Синтез речи

Лекция №1

Гриша Стерлинг, SberDevices

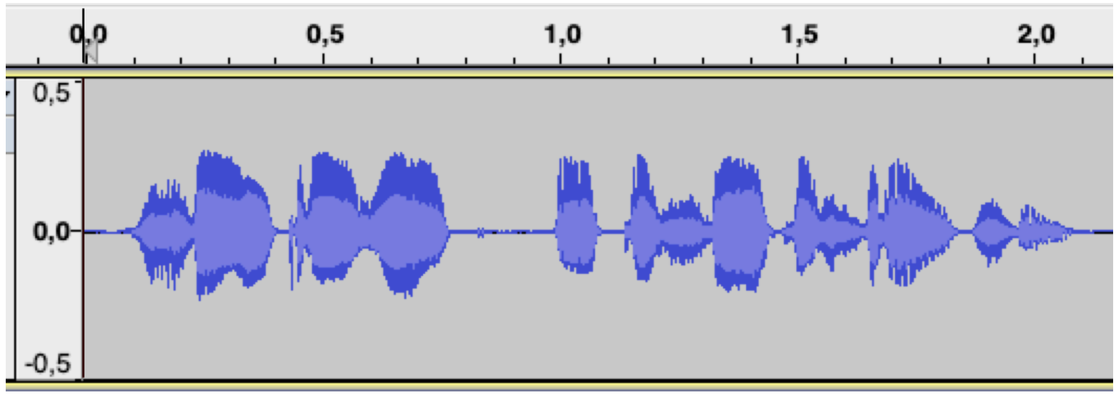
1. Введение
2. Метрики качества
3. Обзор задачи
4. Вокодеры (ч. 1)

Синтез речи

Задача:
- озвучить заданный текст голосом

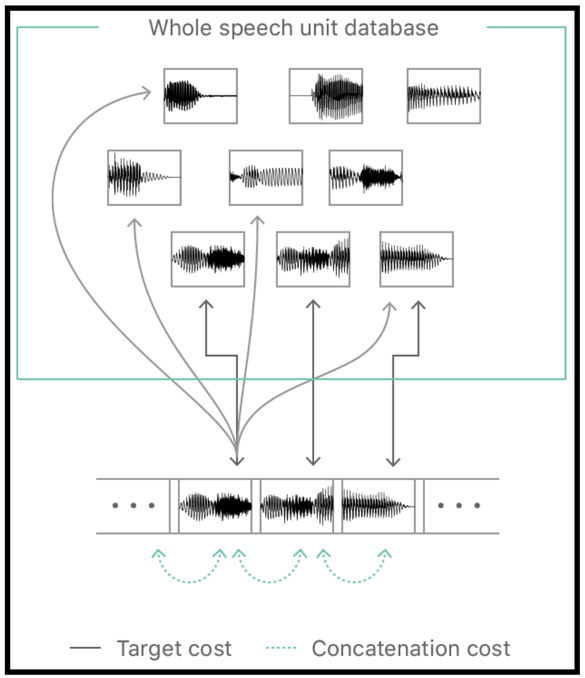
Задачи со звездочкой:
- style transfer
- несуществующим голосом
- controllable speech synthesis
- на другом языке
- эмоции
- шепот
- субвокализации, смех

«Всем привет, это синтез речи»
->



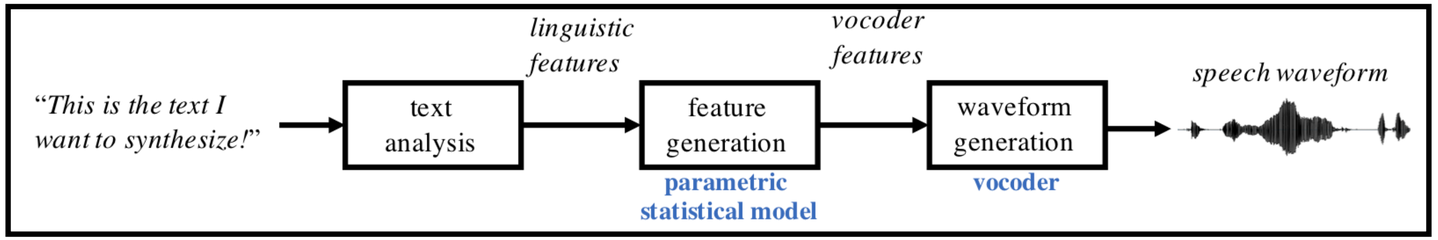
Синтез речи

Concatenative
(unit selection)



Parametric

2 стадии:
- акустическая модель
- вокодер (**vo**ice **coder**)



Метрики качества

Crowdsourced:

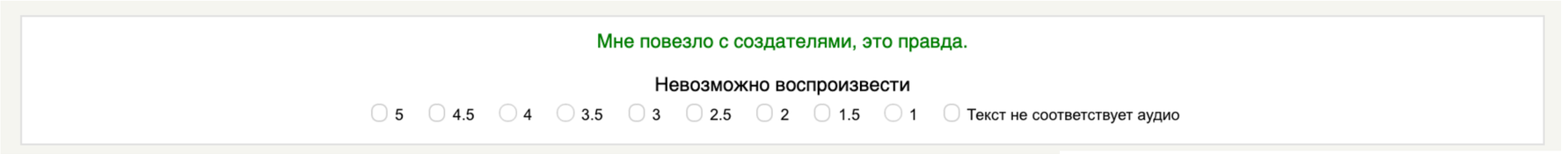
- MOS (mean opinion score)
- MUSHRA (Multiple Stimuli with Hidden Reference and Anchor)
- PSER (Pronunciation sentence error rate)
- SBS

Automatic:

- SNR (sound to noise ratio)
- Frechet distance
- nn-based mos
- MCD (mel-cepstral distortion)

Метрики качества

MOS:

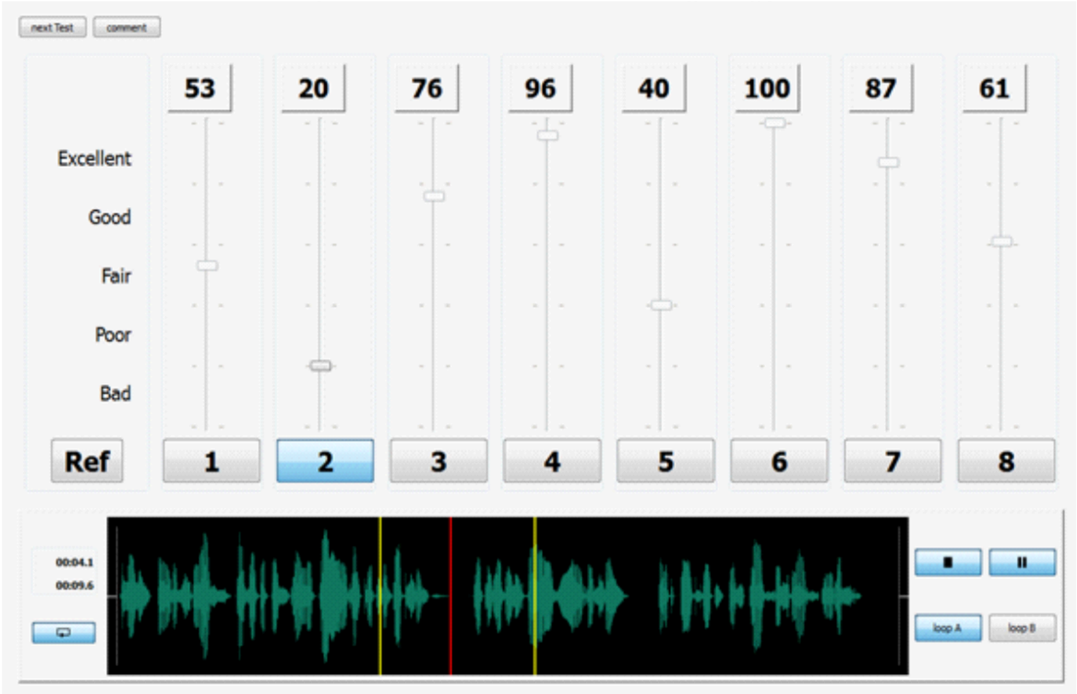


$$MOS = \frac{\sum_{n=1}^{N} R_n}{N}$$

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

MUSHRA:

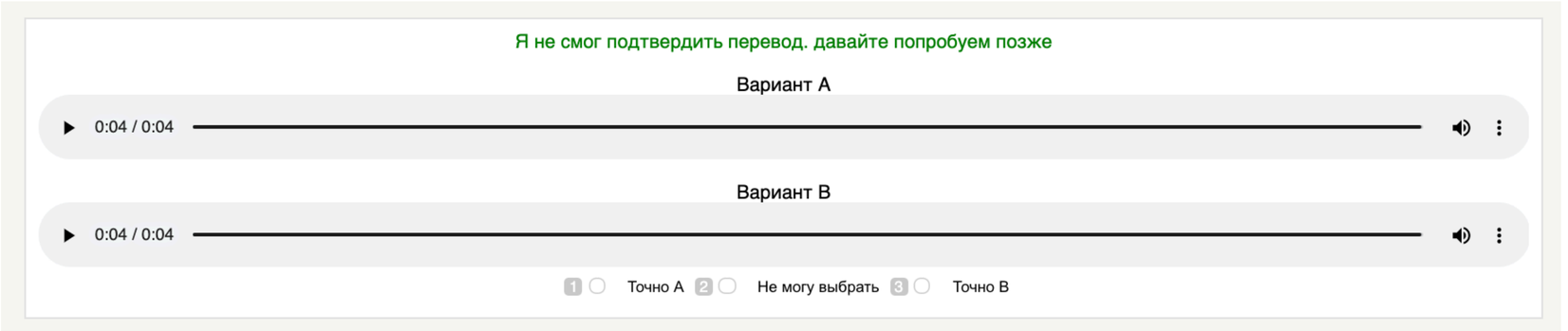
FIGURE 4
Example of a user interface used in the blind grading phase



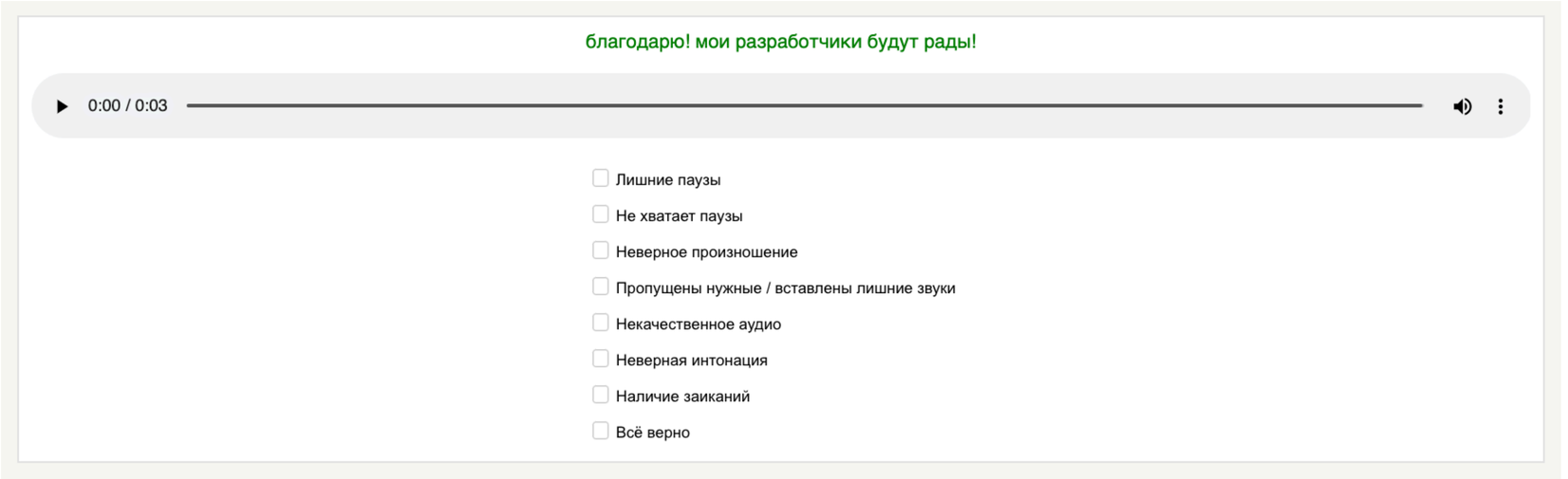
+ reference
+ hidden reference
+ few samples
+ 1-2 anchors

Метрики качества

SBS:

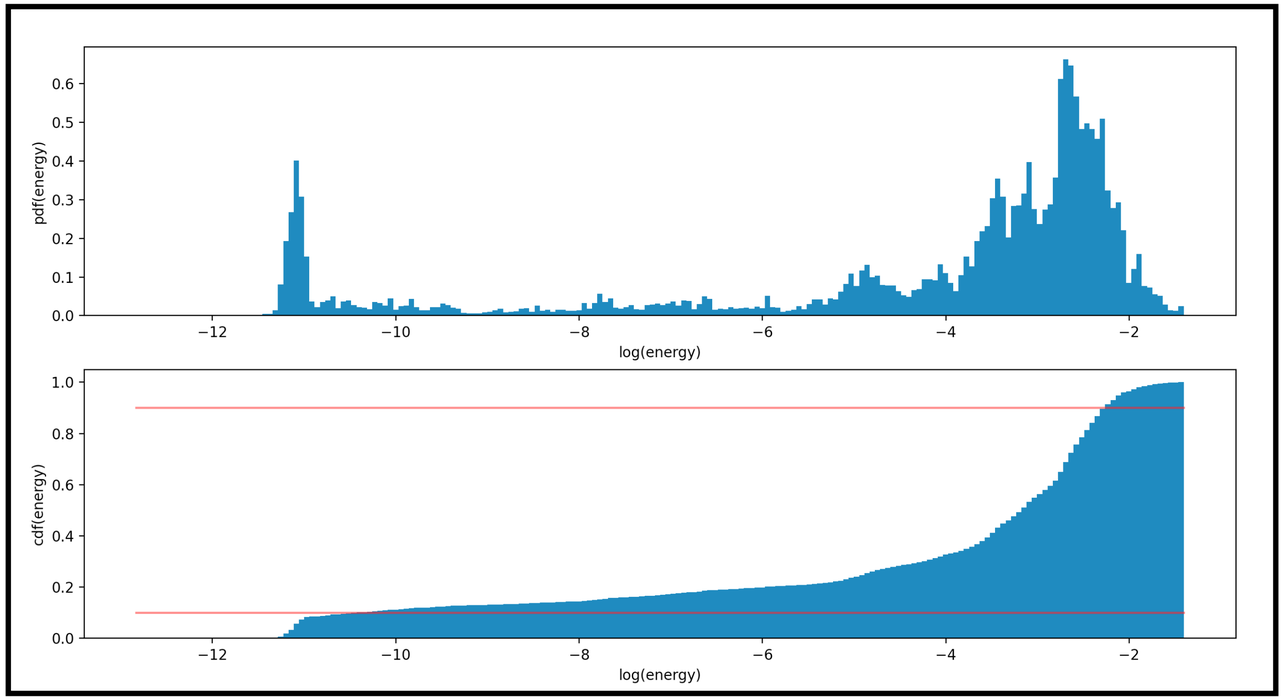


PSER:

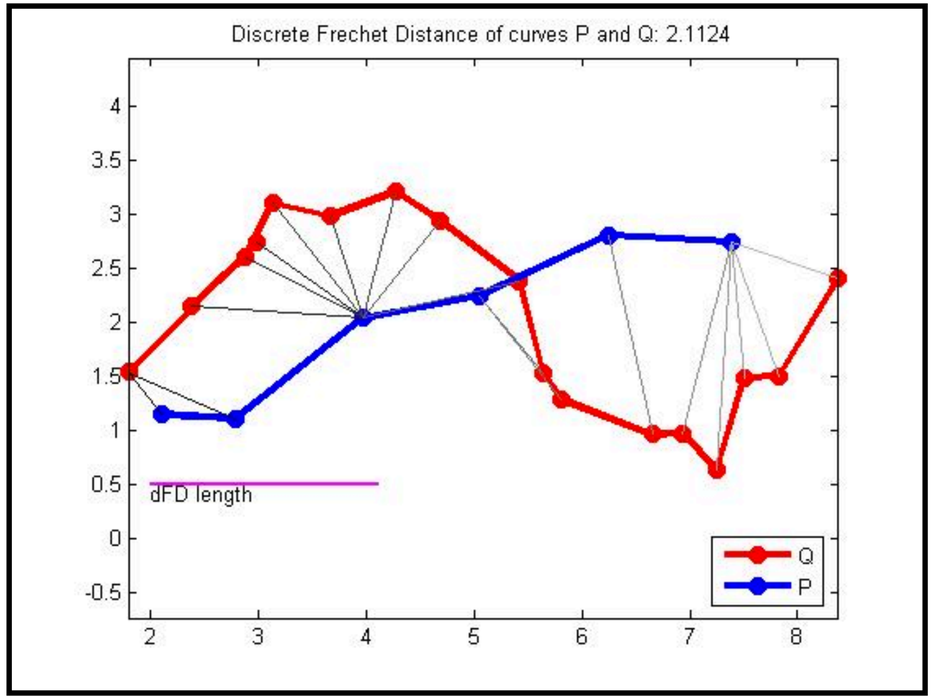


Метрики качества

SNR:



Frechet distance:



Метрики качества

MCD:

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

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

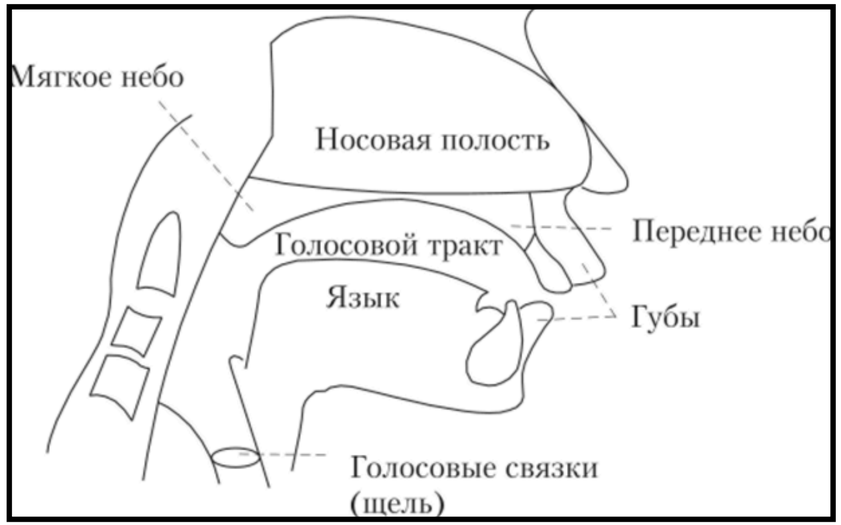
Yeunju Choi, Youngmoon Jung, Hoirin Kim

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

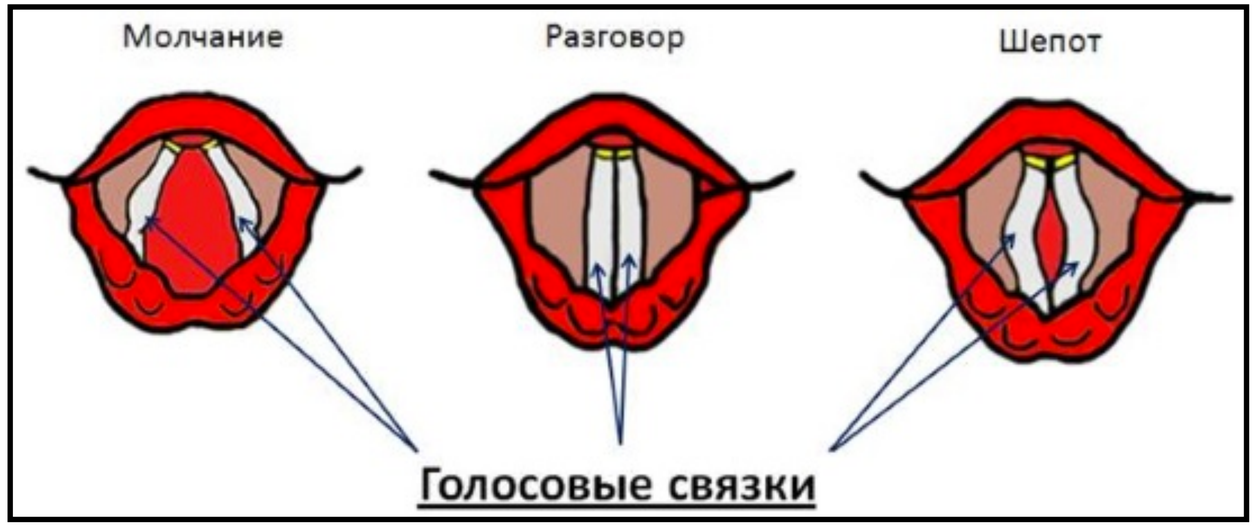
Что такое речь

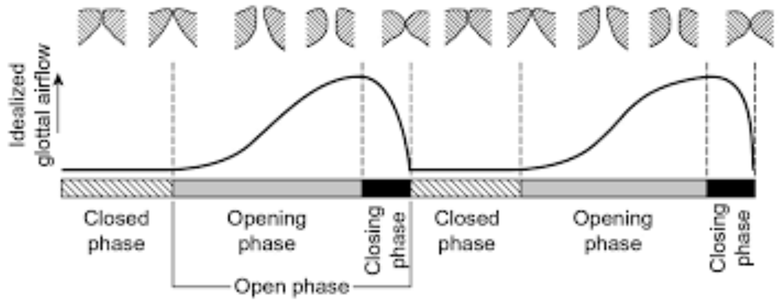
Речевой тракт человека



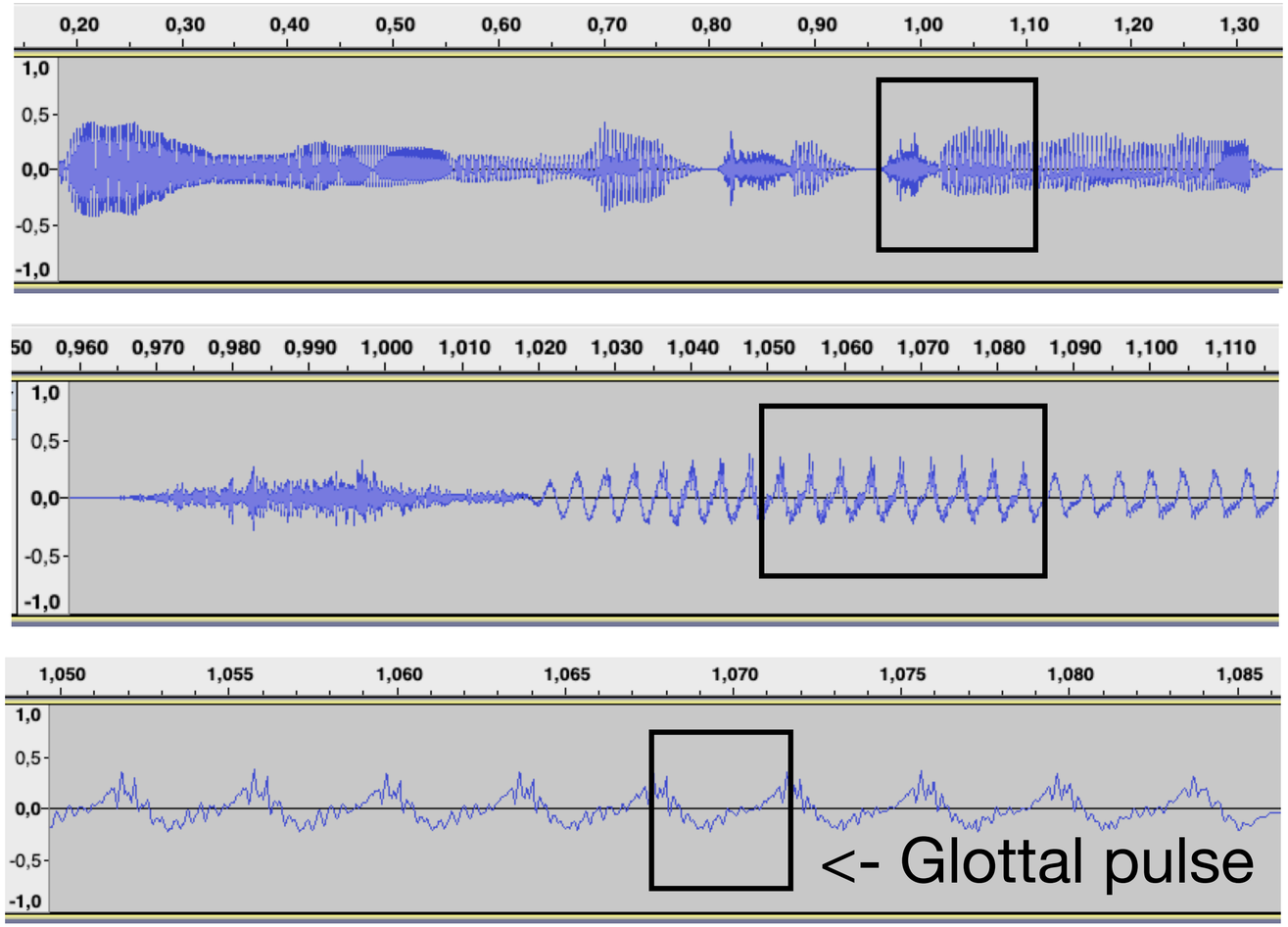
$$y(t) = h(t) * x(t),$$

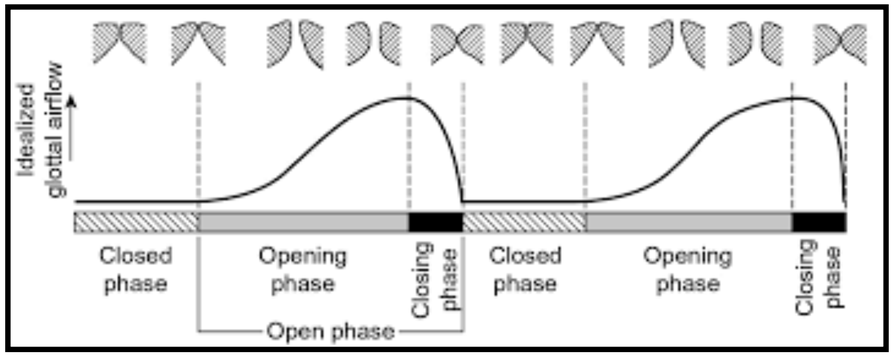
Голосовые связки

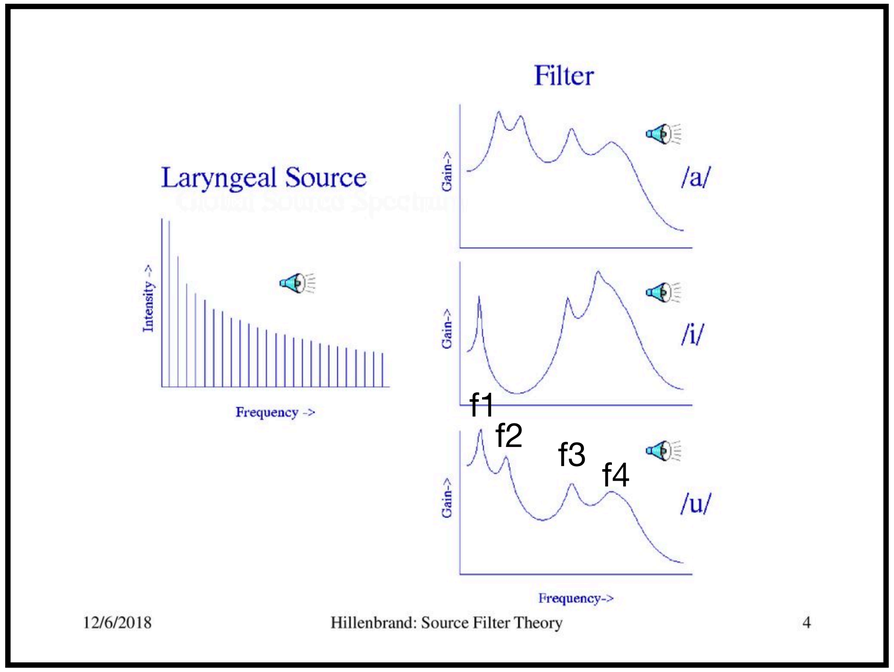




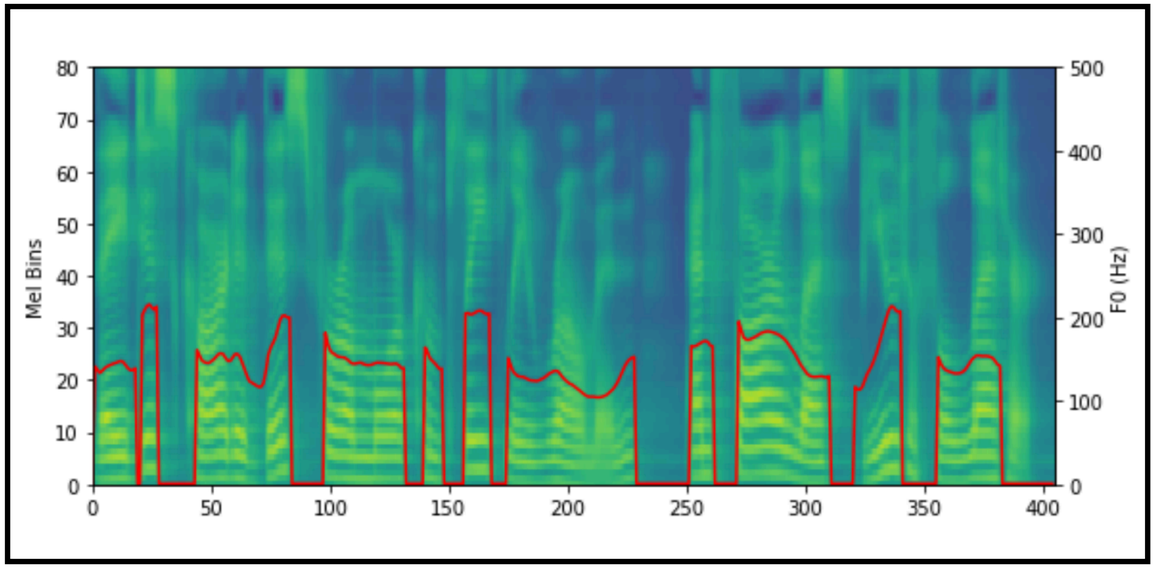
Что такое речь





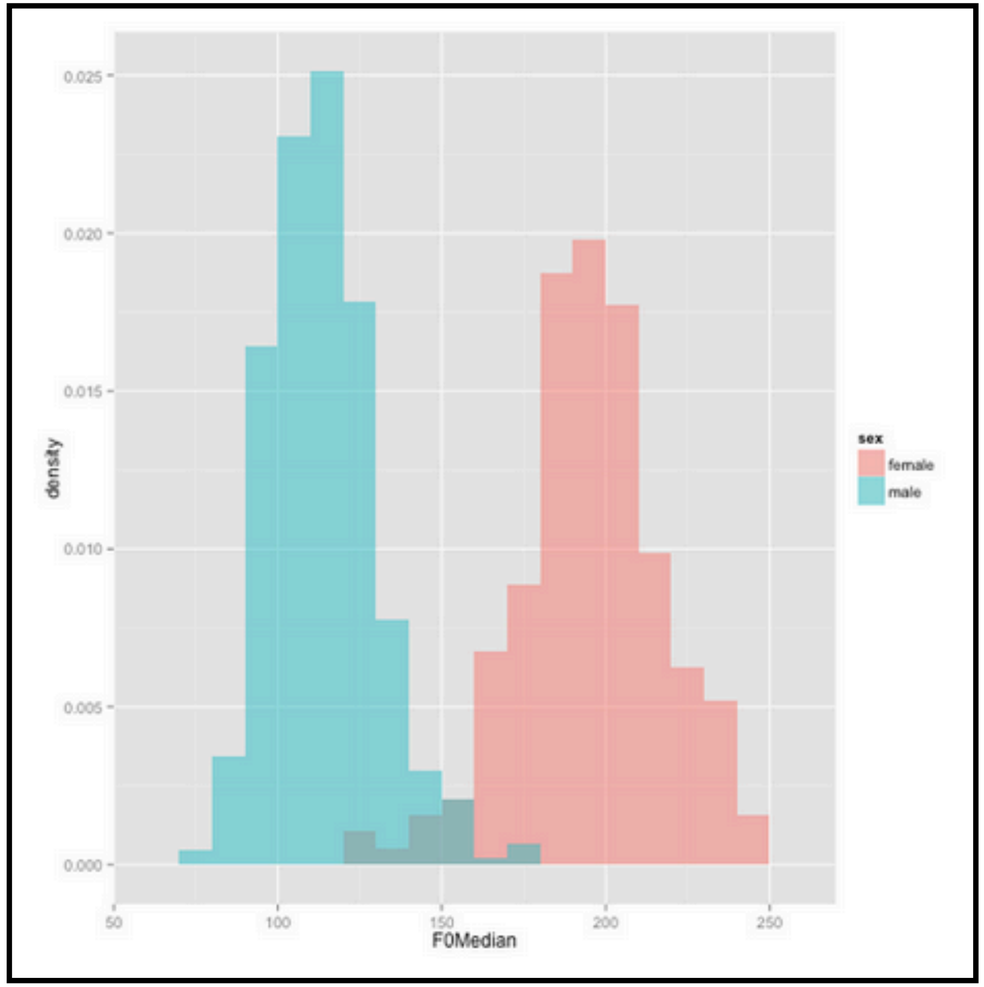


pitch (f0, частота основного тона)

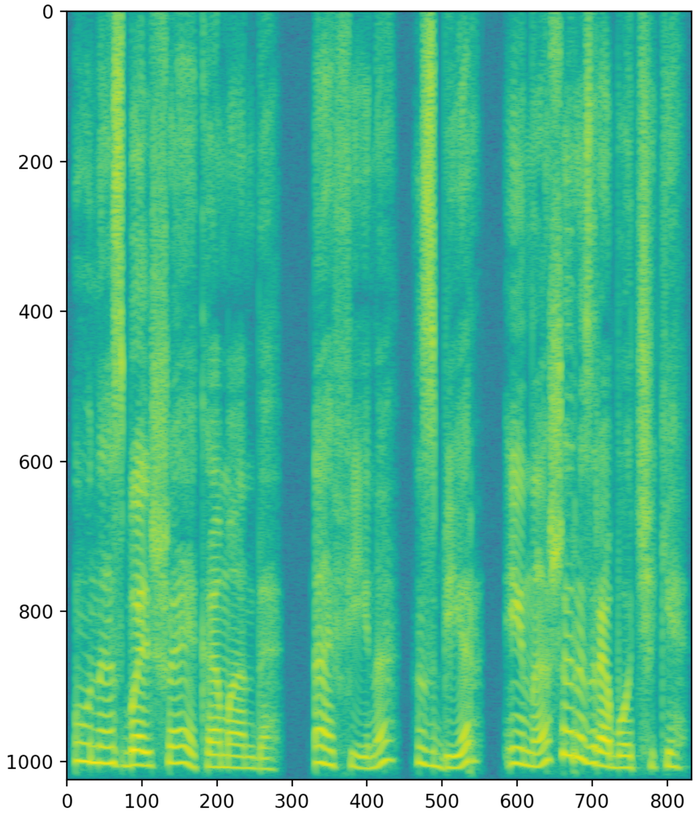


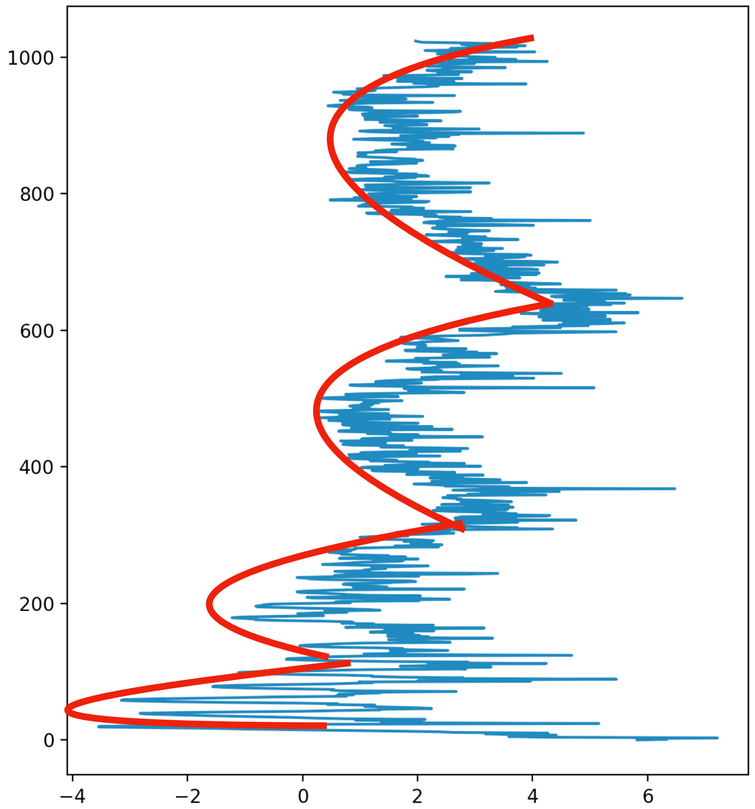
voiced/unvoiced

- гласные, звонкие согласные
- глухие согласные

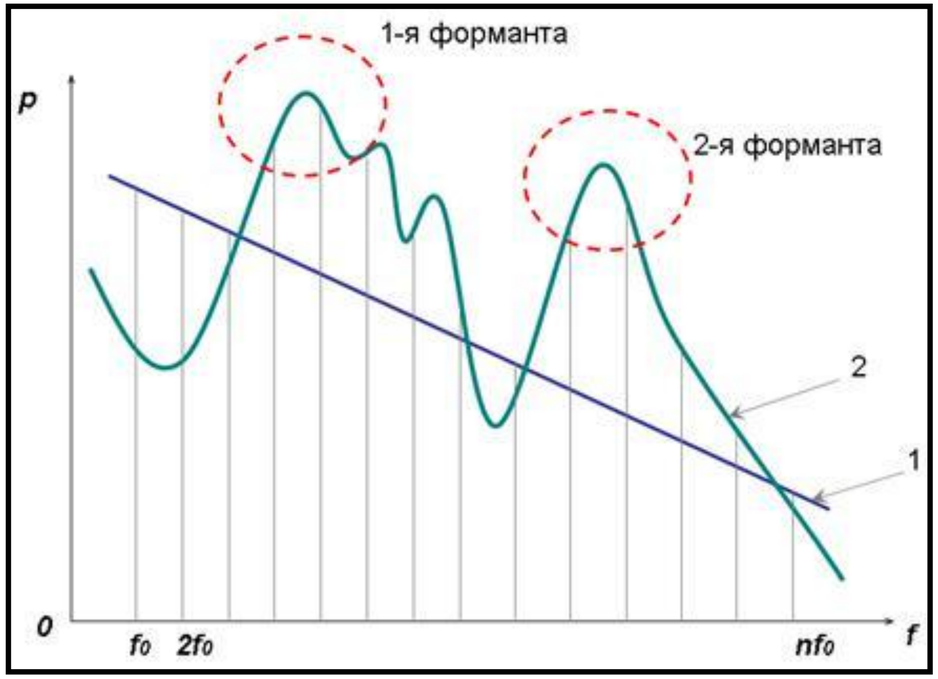


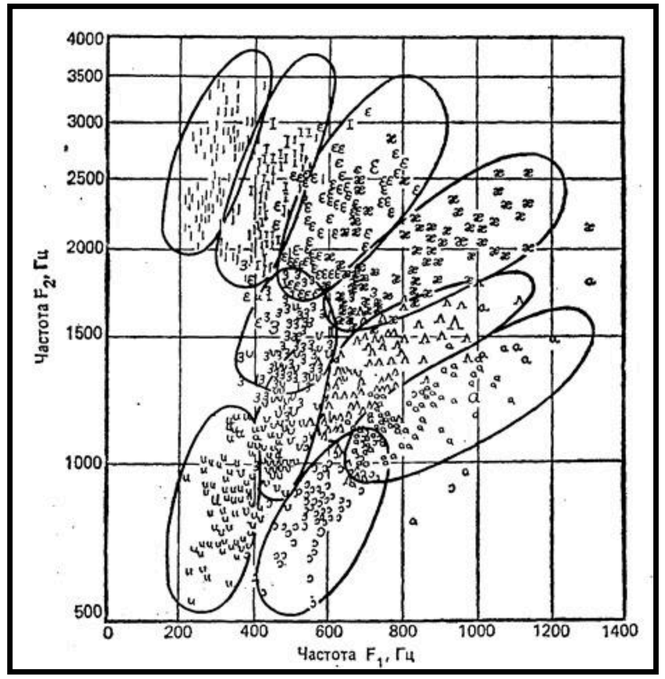
Форманты


0
200
400
600
800
1000
0
100
200
300
400
500
600
700
800


1000
800
600
400
200
0
−4
−2
0
2
4
6

Форманты


1-я форманта
2-я форманта
P
0
f
f0
2f0
nf0
2
1


4000
3500
3000
2500
2000
1500
1000
500
Частота F2, Гц
0
200
400
600
800
1000
1200
1400
Частота F1, Гц

Вокодеры

Не авторегрессионные:

- Griffin-Lim algorithm
- WORLD
- WaveGlow
- GAN-based

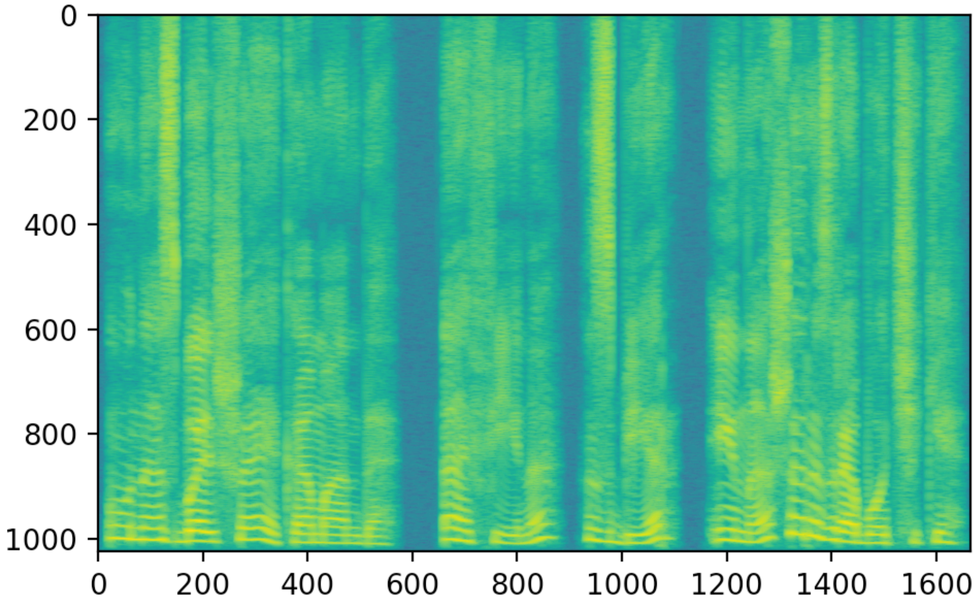
Авторегрессионные:

- WaveNet
- WaveRNN
- LPCNet

Griffin-Lim algorithm

Phase reconstruction:

a - искомый сигнал
X = |FFT(a)| - спектрограмма

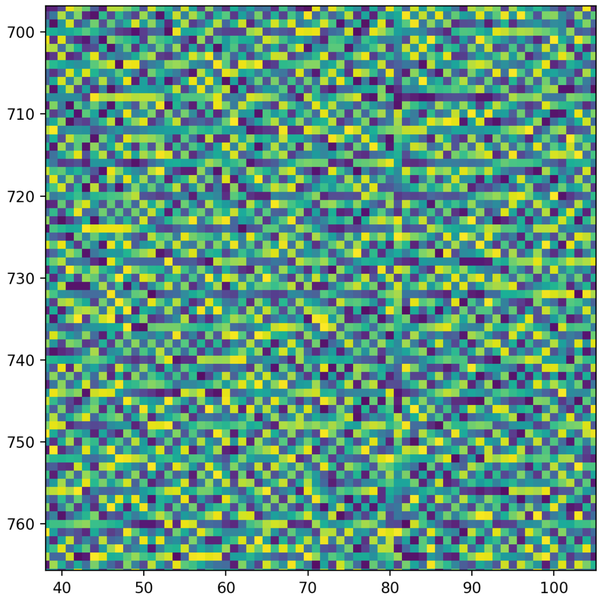


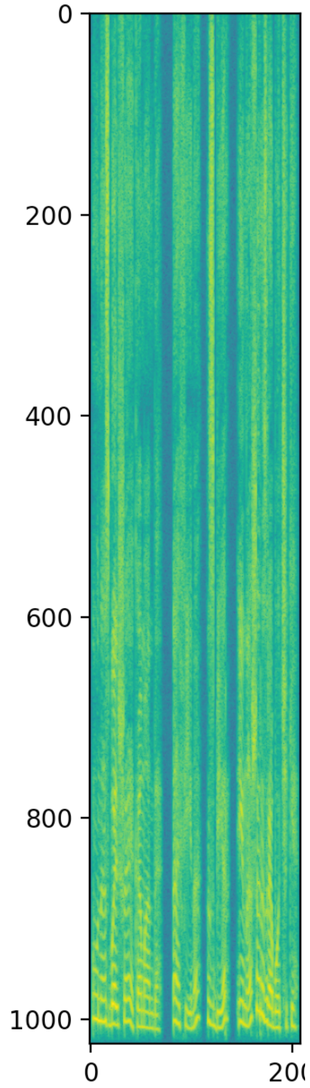
Griffin-Lim algorithm

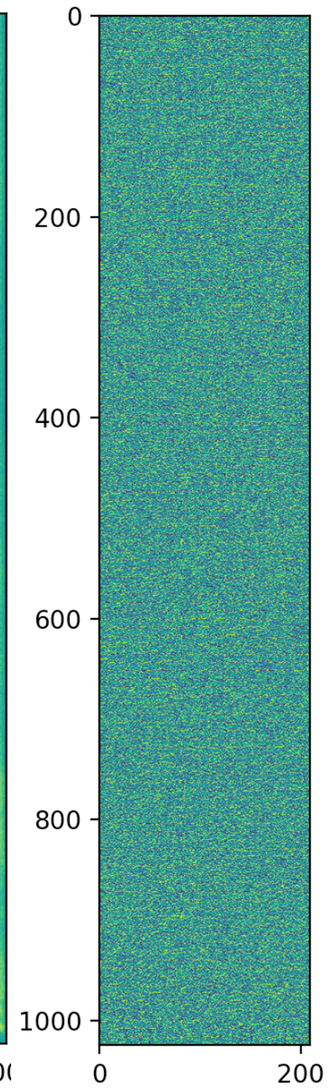
Алгоритм:

1. Инициализируем фазу Y_0 случайно
2. a_1 = iFFT(X e^ iY_0)
3. Y_1 = phase (FFT(a_1))
4. Итерируемся до сходимости
   (a_j = iFFT(X e^ iY_j-1))

$$\min_{\mathbf{X}} ||\mathbf{X} - P_{\mathcal{C}}(\mathbf{X})||_{\text{Fro}}^2 \text{ s.t. } \mathbf{X} \in \mathcal{A}$$







LPC (linear predictive coding)

Вейвформа - это свертка импульса с фильтром:

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

2. $x_t = \sum_{p=1}^{P} a_p x_{t-p} + \epsilon_t$

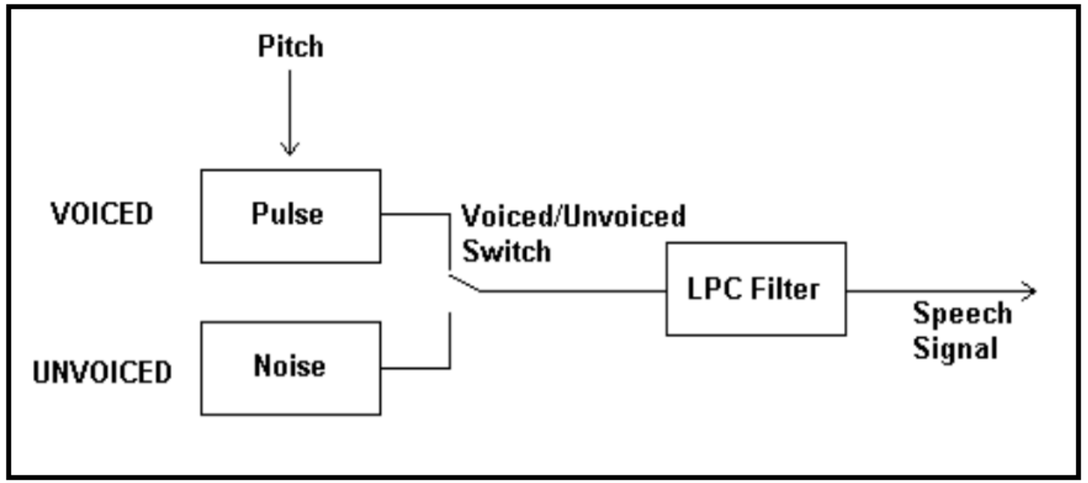
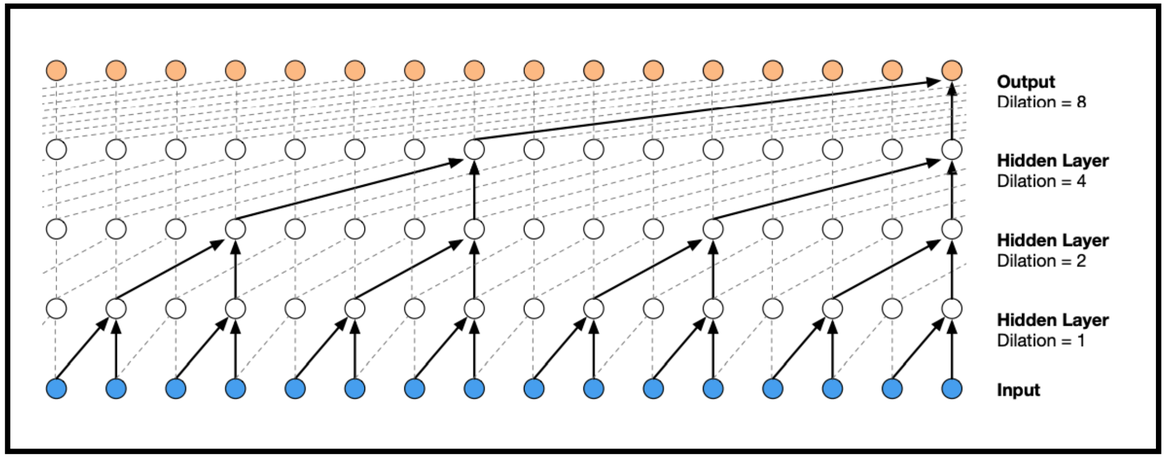


Figure 11 : LPC Decoder

Wavenet

2. $$x_t = \sum_{p=1}^{P} a_p x_{t-p} + \epsilon_t$$

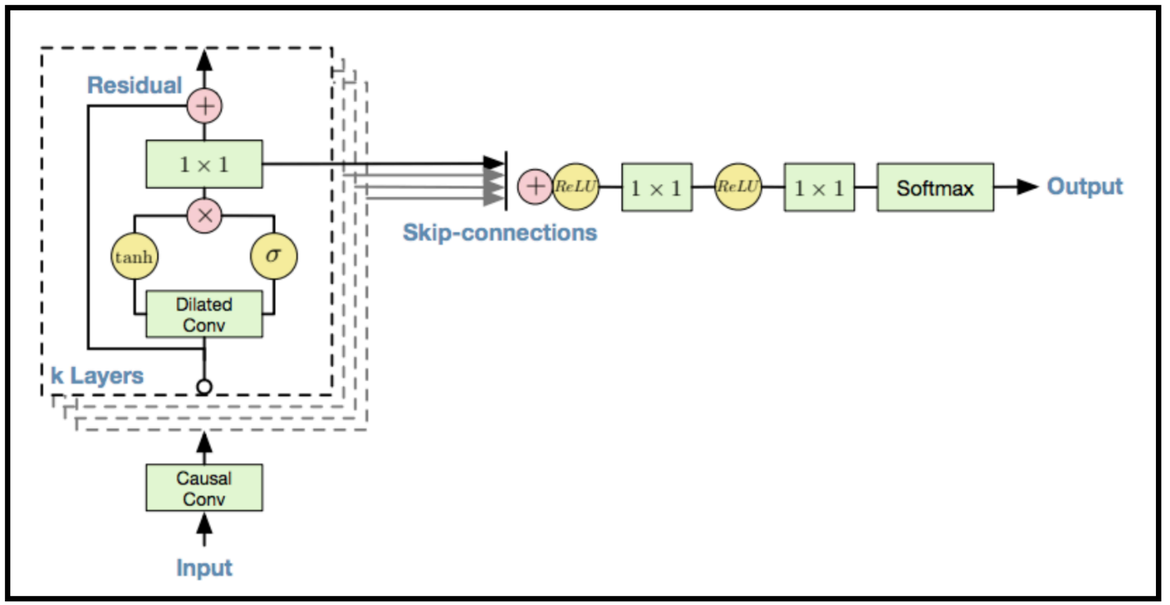
P = ?
eps = ?



3. $$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h}).$$

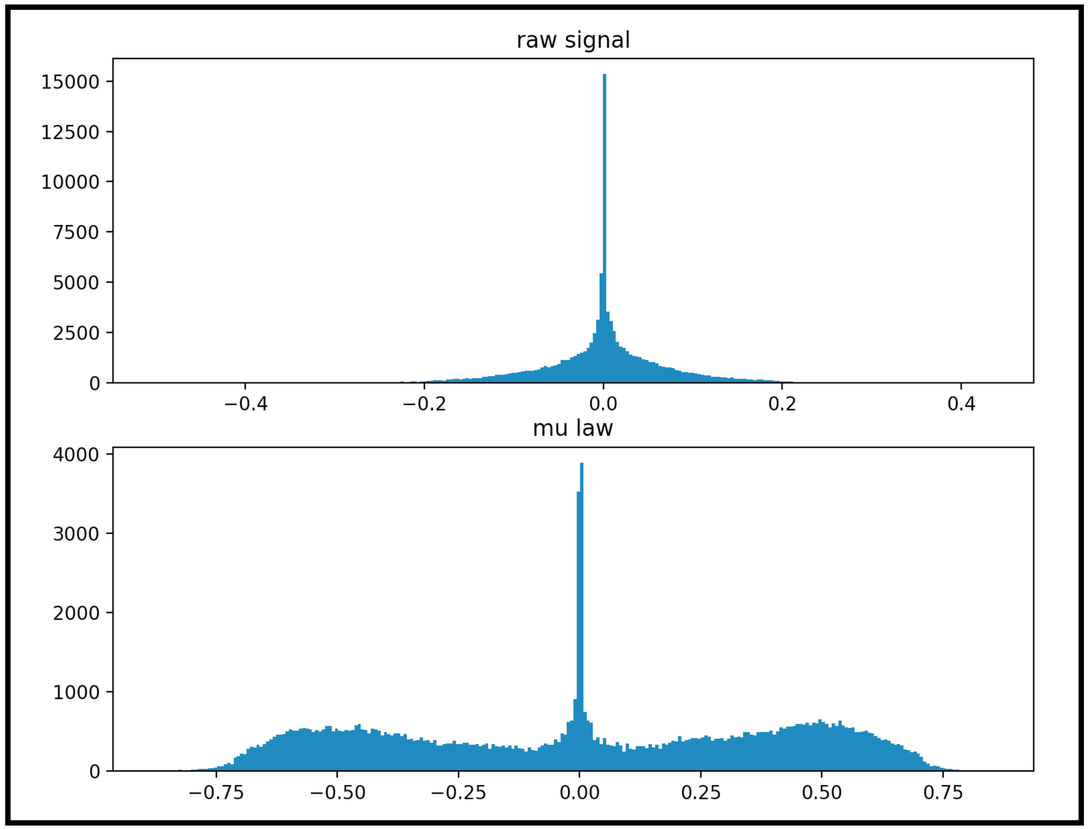
4. $$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k}^T \mathbf{h}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k}^T \mathbf{h}\right).$$

h - global conditioning:
speaker + phonemes



Wavenet

$$f(x_t) = \text{sign}(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)},$$



Inference: