LING 1340/2340

(MULTIMODAL) DATA

AGENDA

- Progress report: How goes?
- Speech data recap and issues
  - More tools: Parselmouth, SpeechRecognition
- Multimodal data

RECAP: SPEECH DATA

- Sourcing/sifting/cleaning/organizing data in the wild
- Common task: convert to text
- Default assumptions (*noisy-channel model, HMM*)
- Issues with ASR (variation, decoding)

ISSUE: LEVELS OF COMPLEXITY

- Forced alignment - no word-level inference
- Task-specific data - few reasonable competitors
- Large Vocabulary Continuous Speech Recognition (LVCSR)
  - *a.k.a.* speech analytics

APPROACHES TO LVCSR

- Topic analysis
- Speaker-dependent training
- *n*-gram modeling (for phones and words)
- Deep learning (Deep/Recurrent neural networks)
- Adaptive training

DEEP NEURAL NETWORKS

- Successive layers of a neural network; multiple levels of representation (e.g. of linguistic structure)
  - See Anish's slides from last semester
- Recurrent neural networks include temporal states
- Both require a LOT of training data

ISSUE: HOW MUCH DATA?

- In principle: *enough to be able to distinguish the signal from the noise*
- Enough to inform enough feature layers
- Pre-training can compensate for low training resources (Thomas et al. 2013; Vu et al. 2011)

MORE TOOLS

- Praat script repositories
- Parselmouth: Access Praat code through Python
  - Also not very well documented!
- SpeechRecognition package: Use ASR APIs through Python
- aeneas: Forced alignment through Python

ELAN: ANNOTATION FOR VIDEO + AUDIO

- (link)
- Projects using ELAN: https://tla.mpi.nl/past-projects/
- Example: BU ASL corpus