

Генеративные модели в машинальном обучении

Лекция 7
Синтез речи (Text-To-Speech)

Михаил Гущин

mhushchyn@hse.ru

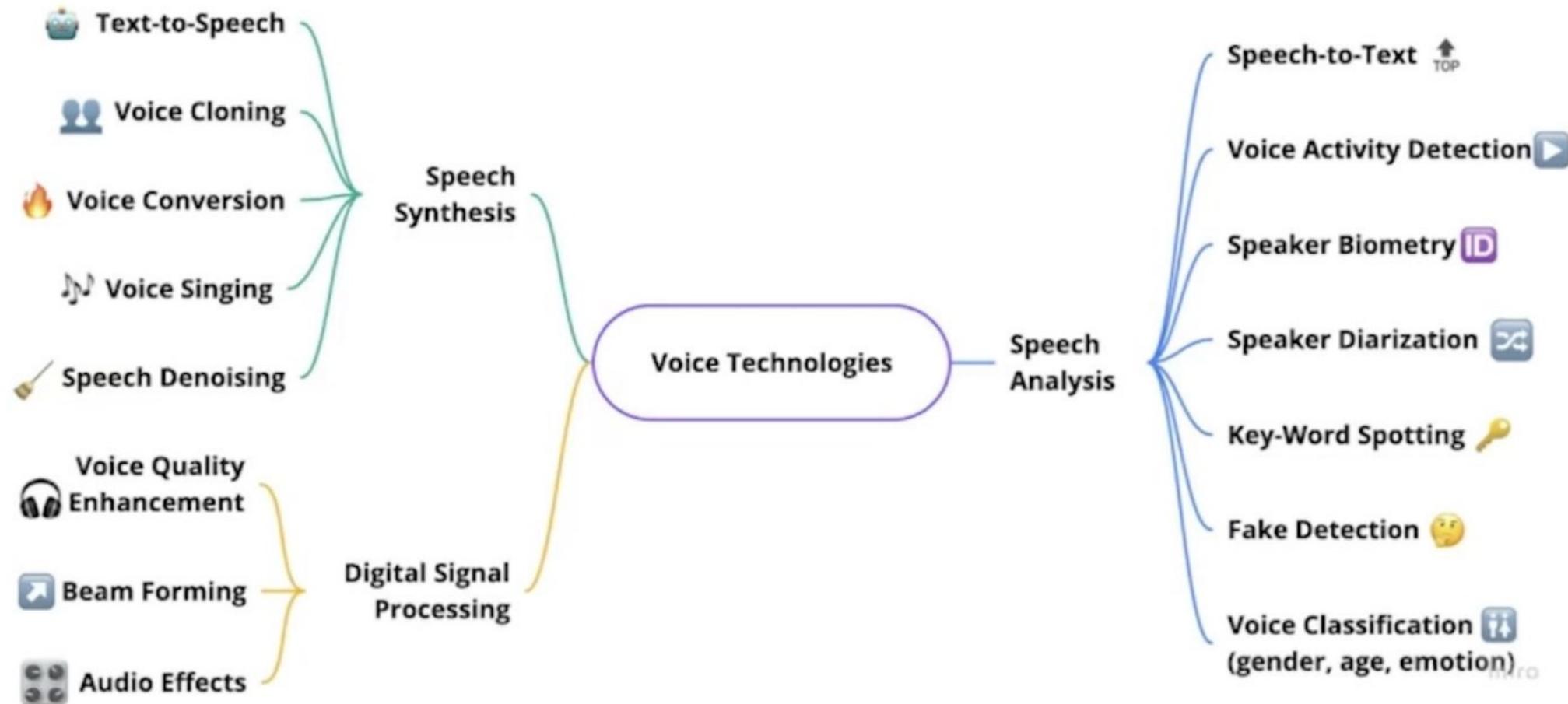
НИУ ВШЭ, 2024



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

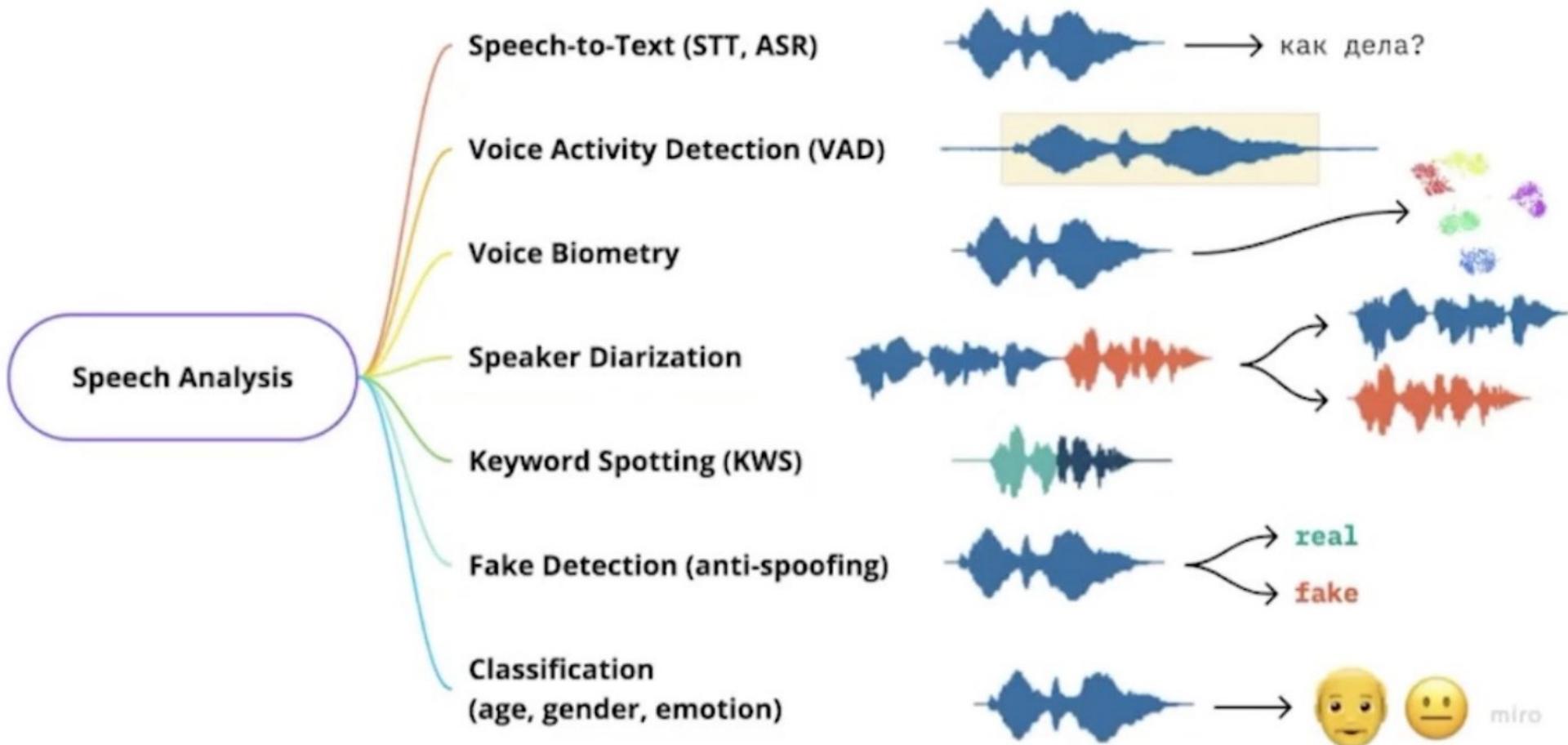
Голосовые технологии

Голосовые технологии



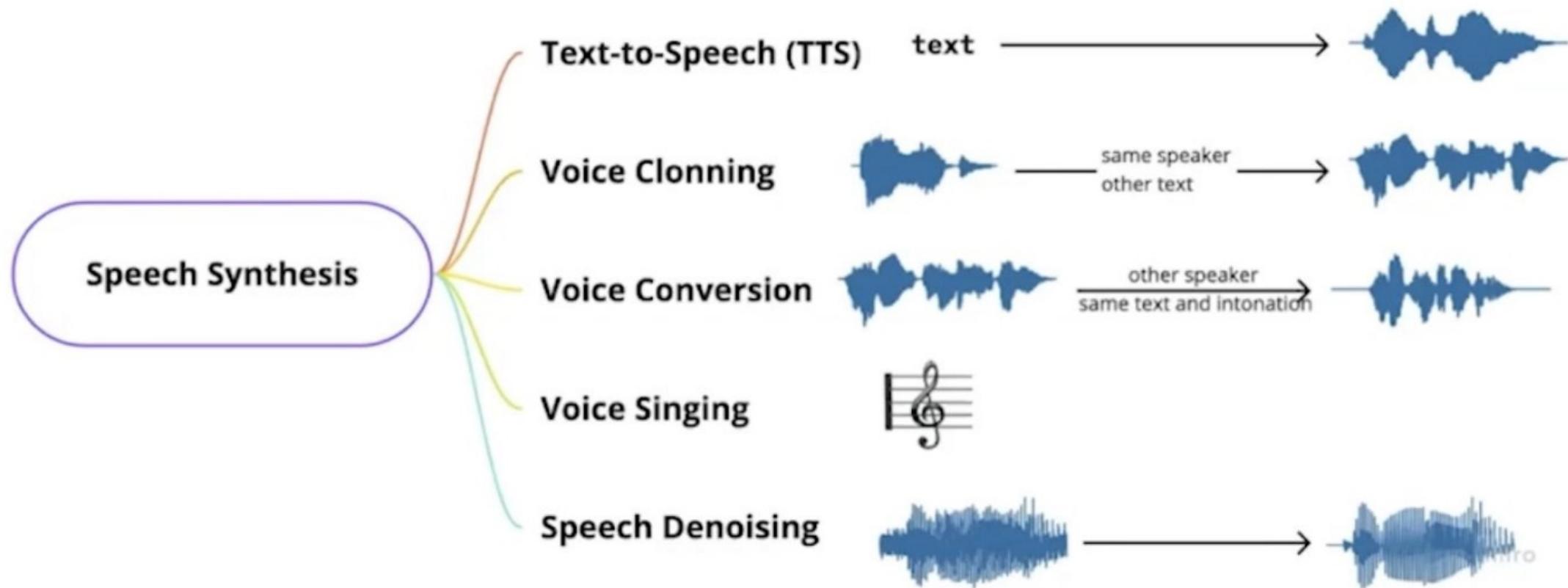
Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

Голосовые технологии



Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

Голосовые технологии



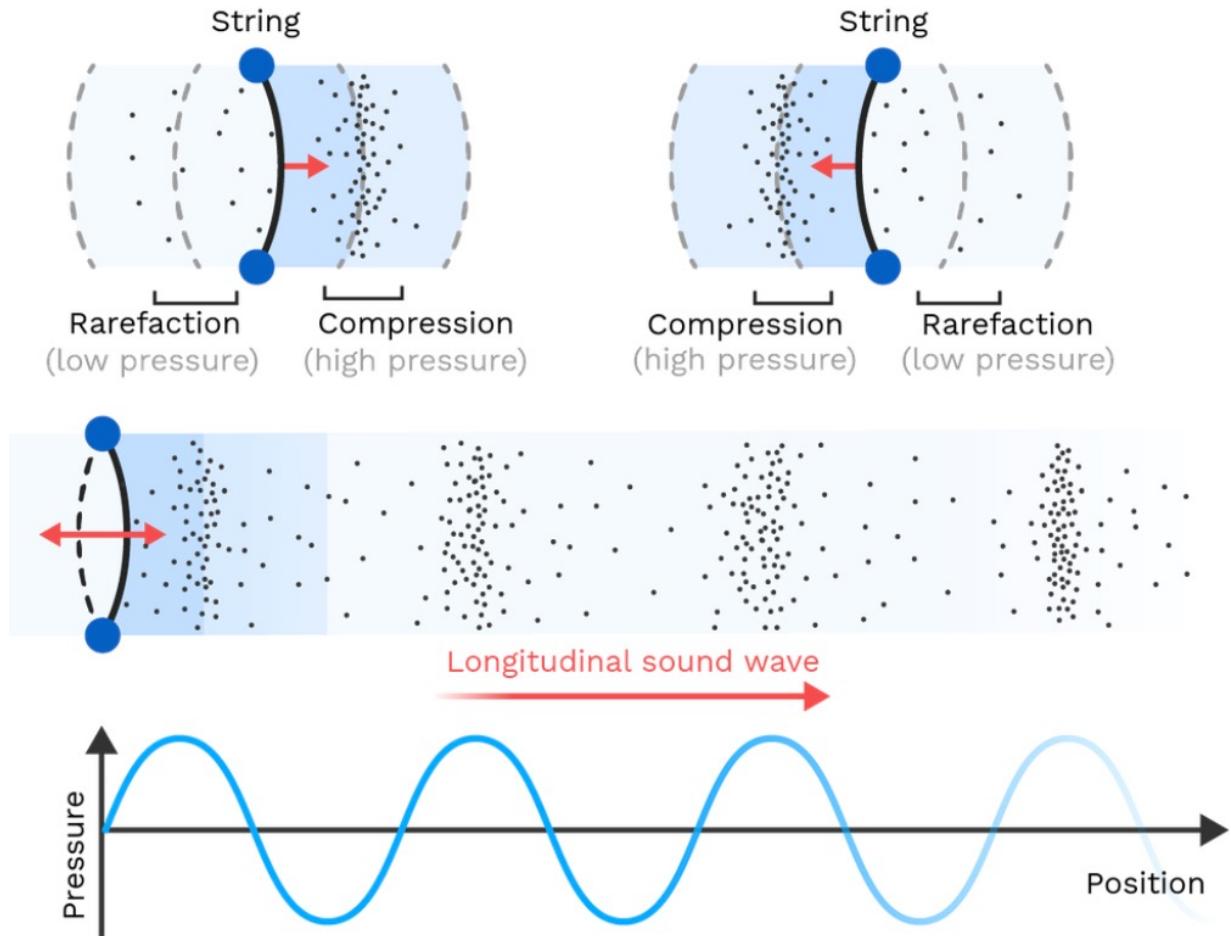
Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

A close-up photograph of a DJ mixer. In the foreground, a pair of white over-ear headphones is resting on the mixer's surface. The mixer features several illuminated buttons in yellow and green, and two black faders with white markings. The background is dark, with blurred lights and other equipment visible, creating a typical night club atmosphere.

Что такое звук

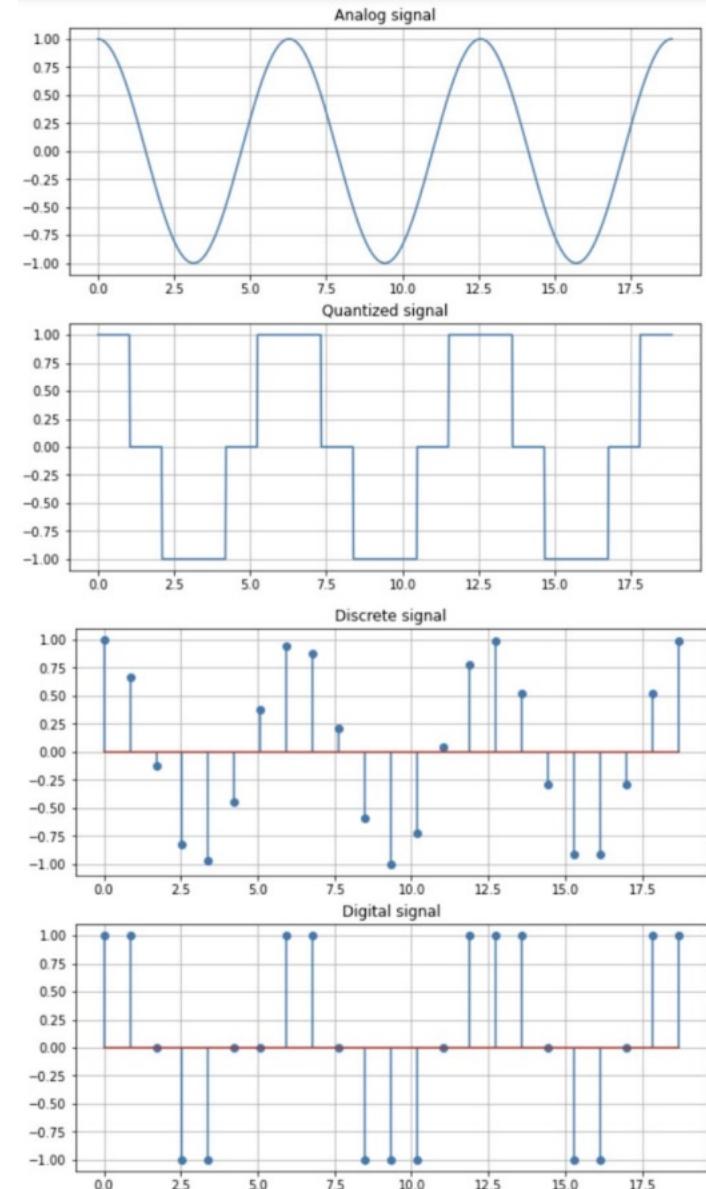
Звук

- ▶ Звуковая волна — это колебания, вызванные движением энергии, проходящей через воздух
- ▶ Микрофон улавливает эти колебания воздуха и преобразует их в электрические колебания
- ▶ Эти колебания преобразуются в аналоговый сигнал



Звуковые сигналы

- ▶ Аналоговый сигнал дискретизируется, квантуется и кодируется
- ▶ Аналоговый сигнал дискретизируется в том смысле, что сигнал представлен в виде последовательности значений, взятых в дискретные моменты времени t с шагом d
- ▶ Квантование сигнала заключается в разбиении диапазона значений сигнала на N уровней с шагом d и выборе для каждой ссылки уровня, который ему соответствует
- ▶ Кодирование сигнала — это всего лишь способ представления сигнала в более компактной форме



Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

Характеристики сигнала

- ▶ Sample rate (SR) - number of audio samples per one second (e.g. 8 kHz, 22.05 kHz, 44.1 kHz)
- ▶ Sample size - number of bits per one sample (e.g. 8, 16, 25, 32 bits)
- ▶ Number of channels -- how many signals we record in parallel (e.g. mono(1), stereo(2))

8000 Hz

The international [G.711](#) standard for audio used in telephony uses a sample rate of 8000 Hz (8 kHz). This is enough for human speech to be comprehensible.

44100 Hz

The 44.1 kHz sample rate is used for compact disc (CD) audio. CDs provide uncompressed 16-bit stereo sound at 44.1 kHz. Computer audio also frequently uses this frequency by default.

48000 Hz

The audio on DVD is recorded at 48 kHz. This is also often used for computer audio.

96000 Hz

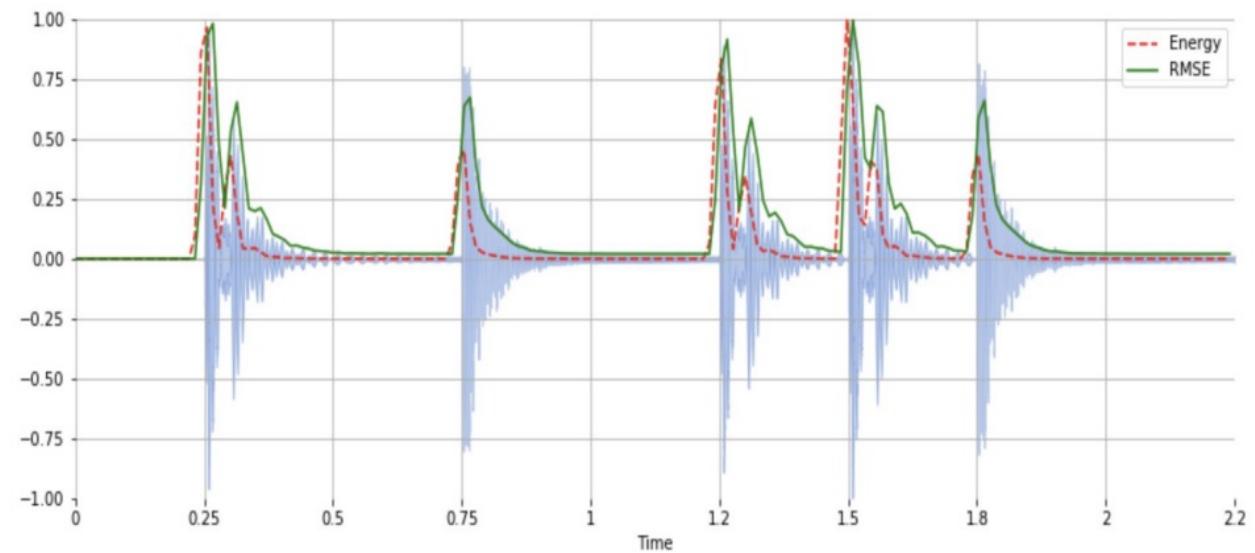
High-resolution audio.

192000 Hz

Ultra-high resolution audio. Not commonly used yet, but this will change over time.

Характеристики сигнала

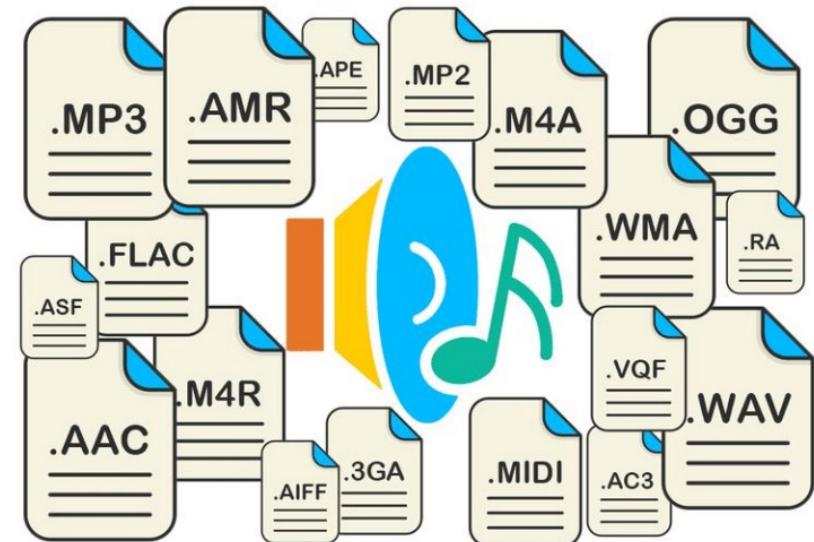
- Assume $f(n)$ is our signal where n is time
- Power of signal is $f^2(n)$
- Energy of signal is $\sum f^2(n)$
- In practice estimated by some **window**
- Energy in **decibels**: $10 \log_{10} E$
- $\text{SNR}_{dB} = 10 \log_{10} \frac{E_{\text{signal}}}{E_{\text{noise}}}$



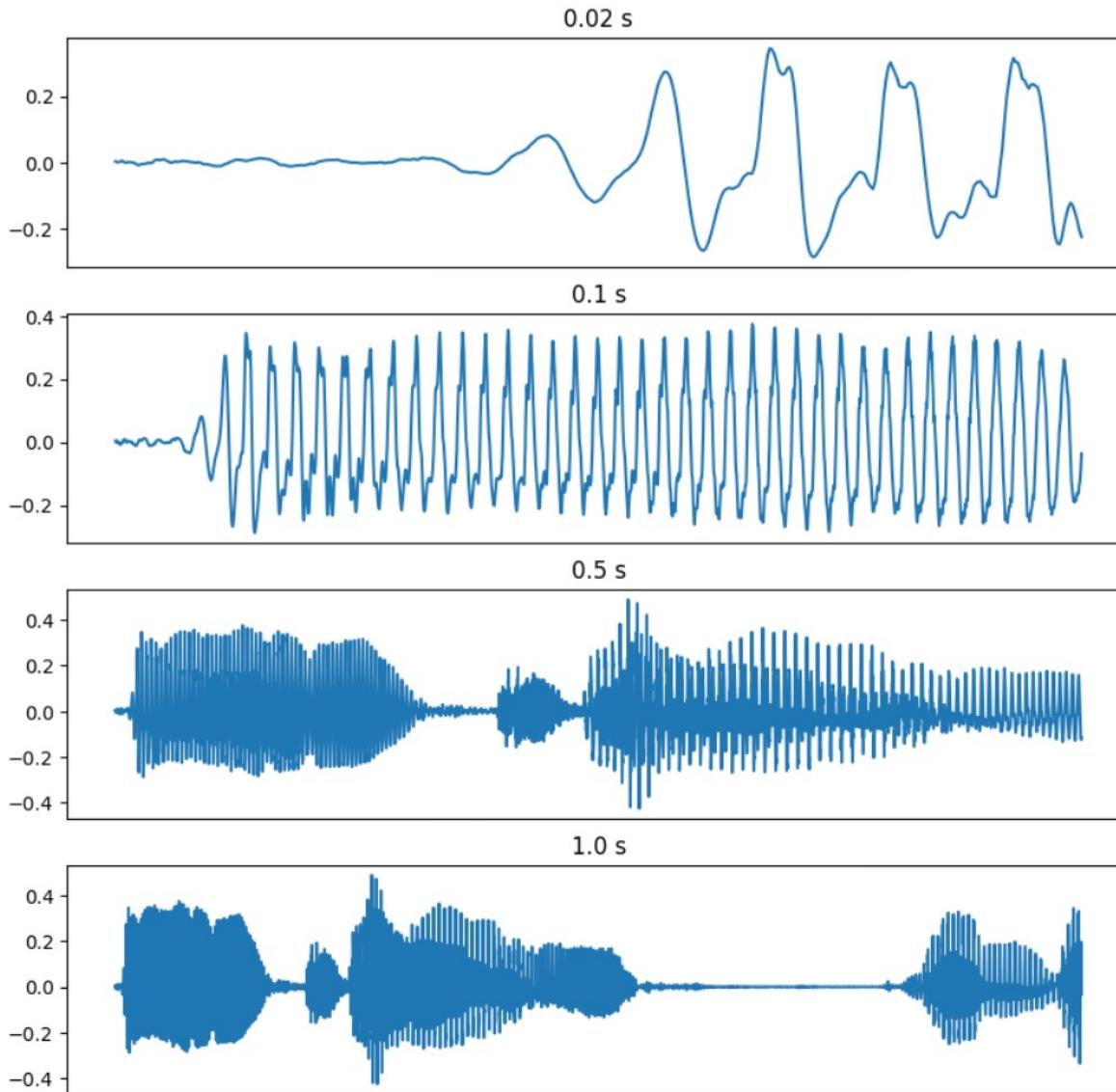
Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

Форматы данных

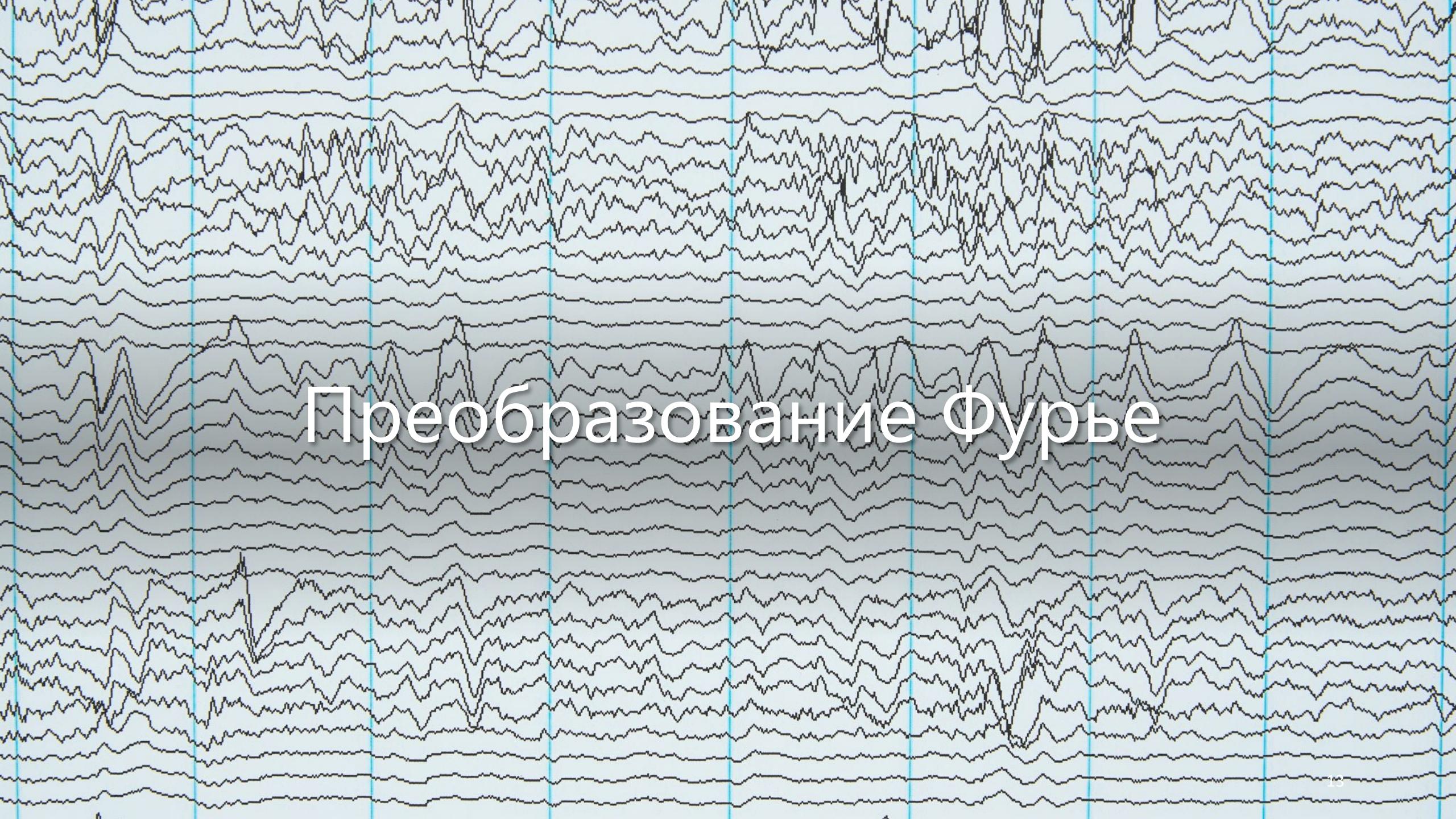
- ▶ Несжатые форматы: WAV, AIFF и т. д.
- ▶ Сжатие без потерь (2:1): FLAC, ALAC и т. д.
- ▶ Сжатие с потерями (10:1): MP3, Opus и т. д.
- ▶ Скорость передачи данных измеряет степень сжатия. Количество бит, которые передаются или обрабатываются за единицу времени.



Пример сигнала (waveforms)



Источник: https://github.com/yandexdataschool/speech_course



Преобразование Фурье

Преобразование Фурье

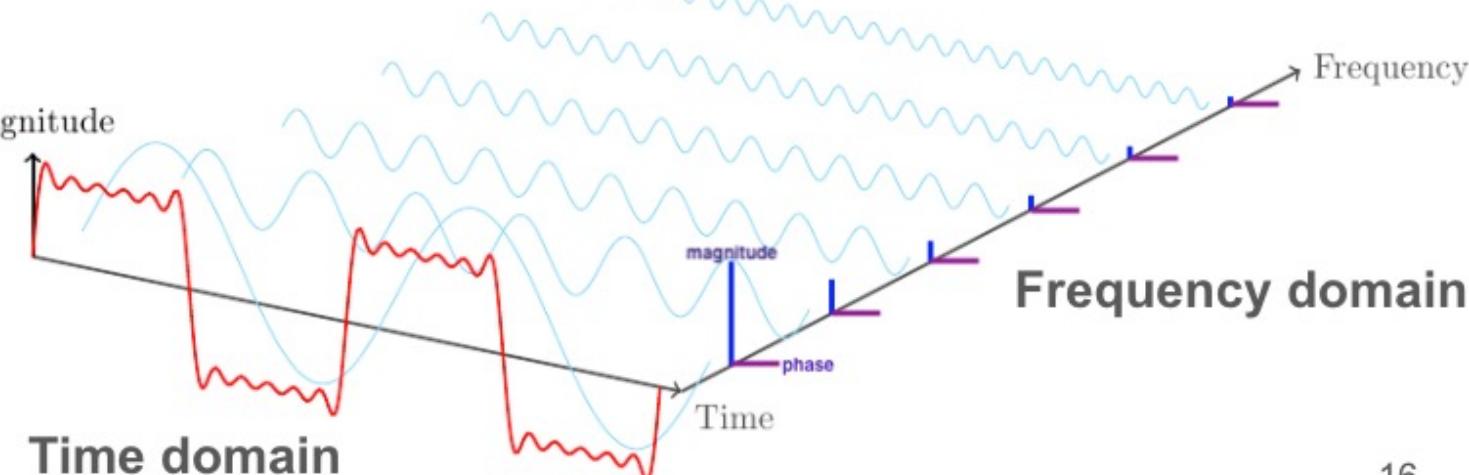
Any absolutely integrable **periodic** function $x(t)$ with periodicity T can be represented as

$$x(t) = A_0 + \sum_{n=1}^{\infty} \left(A_n \cos \left(\frac{2\pi n t}{T} \right) + B_n \sin \left(\frac{2\pi n t}{T} \right) \right) =$$

Use formula for cosine of the difference

$$= \frac{A_0}{2} + \sum_{n=1}^{\infty} \boxed{A_n} \cos \left(2\pi \frac{n}{T} t - \boxed{\phi_n} \right)$$

magnitude frequency phase



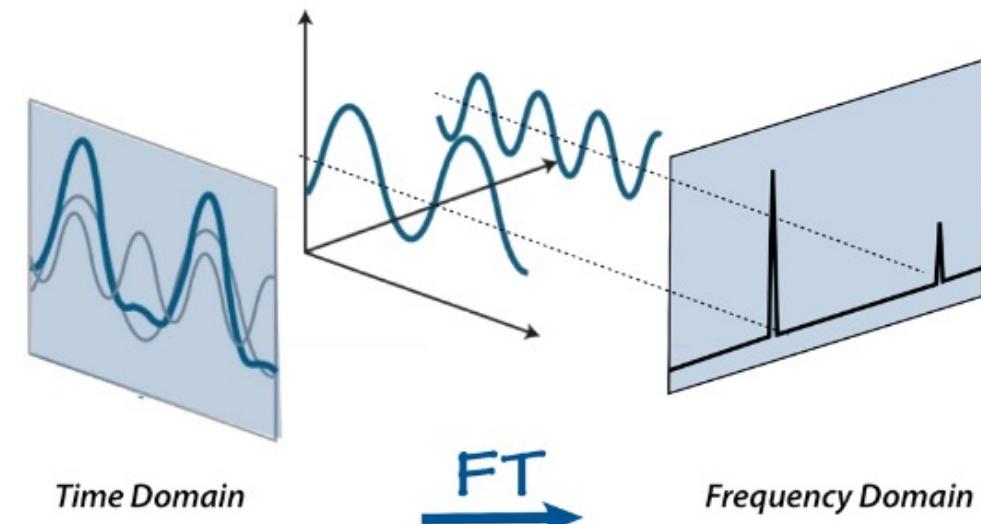
Преобразование Фурье (другая форма)

FT transfers a signal from real-valued function in the **time domain** to a complex-valued function in **frequency domain**:

$$X(f) : \mathbb{R} \rightarrow \mathbb{C}$$

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi i f t} dt$$

frequency original signal



$$X(f) = x + iy = \rho e^{i\phi}, \rho = \sqrt{x^2 + y^2}, \phi = \arctan\left(\frac{y}{x}\right)$$

magnitude phase

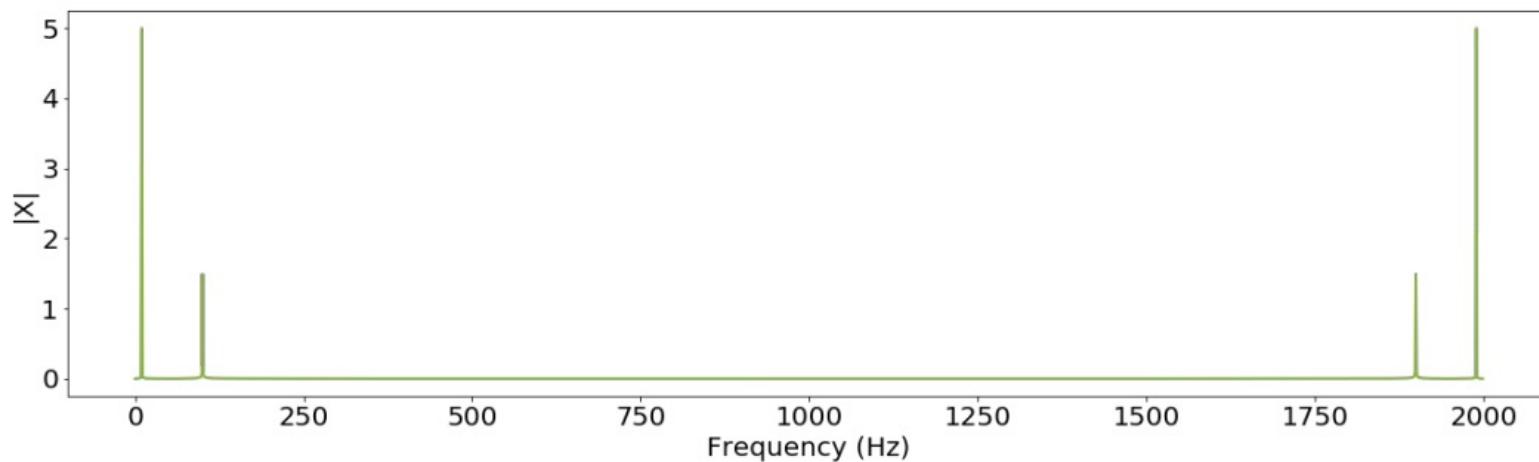
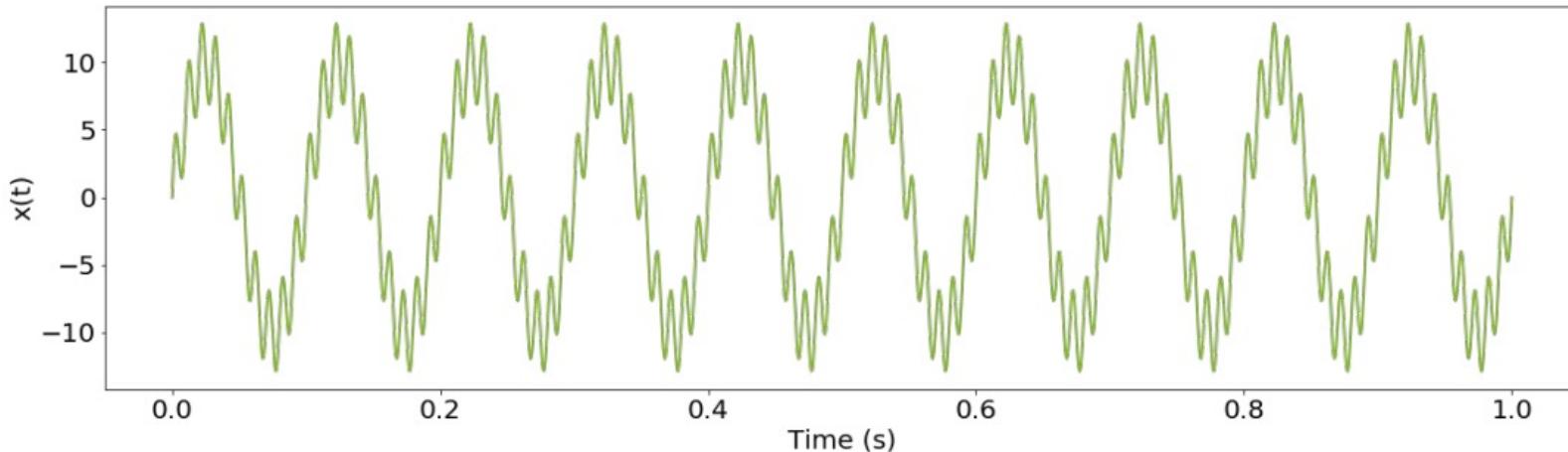
<https://www.youtube.com/watch?v=spUNpyF58BY>

<https://stemporium.blog/2023/04/13/what-is-the-fourier-transform-and-how-is-it-used-in-image-processing/>

Пример

$$F = 2kHz$$

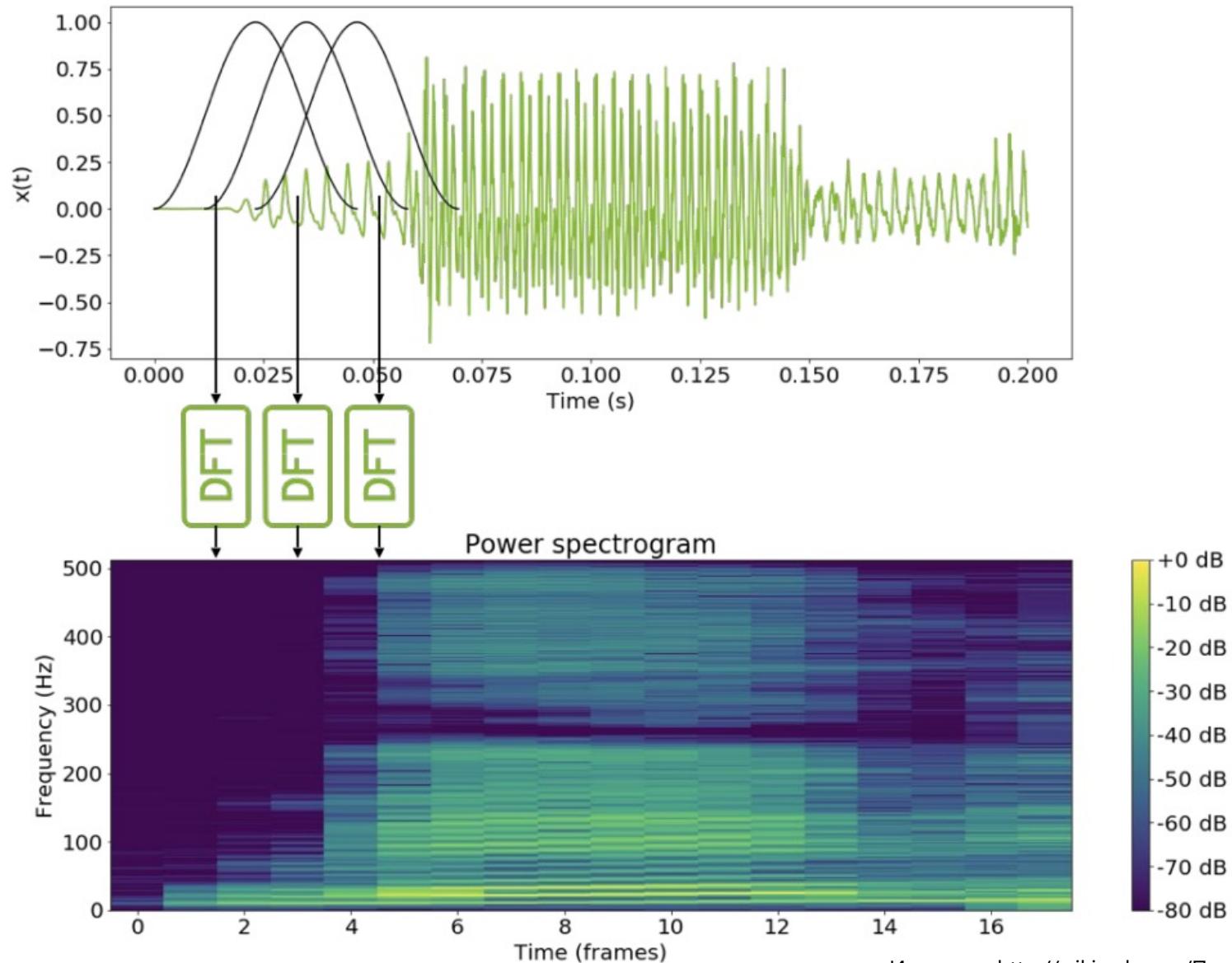
$$f(t) = 10 \sin(2\pi 10t) + 3 \sin(2\pi 100t)$$



Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

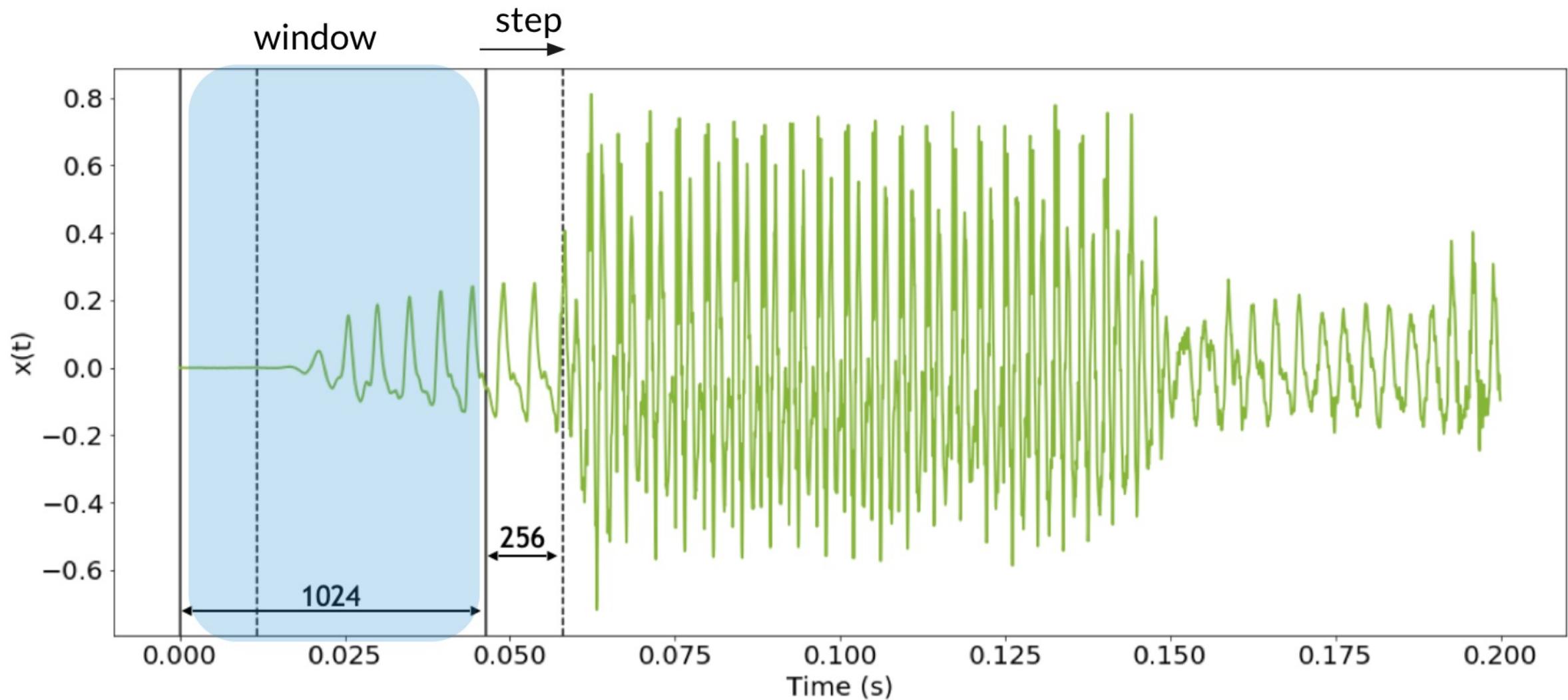
Спектограмма

Спектрограмма



Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

Спектрограмма

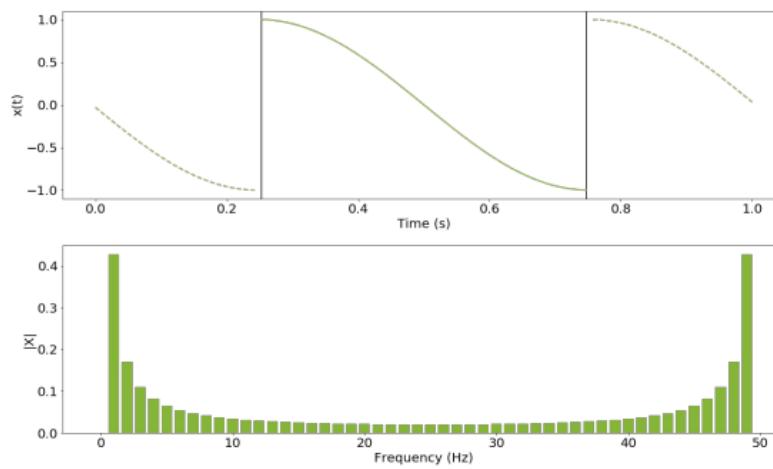


Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

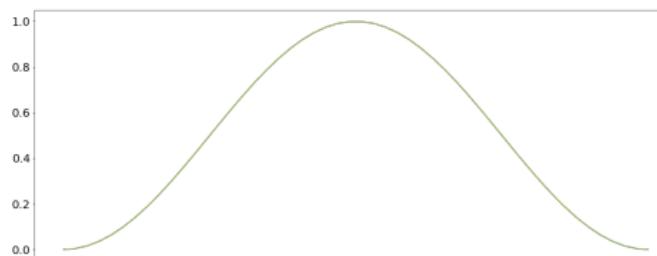
Спектрограмма

FFT + Windowing

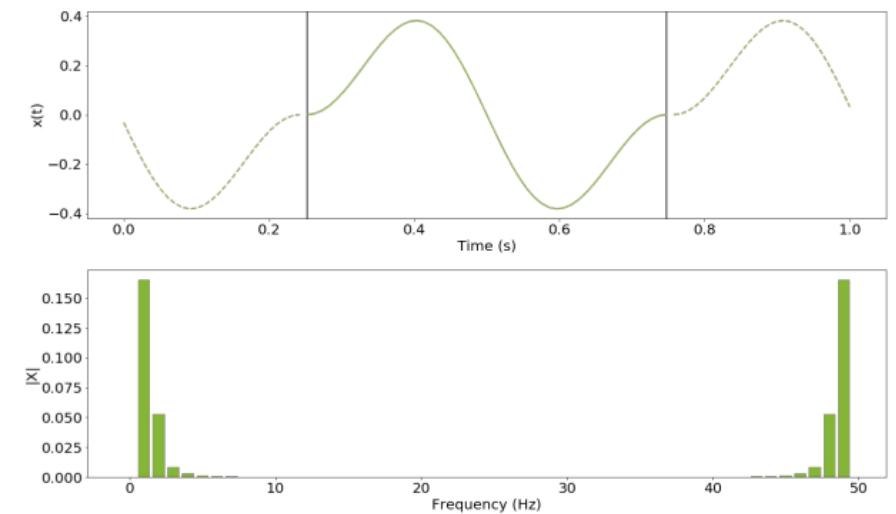
Sliced signal



Window

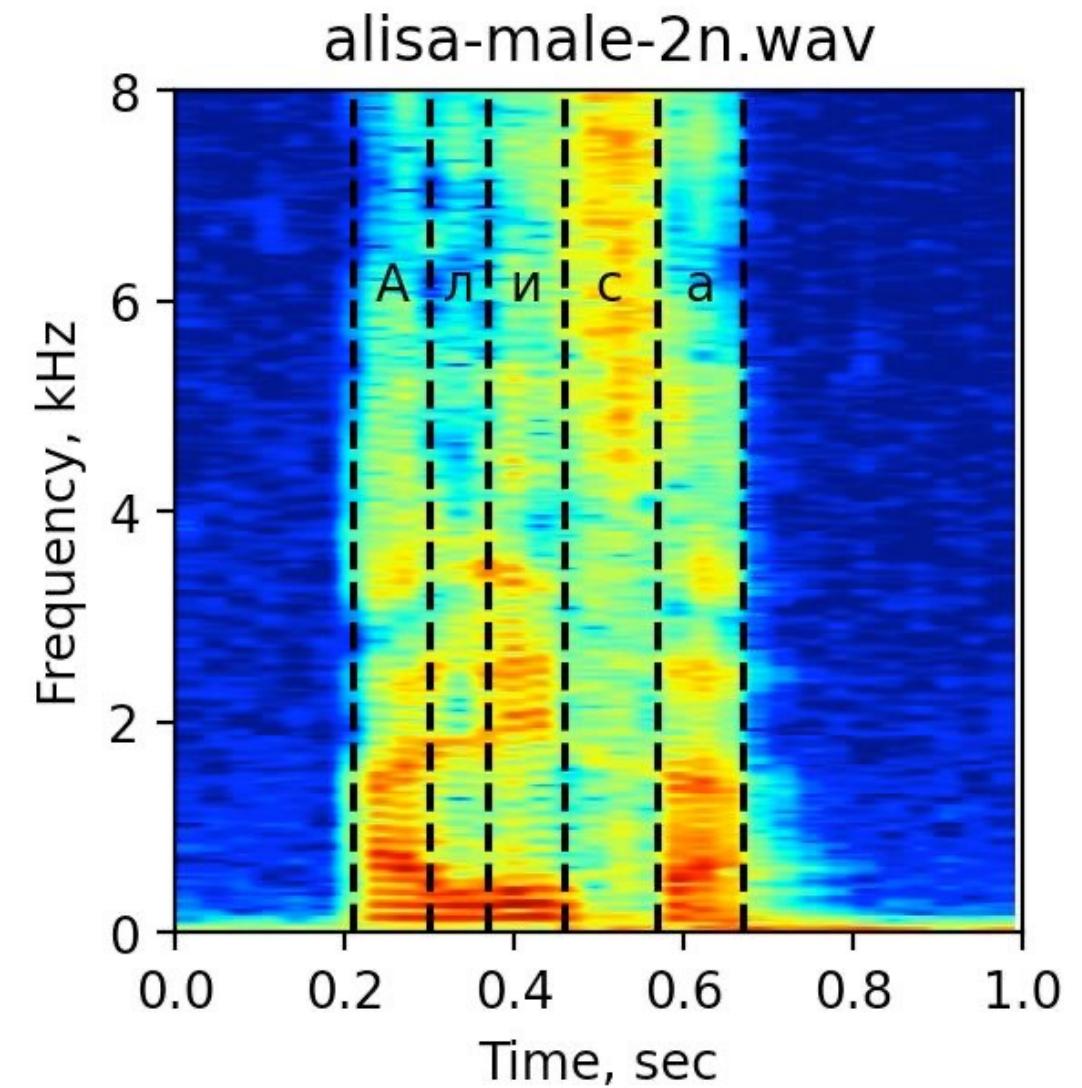
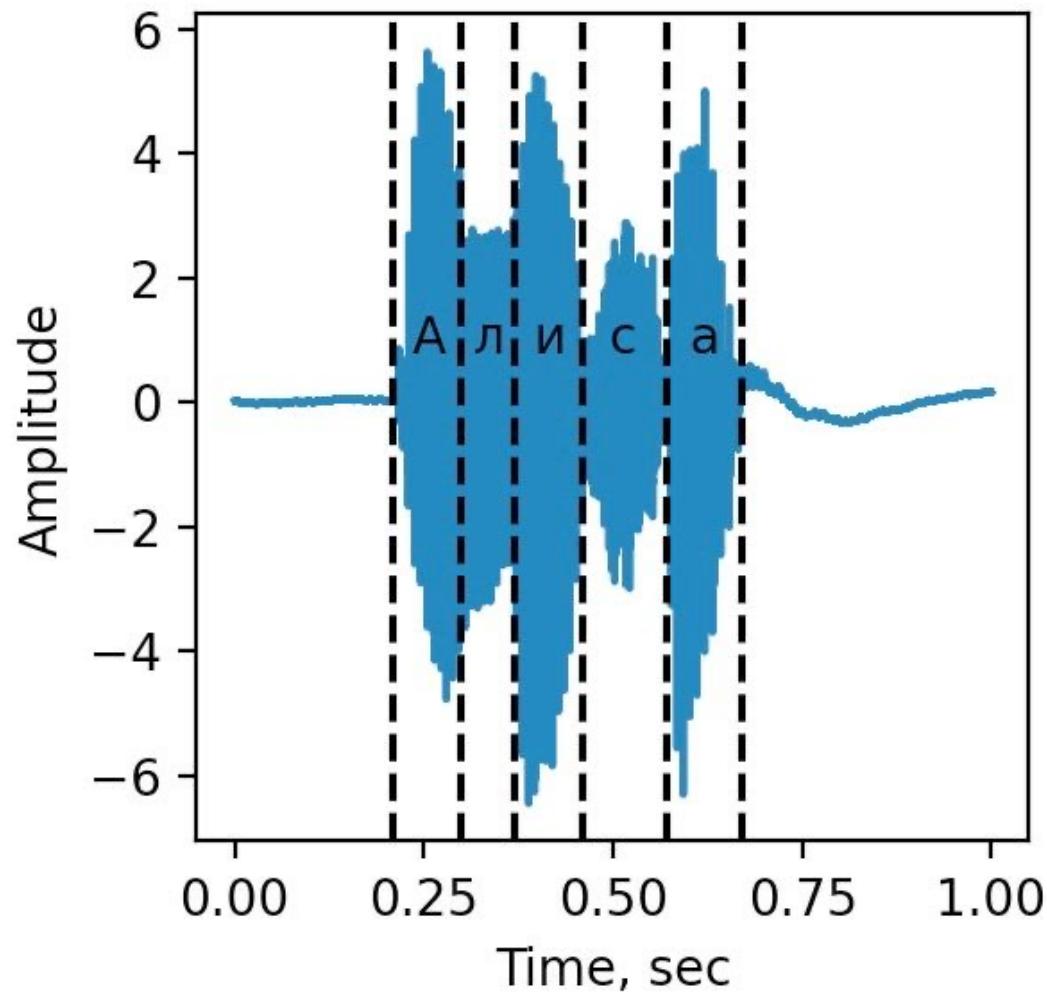


Windowed signal



Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

Пример

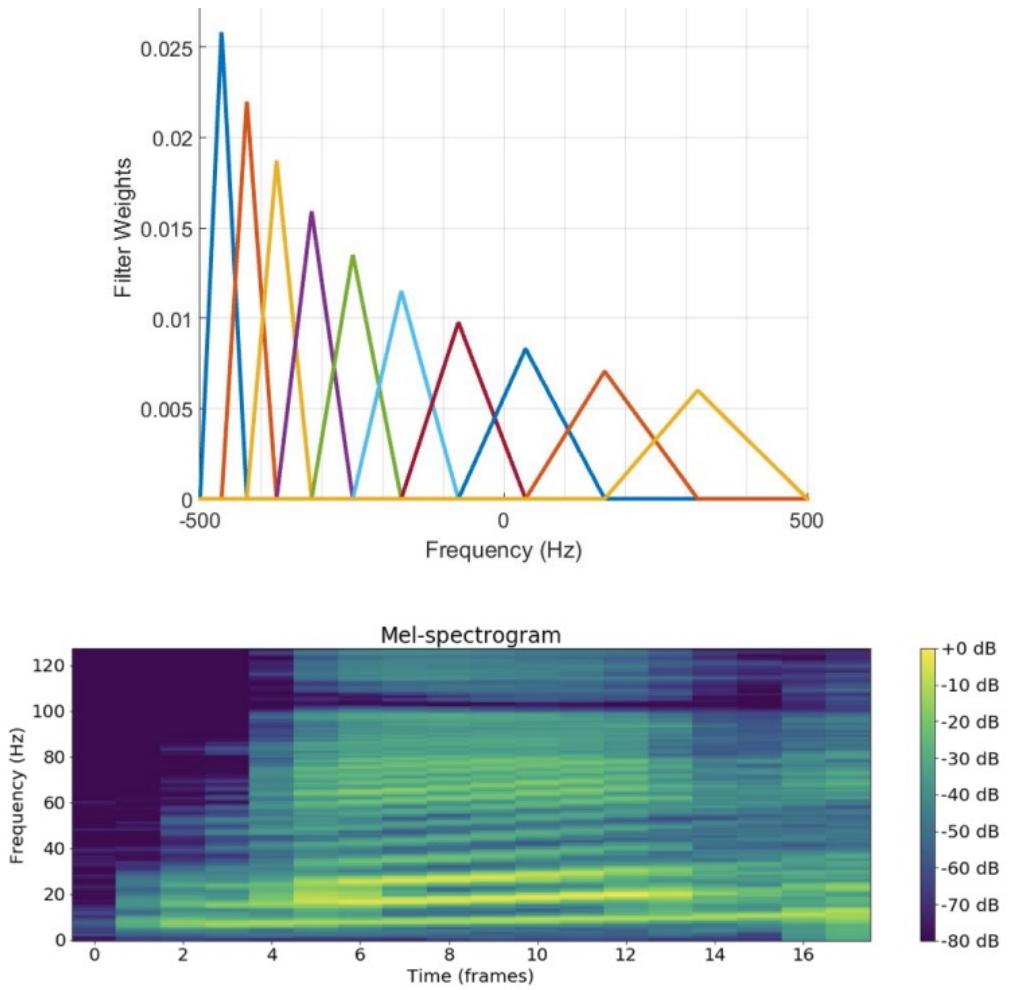
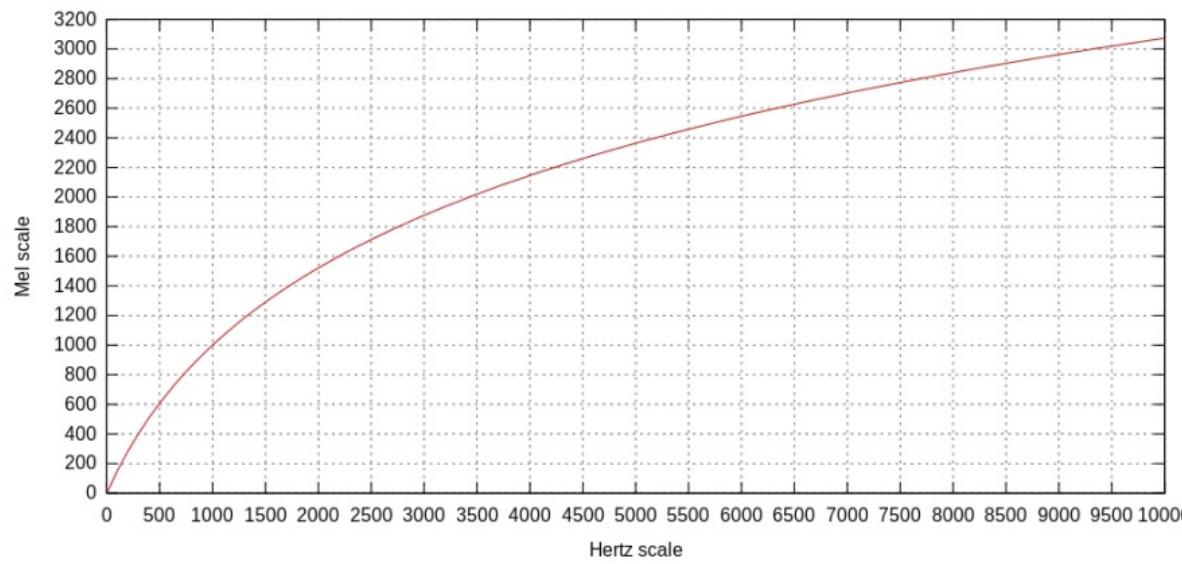


Источник: https://github.com/yandexdataschool/speech_course

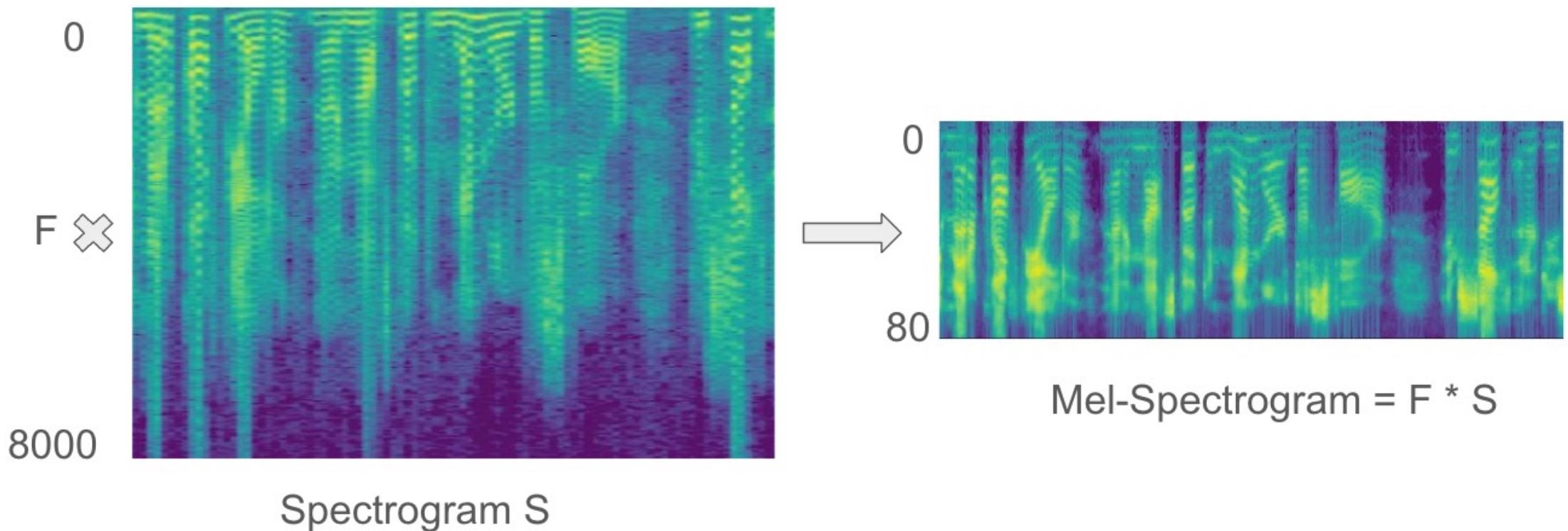
Мел-спектrogramма

$$m = 2595 \log_{10} \left(1 + \frac{f}{700} \right) = 1127 \ln \left(1 + \frac{f}{700} \right)$$

$$f = 700 \left(10^{\frac{m}{2595}} - 1 \right) = 700 \left(e^{\frac{m}{1127}} - 1 \right)$$



Пример



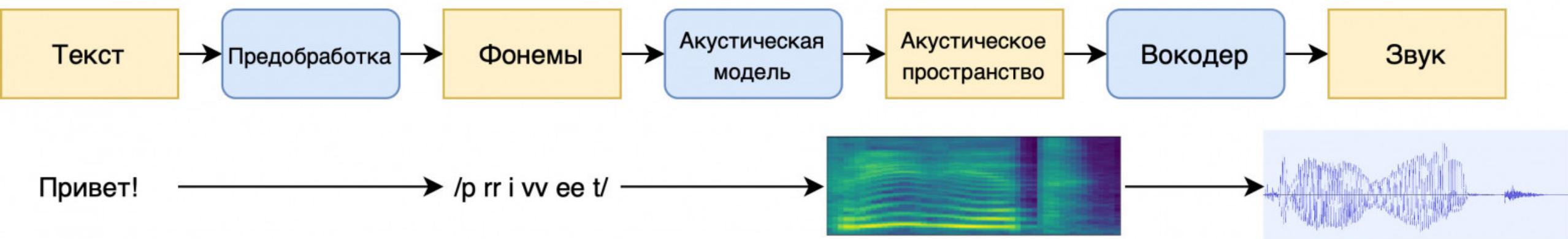


Text To Speech

Задачи TTS

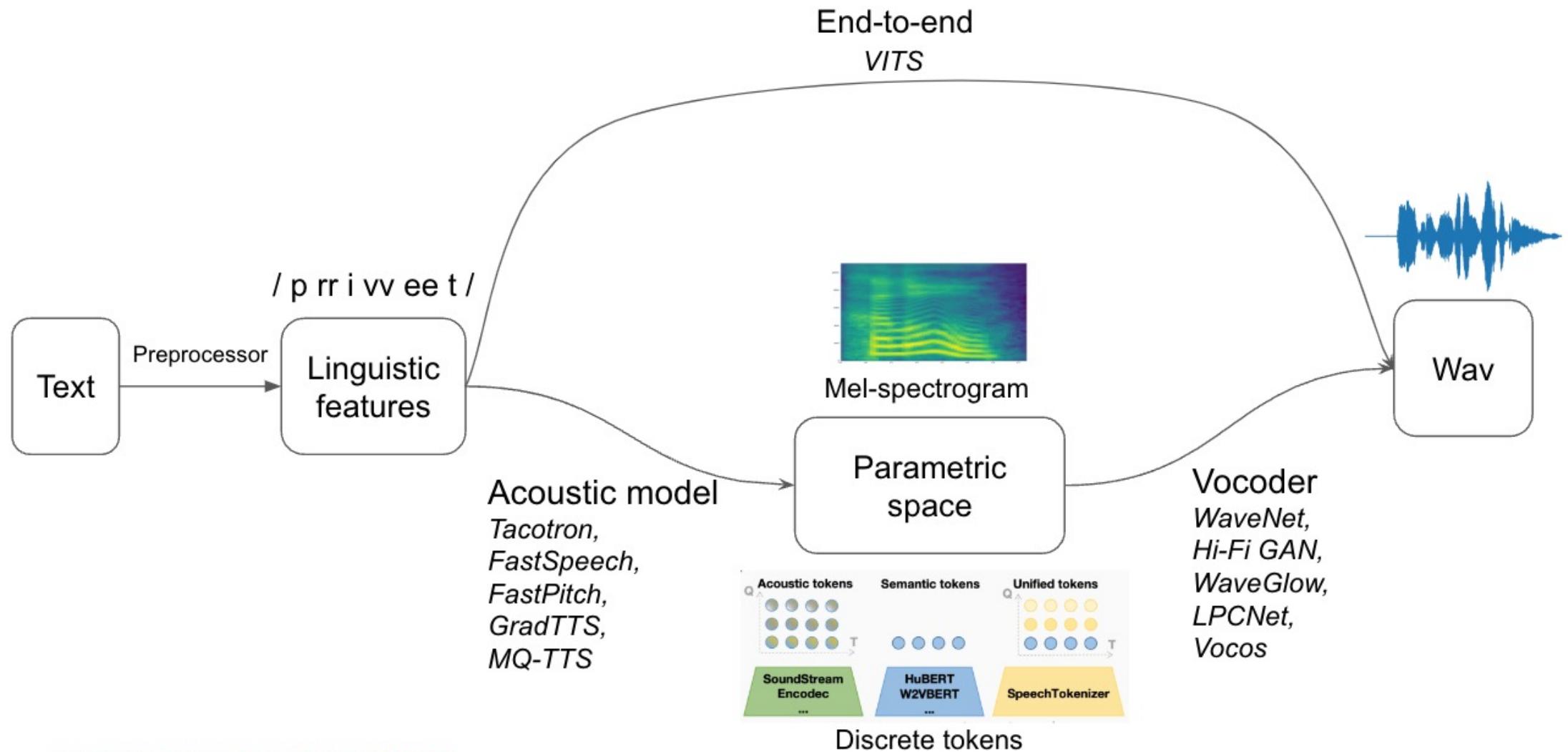
- ▶ По заданному тексту хотим получить озвучку голосом
- ▶ Можем принимать на вход множество доп параметров (голос, стиль,etc)
- ▶ Ожидаемое качество может сильно различаться от задачи:
обзвонщик vs голосовой помощник vs аудиокниги vs озвучка фильмов

Общий подход решения



Источник: https://github.com/yandexdataschool/speech_course

Модели TTS



<https://github.com/ZhangXInFD/SpeechTokenizer>

Источник: https://github.com/yandexdataschool/speech_course

The background of the slide features a dark gray or black surface with a subtle, flowing texture. This texture is composed of numerous thin, light-colored lines that create a sense of depth and motion, resembling waves or ripples on water. The lines are more concentrated in the center and spread out towards the edges.

WaveNet

WaveNet

- ▶ Представлено DeepMind в 2016 году
- ▶ Современный уровень (SOTA) на момент выпуска
- ▶ Генеративная глубокая нейронная сеть для создания необработанных звуковых волн (не обязательно вокодер)
- ▶ Как GPT, только для звука ☺

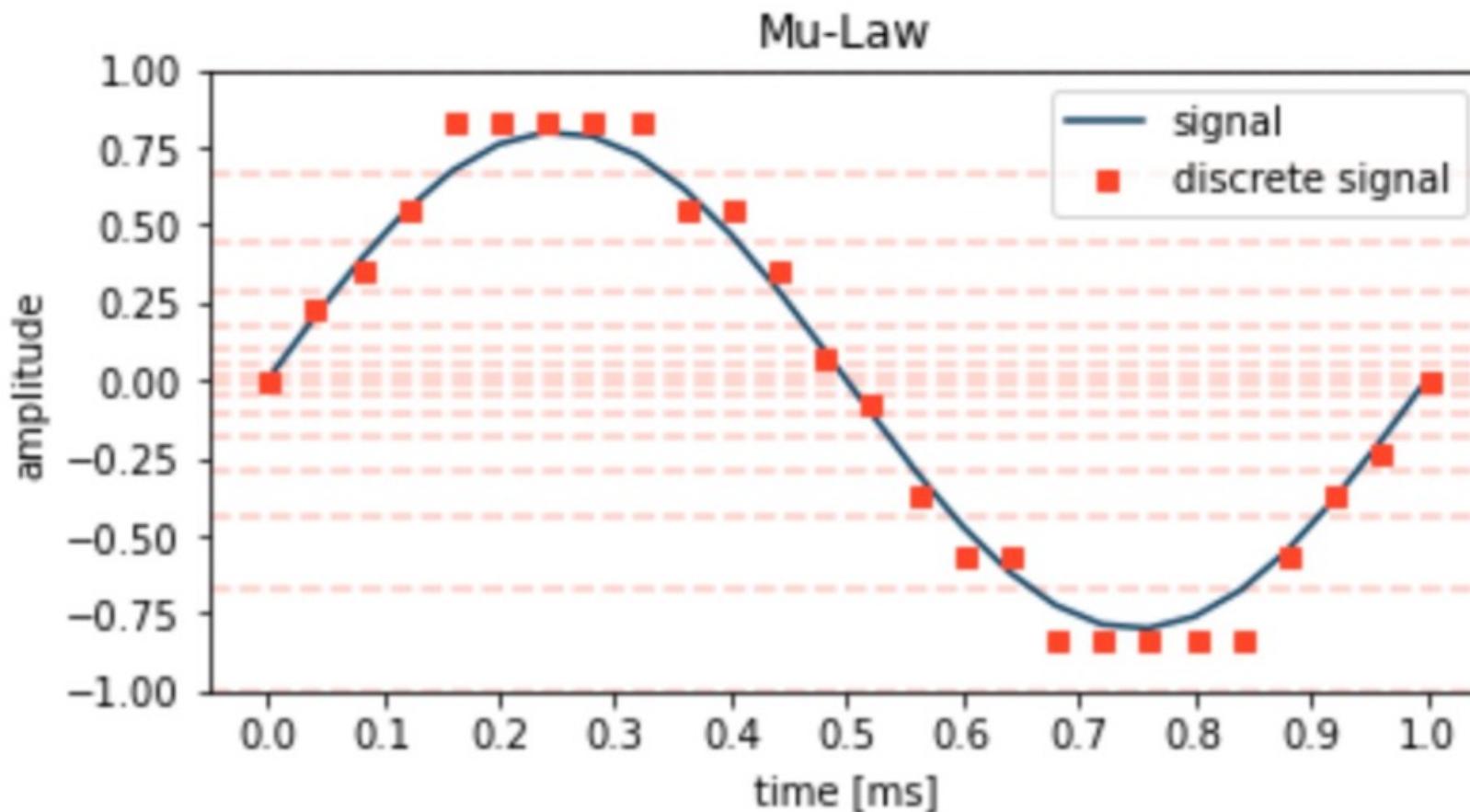
Источник: https://github.com/yandexdataschool/speech_course

Mu-law encoding

- ▶ WaveNet обрабатывает wav как дискретный сигнал: все значения амплитуды квантуются в дискретные ячейки, а затем модель предсказывает номер ячейки.
- ▶ Аудиосэмплы квантуются с использованием кодирования по закону МЮ:
 - более низкие амплитуды дискретизируются чаще, чем более высокие, что разумно для человеческой речи
 - один аудиосэмпл квантуется в 256 ячеек (вместо 65536 ячеек для 16-битного целого числа)

Источник: https://github.com/yandexdataschool/speech_course

Mu-law encoding

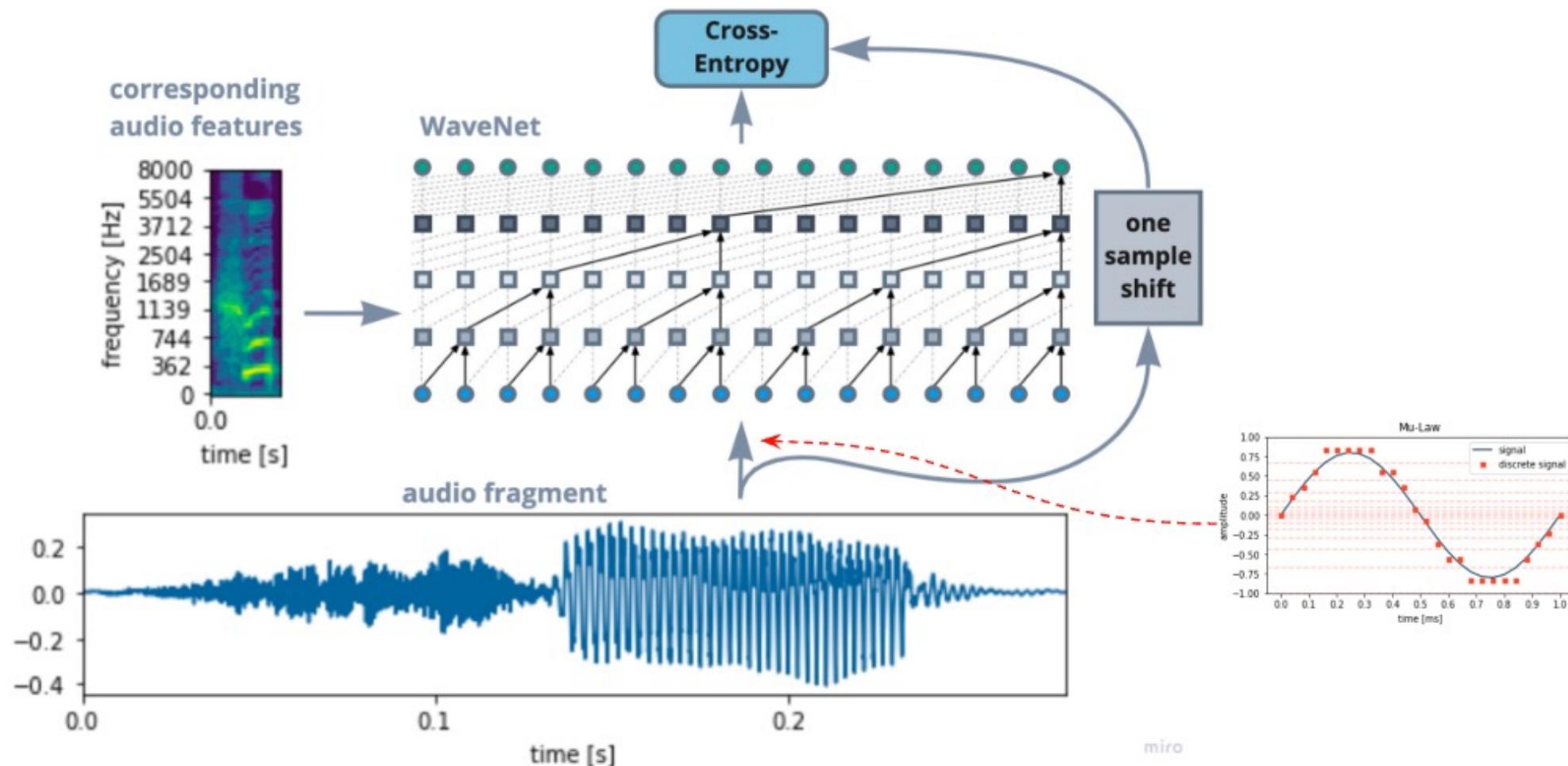


MSE loss → Cross-Entropy loss

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

Источник: https://github.com/yandexdataschool/speech_course

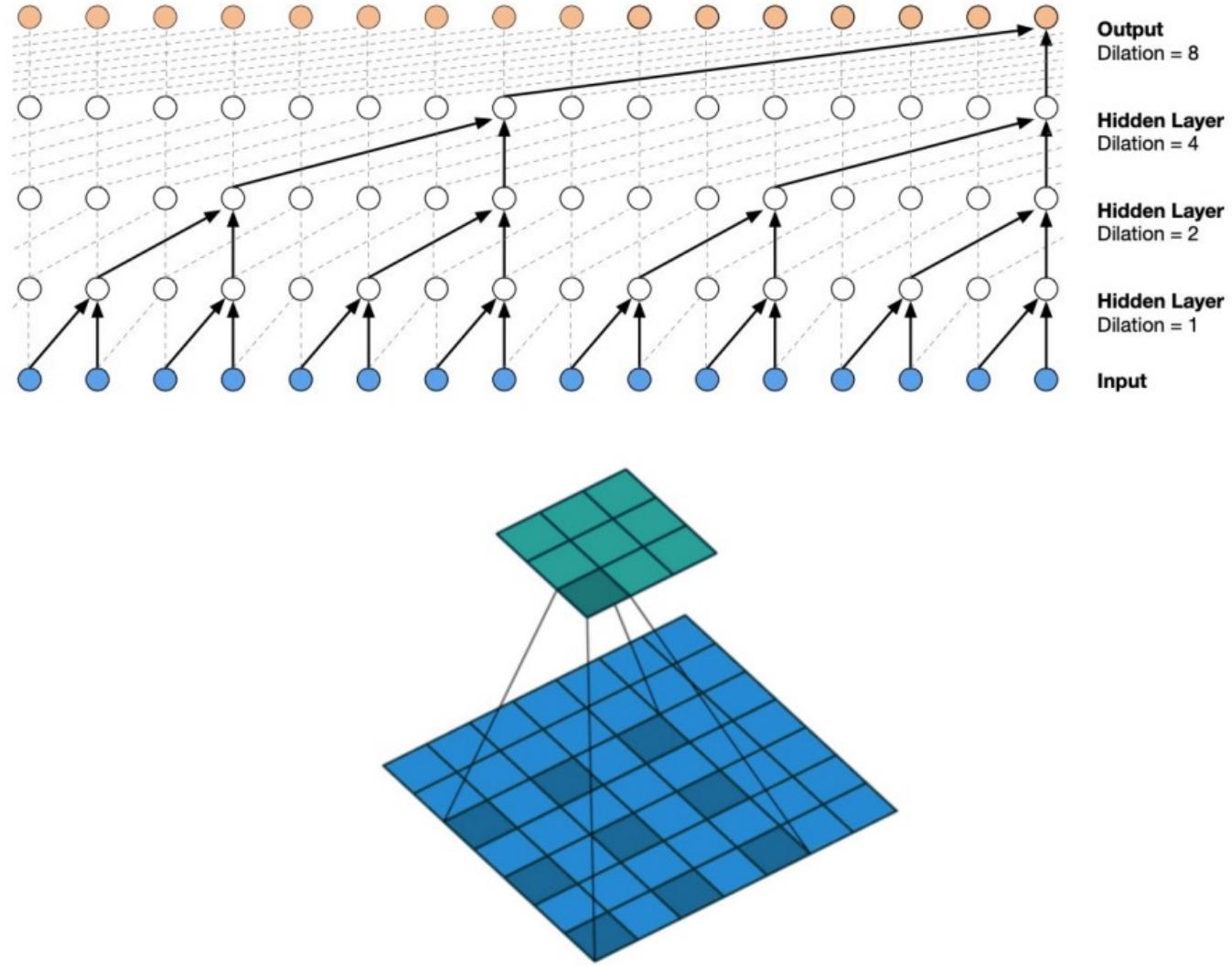
WaveNet



Источник: https://github.com/yandexdataschool/speech_course

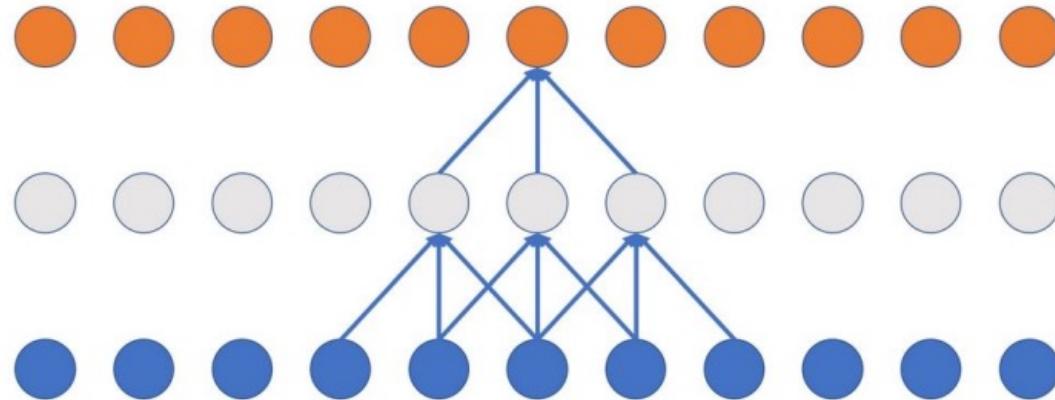
Dilated Convolution

- ▶ Увеличить поле восприятия
- ▶ Позволяет проводить моделирование долговременных зависимостей
- ▶ Причинно-следственная природа: не заглядывает в будущее
- ▶ Экспоненциальное увеличение расширения: расширение растет экспоненциально: 1, 2, 4, 8, ...
- ▶ Размер ядра: обычно 2, хотя 3 также возможно

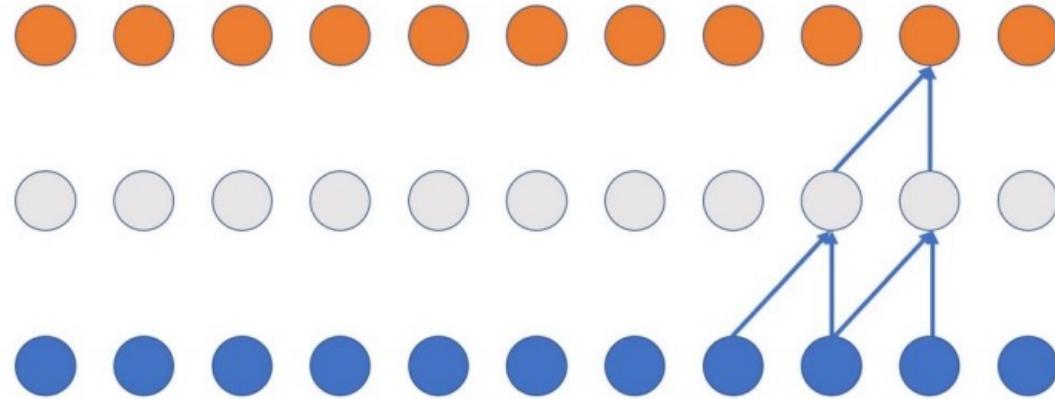


Causal Convolution

Standard Convolution

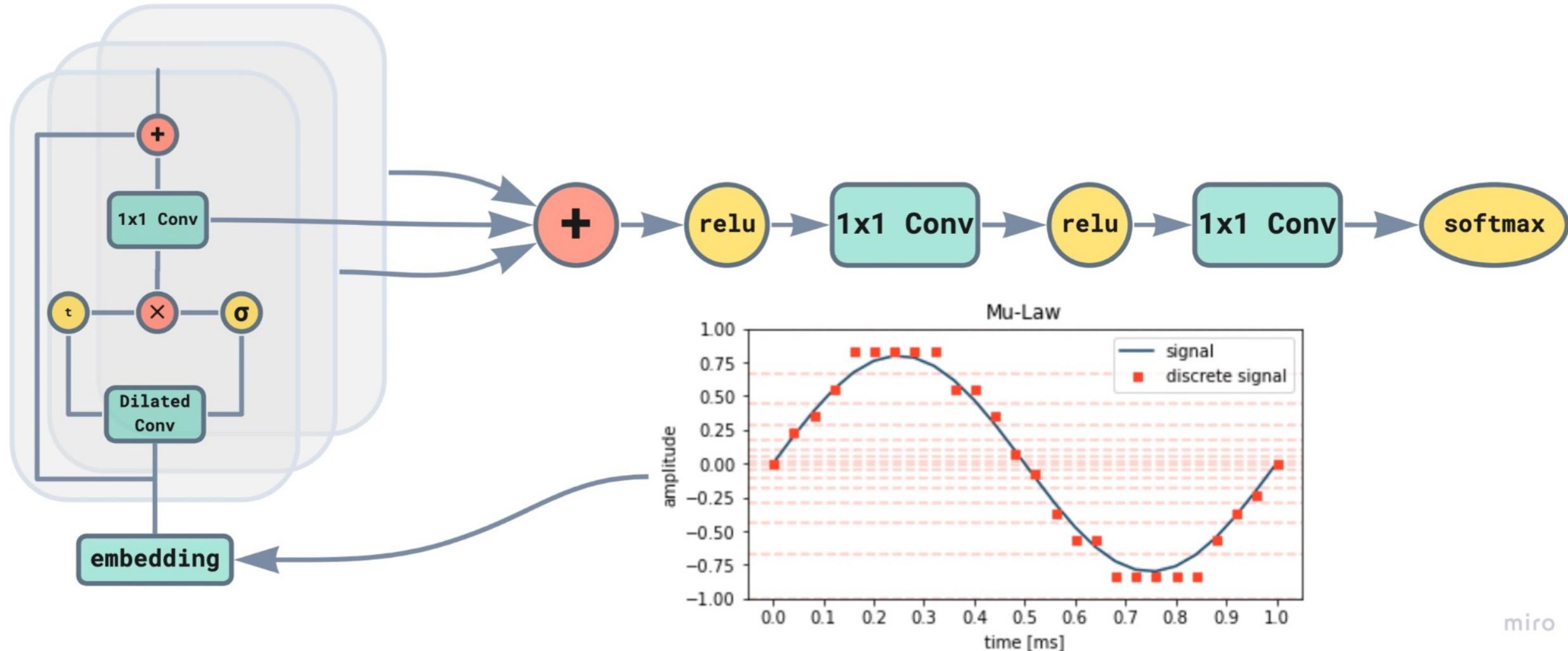


Causal Convolution



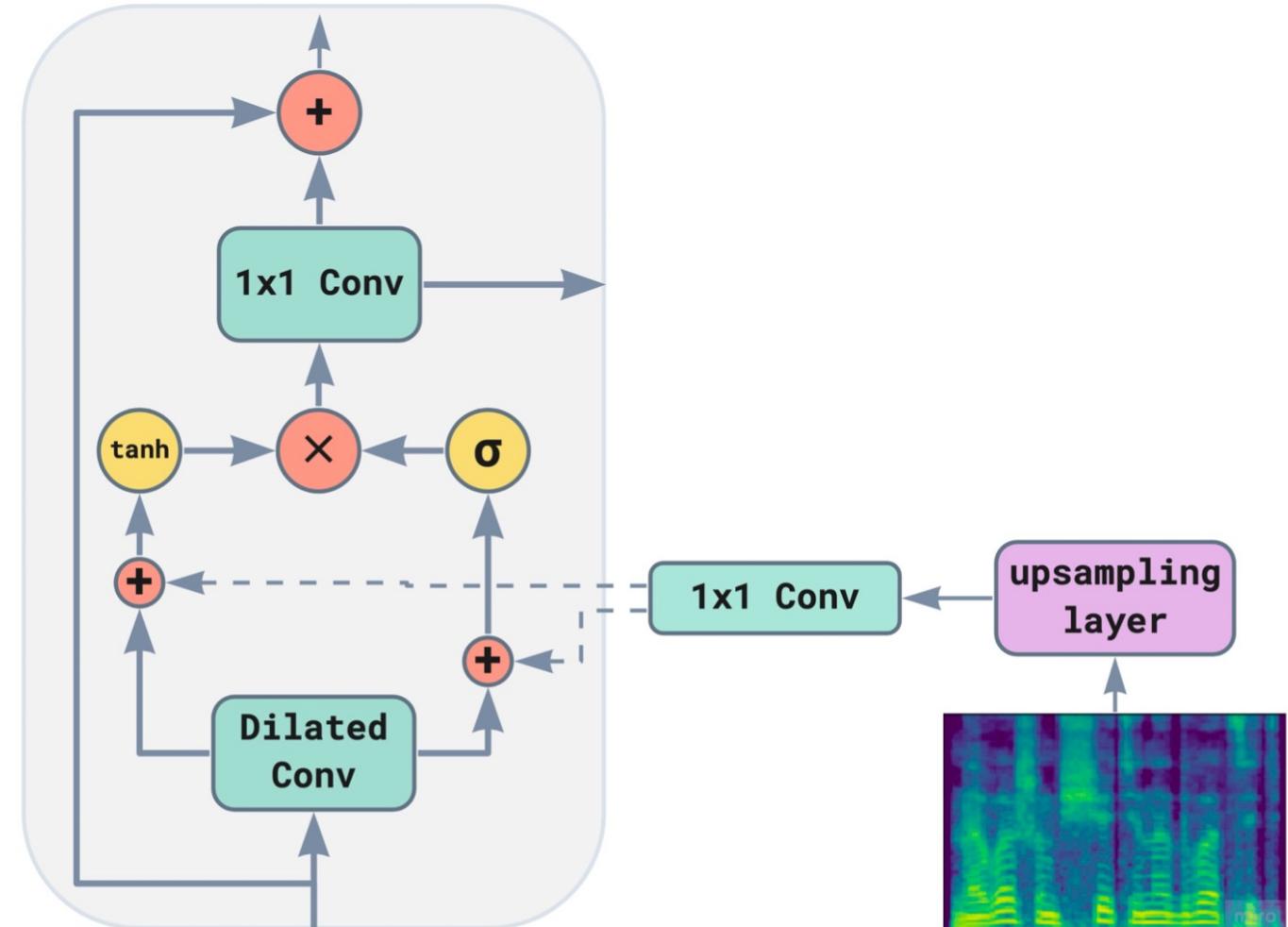
Архитектура WaveNet

all layers



Условный WaveNet

- ▶ Сигнал и Спектrogramма имеют разное временное разрешение.
- ▶ Повышение дискретизации по методу ближайшего соседа

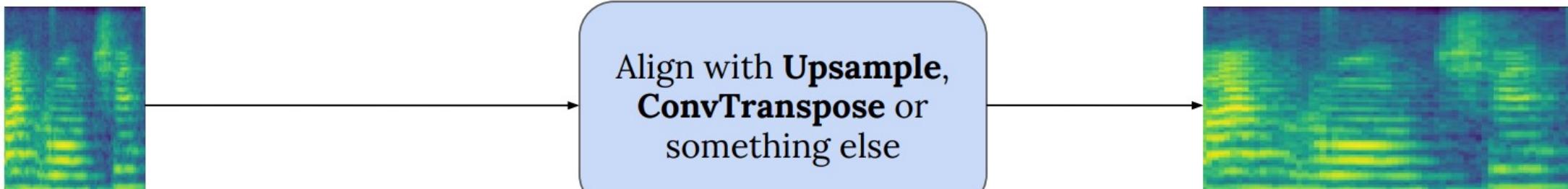


Условный WaveNet



(x_1, \dots, x_{i-1})

i is 256 times larger than j



(c_1, \dots, c_{j-1})

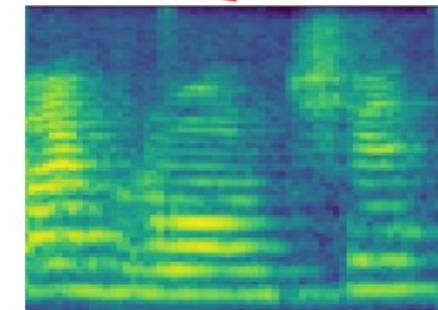
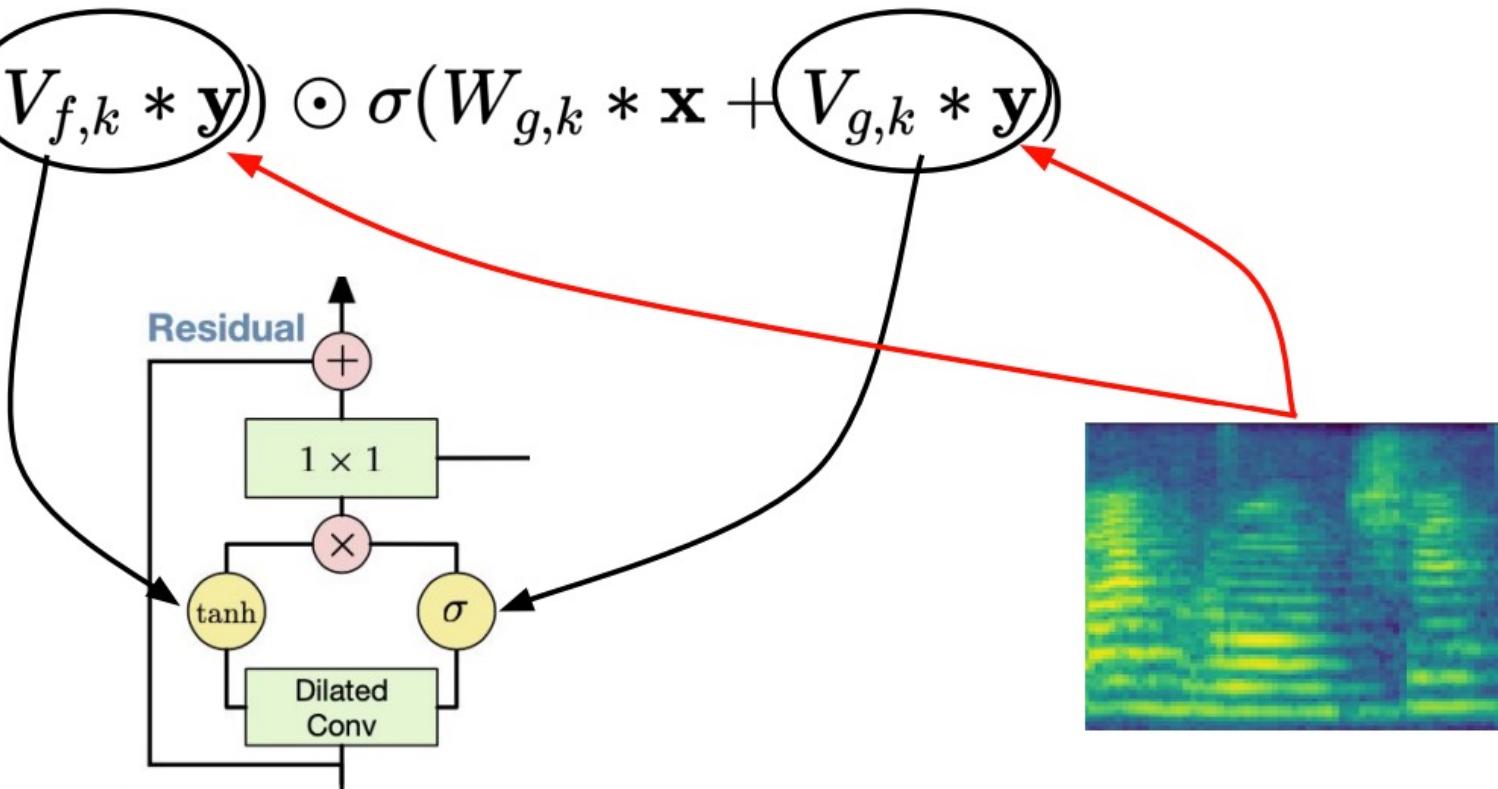
Align with **Upsample**,
ConvTranspose or
something else

Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

Условный WaveNet

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

$$\mathbf{z} = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y})$$



Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

Условный WaveNet

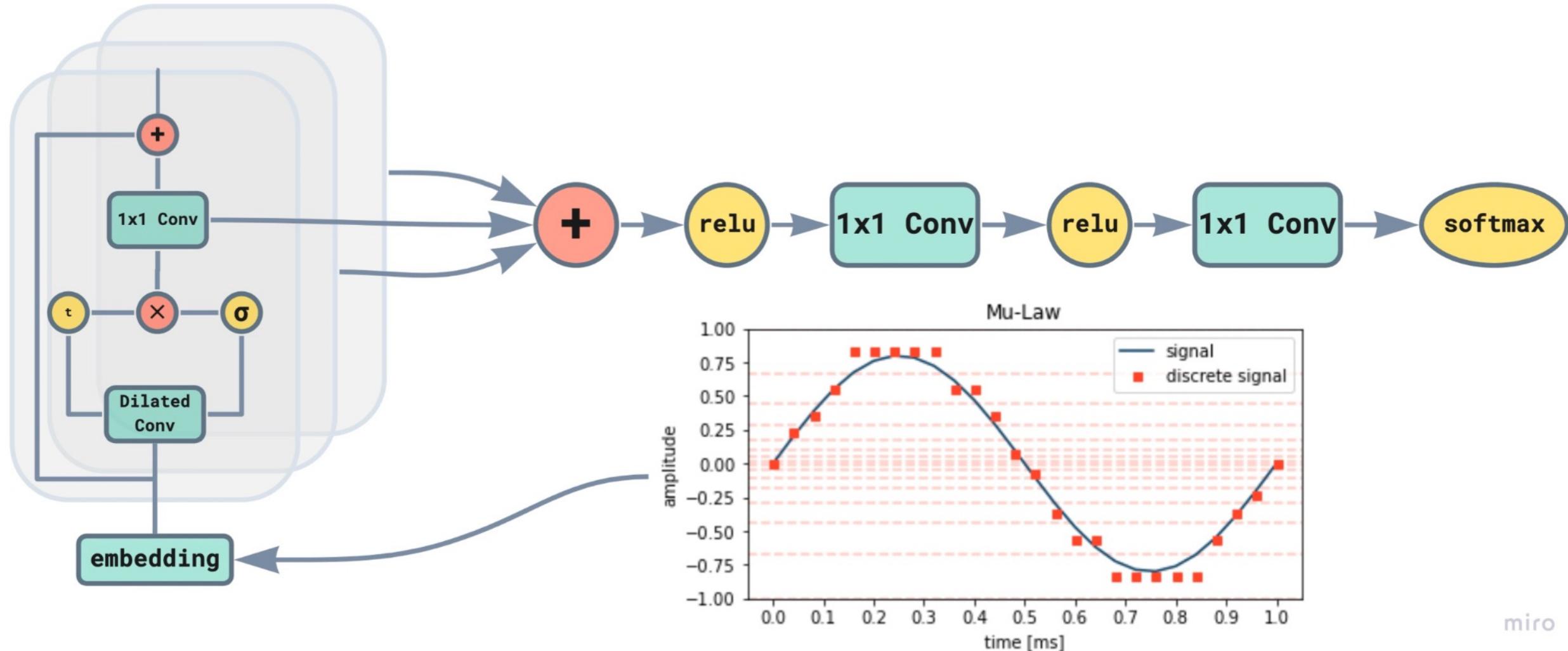
$$\mathbf{z} = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y})$$

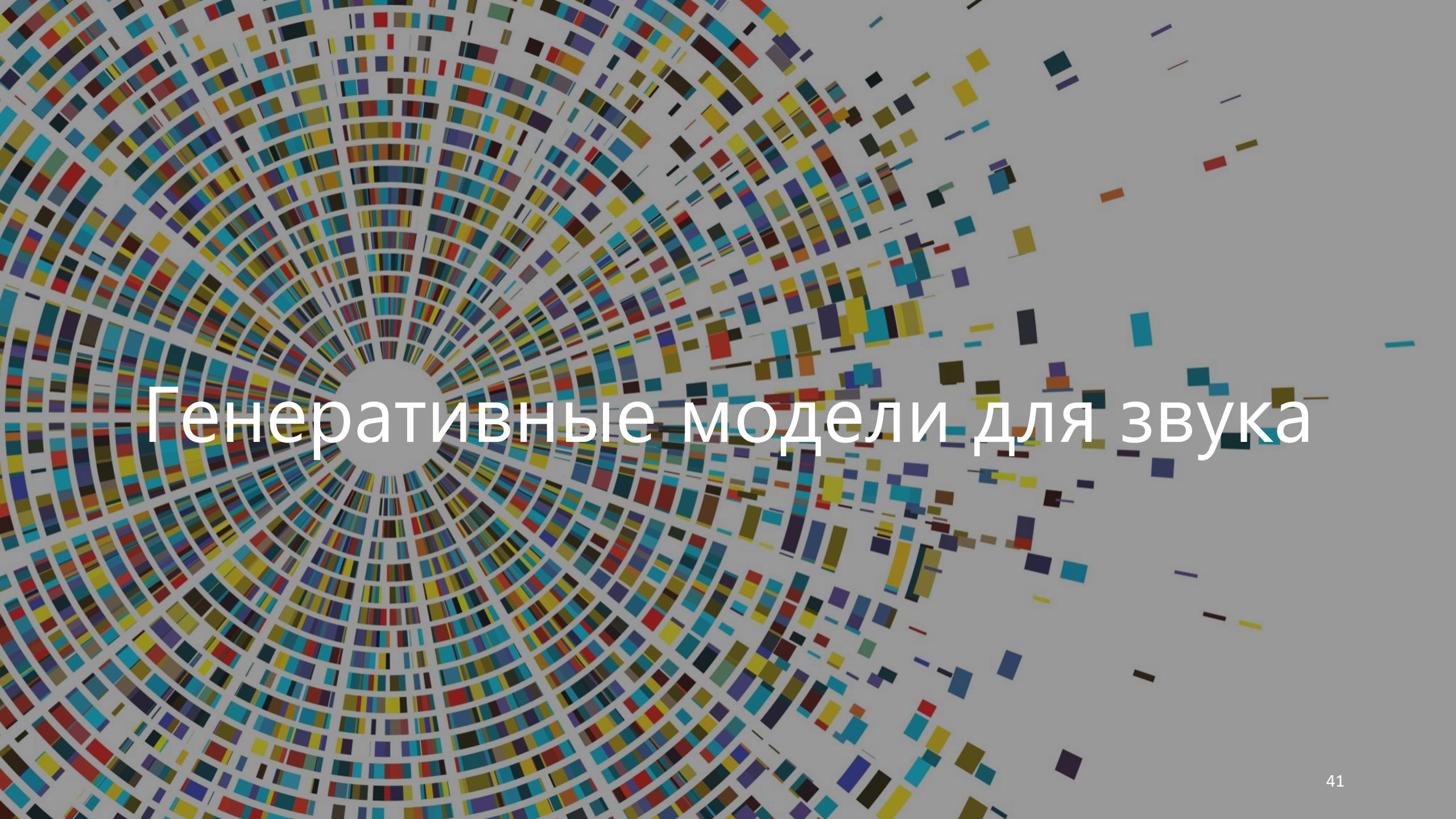
x_i (x_1, \dots, x_{i-1}) (c_1, \dots, c_{i-1})

Источник: http://wiki.cs.hse.ru/Прикладные_задачи_анализа_данных

Архитектура WaveNet

all layers

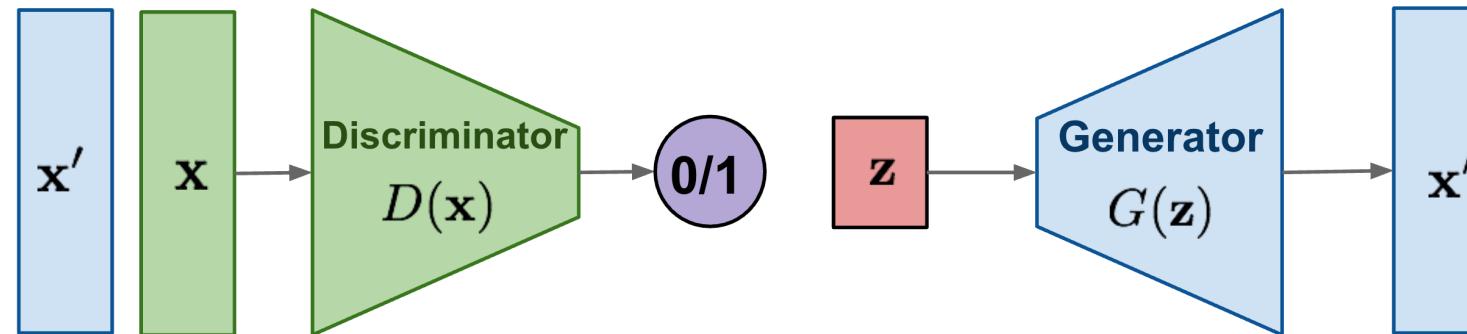




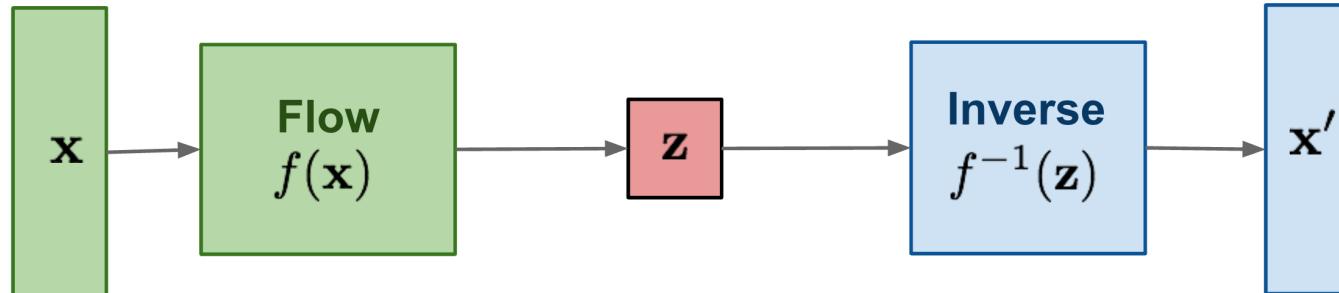
Генеративные модели для звука

Генеративные модели

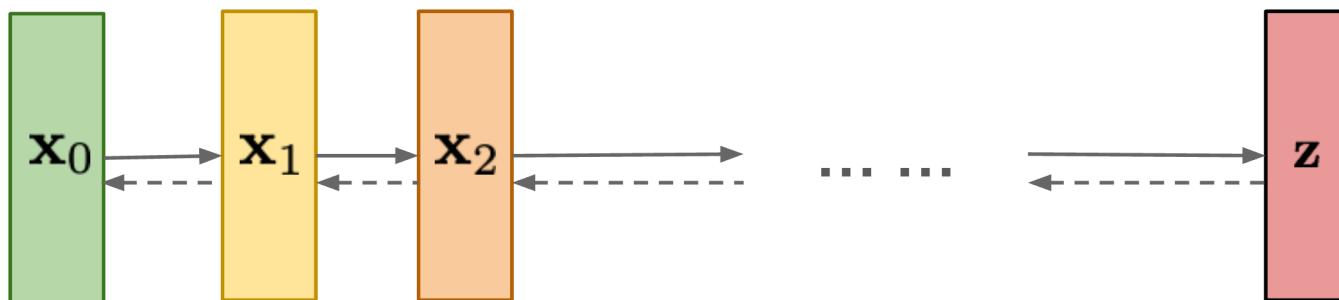
GAN: Adversarial training



Flow-based models:
Invertible transform of distributions

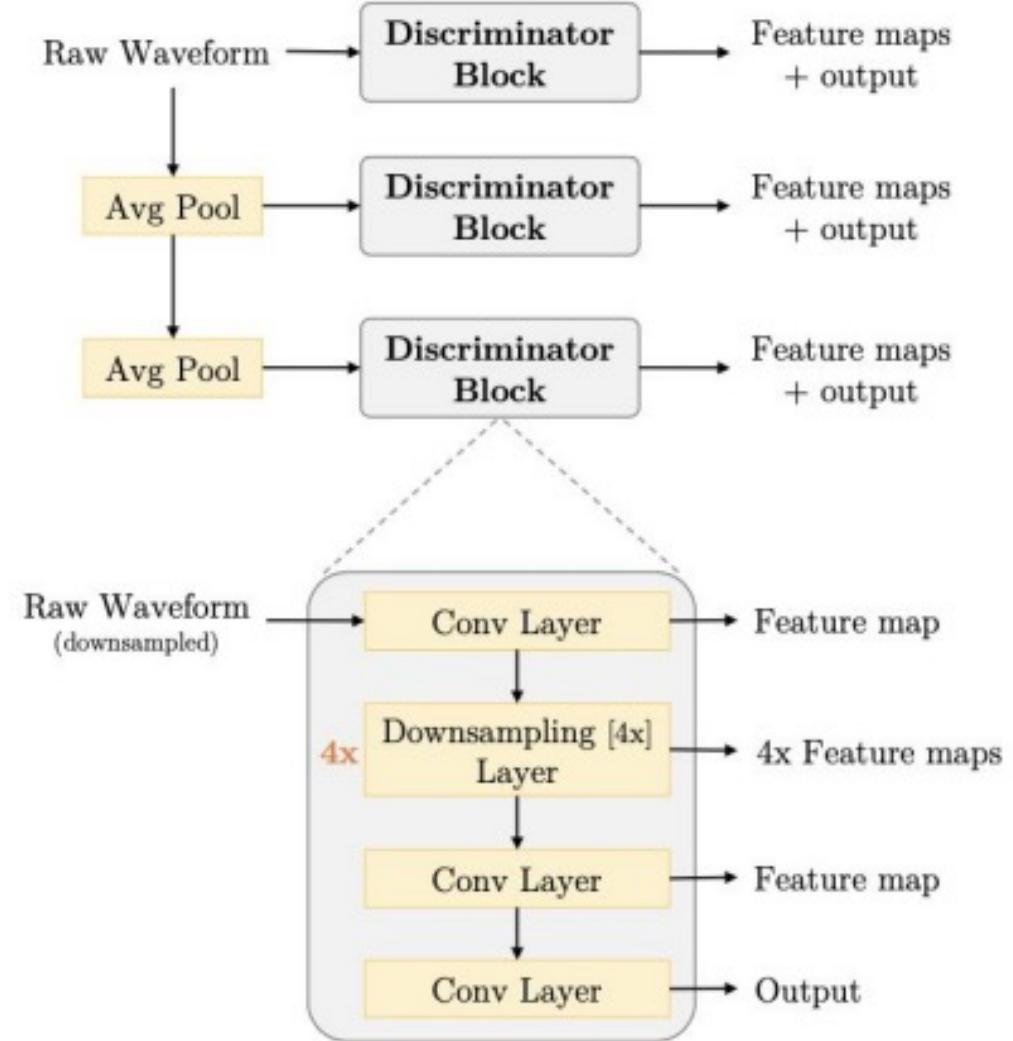
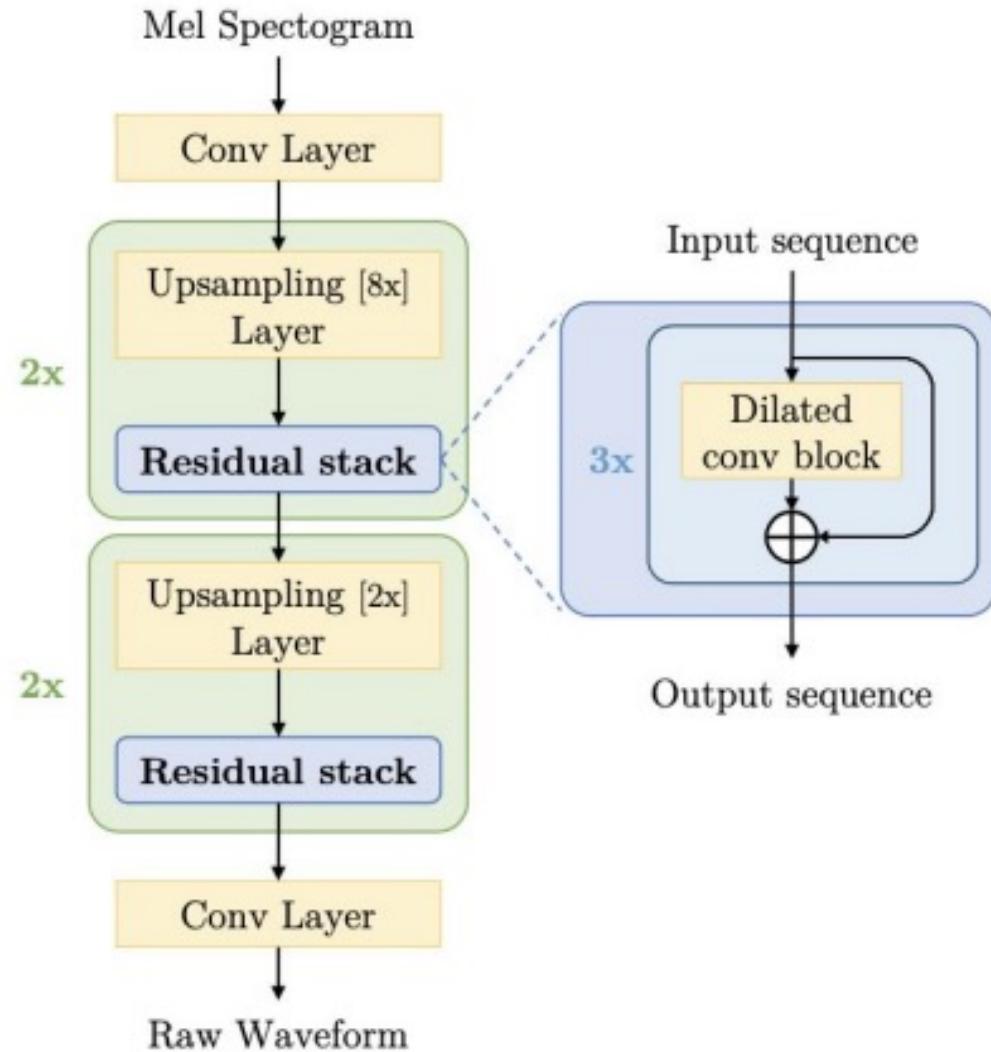


Diffusion models:
Gradually add Gaussian noise and then reverse



Source: <https://lilianweng.github.io/posts/2021-07-11-diffusion-models>

MeIGAN

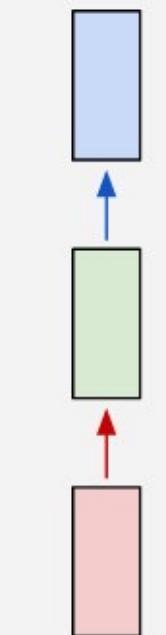




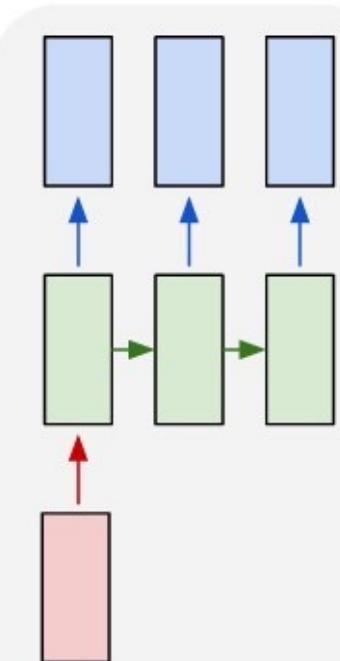
Акустические модели

Типы задач

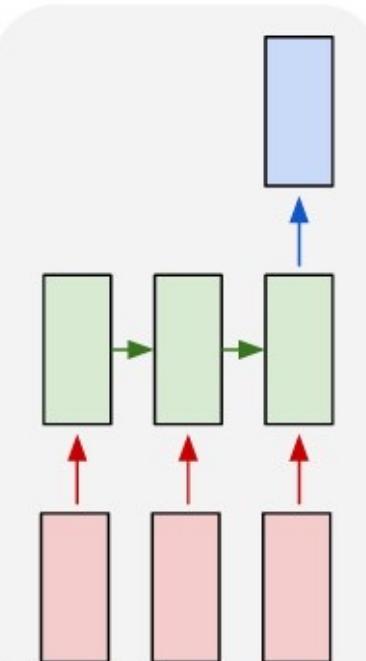
one to one



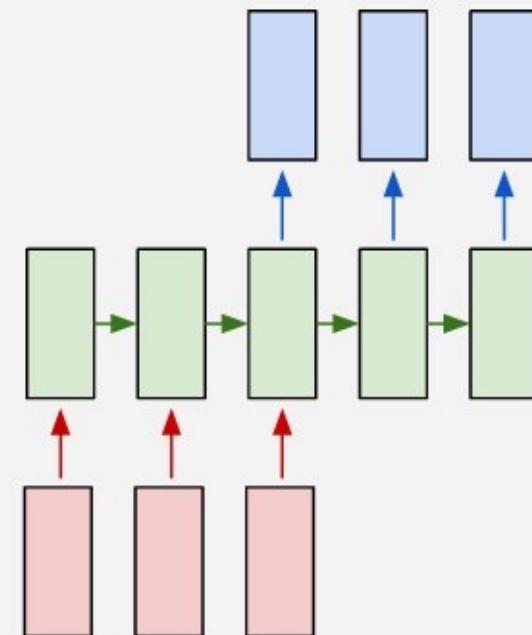
one to many



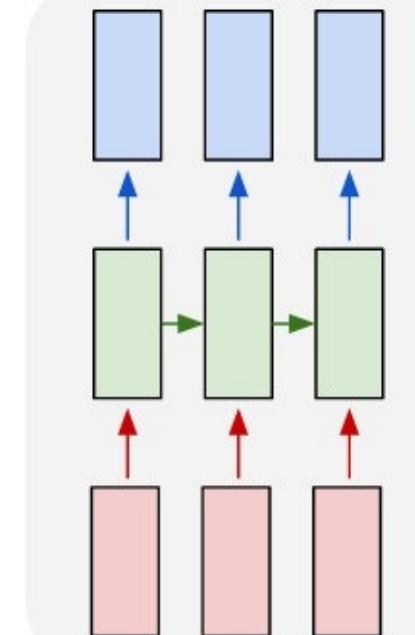
many to one



many to many

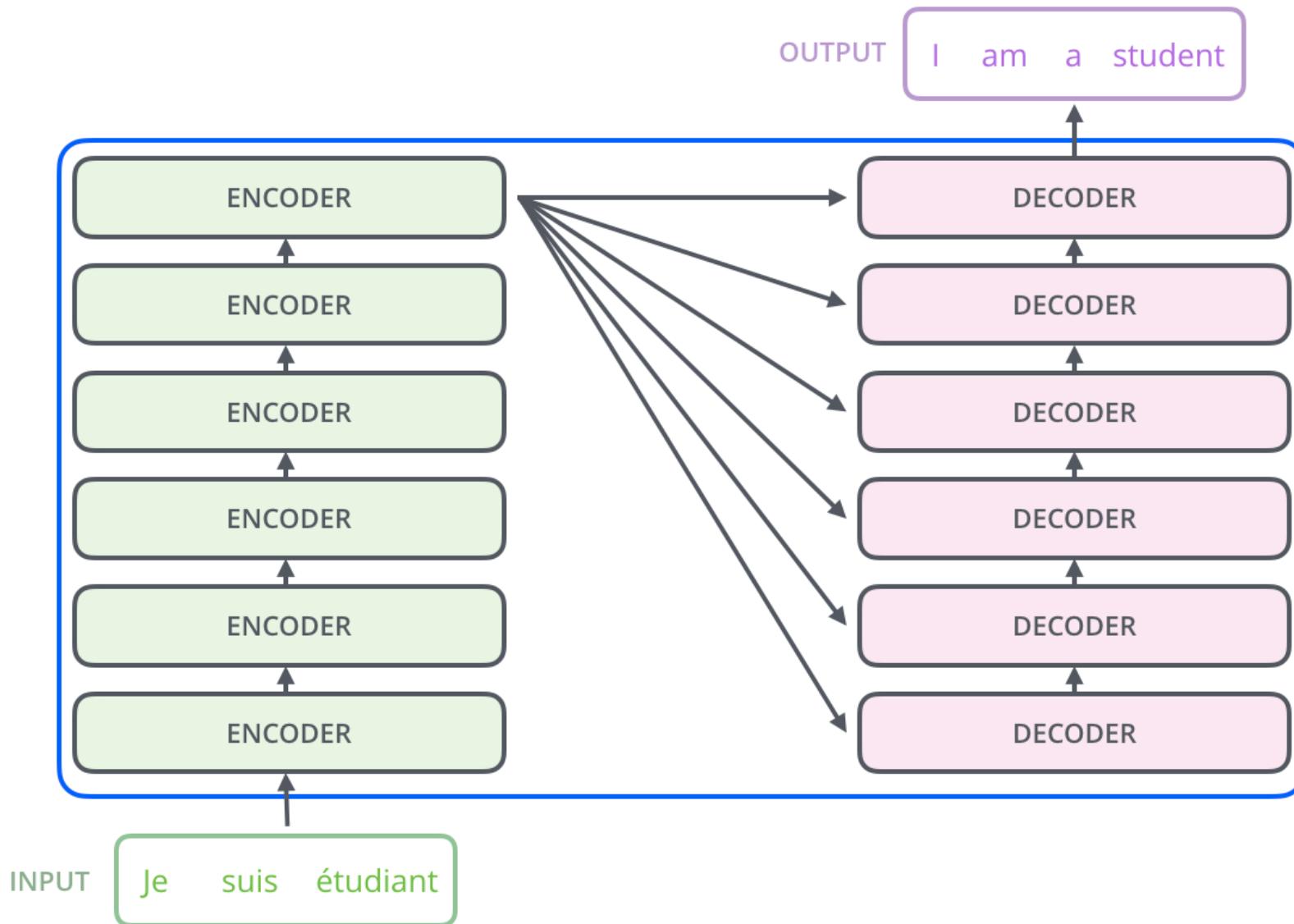


many to many



Источник: <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Общая схема трансформера



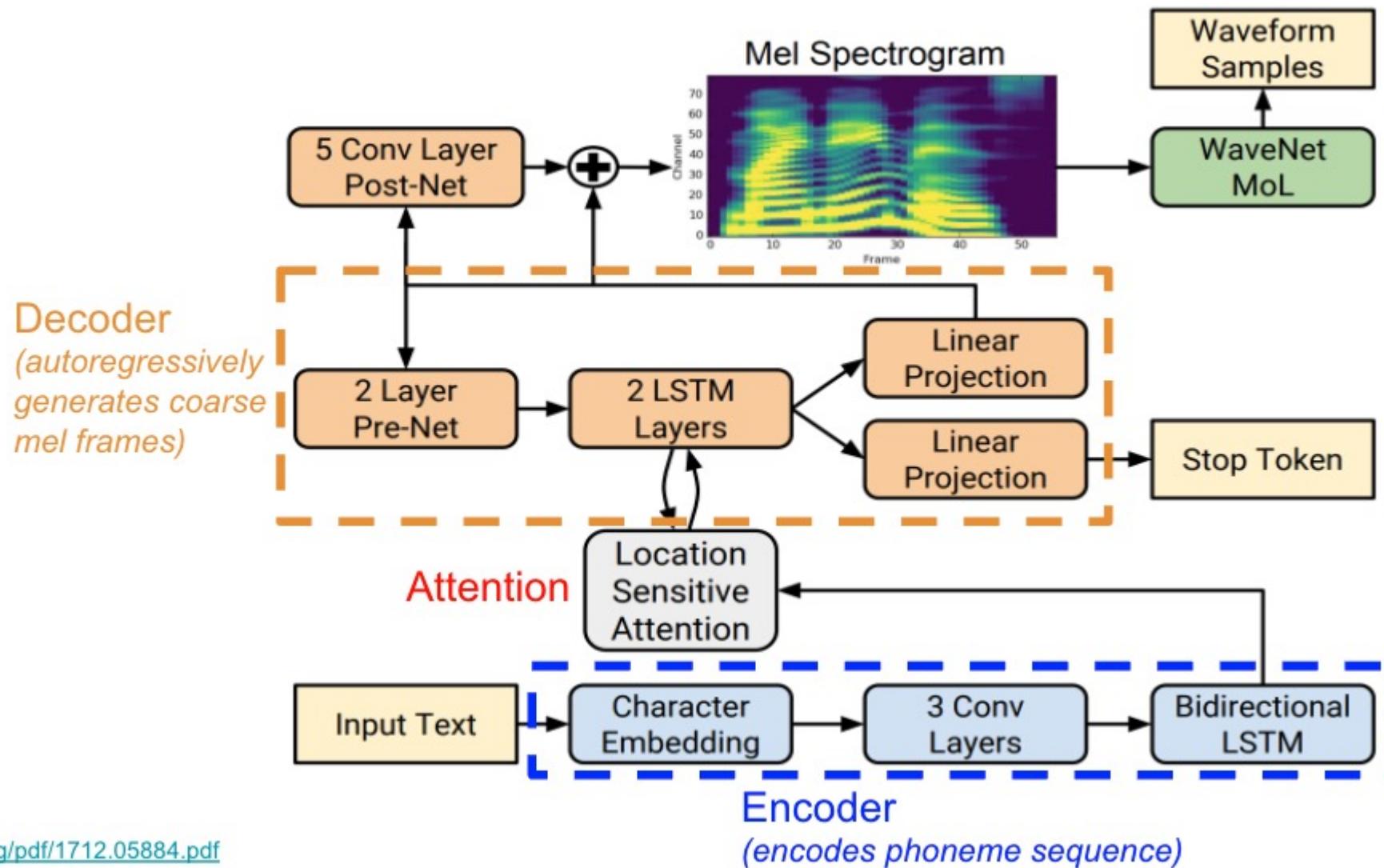
Акустическая модель

- ▶ Акустические модели позволяют получить мел-спектrogramму для данного текста
- ▶ Это seq2seq задача (текст → мел)

Tacotron 2

- ▶ Google разработал Tacotron2 в 2017 году
- ▶ LSTM-только в базовой версии (тогда Трансформеры еще не захватили мир)
- ▶ Не требует больших объемов данных для начала воспроизведения речи (~20 часов студийных записей может быть достаточно для конкурентоспособной базовой версии)
- ▶ Все еще работает на удивление хорошо

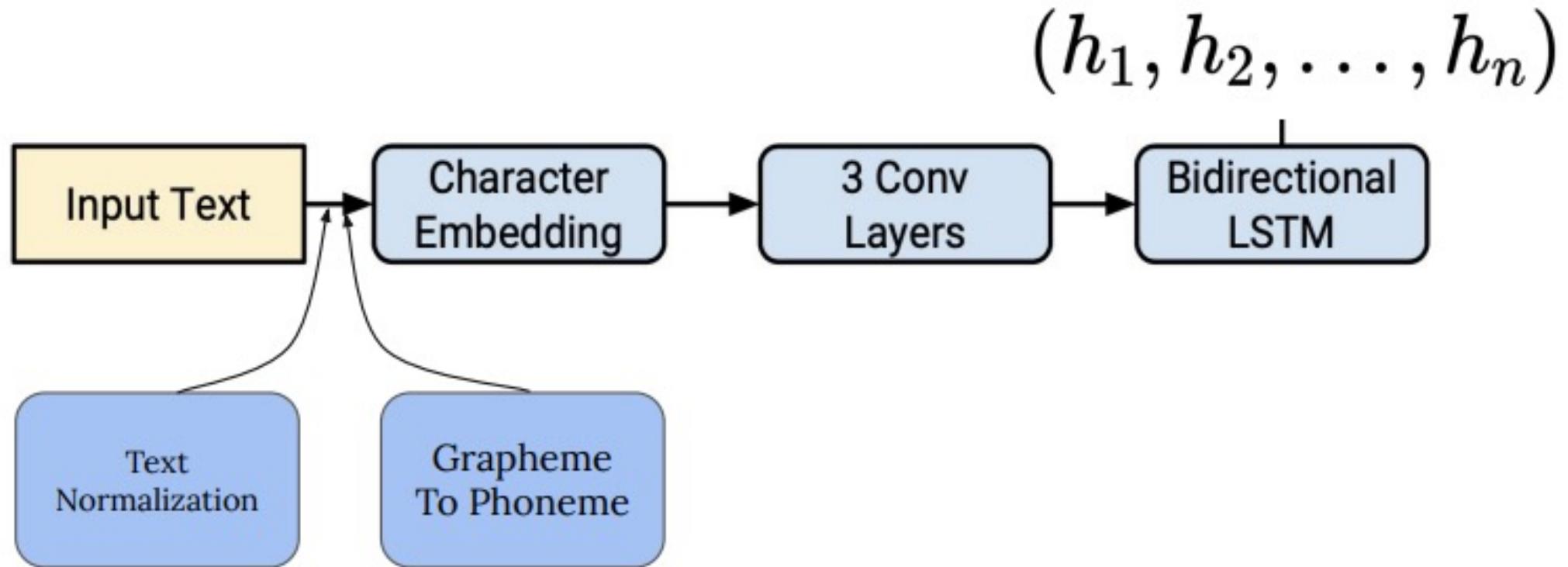
Tacotron 2



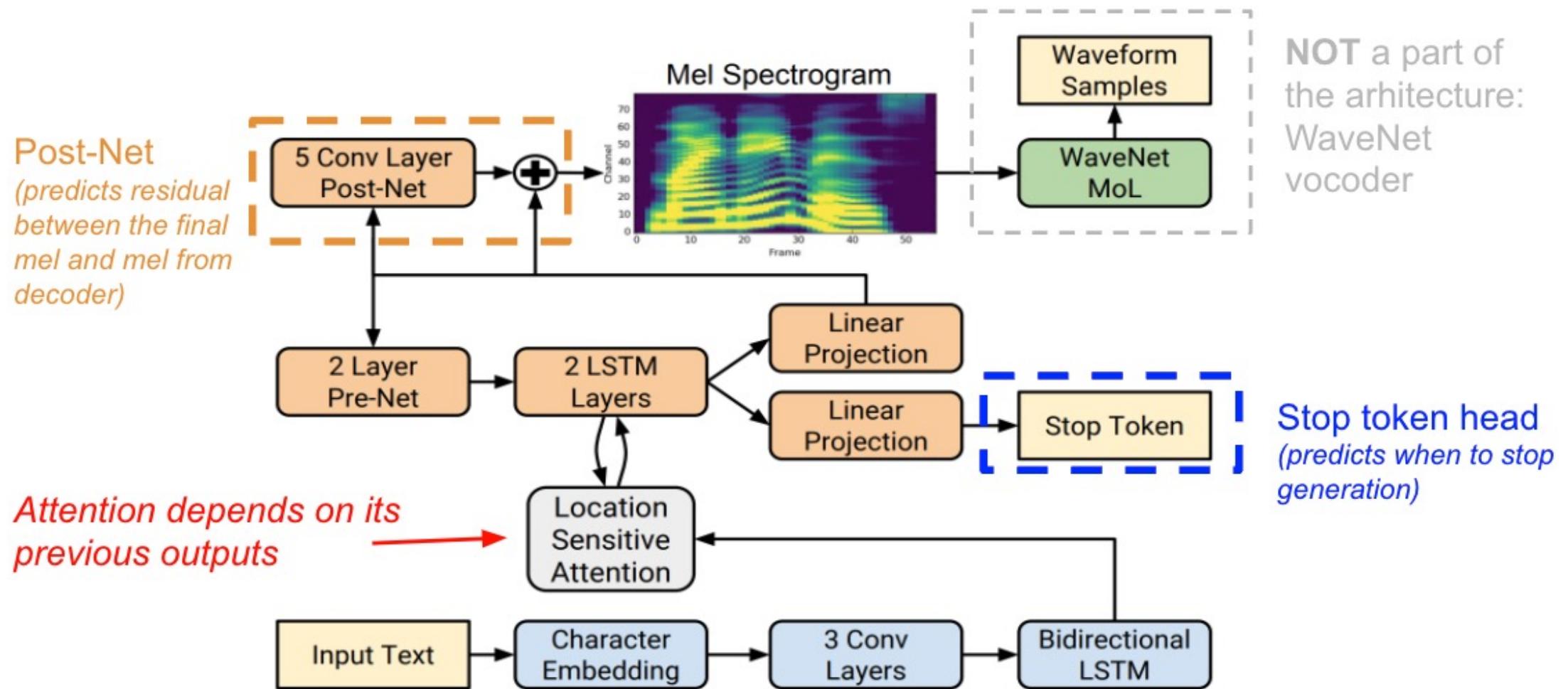
<https://arxiv.org/pdf/1712.05884.pdf>

Источник: https://github.com/yandexdataschool/speech_course

Tacotron 2



Tacotron 2



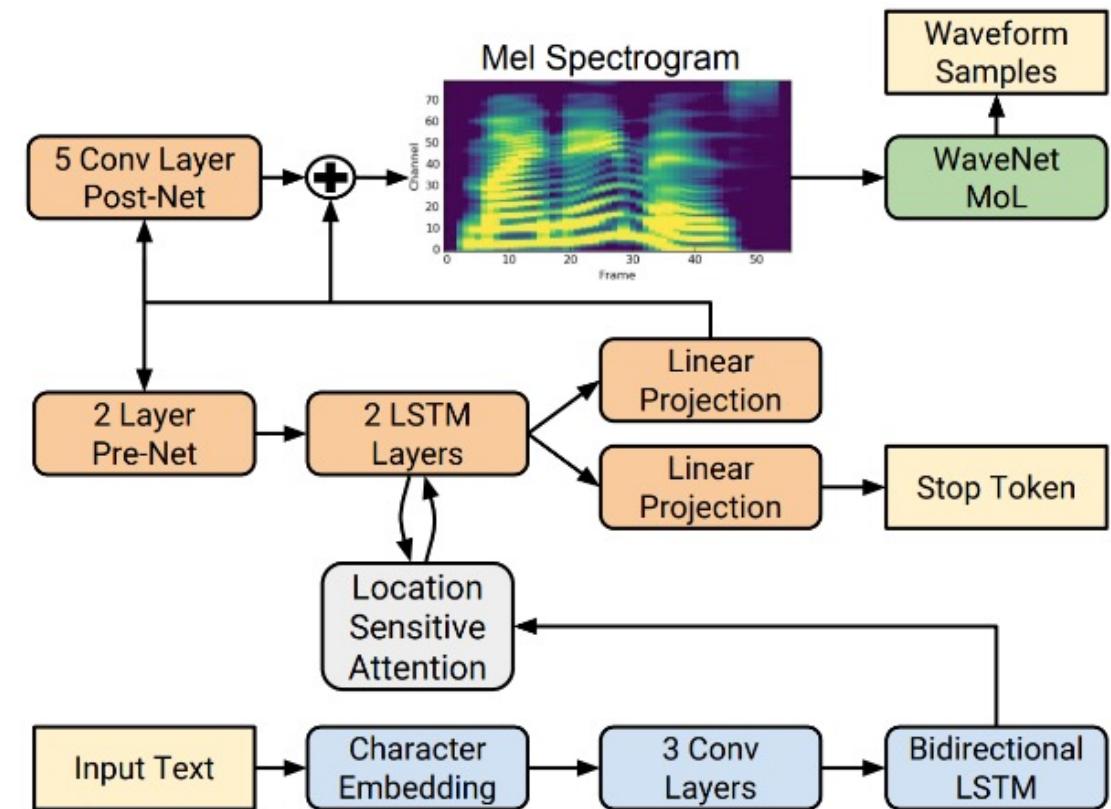
Tacotron 2

$$\mathcal{L} = \mathcal{L}_{\text{pre}} + \mathcal{L}_{\text{post}} + \text{StopToken}$$

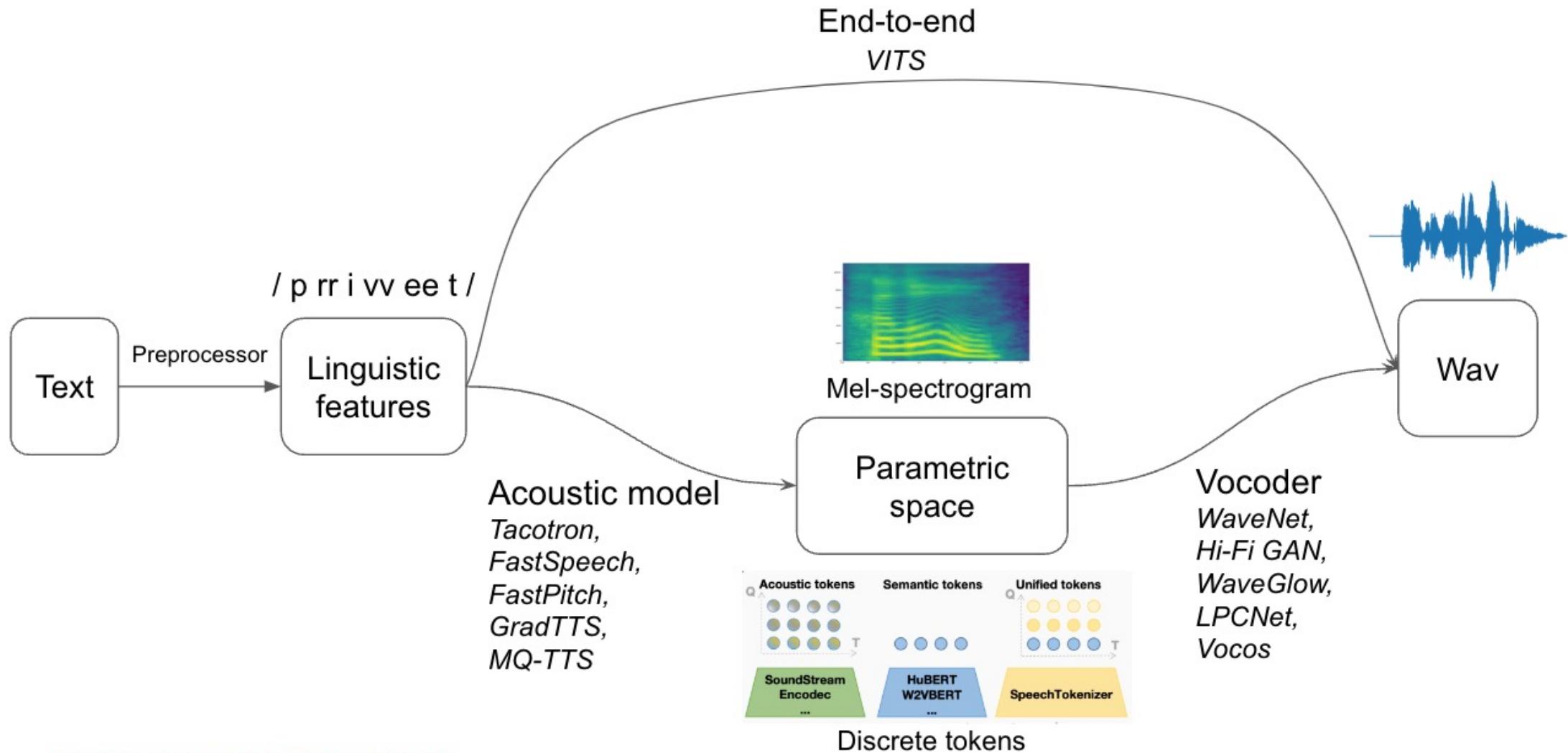
$$\mathcal{L}_{\text{pre}} = \text{MSE}(x, \hat{x}_{\text{pre}})$$

$$\mathcal{L}_{\text{post}} = \text{MSE}(x, \hat{x}_{\text{post}})$$

$$\text{StopToken} = \text{CE}(h, \mathbb{I}[h = \text{stop}])$$



Модели TTS



Источник: https://github.com/yandexdataschool/speech_course