

CMPT 413

Computational Linguistics

Anoop Sarkar

<http://www.cs.sfu.ca/~anoop>

2/9/07

1

Automatic Speech Recognition

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

2/9/07

2

Noisy Channel Model

- Finding the best transcription \mathbf{w}^* given an observation sequence \mathbf{o}

$$\begin{aligned}\mathbf{w}^* &= \arg \max_{\mathbf{w}} P(\mathbf{w} | \mathbf{o}) = \arg \max_{\mathbf{w}} \frac{P(\mathbf{o} | \mathbf{w})P(\mathbf{w})}{P(\mathbf{o})} \\ &= \arg \max_{\mathbf{w}} \underbrace{P(\mathbf{o} | \mathbf{w})}_{\substack{\text{generative} \\ \text{model}}} \underbrace{P(\mathbf{w})}_{\substack{\text{language} \\ \text{model}}}\end{aligned}$$

2/9/07

3

Generative Models of Speech

- Speech recognition:** find word sequence \mathbf{w} that maximizes $P(\mathbf{w} | \mathbf{o})$, where \mathbf{o} is a sequence of time dependent acoustic features (output of signal processing on speech signal)
- Typical decomposition of $P(\mathbf{w} | \mathbf{o})$ into a cascade of generative models:
 - Acoustic Model:
 $P(\mathbf{o} | \mathbf{p})$ predict observation sequence \mathbf{o} given phone sequence \mathbf{p}
 - Pronunciation Model:
 $P(\mathbf{p} | \mathbf{w})$ predict phone sequence \mathbf{p} given a word sequence \mathbf{w}
 - Language Model:
 $P(\mathbf{w})$ predict word sequence \mathbf{w}

2/9/07

4

Generative Models of Speech

- $P(\mathbf{w} \mid \mathbf{o}) = P(\mathbf{o} \mid \mathbf{w}) * P(\mathbf{w})$ using Bayes Rule
 - Decomposition of $P(\mathbf{o} \mid \mathbf{w})$ into a cascade of models:
 - Acoustic Model** $P(\mathbf{o} \mid \mathbf{p})$ (model trained on the TIMIT corpus):
 - Pronunciation Model** $P(\mathbf{w} \mid \mathbf{p})$ (model trained using TIMIT and the CMU pronunciation dictionary):
 - Language Model**: $P(\mathbf{w})$ (model trained using large amounts of text in the same domain)
- cf. *Fundamentals of Speech Recognition*, Rabiner and Juang

2/9/07

5

Generative Models of Speech

- Further decomposition of the acoustic model: $P(\mathbf{o} \mid \mathbf{p})$
 - $P(\mathbf{o} \mid \mathbf{d})$ observation vectors given distribution sequences (quantitative given symbolic)
 - $P(\mathbf{d} \mid \mathbf{m})$ distribution sequences given model sequences (model dependent phone sequences)
 - $P(\mathbf{m} \mid \mathbf{p})$ model sequences given phone sequences

2/9/07

6

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

2/9/07

7

Brief History of ASR

- 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

2/9/07

8

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

2/9/07

9

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

2/9/07

10

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

2/9/07

11

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

2/9/07

12