Course web page

- http://www.utdallas.edu/~assmann/hcs6367
  - Course information
  - Lecture notes
  - Speech demos
  - Assigned readings
  - Additional resources in speech & hearing

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.

Acoustic properties of speech

- Time
- Frequency
- Amplitude

Spectral analysis

- Amplitude spectrum: sound pressure levels associated with different frequency components of a signal
  - Power / intensity
  - Amplitude / magnitude
  - Log units and decibels (dB)

![Waveform](image1.png)  ![Spectrum](image2.png)
Spectral analysis

• 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

Spectral analysis of speech

• Waveforms

Spectral analysis of speech

• Spectrogram (time x frequency x amplitude)

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.
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

Speech spectrogram

- *running amplitude spectra* (codes amplitude changes in different frequency bands over time).
Vowel perception

- What is a vowel?
  - Vowels are sounds produced with a relatively open vocal tract configuration.
- Vowel production
  - Sound source at or near the glottis
  - Supralaryngeal vocal tract filter
  - Pharyngeal / oral cavities relatively unconstricted

Vowel perception

- Oral vowels
  - Velopharyngeal port (velum) is closed
  - Vocal tract transfer function normally appears as an all-pole filter (Fant, 1960).
  - The poles are the natural resonant frequencies of the vocal tract, also called formants.
  - The formant frequencies are determined by the shape of the airway between the glottis and lips
    - Area function = cross sectional area vs. distance
Vocal tract model

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

Resonance

- A quarter-wave resonator produces standing wave at each of its resonance frequencies, which occur at odd integer multiples (1, 3, 5, ...) of the lowest resonance frequency.
- Distribution of volume velocity (inverse: pressure) in the tube reflects particle vibrations at different locations
- Nodes (maxima) and antinodes (minima)

Perturbation Theory

- Relationship between location of vocal tract constrictions and formant frequencies.

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)
Vowel perception

- Nasal vowels
  - Velopharyngeal port opens to the nasal cavity
  - Nasal coupling introduces additional poles and zeros into the vocal tract transfer function
  - These corresponding to nasal formants and antiformants and appear as additional peaks and valleys in the spectrum of a nasal vowel.

Oral vs. nasal vowels

Oral vowel /ɑ/ vs. Nasal vowel /ɑ̃/

Vowel perception

- Vowels are traditionally described in terms of phonetic features:
  - Height (openness)
  - Advancement
  - Rounding
  - Length

American English vowel space

- Articulatory dimensions of tongue height (or jaw opening) and tongue advancement (front-back position) are closely linked to the center frequencies of the formants.
- Height is inversely related to F1 (higher vowels – lower F1)
- Advancement is related to F2 (front tongue position – higher F2)
**Peterson and Barney (1952)**

- Acoustic measurements (made from broadband spectrograms) of formant center frequencies (F1, F2, and F3) in vowels spoken by 76 men, women and children.
- **Vowel space**: projection of a given talker's vowels in an F1 x F2 plane
- **Simple target model**: vowels are differentiated (perceptually) by F1 and F2 frequencies measured near the middle of the vowel (vowel target).
- F3 helps to separate the vowel /ɚ/ (as in American English "bird") from all other vowels.

**Vowel perception**

- Broadband spectrograms of speech show dark bands (indicating spectral regions with high intensity) near formant peaks. Narrowband spectrograms tend to highlight individual harmonics.

**American English vowel space**

- Close match between formant frequencies and traditional impressionistic phonetic descriptions of tongue height and advancement (Joos, 1948).
- Formant plots show a rotated/flipped version of the traditional (impressionistic) vowel space defined by height and advancement.

**Formant specification of vowels**

- Formants (resonances) are defined by their center frequencies, amplitudes and bandwidths.
Formant specification of vowels

- Fant (1956) showed that the entire spectrum of a vowel could be uniquely described by a low-dimensional description incorporating only the formant frequencies and bandwidths.

Formant synthesizers

- Fant’s source-filter model is incorporated into models of formant synthesis (Klatt 1980, 1988).

Pattern Playback

- One of the earliest speech synthesizer was the Pattern Playback, developed at Haskins Laboratories by Franklin Cooper.
- Pattern Playback uses a optical representation of speech, in the form of a sound spectrogram.

Pattern Playback

- A tone wheel generates 50 harmonics, multiples of a fixed frequency of 120 Hz (120, 240, 360, ... 6000).
- For this reason, Pattern Playback could only produce monotone speech in the adult male register.
Speech synthesis experiments

- Delattre et al. (1952) found that two formants were sufficient to synthesize vowel-like sounds matching natural vowels in languages like English and French.
- In some cases (e.g., back vowels, as in “hod”) vowels could be approximated with a single formant.
- Three formants appeared to be necessary to approximate the vowel /ə/.

Vowel perception

- Formant frequencies are believed to provide the most important cues for vowel identification.
- Formants have traditionally been defined as vocal-tract resonances. However, a more general way of describing the role of formants in vowel perception is in terms of spectral peaks and shoulders that are associated with vocal tract resonances.

Formant Estimation

Vowel spectra have peaks corresponding to the center frequencies of formants

Spectral shape models

- Principal components analysis (PCA; Pols et al., 1969)
- mel cepstra (Davis and Mermelstein, 1980)
- Perceptual linear prediction (PLP; Hermansky, 1990)
- Discrete cosine transform (DCT; Zahorian and Jagharghi, 1993)
Perceptual role of formants

- Formants can be hard to extract and measure (especially in children’s voices)
- Vowel classification using spectral shape features sometimes produce higher classification accuracy (but they may be relying on similar information)
- When formant peaks are removed from the spectrum, listeners still hear the same vowel (Ito et al., 2001)

Perceptual role of formants

- Re-examining the findings by Ito et al. (2001), Kiefte and Kluender (2005) note that higher formant amplitudes and spectral slope were preserved when formant peaks were removed. These features might have provided the information needed for vowel identification.

Perceptual role of formants

- However, when Kiefte and Kluender replicated the Ito et al. experiment using more realistic stimuli (diphthongs, with time-varying formants) they found that listeners no longer relied on amplitude and spectral slope cues.

Perceptual role of formants

- They concluded that listeners will use shape cues when formant peaks are unavailable, but do so only for simple steady vowels (monophthongs) which do not resemble natural vowels.

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

Dual-target model
Dual-target model