

Spoken Dialogue Processing for Multimodal Human-Robot Interaction

Tatsuya Kawahara
(Kyoto University, Japan)

<http://www.sap.ist.i.kyoto-u.ac.jp/~kawahara/pub/ICMI19-tutorial.pdf>

1

Spoken Dialogue Systems (SDS) are prevailing

- Smartphone Assistants

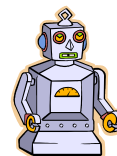


- Smart Speakers



What about Social Robots?

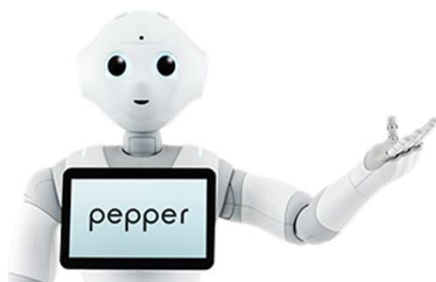
- Social Robots
Intended for interaction with human



2

A majority of Peppers are returned without renewing rental contracts

2015



© Softbank

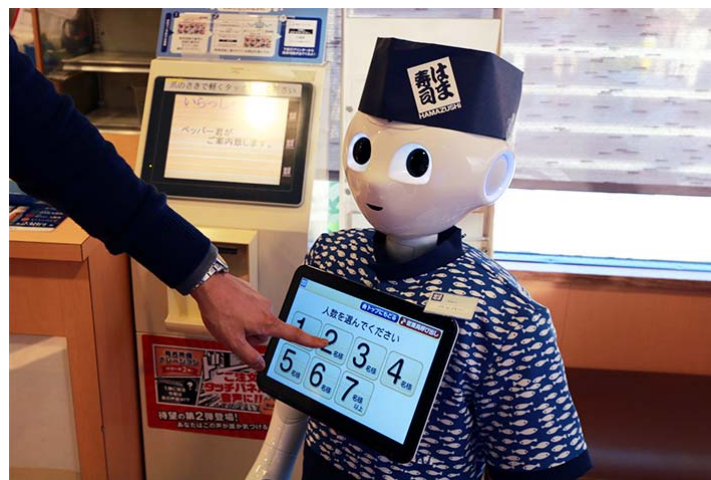
3 years later

2018



3

In successful cases, speech input is not used



© Softbank

4

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>

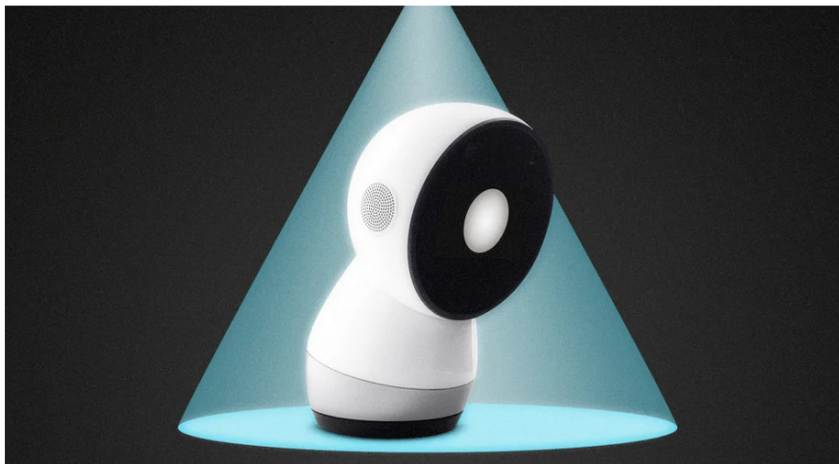
©価格.com

5

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>

Aibo came back

- First generation shipped in 1999
- SONY terminated the product in 2006



© SONY

- New generation shipped 2018
- ...?



© SONY

7

Agenda (Research Questions)

1. Why robots are not prevailing in society?
2. What kind of **tasks** are robots expected to conduct?
3. What kind of **robots** are suitable (for the task)?
4. Why **spoken dialogue** (speech input) is not working with robots?
5. What kind of **other modalities** and interactions are useful?
6. What kind of **evaluations** should be conducted?

What?

Who?

How?

How (well)?

8

Agenda (Research Questions)

1. Why robots are not prevailing in society?
2. What kind of **tasks** are robots expected to conduct?
 1. Who are typical users? Whom?
 2. Where are they served? Where?
3. What kind of **robots** are suitable (for the task)? Who?
 1. Difference between virtual agents vs. humanoid robots?
 2. Does physical presence or multi-modality matter?
4. Why **spoken dialogue** (speech input) is not working with robots?
 1. Speech and Language processing (ASR, TTS, SLU, DM)
 2. Non-verbal issues...turn-taking, backchannel How?
5. What kind of **other modalities** and interactions are useful?
6. What kind of **evaluations** should be conducted? How (well)?

1. Why robots are not prevailing in society?

- Basically cost issue
 - Hardware fragile → maintenance
 - Much more expensive (>10 times) than smart speakers
- Performance to meet the price?
- Unused → Big useless hardware

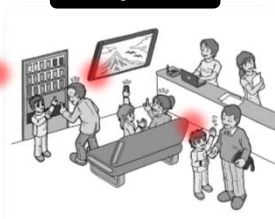
2. What kind of tasks are robots expected to conduct?

11

Expected Roles by Robots

Physical presence & Face-to-Face interaction matters

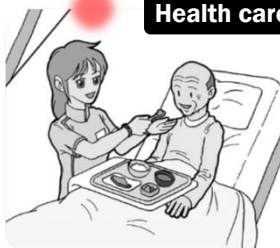
Receptionist receive=welcome



Attendant attend=care



Health care → Senior



Teaching → Children



12

Other Scenarios?

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

13

Dialogue Category (Tasks)

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

14

Dialogue Category (Tasks)

- User initiative
- System initiative
- Mixed initiative

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

15

Dialogue Category (Tasks)

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

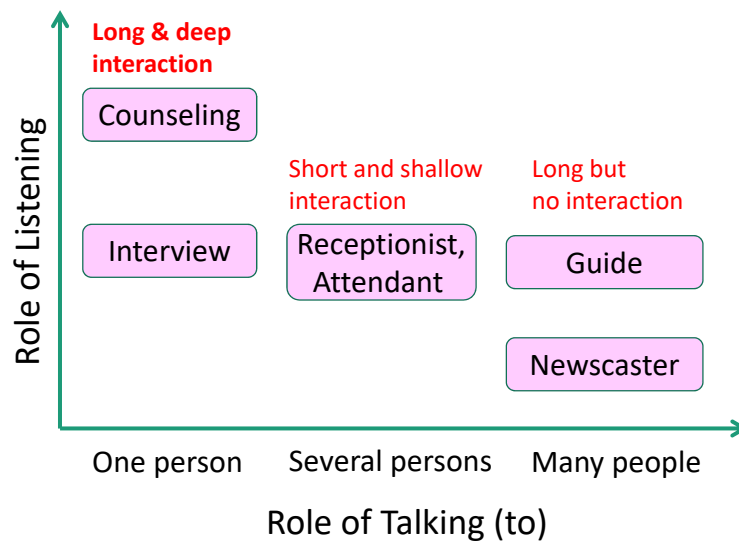
Agent is OK?

Adult android effective

Mechanical Robot

16

Dialogue Roles of Adult Androids



17

Chatting function desired

- In many cases (most of the tasks)
- Ice-breaking in the first meeting
- Relaxing during a long interaction
- Keeping engagement

18

3. What kind of robots are suitable (for the task)?

19

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]

20

Animal Robots Stuffed Animals Talking (some listening)

- Aibo



© SONY

- Paro



© Daiwa House

- ????



Substitute of a pet

21

Child-looking or Child-size Humanoid Robots

- CommU



©VSTONE, Osaka U

- Nao



© Softbank robotics

- Palro



© Fuji soft

Substitute of a grandchild

22

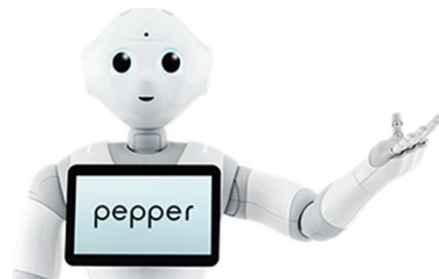
Adult-size Humanoid Robots

- Asimo



© HONDA

- Pepper



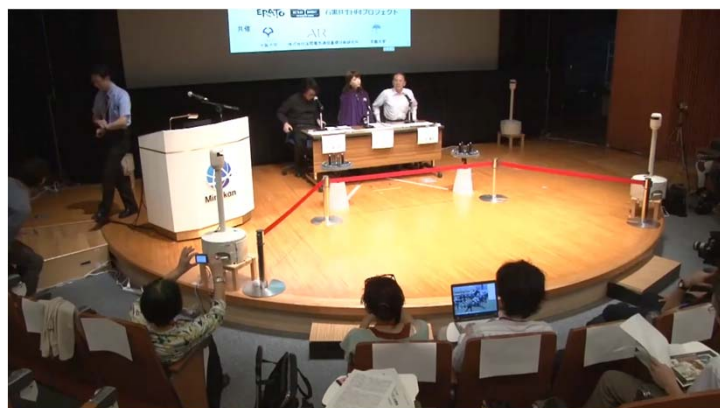
© Softbank

Still child-like! → Implying not so intelligent

23

Adult Androids

- ERICA



Debut in 2015

24

How long can you keep talking?

- Smart Speaker

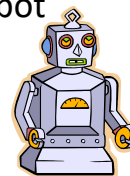


- Virtual Agent



MMD Agent ©NITECH

- Humanoid Robot



- Human

(A person you meet for the first time)



25

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

- Pet



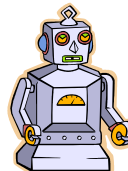
- Baby



- Kid (~10 year old)



- Humanoid


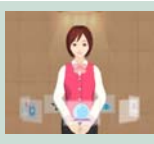

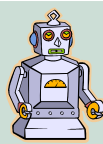
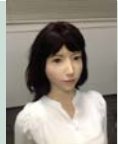


- Android




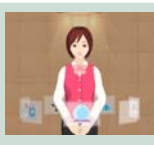

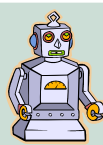
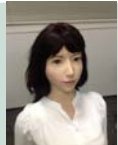
26

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					

27

Comparison of Dialogue Interfaces

	Smart Speaker	Virtual Agent	Pet Robot	Humanoid Robot	Adult Android
					
???					
???					
???					
???					
???					

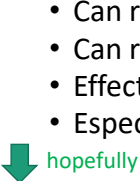

28

Difference between Virtual Agents vs. Humanoid Robots/Androids?

- Physical presence + moving
- Fully multi-modal including eye-gaze
 - Hard to make mutual gaze with agents
- Robots are deemed to be more autonomous than agents
 - Move and act autonomously
 - Can be a partner
- ???

29

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
- 
 • Attachment
- 
 • Bullying esp. by group of kids
 - (cf.) Virtual agents cursed

30

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, teaching) → virtual agents
- Not 'collaborative' users
 - Kids and senior people do not follow the protocol

31

Dialogue Category (Tasks)

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

Agent is OK?

Adult android effective

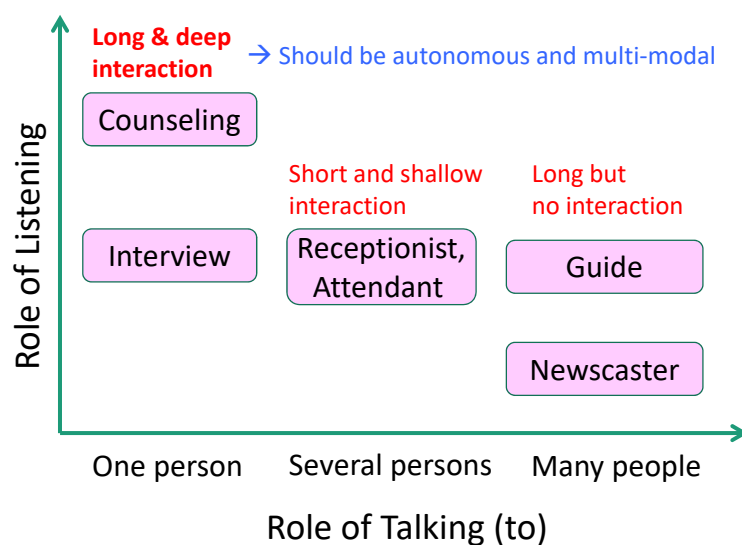
32

Face-to-Face (F2F) 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

33

Dialogue Roles of Adult Androids



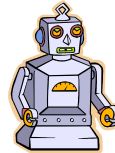
34

Tool ↔ Companion, Partner

- Smartphone Assistants



- Communicative Robots



- Smart Speakers

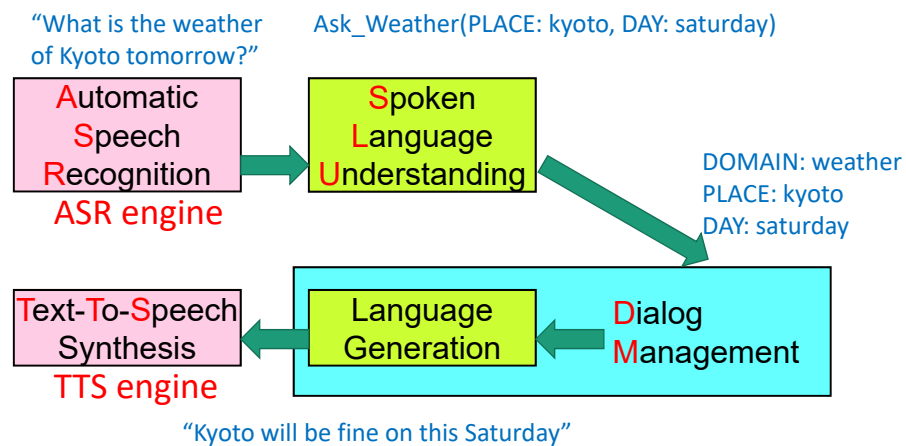


35

4. Why spoken dialogue
(speech input) is not working
with robots?

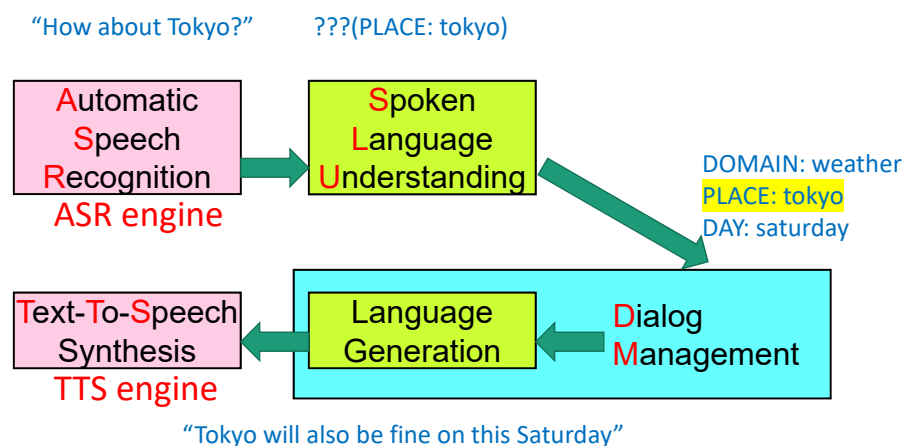
36

Architecture of Spoken Dialogue System (SDS)



37

Architecture of Spoken Dialogue System (SDS)



38

Automatic Speech Recognition (ASR)

39

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

40

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

41

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

42

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

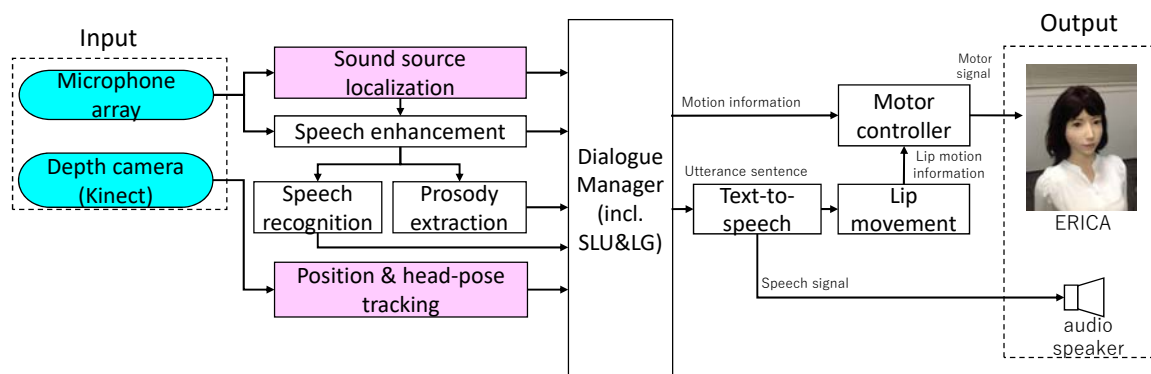
43

Example Implementation for ERICA



44

Example Implementation for ERICA



45

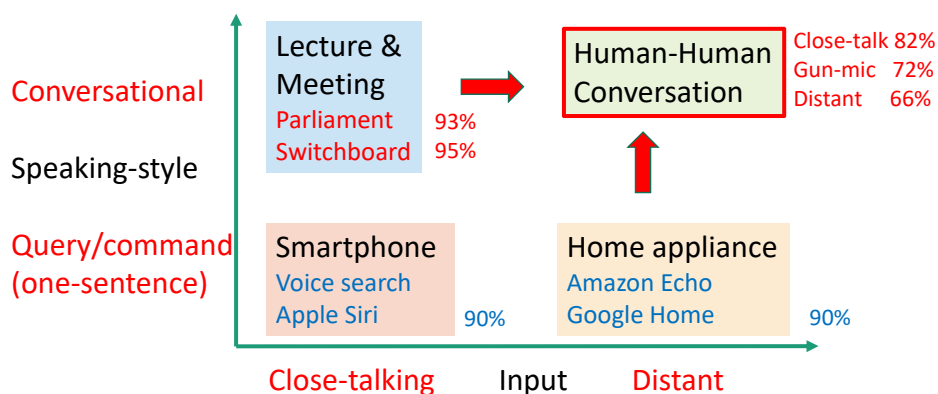
Real Problem in Distant Talking

- 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

46

Distant & Conversational Speech Recognition

Accuracy is degraded with the synergy of two factors



47

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

48

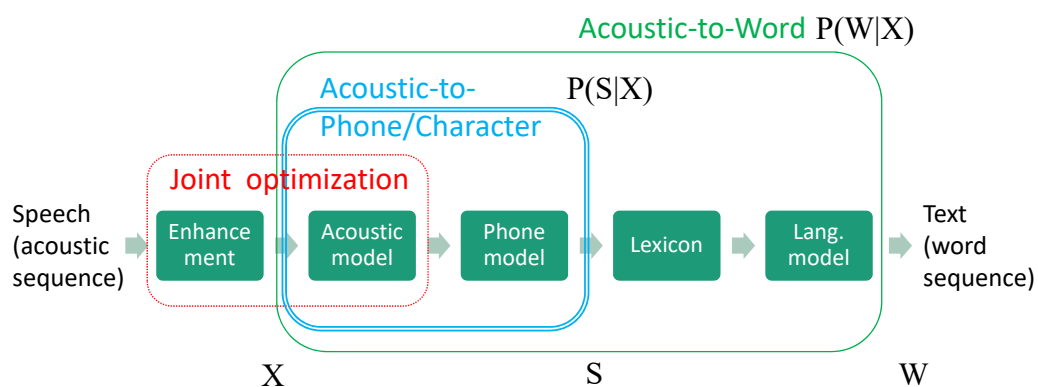
Review of ASR

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
 - Cloud-based ASR can hardly meet requirement
- ↓
- Recent End-to-End (acoustic-to-word) ASR
 - 0.03xRT though still need to wait for the end of utterances
 - All downstream NLP modules must also be tuned

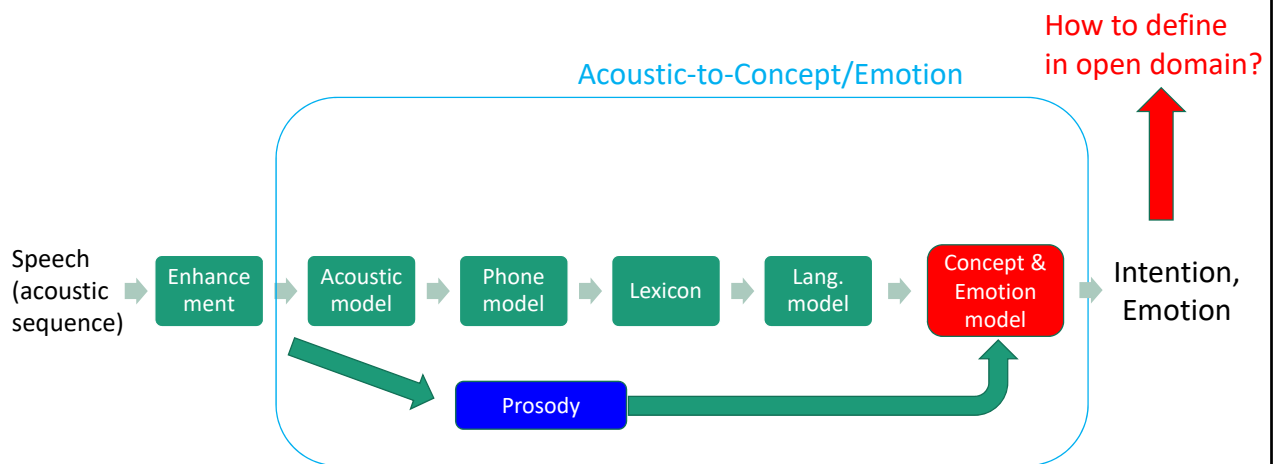
49

End-to-End Automatic Speech Recognition (ASR)



50

End-to-End Speech Understanding




51

Text-To-Speech Synthesis (TTS)

52

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

53

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?

54

Voice of Android ERICA

Conversation-oriented

- Backchannels
- Filler
- Laughter



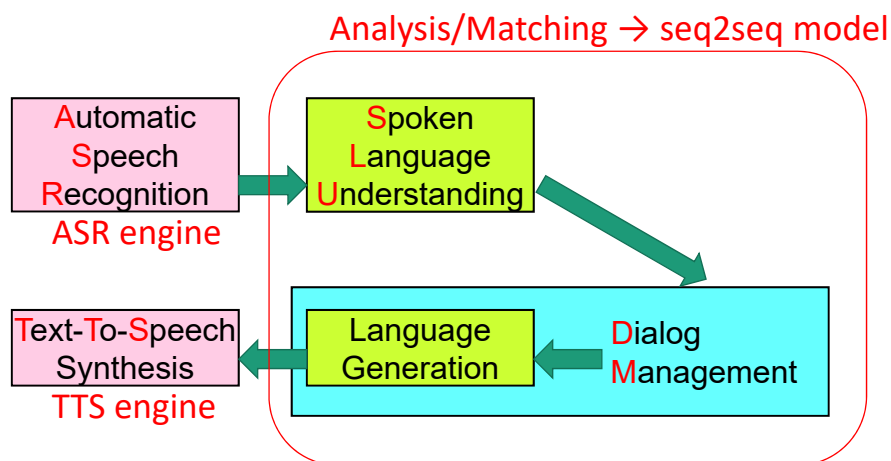
<http://voicetext.jp> (ERICA)

55

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

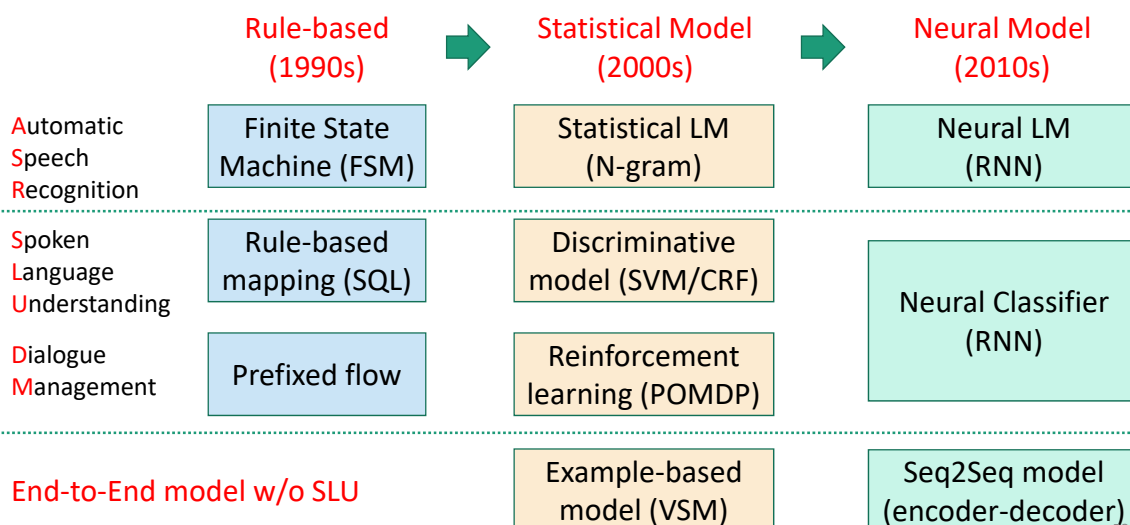
56

Architecture of Spoken Dialogue System (SDS)



57

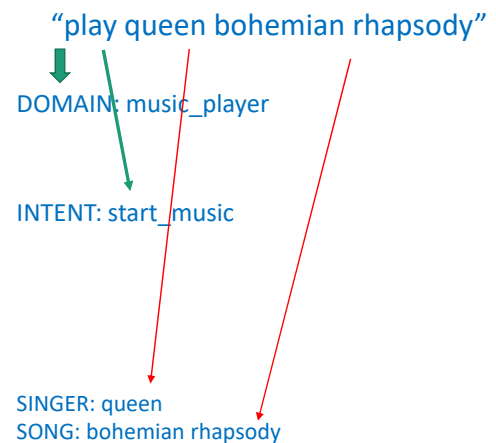
Historical Shift of Methodology



58

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
- **Slot/Entity**
 - Named Entity (NE) tagger
 - Numerical values



59

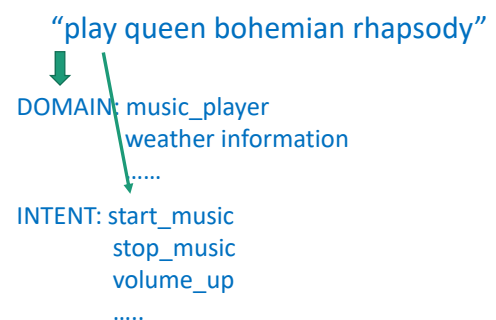
Semantic Analysis for SLU

- **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



60

Semantic Analysis for SLU

Sequence labeling problem

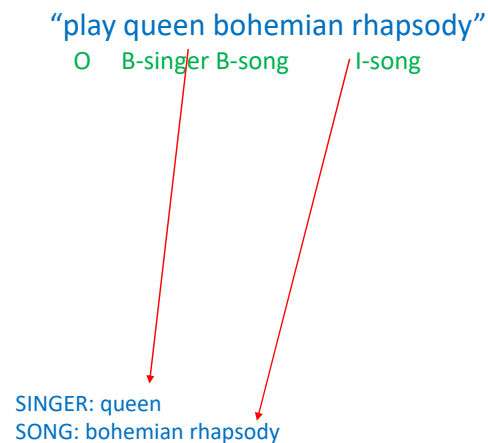
- Statistical Discriminative Model: CRF
- Neural Tagger: RNN

Domain-independent NE tagger



• Slot/Entity

- Named Entity (NE) tagger
- Numerical values



61

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

62

Dialogue Management

- Decide proper **Action**

- Make query/command
- Present results

“What is the weather of Kyoto tomorrow?”

Ask_Weather(PLACE: kyoto, DAY: saturday)

DOMAIN: weather
PLACE: kyoto
DAY: saturday

“Kyoto will be fine on this 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?**

63

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)

64

Disambiguation by Dialogue SHRDLU [Winograd1972]

Operating blocks in virtual world

U: PICK UP A BIG RED BLOCK.

S: OK. (does it)

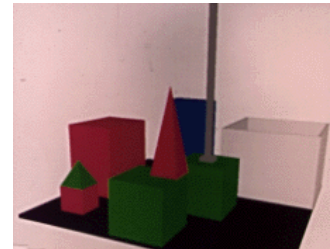
U: GRASP THE PYRAMID.

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)



65

Review of Orthodox SLU+DM

- Show intelligent behaviors
- Complex and not easy to implement

- Hand-crafted rules → Only works for small tasks
 - Easy and practical solution
 - Not flexible and scalable
- Machine learning (Statistical & Neural) → Only possible for big companies
 - 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?

66

Alternatives of Orthodox SLU+DM

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

67

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.

68

Dialogue based on Pattern Matching ELIZA [Weizenbaum1966]

Word:Rank	Outputs
alike:10	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?

69

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"

70

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
 - Elaborate
 - Needs a training data set

71

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

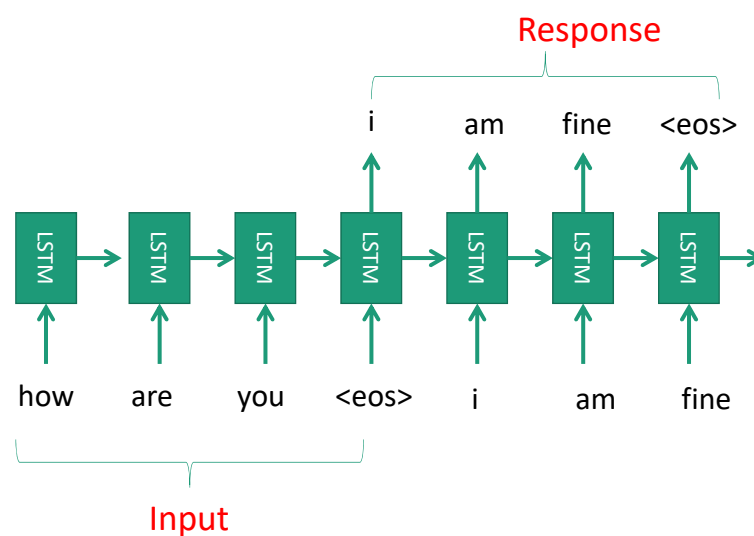
72

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

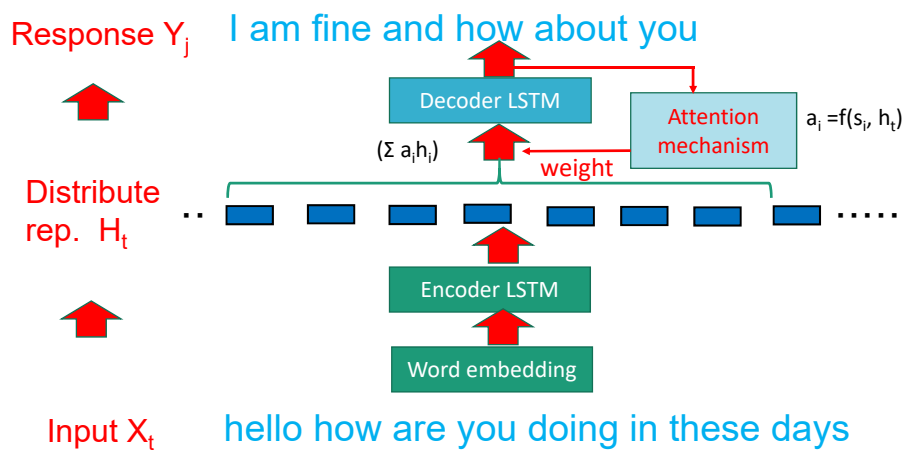
73

Neural Conversation Model



74

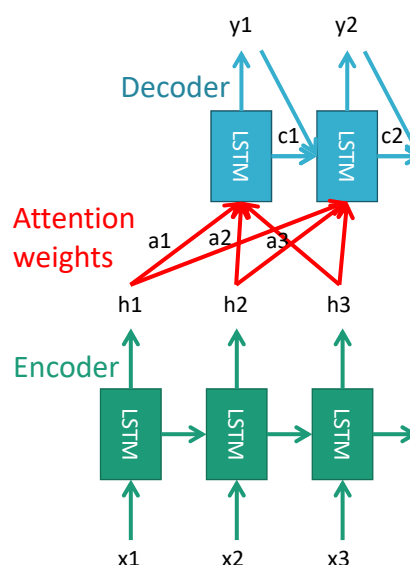
Encoder-Decoder (Seq2Seq) Model with Attention Mechanism



75

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



76

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?”

77

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

78

(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

} combination

79

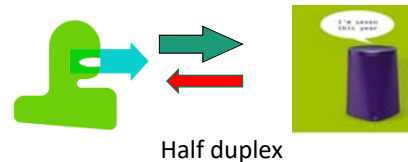
Non-verbal Issues in Dialogue

80

Protocol of Spoken Dialogue

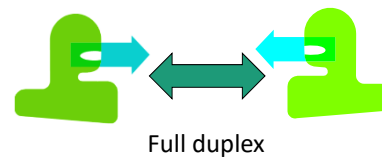
• Human-Machine Interface

- 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



81






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

82

Comparison of Dialogue Interfaces

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

83

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

84

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



85

Generating Backchannels

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

86

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**

87

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

88

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



89

Generating Fillers

- No filler 
- Filler before moving to next question 

90

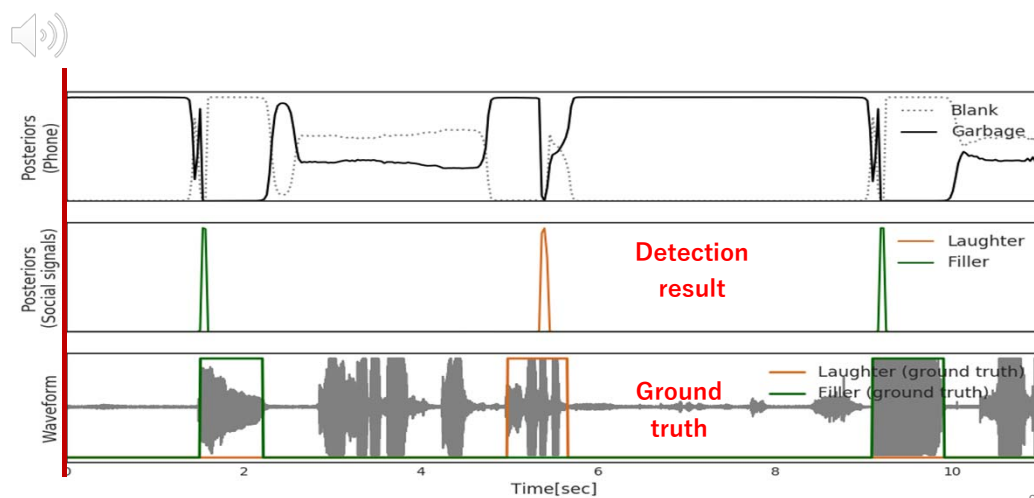
Generating Laughter

- 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



91

Detection of Laughter, Backchannels & Fillers



92

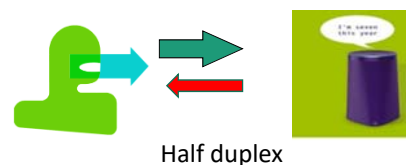
Turn-taking

93

Protocol of Spoken Dialogue

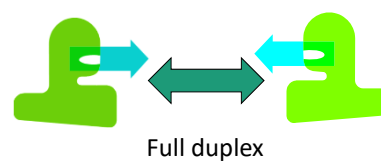
- **Human-Machine Interface**

- 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



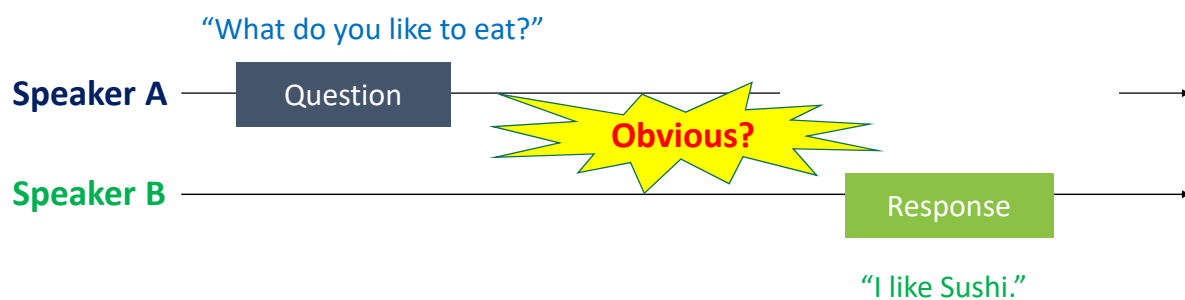
94

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

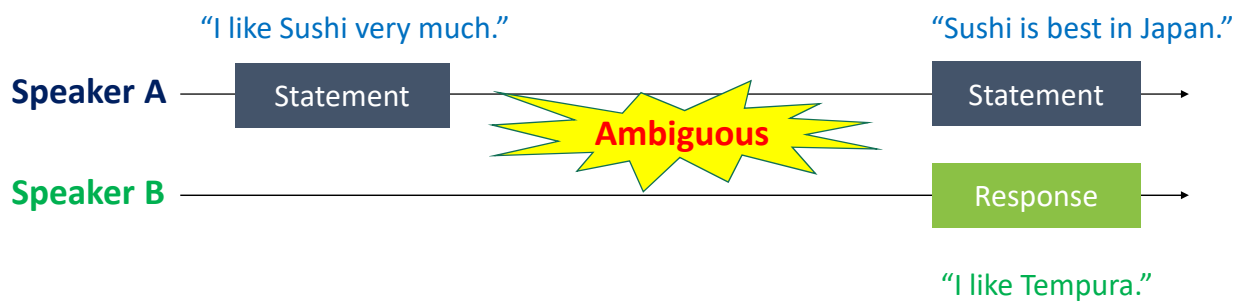
95

Turn-switch after Question



96

Turn-keep/switch after Statement?



97

Turn-keep/switch after Response?



98

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

99

Proactive Turn-taking System

- Fuzzy decision ← Binary decision

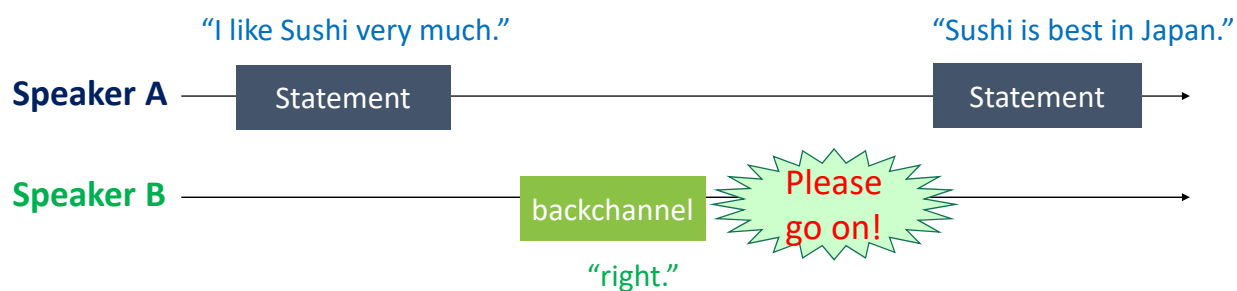


- Use fillers and backchannels when ambiguous

	User status	System action
confidence ↑	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

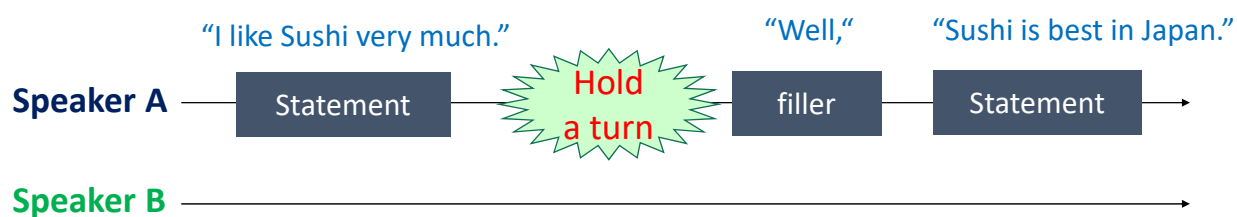
100

Turn-keep/switch after Statement?



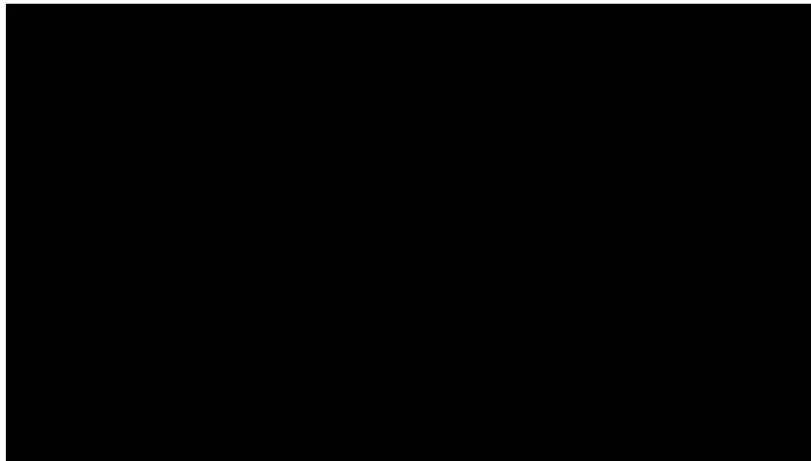
101

Turn-keep/switch after Statement?



102

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



103

Initiative Management

- **System-initiative**
 - System mostly (talks OR) asks questions to users before service
 - Adopted by call centers
(ex.) form-filling, questionnaire, interview, guide
- **User-initiative**
 - User mostly asks questions/queries (OR talks) to system
 - Adopted by smartphone assistants and smart speakers
(ex.) assistant, receptionist, attentive listening
- **Mixed-initiative**
 - Adopted by chat bots
(ex.) chatting, speed dating, negotiation, debate, counseling

104

Dialogue Category (Tasks)

- User initiative
- System initiative
- Mixed initiative

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

105

5. What kind of other modalities and interactions are useful?

(some of them already mentioned)

106

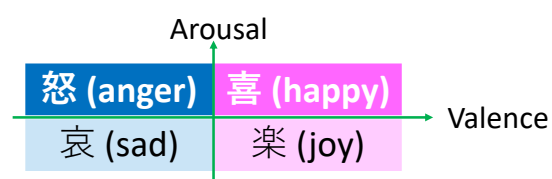
Recognition of Mental States during Dialogue

- Valence
 - Positive/negative feeling on what is talked about
 - proper assessment by the system
- Engagement
 - Positive/negative attitude to keep the current dialogue
 - change topics and manner of the system response
- Rapport
 - Trust/attachment to the robot

107

Emotion Recognition

- Arousal-Valence Model



- **Arousal** recognition is easy
 - Prosody (power)
 - But people are not so often angry or happy
- **Valence** recognition is difficult but important
 - Prosody...not reliable
 - Lexical (→sentiment analysis)...ASR error prone
 - proper assessment by the system
 - **empathy to positive/negative feeling**

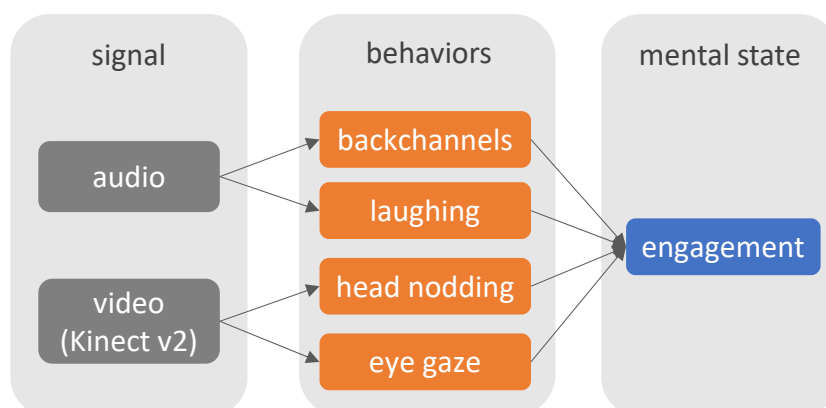
108

Engagement Recognition

- Engagement
 - Willingness to start and continue the dialogue
 - Important for system's dialogue action
 - change topics and manner of the system response
- Cue (features)
 - Audio: backchannel (BC), laughter
 - Visual: eye-gaze, Nodding
- Machine learning
 - Annotation is NOT easy in both quantity and quality (subjective)

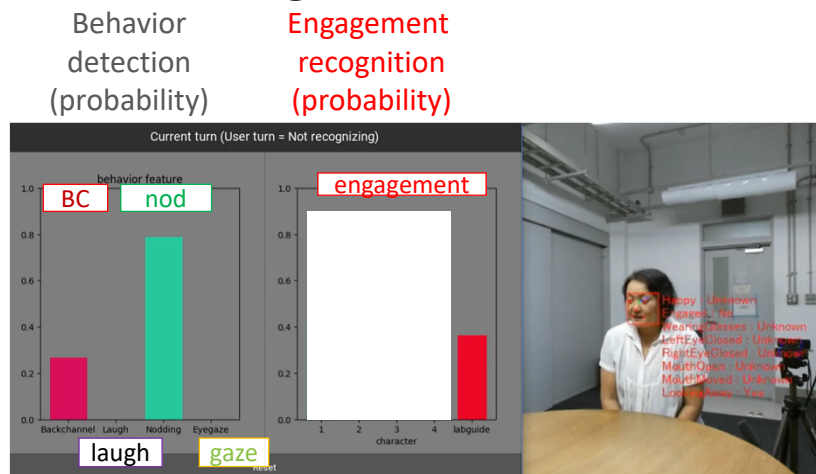
109

Engagement Recognition via User Behaviors



110

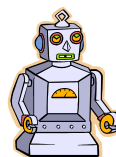
Demonstration of Behavior Detection and Engagement Recognition



111

Other Non-verbal Interaction Modalities

- Facial expression
- Gesture
- Posture and movement
- Touch



112

Character Modeling

- Appropriate Character
 - Counselor: attentive and introvert
 - Receptionist: attentive and formal
 - Guide to VIP: extrovert and formal
 - Guide to kids: extrovert and casual
- Character modeling
 - Big Five
- Behavior modeling
 - backchannels and fillers
 - turn-switch time
 - amount of utterances
 - prosody and speaking delivery

113

Case Studies

114

Demonstration of Two-robot System



115

Demonstration of Attentive Listening System



ERICA can converse for five minutes with naïve users!!

116

Demonstration of Job Interview (English)



ERICA can converse for five minutes with naïve users!!

117

6. What kind of **evaluations** should be conducted?

118

Experiments

- Lab experiments
 - Subjects are collected, paid, and well-prepared
 - Controlled environment
 - Necessary for writing papers
- Field experiments
 - Real (ad-hoc) users
 - Real environment
 - Necessary for feasibility study

119

Evaluation Criteria

- Objective evaluation
 - Responses/behaviors are appropriate or not
- User reaction
 - Positive/negative behaviors
- Subjective evaluation
 - Comparison in different settings
 - User experience OR Third person's viewpoint
- Total Turing Test
 - Comparable to WOZ setting
 - Comparable to "human-like interaction experience"
 - measured by engagement level

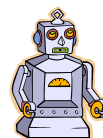
120

Ethical Issues

- Can robot be a counselor?
- Can robot assess a human?
 - Can AI assess a human?
- Can robot be a soul mate of a senior person?
 - Can AI agent be a soul mate (lover) of a young person?

121

Thank you for your attention



122

References

1. Christoph Bartneck, Tony Belpaeme, Friederike Eysel, Takayuki Kanda, Merel Keijsers, Selma Sabanovi.
Human-Robot Interaction — An Introduction.
<https://www.human-robot-interaction.org/>
2. Tatsuya Kawahara.
Spoken dialogue system for a human-like conversational robot ERICA.
In Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS), (keynote speech), 2018.