AUD 6306
Speech Science

Dr. Peter Assmann
Spring semester 2017

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; all readings online
  • 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 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
- Audiospatial 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 online journals

http://www.utdallas.edu/library/resources/journals.html

- Click on link A – Z List of eJournals or search directly by journal name
  
  Journal of the Acoustical Society of America

Speech Science

![Speech Science Diagram]

Primate vocal tract

![Primate vocal tract Diagram]

Source-filter theory of speech production

![Source-filter Theory Diagram]

Human vocal tract

![Human Vocal Tract Diagram]
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 $F_1$, $F_2$, $F_3$, ... 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 $\otimes$ filter $\otimes$ radiation = output sound

![Glottal source](image) 🖋️ 🛠️ 🎵 🖋️
**Speech terminology...**

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

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

**Harmonicity and Periodicity**

- **Harmonic**: regularly repeating peak in the amplitude spectrum

  $F_0 = 1000 / 6 = 166$ Hz
  $F_0 = 1 / T_0$

**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.
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.

Harmonic singing

- **Tuvan throat singing**
- **Amazing Grace**
- **Overtone singing**

Effects of F0 changes

- **Source-filter independence**
- **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 diagram](image)

Uniform tube model (schwa)

- 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 diagram](image)

Vocal tract model

- 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 diagram](image)

Acoustic vowel space

![Acoustic vowel space diagram](image)

Vocal tract model

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

![Vocal tract model diagram](image)
Vocal tract model

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

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.

- **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 \( c \approx 93000 \text{ cm/s} \))

  \[ F_n = \left( 2n - 1 \right) \frac{c}{4L} \]
  - \( F_1 = (2(1) - 1) \times 93000/(4 \times 17.5) = \)
  - \( F_2 = (2(2) - 1) \times 93000/(4 \times 17.5) = \)
  - \( F_3 = (2(3) - 1) \times 93000/(4 \times 17.5) = \)

Helium speech

- **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-XbjFn3aQgE](http://www.youtube.com/watch?v=d-XbjFn3aQgE)

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/.

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).

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

**Digital representations of signals**

**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.

Formant Estimation

Vowel spectra have peaks corresponding to the center frequencies of formants

Spectrum of natural vowel /a/

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.

Representations of speech signals

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 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

American English vowel space

“The watchdog” waveform

“The watchdog” waveform
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

<table>
<thead>
<tr>
<th>American</th>
<th>English vowel space</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td></td>
</tr>
<tr>
<td>front</td>
<td>“heed”</td>
</tr>
<tr>
<td>high</td>
<td>“hid”</td>
</tr>
<tr>
<td>mid</td>
<td>“hayed”</td>
</tr>
<tr>
<td>low</td>
<td>“had”</td>
</tr>
<tr>
<td></td>
<td>“schwa”</td>
</tr>
<tr>
<td></td>
<td>“hut”</td>
</tr>
<tr>
<td></td>
<td>“hod”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Advancement</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dynamic cues in vowel perception

- Formant frequency changes over time:
  - Miller (1989): Formant ratio theory
  - Nearey (1989): Extrinsic and intrinsic factors

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:

Vowel-inherent spectral change

Dual-target model

“noisy” formant tracks
FFs as a function of age and sex

Vowel formant space: F1 x F2

Boys
Girls

Vowel formant space: F1 x F2

Males
Females

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

Nearey & Assmann, 2006

Vowel formant space: F1 x F2

Boys
Girls

Medians of 75 vowels per talker

Boys
Girls

Graphical interpretation of CLIH (sliding template) model

Men
Women
Children

Nearey & Assmann, 2006
F₀ as a function of age and sex

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F₀ distribution – child talkers

F₀ distribution – males (blue)

F₀ distribution – females (red)
**Wavesurfer**

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

- Wavesurfer User Manual  
  www.speech.kth.se/wavesurfer/man.html