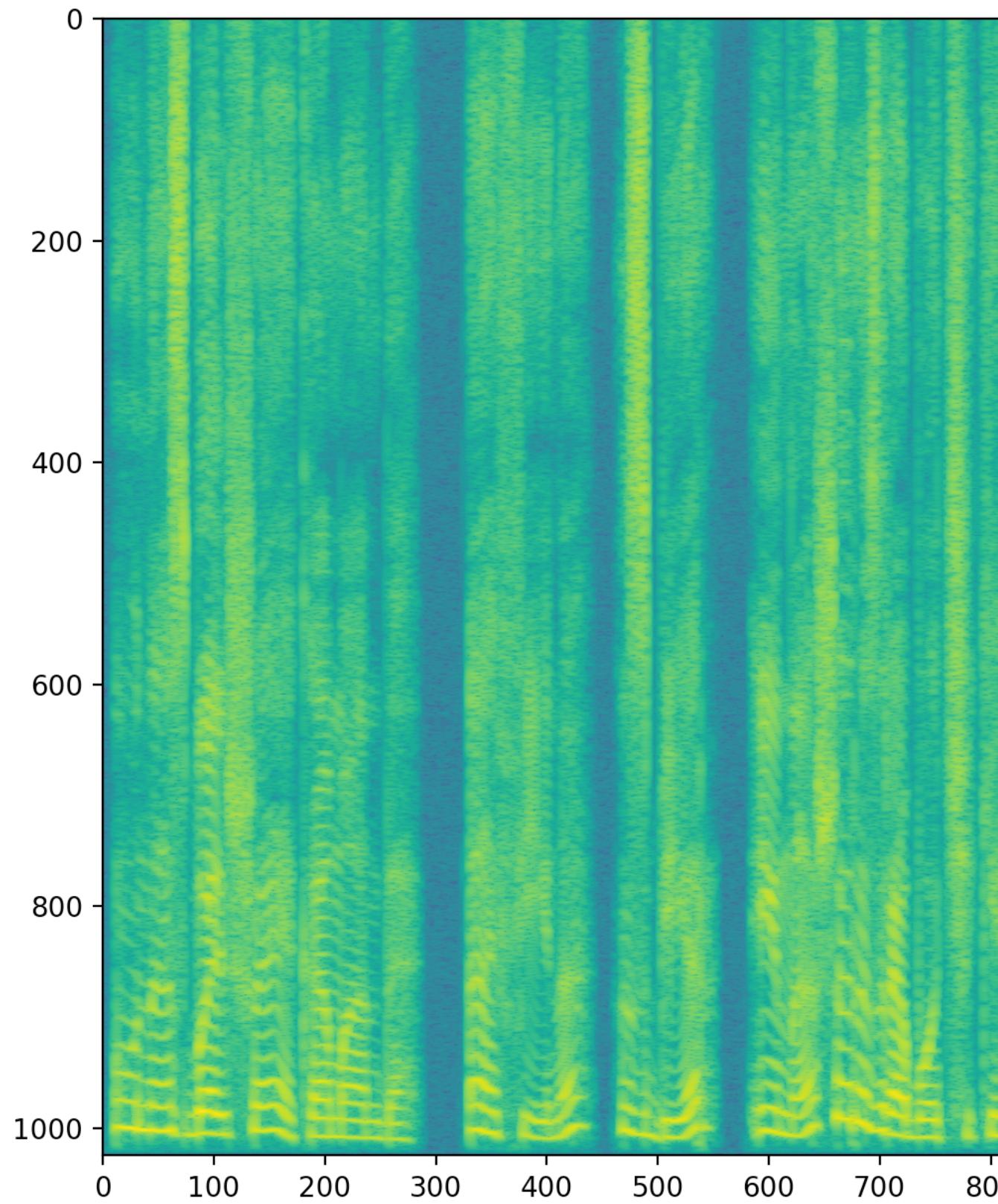
гласные
звуковые согласные

глухие согласные

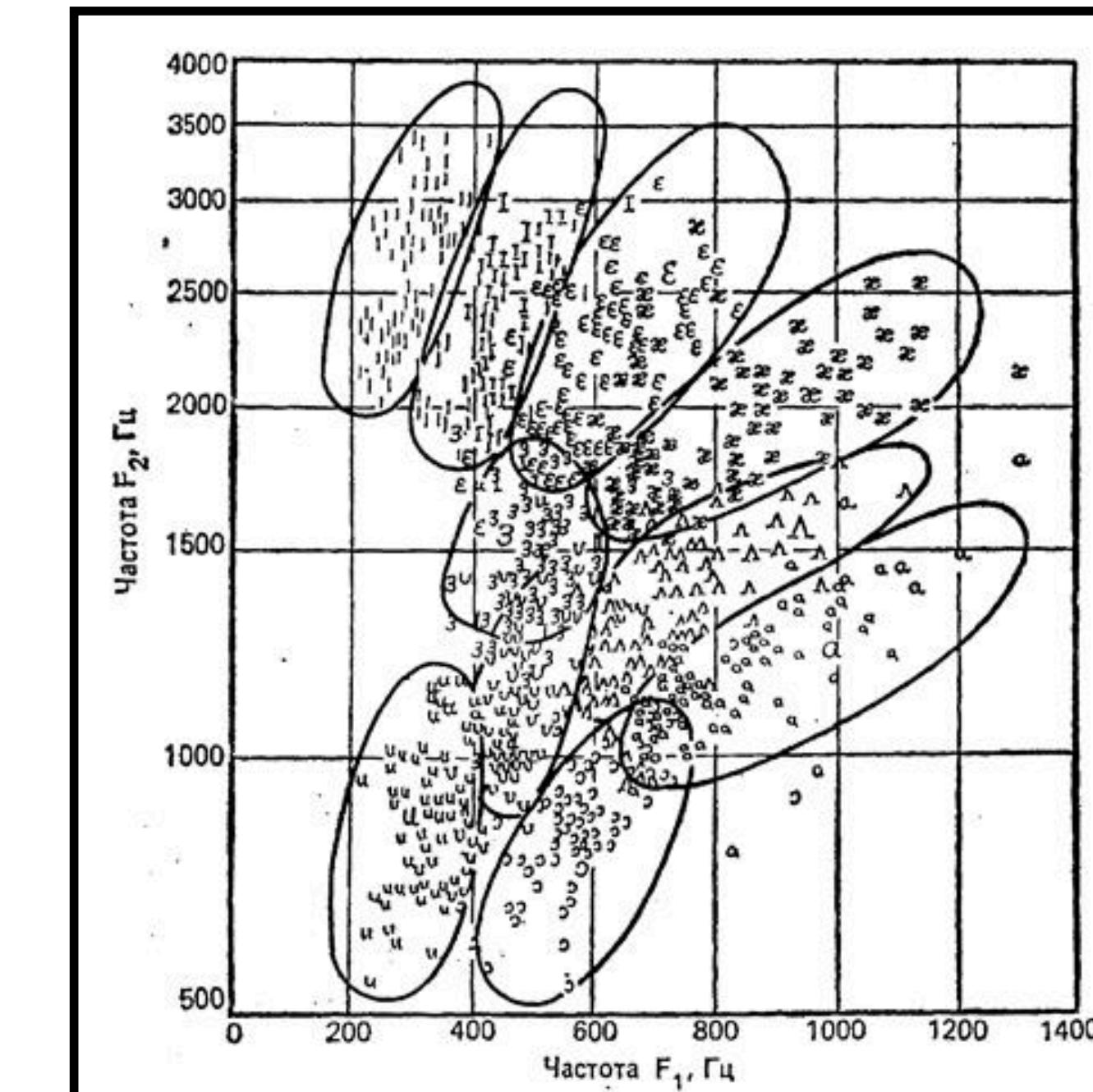
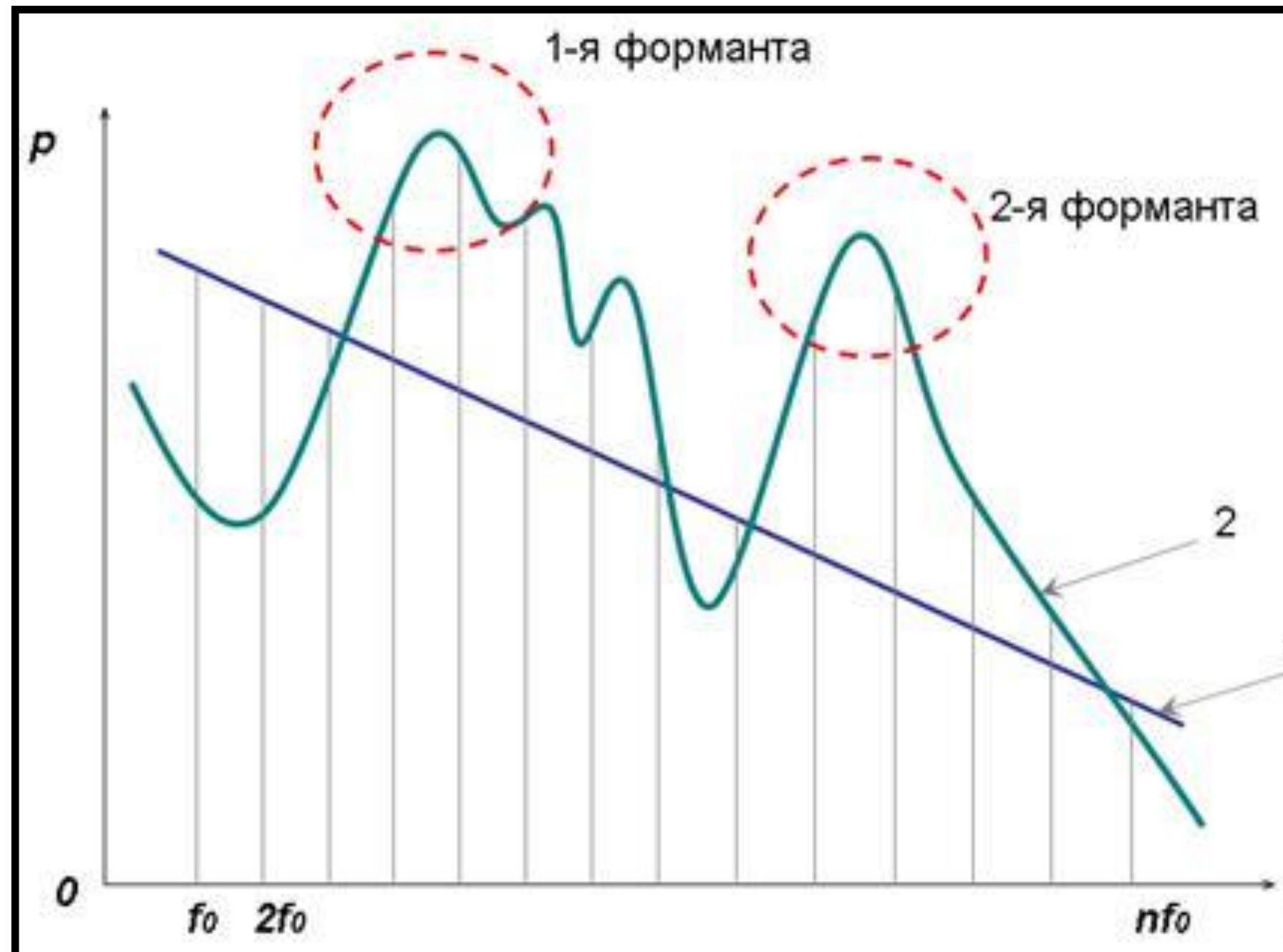


Yin algorithm, crepe, pyreaper

Форманты



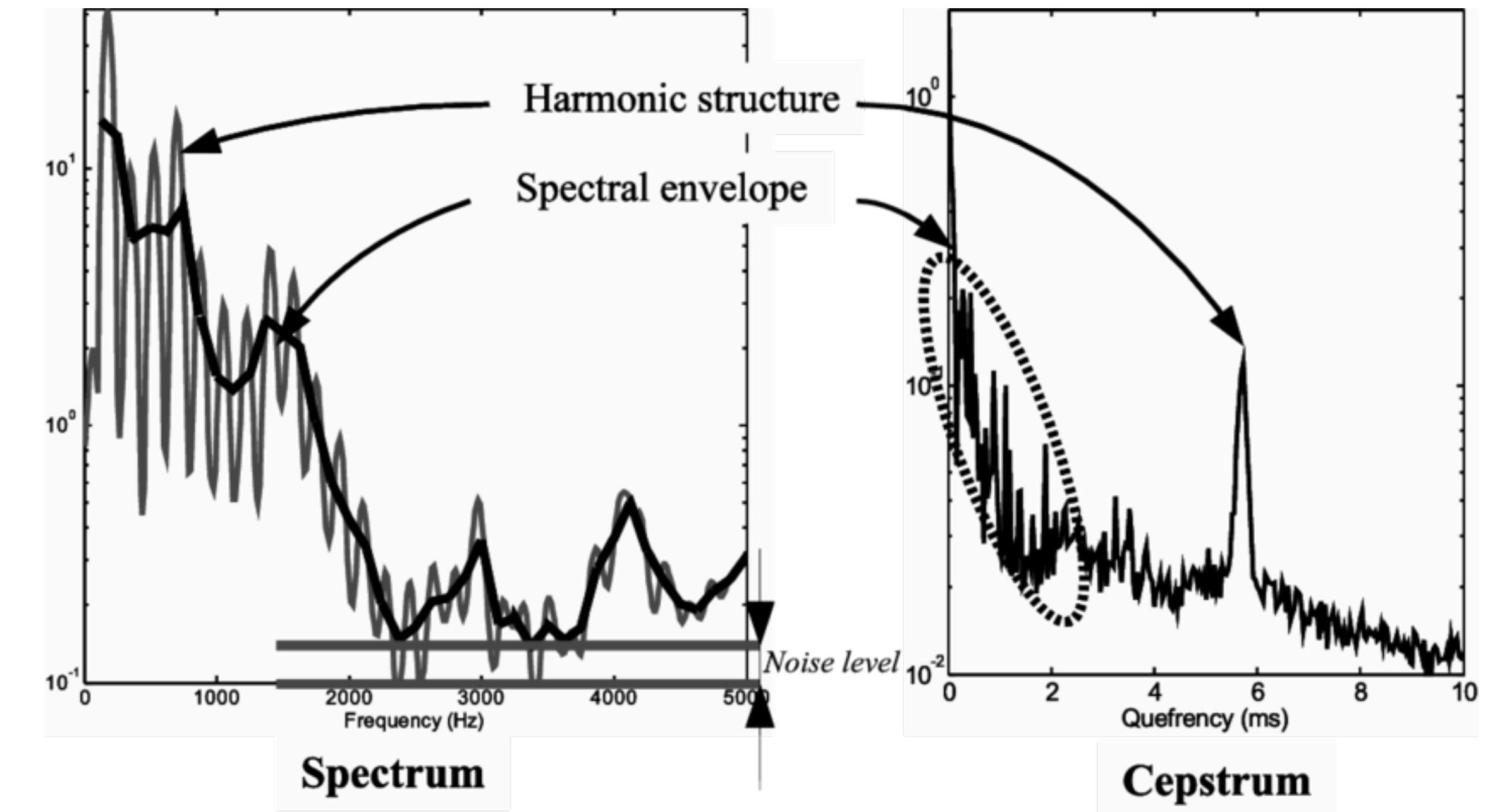
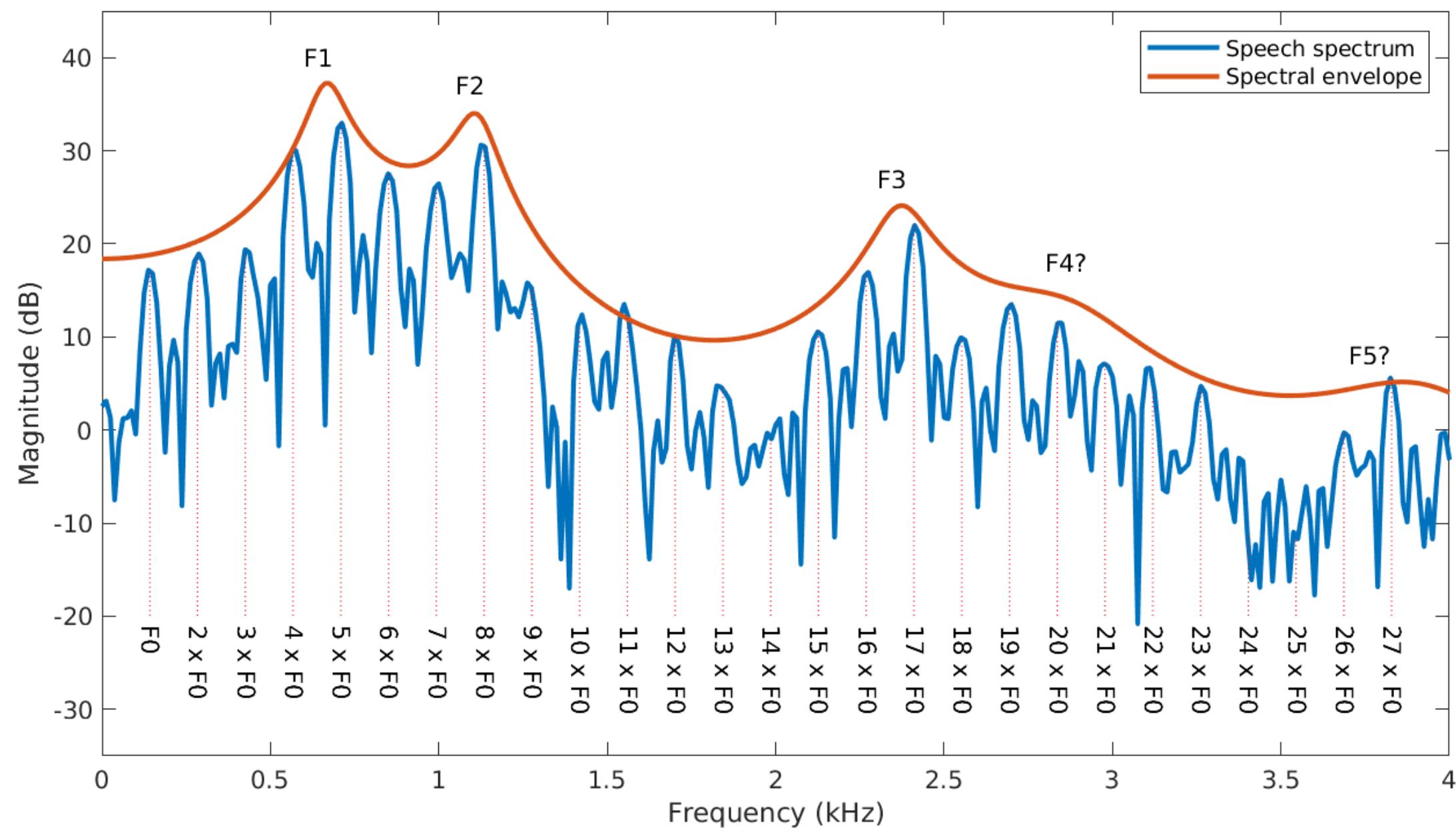
Форманты



Кепстр

- Спектр выглядит как периодический сигнал -> FFT

$$Db = 20 \lg (A / A_{ref})$$

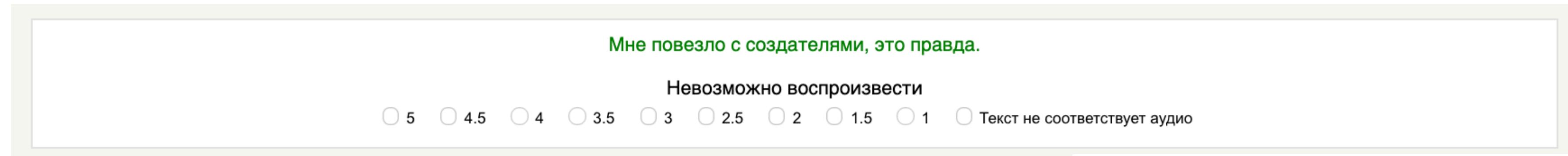


F0, F1, F2, F3, ...

- 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:



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

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

FIGURE 4

Example of a user interface used in the blind grading phase

MUSHRA:



- + 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:

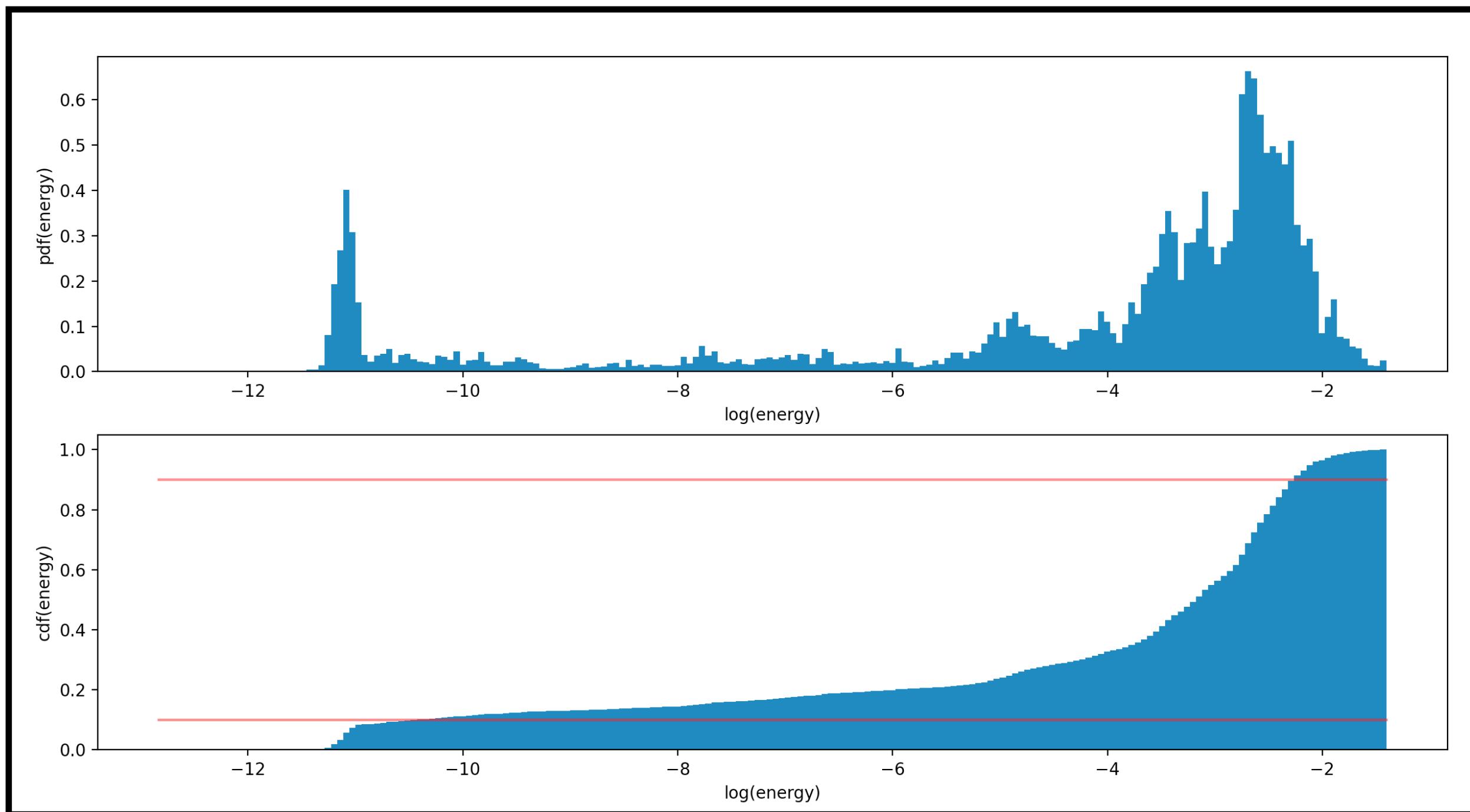
благодарю! мои разработчики будут рады!

▶ 0:00 / 0:03 ————— 🔍 ⋮

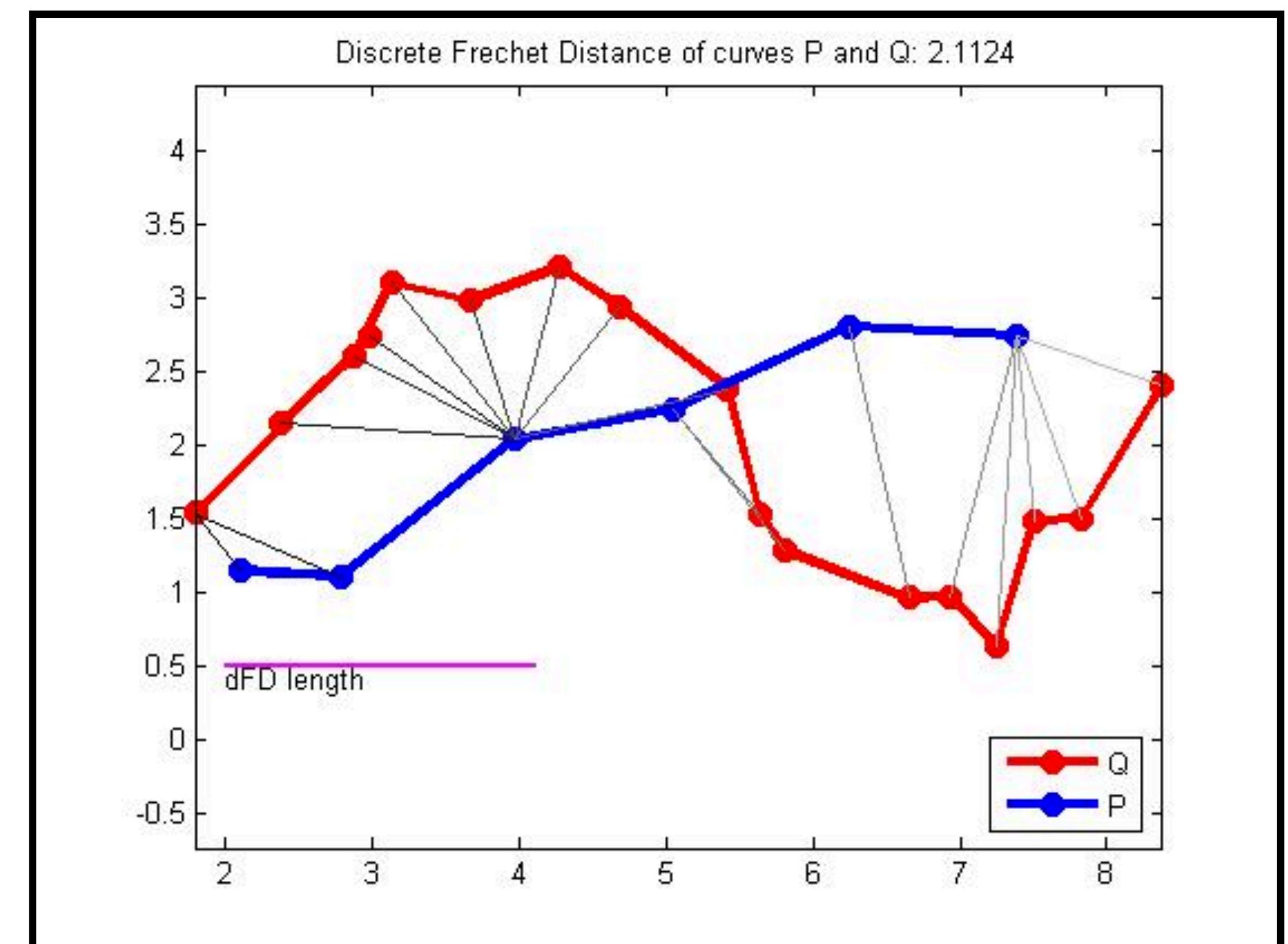
- Лишние паузы
- Не хватает пауз
- Неверное произношение
- Пропущены нужные / вставлены лишние звуки
- Некачественное аудио
- Неверная интонация
- Наличие заиканий
- Всё верно

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](#)