Signal Representation

EE698V - Machine Learning for Signal Processing

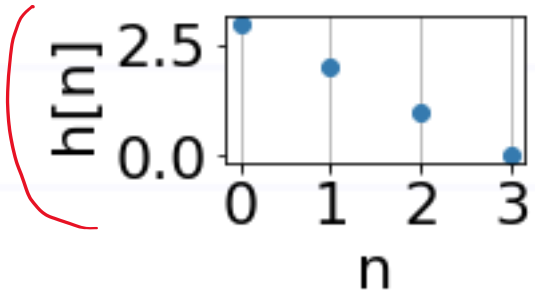
Vipul Arora

Tensors

- 0 order: Scalar, x
  - Temperature
  - time, t
  - frequency, f
  - amplitude, a
  - phase, ϕ

Tensors

- 1st order: Vector, **x**
  - **t** or $[t_1, t_2, \ldots]$
  - f[n] or **f** or $[f_1, f_2, \ldots]$
  - a[n] or **a** or $[a_1, a_2, \ldots]$
  - ϕ[n] or $\boldsymbol{\phi}$



h = [3,2,1,0]

INDEXED with 1 index

$i \in \mathbf{I}$

Tensors

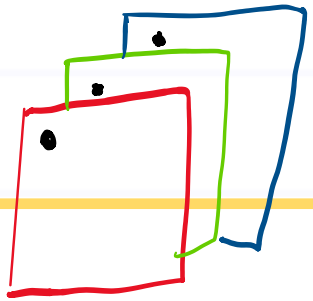
- 2nd order: Matrix, X
  - a grayscale image, X[m,n]
  - a spectrogram, X[f,t]

$$X = \begin{bmatrix} 2 & 7 & 5 & -1 \\ 5.1 & -4 & 3.7 & 2.9 \\ 6.9 & 5.1. & 9.1 & 8.1 \end{bmatrix}$$

INDEXED with 2 indices

$i_1, i_2 \in \mathbf{I}$

Tensors



- 3rd order, **X**
  - A grayscale video, X[i,m,n]
  - A colored image, X[m,n,c]
  - A collection of spectrograms, X[i,f,t]

$$\left[\begin{array}{cccc|c} -2 & 9 & -3 & 8 & \\ & & & & 9 \\ \hline 2 & 7 & 5 & -1 & \\ 5.1 & -4 & 3.7 & 2.9 & 1.3 \\ 6.9 & 5.1. & 9.1 & 8.1 & \end{array}\right]$$

INDEXED with 3 indices

$i_1, i_2, i_3 \in \mathbf{I}$

Tensors

- Nth order
  - $x[i_1, i_2, \ldots, i_N]$

Hard to visualize !!

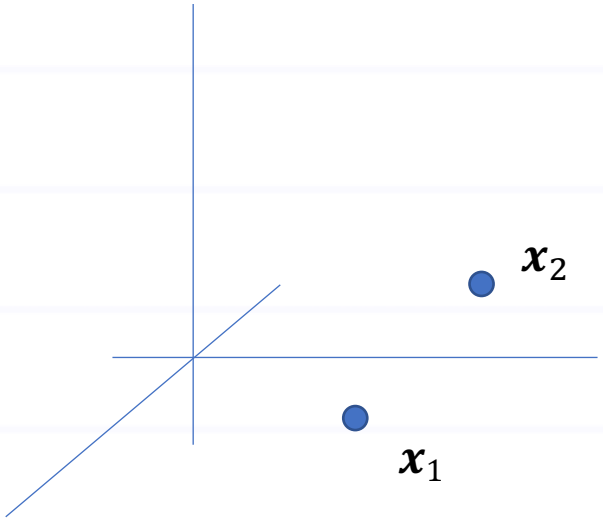
INDEXED with N indices

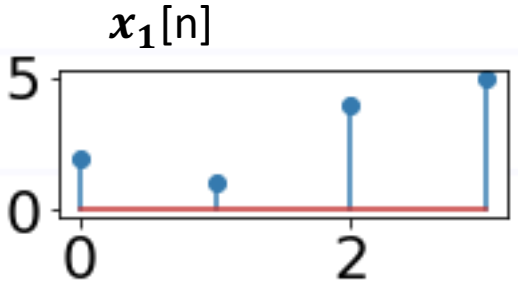
$i_1, i_2, \ldots, i_N \in \mathbf{I}$

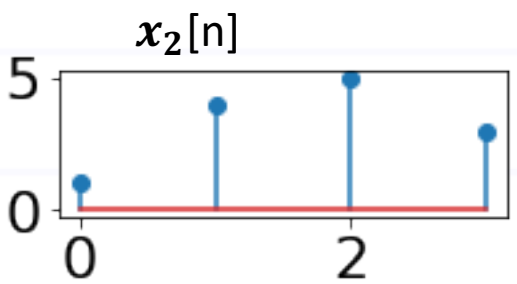
How do we measure difference?

- Euclidean distance
- $d(\boldsymbol{x_1}, \boldsymbol{x_2}) = |\boldsymbol{x_1} - \boldsymbol{x_2}|$

$$= \sqrt{\sum_n (\boldsymbol{x_1}[n] - \boldsymbol{x_2}[n])^2}$$







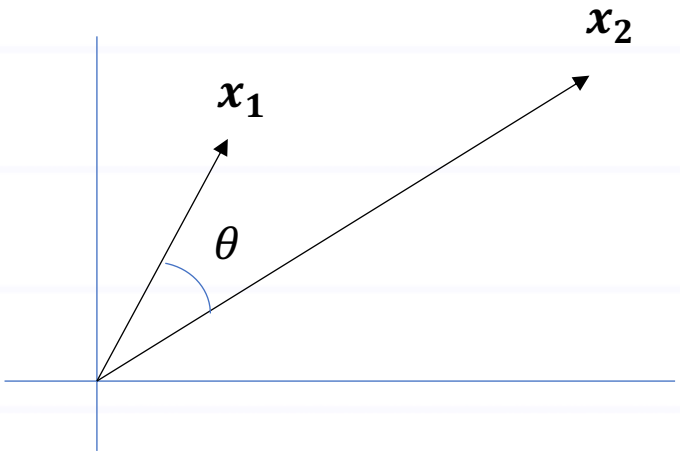
$d(\boldsymbol{x_1}, \boldsymbol{x_2})$ = 3.87

How do we measure difference?

- On a hypersphere: Cosine similarity

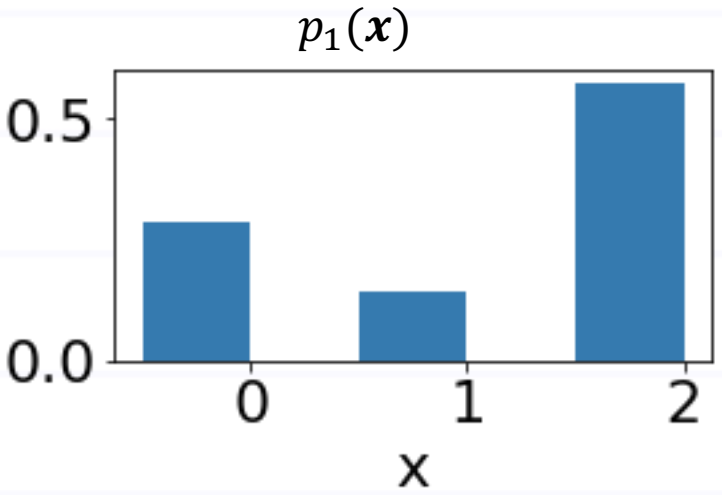
- $d(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{\boldsymbol{x_1^T} \cdot \boldsymbol{x_2}}{|x_1|\,|x_2|}$

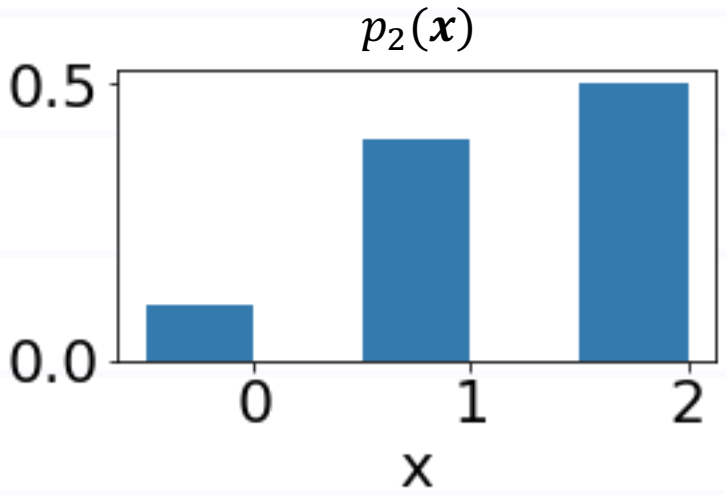
$= \cos(\theta)$



How do we measure difference?

- Between probability distributions: KL Divergence
- $d(p_1(\boldsymbol{x}), p_2(\boldsymbol{x})) = -\sum_x p_1(x) \log\left(\frac{p_2(x)}{p_1(x)}\right)$



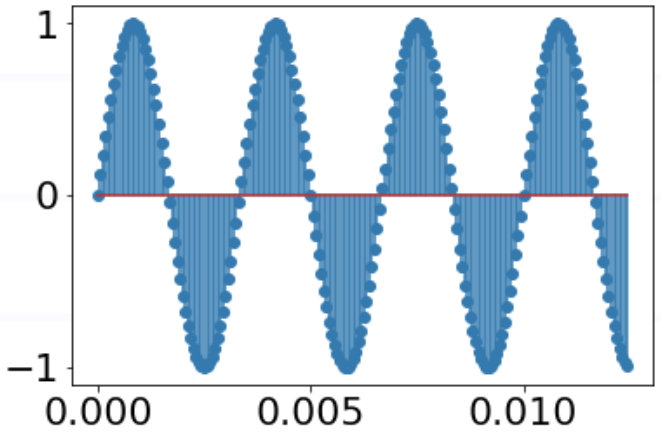


$d(p_1(\boldsymbol{x}), p_2(\boldsymbol{x})) = \mathbf{0.23}$

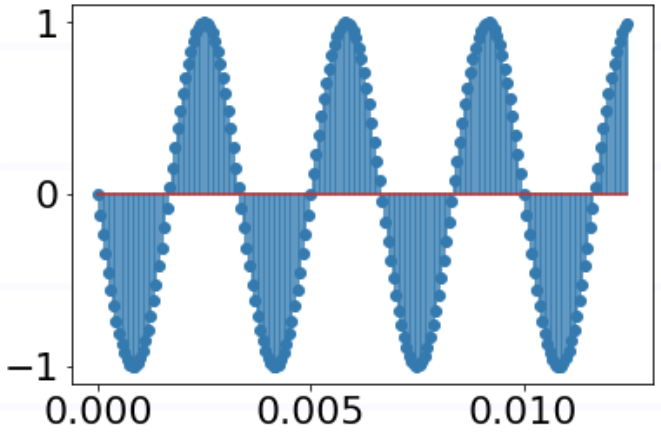
Does it really make sense?

$$\sqrt{\sum_{n=0}^{N-1}\left(x_1[n]-x_2[n]\right)^2\ \frac{1}{N}}$$

$x_1[n]$



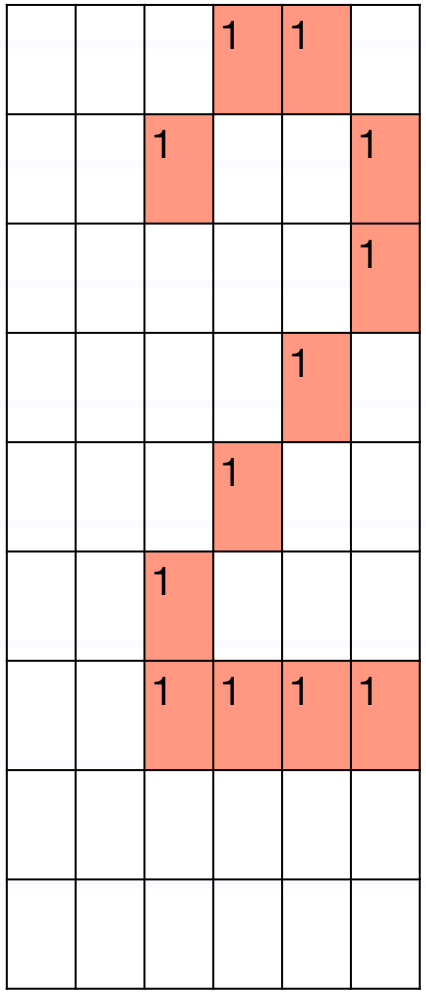
$x_2[n]$



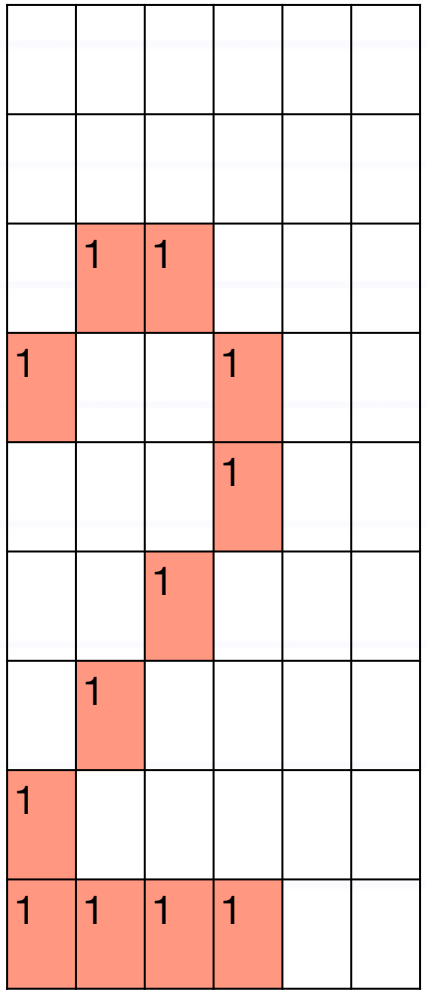
$d(x_1, x_2) = 19.9$

Does it really make sense?

$\boldsymbol{x_1} =$



$\boldsymbol{x_2} =$



$d(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sqrt{10}$

It gets more complicated

What is lacking?

SPACE

What is Space?

- Space provides support and a distinct identity to every object
- You have heard of
  - Cartesian space
  - Spherical space
  - Riemann space
  - …

Examples of Space

- Food you eat
  - apple vs banana
- Sounds you listen to
  - Sa vs Re
  - flute vs violin
  - English vs Punjabi

Examples of Space

- Images you see
  - red vs green
  - marigold vs jasmine
- Scents you smell
  - sandal vs saffron

Examples of Space

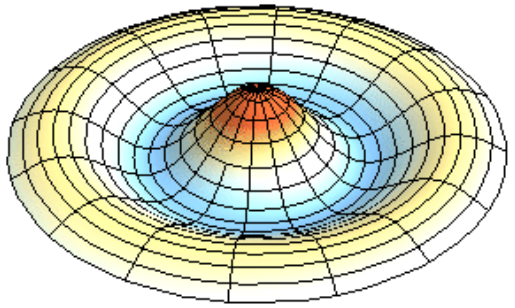
- Touch experiences
  - cotton vs brick
  - hot vs cold

But that is sense perception

- How can machines do it?
- Detectors or sensors
  - They correlate with what we experience

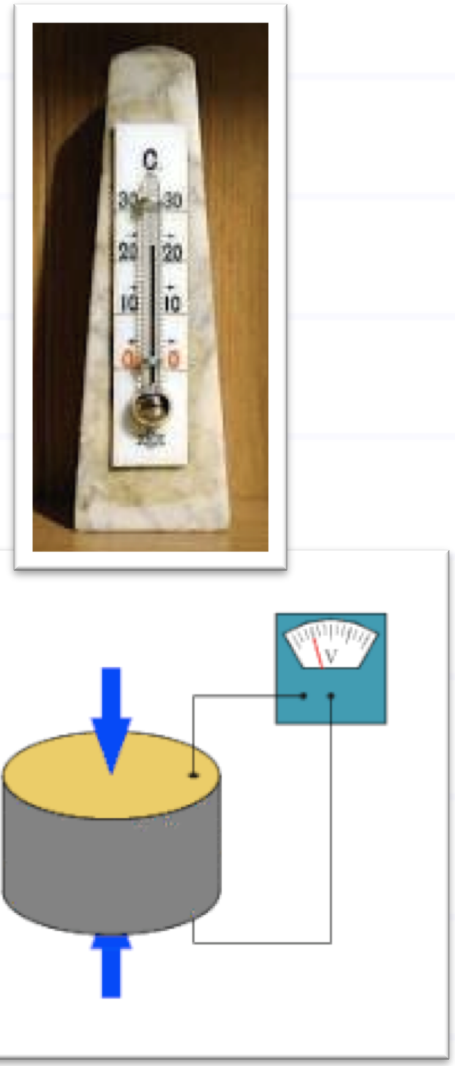
Detectors

- A vibrating membrane correlates with what we hear
- A photosensitive chemical correlates with what we see



Detectors

- Mercury volume correlates with what we experience as heat/cold
- Piezoelectric crystals correlate with what we experience as pressure



Space again!

- Raw readings of detectors not enough
- We need to formulate a mathematical space that correlates with what we experience

Feature Extraction

- Important steps:
  - identify what we are interested in
  - identify which detectors correlate with it
  - design features (space)

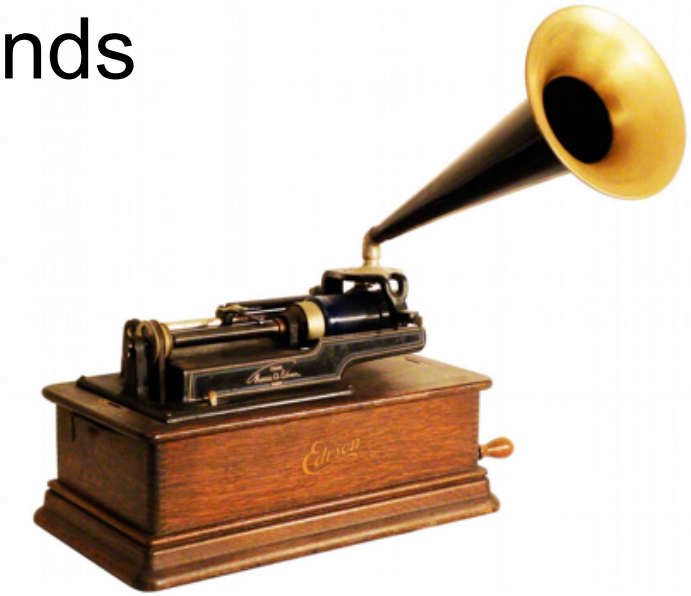
Feature Extraction

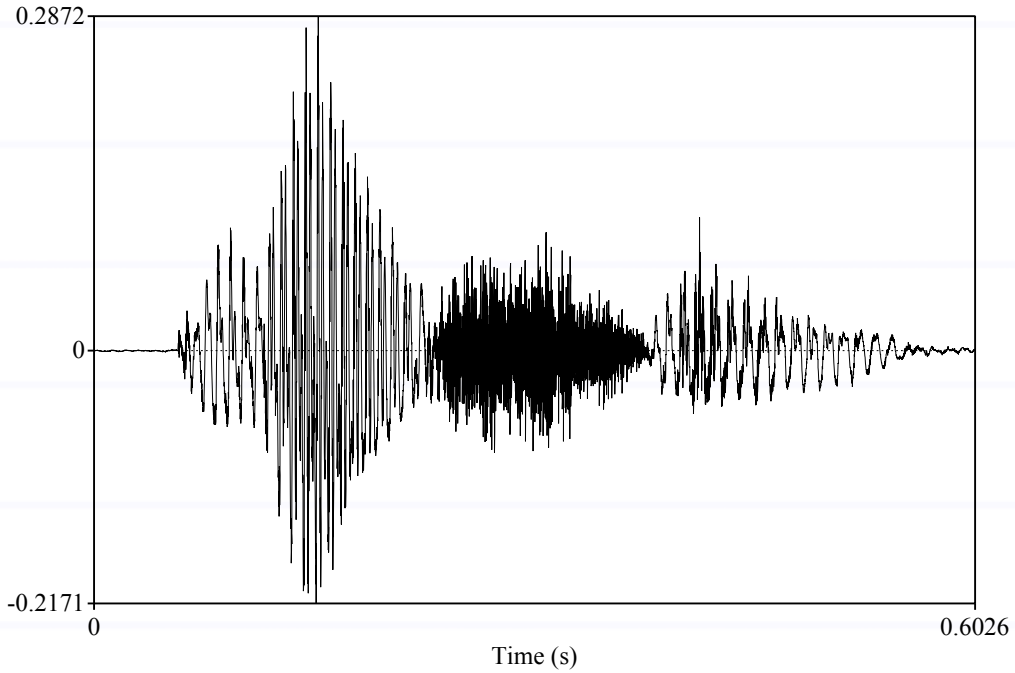
- Features should be
  - distinguishing for what we are interested in
  - invariant to what we are not interested in
  - minimal in size

Audio Signals

Audio

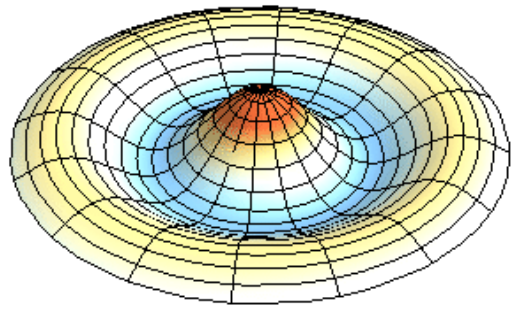
- Edison invented gramophone
- Recording and playing back of sounds



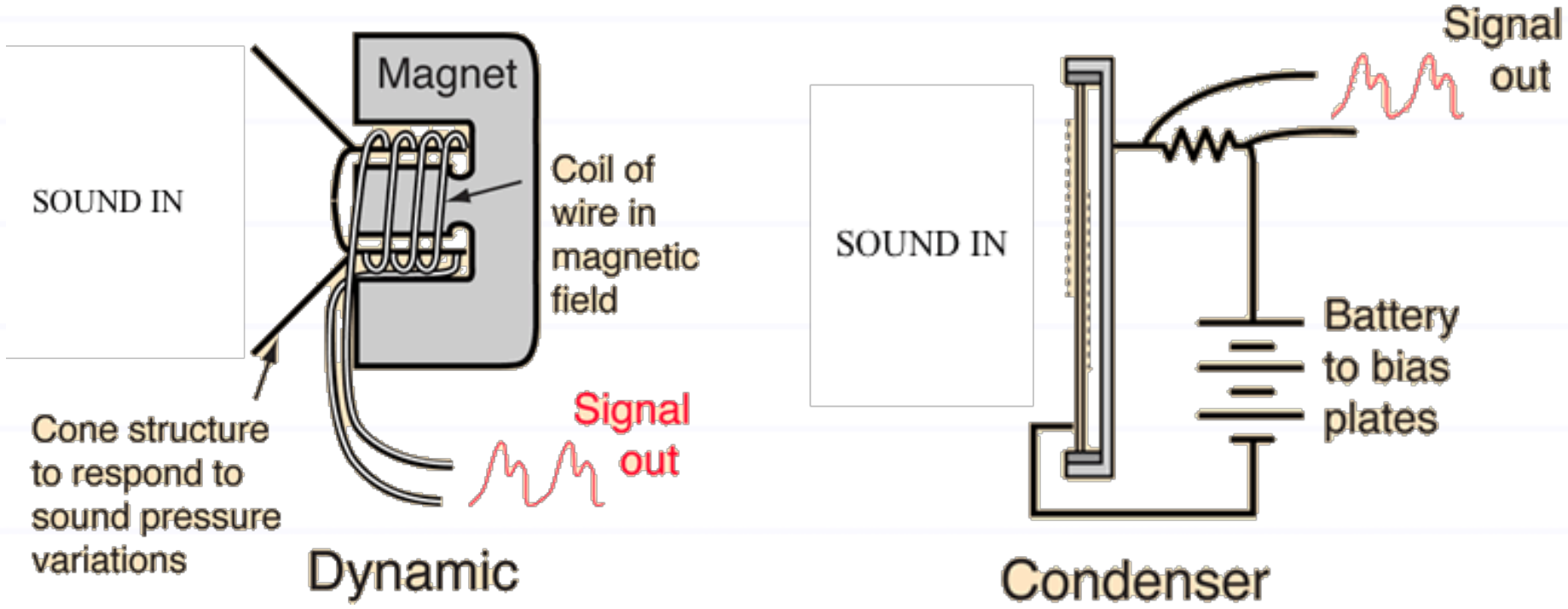


Audio

- Sound makes the membrane vibrate
- How to store the signal?
- In electric form …
- Convert physical motion into electric signals



Audio



http://hyperphysics.phy-astr.gsu.edu/hbase/Audio/mic.html

Audio

- We get a continuously varying signal, how to save it in a computer?
- Sample and quantize

Sampling

- Humans can hear in the range 20Hz to 20kHz
- Popular sampling rates:
  - 44.1kHz for CD recordings

Quantization

0 1 1 0 ... 16 bits

- Converting $x \in \mathbb{R}$ to a digital number
- Q bits per sample => 2^Q possible integer values per sample
  - 16 bits per sample for CD recordings

Audio

- Finally, audio is a stream of numbers
- Represent it as a vector of shape (N,1) with N as the number of samples

Pitch

- A perceptual attribute of sound
  - musical notes
- Correlated with F0 (Hz) non-linearly
  - mel scale experimentally derived from perceptual pitch
    - 2 times mel = perceptual pitch doubling
- Range: ~ 30 Hz to 4kHz (music)
- Resolution: different at different F0

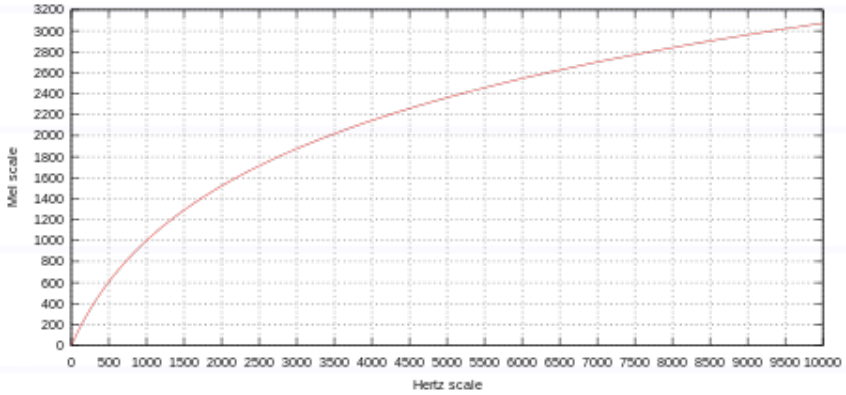


Image source: Wikipedia

Loudness

$20 \log_{10} \frac{p}{p_0}$

- A perceptual attribute of sound
  - quite to loud
- Correlated with pressure level (dB) non linearly https://en.wikipedia.org/wiki/Sound_pressure#Sound_pressure_level
- Range: varies with pitch

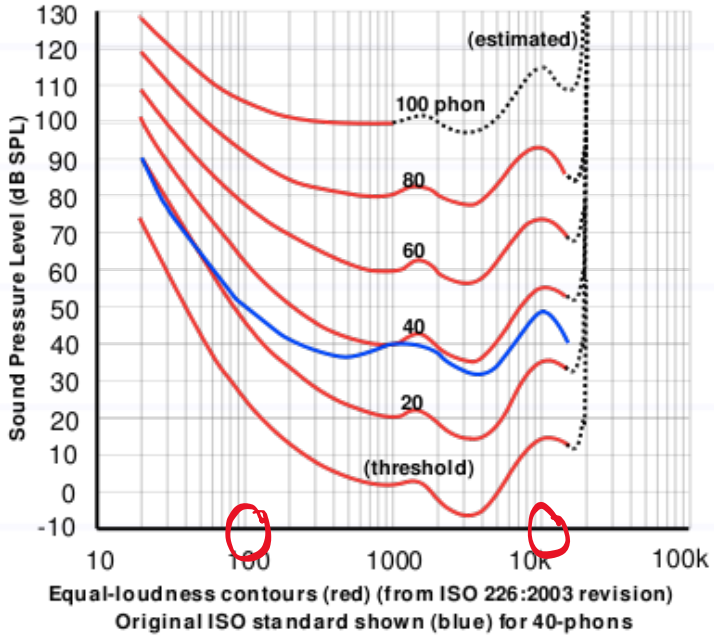
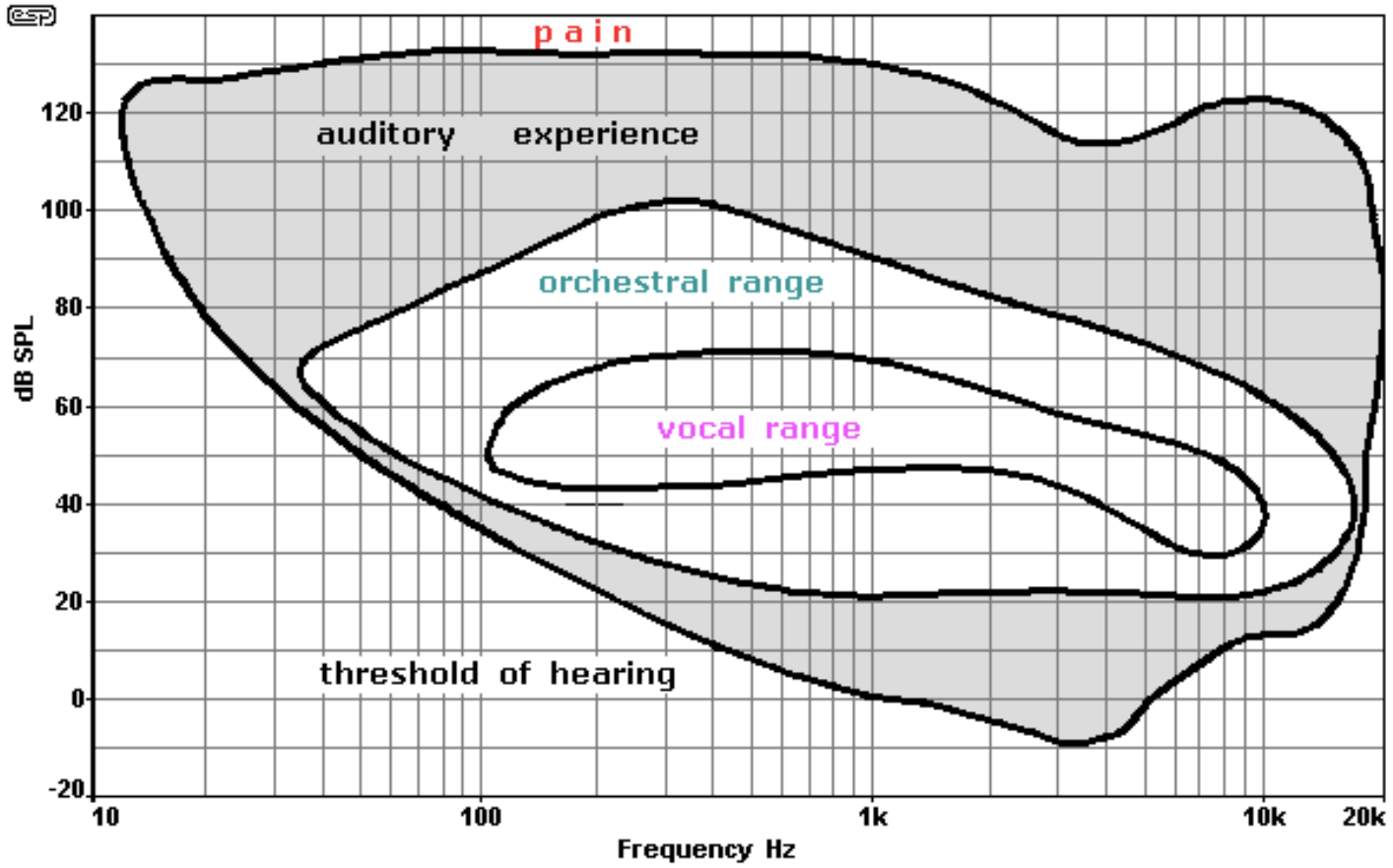


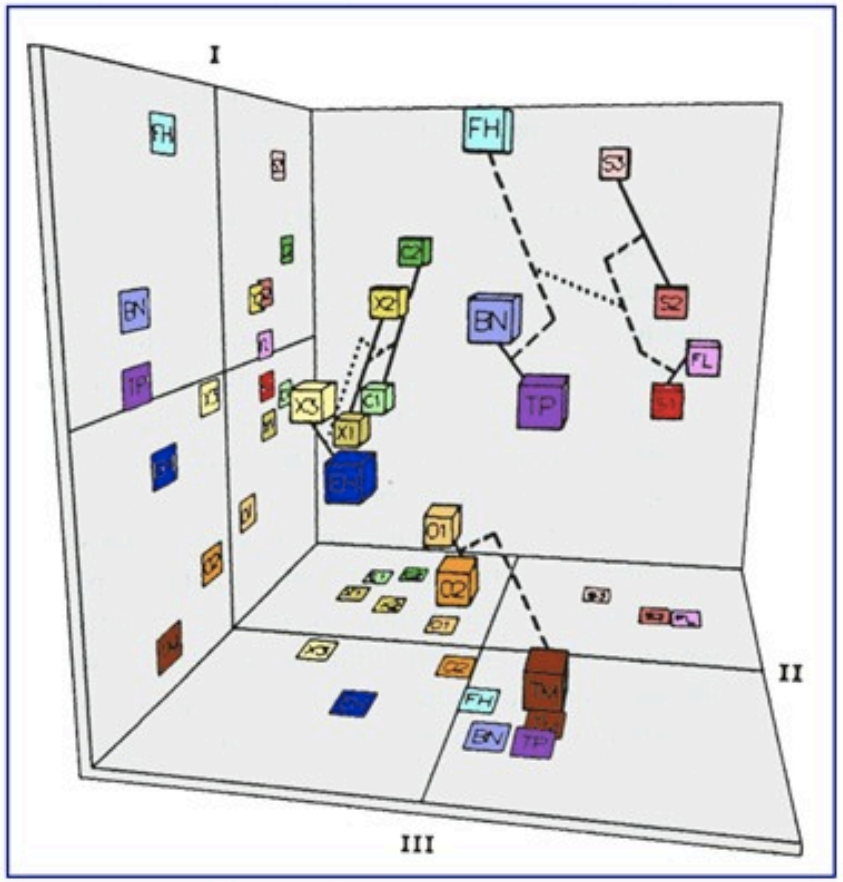
Image source: Wikipedia

Loudness and Pitch



Source: https://www.joyaudiokits.com/calculator

Timbre



- BN - Bassoon
- C1 - E flat Clarinet
- C2 - B flat Bass Clarinet
- EH - English Horn
- FH - French Horn
- FL - Flute
- O1 - Oboe
- O2 - Oboe (different instrument and player)
- S1 - Cello, muted *sul ponticello*
- S2 - Cello
- S3 - Cello, muted *sul tasto*
- TM - Muted Trombone
- TP - B flat Trumpet
- X1 - Saxophone, played *mf*
- X2 - Saxophone, played *p*
- X3 - Soprano Saxophone

- Dimension I: spectral energy distribution, from broad to narrow
- Dimension II: timing of the attack and decay, synchronous to asynchronous
- Dimension III: amount of inharmonic sound in the attack, from high to none

- A perceptual attribute of sound
  - Correlated with source of sound
  - Mathematically??

Image source: https://slideplayer.com/slide/7070982/

Binaural Hearing

- Two ears
- We can localize the source of sound
  - Interaural Time Difference
  - Interaural Intensity Difference
  - Filtering by the body
- Uses:
  - in microphone arrays for source localization/separation (Alexa!)
  - 3D sound effects

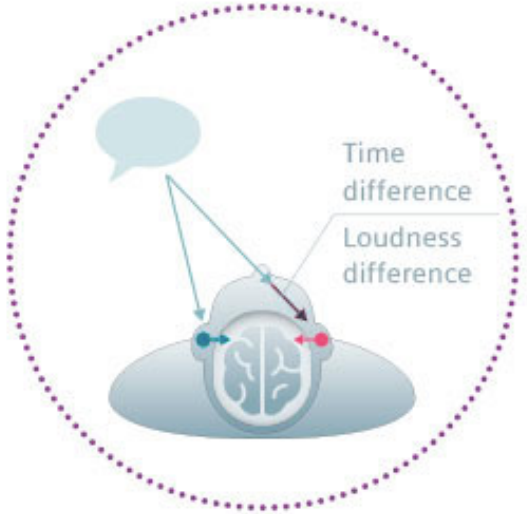
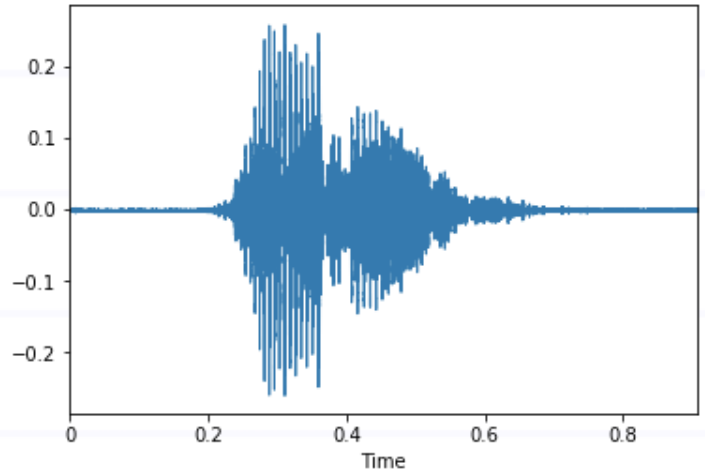


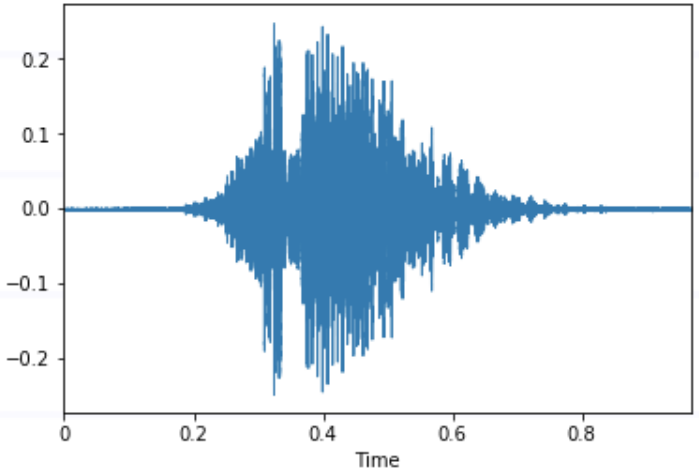
Image: http://ear.skewsoft.com/binaural-hearing/

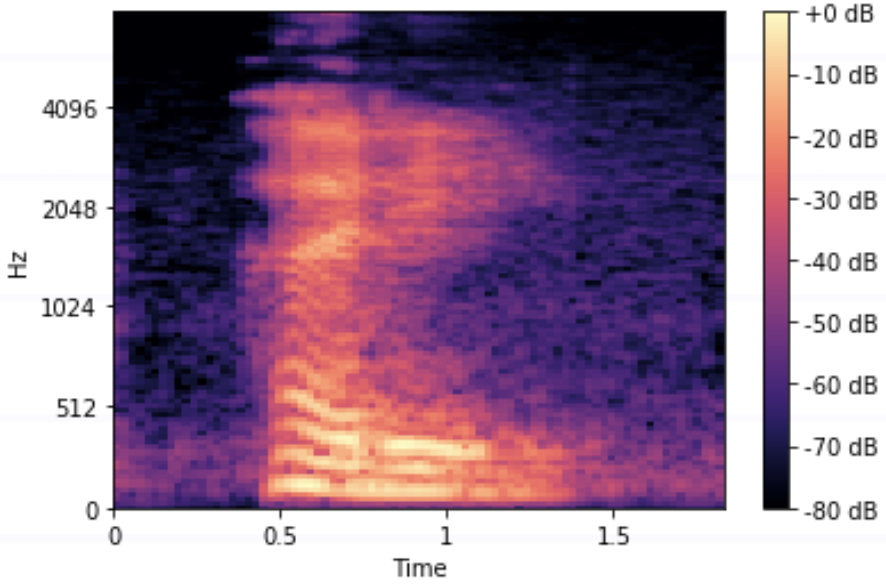
Lessons for feature design

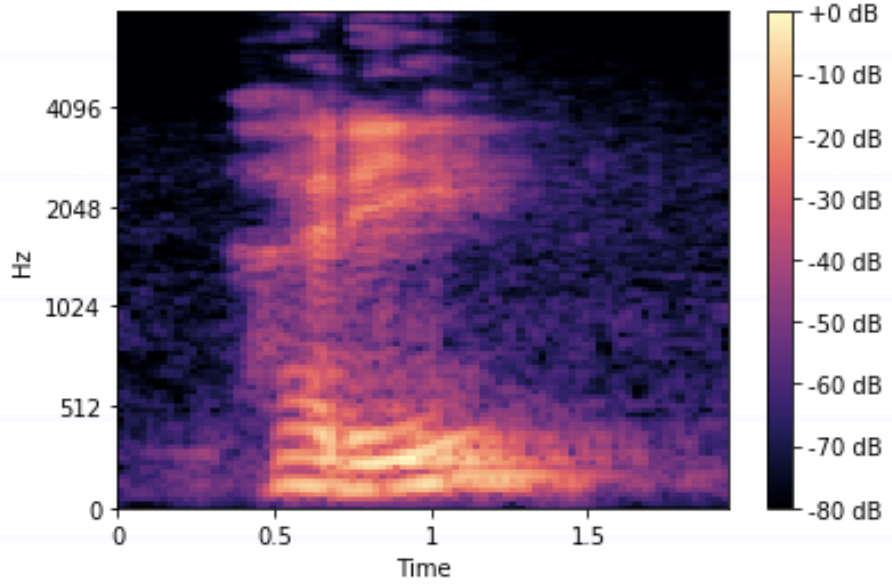
- Compare sounds using phase invariant representations
  - |X(k,n)| : magnitude spectrograms
- Compare melodies using pitch
- Compare magnitudes on log scale (dB)
- Compare spectra on mel (-like) frequency scale

Audio Representation

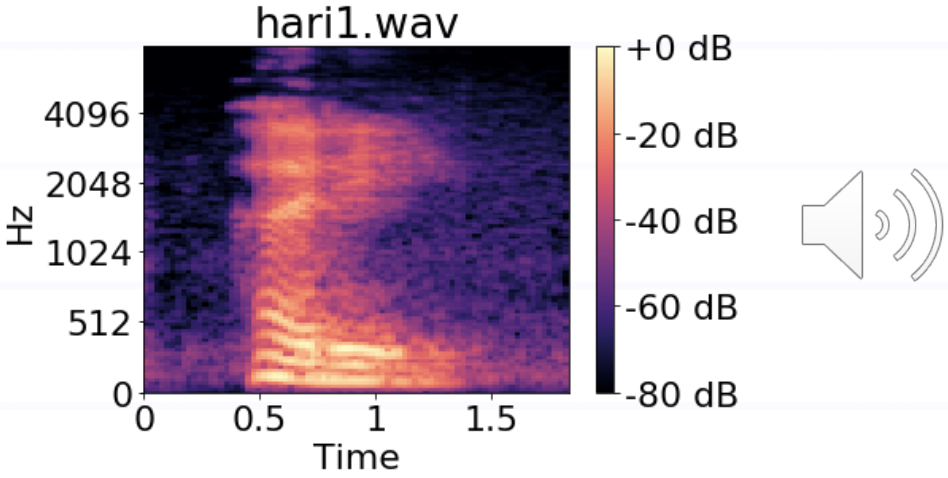


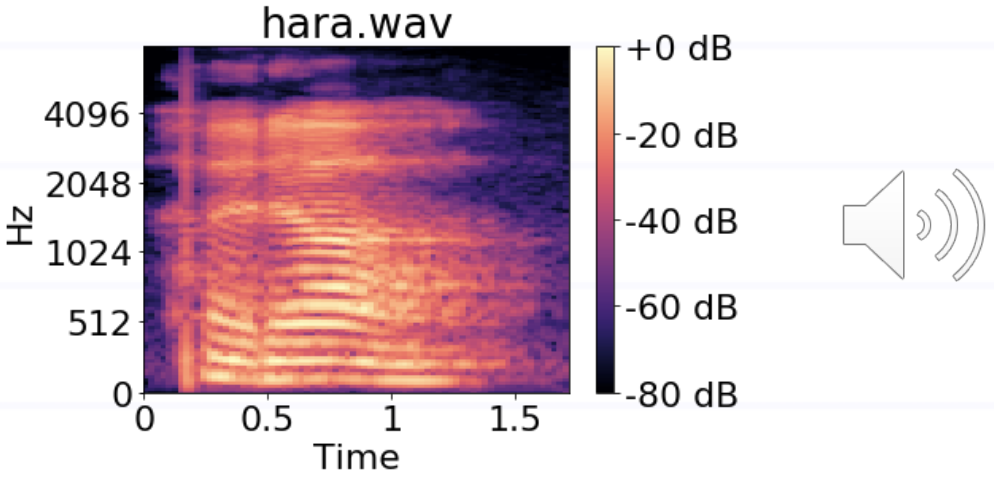


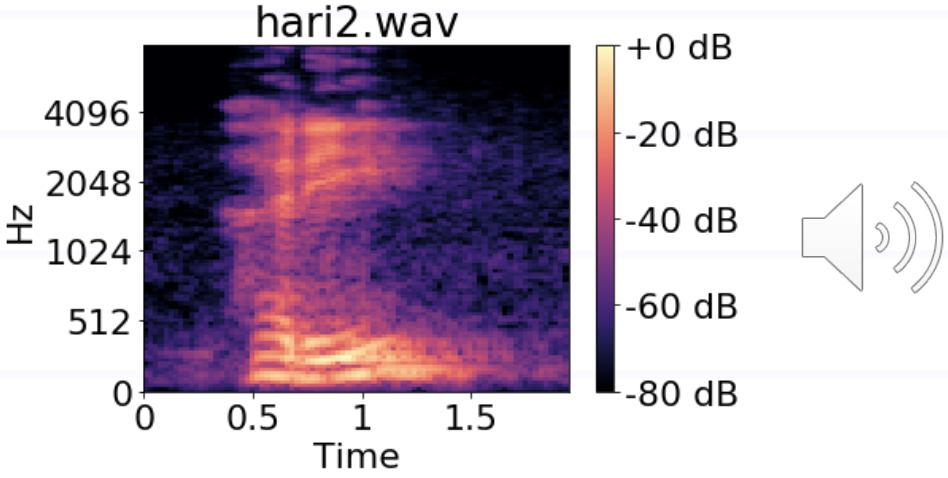


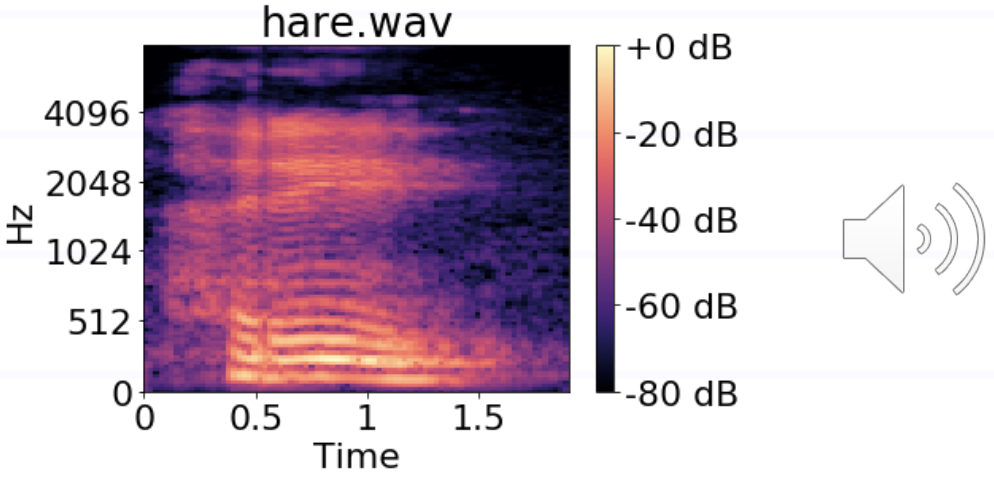


Audio Representation - phonemes

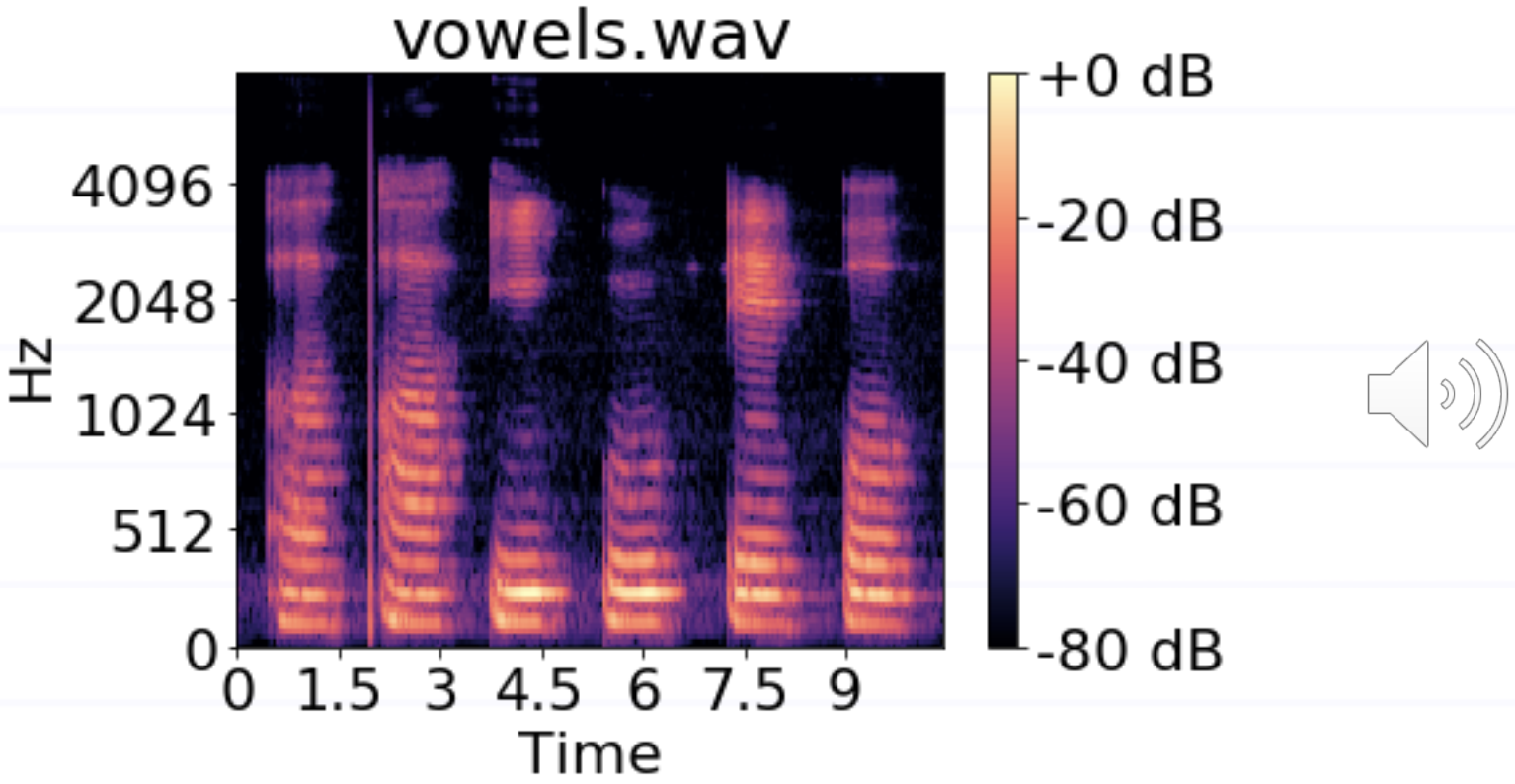




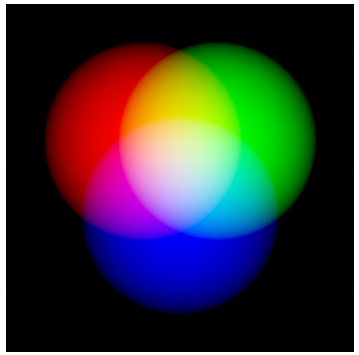


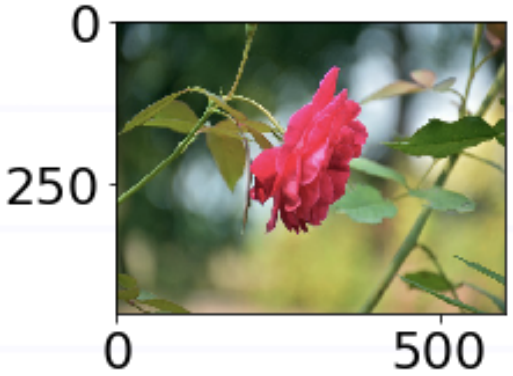


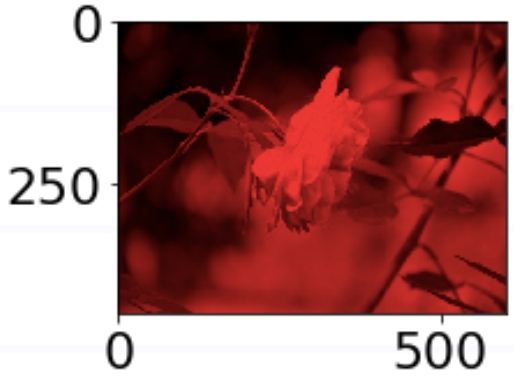
Audio Representation - phonemes

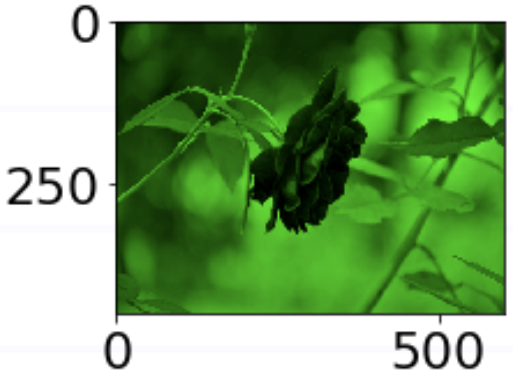


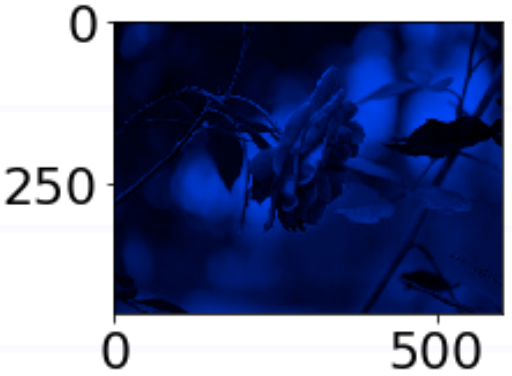
Visual Perception











Eyes are sensitive to CHANGES

- Edges
- Corners
- Blobs
- etc.

How to detect them in images…

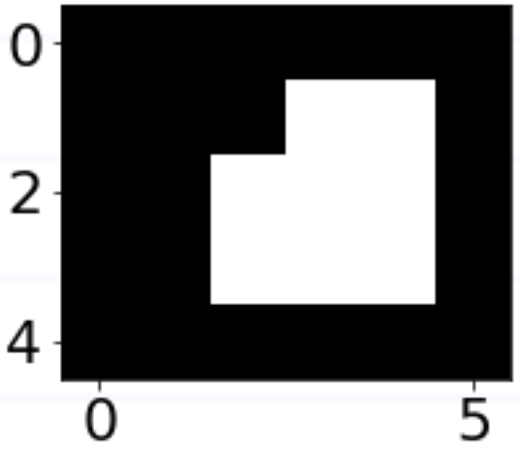
Filters – horizontal edge

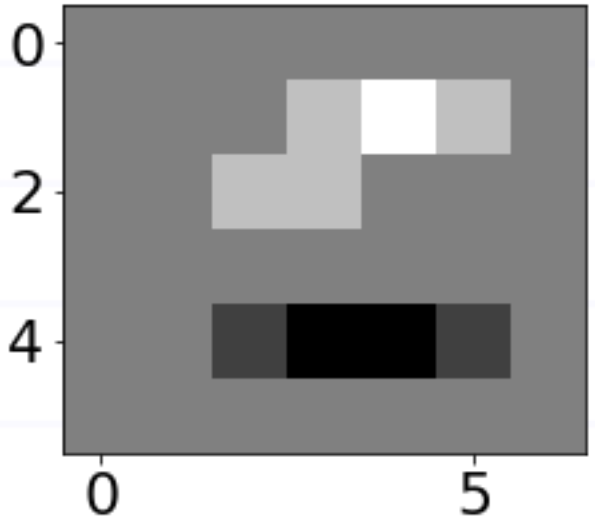
```
[[0  0  0  0  0  0]
 [0  0  0  1  1  0]
 [0  0  1  1  1  0]
 [0  0  1  1  1  0]
 [0  0  0  0  0  0]]
```

```
[[-1 -1]
 [ 1  1]]
```

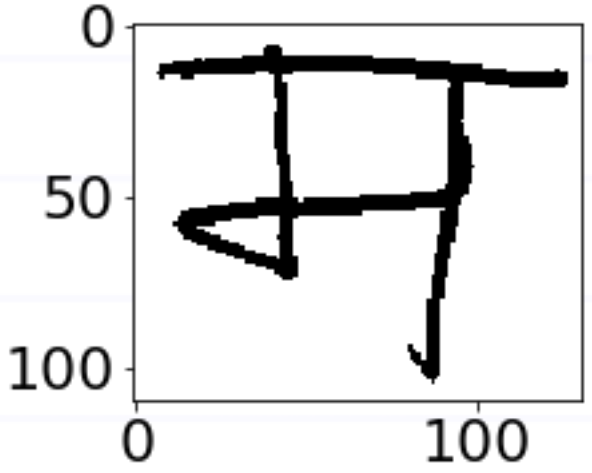
```
[[ 0  0  0  0  0  0  0]
 [ 0  0  0  1  2  1  0]
 [ 0  0  1  1  0  0  0]
 [ 0  0  0  0  0  0  0]
 [ 0  0 -1 -2 -2 -1  0]
 [ 0  0  0  0  0  0  0]]
```

Filters – horizontal edge

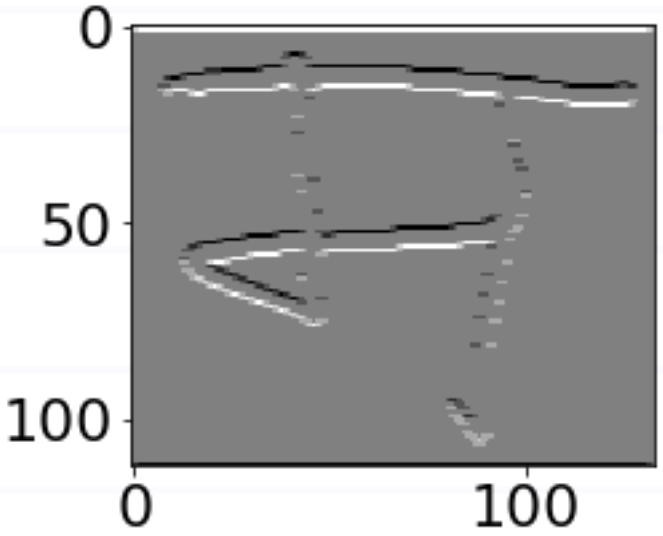




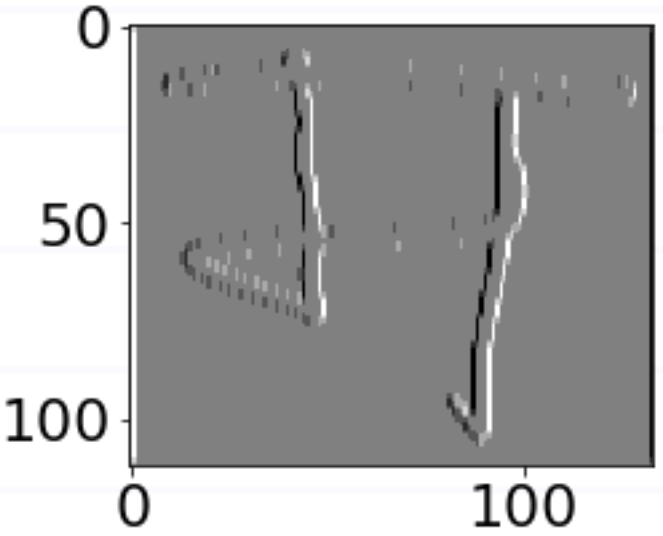
Filtering




```
[[-1 -1 -1]
 [ 1  1  1]
 [ 0  0  0]]
```




```
[[-1  1  0]
 [-1  1  0]
 [-1  1  0]]
```



Essence

- We are trying to mimic
  - ear detectors
  - eye detectors
- So, better take lessons from them to formulate the space

Topics discussed

- Space
- Detectors
- Features Extraction
- Sound features