Speech Processing
Overview & Supporting Concepts
Professor Marie Roch
Readings are listed in the schedule

What is speech processing?

• Specialized branch of machine learning
• Interdisciplinary
  – signal processing
  – statistics
  – linear algebra
  – linguistics
  – optimization
  – perception & production
Basic ideas

• Acquire an audio signal
• Extract features
• Classify features to labels depending on goals, e.g.
  – text to speech
  – identify a person
  – recognize directives

Probability

• A measure of uncertainty.
• Sources of uncertainty:
  – Process is stochastic (nondeterministic).
  – We cannot observe all aspects of process.
  – Incomplete modeling of process.
Probability

Two ways to think about:
• frequentist – Related to the proportion of events occurring
• Bayesian – Related to degree of belief

Both types of probability are treated using the same rules.

Random variables

• Variables that can be bound to values non-deterministically.

• Common to use notation to distinguish between*
  – a random variable X (capitalized)
  – and an observed value x (lower case, e.g. x=5)

*Goodfellow et al. use bold in place of capitalization.
Discrete random variables

• Values from a discrete set of labels:

\[ X \in \{ \text{red, orange, yellow, blue, indigo, violet} \}, Y \in \{0,1,2\} \]

• Probability density (mass) function (pdf/pmf) is the probability of something occurring, \( P(X=x) \):

\[ P(X = \text{red}) = 0.8, P(Y = 2) = \frac{1}{3} \]

• \( \forall x \in \text{domain}(X), 0 \leq P(X = x) \leq 1 \) and \( \sum_{x \in \text{domain}(X)} P(X = x) = 1 \)

Goodfellow et al. use PMF for discrete distributions. \( P(X=x) \) is commonly abbreviated to \( P(x) \)

Distribution

• Set of probabilities associated with the values of a random variable.

• Uniform distribution – Special distribution where all probabilities are equal.
Example

• Let X be a die roll
  – P(X=3) denotes the probability of rolling a 3
  – Fair die \( \Rightarrow P(X=3) = \frac{1}{6} \)

• If we don’t know if the die is fair, how would we approximate \( P(X=x) \)?

• Estimates of probability are denoted with a hat: \( \hat{P}(X = x) \)

Continuous random variables

• Values from a continuous domain
• Satisfies the following:
  \[
  \forall x \in \text{domain}(X), 0 \leq P(X = x)
  \]
  \[
  P(a \leq X \leq b) = \int_a^b P(X = x)dx
  \]
  \[
  \int_x P(X = x)dx = 1
  \]
• Worth noting:
  – no max on \( P(X=x) \)
  – What is? \( P(a \leq X \leq a) = \int_a^a P(X = x)dx \)
Continuous uniform distribution

\[ X \sim U(a,b) \]

implies \( P(X = x) = \begin{cases} 
0 & x < a \text{ or } x > b \\
\frac{1}{b-a} & a \leq x \leq b 
\end{cases} \)

In general, \( \sim \) means “has a distribution of.”
U(a,b) is a uniform distribution between values and b

Expectation

discrete: \( E[u(X)] = \sum_x u(x)P(x) \)

continuous: \( E[u(X)] = \int_x u(x)P(x)dx \)

When \( u(X) = X \), we call this the mean, or average value:
\[ \mu = E[X] \]
Sample mean $\hat{\mu}$

- What if we don’t know $P(X=x)$?
- If we roll a die many times, we can estimate $P(X=x)$ by counting the number of times each $x$ occurs and dividing by the number of rolls.

  - To think about: How does this relate to our traditional idea of an average?

Variance

- Answers the question: What is the expected squared deviation from the mean?

- Can be defined with expectation operator

$$ E[(X - \mu)^2] = \sum_x (x - \mu)^2 P(X = x) $$

- Sample variance (unbiased)

$$ \hat{E}[(x - \mu)^2] = \frac{1}{N-1} \sum_x (x - \mu)^2 $$
Normal distribution

• Normal or Gaussian distributions are the so-called “bell curve” distributions

• Parameters: mean & variance:  \( X \sim n(\mu, \sigma^2) \)

\[
P(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

\[X \sim n(0,1), \text{ the standard normal}\]

• Mean controls the center, variance the breadth of the distribution.

• A normal distribution can be fit with sample mean & variance
Application: Speech/noise segmentation

How can we determine if someone is talking?

1. Record: Acquire pressure signal
2. Frame: Break into pieces
3. Compute features: root mean square intensity
4. Train: Fit speech/noise distribution models
5. Classify: See which distribution matches new samples

Acquisition: pressure sensors

Basic ideas for microphones:

• Deformable membrane
  – pushed by compression
  – pulled by rarefaction
• Deformation is converted to voltage
• Sample: voltage measured N times/second
  – called sample rate (Fs) and
  – measured in Hertz (Hz)
Pressure and Intensity

- Pressure
  - Amount of force per area
  - Typically measured in Pascals if the microphone is calibrated otherwise in counts
- Intensity
  - Product of sound pressure and particle velocity per area
  - $\text{Pressure}^2 \alpha \text{Intensity}$

Framing

- Most audio signals vary over time
- Analyze small segments that are roughly static

chimpanzee pant hoot – courtesy Cat Hobaiter
decibel (dB) scale

- Human auditory sensitivity to pressure:
  20 µPa – 200 Pa (10^7 range!)
- The decibel scale allows us to compare the intensity between two sounds in a compressed range:
  \[
  \text{Intensity} = 10 \log_{10} \left( \frac{I}{I_0} \right)
  \]
  \[
  \sim \text{ with pressure} = 10 \log_{10} \left( \frac{P}{P_0} \right)^2 = 20 \log_{10} \left( \frac{P}{P_0} \right)
  \]

What value of \( P_0 \)?

- Calibrated terrestrial systems:
  \( P_0 = 20 \mu \text{Pa} \) (threshold of hearing)
  denoted dB re 20 µPa
- Uncalibrated systems:
  \( P_0 = 1 \) (for convenience)
  denoted dB rel.
Computing intensity

- Computed for each frame
- Samples (pressure) are
  - converted to root mean square (RMS):
    - squared
    - averaged
  - converted to dB: $10\log_{10}(\text{RMS})$

Our first speech problem

- Given a recording, separate speech from constant background noise
Relative Intensity

Intensity histogram
So how can we characterize speech and silence?

Silence estimator

- Beginning of recording vs.
- Entire recording
Normal fit of silence

Similar for speech 1.3 – 2.8 s
Intensity distribution with pdfs

Where should we draw the boundary?

Conditional probability

The probability of something occurring can change when you know something…

\[ P(X = "nice to meet you") \]

vs

\[ P(X = "nice to meet you" \mid \text{meeting new person}) \]
Conditional probability

The probability of something occurring can change when you know something…

\[ P(X = "\text{nice to meet you}") \]

vs

\[ P(X = "\text{nice to meet you}| \text{meeting new person}) \]

A problem…

Our intensity pdfs showed

\[ P(X | \text{speech}) \text{ and } P(X | \text{noise}) \]

but we would like to know

\[ P(\text{speech} | X) \text{ and } P(\text{noise} | X) \]

which is known as the **posterior distribution**
Bayes’ decision rule

The optimal classification rule is the one that maximizes the posterior probability:

\[
\arg \max_{\omega \in \{\text{speech, noise}\}} P(\omega | X)
\]

Regrettably, we don’t know \( P(\omega | X) \), but we do know \( P(X | \omega) \)

Bayes’ rule

(of conditional probability)

\[
P(A | B) = \frac{P(A, B)}{P(B)} \triangleq \text{conditional probability}
\]

note that swapping names A and B

\[
P(B | A) = \frac{P(B, A)}{P(A)} \Rightarrow P(B, A) = P(B | A)P(A)
\]

and \( P(A, B) = P(B, A) \)

\[
\therefore P(A | B) = \frac{P(B | A)P(A)}{P(B)}
\]

also known as Bayes’ Theorem/Bayes’ Law)
Bayes’ decision rule

Bayes’ rule to the rescue!

\[
\arg \max_{\omega} P(\omega \mid X) = \frac{P(X \mid \omega)P(\omega)}{P(X)}
\]

We know \(P(X \mid \omega)\). We don’t need to know \(P(X)\) as

\[
\max \left( \frac{P(X \mid \text{speech})P(\text{speech})}{P(X)}, \frac{P(X \mid \text{noise})P(\text{noise})}{P(X)} \right)
= \max \left( P(X \mid \text{speech})P(\text{speech}), P(X \mid \text{noise})P(\text{noise}) \right)
\]
Prior probability

- Probability of an observation without any prior information.
- When we don’t know this, we frequently assume a uniform (equally likely) distribution:

\[
= \max \left( P(X \mid \text{speech}) \frac{1}{2}, P(X \mid \text{noise}) \frac{1}{2} \right)
\]

Bayes decision rule revisited

Assuming a uniform prior, we can now make decisions about speech or noise:

\[
\text{decision}(x) = \arg \max_{\omega \in \{\text{speech, noise}\}} \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu_\omega)^2}{2\sigma^2}}
\]

How can we solve the boundary threshold?
Threshold for our minimal speech detector

Solve for x (2 roots):

\[
\frac{1}{\sqrt{2\pi\sigma^2_{\text{noise}}}} e^{-\frac{(x-\mu^2_{\text{noise}})}{2\sigma^2_{\text{noise}}}} = \frac{1}{\sqrt{2\pi\sigma^2_{\text{speech}}}} e^{-\frac{(x-\mu^2_{\text{speech}})}{2\sigma^2_{\text{speech}}}}
\]

Result

Caveat: Example is for pedagogical purposes, we can do better than this…