This lecture

• Acoustics.
• Speech production.
• Speech perception.

• Some images from Gray’s Anatomy, Jim Glass’ course 6.345 (MIT), the Jurafsky & Martin textbook, Encyclopedia Britannica, the Rolling Stones, the Pink Floyds.
acoustics
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).

![Image of a Newton's cradle](image-url)
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’.
    Often measured in **Decibels (dB)**.
  - **Frequency**: *n.* The number of cycles within a unit of time.
    e.g., 1 Hertz (Hz) = 1 oscillation/second
Speech waveforms

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

• **Superposition**: *n.* the adding of sinusoids together.

• **Phase**: *n.* The horizontal offset of a sinusoid ($\phi$).
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: 5—10 ms
  - E.g. frame length: 10—25 ms
Window types

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
- Frequency (Hz)
- Amplitude
Aside – Euler’s formula

- Extracting sinusoids is possible because of a relationship between $e$ and sinusoids expressed in Euler’s formula:

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

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

- **Input:** Continuous signal $x(t)$.

- **Output:** Spectrum $X(F)$

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

- 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 $H(F) = aX(F) + bY(F)$

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

Fun fact: Fourier instructed Champollion.
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, \( \therefore \) 8 kHz sampling.
Discrete Fourier transform (DFT)

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

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

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

- **Algorithm(s):** the Fast Fourier Transform (FFT) with complexity \( O(N \log N) \).
  - (Aside) The Cooley-Tukey algorithm divides-and-conquers by breaking the DFT into smaller ones \( N = N_1N_2 \).
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.

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?
Aside – Filtering

• Sometimes you only want part of a signal.
  • E.g., you have measurements of lip aperture over time – you know that they can’t move > 5-10 Hz.
  • E.g., you know there’s some low-frequency Gaussian noise in either the environment or transmission medium.

• Low- and high-pass filters can be combined in series, yielding a band-pass filter.
speech production
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.
The neurological origins of speech

- Studying how systems break down can indicate how they work.

- **Reduced** hierarchical syntax.
- **Anomia**.
- **Reduced** “mirroring” between observation and execution.
- **Normal** intonation/rhythm.
- **Meaningless** words.
- ‘**Jumbled**’ syntax.
- **Reduced** comprehension.
The neurological origins of speech

• Cranial nerves carry messages from the brain to the various articulators.

• **Cranial nerves** carry messages from the brain to the various articulators.
  • Damage to these nerves can result in neuro-motor disorders such as cerebral palsy.
  • These may be another example of the noisy channel.
Fundamental frequency

- \( F_0 \): n. (fundamental frequency), the rate of vibration of the glottis – often very indicative of the 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>Men</td>
<td>125</td>
<td>80</td>
<td>200</td>
</tr>
<tr>
<td>Women</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>
Prosody

- **Sonorant**: *n.* Any **sustained** sound in which the **glottis** is vibrating (i.e., the sound is ‘**voiced**’).
  - Includes some consonants (e.g., /w/, /m/).

- **Prosody**: *n.* the **modification** of speech acoustics in order to convey some **extra-lexical** meaning:
  - **Pitch**: Changing of $F_0$ over time.
  - **Duration**: The length in time of sonorants.
  - **Loudness**: The amount of **energy** produced by the **lungs**.
Pitch prosody example

\[ F_0 \text{ (Hz)} \]

\[ F_0 \text{ (Hz)} \]
Pitch can modify meaning

• e.g., I ask you “who is that?”

• e.g., I ask you “what is his job?”

Pitch tends to rise when uttering novel or important information.
Pitch can modify meaning

• *I* never said she stole my money.  (Someone else said it)
• *never* said she stole my money.  (It never happened)
• I never *said* she stole my money.  (I just hinted at it)
• I never said *she* stole my money.  (Someone else stole it)
• I never said she *stole* my money.  (She just borrowed it)
• I never said she stole *my* money.  (She stole someone else’s)
• I never said she stole my *money*.  (She stole my heart).
Phonemes

• **Phoneme**: *n.* a distinctive unit of speech sound.

• Phonemes can be partitioned into **manners of articulation**:
  - **Vowels**: open vocal tract, no nasal air.
  - **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.
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/)
  - **Labiodental:** constriction between lip and teeth (e.g., /f/)
  - **Velar:** constriction at or near the velum (e.g., /k/).
Phonemic 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 phonemic alphabet will be used by default in this course.

• Other popular alphabets include ARPAbet, Worldbet, and OGIbet, usually adding special cases.
  - E.g., /pcl/ is the period of silence immediately before a /p/.
TIMIT Phonemic alphabet *(incomplete)*

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

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

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

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

*(Incomplete)*
Phoneme sequences

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

What are some characteristics of the six **manner**s of articulation?
Vowels (1/6)

• 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>/ax/</td>
<td>about</td>
</tr>
<tr>
<td>/ux/</td>
<td>suit</td>
</tr>
<tr>
<td>/ix/</td>
<td>roses</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>/axr/</td>
<td>butter</td>
</tr>
</tbody>
</table>

- There are approximately 19 vowels in Canadian English, including diphthongs in which the articulators move over time.
- Vowels are distinguished primarily by their formants. (?)

\[\text{Vowels (1/6)}\]

\[\text{• There are approximately 19 vowels in Canadian English, including diphthongs in which the articulators move over time.}\]

\[\text{• Vowels are distinguished primarily by their formants. (?)}\]
The uniform tube

- The positions of the tongue, jaw, and lips change the shape and cross-sectional area of the vocal tract.
Uniform tubes in practice

• 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 ½ the length of another will be 1 octave higher.
Vowels as concatenated tubes

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

- We model the **volume velocity** $U_k$ and the **pressure variation** $p_k$ at position $x$ in the $k^{th}$ lossless tube (whose area is $A_k$) at time $t$

\[
U_k(x, t) = U_k^+ \left( t - \frac{x}{c} \right) - U_k^- \left( t + \frac{x}{c} \right)
\]

\[
p_k(x, t) = \frac{\rho c}{A_k} \left[ U_k^+ \left( t - \frac{x}{c} \right) + U_k^- \left( t + \frac{x}{c} \right) \right]
\]

where
- $c$ is the speed of sound,
- $\rho$ is the density of air.
Waves in concatenated tubes

• Because of partial wave \textit{reflections} that occur at tube boundaries, we can generate spectra with particular \textit{resonances}.
Formants and vowels

• **Formant**: *n.* A concentration of energy within a frequency band. Ordered from low to high bands (e.g., $F_1$, $F_2$, $F_3$).
The vowel trapezoid

$F_1$ increases

$F_2$ increases
Tongues and formants

Front/low

Front/high

Back/high
Fricatives (2/6)

- 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 glottis can be vibrating), and an **unvoiced** fricative.

<table>
<thead>
<tr>
<th></th>
<th>Unvoiced</th>
<th>Voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labial</td>
<td>/f/</td>
<td>/v/</td>
</tr>
<tr>
<td></td>
<td><strong>fee</strong></td>
<td><strong>V</strong>endetta</td>
</tr>
<tr>
<td>Dental</td>
<td>/th/</td>
<td>/dh/</td>
</tr>
<tr>
<td></td>
<td><strong>thief</strong></td>
<td><strong>Th</strong>ee</td>
</tr>
<tr>
<td>Alveolar</td>
<td>/s/</td>
<td>/z/</td>
</tr>
<tr>
<td></td>
<td><strong>see</strong></td>
<td><strong>Z</strong>ardo<strong>z</strong></td>
</tr>
<tr>
<td>Palatal</td>
<td>/sh/</td>
<td>/zh/</td>
</tr>
<tr>
<td></td>
<td><strong>she</strong></td>
<td><strong>Zh</strong>a-<strong>zh</strong>a</td>
</tr>
</tbody>
</table>
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**.

![Diagram of Plosive Articulation](image-url)
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>porpoise</td>
</tr>
<tr>
<td></td>
<td>/b/</td>
<td>baboon</td>
</tr>
<tr>
<td>Alveolar</td>
<td>/t/</td>
<td>tort</td>
</tr>
<tr>
<td></td>
<td>/d/</td>
<td>dodo</td>
</tr>
<tr>
<td>Velar</td>
<td>/k/</td>
<td>kick</td>
</tr>
<tr>
<td></td>
<td>/g/</td>
<td>Google</td>
</tr>
</tbody>
</table>

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

The “voice bar”
Despite a common vowel, the motion of $F_2$ and $F_3$ into (and out of) the vowel helps identify the plosive.
Nasals (4/6)

- **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

- Despite a common vowel, the motion of $F_2$ and $F_3$ before and after each nasal helps to identify it.

Nasals often appear as two formants
Semivowels (5/6)

• **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

- Semivowels are often sub-classified as glides or liquids.

<table>
<thead>
<tr>
<th></th>
<th>Semivowel</th>
<th>Nearest vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glides</td>
<td>/w/</td>
<td>Wow</td>
</tr>
<tr>
<td></td>
<td>/y/</td>
<td>yoyo</td>
</tr>
<tr>
<td>Liquids</td>
<td>/r/</td>
<td>rear</td>
</tr>
<tr>
<td></td>
<td>/l/</td>
<td>Lulu</td>
</tr>
</tbody>
</table>

- Semivowels are more constricted versions of corresponding vowels.
  - Similar formants, though generally weaker.
Semivowels

- Note the drastic formant transitions which are more typical of semivowels.
Affricates and aspirants (6/6)

• 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
Alternative pronunciations

- Pronunciations of words can vary significantly, but with observable frequencies.
- The Switchboard corpus is a phonetically annotated database of speech recorded in telephone conversations.

<table>
<thead>
<tr>
<th></th>
<th>because</th>
<th></th>
<th>about</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ARPAbet %</td>
<td>ARPAbet %</td>
<td>ARPAbet %</td>
<td>ARPAbet %</td>
<td>ARPAbet %</td>
</tr>
<tr>
<td>b iy k ah z 27%</td>
<td>k s 2%</td>
<td>ax b aw 32%</td>
<td>b ae 3%</td>
<td></td>
</tr>
<tr>
<td>b ix k ah z 14%</td>
<td>k ix z 2%</td>
<td>ax b aw t 16%</td>
<td>b aw t 3%</td>
<td></td>
</tr>
<tr>
<td>k ah z 7%</td>
<td>k ih z 2%</td>
<td>b aw 9%</td>
<td>ax b aw dx 3%</td>
<td></td>
</tr>
<tr>
<td>k ax z 5%</td>
<td>b iy k ah zh 2%</td>
<td>ix b aw 8%</td>
<td>ax b ae 3%</td>
<td></td>
</tr>
<tr>
<td>b ix k ax z 4%</td>
<td>b iy k ah s 2%</td>
<td>ix b aw t 5%</td>
<td>b aa 3%</td>
<td></td>
</tr>
<tr>
<td>b ih k ah z 3%</td>
<td>b iy k ah 2%</td>
<td>ix b ae 4%</td>
<td>b ae dx 3%</td>
<td></td>
</tr>
<tr>
<td>b ax k ah z 3%</td>
<td>b iy k aa z 2%</td>
<td>ax b ae dx 3%</td>
<td>ix b aw dx 2%</td>
<td></td>
</tr>
<tr>
<td>k uh z 2%</td>
<td>ax z 2%</td>
<td>b aw dx 3%</td>
<td>ix b aa t 2%</td>
<td></td>
</tr>
</tbody>
</table>
Known effects of pronunciation

- Speakers tend to **drop** or **change** pronunciations in **predictable** ways in order to reduce the effort required to **co-ordinate** the various articulators.
  - **Palatalization** generally refers to a **conflation** of phonemes closer to the frontal palate than they ‘should’ be.
  - **Final t/d deletion** is simply the **omission** of alveolar plosives from the ends of words.

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Palatalization</th>
<th>Reduced</th>
<th>Phrase</th>
<th>Final t/d Deletion</th>
<th>Reduced</th>
</tr>
</thead>
<tbody>
<tr>
<td>set your</td>
<td>s eh t y ow r</td>
<td>s eh ch er</td>
<td>find him</td>
<td>f ay n d h ih m</td>
<td>f ay n ix m</td>
</tr>
<tr>
<td>not yet</td>
<td>n aa t y eh t</td>
<td>n aa ch eh t</td>
<td>and we</td>
<td>ae n d w iy</td>
<td>eh n w iy</td>
</tr>
<tr>
<td>did you</td>
<td>d ih d y uw</td>
<td>d ih jh y ah</td>
<td>draft the d r ae f t dh iy</td>
<td>d r ae f dh iy</td>
<td></td>
</tr>
</tbody>
</table>
Variation at syllable boundaries

[Image of a spectrogram with the words 'nitrate' and 'night rate']
Phonological variation

- The acoustics of a phoneme depend strongly on the context in which that phoneme occurs.

That must make recognizing phonemes hard, right? How do humans do it?
The inner ear

- Time-variant waves enter the ear, vibrating the **tympanic membrane**.
- This membrane causes tiny bones (incl. **malleus**) 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**.

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

(No need to memorize this either)
Aside – Challenges of perception

- **Cochlear implants** replace the basilar membrane and stimulate the auditory nerve directly.
Next...

- How the Mel scale is used in ASR.
- Automatic speech recognition.