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

\[ P(\text{words} | \text{signal}) \]

Use Bayes' rule:

\[ P(\text{words} | \text{signal}) = P(\text{signal} | \text{words}) P(\text{words}) \]

I.e., decomposes into acoustic model + language model.

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

Phone models

Frame features are typically formants—peaks in the power spectrum.

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

- An integer in \( [0:255] \) (using vector quantization); or
- The parameters of a mixture of Gaussians.

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

E.g., \([t]\) has silent Onset, explosive Mid, hissing End.

Triphone context: each phone becomes \( n^2 \) distinct phones, depending on the phones to its left and right.

E.g., in \( \text{star} \), \([t]\) is written \([t(s,aa)]\) (different from \( \text{tar} \)!)

Triphones useful for handling coarticulation effects: the articulators have inertia and cannot switch instantaneously between positions.

E.g., \([t]\) in \( \text{eight} \) has tongue against front teeth.

Phone sounds

Raw signal is the microphone displacement as a function of time.

Proceeds into overlapping 30ms frames, each described by features.

ARPA-Be designed for American English.

All human speech is composed from 40-50 phones, determined by the configuration of articulators (lips, tongue, vocal cords, air flow).

Form an intermediate level of hidden states between words and signal.

Acoustic model = pronunciation model + phone model.
Phone model example

Phone HMM for \[m\]:

\[
\begin{align*}
\text{C1:} & \quad 0.5 \\
\text{C2:} & \quad 0.2 \\
\text{C3:} & \quad 0.3 \\
\text{C4:} & \quad 0.2 \quad (\text{End}) \\
\text{C5:} & \quad 0.1 \quad (\text{Onset}) \\
\text{C6:} & \quad 0.5 \\
\text{C7:} & \quad 0.4 \\
\end{align*}
\]

Output probabilities for the phone HMM:

\[
\begin{align*}
\text{Onset:} & \quad 0.3 \\
\text{Mid:} & \quad 0.6 \\
\text{End:} & \quad 0.1 \\
\end{align*}
\]

---

Chapter 15, Section 6

Word pronunciation models

Each word is described as a distribution over phone sequences.

A Phone model is represented as an HMM transition model.

\[
P(\text{towmeytow}) = P(\text{towmeytow}) = P(\text{towmeytow}) = 0
\]

Isolated words

Phone models + word models improve likelihood.

\[
P(\text{e1:tword}) = P(\text{e1:tword})
\]

Prior probability

\[
P(\text{word})
\]

Can be computed recursively:

\[
P(\text{e1:tword}) = P(\text{e1:tword})
\]

Isolated-word dictation systems with training reach 95{99\% accuracy.

Continuous speech

Not just a sequence of isolated-word recognition problems!

- Adjacent words highly correlated
- Sequence of most likely words
- Segmentation: there are few gaps in speech
- Cross-word coarticulation (e.g., "next thing")

Continuous speech systems have 60–80\% accuracy on good day.

Structure is created manually; transition probabilities learned from data.

Continuous speech example with training model accuracy:

\[
(1^{10} X | Y)^{10} = (1^{10} X | Y)^{10}
\]

Chapter 15, Section 6

Language model

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

\[
P(w_1 w_n) = \prod_{i=1}^{n} P(w_i | w_1 \ldots w_{i-1})
\]

Bigram model:

\[
P(w_i | w_{i-1}) = \prod_{i=1}^{n} P(w_i | w_{i-1})
\]

Each word is described as a distribution over phone sequences.

Combined HMM

States of the combined language+word+phone model are labelled 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.

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 most likely word sequence.
DBNs for speech recognition

Also easy to add variables for, e.g., gender, accent, speed.
Zweig and Russell (1998) show up to 40% error reduction over HMMs

Summary

Since the mid-1970s, speech recognition has been formulated as probabilistic inference.
Evidence = speech signal, hidden variables = word and phone sequences.
"Context" effects (coarticulation etc.) are handled by augmenting state.
Variability in human speech (speed, timbre, etc., etc.) and background noise make continuous speech recognition in real settings an open problem.