

Multi-Modal User Interaction 2010

Lecture 1: Introduction and Talking to Computers (I)



Zheng-Hua Tan

Department of Electronic Systems
Aalborg University, Denmark
zt@es.aau.dk



MMUI, I, Zheng-Hua Tan

1

Human interaction speeds

- Human interaction speeds (potential text entry bandwidth) for a number of methods

Interaction method	Word per minute
Multi-tap (timeout kill)	
T9	
Handwriting	
Keyboard touch-typing	
Eye-gaze tracking	
Speaking	
Dictation	



MMUI, I, Zheng-Hua Tan

2

Computer as dream of human being

HAL talks, listens, reads lips and solves problems

- Nature and effortless for human
- Hard for computer
- Dream of AI scientists and human
- True in *2001: A Space Odyssey*



(After *2001: A Space Odyssey*, 1968)



MMUI, I, Zheng-Hua Tan

3

Computer as a reality: state-of-the-art

- Man against machine

Competition results

	Eli Champion texter	Sean Speech technology
Texting	00:50.17	00:21.83
Texting 2	00:24.72	00:13.49
	Perry Champion driver	Sean Speech technology
Driving		
⚡ = Crash		

NUANCE

- Text to speech (TTS) @ AT&T



MMUI, I, Zheng-Hua Tan

4

Computer as a reality: state-of-the-art

■ Dragon Naturally speaking 10

- ❑ It's three times faster than most people type (typing average WPM __; reading WPM __)
- ❑ Up to 99% accurate right out of the box!
- ❑ The latency between speaking and seeing words on the PC has nearly been eliminated.
- ❑ Let you find files on your PC, search web maps, shop on eBay, set appointments and more, all with simple voice commands.
- ❑ Demo, CD

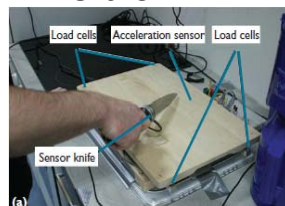


MMUI, I, Zheng-Hua Tan

5

Interaction

- Interaction with daily objects includes a few new elements, e.g.,
 - ❑ no keyboard and mouse available;
 - ❑ the user focusing on other tasks in hand and leaving reduced attention for interaction
- Interaction should be natural, effortless and even invisible.



Embedded interaction



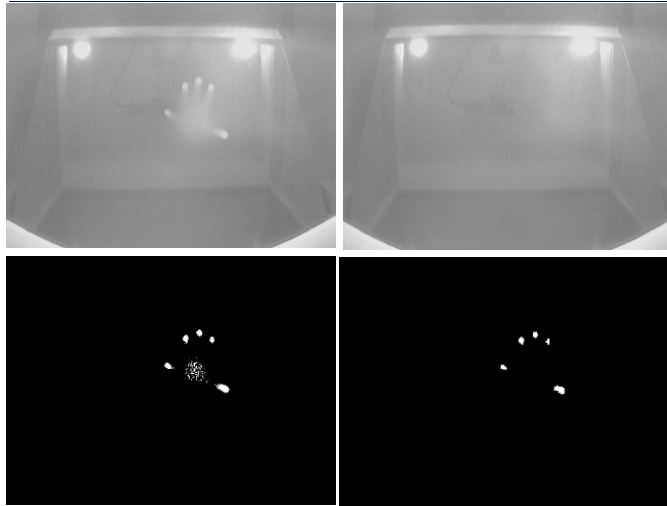
(Kranz et al. 2010)



MMUI, I, Zheng-Hua Tan

6

Touch table and scratch surface



Stethoscope
to enable
scratch-input



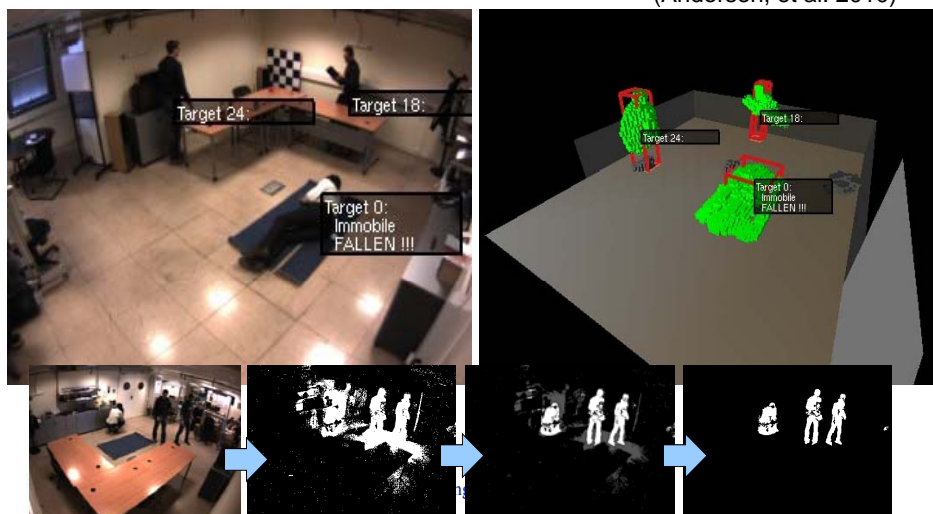
MMUI, I, Zheng-Hua Tan

7

3D sensing

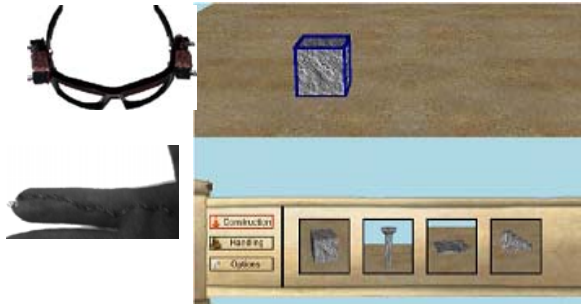
- a) an image with persons and information overlay
- b) detected foreground and information

Elderly care, surveillance
(Andersen, et al. 2010)



Egocentric interaction

- Exploits the spatial relation between user and device and uses changes in this relation as input commands.



(T. Luel and F. Mazzone, 2009)



(M.H. Justesen, et al. 2010)



MMUI, I, Zheng-Hua Tan

9

Interaction through tagging

- (RFID, barcode) tagging can make interaction and finding information shadows much easier by eliminating the need for human inputs or interferences.
- Nabaztag:tag and Mir:ror are two interesting examples of interaction through tagging developed by <http://www.violet.net>.



MMUI, I, Zheng-Hua Tan

10

Finding information

Google it! ☺ ☹

Layar First Mobile
Augmented Reality
Browser

Hyperlinked buildings in the two worlds



(Quack, et al., 2008)



The world is the interface!

MMUI, I, Zheng-Hua Tan

11

Course info

- MM1~5: Ann Morrison
- MM6 ~10: Zheng-Hua Tan
 - Tel. 99 40 86 86
 - Room A6-319, NJ12
 - Speech interaction, lip-reading
 - Eye-gaze tracking
 - Multimodal design (speech, eye-gaze, gesture)
 - Multimodal fusion



MMUI, I, Zheng-Hua Tan

12

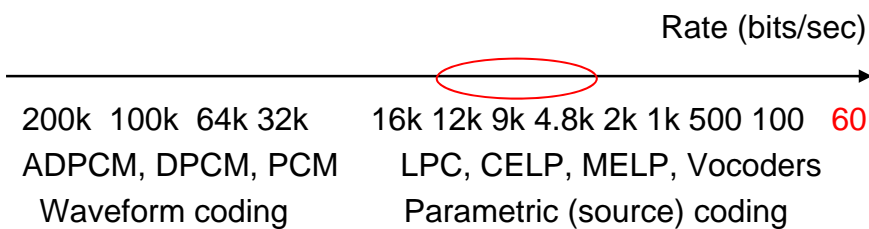
Part II: Basic about speech

- Introduction
- Basics about speech – a short introduction
- Template based approach – DTW
- Statistical model based approach – HMM
- Types of speech recognizers
- Applications



Information in Speech

- Speech coding data rates



Human can understand text:

$$10 \text{ char/sec} \times 6 \text{ bits/ASCII char} = 60 \text{ bits/sec}$$

Is content in speech more than 60 bits/sec?



Information in Speech – cont.

🔊 “That's one **small step for man**; one **giant leap for mankind**.”
-- Neil Armstrong, *Apollo 11 Moon Landing Speech*

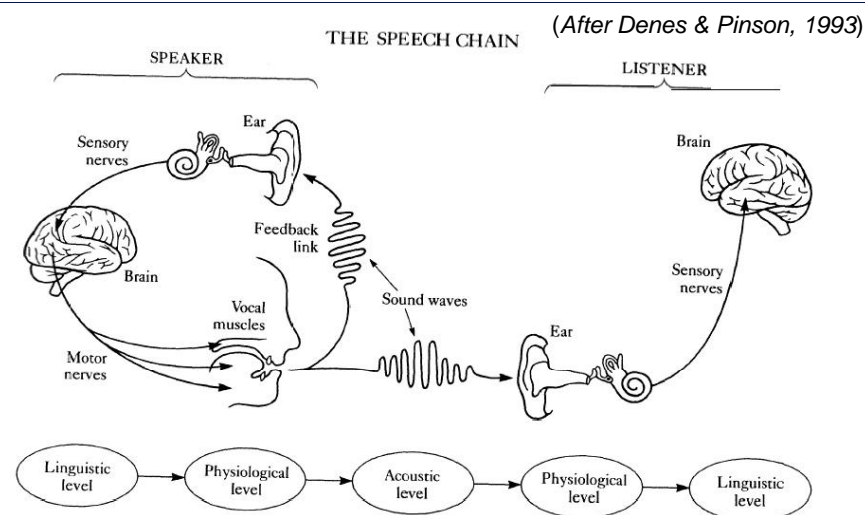
Speech contains **speaker identity, emotion, meaning, text, language, sex and age, channel characteristics**. → speech techniques



MMUI, I, Zheng-Hua Tan

15

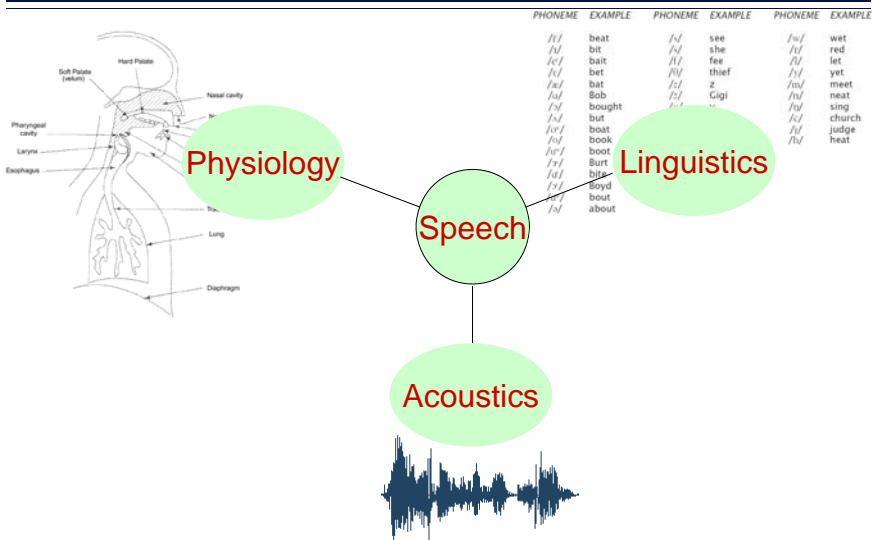
The speech chain



MMUI, I, Zheng-Hua Tan

16

Speech is a complex process



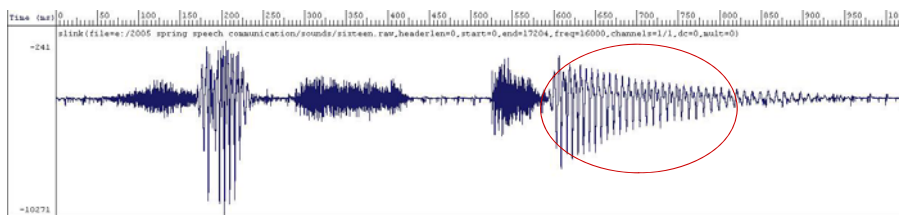
MMUI, I, Zheng-Hua Tan

17

Speech sounds and waveforms

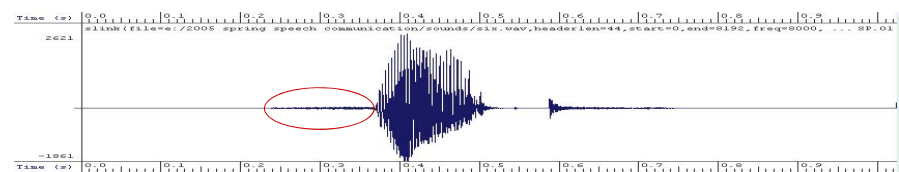


sixteen /s/ /i/ /k/ /s/ /t/ /ee/ /n/



six

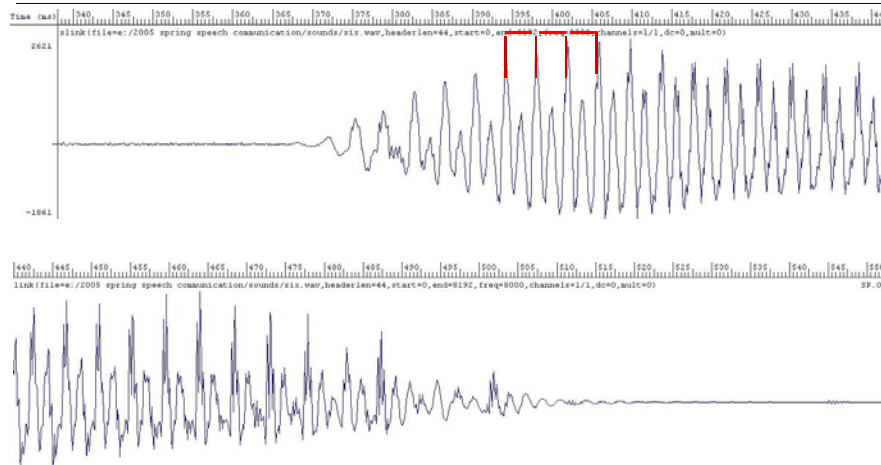
periodicity, intensity, duration, boundary, etc



MMUI, I, Zheng-Hua Tan

18

Observing pitch from waveforms

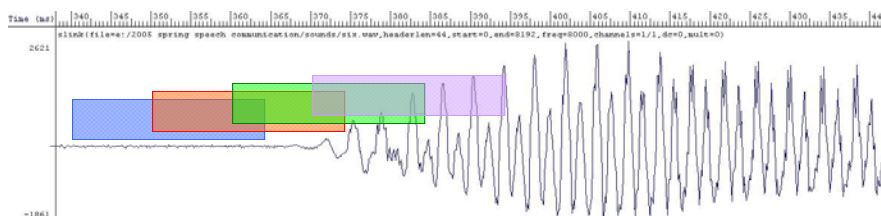


MMUI, I, Zheng-Hua Tan

19

Dimension & speech representation

- **The curse of dimension** – the computational cost increases exponentially with the dimension of the problem
- The frame-based analysis yields a sequence as a new representation of the speech signal
 - samples at 8000/sec → **vectors** at 100/sec



MMUI, I, Zheng-Hua Tan

20

Spectrogram

■ Spectrogram

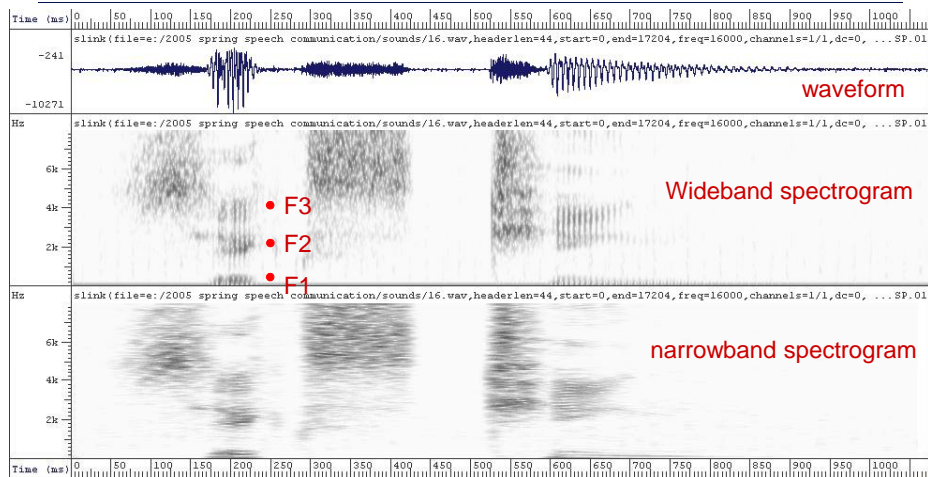
- ❑ 2-D waveform (amplitude/time) is converted into a 3-D pattern (amplitude/frequency/time)
- ❑ Wideband spectrogram: analyzed on 15ms sections of waveform with a step of 1ms
 - Voiced regions with vertical striations due to the periodicity of the time waveform (each vertical line represents a pulse of vocal folds) while unvoiced regions are 'snowy'.
- ❑ Narrowband spectrogram: analyzed on 50ms sections
 - Pitch for voiced intervals in horizontal lines



MMUI, I, Zheng-Hua Tan

21

Sound Spectrogram: an example



MMUI, I, Zheng-Hua Tan

22

Speech Tool

- **Speech Filing System- Tools for Speech Research**

- It performs standard operations such as recording, replay, waveform editing and labelling, spectrographic and formant analysis and fundamental frequency estimation.
- <http://www.phon.ucl.ac.uk/resource/sfs/>

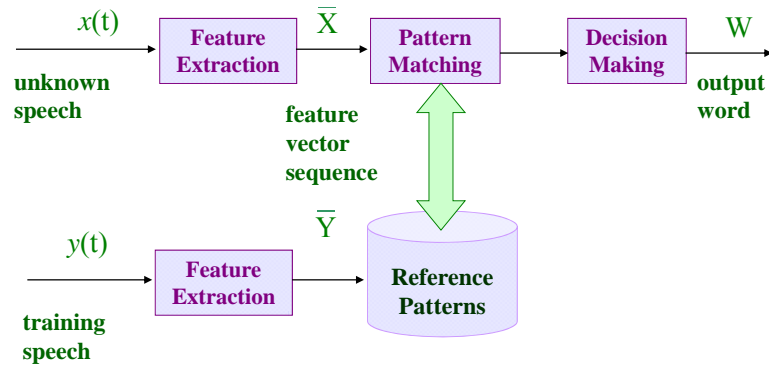


Part III: DTW

- Introduction
- Basics about speech – a short introduction
- Template based approach – DTW
- Statistical model based approach – HMM
- Types of speech recognizers
- Applications



Template based ASR



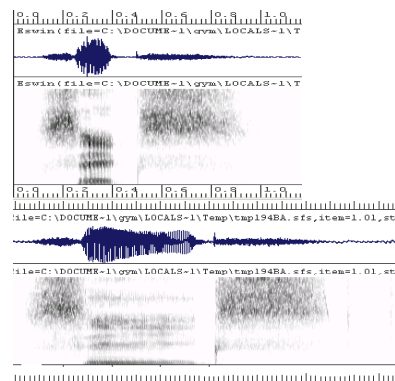
Template matching mechanism

- ❑ Calculate the distance between two patterns
- ❑ Dynamic time warping (DTW)



Speaking rate and time-normalization

- Speaking rate variation causes nonlinear fluctuation in a speech pattern time axis



- Time-normalization is needed.



DP based time-normalization

- **Dynamic programming** is a pattern matching algorithm with a nonlinear time-normalization effect.
 - Time differences btw two speech patterns are eliminated by warping the time axis of one so that the maximum coincidence is attained with the other, also called dynamic time warping (DTW)
 - The time-normalized distance is calculated as the minimized residual distance between them, remaining still after eliminating the timing differences.



MMUI, I, Zheng-Hua Tan

27

Dynamic programming

- Consider two speech patterns expressed as a sequence of feature vectors :

$$A = a_1, a_2, \dots, a_i, \dots, a_I$$

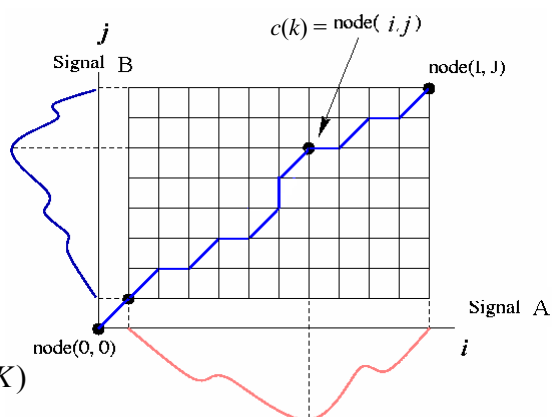
$$B = b_1, b_2, \dots, b_j, \dots, b_J$$

- Consider an i - j plane, then time differences can be depicted by a sequence of points $c = (i, j)$:

where

$$F = c(1), c(2), \dots, c(k), \dots, c(K)$$

$$c(k) = (i(k), j(k))$$



MMUI, I, Zheng-Hua Tan

28

Dynamic programming (cont'd)

- The sequence c is called a **warping function**.
- A distance btw two feature vectors is

$$d(c) = d(i, j) = \|a_i - b_j\|$$
- The weighted summation of distances on warping function F becomes

$$E(F) = \sum_{k=1}^K d(c(k)).w(k)$$
- The time-normalized distance btw A and B is defined as the minimum residual distance btw them

$$D(A, B) = \min \left[\frac{\sum_{k=1}^K d(c(k)).w(k)}{\sum_{k=1}^K w(k)} \right]$$



MMUI, I, Zheng-Hua Tan

29

Restrictions on warping function

- Warping function F (or points $c(k)$), as a model of time-axis fluctuation in speech, has restrictions:
 - 1) Monotonic conditions :

$$i(k-1) \leq i(k) \text{ and } j(k-1) \leq j(k)$$
 - 2) Continuity conditions :

$$i(k) - i(k-1) \leq 1 \text{ and } j(k) - j(k-1) \leq 1$$
 - 3 Boundary conditions :

$$i(1) = 1, j(1) = 1 \text{ and } i(K) = I, j(K) = J.$$
 - 4) Adjustment window condition

$$|i(k) - j(k)| \leq r$$
 - 5) Slope constraint condition :

A gradient should be neither too steep nor too gentle.



MMUI, I, Zheng-Hua Tan

30

The simplest DP of symmetric form

- Step 1: Initialisation:

$$g(1, 1) = 2d(1, 1)$$

- Step 2: Iteration (DP equation):

$$g(i, j) = \min \begin{bmatrix} g(i, j-1) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i-1, j) + d(i, j) \end{bmatrix}$$

Adjustment window:

$$j - r \leq i \leq j + r$$

- Step 3: Termination:

Time-normalised distance

$$D(A, B) = \frac{1}{N} g(I, J), \text{ where } N = I + J$$



MMUI, I, Zheng-Hua Tan

31

From template to statistical method

- The template method with DP alignment is a simplified, **non-parametric method** which is hard to characterise the **variation** among utterances
- Hidden Markov model (HMM) is a powerful **statistical method** of characterising the observed data samples of a discrete-time series
- The underlying assumption of the HMM is
 - The speech signal can be well characterised as a parametric random process
 - The parameters of the stochastic process can be estimated in a precise, well-defined manner



MMUI, I, Zheng-Hua Tan

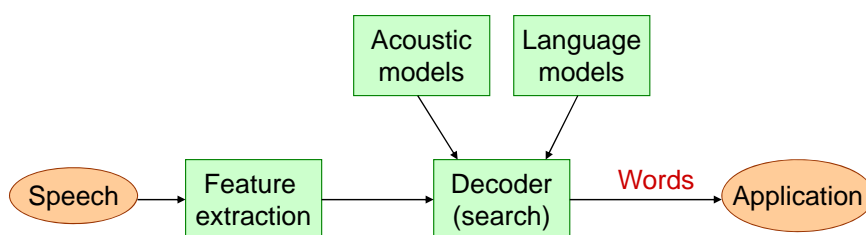
32

Part IV: HMM – conceptual intro

- Introduction
- Basics about speech – a short introduction
- Template based approach – DTW
- Statistical model based approach – HMM
- Types of speech recognizers
- Applications



Key components of LVCSR system

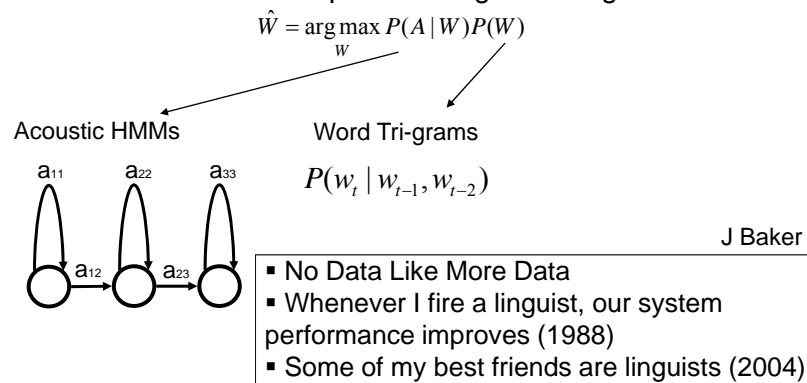


- Speech recognition involves:
 - How to **represent** the signal
 - How to **model** both acoustic and language constraints
 - How to **search** for the optimal answer



The Statistical Approach

- Hidden Markov Models based statistical approach (Fred Jelinek and Jim Baker, IBM)
- Foundations of modern speech recognition engines

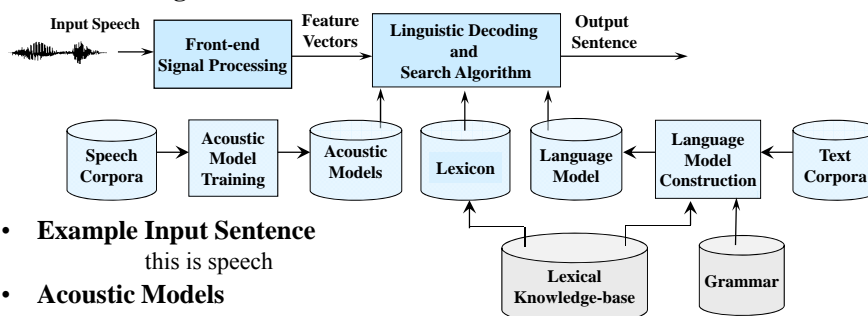


MMUI, I, Zheng-Hua Tan

35

Large vocabulary speech recognition

• A Block Diagram



• Example Input Sentence

this is speech

• Acoustic Models

(th-ih-s-ih-z-s-p-ih-ch)

• Lexicon

(th-ih-s) → this
(ih-z) → is
(s-p-iy-ch) → speech

• Language Model

(this) – (is) – (speech)

$P(\text{this}) P(\text{is} | \text{this}) P(\text{speech} | \text{this is})$

$P(w_i | w_{i-1})$ bi-gram language model

$P(w_i | w_{i-1}, w_{i-2})$ tri-gram language model etc

L.S. Lee, 2007

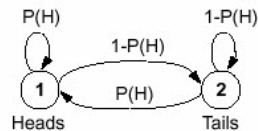
36

“Hidden” Markov model

Consider the problem of predicting the outcome of a coin toss experiment.
You observe the following sequence:

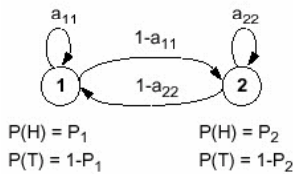
$$\bar{O} = (HHTTTHTH...H)$$

What is a reasonable model of the system?



1-Coin Model
(Observable Markov Model)

O = H H T T H T H H T T H ...
S = 1 1 2 2 1 2 1 1 2 2 1 ...



2-Coins Model
(Hidden Markov Model)

O = H H T T H T H H T T H ...
S = 2 1 1 2 2 2 1 2 2 1 2 ...



MMUI, I, Zheng-Hua Tan

37

The Urn-and-Ball model

The Urn-and-Ball Model

doubly stochastic systems



P(red) = $b_1(1)$
P(green) = $b_1(2)$
P(blue) = $b_1(3)$
P(yellow) = $b_1(4)$
...



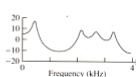
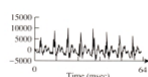
P(red) = $b_2(1)$
P(green) = $b_2(2)$
P(blue) = $b_2(3)$
P(yellow) = $b_2(4)$
...



P(red) = $b_3(1)$
P(green) = $b_3(2)$
P(blue) = $b_3(3)$
P(yellow) = $b_3(4)$
...

$\bar{O} = \{\text{green, blue, green, yellow, red, ..., blue}\}$

How can we determine the appropriate model for the observation sequence given the system above?



MMUI, I, Zheng-Hua Tan

38

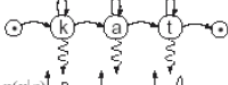
Elements of HMM

- HMM is specified by:

- states q^i 

- transition probabilities a_{ij} 

$$p(q_n^j | q_{n-1}^i) \equiv a_{ij}$$

- emission distributions $b_i(x)$ 

$$p(x | q^i) \equiv b_i(x)$$

- + (initial state probabilities $p(q_1^i) \equiv \pi_i$)

	k	a	t	*
*	1.0	0.0	0.0	0.0
k	0.9	0.1	0.0	0.0
a	0.0	0.9	0.1	0.0
t	0.0	0.0	0.9	0.1

From Dan Ellis, 2004.

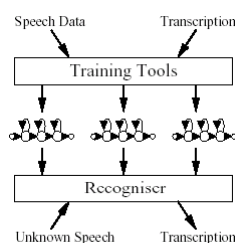


MMUI, I, Zheng-Hua Tan

39

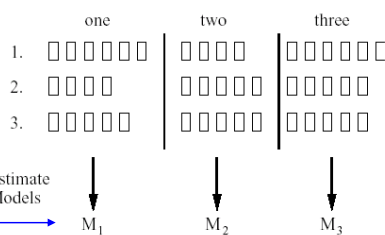
HMM for IWR

(Young et al. 1996)

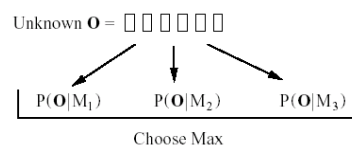


(a) Training

Training Examples



(b) Recognition



MMUI, I, Zheng-Hua Tan

40

Part V: Types of recognizers

- Introduction
- Basics about speech – a short introduction
- Template based approach – DTW
- Statistical model based approach – HMM
- Types of speech recognizers
- Applications

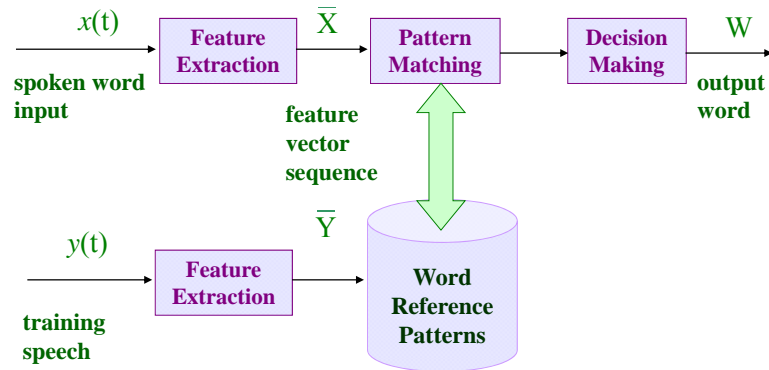


Types of speech recognisers

- Isolated word recognition
- Grammar based recognition
- Large vocabulary continuous speech recognition (N-gram)



Template based method for IWR



Template matching mechanism

- Calculate the distance between two patterns
- Dynamic time warping (DTW)

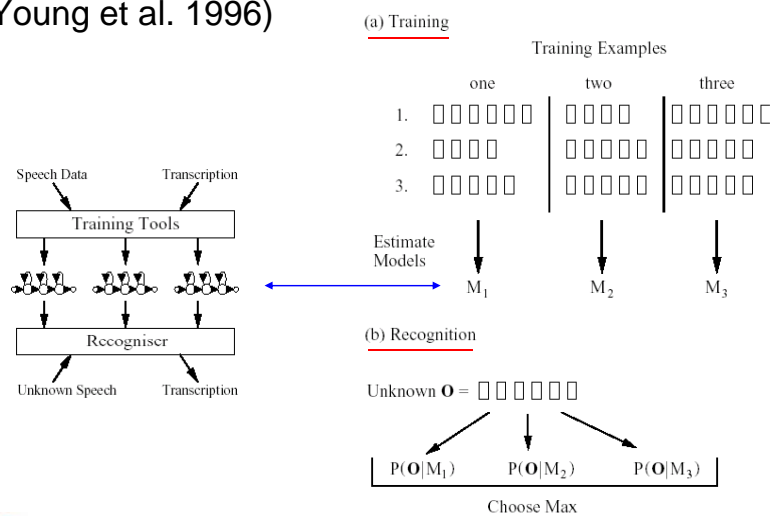


MMUI, I, Zheng-Hua Tan

43

HMM for IWR

(Young et al. 1996)



MMUI, I, Zheng-Hua Tan

44

Language modelling – word looping?

- The allowed sequence of phoneme-based HMMs is defined by a finite state network and all of the words are placed in a loop

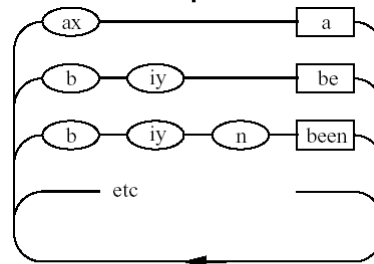


Fig. 1.7 Recognition Network for Continuously Spoken Word Recognition



MMUI, I, Zheng-Hua Tan

45

Grammar – constraining search space

IWR, grammar-based ASR, N-grams

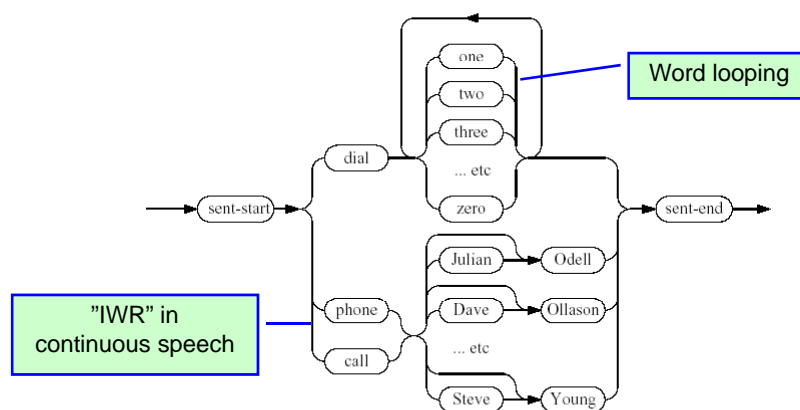


Fig. 3.1 Grammar for Voice Dialling



MMUI, I, Zheng-Hua Tan

46

N-grams

- LM is estimating the probability of word in an utterance given the preceding words.
- N-grams (bigrams, trigrams, etc.)

$$P(w_k | w_1 \dots w_{k-1}) = P(w_k | w_{k-n+1} \dots w_{k-1})$$

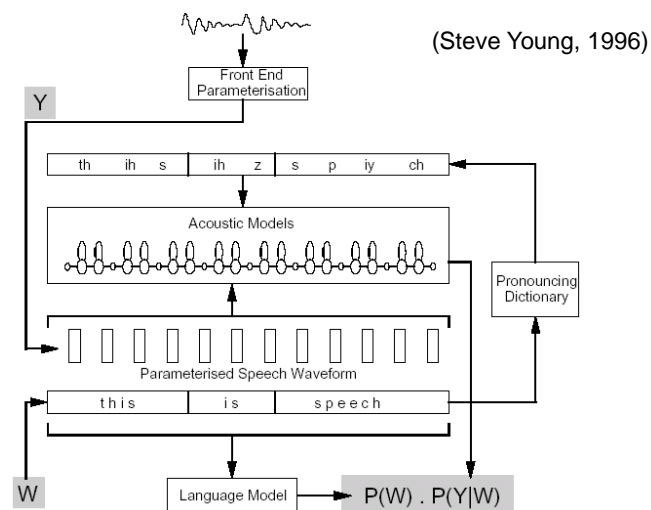
- Discounting and backing-off



MMUI, I, Zheng-Hua Tan

47

LVCSR system overview



MMUI, I, Zheng-Hua Tan

48

Attributes of ASR systems

- **Vocabulary:** small (<20 words) to large (>50K words)
- **Perplexity:** small (< 10) to large (> 200)
- **Enrollment:** speaker-dependent to speaker-independent
- **Speaking mode:** isolated-word to continuous-speech
- **Speaking style:** read speech to spontaneous speech
- **SNR:** high (> 30 dB) to low (< 10 dB)
- **Transducer:** noise-cancelling microphone to cell phone



Part VI: Applications

- Introduction
- Basics about speech – a short introduction
- Template based approach – DTW
- Statistical model based approach – HMM
- Types of speech recognizers
- Applications



Typical applications

- Broad classes that require different UI design [Huang]
 - Office: Desktop applications
 - Home: TV and kitchen
 - Mobile: Cell phone and car
- Applications
 - Command and control
 - Data entry
 - Getting information
 - Conversational systems
 - Dictation (nuance.com, Microsoft, IBM.com)
 - Reading tutor (rosettastone.com, saybot.com)
 - IVR (ferry ticket booking)



MMUI, I, Zheng-Hua Tan

51

Command and control

- Either developers or users define grammars
- Associate with each legal path in the grammar is a corresponding executable event.
- Useful in situations
 - Answering questions
 - Accessing large lists
 - Providing hands-free computing
 - Humanizing the computer
 - Game and entertainment
 - Handheld devices and cars



MMUI, I, Zheng-Hua Tan

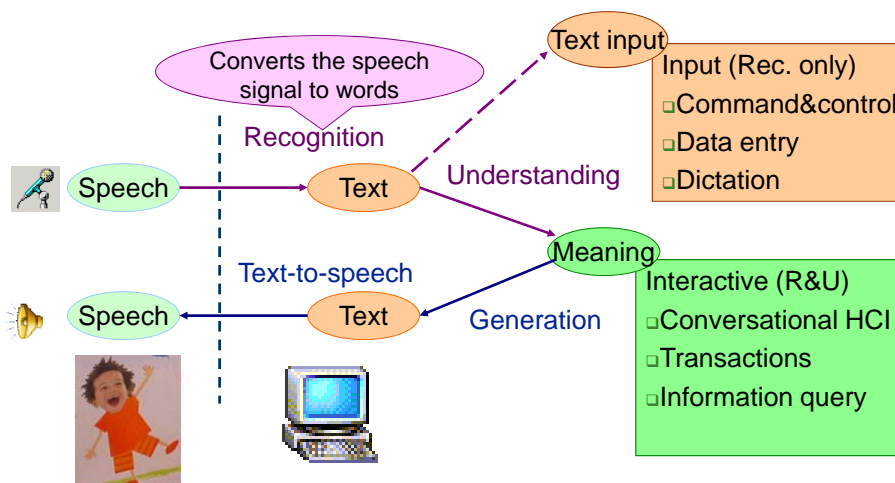
52

Dictation

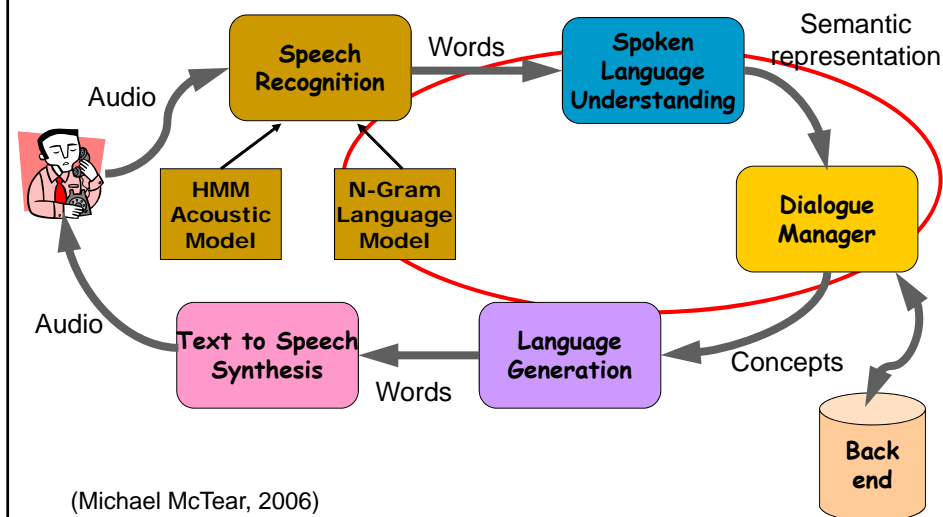
- Dictation should not be considered “general recognition”, as it is dependent on the “topic” of the text data used for LM-training
- Dictation performs better after adaptation to the user
 - Though it can be used as speaker-independent.



Human-computer interaction via speech



Basic dialog system architecture



(Michael McTear, 2006)



MMUI, I, Zheng-Hua Tan

55

Kitchen scenario – fact or fiction?

- Rachel goes into the kitchen, takes a piece of bread and puts it into the toaster. “Not so well done this time.” She goes to the fridge, takes out a carton of milk, and notices that it is almost empty. “Don’t forget to order another carton of milk”, she says to the fridge. “You’re having some friends round for hot chocolate later, maybe I should order two cartons”, says the fridge. “Okay”, says Rachel.

(McTear)



MMUI, I, Zheng-Hua Tan

56

Summary

- Introduction
- Basics about speech – a short introduction
- Template based approach – DTW
- Statistical model based approach – HMM
- Types of speech recognizers
- Applications

