Speech Recognition

Mitch Marcus
CIS 421/521
Artificial Intelligence
A Sample of Speech Recognition

Today's class is about:

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

Second we will look at how Hedin Mark off 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 on November 11, 2017, into the email app on my iPhone.
I. Why is Speech Recognition Hard??
A Speech Spectrogram

- Represents the varying short term amplitude spectra of the speech waveform
- Darkness represents amplitude at that time & frequency.
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 $F_1$, $F_2$, and $F_3$ – the first three resonances of the vocal tract – largely determine the perceived vowel
Consonants are determined by \textit{(inter alia)}:

- Formant motion

\textbf{Length of Silence (“Voice Onset Time”)}

\textbf{short VOT}

\textbf{long VOT}

http://www.frontiersin.org/files/Articles/76444/fpsyg-05-00549-HTML/image_m/fpsyg-05-00549-g001.jpg
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 from Chapter 9, Jurafsky & Martin)
Speech Recognition Architecture

Signal Processing

Deep Neural Network - 2012

Everything we’ve learned

HMM Lexicon

Viterbi Decoder

N-gram Grammar

if music be
Speech recognition via Bayes Rule!

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

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

\textit{Signal} is the (speech) output from a

“noisy channel”
The noisy channel model: another view of HMMs

- Ignoring the denominator leaves us with two factors: $P(W)$ and $P(\text{Signal}|W)$
Speech Architecture meets Noisy Channel

Outputs of HMM:
Vectors encoding 10 msec of sound

States of HMM:
phonemes

P(OIW)

Acoustic Model + Lexicon

Decoding Search

Language Model

P(W)

Feature Extraction

O

W

CIS 421/521 - Intro to AI
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

Lexicon

one  w ah n
two  t uw
three thr iy
four f ao r
five f ay v
six  s ih k s
seven s eh v ax n
eight ey t
nine  n ay n
zero  z iy r ow
oh   ow

Phone HMM

Start

End

p(“one”) p(“two”) p(“zero”) p(“oh”)
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>
Structure of Phoenix
A Spoken Language Understanding System

speech: "show me...ah...I want to see all the flights to Denver after two pm"

digitize: 16 KHz, 16 bit samples

VQ codes: A vector of 3 bytes, each 10 ms

words:
"show me I want to see all flights to Denver after two pm"

frame:
[list]: I want to see
[flights]: all flights
[arrive_loc]: to Denver
[depart_time_range]: after two pm

canonical frame:
[list]: flights
[arrive_loc]: "DEN"
[depart_loc]: "PIT"
[depart_time_range]: 1400 2400

ATIS Application

SQL:
select airline_code, flight_number
from flight_table
where (from_airport = 'PIT' and to_airport = 'DEN')
and (departure_time > 1400)
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>
Progress in Automatic Speech Recognition

http://recognize-speech.com/acoustic-model/knn/benchmarks-comparison-of-different-architectures