

# **Определение произношение корректности речи.**

**Смаджаров Берик**

# План

1. Аудиофайлы, как мы привыкли
2. Предварительная обработка речевого сигнала и извлечение признаков
3. Акустическая модель
4. Языковая модель
5. Глубокие нейронные сети
6. Оценка корректности

# Аналоговый сигнал

Каждый из представляющих  
параметров  
описывается непрерывным  
множеством значений.



# Цифровой сигнал

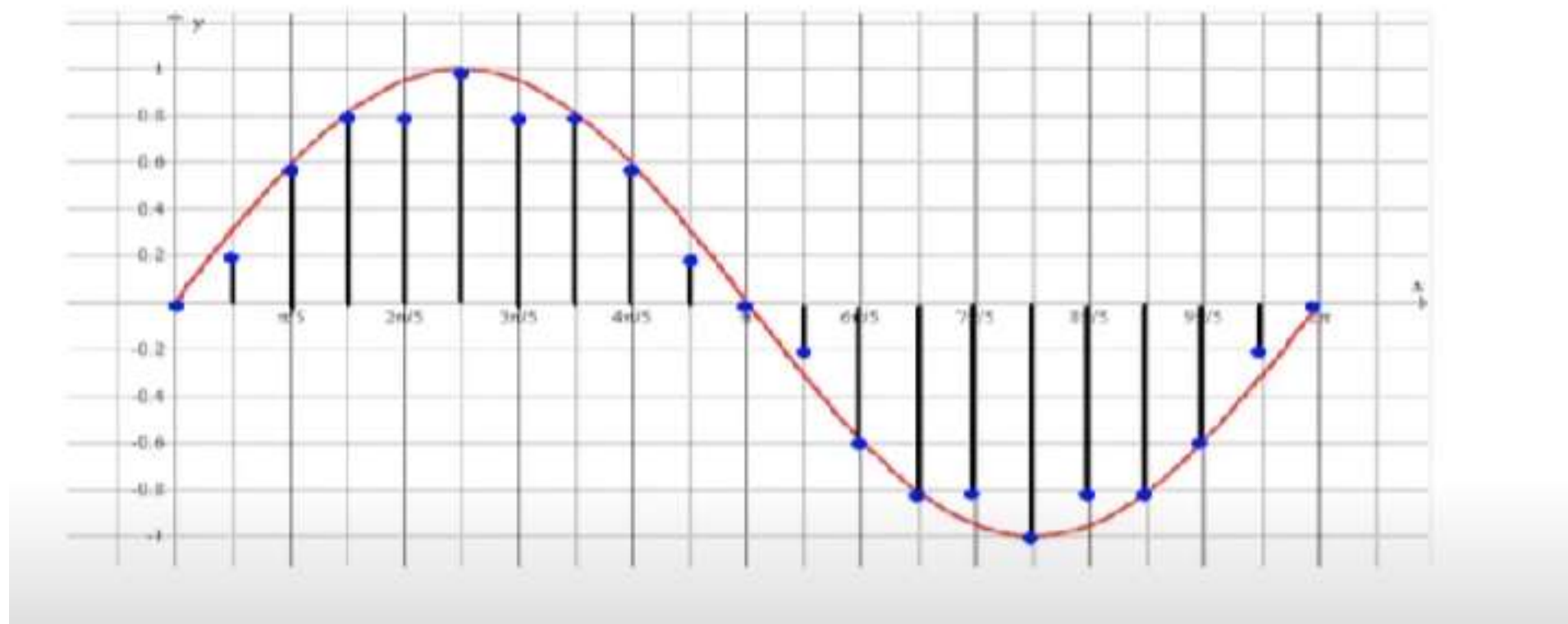
Можно представить в виде  
последовательности дискретных  
значений

- компактнее
- точнее



# Как хранятся аудиофайл ?

Sample rate - число отсчетов в секунду. Типичные значение: 1600, 220500



# Основная часть

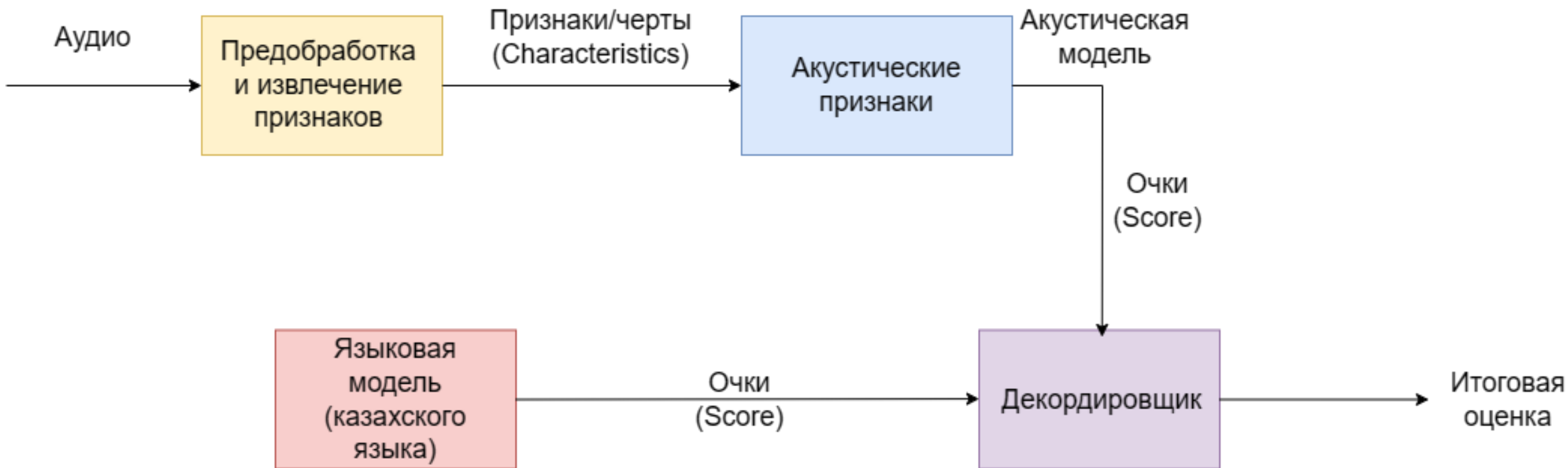


Рис 1.

# Основная часть

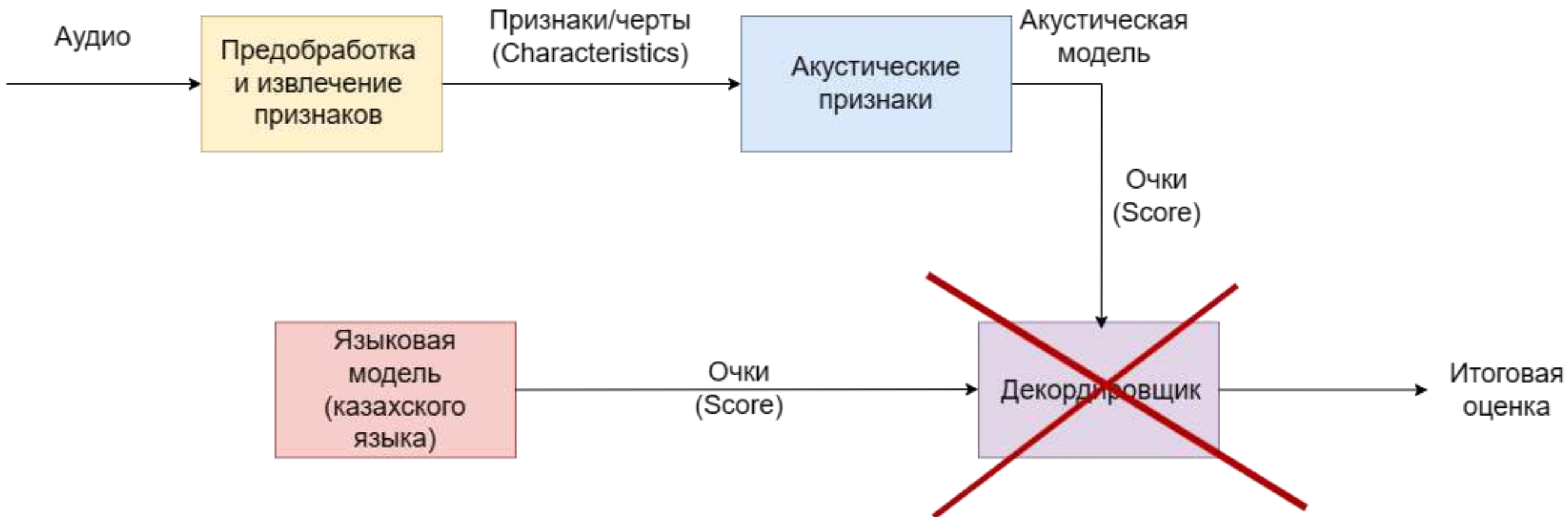


Рис 2.

# Предварительная обработка сигнала и извлечение признаков





# 1. Предварительная обработка сигнала

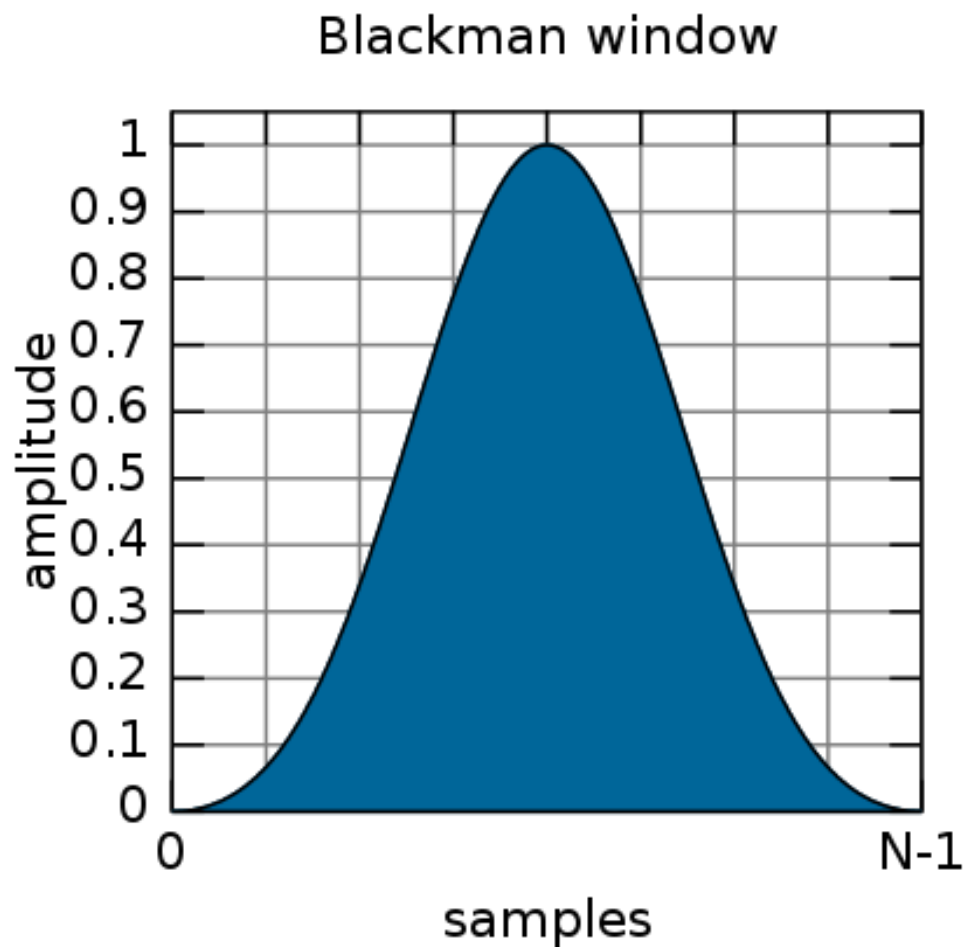
Предварительная обработка речевого сигнала – является важнейшим этапом в построении эффективной и надежной системы распознавания речи. Она состоит из нескольких этапов:

- Предварительное усиление (pre-emphasis)

$$y[n] = x[n] - \alpha \cdot x[n - 1]$$

$$H(z) = 1 - \mu z^{-1}.$$

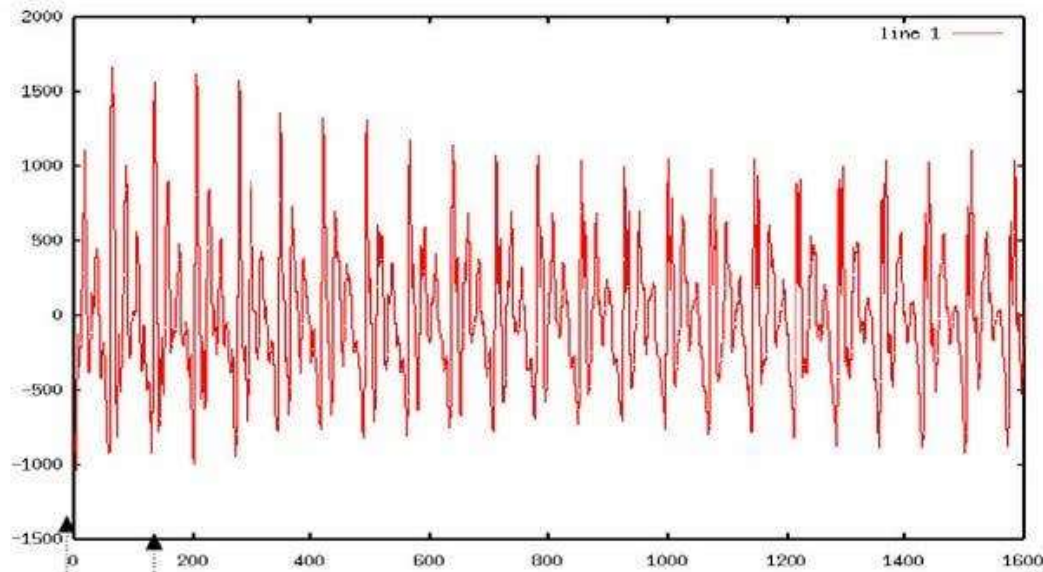
# Framing and windowing



$$w(n) = \begin{cases} 0.56 - 0.47 \cos\left(\frac{2\pi n}{N-1}\right), & 0 \leq n \leq N-1 \\ 0, & \text{other} \end{cases}.$$

$$s_w(n) = w(n) \cdot s(n).$$

Рис 2.



**A**  $\sim 20 - 25$  ms

**B**  $\sim 10$  ms

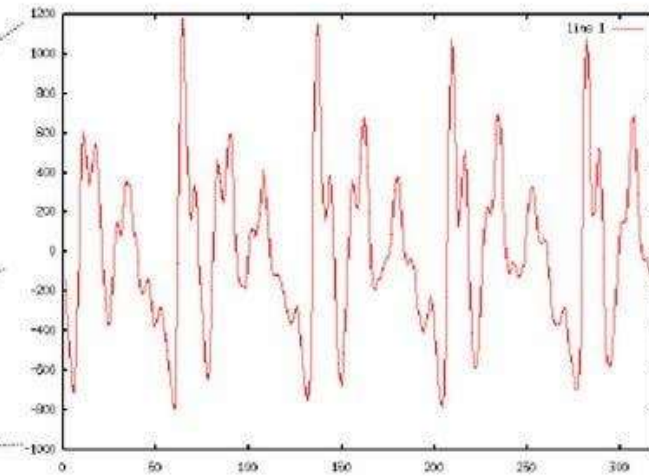
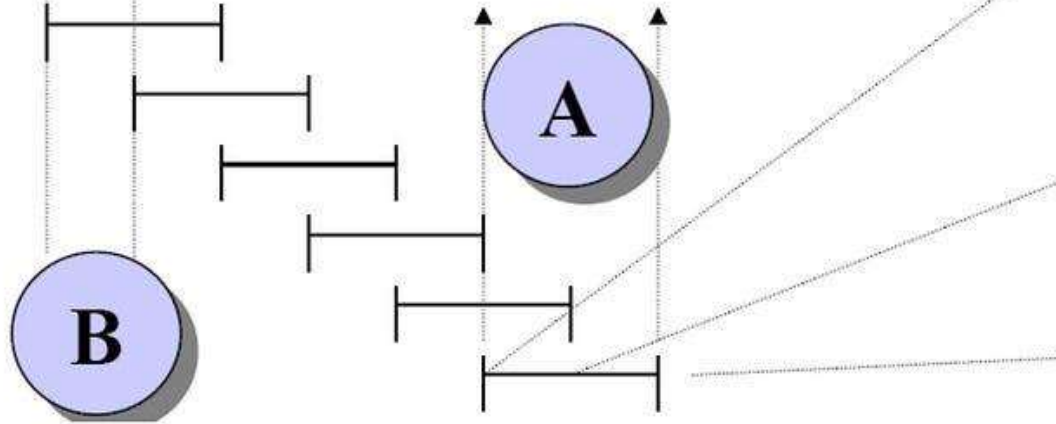
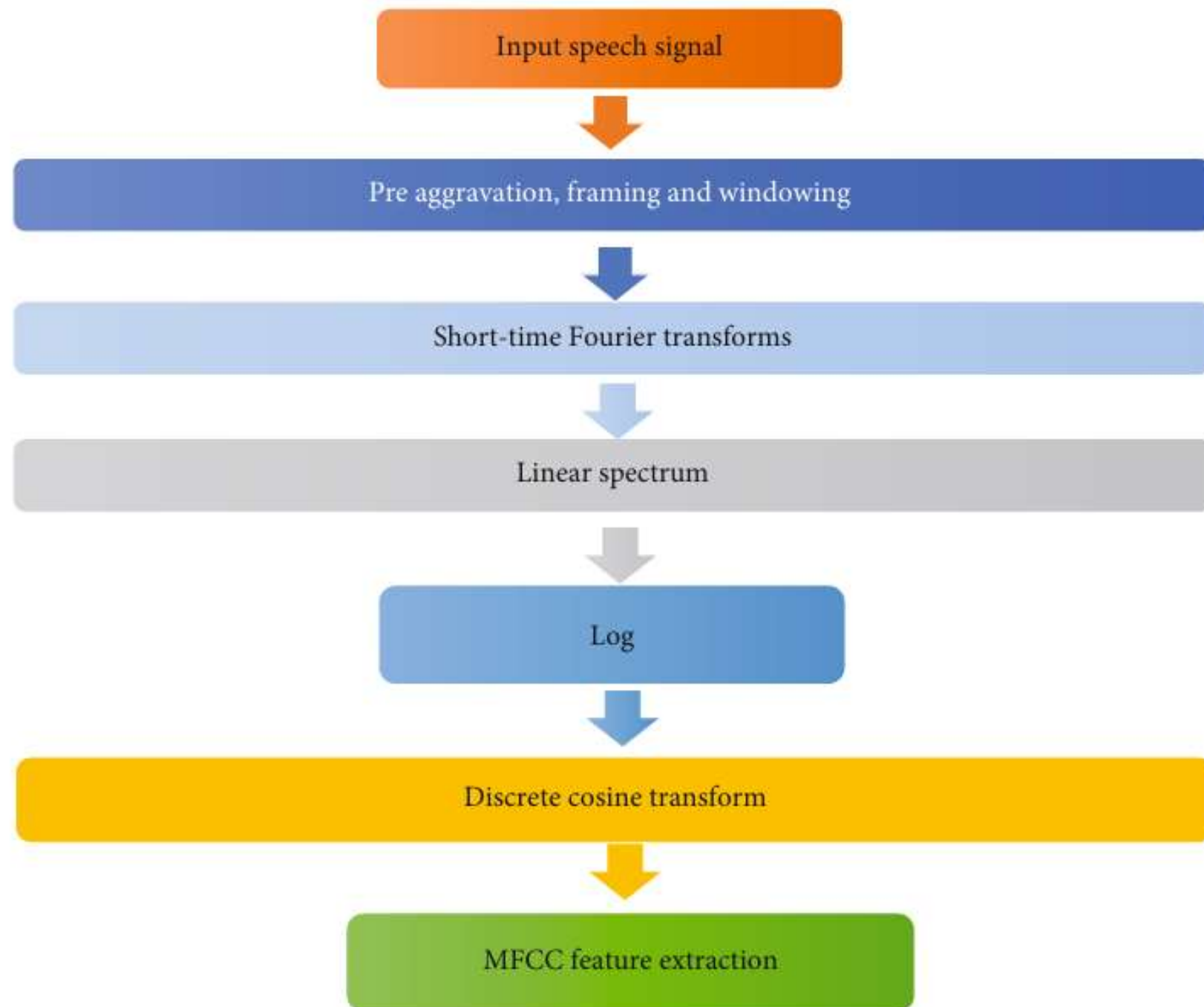


Image from Bryan Pellom

## 2. Извлечение признаков



# Мел спектрограмма

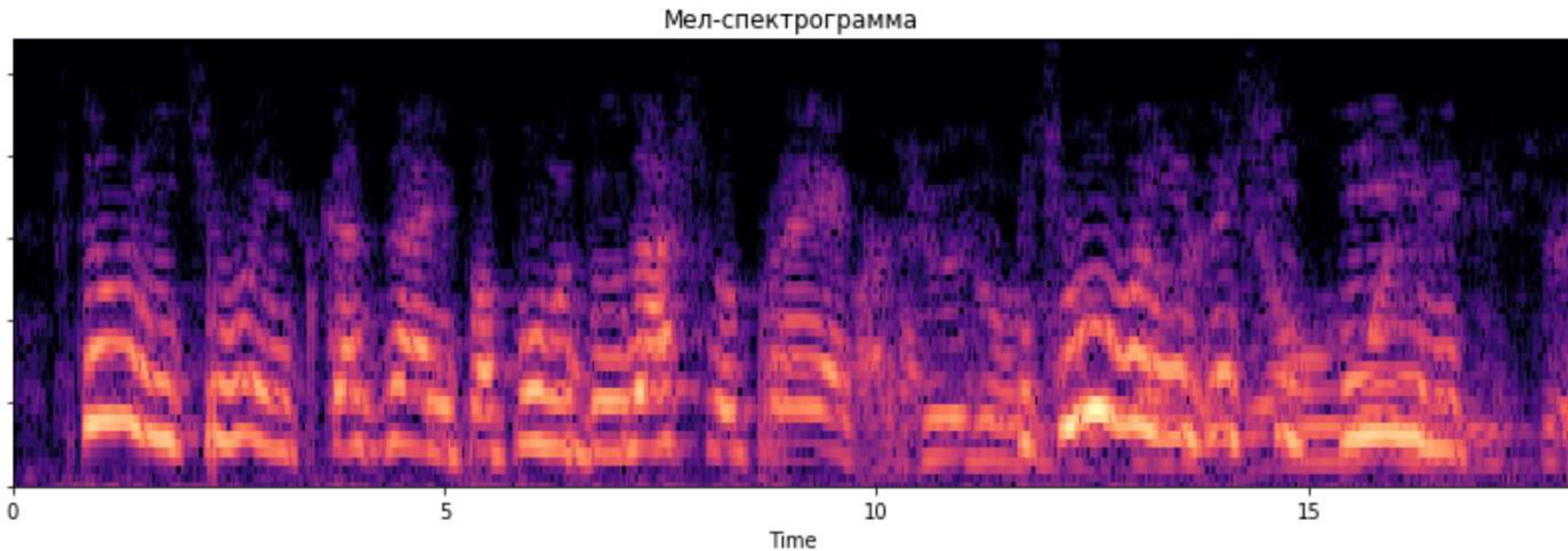
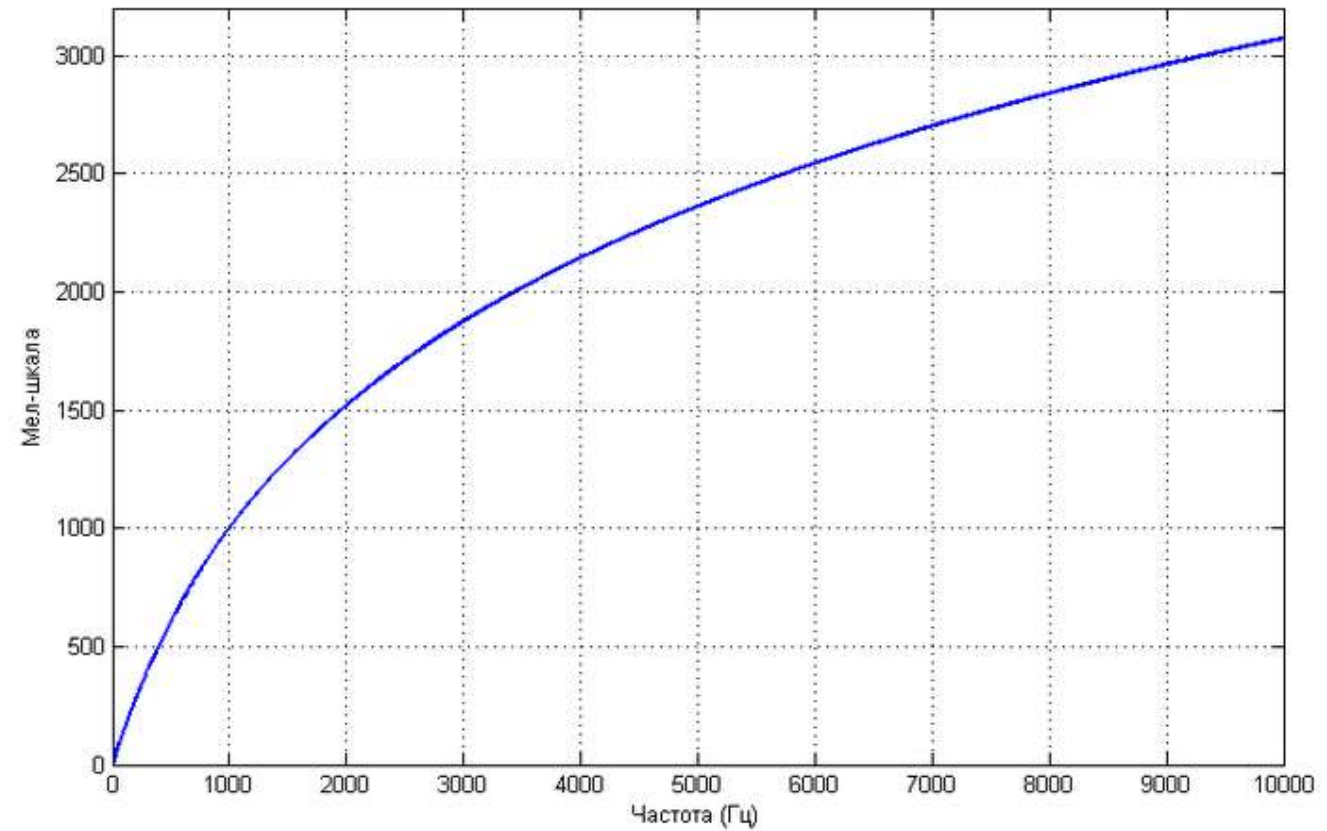
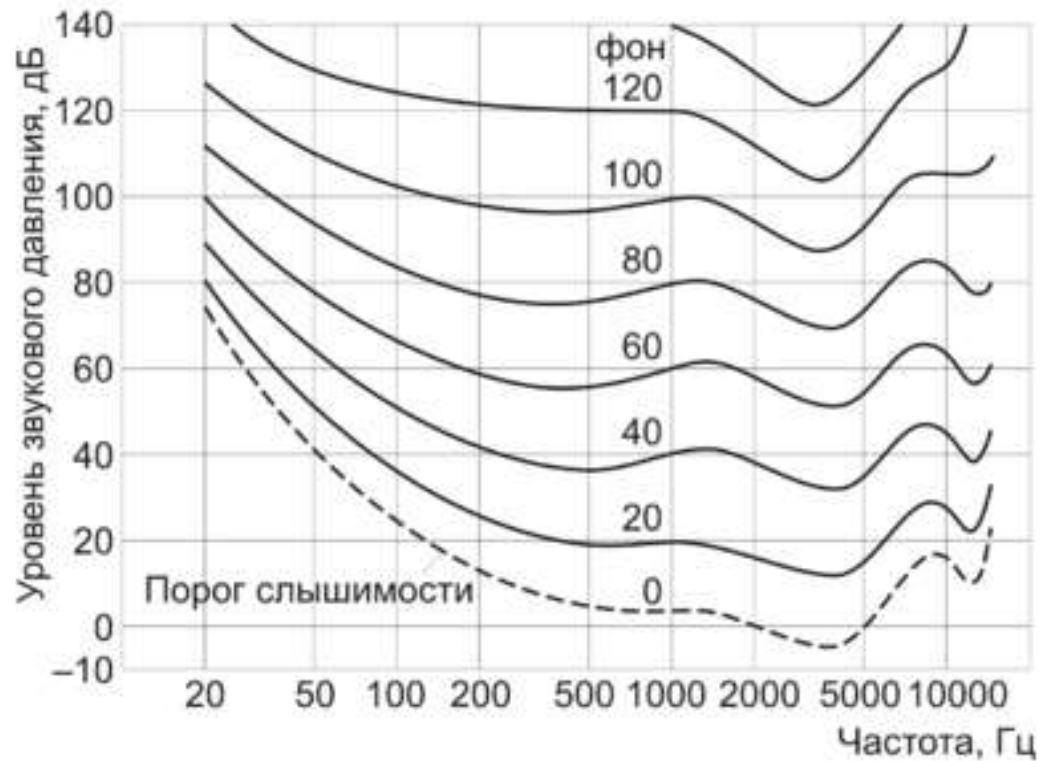


Рис 2.



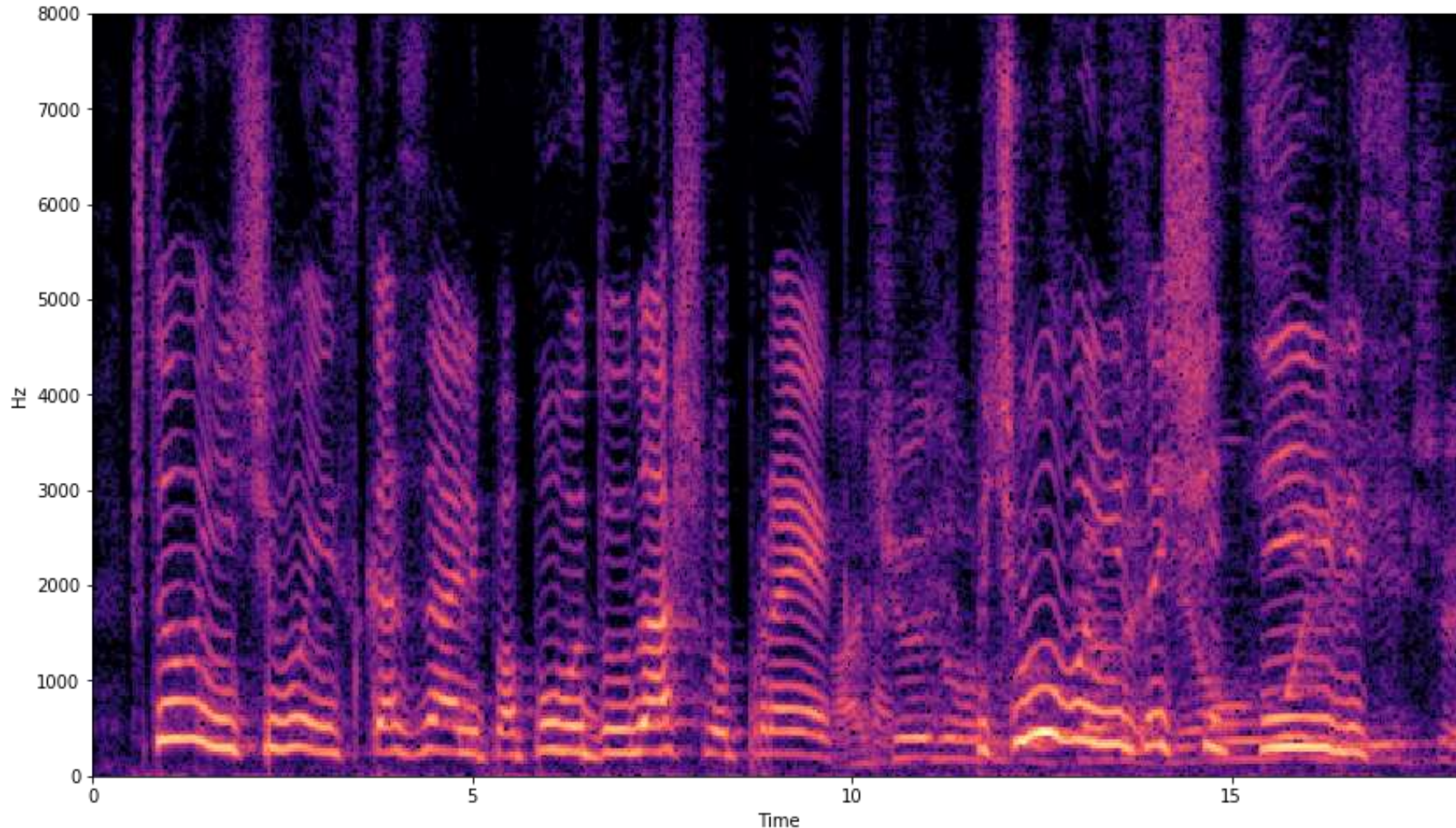
**Мел** - единица высоты звука, основанная на восприятии этого звука нашими органами слуха



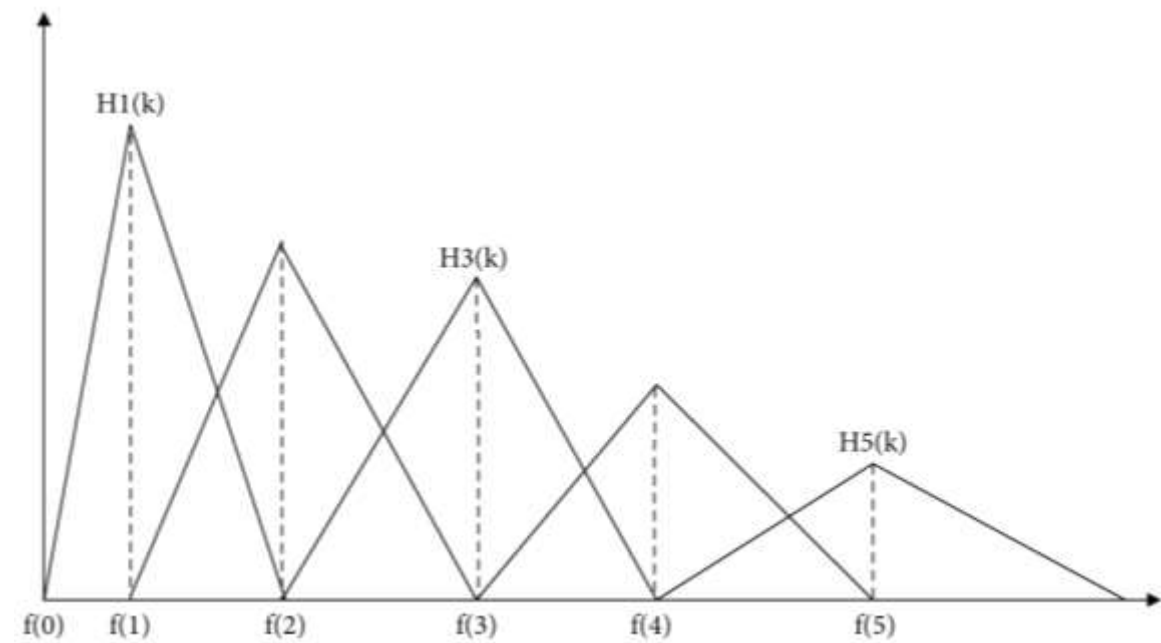
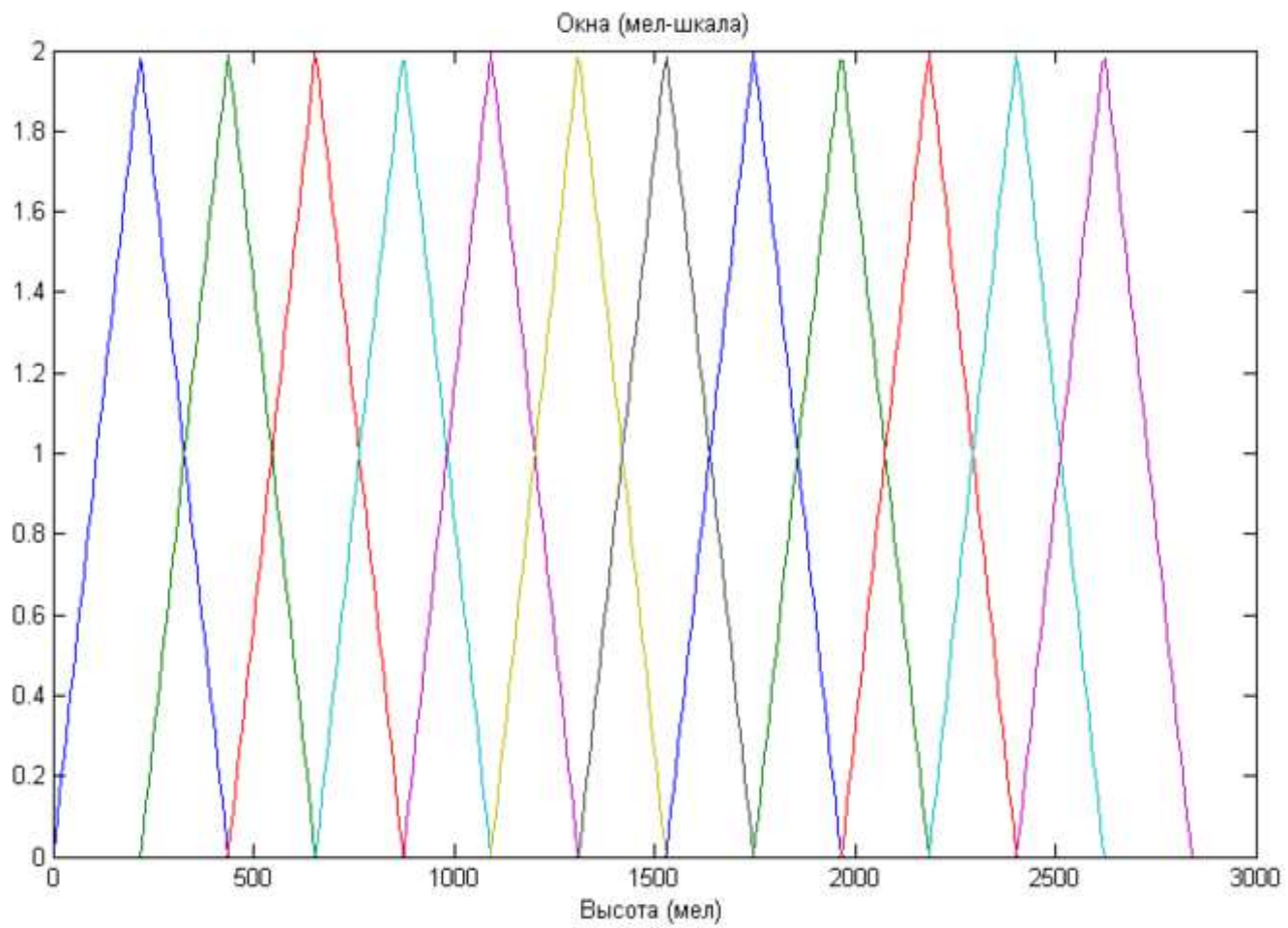
$$\text{mel}(f) = 2595 \lg \left( 1 + \frac{f}{700} \right).$$

# Быстрое-преобразование-Фурье (STFT)

$$S_i(k) = \sum_{n=0}^{N-1} s_i(n) e^{-j\frac{2\pi nk}{N}} (0 < K < N).$$



# Фильтры





Фильтры задаются следующими выражениями:

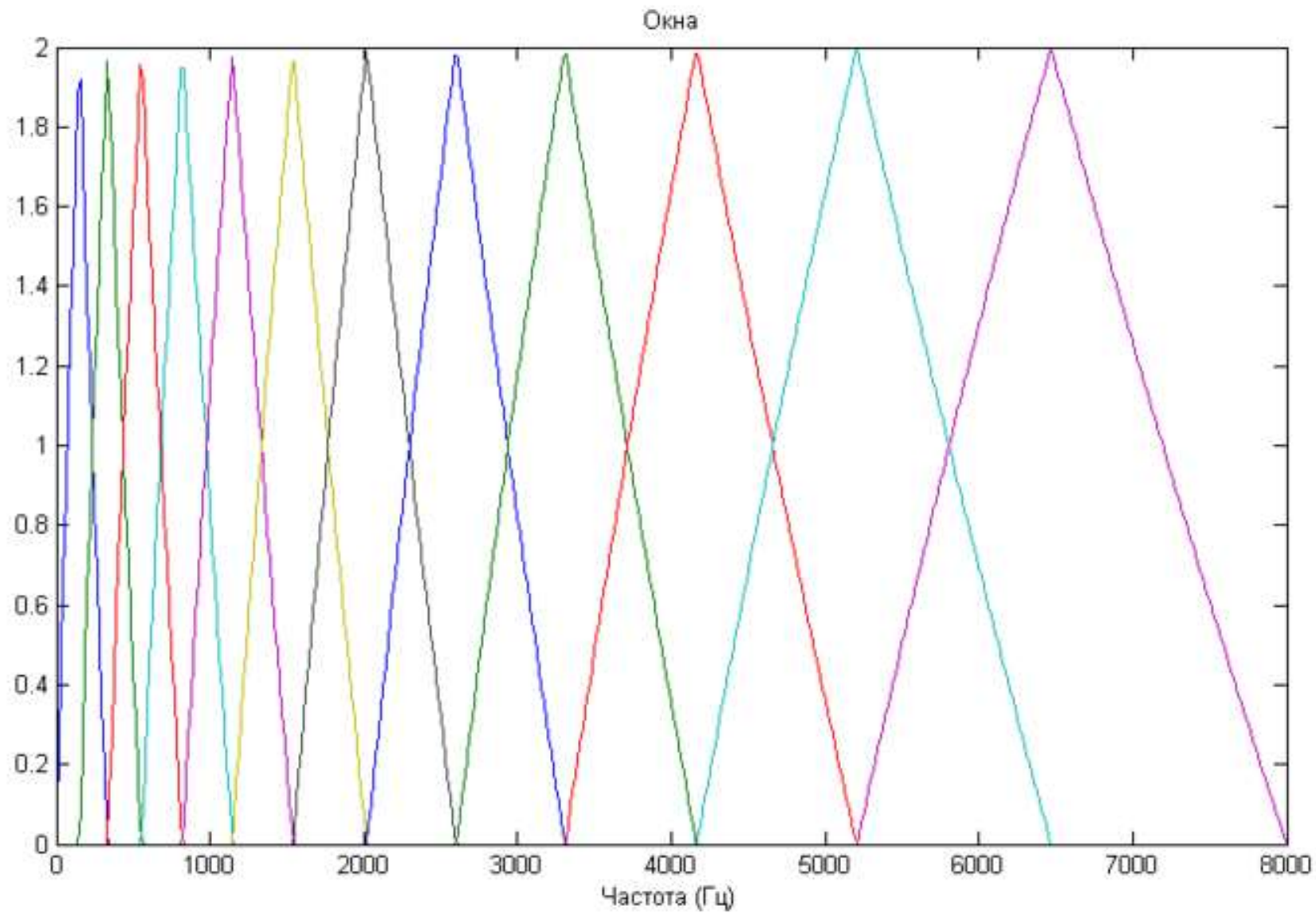
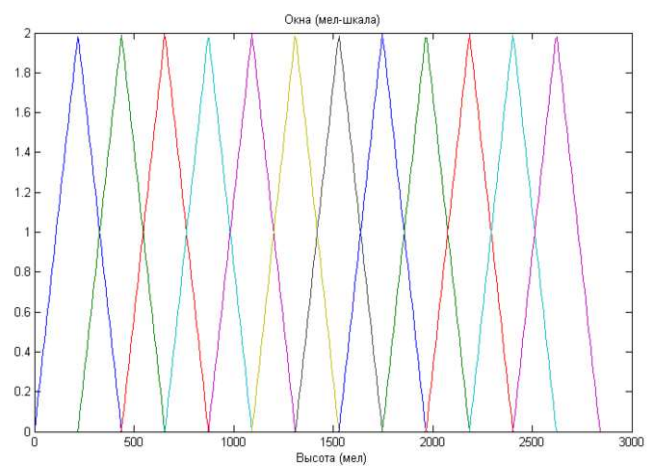
$$H_m = \begin{cases} 0 & k < f[m-1] \\ \frac{(k-f[m-1])}{(f[m]-f[m-1])} & f[m-1] \leq k < f[m] \\ \frac{(f[m+1]-k)}{(f[m+1]-f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & k > f[m+1] \end{cases}$$

где  $f[m]$  получаем из равенства

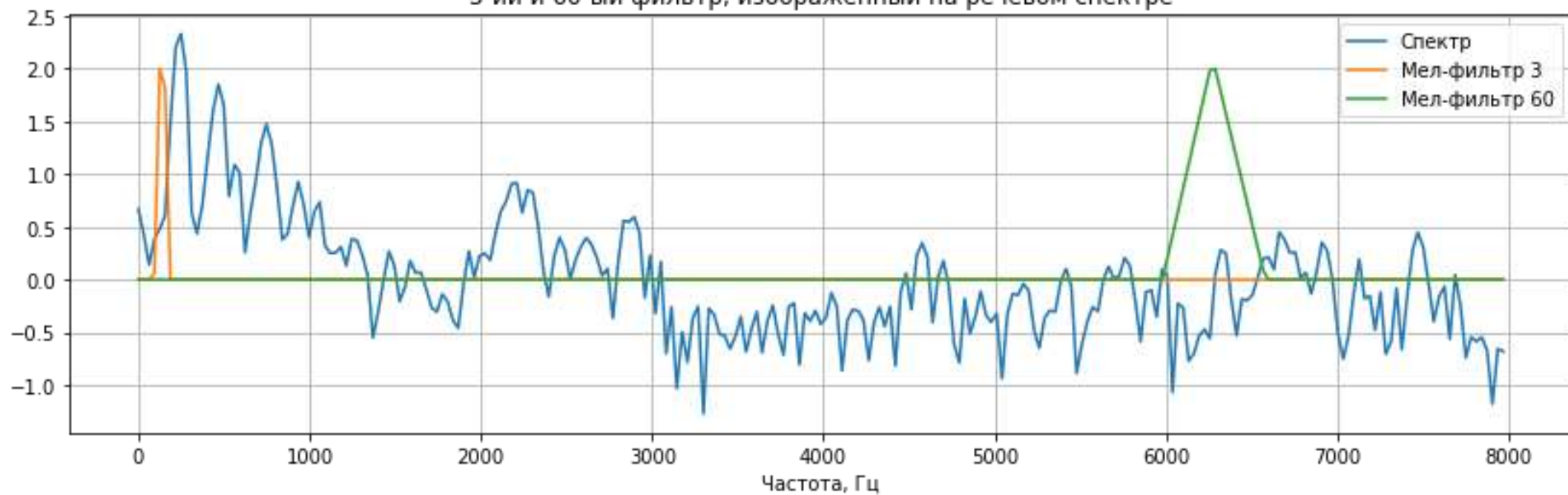
$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_1) + m \frac{B(f_h) - B(f_1)}{M+1}\right)$$

где  $B(b)$  – преобразование значения частоты в мел-шкалу, соответственно,

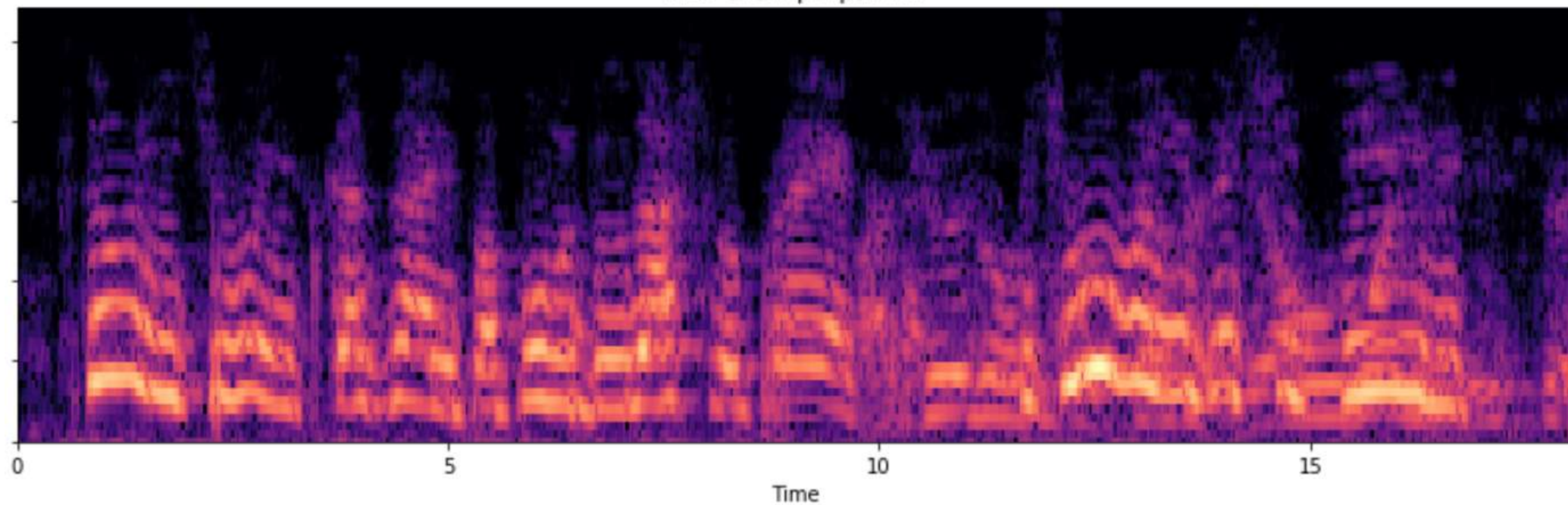
$$B^{-1}(b) = 700(\exp(b/1125) - 1)$$



3-ий и 60-ый фильтр, изображенный на речевом спектре

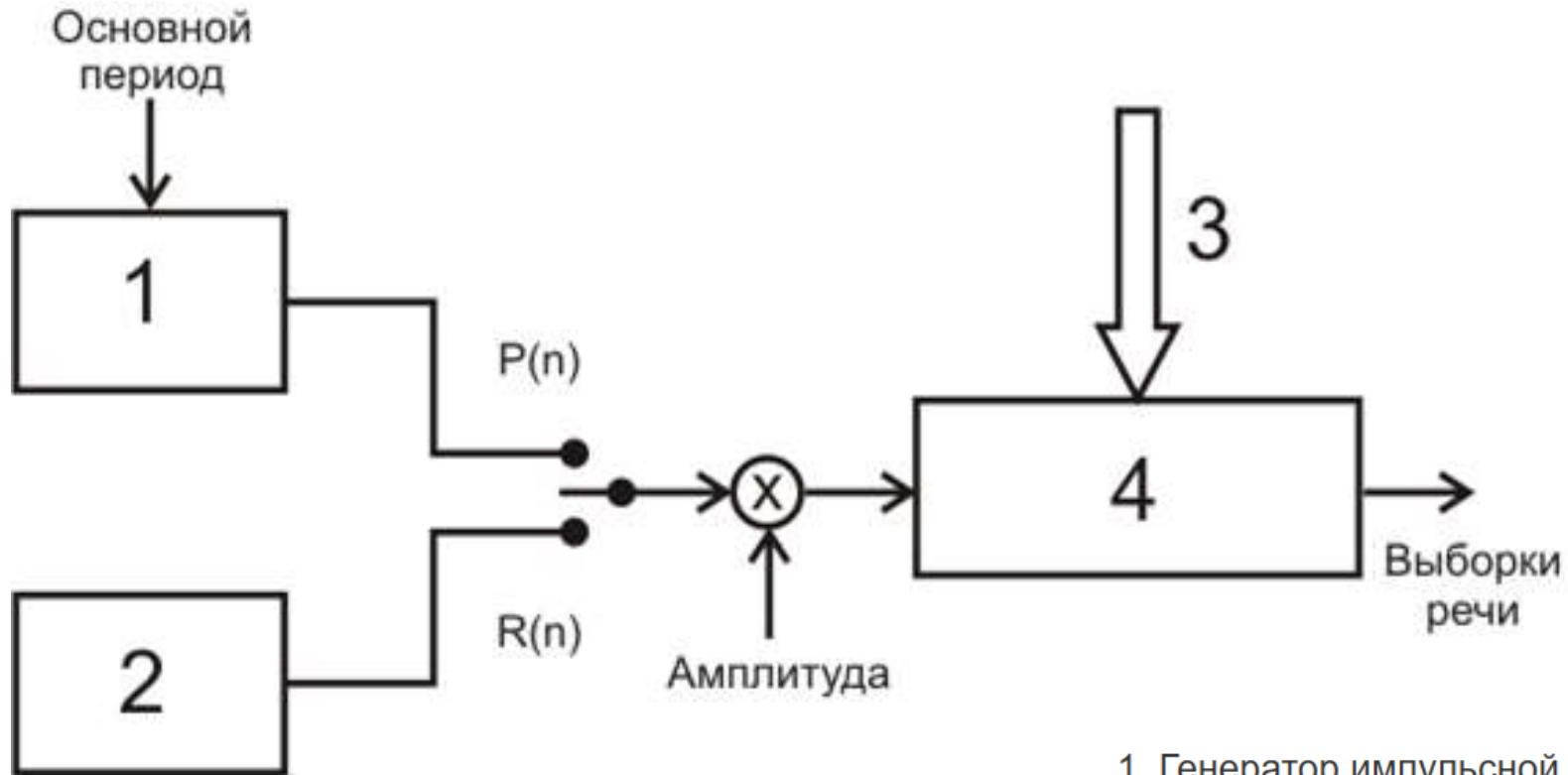


Мел-спектрограмма



# Кепстр

В соответствии с теорией речеобразования речь представляет собой акустическую волну, которая излучается системой органов: легкими, бронхами и трахеей, а затем преобразуется в голосовом тракте



1. Генератор импульсной последовательности (тонов)
2. Генератор случайных чисел (шумов)
3. Коэффициенты цифрового фильтра (параметры голосового тракта)
4. Нестационарный цифровой фильтр

# Акустическая модель



# Определение

**Акустическая модель** — это функция, принимающая на вход признаки на небольшом участке акустического сигнала (фрейме) и выдающая распределение вероятностей различных фонем на этом фрейме.

**Скрытая Марковская модель (СММ)** — это модель которая представляет из себя марковскую цепь.

# Цепь Маркова

**Цепь Маркова** — последовательность случайных событий с конечным или счётным числом исходов, характеризующаяся тем, что при фиксированном настоящем будущее независимо от прошлого.

Они характеризуется тем что:

- Процесс в каждый момент времени находится в одном из  $n$  состояний;
- При этом, если мы находимся в состоянии с номером  $i$ , то мы можем перейти в состояние  $j$  с вероятностью  $p_{ij}$ .

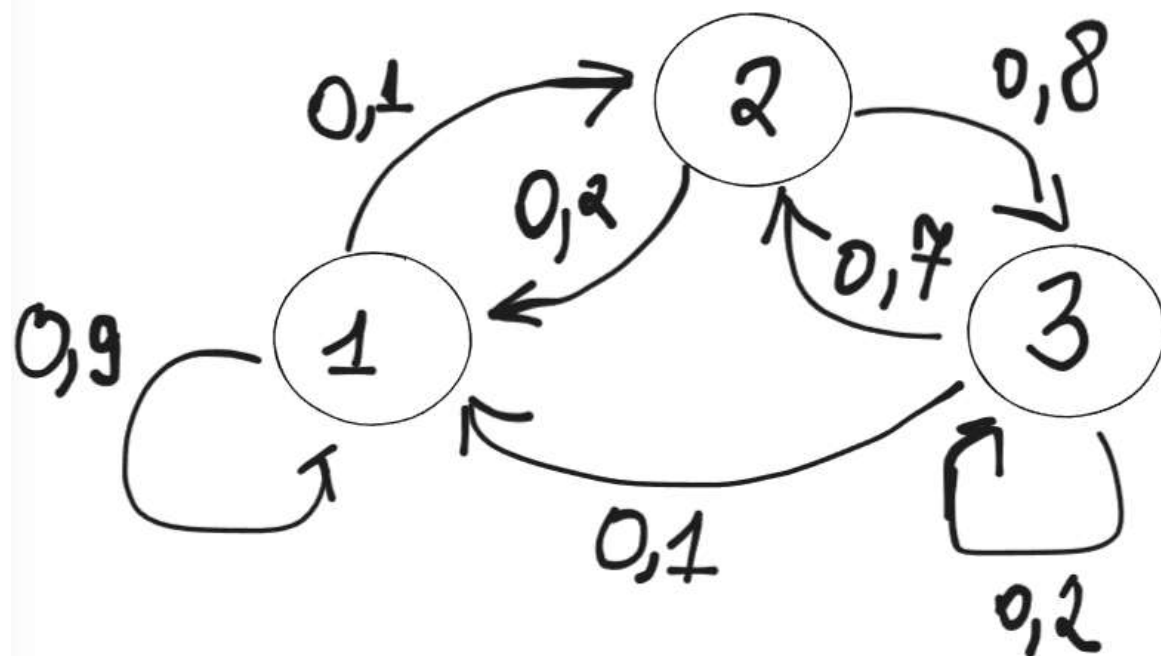
Переходы из одного состояние в другое записаны в матрицу такую матрицу называют **матрица переходов**

$P = ||p_{ij}||$  и на неё накладываются следующие условия

1.  $p_{ij} \geq 0$
2.  $\forall i \sum_j p_{ij} = 1$



# Цепь Маркова



1 : Чтение книги

2 : Разговор по телефону

3 : Работа над задачей



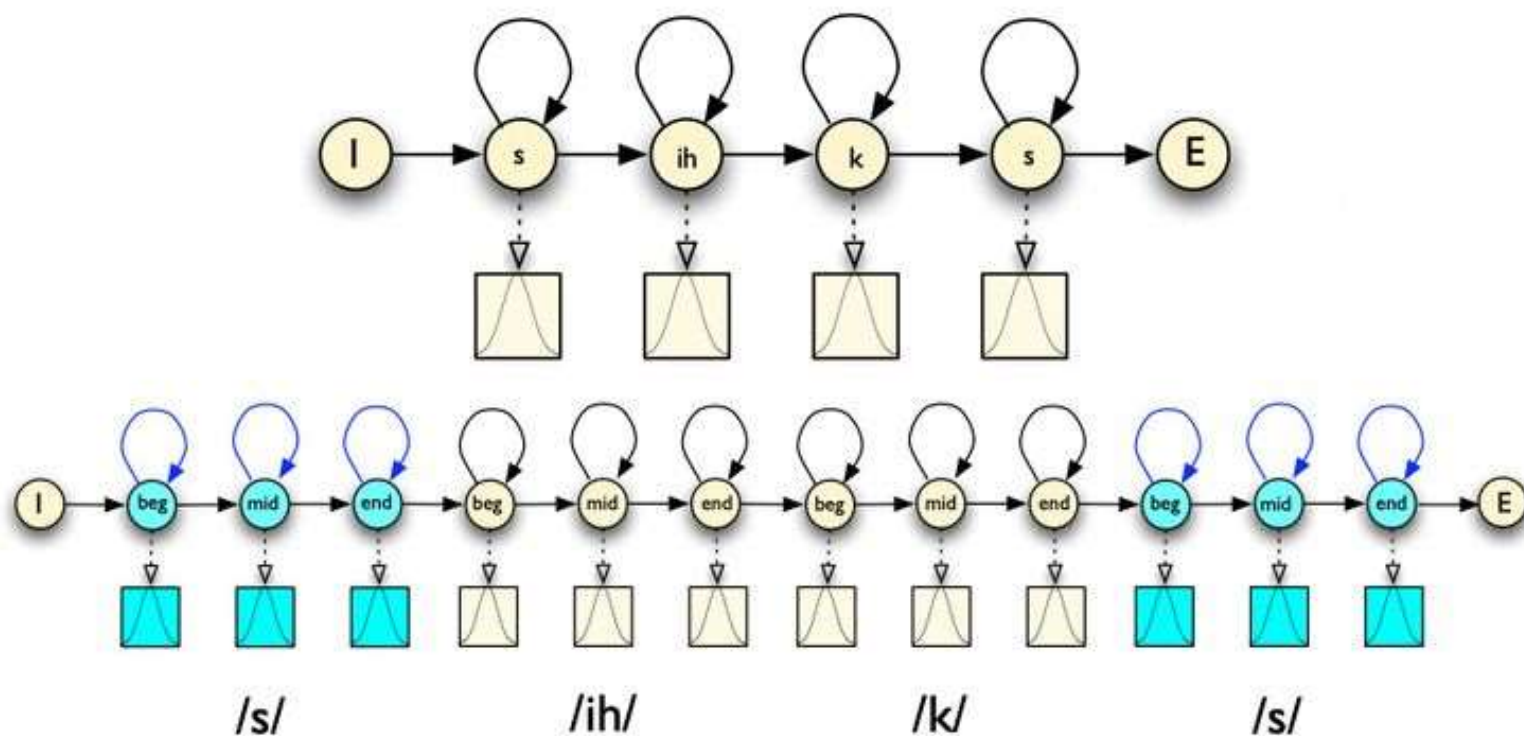
# Марковская модель

**Скрытая Марковская модель** (англ. A hidden Markov model) — модель процесса, в которой процесс считается Марковским, причем неизвестно, в каком состоянии  $s_i$  находится система (состояния скрыты), но каждое состояние  $s_i$  может с некоторой вероятностью  $b_{i\omega_j}$  произвести событие  $\omega_j$ , которое можно наблюдать.

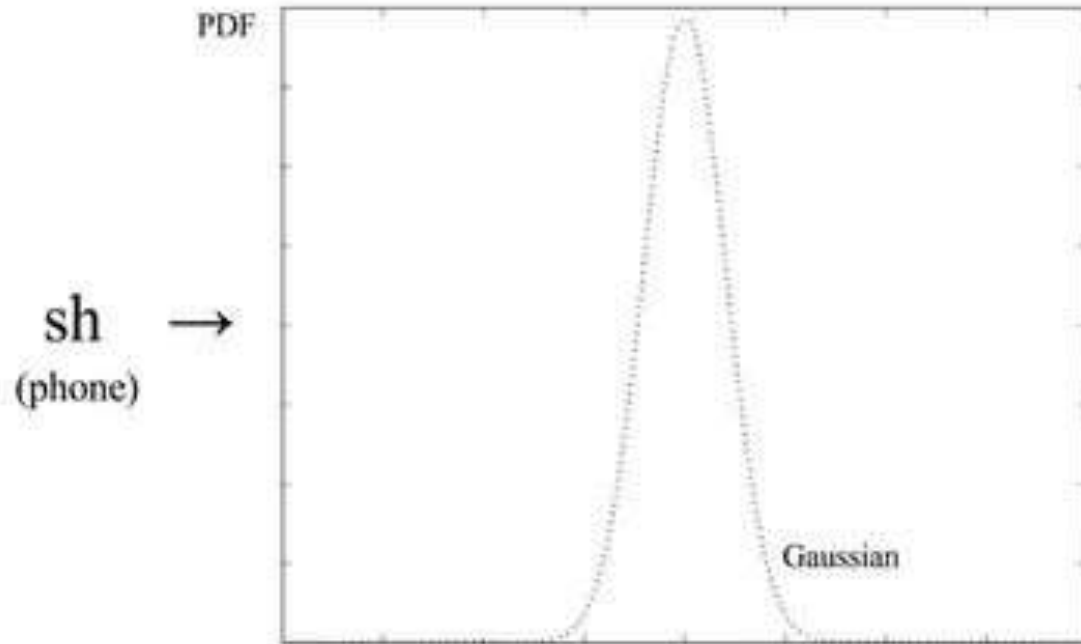
Марковская модель  $\lambda$  задается как  $\lambda = \{S, \Omega, \Pi, A, B\}$ , где  $S = \{s_1 \dots s_n\}$  — состояния,  $\Omega$  — возможные события,  $\Pi = \{\pi_1 \dots \pi_n\}$  — начальные вероятности,  $A = \{a_{ij}\}$  — матрица переходов, а  $B = \{b_{i\omega_k}\}$  — вероятность наблюдения события  $\omega_k$  после перехода в состояние  $s_i$ .

# Акустическая модель

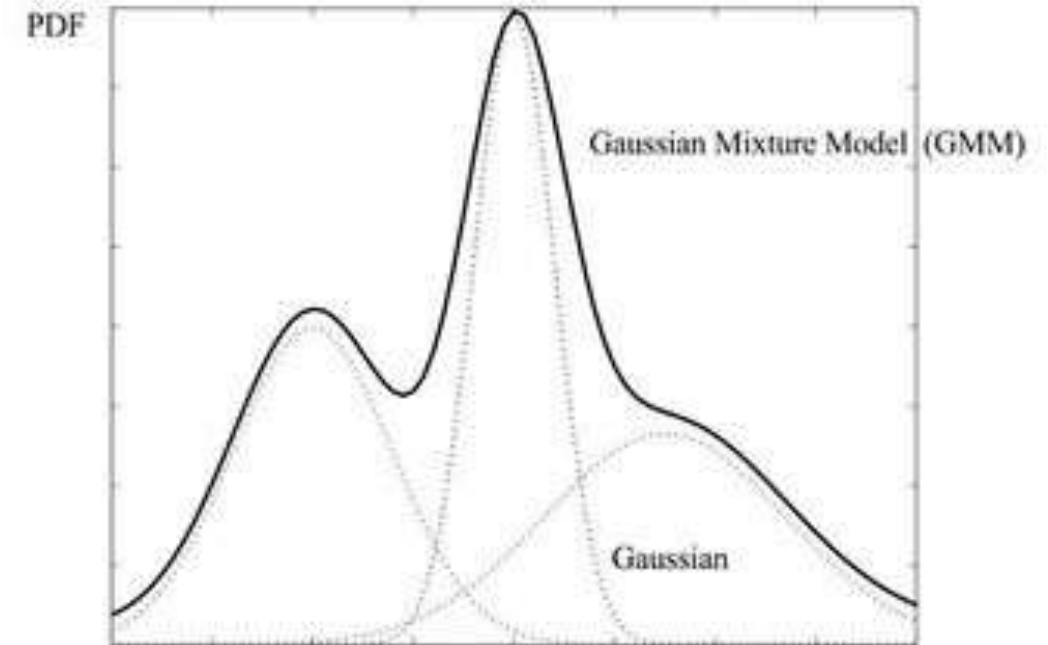
**Акустическая модель** — это функция, принимающая на вход признаки на небольшом участке акустического сигнала (фрейме) и выдающая распределение вероятностей различных фонем на этом фрейме.



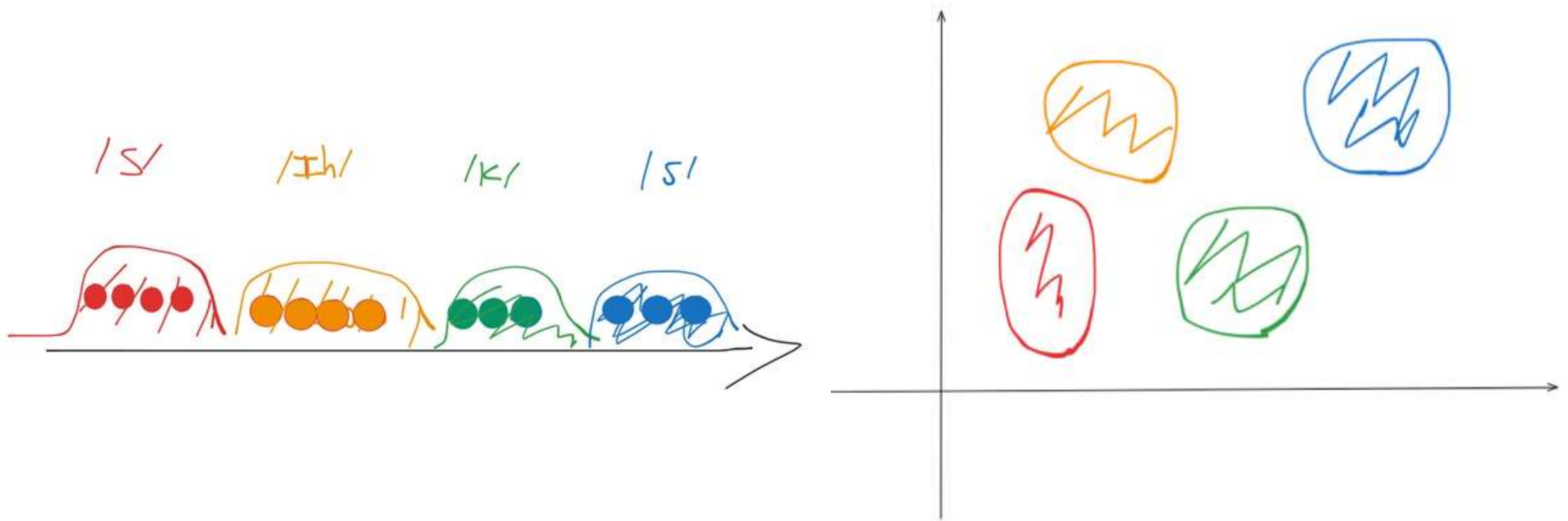
# Акустическая модель



acoustic model

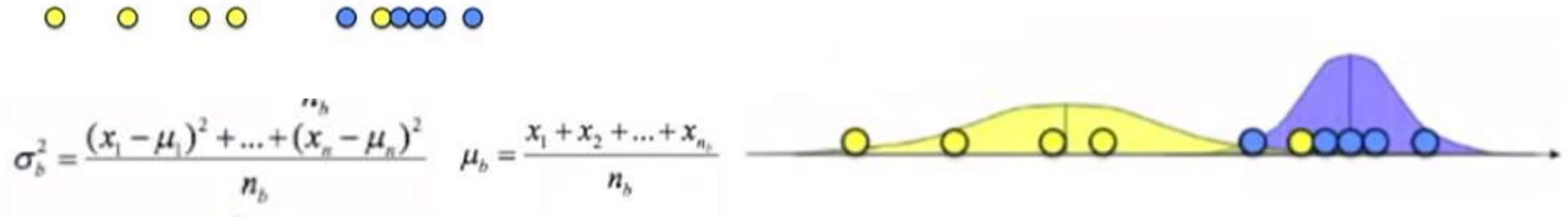


# Акустическая модель



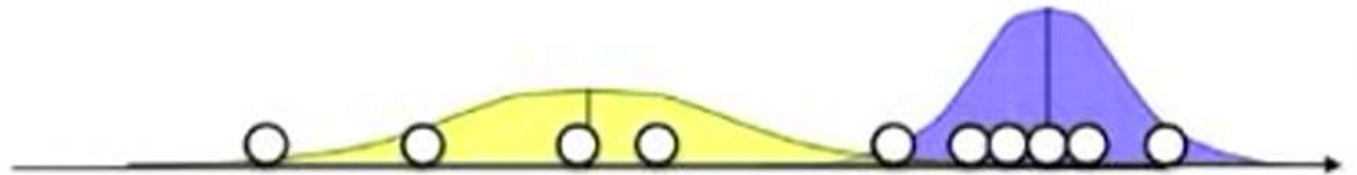
# ЕМ алгоритм

При СГМ есть все основание пользоваться мягкой кластеризацией.



$$P(b | x_i) = \frac{P(x_i | b)P(b)}{P(x_i | b)P(b) + P(x_i | a)P(a)}$$

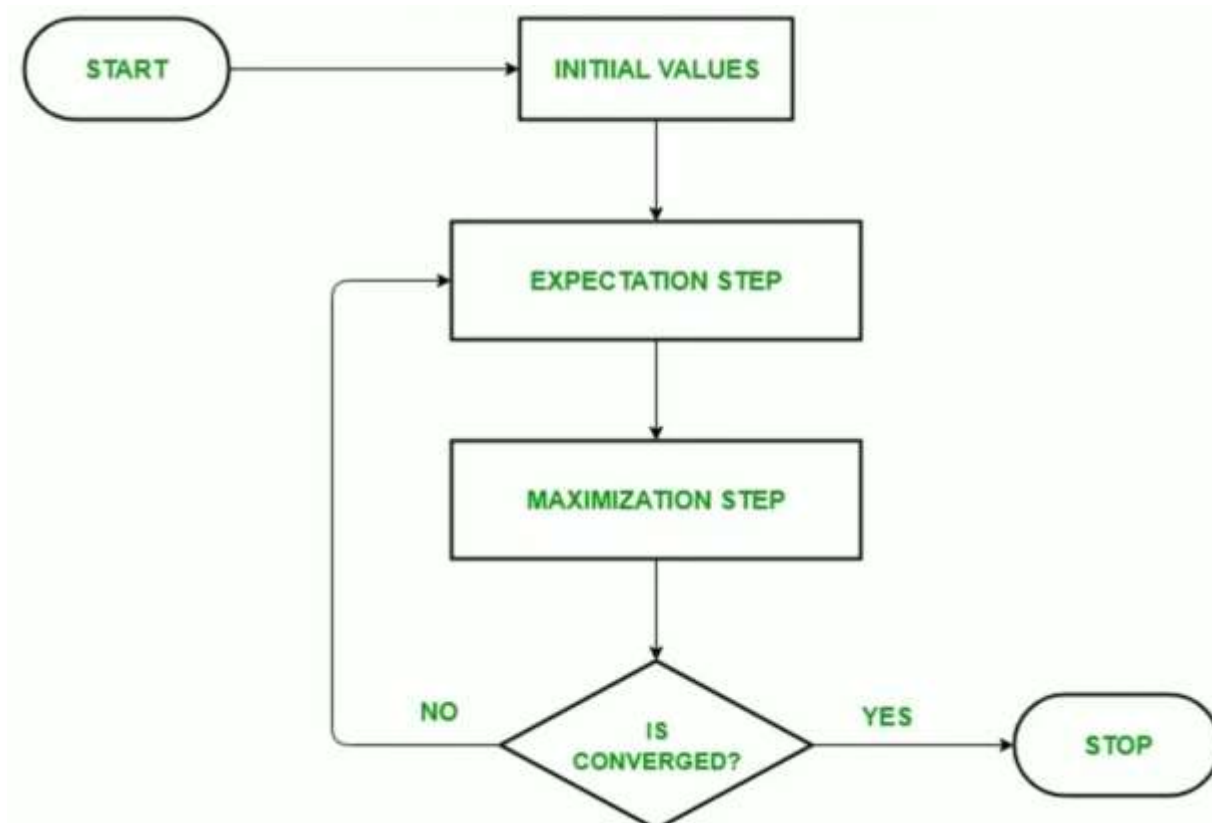
$$P(x_i | b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right)$$



Курица и яйцо ?

НЕТ

ЕМ алгоритм !



# Акустическая модель

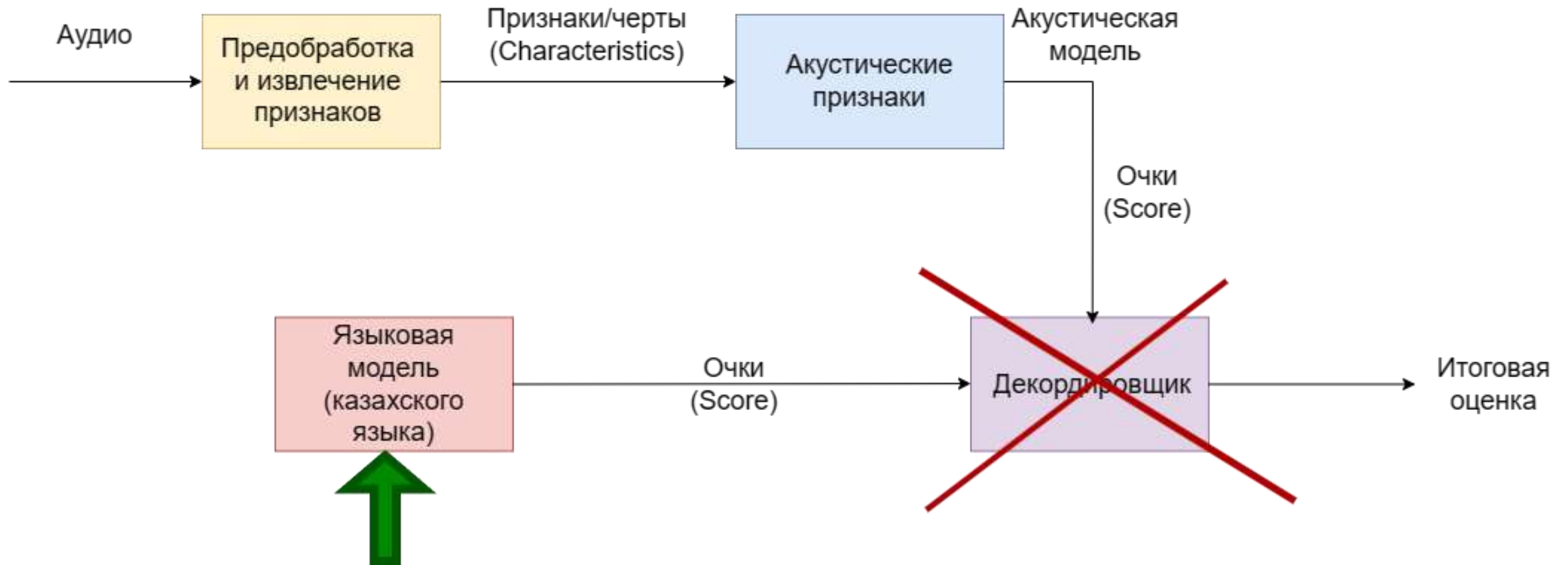
Вероятностное распределение непрерывных наблюдений объясняется с помощью функции плотности вероятности СГМ.

$$b_i(o_t) = \sum_{m=1}^M \frac{c_{i,m}}{(2\pi)^{D/2} |\Sigma_{im}|^{1/2}} \exp \left[ -\frac{1}{2} (o_t - \mu_{i,m})^T \sum_{i,m}^{-1} (o_t - \mu_{i,m}) \right].$$

Когда компонент смеси  $m$  уменьшается до 1, вероятностное распределение выходных данных на основе этого состояния дегенерируется в гауссовское распределение, как показано в уравнении

$$b_i(o_t) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp \left[ -\frac{1}{2} (o_t - \mu_i)^T \sum_i^{-1} (o_t - \mu_i) \right].$$

# Языковая модель



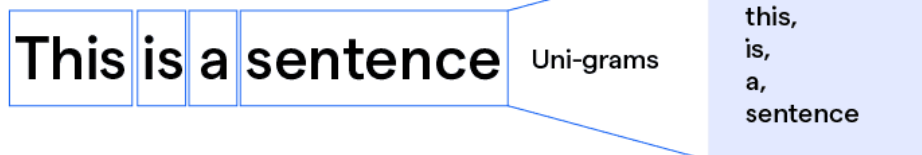


# Определение

**Языковая модель** — позволяет узнать, какие последовательности слов в языке более вероятны, а какие менее. Здесь в самом простом случае требуется предсказать следующее слово по известным предыдущим словам

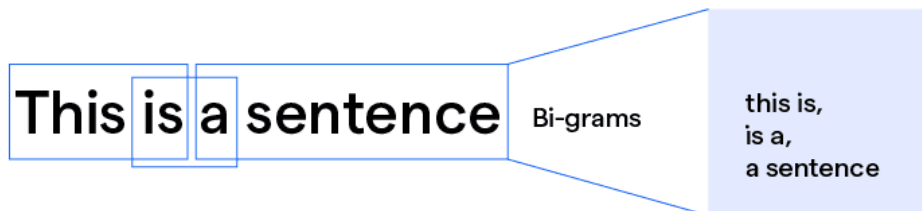
## N-Gram

N=1:



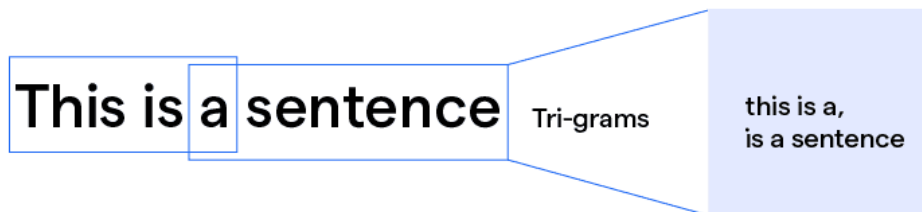
$$P(w|h) = \frac{c_1}{c_2}.$$

N=2:



$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2) \cdots P(w_n|w_1^{n-1}) = \prod_{k=1}^n P(w_k|w_1^{k-1}).$$

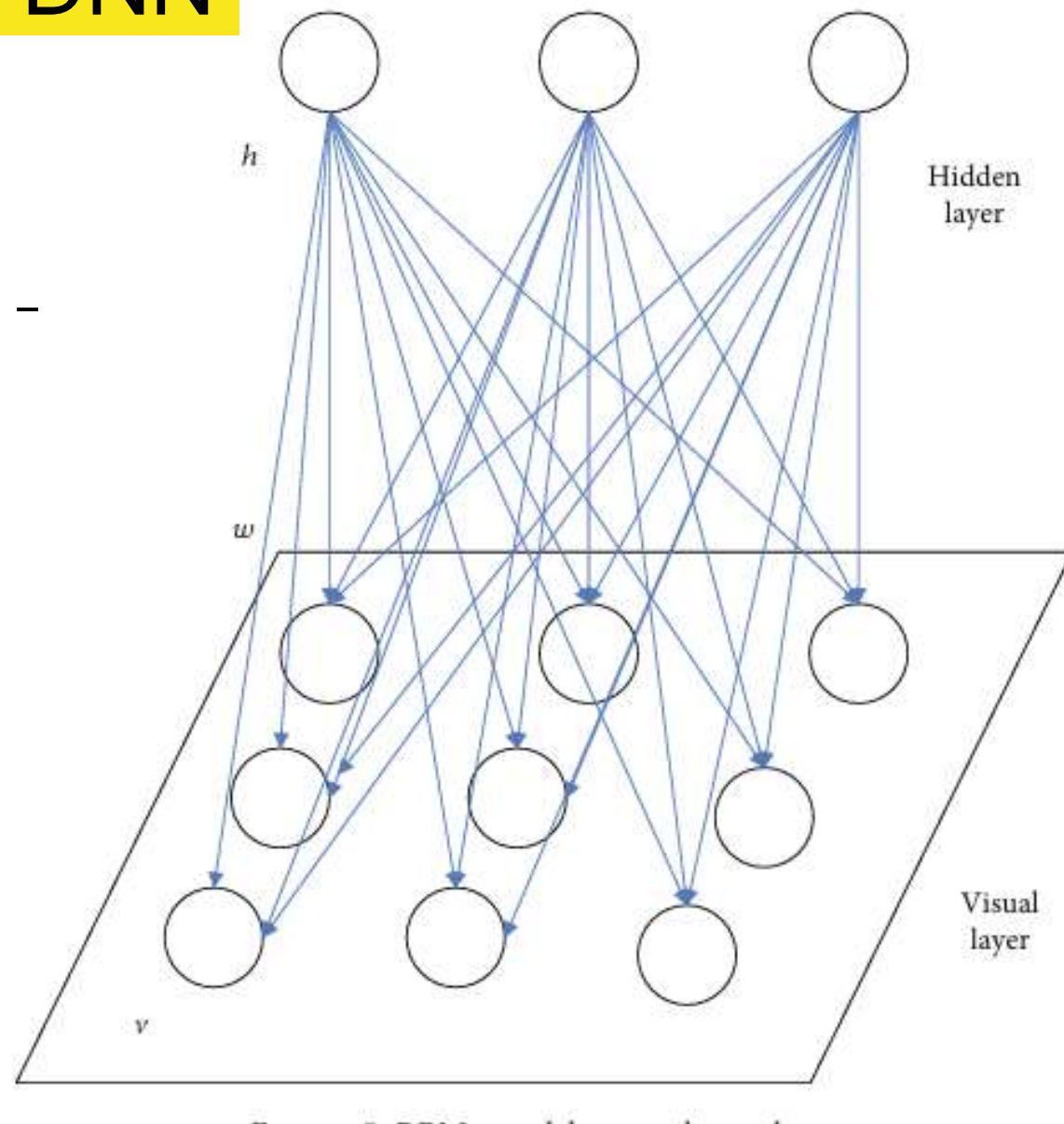
N=3:



# Глубокие нейронные сети. DNN

## I . Ограниченная машина Больцмана (RBMs) –

это особый тип нейронных сетей, которые используют вероятностные методы для создания состояний нейронов. Они генерируют эти состояния, опираясь на различные вероятностные подходы. Из этого можно сделать вывод, что поведение нейронов в сети можно описать определенным вероятностным распределением

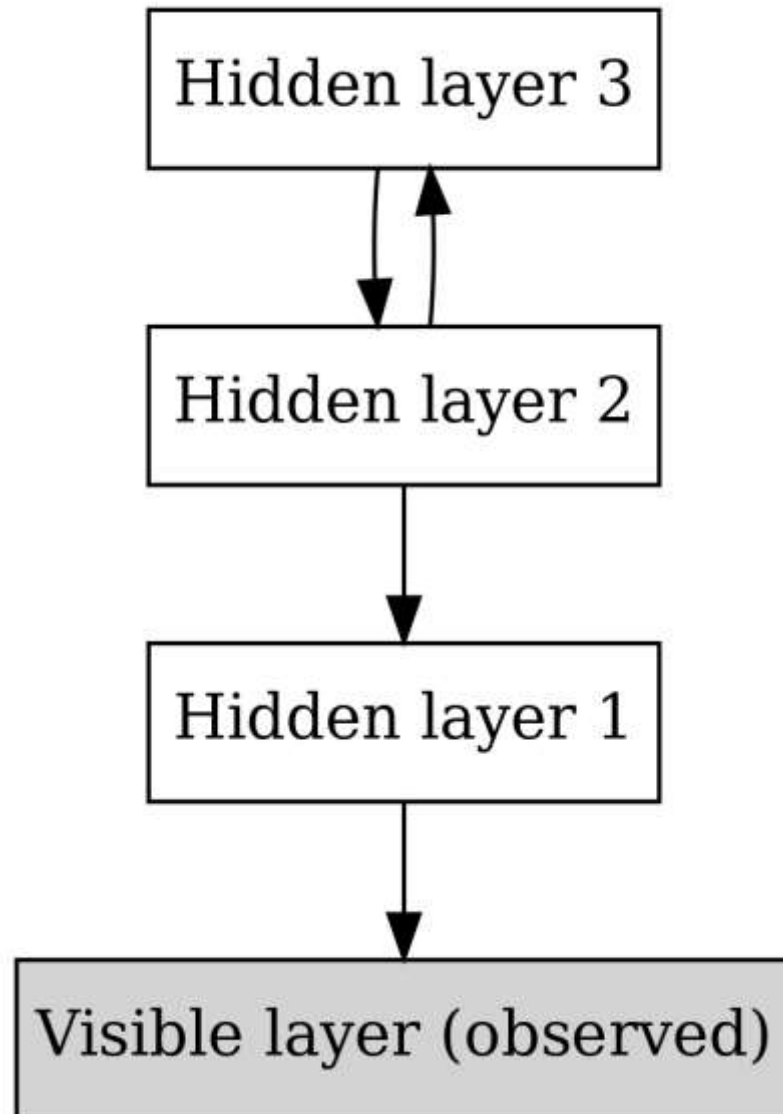


# Глубокие нейронные сети. DNN



**II. DBN (Deep belief NN) или «Сеть глубокого доверия»** – также являются случайными глубокими нейронными сетями, и поэтому они сочетают в себе неконтролируемое обучение

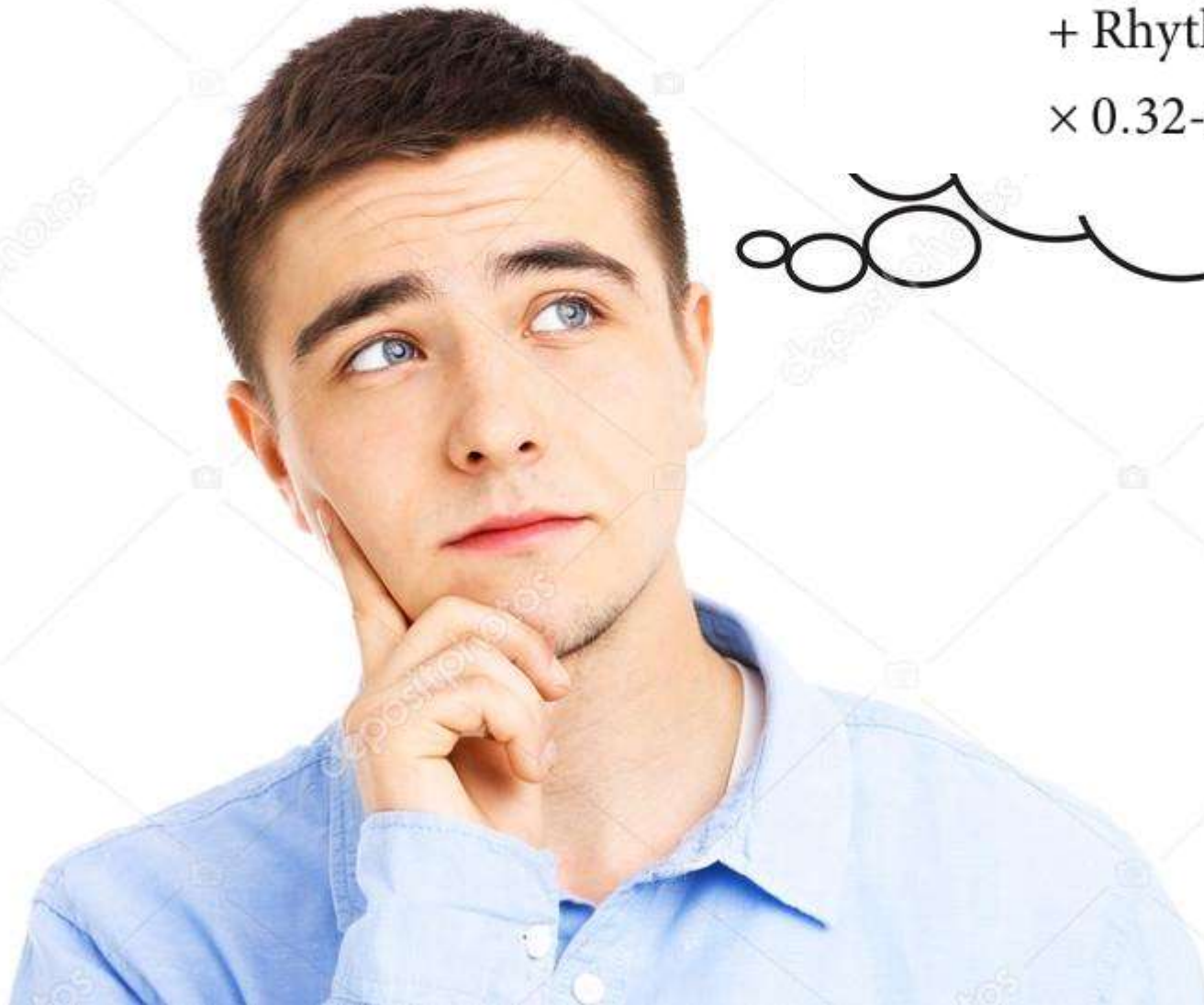
**STACKING + RBM => DBN**

Wikipedia: [Deep belief network - Wikipedia](#)



# Оценка


$$\text{Score} = \text{AccuracyScore} \times 0.45 + \text{SpeedScore} \times 0.17 \\ + \text{RhythmScore} \times 0.36 + \text{IntonationScore} \\ \times 0.32 - 0.402.$$




# Список литературы

- E. Bocchieri, "System and method for speech recognition modeling for mobile voice search," Jersey City nj Usphiladelphia Pa Uschatham Nj Us, vol. 47, no. 10, pp. 4888–4891, 2017.
- M. Telmem and Y. Ghanou, "Estimation of the optimal HMM parameters for amazigh speech recognition system using CMU-Sphinx," Procedia Computer Science, vol. 127, pp. 92–101, 2018.
- He, G. Jin, and S. B. Tsai, "Design and implementation of embedded real-time English speech recognition system based on big data analysis," Mathematical Problems in Engineering, vol. 2021, 2021.
- W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," Speech Communication, vol. 67, pp. 154–166, 2015.
- Z. R. Shi and J. X. Chen, "Event detection via recurrent and convolutional networks based on language model," Journal of Xiamen University (Natural Science), vol. 58, no. 3, pp. 442–448, 2019.
- J. Yang, Y. D. Sun et al., "Weakly supervised learning with denoising restricted Boltzmann machines for extracting features," Acta Electronica Sinica, vol. 12, pp. 2365–2370, 2014.



**Спасибо большое за внимание !**