acoustics of speech perception
This lecture

• Introduction to acoustics and speech perception.
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.
Phonemes

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

• English phonemes can be partitioned into **groups**, *e.g.*,:
  - **Stops/plosives**: complete vocal tract constriction and burst of energy (*e.g.*, ‘*papa*’).
  - **Fricatives**: noisy, with air passing through a tight constriction (*e.g.*, ‘*shift*’).
  - **Nasals**: involve air passing through the *nasal* cavity (*e.g.*, ‘*mama*’).
  - **Vowels**: open vocal tract, no nasal air.
  - **Glides/liquids**: similar to vowels, but typically with more constriction (*e.g.*, ‘*wall*’).
### 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)*

*How are these phonemes manifested in sound?*
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 $F = 1/T$

Given $\omega = 2\pi/T$, and phase $\phi = \pi/2$,

$$f(t) = A \sin(\omega t + \phi) = A \cos(\omega t)$$
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**

![Diagram](image.png)

**Lower frequency, higher amplitude**

**Higher frequency, lower amplitude**
Speech waveforms

“Two plus seven is less than ten”
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.
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 phoneme being uttered.
  • ∴ If we could separate the waveform into its component sinusoids, it would help us classify phonemes being uttered.
  • But the shape of the signal changes over time ...
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

Amplitude

Frequency (Hz)
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 $X(F)$ \quad ($\omega = 2\pi F$)

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

- It’s invertible, i.e., $x(t) = \int_{-\infty}^{\infty} X(F) e^{i2\pi Ft} \, dF$.
- 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.
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 signal representation

• **Quantization:** *n.* the conversion of **floating point** amplitude values of each sample to **integers**.

• **PCM:** *n.* (Pulse Code Modulation) A method of quantization in which the analog amplitude is quantized at **uniform intervals**.
  (e.g., 8 bit (−128..127), 16 bit (−32768..32767).
Discrete Fourier transform (DFT)

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

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

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

- **Algorithm(s):** the Fast Fourier Transform (FFT) with complexity $O(N \log N)$.
  - The Cooley-Tukey algorithm divides-and-conquers by breaking the DFT into smaller ones $N = N_1 N_2$. 
Discrete Fourier transform (DFT)

- Below is a 25 ms Hamming-windowed signal from /iy/, 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”
Formants and phonemes

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

- The acoustics of a phoneme depend strongly on the context in which that phoneme occurs.
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** phoneme in which the **glottis** is vibrating (i.e., the phoneme is ‘**voiced**’).
  - Includes some consonants (e.g., /w/, /m/, /g/).

- **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

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

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

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

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

• e.g., Mr. X asks you the name of the baker, whose name is ‘Jones’.

• e.g., Mr. Jones asks you the profession of Mr. X.

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)
• *I* **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).
• How are acoustic waves perceived by the human ear?
  • (this generalizes to other animals).

• This will help us to define the input data parameters that will be used in automatic speech recognition.
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.
Challenges of perception

- **Cochlear implants** replace the basilar membrane and stimulate the auditory nerve directly.
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 formula) 😊
The Mel-scale filter bank

• To mimic the response of the human ear (and because it often improves speech recognition), we often discretize the spectrum using $M$ triangular filters.
  • Uniform spacing before 1 kHz, logarithmic after 1 kHz
Next...

• How the Mel scale is used in ASR.
• Articulatory phonetics.
• Automatic speech recognition.

• Some images from Gray’s Anatomy, Jim Glass’ course 6.345 (MIT), Encyclopaedia Britannica, Pink Floyd’s Dark Side of the Moon(*).

(*) As a matter of fact, it’s all dark.
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.