01 Speech/Language Technology & Automatic Speech Recognition : Overview

The Motivation

Motivation of Automatic Speech Recognition

Human use “speech” as the major mean of communication with other people.

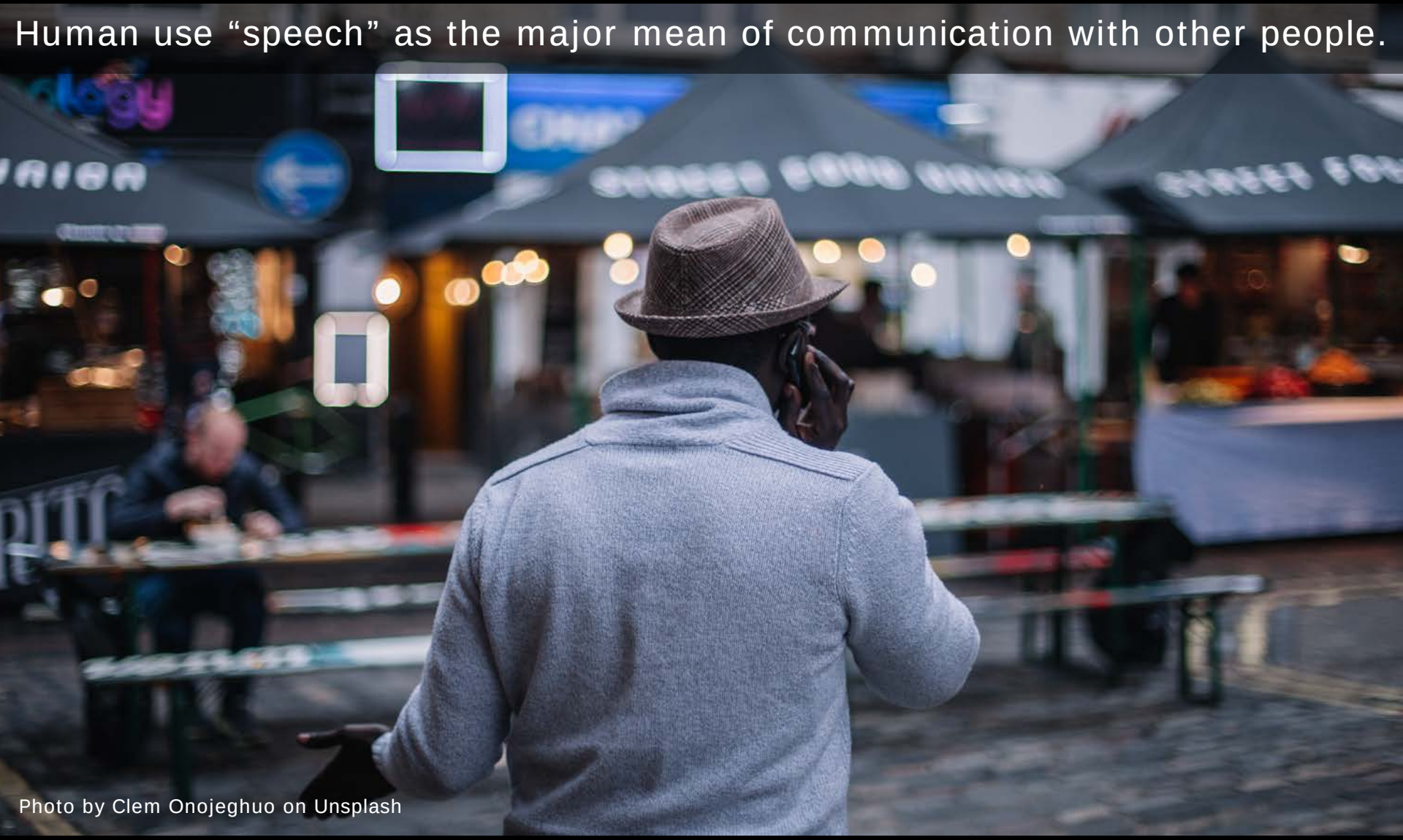


Photo by Clem Onojeghuo on Unsplash

Motivation of Automatic Speech Recognition

Human use “speech” as the major mean of communication with other people.

The goal is to make “Machines” that understand “Speech Communication”

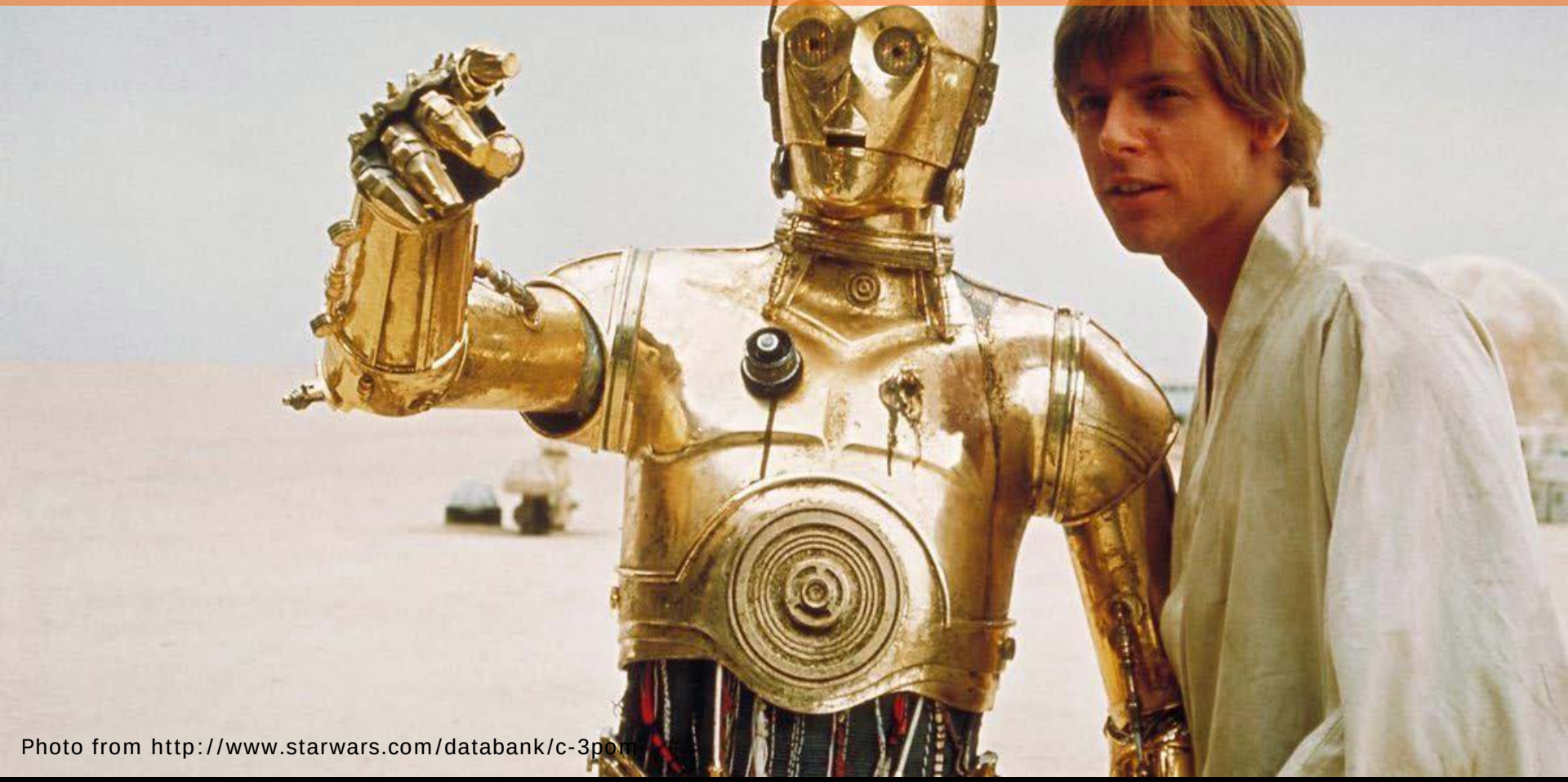


Photo from http://www.starwars.com/databank/c-3po

Speech as Computer Inputs

Photo by Kaitlyn Baker on Unsplash

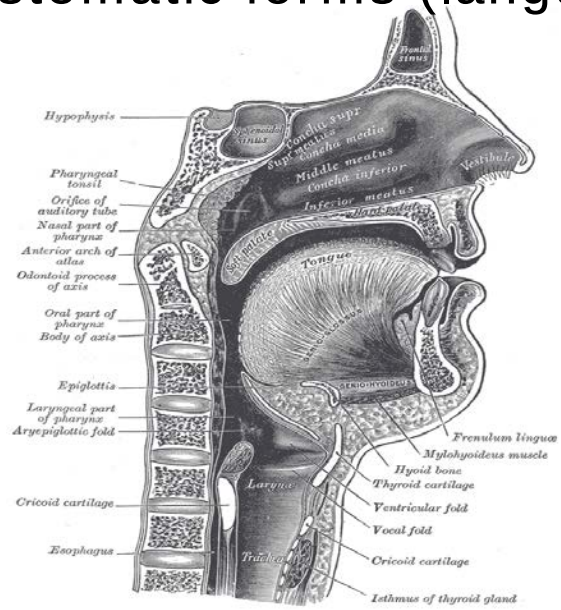
Artificial “Sentient Being”

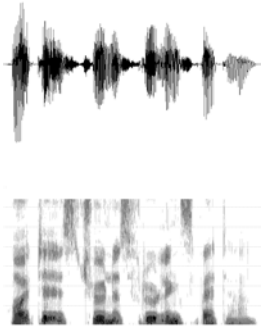
Sentient
= able to perceive or feel things

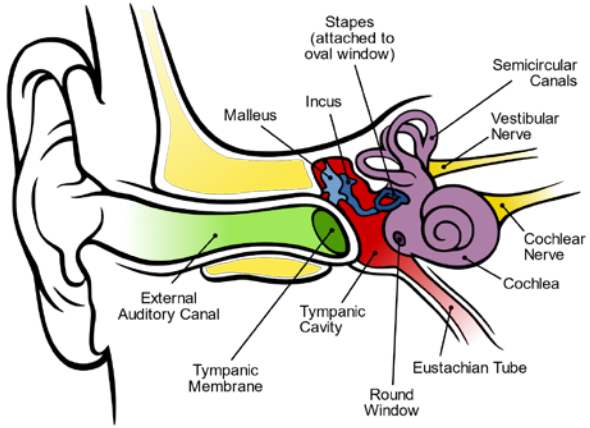
Photo by Atiwong Suchato

Speech Communication

- The brain controls various articulators in the vocal tract to create the sound wave in some systematic forms (language).
- Auditory system sends nervous signal according to the received sound wave to the brain.







Public Domain, https://commons.wikimedia.org/w/index.php?curid=504492

By Chittka L, Brockmann - Perception Space—The Final Frontier, A PLoS Biology Vol. 3, No. 4, e137 doi:10.1371/journal.pbio.0030137 (Fig. 1A/Large version), vectorised by Inductiveload, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=5957984

Behind the Mic: The Science of Talking with Computers

Google
Published on Oct 17, 2014

Language. Easy for humans to understand (most of the time), but not so easy for computers. This is a short film about speech recognition, language understanding, neural nets, and using our voices to communicate with the technology around us.

https://www.youtube.com/watch?v=yxxRAHVtafI

Behind the Mic: The Science of Talking with Computers

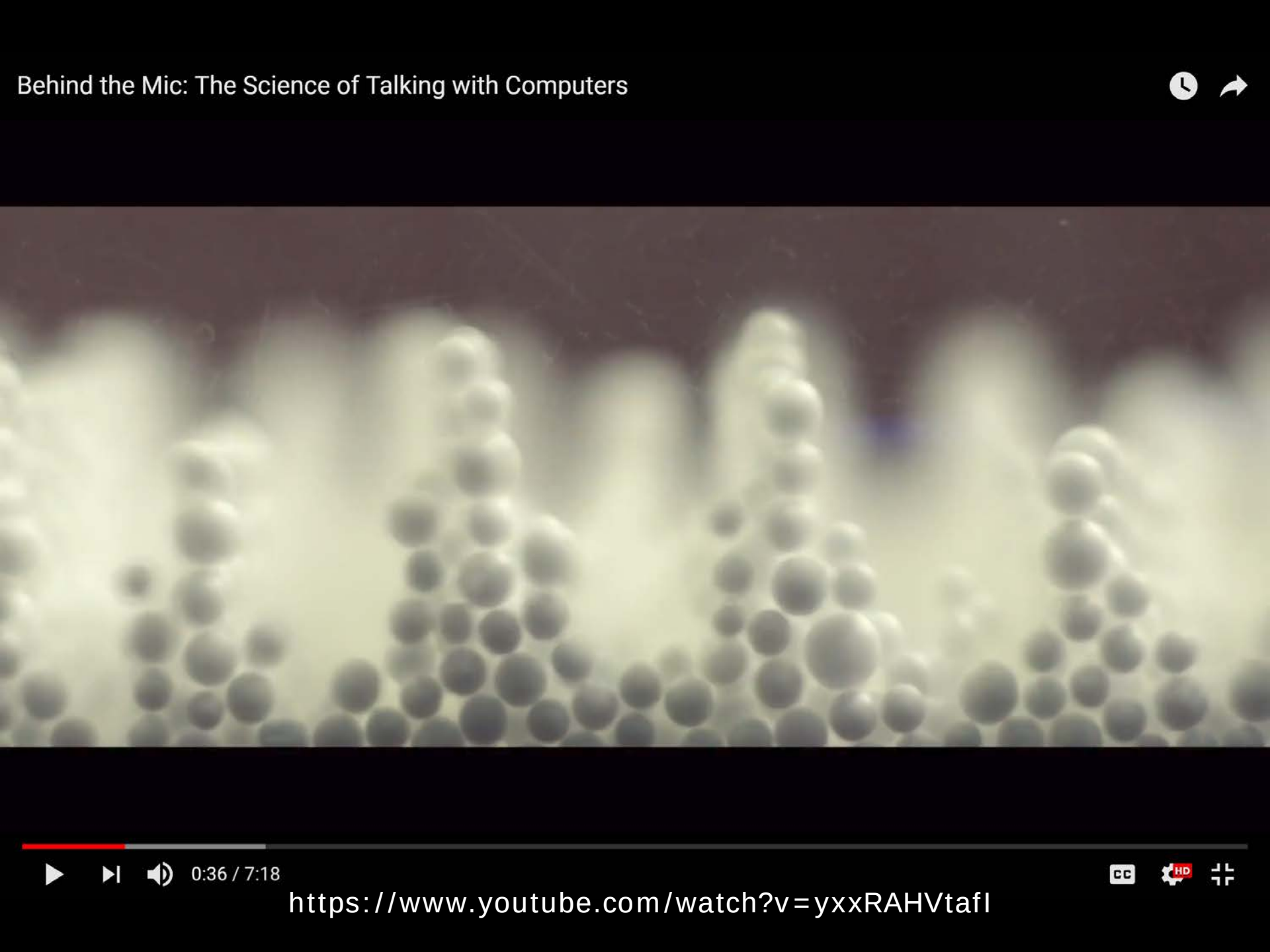
Google
Published on Oct 17, 2014

0:00 – 0:46

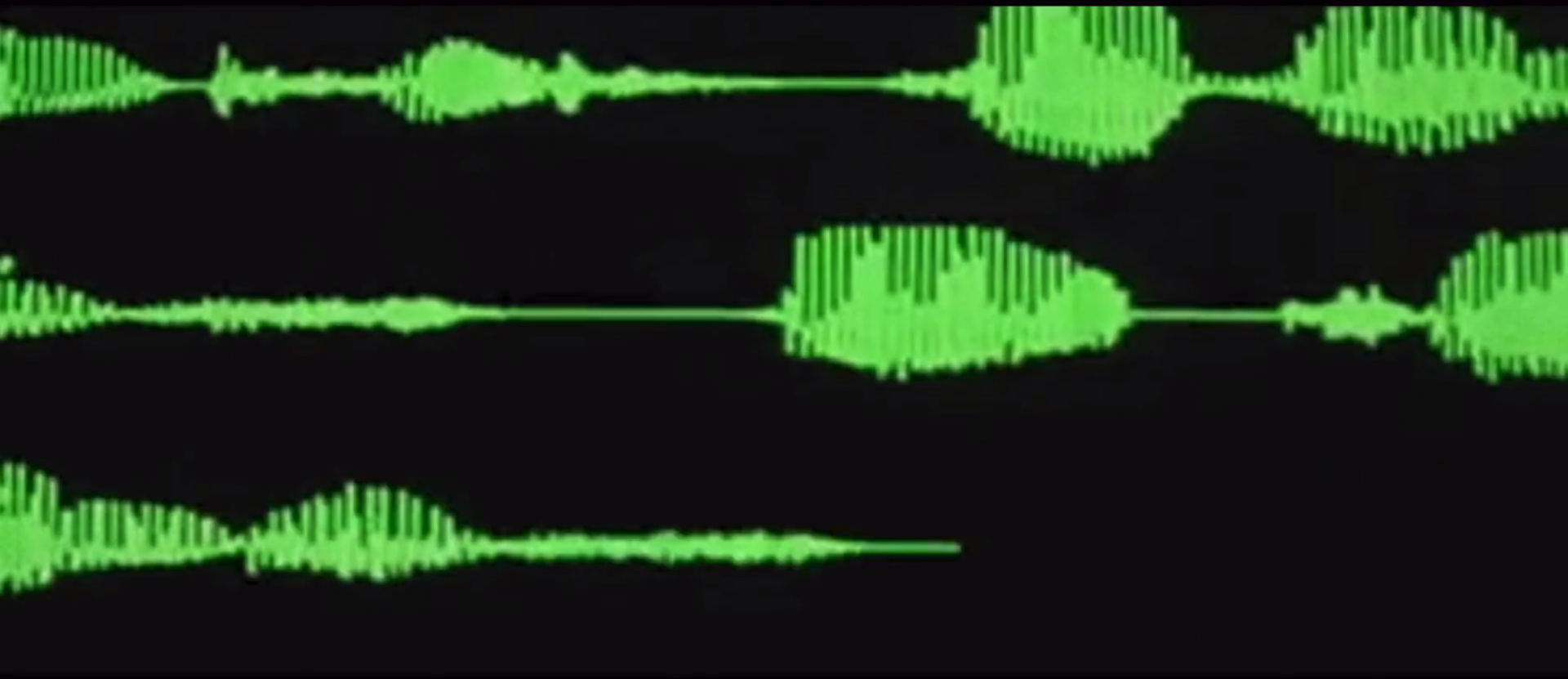
“Speech Communication”

https://www.youtube.com/watch?v=yxxRAHVtafI

https://www.youtube.com/watch?v=yxxRAHVtafI


Behind the Mic: The Science of Talking with Computers
0:36 / 7:18
https://www.youtube.com/watch?v=yxxRAHVtafI

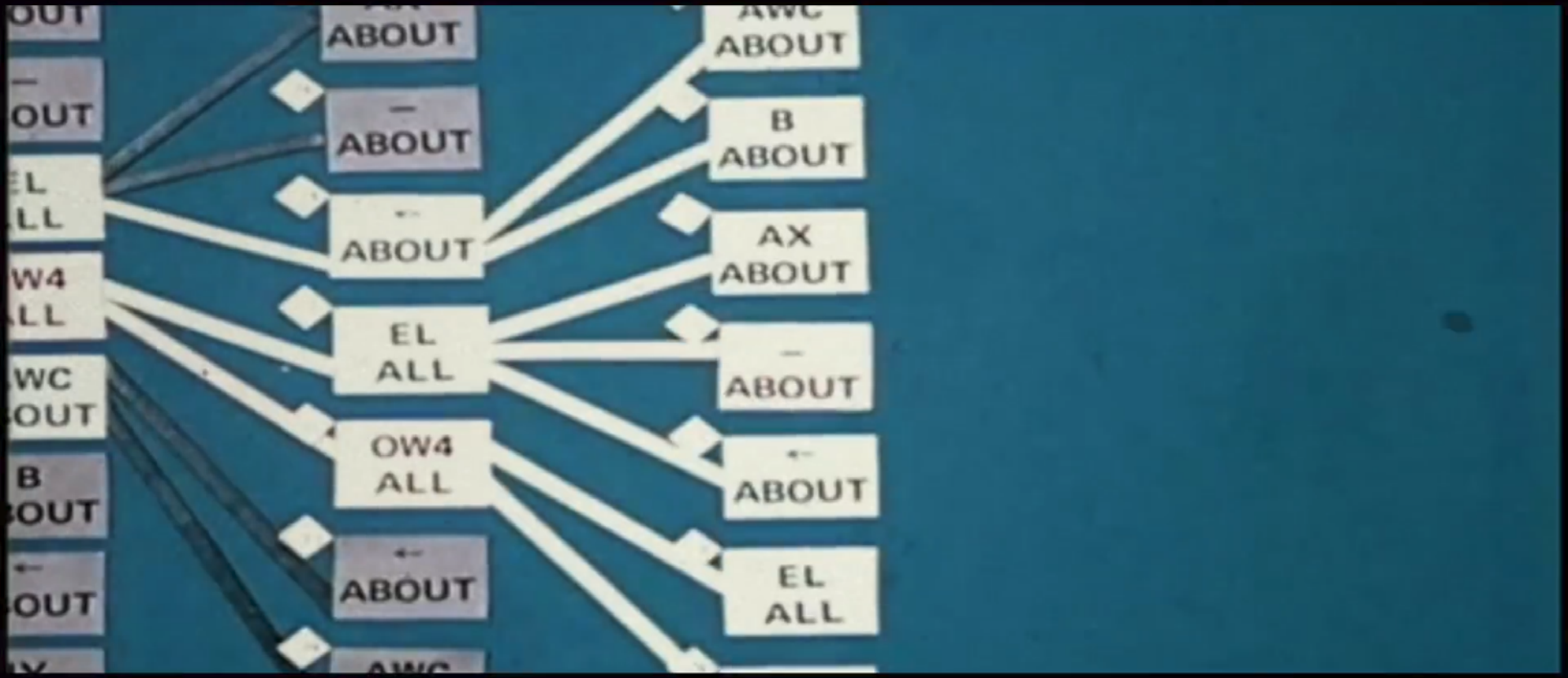
Behind the Mic: The Science of Talking with Computers



0:38 / 7:18

https://www.youtube.com/watch?v=yxxRAHVtafI

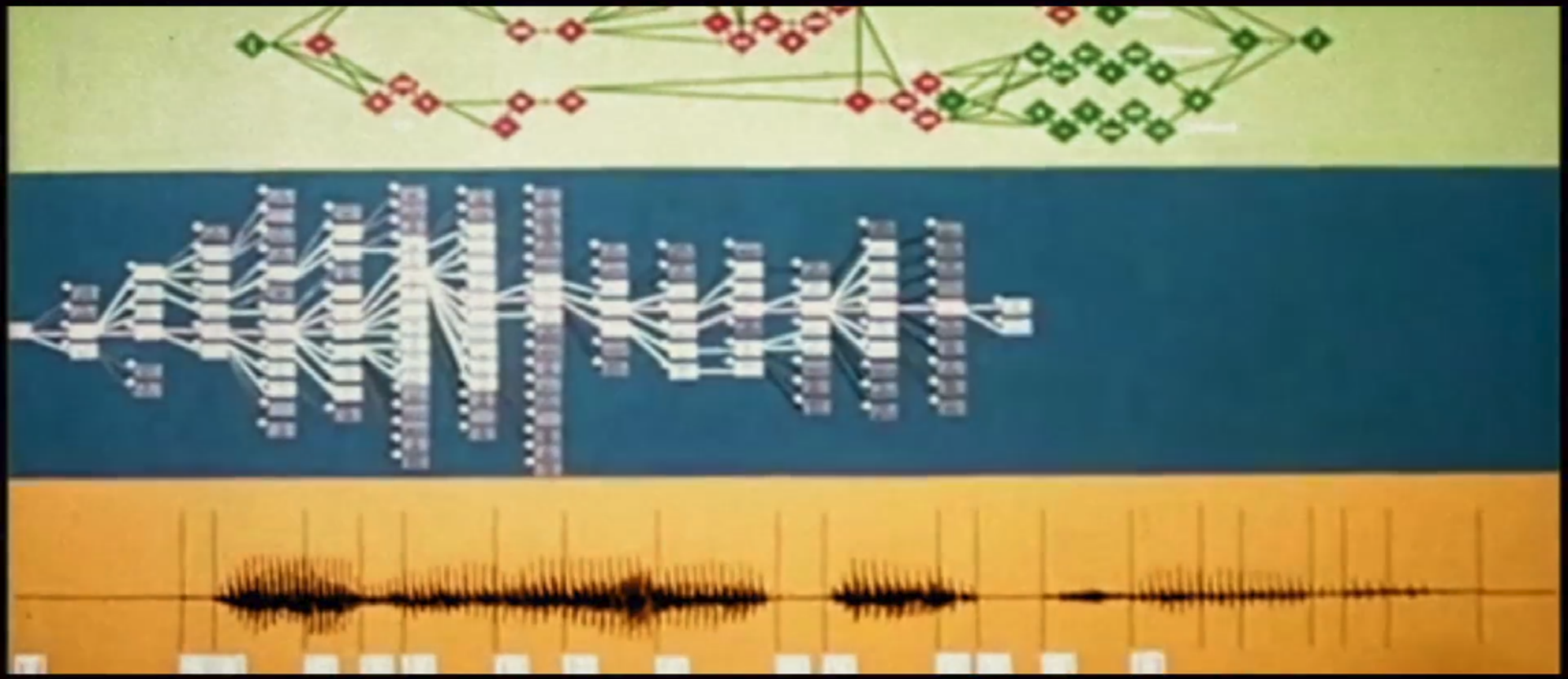
Behind the Mic: The Science of Talking with Computers



0:40 / 7:18

https://www.youtube.com/watch?v=yxxRAHVtafI

Behind the Mic: The Science of Talking with Computers



0:41 / 7:18

https://www.youtube.com/watch?v=yxxRAHVtafI

Behind the Mic: The Science of Talking with Computers

0:42 / 7:18

https://www.youtube.com/watch?v=yxxRAHVtafI

Behind the Mic: The Science of Talking with Computers

Google
Published on Oct 17, 2014

0:48 – 3:00

“Automatic Speech Recognition”
+ The need for “Natural Language Understanding”

https://www.youtube.com/watch?v=yxxRAHVtafI

- Developed since 1952
- Increasing capability:
  - Number of registered speakers
  - Number of Vocab
  - Pronunciation variation
- Hidden Markov Models in the 80's
- General Steps:
  - Raw waveforms → segmenting → identify phonemes
- Transcribing Vs. Understanding

Behind the Mic: The Science of Talking with Computers

Google
Published on Oct 17, 2014

3:01 – 4:16

Natural Language Understanding is hard
(and we are not sure how it works)

https://www.youtube.com/watch?v=yxxRAHVtafI

- Complex
- Challenges in semantics
  - Irony / Emotion / Humor
- Human brains can do:
  - Language understanding
  - Pattern recognition

Behind the Mic: The Science of Talking with Computers

Google
Published on Oct 17, 2014

4:16 – 6:10

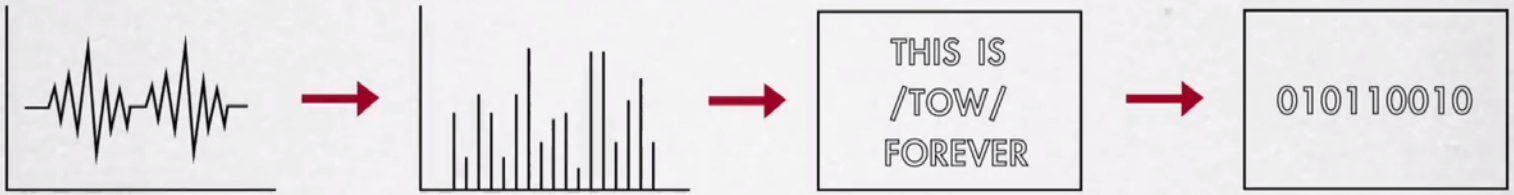
Neural Networks can be (are) the solution (?)

https://www.youtube.com/watch?v=yxxRAHVtafI

- Knowledge = Connection between neurons
- Neural Networks ← Computer simulation of the brains
- Features:
  - Hand-engineered
  - Neural Nets-learned

Behind the Mic: The Science of Talking with Computers

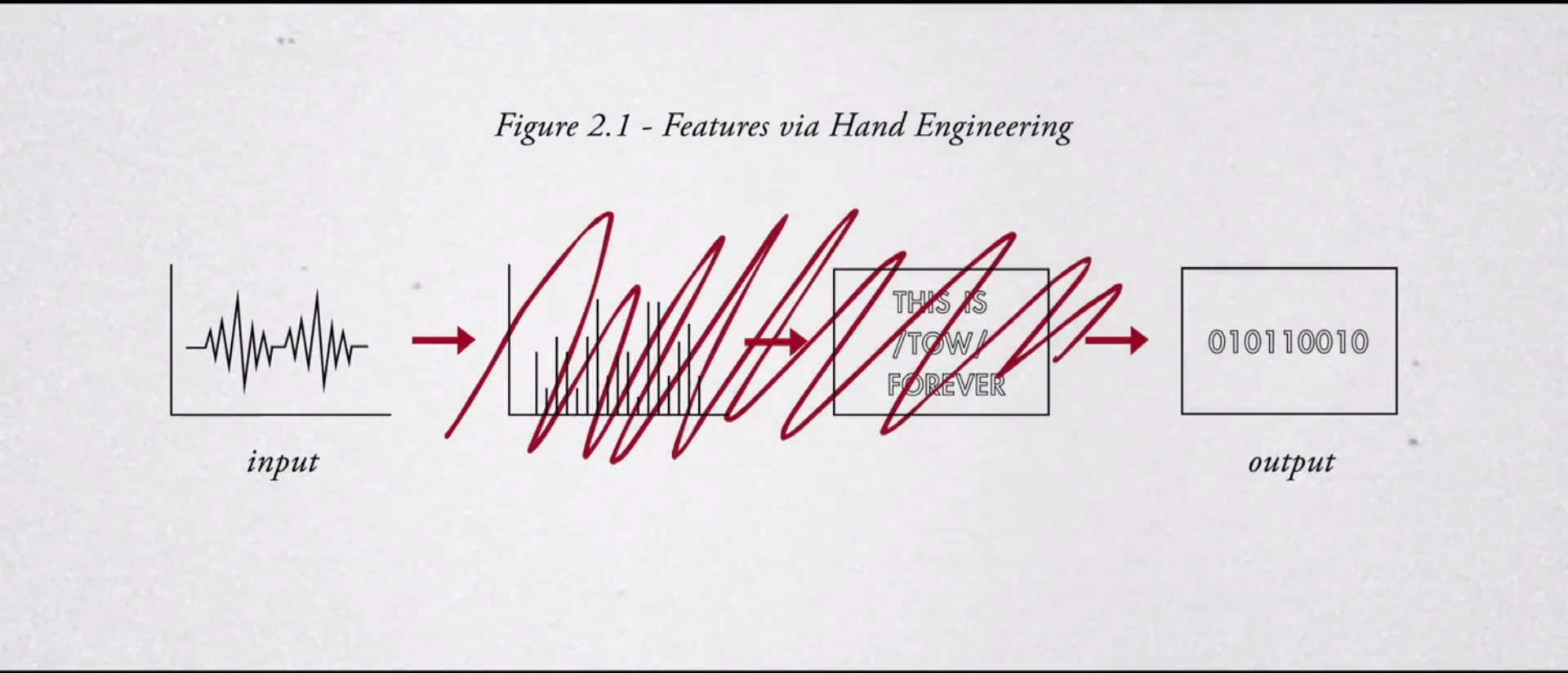
Figure 2.1 - Features via Hand Engineering



5:10 / 7:18

https://www.youtube.com/watch?v=yxxRAHVtafI

Behind the Mic: The Science of Talking with Computers



5:13 / 7:18

https://www.youtube.com/watch?v=yxxRAHVtafI


Behind the Mic: The Science of Talking with Computers
PARIS - FRANCE + ITALY = ROME
5:52 / 7:18
https://www.youtube.com/watch?v=yxxRAHVtafI

The Technologies

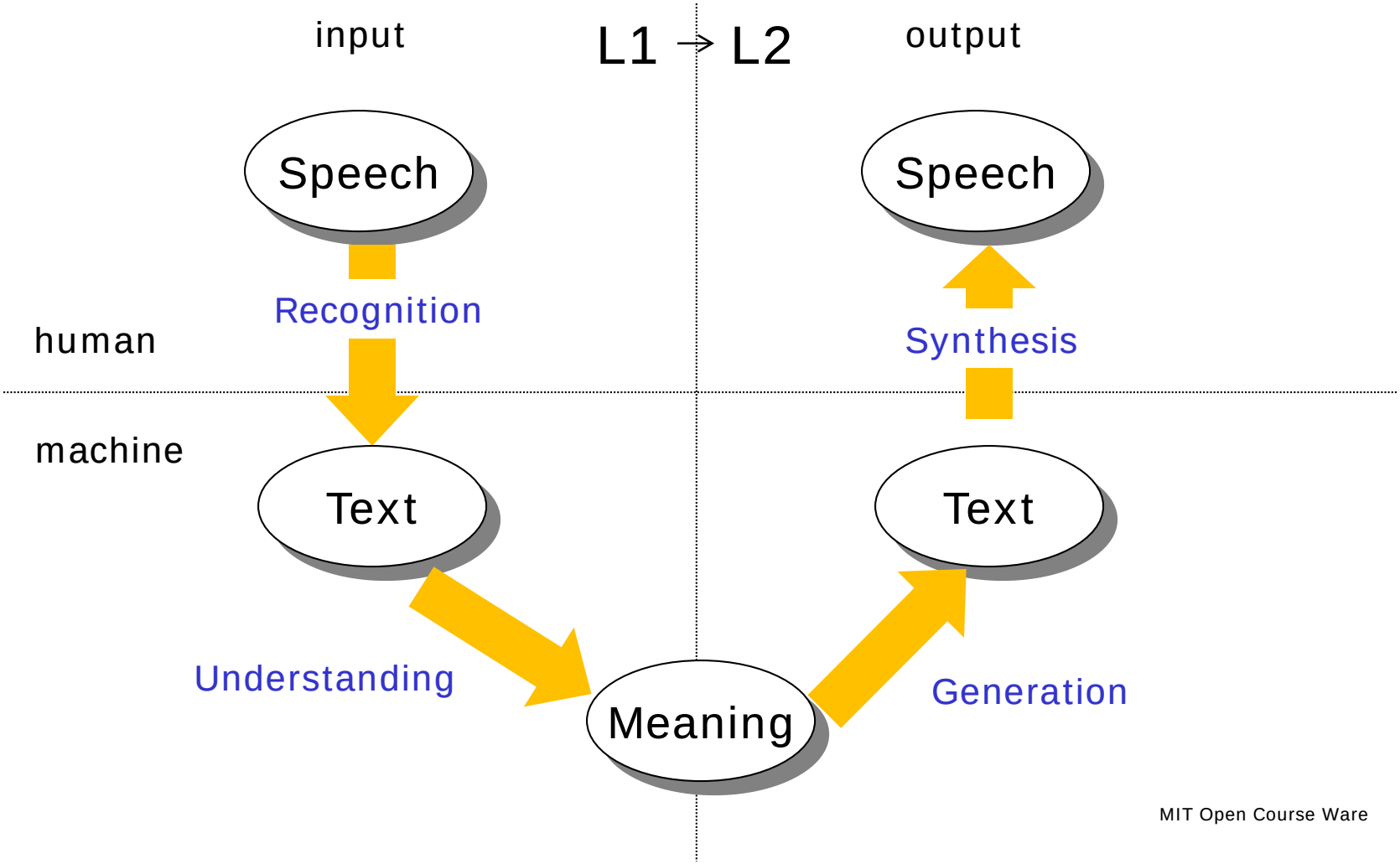
Speech and Language-related Technologies

- Human-machine interaction
  - Automatic Speech Recognition
  - Speech Synthesis / Text-to-Speech (TTS)
  - Natural Language Understanding (NLU)
  - Natural Language Generation (NLG)
- Telecommunication
  - Speech Coding
- Language
  - Statistical Machine Translation (SMT)
  - Language ID
- Language Acquisition
  - Pronunciation Training

Speech and Language-related Technologies

- Security/Forensics
  - Speaker ID
  - Speaker Verification
- Medical Applications
  - Diagnosis of Diseases
- Information Retrieval
  - Video/Audio Transcribing
  - Audio/Text Summarizing
- Speech Manipulation
  - Speaking Rate Adjusting
  - Voice Disguiser

Human-machine comm. via spoken language



MIT Open Course Ware

Speech-based Interface Applications

ASR Only

Transcribing

- Simple Command and Control
- Simple Data Entry (Over the phone)
- Dictation

ASR + NLU

Understanding

- Interactive Conversation
  - Information kiosks
  - Transactional processing
  - Intelligent agents

Benefits of Speech Interface

Limitation of Speech Interface

Some consideration in Speech UI Design

- Speech interfaces have two properties not normally found in more mature interface technologies:
  - They are **errorful**
  - Their **states are often opaque** to users.

Source: Speech Interface Guideline
(http://www.speech.cs.cmu.edu/air/papers/SpInGuidelines/SpInGuidelines.html)

Opaqueness of States

- Shopping Cart
- Check Out
- Make Payment
- Shipping Options

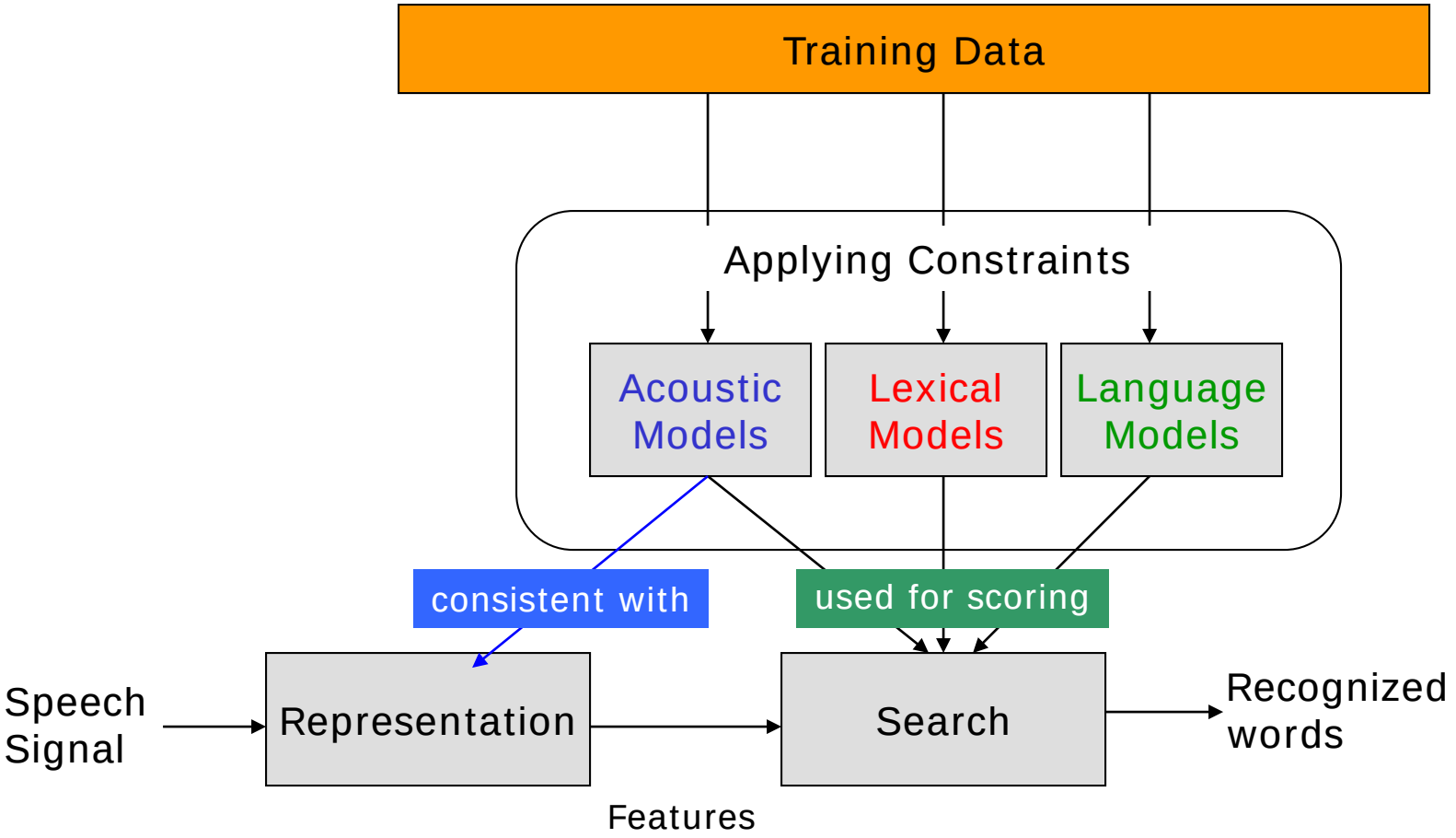
Photo by Patrick Pierre on Unsplash

Automatic Speech Recognition

Automatic Speech Recognition

- Uncover words from acoustic signal

Major Components of an ASR System

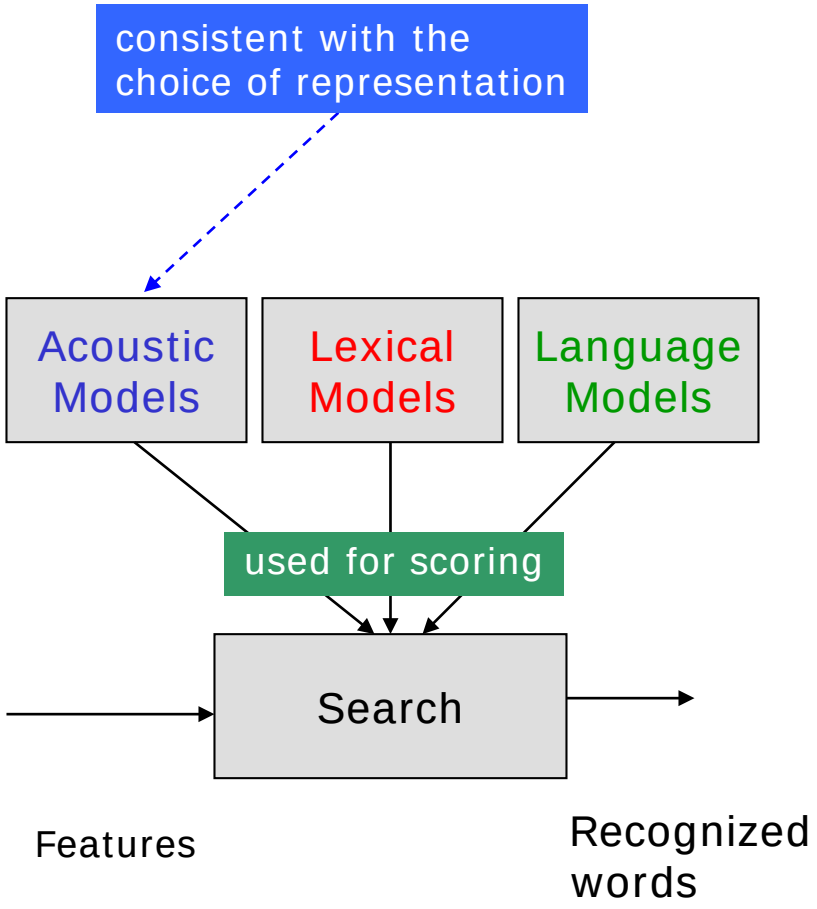


MIT Open Course Ware

Representation

- must select the right features for the representation used for the desired task
- tasks:
  - recognition of fruits in a basket containing mangos, guavas, durians, and coconuts
    - weight?
    - skin color?
    - meat color?
    - shape?
  - recognition of different sound classes?

Constraints

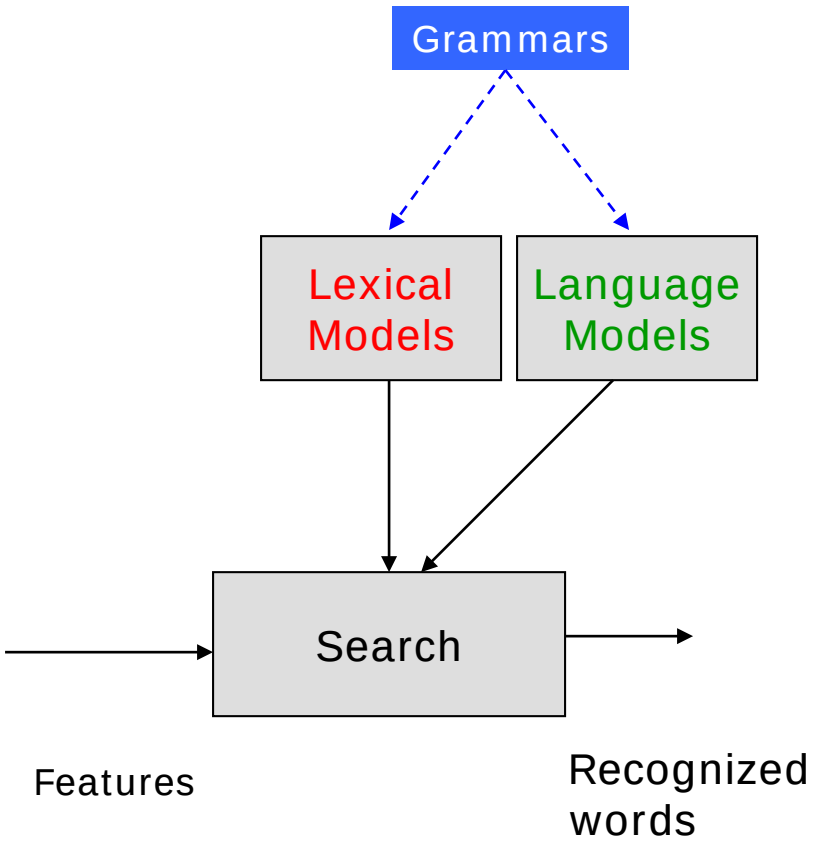


- Acoustic
  - signal generated by human vocal apparatus
- Lexical
  - Words in a language are limited.
- Language
  - A sentence must be syntactically and semantically well formed.

Domain Constraint

- Domain example
  - Movie listing
  - Phone directory
  - Weather information
- Specifying domain makes **lexical** and **language** constraint more specific.

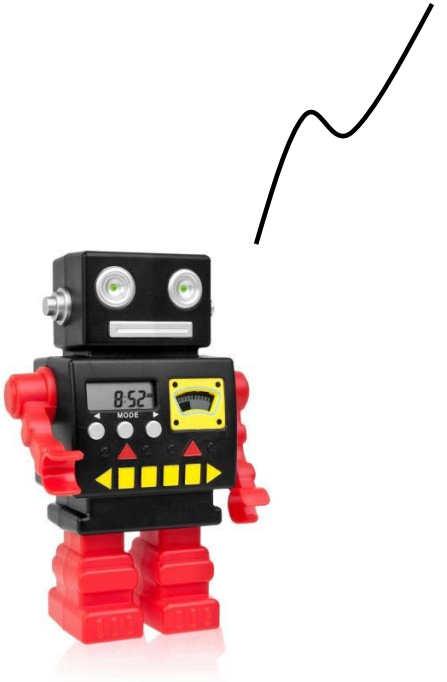
Speech Recognition Grammars



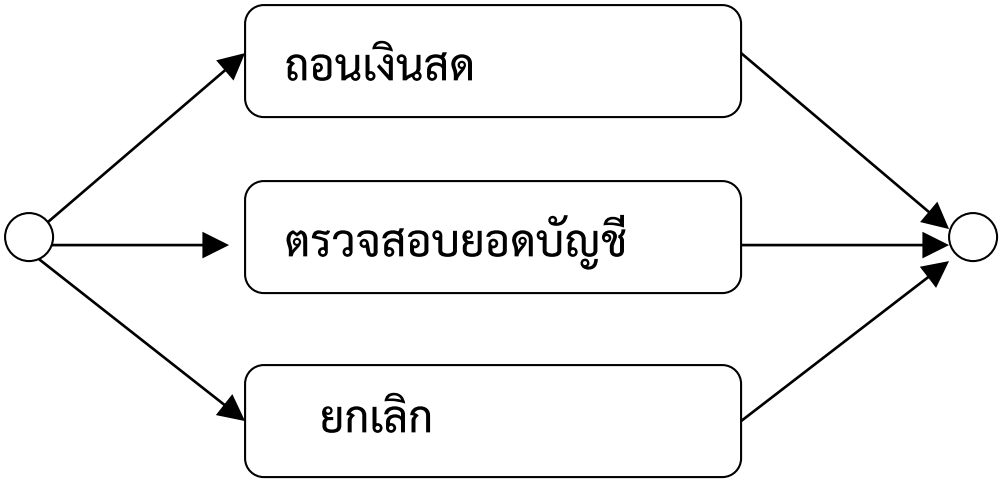
- Speech Recognition Grammars tell the system what to expect a human user to say.
- Can be graphically represented as directed graphs.
- Each possible spoken utterance is conformed with a path traversing the graph.

Speech Recognition Grammars

“กรุณาเลือกรายการที่ต้องการโดยพูดว่า ถอนเงินสด หรือ ตรวจสอบยอดบัญชี หรือพูดว่า ยกเลิก ถ้าต้องการหยุดทำรายการ ”



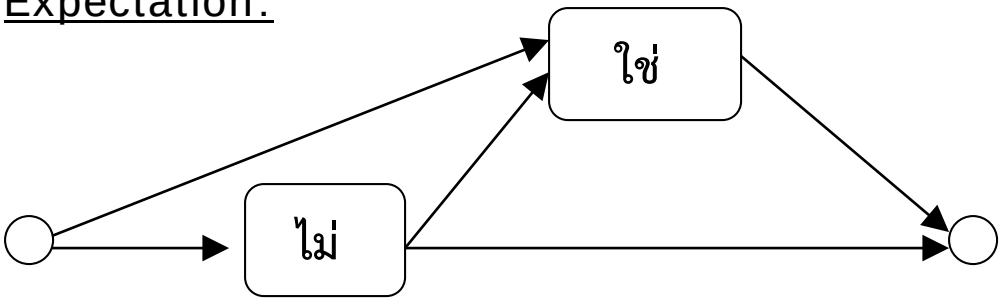
Expectation:



Speech Recognition Grammars

“คุณพูดว่า ถอนเงินสด ใช่ หรือ ไม่ใช่ ”

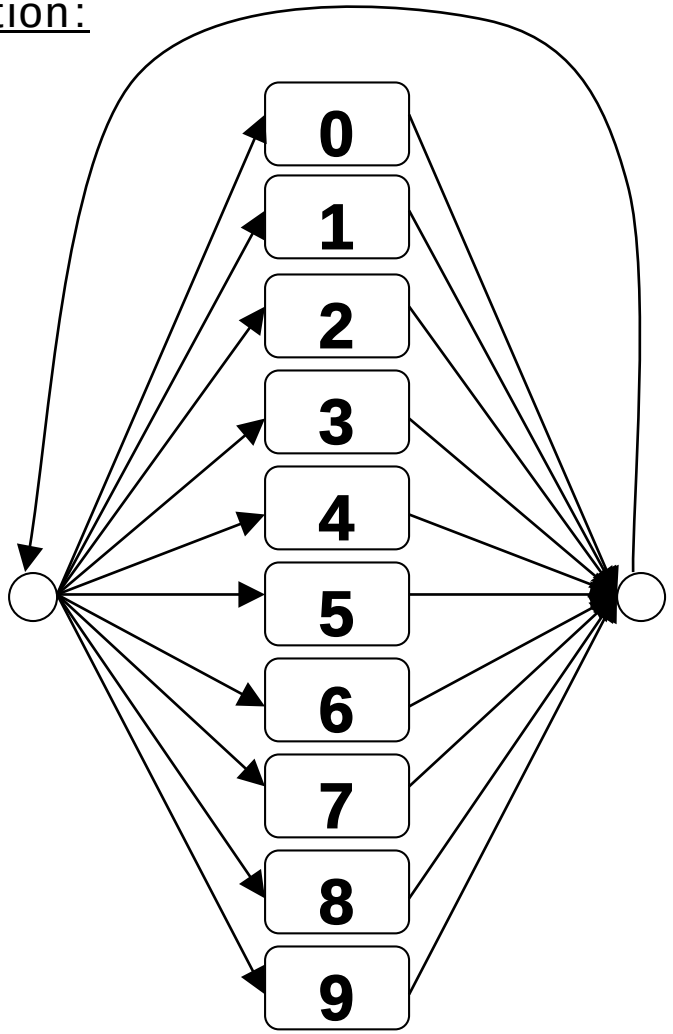
Expectation:



Speech Recognition Grammars

“กรุณาบอกรหัสเอทีเอ็ม (หุหุหุ)”

Expectation:



Speech Recognition Accuracy

Spoken Utterance 1:	ศูนย์ สอง สอง หนึ่ง แปด หก เก้า ห้า เก้า
Recognition Result:	ศูนย์ สอง สอง หนึ่ง แปด หก ห้า ห้า ห้า
Spoken Utterance 2:	ศูนย์ สอง สอง หนึ่ง แปด หก แปด แปด เจ็ด
Recognition Result:	ศูนย์ สอง สาม หนึ่ง แปด หก แปด แปด เจ็ด
Spoken Utterance 3:	ศูนย์ สอง สอง หนึ่ง แปด หก เก้า ห้า ห้า
Recognition Result:	ศูนย์ สอง สอง หนึ่ง แปด หก เก้า ห้า ห้า

$$\text{Utterance Recognition Accuracy} = \frac{\text{1 correct utterances}}{\text{Total 3 utterances}} \times 100\% = 33.33\%$$

Speech Recognition Accuracy

Spoken Utterance 1:	ศูนย์ สอง สอง หนึ่ง แปด หก เก้า ห้า เก้า
Recognition Result:	ศูนย์ สอง สอง หนึ่ง แปด หก ห้า ห้า ห้า
Spoken Utterance 2:	ศูนย์ สอง สอง หนึ่ง แปด หก แปด แปด เจ็ด
Recognition Result:	ศูนย์ สอง สาม แปด แปด หก แปด แปด เจ็ด
Spoken Utterance 3:	ศูนย์ สอง สอง หนึ่ง แปด หก เก้า ห้า ห้า
Recognition Result:	ศูนย์ สอง สอง หนึ่ง แปด หก เก้า ห้า ห้า

$$\text{Word Recognition Accuracy} = \frac{\text{23 correct words}}{\text{Total 27 words}} \times 100\% = 85.19\%$$

Word Error Rate (WER) = 14.81%

Recognition Speed

- Real-Time Factor (RTF)

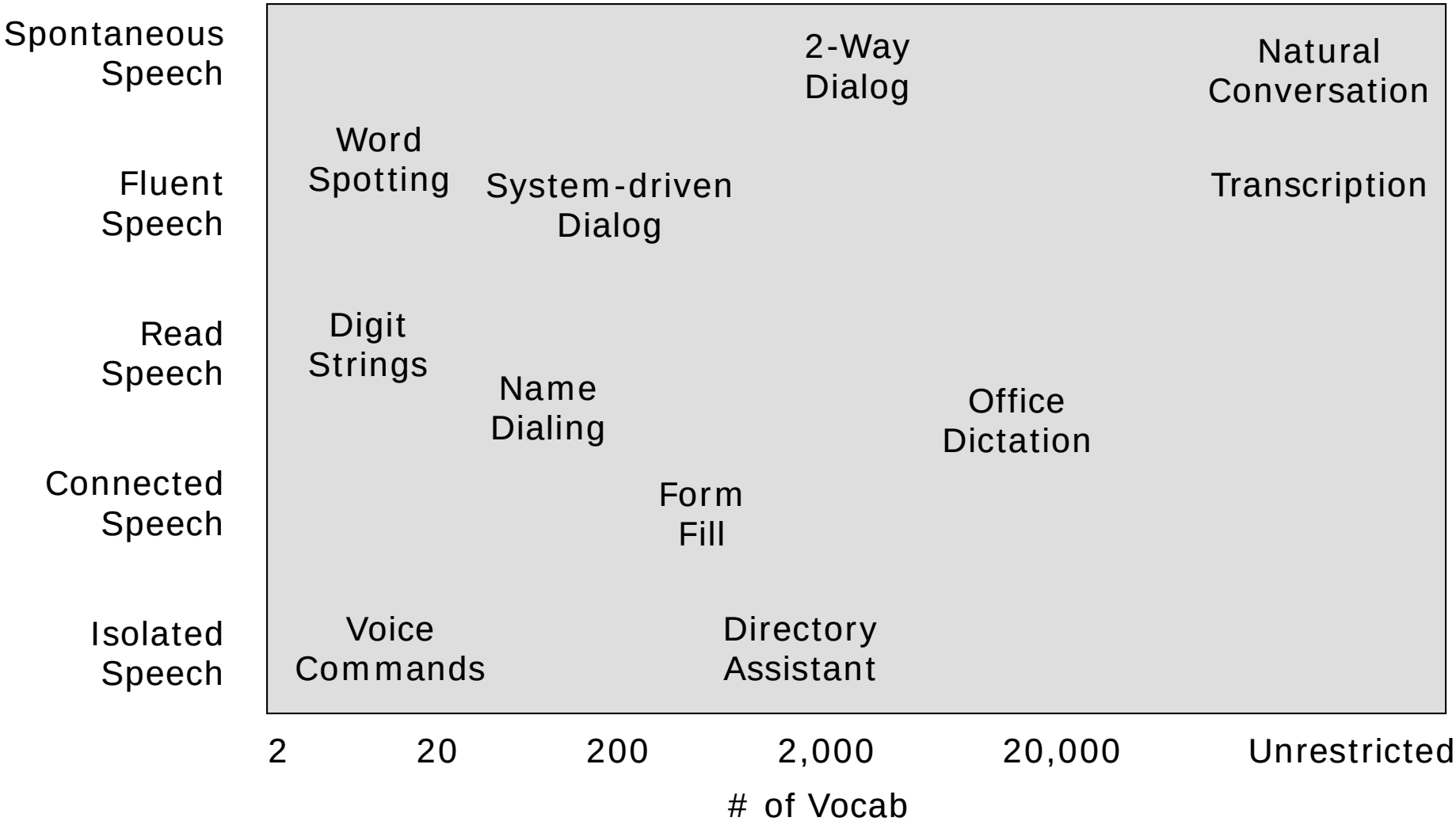
$$\frac{\text{Processing Time}}{\text{Utterance Length}}$$

Characterizing ASR System Capability

Parameters	Range
Speaking Mode	Isolated Word – Continuous Speech
Speaking Style	Read Speech – Spontaneous Speech
Enrollment	Speaker Dependent – Speaker Independent
Vocabulary	Small (<20words) – Large (>50,000 words)
Perplexity	Small (<10) – Large (>200)
SNR	High (>30dB) – Low (<10dB)
Transducer	Noise-canceling mic. – cell phone

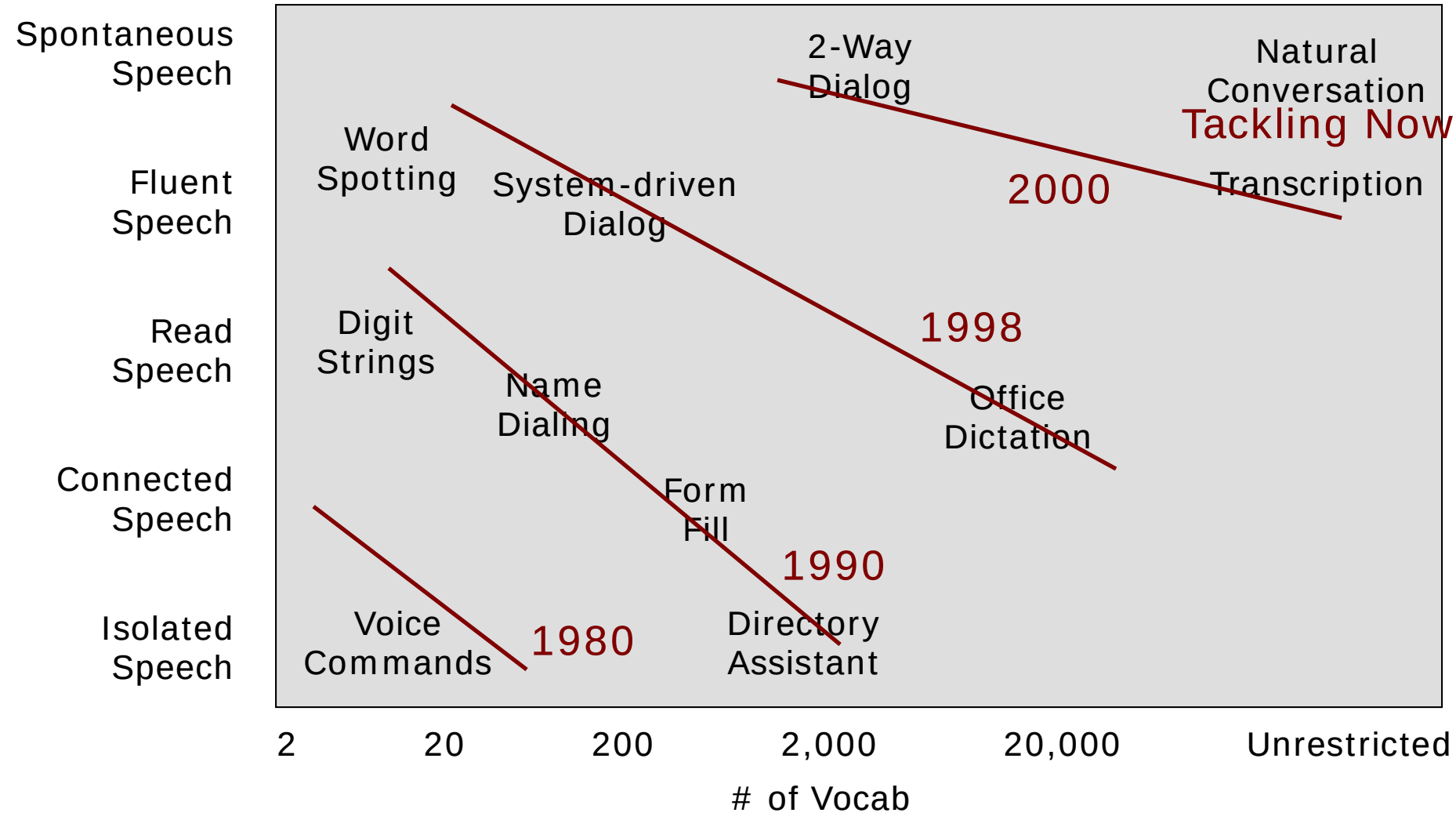
MIT Open Course Ware

ASR Difficulties



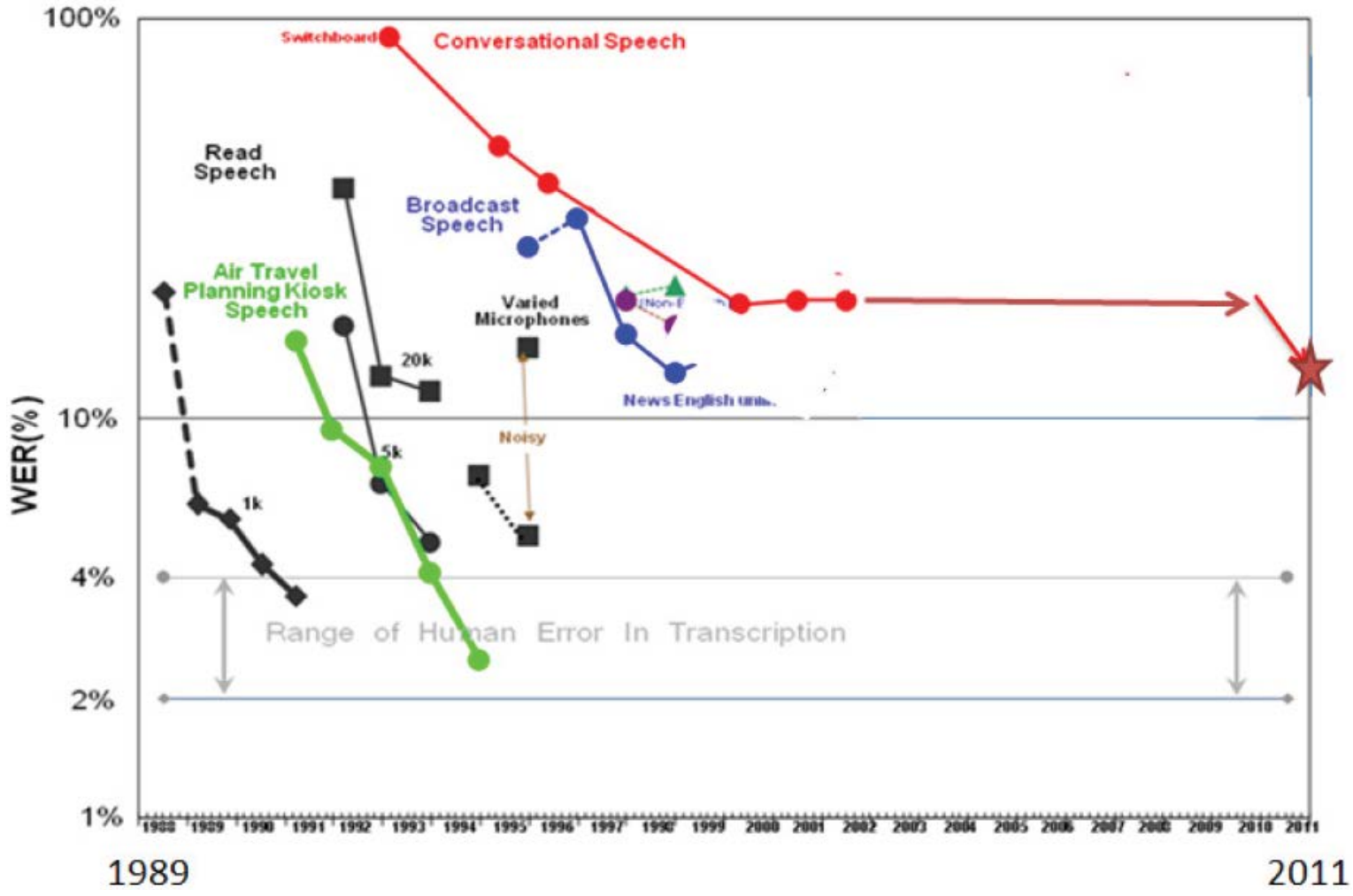
Adapted from Juang and Furui, Proc. of IEEE, 2000

ASR Status (English)



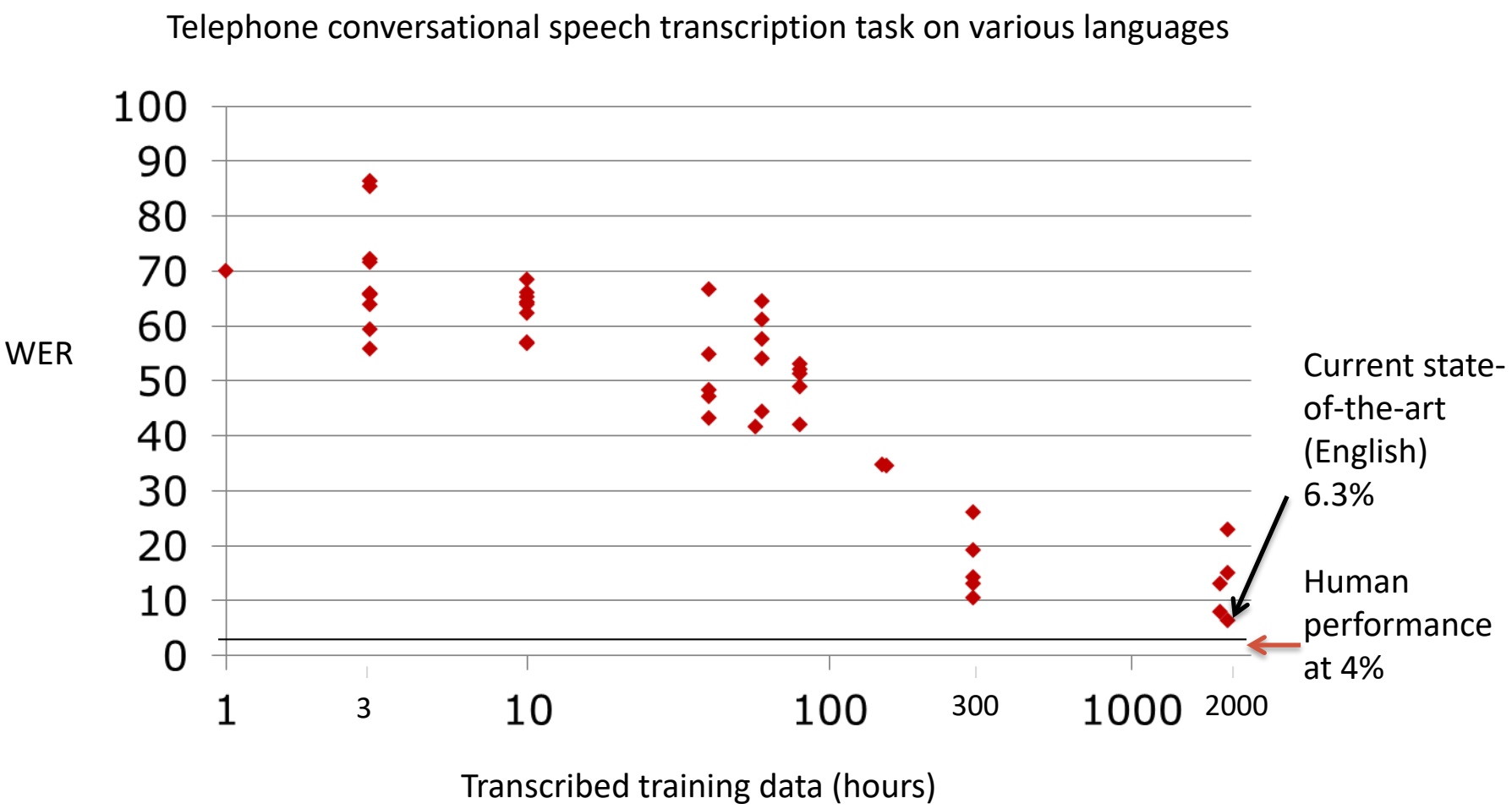
Adapted from Juang and Furui, Proc. of IEEE, 2000

ASR Status (English)



http://recognize-speech.com/images/MichaelT/Benchmarks/Gartner_Switchboard_DNN_breakthrough_2.png

ASR Performance



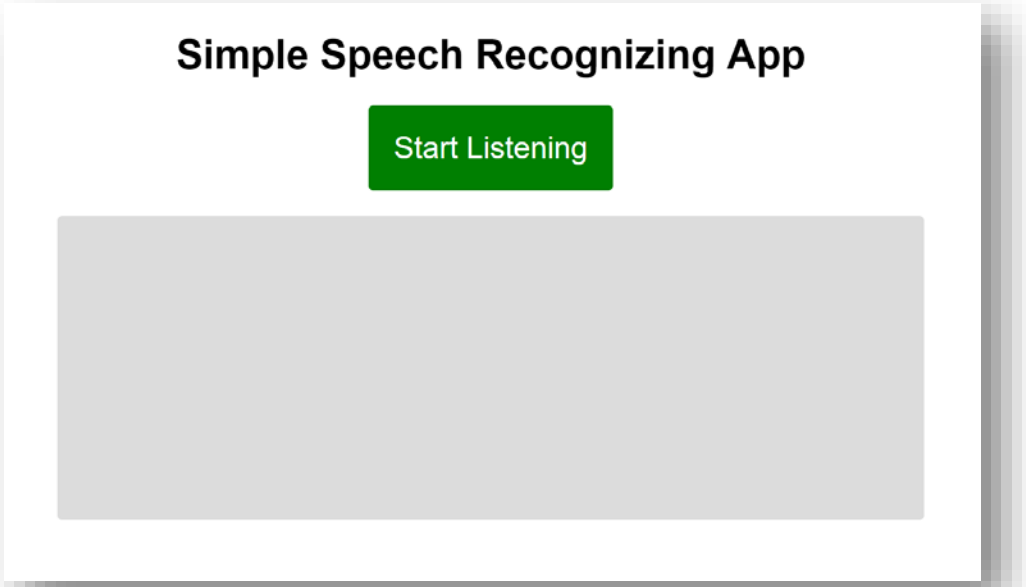
ASR of Under-Resourced Languages

A language with some of (if not all) the following aspects:
- Lack of a unique writing system or stable orthography
- Limited presence on the web
- Lack of linguistic expertise
- Lack of electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, vocabulary lists, etc.

Krauwer, S., 2003. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In: Proceedings of the 2003 International Workshop Speech and Computer SPECOM-2003, Moscow, Russia, pp. 8–15.

Sample ASR Implementations

- Speech API (Web API)
- Annyang.js



Challenges

- Co-articulation
- Speaker Independent
  - Gender / Children
  - Dialect
  - Non-native
- Spontaneous Speech
  - Disfluencies
- Out-of-Vocabulary Word
  - Fill words
  - Proper names
- Noise / Channel Robustness

Increase in Acoustic Variation

Roadmap in This Course