Automatic Speech Recognition

- Acoustic observations: signal processing to extract energy levels at each frequency level
- Observation sequence $o$ is composed of acoustic features extracted from the waveform at regular (10msec) intervals
- ASR is the task of converting the observation sequence $o$ into a transcription $w$
Noisy Channel Model

• Finding the best transcription $w^*$ given an observation sequence $o$

$$w^* = \arg \max_w P(w | o) = \arg \max_w \frac{P(o | w)P(w)}{P(o)}$$

= \arg \max_w \frac{P(o | w)P(w)}{P(o)}$

generative model

language model

Generative Models of Speech

• **Speech recognition**: find word sequence $w$ that maximizes $P(w | o)$, where $o$ is a sequence of time dependent acoustic features (output of signal processing on speech signal)

• Typical decomposition of $P(w | o)$ into a cascade of generative models:
  - Acoustic Model:
    $P(o | p)$ predict observation sequence $o$ given phone sequence $p$
  - Pronunciation Model:
    $P(p | w)$ predict phone sequence $p$ given a word sequence $w$
  - Language Model:
    $P(w)$ predict word sequence $w$
Generative Models of Speech

- \( P(w \mid o) = P(o \mid w) \ast P(w) \) using Bayes Rule
- Decomposition of \( P(o \mid w) \) into a cascade of models:
  - **Acoustic Model** \( P(o \mid p) \) (model trained on the TIMIT corpus):
  - **Pronunciation Model** \( P(p \mid w) \) (model trained using TIMIT and the CMU pronunciation dictionary):
  - **Language Model**: \( P(w) \) (model trained using large amounts of text in the same domain)

  *cf. Fundamentals of Speech Recognition, Rabiner and Juang*

Generative Models of Speech

- Further decomposition of the acoustic model: \( P(o \mid p) \)
  - \( P(o \mid d) \) observation vectors given distribution sequences (quantitative given symbolic)
  - \( P(d \mid m) \) distribution sequences given model sequences (model dependent phone sequences)
  - \( P(m \mid p) \) model sequences given phone sequences
Brief History of ASR

• 1909: Universal service AT&T
• 1920s: Radio Rex
  – 500 Hz of energy in the word “Rex” caused the toy dog to move
• 1950s: Digit Recognition
  – 1952: Davis, Biddulph and Balashek (Bell Labs)
• Theory: 1967, Hidden Markov Models (HMMs) and Viterbi algorithm

• 1960s: Advances in Signal Processing and Neural Nets (not much progress in ASR)
• 1969: Advances in discrete word recognition
  – Vicens system (500 words)
  – Medress system (100 words)
• 1969: John Pierce letter
• 1970s: Despite large ARPA funding, not much success. Theory: dynamic programming methods
Brief History of ASR

• End of 1970s: Small vocabulary speech recognition
  – Heuristics’ $259 H-2000 Speech link
  – Verbex, Nippon, Threshold, Scott, Centigram and Interstate systems for between $2000 - $100,000
• Theory: 1977, the EM algorithm

2/20/08

Brief History of ASR

• 1980: IDA Symposium at Princeton
• 1980s: Discrete ASR, Language Models, corpus collection efforts
  – TIMIT corpus (phonetics)
  – ATIS corpus (Air Travel Information System)
  – Focus on language understanding dialog systems

2/20/08
Brief History of ASR

• 1990s: Large Vocabulary Continuous ASR
  – Dynamic Time Warping (edit distance)
  – Better phonetic models using classifiers (decision trees and neural nets)
  – Better language models using smoothing
  – Larger corpora: $10^7$ and $10^9$ in size

Brief History of ASR

• Current Work
  – Other languages and dialects
  – Multiple speakers, Speaker adaptation
  – Speaker identification
  – Noise resistant (telephone speech)
  – Open source software: HTK, Sphinx, CMU LM toolkit, SRI LM toolkit