CSCI 544, Lecture 15: Language models and speech recognition

Ron Artstein

2022-10-11

Administrative notes

No class on October 13

Coding Assignment 3 was due today

- ☛ Late sumbission by tomorrow with 10% penalty

Article selection for presentation due October 18

Coding Assignment 4 due October 25

Project:

Due Date	Task
November 3	Project status report
Nov 29/Dec 1	Poster presentations (in class)
December 1	Final report
December 3	Self-evaluation and peer grading

Article selection

- Each group selects an article for presentation.
- Select from top-tier NLP venues of 2022: ACL, NAACL, TACL, Findings of ACL, or Findings of NAACL.
- Put title and link on shared spreadsheet; check for conflicts.
  - Sheet accessible to USC Google accounts.
- Article does not need to be related to project.
- Pick article with interesting theme, point, or result.
  - If you cannot identify something interesting, choose a different article.
- Presentation does not have to cover the entire article.
- Presentation aimed for audience with the background covered in class.

Coding assignment 4

Write a Perceptron classifier

From scratch!

Corpus of hotel reviews

- Two classification problems: true/fake and positive/negative
- Two models: vanilla and averaged

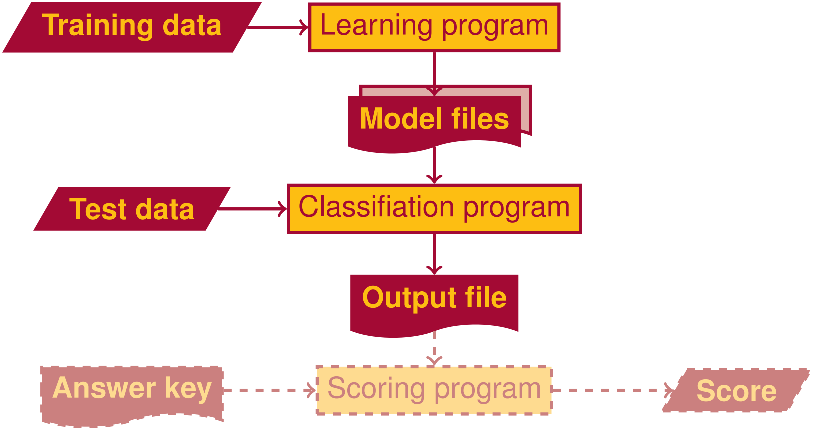
Graded on performance

Programming in Python

Submit on Vocareum

- Automatic feedback
- Submit early, submit often!

Coding assignment 4: programs



Coding assignment 4: notes

Problem formulation

Two binary classifiers: two sets of parameters

Features and tokenization

Experiment to see what works best!

. . . after you get the basic program working

Runtime efficiency

Compact representation

Don't multiply zeros

Over/underfitting

Choose number of iterations

Probability of a sentence

Does it make sense to talk about the probability of a sentence?

> or 'situations'. But it must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term. On empirical grounds, the probability of my producing some given sentence of English – say, this sentence, or the sentence "birds fly" or "Tuesday follows Monday", or whatever – is indistinguishable from the probability of my producing a given sentence of Japanese.

Noam Chomsky. Quine's empirical assumptions.
Synthese 19(1/2): 53–68, 1968.

Language models

Probability of a sentence/utterance $\approx$
Probability of word given previous words

Unigram language model (bag of words): Independence assumption (similar to naïve Bayes)

$$P(w_1 \dots w_n) \approx \prod_i P(w_i) = P(w_1) \cdot P(w_2) \cdots P(w_n)$$

Bigram language model: Markov assumption

$$P(w_1 \dots w_n) \approx \prod_{i=0}^{n} P(w_{i+1}|w_i) = P(w_1|\#) \cdot P(w_2|w_1) \cdots P(\#|w_n)$$

N-gram language models

Trigram, 4-gram etc. use probabilities for longer word sequences

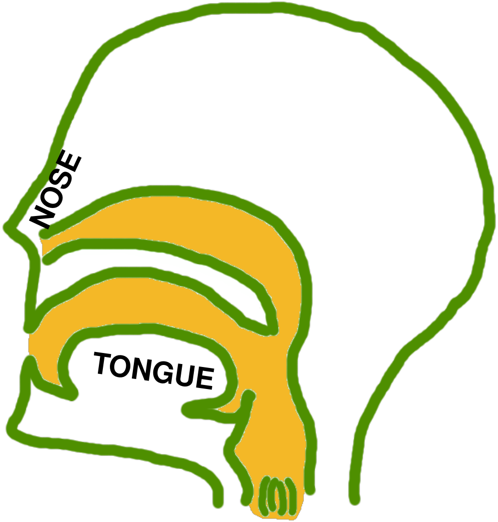
Some smoothing methods:

- Backoff: Use probability of highest available n-gram
- Interpolation: weighted mixture of trigram, bigram, unigram probabilities
- Etc.

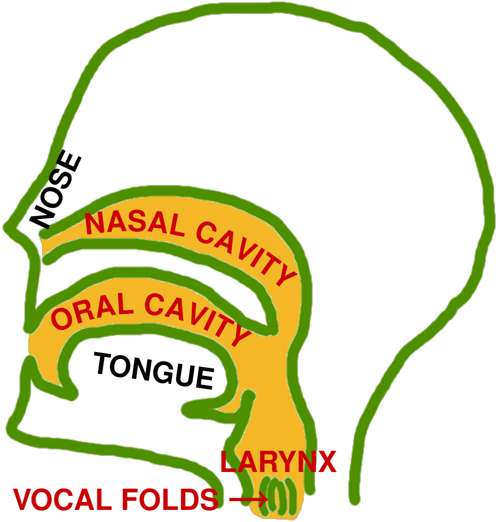
Language models can give a **prior** probability on sentences

Human vocal tract (and some phonetics)

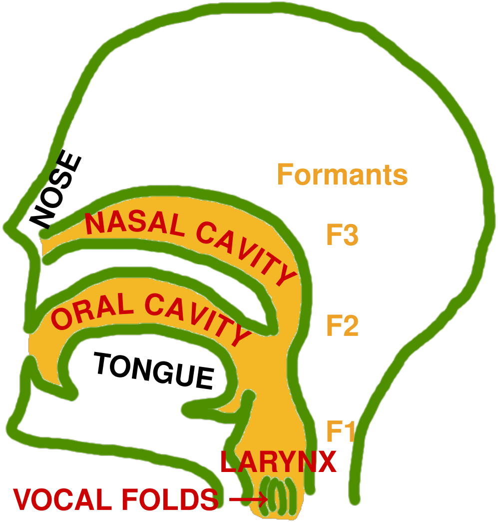
Human vocal tract (and some phonetics)

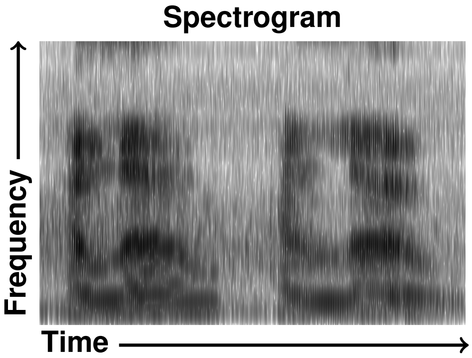


Human vocal tract (and some phonetics)

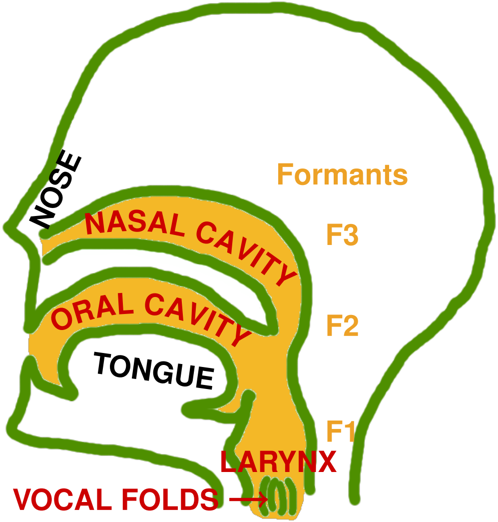


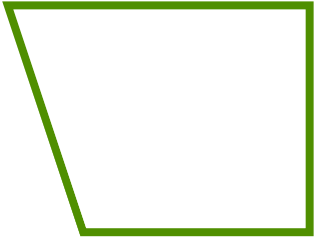
Human vocal tract (and some phonetics)



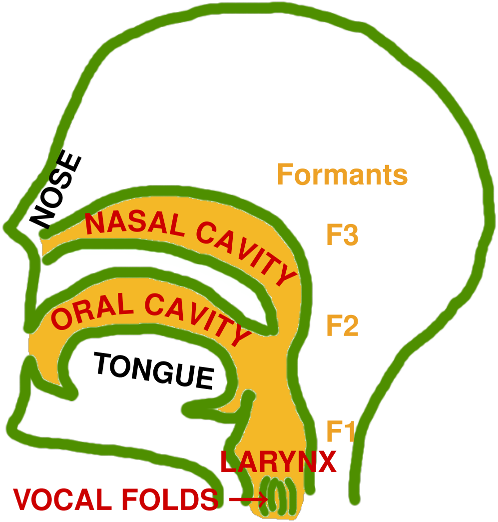


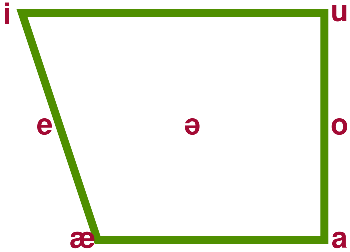
Human vocal tract (and some phonetics)



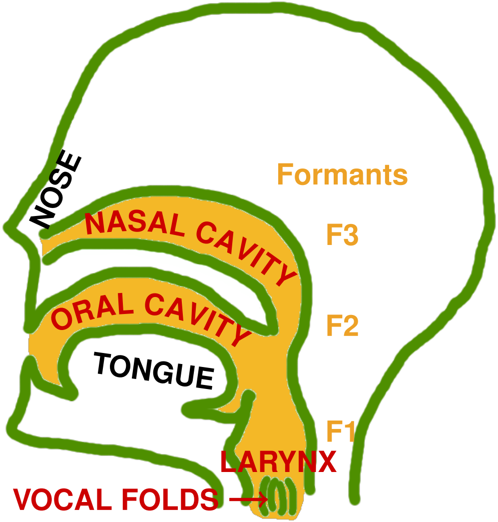


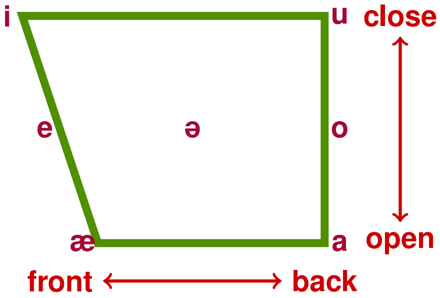
Human vocal tract (and some phonetics)



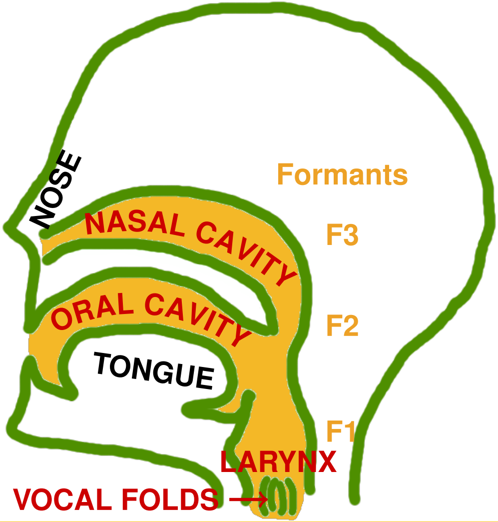


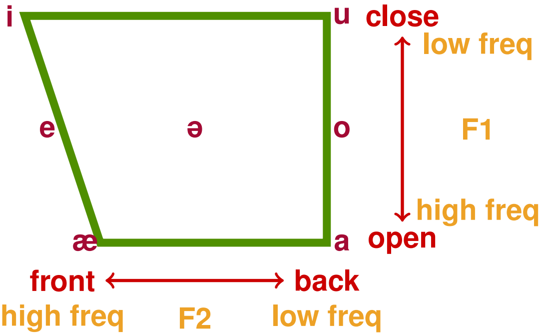
Human vocal tract (and some phonetics)





Human vocal tract (and some phonetics)





Speech recognition basics

Sample the waveform (e.g., every 10 msec)

Extract features

- Mel Frequency Cepstral Coefficients: MFCC
- Mel scale: pitch $\rightarrow$ perceptual pitch (works better)

Calculate $P(\text{word}|\text{sound})$

[word string] [time series of acoustic vectors]

Easier to calculate $P(\text{sound}|\text{word})$;
need prior distribution on word strings

$$P(\text{word}|\text{sound}) \propto \underbrace{P(\text{sound}|\text{word})}_{\text{[acoustic model]}} \cdot \underbrace{P(\text{word})}_{\text{[language model]}}$$

Words are too big…

Predict acoustic model from shorter units of sound: **phones**

	feeling					
IPA	f	i	l	ɪ	ŋ	(One character per phone)
ARPABET	F	IY	L	IH	NG	(Phones separated by spaces)

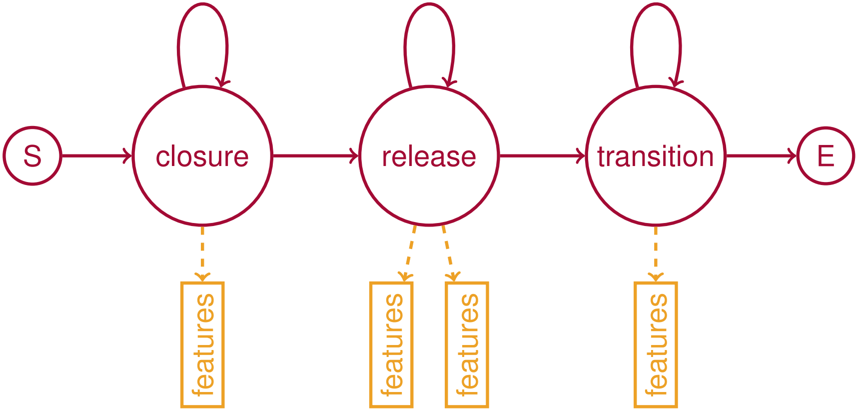
Vowels and voiced sounds have formants throughout

Stops are silent, affect formants before closure and after release

Phones modeled as 3-state HMMs: closure, release, transition

- States, transitions are discrete
- Observations are continuous:
  $P(\text{observation}|\text{state}) = \text{Normal}(\text{mean}, \text{variance})$

HMMs for phone recognition



Decoding speech

Find the most likely sequence of words, given a sequence of feature vectors

- Transitions *within* phones are for structure
- Transitions *between* phones constrained by language model and lexicon

Search through a lattice of hypotheses, with pruning (e.g. Viterbi decoding, beam search)

Speech recognition basics (review)

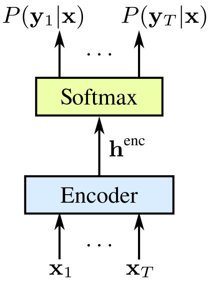
$$P(\text{word}|\text{sound}) \propto \underbrace{P(\text{sound}|\text{word})}_{\text{[acoustic model]}} \cdot \underbrace{P(\text{word})}_{\text{[language model]}}$$

- Acoustic model conditioned on phones
- Language model = distribution on words
- Phones and words linked by dictionary (lexicon)

Splitting the problem into acoustic and language models

- Acoustic model may vary by speaker
- Language model may vary by domain

Neural acoustic modeling: CTC



(a.) CTC

Graves et al., ICML 2006

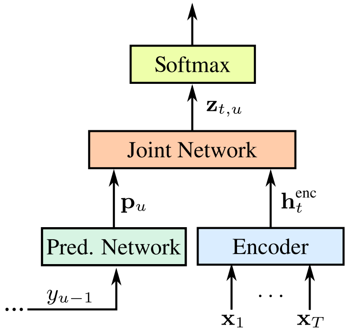
- Acoustic feature vectors $\rightarrow$ transcription symbols
- Output includes blank symbol
- ✔ Works well for phonetic transcription

From phonetic symbols to words: dictionary + LM

Prabhavalkar et al., Interspeech 2017

- ✘ Mapping acoustic feature vectors directly to characters doesn't work so well

Neural language modeling: RNN-Transducer



(b.) RNN-Transducer

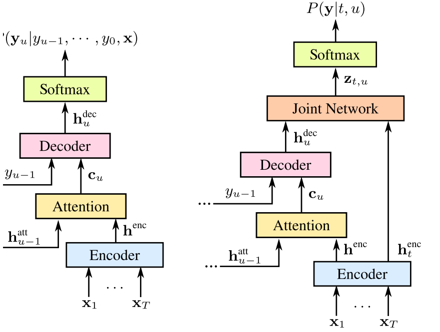
Graves, ICML workshop 2012

- Prediction network: dependencies between output symbols
- Analogous to a language model
- ✔ Predict phonetic transcription

Prabhavalkar et al., Interspeech 2017

- ✔ Mapping acoustic vectors to characters
- **Streaming** with unidirectional networks

Neural modeling: attention



c.) Attention-based Model (d.) RNN-Transducer with Attention

Chan et al., ICASSP 2016

- End-to-end: Listen-attend-spell

Attention: non-local dependencies

Prabhavalkar et al., Interspeech 2017

- ✔ Acoustic vectors to characters
- ✘ Difficulty with streaming

Chiu et al., ICASSP 2018

- Wordpieces instead of characters
- Other oprimizations

Evaluating speech recognizers

Word Error Rate: Levenshtein edit distance between reference and hypothesis

$$\text{WER} = \frac{\text{Deletions} + \text{Insertions} + \text{Substitutions}}{\text{Length of reference string}}$$

Ref: She is going to the beach
Hyp: She is **gone** to **am** the beach

$$\text{WER} = \frac{2}{6} = 33.3\%$$

Ref: I'm
Hyp: I am

$$\text{WER} = \frac{2}{1} = 200\%$$

Other measures: Concept Error Rate (what's a concept?)
Morpheme Error Rate
Phone Error Rate

Phones can match better than words

Speech:	Are you **married**	Word Error Rate
ASR:	Are you **Mary**	1/3 = 0.33

Speech:	AA R Y UW M EH R IY **D**	Phone Error Rate
ASR:	AA R Y UW M EH R IY	1/9 = 0.11

Speech:	Are **all** **soldiers** **deployed**	Word Error Rate
ASR:	Are **also** **just** **avoid**	3/4 = 0.75

AA R AO L S OW **L** JH **ER** **Z** **D** **IH** **P** **L** OY D	Phone Error Rate
AA R AO L S OW JH **IH** **S** **T** **AH** **V** OY D	7/16 = 0.44

Wang, Artstein, Leuski, Traum FLAIRS 2011