Lecture 20: ASR

LING 1340/2340: Data Science for Linguists

Na-Rae Han

Objectives

- ASR!
  - ASR demo
  - ASR theory

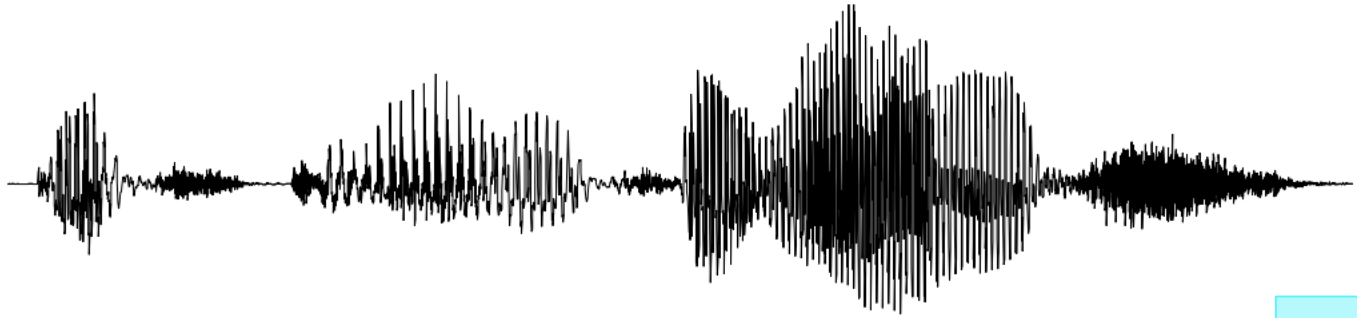
Forced Alignment, but no transcripts?

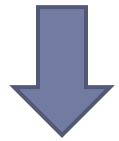
- Forced alignment needs:
  - Audio recording (wav file)
  - Transcript (txt file)
- But wait! What if we don't even have a transcript file?
- We can auto-transcribe… ASR!
- ASR demo using `SpeechRecognition` Python library

  ➔ In Jupyter Notebook

Backing up: ASR

- Forced alignment is based on ASR technology.
- This is NOT an NLP class, but we should at least have some sense of how ASR works…





It's time for lunch

Is **processing speech** going to be entirely different from **text processing technologies**?

In which we skim through blog articles in lieu of proper academic textbook (... and point at things we recognize)

- Proper academic textbook chapter on ASR/TTS:
  - Jurafsky & Martin (2020) *Speech and Language Processing* Ch. 16 Automatic Speech Recognition and Text-to-Speech

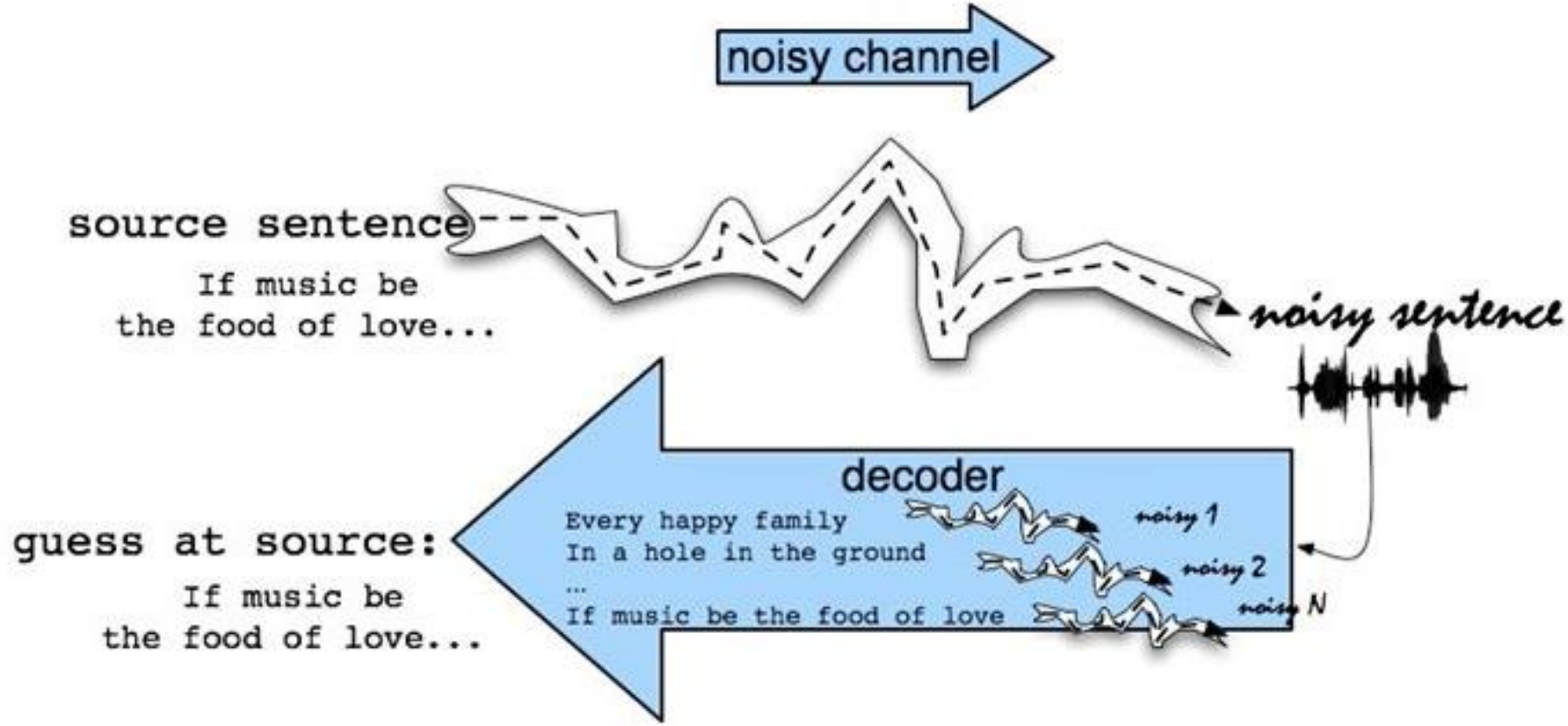
- More accessible:
  - Speech Recognition – ASR Model Training (by Jonathan Hui)
  - Introduction to ASR (by Maël Fabien, with IPA!!)

All the building blocks…

- English:
  - ARPAbet
  - CMU Pronouncing Dictionary
- World languages:
  - G2P (grapheme-to-phoneme)
- HMM (Hidden Markov Model), HTK (HMM ToolKit)
- Kaldi (ASR toolkit, built on HTK)
- Finite-State Transducer (OpenFST)
- N-gram language models

Many of them look familiar…
from LING 1330
Intro to CompLing!

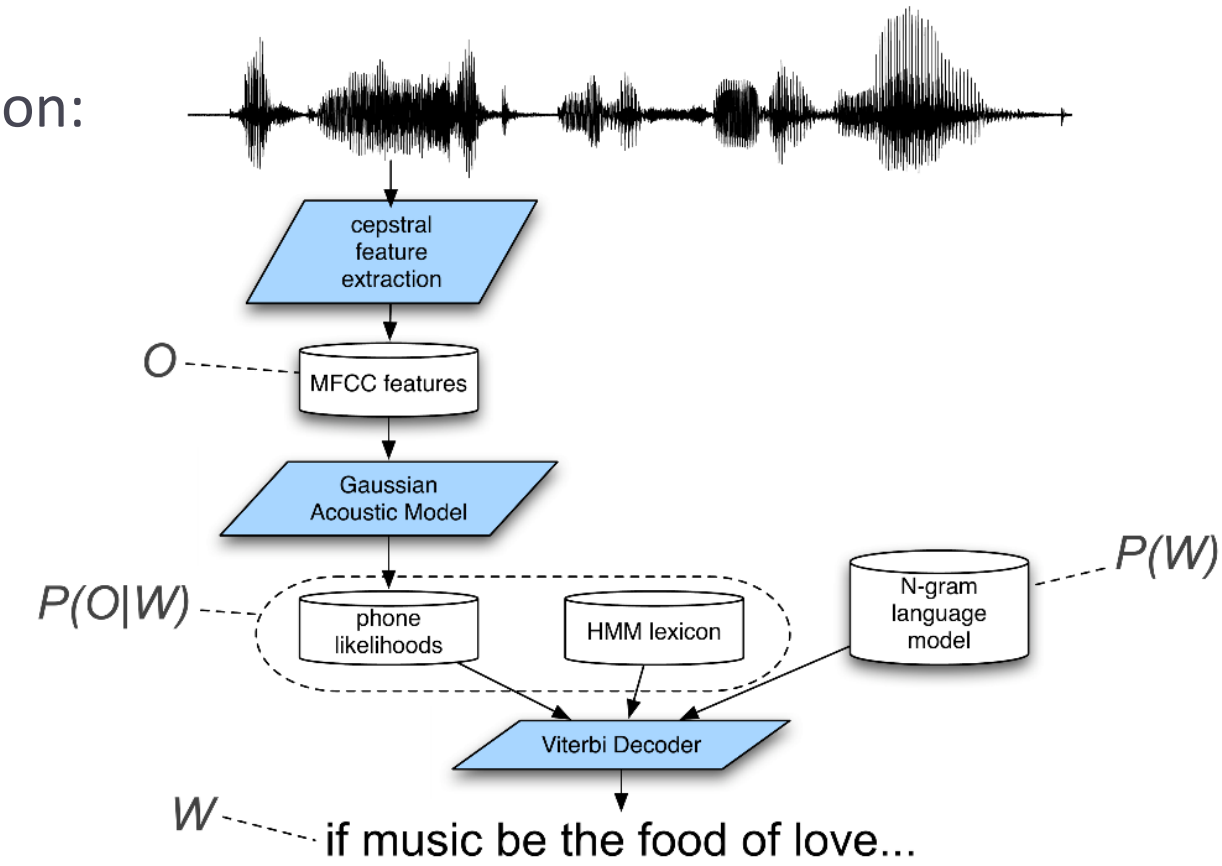
The Noisy Channel Model



SLP, Jurafsky & Martin
https://web.stanford.edu/~jurafsky/slp3/B.pdf

Speech recognition architecture (classic)

- ASR components
  - Lexicons and pronunciation:
    - Hidden Markov Models
  - Feature extraction
  - Acoustic modeling
  - Decoding
  - Language modeling:
    - N-gram models



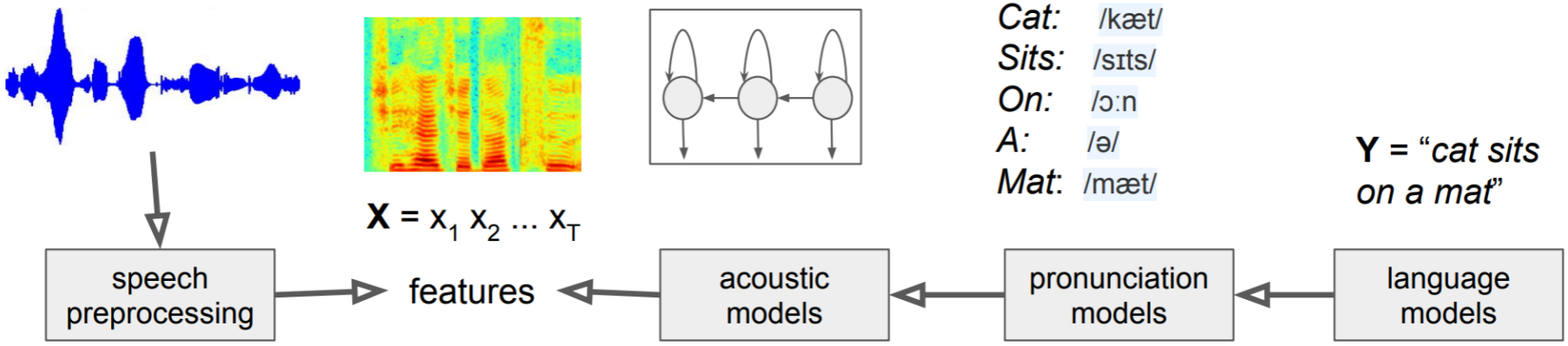
- But: why "classic"?

Because **DEEP LEARNING** (what else?)

SLP, Jurafsky & Martin

Speech recognition architecture (classic)

- Inference: Given audio features **X** = $x_1 x_2 \ldots x_T$ infer most likely text sequence **Y*** = $y_1 y_2 \ldots y_L$ that caused the audio features

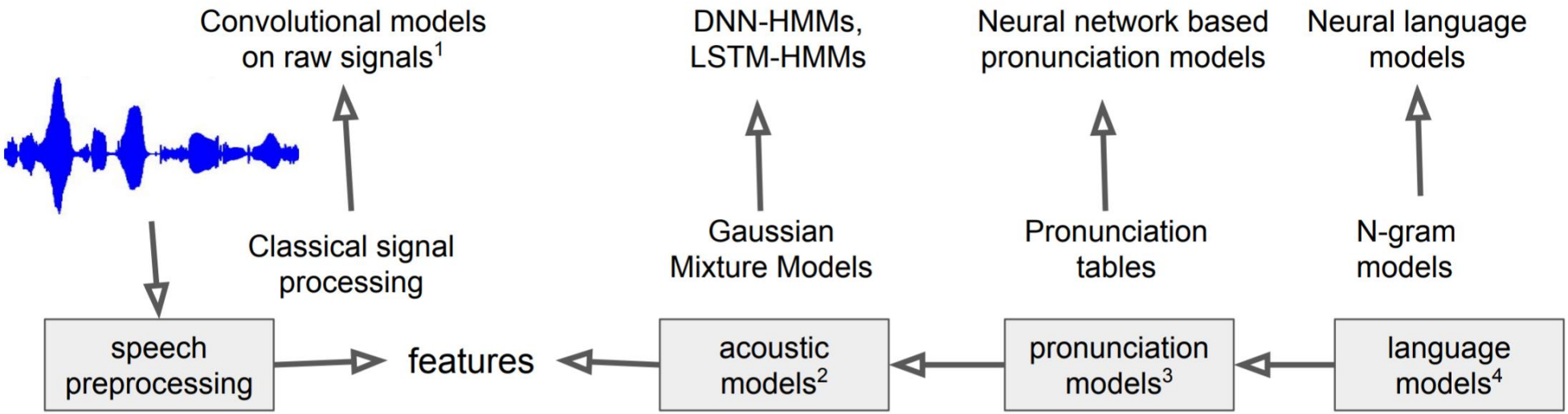


$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} p(\mathbf{X}|\mathbf{Y})\, p(\mathbf{Y})$$

SLP, Jurafsky & Martin

Speech recognition architecture (neural net)

- Each of the components seems to be better off with a neural network



SLP, Jurafsky & Martin

Course Wrap

- That was a course! (Na-Rae's lecture portion anyway.)
- Keep on learning – what's next?
  - Classes to take at Pitt?
  - DataCamp: Educational premium access effective until July 6
  - Keep on supercomputing! We used less than 1% of our 50,000 SU, our allocation is active until September 3
  - Join PyLing! (Pitt Python Linguistics Group)

Wrapping up

- Let's go over:
  - Presentation guidelines
  - Final project submission guidelines

- Next class:
  - Riley: Introduction to SQL
  - Qidu project presentation