Today's class is about:

First, Weiss speech recognition is difficult. As you'll see, the impression we have speech is like beads on a string just wrong.

Second we will look at how hidden Markov models are used to do speech recognition.

And finally, we will look at how the speech dialogue technology behind systems like Siri might be configured.

This was dictated one November 11, 2016, into the email app on my iPhone.

I. Why is Speech Recognition Hard??

A trained person can “read” a spectrogram

Therefore, the spectrogram contains all the information a machine needs as well....

Prof. Victor Zue, MIT

Vowels are determined by their formants

The frequencies of F₁, F₂, and F₃ – the first three resonances of the vocal tract – largely determine the perceived vowel

Represents the varying short term amplitude spectra of the speech waveform

Darkness represents amplitude at that time & frequency.
Consonants are determined by *(inter alia)*:

- **Formant motion**
- **Length of Silence** (“Voice Onset Time”)

---

Coarticulation

- The same abstract phoneme can be realized very differently in different phonetic contexts: *coarticulation*
- \( F_2 \) in the vowel /\( u/\), crucial to its identification, varies significantly due to surrounding consonants in the syllables:

<table>
<thead>
<tr>
<th>Context</th>
<th>( F_2 ) (kHz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“kook”</td>
<td>1.0</td>
</tr>
<tr>
<td>“moom”</td>
<td>0.8-1.0</td>
</tr>
<tr>
<td>“toot”</td>
<td>1.2</td>
</tr>
</tbody>
</table>

---

Speech Information is not local

- The identity of speech units, *phones*, cannot be determined independently of context.
- Sometimes two phones can best be distinguished by examining properties of neighboring phones:

---

Speech Information is not local

- /\( s/\) and /\( z/\) are often acoustically identical...
- They are differentiated by the length of the preceding vowel:

---

Words are constant, but utterances aren’t

- Spectrograms of similar words pronounced by the *same* speaker may be more alike than
  - Spectrograms of the *same* word pronounced by different speakers:

    - “wait” – MM (m)
    - “wait” – JH (f)
    - “wait” – whispered(MM)

---

II. HMMs for Speech Recognition

(Illustrations in II from Chapter 9, Jurafsky & Martin)
Speech Recognition Architecture

Speech recognition via Bayes Rule!

\[ \hat{W} = \arg \max_{W \in \Omega} P(\text{Signal} \mid W) P(W) \]

Where: \( W \) is a (text) string from a Source

Signal is the (speech) output from a "noisy channel"

Speech Architecture meets Noisy Channel

Schematic HMM for the word *six*

- Simple one state per phone model
- Left to right topology with self loops and no skips
- Start and End states with no emissions
- States output 10 msec spectral slices or DNN vectors
Phones have dynamic structure

- Wait (said by Mitch Marcus), pronounced [w ey t]
- The formants of the dipthong ey move continually
- T consists of (a) a silence, (b) a burst

A 3-state HMM phone model

- Three emitting states
- Two non-emitting states
- Usually includes skip states

The word six [siks] using 3-state HMM phone models

A simple full HMM for digit recognition

III. Speech Dialogue Understanding

Multiple knowledge sources provide redundancy

- Grammatical, semantic and pragmatic information can be used to make recognition robust.
- A first experiment: AT&T Bell Labs airline reservation system (Levinson-1977)

Multiple knowledge sources provide redundancy

<table>
<thead>
<tr>
<th>Processing level</th>
<th>Sentences correct</th>
<th>Errors detected</th>
<th>Word Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustic</td>
<td>Na</td>
<td>0</td>
<td>88%</td>
</tr>
<tr>
<td>Syntactic</td>
<td>330</td>
<td>0</td>
<td>99%</td>
</tr>
<tr>
<td>Pragmatic</td>
<td>345</td>
<td>6</td>
<td>&gt;99%</td>
</tr>
</tbody>
</table>
Speech Recognition: Task Dimensions

- Speaker Dependent, Independent, Adaptive
  - Speaker dependent: System trained for current speaker
  - Speaker independent: No modification per speaker
  - Speaker Adaptive: adapts an initial model to speaker
- Read vs. dictation vs. conversational
- Quiet Conditions vs. various noise conditions
- Known microphone vs. unknown microphone
- Perplexity level
  - Low perplexity: Average expected branching factor of grammar < 10-20
  - High perplexity: Average expected branching factor of grammar > 100

Perplexity (average branching factor of LM): Why it matters

- Experiment (1992): read speech, Three tasks
  - Mammography transcription (perplexity 60)
    "There are scattered calcifications with the right breast"
    "These too have increased very slightly"
  - General radiology (perplexity 140)
    "This is somewhat diffuse in nature"
    "There is no evidence of esophageal or gastric perforation"
  - Encyclopedia dictation (perplexity 430)
    "Czechoslovakia is known internationally in music and film"
    "Many large sulphur deposits are found at or near the earth’s surface"

<table>
<thead>
<tr>
<th>Task</th>
<th>Vocabulary</th>
<th>Perplexity</th>
<th>Word error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammography</td>
<td>837</td>
<td>66</td>
<td>3.4%</td>
</tr>
<tr>
<td>Radiology</td>
<td>4447</td>
<td>141</td>
<td>5.8%</td>
</tr>
<tr>
<td>Encyclopedia</td>
<td>3021</td>
<td>433</td>
<td>14.6%</td>
</tr>
</tbody>
</table>