DT2118
Speech and Speaker Recognition
Introduction

Giampiero Salvi

KTH/CSC/TMH giampi@kth.se

VT 2016
Outline

Course Organization

Introduction
    The Big Picture
    Challenges

Models of Speech Production
    Source/Filter Model: Vowel-like sounds
    Source/Filter Model, General Case
Outline

Course Organization

Introduction
   The Big Picture
   Challenges

Models of Speech Production
   Source/Filter Model: Vowel-like sounds
   Source/Filter Model, General Case
Contact Info

Giampiero Salvi (giampi@kth.se)

All communications handled through the course web:
https://www.kth.se/social/course/DT2118/
Course Objectives

after the course you should be able to:

- implement simple training and evaluation methods for Hidden Markov Models
- train and evaluate a speech recogniser using software packages
- compare different feature extraction and training methods
- document and discuss specific aspects related to speech and speaker recognition
- with the help of the literature, review and criticise other students’ work in the subject
Schedule

Part 1  Introduction, Speech Signal, Features, Statistics (ca 4 hours)

Part 2  Hidden Markov Models, Training and Decoding, Acoustic Models (ca 4-6 hours)

Part 3  Decoding and Search Algorithms (ca 2 hours)

Part 4  Language Models (Grammars) (ca 2 hours)

Part 5  Noise robustness and Speaker Recognition (ca 2-4 hours)
Literature

- **Spoken Language Processing: A Guide to Theory, Algorithm, and System Development**
  *Xuedong Huang, Alex Acero, Hsiao-Wuen Hon*, Prentice Hall
  - 3 at KTH library,
  - 6 at TMH library (against 300 SEK deposit)

- **Automatic Speech Recognition: A deep learning approach**
  *Dong Yu and Li Deng*, Springer 2015
  Available in PDF from SpringerLink (via KTH Biblioteket)

- **HTK manual** version 3.4

- selected research articles
## Reading Instructions (course book)

These are indicative, check the schedule for more updated instructions

<table>
<thead>
<tr>
<th>Part</th>
<th>Topic</th>
<th>Pages</th>
<th># pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part 1</td>
<td>(Spoken Language Structure)</td>
<td>(19–71)</td>
<td>(52)</td>
</tr>
<tr>
<td></td>
<td>Digital Signal Processing</td>
<td>(201–273)</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>Probability, Statistics and Inform. Theory</td>
<td>73–131</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>Pattern Recognition</td>
<td>133–197</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>Speech Signal Representations</td>
<td>275–336</td>
<td>62</td>
</tr>
<tr>
<td>Part 2</td>
<td>Hidden Markov Models</td>
<td>377–413</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Acoustic Modeling</td>
<td>415–475</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>Environmental Robustness</td>
<td>477–544</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>HTK tutorial (HTK book)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part 3</td>
<td>Basic Search Algorithms</td>
<td>591–643</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>(Large-Vocabulary Search Algorithms)</td>
<td>(645–685)</td>
<td>(41)</td>
</tr>
<tr>
<td></td>
<td>(Applications and User Interfaces)</td>
<td>(919–956)</td>
<td>(38)</td>
</tr>
<tr>
<td>Part 4</td>
<td>Language Modeling</td>
<td>545–590</td>
<td>46</td>
</tr>
<tr>
<td>Part 5</td>
<td>Speaker Recognition literature</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Optional chapters in parentheses)
Requirements/Activities

Grades: Pass/Fail
In order to pass you have to:

1. carry out three labs and hand in the report
2. carry out mini-project in groups and write report (paper)
3. review other students’ report
4. present your work at final seminar
5. discuss other student’s work at final seminar
6. submit final version of the paper
Lab 1: Speech Feature Extraction

➤ implement extraction for typical speech features
➤ analyse the features on speech data
➤ compare utterances with Dynamic Time Warping
Lab 2: Gaussian Hidden Markov Models

- implement the decoding algorithms for HMMs
- implement the training algorithms for HMMs
- test the algorithms on isolated digits
Lab 3: Continuous Speech Recognition and Deep Learning

- Extend the training and testing algorithms to continuous speech
- Test the algorithms on the TIDIGIT database (connected digits)
- Optional: implement DNNs using Theano, compare with GMM-HMMS
Project report

- Suggest a title or choose a topic from a list
- Project report in form of research paper
- Suggested topics:

| Own work and experiments after discussion with the teacher |
| Limitations in standard HMM and a survey of alternatives |
| Pronunciation variation and its importance for speech recognition |
| Language models for speech recognition |
| New search methods |
| Techniques for robust recognition of speech |
| Confidence measures in speech recognition |
| The role of prosody for speech recognition |
| Speaker variability and methods for adaptation |
Computational Resources at PDC

For Lab 3 and the Project

- apply for an account at https://www.pdc.kth.se/support/accounts/user
- use edu16.DT2118 when asked for time allocation
Important dates

Thu 14 April: submit Lab 1 report
Tue 19 April: decide topic for project
Thu 28 April: submit Lab 2 report
Thu 12 May: submit Lab 3 report
Sun 29 May: submit review of other report
Tue 31 May: Final seminar: present own project results, and discuss others’
Mon 6 Jun: Final report

KTH Social deadlines are set at 23:55
Part 1
Outline

Course Organization

Introduction
  The Big Picture
  Challenges

Models of Speech Production
  Source/Filter Model: Vowel-like sounds
  Source/Filter Model, General Case
Motivation

- Natural way of communication (No training needed)
- Leaves hands and eyes free (Good for functionally disabled)
- Effective (Higher data rate than typing)
- Can be transmitted/received inexpensively (phones)
The dream of Artificial Intelligence

2001: A space odyssey (1968)
A very long endeavour

1952, Bell laboratories, isolated digit recognition, single speaker, hardware based [1]

A very long endeavour

1952, Bell laboratories, isolated digit recognition, single speaker, hardware based [1]

An underestimated challenge:
60 years of bold announcements

Today’s Reality

I Now Pronounce You Chuck & Larry (2007)
The ASR Goal (for this course)

Convert speech into text

Automatic Speech Recognition

“My name is ...”
[confidence score]
The ASR Goal (for this course)

Convert speech into text

Automatic Speech Recognition

“My name is . . .”
[confidence score]

CC  Please tell me your name
LV  Larry Valentine
CC  I’m sorry, I didn’t quite get that
LV  Larry Valentine
CC  You said “Berry Schmallenpine” . . . is that right?
LV  Schmallenpine?!?!?
CC  You said “Schmallenpine” . . . is that right?
The ASR Goal (for this course)

Convert speech into text

Automatic Speech Recognition

“My name is . . .”
[confidence score]

CC Please tell me your name
LV Larry Valentine
CC I’m sorry, I didn’t quite get that
LV Larry Valentine
CC You said “Berry Schmallenpine” . . . is that right?
LV Schmallenpine?!?!?
CC You said “Schmallenpine” . . . is that right?
The ASR Goal (for this course)
Convert speech into text

Automatic Speech Recognition

“My name is . . .” [confidence score]

CC Please tell me your name
LV Larry Valentine
CC I’m sorry, I didn’t quite get that
LV Larry Valentine
CC You said “Berry Schmallenpine” . . . is that right?
LV Schmallenpine?!?!?
CC You said “Schmallenpine” . . . is that right?
The ASR Goal (for this course)
Convert speech into text

Automatic Speech Recognition

“My name is . . .”
[confidence score]

CC  Please tell me your name
LV  Larry Valentine
CC  I’m sorry, I didn’t quite get that
LV  Larry Valentine
CC  You said “Berry Schmallenpine” . . . is that right?
LV  Schmallenpine?!?!
CC  You said “Schmallenpine” . . . is that right?
ASR in a Broader Context

- Automatic Speech Recognition
- Dialogue Manager
- Spoken Language Understanding
- Text to Speech
The Speech Chain

Peter Denes, Elliot Pinson, 1963
ASR versus Computer Vision

SCENE

OBSERVER

Optical level

Physiological level

Linguistic level
### ASR versus Computer Vision

<table>
<thead>
<tr>
<th>Property</th>
<th>ASR</th>
<th>Computer Vision</th>
</tr>
</thead>
<tbody>
<tr>
<td>signal originates from:</td>
<td>cognition + physics</td>
<td>physics</td>
</tr>
<tr>
<td>persistence:</td>
<td>disappears as soon as heard</td>
<td>continually available (active perception)</td>
</tr>
<tr>
<td>across countries:</td>
<td>different languages</td>
<td>same objects</td>
</tr>
<tr>
<td>type of interaction:</td>
<td>two-way</td>
<td>one-way</td>
</tr>
</tbody>
</table>
The Speech Chain (from the book)
Not covered in this course:

- multimodality
- interaction (bi-directional)
- incrementality
- non-verbal communication
Main variables in ASR

- Speaking mode: isolated words vs continuous speech
- Speaking style: read speech vs spontaneous speech
- Speakers: speaker dependent vs speaker independent
- Vocabulary: small (<20 words) vs large (>50,000 words)
- Robustness: against background noise
Challenges — Variability

**Between speakers**
- Age
- Gender
- Anatomy
- Dialect

**Within speaker**
- Stress
- Emotion
- Health condition
- Read vs Spontaneous
- Adaptation to environment (Lombard effect)
- Adaptation to listener

**Environment**
- Noise
- Room acoustics
- Microphone distance
- Microphone, telephone
- Bandwidth

**Listener**
- Age
- Mother tongue
- Hearing loss
- Known / unknown
- Human / Machine
Example: spontaneous vs hyper-articulated

Va jobbaru me

Vad jobbar du med

“What is your occupation”
(“What work you with”)

32 / 54
### Examples of reduced pronunciation

<table>
<thead>
<tr>
<th>Spoken</th>
<th>Written</th>
<th>In English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesempel</td>
<td>Till exempel</td>
<td>for example</td>
</tr>
<tr>
<td>åhamba</td>
<td>och han bara</td>
<td>and he just</td>
</tr>
<tr>
<td>bafatt</td>
<td>bara för att</td>
<td>just because</td>
</tr>
<tr>
<td>javende</td>
<td>jag vet inte</td>
<td>I don’t know</td>
</tr>
</tbody>
</table>
Microphone distance

Headset

2 m distance
Applications today

Call centers:
- traffic information
- time-tables
- booking...

Accessibility
- Dictation
- hand-free control (TV, video, telephone)

Smart phones
- Siri, Android...
Course Organization

Introduction
  The Big Picture
  Challenges

Models of Speech Production
  Source/Filter Model: Vowel-like sounds
  Source/Filter Model, General Case
Speech Examples

TIMIT database (American English)

example of “clean” speech
Elements of Signal Processing

- continuous/digital signals
- Linear and Time Invariant (LTI) systems
- impulse response and convolution
- Fourier transform and transfer function
- sampling theorem
- short-time Fourier transform

(Chapter 5 in the book)
live examples
Physiology

- Soft palate (velum)
- Hard palate
- Nasal cavity
- Nostril
- Lip
- Tongue
- Teeth
- Oral cavity
- Jaw
- Trachea
- Lung
- Diaphragm
- Pharyngeal cavity
- Larynx
- Esophagus
- Vocal folds
- Glottis
- Muscle Force and Relaxation

Diagram showing the physiological pathways involving nasal cavity, nostrils, lip, tongue, teeth, oral cavity, jaw, trachea, lungs, and diaphragm. Connections are made to the nasal cavity, oral cavity, and trachea, indicating muscle force for relaxation.
Source/Filter Model, Vowel-like sounds

Vowels

- Source (periodic)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)
Glottal Flow

\[ G(z) = \frac{1}{(1 - \beta z)^2}, \quad \beta < 1 \]
Radiation form the Lips/Nose

Problem of radiation at the lips plus diffraction about the head too complicated.
Radiation form the Lips/Nose

Approx. with a piston in a rigid sphere: solved but not in closed form
Radiation from the Lips/Nose

2nd approx: piston in an infinite wall

\[ R(z) \approx 1 - \alpha z^{-1} \]
Tube Model of the Vocal Tract
Tube Model (cntd.)

▶ assume planar wave propagation and lossless tubes
▶ solve pressure \( p(x, t) \) and velocity \( u(x, t) \) in each tube according to wave equation
▶ impose continuity of pressure and velocity at the junctions

⇒ all-pole transfer function (\( N = \) number of tubes)

\[
V(z) = \frac{A z^{-N/2}}{1 - \sum_{k=1}^{N} a_k z^{-k}}
\]
 assume planar wave propagation and lossless tubes

solve pressure $p(x, t)$ and velocity $u(x, t)$ in each tube according to wave equation

impose continuity of pressure and velocity at the junctions

⇒ all-pole transfer function ($N =$ number of tubes)

$$V(z) = \frac{A z^{-N/2}}{1 - \sum_{k=1}^{N} a_k z^{-k}}$$
Source/Filter Model: vowel-like sounds
Source/Filter Model: vowel-like sounds

\[ p[n] \]

- Waveform
- Spectrum (log)

(time (msec))
(freqency (kHz))
Source/Filter Model: vowel-like sounds

\[ p[n] \]

\[ p[n] \ast g[n] \]
Source/Filter Model: vowel-like sounds

\[ p[n] \]

\[ p[n] \ast g[n] \]

\[ p[n] \ast g[n] \ast r[n] \]
Source/Filter Model: vowel-like sounds

\[ p[n] \]

\[ p[n] \ast g[n] \]

\[ p[n] \ast g[n] \ast r[n] \]

\[ p[n] \ast g[n] \ast r[n] \ast v[n] \]
$F_0$ and Formants

- **Varying $F_0$ (vocal fold oscillation rate)**

![Graph showing spectrum (log) for $f_0 = 100$Hz and $f_0 = 250$Hz with varying formants for vowel $\varepsilon$ and $u$.](image_url)
$F_0$ and Formants

- **Varying $F_0$ (vocal fold oscillation rate)**

  ![Spectrum](image1.png)

- **Varying Formants (vocal tract shape)**

  ![Spectrum](image2.png)
Source/Filter Model, General Case

Vowels

- Source (periodic)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)
Source/Filter Model, General Case

Fricatives (e.g. sh) or Plosive (e.g. k)

- Soft palate (velum)
- Hard palate
- Nasal cavity
- Nostril
- Lip
- Tongue
- Teeth
- Oral cavity
- Jaw
- Trachea
- Lung
- Esophagus
- Pharyngeal cavity
- Larynx

- Source (noise or impulsive)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)
Fricatives (e.g. s) or Plosive (e.g. t)
Source/Filter Model, General Case

Nasalised Vowels

- **Source (periodic)**
- **Front Cavity**
- **Back Cavity**
- **Back Cavity (2nd approx.)**

Diagram showing various anatomical structures including:
- Soft palate (velum)
- Hard palate
- Nasal cavity
- Nostril
- Lip
- Tongue
- Teeth
- Oral cavity
- Jaw
- Trachea
- Lung
- Diaphragm
- Esophagus
- Larynx
- Pharyngeal cavity
Source/Filter Model: fricative sounds

\[ p[n] \]

\[ p[n] \ast r[n] \]

\[ p[n] \ast r[n] \ast v[n] \]
Complete Source/Filter Model
THE INTERNATIONAL PHONETIC ALPHABET (2005)

CONSONANTS (PULMONIC)

<table>
<thead>
<tr>
<th>IPA</th>
<th>LABIAL</th>
<th>CORONAL</th>
<th>DORSAL</th>
<th>RADICAL</th>
<th>LARYNGEAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bilabial</td>
<td>Labiodental</td>
<td>Dental</td>
<td>Alveolar</td>
<td>Palato-alveolar</td>
</tr>
<tr>
<td>Nasal</td>
<td>m</td>
<td>m</td>
<td>n</td>
<td>η</td>
<td>η</td>
</tr>
<tr>
<td>Plosive</td>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
<td>t</td>
</tr>
<tr>
<td>Fricative</td>
<td>f</td>
<td>v</td>
<td>θ</td>
<td>δ</td>
<td>s</td>
</tr>
<tr>
<td>Approximant</td>
<td>u</td>
<td>u</td>
<td>l</td>
<td>j</td>
<td>w</td>
</tr>
<tr>
<td>Trill</td>
<td>B</td>
<td>r</td>
<td>R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tap, Flap</td>
<td>v</td>
<td>r</td>
<td>l</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral fricative</td>
<td></td>
<td>l</td>
<td>r</td>
<td>l</td>
<td>k</td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>l</td>
<td>l</td>
<td>x</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>Lateral flap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured ŋ. Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial extensions of the IPA.
IPA Chart: Vowels

THE INTERNATIONAL PHONETIC ALPHABET (2005)

VOWELS

Front    Near front    Central    Near back    Back

Close   i       y       i       u       u       u
Near close   I   • Y   I   • ð   ñ   ð
Close mid  e   ø       è       ð       ð       ø
Mid   e   ø       è       ð       ð       ø
Open mid  ε   øε   ™   øε   ™   øε
Near open  æ   æ   æ   æ   æ   æ
Open  a   øæ   ææ   ææ   ææ   ææ

Vowels at right & left of bullets are rounded & unrounded.
Phonology vs Phonetics

Phonemes

- co-articulation
- speaking style
- dialogue
- reduction
- assimilation
- speaker differences
- environment
  (loudness, channel, room acoustics, noise)

Phones
Phonology vs Phonetics

Phonemes

co-articulation, speaking style, dialogue, reduction, assimilation, speaker differences, environment (loudness, channel, room acoustics, noise)

Words

Phones

Sounds
Components of ASR System

Speech Signal → Spectral Analysis → Feature Extraction

Representation

Constraints - Knowledge

Acoustic Models → Lexical Models → Language Models

Search and Match

Decoder

Recognised Words