Speech Signal Processing

柏林陳 2003
Course Contents

• Both the theoretical and practical issues for spoken language processing will be considered
• Technology for **Automatic Speech Recognition (ASR)** will be further emphasized
• Topics to be covered
  – Statistical Modeling Paradigm
    • Spoken Language Structure
    • Hidden Markov Models
    • Speech Signal Analysis and Feature Extraction
    • Acoustic and Language Modeling
    • Search/Decoding Algorithms
  – Systems and Applications
    • Keyword Spotting, Dictation, Speaker Recognition, Spoken Dialogue, Speech Information Retrieval etc.
Textbook and References

• Textbook:

• References:
Grading

• Midterm or Final: 30%
• Exercises: 30%
• Project: 15%
• Attendance/Presentation/Report: 20%
• Others: 5%
Introduction

References:
Historical Review

1952, Isolated-Digit Recognition, Bell Lab.

1956, Ten-Syllable Recognition, RCA

1959, Ten-Vowel Recognition, MIT Lincoln Lab

1959, Phoneme-sequence Recognition using Statistical Information of context, Fry and Denes

1960s, Dynamic Time Warping to Compare Speech Events, Vintsyuk

1960s-1970s, Hidden Markov Models for Speech Recognition, Baum, Baker and Jelinek

1970s ~

Voice-Activated Typewriter (dictation machine, speaker-dependent), IBM

Telecommunication (keyword spotting, speaker-independent), Bell Lab

SRI

BBN Technologies

Speech at CMU

LIMSI

MIT SLS

Cambridge HTK

Microsoft

Philips

Gestation of Foundations
Progress of Technology

• US. National Institute of Standards and Technology (NIST)

http://www.nist.gov/speech/
Progress of Technology

- Generic Application Areas (vocabulary vs. speaking style)
Progress of Technology

- Benchmarks of ASR performance: Overview
Progress of Technology

- Benchmarks of ASR performance: Broadcast News Speech
Progress of Technology

• Benchmarks of ASR performance: Conversational Speech

Figure 4: History of lowest word error rates (WER) obtained in NIST conversational speech evaluations on Switchboard and CallHome type conversations in English [26].

Figure 5: Chinese Character error rates of the best performing evaluation system in NIST Mandarin conversational speech evaluations 1995-2000 [26].
Determinants of Speech Communication

Speech Generation

Message Formulation

Language System

Neuromuscular Mapping

Vocal Tract System

Speech Generation

Phone, Word, Prosody

Feature Extraction

Articulatory Parameter

Speech Analysis

Speech Understanding

Message Comprehension

Language System

Neural Transduction

Cochlea Motion

Application Semantics, Actions

P(M)

P(W|M)

P(S|W,M)

P(A|S,W,M)

P(X|A,S,W,M)
Statistical Modeling Paradigm

- The statistical modeling paradigm used in speech and language processing
Statistical Modeling Paradigm

• Approaches based on Hidden Markov Models (HMMs) dominate the area of speech recognition
  – HMMs are based on rigorous mathematical theory built on several decades of mathematical results developed in other fields
  – HMMs are generated by the process of training on a large corpus of real speech data
Difficulties: Speech Variability

- Pronunciation Variation
- Intra-speaker variability
- Inter-speaker variability
- Variability caused by the context
- Variability caused by the environment
- Robustness Enhancement
- Speaker-independency
- Speaker-adaptation
- Speaker-dependency
- Context-Dependent Acoustic Modeling
Large Vocabulary Continuous Speech Recognition

Feature Extraction

Linguistic Decoding and Search Algorithm

Speech Corpora

Acoustic Modeling

Acoustic Models

Lexicon

Language Models

Language Modeling

Text Corpora

文字輸出

Feature Vectors

可能詞句

語音輸入

\[ \hat{W} = \arg \max_w P(W|X) \]

\[ = \arg \max_w \frac{P(X|W)P(W)}{P(X)} \]

\[ = \arg \max_w P(X|W)P(W) \]

聲學模型機率

語言模型機率

貝氏定理

詞彙網路搜尋

音頻資料庫

語音資料庫

語音輸入
### Large Vocabulary Continuous Speech Recognition

- Transcription of Broadcast News Speech

<table>
<thead>
<tr>
<th>Segment</th>
<th>Duration</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.0000</td>
<td>SIL</td>
</tr>
<tr>
<td>1</td>
<td>1.0000</td>
<td>71695</td>
</tr>
<tr>
<td>2</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.7571</td>
<td>820</td>
</tr>
<tr>
<td>4</td>
<td>0.7571</td>
<td>422</td>
</tr>
<tr>
<td>5</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>33</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>34</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>39</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>41</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>42</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>43</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>44</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>46</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>47</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>48</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>49</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>51</td>
<td>0.7571</td>
<td>1</td>
</tr>
<tr>
<td>52</td>
<td>0.7571</td>
<td>1</td>
</tr>
</tbody>
</table>
Spoken Dialogue

• Spoken language is attractive because it is the most natural, convenient and inexpensive means of exchanging information for humans

• In mobilizing situations, using keystrokes and mouse clicks could be impractical for rapid information access through small handheld devices like PDAs, cellular phones, etc.
Spoken Dialogue

- Flowchart
Spoken Dialogue

- Multimodality of Input and Output

Experimental client workstation incorporating sight, sound, and touch modalities for human/machine communication. The eye tracker provides a gaze-controlled cursor for indicating objects in the display. The tactile force-feedback glove allows displayed objects to be grasped, “felt,” and moved. Hands-free speech recognition and synthesis provides natural conversational interaction [7].
### Spoken Dialogue

- **Deployed Dialogue Systems**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Language</th>
<th>Vocabulary Size</th>
<th>Average Words/Ut</th>
<th>Utts/Dialogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSELT Train Timetable Info</td>
<td>Italian</td>
<td>760</td>
<td>1.6</td>
<td>6.6</td>
</tr>
<tr>
<td>SpeechWorks Air Travel Reservation</td>
<td>English</td>
<td>1000</td>
<td>1.9</td>
<td>10.6</td>
</tr>
<tr>
<td>Philips Train Timetable Info</td>
<td>German</td>
<td>1850</td>
<td>2.7</td>
<td>7.0</td>
</tr>
<tr>
<td>CMU Movie Information</td>
<td>English</td>
<td>757</td>
<td>3.5</td>
<td>9.2</td>
</tr>
<tr>
<td>CMU Air Travel Reservation</td>
<td>English</td>
<td>2851</td>
<td>3.6</td>
<td>12.0</td>
</tr>
<tr>
<td>LIMSI Train Timetable Info</td>
<td>French</td>
<td>1800</td>
<td>4.4</td>
<td>14.6</td>
</tr>
<tr>
<td>MIT Weather Information</td>
<td>English</td>
<td>1963</td>
<td>5.2</td>
<td>5.6</td>
</tr>
<tr>
<td>MIT Air Travel Reservation</td>
<td>English</td>
<td>1100</td>
<td>5.3</td>
<td>14.1</td>
</tr>
<tr>
<td>AT&amp;T Operator Assistance</td>
<td>English</td>
<td>4000</td>
<td>7.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Air Travel Reservations (human)</td>
<td>English</td>
<td>?</td>
<td>8.0</td>
<td>27.5</td>
</tr>
</tbody>
</table>
Spoken Dialogue

• Topics vs. Dialogue Terms

![Bar chart showing average number of turns per dialogue for different topics: Movies, Restaurants, Flights, Cars, Jobs.](chart.png)
Speech-based Information Retrieval

- **Task:**
  - Automatically indexing a collection of spoken documents with speech recognition techniques
  - Retrieving relevant documents in response to a text/speech query
Speech-based Information Retrieval (cont.)

在四種不同時機下的資訊檢索過程。使用聲音問句(VQ，Voice Queries)或文字問句(TQ，
Text Queries)去檢索聲音資訊(VI，Voice Information)或者是傳統的文字資訊(TI，Text
Information)。
Speech-based Information Retrieval (cont.)

輸入聲音問句：“請幫我查總統府升旗典禮”

中文語音資訊搜尋雜形展示系統

可以選擇同時使用音節、字、同義字等多種索引特徵

聲音問句的語音辨識結果

檢索到新聞的語音辨識結果

檢索到新聞的影音
Speech-based Information Retrieval (cont.)

- Overlapping character bigrams
- Overlapping syllable bigrams

- Vector space model
- Character based indexer
- Syllable based indexer

- PDA, microphone, cellular phone
- Pocket PC
- Search client

- SAPI
- Mandarin LVCSR engine
- LVCSR or syllable decoding
Applications

Multimedia Technologies
Speech-based Information Retrieval
Spoken Dialogue
Dictation & Transcription
Distributed Speech Recognition and Wireless Environment
Multilingual Speech Processing

Emerging Technologies
Information Indexing & Retrieval
Text-to-speech Synthesis
Speech/ Language Understanding

Basic Technologies
Acoustic Processing: features, modeling, pronunciation variation, etc.
Decoding & Search Algorithms
Linguistic Processing & Language Modeling

Integrated Technologies
Speech Recognition Core
Keyword Spotting
Robustness: noise/channel feature/model

Hands-free Interaction: acoustic reception microphone array, etc.
Speaker Adaptation & Recognition

Adopted from Prof. Lin-shan Lee
Speech Processing Toolkit

• HTK (Hidden Markov Model ToolKit)
  – A toolkit for building Hidden Markov Models (HMMs)
  – The HMM can be used to model any time series and the core of HTK is similarly general-purpose
  – In particular, for the acoustic feature extraction, HMM-based acoustic model training and HMM network decoding
Speech Processing Toolkit

- HTK (Hidden Markov Model Toolkit)

```
HLEd
HLStats

HSLAB
HCOPY
HLIST
HQUANT

Transcriptions
Speech

HCOMP, HINIT, HRES, HEREST
HSMOOTH, HHED, HEADAPT

HDMAN

HMMs

HVITE

Networks
HBUILD
HPARSE

Dictionary

Transcriptions
HRESULTS

Data Prep
Training
Testing
Analysis
```
Speech Industry

- Telecommunication
- Information Appliance
- Interactive Voice Response
- Voice Portal
- Multimedia Database
- Education
- .....
# 2003 Speech Workshop

- **Date:** 19 September, 2003  
- **Place:** National Tsing Hua Univ., Hsinchu  
- **Webpage:** [http://140.114.75.26/rocling/](http://140.114.75.26/rocling/)

<table>
<thead>
<tr>
<th>日期</th>
<th>時間</th>
<th>會議內容</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/19</td>
<td>13:30–14:20</td>
<td>演講：當前語音技術研究之趨勢與展望</td>
</tr>
<tr>
<td></td>
<td>14:30–15:20</td>
<td>演講：車內環境之對話系統</td>
</tr>
<tr>
<td></td>
<td>15:40–16:20</td>
<td>座談會：電信學門語音處理領域發展與國際合作</td>
</tr>
<tr>
<td></td>
<td>註：13:30–17:00 國科會研究計畫成果發表(壁報)</td>
<td></td>
</tr>
</tbody>
</table>