Fundamentals of Automatic Speech Recognition

Britta Wrede
Gernot A. Fink
Applied Computer Science Group, Bielefeld University

July 2005
Fundamentals of Automatic Speech Recognition

Britta Wrede
Gernot A. Fink
Applied Computer Science Group, Bielefeld University

July 2005

• Introduction
  Why use speech recognition?
• Statistical Speech Recognition
  General Framework
• Feature Extraction
  Short-Time Analysis
• Acoustic Modeling
  Hidden Markov Models
• Language Modeling
  n-Gram Models
• Summary
Motivation

Application Areas for Automatic Speech Recognition (ASR)

- telephone-based information systems
- dictating machine
- control of machines e.g. for medical applications (OP)
- long-term vision: interaction with robots
- related application areas:
  - speech therapy
  - (second) language acquisition
Introduction

Why automatic speech recognition?

Spoken speech is:

- natural method of interaction for humans
- important modality in human-human communication
- efficient and easy to use
- ...and requires little/no additional training
Why is speech recognition difficult?

**Complexity**

- high data rate (16,000+ samples/second, 100+ words/minute)
- large inventory of units (≈ 50 phones, 100,000+ words)

**Variability**

- production of sounds influenced by context (coarticulation/assimilation)
- between different speakers, however: even for single speaker!
  (speaker dependent vs. independent)
- due to speaking style (controlled, formal, spontaneous)
- with respect to recording environment/equipment
  (close talking microphone, quite office room, driving car, ...)

**Continuity**

no segment boundaries present between phones, words
(isolated word recognition vs. continuous speech speech recognition)
Application Areas

- **Voice command systems / numbers recognition**
  e.g. in cars, for telephony-based services
  (small vocabulary 2-100, speaker independent, isolated words / short, well defined phrases, robust to noise)
  Error Rate < 5%

- **Dictation systems**
  e.g. for physicians or lawyers, also private users
  (large vocabulary 10,000-100,000, speaker dependent, controlled speech, “sensitive”)
  Error Rate 5 – 10%

- **Research systems**
  (average to large vocabulary 3,000-20,000, speaker independent, spontaneous speech, adaptive)
  Error Rate 15 – 50%
Theory: Channel Model

Diagram:

- LINGUISTIC SOURCE
  - text production
  - $P(w)$

- ACOUSTIC CHANNEL
  - word articulation
  - feature extraction
  - $P(X|w)$

- SPEECH RECOGNITION
  - model decoding
  - $\hat{w}$
  - $\arg\max_w P(w|X)$
**Theory:** Channel Model

- 2 components: “acoustic” model $P(X|w)$ & language model $P(w)$
- Assumption: strong relation between articulation and acoustics
Feature Extraction: description of relevant characteristics of the signal

⇒ short-time analysis (Mel-Cepstrum)
Modeling for Speech Recognition

**Feature Extraction:** description of relevant characteristics of the signal

⇒ short-time analysis (Mel-Cepstrum)

**Acoustic Modeling:** description of acoustic units, e.g. speech sounds, words

⇒ Hidden Markov Models ≈ statistical pattern matching
Modeling for Speech Recognition

**Feature Extraction:** description of relevant characteristics of the signal

⇒ short-time analysis (Mel-Cepstrum)

**Acoustic Modeling:** description of acoustic units, e.g. speech sounds, words

⇒ Hidden Markov Models ≈ statistical pattern matching

**Language Modeling:** restriction of potential word sequences using e.g.

- formal grammars

  valid vs. invalid

- stochastic grammars

  likely … unlikely vs. invalid

- “purely statistical”: calculation of $P(w)$

⇒ $n$-gram models
Feature Extraction

Short-Time Analysis
parametric representation of short speech segments (approx. 10-30 ms)

Assumption: characteristic (= spectral?) features are stationary within segments

Most widely used method: spectral analysis $\rightarrow$ Mel-Cepstrum

- warping of the frequency axis similar to human hearing (filter bank)
- separation of coarse and fine structure of the log-power spectrum

$$\text{signal} \rightarrow \text{DFT} \rightarrow |.| \rightarrow \text{Mel} \rightarrow \log \rightarrow \text{DCT} \rightarrow \text{Mel-Cepstrum}$$

Dynamic Features:
capture spectral variations by calculating time derivatives
• ‘windowing’ of signal (10-30 ms)

• computation of cepstrum containing:
  – coarse spectral structure (slope, formants)
  – spectral fine structure (jitter, shimmer, harmonics)

• removal of spectral fine structure
Feature Extraction: Dynamic Features

Dynamic Features contain:

- contain acoustic changes (e.g. of formants) and thus articulatory movements over time

- are computed as 1st and 2nd order derivatives over time
Summary: Feature Extraction

Every 10 ms a 39-dimensional feature vector is computed:

- 12 static MFCCs + 1 energy
- 13 first order derivatives
- 13 second order derivatives
Hidden Markov Models (HMM)

What units should be modelled?

- **Phonemes**, syllables, words...

- Phonemes are too variable due to coarticulation

- Triphones = phonemes in context:
  - capture coarticulation
  - while keeping non-variable information of phoneme

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phonemes</th>
<th>Triphones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisch</td>
<td>fIS</td>
<td>#/f/I f/I/S I/S/#</td>
</tr>
<tr>
<td>Kit</td>
<td>klt</td>
<td>#/k/I k/I/t I/t/#</td>
</tr>
</tbody>
</table>
models for complete words (i.e. inflected forms) can generally not be used
⇒ smaller **sub-word units**

- models for speech sounds (“phoneme models”)
  usually *linear* models, 3–6 states for “phases”

- models for groups of sounds (e.g. for syllables or words)

- context dependent (phoneme) models
  usually tri-phones, e.g. **p/i/t** in */spitS*/

  ✓ very flexible, can easily be combined

  ⌈ trainability ⇒ generalization necessary!

- speech pauses also need to be modeled!
Acoustic Modeling: Model Structure

**Goal:** segmentation

- segmentation units = words ...
- ... represented as sequence of phoneme models (i.e. states)
- lexicon = set of words to recognize (also: phonetic prefix-tree)
- utterance = arbitrary sequence of words from the lexicon

⇒ decoding the model produces segmentation (i.e. determining the optimal state/model sequence)
Hidden Markov Models (HMM)

How should units be modelled? ⇒ HMMs

- HMM consists of states and transitions
- each state describes a (hopefully) stationary phase of a phoneme
- emission-probabilities describe acoustic features of this phase
- transition-probabilities describe temporal structure of phoneme

**Diagram:**

- persevering coarticulation
- stationary phase
- anticipatory coarticulation
Hidden Markov Models

How can emission- and transition-probabilities be estimated?

- initial segmentation of training data into phonemes needed
- assignment of speech samples (= feature vectors) to triphone states
- computation of statist. parameters from feature vectors (e.g. mean, variance)
A 1st order Hidden Markov Model $\lambda$ is defined by:

- a finite set of states $\{s | 1 \leq s \leq N \}$

- a matrix of state transition probabilities
  $$A = \{a_{ij}|a_{ij} = P(s_t = j | s_{t-1} = i)\}$$

- a vector of initial state probabilities
  $$\pi = \{\pi_i|\pi_i = P(s_1 = i)\}$$

- and state specific emission probability distributions
  $$\{b_j(O_t)|b_j(O_t) = p(O_t|s_t = j)\}$$
Hidden Markov Models

*How can HMMs be applied for pattern recognition?*
How can HMMs be applied for pattern recognition?

Assumption:

patterns (e.g. speech signals) are generated by a stochastic model with \textit{principally equivalent} behavior!
Hidden Markov Models

How can HMMs be applied for pattern recognition?

Assumption:
patterns (e.g. speech signals) are generated by a stochastic model with princi-
pally equivalent behavior!

Evaluation: determining quality of modeling
→ calculate production probability $P(O|\lambda)$
Hidden Markov Models

How can HMMs be applied for pattern recognition?

Assumption:
patterns (e.g. speech signals) are generated by a stochastic model with *principally equivalent* behavior!

Evaluation: determining quality of modeling
→ calculate production probability $P(O|\lambda)$

Decoding: uncovering the “internal structure” of the model (≡ “recognition”)
→ determine optimal state sequence $s^* = \arg\max_s P(O, s|\lambda)$
Hidden Markov Models

How can HMMs be applied for pattern recognition?

Assumption: patterns (e.g. speech signals) are generated by a stochastic model with principally equivalent behavior!

Evaluation: determining quality of modeling

→ calculate production probability $P(O|\lambda)$

Decoding: uncovering the “internal structure” of the model (≡ “recognition”)

→ determine optimal state sequence $s^* = \arg\max_s P(O, s|\lambda)$

Training: creating the “optimal” model

→ improve a given model $\lambda$ so that $P(O|\hat{\lambda}) \geq P(O|\lambda)$
Hidden Markov Models: Other applications

- recognition of phoneme quality e.g. for language acquisition: how well does the spoken utterance map the target utterance?

- visualisation of articulatory features in spoken utterance

- could also be used for intonation recognition and emotion recognition
Hidden Markov Models: Summary

✓ parameters can be estimated automatically from training samples (e.g. pre-recorded utterances)

✓ models “capture” substantial amount of variation in realization and duration

⚡ for robust, large vocabulary, speaker independent systems considerable amounts of training data necessary (several hours of speech data)

⚡ model configurations have to be specified by experts (i.e. number of mixture densities and model states, type and structure of sub-word units)
Overview

LINGUISTIC SOURCE: text production

ACOUSTIC CHANNEL: word articulation, feature extraction

SPEECH RECOGNITION: model decoding

\[ P(w) \rightarrow P(X|w) \rightarrow \text{argmax } P(w|X) \]
Why Language Modeling?

Typical Speech Recognition problems:

- They are leaving in about fifteen **minuets** to go to her house.
- The study was conducted mainly **be** John Black.
- The design **an** construction of the system will take more than a year.
- Hopefully, all **with** continue smoothly in my absence.
- Can they **lave** me a message?
- I need to **notified** the bank of this problem.
- He is trying to **fine** out.
Why Language Modeling?

- acoustic cues alone do not convey enough information
- human performance on speech recognition for unknown language is also not good

⇒ Use other information sources: Knowledge about which words are likely to occur together

⇒ Statistical solution: N-gram models
### What are N-gram models?

#### Example Bi-grams for: I want to eat dinner

<table>
<thead>
<tr>
<th>Tag</th>
<th>Word</th>
<th>Probability</th>
<th>Next Word</th>
<th>Probability</th>
<th>Next Word</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;S&gt;</td>
<td>I</td>
<td>0.25</td>
<td>I</td>
<td>0.32</td>
<td>want</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>want</td>
<td>0.32</td>
<td>to</td>
<td>0.65</td>
<td>to eat</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>to eat</td>
<td>0.26</td>
<td>dinner</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;S&gt;</td>
<td>I'd</td>
<td>0.06</td>
<td>I</td>
<td>0.29</td>
<td>would</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>would</td>
<td>0.29</td>
<td>want a</td>
<td>0.42</td>
<td>to</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>want a</td>
<td>0.42</td>
<td>to have</td>
<td>0.48</td>
<td>lunch</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>to have</td>
<td>0.48</td>
<td>eat lunch</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;S&gt;</td>
<td>Tell</td>
<td>0.04</td>
<td>I</td>
<td>0.08</td>
<td>don’t</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>don’t</td>
<td>0.08</td>
<td>want some</td>
<td>0.04</td>
<td>to spend</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>want some</td>
<td>0.04</td>
<td>to spend</td>
<td>0.09</td>
<td>e some</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>to spend</td>
<td>0.09</td>
<td>e some</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;S&gt;</td>
<td>I’m</td>
<td>0.02</td>
<td>I</td>
<td>0.04</td>
<td>have</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>have</td>
<td>0.04</td>
<td>want thai</td>
<td>0.14</td>
<td>be</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>want thai</td>
<td>0.14</td>
<td>to be</td>
<td>0.02</td>
<td>a</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>to be</td>
<td>0.02</td>
<td>eat a</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### N-grams:

- **Uni-gram:**
  - dinner
  - dinner

- **Bi-gram:**
  - W1 dinner
  - eat dinner

- **Tri-gram:**
  - W2 W1 dinner
  - to eat dinner

- **4-gram:**
  - W3 W2 W1 dinner
  - want to eat dinner
How to estimate N-grams

- select a corpus that represents your application area

- for every word in the lexicon count its occurrence in a Bi-gram context, e.g. eat:

| Bi-gram    | count | p(∗|eat) |
|------------|-------|---------|
| eat on     | 16    | .49     |
| eat some   | 6     | .18     |
| eat lunch  | 6     | .18     |
| eat dinner | 5     | .15     |

- compute probabilities p(W2|W1)
Overview

LINGUISTIC SOURCE

- Text production

ACOUSTIC CHANNEL

- Word articulation
- Feature extraction

SPEECH RECOGNITION

- Model decoding

\[ P(w) \quad \rightarrow \quad P(X|w) \quad \rightarrow \quad \arg\max_w P(w|X) \]
ESMERALDA: System Architecture

- Feature extraction
- Codebook evaluation
- Integrated path-search
- Language model design
- Heuristic methods
- Psycho-acoustic knowledge
- Vector quantisation
- HMM training
- P(z|x y)
- S → NP VP
  NP → N
- Linguistic knowledge
- Best word chain
Integrated Parsing and Recognition

Goal:

- use declarative grammar as a language model
  (especially useful for artificial domains with limited or no training data)
- apply grammatical restrictions robustly

Problems:

- grammar decisions are binary: valid vs. invalid utterance
- grammars decide about complete sentences

Solutions:

- use penalty scores for ungrammatical input
- allow for partial parses i.e. phrases or constituents
Integration of Speech Recognition & Understanding

Speech recognition

Speech understanding

Grammar

Acoustic model

P(w)

Statistical language model

Linguistic and pragmatic knowledge

Britta Wrede
Open Challenges for ASR

• open vocabulary (understanding of unknown words)

• ASR in noisy environments

• closer coupling with speech understanding and dialog context

• gather more information from speech signal that may be important:
  – prosodic information (F0, speech rate, articulation style..)
  – emotional state
References

Phonetics:


ASR and Language Modeling:
