Speech Processing 15-492/18-492

Speech Recognition
Systems
Other ASR techniques
How good are they?
- Expected ASR
- Factors that make things worse

How good do they need to be?
- What can you do with low WER?
Continual Progress in Speech Recognition at (D)ARPA Evaluations
What makes it worse

- **Channel**
  - Telephone vs Wide band
  - Close-talking vs far-field

- **Style:**
  - Command and Control
  - Limit information getting
  - Limit domain but general speech
  - Machine directed vs Human directed speech
  - Broadcast (performance) vs Conversational
  - Single vs Dialog vs Multiperson
Expected WER: Real-time

- **Command and Control**
  - Limited vocabulary and directed speech
  - < 10% (< 5% for some users)

- **Simple Dialog**
  - Machine directed speech with interested users
  - < 20% (but sometimes works with < 30%)

- **Dictation**
  - Single speaker, well performed
  - <5% for some users > 30% for (short term) users

- **Speech-to-Speech Translation**
  - Machine mediated, target domain
  - <20% (but will vary for different people)
Expected WER: offline

- **Broadcast News**
  - Large vocabulary, well performed
  - <10% but not real-time (maybe 100 times real time)

- **Conversational Speech (Call Home)**
  - Large vocabulary, not well performed
  - > 40% WER (depends on particular users and conversations)

- **Information retrieval**
  - Large vocabulary very varied content
  - > 60% can still give useful results
Other uses

- TV show subtitling for the deaf
- Court transcription
- Medical dictation
- Air traffic control transcription
Other ASR techniques

- Including Articulatory/Phonetic Features (Metze)
- Build recognizers for
  - Voiced/unvoiced
  - Nasality
  - Closures (quiet part of stops)
  - Aspiration (Fricatives)
  - Tongue position
- Run all in parallel and “join” them
- Combine with more standard approaches
- Can be more robust to speaking style
Multi-engine Recognition

- **Use three recognizers and combine results**
- **Rover**
  - Combine scores per-sentence
- **Combine lattices**
  - Confusion networks
- **Cross adaptation**
  - Interleave systems with adaptation
- **It usually works better when system different**
  - (and both of them good)
Whispered Speech

- Doesn’t disturb other people
- Can use throat mike
- Works in noisy environment
Muscle Movement

- **EMG: Electromyographic Signals**
  - Recognize muscle impulses
- **Can work in noisy environments**
- **Can work without you making a noise**
Articulatory Movement

- **Attach metal studs to:**
  - Lips, teeth, tongue, velum

- **Record movement in magnetic field**
  - Non-intrusive
EMA: Electromagentoarticulatograph
ASR requires:

- **Acoustic model**
  - HMMs trained from lots of data

- **Pronunciation lexicon**
  - List of pronunciations for words

- **Language model**
  - Trigrams trained from lots of data
More/better training data
  - Well transcribed and closest to target system

Better signal
  - Better microphone, no noise

Better speaker
  - Interested party, know how to speak

Time and memory
  - Bigger systems do better
  - Greater CPU does better