Prologue: Introduction to Intelligent Speech Technology

Yanmin Qian

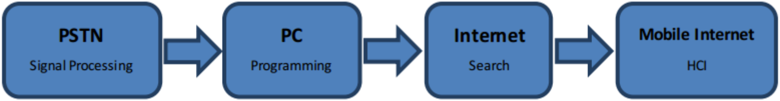
SpeechLab
Department of Computer Science & Engineering
Shanghai Jiao Tong University

Autumn 2017

Syllabus

- **Prologue: Introduction to Intelligent Speech Technology**
- **Part 1: Basic Concepts and Theories**
  - Probability and stochastic process
  - Pattern recognition and machine learning
- **Part 2: Fundamental of Speech Recognition**
  - Speech signal processing
  - Acoustic modelling (Hidden Markov Models)
  - Language modelling (n-grams)
  - Decoding algorithm
  - Large vocabulary continuous speech recognition (LVCSR)
- **Part 3: Advanced Topics of Speech Recognition**
  - Deep neural network for speech recognition
  - Discriminative training and adaptation

Intelligent Speech Technology



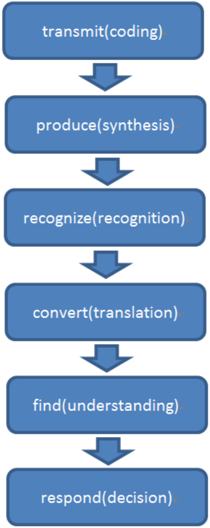
Speech interaction is one of the most important forms of Human Computer Interaction (HCI).

Speech and Language Processing

Speech and language processing aims at modelling and manipulating information from speech and text to

- **transmit (coding)** speech signal efficiently
- **produce (synthesis)** human-like natural text and/or speech
- **recognize (recognition)** underlying text and/or other information from speech signals
- **convert (translation)** text from language to another language
- **find (understanding)** semantic and syntactic content from recognized text and other info.
- **respond (decision)** to incoming semantic content to form conversation

Name a few applications?



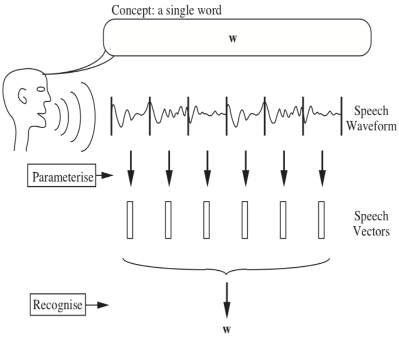
Speech Recognition

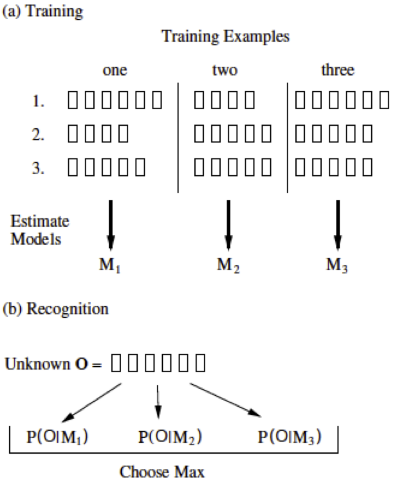
Speak now

What can I help you with?

Statistical Speech Recognition

Diagram of Isolated Word Recognition

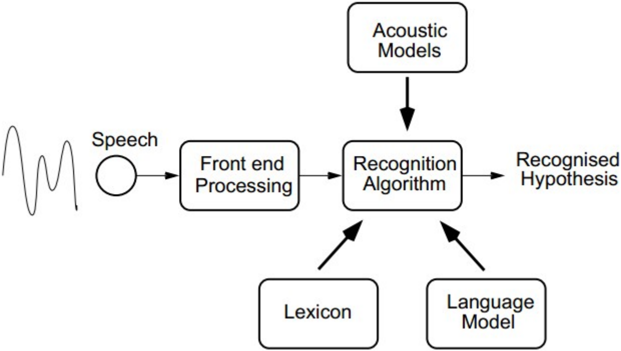




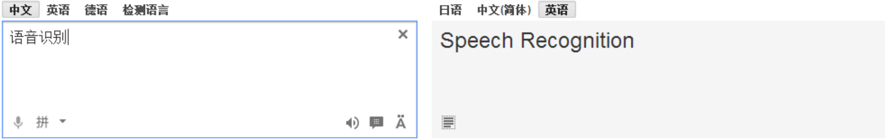
Statistical Speech Recognition

Full Architecture

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} p(\mathbf{A}|\mathbf{O})p(\mathbf{O}|\mathbf{L})P(\mathbf{L}|\mathbf{W})P(\mathbf{W})$$



Speech Synthesis

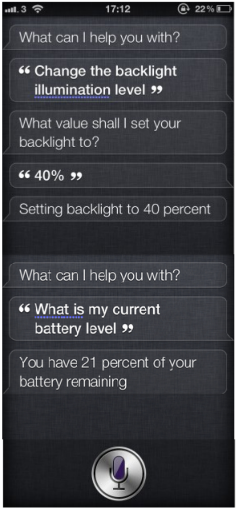


语音识别

Speech Recognition

Stephen Hawking is one of the most famous people using speech synthesis to communicate

Spoken Dialogue System



What can I help you with?

“ Change the backlight illumination level ”

What value shall I set your backlight to?

“ 40% ”

Setting backlight to 40 percent

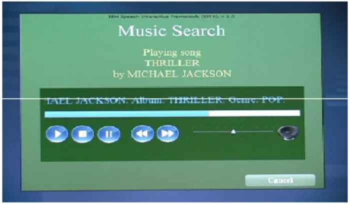
What can I help you with?

“ What is my current battery level ”

You have 21 percent of your battery remaining

Raynor talking to Adjutant in Starcraft 2

In-car spoken dialogue system



IBM Speech Interactive Framework (SPIN), v 1.0

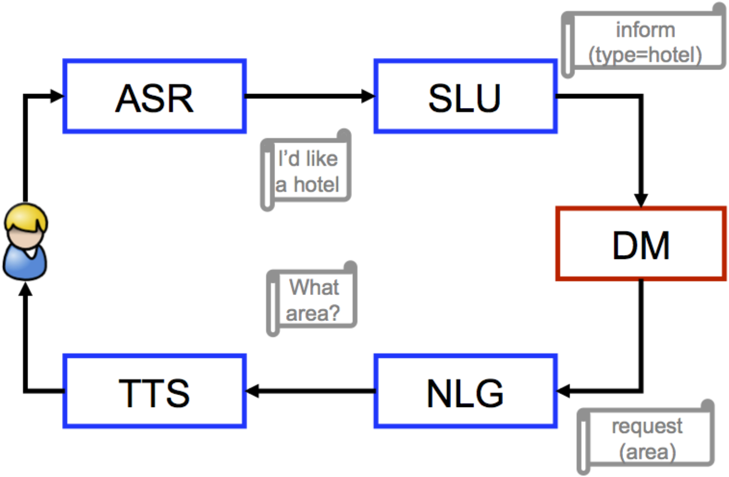
Music Search

Playing song
THRILLER
by MICHAEL JACKSON

IAEL JACKSON. Album: THRILLER. Genre: POP.

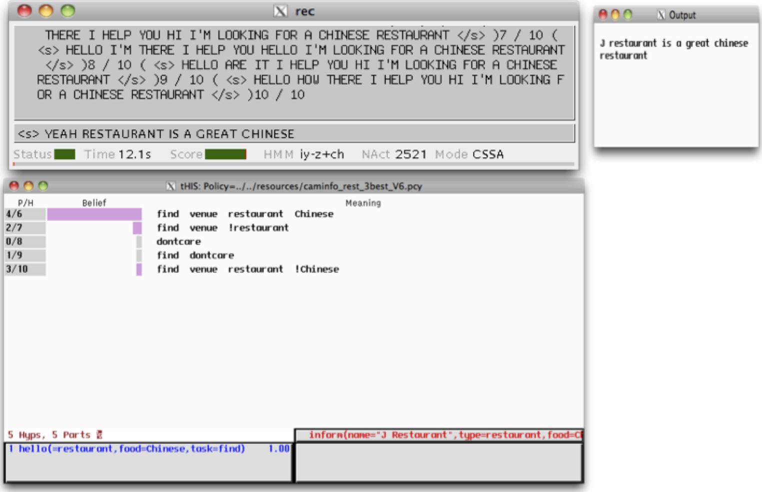
Spoken Dialogue System

Full Architecture



Spoken Dialogue System

Task-oriented spoken dialogue system



```
THERE I HELP YOU HI I'M LOOKING FOR A CHINESE RESTAURANT </s> )7 / 10 (
<s> HELLO I'M THERE I HELP YOU HELLO I'M LOOKING FOR A CHINESE RESTAURANT
</s> )8 / 10 ( <s> HELLO ARE IT I HELP YOU HI I'M LOOKING FOR A CHINESE
RESTAURANT </s> )9 / 10 ( <s> HELLO HOW THERE I HELP YOU HI I'M LOOKING F
OR A CHINESE RESTAURANT </s> )10 / 10

<s> YEAH RESTAURANT IS A GREAT CHINESE
```

Status 🟩 Time 12.1s Score 🟩 HMM iy-z+ch NAct 2521 Mode CSSA

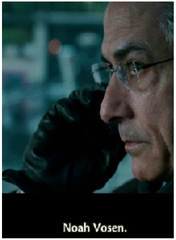
```
J restaurant is a great chinese
restaurant
```

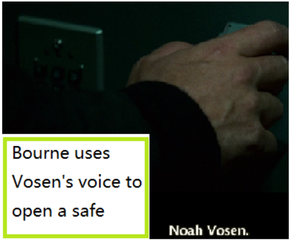
P/H	Belief	Meaning
4/6		find venue restaurant Chinese
2/7		find venue !restaurant
0/8		dontcare
1/9		find dontcare
3/10		find venue restaurant !Chinese

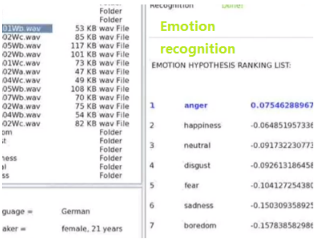
```
5 Hyps, 5 Parts
1 hello(=restaurant,food=Chinese,task=find)       1.00
```

```
inform(name="J Restaurant",type=restaurant,food=C
```

Recognition of Non-text Information (Speaker, Language, Emotion, Humming etc.)




Bourne uses Vosen's voice to open a safe



		Folder
		Folder
01Wb.wav	53 KB	wav File
02Wc.wav	85 KB	wav File
05Wb.wav	117 KB	wav File
02Wb.wav	101 KB	wav File
01Wc.wav	73 KB	wav File
02Wa.wav	47 KB	wav File
04Wc.wav	49 KB	wav File
05Wb.wav	108 KB	wav File
07Wb.wav	70 KB	wav File
02Wa.wav	75 KB	wav File
04Wb.wav	54 KB	wav File
02Wc.wav	82 KB	wav File
om		Folder
t		Folder
		Folder
ness		Folder
l		Folder
ss		Folder

guage = German

aker = female, 21 years

Recognition Done!

Emotion recognition

EMOTION HYPOTHESIS RANKING LIST:

1	anger	0.07546288967
2	happiness	-0.06485195733€
3	neutral	-0.09173223077ǝ
4	disgust	-0.09261318645ɛ
5	fear	-0.10412725438(
6	sadness	-0.15030935892ɛ
7	boredom	-0.15783858298€

Syllabus
Recap

- Prologue: Introduction to Intelligent Speech Technology
- **Part 1: Basic Concepts and Theories**
  - Probability and stochastic process
  - Pattern recognition and machine learning
- **Part 2: Fundamental of Speech Recognition**
  - Speech signal processing
  - Acoustic modelling (Hidden Markov Models)
  - Language modelling (n-grams)
  - Decoding algorithm
  - Large vocabulary continuous speech recognition (LVCSR)
- **Part 3: Advanced Topics of Speech Recognition**
  - Neural network in speech recognition
  - Discriminative training and adaptation

Key Actions and Course Works

- **Evaluation**: No Examination
  - Attendance or request for leave (10%)
  - Projects + Talks (90%)

- **Part 1-Project 1: Basic concepts and theories (40%)**
  - `Memerize and Instantiate`
  - `Implement EM algorithm for GMM - Individual`
- **Part 2-Project 2: Automatic Speaker Verification Spoofing Detection (30%)**
  - `Derive and Practise`
  - `Design the complete system - Challenge (Group)`
- **Part 3-Talk: Advances on Speech Processing (20%)**
  - `Read and Summarize`
  - `Paper/Derivation/Tool presentation - Group`

Project 1 - EM algorithm for GMM

- **EM algorithm for GMM (40%)**
  - A simple binary classification task
    - 2-dimensional feature are provided
    - Three sets are provided: training, dev and test
    - GMM training and evaluation
  - Detailed report + Classification results + Source code

Project 2 - Automatic Speaker Verification Spoofing Detection

- **Automatic Speaker Verification Spoofing Detection Challenge (30%)**
  - 1st ASVspoofing Challenge in SJTU
  - A detection and recognition task
    - Training set and test set are provided
  - The complete system design
    - Feature extraction
    - Model training
    - Recognition or classification
  - The data and rules are the same as the ASVspoof 2017 Challenge
    - http://www.asvspoof.org/
  - Detailed technical report with results (formal paper style)
    - The formal ICASSP template will be provided
  - Rank & Reward

Talk - Advances on Speech Processing

- **Advanced Talks on Speech Processing (20%)**
  - New Progress in Speech & Language processing
  - First Read and then Present
  - Any topics are welcomed
    - Speech Enhancement
    - Speech Recognition
    - Speaker / Language Identification
    - Speech Emotion Recognition
    - Language Modeling
    - ......

References

- **Probability and pattern recognition basics**

  Pattern Recognition and Machine Learning.
  Christopher M. Bishop, Springer.

- **Speech recognition theory and tools**

  HTK Book, Steve Young, et al. Cambridge University

- **Speech and language technology**

  Spoken Language Processing.
  A Guide to Theory, Algorithm and System Development
  Xuedong Huang, Alex Acero, Hsiao-Wuen Hon

- **Deep Learning for Speech Processing**

  Automatic Speech Recognition-A Deep Learning Approach.
  Dong Yu & Li Deng. Springer

Information

- Yanmin Qian
  - Computer Science Department-SEIEE 3#515
  - Email: yanminqian@sjtu.edu.cn
  - Cell-Phone: +86-18516056597
- Teaching Assistant: Heinrich Dinkel
  - Computer Science Department-SEIEE 3#225
  - Email: richman@sjtu.edu.cn
  - Cell-Phone: +86-13262830583
- Speech-Lab in SJTU (Speech-Lab in SJTU)
  - Homepage: http://speechlab.sjtu.edu.cn/

FTP

- Materials for download and upload
  - FTP: ftp://202.120.38.125
  - Port: 8821
  - Username: speech2017m
  - Password: speech2017m