Speech Recognition
Final Project Information

Cyril Allauzen
Google, NYU Courant Institute
allauzen@cims.nyu.edu

Slide credit: Eugene Weinstein
Logistic

- Project proposal due on April 24th (brief description of idea in email form)
- Final project writeup due on May 11th at 7pm.
  - 2-4 page description of the background, technique used, and result
- Project presentations: May 13th, 7-9pm, WWH 201
- Collaboration is allowed/encouraged (expectations grow with group size)
Goal

• Goal is to do a substantial exploration of some non-trivial idea

• Idea complexity: more complex than those in homeworks
  – Need to explore idea in more detail
  – Should explore more than one technique

• Project ideas not restricted to those in this presentation
Idea 1: Vowel Restoration

• Followup on HW1.
• A real NLP task for languages such as Arabic or Hebrew.
• With English: easy to create a training and test sets from any text corpora (some text normalization probably needed)
• Explore the ideas suggested in HW1.
• Investigate language modeling questions: character-based or word based LM, smoothing methods, $n$-gram order, shrinking,...

• A lot of related NLP tasks:
  – Diacritic restoration
  – Capitalization
  – Segmentation
  – Input methods/error corrections
Idea 2: Explore Kaldi

- Kaldi is a complete state-of-the-art speech recognition system.
  - Many tools implemented by Kaldi.
  - Several experimental setups provided in the `egs` directory, many of them using freely available dataset (librispeech, vys-tadial, voxforge, ...).
  - Several extra packages easy to install from the `tools` directory.

- Examples of ideas:
  - Acoustic modeling: GMMs, DNNs, features, speaker-adaptation, ...
  - Language modeling rescoring: $n$-gram models, MaxEnt models, RNNs, lattice vs. $n$-best rescoring.
Idea 3: Multi-Domain Language Models

• Download text corpora from different domains
  – e.g., Europarl and OANC (many domains)
  – Most corpora will need some cleanup - e.g., remove caps, punctuation, tags.

• Split each into training, test, and dev sets

• Compare in-domain to cross-domain language model quality

• What if you interpolate models from multiple domains?

• Or maybe combine training data sets?
Idea 4: Acoustic Model Analysis

- Goal: empirically explore closeness of different phonemes and their models.
- You can use a corpus in Kaldi for this (HW0).
- Compute and plot closeness of all pairs of CI phonemes based on their average MFCC vectors. Repeat for CD phonemes.
- Now, compute distances between their single-Gaussian models. What about GMMs?