Lecture 17: ASR Theory

LING 1340/2340: Data Science for Linguists

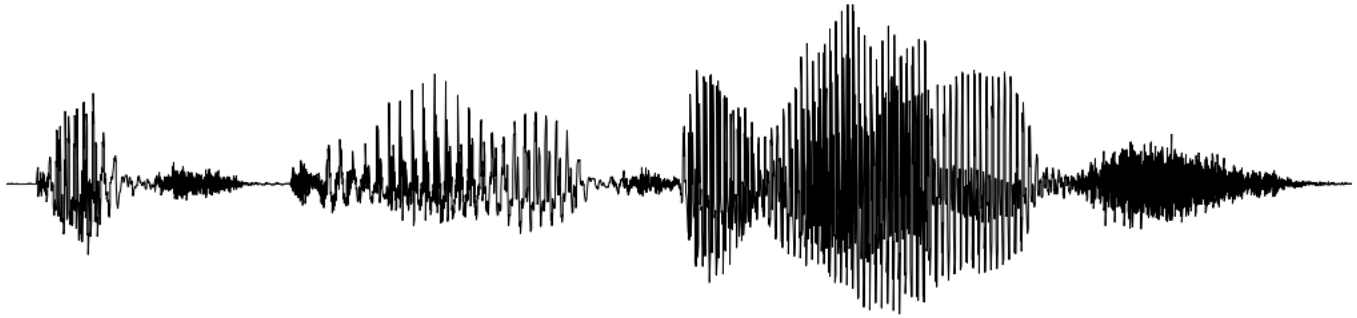
Na-Rae Han

Objectives

- ASR theory!

ASR

- ASR technology is fairly mature, especially for languages like English.
- 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 (Again) In Lieu of Proper Academic Textbook

- Proper academic textbook chapter on ASR/TTS:
  - Jurafsky & Martin (2020) *Speech and Language Processing* Ch. 26 Automatic Speech Recognition and Text-to-Speech
- More accessible:
  - Speech Recognition – ASR Model Training (by Jonathan Hui)
    - Part of a series, "Forced alignment" at bottom!
  - Introduction to ASR (by Maël Fabien, less technical, with IPA!!)

  ← Let's take a quick look at this one…

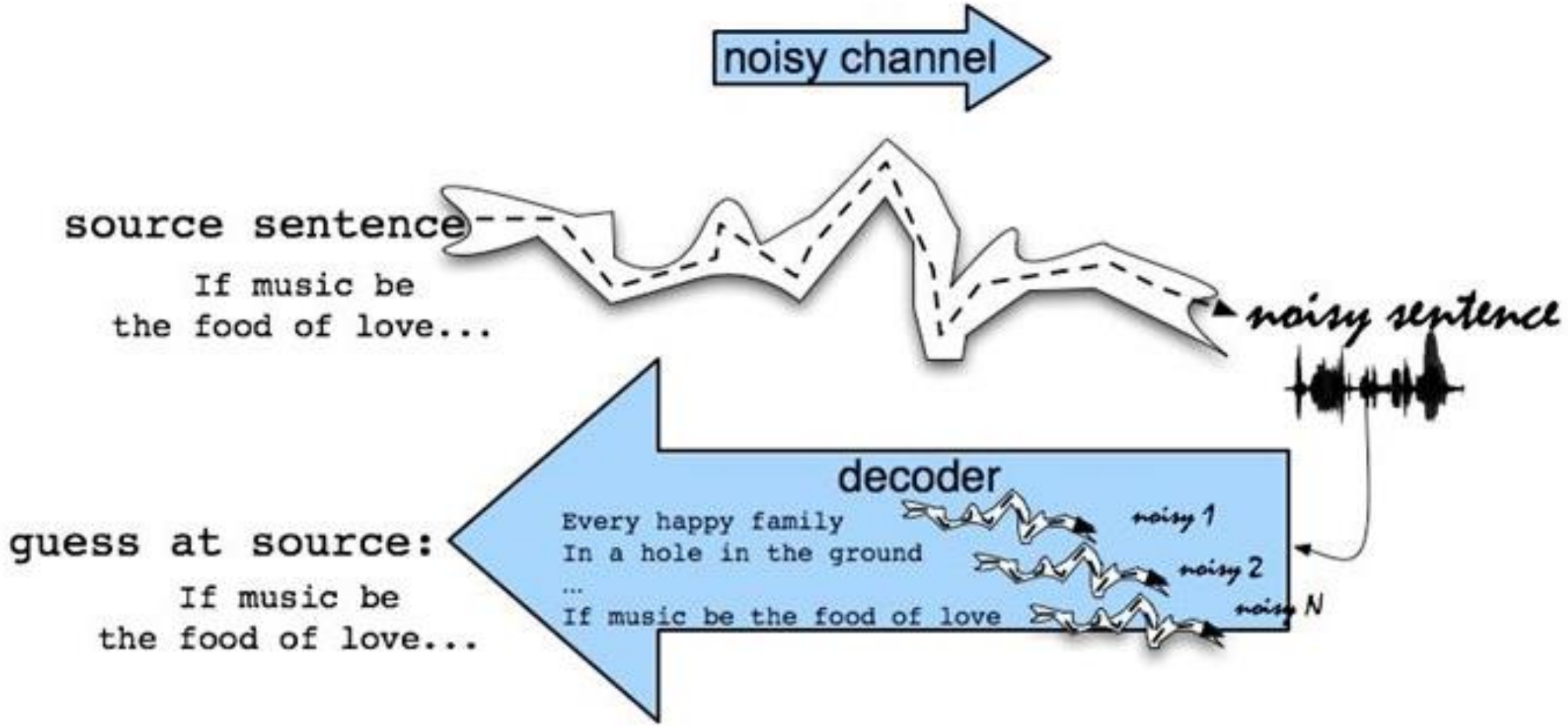
All the building blocks…

- English:
  - ARPAbet
  - CMU Pronouncing Dictionary
- World languages:
  - G2P (grapheme-to-phoneme)
- HMM (Hidden Markov Model), HTK (HMM ToolKit)
- Kaldi (ASR toolkit, built on HTK)
- (Weighted) 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

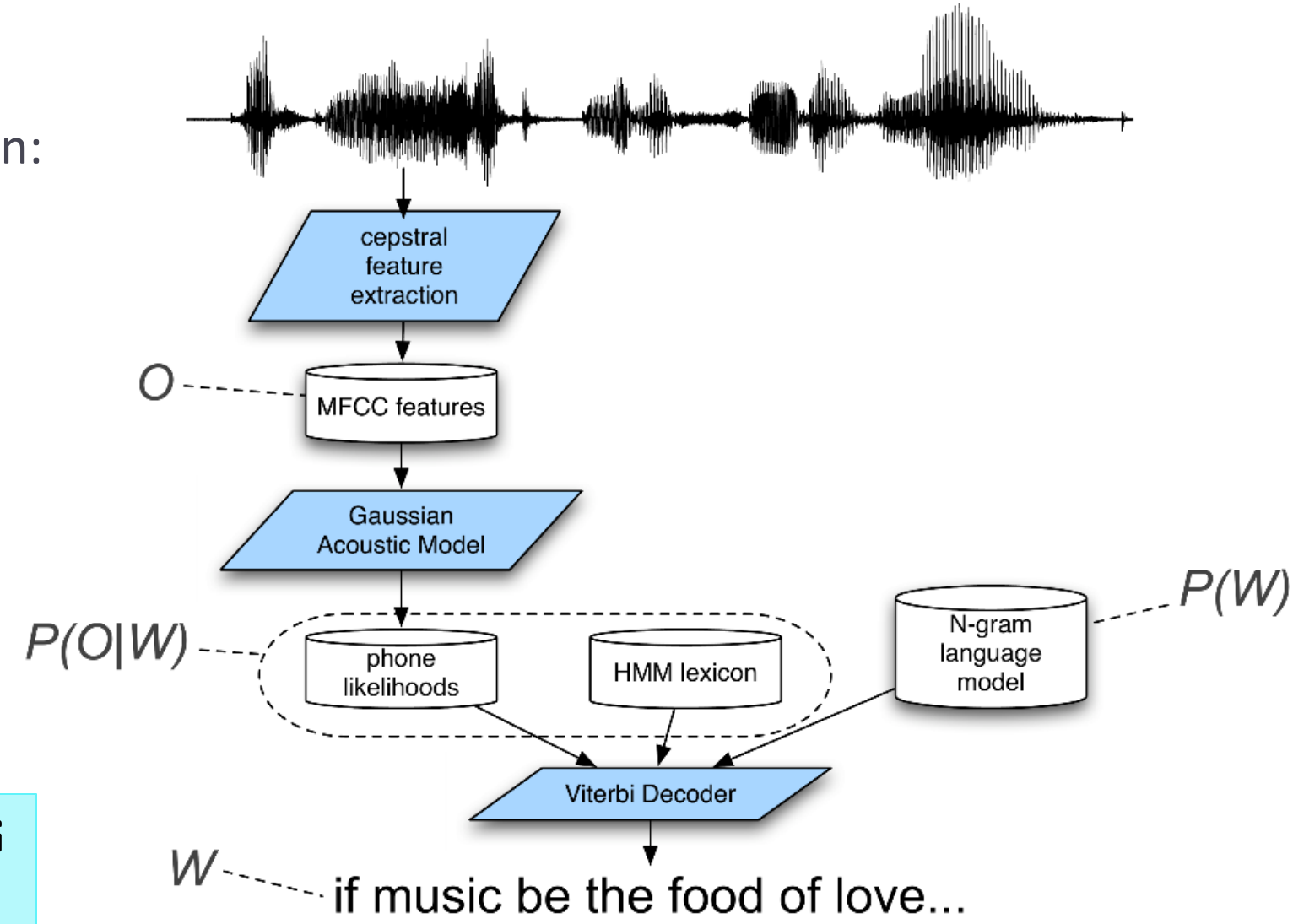


Speech recognition architecture (classic)

SLP, Jurafsky & Martin

- ASR components
  - Lexicons and pronunciation:
    - Hidden Markov Models
  - Feature extraction
  - Acoustic modeling
  - Decoding
  - Language modeling:
    - N-gram models
- But: why "classic"?

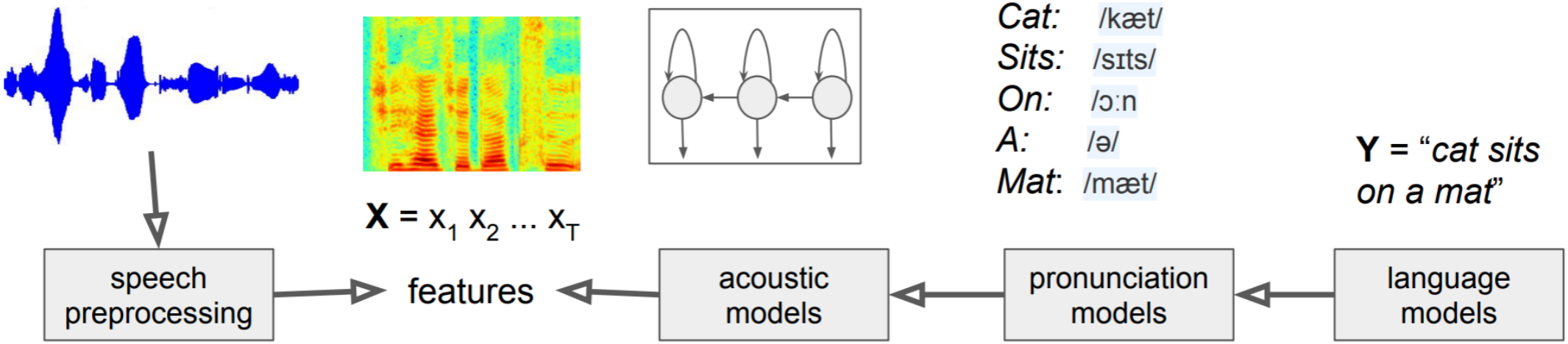
Because **DEEP LEARNING** (what else?)



Speech recognition architecture (classic)

SLP, Jurafsky & Martin

- 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

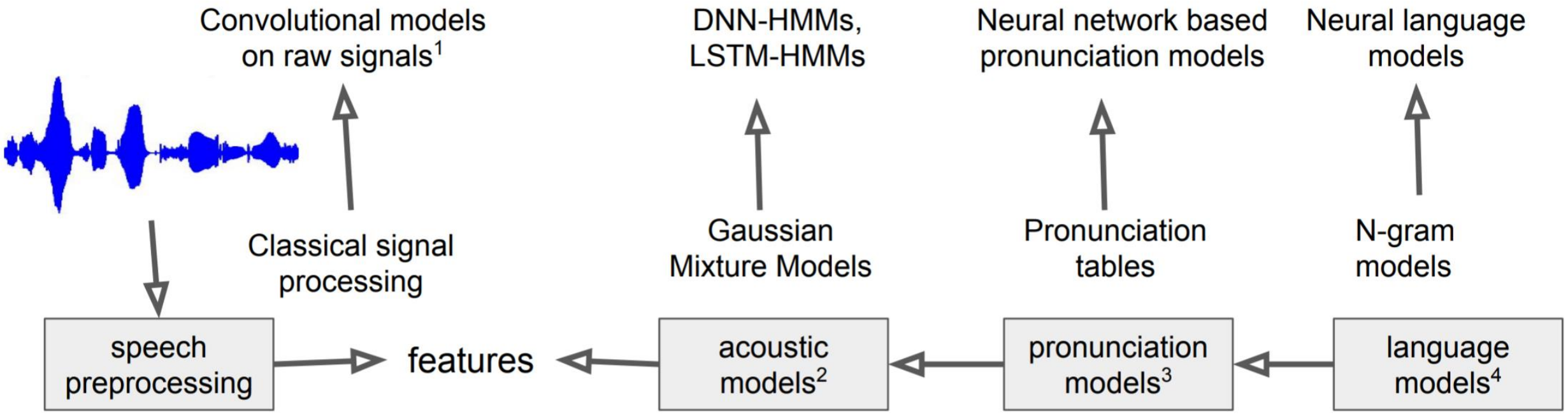


$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} p(\mathbf{X}|\mathbf{Y})\, p(\mathbf{Y})$$

Speech recognition architecture (neural net)

SLP, Jurafsky & Martin

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



Wrapping up

- Next class
  - Presentations: Ben, Kinan, Manho, Alejandro

- Final project submission:
  - May 1 (Sun) 6pm
  - If using 2-day late pass, email and LET ME KNOW before SUNDAY! (Final grade is due on Wed)