LING 1340/2340

# SPEECH DATA SCIENCE

# RECAP: "DATA SCIENCE"

- "Bringing structure to large quantities of formless data" (Davenport & Patil 2012)
- Sourcing/sifting/cleaning/organizing data in the wild

## SPEECH VS. WRITING

- Speech is ubiquitous to human communities
- Writing was invented
- Speech is spontaneous
- Writing is deliberate
- Humans acquire speech without instruction
- Writing requires instruction to learn

# SPEECH CORPORA

- Ubiquitous:
  - All communities, all languages
- Not deliberate:
  - Different audience design considerations (Bell 1984)
  - More plentiful; more contexts
- No instruction needed:
  - Less formal* constraints

# WHAT TO DO WITH SPEECH DATA?

- Analyze it directly.
    - Language identification
    - Phonetic research
    - Informing models (such as the following)
- Convert it to text, then do other things with it…
    - ASR (Automatic Speech Recognition) and ASU (Understanding)
    - Automatic closed-captioning
- Make it!
    - Speech Synthesis / Text-to-Speech (TTS)
    - Conversational Agents

# POPULAR SPEECH CORPORA

- Buckeye Corpus (Pitt et al. 2005)
- TIMIT (Garofolo et al. 1993)
- TalkBank links

# POPULAR SPEECH DATA ANALYSIS TOOLS FOR LINGUISTS

- Praat (Boersma & Weenink 2019)
- Klatt formant synthesizer (Klatt 1975, 1984)
- Penn forced aligner (Yuan & Liberman 2009)
  - FAVE–align (Rosenfelder et al. 2011)
  - Montreal Forced Aligner (McAuliffe et al. 2017)
  - EasyAlign (Goldman 2011 – Windows only)
- ELAN multimodal annotator (Wittenburg et al. 2006)

# SPEECH RECOGNITION BASICS

- Assume that all speech data is noisy (*“noisy-channel” model*)
- Compare every possible sentence to the target waveform, and select the best match (*decoding/search/inference*)
  - *What is the “best match”?* Bayesian inference.
  - *Every possible sentence?!* Hidden Markov Models.

# THE HIDDEN MARKOV MODEL AND SPEECH - ASSUMPTIONS

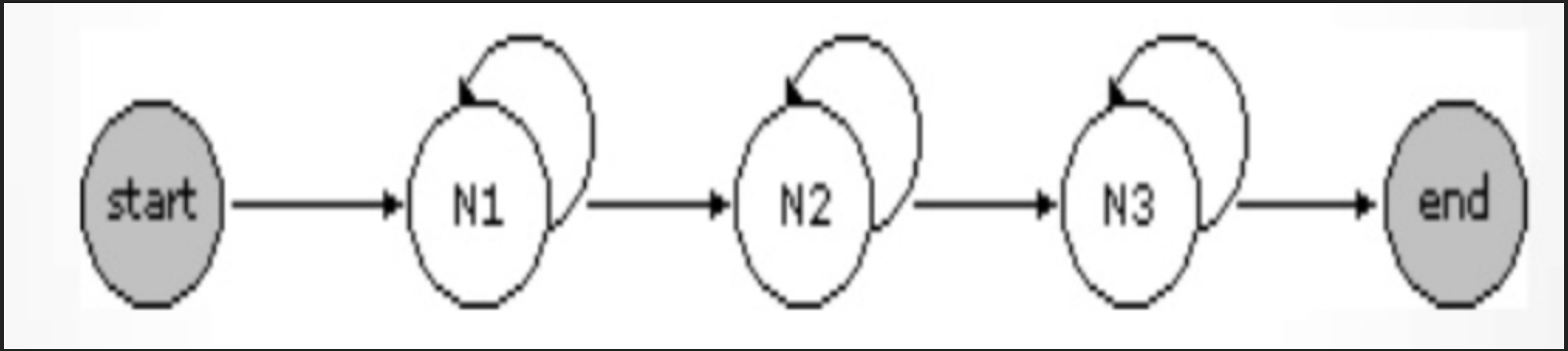
- The speech stream is a sequence of steady states
- Transitions between states are not arbitrary
  - Simple assumption: any state (phone) transitions only to itself or to a specific following state
  - Phonemes are encoded as a series of states (*Why?*)
- Each word is a different HMM composed of phone HMMs

# ASR: ISSUES

- Speaker variation
- Genre variation
- Noise/environmental variation
- Disfluencies
- [Predictive text issues]
- Decoding

# FORCED ALIGNMENT

- Task is to determine when N1, N2, N3 begin



- Is there still inference?

# FOR THURSDAY

- To-do #15:
  - Install an *updated version* of Praat
  - Download TIMIT corpus (Licensed-Data-Sets)
- 3rd progress report next Tuesday…