Automatic Speech Recognition

Instructor: Wei Xu
Ohio State University

Many Slides from Julia Hirschberg
Recreating the Speech Chain

SPOKEN LANGUAGE UNDERSTANDING

DIALOG MANAGEMENT

SPEECH RECOGNITION

SPEECH SYNTHESIS

DIALOG

SEMANTICS

SYNTAX

LEXICON

MORPHOLOGY

PHONETICS

INNER EAR
ACOUSTIC NERVE

VOCAL-TRACT ARTICULATORS

SPEECH RECOGNITION

SPEECH SYNTHESIS

DIALOG MANAGEMENT

INNER EAR
ACOUSTIC NERVE

VOCAL-TRACT ARTICULATORS
Speech Recognition: the Early Years

• 1952 – Automatic Digit Recognition (AUDREY)
  – Davis, Biddulph, Balashek (Bell Laboratories)
Speech Recognition: the Early Years

- 1952 – Automatic Digit Recognition (AUDREY)
  - Davis, Biddulph, Balashek (Bell Laboratories)
1960’s – Speech Processing and Digital Computers

- AD/DA converters and digital computers start appearing in the labs

James Flanagan
Bell Laboratories
The Illusion of Segmentation... or...
The Illusion of Segmentation... or...
Why Speech Recognition is so Difficult
The Illusion of Segmentation... or...

Why Speech Recognition is so Difficult
The Illusion of Segmentation... or...

Why Speech Recognition is so Difficult
The Illusion of Segmentation... or...

Why Speech Recognition is so Difficult
The Illusion of Segmentation... or...

Why Speech Recognition is so Difficult

(user:Roberto (attribute:telephone-num value:7360474))
The Illusion of Segmentation... or...
Why Speech Recognition is so Difficult
The Illusion of Segmentation... or...

Why Speech Recognition is so Difficult

Ellipses and Anaphors
Limited vocabulary
Multiple Interpretations
Speaker Dependency
Word variations
Word confusability
Context-dependency
Coarticulation
Noise/reverberation
Intra-speaker variability
General purpose speech recognition seems far away. Social-purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish…

It would be too simple to say that work in speech recognition is carried out simply because one can get money for it. That is a necessary but not sufficient condition. We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn’t attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamour…

Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve “the problem.” The basis for this is either individual inspiration (the “mad inventor” source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach).

The Journal of the Acoustical Society of America, June 1969
1969 – Whither Speech Recognition?

General purpose speech recognition seems far away. Social-purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish…

It would be too simple to say that work in speech recognition is carried out simply because one can get money for it. That is a necessary but not sufficient condition. We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn’t attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamour…

Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve “the problem.” The basis for this is either individual inspiration (the “mad inventor” source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach).

The Journal of the Acoustical Society of America, June 1969
1971-1976: The ARPA SUR project

- Despite anti-speech recognition campaign led by Pierce Commission ARPA launches 5 year Spoken Understanding Research program
- Goal: 1000-word vocabulary, 90% understanding rate, near real time on 100 mips machine
- 4 Systems built by the end of the program
  - SDC (24%)
  - BBN’s HWIM (44%)
  - CMU’s Hearsay II (74%)
  - CMU’s HARPY (95% -- but 80 times real time!)
- Rule-based systems except for Harpy
  - Engineering approach: search network of all the possible utterances
1971-1976: The ARPA SUR project

- Despite anti-speech recognition campaign led by Pierce Commission ARPA launches 5 year Spoken Understanding Research program
- Goal: 1000-word vocabulary, 90% understanding rate, near real time on 100 mips machine
- 4 Systems built by the end of the program
  - SDC (24%)
  - BBN’s HWIM (44%)
  - CMU’s Hearsay II (74%)
  - CMU’s HARPY (95% -- but 80 times real time!)
- Rule-based systems except for Harpy
  - Engineering approach: search network of all the possible utterances

LESSON LEARNED:
Hand-built knowledge does not scale up
Need of a global “optimization” criterion

• Lack of clear evaluation criteria
  – ARPA felt systems had failed
  – Project not extended
• Speech Understanding: too early for its time
• Need a standard evaluation method

DARPA was deeply disappointed with researchers working on the Speech Understanding Research program at Carnegie Mellon University. DARPA had hoped for, and felt it had been promised, a system that could respond to voice commands from a pilot. The SUR team had developed a system which could recognize spoken English, but only if the words were spoken in a particular order. DARPA felt it had been duped and, in 1974, they cancelled a three million dollar a year grant.[24]

Many years later, successful commercial speech recognition systems would use the technology developed by the Carnegie Mellon team (such as hidden Markov models) and the market for speech recognition systems would reach $4 billion by 2001.[25]
1970’s – Dynamic Time Warping
The Brute Force of the Engineering Approach

T.K. Vyntsyuk (1968)
H. Sakoe,
S. Chiba (1970)
1970’s – Dynamic Time Warping
The Brute Force of the Engineering Approach

T.K. Vyntsyuk (1968)
H. Sakoe,
S. Chiba (1970)

Isolated Words
Speaker Dependent

Connected Words
Speaker Independent

Sub-Word Units
1980s -- The Statistical Approach

- Based on work on Hidden Markov Models done by Leonard Baum at IDA, Princeton in the late 1960s
- Purely statistical approach pursued by Fred Jelinek and Jim Baker, IBM T.J.Watson Research
- Foundations of modern speech recognition engines

\[ \hat{W} = \arg \max_{W} P(A \mid W)P(W) \]

Acoustic HMMs

Word Tri-grams

\[ P(w_t \mid w_{t-1}, w_{t-2}) \]
1980s -- The Statistical Approach

- Based on work on Hidden Markov Models done by Leonard Baum at IDA, Princeton in the late 1960s
- Purely statistical approach pursued by Fred Jelinek and Jim Baker, IBM T.J. Watson Research
- Foundations of modern speech recognition engines

\[ \hat{W} = \arg \max_W P(A \mid W)P(W) \]

- Acoustic HMMs
- Word Tri-grams

- No Data Like More Data
- Whenever I fire a linguist, our system performance improves (1988)
- Some of my best friends are linguists (2004)
1980-1990 – Statistical approach becomes ubiquitous


Pros and Cons of DARPA programs

+ Continuous incremental improvement
- Loss of “bio-diversity”
State of the Art before Deep Learning

• Low noise conditions
• Large vocabulary
  – ~20,000-60,000 words or more…
• Speaker independent (vs. speaker-dependent)
• Continuous speech (vs isolated-word)
• Multilingual, conversational
• World’s best research systems:
  • Human-human speech: ~13-20% Word Error Rate (WER)
  • Human-machine or monologue speech: ~3-5% WER
Building an ASR System

• Build a statistical model of the speech-to-words process
  – Collect lots of speech and transcribe all the words
  – Train the model on the labeled speech

• Paradigm:
  – Supervised Machine Learning + Search
  – The Noisy Channel Model
The Noisy Channel Model

- Search through space of all possible sentences.
- Pick the one that is most probable given the waveform
The Noisy Channel Model (II)

• What is the most likely sentence out of all sentences in the language L, given some acoustic input O?

• Treat acoustic input O as sequence of individual acoustic observations
  – \( O = o_1, o_2, o_3, \ldots, o_t \)

• Define a sentence as a sequence of words:
  – \( W = w_1, w_2, w_3, \ldots, w_n \)
Noisy Channel Model (III)

• Probabilistic implication: Pick the highest probable sequence:

\[ \hat{W} = \arg \max_{W \in L} P(W | O) \]

• We can use Bayes rule to rewrite this:

\[ \hat{W} = \arg \max_{W \in L} \frac{P(O | W)P(W)}{P(O)} \]

• Since denominator is the same for each candidate sentence \( W \), we can ignore it for the argmax:

\[ \hat{W} = \arg \max_{W \in L} P(O | W)P(W) \]
Speech Recognition Meets Noisy Channel: Acoustic Likelihoods and LM Priors
Components of an ASR System

• Corpora for training and testing of components
• Representation for input and method of extracting
• Pronunciation Model
• Acoustic Model
• Language Model
• Feature extraction component
• Algorithms to search hypothesis space efficiently
Speech Recognition (HMM+GMM)
Training and Test Corpora

• Collect corpora appropriate for recognition task at hand
  – Small speech + phonetic transcription to associate sounds with symbols (Acoustic Model)
  – Large (>= 60 hrs) speech + orthographic transcription to associate words with sounds (Acoustic Model)
  – Very large text corpus to identify ngram probabilities or build a grammar (Language Model)
Building the Acoustic Model

• Goal: Model likelihood of sounds given spectral features, pronunciation models, and prior context
• Usually represented as Hidden Markov Model
  – States represent phones or other subword units
  – Transition probabilities on states: how likely is it to see one sound after seeing another?
  – Observation/output likelihoods: how likely is spectral feature vector to be observed from phone state i, given phone state i-1?
Word HMM
Speech Recognition (HMM+GMM)

Transcription: Samson
Sub-phones: 942 – 6 – 37 – 8006 – 4422 ...

Hidden Markov Model (HMM):
- Transition from 942 to 942
- Transition from 942 to 6

Acoustic Model:
- Three acoustic models for each sub-phone

Audio Input:
- Features for each sub-phone

GMM models: P(x|s)
x: input features
s: HMM state
Building the Pronunciation Model

- Models likelihood of word given network of candidate phone hypotheses
  - Multiple pronunciations for each word
  - May be weighted automaton or simple dictionary
- Words come from all corpora (including text)
- Pronunciations come from pronouncing dictionary or TTS system
ASR Lexicon: Markov Models for Pronunciation
Building the Language Model

• Models likelihood of word given previous word(s)
• Ngram models:
  – Build the LM by calculating bigram or trigram probabilities from text training corpus: how likely is one word to follow another? To follow the two previous words?
  – Smoothing issues
• Grammars
  – Finite state grammar or Context Free Grammar (CFG) or semantic grammar
• Out of Vocabulary (OOV) problem
Search/Decoding

• Find the best hypothesis $P(O|W) P(W)$ given
  – A sequence of acoustic feature vectors ($O$)
  – A trained HMM (AM)
  – Lexicon (PM)
  – Probabilities of word sequences (LM)

• For $O$
  – Calculate most likely state sequence in HMM given transition and observation probs
  – Trace back thru state sequence to assign words to states
  – N best vs. 1 best vs. lattice output

• Limiting search
  – Lattice minimization and determinization
  – Pruning: beam search
Evaluating Success

• Transcription
  – Low WER (Subst+Ins+Del)/N * 100
    Thesis test vs. This is a test. 75% WER
    Or That was the dentist calling. 125% WER

• Understanding
  – High concept accuracy
    • How many domain concepts were correctly recognized?
      I want to go from Boston to Baltimore on September 29
Domain concepts | Values
---|---
source city | Boston
target city | Baltimore
travel date | September 29
Score recognized string | “Go from Boston to Washington on December 29” vs. “Go to Boston from Baltimore on September 29”
(1/3 = 33% CA)
Summary

• ASR today
  – Combines many probabilistic phenomena: varying acoustic features of phones, likely pronunciations of words, likely sequences of words
  – Relies upon many approximate techniques to ‘translate’ a signal
  – Finite State Transducers

• ASR future
  – Can we include more language phenomena in the model?