Yandex

SCHOOL OF DATA ANALYSIS

Speech Course

Andrey Malinin, Vladimir Kirichenko, Sergey Dukanov

7th February 2022

Introducing lecturers

Andrey Malinin

Senior Researcher at Yandex

Vladimir Kirichenko

Head of TTS at Yandex

Sergey Dukanov

TTS TeamLead at Yandex

Introducing seminarists

Evgenia Elistratova

Anastasia Demina

Ekaterina Ermishkina

Contacts

https://t.me/+MmobByonc7AzMDQ0

Science-fiction and Science Reality

Star Trek: Ships Computer
1966

2001 - A Space Odyssey: HAL-9000
1968

Film "Her"
2013

Yandex Alisa
2017

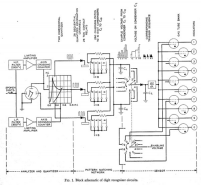
Audio

A Brief History of Speech Recognition

- *isolated words*
- filter-bank analysis
- time-normalization
- dynamic programming

Acoustic phonetics-based

50s – 60s



70s

- IBM Labs
- AT&T Bell Labs
- DARPA program

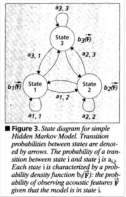
Template-based

- *connected digits*
- pattern recognition
- LPC analysis
- clustering algorithms

80s

- *connected words*
- hidden Markov models
- stochastic language modeling

Statistical-based



■ **Figure 3.** *State diagram for simple Hidden Markov Model. Transition probabilities between states are denoted by arrows. The probability of a transition between state* i *and state* j *is* a_{ij}*. Each state* i *is characterized by a probability density function* $b_i(\vec{y})$*: the probability of observing acoustic features* $\vec{y}$ *given that the model is in state* i*.*

90s

- *continuous speech*
- statistical learning

Syntax semantics

00s

- very large vocabulary
- CTC loss [Graves]: 2006

Machine learning

10s

- data-driven quality
- deep recurrent architectures
- Listen-attend-spell
- RNN-transducer

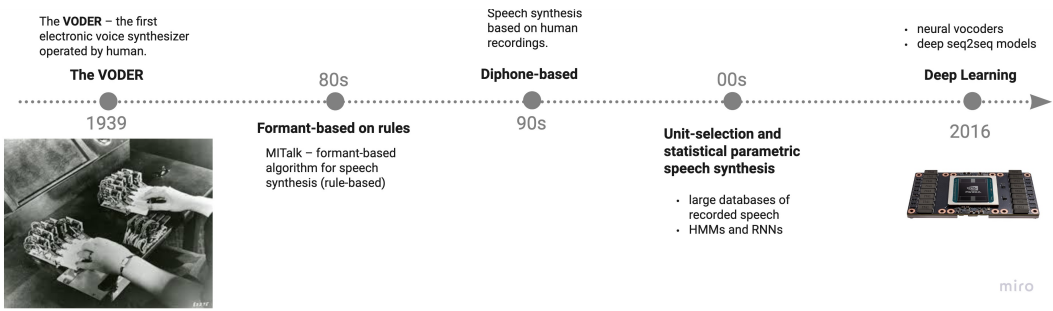
Deep learning

miro

▸ Video 1961 *Shoebox* IBM

▸ Paper *Fifty years of progress in speech and speaker recognition*

A Brief History of Speech Synthesis

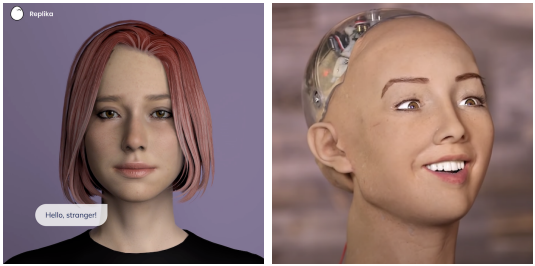


Video The VODER 1939

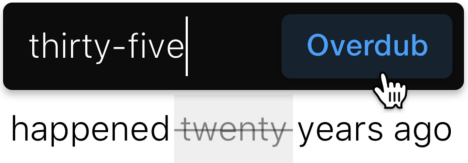
Video Ordering pizza in 1974

Video Signing up for a haircut in 2018

Voice Technologies Applications



- Virtual voice assistants in smartphones and smart speakers
- Interaction interfaces for persons with disabilities
- Video voice-over, voice acting in games, anthropomorphic robots, tools for voice processing (e.g. podcasting tools as descript.com Link)



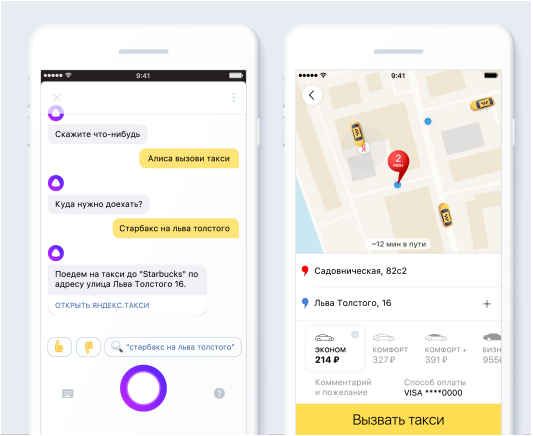
Smart Speakers Market and Voice Technologies Engagement

- Estimated number of sold **smart speakers** in the world by 2021 is up to **175 million**, in Russia it's up to **4.5 million** devices.[1]
- **41%** of millennials and **34%** of Gen X use smart speakers in the USA. [2]
- **4 Billion** pairs *user* $\leftrightarrow$ *virtual assistant* by the end of 2020 (estimated growth is **2x** in 4 years). [3]

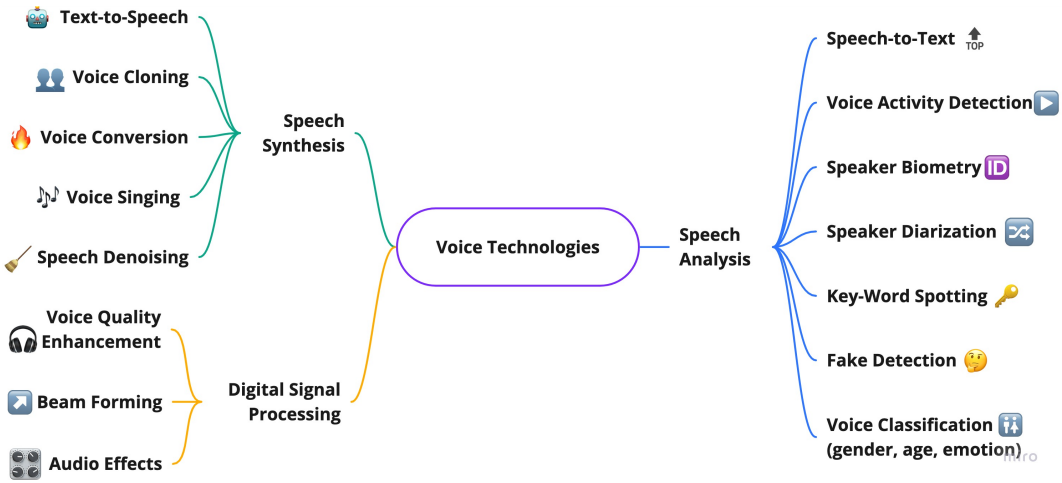
[1] Link JustAI report
[2] Link Emarketer report
[3] Link Juniper Research

Yandex Alisa

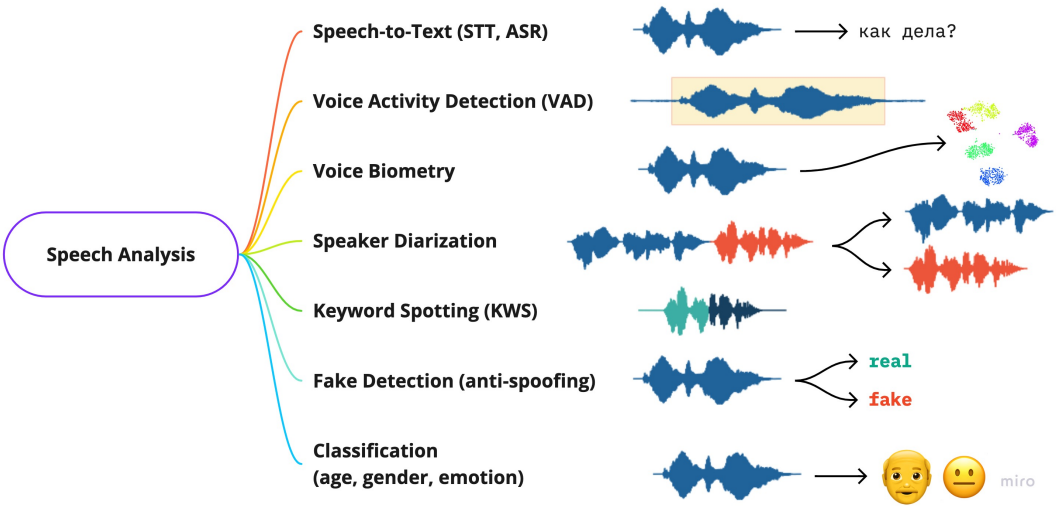
- **1 billion** user queries per month
- **45 million** monthly active users
- 3 000 skills (games, smart home, retail)



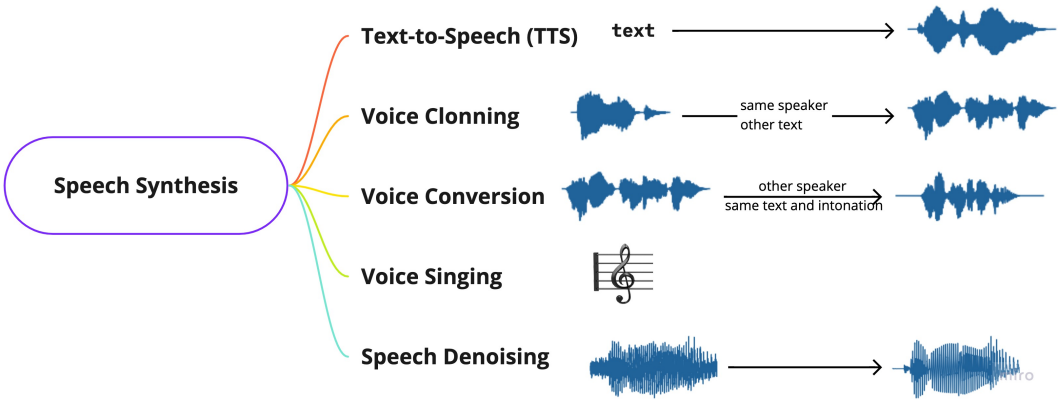
Voice Technologies Mind-map



Speech Analysis Mind-map



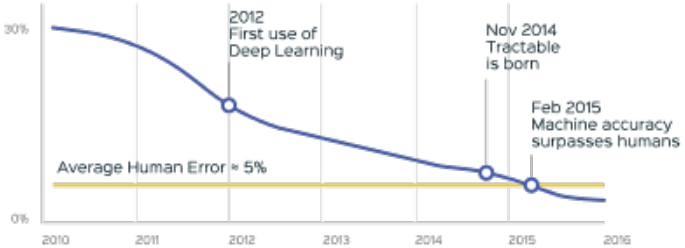
Speech Synthesis Mind-map



Voice Technologies Research Contributors

- **DeepSpeech** (*baidu, 2014*) – a large margin in quality due to NN
- **LAS** (*google, 2015*) – end-to-end speech recognition
- **WaveNet** (*deepmind, 2016*) – first neural vocoder (human-level quality)
- **Tacotron** (*google, 2017*) – first end-to-end speech synthesis model
- **Transformer** (*google, 2017*) – *"attention is all you need"*

Speech recognition



There are interesting domains **waiting for your efforts**!

Speech synthesis

Side-by-side comparison with the original speech sometimes gains in favor of TTS.

Speech Recognition State-of-the-Art

не подходи сюда между нами водораздел
если бы моя башка не висела на волоске
столько мяса 5 бы делал то что хотел
ему реплик чтобы не член если бы не город
вон напротив всех с копеечных мониторов
08 вовсе чего позор poser
но сколько рапир копро где-то до сих пор косяк
я жил где мечети и mighty и сенегал
теперь помоем приключением снимается сериал
я им сину пойманы в теле авантюриста
за болтаю всех долетели евангелистов
сколько слов было сколько зубов выбито
сколько всего выпито дома у ноу лимита
город flow мимика голос beata undine
каналов тебе не было фристайлистом права принято

Все, подходи сюда! Между нами водораздел.
Если бы моя башка не висела на волоске
столько гася вряд ли бы делал то, что хотел,
и мой рэп, что бы ничем, если бы не город,
он, напротив всех из копеечных мониторов
ноль восемь окси, чего позор, позер!
Но сколько рэперков под это до сих пор косят
я жил, где мечети ямайцы и сенегал.
Теперь по моим приключениям снимается сериал.
Я им, сину, поманите авантюриста,
заболтаю всех дарителей евангелистов.
Сколько слов было, сколько зубов и бита,
сколько всего выпито дома у нолимит
город флоу мимика, голос бетон или
канала к глине было приставить замру принято.

Audio – audio for transcription by Google ASR (**left**) and Russian ASR (**right**)

Speech Synthesis State-of-the-Art

- **Unconditioned speech synthesis**
  Audio What happens if generative model is not conditioned on text.
- **Unsupervised learning of emotions and styles** [4]
  Audio *United Airlines 563 from Los Angeles to New Orleans has Landed.*
- **Speaking rate control (at 10% supervision)** [5]
  Audio *So many people disregarded but it really great in so many different ways.*
- **Pitch variation control (at 10% supervision)**
  Audio *I'm gonna make a cake and lots of smoothies for these kids.*

[4] https://arxiv.org/abs/1803.09017
[5] https://arxiv.org/abs/1910.01709

Speech Synthesis State-of-the-Art

- **Unseen speaker voice cloning by single reference audio** [6]
  Audio *There were many editions of these works still being used in the nineteenth century.*
- **Cross-language voice cloning** [7]
  Audio English $\longrightarrow$ Spanish and Mandarin
- **Voice Conversion**
  Video Change speaker identity to another one.
- **Songs generatetion**
  Demo Synthesis song with 10-12s fragment seed.

[6] https://arxiv.org/abs/1806.04558
[7] https://arxiv.org/abs/1907.04448

Course Learning Objectives

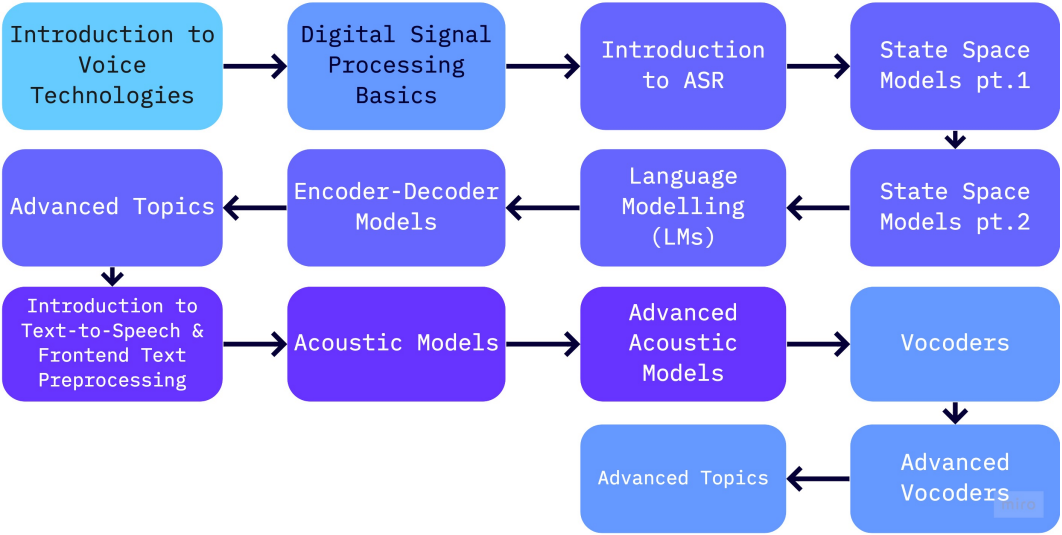
This course aims you to:

- delve into with speech as a new modality,
- understand key concepts of all up-to-date ASR and TTS algorithms,
- be able to implement such algorithms and build end-to-end ASR/TTS systems,
- know how to take apart algorithms in new publications for being up-to-date;

Organization Matters

- **15** lectures: **1** – Intro, **1** DSP, **5** – ASR, **6** – TTS, **2** – Keynote Lectures
- Seminar after **each** lecture:
  - Practical tasks (laboratory work)
  - Discussion of difficult topics
  - Questions about homework
- Homework - after **every second** lecture
- Deadline - soft, but with point loss.

Course Structure

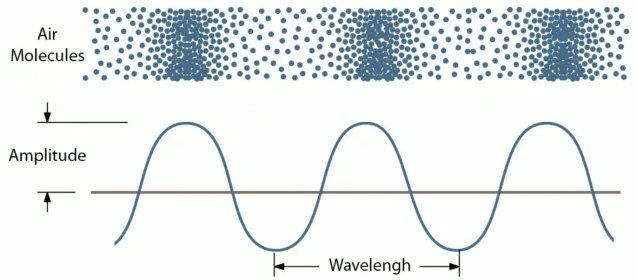


Five Minute Break!

Audio – cyberpunk table reservation in Russia

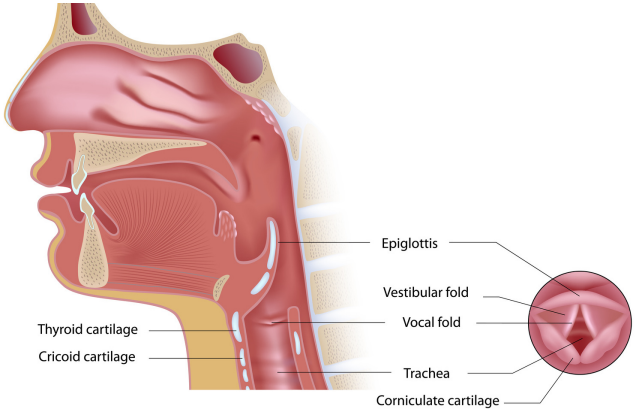
What is speech, how is it produced and how do we hear it?

Acoustic Physics



- Sound waves – longitudinal vibrations waves travelling through the air.
  - Defined by **Amplitude** (pressure) and **Frequency** (inverse of Wavelength)
- Humans produce sound using the vocal tract and hear using ears
- Can convert between electric signals and sound via speakers and microphones

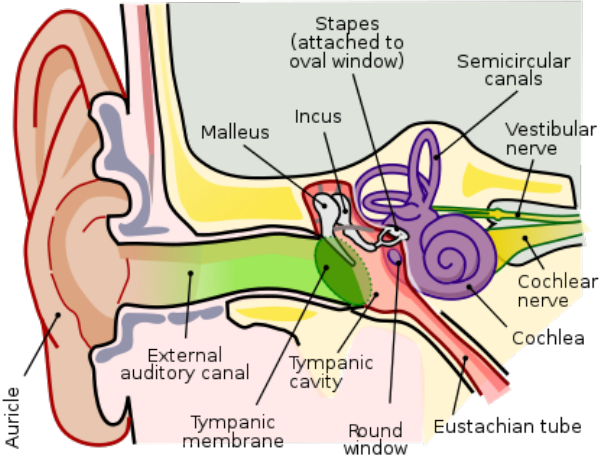
Human Vocal Tract



- Lets try to say something with this model:
  Pink Trombone.

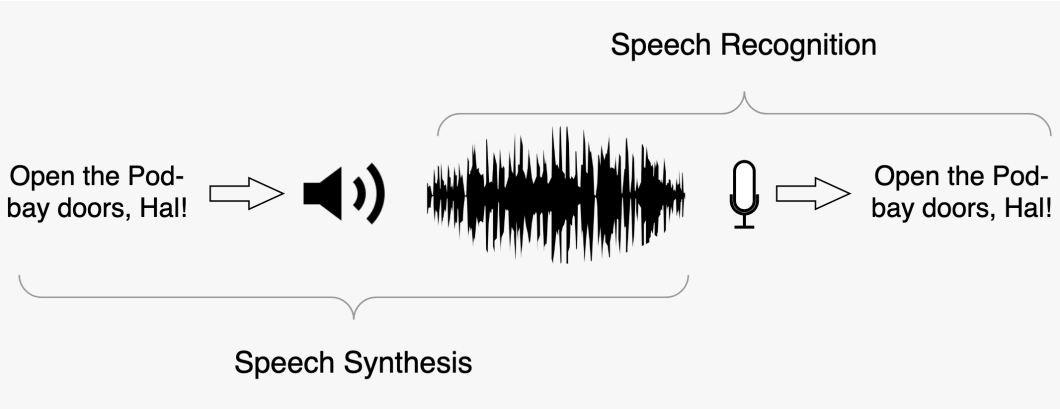
- **Source–filter model.**
- Tensioned *vocal folds* under air pressure from lungs is the **primary source** of oscillations (*voiced sounds*).
- Nasal and oral cavities act as a sound **filter**.
- The *tongue* and the *uvula* are **secondary sources** of oscillations.
- *Unvoiced sounds* are caused by turbulence of the airflow (near static constrictions).

Auditory system

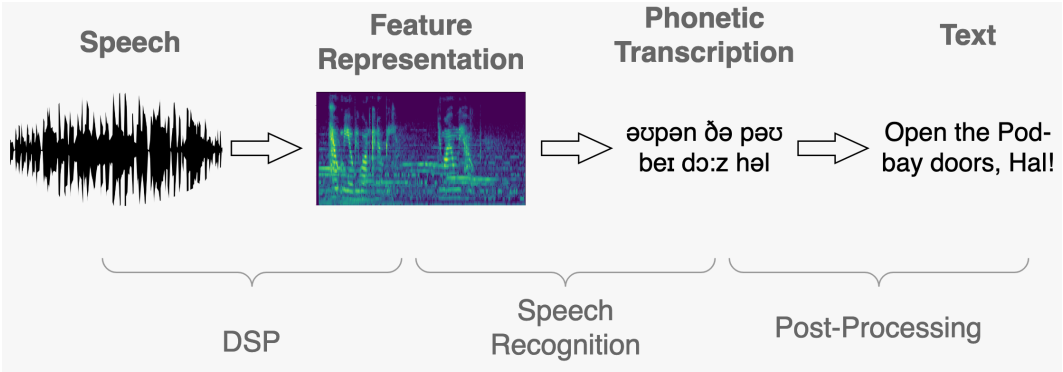


- The human ear can nominally hear sounds in the range **20 Hz** to **20 kHz** and **0 dB** ($\simeq$ 20 μPa) to **120 dB**.
- Perception of human ear is **logarithmic** both in *frequency* and *volume*.
- Also our volume perceptions depends on frequency of signal.

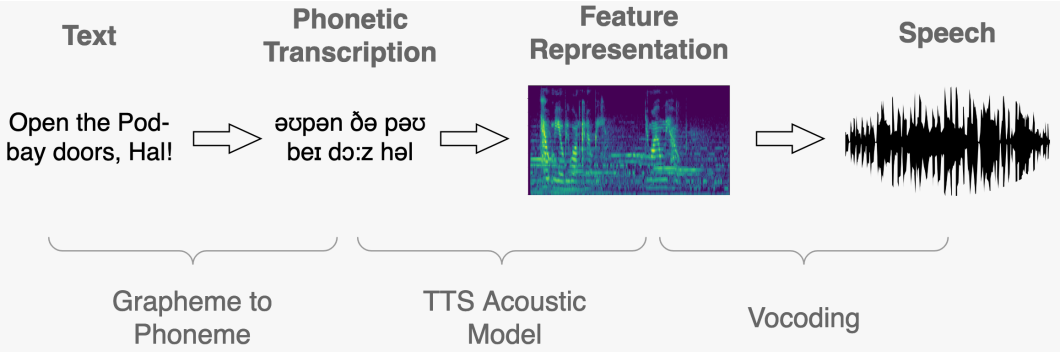
How can we recognize and synthesize speech using machines?


Speech Recognition
Open the Pod-
bay doors, Hal!
Open the Pod-
bay doors, Hal!
Speech Synthesis

High-level Speech Recognition Pipeline



High-Level Speech Synthesis Pipeline

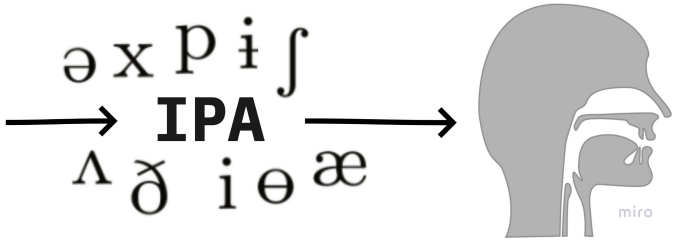


How can we write down sounds?

Phonetic Systems (IPA)

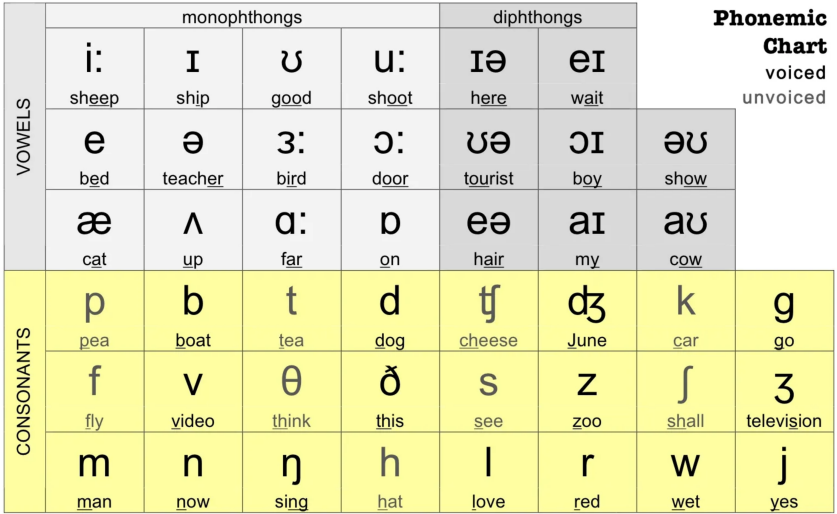
International Phonetic Alphabet (*IPA*) is notation for representation of speech sounds – it allows to formalize speech (all languages and accents).

Lorem ipsum dolor sit amet, consectetur adipiscing elit.



- **speech synthesis**: to resolve pronunciation ambiguities, improve model-data fitting;
- **speech recognition**: better understand different accents and better handle mispronunciations;

Phonetic Systems (IPA)



Grapheme-to-Phoneme (G2P)

это очень сложно для меня
ɛ+ t ə o+ t͡ɕ ɪ nʲ s lˠ o+ ʐ n ə d lʲ a+ mʲ ɪ nʲ a+

schwa

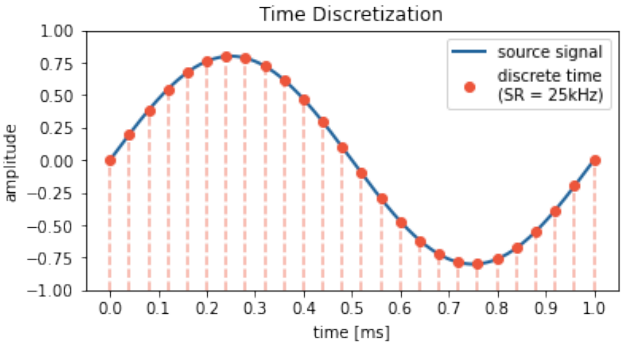
для назначения групповой встречи требуется
d lʲ a+ n ə z n ɐ t͡ɕ e+ nʲ ɪ j ə g r ʊ p ɐ v o+ j f s t rʲ e+ t͡ɕ ɪ t rʲ e+ b ʊ j ɪ t͡s ə

человек который любит выражать свои чувства
t͡ɕ ɪ lˠ ɐ vʲ e+ k k ɐ t o+ r ɨ j lʲ ʉ+ bʲ ɪ t v ɨ r ɐ ʐ a+ tʲ s v ɐ i+ t͡ɕ u+ s t v ə

хороший рейтинг
x ɐ r o+ ʂ ɨ j r ɛ+ j tʲ ɪ n k

- ə [schwa] – mid central vowel
- consonant unvoicing: g/k
- consonant palatalization
- unpronounceable graphemes

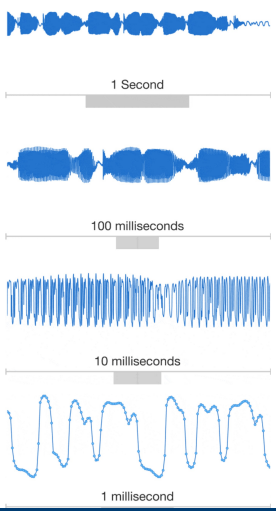
How can we efficiently represent speech?

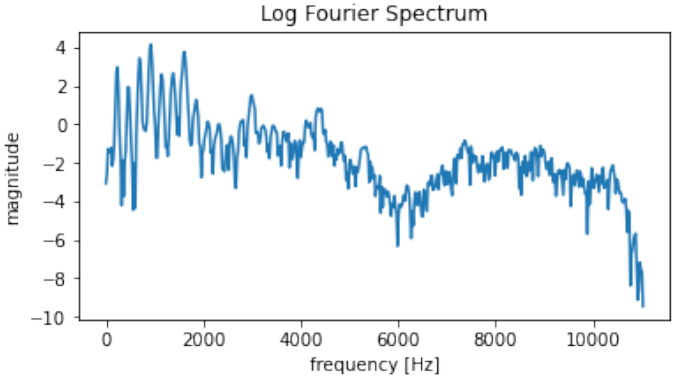


- Need to convert from continuous signal into a discrete series of amplitude values
  - We must sampling the signal with an appropriate **Sampling Rate**.
- Human speech contains frequencies from 20Hz to 20kHz
  - We need to sample the audio with a sample rate of at least 40kHz (typically 44kHz).

Digital Signal Processing (DSP) basics - Spectral Representation

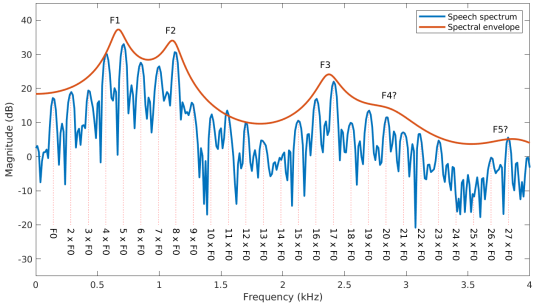
- Each second of speech contains of 44 thousand timestamps
  - This is very difficult to work with!
- Working in time-amplitude domain is also inconvenient.
  - Hard to separate signals
  - Periodic structure at different time-scales
  - Non-linear sensitivity to amplitude and frequency
  - Not robust to inaudible variations
- Need to convert sound into a better representation!
  - Mel-Spectogram
  - Mel-Frequency Cepstral Coefficients





- Take the Discrete Fourier Transform of a small signal widows
  - Yields the **frequency spectrum** of speech within the window
  - Vary window width for better frequency vs. time resolution

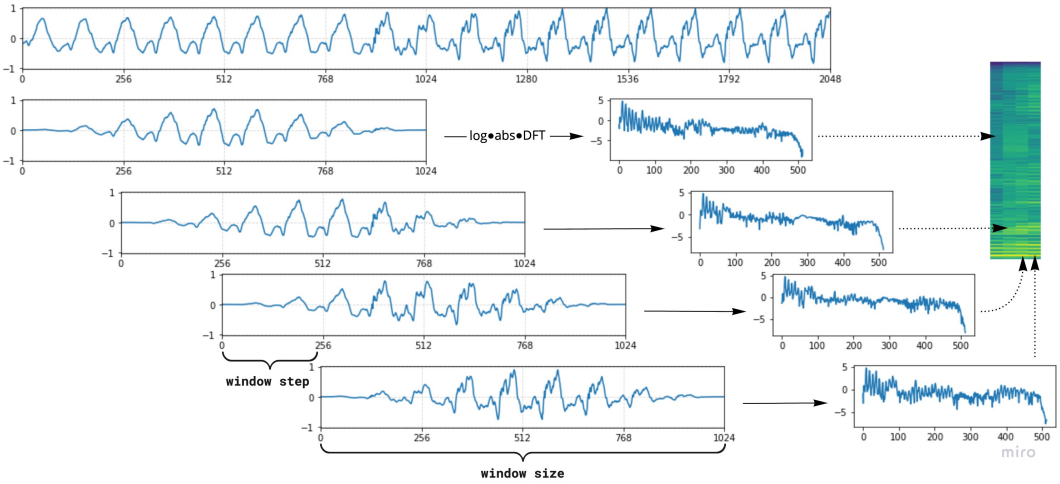
Fundamental frequency (F_0) and formants



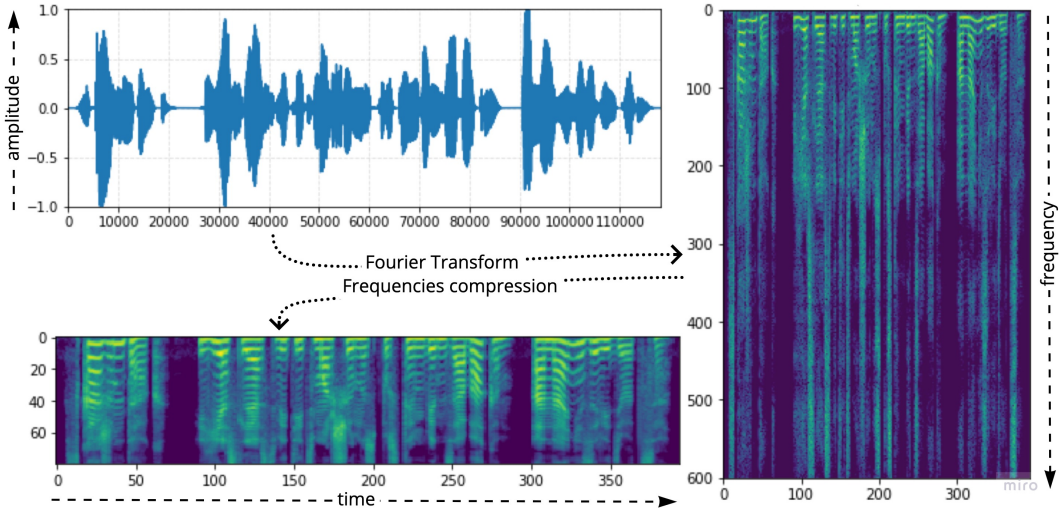
- **Blue curve** – spectrum (power of each of the frequencies).
- **Red curve** – spectrum **envelope**.

- F_0 – *fundamental frequency*.
- $k \times F_0$ – harmonics
- *Formants* produced by the resonances in the vocal tract.
- Auditory illusions: ▸ Auditory illusions

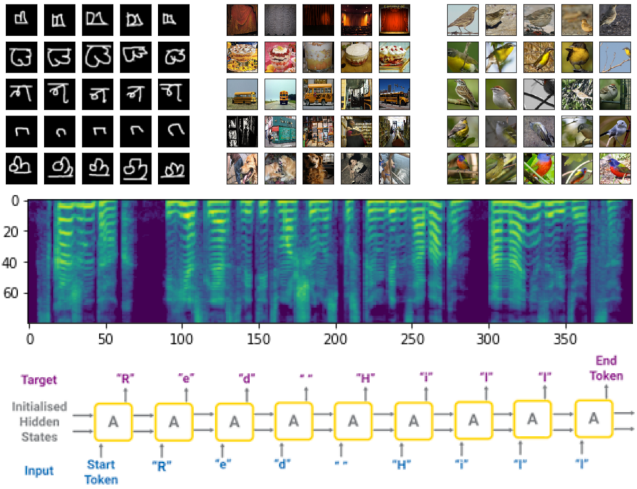
DSP basics - Spectograms



DSP basics - Time and Frequency Compression

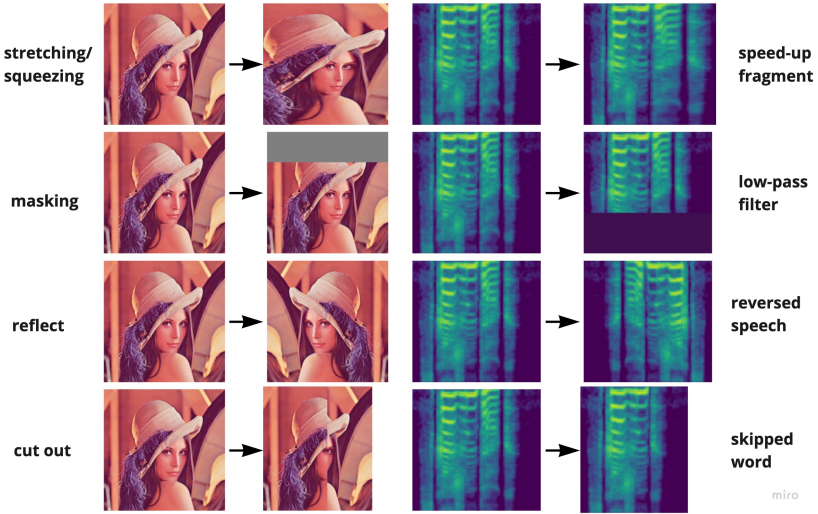


How is speech different from text and images?

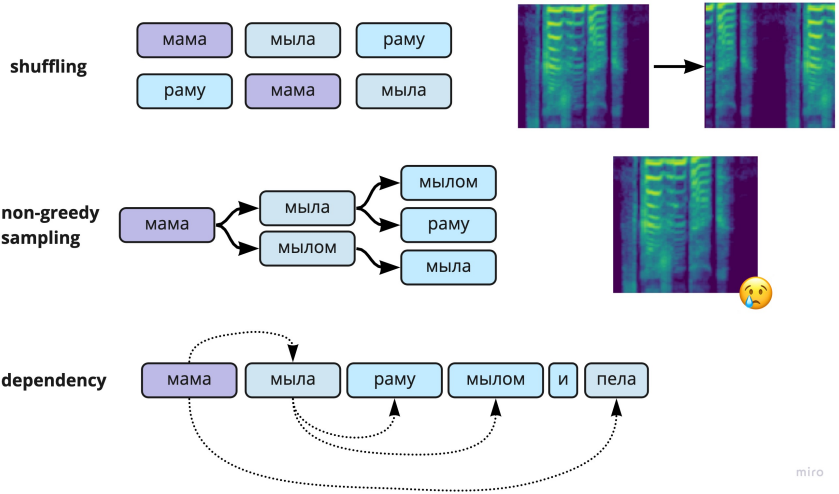


- Repeating patterns, spatial information, information redundancy → using CNN techniques
- Sequential information → using RNNs
- *But!* Applying these techniques "as is" will not lead to good results.

Speech compared to natural images

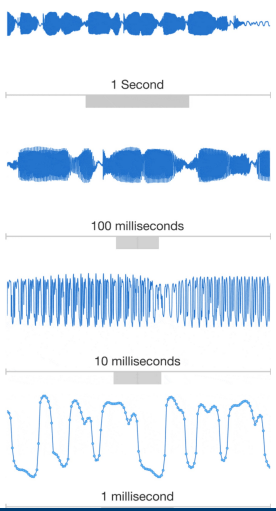


Speech compared to natural texts



Dimensionality

- Hundreds of thousands timestamps $\rightarrow$ hard to analyse
- We are trying to minimize information loss while converting to previously described representations
- This compression is lossy $\rightarrow$ use specific models (*vocoders*) for reconstruction



Lecture summary

- Speech recognition and synthesis retrospective.
- Speech technologies are being in active development.
- Speech is another modality (different from images and texts).