Spoken Dialogue System (SDS) for a Human-like Conversational Robot ERICA

Tatsuya Kawahara

(Kyoto University, Japan)

Limitation of Current (deployed) SDS

- Machine-oriented constrained dialogue
  - Think over what system can [conceptual constraint]
  - Utter one simple sentence [linguistic constraint]
  - with clear articulation [acoustic constraint]
  - and wait for response [reactive model]

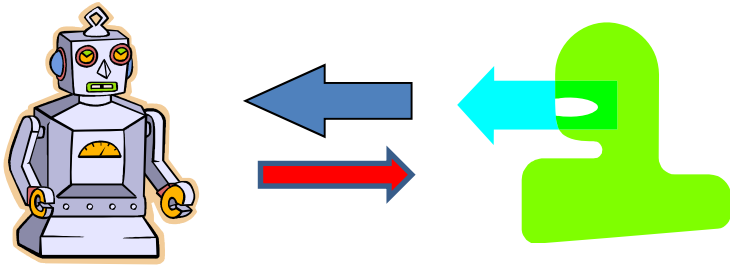
- Big gap from human (or ideal) dialogue
  - Human tourist guide, Concierge at hotels

Robot

Human-Machine Interface (Current SDS)

constrained speech/dialog

- Half duplex and reactive
- One sentence per one turn
- System responds only when user asks

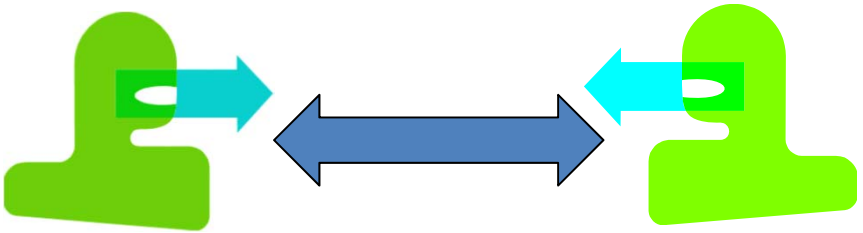


People are aware they are talking to a machine.

Human-Human Communication

natural speech/dialog

- Duplex and interactive
- Many sentences per one turn
- Backchannels



Human is the most natural interface! → Human-like Robot

Android ERICA Project started in 2016

http://sap.ist.i.kyoto-u.ac.jp/erato/

JST ERATO Symbiotic Human-Robot Interaction Project (2014-2020)

- Goal: Autonomous android who behaves and interacts just like a human
  - Facial look and expression
  - Gaze and gesture
  - Natural spoken dialogue
- Criterion: Total Turing Test
  - Convince people it is comparable to human, or indistinguishable from remote-operated android
- Science:
  - Clarify what is missing or critical in natural interaction
- Engineering Applications:
  - Replace social roles done by human（感情労働）
  - Conversation skill training

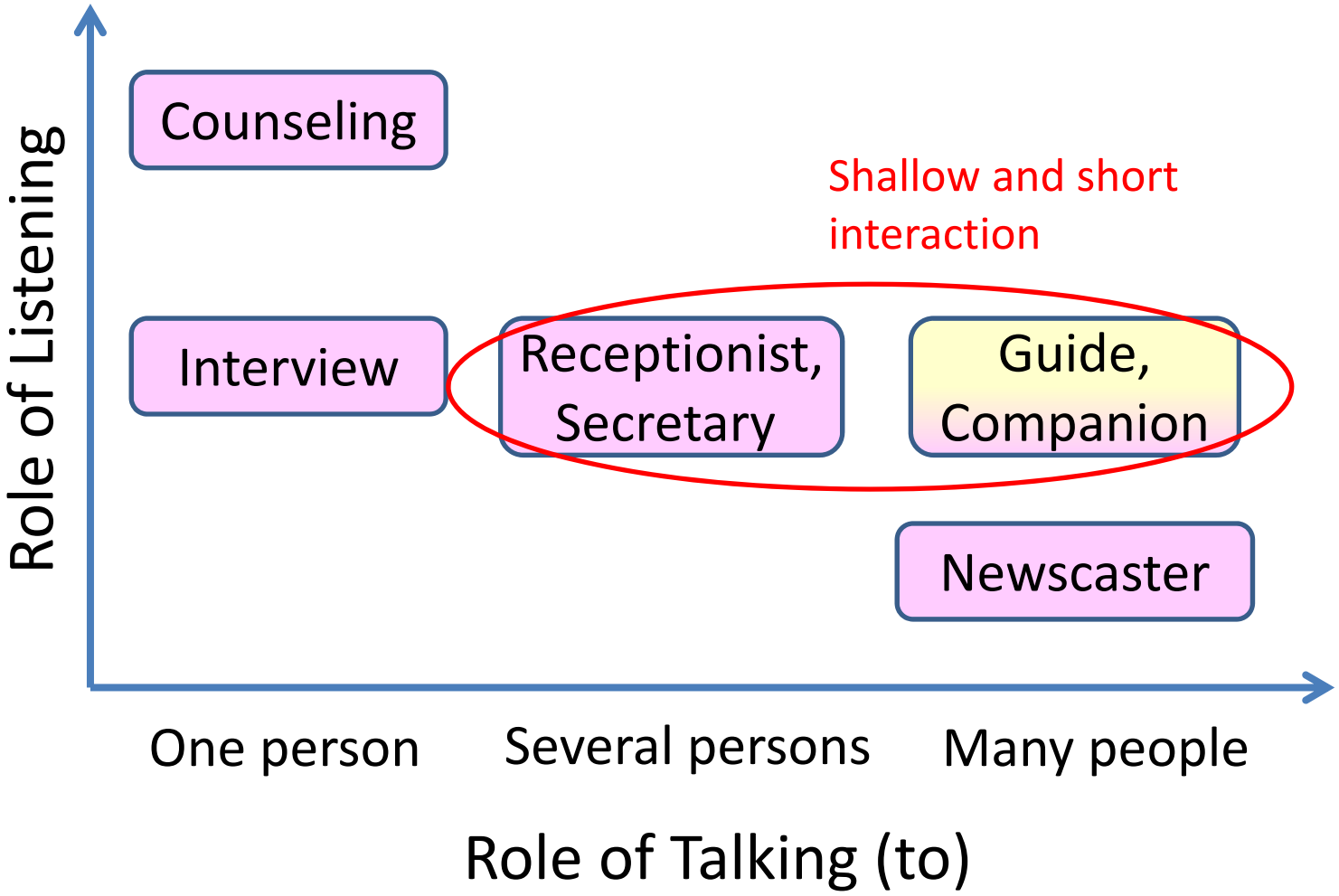
Android ERICA
with flowers with microphones & camera

Tasks of ERICA

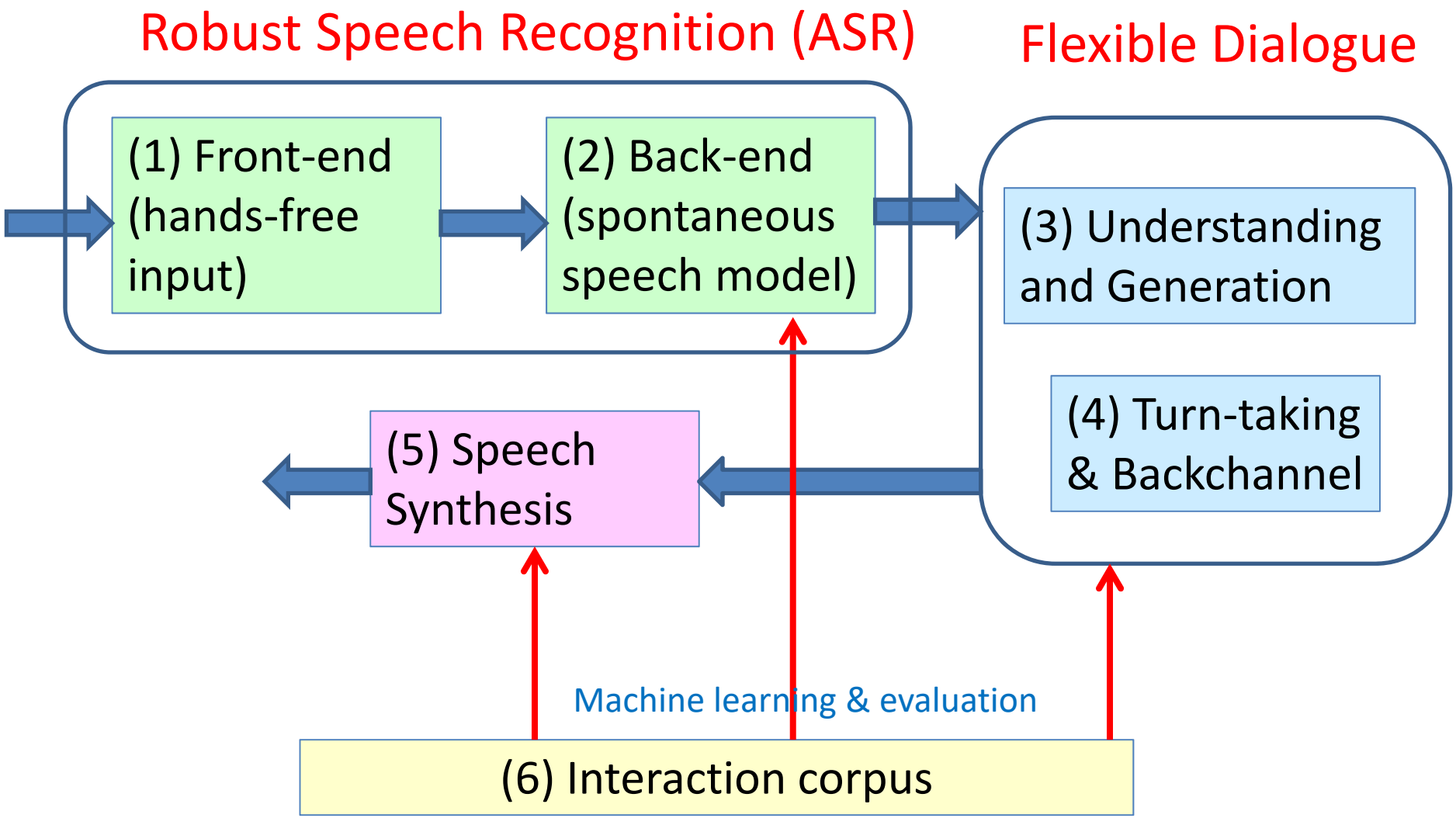
- × Information services → smart phones
- × Move objects → conventional robots
  - × ERICA cannot move except for gestures
- × Chatting → ChatBot
  - × Should involve physical presence and non-verbal communication

- Social Interaction

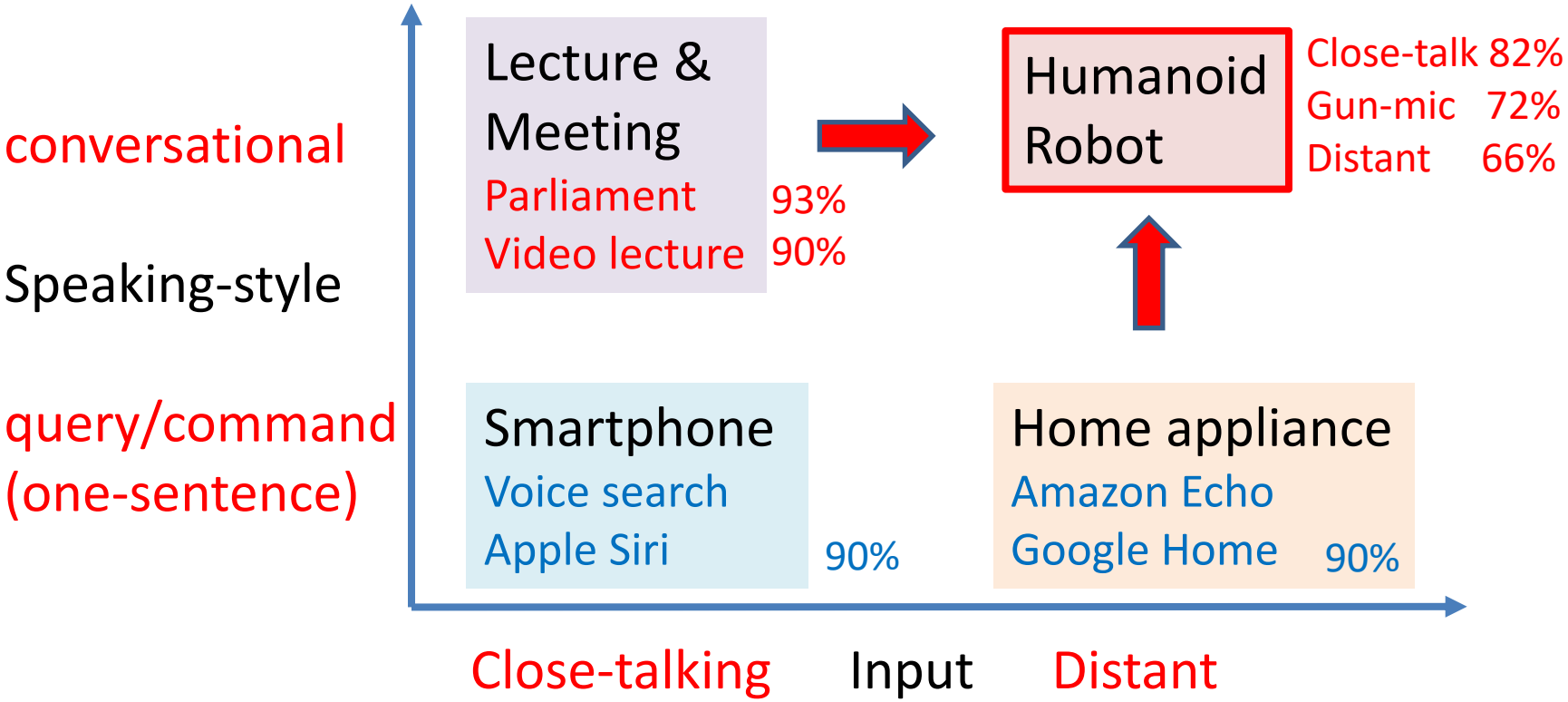
Social Roles of ERICA



Research Topics



Challenge in Speech Recognition



Real Problem in Distant Talking

- When people speak without microphone, speaking style becomes so casual that it is not easy to detect utterance units.
  - Not addressed in conventional “challenges”
  - Circumvented in conventional products
    - Smartphones: push-to-talk
    - Smart speakers: magic word “Alexa”, “OK Google”
    - Pepper: talk when flash

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
- Cloud-based ASR cannot meet requirement

- Recent End-to-End (acoustic-to-word) ASR
  - 0.03xRT [ICASSP18]
- All downstream NLP modules must be tuned

Features in Speech Synthesis

- Very high quality
- Conversational style rather than text-reading
  - Questions (direct/indirect)
- A variety of non-lexical utterances with a variety of prosody
  - Backchannels
  - Fillers
  - Laughter
- http://voicetext.jp (ERICA)

Human-like Dialogue Features

- Hybrid Dialogue Structure
- Mixed-initiative

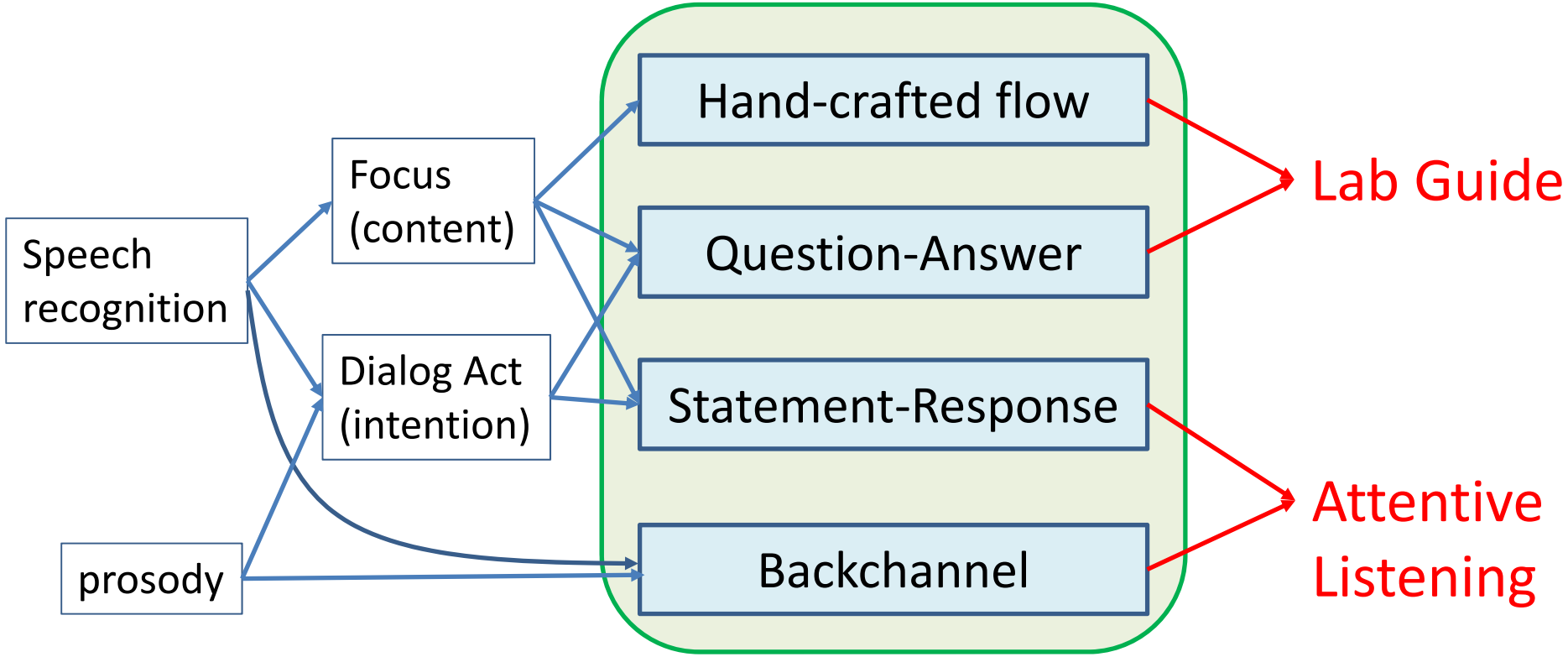
- Natural turn-taking
- Backchanneling

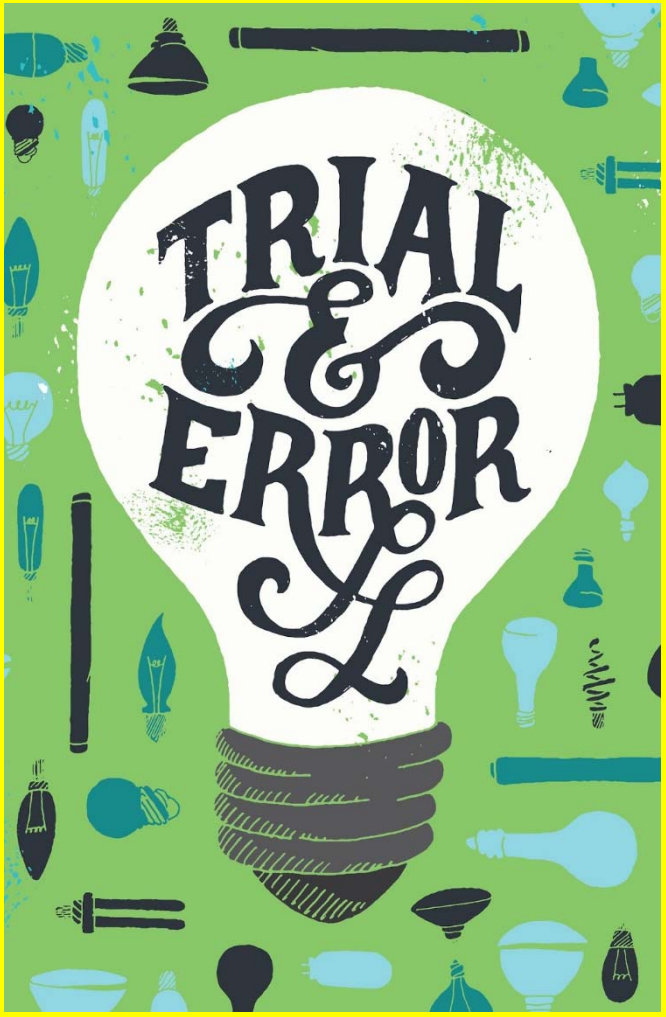
- Non-lexical utterances
- Non-verbal information (in spoken dialogue)

Hybrid of Different Dialogue Modules

- State-transition flow (hand-crafted)
  - Used in limited task domain
  - Deep interaction but works only in narrow domains
  - Cannot cope beyond the prepared scenario
- Question-Answering
  - Used in smartphone and smart speakers
  - Wide coverage but short interaction
  - Cannot cope beyond the prepared DB
- Statement-Response
  - Used in ChatBot
  - Wide coverage but shallow interaction
  - Many irrelevant OR only short formulaic responses

Spoken Dialog System of ERICA





- Systems were not convincing and engaging!
- Dialogues were not realistic!

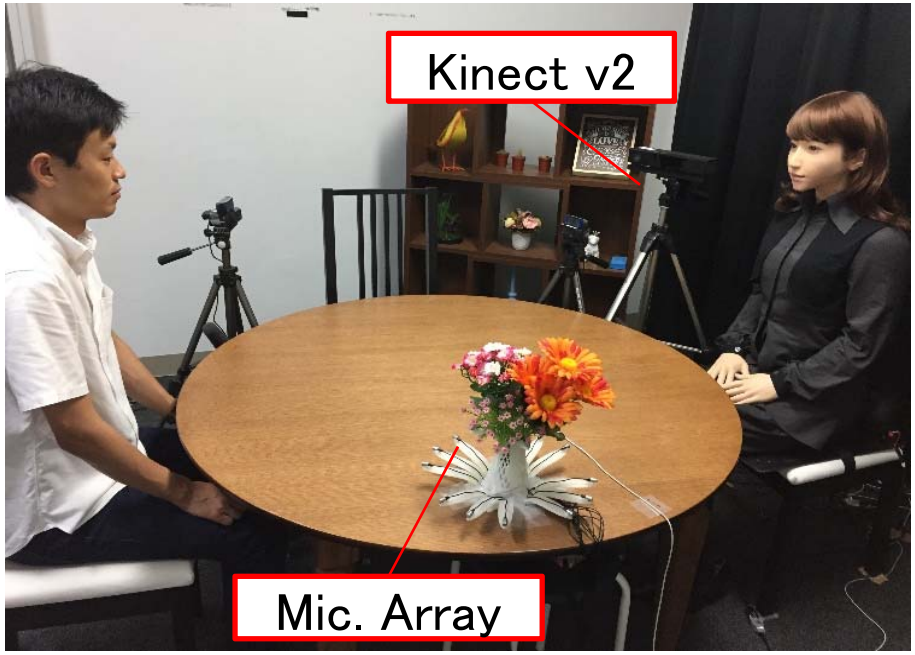
Real Problems in non-task-oriented SDS

- System often generates boring (safe) OR irrelevant (challenging) dialogue.
- Sensible adults (college students) hesitate to talk to robots.
- Attendants and Receptionists involve shallow interaction for easy task.
  - These robots are being deployed.

Our Solutions

- Realistic social role given to ERICA
- So matched users will be seriously engaged
- "Social interaction" task
  - Dialogue itself is task
    - Mutual understanding or appealing
  - (cf.) tasks solved via spoken dialogue
    - query or transaction
  - Not just chatting
  - Must be engaged by users as well as the robot
  - Face-to-face (physical presence) is important

Dialogue with Android ERICA in WOZ setting



control

operator

Task 1: **Attentive Listening**

- ERICA mostly listens to senior people
  - Topics on memorable travels and recent activities
  - Encourages users to speak

Task 2: **Job Interview (Practice)**

- ERICA plays a role of interviewer
  - asks questions, which are answered by users
  - makes additional questions according to initial answers
  - provides a realistic simulation, or replace human
- Users need to appeal themselves

Very strained

Physical presence and face-to-face is important!

Task 3: **Speed Dating (Practice)**

- ERICA plays a role of female participant
  - asks questions to users AND answers questions by users on topics such as hobbies, favorite foods and music
  - provides a realistic simulation by not being too friendly
  - gives proper feedbacks according to the dialogue
- Users need to not only appeal but also listen

Relaxed, but somewhat nervous

Physical presence and face-to-face is important!

Comparison of 3 Tasks

	Attentive Listening	Job interview	Speed Dating
Dialogue Initiative	User	System	Both (mixed)
Utterance mostly by	User	User	Both
Backchannel by	System	System	Both
Turn-switching	Rare	Clear	Complex
# dialogue sessions	19	30	33

Comparison of 3 Tasks

	Attentive Listening	Job interview	Speed Dating
%Utterance by User	64%	53%	49%
%Occurrence of system backchannel	38%	19%	19%
%Turn-switching	19%	30%	37%
Turn-switch time	454msec	629msec	548msec

Challenge: Total Turing Test

1. **Can we generate same responses for a corpus collected via WOZ?**
   [objective evaluation]
2. **Can autonomous ERICA satisfy subjects in a same level as WOZ?**
   [subjective evaluation]

Attentive Listening System

Attentive Listening

- People, esp. senior, want someone to listen.
- Talking by remembering is important for maintaining communication ability.

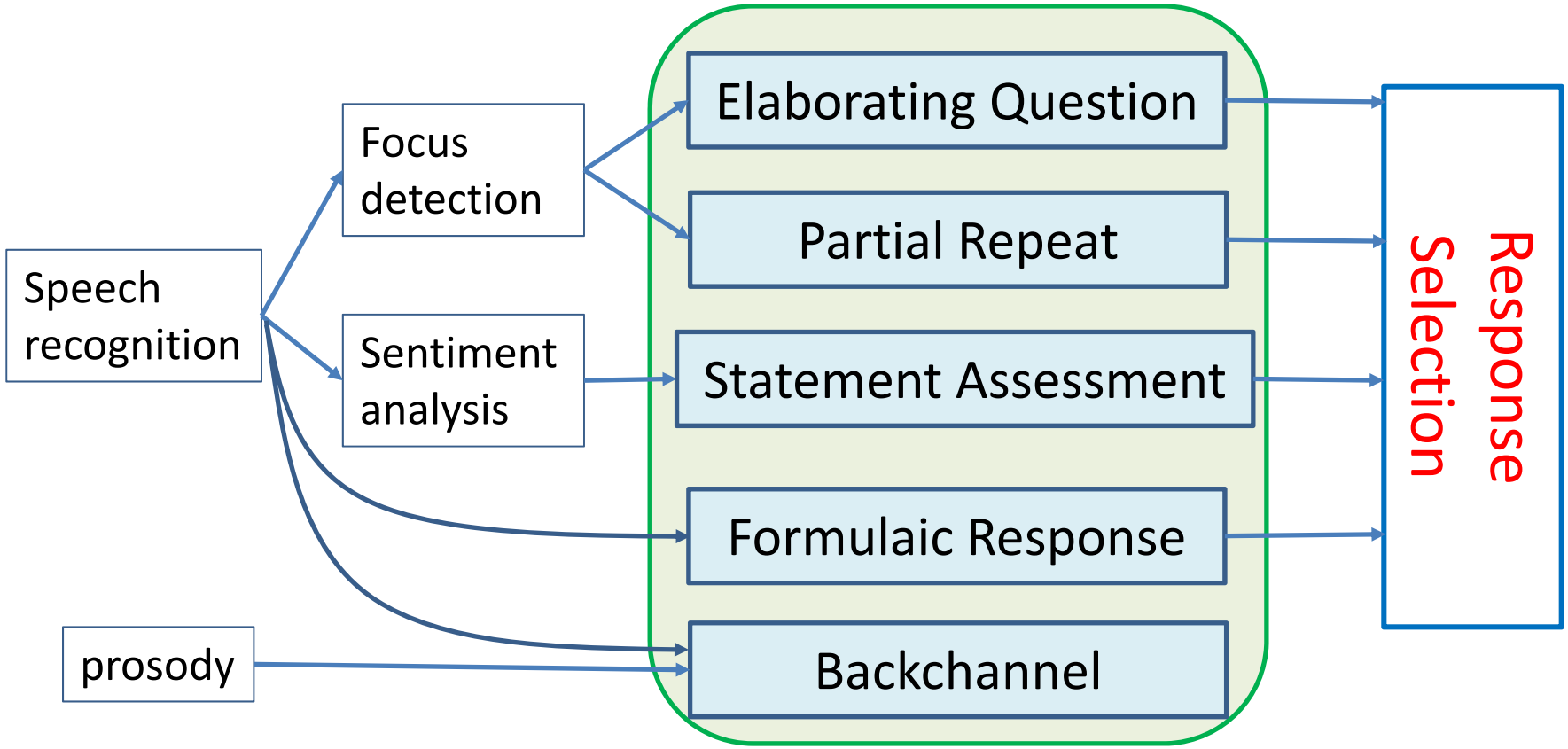
- System (robot), which listens and encourages the subject to talk more
  - Need to respond to anything
  - Does not require large knowledge base
  - Empathy and entrainment is important

Challenge: Total Turing Test of Attentive Listening System

- Can robot be a counselor?
  - Ishiguro thinks so
- Almost all senior subjects believed to be talking to ERICA during data collection in WOZ setting.

1. **Can we generate same responses for a corpus collected via WOZ? [objective evaluation]**
2. **Can autonomous ERICA satisfy subjects in a same level as WOZ? [subjective evaluation]**

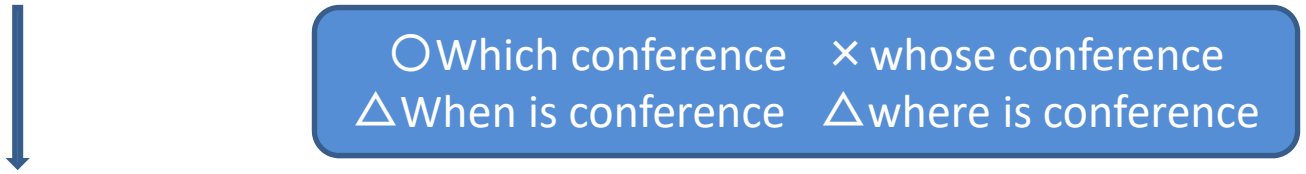
Flow of Attentive Listening System



Elaborating Question and Partial Repeat based on Focus Word

- Detect a focus word
- Try to combine with WH phrases for a plausible question

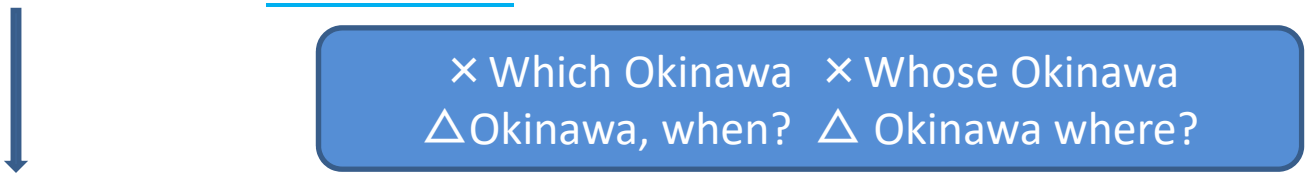
"I went to a conference."



"Which conference?" [Elaborating question]

- Or simply repeat the focus word

"I went to Okinawa."



"Okinawa?" [Partial repeat]

Statement Assessment based on Sentiment Analysis

- Sentimental attribute annotated for each word
- Assessment selection based on (summed) attribute values

	Positive	Negative
Objective (fact)	That's nice 素敵ですね	That's bad 大変ですね
Subjective (comment)	Wonderful いいですね	That's a pity 残念ですね

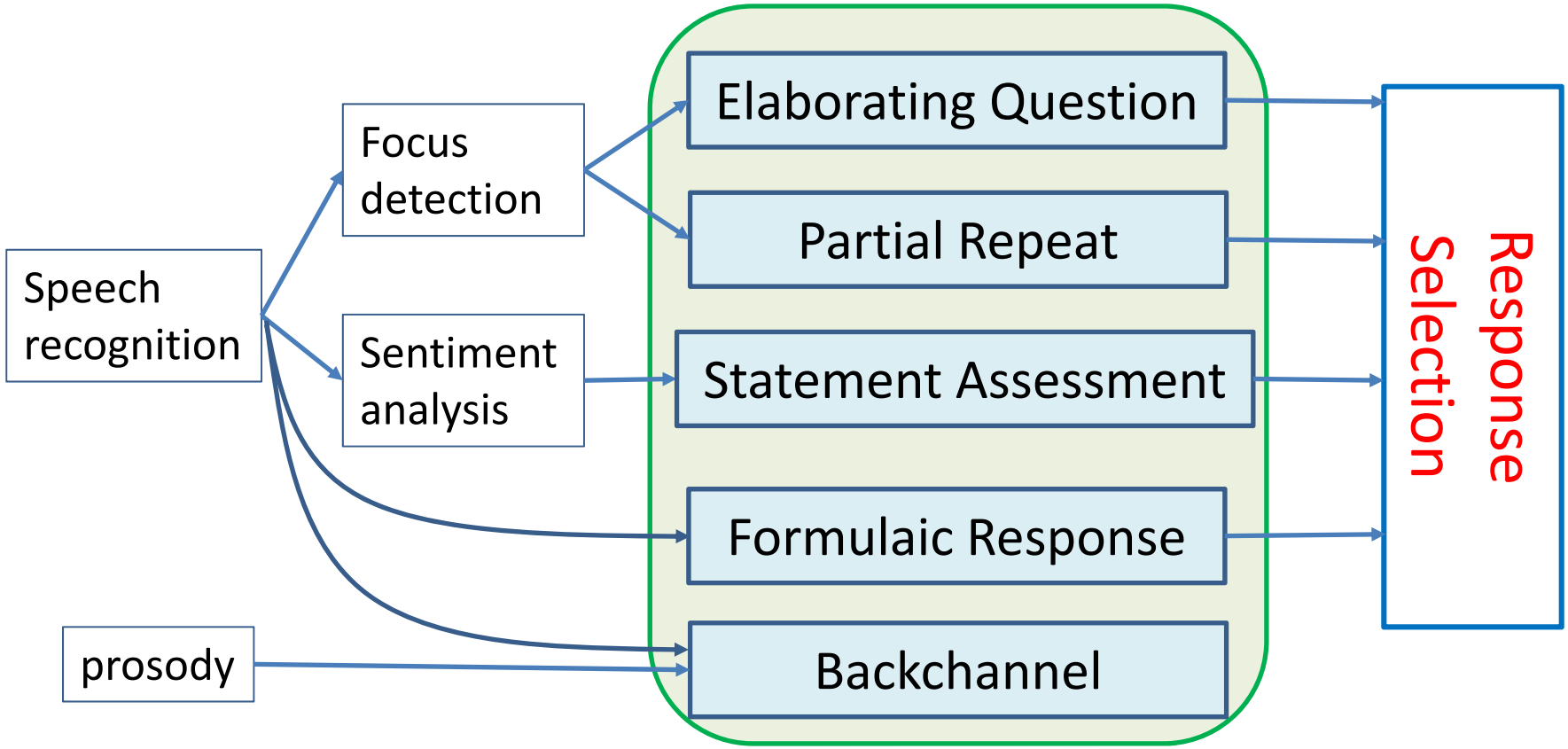
"I went a party." → "That's nice"

"But I was tired." → "That's a pity"

Formulaic Response

- Used as a back-off
  - “I see.”
  - “Really?”
  - “Isn’t it?”
- Function similar to backchannels

Flow of Attentive Listening System



Response Selection among Candidates

- There are many possible responses
- No ground truth (Even the corpus is not ground truth)

"Last Sunday, I went to a high-school reunion."

Formulaic response	"Really?"	○
Assessment	"That's nice"	○
Partial repeat	"High-school reunion?"	○
Elaborating question	"Which reunion?"	×

Not selection problem, but validation problem
(acceptable given linguistic & dialogue context?)

Response Selection among Candidates

- Many possible responses other than corpus occurrence
- Annotated acceptable responses

	Corpus occurrence	Acceptable ratio
Formulaic response	45%	90%
Assessment	21%	60%
Partial repeat	22%	64%
Elaborating question	11%	28%

- Formulaic responses are mostly acceptable.
- Assessments and partial repeats are possible in a majority case.

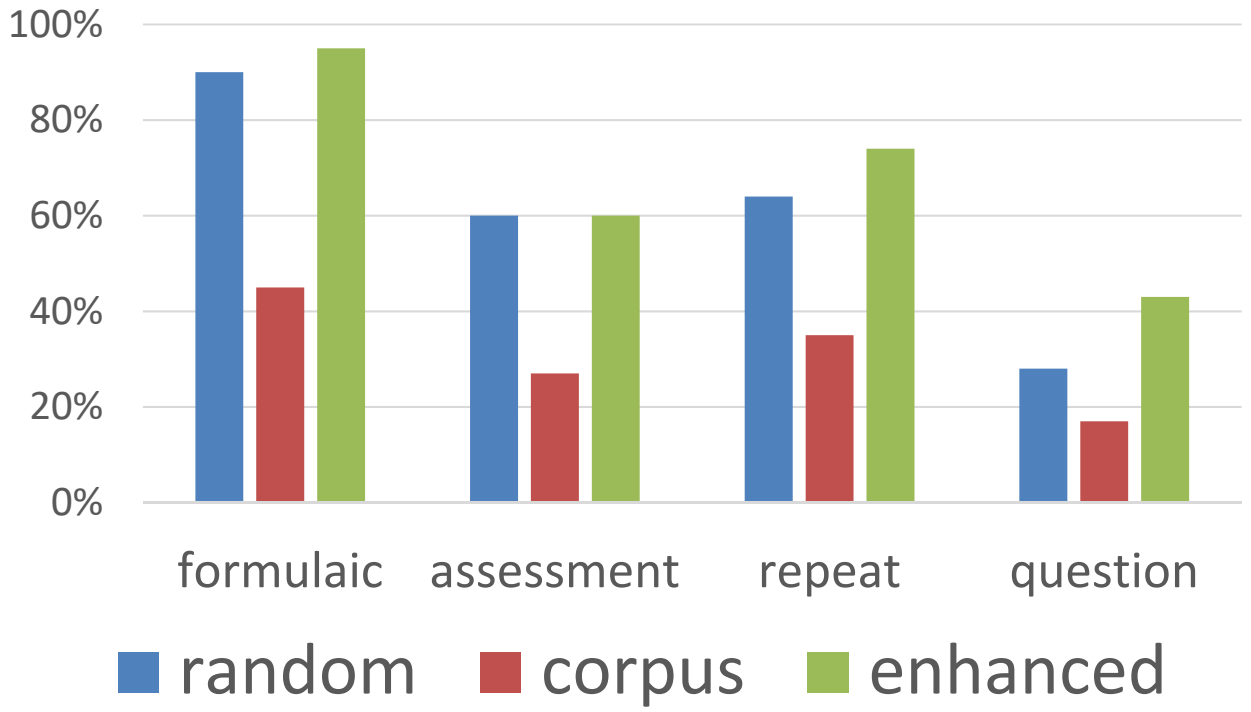
Evaluation of Generated Responses

	recall	precision	F-measure
Formulaic response	99%	91%	0.95
Assessment	51%	73%	0.60
Partial repeat	68%	80%	0.74
Elaborating question	46%	41%	0.43
Weighted average	70%	73%	0.71

- Significantly better than the chance rate
- Still many irrelevant elaborating questions

Comparison with Standard Corpus-based Training

- Randomly generated according to distribution in the corpus
- Training with the corpus occurrence only
- Training with the enhanced annotation of acceptance



Challenge: Total Turing Test of Attentive Listening System

- Almost all senior subjects believed to be talking to ERICA during data collection in WOZ setting.

1. **Can we generate same responses for a corpus collected via WOZ? [objective evaluation]**

   → 70%

1. **Can autonomous ERICA satisfy subjects in a same as WOZ? [subjective evaluation]**

   2.1 Offline video/audio evaluation → ???

   2.2 Online system experience

(Preliminary) Subjective Offline Video/Audio Evaluation

- Video (Audio) prepared by replacing the operator's voice with the system's response 
- Third-party subjects evaluated several questionnaire items, and compared against the baseline

- Overall evaluation is not good (around 0 [-3～3 scale])
  - Precision of 70% is not sufficient.
    - Irrelevant questions and assessments give bad impression.
  - Response is monotonous.
  - TTS and turn-taking is not natural enough?
  - No backchannels in this experiment!!

Conclusions & Practical Issues

- Considering arbitrary nature is important
- Enhanced annotation requires much effort
- Machine learning gives some improvement
- 70% in recall & precision
- But the system is not yet in a satisfactory level

Generation of Backchannels

Non-lexical utterances

--“Voice” beyond “Speech”--

- Continuer Backchannels: “うん”
  - listening, understanding, agreeing to the speaker
- Assessment Backchannels: “はー”、“ふーん”
  - Surprise, interest and empathy
- Fillers: “あのー”、“えーと”
  - Attention, politeness
- Laughter
  - Funny
  - Socializing
  - Self-pity

Backchannels (BC)

- 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) ← Many previous works
  - Usually at the end of speaker's utterances
  - Should predict before end-point detection
- Lexical form (what)
  - Machine learning using prosodic and linguistic features [Interspeech16]
- Prosody (how)
  - Adjust according to preceding user utterance [IWSDS15]

  - **Many systems use same recorded pattern**
  - **giving monotonous impression to users**

Categories and Occurrence Counts of Backchannels

category	occurrence at IPU (clause) boundaries
Un うん	12% (10%)
Un **x2** うんうん	7% (9%)
Un **x3** うんうんうん	13% (19%)
Assessments	8% (14%)
None	60% (47%)

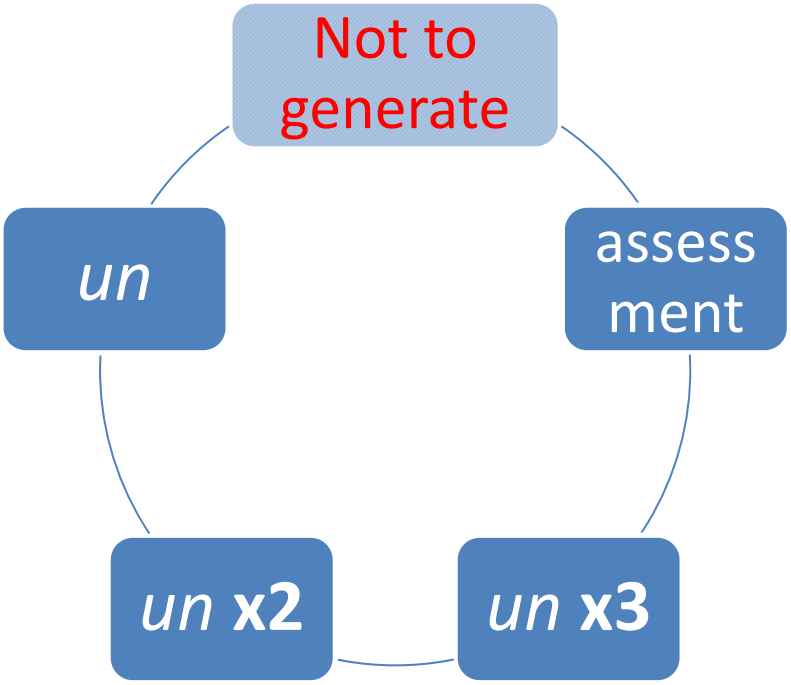
Backchannels are observed at 40% of IPUs with different forms in a good balance

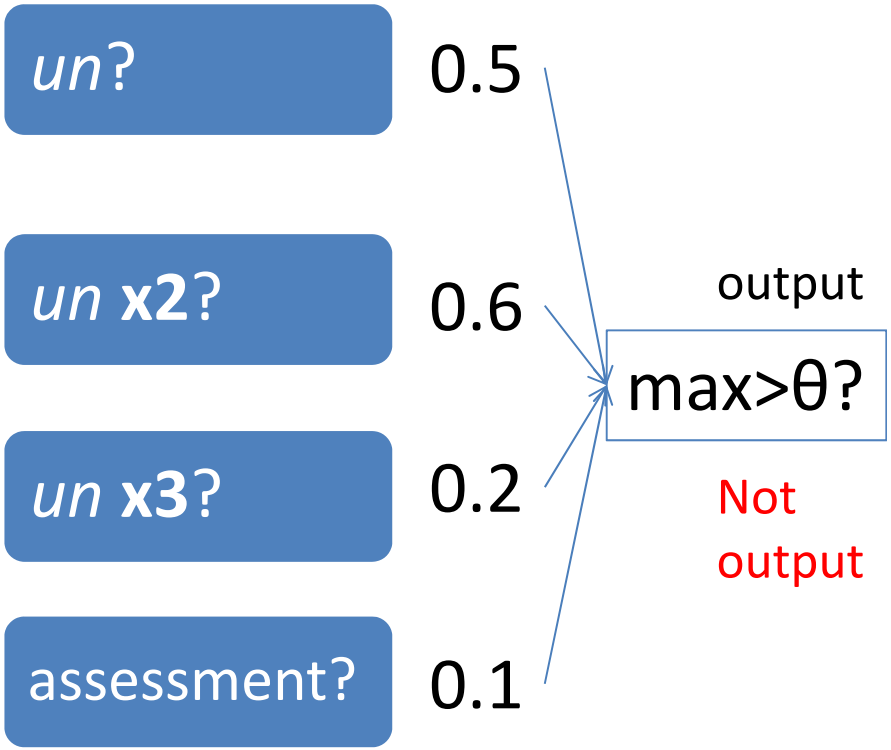
Additional Annotation of Backchannels

- Generation of backchannels and choice of their form are arbitrary
- Evaluation with exactly observed patterns may not be meaningful

- Augment the annotation
  - Three human annotators judge which backchannel forms are acceptable, given dialogue context
  - Accept only when ALL three annotators agree
  - The added forms are regarded as correct in evaluation

Selection Problem → Validation Problem





Prediction Performance by using Linguistic and Prosodic Features

Category	Recall	Precision	F-measure
un	0.311	0.657	0.422
un **x2**	0.382	0.820	0.521
un **x3**	0.672	0.333	0.454
assessments	0.467	0.342	0.405
not-to-generate	0.775	0.769	0.772
average	0.643	0.643	0.643

- Precision of simple continuers (*un*, *un* **x2**) is very high because they are acceptable in many cases.
- Reasonable performance for "not-to-generate" decision.

Subjective Offline Evaluation of Generated Backchannels

- Voice files of backchannels (one for each category) recorded by a voice actress (for TTS)
- Audio channel of counselors replaced by the generated backchannels

- 9 subjects listened 8 segments of dialogue, and evaluated on 7 items with 7-point scales
- Compared with
  - Weighted random generation
  - Counselors' choice (voice replaced)

Subjective Evaluation of Backchannels

	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
  - Tuning of precise timing is also needed

Challenge: Total Turing Test of Backchanneling System

1. **Can we generate same responses for a corpus collected via WOZ? [objective evaluation]**

   → 64%

1. **Can autonomous ERICA satisfy subjects in a same level as WOZ? [subjective evaluation]**

   2.1 Offline video/audio evaluation

   → ○ backchannel forms

   → × prosody & precise timing

   2.2 Online system experience

   → ??? (demo)

Generating Fillers

IWSDS18

- No filler 
- Filler before moving to next question  

Demonstration of Attentive Listening System

Current Lessons Learned

- Backchannels are effective, but proper precise timing is critical. (<200ms)

- Repeat of named entities is effective for showing understanding, but vulnerable to ASR errors.
- Proper assessment is expected at the end of talk, but often difficult.
  - People want to share their joy/sadness
- When above two works, dialogue is engaging.

Job Interview System

Job Interview

- Interview is an essential process in hiring persons and accepting (graduate) students
- Purpose
  - Check communication skill (when inclined to hire)
  - Find something special (when uncertain to hire)
- Face-to-face is norm
- Currently,
  - Students (and Companies) spend a lot in rehearsal and preparation

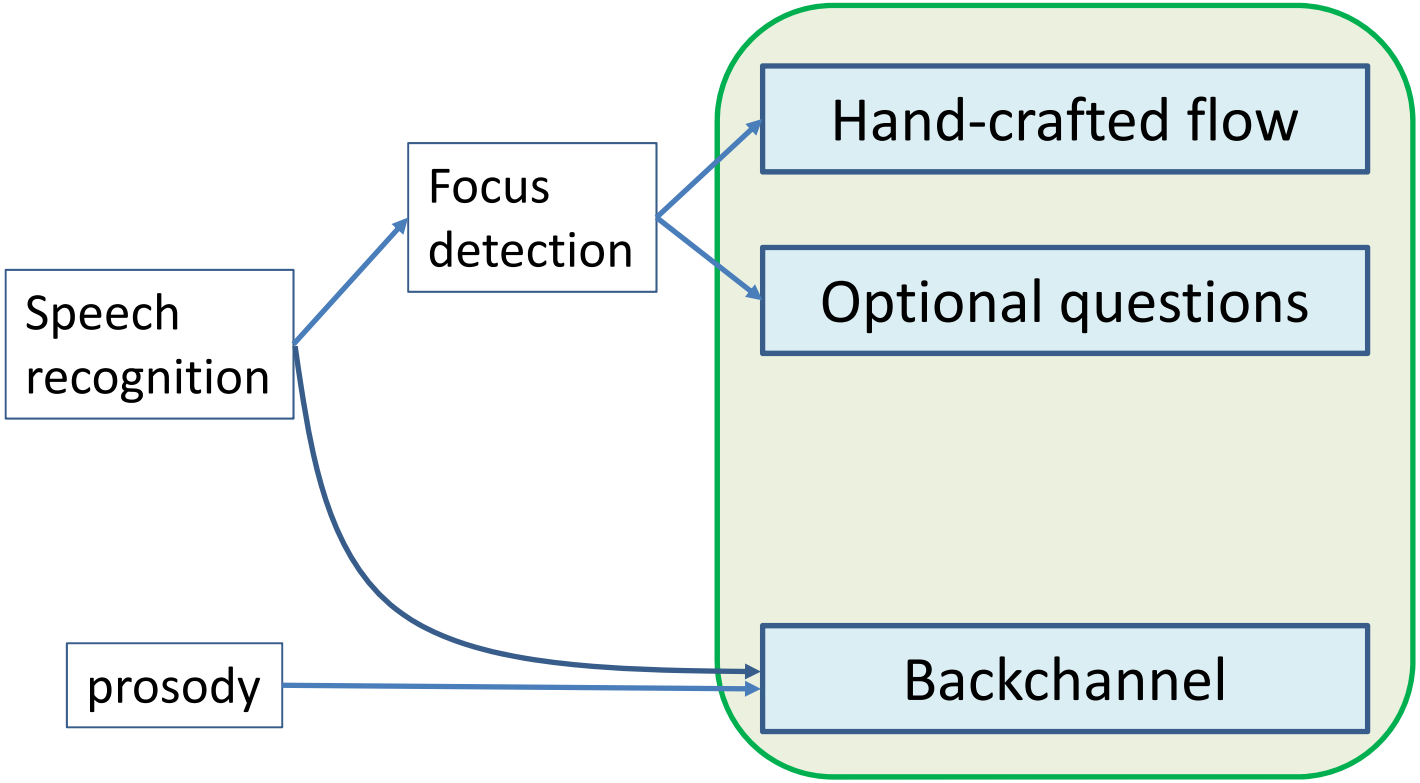
Challenge: Total Turing Test of Job Interview System

- Can robot be an interviewer?

- Some Japanese companies are introducing robots for interview in the initial stage
  - But mostly based on prepared question scenario
  - Interviewee can easily prepare (rehearse) well

1. **Can we generate adaptive (non-scenario-based) questions? [corpus-based evaluation]**
2. **Can autonomous ERICA make subjects feel like real interview? [subjective evaluation]**

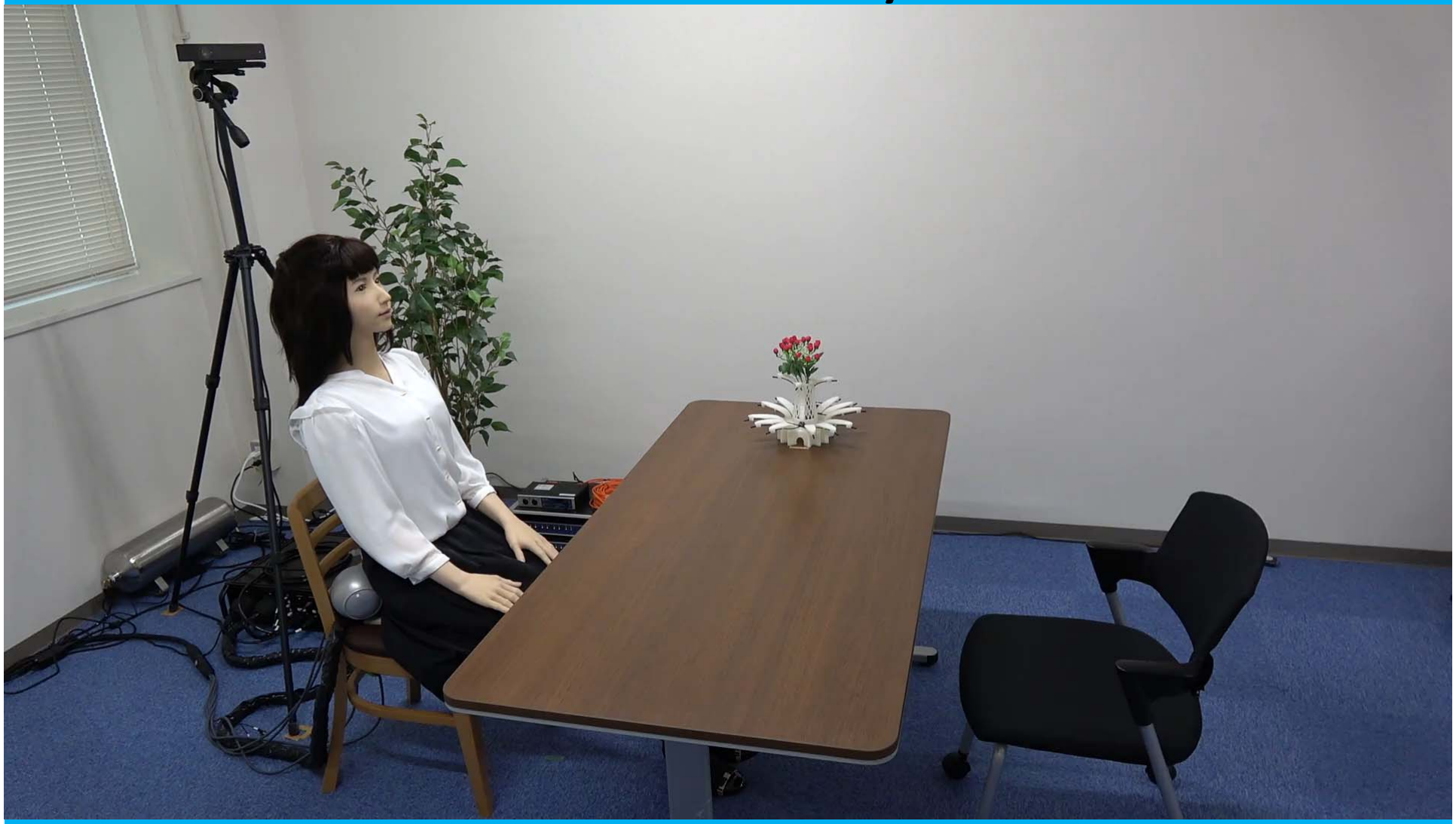
Flow of Job Interview System



Current Implementation

- Flow of basic questions
  - Motivation for application
  - strong/weak points of the interviewee…
- Optional additional questions
  - “Why our company instead of other companies?”
  - “Can you tell me a specific example?”
- Selection of optional questions
  - Machine learning is difficult
  - Heuristics based on duration of turns

Demonstration of Job Interview System



Other Topics

Flexible Turn-taking

- Natural ← push-to-talk, magic words
  - TRP predictor (pause / prosody)
- Fuzzy decision ← Binary decision
  - Use fillers and backchannels when ambiguous
  - TTS output cannot be stopped

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

Non-verbal information

- Valence Recognition
  - Positive/negative feeling on what is talked about

  → proper assessment (including prosody) in attentive mode

- Engagement Recognition

  IWSDS18
  - Positive/negative attitude to keep the current dialogue

  → change topics, turn-taking behaviors,
  manner of system reply (including prosody)

- Ice-breaking
  - Rapport with the first-comer

  → Switch dialogue to main topic

Character Modeling → Desire

- Attentive / Inattentive
- Extrovert / Introvert
- Polite / Casual

(cf.) Big Five

Myers-Briggs Type Indicator (MBTI)

Evaluation Criteria

- Total Turing Test (Level 1)
  - Comparable to WOZ setting

← **Current Work**

- Total Turing Test (Level 2)
  - Comparable to “human-like interaction experience”
  - measured by Engagement level

References

1. D.Lala, P.Milhorat, K.Inoue, M.Ishida, K.Takanashi, and T.Kawahara.
   Attentive listening system with backchanneling, response generation and flexible turn-taking.
   In Proc. SIGdial 2017.
2. P.Milhorat, D.Lala, K.Inoue, Z.Tianyu, M.Ishida, K.Takanashi, S.Nakamura, and T.Kawahara.
   A conversational dialogue manager for the humanoid robot ERICA.
   In Proc. IWSDS 2017.
3. T.Kawahara, T.Yamaguchi, K.Inoue, K.Takanashi, and N.Ward.
   Prediction and generation of backchannel form for attentive listening systems.
   In Proc. INTERSPEECH 2016.
4. R.Nakanishi, K.Inoue, S.Nakamura, K.Takanashi, and T.Kawahara.
   Generating fillers based on dialog act pairs for smooth turn-taking by humanoid robot.
   In Proc. IWSDS2018.
5. K.Inoue, D.Lala, K.Takanashi, and T.Kawahara.
   Latent character model for engagement recognition based on multimodal behaviors.
   In Proc. IWSDS2018.