Advanced Topics in Audio Processing using Deep Learning

Tal Rosenwein, Oct 2025

About Myself

- **Personal info:** Married + 3, Herzliya,
- **Hobbies:**
  - Martial arts (since 5 YO), Running
- **IDF**
- **Education:**
  - Bsc in Biomedical Engineering (signal processing) @BGU.
  - Msc in Biomedical Engineering (signal processing and ML) @BGU.
- **Experience:**
  - Co-Founder - Stealth
  - OrCam
    - Algorithms developer / researcher in CV
    - ASR team lead
    - Group lead (ASR + OCR + tech transfer)
    - Initiated the NLP, speech enhancement and source separation
    - VP AI and Algo (Research)
    - CEO - Hear
  - Consulting
  - Lecturer (TAU)

LinkedIn

Motivation

- Knowledge distillation
- Make sure you apply it to good causes only, and understand the border impact of such technology.

source

Course Goals

- Create a positive and supportive environment where you can experiment, grow, and achieve accomplishments you can be proud of.
- “Observe” audio instead of “listen” to audio
- Be able to understand and implement relevant papers

Syllabus

Home assignments (15%):

- Due in groups of 4
- Binary score

Final project (85%):

- Groups of 4
- Grades 0-100
- +presentation (TBD)

Office hours: TBD (contact me)

*One cannot pass the course if the final project grade is below passing criteria.

Syllabus

Main Topics (12 lectures):

• Motivation and Course Overview	1 lecture
• Digital Signal Processing (DSP)	4 lectures
• Traditional DSP for Audio	1 lecture
• Automatic Speech Recognition (ASR)	2 lectures
• Speech Enhancement/Source Separation	1 lecture
• Text-to-Speech (TTS)	1 lecture
• Generative AI: Audio Applications	1 lecture
• Conversational AI	1 lecture

Link

*each lecture will have its own supplementary materials. I will send the references.
*This course covers numerous subjects, but time constraints necessitate brief presentations for some topics.


FUN
AHEAD

Audio Research





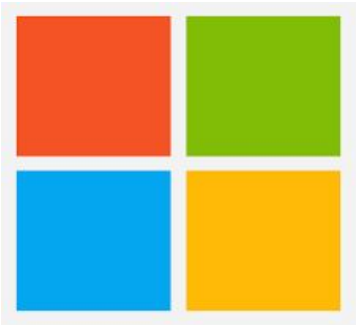
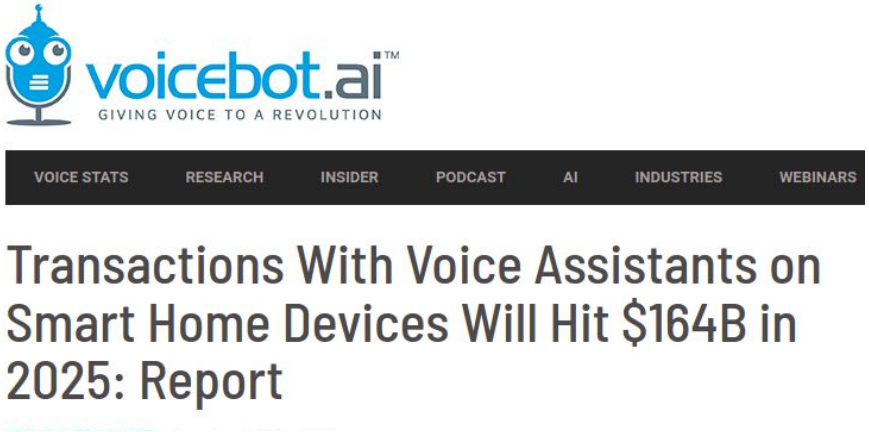



Image source [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#)

Audio - Market Size

Image source1 source2 source3

Audio - Market Size



With the growth of LLMs and rapid adoption of generative AI innovations, digital-native consumers are increasingly demanding a more natural and intuitive way to interact with applications. Gartner predicts that by 2025, 70% of digital and marketing communications will be supported by Avatars using text-to-video Generative AI, up from less than 5% in 2022.

Image source1 source2 source3

Technology Strikes in Several Verticals

Image Source (1) (2) (3)

Technology Strikes in Several Verticals

VNI Complete Forecast Highlights

cisco

- Globally, consumer Internet video traffic was 56.4 EB per month in 2017, the equivalent of 14 billion DVDs per month, or 19 million DVDs per hour.
- Global consumer Internet video traffic grew 40% in 2017.
- Globally, Internet video traffic will be 82% of all consumer Internet traffic by 2022, up from 73% in 2017.

Image Source (1) (2) (3) (4)

Technology Strikes in Several Verticals

- Availability of data
- Computational power
- Progress in research

Image Source (1) (2) (3)

Data-Driven Approach

Shifting from Expert Craftsmanship to Data-Driven Mastery: AI and Deep Learning Revolutionizing Solutions.

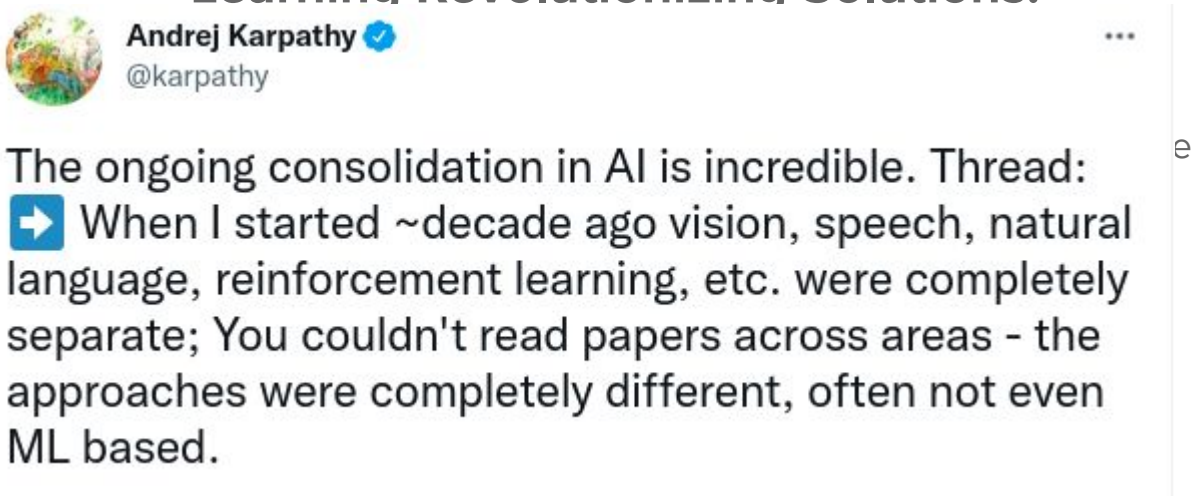
- Domain experts:
  - Hand craft the features / simulate physics to ‘learn’ statistical behaviour of the world
  - Rule-based models (in some cases really good ones -ASR, codecs, etc.)
- Data-driven
  - Clean data set + Computational power (expressive models) = good results
  - Reduce domain knowledge

Data-Driven Approach

Shifting from Expert Craftsmanship to Data-Driven Mastery: AI and Deep Learning Revolutionizing Solutions.

- Domai
  - Ha
  - Ru
- Data-d
  - Cle
  - Re

e world



Data-Driven Approach

Shifting from Expert Craftsmanship to Data-Driven Mastery: AI and Deep Learning Revolutionizing Solutions.

And still, in real-world applications, engineering is a must.

- This is why this course will cover some classic methods
  - In many real world applications, there is not enough data to do ‘data driven’ approaches: low resources languages, applications where there is not enough data, etc.

Data-Driven Approach

Shifting from Expert Craftsmanship to Data-Driven Mastery: AI and Deep Learning Revolutionizing Solutions.

And still, in real-world applications, engineering is a must.

- This is why this course will cover some classic methods
  - In many real world applications, there is not enough data to do ‘data driven’ approaches: low resources languages, applications where there is not enough data, etc.
- When one needs to be SoTA, the combination of classic and E2E methods sometimes outperforms each individual approach.

APDL, Tal Rosenwein

source

Data-Driven Approach

Shifting from Expert Craftsmanship to Data-Driven Mastery: AI and Deep Learning Revolutionizing Solutions.

To deploy models that serves products, several factors must be achieved:

- **On-cloud vs on-prem costs**- Hi Siri, Alexa, etc. runs on edge

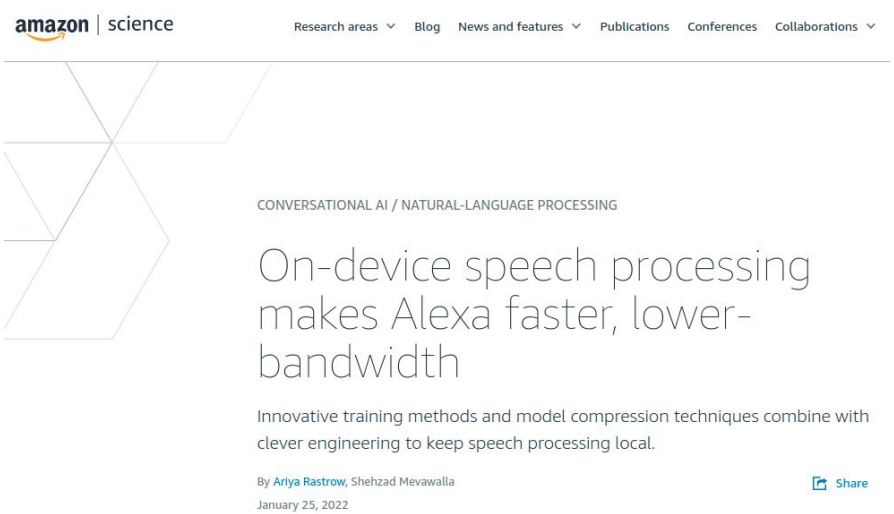


Image Source1 source2

Data-Driven Approach

Shifting from Expert Craftsmanship to Data-Driven Mastery: AI and Deep Learning Revolutionizing Solutions.

To deploy models that serves products, several factors must be achieved:

- **On-cloud vs on-prem costs**- Hi Siri, Alexa, etc. runs on edge
- **Performances** - Google didn't replace their traditional ASR models for few years because their hand-crafted models outperformed the E2E models
- **Explainability** - regulations such as in healthcare, credit score (explainable AI hype).
- **Safety**

Course Overview- Bites

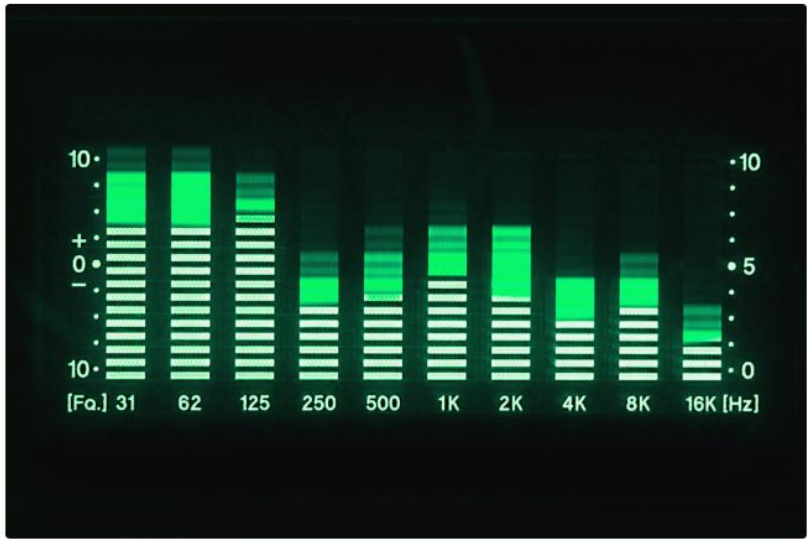
Image Source

Course Overview- Frequencies

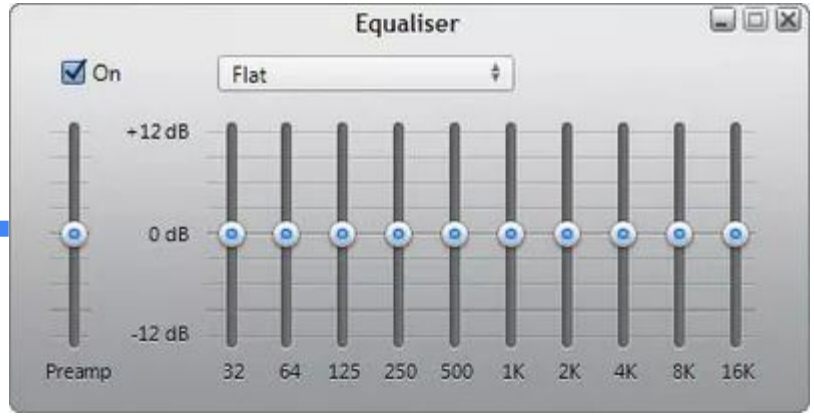
https://www.szynalski.com/tone-generator/

Course Overview- Frequencies

https://www.szynalski.com/tone-generator/



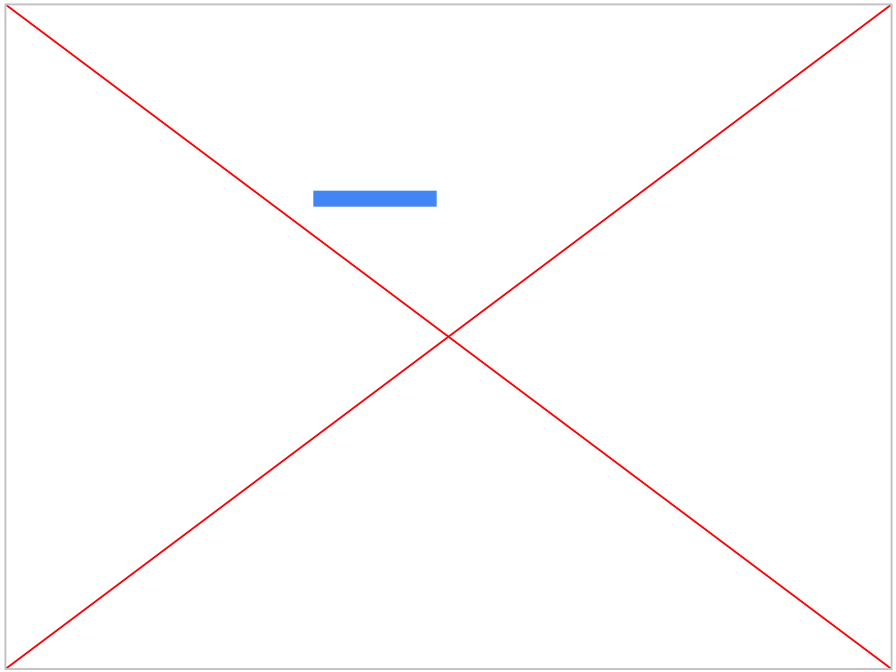
Steven Puetzer/Getty Images



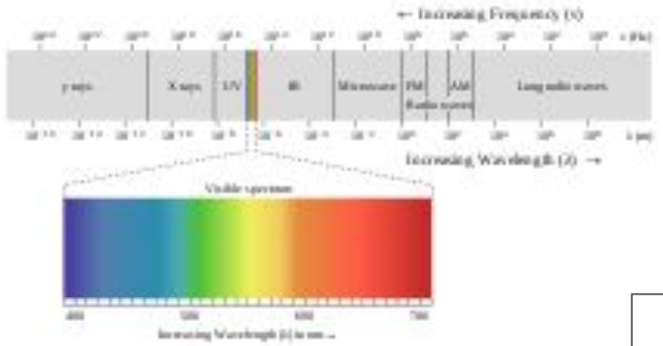
Source 1, 2

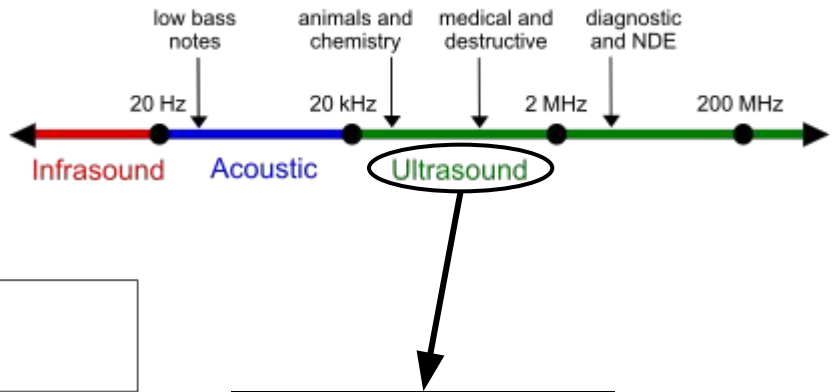
Course Overview- Frequencies

https://www.tiktok.com/@lewitt.audio/video/7213768652880022789



Course Overview- Frequencies

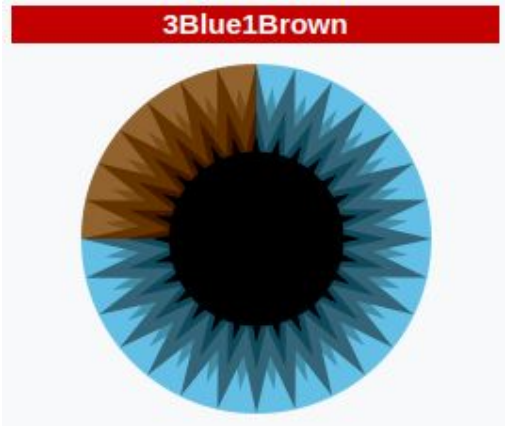




Source [1](), [2](), [3]()

Course Overview- Frequencies

https://www.youtube.com/embed/spUNpyF58BY?start=50&end=150

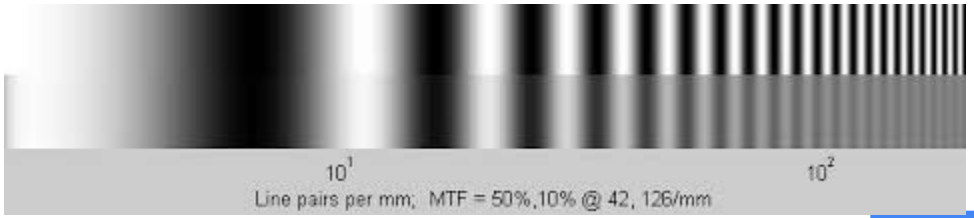


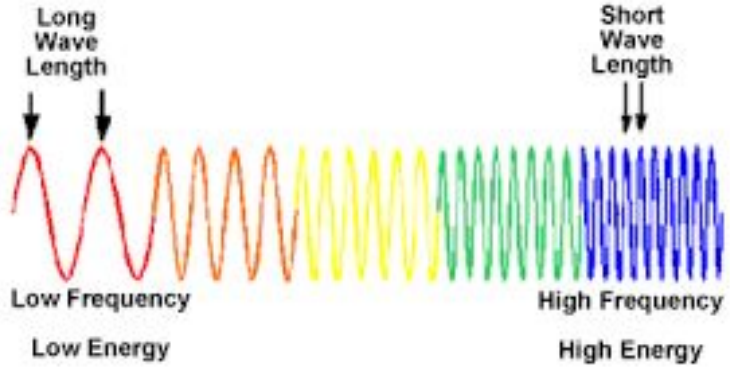
Source 1

Course Overview- Frequencies

The frequency of a sinusoid $\cos 2\pi f_0 t$ or $\sin 2\pi f_0 t$ is f_0, and the period is $T_0 = 1/f_0$. These sinusoids can also be expressed as $\cos \omega_0 t$ or $\sin \omega_0 t$, where $\omega_0 = 2\pi f_0$ is the *radian frequency*,

Course Overview- Frequencies





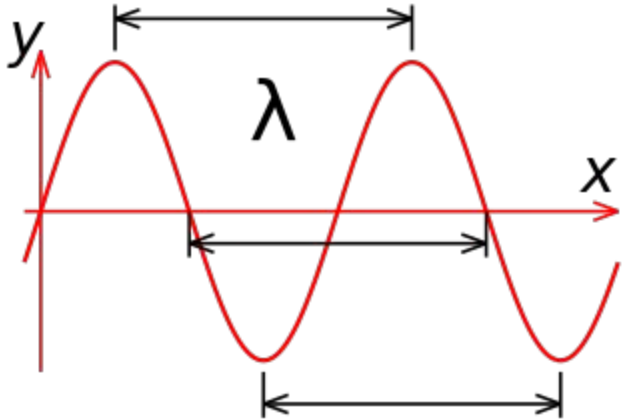
(NOTE: Frequency refers to number of crests of waves of same wavelength that pass by a point in one second.)

Source 1, 2

Course Overview- Frequencies

Cos/Sin with frequency of 10Hz has a **wavelength** of ~30 meters.

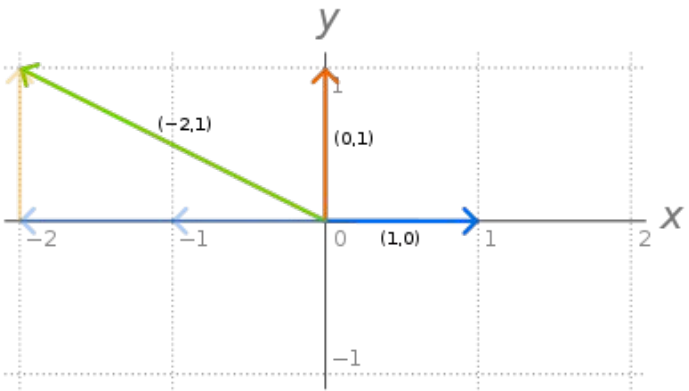
- Recall that the speed of sound is *c*=343 [m/s]
- *c[m/s]=λ[m]f[1/s]*
- If we set *f=10Hz; Hz=1/sec*
- 343[m/s]=λ[m]f[1/s] ->f = 10[1/s] →λ = 34.3[m]



Source

Course Overview- Frequencies

Basis of vector space



Each sample *Si* can be represented as a linear combination of the basis vectors.

$$S_i = \Sigma \alpha_i V_i$$
$$V_i - basis\ vector$$
$$\alpha_i - weight\ of\ V_i$$

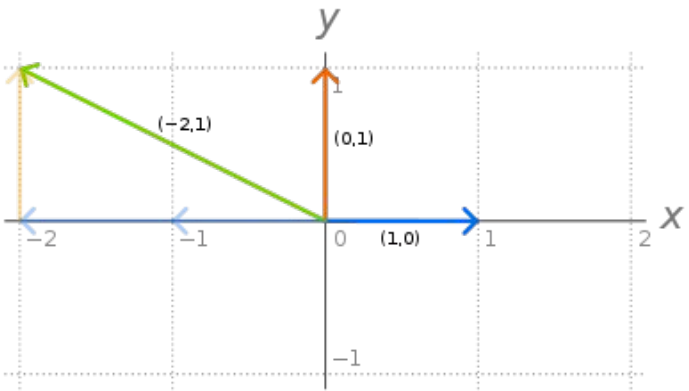
In order to find alpha_i we PROJECT the sample over Vi

$$\alpha_i =< V_i, S_j >= \Sigma_l V_{il} S_{jl}$$

Source

Course Overview- Frequencies

Basis of vector space



Each sample *Si* can be represented as a linear combination of the basis vectors.

$$S_i = \Sigma \alpha_i V_i$$
$$V_i - basis\ vector$$
$$\alpha_i - weight\ of\ V_i$$

In order to find alpha_i we PROJECT the sample over Vi

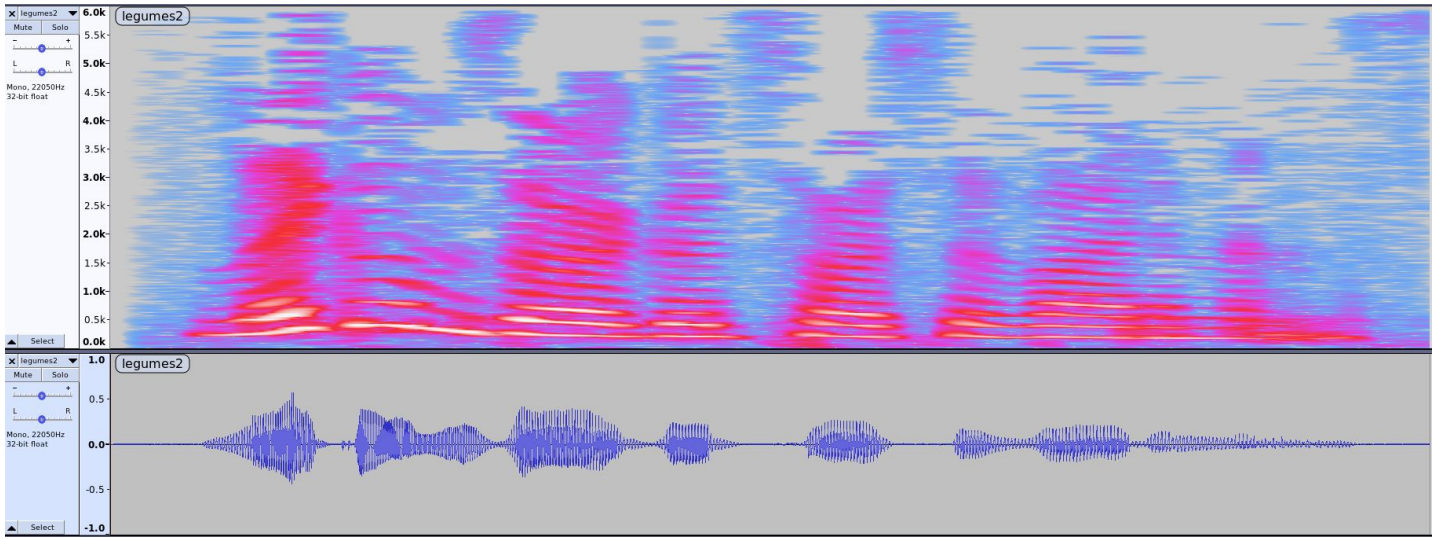
$$\alpha_i =< V_i, S_j >= \Sigma_l V_{il} S_{jl}$$

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i 2\pi \frac{k}{N} n}$$

Source

Course Overview- Audio DSP

- Audio (Zoom in, Zoom out)
- Spectral View (Heatmap)
- Pitch (voiced Vs unvoiced)
- Formants



Course Overview- Audio DSP

- Listening Vs Recording using a Microphone

Course Overview- ASR Levels Of Difficulty

Keywords spotting

Closed Set Voice Commands

Open Domain Questions

'Sterile' LVCSR

Real Life LVCSR

Hey Siri

Say "Hey Siri, send a message."

Image source 1 2 3 4 5

Course Overview- ASR As a System

So we have the best ASR engine, but now what?

“you know there's something interesting about you i had to come meet you hmm huh what no there's something interesting about you i just wanted what to come meet you i don't know i just got that vibe that's weird is that okay no it's not oh oops”

Course Overview- ASR As a System

Making The Transcript More Intelligible

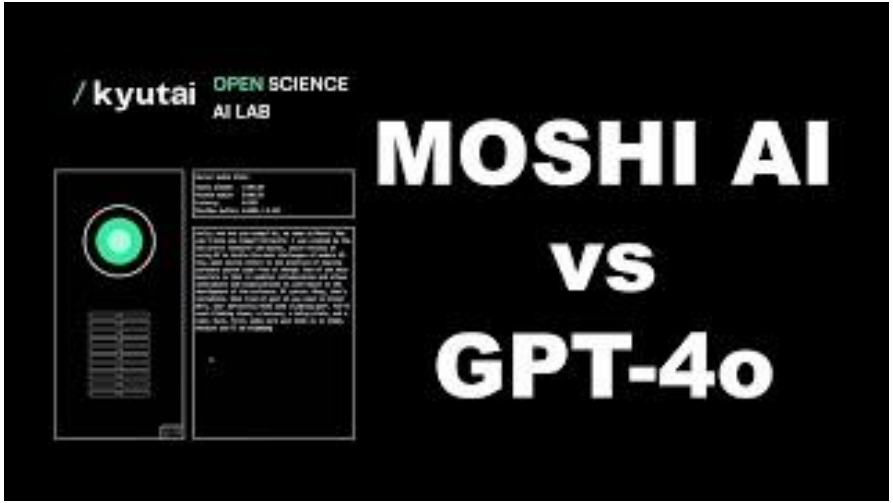
A: You now there's something interesting about, you I had to come meet you.
B: Hmm?
A: Huh?
B: What?
A: No, there's something interesting about you I just wanted
B: What?
A: to come meet you. I don't know, I just got that vibe
B: That's weird.
A: Is that okay?
B: No.
A: it's not. Oh, oops.

Course Overview- ASR: Current Challenges

Speech Impairment

Accents

Course Overview- NLU: Conversational AI



Course Overview- NLU: Conversational AI

Course Overview- TTS

Course Overview- Speech Enhancement

Link (SoTA blind source separation)

Course Overview- GenAI: Music Generation

Link (Melody conditioning)

Questions?

Tal Rosenwein, Jan 2024