

Spoken Dialogue for Social Robots

Tatsuya Kawahara
(Kyoto University, Japan)



Kristiina Jokinen
(AIST AI Center, Japan)



Spoken Dialogue Systems (SDS) are prevailing

- Smartphone Assistants

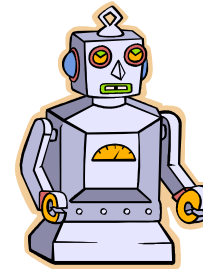


- Smart Speakers



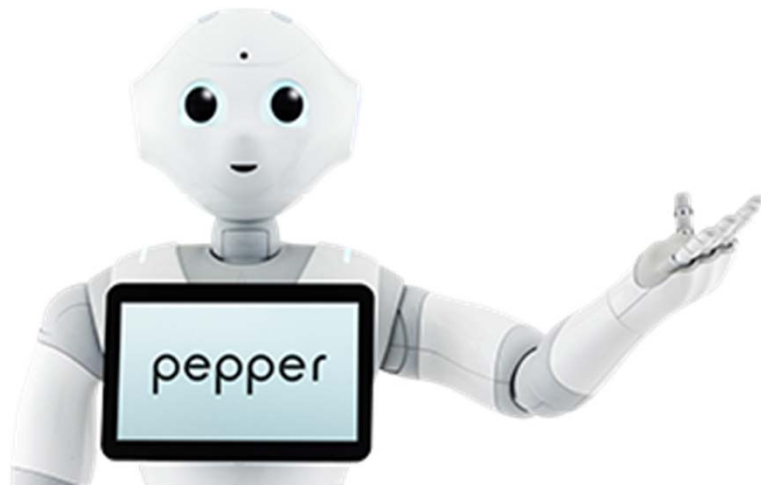
What about Social Robots?

- Social Robots
Intended for interaction with human



A majority of Peppers are returned without renewing rental contracts

2015



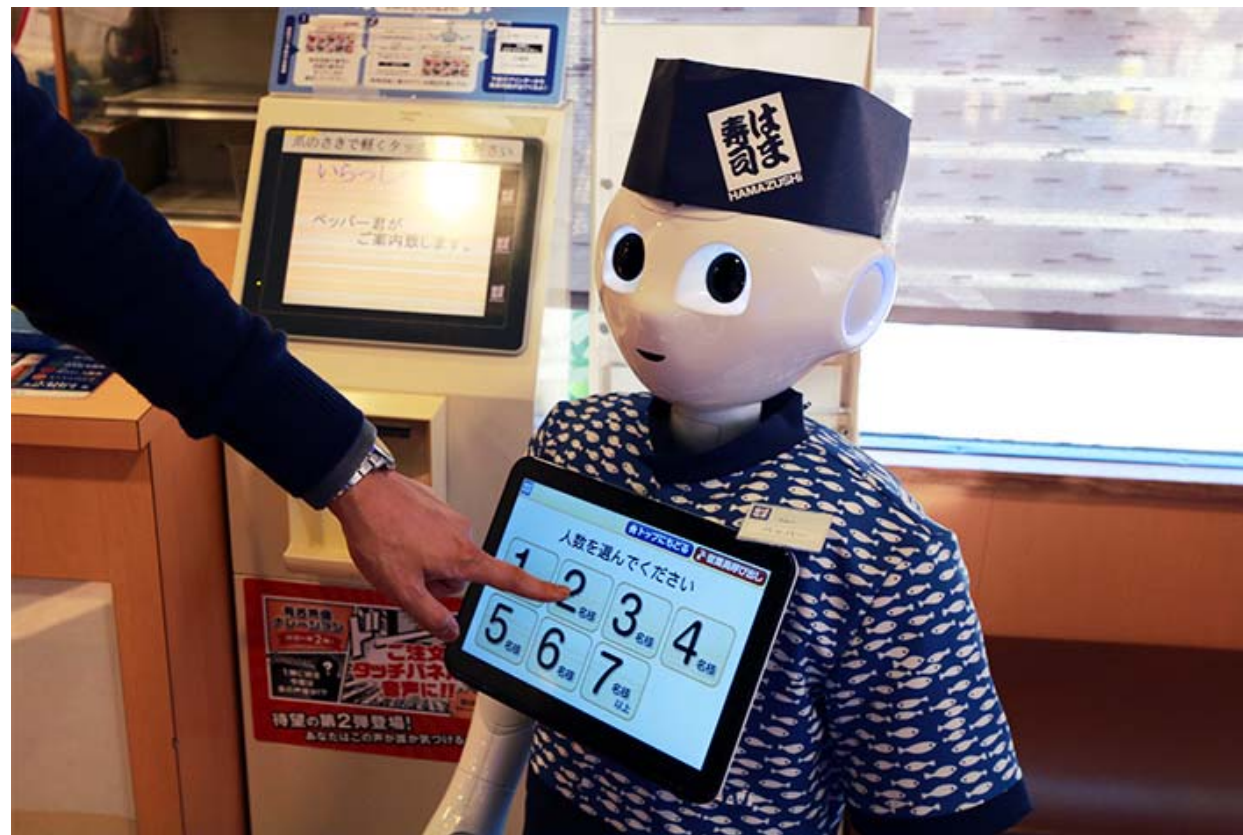
© Softbank

3 years later



2018

In successful cases, speech input is not used



© Softbank

Hen na Hotel with robot receptionists



Female android

Dinosaur robot

Critical interaction
such as check-in is
done with touch panel

<https://youtu.be/zx13fyz3UNg>

©価格.com

Robots are in many nursing homes,
but do not make speech interaction (effectively)

Paro



© Daiwa House

Palro



© Fuji soft

03.06.19

One of the decade's most hyped robots sends its farewell message

"Thank you very, very much for having me around," the social robot Jibo told its users this week.



<https://www.fastcompany.com/90315692/one-of-the-decades-most-hyped-robots-sends-its-farewell-message>₈

Still 5 Robots are Chosen in [TIME Magazine 100 Best Inventions 2019](#)

Tutor



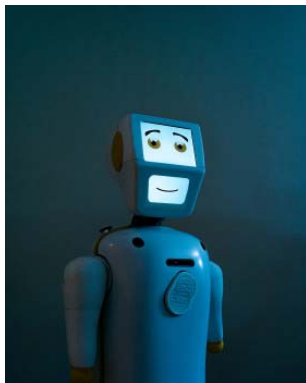
Delivery



Porter in Hospital



Companion for Elderly



Home Robot



Under COVID-19 Robots Became Essential Workers (IEEE Spectrum)



Delivering goods



Checking patients (online)



Monitoring visitors



Agenda (Research Questions)

0. Why social robots are not prevailing in society?
1. What kind of **tasks** are social robots expected to conduct?
2. What kind of social **robots** are suitable for the tasks?
3. Why **spoken dialogue** is not working well with robots?
4. What kind of **non-verbal and other modalities** are useful?
5. What kind of system **architectures** are suitable?
6. What kind of **ethical** issues must be considered?

What?

Who?

How?

Agenda (Research Questions)

0. Why social robots are not prevailing in society?
1. What kind of **tasks** are social robots expected to conduct?
2. What kind of social **robots** are suitable for the tasks?
3. Why **spoken dialogue** is not working with robots?
 1. ASR and TTS
 2. SLU+DM (end-to-end?)

4. What kind of **non-verbal and other modalities** are useful?
 1. Backchannel, turn-taking
 2. Eye-gaze

5. What kind of system **architectures** are suitable?
6. What kind of **ethical** issues must be considered?


} optional

break

Kawahara

Jokinen

0. Why social robots are not prevailing in society?

- Basically cost issue
 - Hardware expensive & fragile → maintenance
 - Much more expensive (>10 times) than smart speakers
 - Benefit does NOT meet the cost
- 
- **Tasks (=what robots can do)** are limited or irrelevant
 - Many tasks can be done via smartphones and smart speakers
 - **Spoken Language Interaction experience** is poor
 - Compared with smartphones and smart speakers
 - While expectation is high

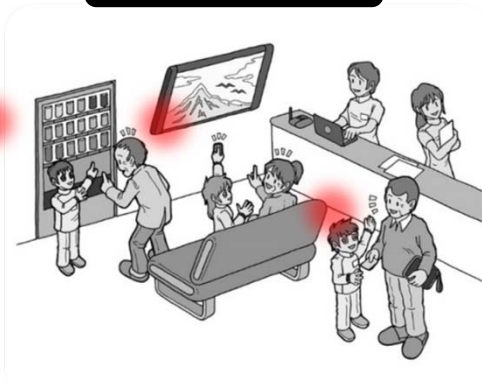
1. What kind of tasks are social robots expected to conduct?

Expected Roles by Robots

Physical presence &
Face-to-Face interaction
matters

Receptionist

receive=welcome



Attendant

attend=care



Health care

→ Senior



Teaching

→ Children



Still 5 Robots are Chosen in [TIME Magazine 100 Best Inventions 2019](#)

Tutor



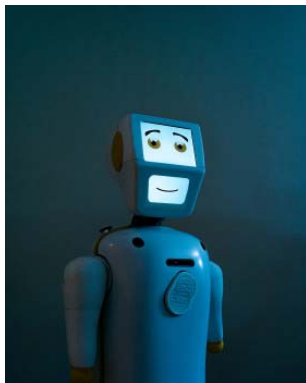
Delivery



Porter in Hospital



Companion for Elderly



Home Robot



Other Scenarios?

1. Who are typical users?
2. Where are they served?

Dialogue Category (Tasks)

Smartphone
Smart speaker

	No Resource (Dialog is task)	Information Services	Physical Tasks
Goal observable	Negotiation	Search, Order Receptionist	Manipulation Porter, Cleaner
Content definite	Debate Interview	Newscaster Tutor, Guide	
Objective shared	Counseling Speed dating	Attendant	Helper
No clear objective (socialization)	Chatting Companion		

Dialogue Category (Tasks)

- User initiative
- System initiative
- Mixed initiative

	No Resource (Dialog is task)	Information Services	Physical Tasks
Goal observable	Negotiation	Search, Order Receptionist	Manipulation Porter, Cleaner
Content definite	Debate Interview	Newscaster Tutor, Guide	
Objective shared	Counseling Speed dating	Attendant	Helper
No clear objective (socialization)	Chatting Companion		

Dialogue Category (Tasks)

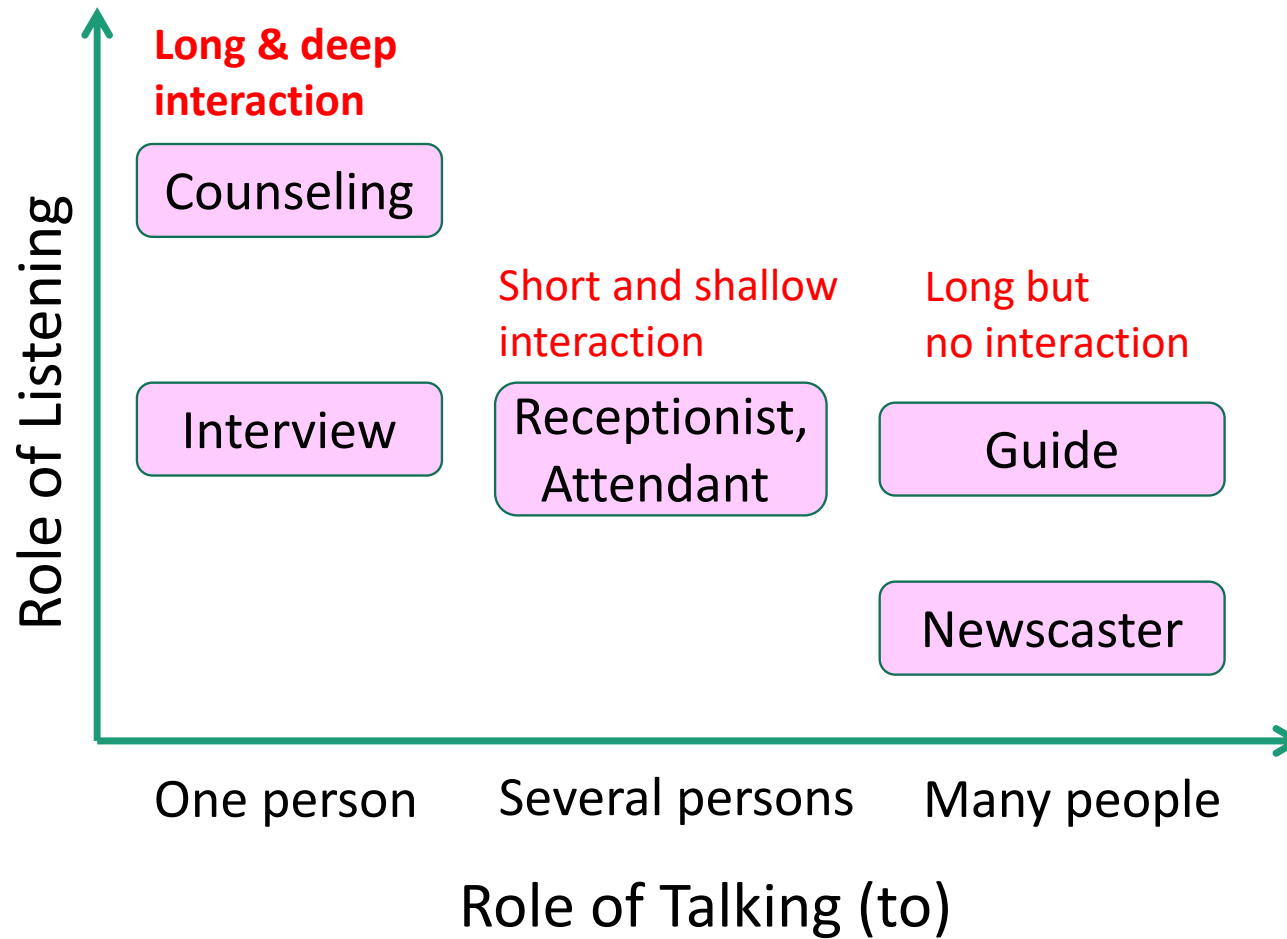
	No Resource (Dialog is task)	Information Services	Physical Tasks
Goal observable	Negotiation	Search, Order Receptionist	Manipulation Porter, Cleaner Helper
Content definite	Debate Interview	Newscaster Tutor, Guide	
Objective shared	Counseling Speed dating	Attendant	
No clear objective (socialization)	Chatting Companion		

Agent is OK?
↑

↓
Android effective?

↓
Mechanical Robot

Dialogue Roles of Adult Androids



Chatting function

- Desired in many cases (most of the tasks)
 - Ice-breaking in the first meeting
 - Relaxing during a long interaction
 - Keeping engagement
- Can be done without robots/agents (cf.) chatbot
- Will be more engaging with robots/agents

2. What kind of robots are suitable for the tasks?

Robot's Appearance → Affordance

People assume robot's capabilities based on its appearance

- Looks like a human → expected to act like a human
- Has eyes → expected to see
- Speaks → expected to understand human language and converse
 - Speaks fluently → expected to communicate smoothly
- Expresses emotion with facial expressions → expected to read emotions

[Human Robot Interaction <https://www.human-robot-interaction.org/>

Chapter 4]

Animal (Non-Humanoid) Robots Stuffed Animals Talking (some listening)

- Aibo



© SONY

- Paro



© Daiwa House

- ???



Substitute of a pet

Child-like or Child-size Humanoid Robots

- CommU



©VSTONE, Osaka U

- Nao



© Softbank robotics

- Palro



© Fuji soft

Substitute of a grandchild

Adult-size Humanoid Robots

- Robovie



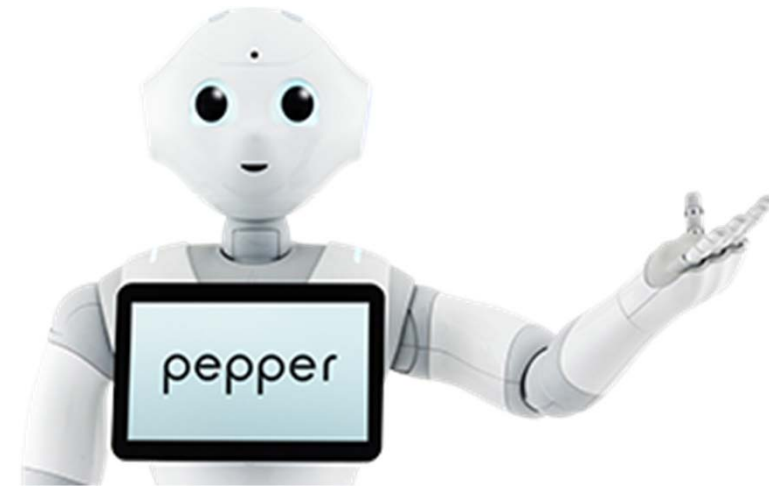
© ATR

- Asimo



© HONDA

- Pepper



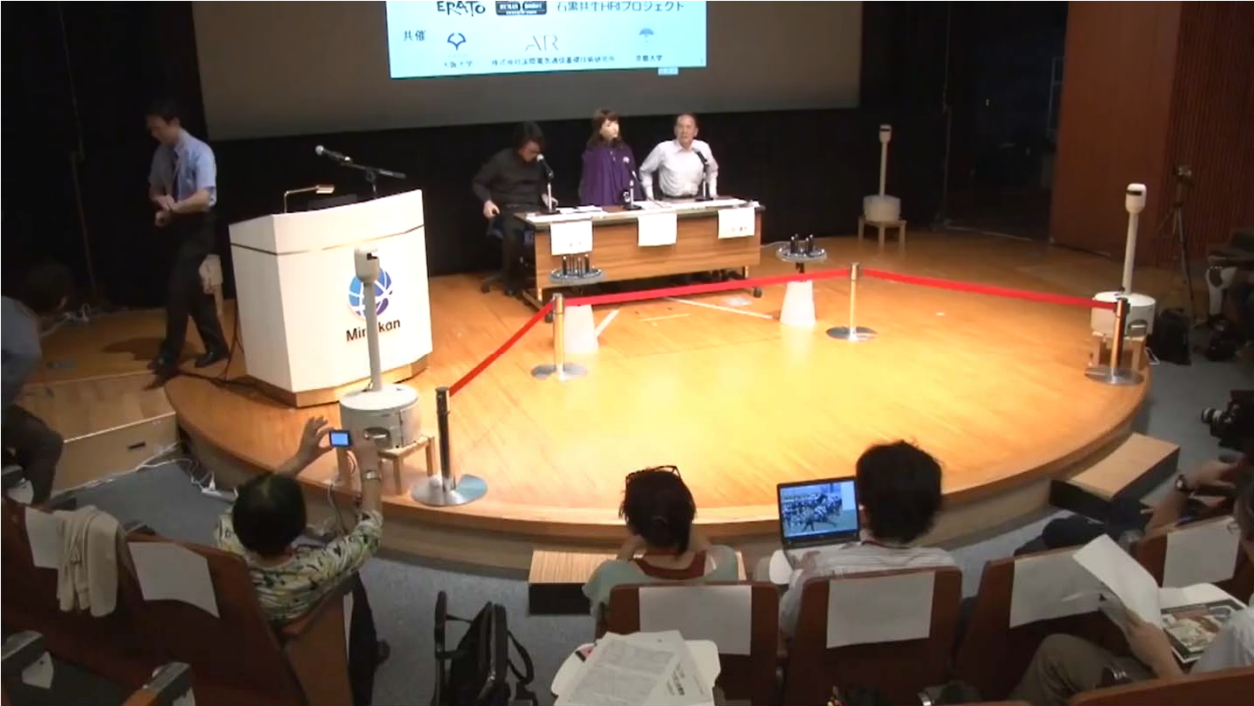
© Softbank

Expected to do something

But still child-like! → Implying not so intelligent

Adult Androids

- ERICA



Debut in 2015

How long can you keep talking?

- Smart Speaker

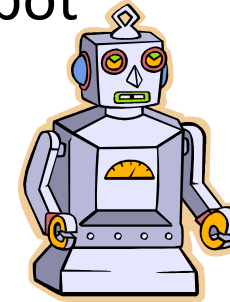


- Virtual Agent



MMD Agent ©NITECH

- Humanoid Robot



- Human

(A person you meet for the first time)

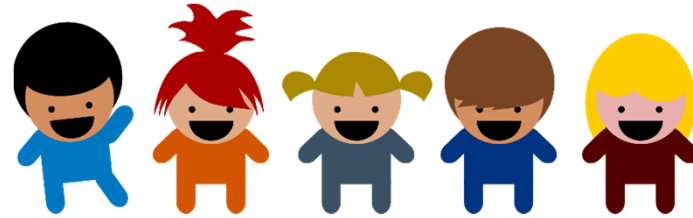


How long can you keep talking (about one story)?

- Pet



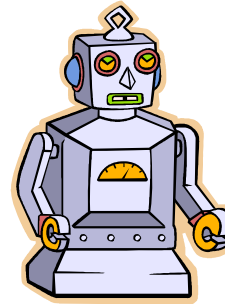
- Baby



- Kid (~10 year old)






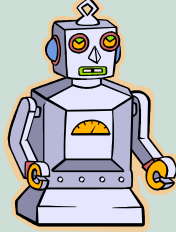

- Humanoid



- Android



Comparison of Dialogue Interfaces

	Smart Speaker	Virtual Agent	Pet Robot	Humanoid Robot	Adult Android
					
Would like to have at home?					
Would like to have at office?					
Asking today's schedule					
Talking about your life					
Companion for senior					

Would like to have at home?




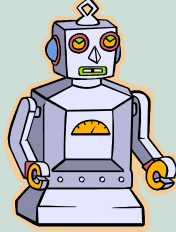

Would like to have at office?

Asking today's schedule

Talking about your life

Companion for senior

Comparison of Dialogue Interfaces

	Smart Speaker	Virtual Agent	Pet Robot	Humanoid Robot	Adult Android
					
Would you give a nickname?					
???					
???					
???					
???					

Would you give a nickname?

???

???

???

???

Difference between Virtual Agents vs. Humanoid Robots/Androids?

- **Physical presence** + mobility
- **Multi-modality** + flexibility
 - Hard to make mutual gaze with virtual agents
- Robots are deemed to be more autonomous than agents
 - Move and act autonomously
 - Can be a partner
- ???

BUT

- Robots are expensive and difficult to install and maintain

Physical Presence of Robots

- Attract people
 - Can robots hand out flyers on the street better than human?
 - Can robots attract people to (izakaya) restaurant better than human?
 - Effective in the beginning
 - Especially for kids and senior people
- ↓ hopefully
- Attachment
- unfortunately
- Bullying by group of kids
(cf.) Virtual agents cursed

Physical Presence of Robots **NOT NECESSARY**

- When the task goal is information exchange and the user is collaborative
- Information exchange tasks
 - Must be done efficiently/ASAP
 - Short interaction (command, query) → smartphones, smart speakers
 - Long interaction (news, tutor) → virtual agents
- Not 'collaborative' users
 - Kids and senior people who do not understand the protocol

Dialogue Category (Tasks)

	No Resource (Dialog is task)	Information Services	Physical Tasks
Goal observable	Negotiation	Search, Order Receptionist	Manipulation Porter, Cleaner Helper
Content definite	Debate Interview	Newscaster Tutor, Guide	
Objective shared	Counseling Speed dating	Attendant	
No clear objective (socialization)	Chatting Companion		

Agent is OK?
↑

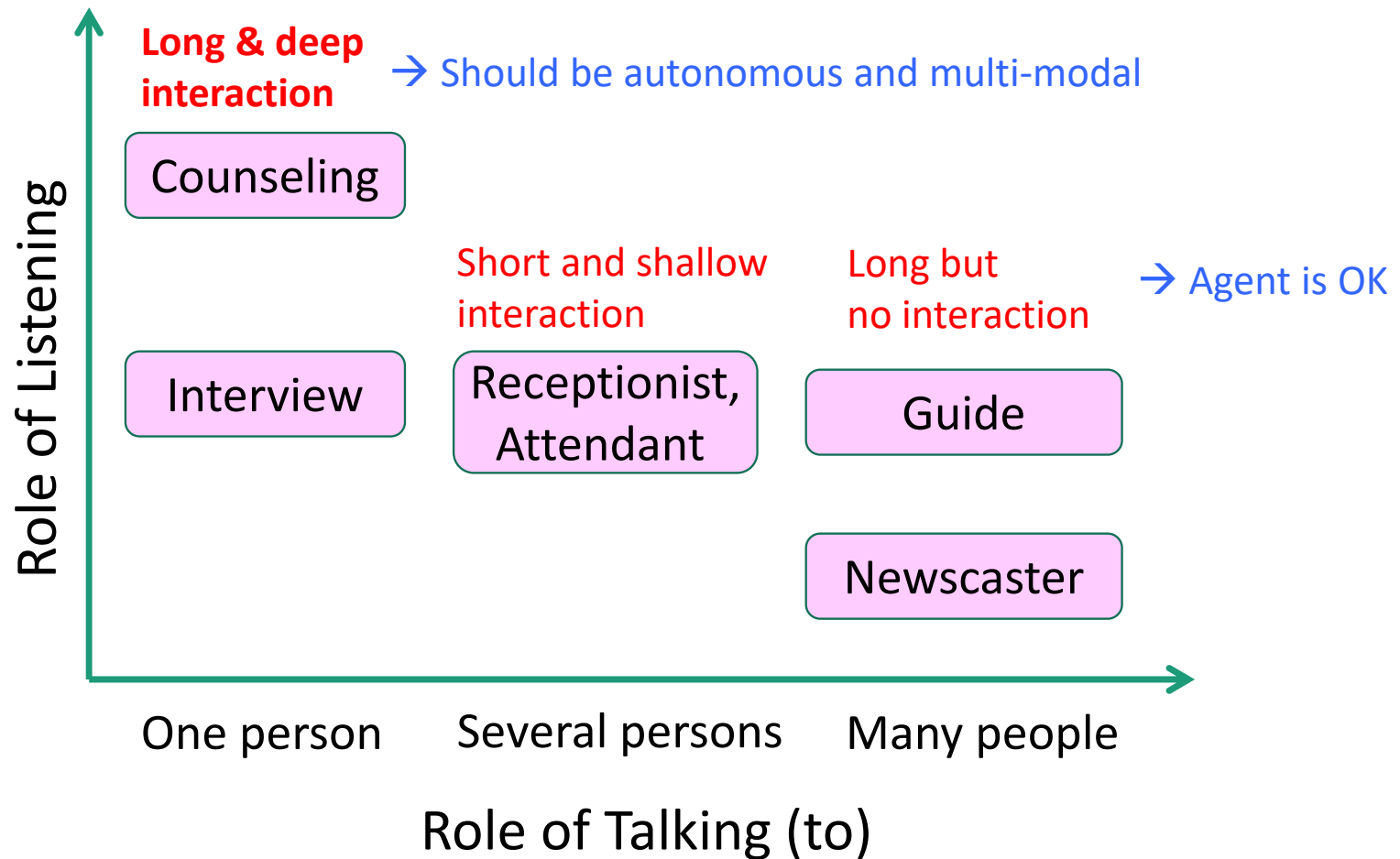
↓
Android effective?

↓
Mechanical Robot

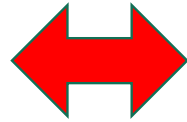
Face-to-Face Multimodal Interaction

- Necessary for long and deep interaction
 - Talk about troubles or life
(ex.) counseling
 - To know communication skills and personality
(ex.) job interview, speed dating
- Multimodality
 - **Mutual gaze**...possible only with **adult androids** (?)
 - Head/body orientation
 - Hand gesture
 - Nodding

Dialogue Roles of Adult Androids



Tool

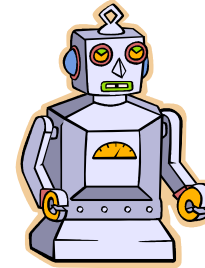


Companion, Partner

- Smartphone Assistants



- Communicative Robots



- Smart Speakers



3. Why spoken dialogue is NOT working well with robots?

Agenda (Research Questions)

0. Why social robots are not prevailing in society?
1. What kind of **tasks** are social robots expected to conduct?
2. What kind of social **robots** are suitable for the tasks?
3. Why **spoken dialogue** is not working with robots?
 1. ASR and TTS
 2. SLU+DM (end-to-end?)

4. What kind of **non-verbal and other modalities** are useful?
 1. Backchannel, turn-taking
 2. Eye-gaze

5. What kind of system **architectures** are suitable?
6. What kind of **ethical** issues must be considered?

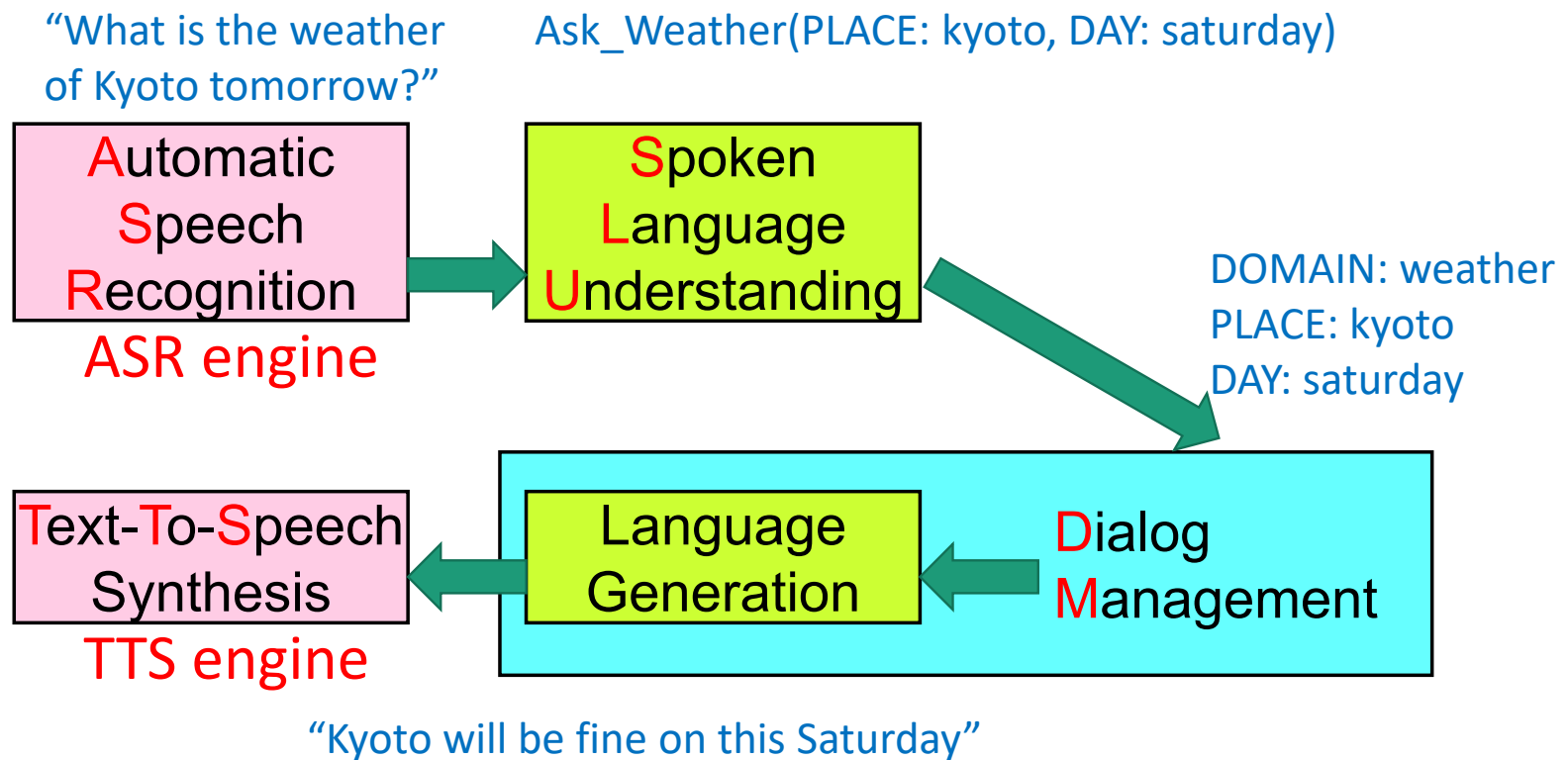
} optional

break

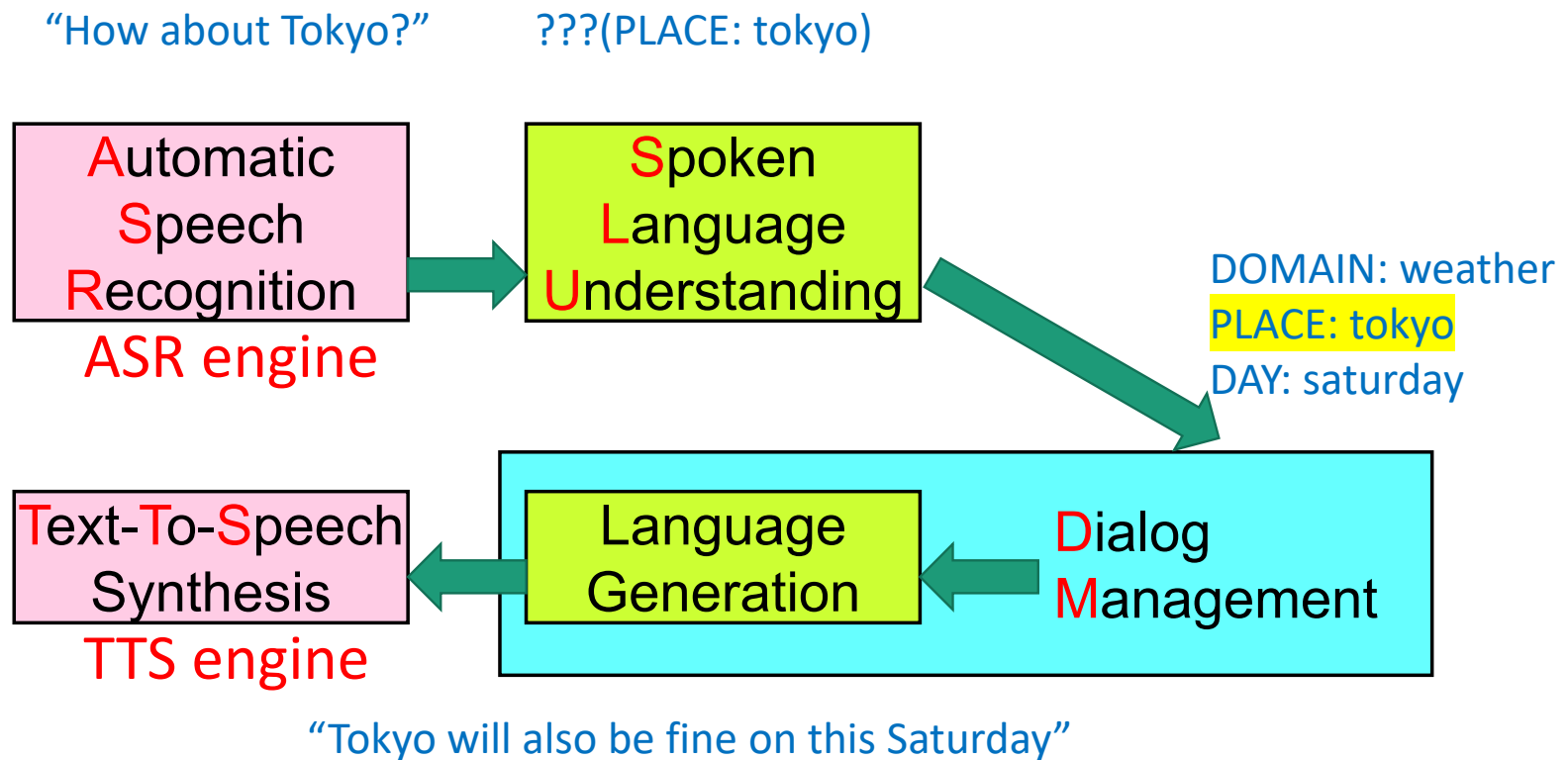
Kawahara

Jokinen

Architecture of Spoken Dialogue System (SDS)



Architecture of Spoken Dialogue System (SDS)



Automatic Speech Recognition (ASR)

Challenges for Automatic Speech Recognition (ASR) for Robots

- **Distant** speech
 - Speaker localization & identification
 - Detection of speech (addressed to the system)
 - Suppression of noise and reverberation
- **Conversational** speech
 - Speech similar to those uttered to human (pets, kids) rather than machines
 - Typical users are kids and senior people
- **Realtime** response
 - Cloud-based ASR servers have better accuracy, but large latency
 - Talking similar to international phone calls

Problems in Distant Speech

- Speaker localization & identification
- Detection of speech (addressed to the system)
- Suppression of noise and reverberation

Smart Speakers

- Don't care
- Use magic words
- Implemented



Maybe applicable to small (personal) robots

- One person
- Not so distant

Problems in Distant Speech

- Speaker localization & identification
- Detection of speech (addressed to the system)
- Suppression of noise and reverberation

Adult humanoid robots

→ with camera

→ ???

→ Implemented



Multi-modal processing

Detection of Speech addressed to the System

- Eye-gaze (head-pose)...most natural and reliable
- Content of speech
- Prosody of speech



- Machine learning
 - Not accurate enough ← must be close to 100%

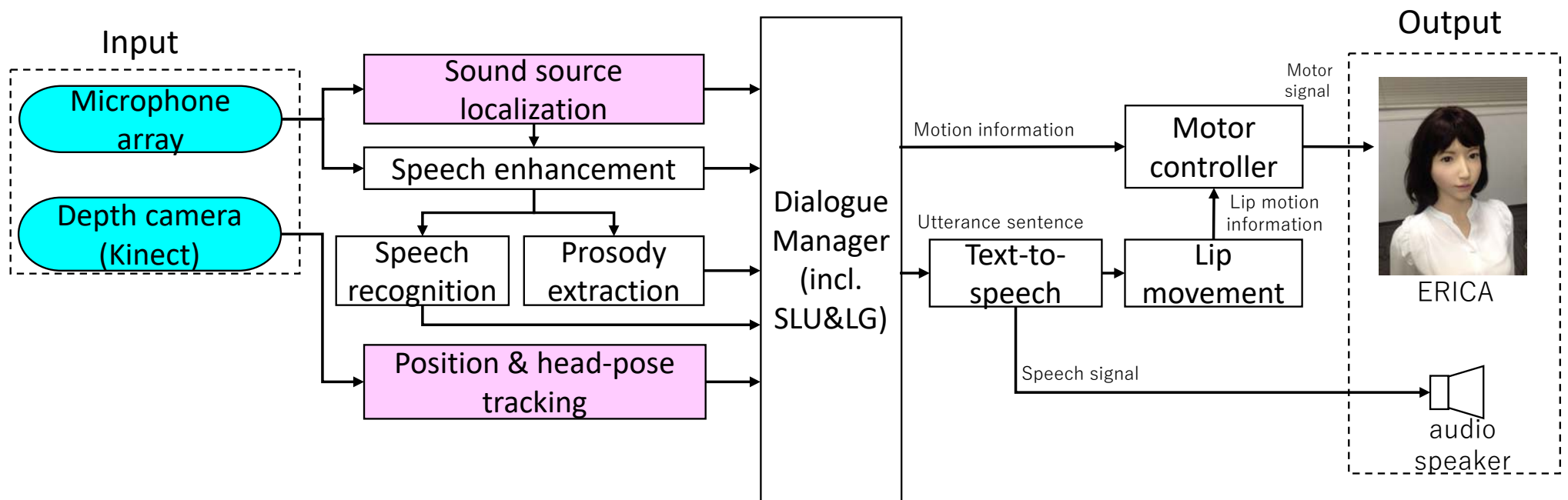


- Incorporation of turn-taking model
 - Context is useful

Example Implementation for ERICA



Example Implementation for ERICA



Real Problem in Distant Talking

- When people speak without microphone, speaking style becomes so casual that it is **NOT easy to detect utterance units**.
 - False starts, ambiguous ending and continuation



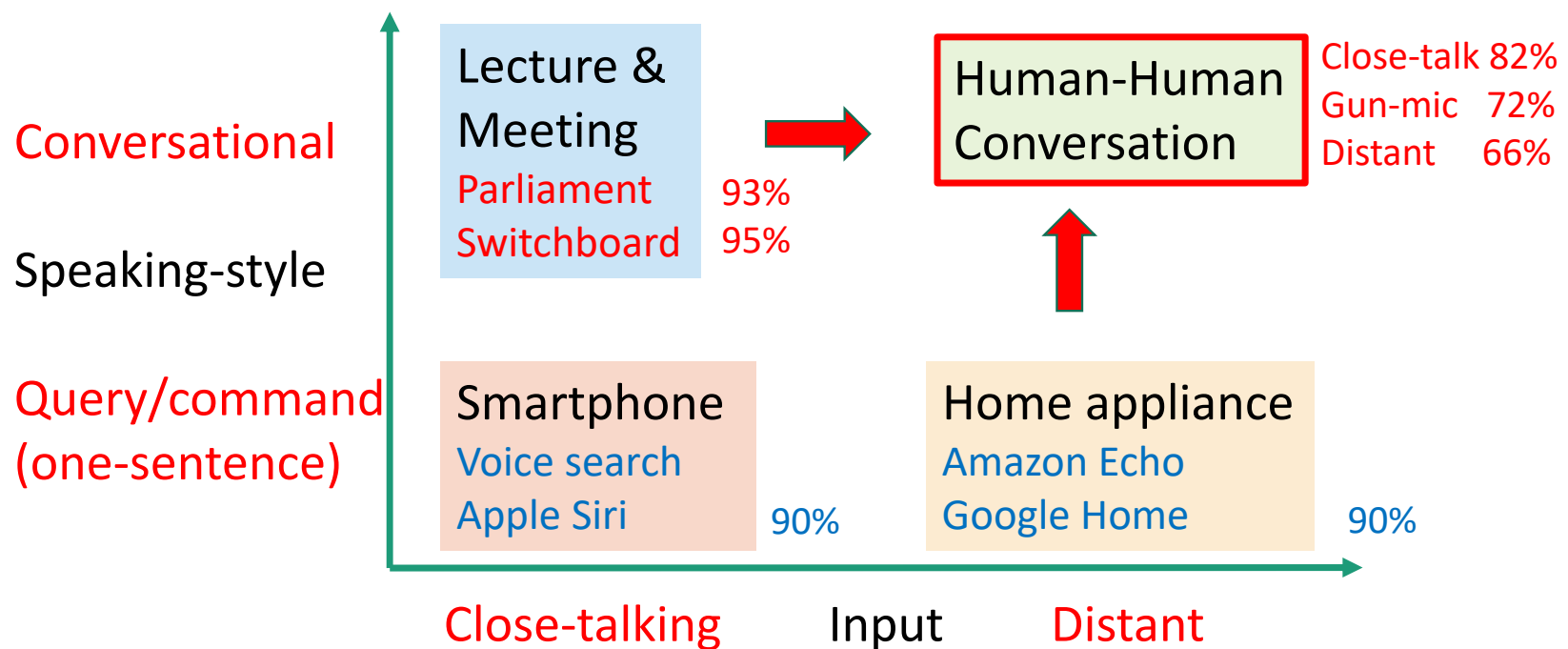
- Not addressed in conventional “challenges”
- Circumvented in conventional products
 - Smartphones: push-to-talk
 - Smart speakers: magic word “**Alexa**”, “**OK Google**”
 - Pepper: talk when flash



- Incorporation of turn-taking model
 - Context is useful

Distant & Conversational Speech Recognition

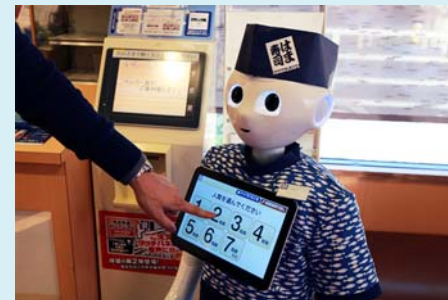
Accuracy is degraded with the synergy of two factors



Review of ASR

Error Robustness and Recovery

- Task and interaction need to be designed to work with low ASR accuracy
 - Attentive listening
- Confirmation of critical words for actions
 - Command & control
 - Ordering
- Error recovery is difficult
 - Start-over is easier for users, too
- Use of GUI?



© Softbank

Review of ASR

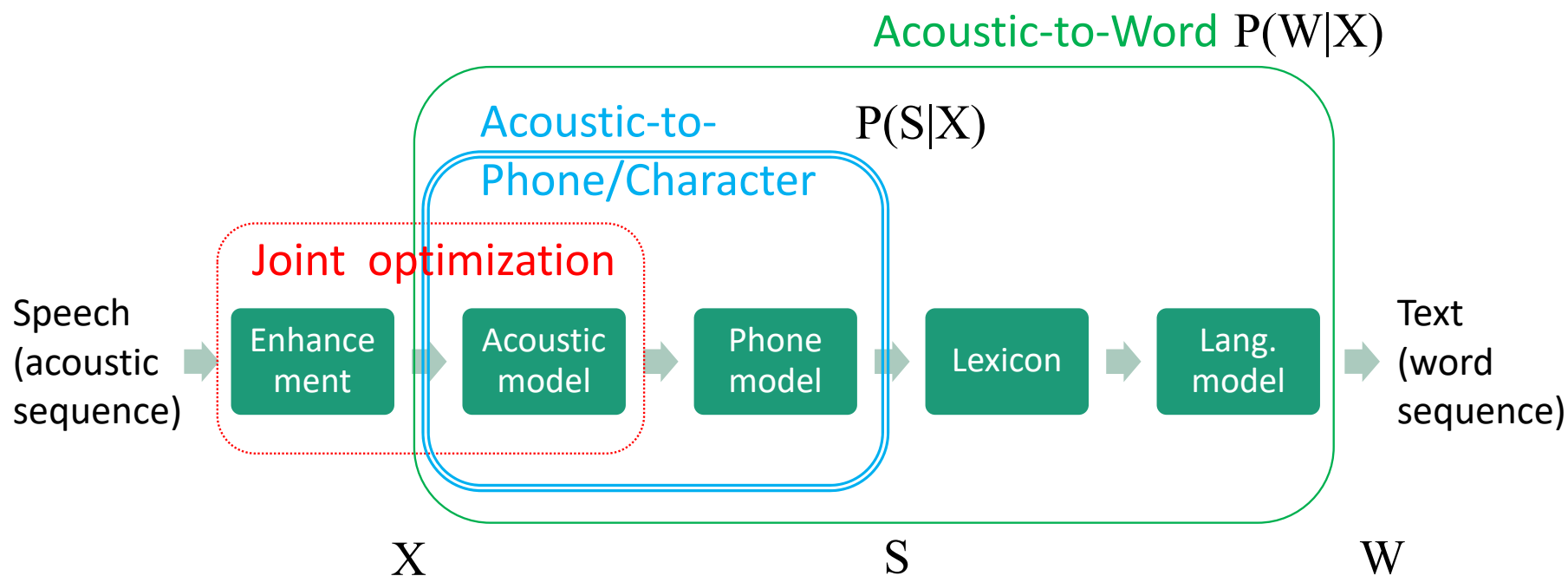
Latency is Critical for Human-like Conversation

- Turn-switch interval in human dialogue
 - Average ~500msec
 - 700msec is too late
 - difficult for smooth conversation (cf.) oversea phone calls
- Many cloud-based ASR hardly meets requirement

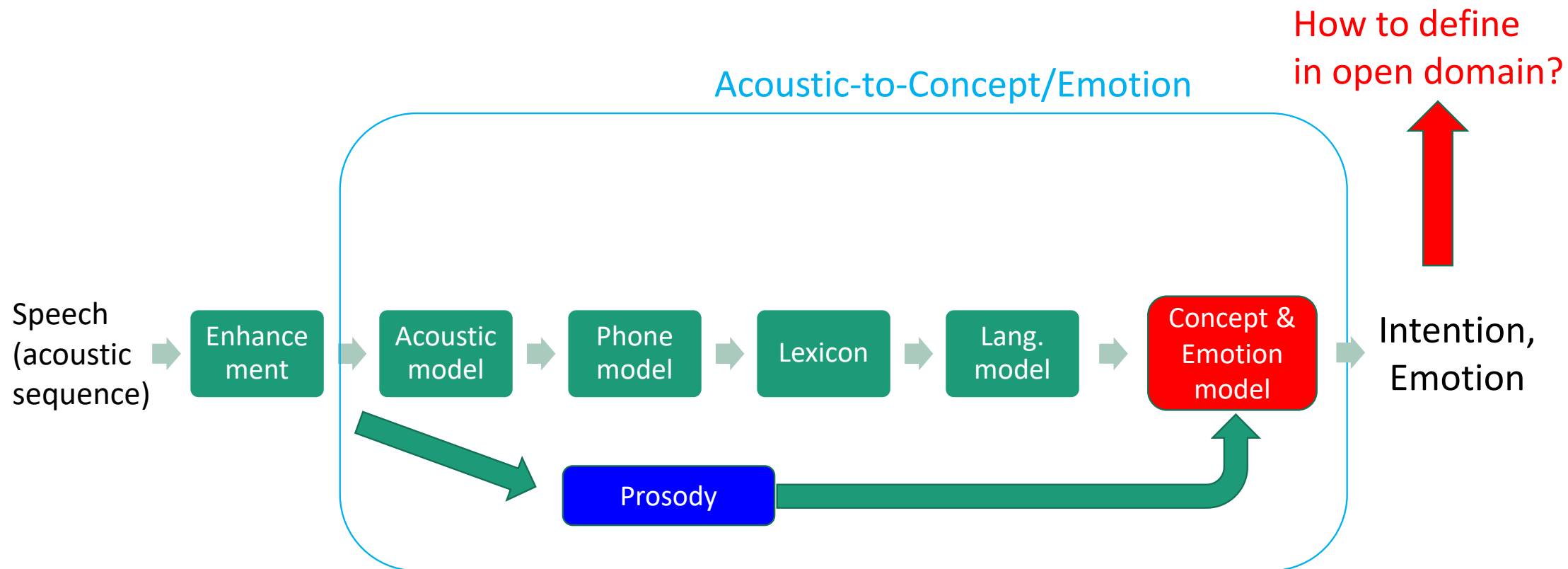


- Recent Development of Streaming End-to-End ASR
- All downstream NLP modules must also be tuned

End-to-End Automatic Speech Recognition (ASR)



End-to-End Speech Understanding



Text-To-Speech Synthesis (TTS)

Requirements in Text-To-Speech Synthesis (TTS)

- Very high quality
 - Intelligibility
 - Naturalness **matched to the character** (pet, kid, mechanical, humanoid)
- **Conversational** style rather than text-reading
 - Questions (direct/indirect)
- A variety of non-lexical utterances with a variety of prosody
 - **Backchannels**
 - **Fillers**
 - **Laughter**



Hardly implemented
in conventional TTS

End-to-End Text-To-Speech Synthesis (TTS)

Tacotron 2 (2017-)

- Seq2seq model: char. seq. → acoustic features
- Wavenet: acoustic features → waveform
- “Comparable-to-Human performance”
 - Mean Opinion Score (MOS) 4.53 vs. 4.58

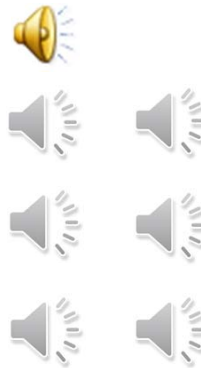
<https://google.github.io/tacotron/publications/tacotron2/>

Turing Test: Tacotron 2 or Human?

Voice of Android ERICA

Conversation-oriented

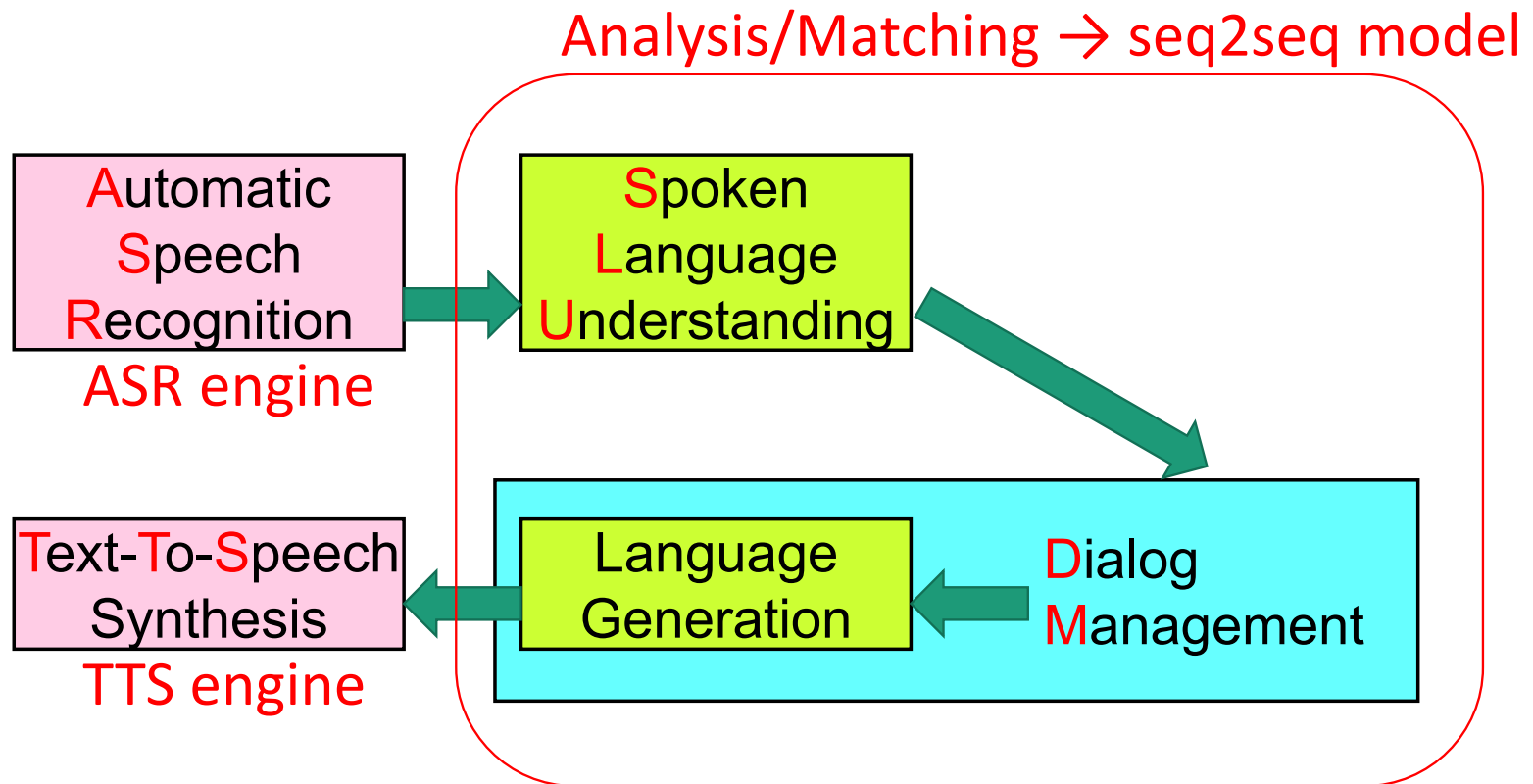
- Backchannels
- Filler
- Laughter



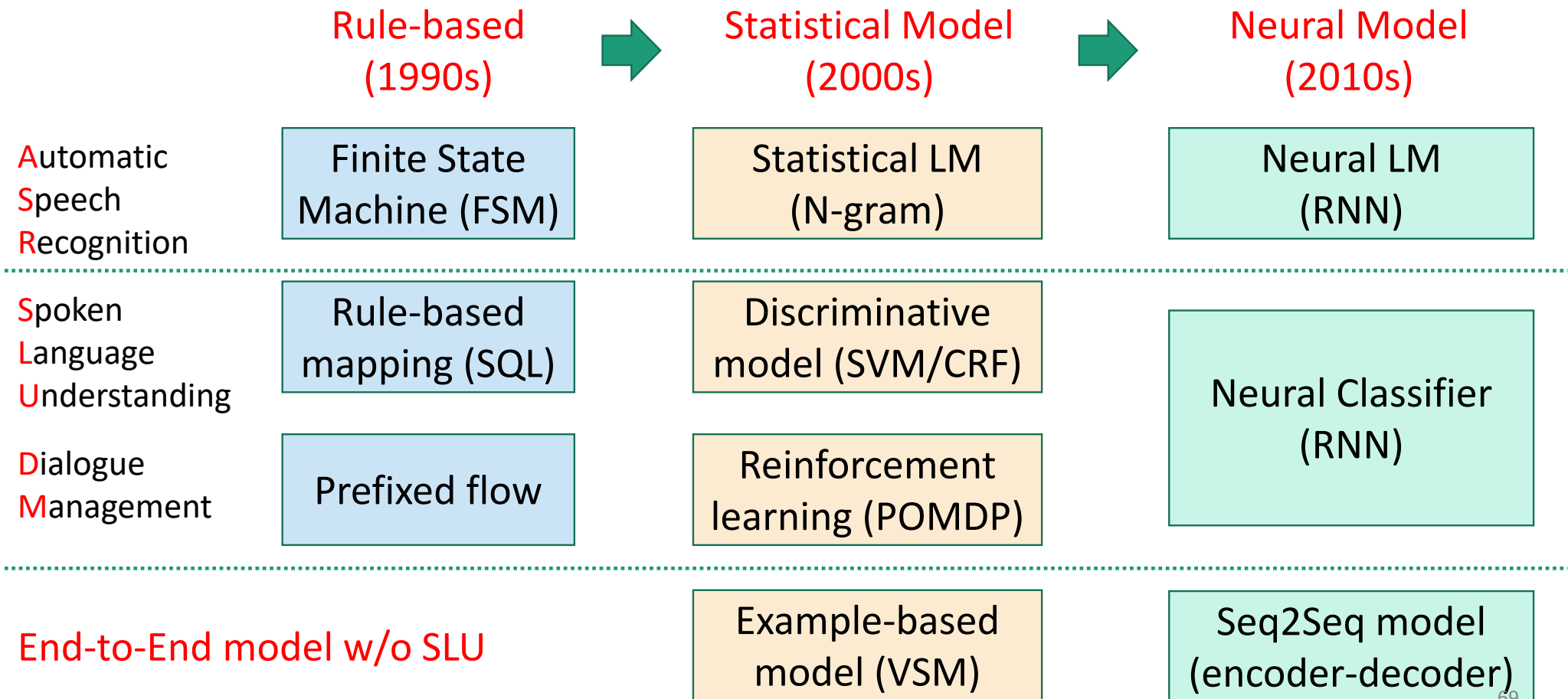
<http://voicetext.jp> (ERICA)

Spoken Language Understanding (SLU)
and
Dialogue Management (DM)

Architecture of Spoken Dialogue System (SDS)

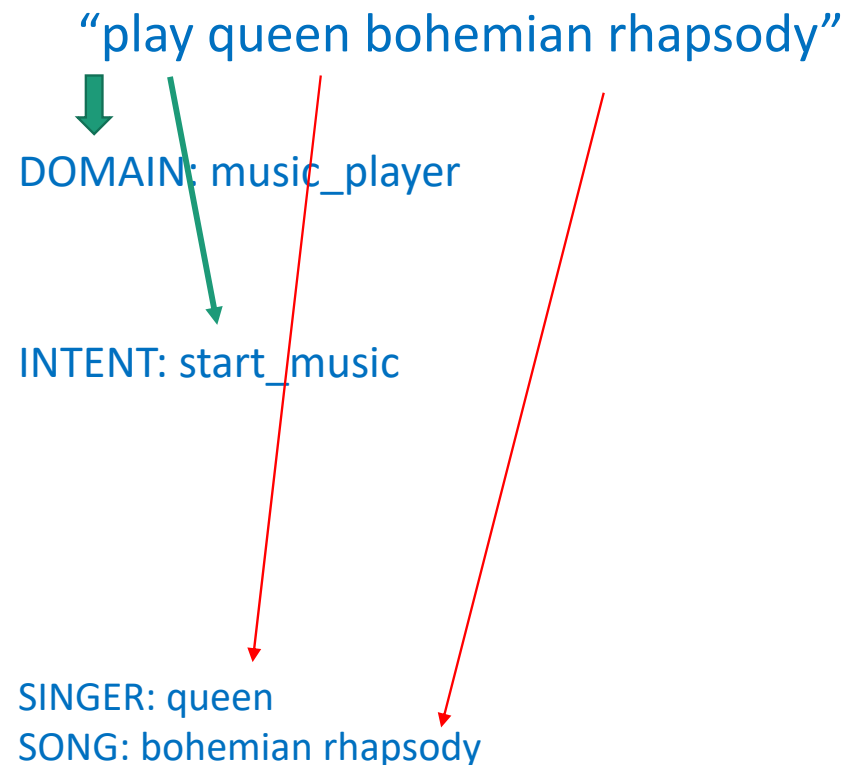


Historical Shift of Methodology



Semantic Analysis for SLU

- **Domain**
(ex.) weather, access, restaurant
- **Intent**
 - Many domains accept only one intent
(ex.) weather, access
 - Some accept many kinds of queries
(ex.) scheduler...where, when
- **Slot/Entity**
 - Named Entity (NE) tagger
 - Numerical values



Semantic Analysis for SLU

- **Domain**
(ex.) weather, access, restaurant
- **Intent**
 - Many domains accept only one intent
(ex.) weather, access
 - Some accepts many kinds of queries
(ex.) scheduler...where, when



Classification problem, given entire sentence

- Statistical Discriminative Model: SVM, Logistic Regression
- Neural Classifier: CNN, RNN



Semantic Analysis for SLU

Sequence labeling problem

- Statistical Discriminative Model: CRF
- Neural Tagger: RNN

Domain-independent NE tagger



• Slot/Entity

- Named Entity (NE) tagger
- Numerical values

“play queen bohemian rhapsody”

O B-singer B-song I-song

SINGER: queen
SONG: bohemian rhapsody

Dialogue Management

- Decide proper **Action**
 - Make query/command
 - Present results

“What is the weather of Kyoto tomorrow?”

Ask_Weather(PLACE: kyoto, DAY: saturday)

“Kyoto will be fine on this Saturday”

DOMAIN: weather
PLACE: kyoto
DAY: saturday

- Maintain **Context**

“How about Tokyo?”

“Tokyo will be cloudy on this Saturday”

DOMAIN: weather
PLACE: tokyo
DAY: saturday

Dialogue Management

“What is the weather of Kyoto tomorrow?”

- Decide proper **Action**
 - Make query/command
 - Present results

Ask_Weather(PLACE: kyoto, DAY: saturday)

“Kyoto will be fine on this Saturday”

DOMAIN: weather
PLACE: kyoto
DAY: saturday



- **Prefixed (hand-crafted) flow**
 - still pragmatic
 - Google Dialogflow, Microsoft LUIS..
- **Reinforcement learning of stochastic model (POMDP)**
 - Considers uncertainty/errors in input/processing
 - Difficult for maintenance, minor fix
- **Neural model?**

Incomplete or Ambiguous Queries

- Majority of actions can be done with required slots
(ex.) Weather ← place (date), Access ← destination, origin,
Take_object ← object (place)

- If some slot is missing,
or some entity is ambiguous,
the system

- needs to ask users

OR

- use a default value
 - current location/time
 - most frequently used one
- present all in GUI

“Tell me the weather?”

“Weather in Cambridge?”

“Which location?”

“Cambridge in UK or MA, USA?”

← Widely used in smartphone assistants,
but not necessarily applicable to robots
working in a real world (w/o GUI)

Disambiguation by Dialogue

SHRDLU [Winograd1972]

Operating blocks in virtual world

U: PICK UP A BIG RED BLOCK.

S: OK. (does it)

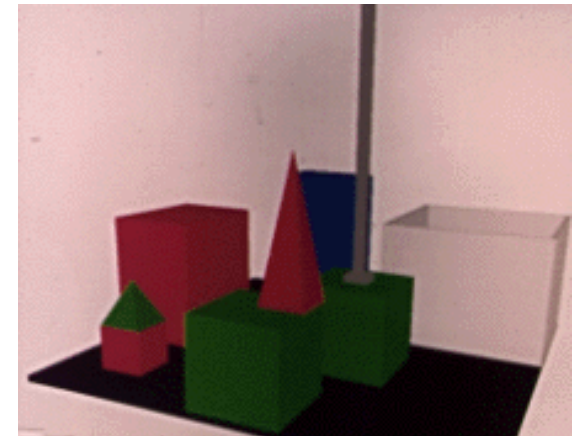
U: GRASP THE PYRAMID.

S: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.

U: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.

S: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

U: OK. (does it)



Review of Orthodox SLU+DM

- Show intelligent behaviors
- Complex and not easy to implement
- Hand-crafted rules
 - Easy and practical solution
 - Not flexible and scalable

→ Only works for **small tasks**
- Machine learning (Statistical & Neural)
 - Needs a lot of data with annotation
 - Not pragmatic for each task domain
 - Except for standard benchmark datasets and big companies
 - Artificial data generation? or Simulation?

→ Only possible for **big companies**

Alternatives of Orthodox SLU+DM

- End-to-End Dialogue Model w/o SLU
 - **Example-Based** Dialogue Model
 - **Neural** Conversation (**Seq2Seq**) model

Dialogue based on Pattern Matching

ELIZA [Weizenbaum1966]

Counseling

User: Men are all alike.

ELIZA: In what way?

User: They're always bugging us about something or other.

ELIZA: Can you think of a specific example?

User: Well, my boyfriend made me come here.

ELIZA: Your boyfriend made you come here.

Dialogue based on Pattern Matching

ELIZA [Weizenbaum1966]

Word:Rank

alike:10

Outputs

In what way?

What resemblance do you see?

always:5

Can you think of a specific example?

Really, always?

what:2

Why do you ask?

Does that interest you?

Example-Based Dialogue Model

“We are going to Tokyo for a meeting”



“we are going to PLACE for a meeting”



Input (example template)	Action / Output
what is the weather of PLACE	Weather(PLACE, today)
is PLACE fine on DAY	Weather(PLACE, DAY)
I am going to PLACE	Access(current, PLACE)
Tell me how to get to PLACE	Access(current, PLACE)
It is hot today	turn_on_airconditioner “Why don’t you have some beer?”



“Here is a direction to get to Tokyo”

Example-Based Dialogue Model

- Vector Space Model (VSM)
 - Feature: Bag-Of-Words model (1-hot vector → word embedding)
 - Metric: cosine distance weighted on content words
- Neural model
 - Compute similarity between input text and example templates (in shortlist)
 - Elaborate matching by considering context
 - Needs a training data set

Incorporation of Information Retrieval (IR) and Question Answering (QA)

- Example database...limited & hand-crafted

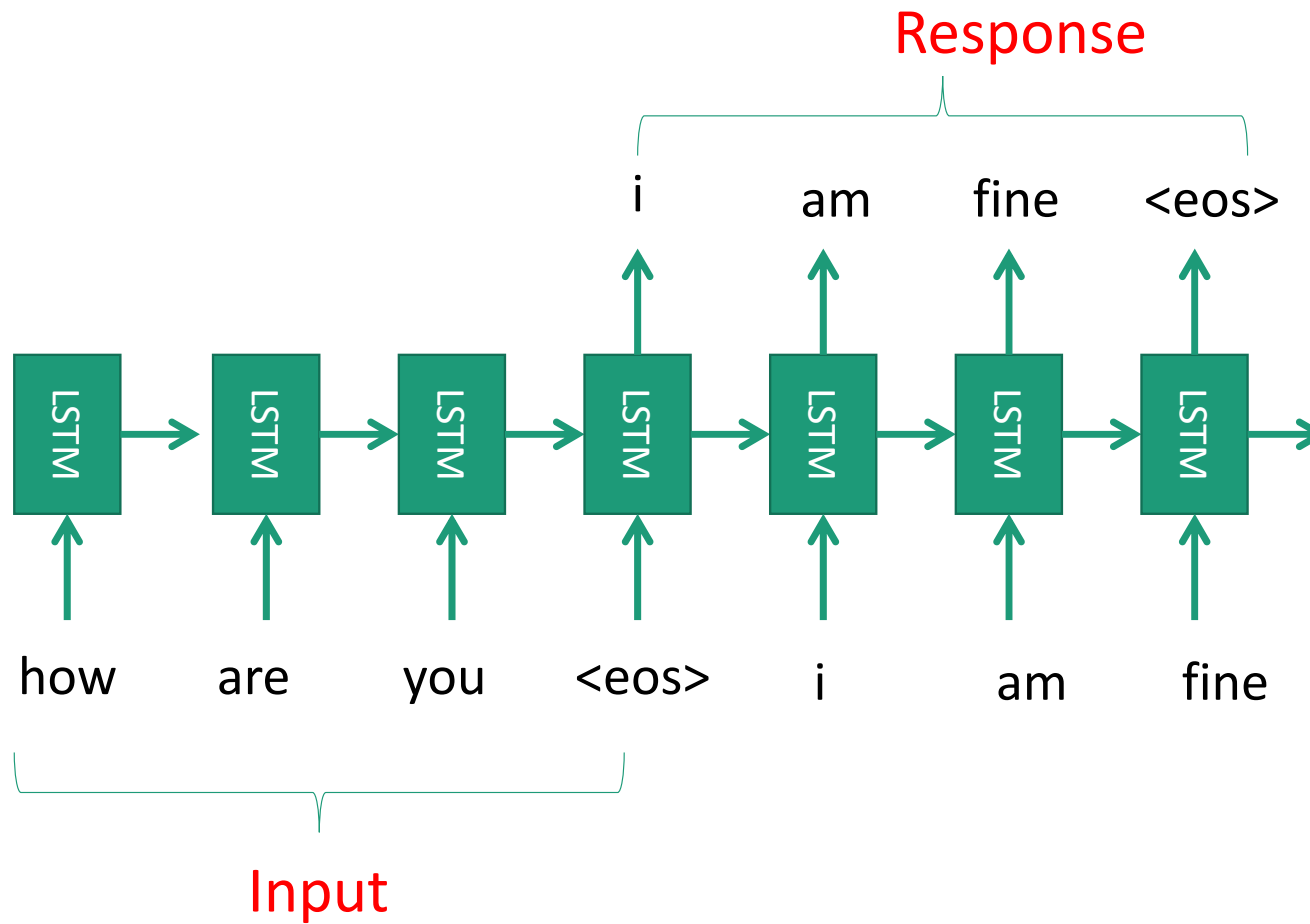


- IR technology to search for relevant text
 - Large documents or Web
 - Manuals, recipe “How can I change the battery?”
 - Wikipedia “I want to visit Kinkakuji temple”
 - news articles “How was New York Yankees yesterday?”
 - Need to modify the text for response utterance
- QA technology to find an answer
 - Who, when, where...
 - When was Kinkakuji temple built?
 - How tall is Mt. Fuji?
 - Works only with limited cases

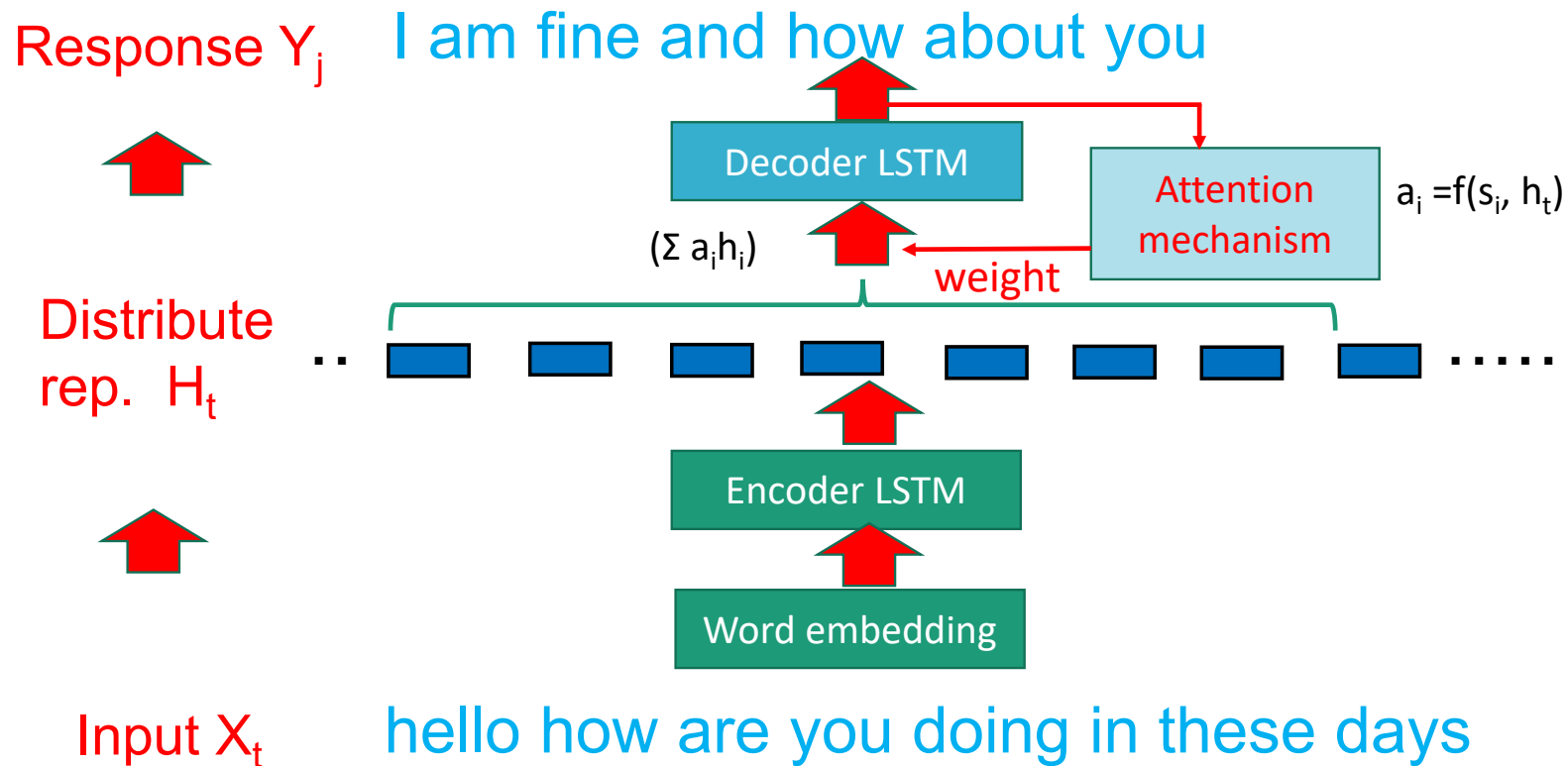
Review of Example-Based Dialogue Model

- **Easy to implement and generate high-quality responses**
 - Pragmatic solution for working systems and robots
- **Applicable only to a limited domain and not scalable**
 - ~hundreds of patterns
- Does not consider dialogue context
 - One query → One response
 - Need an anaphora resolution for “he/she/it”
 - Shallow interaction, Not so intelligent

Neural Conversation Model

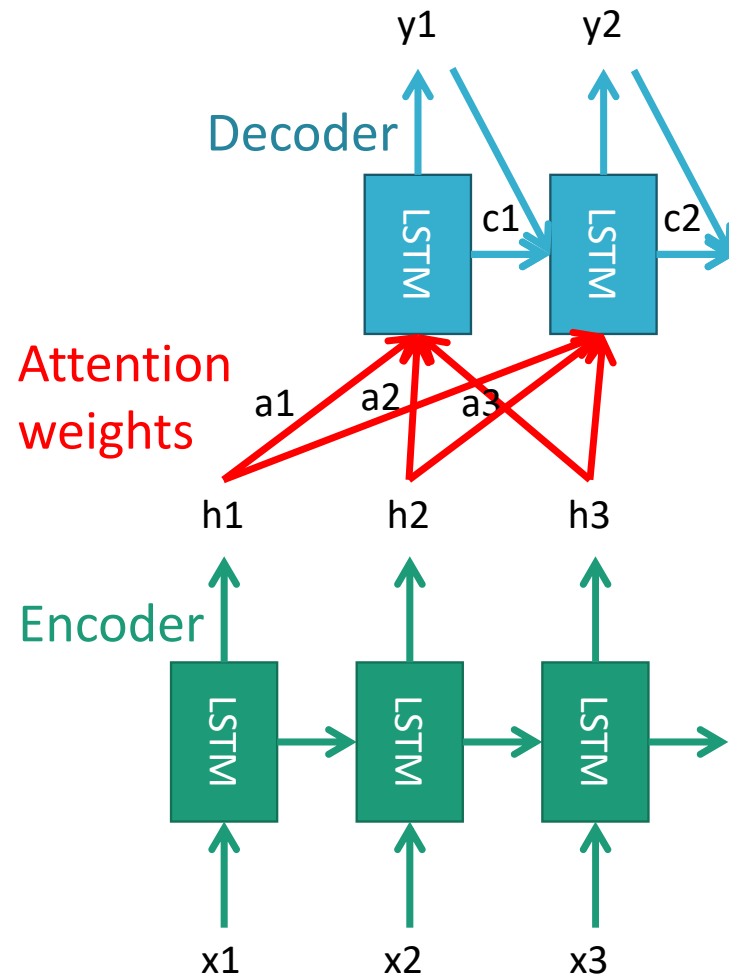


Encoder-Decoder (Seq2Seq) Model with Attention Mechanism



Encoder-Decoder (Seq2Seq) Model with Attention Mechanism

- Encode input sequence via LSTM
- Decode with another LSTM
 - Asynchronous with input
- Weights on encoded LSTM output ($\sum a_i h_i$)
 - Weight a_i are computed based on decoder state and output
- End-to-end joint training



Review of Neural Seq2Seq Model

- **Needs a huge amount of training data**
 - Ubuntu [Lowe et al 15] software support
 - OpenSubtitles [Lison et al 2016] Movie Subtitles
 - Reddit [Yang et al 2018] text on bulletin boards
- Consider dialogue context (by encoding)
- Do NOT explicitly conduct SLU to infer intent and slot values
- NOT straightforward to integrate with external DB & KB
- **Converge to generic responses with little diversity**
 - Frequent and acceptable in many cases
“I see”, “really?”, “how about you?”

Ground-truth in Dialogue(?)

- Many choices in response given a user input
- Trade-off
 - Safe (boring)
 - Elaborate (challenging)
- Simple retrieval or machine learning from human conversations is NOT sufficient



- Filter golden samples
- Need a model of emotions, desire and characters

I like cheese.

(a) That's good. (Reaction)

(b) I like blue cheese. (Statement)

(c) What kind of cheese? (Question)

(Summary) Review of Dialogue Models

- SLU + Dialog Flow
 - Suitable for goal-oriented (complex) dialogue
 - Provide appropriate interactions for limited scenarios
- Example-Based Dialogue and QA
 - Suitable for simple tasks and conversations
 - One response per one query
- Chatting based on Neural Seq2Seq Model
 - Very shallow but wide coverage
 - Useful for ice-breaking, relaxing and keeping engagement

Hybrid
Combination

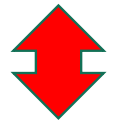
4. What kind of non-verbal and other modalities are useful for human-robot interaction?

Non-verbal Issues in Dialogue

Protocol of Spoken Dialogue

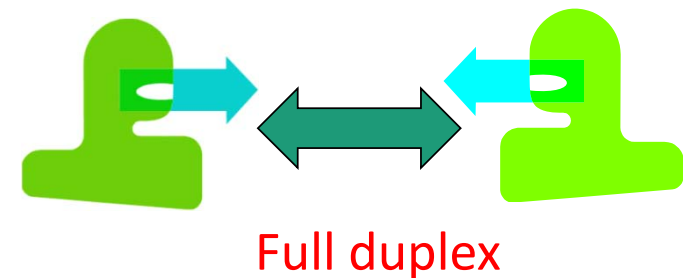
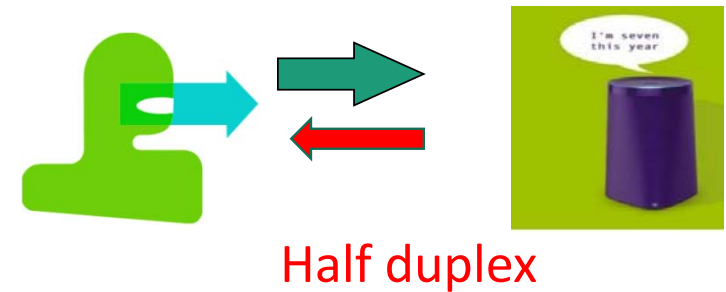
- **Human-Machine Interface**

- Command & Control
- Database/Information Retrieval
- One command/query → One response
- No user utterance → No response



- **Human-Human Dialogue**

- Task goals are not definite
- Many sentences per one turn
- Backchannels








Non-lexical utterances

--“Voice” beyond “Speech”--

- Continuer Backchannels: “right”, “はい”
 - listening, understanding, agreeing to the speaker
- Assessment Backchannels: “wow”, “へー”
 - Surprise, interest and empathy
- Fillers: “well”, “えーと”
 - Attention, politeness
- Laughter
 - Funny, socializing, self-pity

Comparison of Dialogue Interfaces

	Smart Speaker	Virtual Agent	Pet Robot	Child Robot	Adult Android
					
Continuer BC “right”					
Assessment BC “wow”					
Filler “well”					
laughter					
???					

Continuer BC “right”
 Assessment BC “wow”
 Filler “well”
 laughter
 ???

Role of Backchannels

- Feedback for smooth communication
 - Indicate that the listener is listening, understanding, agreeing to the speaker
“right”, “はい”, “うん”
- Express listener’s reactions
 - Surprise, interest and empathy
“wow”, “あー”, “へー”
- Produce a sense of rhythm and feelings of synchrony, contingency and rapport

Factors in Backchannel Generation

- Timing (**when**)
 - Usually at the end of speaker's utterances
 - Should predict before end-point detection
- Lexical form (**what**)
 - Machine learning using prosodic and linguistic features
- Prosody (**how**)
 - Adjust according to preceding user utterance

(cf.) Many systems use same recorded pattern,
giving monotonous impression to users



Generating Backchannels

- Conventional: fixed patterns
- Random 4 kinds
- Machine learning: context-dependent (proposed)



Subjective Evaluation of Backchannels

[Kawahara:INTERSPEECH16]



	random	proposed	counselor
Are backchannels natural ?	-0.42	1.04	0.79
Are backchannels in good tempo ?	0.25	1.29	1.00
Did the system understand well?	-0.13	1.17	0.79
Did the system show empathy ?	0.13	1.04	0.46
Would like to talk to this system?	-0.33	0.96	0.29

- obtained higher rating than random generation
- even comparable to the counselor's choice, though the scores are not sufficiently high
 - Same voice files are used for each backchannel form
 - **Need to change the prosody as well**

Role of Fillers

- Signals thinking & hesitation
- Improves comprehension
 - Provide time for comprehension
- Attracts attention & improves politeness
 - Mitigate abrupt speaking
- Smooth turn-taking
 - Hold the current turn, or Take a turn

Factors in Filler Generation

- Timing (**when**)
 - Usually at the beginning of speaker's utterances
- Lexical form (**what**)
 - Machine learning using prosodic and linguistic features and also dialogue acts
- Prosody (**how**)
 - ???

(cf.) frequent generation of fillers (at every pause) is annoying



Generating Fillers

- No filler



- Filler before moving to next question

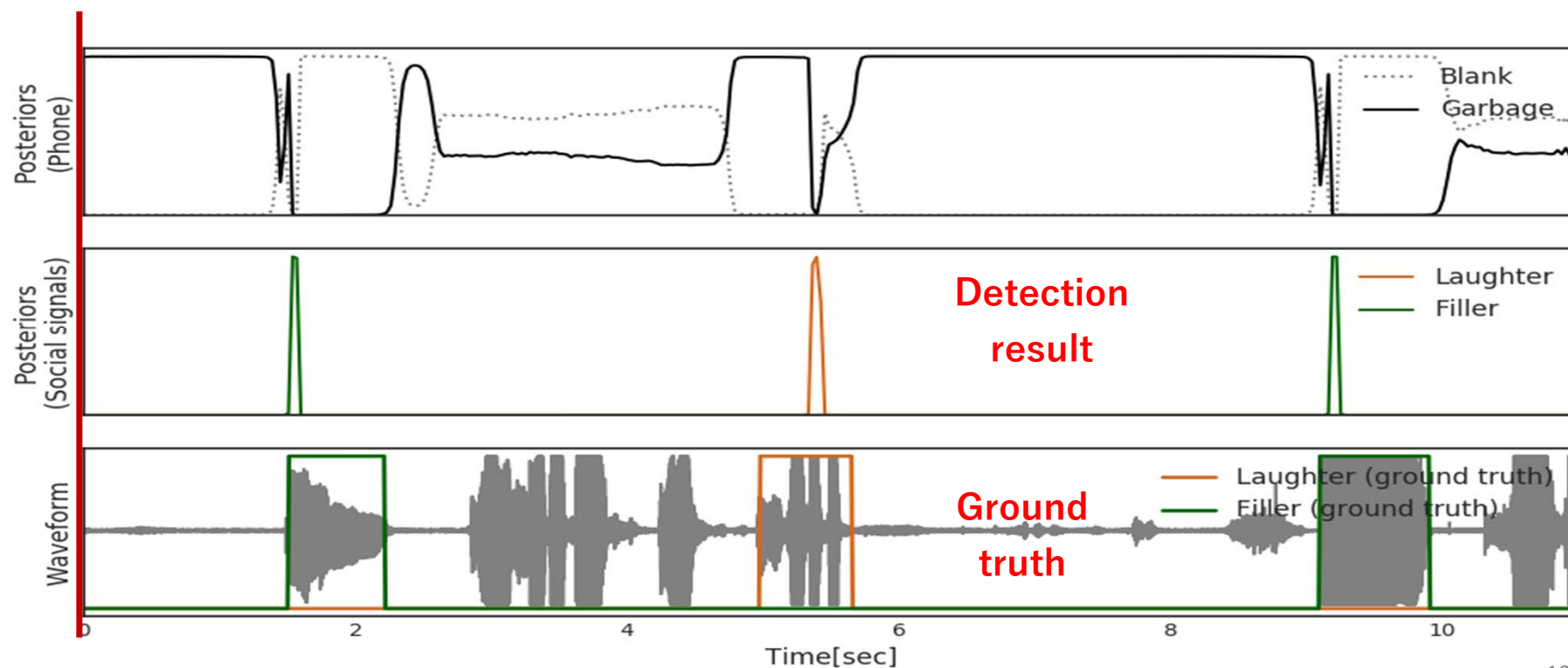


Generating Laughter

- People laugh not necessarily because funny
- But to socialize and relax
 - Should laugh together (**shared-laughter**)
- Sometimes for masochistic
 - Should not respond to negative laughter



Detection of Laughter, Backchannels & Fillers

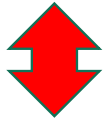


Turn-taking

Protocol of Spoken Dialogue

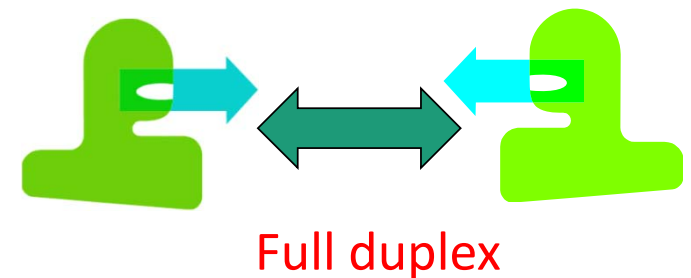
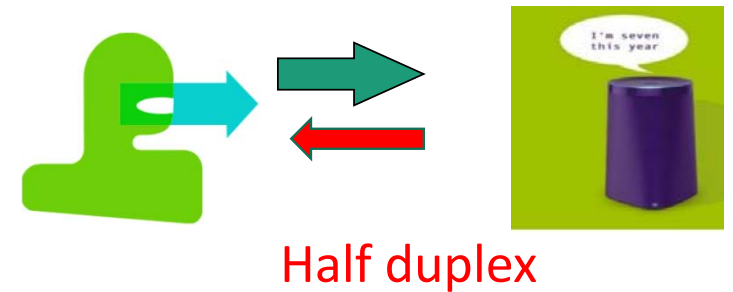
- **Human-Machine Interface**

- One command/query → One response
- No user utterance → No response

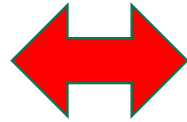


- **Human-Human Dialogue**

- Many sentences per one turn
- Backchannels



Tool

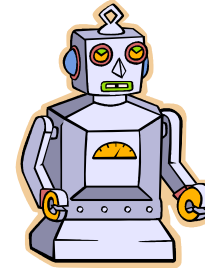


Companion, Partner

- Smartphone Assistants



- Communicative Robots



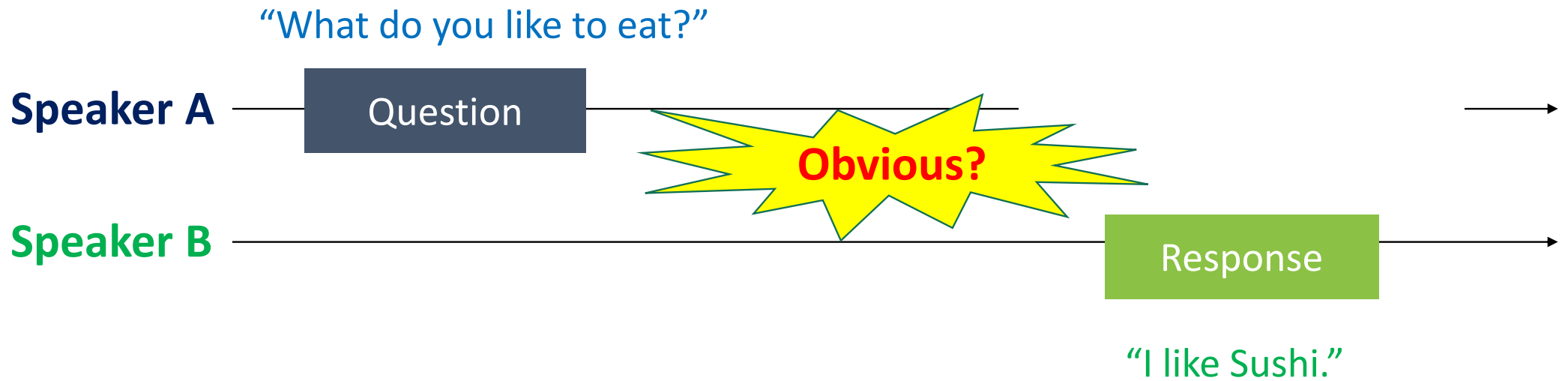
- Smart Speakers



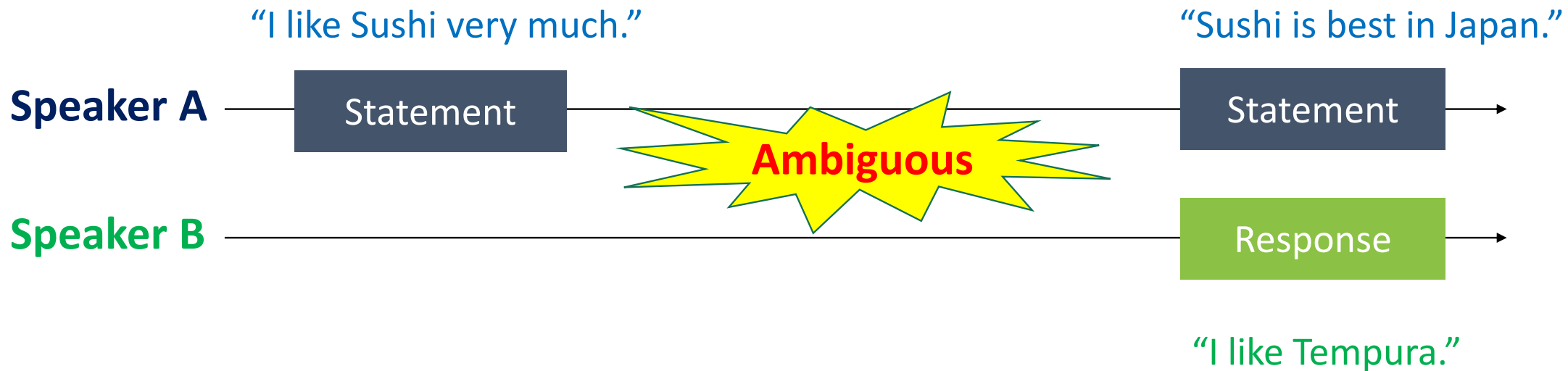
Flexible Turn-taking

- Natural turn-taking ← push-to-talk, magic words
- Avoid speech collision (of system utterance in user utterance) → required
 - Latency of robot's response
- Allow barge-in (user utterance while system speaking)? → challenging
 - ASR and SLU errors
- Machine learning using human conversation is not easy
 - Behavior is different between human-human and human-robot
 - Turn-taking is arbitrary, no ground-truth

Turn-switch after Question



Turn-keep/switch after Statement?



Turn-keep/switch after Response?

“I like Sushi.”

Speaker A

Response

**Very
Ambiguous**

Speaker B

“What do you like?”

- “I like Tempura, too.”
- “Sushi is best in Japan.”
- “What do you like?”

Response	→
Statement	→
Question	→
Question	→
Response	→
Statement	→

- “What kind of Sushi?”
- “I like Sushi, too.”
- “Sushi is best in Japan.”

Turn-taking Prediction Model


- System needs to determine if the user keeps talking or the system can (or should) take a turn
- Turn-taking cue (features) → can be different between human and robot
 - Prosody...pause, pitch, power
 - Eye-gaze
- Machine learning model → ground truth? Turn-taking is arbitrary
 - Logistic regression...decision at each end of utterance
 - LSTM...frame-wise prediction, but decision at each end of utterance

Proactive Turn-taking System

- **Fuzzy** decision ← Binary decision

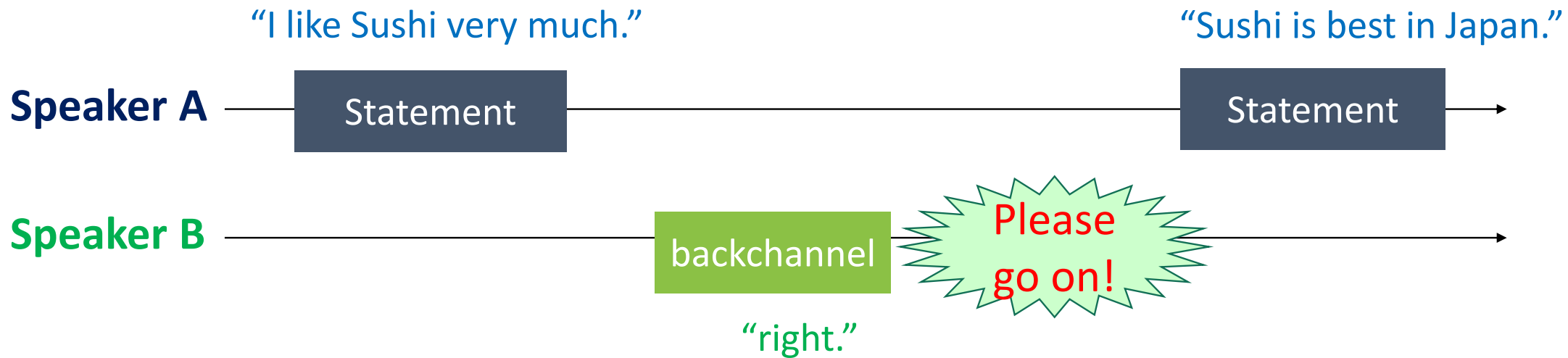


- Use **fillers and backchannels** when ambiguous

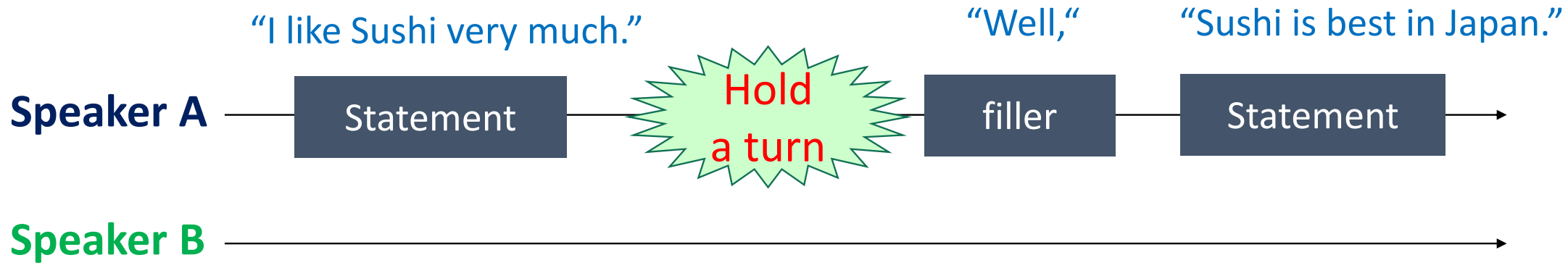


User status	System action
User definitely holds a turn	nothing
User maybe holds a turn	continuer backchannel
User maybe yields a turn	filler to take a turn
User definitely yields a turn	response

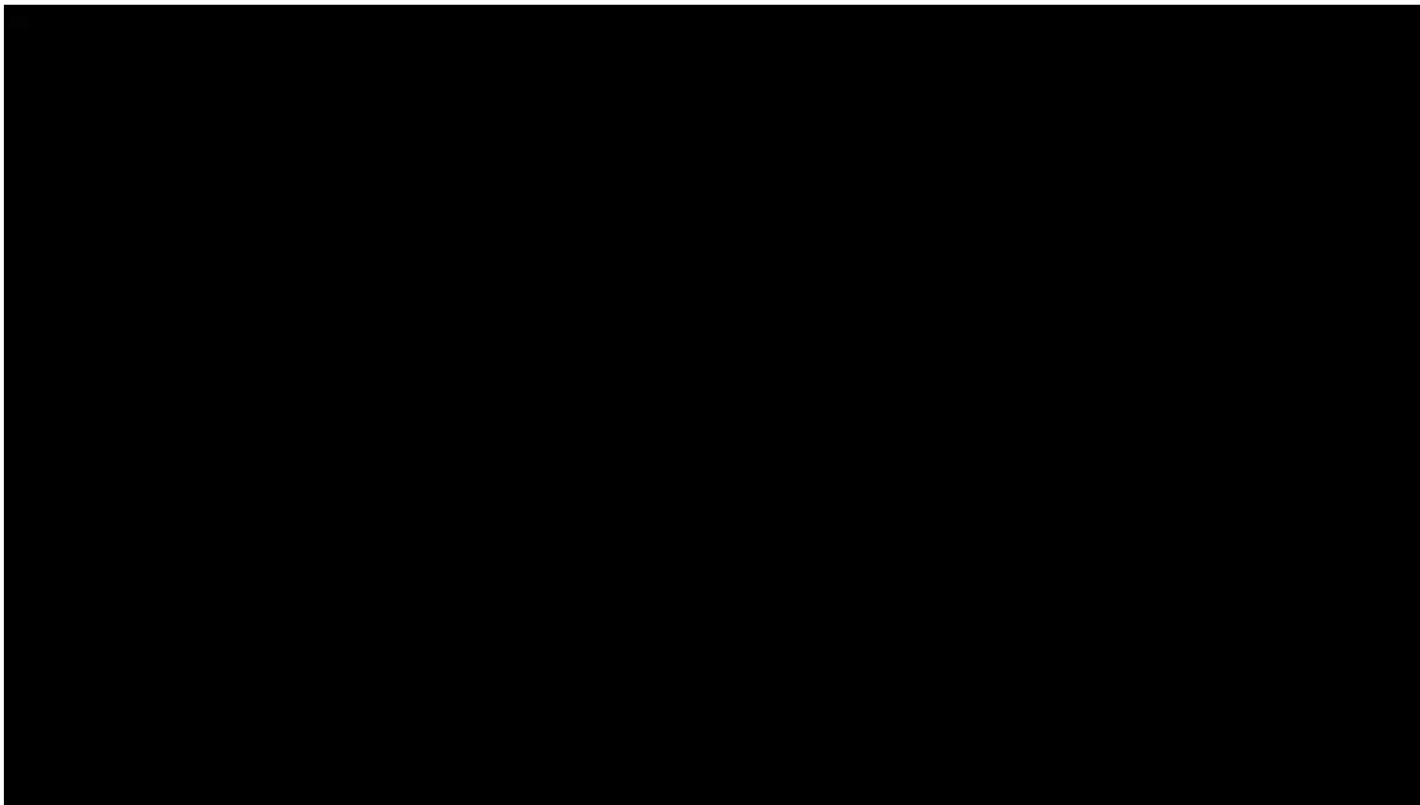
Turn-keep/switch after Statement?



Turn-keep/switch after Statement?



Use Filler (+Gaze Aversion)
for Proactive Turn-taking



References

1. Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, Selma Sabanovi.
Human-Robot Interaction — An Introduction.
<https://www.human-robot-interaction.org/>
2. T. Kawahara.
Spoken dialogue system for a human-like conversational robot ERICA.
In Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS), (keynote speech), 2018.
3. K. Jokinen. Dialogue Models for Socially Intelligent Robots.
In Proc. ICSR, 2018.

Agenda (Research Questions)

0. Why social robots are not prevalent in society?
 1. What kind of tasks are social robots expected to perform?
 2. What kind of social robots are suitable for the tasks?
 3. Why spoken dialogue is not working with robots?
 1. ASR and TTS
 2. SLU+DM (end-to-end?)
-
- 4. What kind of non-verbal and other modalities are useful?**
 1. Backchannel, turn-taking
 2. Eye-gaze, gestures
 5. What kind of system architectures are suitable?
 6. What kind of ethical issues must be considered?

} optional

break

Kawahara

Jokinen

Gaze and Attention

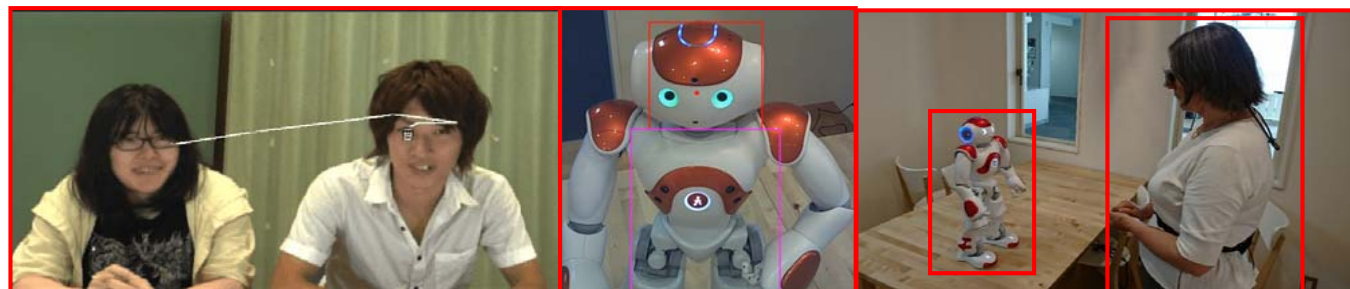
Attention: a process to select the information that enters working memory to be processed (Knudsen, 2007)

- Visual processing systems (Unema et al. 2005, Findlay & Gilchrist 2003)
 - *Global processing*: long saccades and short fixations early in the viewing to get a gist of a scene and the main regions of interest
 - *Local processing*: short saccades and long fixations later in the viewing to examine the focus of attention in more details
- Eye-gaze in everyday activities (Land, 2006)
 - Proactive and preparatory information
 - Gaze and gesture coordination: look ahead before manipulating them

Visual Attention in Interaction

Primary gaze function is to get information. Also, gaze indicates one's attention, engagement, and presence. It coordinates and organises interaction.

- Gaze in human-human and human-agent interactive situations (Kendon 1967, Argyle & Cook 1976, Nakano et al. 2007, Edlund et al. 2000, Mutlu et al. 2009, Jokinen et al. 2009, 2010, Andrist et al. 2014)
 - Focus of shared attention
 - Coordination of turn-taking (mutual gaze)
 - "Pointing device"
 - Conversational feedback
 - Building trust and rapport



Measurements

- **Frequency** of fixations on AOI (area of interest)
- **Duration** of individual fixations on the AOI
- **Accumulated fixations** time on the AOI
- Average **Gazing Ratio** (GR):

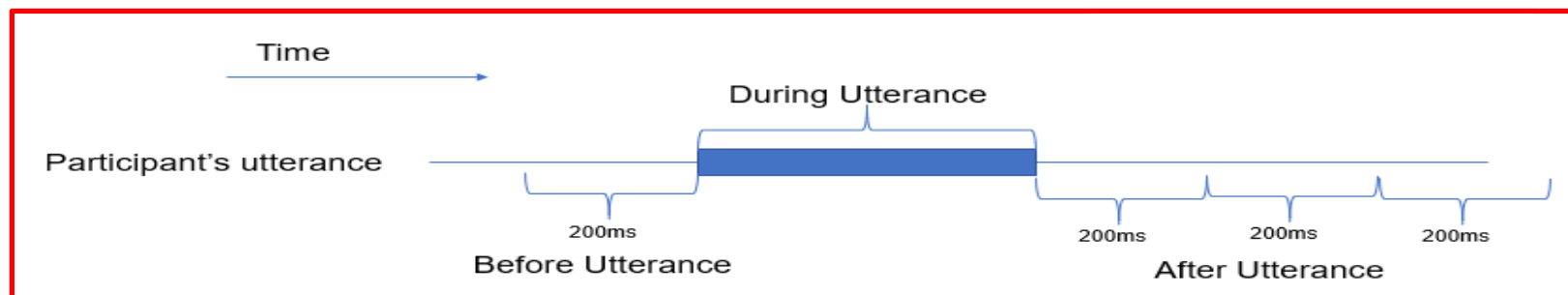
Fixations (stops):

fixation length 80-120 ms,
ave 3 fixations/sec

Saccades (jumps):

fast eye-movemens, during
which we're practically blind

$$\text{Gazing Ratio} = \frac{1}{N} \sum_{i=1}^N \left(\frac{DG_{(i)}}{\text{duration of } i_{th} \text{ window}} \right)$$



Eye-gaze in predicting turn-taking

- Spoken dialogues studies
 - Acoustic correlates at suitable turn-taking places
 - Pause length: usually about 0.2 second pause, but if longer than 0.5 seconds, the current speaker is likely to continue talking
- Eye-gaze in interactions
 - Speakers look at face area 99% of time
 - Mutual gaze to agree on the change of speaker
 - Gaze activity changes more at start of utterance than in middle or end of utterance
 - Gaze wanders off quickly after start of utterance, but fixates on partner a long time before end of utterance
- Eye-gaze helps to distinguish who will speak after pauses
 - Gaze aversion and longer pause signals hesitation
=> the current speaker wants to hold the turn
 - Gaze at the partner and longer pause signals end of utterance
=> the speaker wants to give turn to the partner

Table 9 Mean and standard deviation of gaze offset related to speech. Plus refers to the time from the start of the utterance, minus refers to the time before the utterance ends.

	Beginning	End
Mean	+ 0.38 s	- 1.85 s
std	1.12	1.92



Gaze behavior in three-party conversations

- Interlocutors: speaker, active partner, silent partner
- More gazing to the active partner than to the silent one
- More gaze activity to both partners when speaking than when listening or backchannelling
 - When **speaking**, gaze is **divided between partners**
 - When **listening**, gaze is **directed at the active partner**
- **Silent partner's impact** on the conversation:
 - If silent partner is **passive** (not moving), **Subject gazes** at the silent partner's face **less often but twice as long than when this is engaged** (seems to listen actively)
 - If the silent partner **passive**, Subject **gazes at the background more than engaged** actively

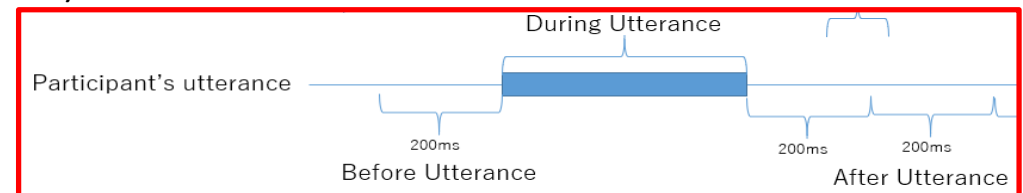


Eye-gaze and unexpected dialogue breakdowns

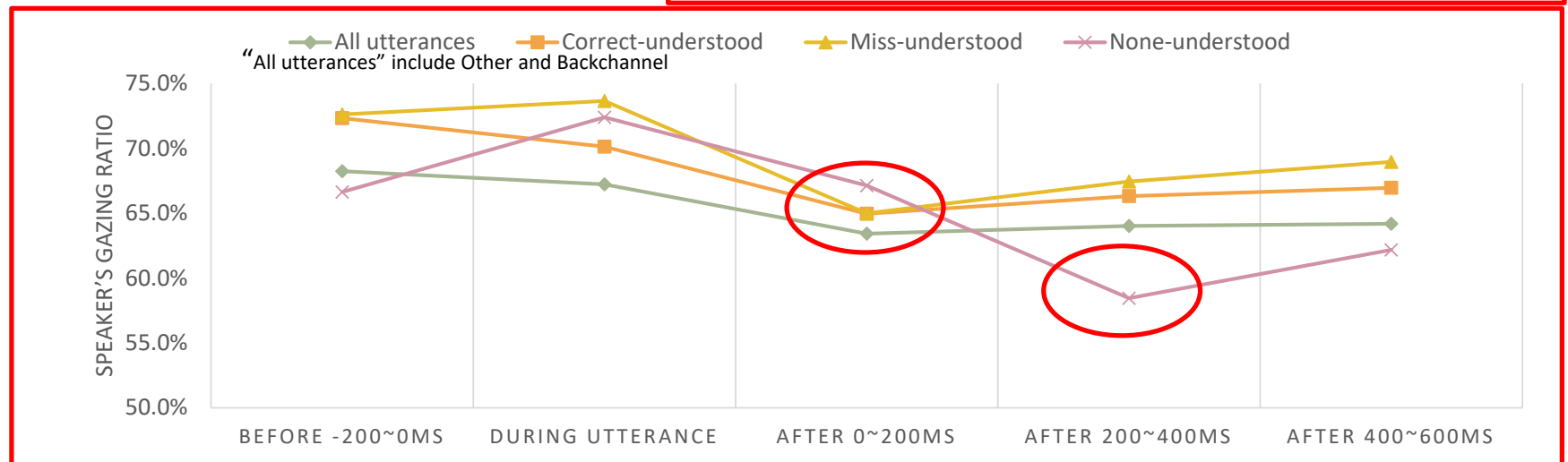
Using the AICO corpus (Jokinen 2020) and gaze ratio measurement:

- Usually participants tend to gaze away from the robot after they finish speaking (up to 200ms) and start to gaze at the robot when the robot gives feedback (after 200ms) -> Correctly understood or Misunderstood utterances
- Unexpected responses (robot gives no feedback or says something unexpected) produce a different gaze pattern: participants gaze at somewhere else than the robot (after 200ms) -> Not-understood utterances

Cognitive processing demands are reflected in the eye-gaze behaviour



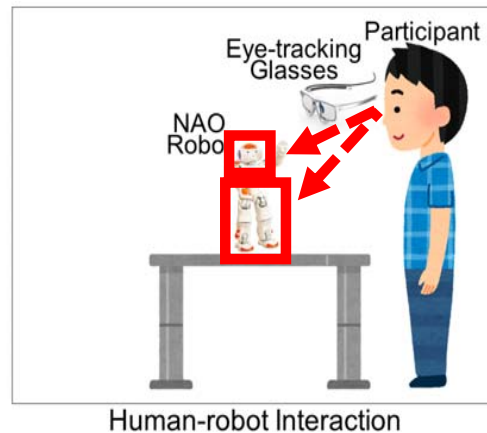
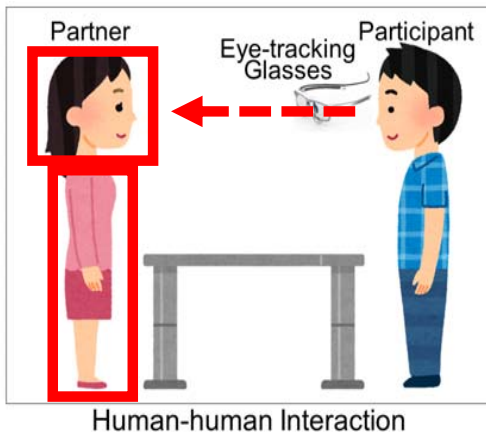
Gaze ratio:
Ave. ratio between duration of looking at the partner within a window and length of the window



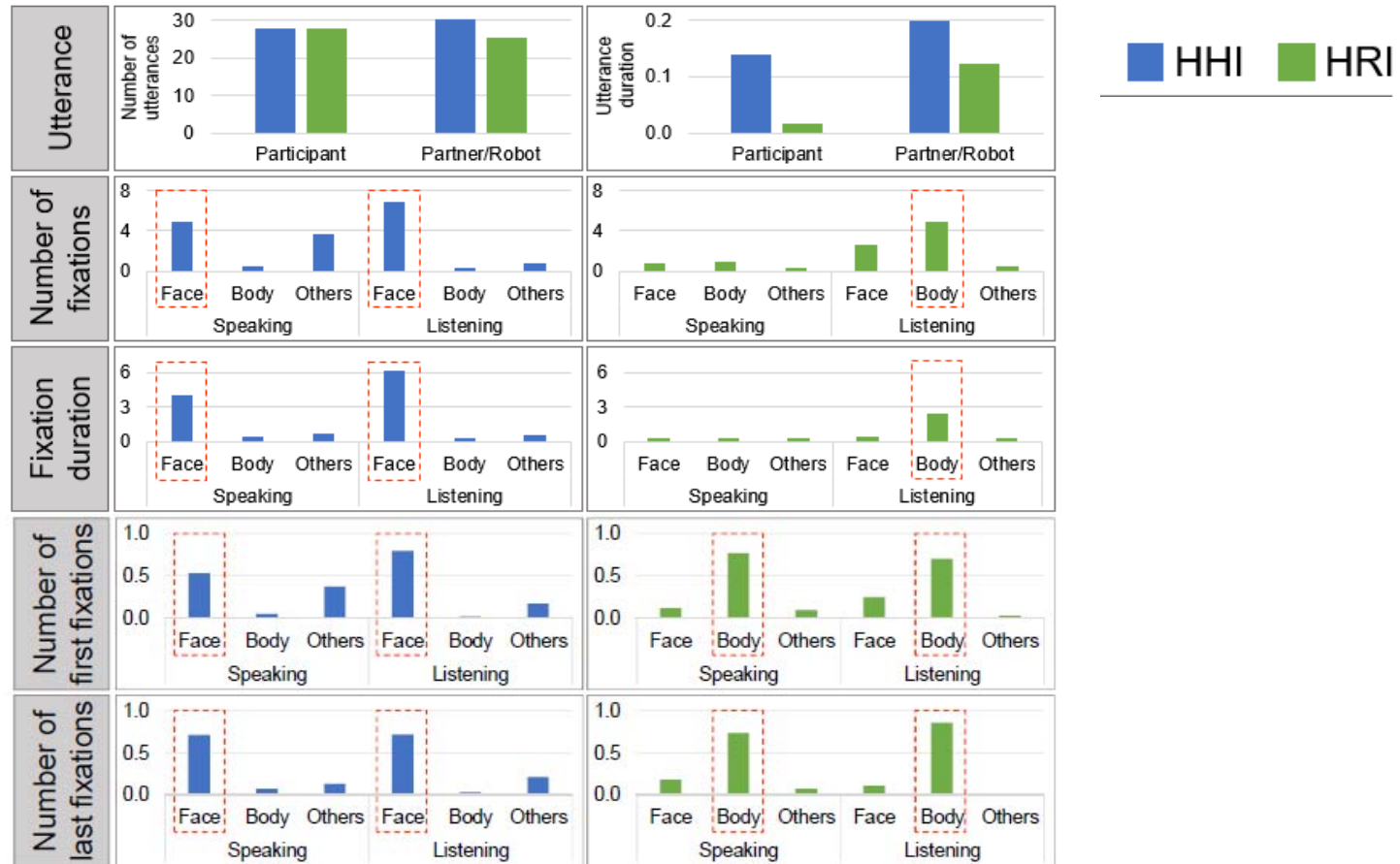
Comparison of Eye Gazes in HHI and HRI

	Human-Human	Human-Robot
Story-telling	Face	Face, Body
Instruction giving	Face	Face

- Participants tend to look at “**face**” in **HHI** and “**body**” in **HRI**.
- In **HHI**, humans monitor partner’s reactions expressed by face and eyes, which provide necessary information.
- In **HRI**, participants focus on the robot’s body rather than on its face, since the face does not provide “live” information about its internal state like its emotion or intention.
- Results concern Nao robotic face, may be different for Erica’s android face



Comparison of Eye Gazes in HHI and HRI (results)



Eye-gaze and perceived personality traits

Questionnaire (7-point Likert scale) composed on 10 questions on perceived personality, paired with Big5 personality scale:

I believe the participant to be:

- | | |
|---------------------------------|------------------------------|
| 1. Extraverted, enthusiastic | 6. Reserved, quiet |
| 2. Critical, quarrelsome | 7. Sympathetic, warm |
| 3. Dependable, self-disciplined | 8. Disorganized, careless |
| 4. Anxious, easily upset | 9. Calm, emotionally stable |
| 5. Open to new experiences | 10. Conventional, uncreative |

Big 5 Model (McCrae 2002):

- Extraversion (1 & 6)
- Agreeableness (2 & 7)
- Conscientiousness (3 & 8)
- Emotional stability (4 & 9)
- Openness (5 & 10)

	Emotion stability	Openness to experience
Number of all gaze-event in HHI (N = 10)	$\rho = .69, p < .05$	n/s
Number of gaze face event in HHI (N = 10)	$\rho = .73, p < .05$	n/s
Number of gaze body event in HHI (N = 10)	n/s	$\rho = -.72, p < .05$
Average duration of gaze body event in HRI (N = 10)	n/s	$\rho = .69, p < .05$

Positive Correlation

Positive Correlation

Negative Correlation

Positive Correlation

- Positive correlation between frequency of gaze to partner's face in HHI & perceived emotional stability of the subject -> relaxed, interest in the partner, deals well with stress
- Positive correlation between average duration of gaze to partner's body in HRI & perceived openness to experience of the subject -> curiosity towards the robot
- Other factors besides personality: social politeness, interest in robot gesturing, type of robot (Nao-robot)

Gesturing and intercultural aspects in HHI and HRI

- Hand, head and body gestures in
 - human-human interaction vs human-robot interaction
 - Japanese vs English interaction
- AICO Corpus (Jokinen, 2020)
 - 60 conversations (30 participants × 2 sessions)
 - 20 Japanese + 10 English speaking participants
 - 30 human-human (HH) + 30 human-robot (HR) interactions
 - Annotated with gesture form and function based on the MUMIN annotation scheme (Allwood et al. 2007)
- **Hand gestures** are 6 times more frequent in English than Japanese HHI, with more varied form and function, but the difference is not big in HRI
- **Head gestures** reduced significantly in HRI, but the difference between Japanese and English speakers is not big
- **Body gestures** increased in HRI, and English speakers produced more body gestures than Japanese
- Gesture detection and activity analysis are important for HRI, but it is important to elicit human gesturing in HRI first

Analysis of Body Behaviours in Human-Human and Human-Robot Interactions

TAIGA MORI *†, KRISTINA JOKINEN†, YASU HARU DEN‡

* Graduate School of Humanities and Studies on Public Affairs, Chiba University, Japan

† AI Research Center, AIST Tokyo Waterfront, Japan

‡ Graduate School of Humanities, Chiba University, Japan

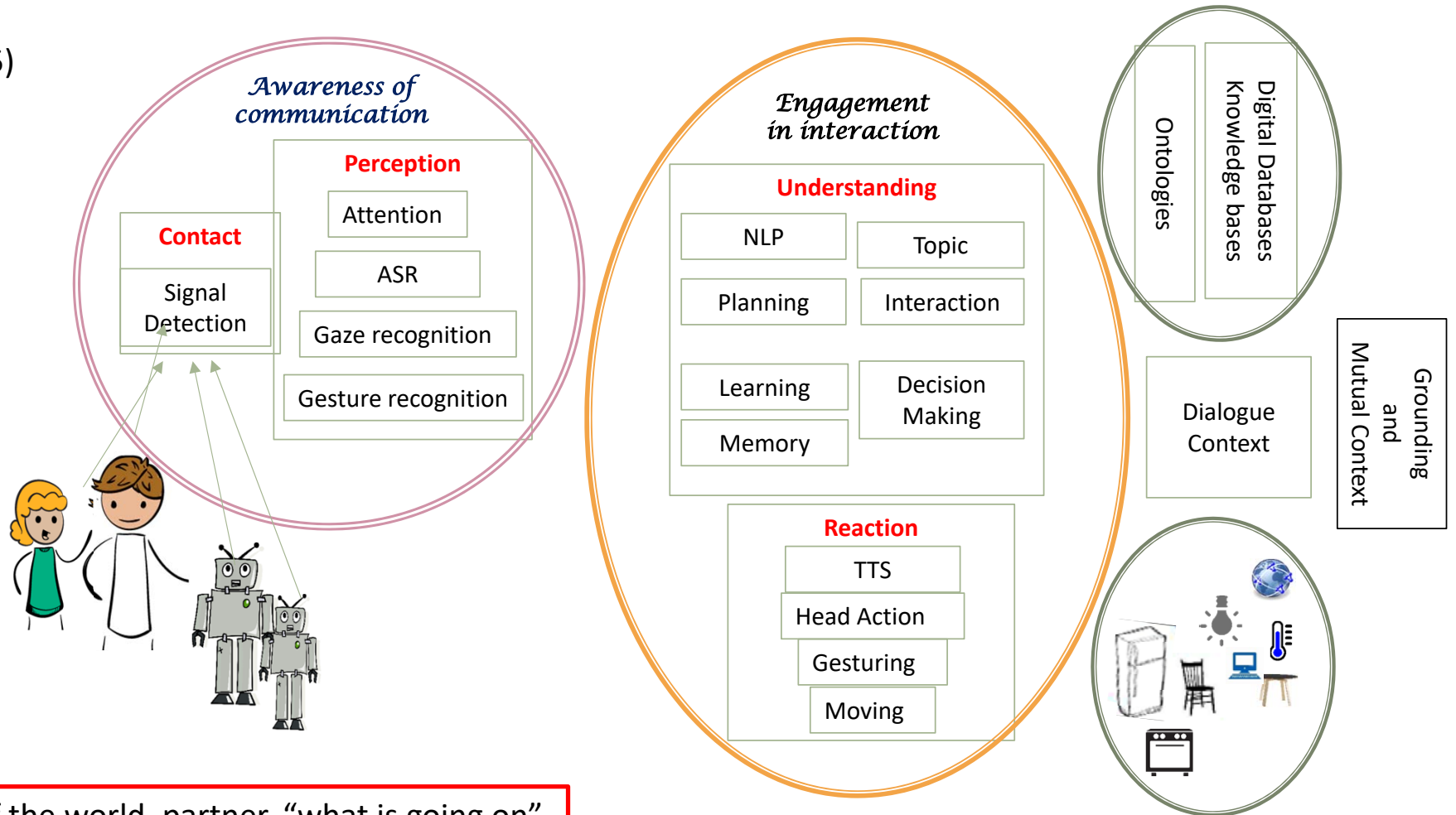


LREC ONION Workshop on peOple in
laNguage, vlsiOn and the miNd.

<https://onion2020.github.io/presentations/>

Situational Awareness

Endsley (1995)



- Knowledge of the world, partner, “what is going on”
- States of knowledge (dialogue states)
- Perception, feedback, and action



Ability to communicate smoothly in a given situation

WikiTalk and WikiListen

Towards listening robots that can join in conversations with topically relevant contributions

WikiTalk: Wikipedia-based talking

- Robots can talk fluently about thousands of topics using Wikipedia information
- Robots make smooth shifts to related topics predicted from Wikipedia links
- Topics are disambiguated using the continuously changing dialogue context

WikiListen: Wikipedia-based listening

- Robots will listen to multiparty human conversations and track changing topics
- Wikification of speech to link mentioned entities and events to Wikipedia articles
- Later, robots will learn to join in with topically relevant dialogue contributions



NAO and WikiTalk

ERICA and WikiTalk



Challenges

- Progress on multiparty speech recognition
- Progress on robust wikification of speech
- Ethical, legal & social issues of listening robots (saving/clearing short/long-term memories)

D. Lala, G. Wilcock, K. Jokinen, T. Kawahara. *ERICA and WikiTalk*. IJCAI 2019.

G. Wilcock, K. Jokinen. *Multilingual WikiTalk: Wikipedia-based talking robots that switch languages*. SIGDial 2015.

Agenda (Research Questions)

0. Why social robots are not prevalent in society?
 1. What kind of tasks are social robots expected to perform?
 2. What kind of social robots are suitable for the tasks?
 3. Why spoken dialogue is not working with robots?
 1. ASR and TTS
 2. SLU+DM (end-to-end?)
-
4. What kind of non-verbal and other modalities are useful?
 1. Backchannel, turn-taking
 2. Eye-gaze
 5. What kind of system **architectures** are suitable?
 6. What kind of ethical issues must be considered?

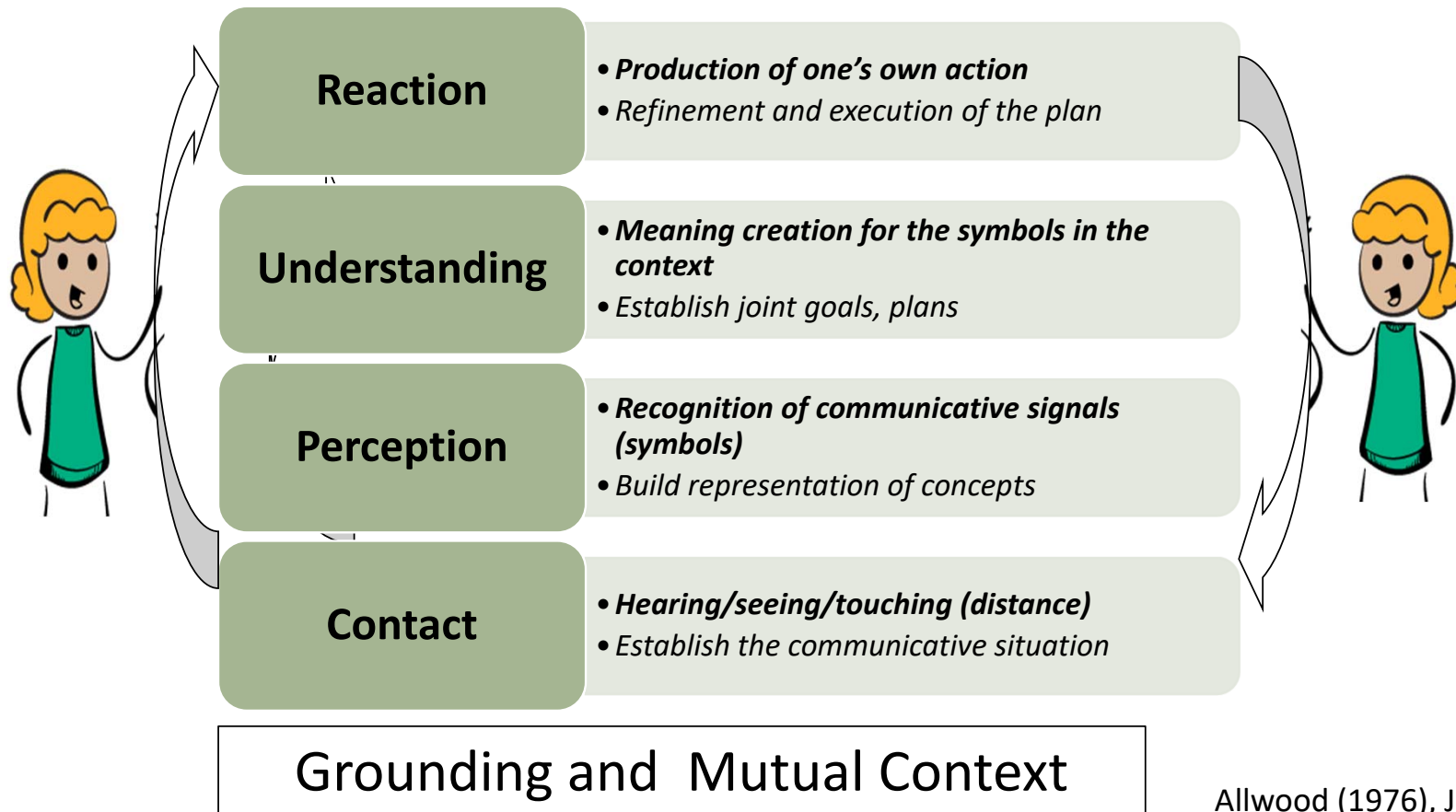
} optional

break

Kawahara

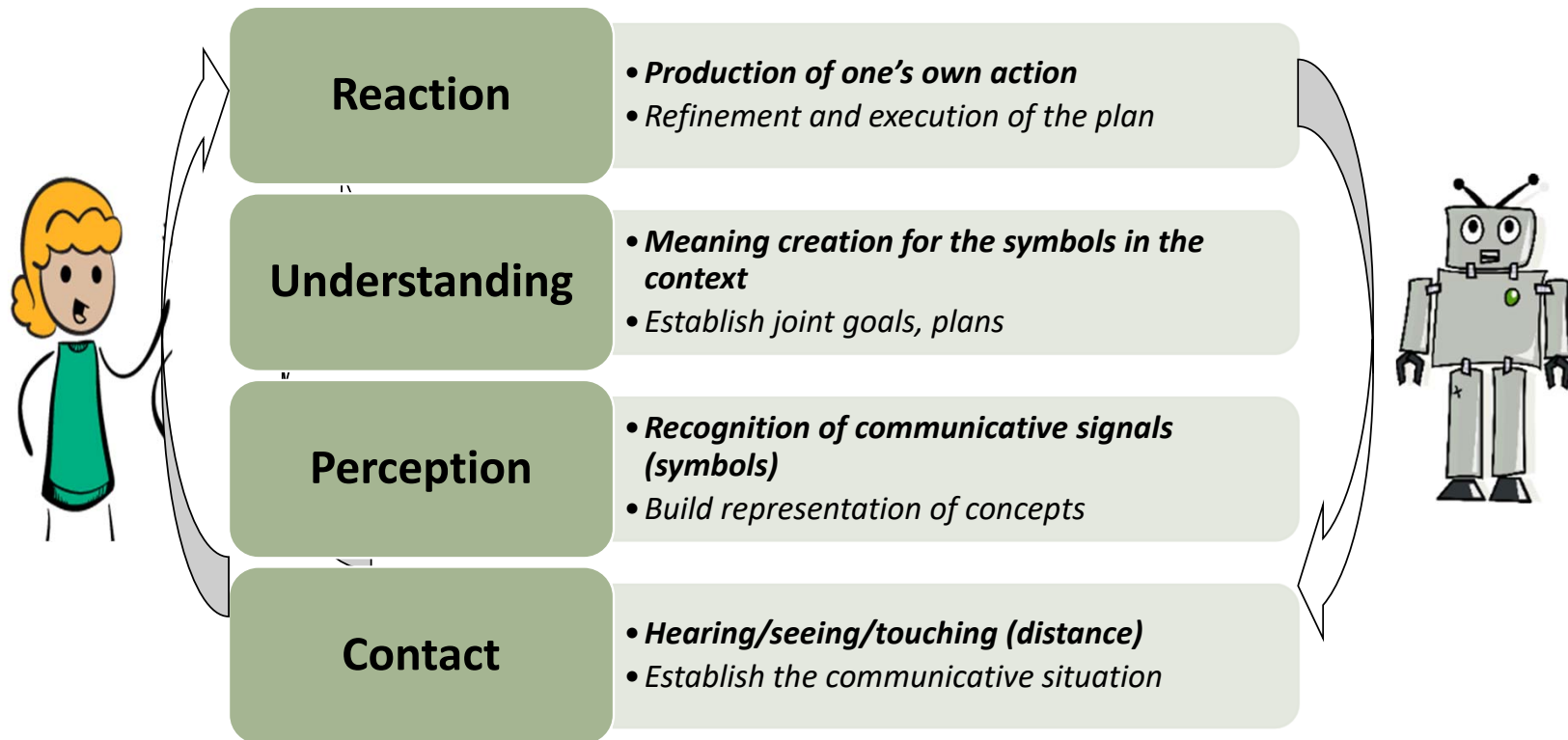
Jokinen

Co-creating Interaction

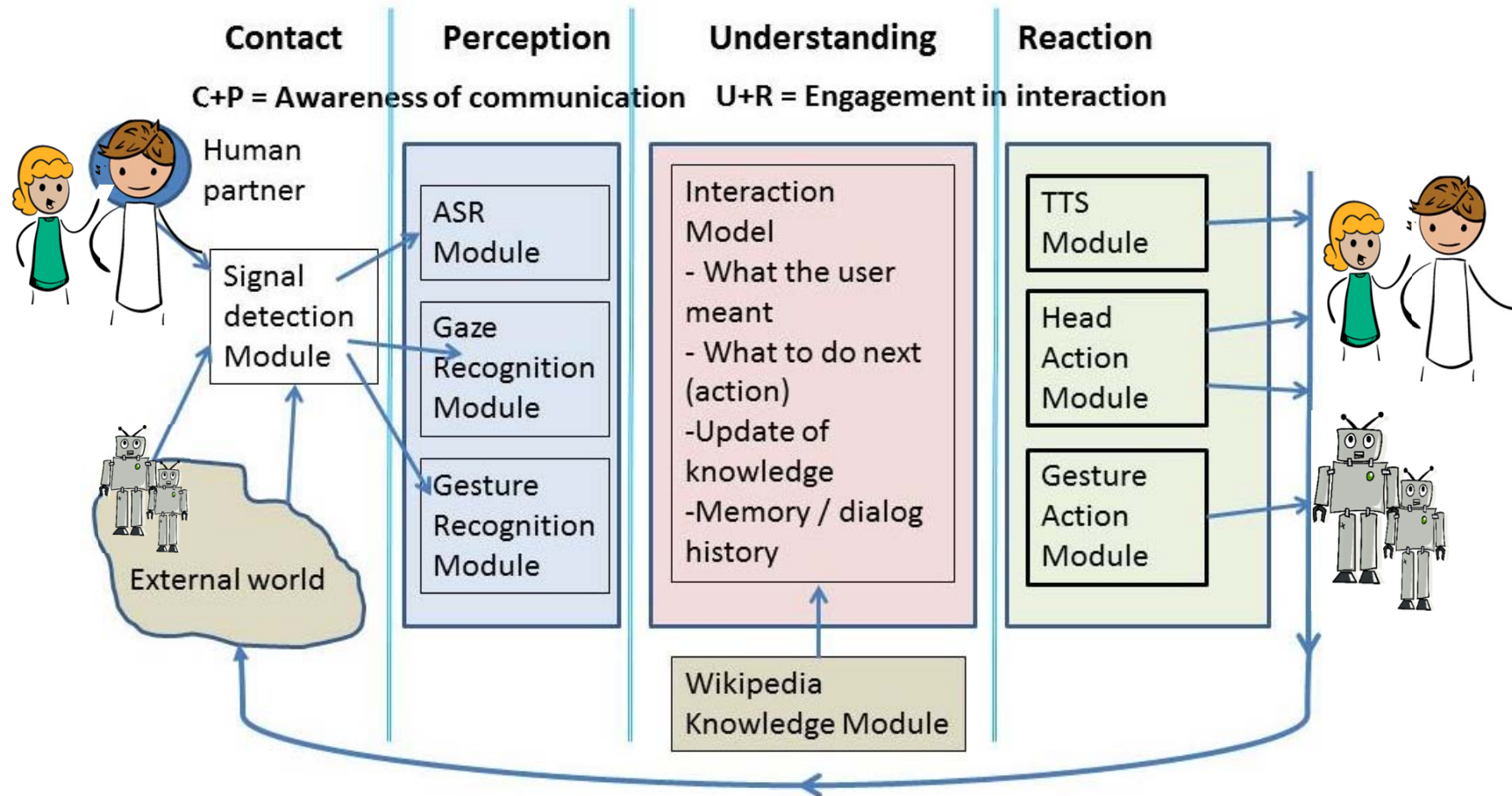


Allwood (1976), Jokinen (2009)

Co-creating Interaction



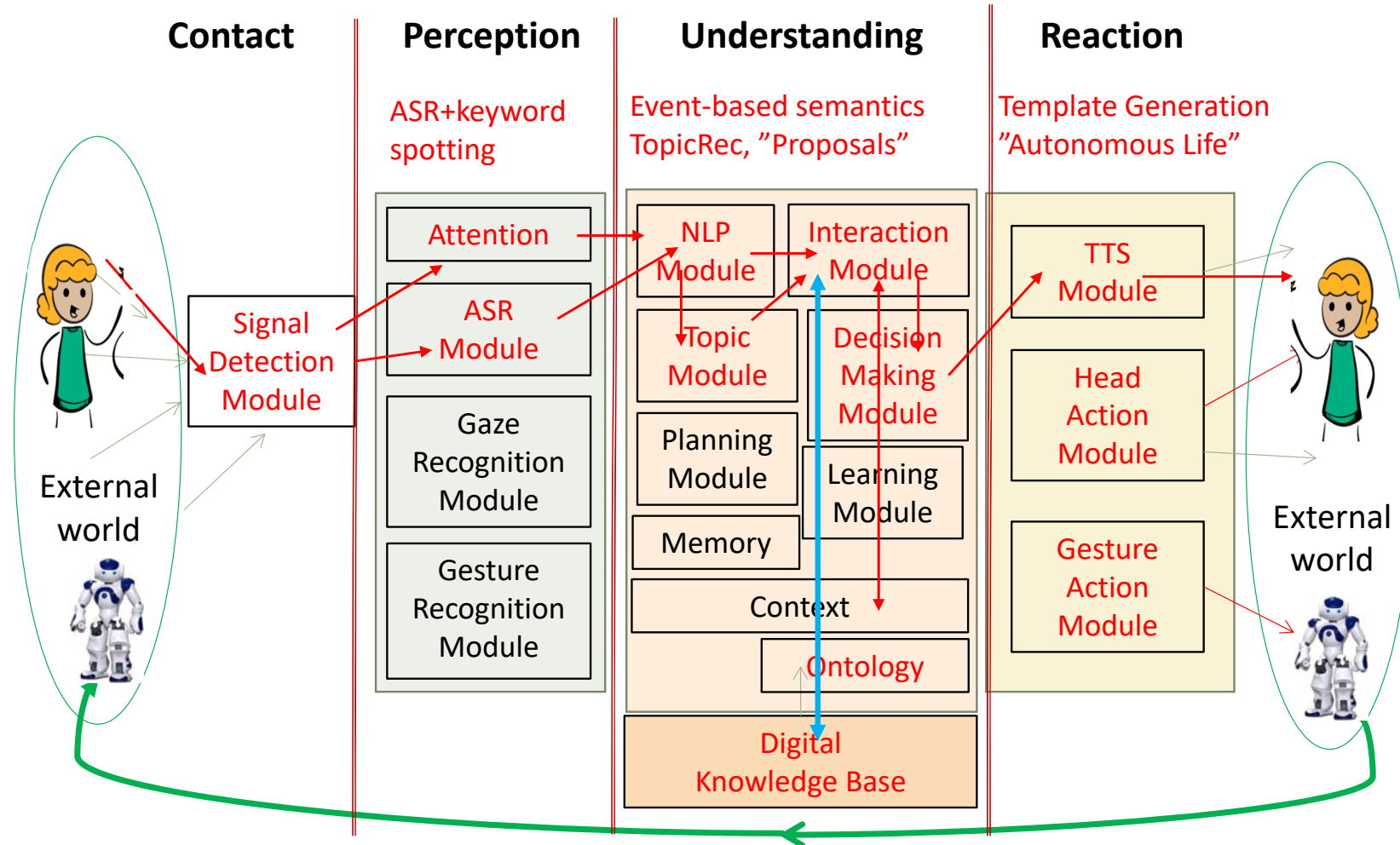
Constructive Dialogue Modelling (CDM)



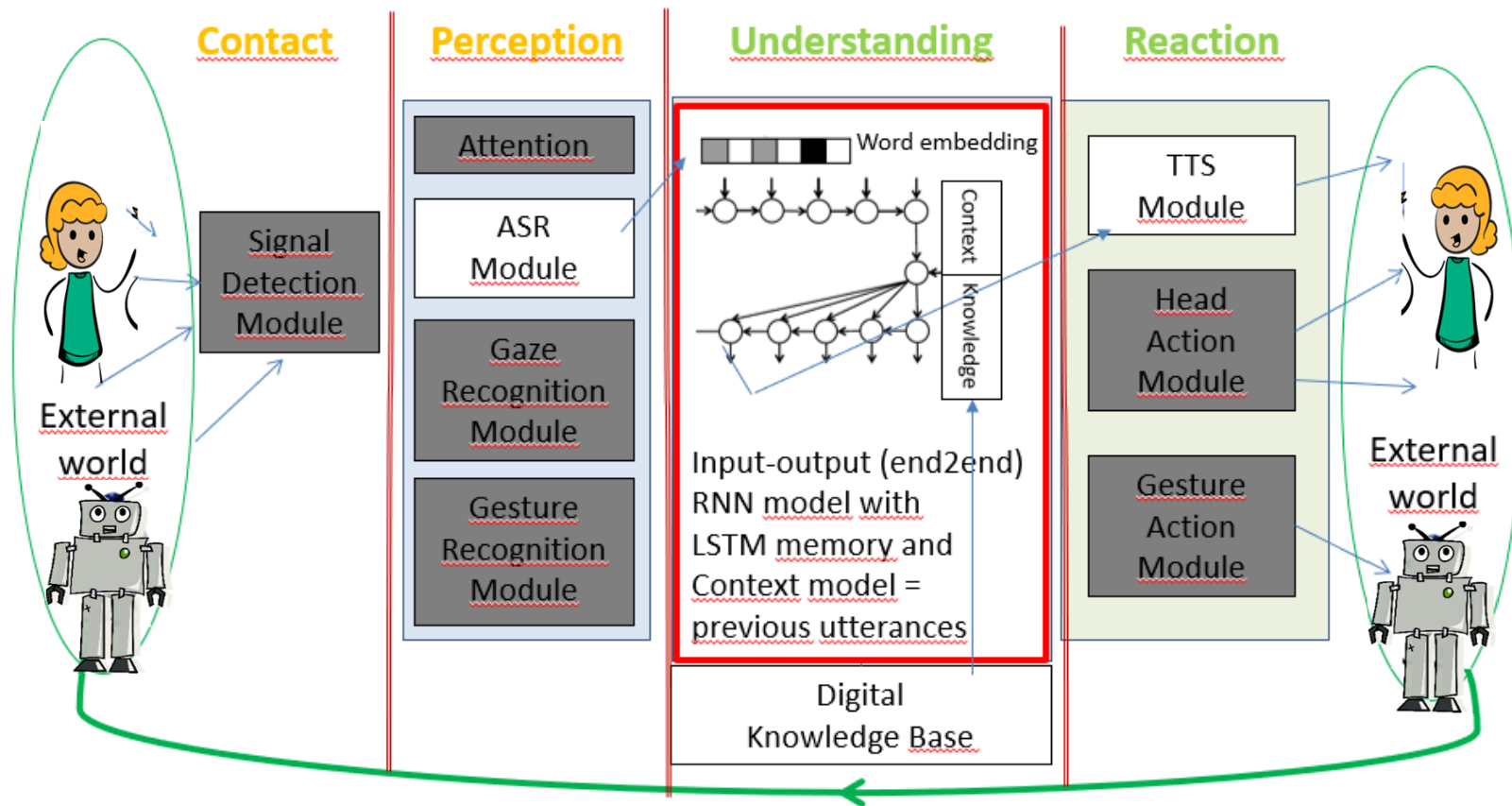
Jokinen (2009, 2019)

Sets of modules which correspond to the four communicative enablements

More modularised system architecture



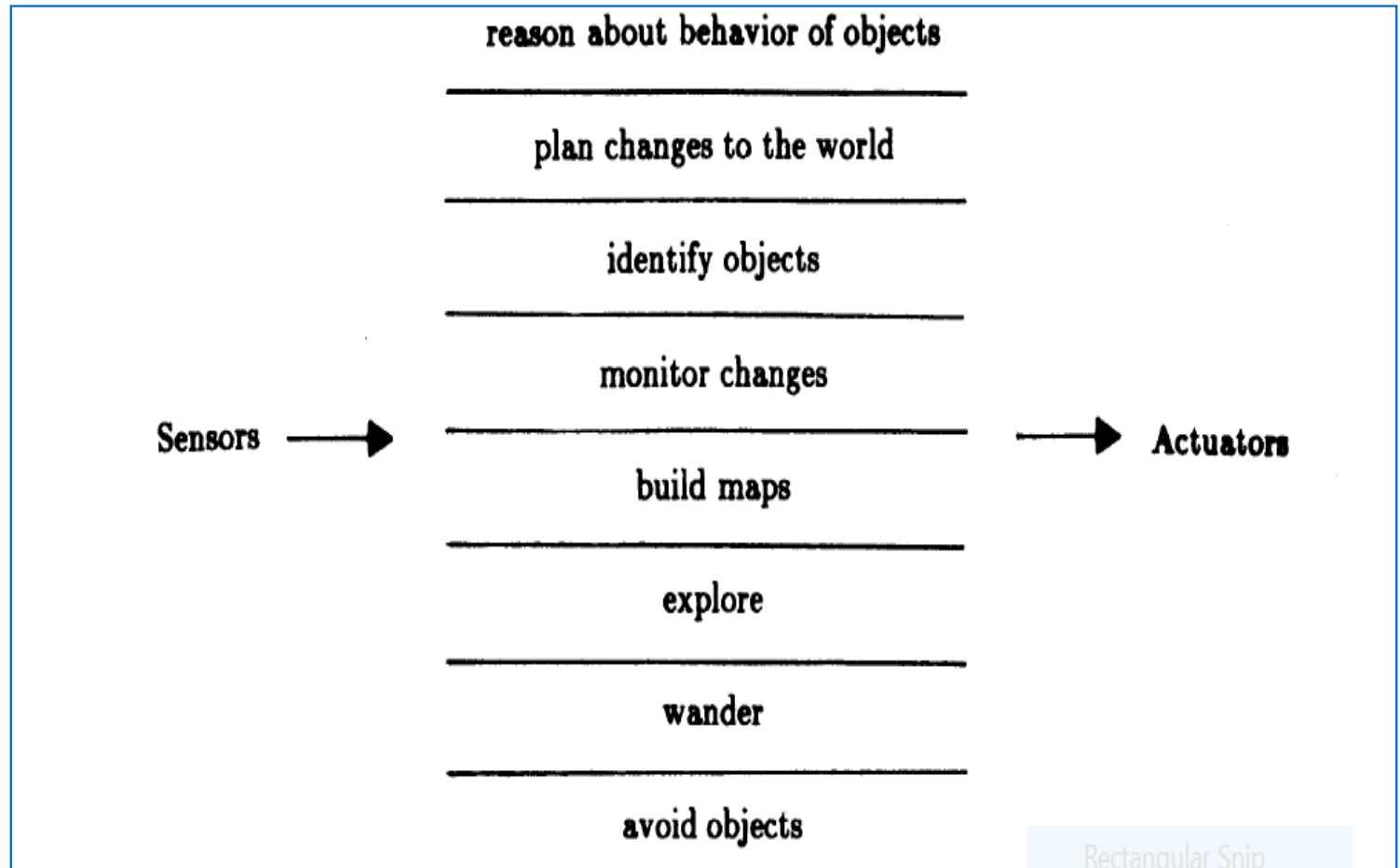
More end2end integration (deep learning)



Yoshua Bengio (IJCAI 2018)

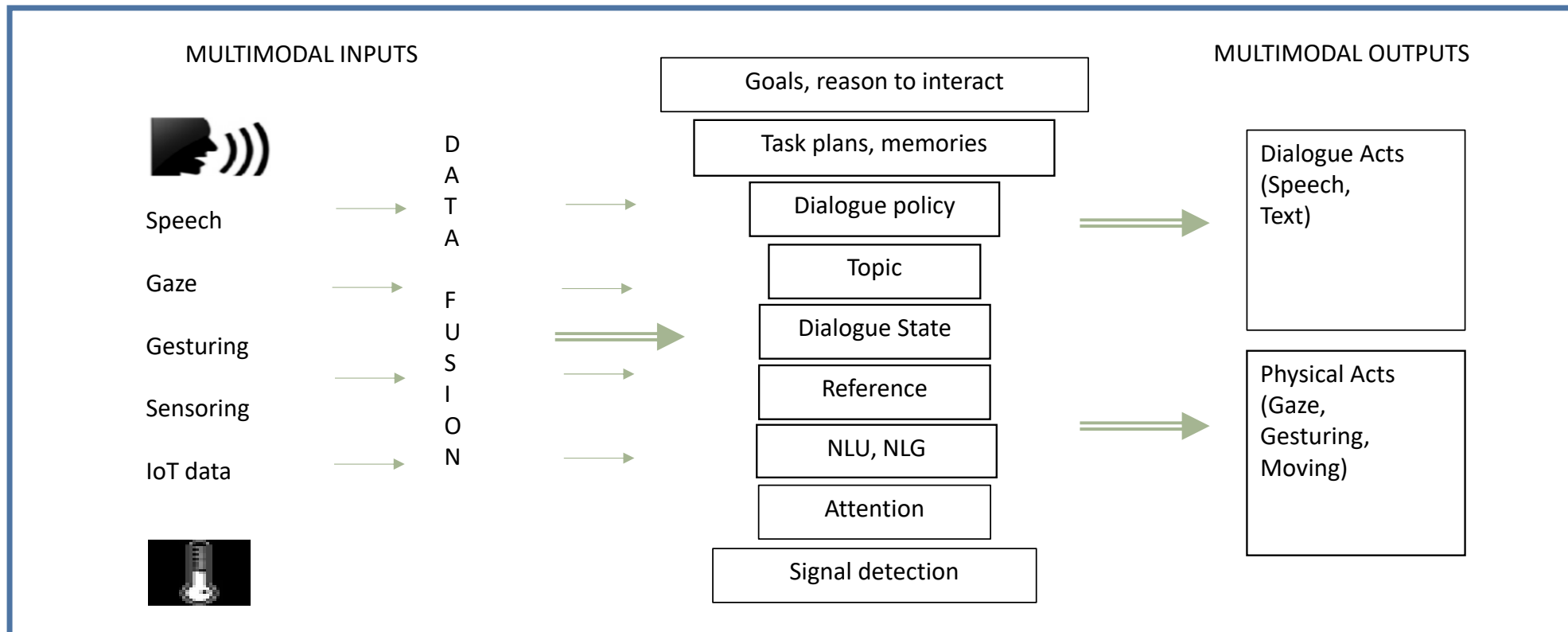
- What's Missing with Deep Learning?
- Answer: Deep Understanding
- Learning « How the world ticks »
 - So long as our machine learning models “cheat” by relying only on superficial statistical regularities, they remain vulnerable to out-of-distribution examples
 - Humans generalize better than other animals thanks to a more accurate internal model of the **underlying causal relationships**
 - To predict future situations (e.g., the effect of planned actions) far from anything seen before while involving known concepts, an essential component of reasoning, intelligence and science
 - Deep learning to expand from perception & system 1 cognition to reasoning & system 2 cognition (Kahneman (2011) *Thinking, Fast and Slow.*)

Subsumption Architecture for Autonomous Robots



Brooks (1986), Li et al. (2016)

Context-Aware Cognitive Agent Architecture



Jokinen (2020). Robotdial. IJCAI.

Subsumption architecture with a hierarchy of layers which relate to behavioural competences = communicative enablements

Agenda (Research Questions)

0. Why social robots are not prevalent in society?
 1. What kind of tasks are social robots expected to perform?
 2. What kind of social robots are suitable for the tasks?
 3. Why spoken dialogue is not working with robots?
 1. ASR and TTS
 2. SLU+DM (end-to-end?)
-
4. What kind of non-verbal and other modalities are useful?
 1. Backchannel, turn-taking
 2. Eye-gaze
 5. What kind of system architectures are suitable?
 6. What kind of **ethical** issues must be considered?

} optional

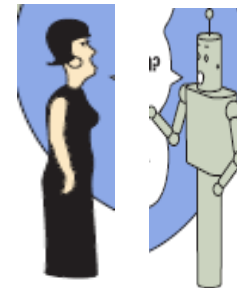
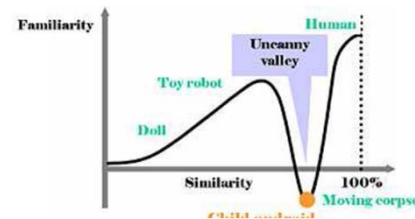
break

Kawahara

Jokinen

Human-like, but not human communication

- Reeves and Nass: anthropomorphize inanimated objects
- Mori: Uncanny valley



- Moore 2012: Bayesian explanation of Uncanny Valley: Cognitive Dissonance between conflicting categories
- Jokinen and Watanabe (2019): Robots as boundary-crossing agents for new services



Robot's Dual Characteristics

Robot as a smart computer

- Processing capability
- Accurate movement/mobility
- Information from Internet

Robot as a smart agent

- Dialogue capability
- Social awareness



Boundary Crossing Robots

- Facilitate interaction and mutual intelligibility between different perspectives
 - Novel ways to interact with social robots as cooperative agents
- Different boundaries to cross:
 - Conceptual categorization
 - Team membership
 - Expectations on understanding
- Experiments with various types of social robots
- Assigning a clear role to the robot agent in the context
- Goal
 - not only to increase user's positive experience with robot
 - but to minimize differences between the user's expectations and experience of social robot agents
- Symbiotic relation between humans and robots



Ethics, trust, and reliability



- **Data storing and privacy**
 - Encryption, secure identification, periodic deletion
- Issues in the **development of dialogue systems**
 - Unintentional biases in the data for the development of dialogue models
 - Data sharing, transfer learning, sensitive info, masking
 - Delivery of sensitive information: critical vs. less important info
 - System evaluation: new evaluation metrics, acceptance and suitability: what is conversational adequacy?
- **Legal issues**
 - Awareness and conscious agreement of recording, logging
 - Ownership of the dialogue data and its use
 - Access rights to interactive situations (family, friends, staff, passers-by, officials)
 - Responsibility for actions and information (inaccurate, unreliable, prejudiced, ...)

Ethical issues for social robot applications

- **Personalisation** and individual preferences
 - Embodiment, appearance, digital
 - True information vs. respecting the person's view-point
- **Long-term interactions**
 - Affection and emotional support
 - Mutual trust
 - Learning interaction strategies through interaction
 - Recorded changes count towards long-term monitoring
 - Tradeoff between security & safety vs. privacy
- **Social norms** and general principles
 - Appropriate, trustworthy and acceptable behaviour
 - Different cultures, different social norms
- **Standards** and standardisation
 - Maximize compatibility, interoperability, safety, quality, explainability
 - Repeatability or creative variation
- **Participatory research**
 - Involving final users

Acceptance and impact

What kind of robot systems are desirable? (function, appearance)

Where can the social robot make a difference?

Future capabilities of social robots?

Can a robot be a counselor?	
Can a robot assess a human?	
Can AI assess a human?	
Can robot have conflicting goals with humans?	
Can a robot have a personality?	
Can a robot be a soul mate of a senior person?	
Can an AI agent be a soul mate (lover) of a young person?	

Future capabilities of social robots?

Can a robot be a counselor?	yes
Can a robot assess a human?	yes
Can AI assess a human?	yes
Can robot have conflicting goals with humans?	yes
Can a robot have a personality?	yes
Can a robot be a soul mate of a senior person?	yes
Can an AI agent be a soul mate (lover) of a young person?	yes

Examples of the capabilities

Can a robot be a counselor?	Eliza (Weizenbaum 1966), Ellie (ICT, 2011, https://ict.usc.edu/prototypes/simsensei/)
Can a robot assess a human?	DiNuovo et al. 2019 https://www.researchgate.net/publication/331673714_Assessment_of_Cognitive_skills_via_Human-robot_Interaction_and_Cloud_Computing
Can AI assess a human?	Personality tests, job interviews, implicit assessment of articles by author information, ...
Can robot have conflicting goals with humans?	HAL2000, navigation obstacles, recommendations unacceptable
Can a robot have a personality?	Affection, humor, appearance, ...
Can a robot be a soul mate of a senior person?	Companion robots
Can an AI agent be a soul mate (lover) of a young person?	Companion robots

Future capabilities and ethics

Should a robot provide counseling?	
Should a robot perform assessment of a human?	
Should AI perform assessment a human?	
Should robot be allowed to have conflicting goals with humans?	
Should a robot have a personality?	
Should a robot be a soul mate of a senior person?	
Should an AI agent be a soul mate (lover) of a young person?	

Future capabilities and ethics

“It is important to reflect how the capabilities and characteristics of current robot agents *can* shape the world and our reality (skills) and how such agents *should* shape the future societies and services (needs)”

Jokinen, K. (2020). Exploring Boundaries among Interactive Robots and Humans. In: D'Haro et al. (Eds.) Conversational Dialogue Systems for the Next Decade. LNEE, Springer.

Future capabilities and ethics

Should a robot provide counseling?	???
Should a robot perform assessment of a human?	???
Should AI perform assessment a human?	???
Should robot be allowed to have conflicting goals with humans?	???
Should a robot have a personality?	???
Should a robot be a soul mate of a senior person?	???
Should an AI agent be a soul mate (lover) of a young person?	???

Development of Interactive Agents



- Oeh Oeh !
- Oeh Oeh ha ?
- Oeh.



- Me hungry! You Food?
- Me Food! You water?
- Come! Me water!



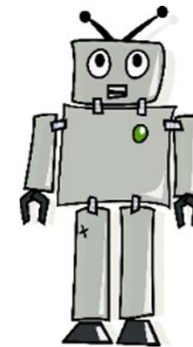
Development of Interactive Agents



- Nice weather, isn't it?
- Yes, very nice.
How is your wife?
- Fine, thank you.



- I just did the Turing test
- And?
- Passed it, no problem!



Some References

- Allwood, J. 1976. Linguistic Communication as Action and Cooperation. Gothenburg Monographs in Linguistics, 2. Göteborg University, Department of Linguistics.
- Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Special Issue, International Journal of Language Resources and Evaluation*, pp. 273–287.
- Andrist, S., Mutlu, B., Gleicher, M. 2013. Conversational gaze aversion for virtual agents. In Proc. IVA 2013, pp 249–262.
- Argyle M., Cook, M. 1976. Gaze and Mutual Gaze. Cambridge University Press.
- Endsley, M.R.: Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors Journal* 37(1), 32-64
- Findlay, J.M., Gilchrist, I.D. 2003. Active Vision: the Psychology of Looking and Seeing. Oxford University Press, Oxford.
- Ijuin, K., Jokinen, K. (2019). Utterances in Social Robot Interactions – correlation analyses between robot’s fluency and participant’s impression. HAI 2019
- Jokinen, K. 2009. Constructive Dialogue Modelling - Speech Interaction with Rational Agents. John Wiley & Sons.
- Jokinen, K. 2018. Dialogue Models for Socially Intelligent Robots. ICSR 2018, Qingdao.
- Jokinen, K., Harada, K., Nishida, M. Yamamoto, S. 2010. Turn-alignment using eye-gaze and speech in conversational interaction. Proceedings of Interspeech-2010. Makuhari, Japan.
- Jokinen, K., Furukawa, H., Nishida, M., Yamamoto, S. (2013). Gaze and Turn-taking behaviour in Casual Conversational Interactions. , Special Issue on Eye-gaze and Conversational Engagement. *ACM Transactions on Interactive Intelligent Systems (TiIS)*, 3:2.
- Jokinen K., Watanabe K. (2019) Boundary-Crossing Robots: Societal Impact of Interactions with Socially Capable Autonomous Agents. In: Salichs M. et al. (eds) Social Robotics. ICSR 2019.

Some References

- Jokinen, K., Hurtig, T. 2006. User Expectations and Real Experience on a Multimodal Interactive System. *Proceedings of Interspeech 2006*, Pittsburgh, US.
- Jokinen, K., Scherer, S. 2012. Embodied Communicative Activity in Cooperative Conversational Interactions - studies in Visual Interaction Management. *Acta Polytechnica Hungarica. Journal of Applied Sciences*. Vol 9, No. 1, pp. 19-40.
- Kendon, A. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press
- Lala, D., Wilcock, G., Jokinen, K., Kawahara, T. 2019. *ERICA and WikiTalk*. IJCAI 2019.
- Laohakangvalvit, T., Jokinen, K. (2019). Eye-gaze Behaviors between Human-Human and Human-Robot Interactions in Natural Scene. ERCEM, 2019.
- Levitski, A., Radun, J., Jokinen, K. 2012. Visual Interaction and Conversational Activity. *The 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye Gaze and Multimodality* Santa Monica, CA, USA
- Mori, T., Jokinen, K., Den, Y. 2020. Analysis of Body Behaviours in Human-Human and Human-Robot Interactions LREC ONION Workshop on peOple in laNguage, vlsiOn and the mind.
- Mutlu, B. Kanda, T. Forlizzi, J. Hodgins, J., Ishiguro, H. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems*, 1(2):12:1–12:33.
- Nakano, Y., Conati, C., Bader, T. 2013. *Eye Gaze in Intelligent User Interfaces*, Springer
- Watanabe, K., Jokinen, K., 2020. Interactive Robotic Systems as Boundary-Crossing Robots – the User’s View. The 29th IEEE International Conference on Robot and Human Interactive Communication (Ro-Man).
- Wilcock, G., Jokinen, K. 2015. *Multilingual WikiTalk: Wikipedia-based talking robots that switch languages*. SIGDial.