WaveNet: A Generative Model for Raw Audio

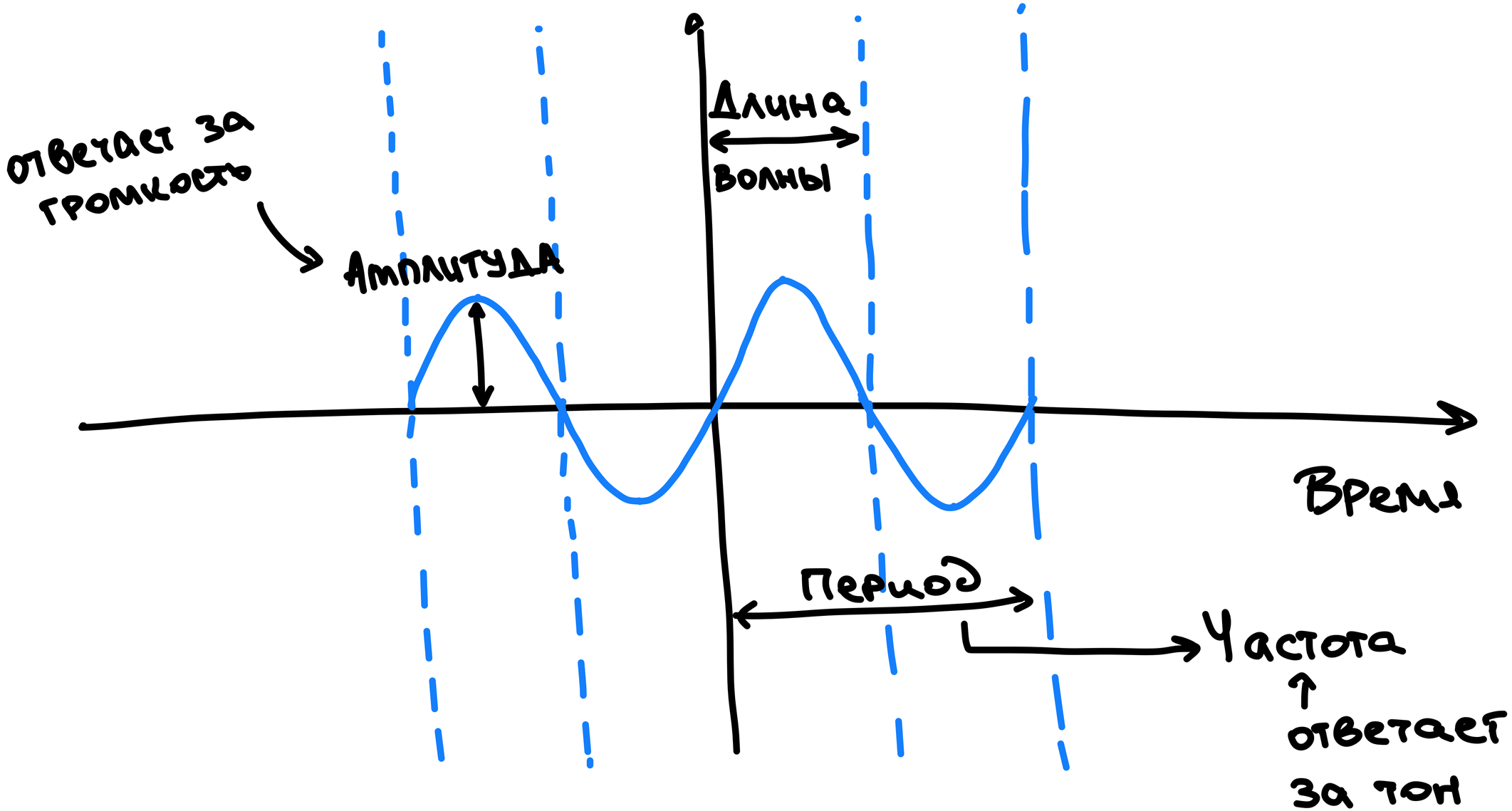
Nadezda Kondrateva, 2023

План

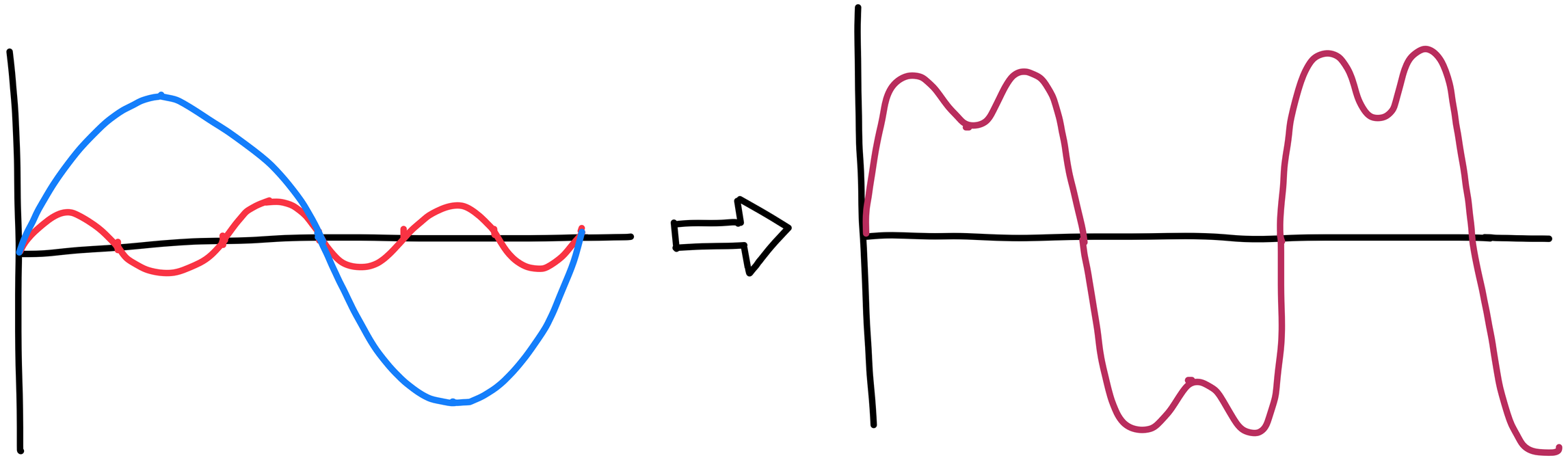
Звук

Звук — колебательное движение частиц упругой среды, распространяющееся в виде волн в газообразной, жидкой или твёрдой средах

Звуковая волна - разрежение и сгущение звука


Отвечает за громкость
Амплитуда
Длина волны
Время
Период
Частота
↑
отвечает за тон

1 Second



vibrations in the air are analog signal

Digital to analog

Incide a computer

DAC

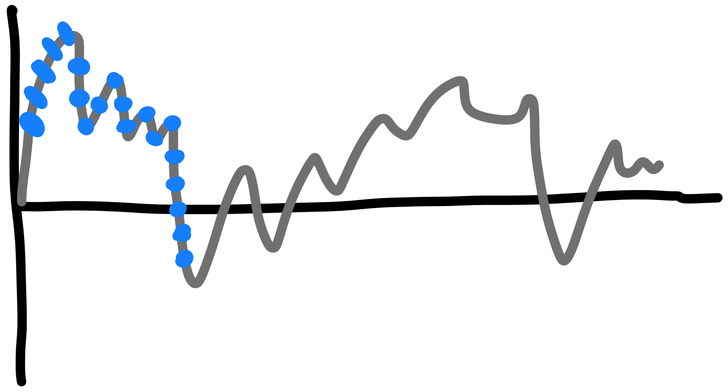
Amplifier

Storage device

Binary signal

ADC

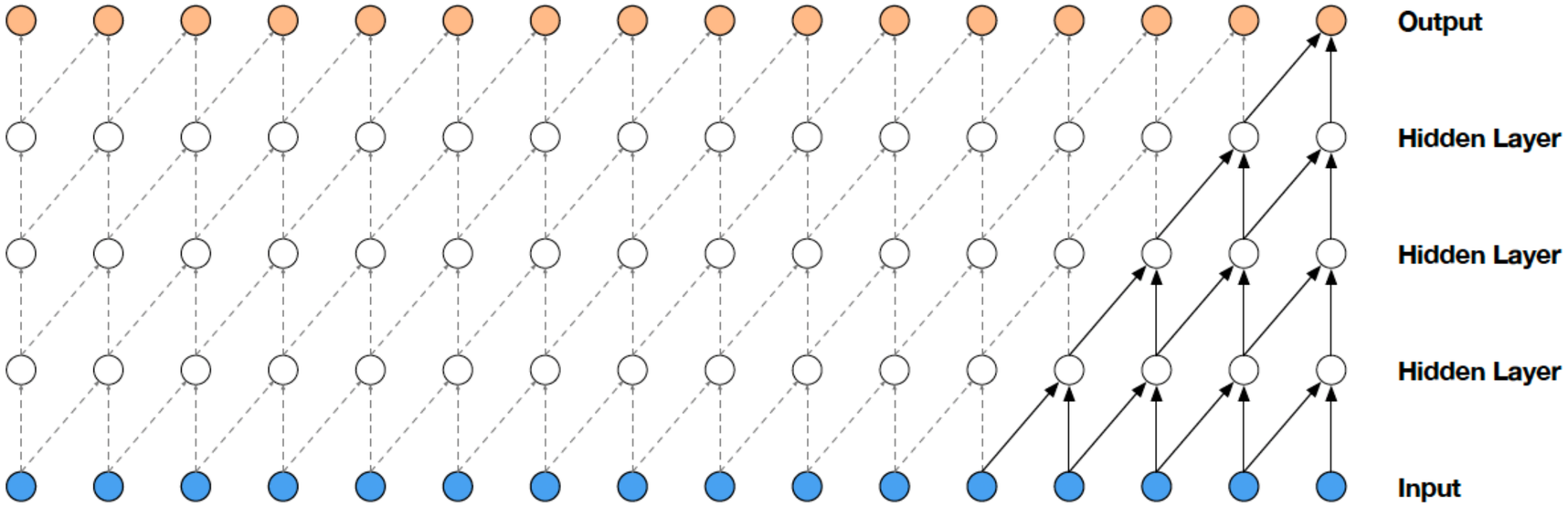
Analog to digital converter measures it all at regular intervals



represent each data point (Delta-Sigma modulation) with 16 bits, ≥44000 rec. per sec.

WaveNet

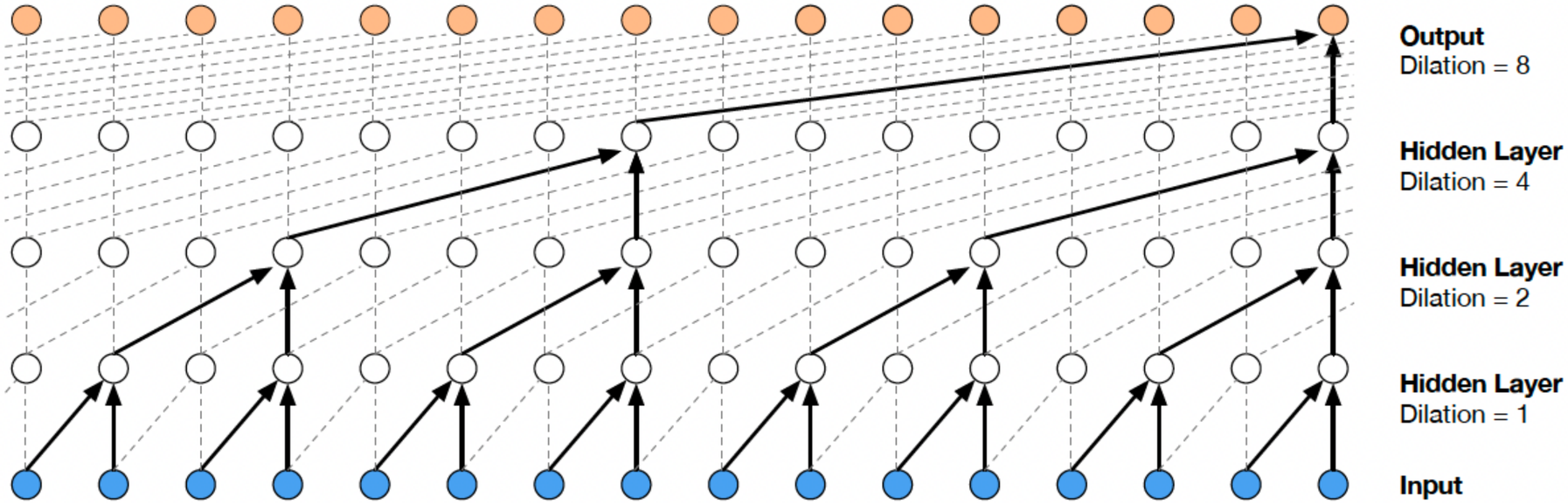
Causal Convolutions



Проблема: нужно много слоев или большая длина фильтра для достаточно большого рецептивного поля

Решение: Dilated Casual Convolutions

Dilated Casual Convolutions



При обучении шаг фильтра удваивается для каждого слоя

Softmax Distributions

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, \ldots, K \text{ and } \mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$$

Проблема: 65,636 **вероятностей для каждого промежутка**

Решение: сжать значения амплитуд:

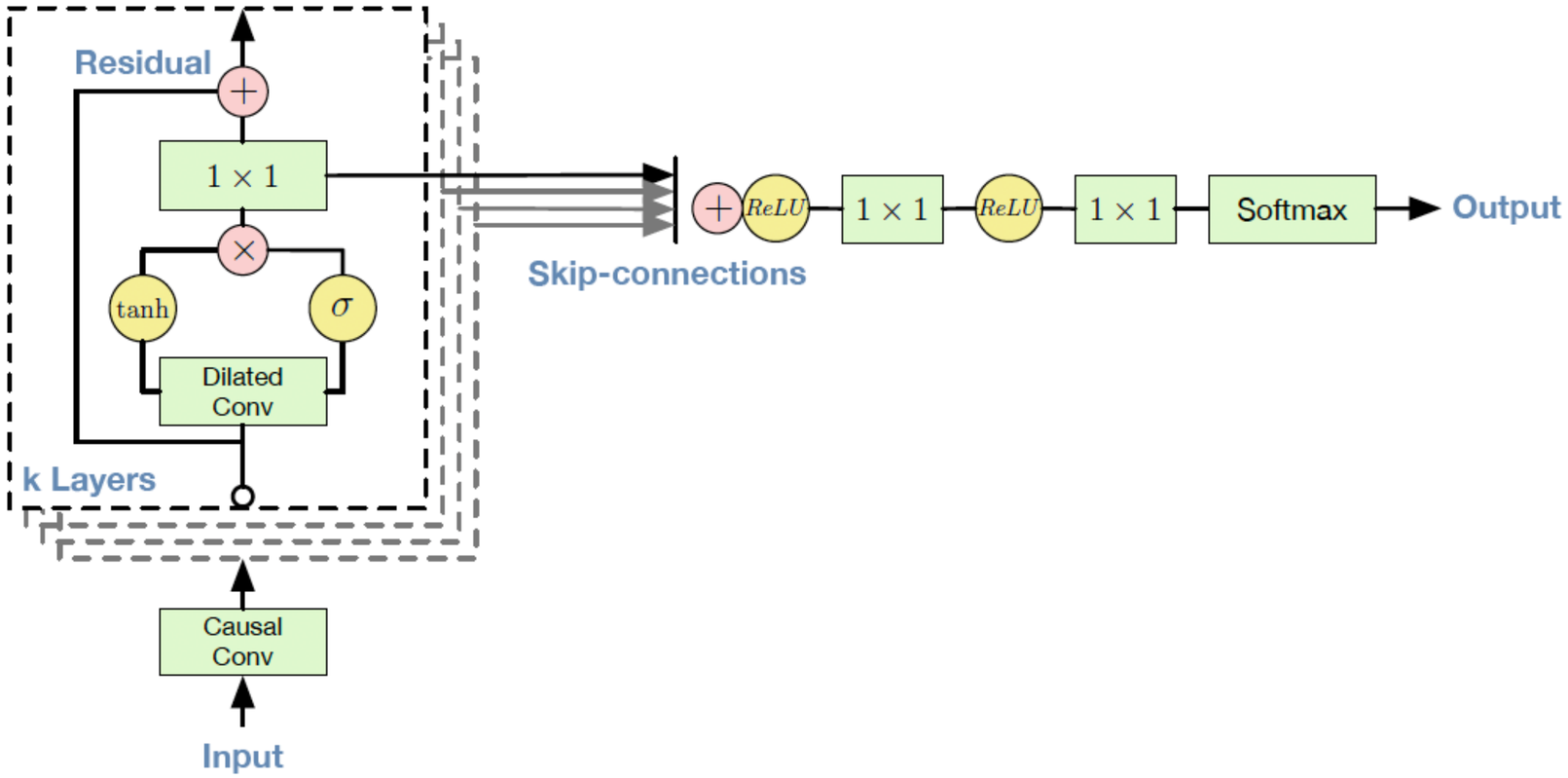
$$f(x_t) = \operatorname{sign}(x_t) \frac{\ln(1 + \mu |x_t|)}{\ln(1 + \mu)} \quad , -1 < x_t < 1 \ , \ \mu = 255$$

Gated Activation Units

обучаемый фильтр

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right)$$

Residual and Skip Connections



А если хотим аудио с определенными характеристиками?

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

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

Эксперименты

Multi-speaker Speech Generation

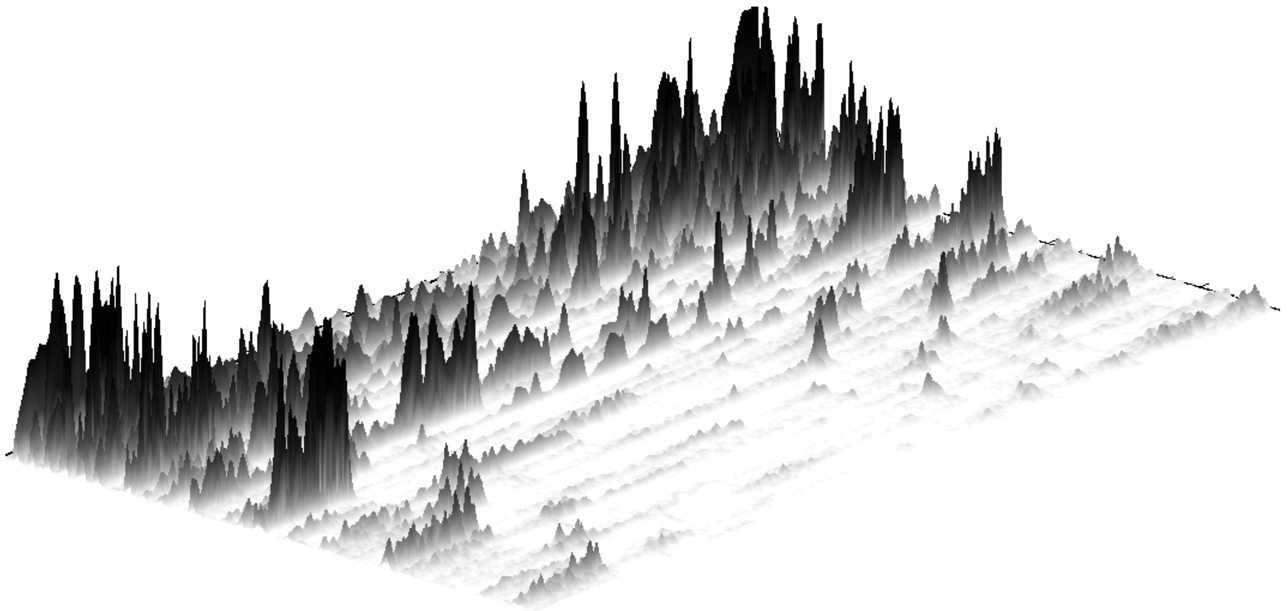
- **Параметризация**: ID **спикера**
- **Генерирует звуки, похожие на человеческие, дополнительные характеристики воспроизводит**
- **Речь несвязная из-за размера рецептивного поля**

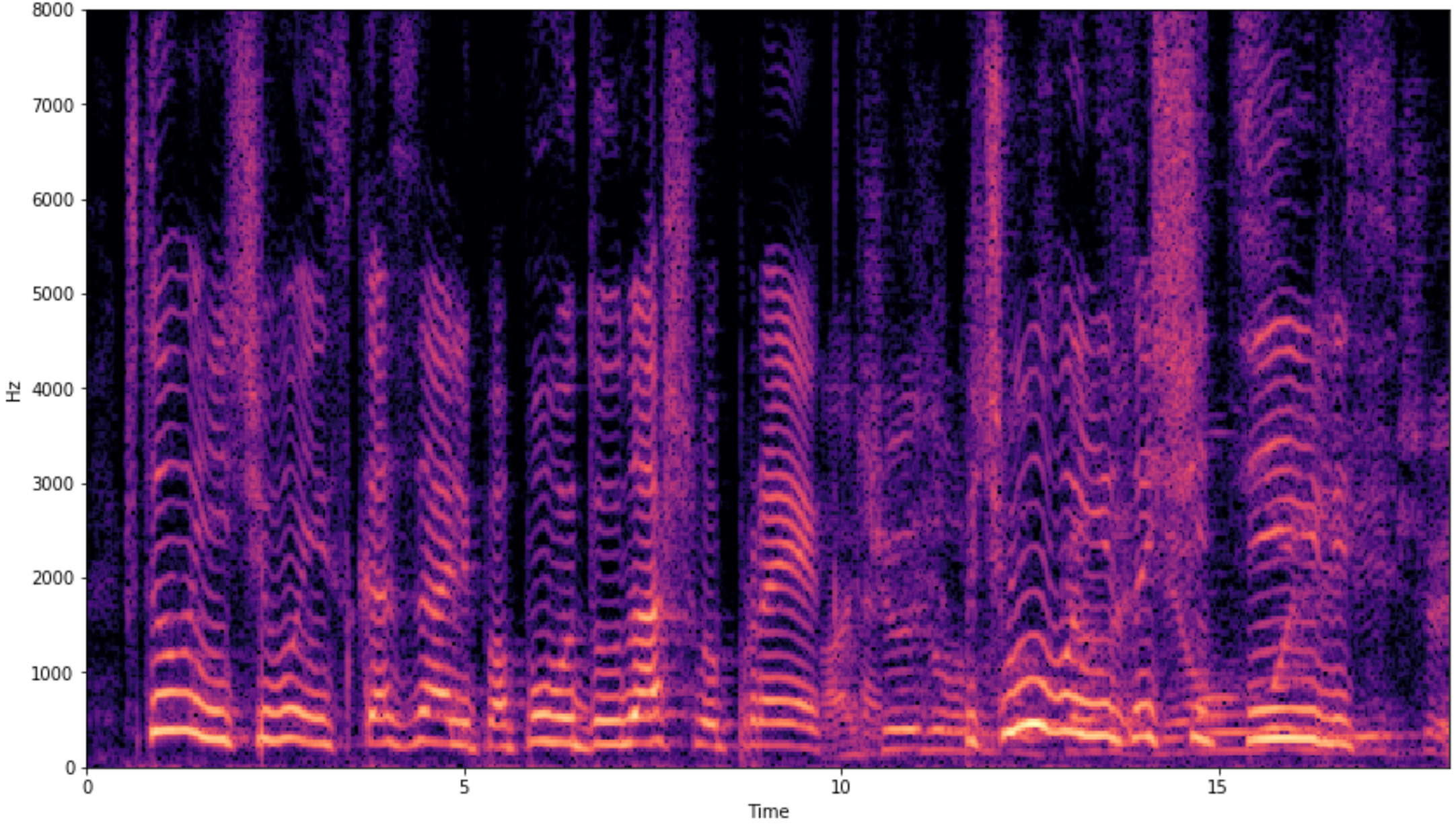
Спектрограмма и мел-спектрограмма

Спектрограмма показывает зависимость амплитуды от времени и частоты

Построение

- Применяется преобразование Фурье для каждого фрагмента звукового сигнала
- По-сути, берет некоторый входной сигнал и выводит соответствующие веса всех частот, составляющих сигнал



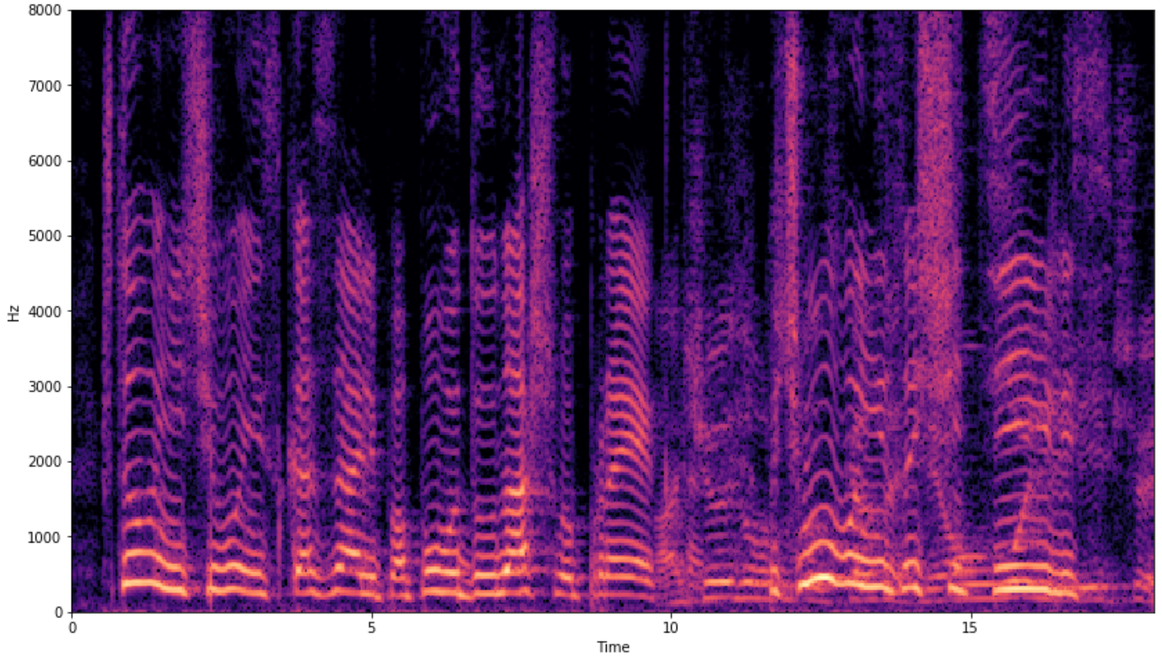

8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
5
10
15
Time

Проблема!

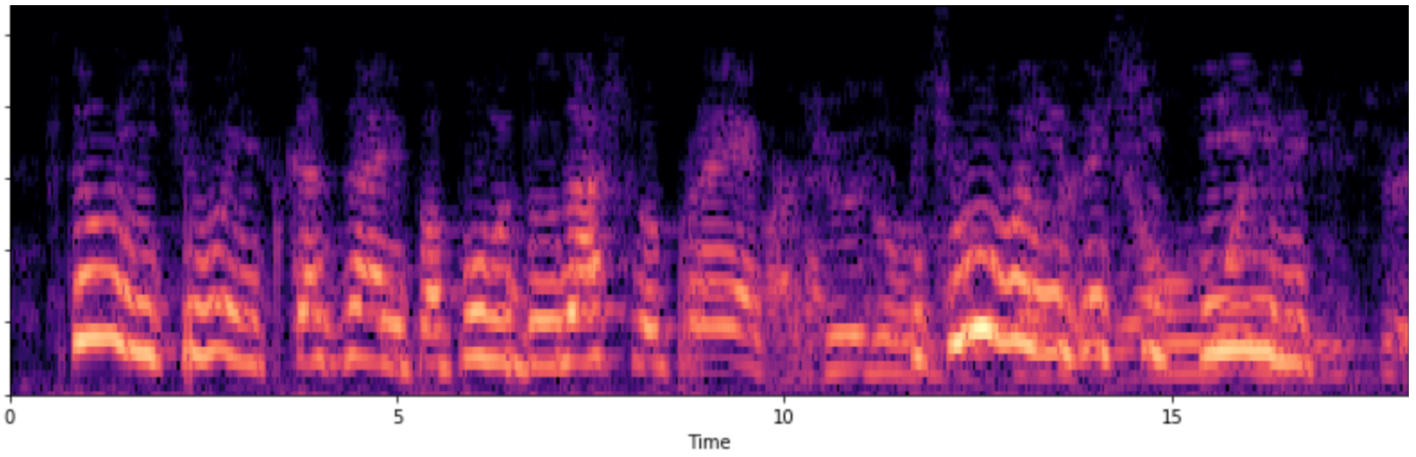
Разные частоты воспринимаются человеком по-разному
Хотим, чтобы одинаковые изменения по шкале одинаково воспринимались для человека

Мел-спектрограмма

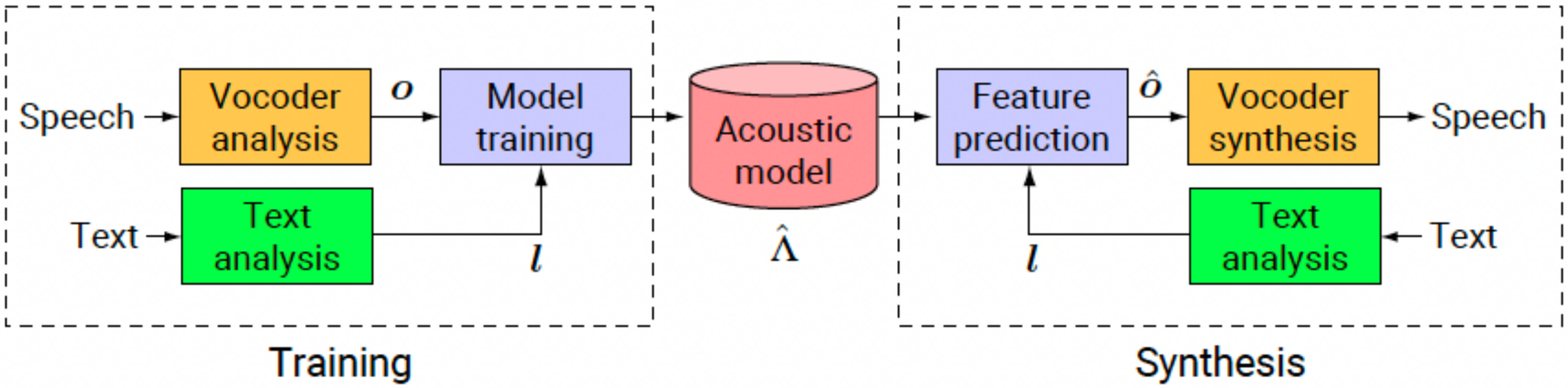
Значения из низких частот спектрограммы остаются практически неизменными на мел-спектре, а в высоких частотах происходит усреднение значений из более широкого диапазона



$$mel = 1127.01048 \ln(1 + \frac{freq}{700})$$



Text-To-Speech



Music

Speech recognition

Итоги

Конец?