Fixed Features and their Origins
Today’s overview

• Features and their importance

• Human perception of signals

• Fixed features for machine perception
Why features?

- Features are a very important area
  - Bad features make problems unsolvable
  - Good features make problems trivial

- Learning how to pick features is the key
  - So is understanding what they mean
A simple example

- How do we compare two numbers?

**Case 1**

\[ x = 3 \]
\[ y = 3 \]

**Case 2**

\[ x = 3 \]
\[ z = 100 \]
A simple example

- We can use their distance:

\[ |x - y| = 0 \quad |x - z| = 97 \]

- \( x, y \) similar but \( x, z \) not so much

- Best way to represent a number is itself!
Moving up a level

- Comparing two vectors:

\[ \mathbf{x}, \mathbf{y}, \mathbf{z} \]
Moving up a level

• Make up a distance measure:

\[ \angle x, y = 4.9^\circ \quad \angle x, z = 48.2^\circ \]

\[ \| x - y \| = 0.12 \quad \| x - z \| = 1.48 \]

• We simply generalize the scalar approach
Moving up again

- Compare two longer vectors:
Look similar but are not!

- Oops!  \( \angle x, y \approx 90^\circ, \quad \|x - y\| = 10.8 \)
• Are these two vectors the same?

• Not if you look at their norm or angle ...
Matrices?

- Comparing:

\[
X = \begin{bmatrix}
1. & 1.98 \\
3.05 & 4.
\end{bmatrix}
\quad \text{and} \quad
Y = \begin{bmatrix}
1.1 & 2. \\
3. & 3.9
\end{bmatrix}
\]

\[
X = \begin{bmatrix}
1. & 1.98 \\
3.05 & 4.
\end{bmatrix}
\quad \text{and} \quad
Z = \begin{bmatrix}
100 & 200 \\
300 & 400
\end{bmatrix}
\]

- Use the norm again:

\[\|X - Y\| = 0.13, \quad \|X - Z\| = 541\]
Image matrices

- Similar images, but the distances don’t say so!
Data norms won’t get you far!

• You need to articulate what matters
  • You need to know what matters

• Features are the means to do so

• Let’s examine what matters to us
  • Sounds and images
Human perception

- It is important to know what we perceive
  - We’ve had a few billion years of MLSP refinement!

- Use that knowledge to design features

- Applies to domains we directly perceive
  - Vision, audition, touch, olfaction
  - But helps in figuring out the rest too
    - bio signals, network data, communications, etc
The bigger picture

Human Perception

Machine Learning Signal Processing

Influences

Explains
A bit of hearing

- Sounds and hearing
  - Easy 1D signals to start with

- Human hearing aspects
  - Physiology and psychology

- Lessons learned
The hardware (outer/middle ear)

• The pinna (auricle)
  • Aids sound collection
  • Does directional filtering
  • Holds earrings, etc ...

• The ear canal
  • About 25mm x 7mm
  • Amplifies sound at ~3kHz by ~10dB
  • Helps clarify a lot of sounds!

• Ear drum
  • End of middle ear, start of inner ear
  • Transmits sound as a vibration to the inner ear
More hardware (inner ear)

- **Ear drum (tympanum)**
  - Excites the ossicles (ear bones)

- **Ossicles**
  - Malleus (hammer), incus (anvil), stapes (stirrup)
  - Transfers vibrations from the ear drum to the oval window
  - Amplify sound by ~14dB (peak at ~1kHz)
  - Muscles connected to ossicles control the acoustic reflex (damping in presence of loud sounds)

- **The oval window**
  - Transfers vibrations to the cochlea

- **Eustachian tube**
  - Used for pressure equalization
The cochlea

- The “mechanical to electrical” converter
  - Translates oval window vibrations to a neural signal
  - Fluid filled with the basilar membrane in the middle
  - Each section of the basilar membrane resonates with a different sound frequency
  - Vibrations of the basilar membrane move sections of hair cells which send off neural signals to the brain

- The cochlea acts like the equalizer display in your stereo system
  - Frequency domain decomposition

- Neural signals from the hair cells go to the auditory nerve
Masking & critical bands

• When the basilar membrane is excited by two tones in the same region, the softest of the two will be masked
  • E.g. two sinusoids at 150Hz and 170Hz, if one sinusoid is loud enough the other will be inaudible

• There are 24 such “critical bands”
  • Simultaneous excitation on a band by multiple sources results in some masking

• There is also temporal masking
  • Preceding sounds mask what’s next
Masking example

**Nearby tones that mask**

**Distant tones don’t mask**
The neural pathways

- **Cochlear nuclei**
  - Prepping/distribution of neural data from cochlea

- **Superior Olivary Complex**
  - Coincidence detection across ear signals
  - Localization functions

- **Inferior Colliculus**
  - Last place where we have most original data
  - Probably initiates first auditory images in brain

- **Medial Geniculate Body**
  - Relays various sound features (frequency, intensity, etc) to the auditory cortex

- **Auditory Cortex**
  - Reasoning, recognition, identification, etc
  - High-level processing
The limits of hearing

- **Frequency**
  - 20Hz to 20kHz (upper limit decreases with age/trauma)
  - Infrasound (< 20Hz) can be felt through skin
  - Ultrasound (> 20kHz) can be “emotionally” perceived (discomfort, nausea, etc)

- **Loudness**
  - Low limit is $2 \times 10^{-10}$ atm
  - 0dB SPL to 130dB SPL (frequency dependent)
    - A dynamic range of 3,000,000 to 1!
  - 130dB SPL threshold of pain
  - 194dB SPL is definition of a shock wave, sounds stops!
How high can you hear?

Or, how good are the class speakers?
Perception of loudness

- Loudness is subjective
  - Perceived loudness changes with frequency
  - Perception of “twice as loud” is not really that!

- Fletcher-Munson curves
  - Equal perceived loudness over frequencies

- Just noticeable difference is about 1dB SL

- 1kHz to 5kHz are best heard frequencies
  - What the ear canal and ossicles amplify!

- Low limit shifts up with age!
Perception of pitch

- Pitch is also a subjective (and arbitrary) measure
- Perceived doubling of pitch doesn’t imply a real doubling of frequency
  - Mel scale is the perceptual pitch scale
  - \(2\times\) Mels correspond to a perceived pitch doubling
- Musically useful range varies from 30Hz to 4kHz
- Just noticeable difference is about 0.5% of frequency
  - Varies with training though

“Pitch is that attribute of auditory sensation in terms of which sounds may be ordered from low to high”
- American National Standards Institute
Perception of timbre

- Timbre is what distinguishes sounds outside of loudness & pitch
  - Another bogus ANSI description!

- Timbre is dynamic and can have many facets which often include pitch and loudness

- There is not a coherent body of literature examining human timbre perception
  - But there is a huge bibliography on computational timbre perception!

Gray's timbre space of musical instruments
Binaural hearing

- Having two ears is good!
  - We can localize sounds around us
  - We can wear glasses ...

- Interaural differences provide the cues
  - Interaural time diff (ITD)
  - Interaural intensity diff (IID)
  - Other extra filtering by pinnae, head, and torso

- Detecting these creates a *spatial* percept

- Borrowing these ideas we can design sensor arrays
So how do we use all that?

- All these processes provide an insight
  - They encapsulate important statistics of sounds
  - They suggest features and scales that we should use

- To make machines that cater to our needs
  - We need to learn from our perception
Lessons learned from the cochlea

• Sounds are not vectors of samples
  • This is not what we process when we hear

• Sounds are “groups of frequencies”
  • That is the perceptual feature used

• Frequency representations are what we should use
Emulating the cochlea

- Using the time/frequency domain
  - Take successive DFTs
  - Keep their magnitude
  - Stack them in time
  - Now you can compare sounds!
Example input

![Waveform Example](image.png)
Example features
A lesson from pitch perception

- Frequencies are not “linear”
  - Perceived scale is called mel

- Use that spacing instead
  - i.e. warp the frequency axis
Warped spectra
A lesson from loudness perception

• We don’t perceive loudness linearly either

• How much louder is the second utterance?

• Representation of magnitude shouldn’t be linear
Linear scale to log scale

- I’ve been cheating and showing logarithmic energy already
- It’s hard to really see the data in linear scale
Sound recap

- Go to time/frequency domain
  - We do so in the cochlea
- Frequencies are not linear
  - We perceive them in another scale
- Amplitude is not linear either
  - Use log scale instead
- Resulting features are the backbone of all speech tech!
  - Further tweaks exist (more later)
Seeing

• Images and videos
  • Typical 2/3D signals

• Bit of physiology and psychology

• Using these to represent images
The hardware

- The human eye
  - Lens in the front
  - Retina in the back
- Lens moves the image in place
- Retina translates to neural code
The retina

- Think of it as the equivalent to the cochlea
  - Does a preliminary feature coding

- Photoreceptor cells
  - Mostly rods and cones

- Acts like a film
Rods and cones

- Neurons that convert light to electricity
  - “Phototransduction”

- Cones
  - Detect color, need lots of light
  - ~8 million of them
  - Mostly in the center of the retina

- Rods
  - Not much color, work with minimal light
  - Located all around the retina
  - Many more than cones, ~120 million
Cones and color

- Three kinds of cones
  - Short/Blue
    - React to short wavelengths
    - Very few \((S/(M+L) = .01)\)
  - Medium/Green
    - React to medium wavelengths
  - Long/Red
    - React to long wavelengths
    - More of them \((L/M = 1.5)\)
Receptive fields

• For each cone/rod bunch there’s a decision
  • Fire signals or not?

• This is governed by a “receptive field”
  • On-center cells fire when light in center
  • Off-center cells fire when light off-center
    • Both don’t fire much when light is uniform

• This favors “informative firing”
  • Essentially capturing edges
Going up the brain (super simplified version)

- **Neural pathway**

  - **Color, Depth and Edges**
    - Retina
    - Integration between eyes, some stream distribution
    - LGN

  - **Various features**
    - V2/V3/V4
    - Orientation detection, “Gabor filters”
    - V1

  - **Movement**
    - V5 or MT

  - **Visual cortex**
The V1

- Collection of cells that respond to orientations
- Different groups fire for different orientations
- “Recognition of orientation patterns”
So why do we care about all that?

- Learn from the master!
  - We can see better than machines
  - We are a pretty good source of inspiration

- Remember who you cater to
  - Your vision algorithms will be evaluated by human eyes
  - They need to “speak the same language”
A lesson from the cones

- The colors to use are Red/Green/Blue
  - Superimposing them can create other colors

- Each color pixel is thus represented by three values \(\{r,g,b\}\)
  - We saw that already when using tensors to represent color images
RGB representation

- A perceptually useful representation for color
  - Think of it as a basis set
    - Additive representation
  - Your color screen uses it!

- An alternative is CMYK
  - Cyan/Magenta/Yellow/Black
    - A subtractive representation

- Various other schemes
A lesson from receptive fields

- It’s the visual changes that convey information
How do we detect the middle part?

• Let’s focus on one row only
  • Can we somehow “measure change”?

\[ x = [0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0] \]
\[ z = [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0] \]

**Simple logic:**

\[
\begin{align*}
\text{if } x[i] \neq x[i-1] \text{ then:} \\
\quad \text{change} = \text{True} \\
\text{else:} \\
\quad \text{change} = \text{False}
\end{align*}
\]

**An alternative continuous “change” measure:**

\[ z[i] = \|x[i] - x[i-1]\| \]

**Using a convolution:**

\[ z = \| [1, -1] \ast x \| \]
How about on two dimensions?

- We need three filters:
  - One for horizontal changes
    \[ f_h = [+1, -1] \]
  - One for vertical changes
    \[ f_v = \begin{bmatrix} -1 \\ +1 \end{bmatrix} \]
  - One for diagonal changes
    \[ f_d = \begin{bmatrix} -1 & 0 \\ 0 & +1 \end{bmatrix} \]
Combining everything in one operation

- Convolutions are linear:
  \[ f_h \ast x + f_v \ast x + f_d \ast x = (f_h + f_v + f_d) \ast x \]
  \[ = \begin{bmatrix} -1 & -1 \\ -1 & +3 \end{bmatrix} \ast x \]

- Change detector is now:
  \[ z = \| f \ast x \| \]
Edge detection

• Construct a filter like retinal receptive fields

\[ f = \begin{bmatrix}
  1 & 1 & 1 \\
  1 & 1 & 1 \\
  1 & 1 & 1 \\
  8 & 8 & 8 \\
  8 & 8 & 8 \\
  8 & 8 & 8 \\
\end{bmatrix} \]

• Convolve images with it to produce a representation that peaks at edges
  • Ignores other parts (like constant values)
Edge detection examples

- Reduces input to relevant information only

![Edge detection examples](image)

Input $x$

Edge detection

Inverted input $1-x$

Edge detection on $1-x$
Why is this important?

- We often care only about the shapes, e.g. text

![Diagram showing inputs and feature representation](image)
A lesson from the V1

• More edge detection
  • Describe image using multiple orientations

• Gabor transform
  • Convolve with filters of varying angles

• Feature representation is the presence of each orientation at each point
Mimicking the V1 with filters

Input

Filter responses

- 0° filter
- 90° filter
- 45° filter
- -45° filter
Extending this idea

- We think of images as collections of:
  - Lines/edges, blobs, corners, ridges, ...

- We can make filters for each type
  - Rich computer vision literature on the subject
    - Interesting neuroscience as well, e.g. the Halle Berry neuron

- We’ll be encountering such features later on when we’ll operate on images
Scale space

• Detail is optional
  • We may want it, or we may not
    • e.g. squinting

• Images are often looked at varying "scales"
  • Remove detailed/fine information and focus at coarse elements
Recap

- Features are needed to represent data
  - We can’t just take norms to express distance

- Selection of features caters to our needs
  - Our senses are an “optimal” feature selector
  - What can we learn from it?

- Know what you want to look for
  - Different features work best in different cases
Ok, but what about …

- Non-sensory signals?
  - Mechanical readings, biomedical signals, finance, ...

- How do we get features for these?
  - We need to see the bigger picture in the features we used in this lecture

- Next lecture:
  - Adaptive feature selection
  - How to use math to derive perceptual features, and learn to “perceive” new modalities
Pointers to more info

• Auditory perception and audio features
  • Wikipedia
    • http://en.wikipedia.org/wiki/Auditory_system
    • http://en.wikipedia.org/wiki/Psychoacoustics
  • Textbook section 7.5

• Visual perception and visual features
  • Wikipedia:
    • http://en.wikipedia.org/wiki/Visual_perception
    • http://en.wikipedia.org/wiki/Feature_detection_(computer_vision)
  • Hubel’s book:
    • http://hubel.med.harvard.edu/index.html
Administrivia

- Yes the class is still full ...

- You can still register late with my approval
  - I will give you approval if there are open slots

- If you plan to drop it, please do so soon
And one more thing ...

- Problem Set 1 is out
  - Check for it on the class web page

- Due Sept 22 at midnight
  - See your TA sooner rather than later
Problem set rules

- We will only accept PDFs!
  - Handwritten submissions will not be graded!

- Don’t cheat!
  - Work on your own, avoid peeking at past solutions (if you do we’ll know)

- Extra credit questions
  - How you get an A+ in class (and a recommendation letter from me)
  - For undergrads, the grad questions are the extra credit