SEEM3510 Human-Computer Interaction
Expressive Human and Command Languages

Helen Meng
Department of Systems Engineering & Engineering Management

Philip Fu
Department of Computer Science & Engineering

Spring 2022
Week 9

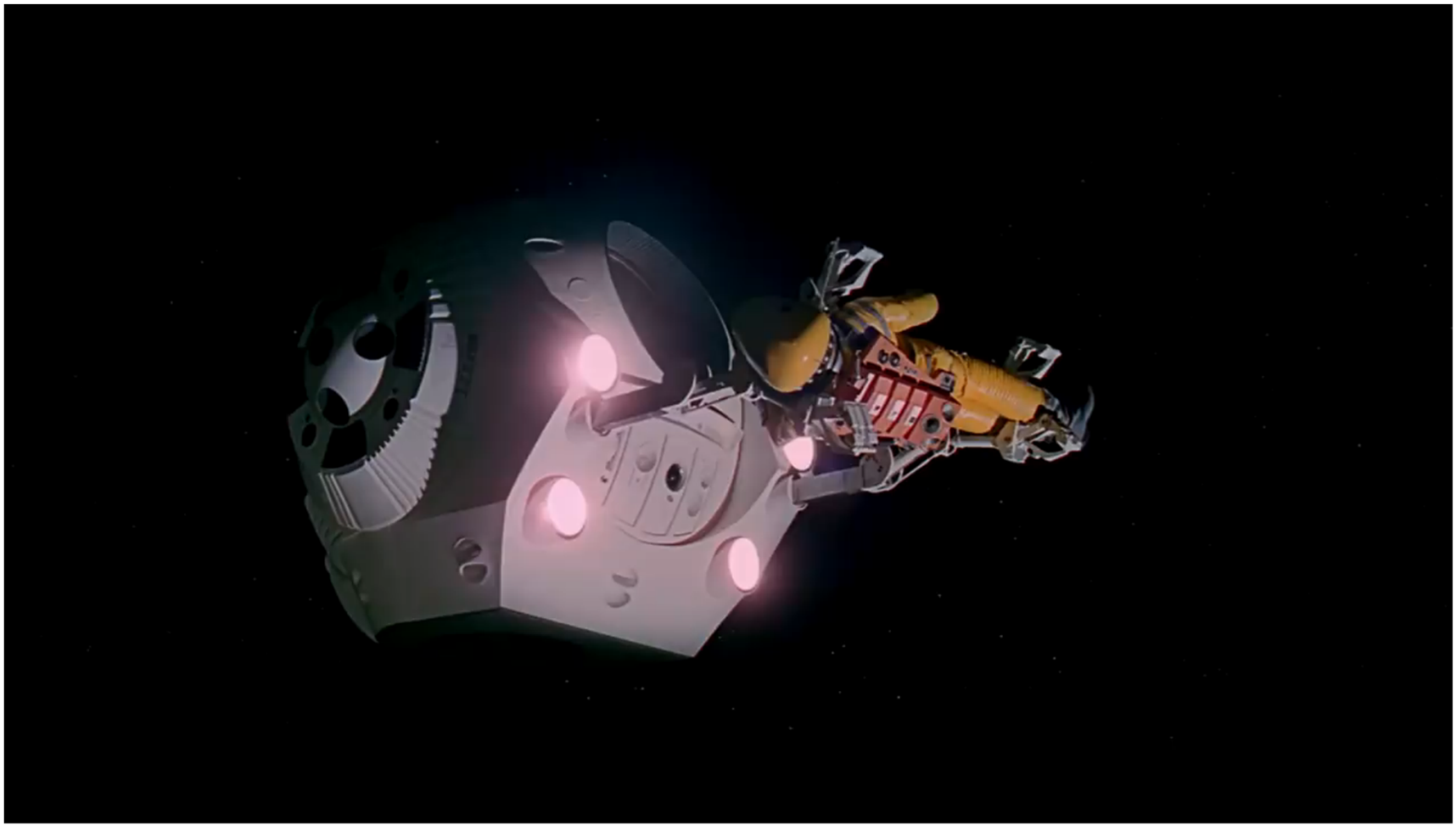
Outline

1. Introduction
2. Input: Speech recognition
3. Output: Speech generation / synthesis
4. Interaction: Natural language processing and human language technologies
5. Traditional Command languages

1. Introduction

- The dream of speaking to computers and having computers speak has long lured researchers and visionaries
- 1968 movie “2001: A Space Odyssey” is a science fiction book and movie about the HAL 9000 computer and has set the standard for computers and developers of natural language systems
- 2013 movie “Her” is a science fiction romantic drama film about a man developing a relationship with an AI virtual assistant personified through a female voice
- The reality is more complex
  - *“Where is the closest coffee shop?”*
  - *“Tell John I will be late”*
  - *“Make space in my drive”*

HAL



HER


THE FOLLOWING PREVIEW HAS BEEN APPROVED FOR
APPROPRIATE AUDIENCES
BY THE MOTION PICTURE ASSOCIATION OF AMERICA, INC.
www.filmratings.com
www.mpaa.org

Turing Test

Alan Turing (1950) – "*Can Machines Think*?"

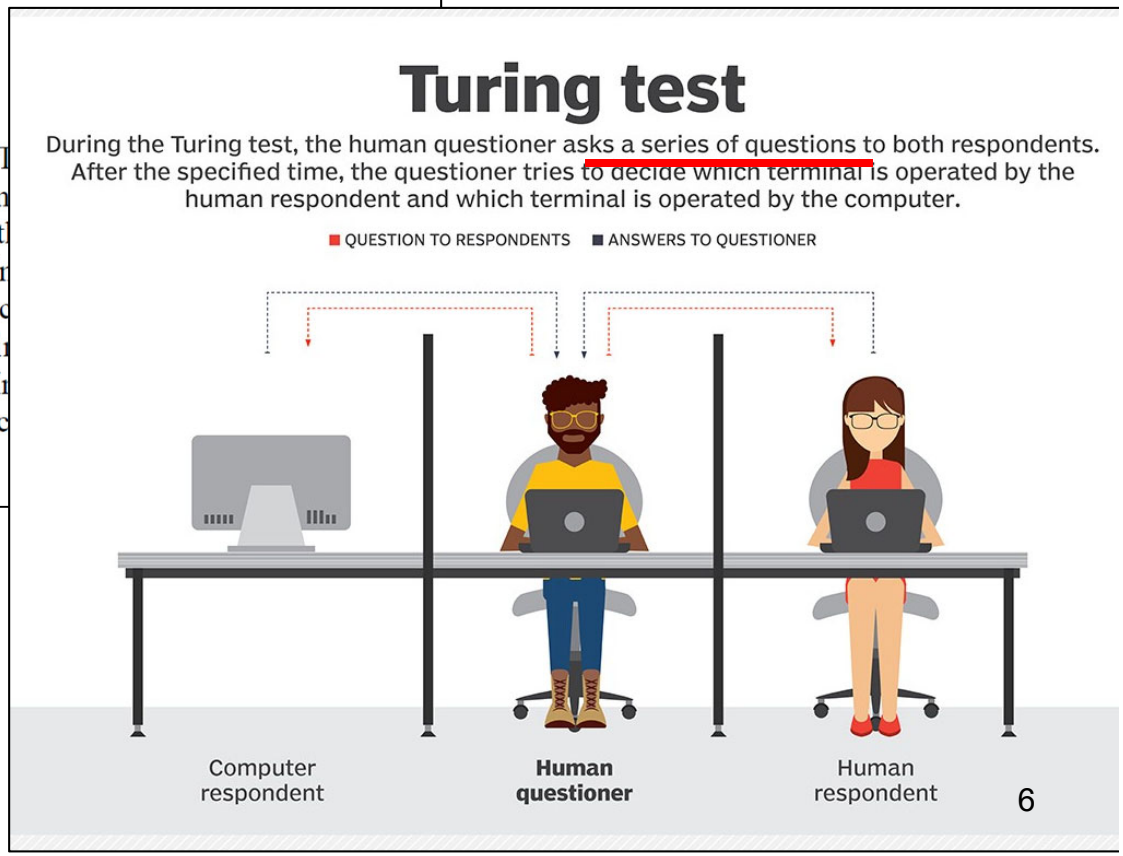
A. M. Turing (1950) Computing Machinery and Intelligence. *Mind 49:* 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" T
definitions of the meaning of the terms "machine" and "thin
framed so as to reflect so far as possible the normal use of t
dangerous, If the meaning of the words "machine" and "thir
examining how they are commonly used it is difficult to esc
meaning and the answer to the question, "Can machines thi
statistical survey such as a Gallup poll. But this is absurd. I
definition I shall replace the question by another, which is c
expressed in relatively unambiguous words.



Branches in AI

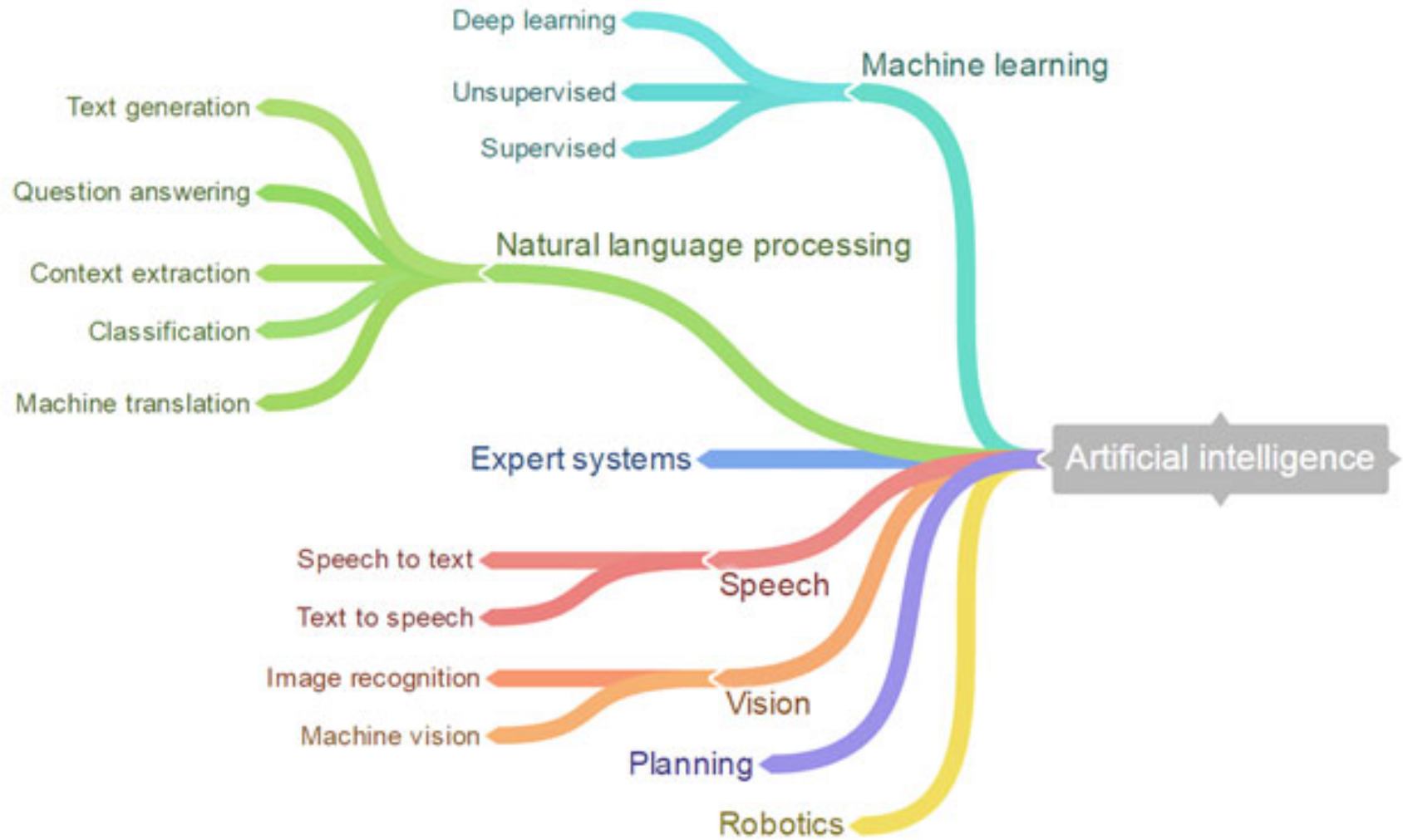
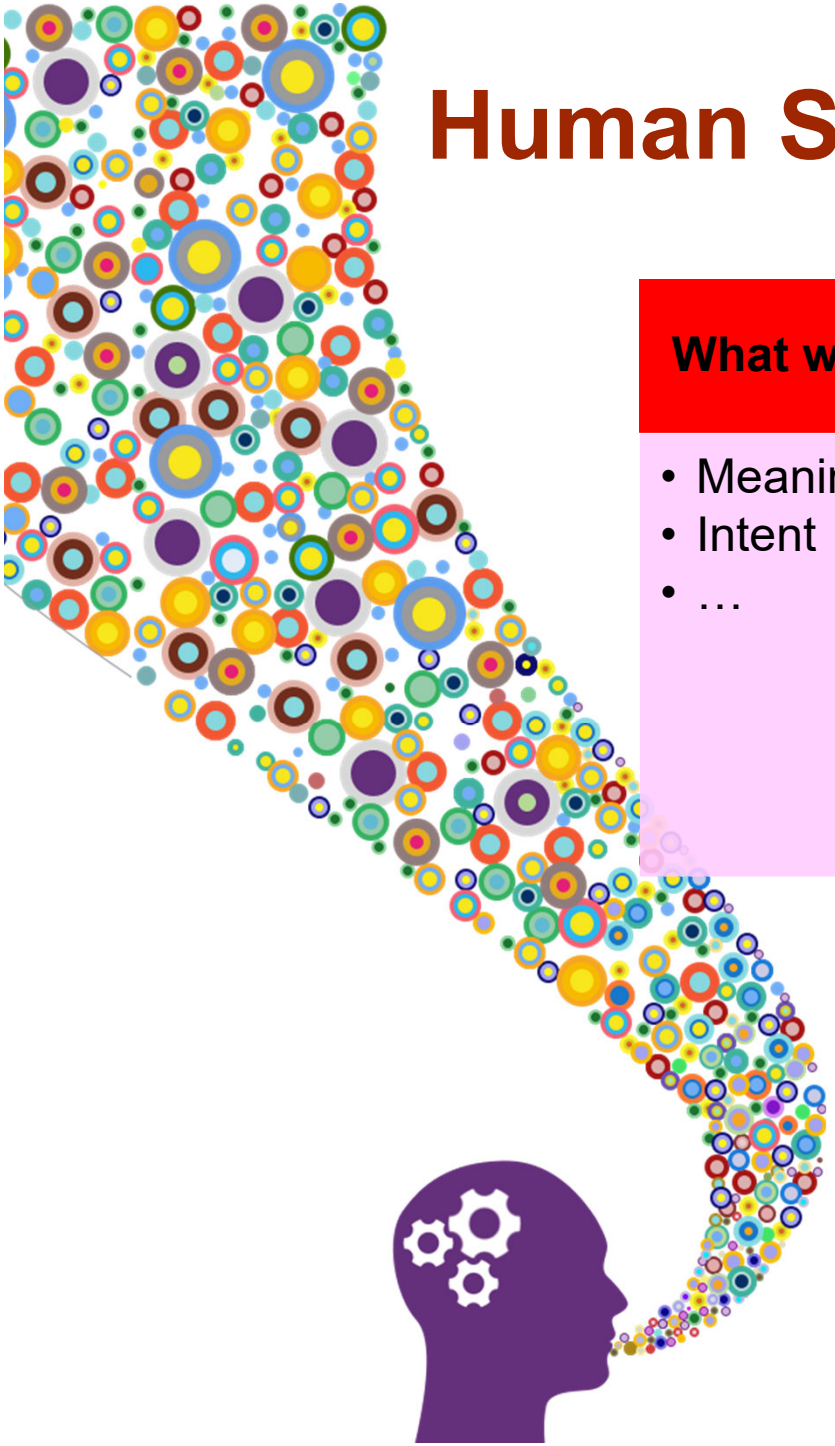


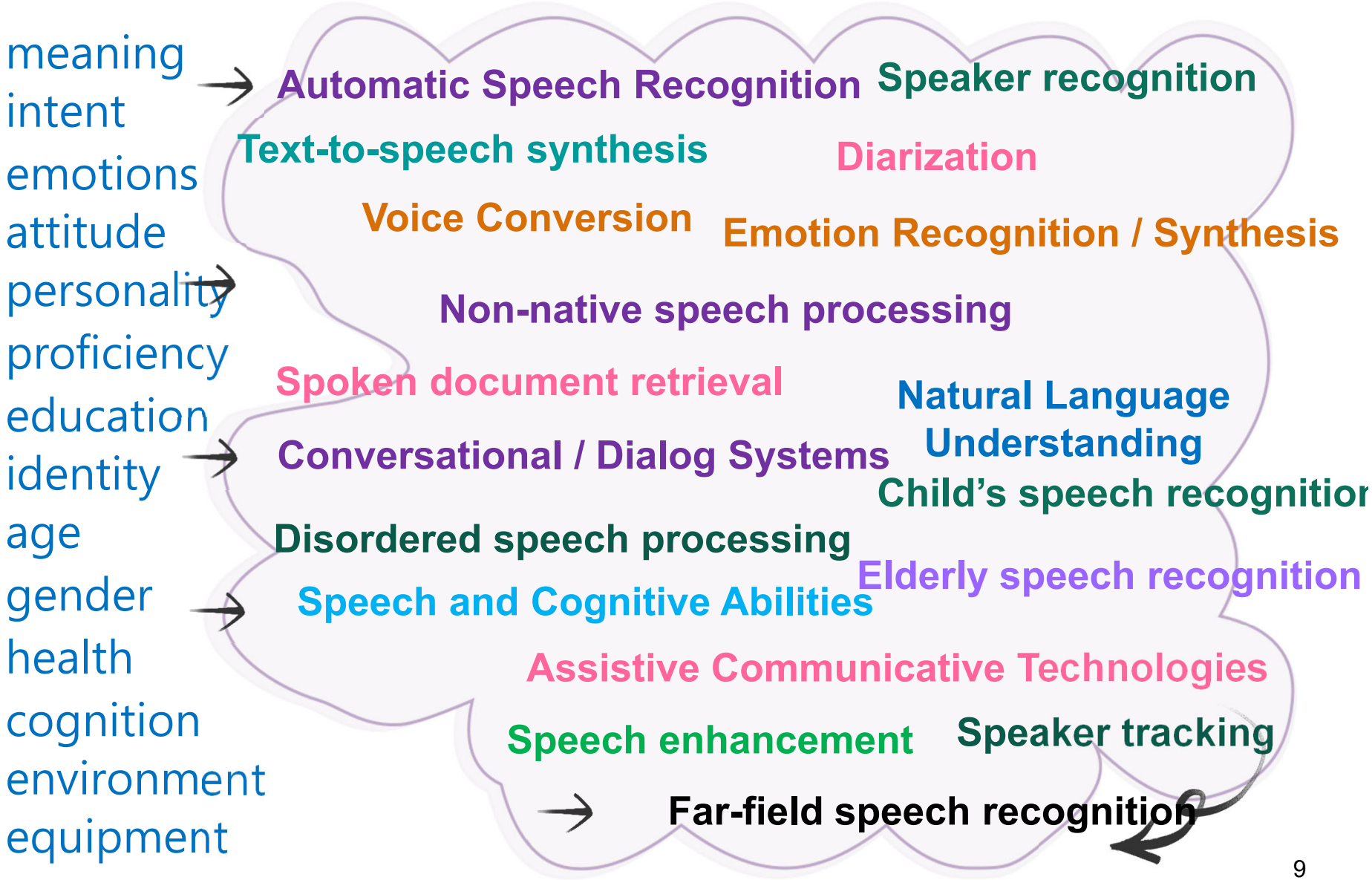
Image source: Google

Human Speech and Language



What we mean	Who we are	How we feel
• Meaning • Intent • …	• Identity • Age • Gender • Personality • Education • …	• Emotion • Attitude • Health • …

AI Technologies for Speech and Language

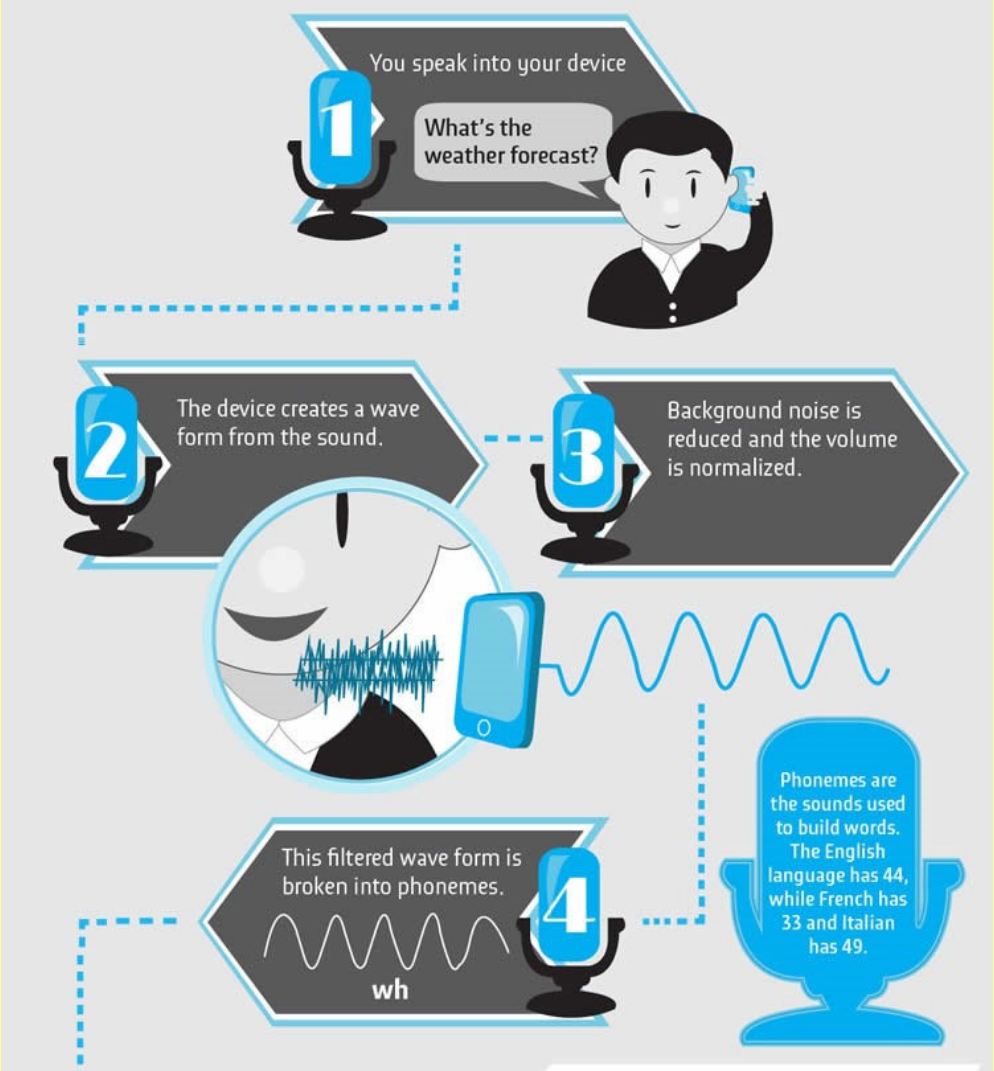


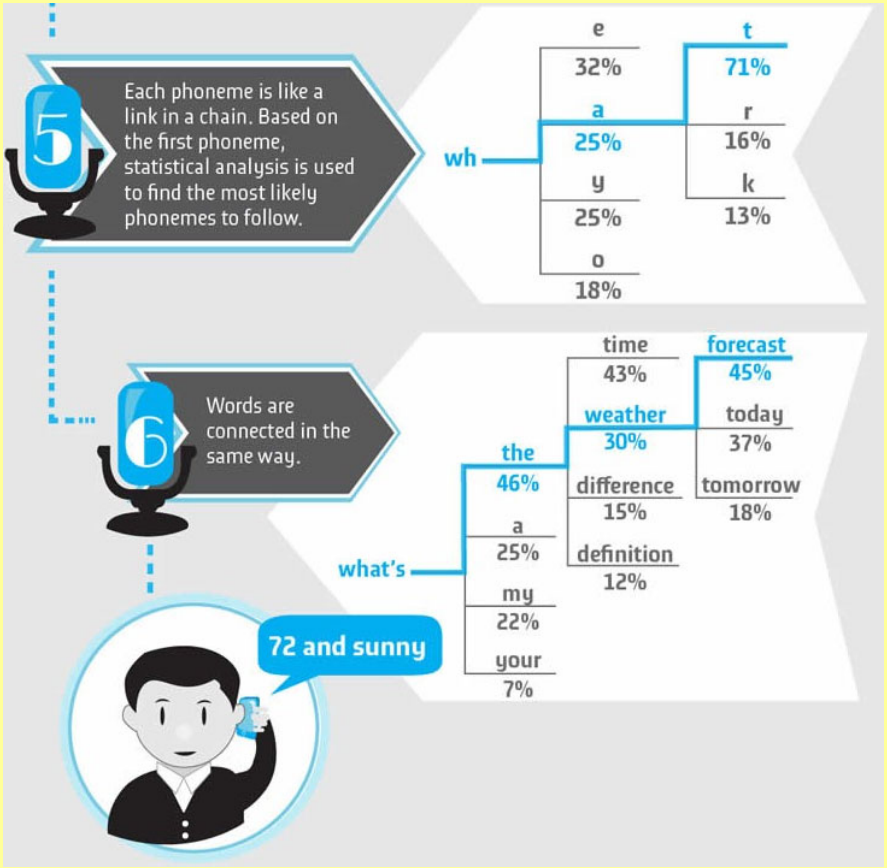
Speech as an Interface Modality

- Natural and efficient
  - Average person speaks 140-160 English words per minute, and types 38-40 words per minute
  - Conveying meaning and intent through what we say and how we speak
- Suitable for users who are hands-busy and/or eyes-busy
- Suitable for users whose mobility is constrained (e.g. due to impairing conditions or physical impairments)
- Suitable for users who is unable to read or write

2. Input: Speech Recognition

- Automatic Speech Recognition (ASR)



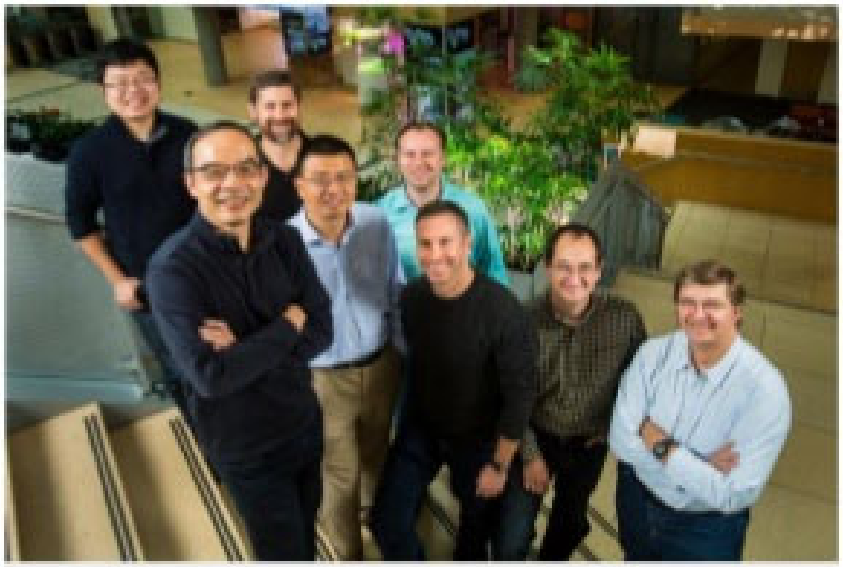


Credits: usabilitygeek

Speech Recognition Breakthrough (2016)

Microsoft's new speech breakthrough

Historic Achievement: Microsoft researchers reach human parity in conversational speech recognition



Microsoft researchers from the Speech & Dialog research group include, from back left, Wayne Xiong, Geoffrey Zweig, Xuedong Huang, Dong Yu, Frank Seide, Mike Seltzer, Jasha Droppo and Andreas Stolcke. (Photo by Dan Deluca)

5.9% word-error rate – Human Parity

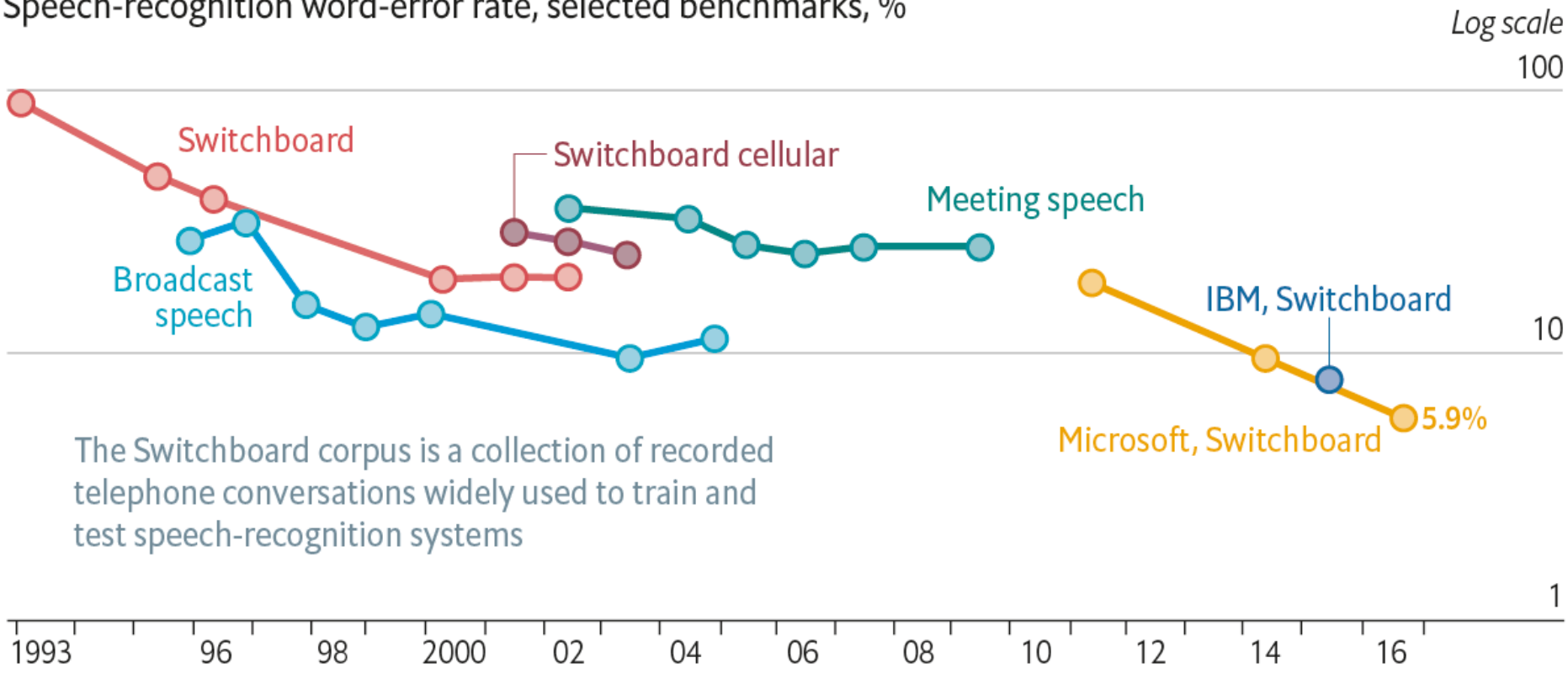
All experiments were run on CNTK

[W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke,
D. Yu, G. Zweig: "The Microsoft 2016 Conversational Speech Recognition System,"
http://arxiv.org/abs/1609.03528]

http://blogs.microsoft.com/next/2016/10/18/historic-achievement-microsoft-researchers-reach-human-parity-conversational-speech-recognition

Loud and clear

Speech-recognition word-error rate, selected benchmarks, %



Sources: Microsoft; research papers

Update: [August 2017] Microsoft announced reaching 5.1% error rate, comparable to professional transcribers

ASR – Example Applications

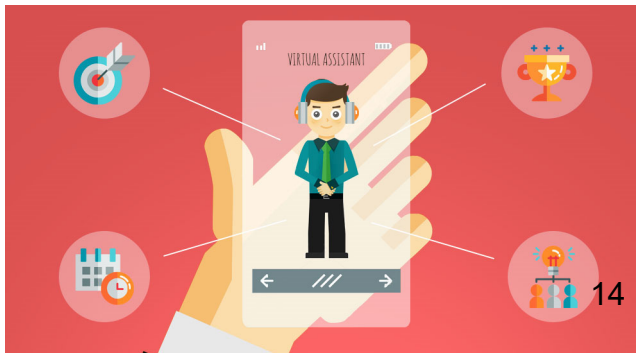
- Voice typing (also known as “dictation” https://www.youtube.com/watch?v=i_Gj2HzEP20

- Audio indexing for information search https://www.youtube.com/watch?v=phGCRNAsPXc

研討會影片檢索系統
Multilingual Symposia Video Indexing System

何鴻燊博士醫療拓展基金會 Dr. Stanley Ho Medical Development Foundation
何鴻燊海量數據決策分析研究中心 The Stanley Ho Big Data Decision Analytics Research Centre

- Hands-free interaction with devices https://www.youtube.com/watch?v=GILvyiWB7xY



Google ASR

Spoken Document Retrieval

研討會影片檢索系統

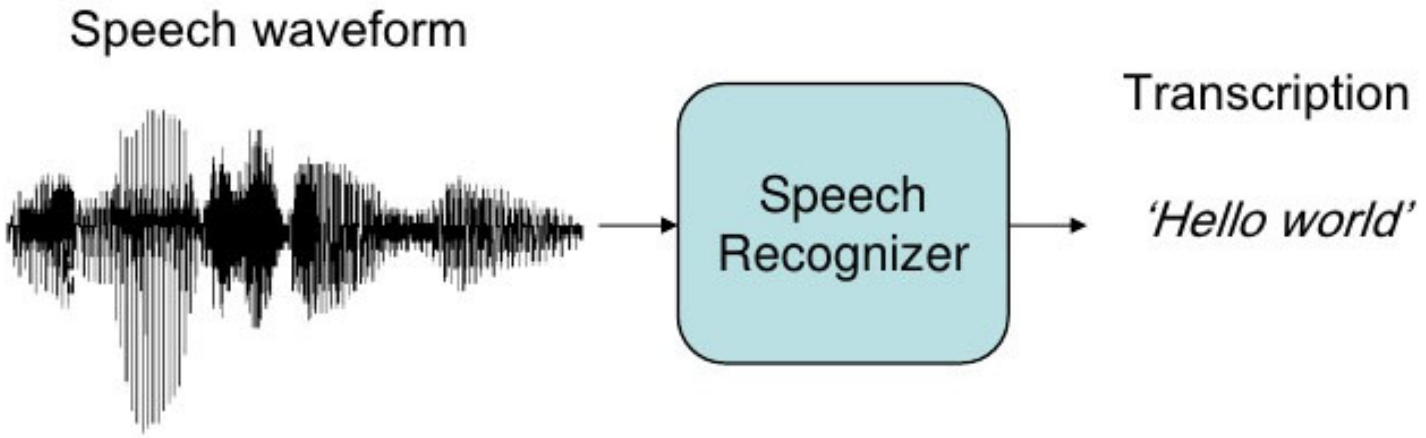
Multilingual Symposia Video Indexing System

何鴻燊博士醫療拓展基金會
Dr. Stanley Ho Medical Development Foundation

何鴻燊海量數據決策分析研究中心
The Stanley Ho Big Data Decision Analytics Research Centre

ASR – Example Applications

Developing ASR for Diverse Users



- Performance of speech recognizers is dependent on the training data
- So they may not work as well for speech types with less data, e.g. accented speech or disordered speech
- Examples in next slide

Developing ASR for Diverse Users

Consider these word error rates (WER):

Language	Non-elderly (WER)	Elderly (WER)	Reference
Japanese	2.34%	4.92%	Kitaoka et al. 2018
French	11.0%	45.7%	Vacher et al. 2015
Portuguese	18.4%	35.3%	Pellegrini et al. 2012
English	36.4%	47.8%	Vipperla et al., 2008

- Does ASR have age bias?
- What about non-native (accented) speech?
- What about disordered speech?
- Other types of speech?

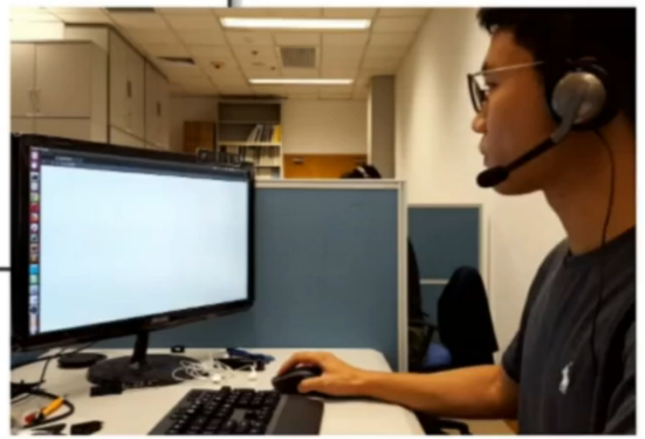
Computer-Aided Pronunciation Training

Problem: mispronunciation detection and diagnosis (MD&D) from non-native learner's speech

香港中文大學
The Chinese University of Hong Kong

Department of Systems Engineering and Engineering Management 香港中文大學 系統工程及工程管理學系

HCCL
Human-Computer Communications Laboratory



Enunciate Demo

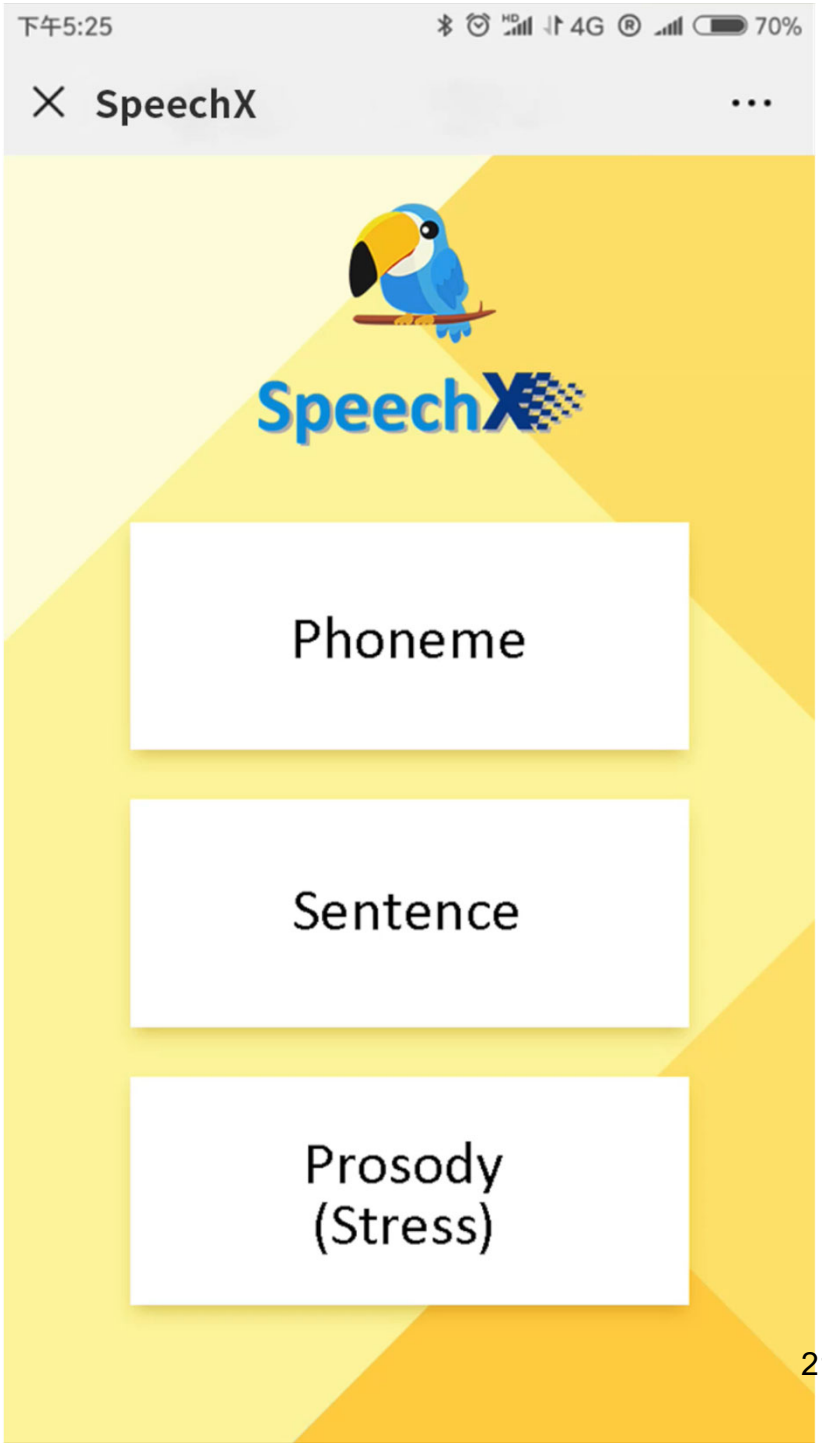
SpeechX

Founders:

Dr. Li Kun, Dr. Lifa Sun

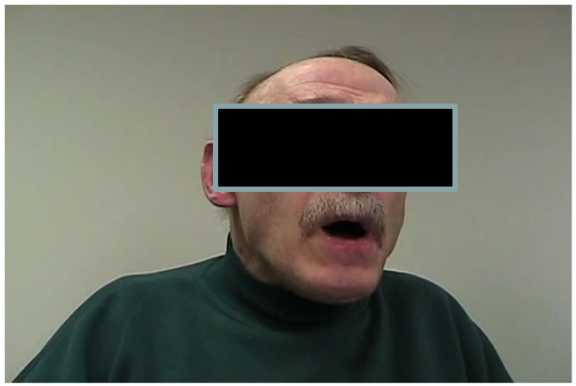
Kun Li, Shao Guang Mao, Xu Li, Zhiyong Wu, Helen Meng, "Automatic Lexical Stress and Pitch Accent Detection for L2 English Speech using Multi-distribution Deep Neural Networks," SPECOM Feb 2018.

Kun Li, Xixin Wu, Helen Meng, "Intonation Classification for L2 English Speech using Multi-distribution Deep Neural Networks," Comp Speech & Lang, May 2017.



Disordered Speech Recognition (DSR)

- Dysarthria
  - Speech disorder due to neuromotor conditions caused by stroke, cerebral palsy, ALS, etc.

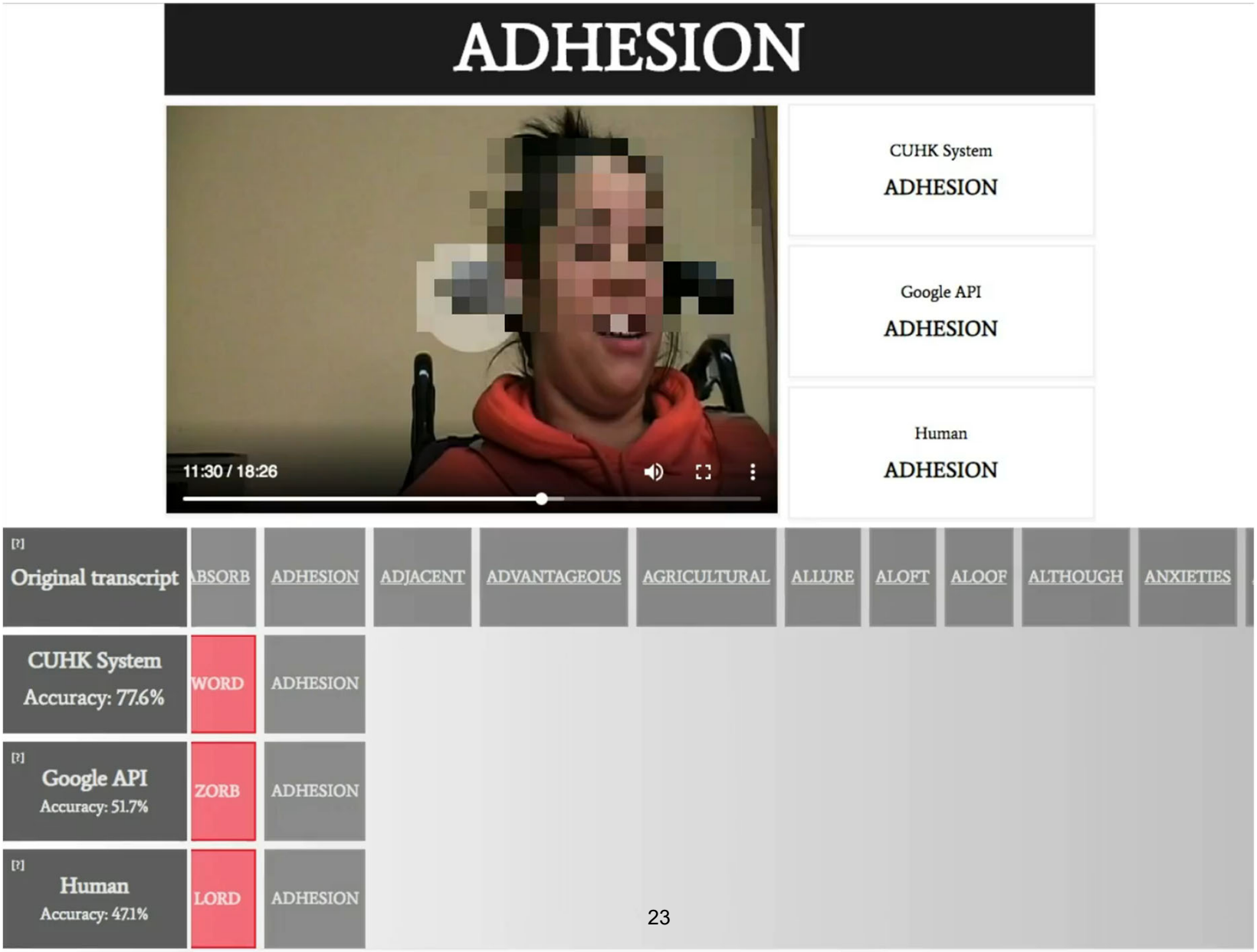


- CUHK's two-pronged approach
  1. Design and collection of CUHK's Dysarthric Speech Corpus – Cantonese Chinese, cultural elements (*since 2015*)

Disordered		*Normal*	

  2. Dysarthric Speech Recognition (DSR) – benchmark on common English corpora (with Professor Xunying Liu)

Dysarthric Speech Recognition



Speech Recognition: Obstacles

- Interference from ambient noise and poor-quality microphones
- Commands need to be learned and remembered
- Recognition challenged by strong accents, specific speech conditions or unusual vocabulary
- Talking is not always acceptable (e.g. in shared office, during meetings)
- Error correction can be time consuming
- Increased cognitive load compared to typing or pointing
- Math or programming difficult without special customization

3. Output: Speech Generation/Synthesis

- Average adult reads at 250-300 words per minute (wpm), cf. preferred listening at 150-160 wpm for comprehension
- But reading (visual modality) may be inconvenient, e.g.
  - User moving around, in-vehicle vibration, illumination conditions, unsuitable visual displays, etc.
- Designers must cope with obstacles of speech
  - Slow pace
  - Ephemeral nature
  - Acceptability and privacy issues in public spaces
  - Difficulty in scanning and searching
  - (commonalities between speech recognition and production)

Speech Production

Canned speech – fixed, digitized speech segments, e.g.

<the next bus will arrive in> <num> <minutes>

https://youtu.be/PeonXfzSEDE

Text-to-speech (TTS) synthesis – several generations of technologies

1. Formant synthesis
2. Concatenated synthesis
3. Parametric synthesis
4. Deep learning synthesis

Listening Test

1. A  B 
2. A  B 
3. A  B 
4. A  B 

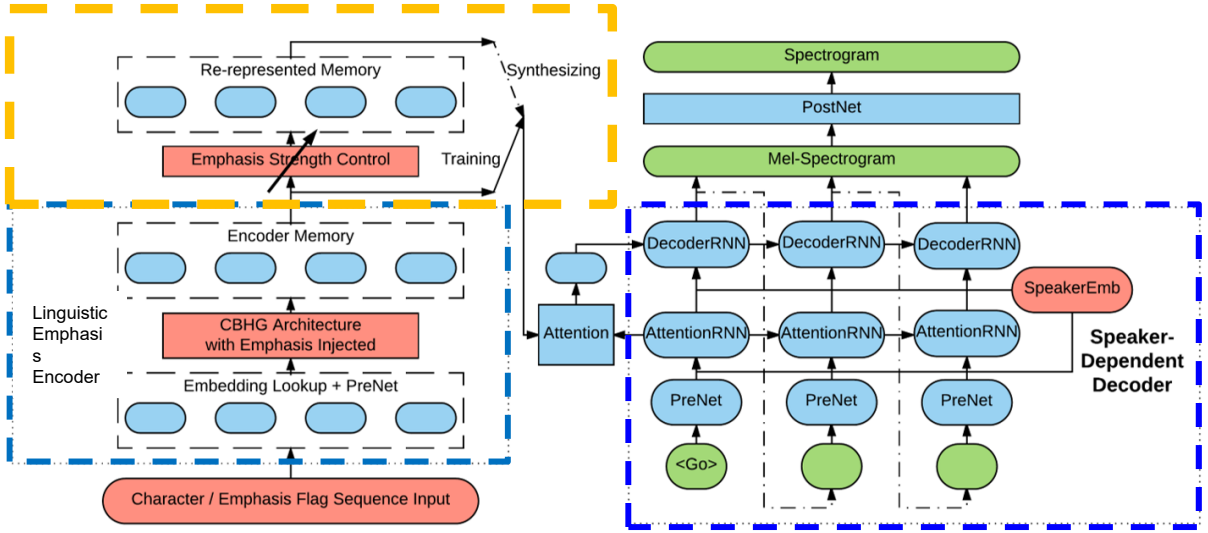
Evaluating Text-to-Speech Synthesis

- Intelligibility
- Naturalness
- Acceptability

Emphasis Generation

Challenge: Controllable Emphatic Speech Synthesis

A number of candidates have consolidated their accommodation with investigation and appreciation.

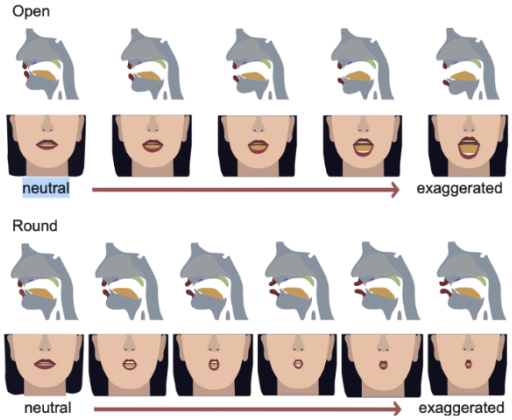


Synthesized Emphatic Speech at Different Controllable Levels					
Alpha=0.0	0.2	0.4	0.6	0.8	1.0

Real Emphatic Recordings (speaker S)	
Neutral	Emphasized

MWang, ZYWu, XXWu, HMeng, SYKang, JJia, LHCai, "Emphatic Speech Synthesis and Control-based Characteristic Transferring End-to-End Speech Synthesis," ACII Asia 2018.

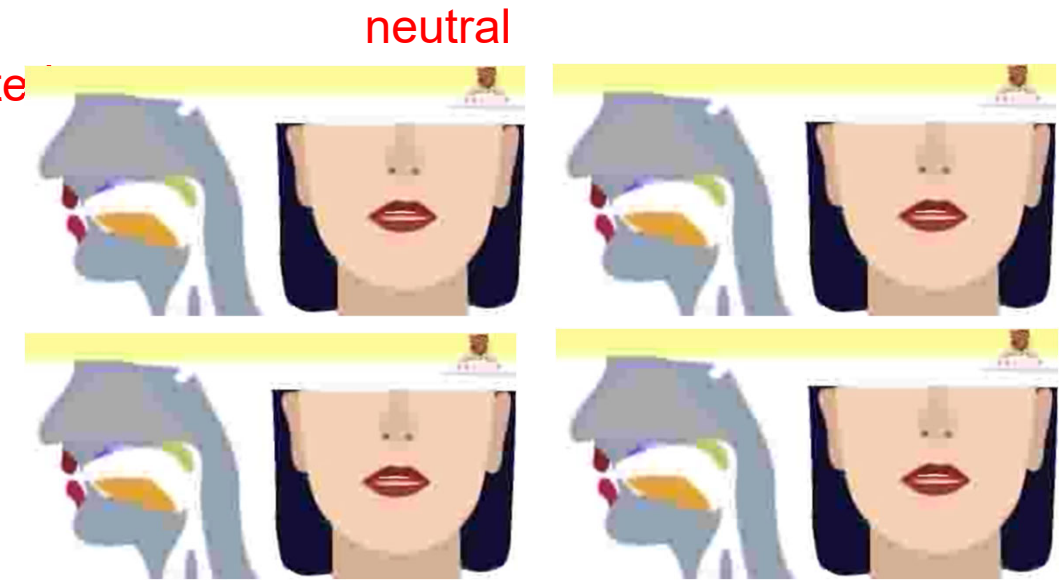
Visual Emphasis Generation



exaggerate

"farmer"

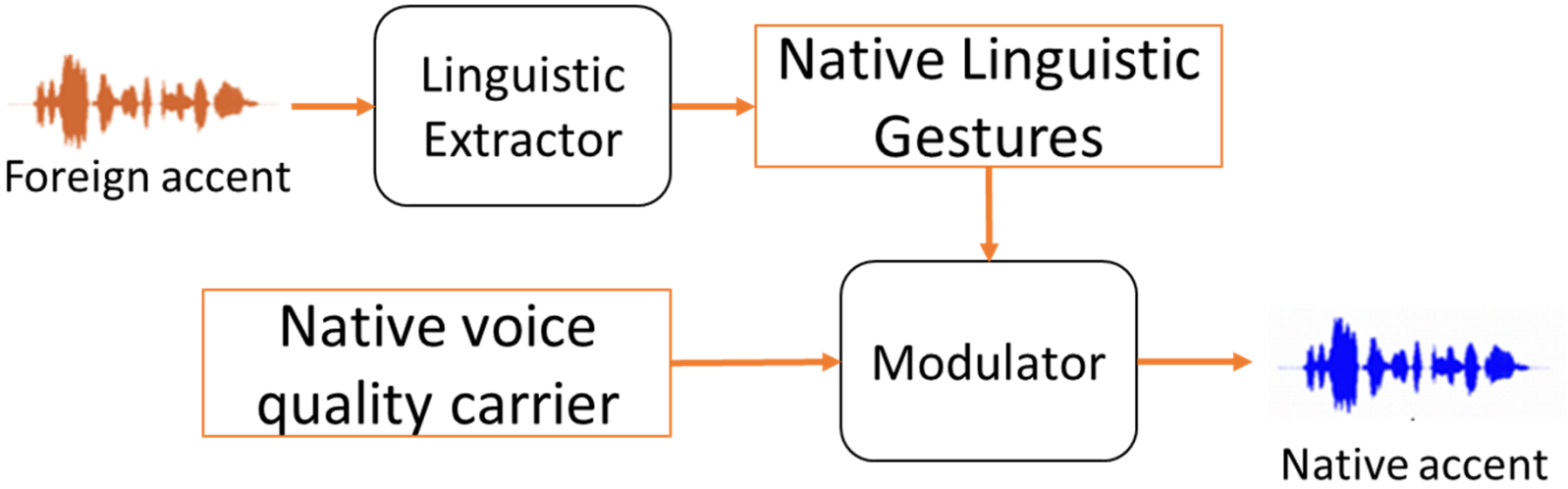
"shoes"



WKLeung, KWYuen, KHWong and HMeng, "Development of Text-to-Audiovisual Speech Synthesis to Support Interactive Language Learning on a Mobile Device" CogInfoCom, 2013

JHZhao, Hyuan, WKLeung, Jliu, SHXia and HMeng, "Audiovisual Synthesis Of Exaggerated Speech For Corrective Feedback In Computer-Assisted Pronunciation Training", Proc. ICASSP 2013.

Accent Reduction



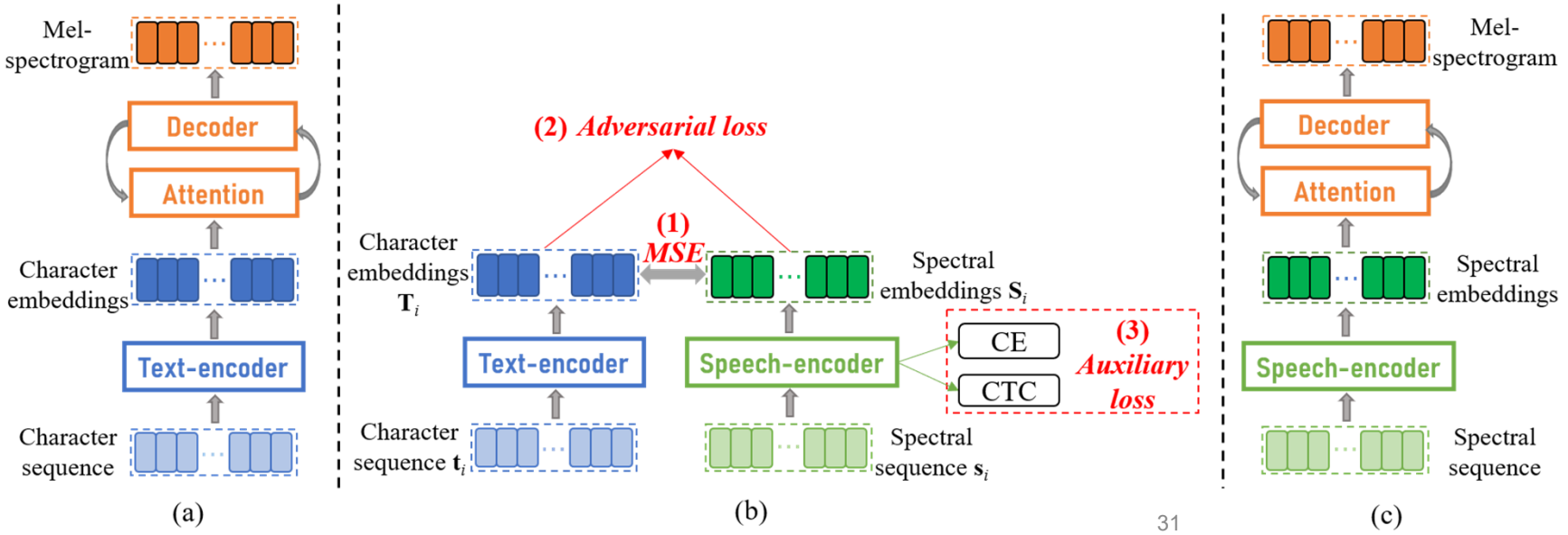
Textual content	Source	Converted
It was more like sugar.		
They are not biologists nor sociologists.		
Without them he could not run his empire.		

SXLiu, DSWang, YWCao, LFSun, XXWu, SYKang, ZYWu, ZYLiu, DSu, DYu, HMeng, "End-to-end Accent Conversion without using Native Utterances", IEEE ICASSP 2020.

Dysarthric Speech Reconstruction

Cross-modal knowledge distillation directly reconstructs disordered speech to become normal

Textual Content	Dysarthric	Reconstructed
paragraph		
sentence		
astounded		



DSWang, JWYu, XXWu, SXLiu, LFSun, XYLiu, HMeng, "End-to-end Voice Conversion via Cross-Modal Knowledge Distillation for Dysarthric Speech Reconstrction, ICASSP 2020.

TTS Applications

- Screen-readers and audio books https://www.youtube.com/watch?v=-gTn4Q-9Lk8&feature=emb_logo (other example: CUHK E-Commu-Book)
- Automatic announcement systems https://www.youtube.com/watch?v=OTmPw4iy0hk
- Text-to-audiovisual speech synthesis (CUHK's technology)

4. Designing Spoken Interactions

Speech in, speech out

a) User initiation – *wake words*
   a) "Hey Siri", "Wake up", "Alexa"
   b) What happens when there is confusions?
   c) https://www.cnet.com/how-to/we-tested-all-4-amazon-echo-wake-words-heres-what-we-learned-about-alexa/

b) Knowing what to say

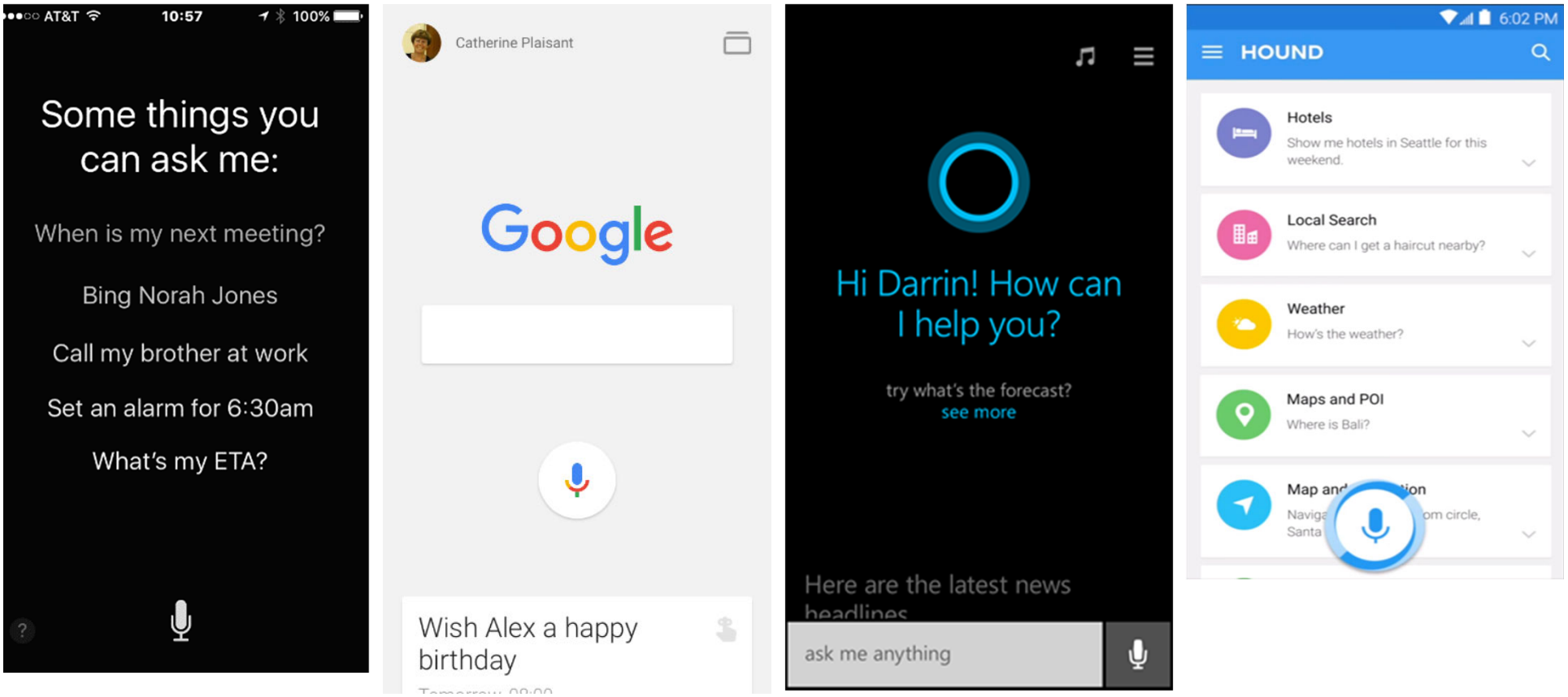
   System: "What service do you need?"

   User: ("account balance" / "bill pay" / "fund transfer")

   Systems: <confirm user's choice, prompt next question>

   – Example (next page)

Example of virtual assistants presenting suggestions



Mobile devices assistants (from left to right: Siri, GoogleNow, Cortana and Hound) all have similar microphone buttons, but different ways of presenting suggestions

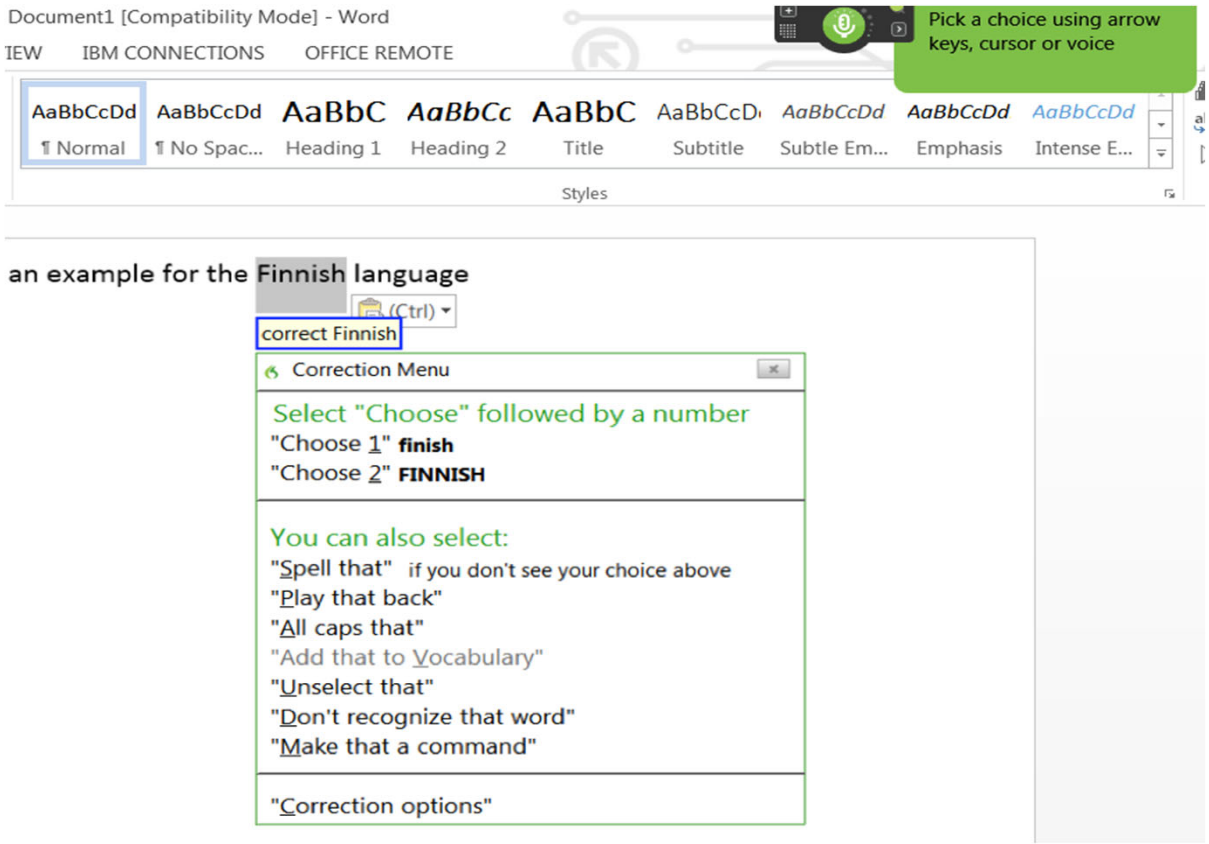
Designing Spoken Interactions

c) Recognition errors
- Confusable words, e.g. *"Boston"* vs *"Austin"*
- Users may stutter, e.g. *"I….I want a chocolate I mean strawberry ice-cream"*
- Errors may affect subsequent interactions

d) Correcting errors
- Pronunciation, grammar corrected by *voice commands*
- ASR needs to separate voice commands from content
- https://www.youtube.com/watch?v=gJ5KCdvBOOo (start at 2:13)
- Example* (see pp36)

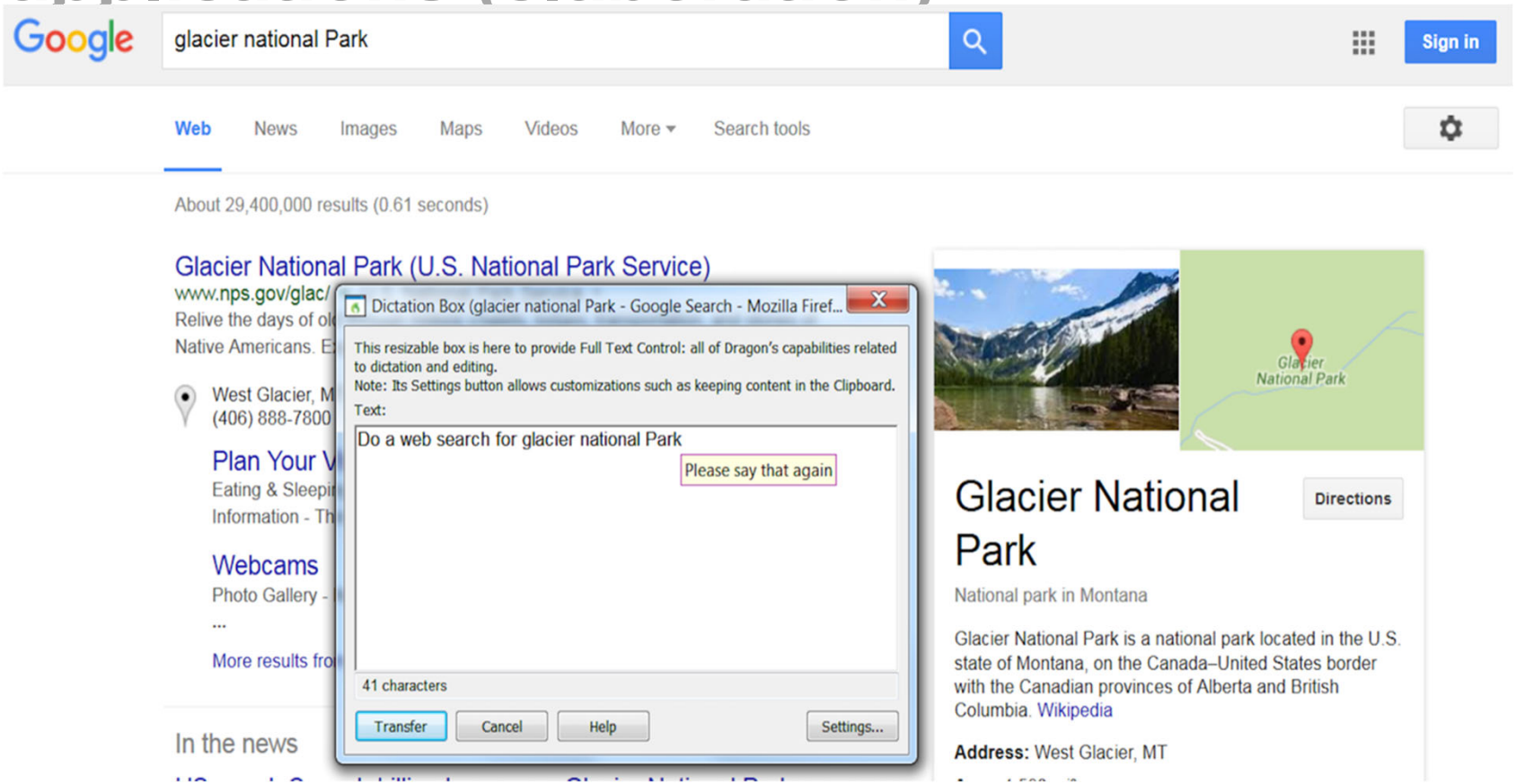
e) Mapping to possible actions
- Dependent on application domains
- Example** (see pp37)

Example*: correcting errors (elaboration)



- Correcting a word during dictation using Nuance Dragon™.
- After saying “Correct finish”, the word is selected and possible corrections displayed, along with additional commands “Spell that”
- Users can use cursor, arrow keys, or voice to choose

Example**: mapping spoken requests to applications (elaboration)



- May be difficult to remember what exact command will accomplish the task *"Search the web for Glacier National Park"* vs. *"Do a web search for Glacier National Park"*

Designing Spoken Interactions

f) Feedback and dialogs

- Confirmation, clarification, execution of commands

User: "Please book a flight from Hong Kong to Boston."
System: You wish to fly from Hong Kong? Please say "yes" or "no".
User: Yes
System: You wish to fly to Austin? Please say "yes" or "no".
User: No. I want to fly to Boston.
System: You wish to fly to Boston? Please say "yes" or no".
User: Yes

- Avoid over-confirmation, consider critical places for confirmation, otherwise execute commands directly, e.g.

"I am ready to delete the file accounts.doc. Should I go ahead?"

Designing Spoken Interactions

g) Designing spoken prompts and commands

- Designing a set of clear and effective commands

```
give me help
give me help on commands
[ ( go | move ) ] ( ( ( back | backward | backwards ) | ( forward | forwards ) ) | ( up | down ) ) ( one | a ) line
[ ( go | move ) ] ( ( ( back | backward | backwards ) | ( forward | forwards ) ) | ( up | down ) ) ( twenty | ... ) lines
( go | move ) ... [ ( ( one | one ) | ( twenty | ... ) ) ]
[ ( go | move ) ] ( ( left | right ) | ( ( back | backward | backwards ) | ( forward | forwards ) ) ) ( one | a ) character
[ ( go | move ) ] ( ( left | right ) | ( ( back | backward | backwards ) | ( forward | forwards ) ) ) ( twenty | ... ) characters
( go | move ) to [ the ] ( bottom | end )
( go | move ) to [ the ] ( bottom | end ) of [ the ] ( line | document )
( go | move ) to [ the ] ( start | top | beginning )
( go | move ) to [ the ] ( start | top | beginning ) of [ the ] ( line | document )
go to sleep
go_to_sleep
help me
```

A rich set of commands used in the Nuance Dragon™ speech recognition system. Synonyms are included and used consistently

Designing Spoken Interactions

g) Designing spoken prompts and commands

- Designing a set of clear and effective commands

```
give me help
give me help on commands
[ ( go | move ) ] ( ( ( back | backward | backwards ) | ( forward | forwards ) ) | ( up | down ) ) ( one | a ) line
[ ( go | move ) ] ( ( ( back | backward | backwards ) | ( forward | forwards ) ) | ( up | down ) ) ( twenty | ... ) lines
( go | move ) ... [ ( ( one | one ) | ( twenty | ... ) ) ]
[ ( go | move ) ] ( ( left | right ) | ( ( back | backward | backwards ) | ( forward | forwards ) ) ) ( one | a ) character
[ ( go | move ) ] ( ( left | right ) | ( ( back | backward | backwards ) | ( forward | forwards ) ) ) ( twenty | ... ) characters
( go | move ) to [ the ] ( bottom | end )
( go | move ) to [ the ] ( bottom | end ) of [ the ] ( line | document )
( go | move ) to [ the ] ( start | top | beginning )
( go | move ) to [ the ] ( start | top | beginning ) of [ the ] ( line | document )
go to sleep
go_to_sleep
help me
```

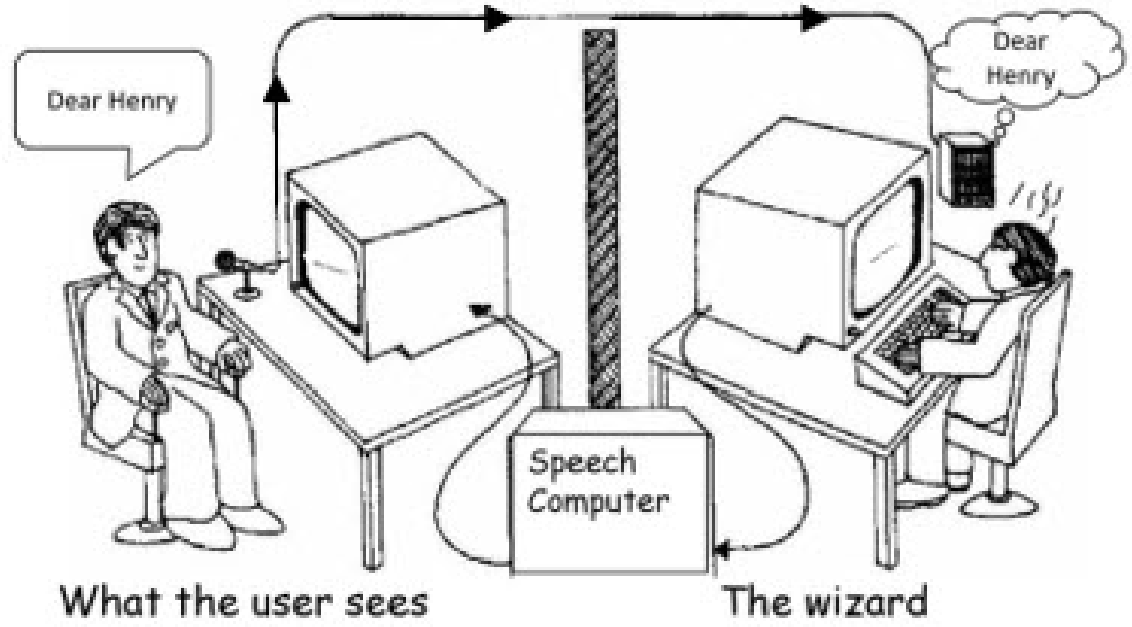
A rich set of commands used in the Nuance Dragon™ speech recognition system. Synonyms are included and used consistently

Designing Spoken Interactions

Another approach: Wizard-of-Oz method

- Allows researchers to test a concept even before development

Wizard of Oz testing – The listening type writer IBM 1984



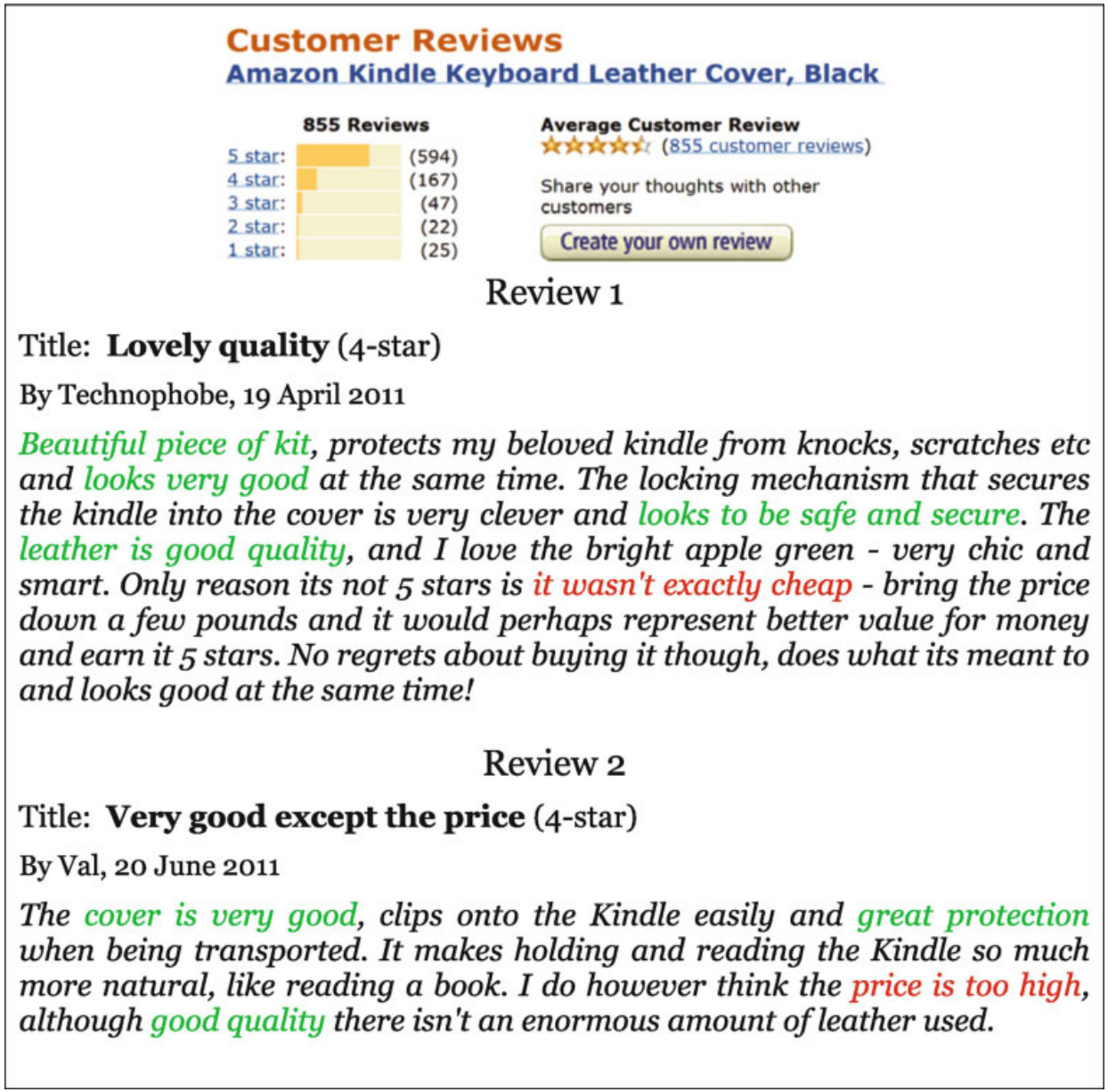
Source: IBM

5. Human Language Technology

- Natural Language Interaction (NLI)
  - Textual input and output

- Many example applications
  - E.g. Sentiment analysis, instructional systems, machine translation, language modelling, caption generation, chatbots, caption generation (see Week 1), etc.

Sentiment Analysis



Customer Reviews

Amazon Kindle Keyboard Leather Cover, Black

855 Reviews

5 star: (594)
4 star: (167)
3 star: (47)
2 star: (22)
1 star: (25)

Average Customer Review

(855 customer reviews)

Share your thoughts with other customers

Create your own review

Review 1

Title: **Lovely quality** (4-star)

By Technophobe, 19 April 2011

Beautiful piece of kit, protects my beloved kindle from knocks, scratches etc and looks very good at the same time. The locking mechanism that secures the kindle into the cover is very clever and looks to be safe and secure. The leather is good quality, and I love the bright apple green - very chic and smart. Only reason its not 5 stars is it wasn't exactly cheap - bring the price down a few pounds and it would perhaps represent better value for money and earn it 5 stars. No regrets about buying it though, does what its meant to and looks good at the same time!

Review 2

Title: **Very good except the price** (4-star)

By Val, 20 June 2011

The cover is very good, clips onto the Kindle easily and great protection when being transported. It makes holding and reading the Kindle so much more natural, like reading a book. I do however think the price is too high, although good quality there isn't an enormous amount of leather used.

Source: [Yusof et al. 2018]

Immersive Training Applications

Source: DTUI

- Immersive Naval Officer Training System (INOTS) new navy officers practice counseling skills in VR environment
- Speech-based interactions with assessment is facilitated https://www.youtube.com/watch?v=K6l9bJSunRg

- Application in education: https://www.youtube.com/watch?v=QXSsAP-EF9E

Machine Translation

Source: Google Translate

Language Modeling

- Pre-trained language model
- Word prediction
- https://demo.allennlp.org/next-token-lm

5. Traditional Command Languages

- Command languages often preferred by expert users who do not want to drag and drop items repeatedly
- Example: Unix command to delete blank lines from a file
  - grep -v ^$ filea > fileb
- Casual users favor GUIs but both styles of interface can be made available successfully
- Other examples that behave like command languages:
  - Web addresses (URLs) can be seen as a form of command language
  - Twitter addresses
  - Database query languages

Command Languages (concluded)

- Programmers or professionals can memorize hundreds of commands and shortcuts, helping mastery of their application
- Histories can be kept and scripts / macros can help automate actions
- Feedback generation esp error messages (e.g. syntax or typos)
- Auto-completion can prevent errors
- May offer brief menus