This lecture

• Speech signals
• Articulatory phonetics

• Some images from Gray’s Anatomy, Jim Glass’ course 6.345 (MIT), the Jurafsky & Martin textbook, Encyclopedia Britannica, the Rolling Stones, the Pink Floyd.
Speech signals
What is sound?

- **Sound** is a time-variant pressure wave created by a vibration.
  - Air particles **hit** each other, setting others in motion.
  - High pressure $\equiv$ **compressions** in the air (C).
  - Low pressure $\equiv$ **rarefactions** within the air (R).
What is sound?

Frequency $F = 1/T$

Phase $\phi$ is displacement of a signal in time. E.g., with $\phi = \pi/2$,

$$\sin(x + \phi) = \cos(x)$$
What is sound?

- A single **tone** is a sinusoidal function of pressure and time.
  - **Amplitude**: \( n \). The degree of the displacement in the air. This is similar to ‘loudness’.
  - **Frequency**: \( n \). The number of cycles within a unit of time. e.g., 1 Hertz (Hz) = 1 oscillation/second
The inner ear

- Time-variant waves enter the ear, vibrating the **tympanic membrane**.
- This membrane causes tiny bones (the **ossicles**) to vibrate.
- These bones in turn vibrate a structure within a shell-shaped bony structure called the **cochlea**.
The cochlea and basilar membrane

- The basilar membrane is covered with tiny hair-like nerves – some near the base, some near the apex.

- High frequencies are picked up near the base, low frequencies near the apex.

- These nerves fire when activated, and communicate to the brain.
The Mel-scale

• Human hearing is **not** equally sensitive to **all** frequencies.
  • We are **less** sensitive to frequencies > 1 kHz.

• A **mel** is a unit of pitch. Pairs of sounds which are **perceptually** equidistant in pitch are separated by an equal number of **mels**.

\[
Mel(f) = 2595 \log_{10} \left( 1 + \frac{f}{700} \right)
\]
Speech waveforms

“Two plus seven is less than ten”
Superposition of sinusoids

- **Superposition**: *n.* the adding of sinusoids together.
Extracting sinusoids from waveforms

- As we will soon see, the relative **amplitudes** and **frequencies** of the sinusoids that combine in speech are often extremely **indicative** of the **speech units** being uttered.
  - ∴ If we could **separate** the waveform into its component sinusoids, it would help us **classify** the speech being uttered.
  - **But the shape of the signal changes over time**
    (it’s not a single repeating pattern)...
Short-time windowing

- Speech waveforms change drastically over time.
- We *move* a short analysis window (assumed to be time-invariant) across the waveform in time.
  - E.g. frame shift: 10-30 ms
  - E.g. frame length: 25-40 ms
Window types

Rectangular window

Hamming window

Hamming eliminates ‘clipping’ at the boundaries of windows.
Extracting a spectrum

White light

Any Colour You Like (track 8)
Extracting a spectrum in a window

Frame -> Spectrum

Spectrum

Amplitude

Frequency (Hz)
Euler’s formula

• Extracting sinusoids uses a relationship between natural exponent $e$ and sinusoids expressed in **Euler’s formula**:

\[ e^{i\psi} = \cos(\psi) + i \sin(\psi) \]

\[ e^{i\pi} = -1 \]
The continuous Fourier transform

- **Input:** Continuous signal $x(t)$.
- **Output:** Spectrum $X(F)$

\[ X(F) = \int_{-\infty}^{\infty} x(t) e^{-i2\pi F t} \, dt \]

- It’s **invertible**, i.e., $x(t) = \int_{-\infty}^{\infty} X(F) e^{i2\pi F t} \, dF$.
- It’s **linear**, i.e., for $a, b \in \mathbb{C}$,
  - if $h(t) = ax(t) + by(t)$,
  - then $H(F) = aX(F) + bY(F)$

Fun fact: Fourier instructed Champollion.

... It needs **continuous** input $x(t)$... *uh oh?*
Discrete signal representation

- **Sampling**: *vbg.* measuring the amplitude of a signal at regular intervals.
  - e.g., 44.1 kHz (*CD*), 8 kHz (*telephone*).
  - These amplitudes are initially measured as **continuous** values at **discrete** time steps.

How quickly should we sample?
Discrete signal representation

- **Nyquist rate:** *n.* the **minimum** sampling rate necessary to preserve a signal’s **maximum** frequency.
  - i.e., **twice** the maximum frequency, since we need $\geq 2$ samples/cycle.
  - Human speech is very informative $\leq 4$ kHz, $\therefore$ At least 8 kHz sampling (16kHz the norm)
Discrete Fourier transform (DFT)

• **Input:** Windowed signal \( x[0] \ldots x[N-1] \).

• **Output:** \( N \) complex numbers \( X[k] \) \((k \in \mathbb{Z})\)

\[
X[k] = \sum_{n=0}^{N-1} x[n] e^{-i2\pi k \frac{n}{N}}
\]

• **Algorithm(s):** the Fast Fourier Transform (FFT) with complexity \( O(N \log N) \).
Discrete Fourier transform (DFT)

- Below is a 25 ms Hamming-windowed signal from /iy/ as in ‘bull sheep’, and its spectrum as computed by the DFT.

Recall: the Fourier transform is invertible.

But this is all just for a small window...
Spectrograms

- **Spectrogram**: *n.* a 3D plot of **amplitude** and **frequency** over **time** (higher ‘redness’ → higher amplitude).
Effect of window length

Wide-band (better time resolution)

Narrow-band (better frequency resolution)
Spectrograms

“Two plus seven is less than ten”

How are these obvious patterns made and perceived?
Articulatory phonetics
Sounds and transcriptions

• We are often interested in the meaning of an utterance
• In English, we often transcribe utterances as word tokens
  • We write: <How to recognize speech>
• Is this “what was said?”
  • We might write instead: <How to wreck a nice beach>
  • We can transcribe (or even adopt) foreign words
    • 沙发 = <sofa>, not <sandy hair>
  • We can even transcribe brand new words
    • <skibidi toilet>
• We can instead transcribe “speech sounds”
Phones and phonetics

- **Phonetics** is the study of speech sounds
- A **phone** is a unit of speech
  - Denoted with square braces: [t], [tʰ], [u]
  - Language-independent
- Phones which are perceived “similarly” are grouped into **phonemes**
  - Denoted with slashes: /t/, /u/
  - [t],[tʰ] ↦ /t/
  - Language-dependent
- Transcriptions are often in-between:
  - [ˈtʰuː] ↦ [tʰu] ↦ [tu] ↦ /tu/
- We will be very loose with the distinction
Phonetic transcription

- Often, we assume that a spoken utterance can be partitioned into a sequence of non-overlapping phones.
  - Demarking the periods during which certain phones are being uttered is called **phonetic transcription**
  - This approach has problems (e.g., when *exactly* does one phoneme end and another begin?), but it’s useful for **classification**.
## Phonetic alphabets

- There are several alphabets that categorize the sounds of speech.
  - The **International Phonetic Alphabet (IPA)** is popular, but it uses non-ASCII symbols.
  - The **TIMIT** phonetic alphabet will be used by default in this course.

- Other popular alphabets include **ARPAbet**, **Worldbet**, and **OGIbet**, usually adding special cases.
  - E.g., [pɛl] is the period of silence immediately before a [p].

### TIMIT | IPA | e.g.
--- | --- | ---
[yi] | [iy] | *beat*
[ih] | [i] | *bit*
[eh] | [ɛ] | *bet*
[ae] | [æ] | *bat*
[aa] | [ɑ] | *Bob*
[ah] | [ʌ] | *but*
[ao] | [ɔ] | *bought*
[uh] | [ʊ] | *book*
[uw] | [u] | *boot*
[ux] | [u] | *suit*
[ax] | [ə] | *about*
## TIMITbet (incomplete)

<table>
<thead>
<tr>
<th>Vowel</th>
<th>e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[iy]</td>
<td>beat</td>
</tr>
<tr>
<td>[ih]</td>
<td>bit</td>
</tr>
<tr>
<td>[eh]</td>
<td>bet</td>
</tr>
<tr>
<td>[ae]</td>
<td>Bat</td>
</tr>
<tr>
<td>[aa]</td>
<td>Bob</td>
</tr>
<tr>
<td>[ah]</td>
<td>But</td>
</tr>
<tr>
<td>[ao]</td>
<td>bought</td>
</tr>
<tr>
<td>[uh]</td>
<td>book</td>
</tr>
<tr>
<td>[uw]</td>
<td>boot</td>
</tr>
<tr>
<td>[ux]</td>
<td>suit</td>
</tr>
<tr>
<td>[ax]</td>
<td>about</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>stop</th>
<th>e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[b]</td>
<td>Bilbo</td>
</tr>
<tr>
<td>[d]</td>
<td>dada</td>
</tr>
<tr>
<td>[g]</td>
<td>Gaga</td>
</tr>
<tr>
<td>[p]</td>
<td>Pippin</td>
</tr>
<tr>
<td>[t]</td>
<td>Toots</td>
</tr>
<tr>
<td>[k]</td>
<td>kick</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nasal</th>
<th>e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[m]</td>
<td>Mama</td>
</tr>
<tr>
<td>[n]</td>
<td>noon</td>
</tr>
<tr>
<td>[ng]</td>
<td>thing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>fricative</th>
<th>e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[s]</td>
<td>Sea</td>
</tr>
<tr>
<td>[f]</td>
<td>Frank</td>
</tr>
<tr>
<td>[z]</td>
<td>Zappa</td>
</tr>
<tr>
<td>[th]</td>
<td>this</td>
</tr>
<tr>
<td>[sh]</td>
<td>Ship</td>
</tr>
<tr>
<td>[zh]</td>
<td>azure</td>
</tr>
<tr>
<td>[v]</td>
<td>Vogon</td>
</tr>
<tr>
<td>[dh]</td>
<td>then</td>
</tr>
</tbody>
</table>

... (Incomplete)
The vocal tract

- Many physical structures are co-ordinated in the production of speech.
- Generally, sound is generated by passing air through the vocal tract.
- Sound is modified by constricting airflow in particular ways.
- We can classify phones by how they are produced.
A taxonomy of phones

• Phones fall into two broad categories

  • **Vowels** are
    • Always periodic
    • Produced with relatively unobstructed airflow
    • Use tongue, lips, and jaw to produce resonances in vocal tract, in turn generating formants

  • **Consonants** are
    • Mostly noisy (not nasals, semivowels)
    • Produced by obstructing airflow
    • Classified by the **place** and **manner** of primary obstruction, as well as **voicing**
Voicing and fundamental frequency

- **Voiced** phones are produced with vibrating **vocal folds**
  - The space between the folds is the **glottis**
- All vowels are voiced; consonants can be **unvoiced**
- $F_0$: *n. (fundamental frequency)*, the rate of vibration (Hz)
  - Very indicative of speaker

<table>
<thead>
<tr>
<th></th>
<th>Avg $F_0$ (Hz)</th>
<th>Min $F_0$ (Hz)</th>
<th>Max $F_0$ (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>125</td>
<td>80</td>
<td>200</td>
</tr>
<tr>
<td>Female</td>
<td>225</td>
<td>150</td>
<td>350</td>
</tr>
<tr>
<td>Children</td>
<td>300</td>
<td>200</td>
<td>500</td>
</tr>
</tbody>
</table>
Vowels

- There are approximately 19 vowels in Canadian English, including diphthongs in which the articulators move over time.

- Vowels are distinguished primarily by their formants. (?)

<table>
<thead>
<tr>
<th>diphthong</th>
<th>e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ey]</td>
<td>bait</td>
</tr>
<tr>
<td>[ow]</td>
<td>boat</td>
</tr>
<tr>
<td>[ay]</td>
<td>bite</td>
</tr>
<tr>
<td>[oy]</td>
<td>boy</td>
</tr>
<tr>
<td>[aw]</td>
<td>bout</td>
</tr>
<tr>
<td>[ux]</td>
<td>suit</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>other</th>
<th>e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[er]</td>
<td>Bert</td>
</tr>
<tr>
<td>[axr]</td>
<td>butter</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Monophthong</th>
<th>e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[iy]</td>
<td>beat</td>
</tr>
<tr>
<td>[ih]</td>
<td>bit</td>
</tr>
<tr>
<td>[eh]</td>
<td>bet</td>
</tr>
<tr>
<td>[ae]</td>
<td>bat</td>
</tr>
<tr>
<td>[aa]</td>
<td>Bob</td>
</tr>
<tr>
<td>[ao]</td>
<td>bought</td>
</tr>
<tr>
<td>[ah]</td>
<td>but</td>
</tr>
<tr>
<td>[uh]</td>
<td>book</td>
</tr>
<tr>
<td>[uw]</td>
<td>boot</td>
</tr>
<tr>
<td>[ax]</td>
<td>about</td>
</tr>
<tr>
<td>[ix]</td>
<td>roses</td>
</tr>
</tbody>
</table>
Uniform tubes

- Formants and resonances can be approximated with tubes
- Many **musical instruments** are based on the idea of uniform (or, in many cases, bent) tubes.
- **Longer** tubes produce ‘**deeper**’ sounds (lower frequencies).
  - A tube \( \frac{1}{2} \) the length of another will be 1 octave higher.
The uniform tube

- The positions of the tongue, jaw, and lips change the shape and cross-sectional area of the vocal tract.
Vowels as concatenated tubes

- The vocal tract can be modelled as the concatenation of dozens, hundreds, or thousands of tubes.
Waves in concatenated tubes

- Reflections at tube boundaries produce resonances which amplify certain frequencies
Formants and vowels

- **Formant**: *n.* A concentration of energy within a frequency band. Ordered from low to high bands (e.g., $F_1$, $F_2$, $F_3$).
Tongues, lips, and formants

Front/low/unrounded

Front/high/unrounded

Back/high/rounded
The vowel trapezoid

$F_1$ increases

$F_2$ increases
Manner of articulation

• Consonants are classified by place and manner of obstruction
• For manner:
  • **Fricatives:** noisy, with air passing through a tight constriction (e.g., ‘shift’).
  • **Stops/plosives:** complete vocal tract constriction and burst of energy (e.g., ‘papa’).
  • **Nasals:** air passes through the nasal cavity (e.g., ‘mama’).
  • **Semivowels:** similar to vowels, but typically with more constriction (e.g., ‘wall’).
  • **Affricates:** Alveolar stop followed by fricative.
  • **Taps:** Quick collision of articulators (‘butter’).
Place of articulation

- The **location** of the *primary constriction* can be:
  - **Alveolar**: constriction near the alveolar ridge (e.g., [t])
  - **Bilabial**: touching of the lips together (e.g., [m], [p])
  - **Dental**: constriction of/at the teeth (e.g., [th])
  - **Palatal**: constriction at the (hard) palate (e.g., [sh])
  - **Labiodental**: constriction between lip and teeth (e.g., [f])
  - **Velar**: constriction at or near the velum (e.g., [k])
  - **Glottal**: constriction of the glottis ([q])
Fricatives

- Fricatives are caused by acoustic turbulence at a **narrow constriction** whose position determines the sound.
Fricatives

- Fricatives have four places of articulation.
- Each place of articulation has a **voiced** fricative (i.e., the folds can be vibrating), and an **unvoiced** fricative.

<table>
<thead>
<tr>
<th>Place of Articulation</th>
<th>Unvoiced</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labio-dental</td>
<td>![f]</td>
<td>![fee]</td>
</tr>
<tr>
<td>Dental</td>
<td>![th]</td>
<td>![thief]</td>
</tr>
<tr>
<td>Alveolar</td>
<td>![s]</td>
<td>![see]</td>
</tr>
<tr>
<td>Palatal</td>
<td>![sh]</td>
<td>![she]</td>
</tr>
</tbody>
</table>

| Voiced | |
|--------| |
| ![Vendetta] | |
| ![Thee] | |
| ![Zardoz] | |
| ![Zhα-zhα] | |
Unvoiced fricatives
Plosives (3/6)

- **Plosives** build pressure behind a complete closure in the vocal tract.
- A sudden release of this constriction results in brief noise.
Plosives

- **Plosives** have three places of articulation:

<table>
<thead>
<tr>
<th></th>
<th>Unvoiced</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labial</td>
<td>[p]</td>
<td><em>porpoise</em></td>
</tr>
<tr>
<td>Alveolar</td>
<td>[t]</td>
<td><em>tort</em></td>
</tr>
<tr>
<td>Velar</td>
<td>[k]</td>
<td><em>kick</em></td>
</tr>
</tbody>
</table>

- **Voiced** stops are usually characterized by a “**voice bar**” during closure, indicating the vibrating vocal folds.
- Formant **transitions** are very **informative** in classification.
Formant transitions in plosives

- Despite a common vowel, the motion of $F_2$ and $F_3$ into (and out of) the vowel helps identify the plosive.
Voicing in plosives

The “voice bar”
Nasals

- Nasals involve lowering the velum so that air passes through the nasal cavity.
- Closures in the oral cavity (at same positions as plosives) change the resonant characteristics of the nasal sonorant.
Formant transitions among nasals

• Nasals often appear as two formants
Semivowels

- **Semivowels** act as consonants in syllables and involve constriction in the vocal tract, but there is **less turbulence**.
  - They also involve slower articulatory motion.
- **Laterals** involve airflow around the **sides** of the tongue.
Semivowels

- Note the drastic formant transitions which are more typical of semivowels.
Affricates and aspirants

- There are two affricates: [jh] (voiced; e.g., judge) and [ch] (unvoiced; e.g., church).
  - These involve an alveolar stop followed by a fricative.
  - Voicing in [jh] is normally indicated by voice bars, as with plosives.

- There’s only one aspirant in Canadian English: [h] (e.g., hat)
  - This involves turbulence generated at the glottis,
  - In Canadian English, there is no constriction in the vocal tract.
Affricates and aspirants

each

huge
Other topics in phonetics

• The grouping of phones into syllables
  • Consisting of a vowel (nucleus), and optionally preceding (onset) and succeeding (coda) consonants
  • Only certain sequences are permissible in English
  • Syllables may be made more prominent via pitch, duration, or loudness

• The prosody, or intonation and rhythm, of an utterance
  • Prominence can also indicate phrase boundaries
  • Gradual F0 movement (tune) can indicate a question or statement

• These are especially important to text-to-speech