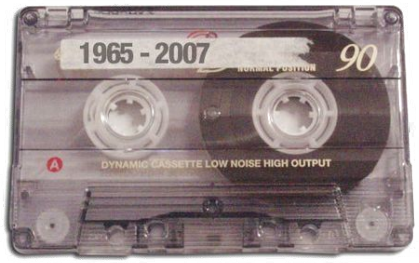
План

1. Аудиофайлы как мы привыкли
2. Feature extraction для обработки звука
    a. STFT
    b. Фрейминг
    c. Мел спектрограммы
    d. wav2vec
3. Примеры больших задач
4. Примеры маленьких (но важных) задач
5. Данные для обучения

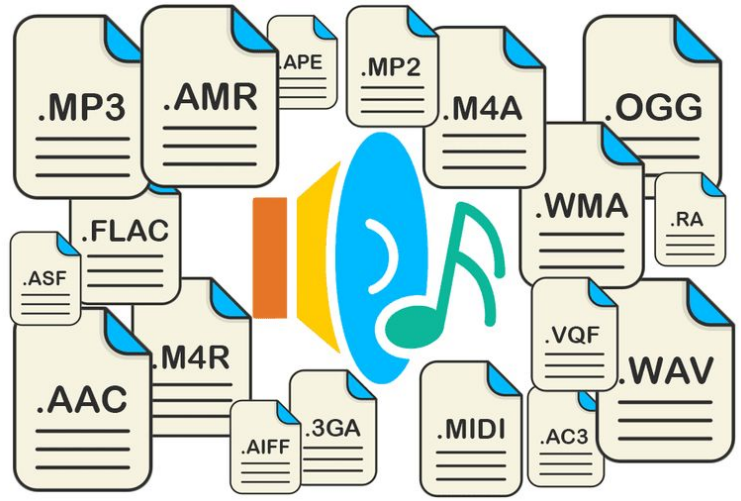
Аналоговый сигнал

Каждый из представляющих параметров описывается непрерывным множеством значений



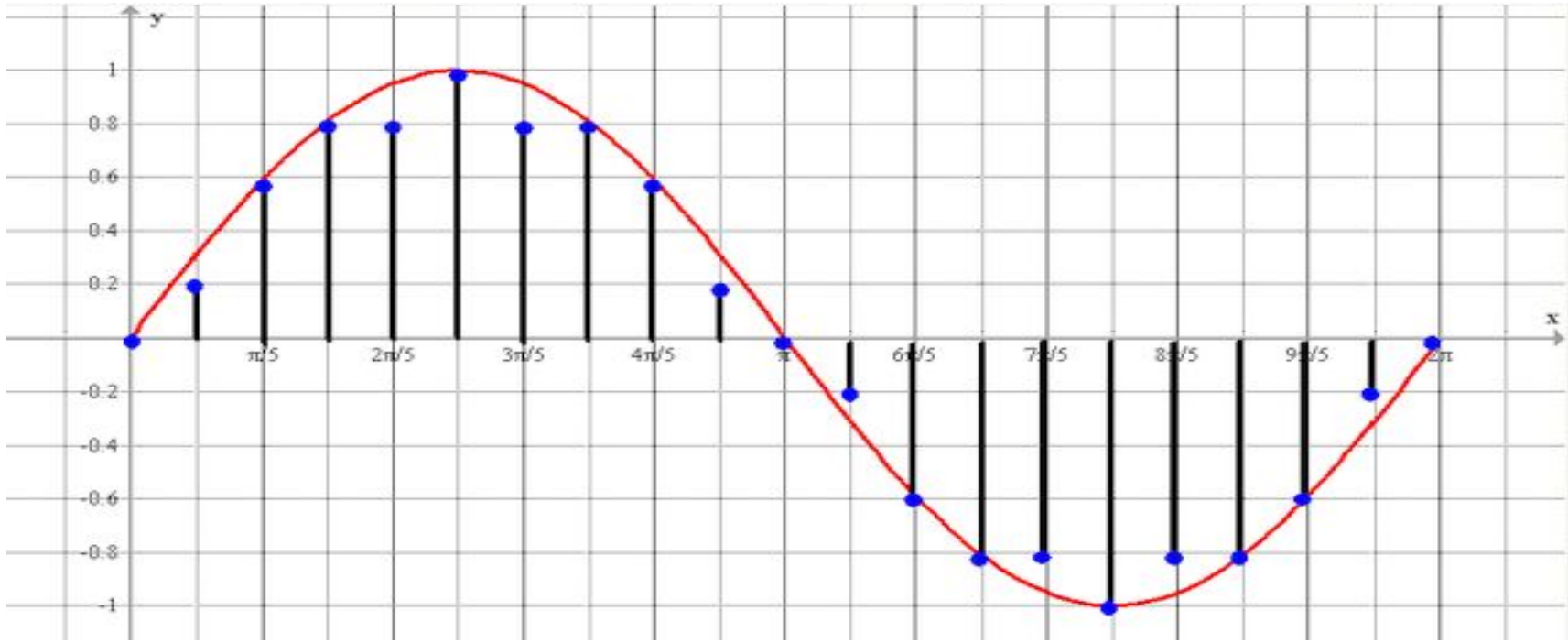
Цифровой сигнал

Можно представить в виде последовательности дискретных значений



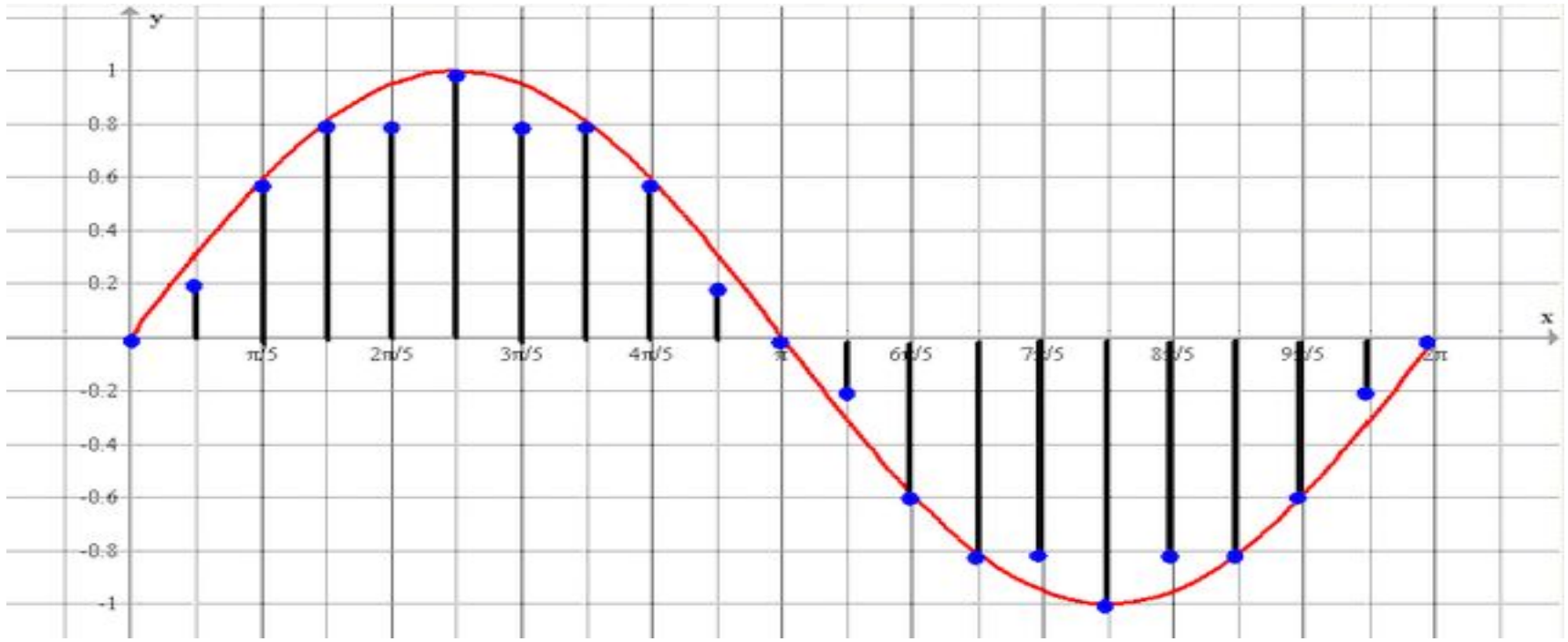
>компактнее

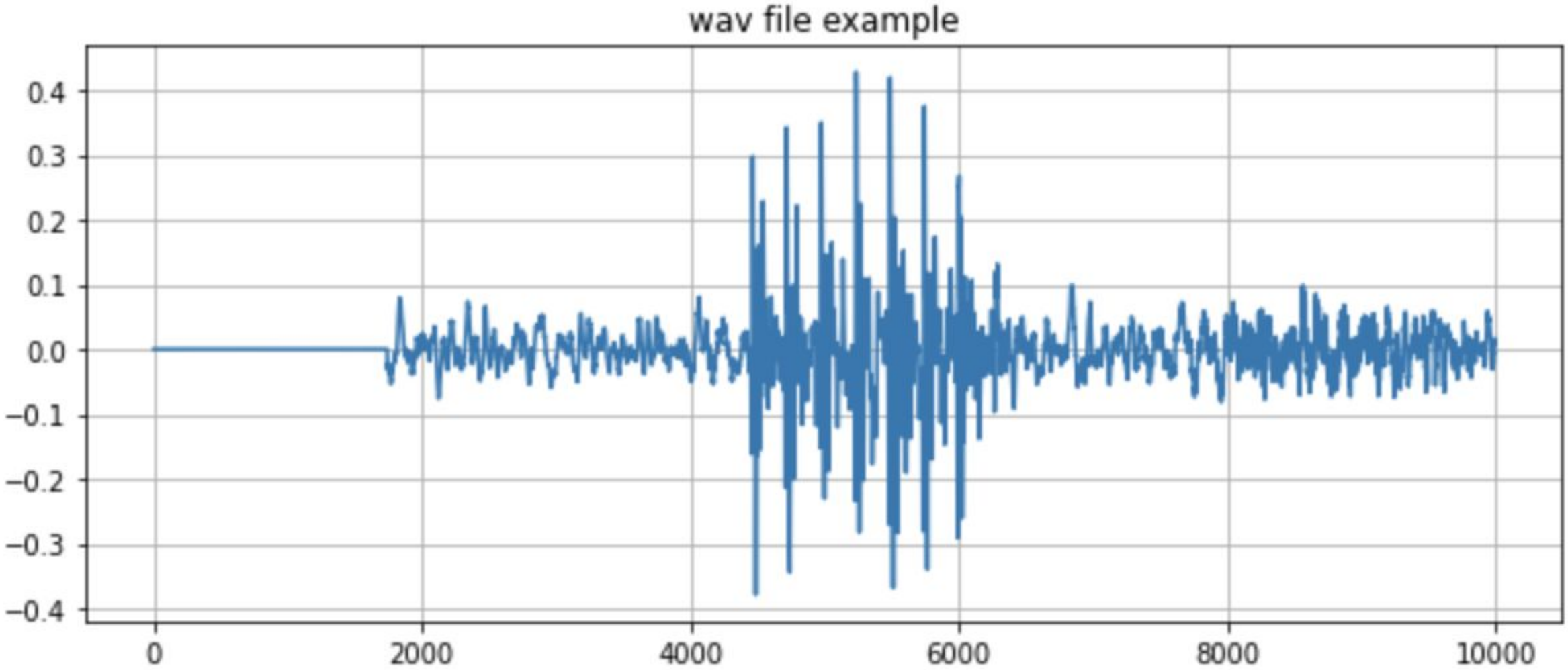
>точнее

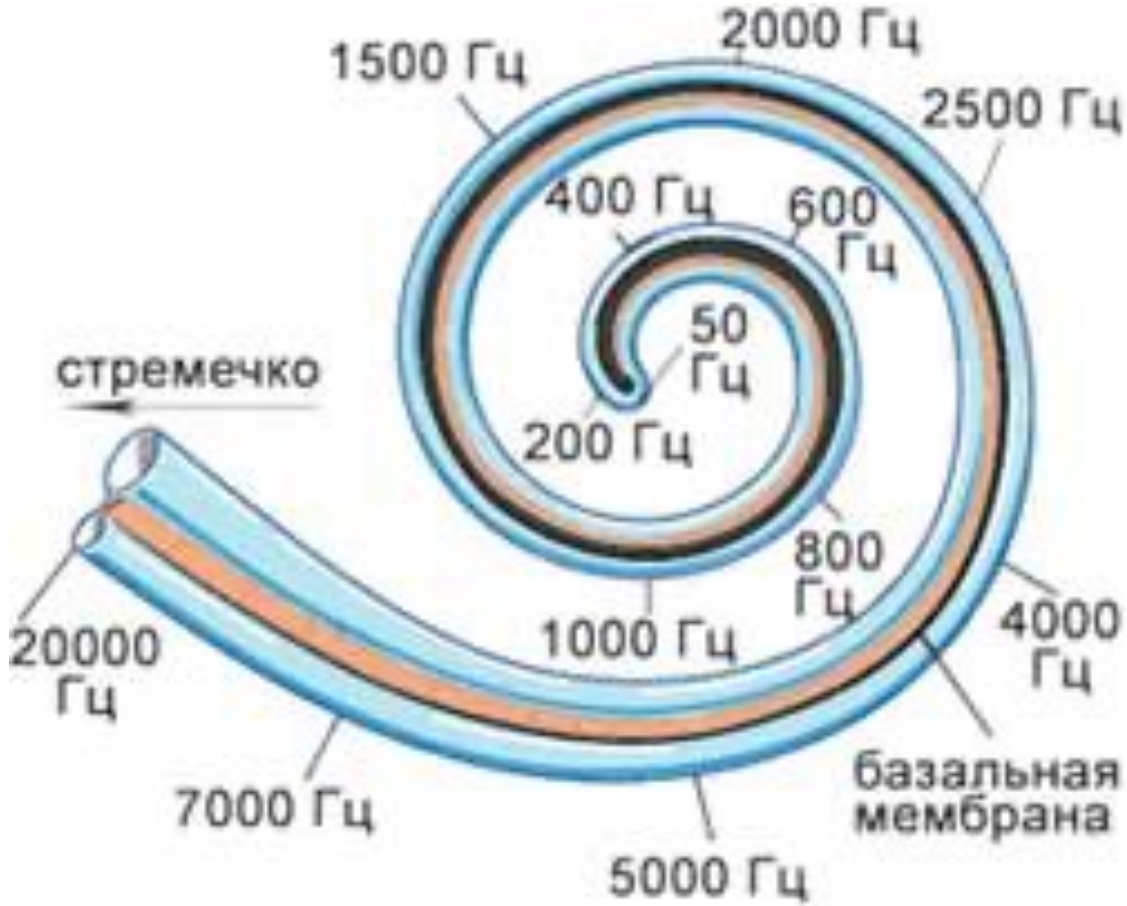

y
x
1
0.8
0.6
0.4
0.2
-0.2
-0.4
-0.6
-0.8
-1
π/5
2π/5
3π/5
4π/5
π
2π

Как хранится аудиофайл?

sample rate — число отсчетов в секунду. Типичные значения: 16000, 22050, 44100




wav file example
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
0
2000
4000
6000
8000
10000


1500 Гц
2000 Гц
2500 Гц
400 Гц
600 Гц
50 Гц
200 Гц
стремечко
800 Гц
1000 Гц
4000 Гц
20000 Гц
7000 Гц
базальная мембрана
5000 Гц

Изображения vs Аудио

1. 2d (чб)
2. 256x256 ~ 65k pix

Изображения vs Аудио

1. 2d (чб)
2. 256x256 ~ 65k pix

1. 1d (mono)
2. 1 sec ~ 44100

Трюк: будем работать с аудио как с картинками

Трюк: будем работать с аудио как с картинками

Вопрос: как перевести аудио в картинку (тензор)?

Definition [edit]

The *discrete Fourier transform* transforms a sequence of *N* complex numbers $\{\mathbf{x_n}\} := x_0, x_1, \ldots, x_{N-1}$ into another sequence of complex numbers, $\{\mathbf{X_k}\} := X_0, X_1, \ldots, X_{N-1}$, which is defined by

$$\begin{aligned} X_k &= \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i2\pi}{N}kn} \\ &= \sum_{n=0}^{N-1} x_n \cdot \left[\cos\left(\frac{2\pi}{N}kn\right) - i \cdot \sin\left(\frac{2\pi}{N}kn\right)\right], \end{aligned} \qquad \textbf{(Eq.1)}$$

where the last expression follows from the first one by Euler's formula.

The transform is sometimes denoted by the symbol $\mathcal{F}$, as in $\mathbf{X} = \mathcal{F}\{\mathbf{x}\}$ or $\mathcal{F}(\mathbf{x})$ or $\mathcal{F}\mathbf{x}$.[A]

Framing

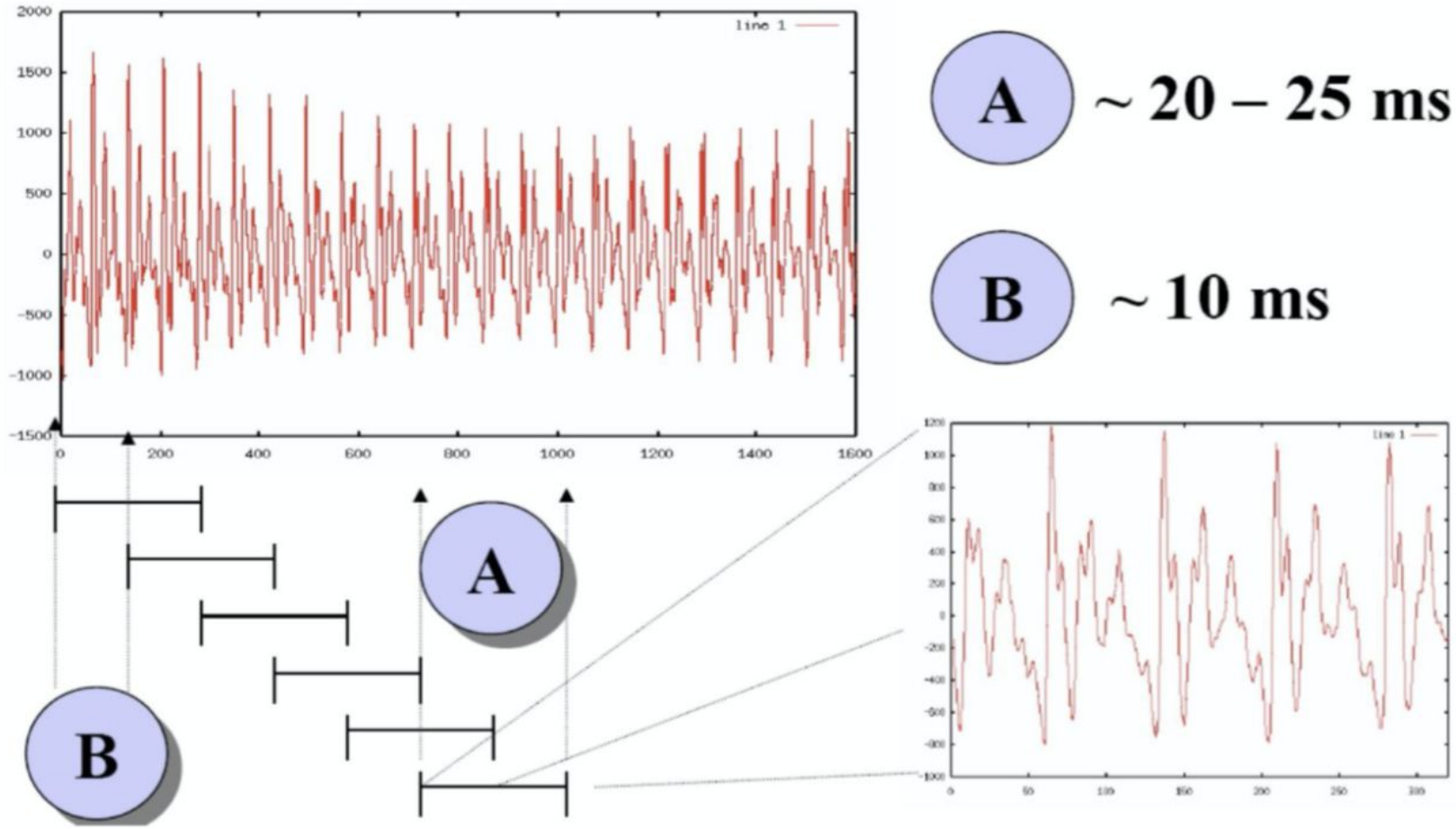
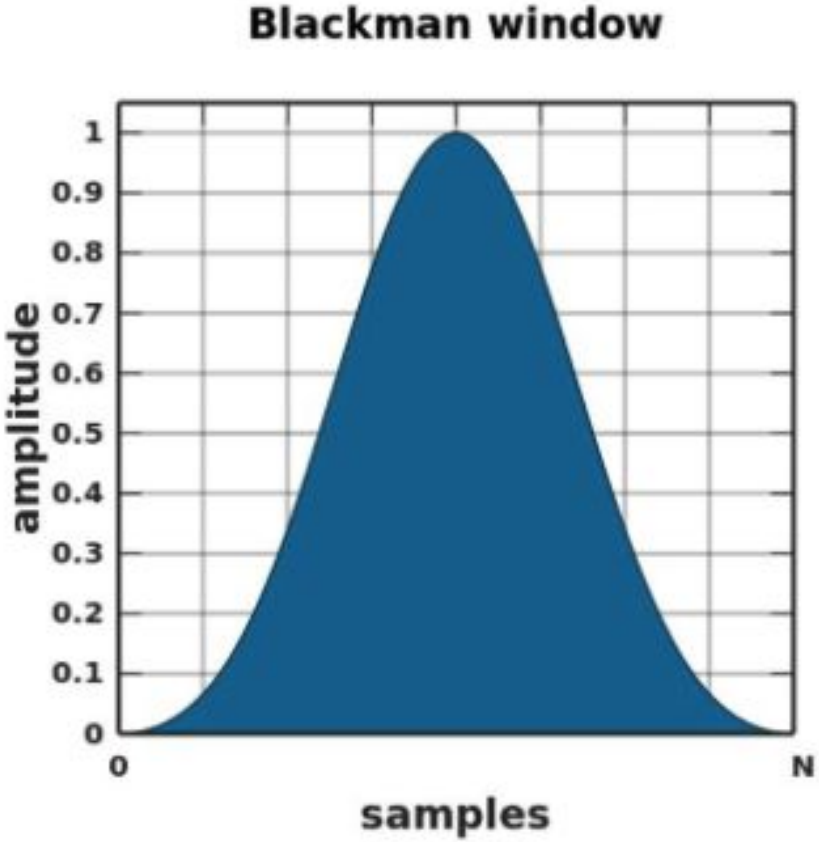
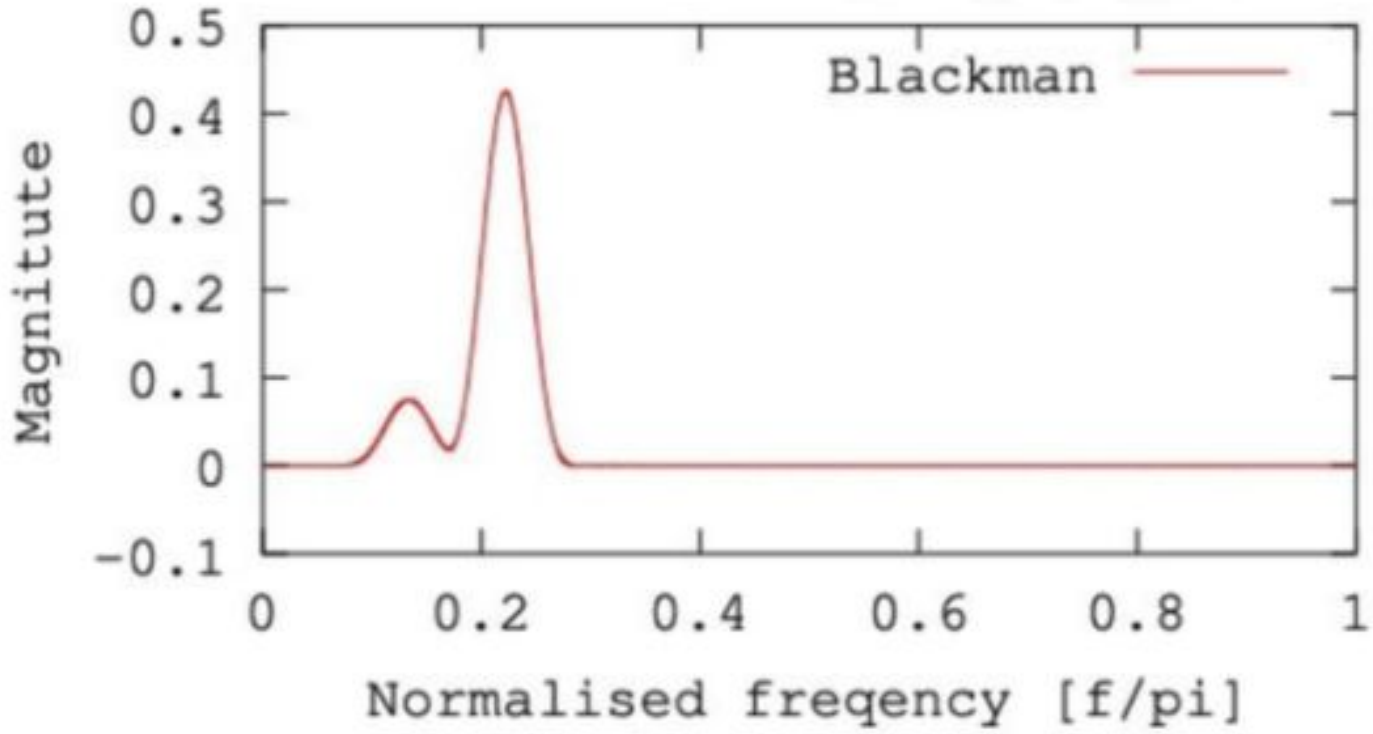


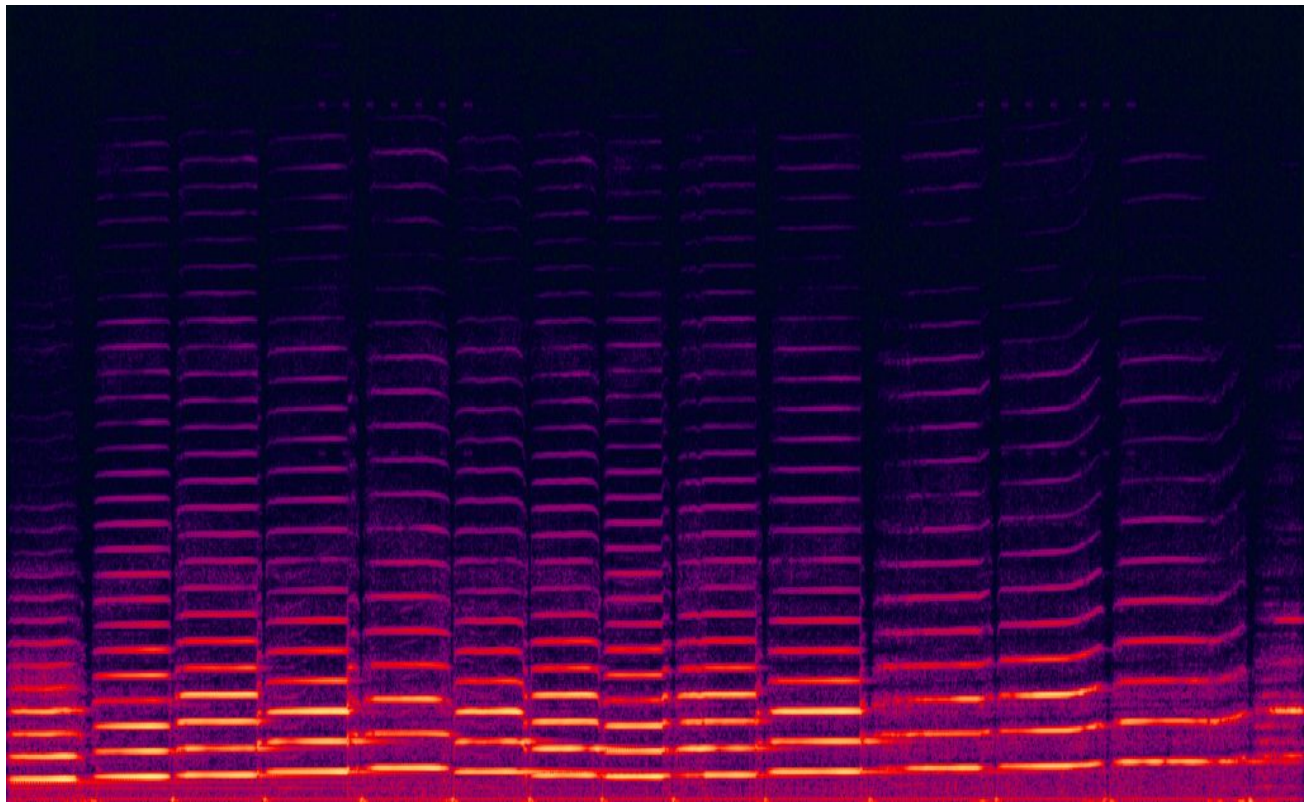
Image from Bryan Pellom

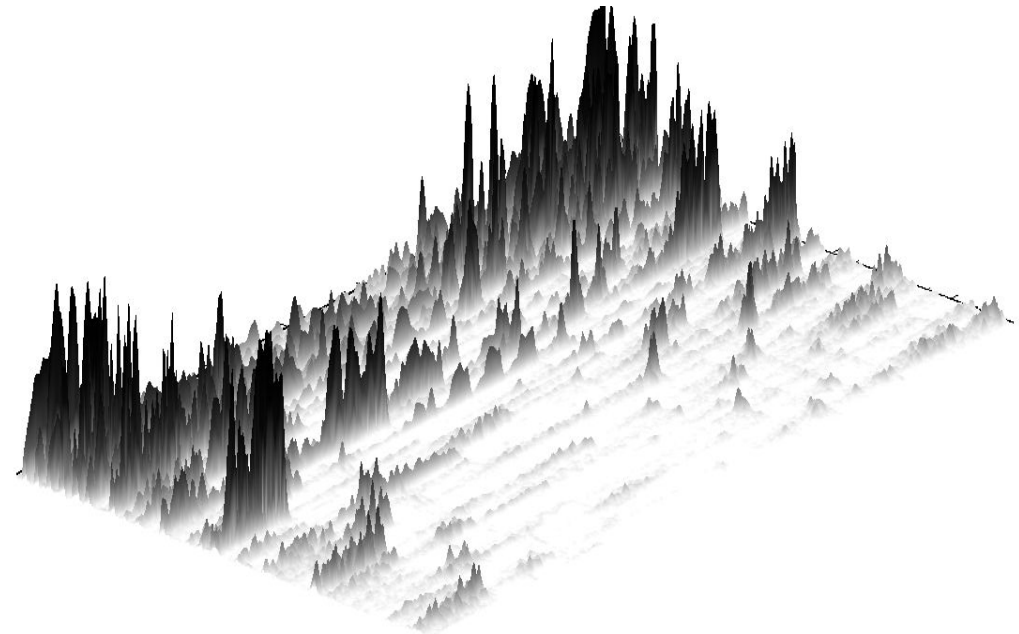
https://medium.com/@jonathan_hui/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9

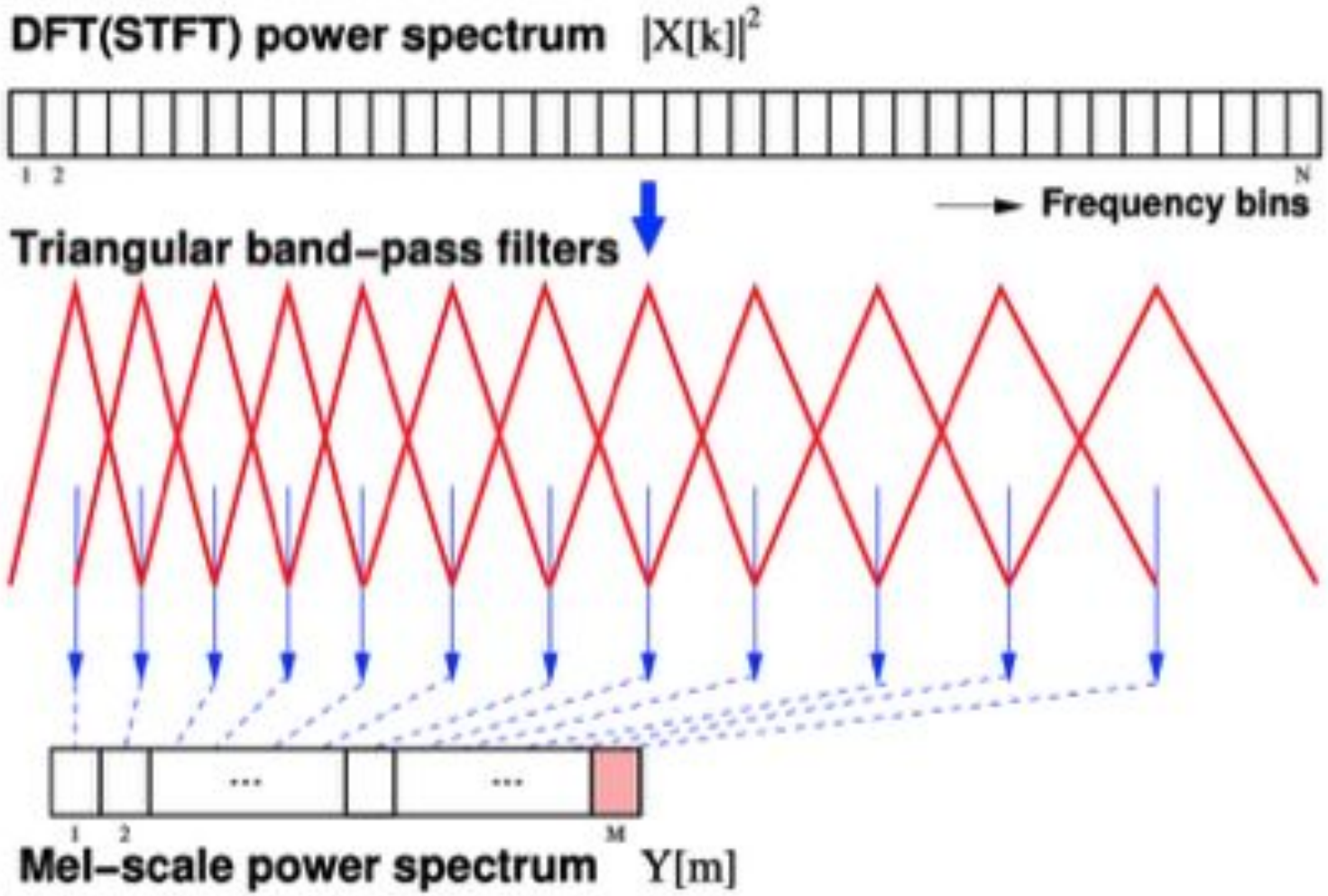
Windowing (with smoothing)










DFT(STFT) power spectrum |X[k]|²
1 2
N
Frequency bins
Triangular band–pass filters
1 2
M
Mel–scale power spectrum Y[m]

Альтернативы?

Wav2vec

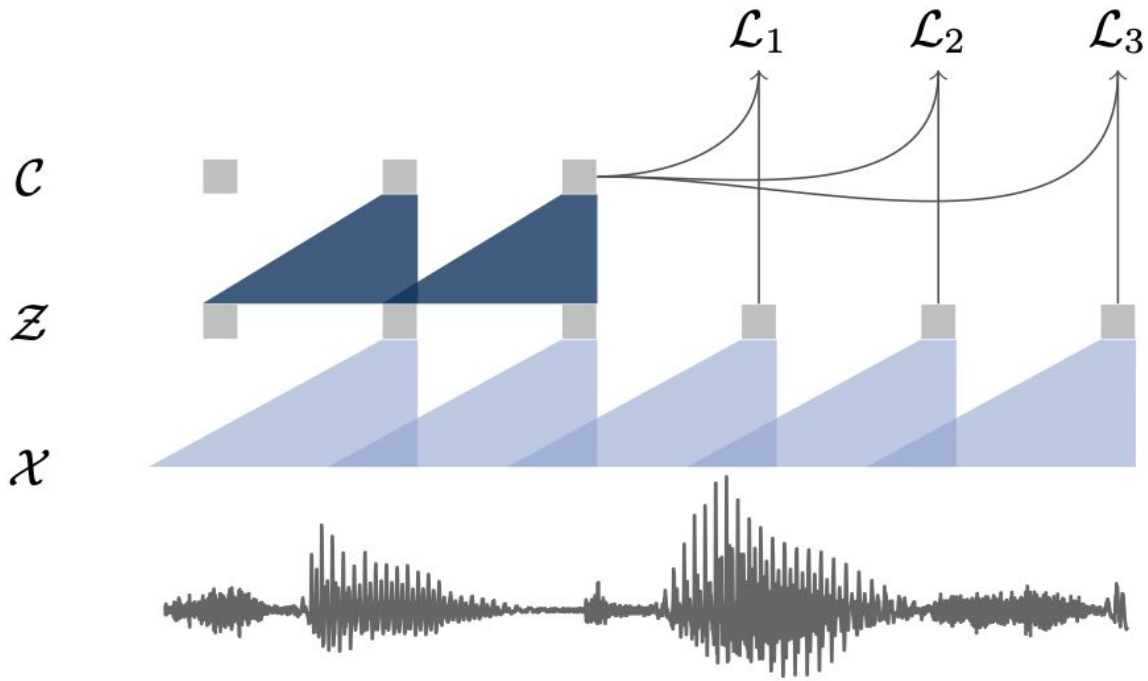


Figure 1: Illustration of pre-training from audio data $\mathcal{X}$ which is encoded with two convolutional neural networks that are stacked on top of each other. The model is optimized to solve a next time step prediction task.

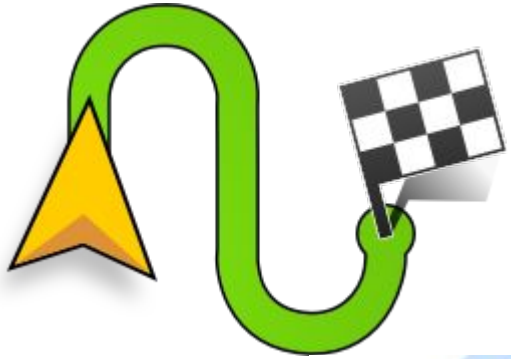
What's hot?

Распознавание речи

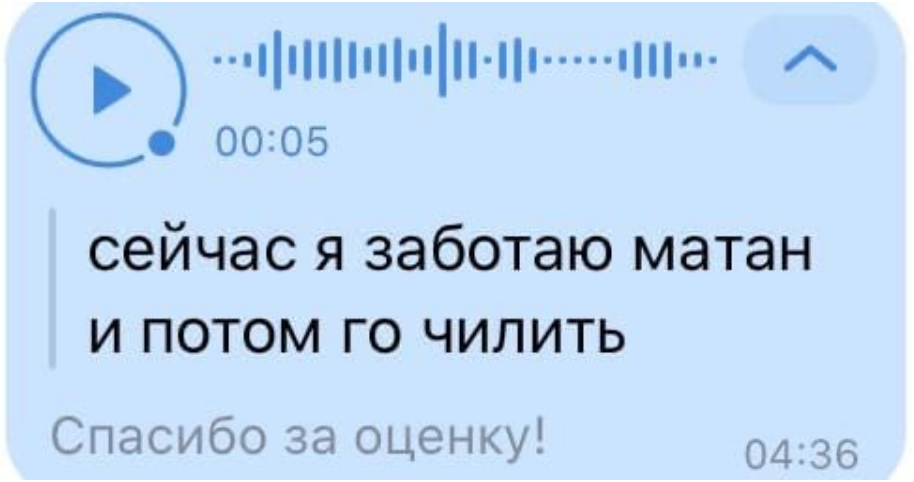
Генерация речи

Speaker features

Перенос стиля голоса

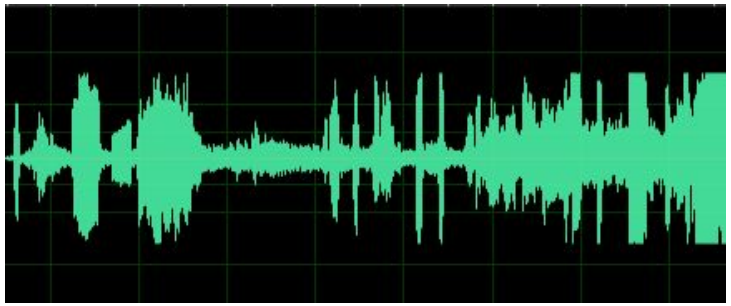


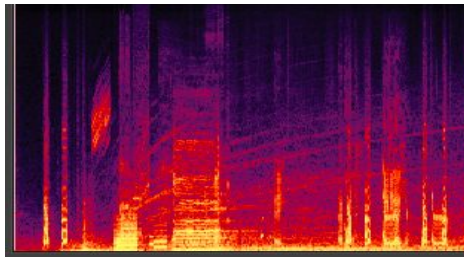

yes your honor but we made i
Subtitles/CC
Английский (автоматические титры)
ВКЛ
Параметры
6:02 / 6:19


00:05
сейчас я заботаю матан
и потом го чилить
Спасибо за оценку!
04:36

Tinkoff

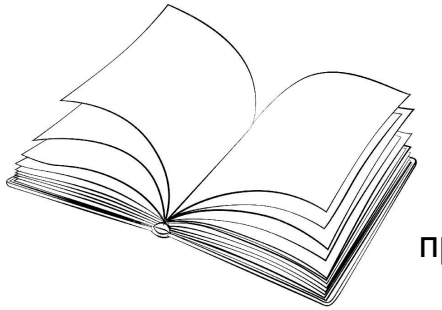
Как работает ASR





приве янадя

акустическая
модель



привет я надя

языковая
модель

profit

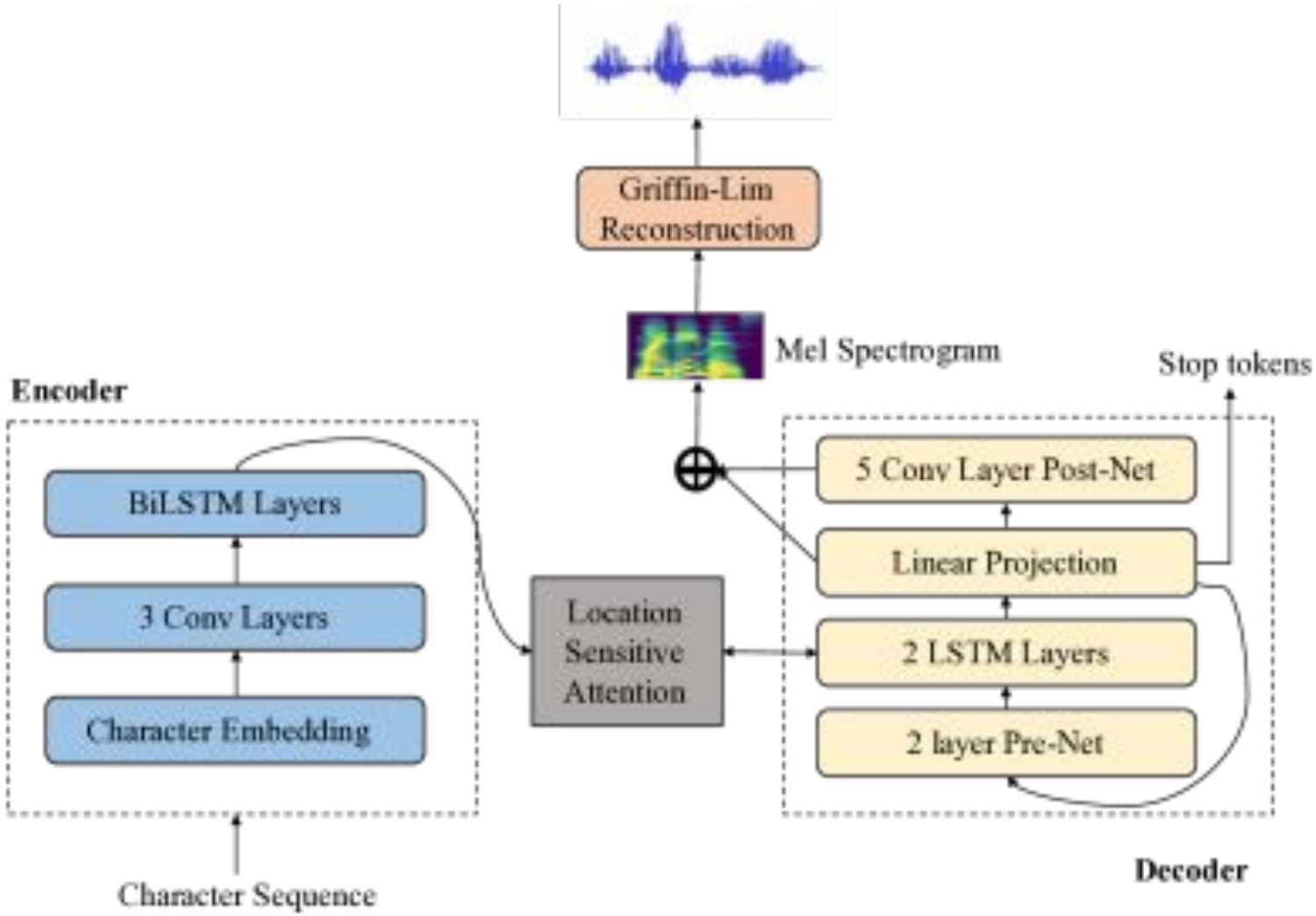
Распознавание речи

Генерация речи

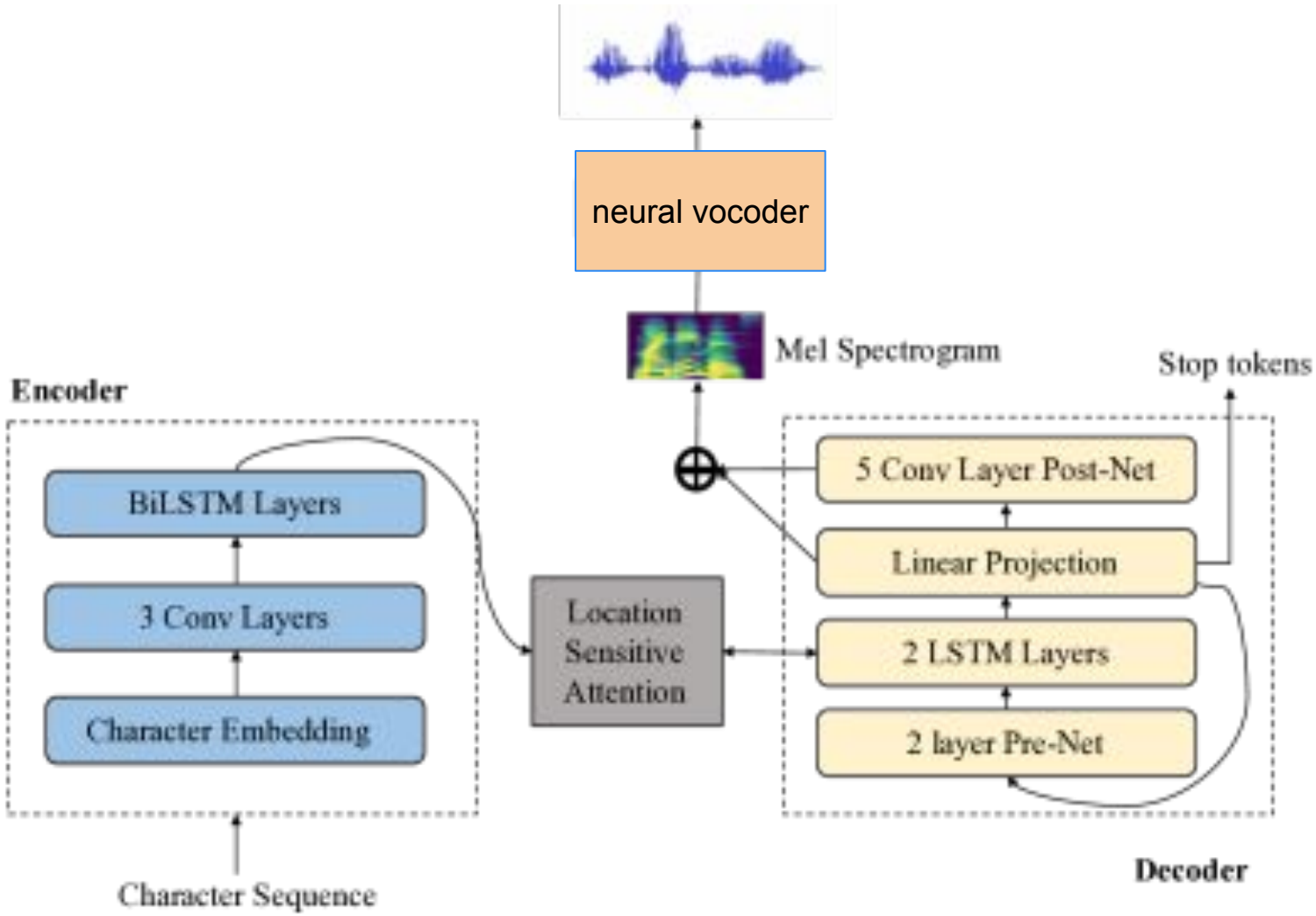
Speaker features

Перенос стиля голоса

Как работает TTS



Как работает TTS



Распознавание речи

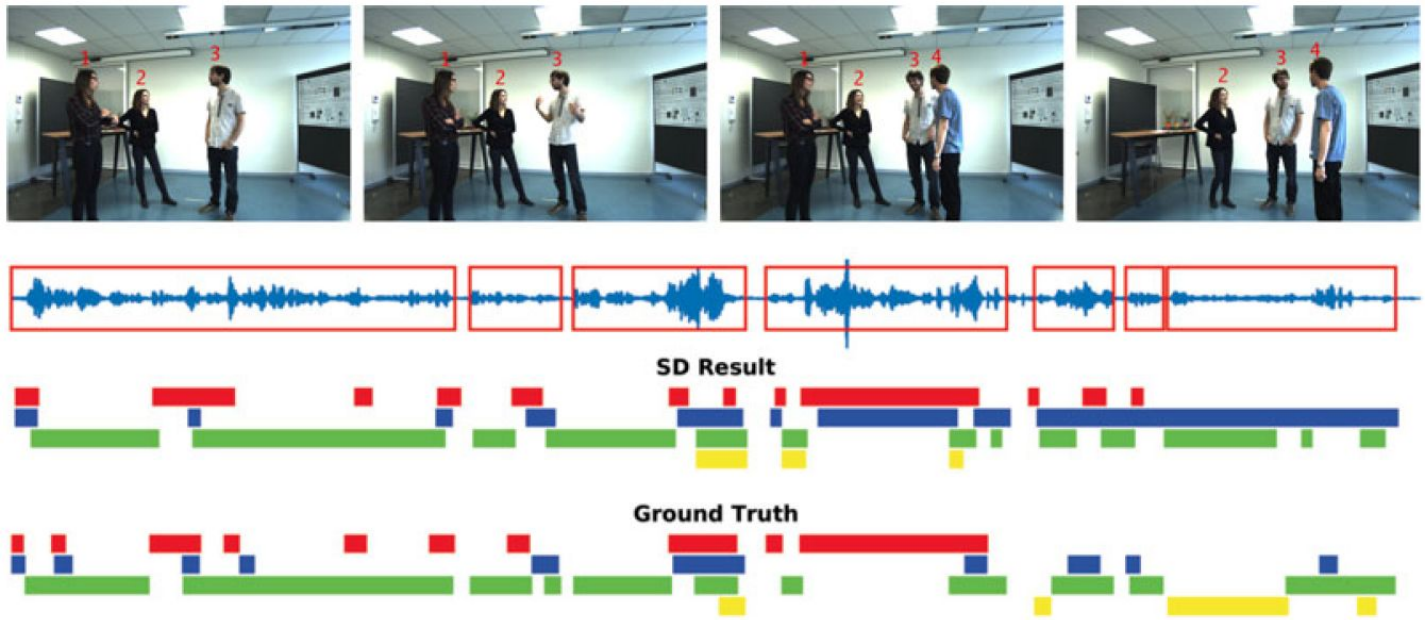
Генерация речи

Speaker features

Перенос стиля голоса

Speaker features

1. Идентификация
2. Диаризация
3. Speaker embeddings
4. ...



Распознавание речи

Генерация речи

Speaker features

Перенос стиля голоса

https://github.com/andabi/deep-voice-conversion

Intro

What if you could imitate a famous celebrity's voice or sing like a famous singer? This project started with a goal to convert someone's voice to a specific target voice. So called, it's voice style transfer. We worked on this project that aims to convert someone's voice to a famous English actress Kate Winslet's voice. We implemented a deep neural networks to achieve that and more than 2 hours of audio book sentences read by Kate Winslet are used as a dataset.

Что еще?

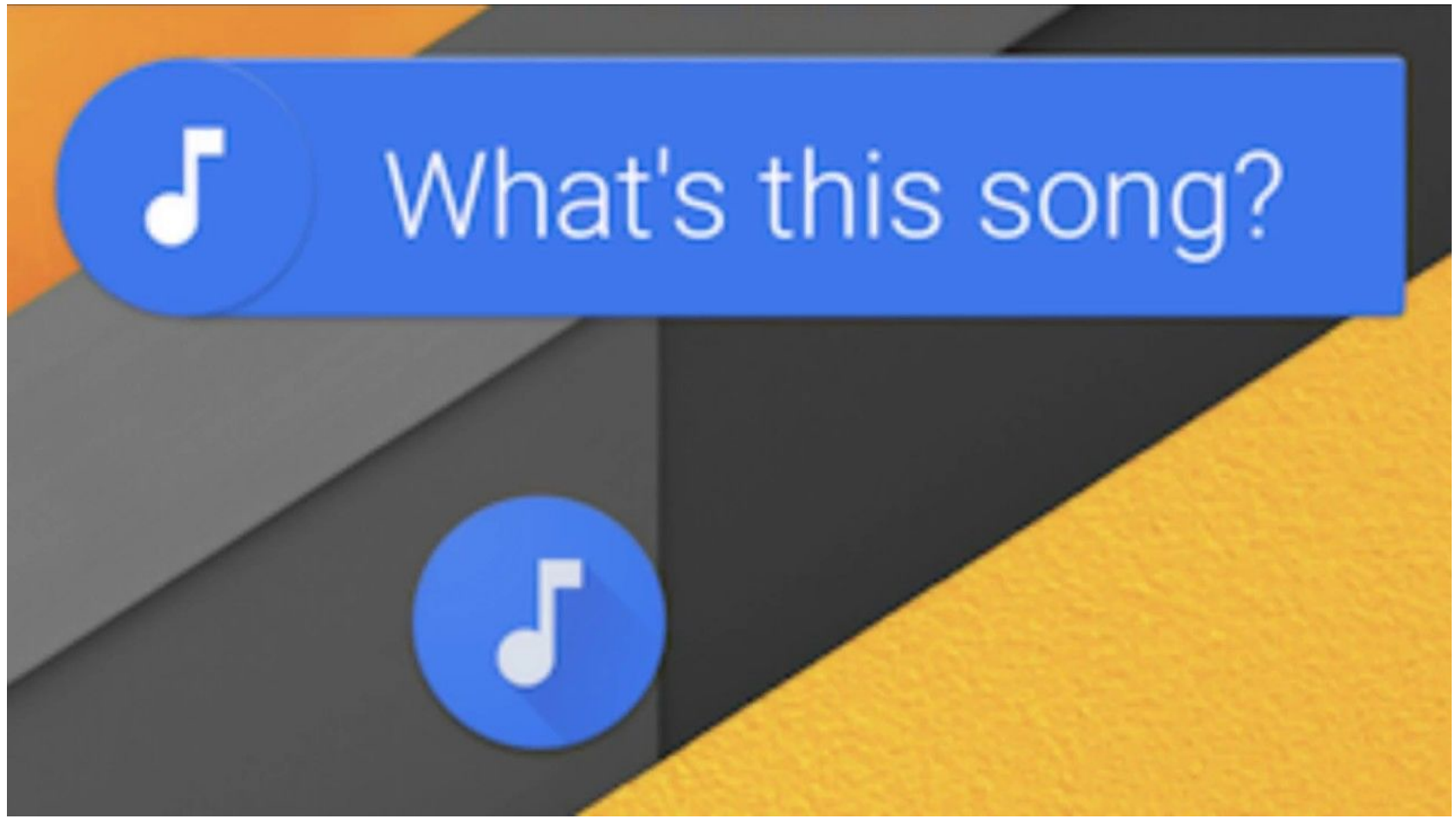
Jukebox

We’re introducing Jukebox, a neural net that generates music, including rudimentary singing, as raw audio in a variety of genres and artist styles. We’re releasing the model weights and code, along with a tool to explore the generated samples.

READ PAPER VIEW CODE

APRIL 30, 2020
12 MINUTE READ, 10 DAY LISTEN

Spleeter by deezer

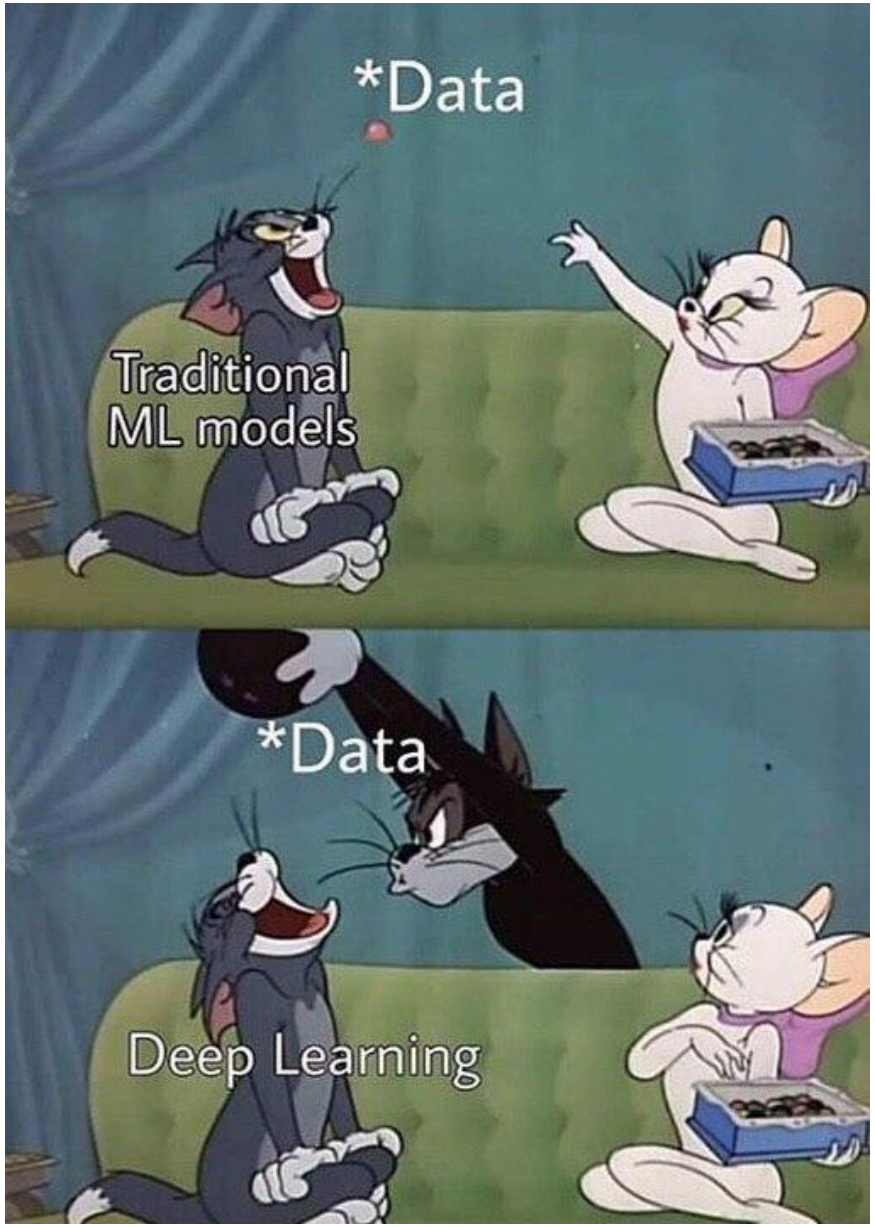

What's this song?

Маленькие, но важные задачи

1. Spotter
2. Voice Activity Detection (VAD)
3. End of Utterance
4. Поиск акустических событий (декомпозиция)
5. Биометрические задачи
    a. пол
    b. возраст
    c. whatever you want

Данные

Больше данных!



Больше данных!

benchmark datasets (EN)

1. TIMIT (4 hours)
2. LibriSpeech (1000 hours)
3. Librivox
4. TED
5. ....

40 hours 1 wav

40*60*6 дорожек по 10 sec (можно получить с помощью aligner)

Russian Open Speech To Text (STT/ASR) Dataset

Arguably the largest public Russian STT dataset up to date:

- ~16m utterances (1-2m with less perfect annotation, see #7);
- ~20 000 hours;
- 2,3 TB (in `.wav` format in `int16`), 356G in `.opus`;
- A new domain - public speech;
- A huge Radio dataset update with **10 000+ hours**;
- (**new!**) Utils for working with OPUS;
- (**new!**) New OPUS torrent;
- (**new!**) New OPUS direct links;

Prove us wrong! Open issues, collaborate, submit a PR, contribute, share your datasets! Let's make STT in Russian (and more) as open and available as CV models.

Planned releases:

- Working on a new project with 3 more languages, stay tuned!


moz://a
deep speech

TORCHAUDIO

This library is part of the PyTorch project. PyTorch is an open source machine learning framework.

https://vk.cc/bVC2Hb

https://colab.research.google.com/drive/1yvCAPJquyDWQBjNPC8Vv96aITUBLI15I?usp=sharing

https://github.com/CorentinJ/Real-Time-Voice-Cloning

https://github.com/auspicious3000/autovc

https://github.com/NVIDIA/waveglow

https://github.com/NVIDIA/NeMo/tree/main/examples/asr/experimental

https://openai.com/blog/jukebox/ (musenet)