ROYAL INSTITUTE
OF TECHNOLOGY

Artificial Intelligence

Speech & Conversational systems

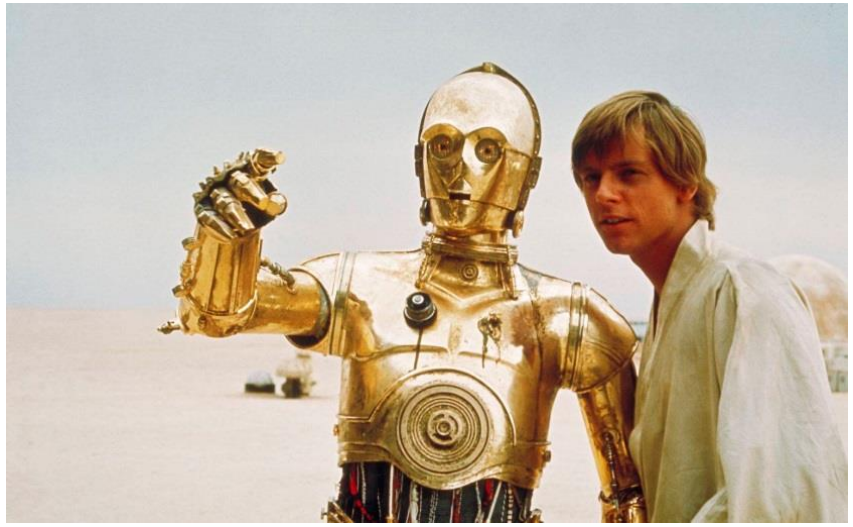
Gabriel Skantze

Professor in Speech Technology
Dept of Speech Music and Hearing
KTH, Stockholm, Sweden

The conversational machine: the vision

2001: A Space Odyssey (1968)

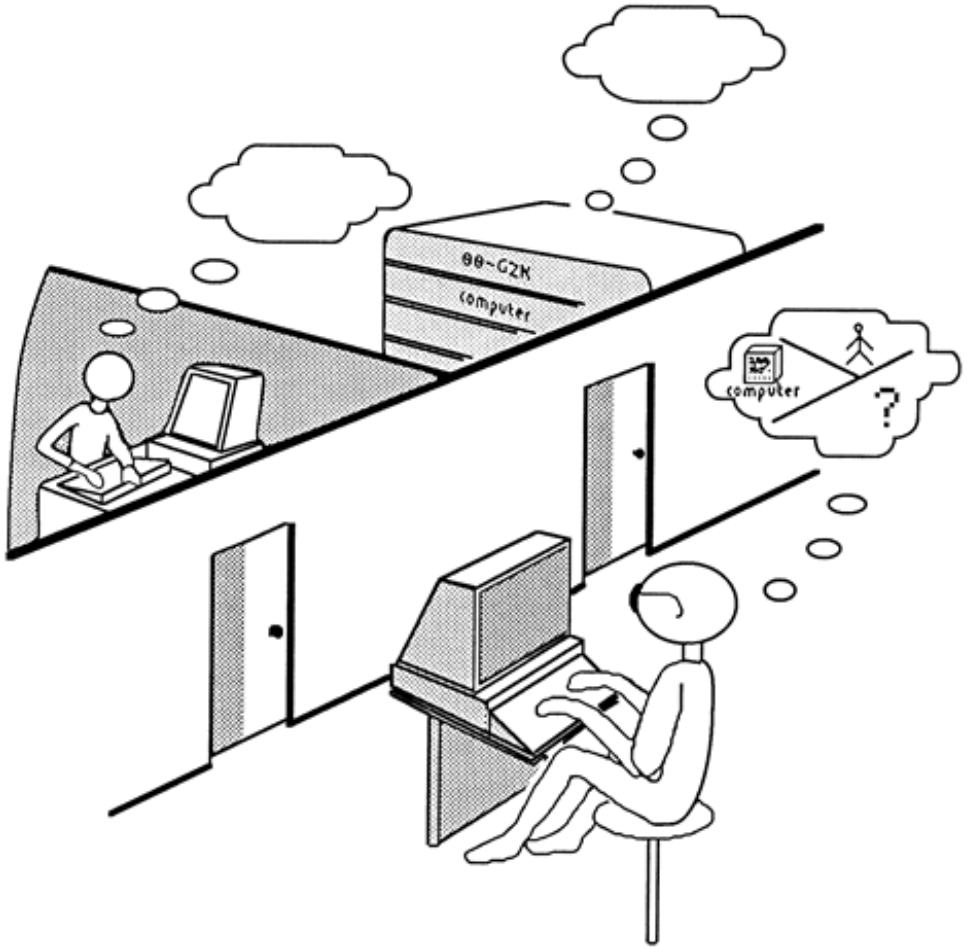
Her (2013)



Star Wars (1977)

Robot and Frank (2012)

”The Turing test” (1950)



Alan Turing

Computing Machinery and Intelligence (1950)

Eliza (1966)

Eliza: *What is bothering you?*

User: I feel sad all the time

Eliza: *Why do you feel sad all the time?*

User: I don't get along with my parents

Eliza: *Tell me about your parents*

User: I don't want to talk about them

Eliza: Why don't you want to talk about them?

Joseph Weizenbaum

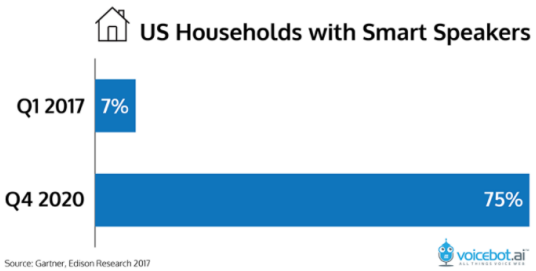
Conversational machines today

Mobile devices

Video games

Chatbots

Smart speakers



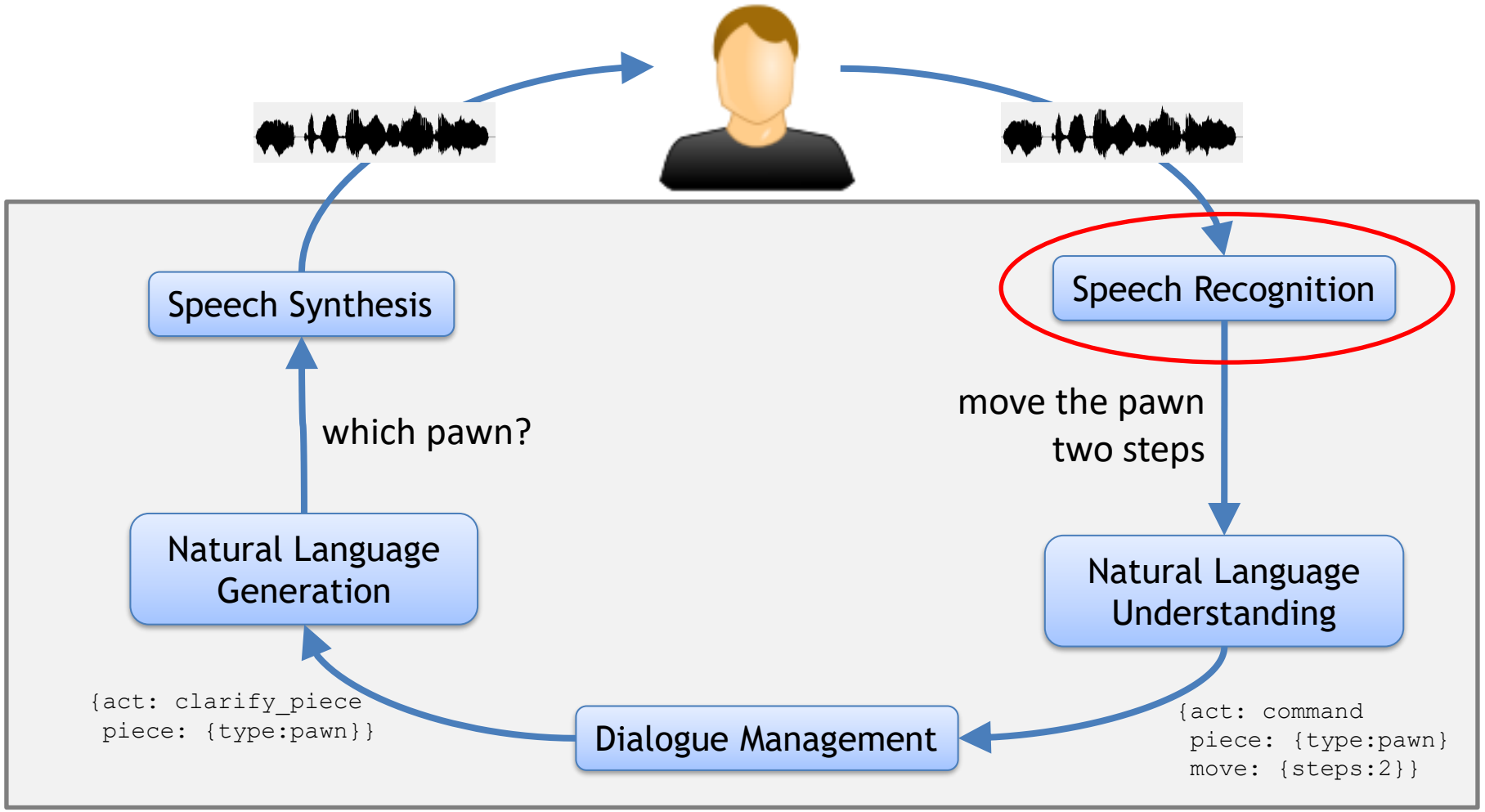
In-car systems

Human-robot interaction

Advantages of spoken language interaction

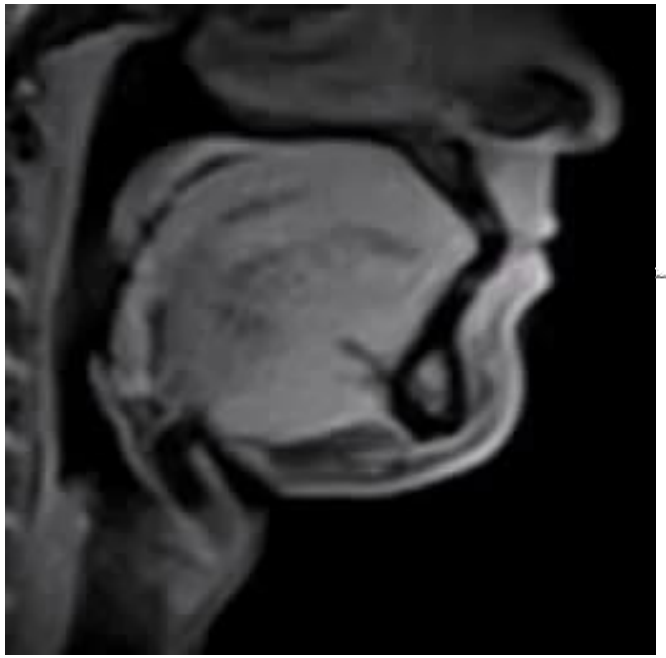
- Can be used over distance
- Can be used when eyes and hands are busy
- Comes natural to us
- A large vocabulary always at hand
- Can be used to express complex information
  - *“show me hotels in san francisco for tomorrow that are less than $300 but not less than $200, and don't include anything that doesn't have wifi.”*
- Exploitation of context allows for efficiency
  - *Will it rain in Paris on the first of July?*
  - *How about London?*
- Has an important social function

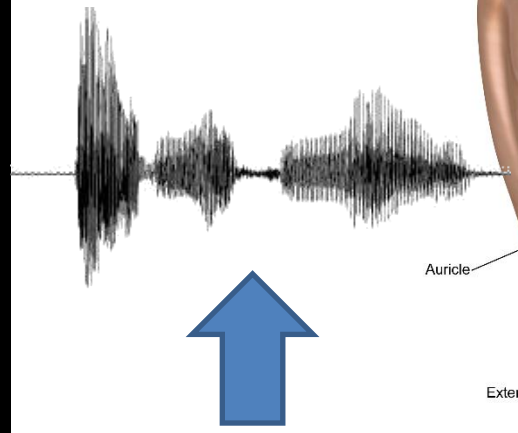
A conversational system

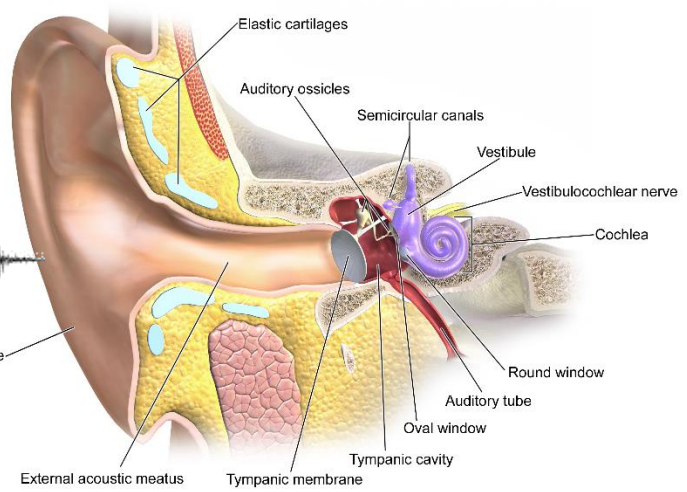


Written Language	Spoken Language
Used since 5000 years	Used since at least 100.000 years
Standardized: Words, letters, spaces, punctuation	Highly variable, ambiguous and noisy
Asynchronous communication	**Real-time** communication
Syntactically **well-formed**	**Disfluent** (Repetitions, hesitations, truncuted words, etc)
Exclusively **symbolic** (*what* we say)	**Non-symbolic** components (*how* we talk: prosody, laughter, breathing, etc)

Speech communication

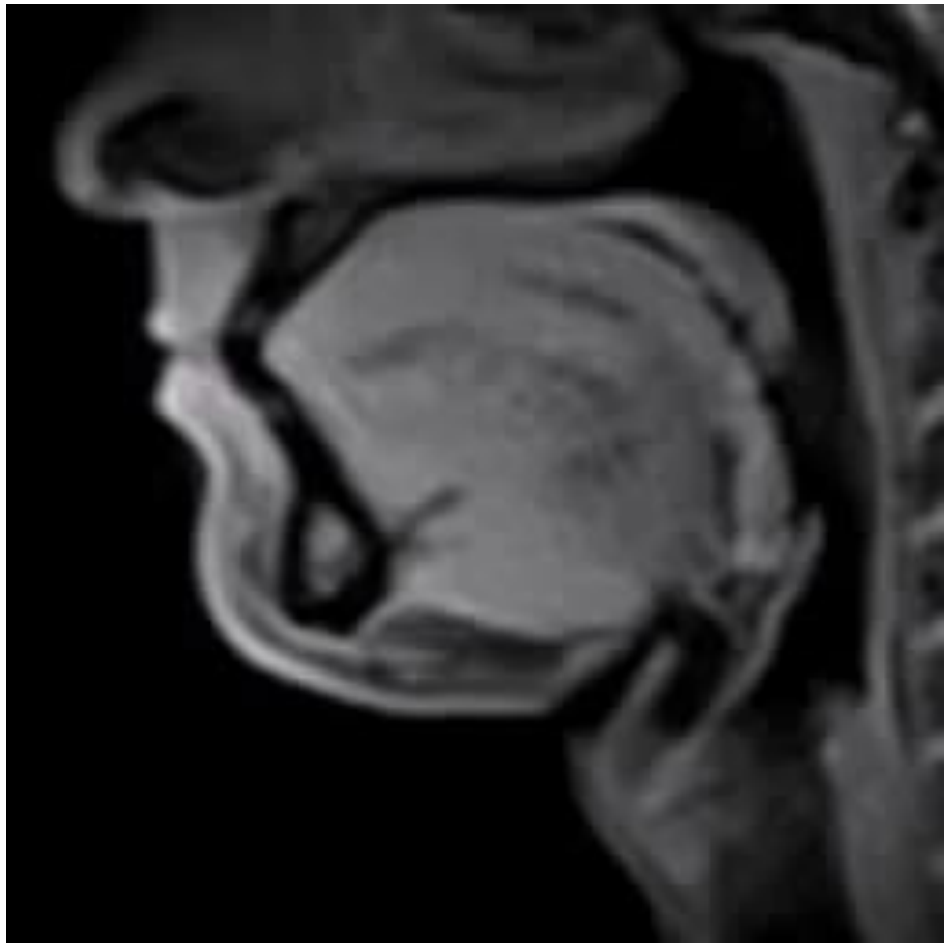






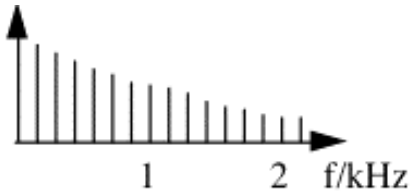
Noise

Speech Production

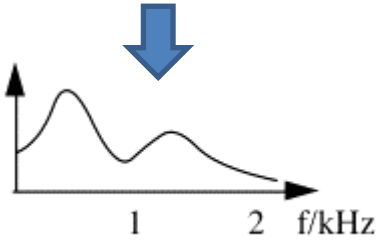


Source-filter model

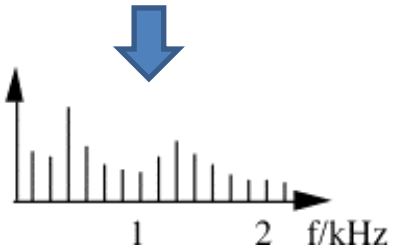
Source:
Vocal Folds



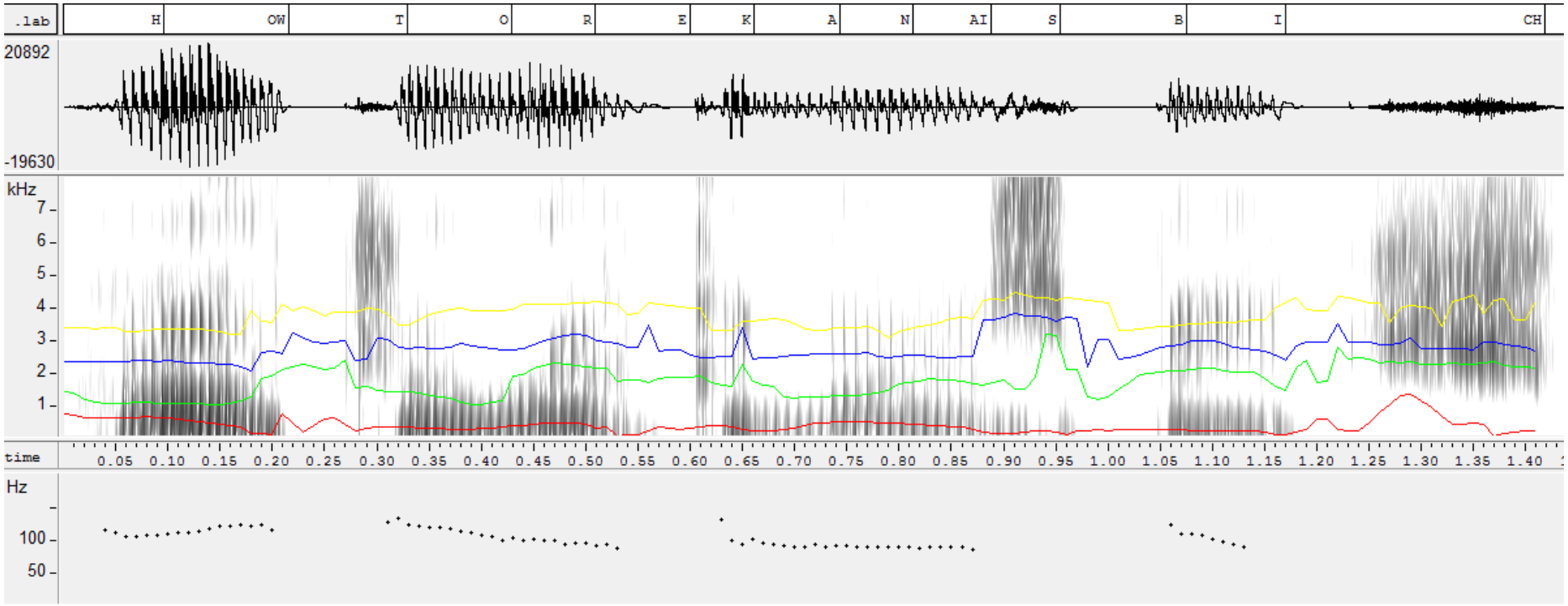
Filter:
Vocal Tract

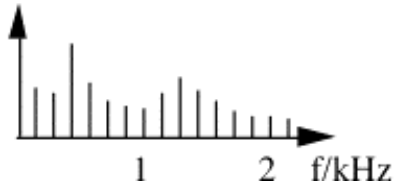


Output:
Speech signal

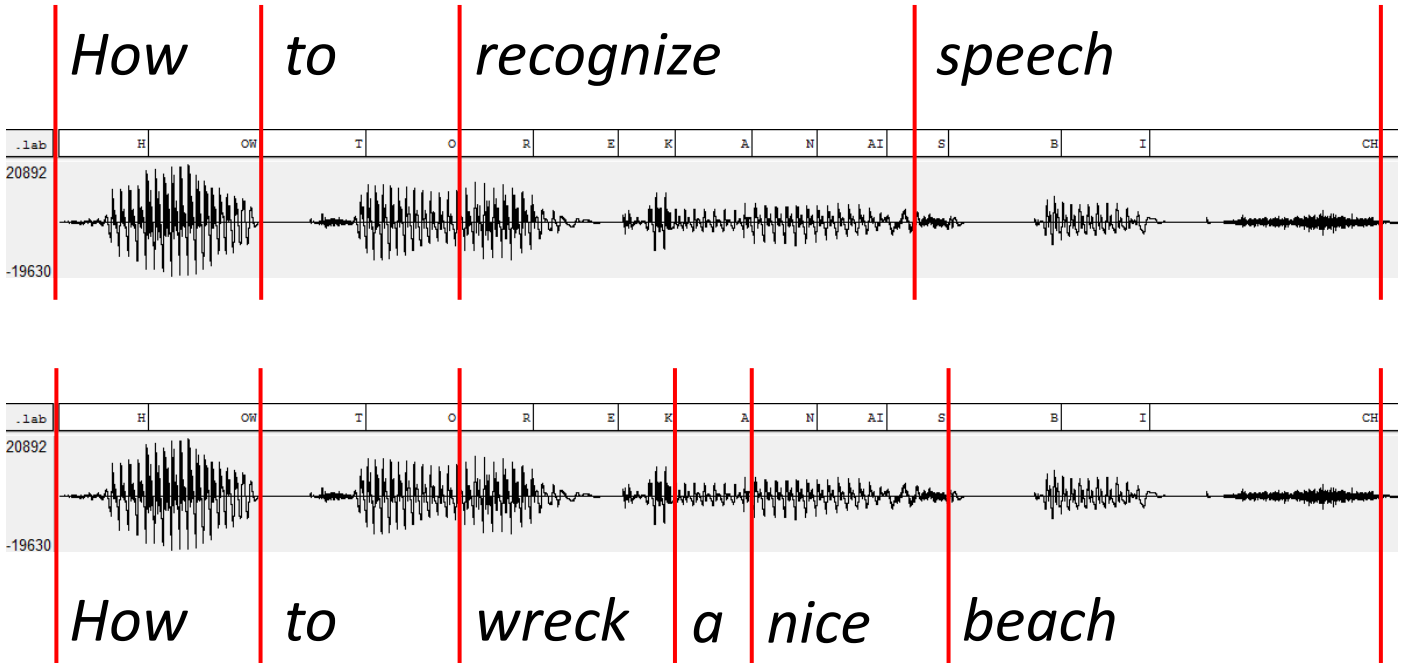


The speech signal visualized





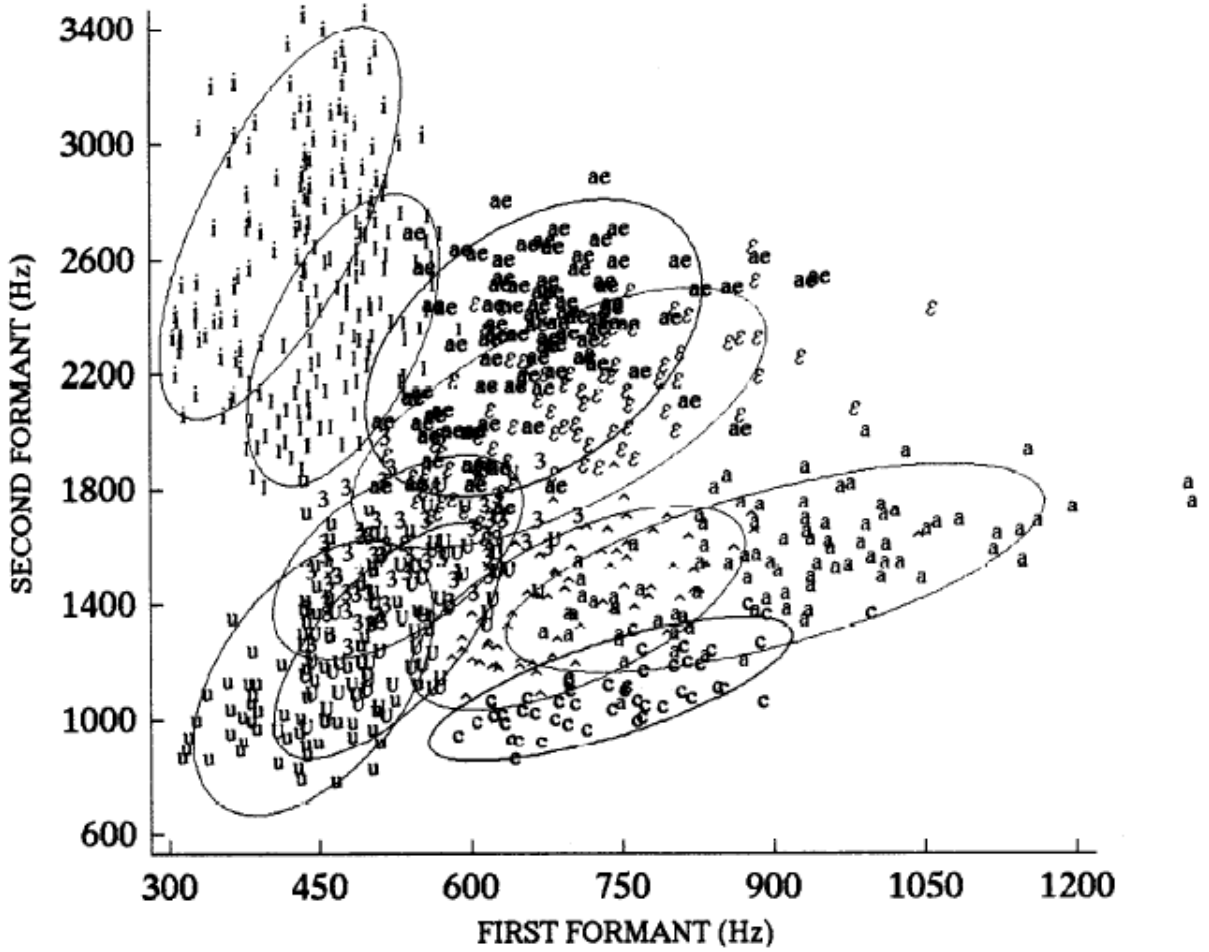
Words in the speech signal



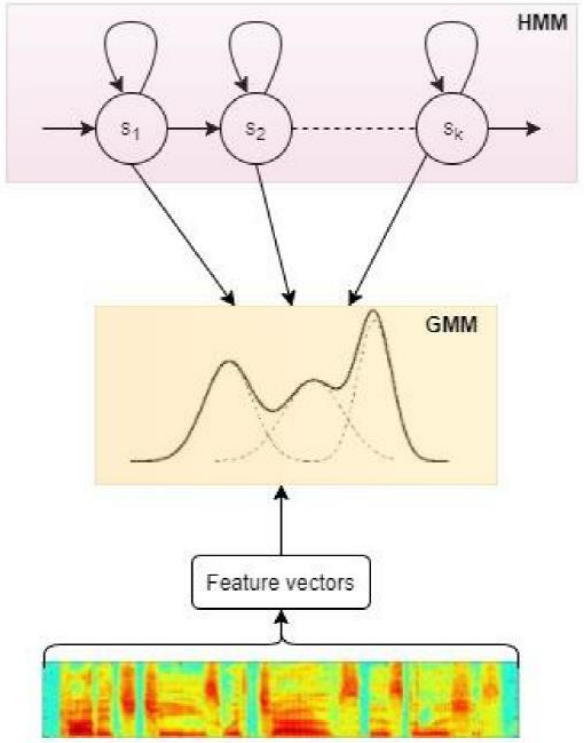
How do we know which one to choose?

The challenge of variability in the signal

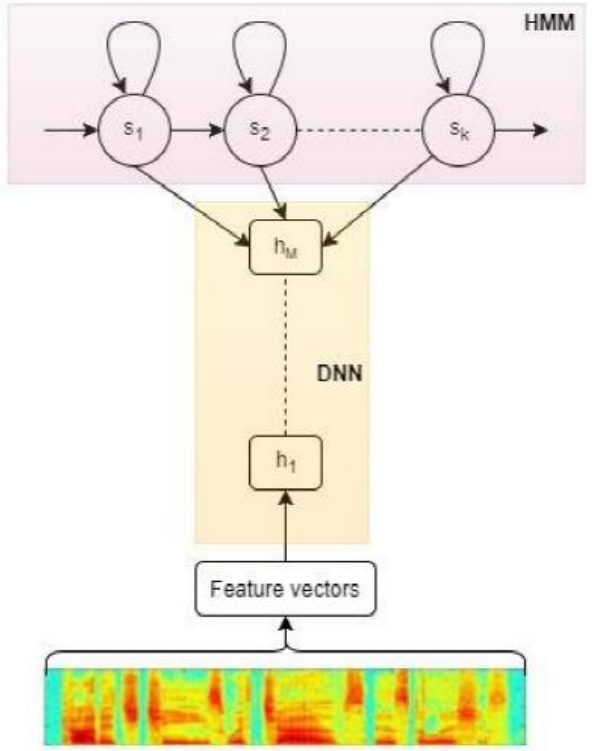
- **Language**
  - dialect, accent
- **Speaking rate**
- **Bodies**
  - sex, size, age
- **Channel**
  - noise, microphone



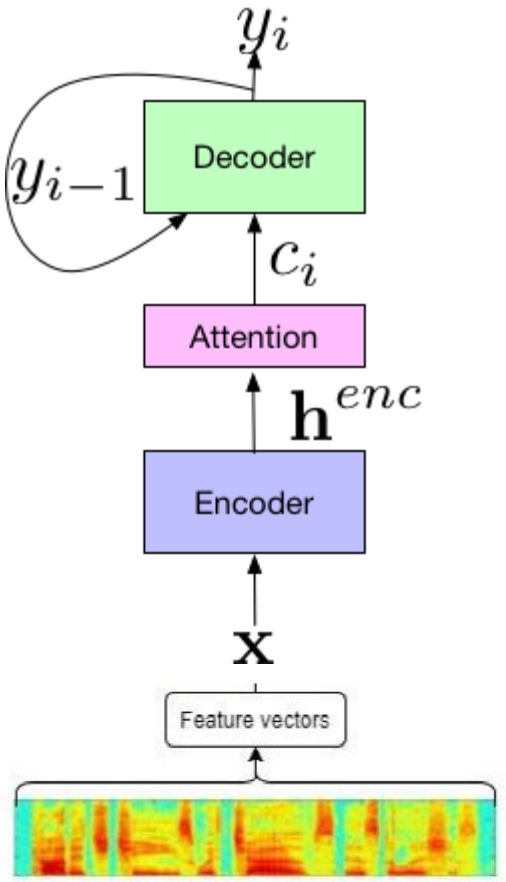
Advances in speech recognition



GMM-HMM

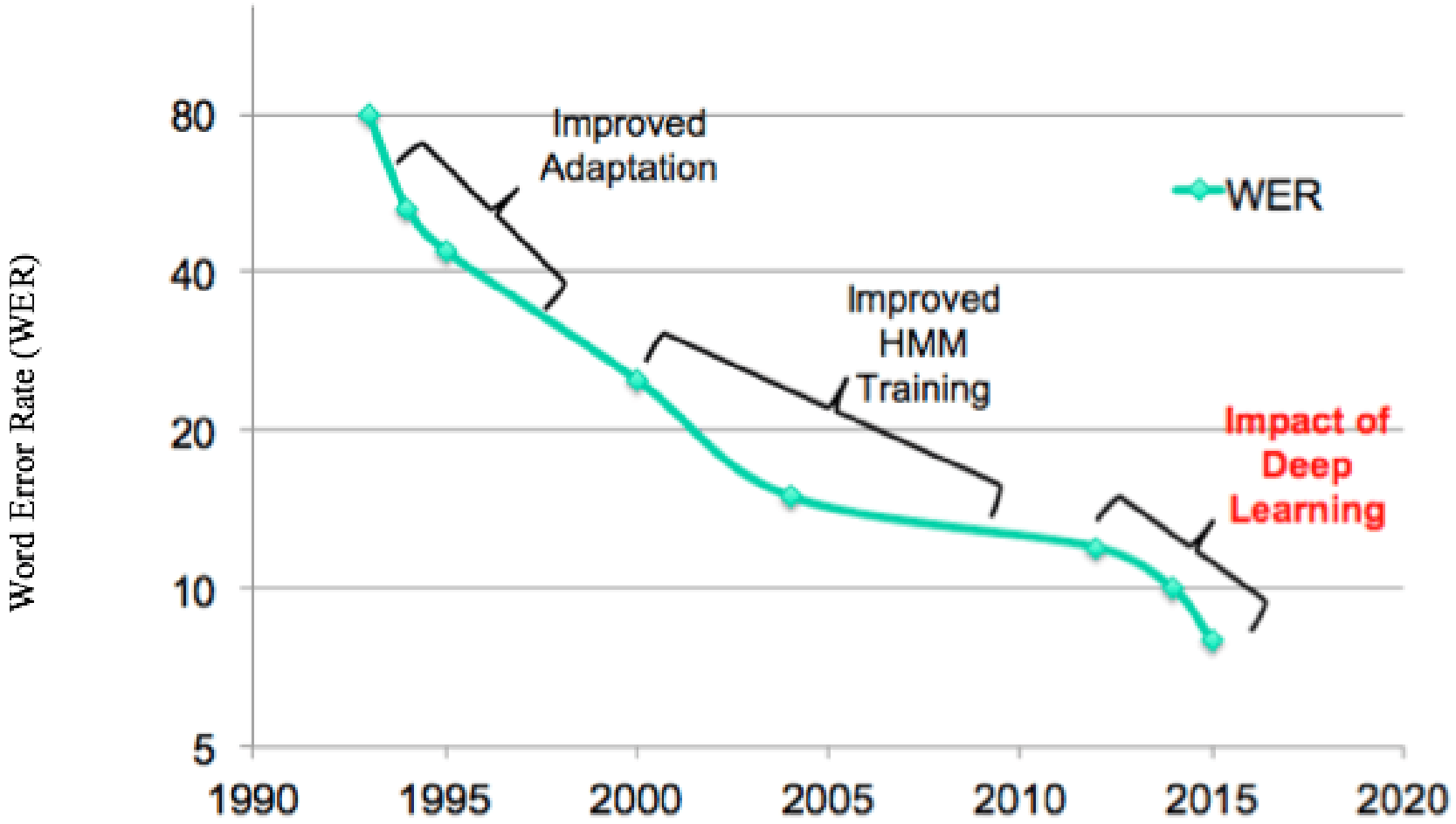


DNN-HMM

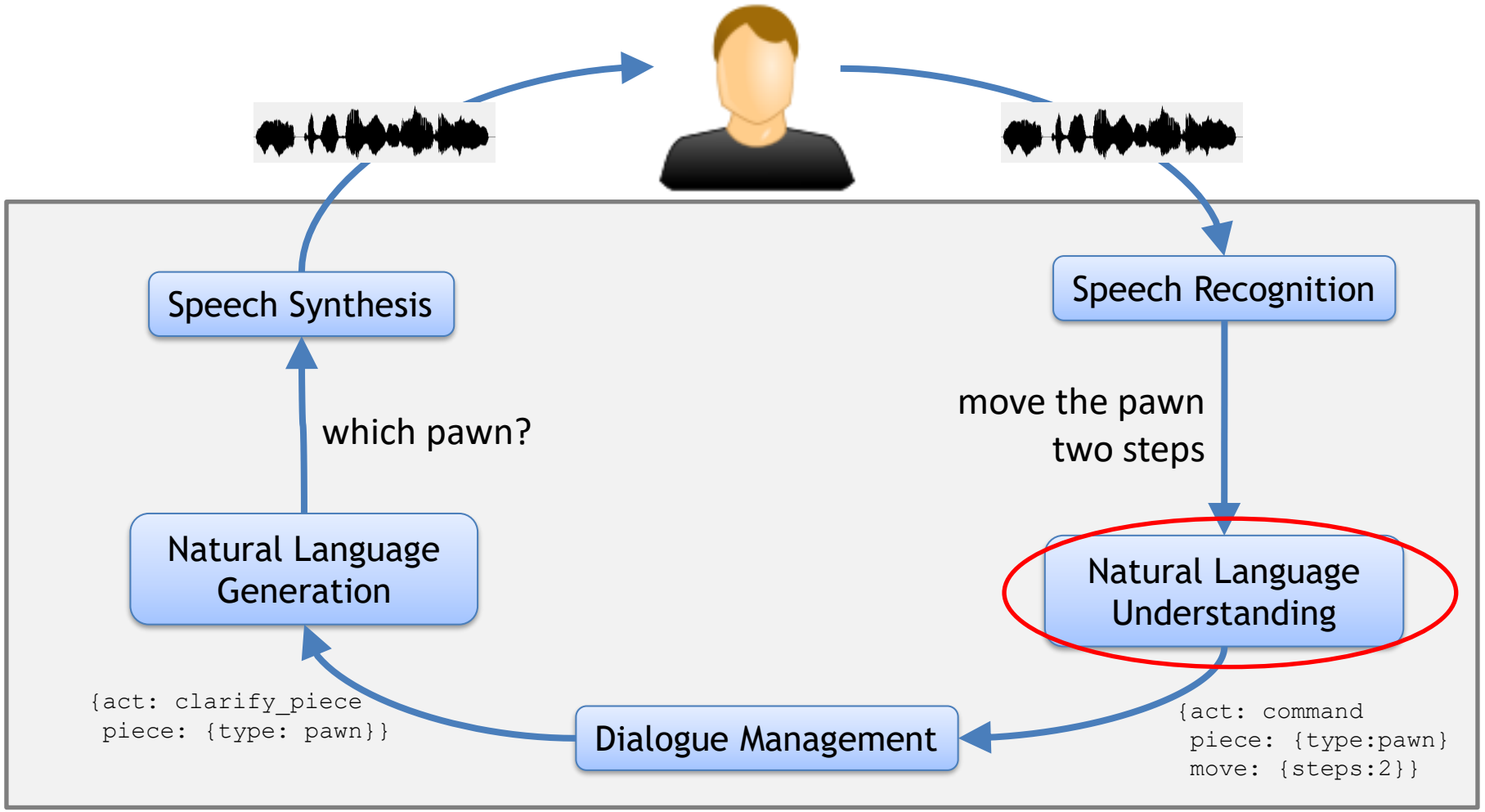


Listen, Attend, Spell (LAS)

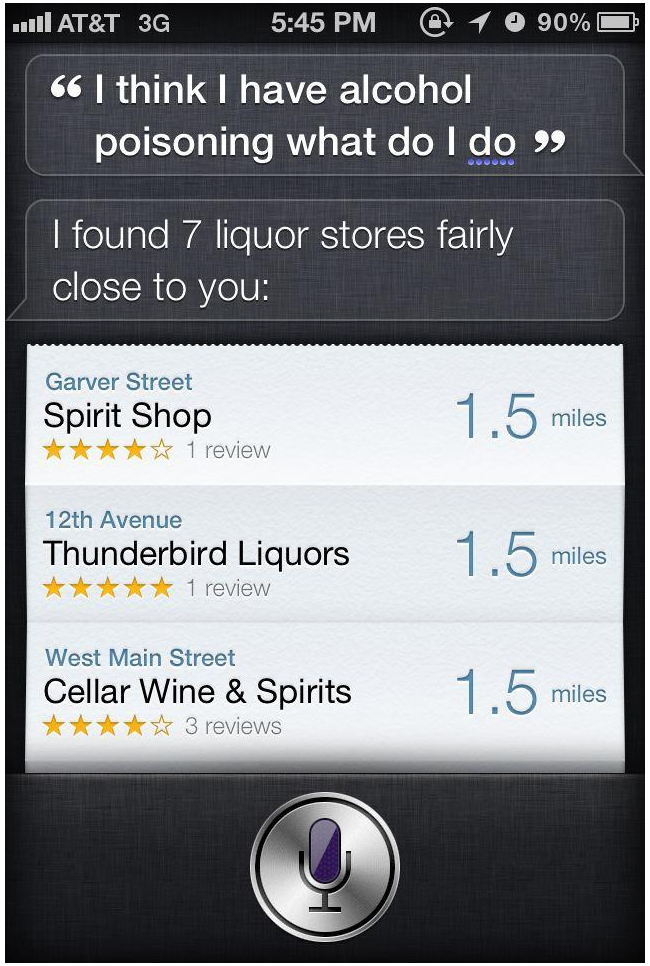
Speech recognition performance

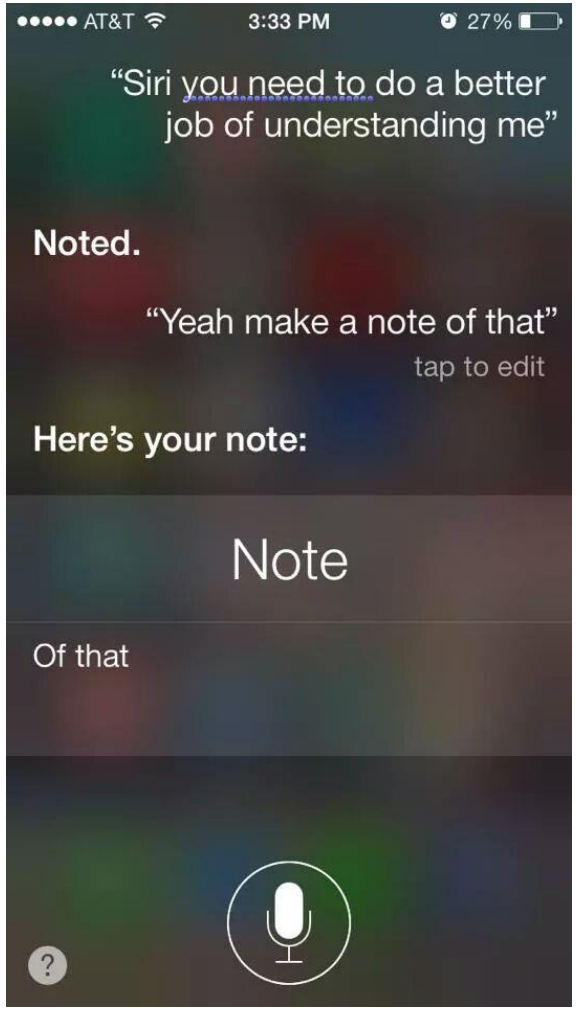


Natural language understanding



When NLU fails





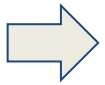
Keyword spotting

Advantages

- Simple to implement and understand
- Some robustness to variability and speech recognition errors

I would like to order a burger with cheese

Can I order a burger and please add some cheese



```
OrderFood(
  type: burger,
  topping: cheese)
```

Disadvantages

- No holistic interpretation or optimisation
- Insensitive to word order
- No structural relations

I would like a burger with cheese and onion

Can I have one burger with cheese and one with onion

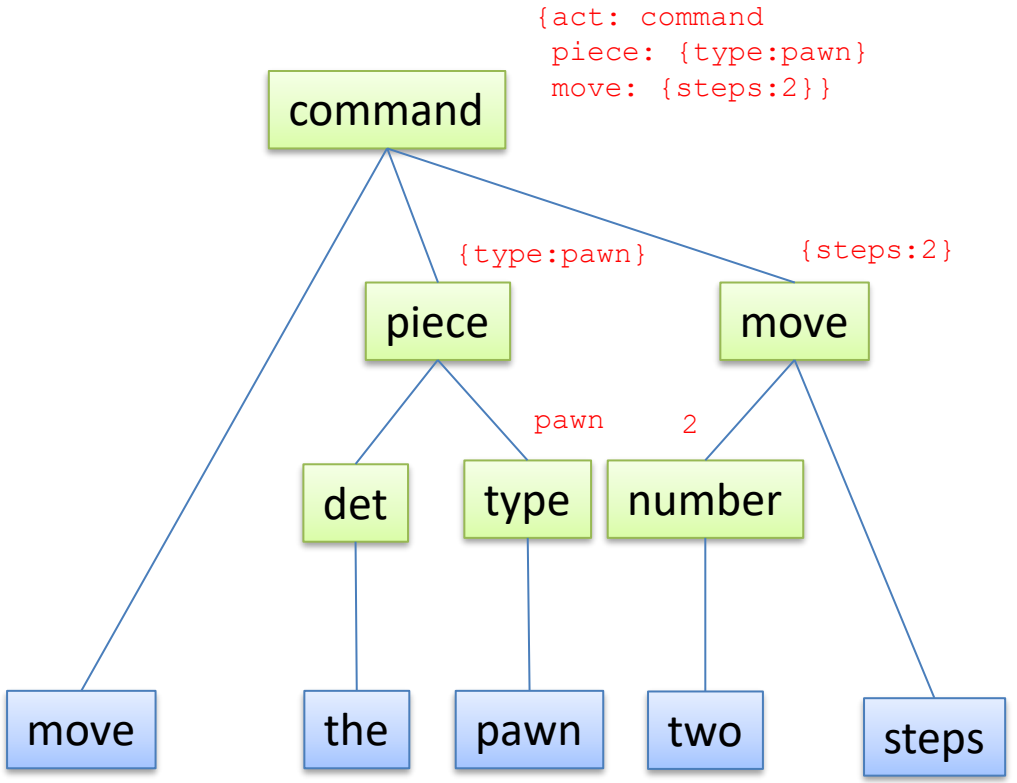
Parsing with grammar

Advantages

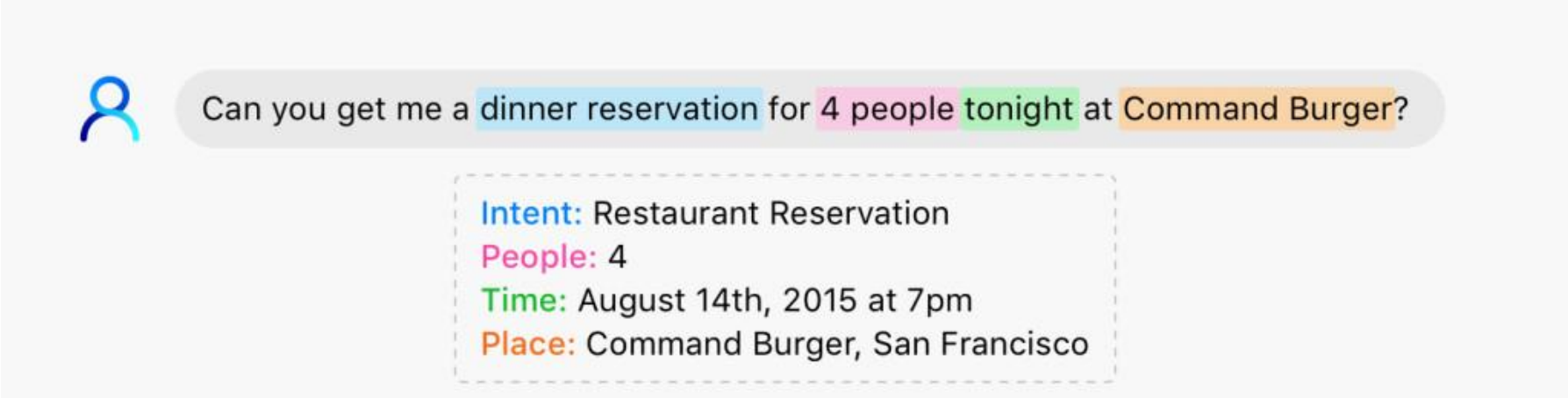
- Captures structural relations

Disadvantages

- Not robust to errors
- Requires linguistic knowledge

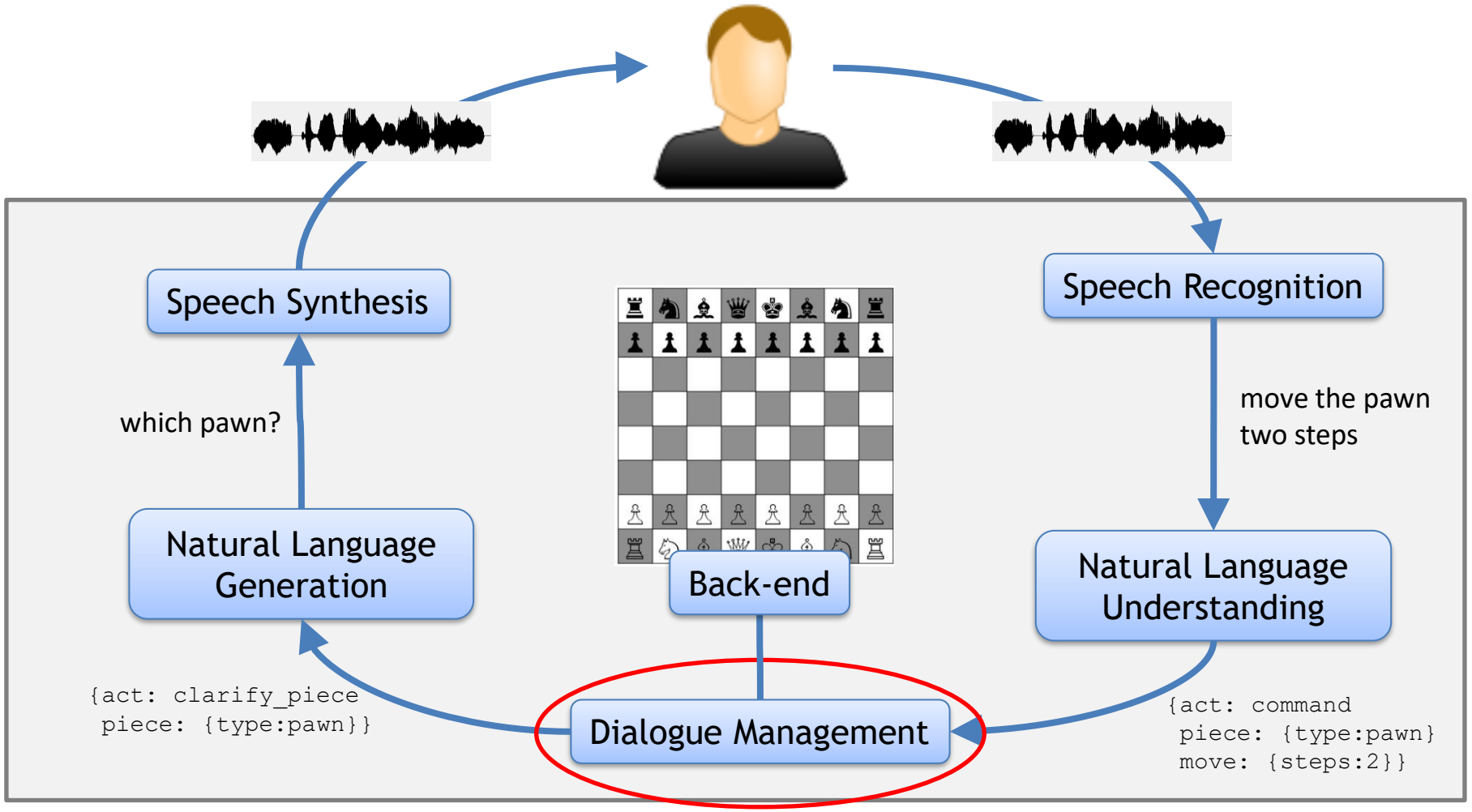


Intent/Entity recognition



- Trained on examples
- Requires little (if any) linguistic knowledge
- Captures some structural relations (but not all!)
- Can be optimised
- Has become a de-facto industry standard for NLU

Dialogue Management



Mirsuku Chatbot (Loebniz winner 2013)

What is biggest, an orange or a grape?

I would say an orange is a little bigger than a grape

How about apples?

Apples?

Do you like them? *Anaphora* *= Do you like apples?*

Do I like what?

Ellipsis Apples *= Do you like apples?*

Was that not a good example?

=> *Modelling context is a very hard challenge!*

Dialog management as form-filling

S: Where are you travelling from?

U: I want to go from **Paris**

S: And where do you want to go?

U: **London**

S: Which date do you want to leave?

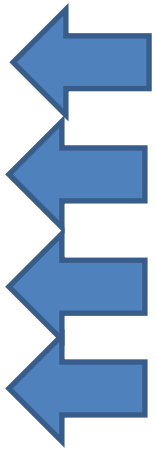
U: On the **13th of January**

S: At what time?

U: **Three** o'clock

S: There are three available flights…

Origin	Paris
Dest	London
Date	13th of January
Time	Three



Mixed initiative

S: Where are you travelling from?

U: I want to book a trip to **London**

S: And where are you travelling from?

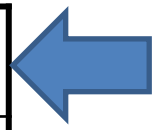
U: I want to leave **Paris** at **three o'clock**

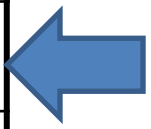
S: On which date?

U: On the **13th of January**

S: Thanks for your reservation …

Origin	Paris
Dest	London
Date	13th of January
Time	Three



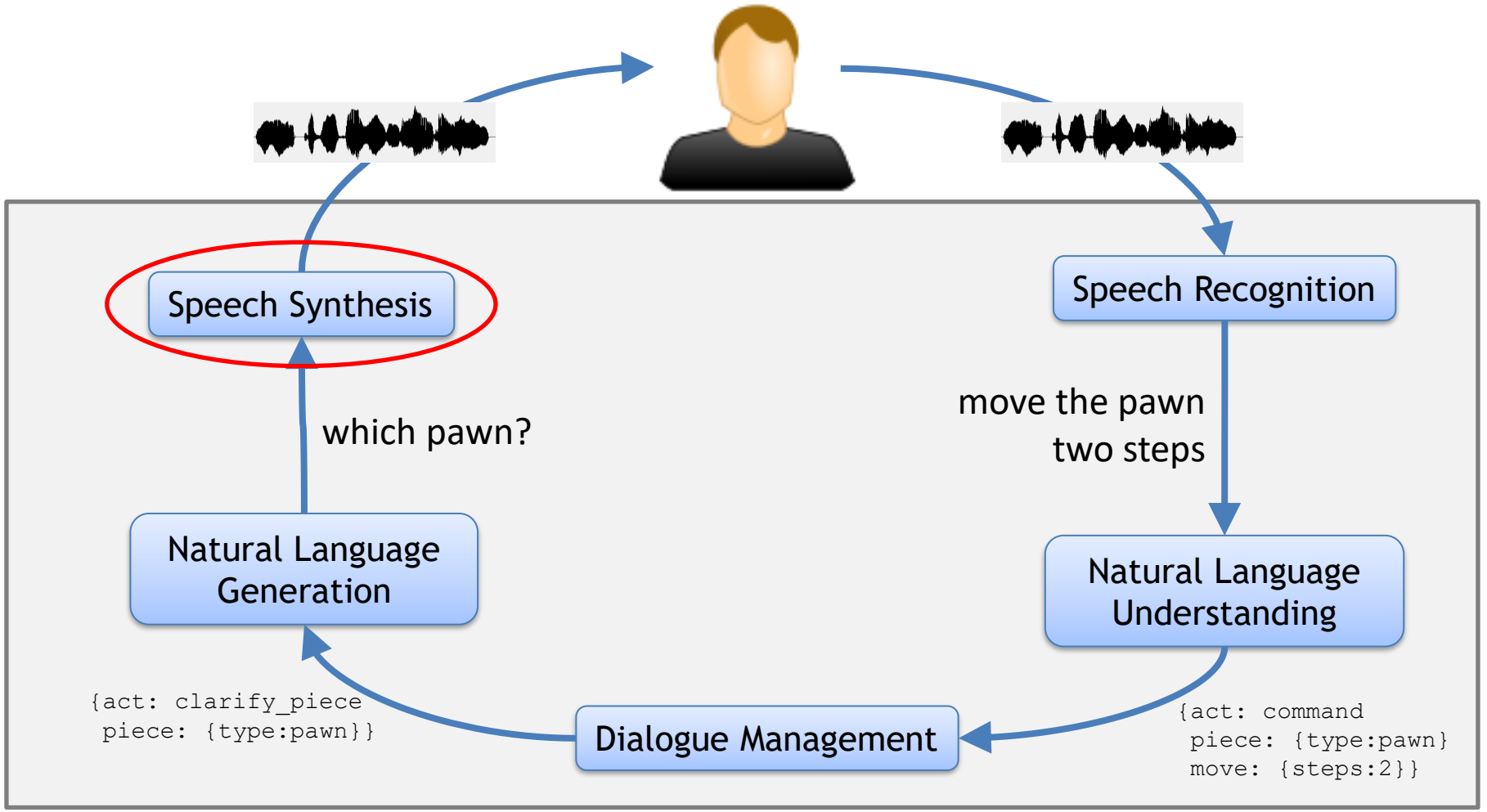


Towards end-to-end learning for conversational systems?

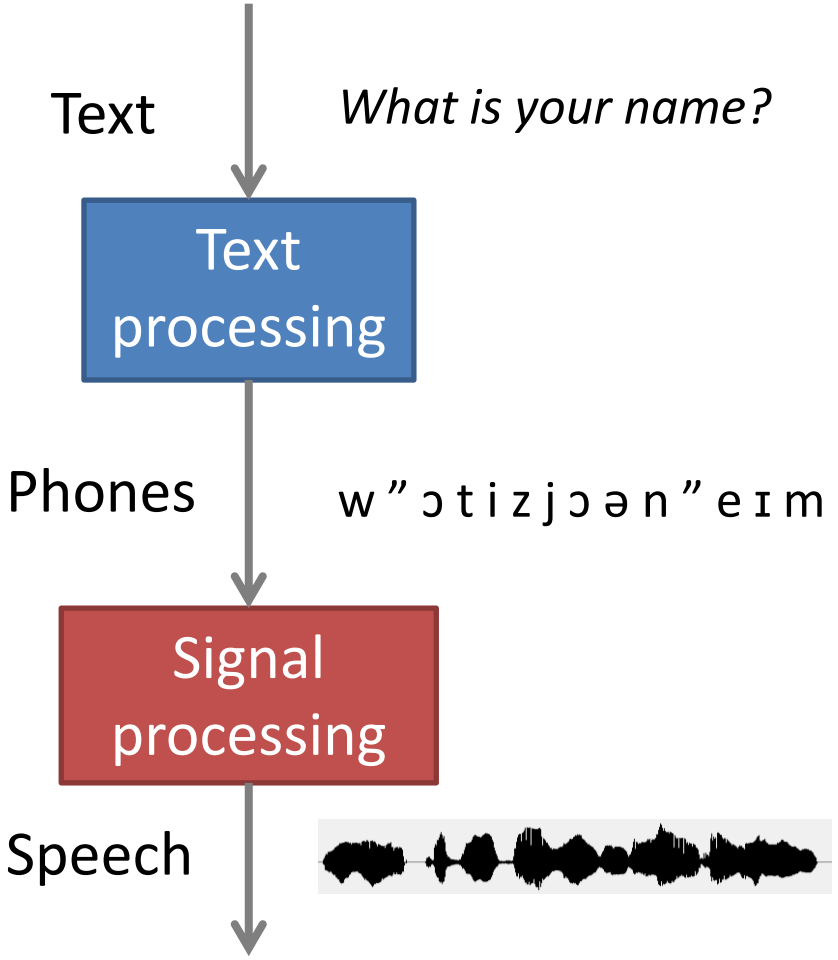
- How can we do this for domains where we don't have so much data?
  - e.g. Transfer learning
- How can we steer the conversation?
  - e.g. Persona
- How can we apply this to task-oriented dialogue?
  - e.g. Hybrid Code Networks

The Google Meena model *has 2.6 billion parameters and is trained on 341 GB of text, filtered from public domain social media conversations.*

Speech Synthesis



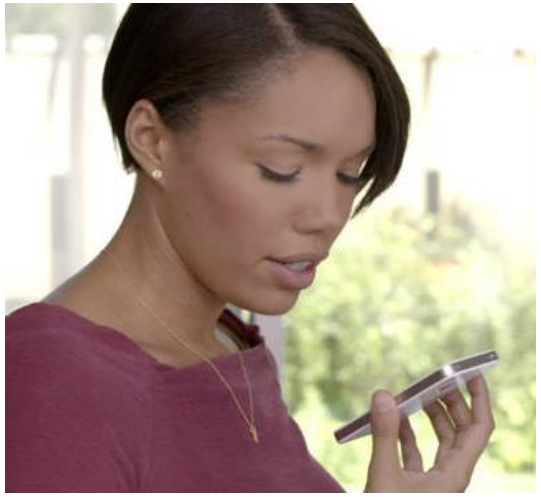
Speech Synthesis



Challenges in Text Processing

- Abbreviations (”St John St”)
- Acronyms (”IBM”)
- Numbers (”Boeing 747”)
- Homographs (”project”)
- Xenophones (”comme il faut”)

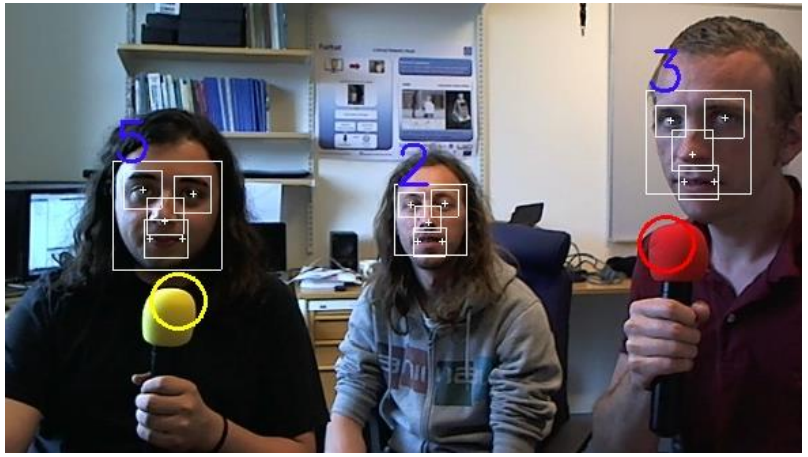
Multi-modal conversational systems



What the face adds to the conversation

Output

- Attention (gaze)
- Facial expressions
- Lip movements

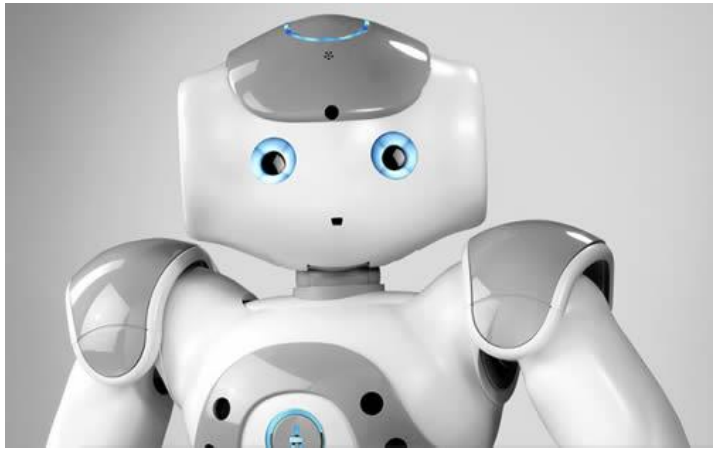


Input

- Speaker detection
- Speaker recognition
- Facial expressions
- Attention (gaze/headpose)

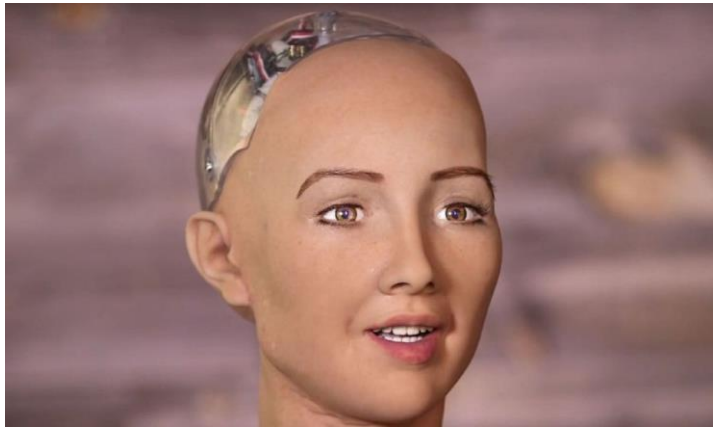
Giving the machine a face

NAO



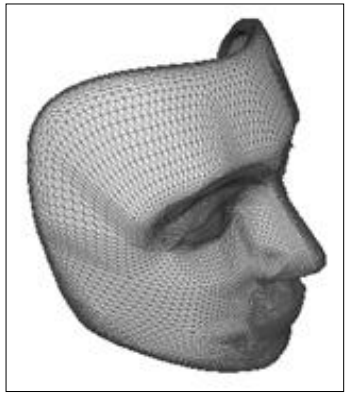
Jibo

Sophia

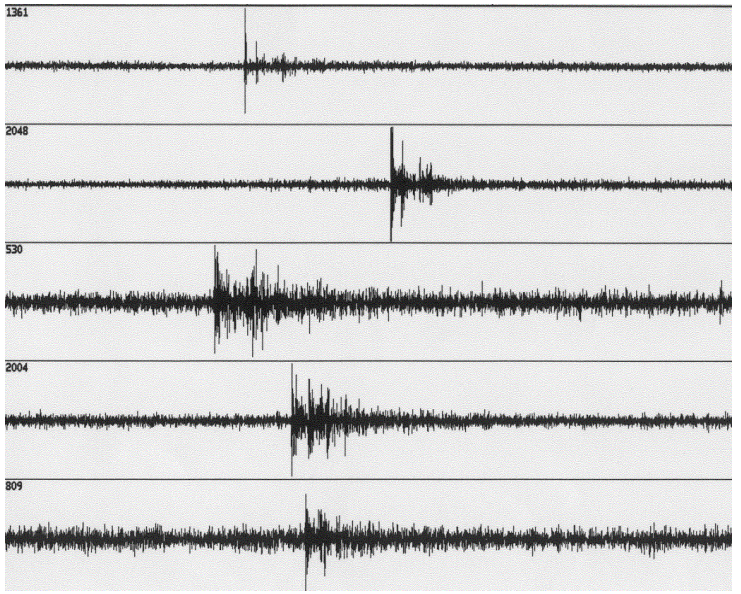


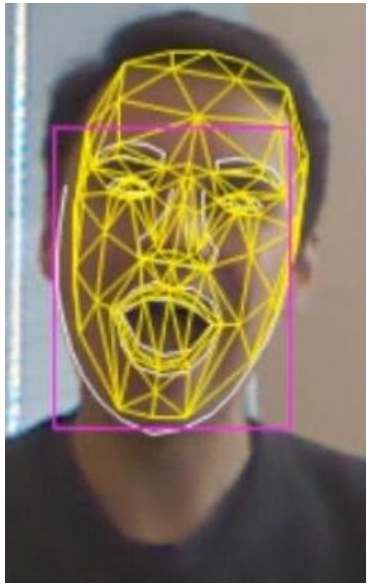
Animated agents

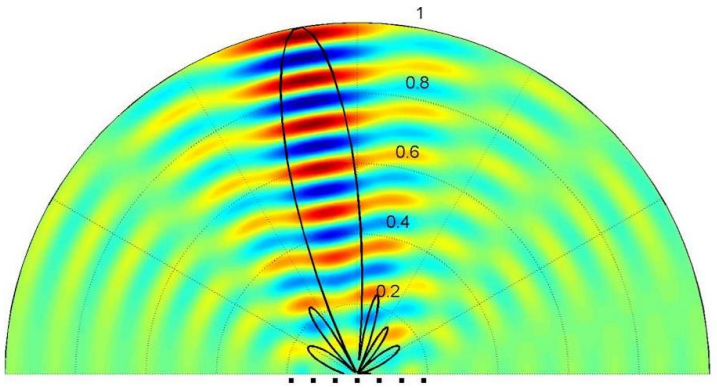
Furhat – a backprojected robot head



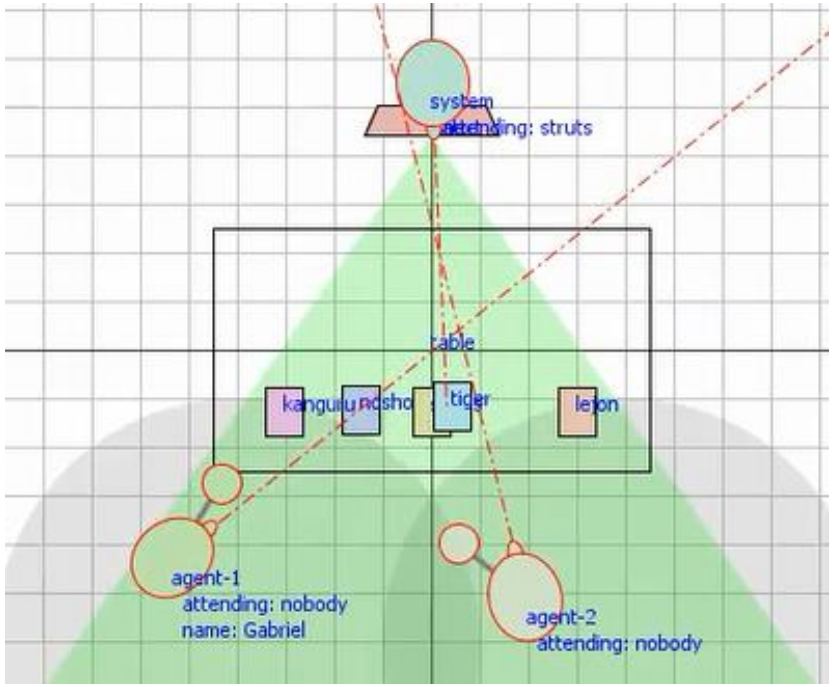
Multi-modal sensors





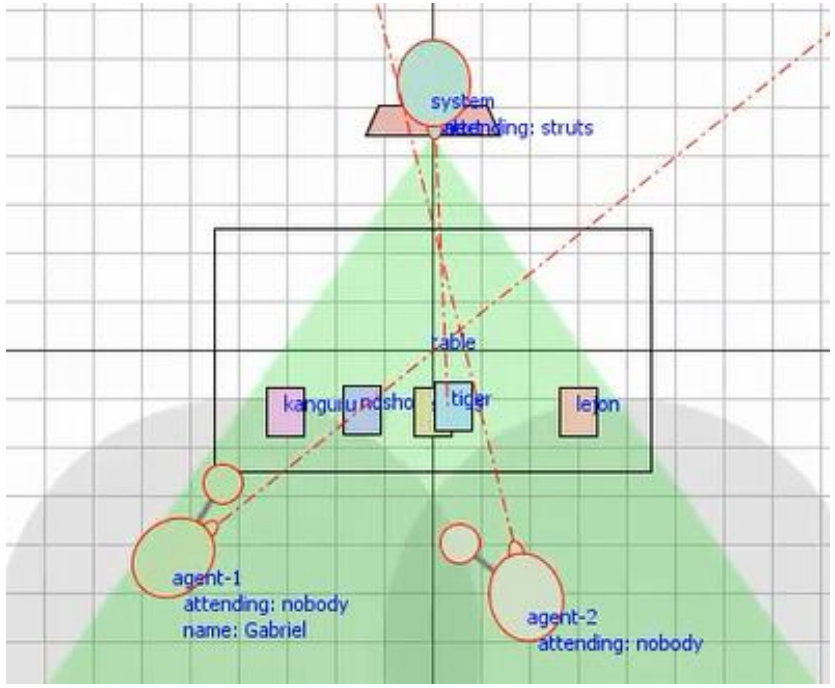


Modelling the situation



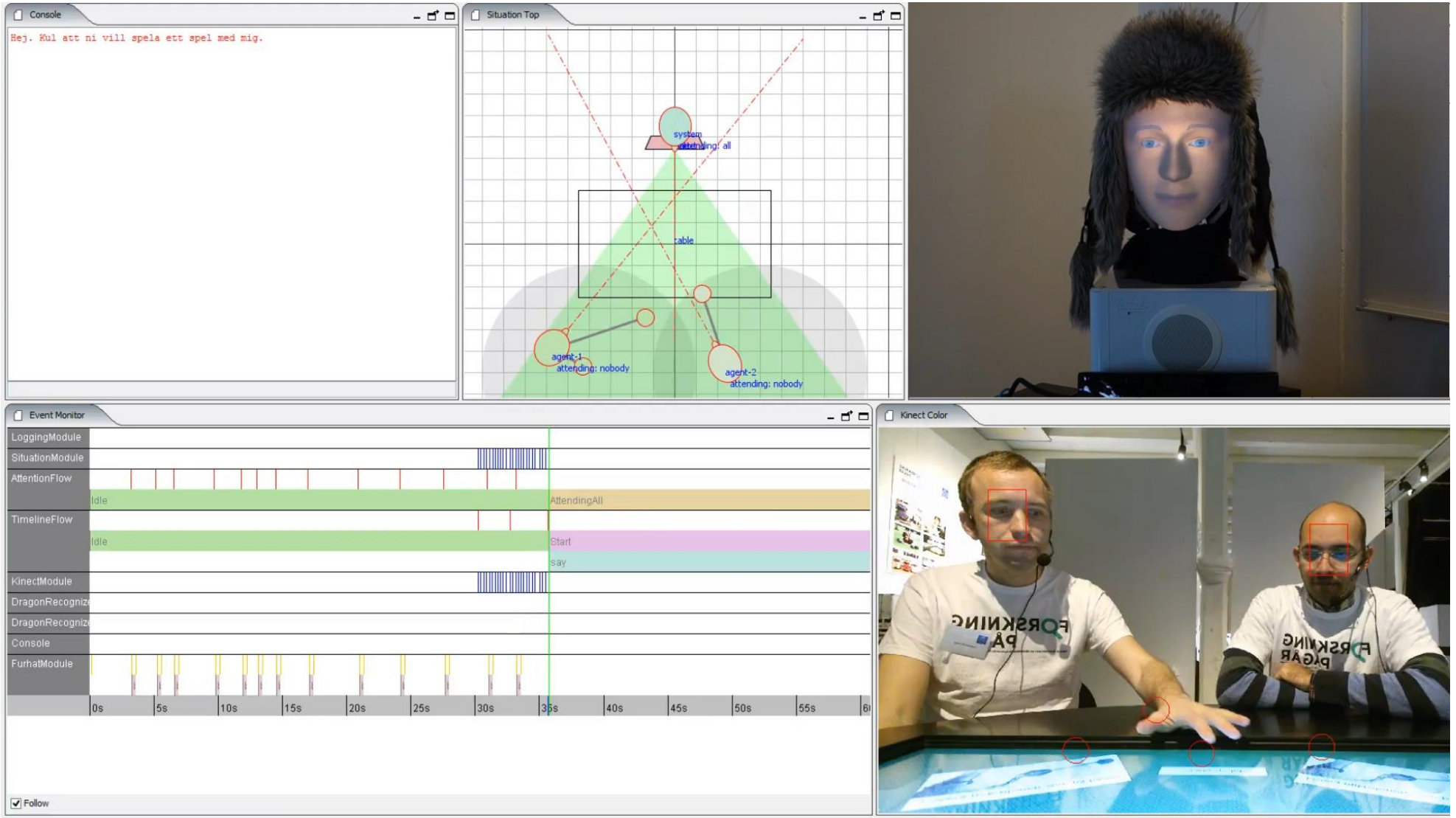
Where are the users located?
Where are the users attending?
Where are objects located?
Who is speaking?
Which object is being talked about?

Controlling the robot



Where should Furhat be looking?
When should Furhat speak?
What should Furhat say?
What facial expressions to use?

Example interaction



ROYAL INSTITUTE
OF TECHNOLOGY

The End

DT2151 – Project in Conversational Systems (period 2)
DT2112 – Speech Technology (period 3)
DT2140 – Multimodal Interaction and Interfaces (period 2)
DT2119 – Speech and Speaker Recognition (period 4)