Practical AI: speech processing

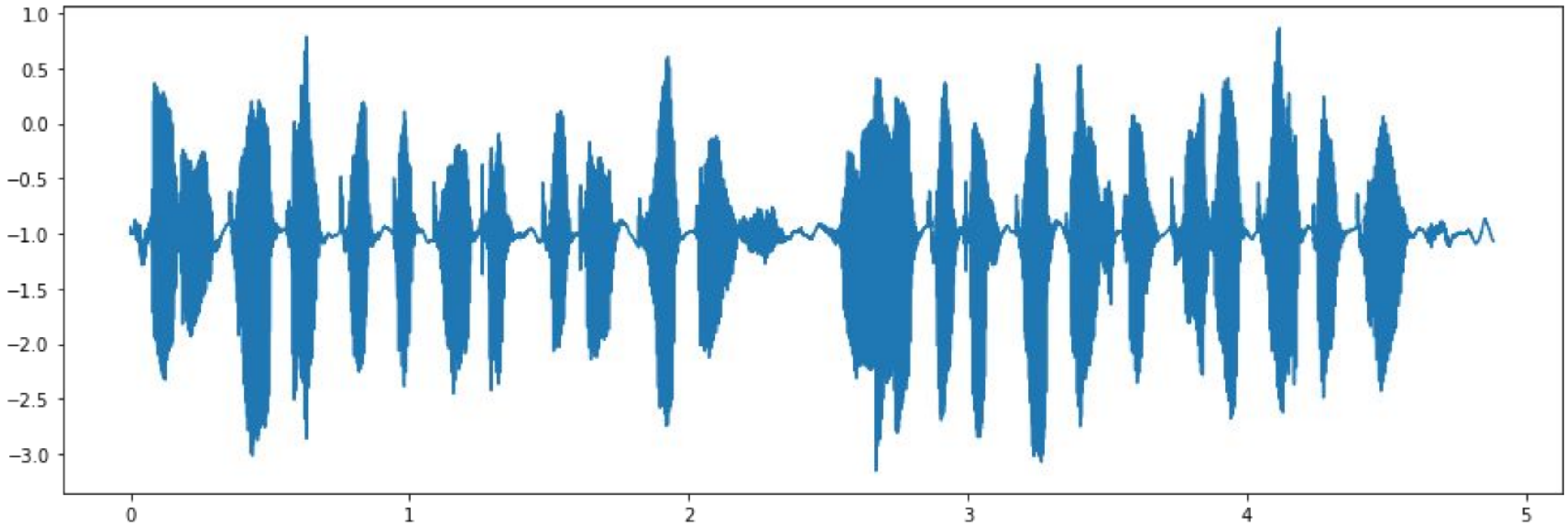
Stanislav Protasov for
Harbour.Space University

Agenda

- Sound as a wave
- Speech recognition
  - Acoustic model
  - Language model
- Speech generation

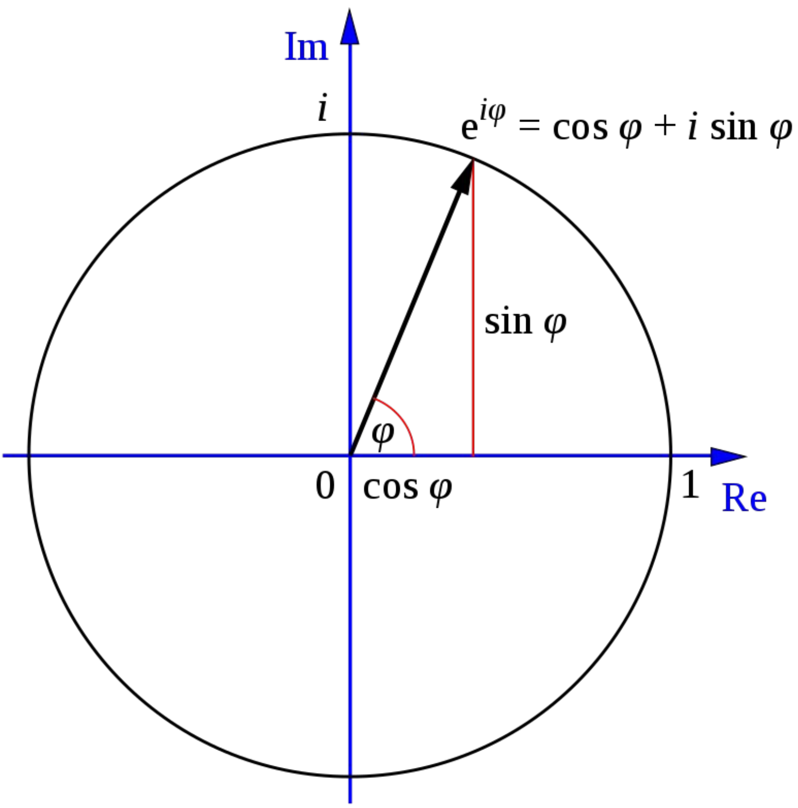
What is the sound?

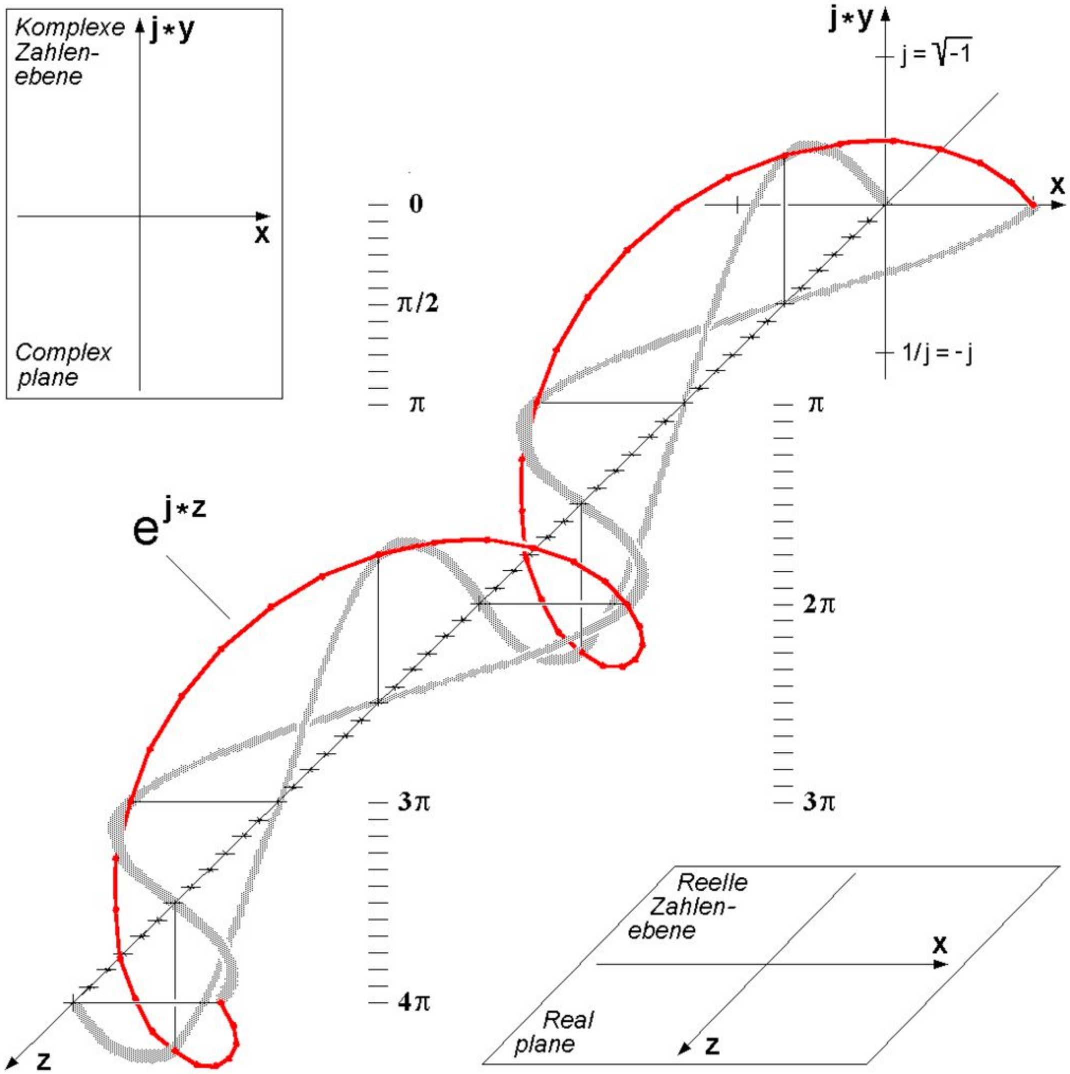
Sound is a **vibration** that propagates through a transmission medium such as a gas, liquid or solid.



Euler's identity

$$e^{ix} = \cos x + i \sin x,$$





https://www.youtube.com/watch?v=ykNtIbtCR-8

FT: $$\hat{f}(w) = \int_{-\infty}^{+\infty} f(x) e^{-2\pi i x w} dx$$

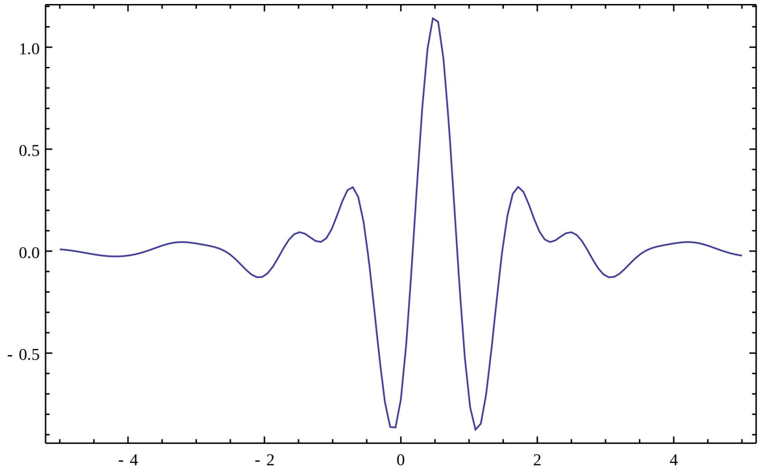
DTFT: $$X_T(w) = \sum_{n=-\infty}^{+\infty} f(nT) e^{-2\pi i w n T}$$

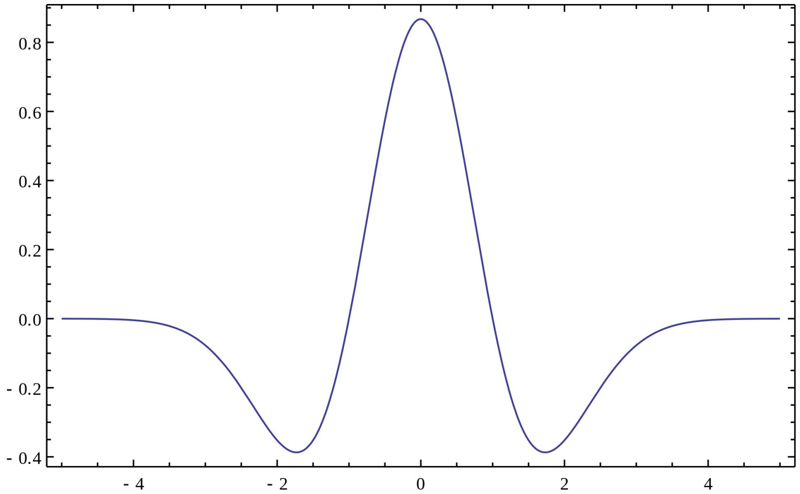
DTFT + window: $$X_T(w) = \sum_{n=0}^{M} f(nT) W\left(\frac{n}{M}\right) e^{-2\pi i w n T}$$

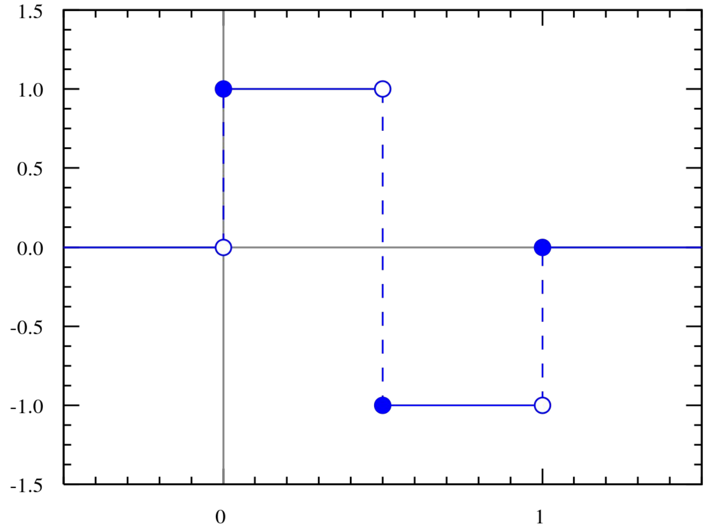
DFT: $$X_{T,N}(k) = X_T\left(\frac{k}{NT}\right) = \quad k = 0, 1, \dots, N-1$$

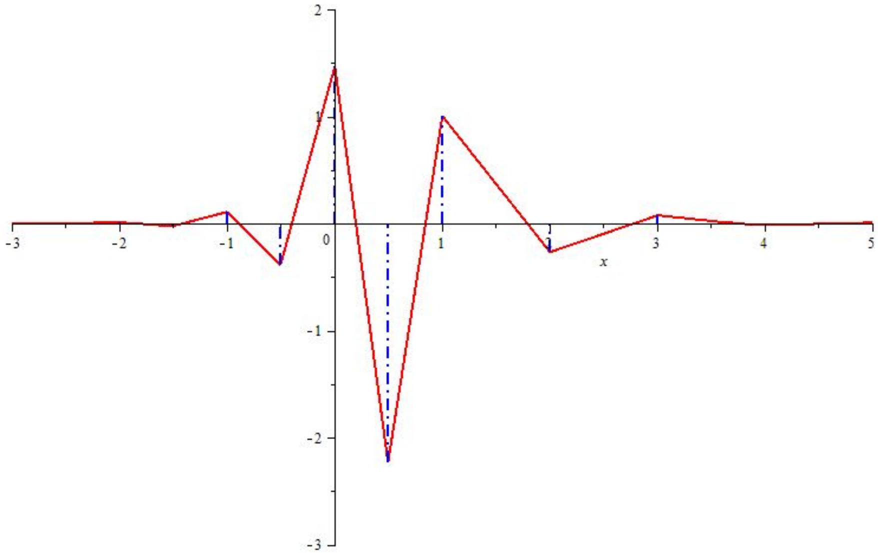
$$= \sum_{n=0}^{M} f(nT) W\left(\frac{n}{M}\right) e^{-2\pi i \frac{kn}{N}}$$

Wavelets



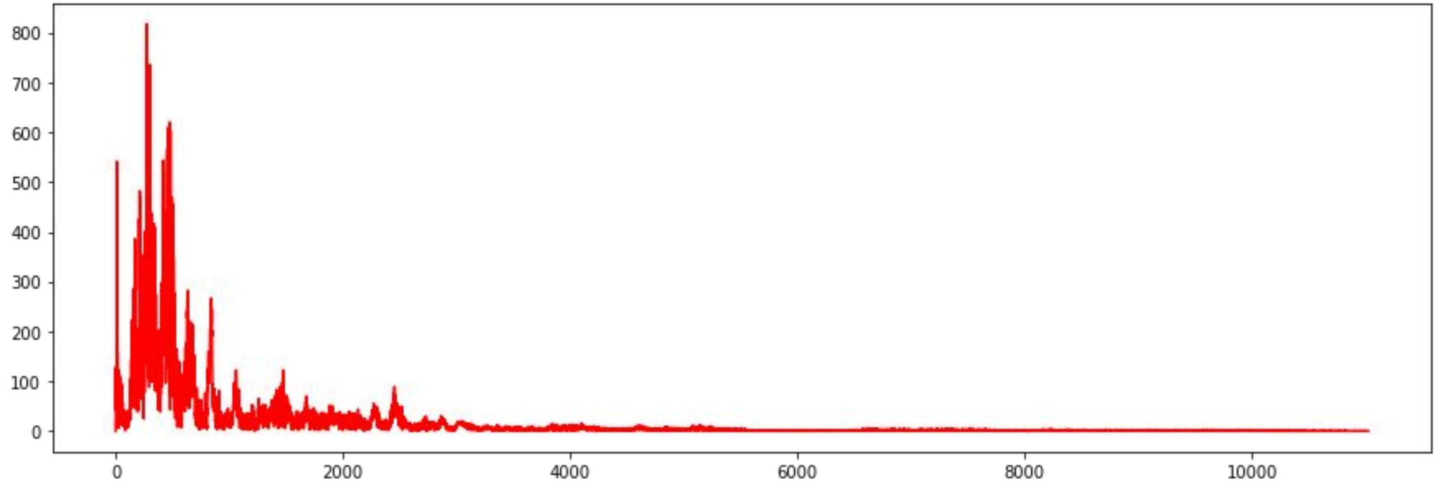




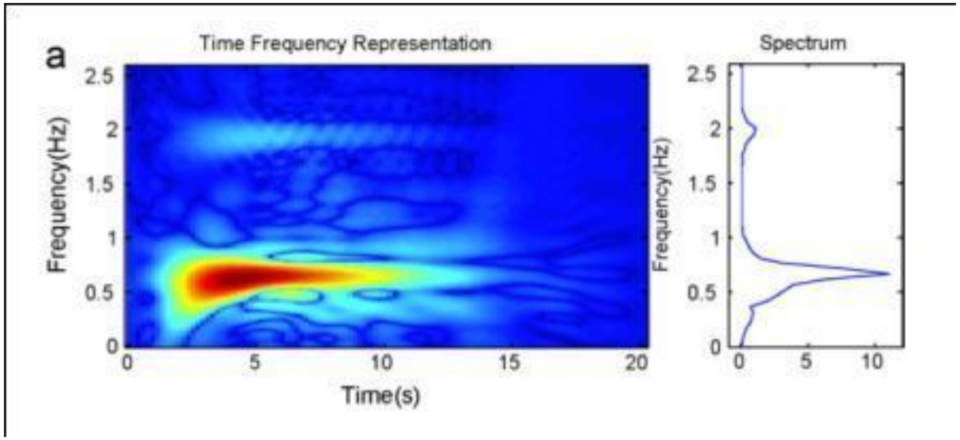


What is the sound for human?

We percept sound using **frequency** receptors. Each moment looks like this:



Timeline is like this:



Sound recording and playback

- Digital uncompressed sound consists of regular measurements of signal.
- Measurement frequency is managed using RATE parameter
  - 22050 means 22050 measurements per second (**discretization**)
- How accurate we measure in managed is tuned with format (**quantization**)
  - How many different amplitude values can be encoded
- Channels — number of inputs/outputs (stereo=2, mono=1)
- BPS = RATE * CHANNELS * FORMAT
- Together this is **PCM**

Lab #0. Make this work

Recording and playing tutorial

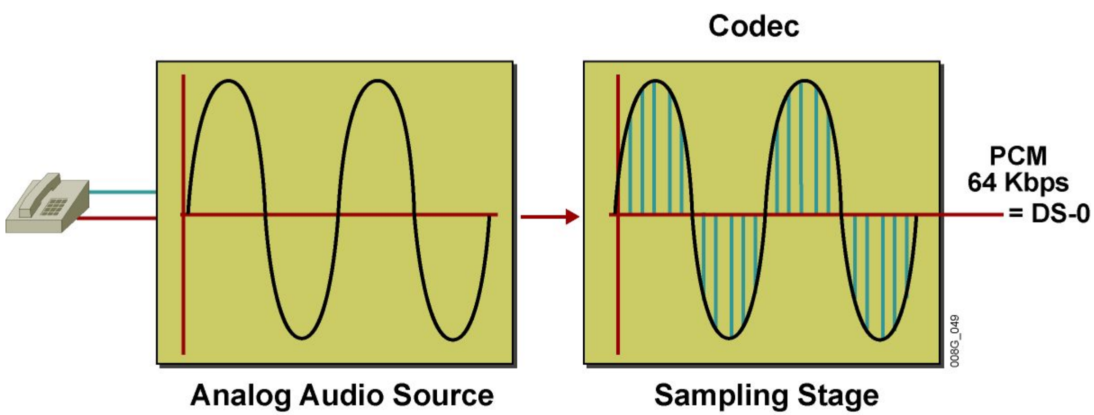
https://github.com/hsu-ai-course/hsu.ai/blob/master/code/06.%20Sound%20record%20and%20play.ipynb

FFT tutorial

https://github.com/hsu-ai-course/hsu.ai/blob/master/code/06.%20Sound%20FFT.ipynb

Nyquist-Shannon (Kotelnikov) theorem

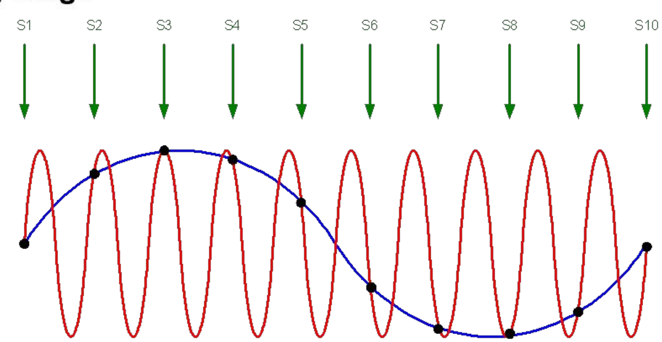
If a function **x(t)** contains no frequencies higher than ***B*** hertz, it is **completely determined** by giving its ordinates at a series of points spaced **1/(2B)** seconds apart.



What if contains? Aliasing. ***n(k)***?

$$\left\{\sin(k\,x) = \sin(n\,x),\; n < k\right\}$$

- `sin(a)+sin(b) = 2·sin(½(a+b))·cos(½(a-b))`



Lab #1

Implement tutorial on chord transformation

https://github.com/str-anger/hsu.ai/blob/master/code/06.%20Chord.ipynb

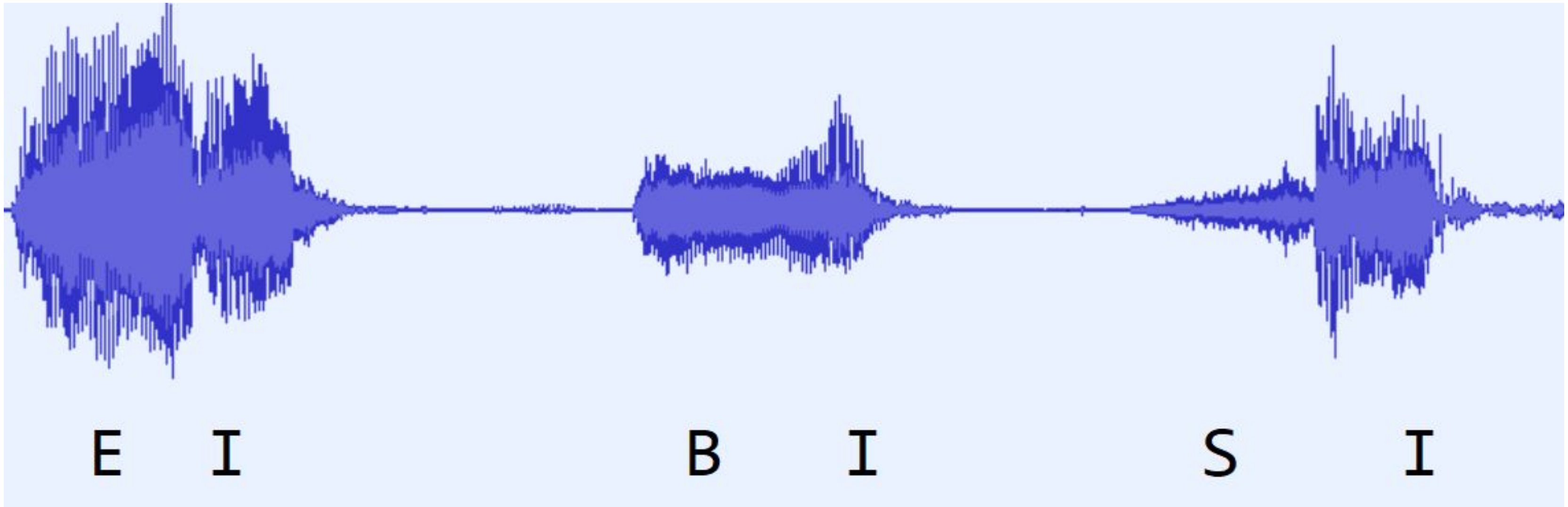
1. Convert to frequencies
2. Find major frequencies
    a. (*) do it automatically (with code, not with your eyes)
3. Can you say what is the chord?

    Chord is a set of pitches played simultaneously

    Refer http://pages.mtu.edu/~suits/notefreqs.html

Acoustic model

As text consist of letters, speech consists of phonemes.



AM: spectrum → phoneme

Language model

Probabilistic model that predicts probability of a word given a sequence of phonemes.

Similar model is used to model sentences of words.

Speech generation

1) Text preprocessing
   a) Number to text
   b) Abbreviations to text
   c) Typo fix
2) Split text into phrases (punctuation, constructions)
3) Phonetic construction (language model)
   a) queue - [kju]
   b) **Арбалетчиков**
      i) a0 r b a0 lj e**1** t ch i0 k o0 v

Speech generation

1) **Accents** are set
   a) Using a dictionary
   b) Using rules
   c) Using statistics (speaker examples)
2) **Reversed acoustic model** is used to consider surrounding
3) **Timbre** is generation with **vocoder**
   a) or RNNs

Lab #2

- Implement **speech generation tutorial.** https://github.com/hsu-ai-course/hsu.ai/blob/master/code/06.%20Speech%20generation.ipynb
  - Register all needed Google Cloud accounts
- (*) Implement **speech recognition tutorial**
  - Download and install CMU Sphinx for you native language (if present)

Hometask

1) Implement **speech recognition from microphone** using Google Cloud Platform.
2) (*) Implement speech-2-speech translation (babel fish)
3) (**) Podcast 2x speed
4) (***) ID recognition by voice