Human interaction speeds

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

<table>
<thead>
<tr>
<th>Interaction method</th>
<th>Word per minute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-tap (timeout kill)</td>
<td></td>
</tr>
<tr>
<td>T9</td>
<td></td>
</tr>
<tr>
<td>Handwriting</td>
<td></td>
</tr>
<tr>
<td>Keyboard touch-typing</td>
<td></td>
</tr>
<tr>
<td>Eye-gaze tracking</td>
<td></td>
</tr>
<tr>
<td>Speaking</td>
<td></td>
</tr>
<tr>
<td>Dictation</td>
<td></td>
</tr>
</tbody>
</table>
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)

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

- Man against machine

<table>
<thead>
<tr>
<th>Competition results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Texting</strong></td>
</tr>
<tr>
<td>Echo 00:55.77</td>
</tr>
<tr>
<td>Parry 00:24.72</td>
</tr>
<tr>
<td><strong>Driving</strong></td>
</tr>
<tr>
<td>1 - Crash</td>
</tr>
</tbody>
</table>

- Text to speech (TTS) @ AT&T
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

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.

( Kranz et al. 2010)
**Touch table and scratch surface**

![Image of touch table and scratch surface with a stethoscope]

Stethoscope to enable scratch-input

---

**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)

**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](http://www.violet.net).
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!

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
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

<table>
<thead>
<tr>
<th>Rate (bits/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200k 100k 64k 32k</td>
</tr>
<tr>
<td>ADPCM, DPCM, PCM</td>
</tr>
<tr>
<td>Waveform coding</td>
</tr>
</tbody>
</table>

Human can understand text:
10 char/sec x 6 bits/ASCII char = 60 bits/sec

Is content in speech more than 60 bits/sec?
“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

The speech chain

(After Denes & Pinson, 1993)
Speech is a complex process

Speech sounds and waveforms

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

six periodicity, intensity, duration, boundary, etc
Observing pitch from waveforms

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
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

Sound Spectrogram: an example
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/](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

unknown speech $x(t)$ → Feature Extraction $\vec{X}$ → Pattern Matching $\vec{Y}$ → Reference Patterns $W$ → Decision Making output word

Training speech $y(t)$ → Feature Extraction $\vec{Y}$

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.

---

**Dynamic programming**

- Consider two speech patterns expressed as a sequence of feature vectors:
  \[
  A = a_1, a_2, \ldots, a_i, \ldots, a_I \]
  \[
  B = b_1, b_2, \ldots, b_j, \ldots, b_J \]
  - Consider an \(i-j\) plane, then time differences can be depicted by a sequence of points \(c=(i,j)\):
  - where
    \[
    F = c(1), c(2), \ldots, c(k), \ldots, c(K) \]
    \[
    c(k) = (i(k), j(k)) \]
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 
  \[ D(A, B) = \min \left[ \frac{\sum_{k=1}^{K} d(c(k)).w(k)}{\sum_{k=1}^{K} w(k)} \right] \]

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:
     \[ i(1) = 1, j(1) = 1 \text{ and } i(K) = J, j(K) = J. \]
  4) Adjustment window condition
     \[ | i(k) - j(k) | \leq r \]
  5) Slope constraint condition:
     A gradient should be neither too steep nor too gentle.
The simplest DP of symmetric form

- **Step 1: Initialisation:**
  \[ g(1, 1) = 2d(1, 1) \]

- **Step 2: Iteration (DP equation):**
  \[
  g(i, j) = \min \left\{ g(i - 1, j - 1) + d(i, j), g(i - 1, j) + 2d(i, j), g(i, j - 1) + d(i, j) \right\}
  \]
  Adjustment window: \( j - r \leq i \leq j + r \)

- **Step 3: Termination:**
  Time-normalised distance
  \[
  D(A, B) = \frac{1}{N} g(I, J), \text{ where } N = I + J
  \]

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.
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

\[ \hat{W} = \arg \max_W P(A|W)P(W) \]

Acoustic HMMs \hspace{1cm} Word Tri-grams

- No Data Like More Data
- Whenever I fire a linguist, our system performance improves (1988)
- Some of my best friends are linguists (2004)

---

**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(this) P(is | this) P(speech | this)
  - P(w_i | w_{i-1}) bi-gram language model
  - P(w_i | w_{i-1}, w_{i-2}) tri-gram language model etc
“Hidden” Markov model

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

\( \tilde{O} = (HHTTHTTHHTTT...H) \)

What is a reasonable model of the system?

1-Coin Model
(Observeable Markov Model)

\( O = HHTTHTTHTHHTT... \)

\( S = 1112212122121... \)

1-Coin Model

\[ \begin{align*}
1 & \quad 2 \\
1 & \quad 1-p_{11} \quad 1-p_{12} \\
2 & \quad 1-p_{21} \quad 2 \\
\end{align*} \]

\[ P(H) = P_1 \quad P(T) = 1 - P_1 \]

\[ P(H) = P_2 \quad P(T) = 1 - P_2 \]

The Urn-and-Ball model

doubly stochastic systems

\[ \begin{align*}
P(\text{red}) & = b_1(1) \quad P(\text{green}) = b_2(2) \\
P(\text{blue}) & = b_3(3) \quad P(\text{yellow}) = b_4(4) \\
\end{align*} \]

\[ \mathbf{b} = (b_1 \); b_2 \); b_3 \); b_4 \); \ldots \); b_k \)

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

- HMM is specified by:
  - states $q^i$
  - transition probabilities $a_{ij}$
    
    \[
p(q_k^i | q_{k-1}^j) = a_{ij}
    \]
  - emission distributions $b_k(x)$
    
    \[
p(x | q_k^i) = b_k(x)
    \]
  - (initial state probabilities $p(q_1^i) = \pi_i$)

From Dan Ellis, 2004.

HMM for IWR

(Young et al. 1996)

(a) Training

<table>
<thead>
<tr>
<th>Training Examples</th>
<th>one</th>
<th>two</th>
<th>three</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Recognition

\[
P(O | M_1) \quad P(O | M_2) \quad P(O | M_3)
\]

Choose Max
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)

HMM for IWR

(Young et al. 1996)
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.

Grammar – constraining search space

IWR, grammar-based ASR, N-grams
N-grams

- LM is estimating the probability of word in an utterance given the preceding words.
- N-grams (bigrams, trigrams, etc.)
  \[ P(w_k \mid w_1\ldots w_{k-1}) = P(w_k \mid w_{k-n+1}\ldots w_{k-1}) \]
- Discounting and backing-off

LVCSR system overview

(Steve Young, 1996)
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-concealing 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)

Command and control

- Either developers of 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
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

- **Speech**
  - Converts the speech signal to words
  - Text input
  - Text
  - Understanding
  - Interactive (R&U)
    - Conversational HCI
    - Transactions
    - Information query

- **Text**
  - Command&control
  - Data entry
  - Dictation

- **Text-to-speech**
  - Meaning
  - Generation

MMUI, I, Zheng-Hua Tan
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)
Summary

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