Speech recognition (briefly)

Chapter 15, Section 6

Outline

- Speech as probabilistic inference
- Speech sounds
- Word pronunciation
- Word sequences

Speech as probabilistic inference

It's not easy to wreck a nice beach

Speech signals are noisy, variable, ambiguous

What is the most likely word sequence, given the speech signal?

I.e., choose words to maximize \( P(\text{words} | \text{signal}) \)

Words are the hidden state sequence, signal is the observation sequence

Phone models

Phone sounds

Frame features are typically formants—peaks in the power spectrum

Frame features in \( P(\text{features} | \text{phone}) \) summarized by:

- an integer in \([0 \ldots 255]\) (using vector quantization); or
- the parameters of a mixture of Gaussians

Three-state phones: each phone has three phases (Onset, Mid, End)

Phone features in \( P(\text{features}) \), summarized by:

Frames with features:

10 15 38
52 47 82
22 63 24
89 94 11
10 12 73

Frame sounds

Phone models

APR Abet designed for American English

<table>
<thead>
<tr>
<th>Artic.</th>
<th>[b]</th>
<th>[d]</th>
<th>[g]</th>
</tr>
</thead>
<tbody>
<tr>
<td>dental</td>
<td>[p]</td>
<td>[t]</td>
<td>[k]</td>
</tr>
<tr>
<td>alveolar</td>
<td>[f]</td>
<td>[θ]</td>
<td>[s]</td>
</tr>
<tr>
<td>labial</td>
<td>[v]</td>
<td>[z]</td>
<td>[h]</td>
</tr>
</tbody>
</table>

Word pronunciation

Word sounds

Speech as probabilistic inference

 Speech Recognition (Briefly)
Isolated word recognition systems with training reach 95–99% accuracy.

Isolated words + word models fix likelihood $P(e_1:t|\text{word})$ for isolated word

$P(\text{word}|e_1:t) = \alpha P(e_1:t|\text{word}) P(\text{word})$

Prior probability $P(\text{word})$ obtained simply by counting word frequencies

$P(e_1:t|\text{word})$ is computed using the HMM for each word.

Continuous speech systems manage 60–80% accuracy on a good day.

Continuous speech recognition works by finding the most likely sequence of words.

- Cross-word confusion—"it’s not that love word combination"—is not
- Probabilistic: there are few rules in speech
- Exceptional: many low frequency words
- Adjacent words highly correlated
- Most likely sequence of words

Not just a sequence of isolated-word recognition problems.

Language model

Prior probability of a word sequence is given by chain rule:

$$P(w_1 \cdot \cdot \cdot w_n) = \prod_{i=1}^{n} P(w_i|w_1 \cdot \cdot \cdot w_{i-1})$$

Bigram model:

$$P(e_1:t|\text{word}) \approx P(w_i|w_{i-1})$$

Train by counting all word pairs in a large text corpus

$P(w_i|w_{i-1})$ is computed using the HMM for each word.

More sophisticated models (trigrams, n-grams, etc.) help a little bit

Structure is created manually, transition probabilities learned from data

$P(\text{word}|e_1:t)$ for isolated word

Each word is described as a distribution over phone sequences

Phone models + word models fix likelihood $P(e_1:t|\text{word})$ for isolated word

Phone model example

Continuous speech

Word pronunciation models
Combined HMM

States of the combined language+word+phone model are labeled by the word we're in + the phone in that word + the phone state in that phone.

Viterbi algorithm finds the most likely phone state sequence.

Does segmentation by considering all possible word sequences and boundaries.

Doesn't always give the most likely word sequence because each word sequence is the sum over many state sequences.

Jelinek invented a way to find the most likely word sequence using

\[
A^* \quad \text{where} \quad a = -\log P(w_i | w_{i-1})
\]

Since the mid-1970s, speech recognition has been formulated as a probabilistic

Summary

Since the mid-1970s, speech recognition has been formulated as probabilistic

Evidence = speech signal; hidden variables = word and phone sequences.

"Context" effects (coarticulation, etc.) are handled by augmenting states.

Variability in human speech (speed, timbre, etc., etc.) and background noise make continuous speech recognition in real settings an open problem.

"Context" effects (coarticulation, etc.) are handled by augmenting states.

Evidence = speech signal; hidden variables = word and phone sequences.

Inference

Since the mid-1970s, speech recognition has been formulated as probabilistic

Summary