Course materials

http://www.utdallas.edu/~assmann/hcs6367/readings

- No required text; all readings online
- Background reading:
  - http://www.speechandhearing.net/library/speech_science.php
- Recommended books:

Course requirements

- Class presentations (15%)
- Written reports on class presentations (15%)
- Midterm take-home exam (20%)
- Term paper (50%)

Class presentations and reports

- Pick two broad topics from the field of speech perception. For each topic, pick a suitable (peer-reviewed) paper from the readings web page or from available journals. Your job is to present a brief (10-15 minute) summary of the paper to the class and initiate/lead discussion of the paper, then prepare a written report.

Suggested Topics

- Speech acoustics
- Vowel production and perception
- Consonant production and perception
- Suprasegmentals and prosody
- Speech perception in noise
- Auditory grouping and segregation
- Speech perception and hearing loss
- Cochlear implants and speech coding
- Development of speech perception
- Second language acquisition
- Audiovisual speech perception
- Neural coding of speech
- Models of speech perception
Finding papers
PubMed search engine:

Finding papers
Journal of the Acoustical Society of America:
http://scitation.aip.org/jasa/

Primate vocal tract
The evolution of speech: a comparative review
W. Tecumseh Fitch

Primate vocal tract

Source-filter theory of speech production

Human vocal tract
Acoustics of speech

- Articulation
- Phonation

Organs of speech

- Lungs: apply pressure to generate air stream (power supply)
- Larynx: air forced through the glottis, a small opening between the vocal folds (sound source)
- Vocal tract: pharynx, oral and nasal cavities serve as complex resonators (filter)

Source-Filter Theory

Audio demo: the source signal

- Source signal for an adult male voice
- Source signal for an adult female voice
- Source signal for a 10-year child

Vocal fold oscillation

- One-mass model
  - Air flow through the glottis during the closing phase travels at the same speed because of inertia, producing lowered air pressure above the glottis.

Vocal fold oscillation

- Three-Mass Model
  - One large mass (representing the thyroarytenoid muscle) and two smaller masses, M1 and M2 (representing the vocal fold surface). All three masses are connected by springs and damping constants.
Source-Filter Theory: Vowels

- Linear systems theory
- Assumptions: (1) linearity (2) time-invariance
- Vowels can be decomposed into two primary components: a **source** (input signal) and a **filter** (modulates the input).

**Time domain version:**

\[ U(t) \otimes T(t) \otimes R(t) = P(t) \]

**Frequency domain version:**

\[ U(f) \cdot T(f) \cdot R(f) = P(f) \]

Source properties

- In voiced sounds the glottal source spectrum contains a series of lines called **harmonics**.
- The lowest one is called the **fundamental frequency** \( F_0 \).

![Amplitude Spectrum](image)

Filter properties

- The vocal tract resonances (called **formants**) produce peaks in the spectrum envelope.
- Formants are labelled \( F_1, F_2, F_3, \ldots \) in order of increasing frequency.

![Amplitude Spectrum](image)

Demo: harmonic synthesis

- Additive harmonic synthesis: vowel /i/.
- Cumulative sum of harmonics: vowel /i/.
- Additive synthesis: "wheel".
- Cumulative sum of partials.
Source-filter theory

Laryngeal Source

Filter

Source: J. Hillenbrand

Source-filter theory

Lip buzz serving as source

Filter

Source: J. Hillenbrand

Speech terminology…

- **Fundamental frequency** (F0): lowest frequency component in voiced speech sounds, linked to vocal fold vibration.
- **Formants**: resonances of the vocal tract.

Fundamental frequency (F0): lowest

Formant

Frequency

Amplitude
Source properties: Pitch

- Fundamental frequency ($F_0$) is determined by the rate of vocal fold vibration, and is responsible for the perceived voice pitch.

Source properties: Pitch

- $F_0$ can be removed by filtering (as in telephone circuits) and the pitch remains the same.
- This is the problem of the missing fundamental, one of the oldest problems in hearing science.
- Pitch is determined by the frequency pattern of the harmonics (or their equivalent in the time domain, the periodicities in the waveform).

Harmonicity and Periodicity

- **Period**: regularly repeating pattern in the waveform
  
  Period duration, $T_0 = 6$ ms

Harmonicity and Periodicity

- **Harmonic**: regularly repeating peak in the amplitude spectrum
  
  $F_0 = 1000 / 6 = 166$ Hz

Harmonicity and Periodicity

- **Period**: regularly repeating pattern in the waveform
  
  Period duration $T_0 = 6$ ms

Harmonicity and Periodicity

- **Harmonic**: regularly repeating peak in the amplitude spectrum
  
  $F_0 = 1 / T_0$

Voicing irregularities

- **Shimmer**: variation in amplitude from one cycle to the next.
- **Jitter**: variation in frequency (period duration) from one cycle to the next.
Voicing irregularities

- **Breathy voice** is associated with a glottal waveform with a steeper roll-off than modal voice. As a result there is less energy in the higher harmonics (steeper slope in the spectrum).

Vocal tract properties

- **Resonating tube model**
  - approximation for neutral vowel (schwa), [ə]
  - closed at one end (glottis); open at the other (lips)
  - uniform cross-sectional area
  - curvature is relatively unimportant

Vocal tract model

- **Quarter-wave resonator:**
  \[ F_n = \frac{(2n-1)c}{4L} \]
  - \( F_n \) is the frequency of formant \( n \) in Hz
  - \( c \) is the velocity of sound (about 35000 cm/sec)
  - \( L \) is the length of the vocal tract (17.5 for adult male)

Acoustic vowel space

- First formant, \( F_1 \) frequency (Hz)
- Second formant, \( F_2 \) frequency (Hz)

\[ \begin{align*}
F_1 &= (2(1) - 1)*35000/(4*17.5) = 500 \text{ Hz} \\
F_2 &= (2(2) - 1)*35000/(4*17.5) = 1500 \text{ Hz} \\
F_3 &= (2(3) - 1)*35000/(4*17.5) = 2500 \text{ Hz}
\end{align*} \]
Vocal tract model

- Quarter-wave resonator:
  \[ F_n = \frac{(2n - 1) c}{4 L} \]
  - \( F_n \) is the frequency of formant \( n \) in Hz
  - \( c \) is the velocity of sound in air (about 35000 cm/sec)
  - \( L \) is the length of the vocal tract (17.5 for adult male)

Helium speech

- The speed of sound in a helium/oxygen mixture at 20°C is about 93000 cm/s, compared to 35000 cm/s in air. This increases the resonance frequencies but has relatively little effect on \( F_0 \). In helium speech, the formants are shifted up but the pitch stays the same.

**Using Matlab as a calculator, find the frequencies of F1, F2 and F3 for a 17.5 cm vocal tract producing the vowel /a/ in a helium/air mixture (velocity \( c \approx 93000 \) cm/s)**

\[ F_n = \frac{(2n - 1) c}{4 L} \]

- \( F_1 = (2(1) - 1) \times 93000 \div (4 \times 17.5) \approx 500 \) Hz
- \( F_2 = (2(2) - 1) \times 93000 \div (4 \times 17.5) \approx 1500 \) Hz
- \( F_3 = (2(3) - 1) \times 93000 \div (4 \times 17.5) \approx 2500 \) Hz

Audio demos

- Speech in air
- Speech in helium
- Pitch in air
- Pitch in helium

http://phys.unsw.edu.au/phys_sounds/PHYSICS/SPEECH_HELIUM/speech.html
Speech in helium

Perturbation Theory
- The first formant (F1) frequency is lowered by a constriction in the front half of the vocal tract (/u/ and /i/), and raised when the constriction is in the back of the vocal tract, as in /a/.

Perturbation Theory
- The second formant (F2) is lowered by a constriction near the lips or just above the pharynx; in /u/ both of these regions are constricted. F2 is raised when the constriction is behind the lips and teeth, as in the vowel /ɪ/.

Perturbation Theory
- The third formant (F3) is lowered by a constriction at the lips or at the back of the mouth or in the upper pharynx. This occurs in /r/ and /r/-colored vowels like American English /ə/.

Perturbation Theory
- F3 is raised when the constriction is behind the lips and teeth or near the upper pharynx.

Perturbation Theory
- All formants tend to drop in frequency when the vocal tract length is increased or when a constriction is formed at the lips.
Perturbation Theory

- F1 frequency is correlated with jaw opening (and inversely related to tongue height).

![Amplitude spectrum](image)

F2 frequency is correlated with tongue advancement (front-back dimension).

![Amplitude spectrum](image)

Spectral analysis

- **Amplitude** spectrum: sound pressure levels associated with different frequency components of a signal
  - Power or intensity
  - Amplitude or magnitude
  - Log units and decibels (dB)
- **Phase** spectrum: relative phases associated with different frequency components
  - Degrees or radians

Spectral analysis of speech

- **Why perform a frequency analyses of speech?**
  - Ear+brain carry out a form of frequency analysis
  - Relevant features of speech are more readily visible in the amplitude spectrum than in the raw waveform

**But:** the ear is not a spectrum analyzer.

- Auditory frequency selectivity is best at low frequencies and gets progressively worse at higher frequencies.
Measuring formants

Formant frequency peak estimation requires an interpolation process.

Formant Estimation

Vowel spectra have peaks corresponding to the center frequencies of formants.

Formant Estimation

But: harmonics also generate spectral peaks; formant frequencies do not necessarily coincide with harmonic frequencies.

Children’s speech

- Children’s voices have high F0s.
- When F0 is 400 Hz (not unusual for 3-year olds), only 4 harmonics appear in the frequency range between 0-1600 Hz.

Sparce sampling problem

- Vowel identity is dependent on the frequencies of formant peaks.
- Formants are difficult to estimate when fundamental frequency is high.
Formants sometimes appear to merge

Short-term amplitude spectrum

Speech spectrograms
- What is a speech spectrogram?
  - Display of amplitude spectrum at successive instants in time ("running spectra")
  - How can 3 dimensions be represented on a two-dimensional display?
    - Gray-scale spectrogram
    - Waterfall plots
    - Animation

Speech spectrograms
- Why are speech spectrograms useful?
  - Shows dynamic properties of speech
  - Incorporates frequency analysis
  - Related to speech production
  - Helps to visually identify speech cues
Waveform spectrogram

Digital representations of signals

TrackDraw: a graphical speech synthesizer
American English vowel space

<table>
<thead>
<tr>
<th>Height</th>
<th>Front</th>
<th>Center</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>i &quot;head&quot;</td>
<td>u &quot;who'd&quot;</td>
<td>a &quot;hut&quot;</td>
</tr>
<tr>
<td>e &quot;hayed&quot;</td>
<td>e &quot;hood&quot;</td>
<td>o &quot;hated&quot;</td>
<td></td>
</tr>
<tr>
<td>mid</td>
<td>e &quot;head&quot;</td>
<td>o &quot;hated&quot;</td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>a &quot;bad&quot;</td>
<td>o &quot;hut&quot;</td>
<td></td>
</tr>
</tbody>
</table>

Peterson and Barney (1952)
- Acoustic measurements (made from spectrograms) of formant frequencies (F1, F2, F3) in vowels spoken by 76 men, women and children.
- Vowel space: projection of a given talker's vowels in a F1 x F2 plane.
- Simple target model: vowels are differentiated (perceptually) by F1 and F2 frequencies measured in the middle of the vowel (vowel target).

Peterson and Barney (1952)

Vowel formant space: F1 x F2
Assmann & Katz JASA 2000