Speech Recognition

MIT 6.893
SMA 5508
Spring 2004
Larry Rudolph (MIT)

A long term goal

- Since 1950, AI researchers claimed
  - Crucial problem
  - Will be solved within the decade
- Finally, it appears true
- Failure rates still too high
  - 90% hit rate is 10% error rate
    - want 98% or 99% success rate

Spectrum of choices

	Constrained Domain	Unconstrained Domain
Speaker Dependent	Voice tags (e.g. phone)	Trained Dictation (Viavoice)
Speaker Independent	Galaxy (we are here)	What everyone wants

Waveform to Phonemes

- Waveform is very fuzzy
- We think there is a large break between words and sentences
  - hard to see from waveform
- Mapping waveform segments to phonemes is not accurate

Phonemes to words

- Group phonemes into words
  - not always 1-1 mapping
    - missing phonemes
    - false phonemes (extra ones)
    - accents
  - many possible choices
- Word should be known to system
  - domain or dictionary

words to sentences

- People do not always speak grammatically correct
  - some invariant rules (for speech)
  - extra or missing words
  - phrases not always sentences
- Easier when sentence is in domain
  - domain specified by grammar

sentences into meaning

- Dictation system: want sentences
- Other system: want to understand
- Integrate high-level processing
  - Most applications need it anyway
  - Helps with recognition
    - useful to disambiguate input

meaning into action

- What happens after meaning?
  - Respond to user (even a beep)
  - Usually generate more substantial response
  - Action should be valid in context

Disambiguation

- Each transformation is rarely highly accurate
- Lots of choices
- Subsequent steps can rule out choices from previous steps

disambiguation strategy

- Select “n-best” choices and pass on
- Each step restricts possible meaning
- Make heavy use of probability
- Viterby search
  - state transitions along with probabilities.
  - push through n choices at once

after domain dependent

- Handling out-of-vocabulary words
- Multimodal input
  - improve recognition rates
    - e.g. lip reading
  - sometimes easier to point than say