AUD 6306
Speech Science

Dr. Peter Assmann
Spring semester 2019

Course web page
http://www.utdallas.edu/~assmann/aud6306/
• Course information
• Speech demos
• Lecture slides
• Assigned reading material
• Additional resources in speech & hearing

Course materials
• No required text
• Some recommended books:

Course requirements
• Class presentations (20%)
• Presentation reports and homework (20%)
• Midterm take-home exam (25%)
• Final take-home exam (35%)

Course theme:
Hearing and speech are closely linked
Class presentations and reports

- Sign up for two presentation dates.
- Pick two broad topics from the field of speech science (topics must be approved).
- Select a suitable (peer-reviewed) paper for each topic and present a brief (10-15 minute) summary of the paper to the class to initiate/lead discussion.
- Prepare a written report (2-3 pages) due about 2-3 weeks after the presentation.

Class presentations and reports

- When you find a paper you’d like to present, email the citation or the PDF version of the paper to me for approval. (In some cases I may suggest an alternative, or more recent paper). I will post the paper on the readings web page and email the link to the class.

Class presentations and reports

- Important note: papers must be selected and approved one week before the presentation to provide others time to read them.
- Everyone is expected to read the assigned articles for each class and be prepared to discuss them.

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/content/asa/journal/jasa

UTD library - online journals
http://www.utdallas.edu/library/resources/journals.html
https://libguides.utdallas.edu/journal-collections

Speech Science

Speech production
- Respiration
- Phonation
- Articulation

Channel properties
- Noise and reverberation
- Transmission properties
- Hearing loss

Speech perception
- Acoustic cues
- Phonetic units
- Lexical access
- Invariance
- Perceptual constancy
- Segmentation
- Neural coding of speech

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

Vocal motor control
- Fitch 2018

Primate vocal tract
Source-filter theory of speech production

Human vocal tract

Acoustics of speech

Organs of speech

Source-Filter Theory

Audio demo: the source signal


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.

  [Source: http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/model.html]

- **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: http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/model.html]

**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** \( F_0 \).

**Demo: harmonic synthesis**

- Additive harmonic synthesis: vowel /i/ 🎤
- Cumulative sum of harmonics: vowel /i/ 🎤
- Additive synthesis: “wheel” 🎤
- Cumulative sum of partials: 🎤
Filter properties

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

Source-filter theory

- Source
- Filter
- Radiation = Output sound

Source: J. Hillenbrand

Source-filter theory

Source: J. Hillenbrand

Source-filter theory

Source: J. Hillenbrand

Source-filter theory

Source: J. Hillenbrand

Source-filter theory

Source: J. Hillenbrand

Source-filter theory

Source: J. Hillenbrand
Source-filter theory

Source: J. Hillenbrand

Speech terminology...

- **Fundamental frequency** ($F_0$): lowest frequency component in voiced speech sounds, linked to vocal fold vibration.
- **Formants**: resonances of the vocal tract.

Harmonicity and Periodicity

- **Period**: regularly repeating pattern in the waveform

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).

Harmonic and Periodicity

- **Harmonic**: regularly repeating peak in the amplitude spectrum
Harmonicity and Periodicity

- **Period**: regularly repeating pattern in the waveform

<table>
<thead>
<tr>
<th>Frequency (kHz)</th>
<th>Amplitude (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-40</td>
</tr>
<tr>
<td>0.5</td>
<td>-20</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1.5</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
</tr>
</tbody>
</table>

Harmonics are integer multiples of $F_0$ and are evenly spaced in frequency.

$F_0 = \frac{1000}{6} = 166$ Hz

$F_n = \frac{1}{T_0}$

Waveform

### Harmonic singing

- **Harmonic singing** (also called overtone singing) involves changing the shape of the vocal tract to align the resonance frequencies (formants) with harmonics of the fundamental. A low, sustained fundamental is produced, similar to the drone of a bagpipe, along with flute-like harmonics that drift in and out.

### Effects of F0 changes

- **Source-filter independence**

### Effects of formant frequency changes

- **Source-filter independence**

Harmonic singing

<table>
<thead>
<tr>
<th>Frequency (kHz)</th>
<th>Amplitude (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-40</td>
</tr>
<tr>
<td>0.5</td>
<td>-20</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1.5</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
</tr>
</tbody>
</table>

Tuvan throat singing

http://www.youtube.com/watch?v=DY1pcEtHI_w&feature=youtu.be

Amazing Grace

http://www.youtube.com/watch?v=qO4UH-M9h40&feature=youtu.be

Overtone singing

https://www.youtube.com/watch?v=UHTF1-9vyO#t=21
Voicing irregularities

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

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

Uniform tube model (schwa)

![Diagram of vocal tract model](image)

- **Quarter-wave resonator**:
  \[ F_n = \left( \frac{2n - 1}{4} \right) \frac{c}{L} \]
  - \( F_n \) 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)

Vocal tract model

- **Quarter-wave resonator**:
  \[ F_n = \left( \frac{2n - 1}{4} \right) \frac{c}{L} \]
  - \( F_1 = \left( \frac{2(1) - 1}{4} \right) \frac{35000}{4 \times 17.5} = 500 \text{ Hz} \)
  - \( F_2 = \left( \frac{2(2) - 1}{4} \right) \frac{35000}{4 \times 17.5} = 1500 \text{ Hz} \)
  - \( F_3 = \left( \frac{2(3) - 1}{4} \right) \frac{35000}{4 \times 17.5} = 2500 \text{ Hz} \)
Acoustic vowel space

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 in air (about 35000 cm/sec)
  - \( L \) is the length of the vocal tract (17.5 for adult male)

Vocal tract model
- Quarter-wave resonator:
  \[ F_n = \frac{(2n - 1)c}{4L} \]
  - \( F_1 = (2(1) - 1) \times 35000/(4 \times 17.5) = 500 \) Hz
  - \( F_2 = (2(2) - 1) \times 35000/(4 \times 17.5) = 1500 \) Hz
  - \( F_3 = (2(3) - 1) \times 35000/(4 \times 17.5) = 2500 \) Hz

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.

Helium speech
- Exercise: Compute the frequencies of \( F_1 \), \( F_2 \) and \( F_3 \) for a 17.5 cm vocal tract producing the vowel /a/ (schwa) in a helium/air mixture (velocity \( c \approx 93000 \) cm/s)
  \[ F_n = \frac{(2n - 1)c}{4L} \]
  - \( F_1 = (2(1) - 1) \times 93000/(4 \times 17.5) = \)
  - \( F_2 = (2(2) - 1) \times 93000/(4 \times 17.5) = \)
  - \( F_3 = (2(3) - 1) \times 93000/(4 \times 17.5) = \)
**Helium speech**

- Exercise: Compute the frequencies of F1, F2, and F3 for a 17.5 cm vocal tract producing the vowel /ә/ (schwa) in a helium/air mixture (velocity $c \approx 93000$ cm/s)

$$F_n = \left(2n - 1\right) \frac{c}{4L}$$

- $F_1 = (2(1) - 1) * 93000 / (4*17.5) = 1328.6$
- $F_2 = (2(2) - 1) * 93000 / (4*17.5) = 3985.7$
- $F_3 = (2(3) - 1) * 93000 / (4*17.5) = 6642.9$

**Audio demos**
- Speech in air
- Speech in helium
- Pitch in air
- Pitch in helium

http://phys.unsw.edu.au/phys_about/PHYSICS/SPEECH_HELIUM/speech.html

**Sulfur Hexafluoride**

- **Helium**
  - density of 0.1786 g/L at sea level
- **Air**
  - density of 1.225 g/L at sea level
- **Sulfur Hexafluoride (SF₆)**
  - density of 6.12 g/L at sea level

Speech production with vocal tract filled with SF₆

http://www.youtube.com/watch?v=d-XbjFp3uqE

**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 /u/.

http://www.youtube.com/watch?v=d-XbjFp3uqE
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 /i/.

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 /ɚ/ (as in "heard").

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).

Perturbation Theory

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

Spectral analysis of speech

- **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.

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

Formants

F1
F2
F3

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

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.

LPC spectrum

Formants sometimes appear to merge
Speech spectrogram
- running amplitude spectra (codes amplitude changes in different frequency bands over time).

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

American English vowel space
- Advancement
- Height

F1 = 281 Hz
F2 = 2196 Hz
F3 = 2755 Hz
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)

American English vowel space

Peterson and Barney (1952)

Invariance problem

- Dynamic cues in vowel perception
- Talker normalization theories
  - Potter and Steinberg (1950): invariant pattern of stimulation shifted up or down along the basilar membrane
  - Miller (1989): Formant ratio theory
  - Joos (1948): Frame of reference theory
  - Nearey (1989): Extrinsic and intrinsic factors

Formant Dynamics

- Formant frequency changes over time:
Graphical interpretation of CLIH (sliding template) model

Movement along diagonal for different speakers

Fixed pattern of ‘holes’ in the template correspond to stored vowel reference pattern

Neary & Assmann, 2006

F₀ as a function of age and sex

F₀ distribution – child talkers

F₀ distribution – males (blue)
$F_0$ distribution – females (red)

Wavesurfer
- Download Wavesurfer: www.speech.kth.se/wavesurfer