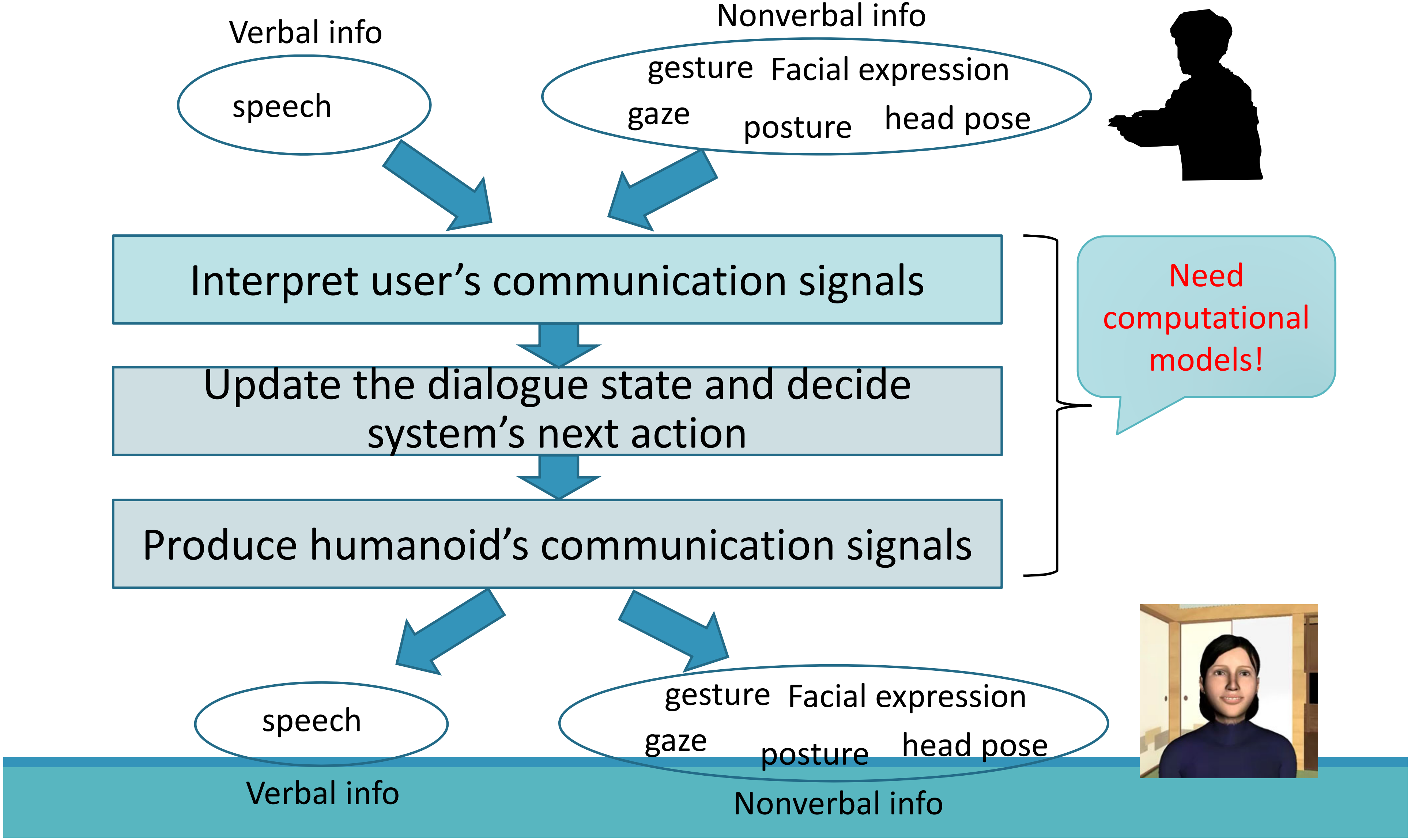
Multimodal Behavior Generation

Yukiko Nakano

Introduction

Multimodal Dialogue System Architecture



Verbal info

speech

Nonverbal info

gesture Facial expression gaze posture head pose

Interpret user's communication signals

Update the dialogue state and decide system's next action

Produce humanoid's communication signals

Need computational models!

speech

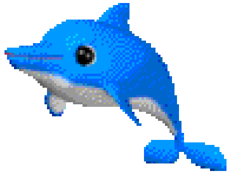
Verbal info

gesture Facial expression gaze posture head pose

Nonverbal info

Conversational agent

- Animation characters or robots that can display humanlike bodily expressions (facial expression, gesture, etc) synchronized with speech
- Autonomous agent: response is decided by the system
  - Embodied conversational agent
  - Believable agent
- Avatar: human user type-in/select response

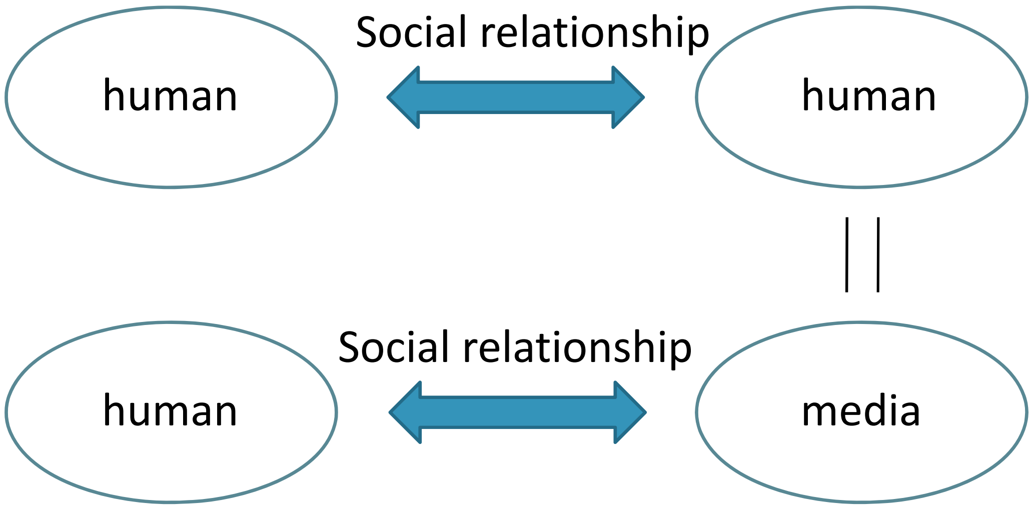


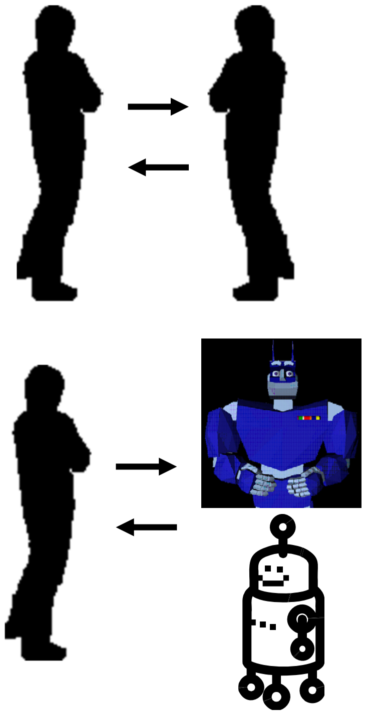
Benefits and advantages of conversational humanoids

- Intuitiveness
  - Users can use a computer by talking to a computer like in face-to-face conversation.
  - Users do not need to learn how to use the interface (manual free)
- Robustness
  - Multimodality contribute to decrease communication failure and increase the robustness of communication
- Naturalness
  - Users tend to treat computers as human (Media equation [Reeves&Nass, 1996)

Media equation

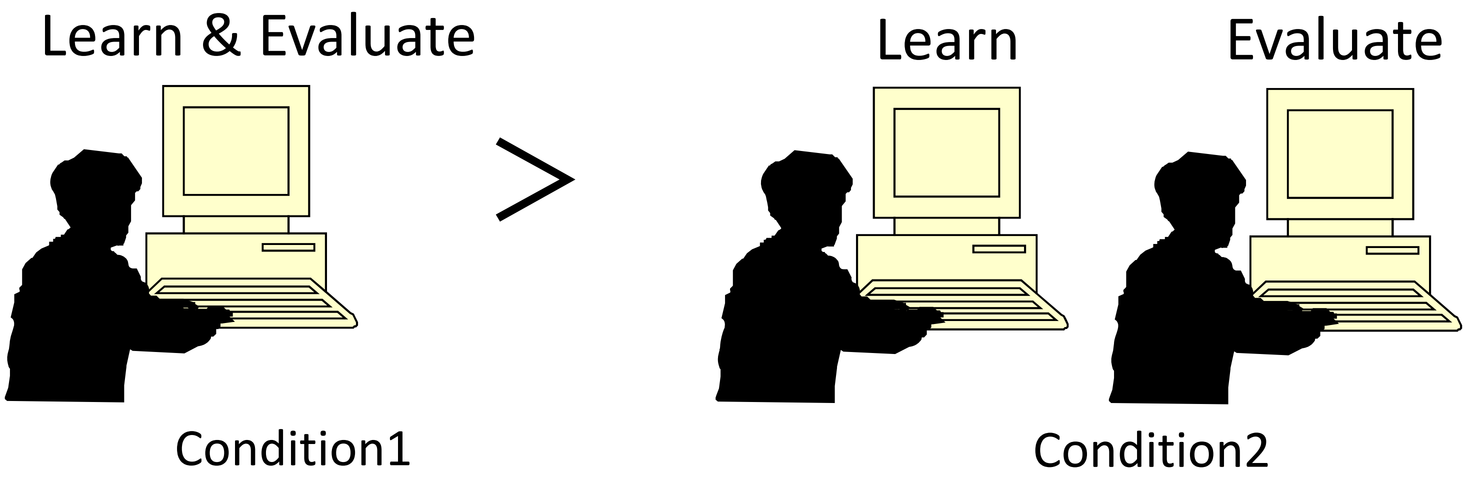
- Media Equation, by Reeves & Nass (1996)
  - Computers Are Social Actors (CASA) paradigm: humans mindlessly apply the same social heuristics used for human interactions to computers and media





Experiment for media equation

- Politeness in human-computer interaction
  - The participants learn about American culture from a computer
  - （Condition1）：Evaluate the system using the same computer
  - （Condition2）：Evaluate the system using a different computer
  - Condition 1 participants gave more positive feedback than Condition 2 participants



Learn & Evaluate

>

Learn

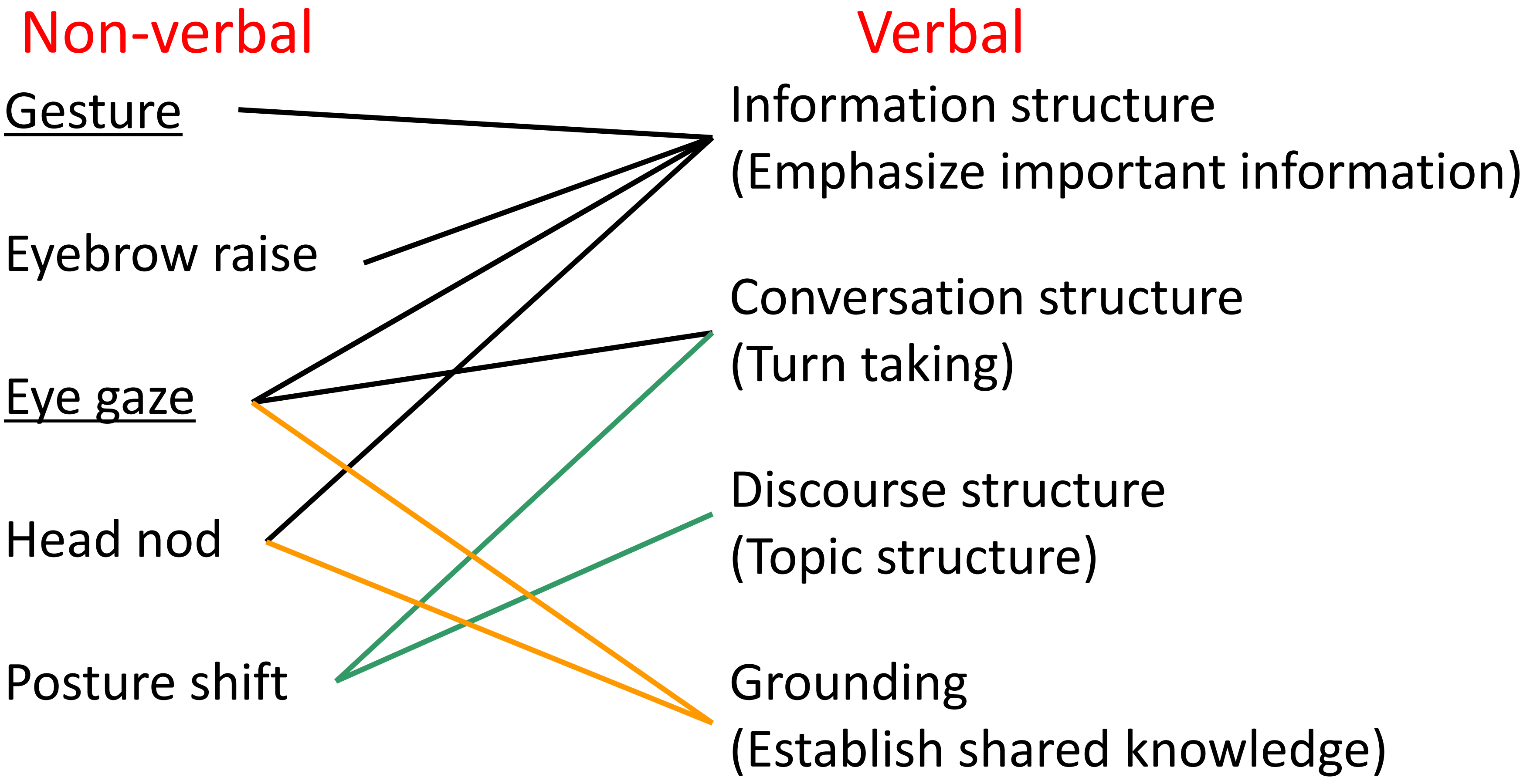
Evaluate

Condition1

Condition2

Behavior Generation

Relationship between Linguistic Structure and Behavioral Cues



Non-verbal

Gesture

Eyebrow raise

Eye gaze

Head nod

Posture shift

Verbal

Information structure
(Emphasize important information)

Conversation structure
(Turn taking)

Discourse structure
(Topic structure)

Grounding
(Establish shared knowledge)

Nonverbal communication signal

Function	Nonverbal behaviors
(1) emphasize utterance	Gesture, eyebrow raise
(2) Give a turn	Stop gesture, gaze at next speaker, mutual gaze, next speaker looks away from the current speaker
(3) Feedback to the speaker	Mutual gaze, acknowledgement
(4) Change topic	Change posture

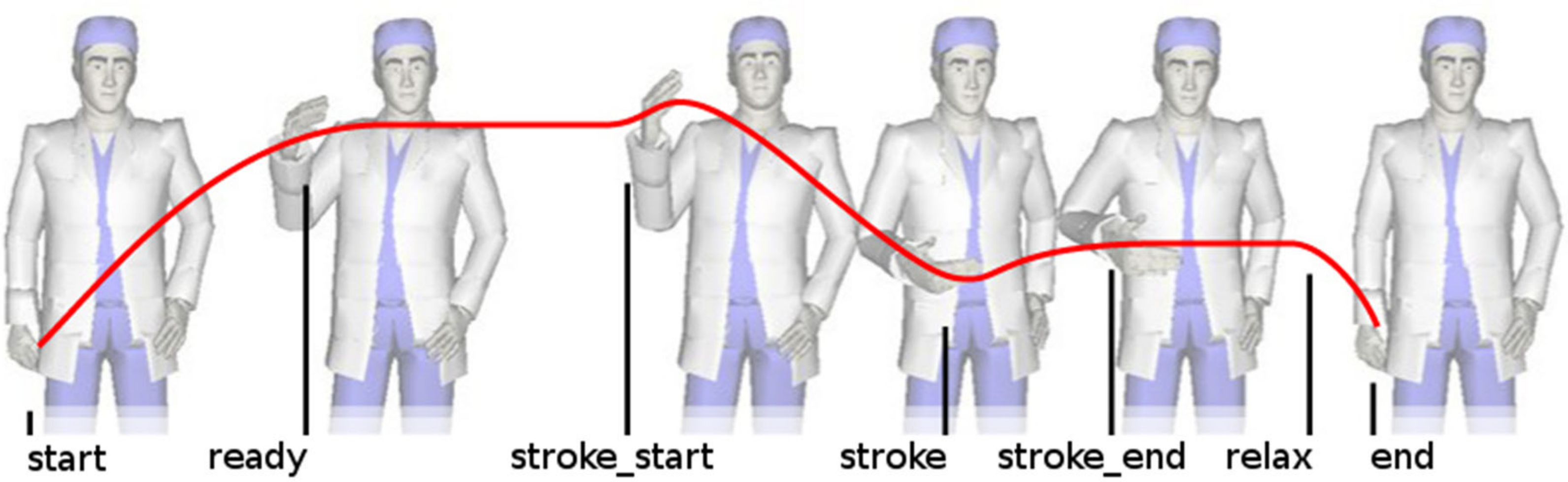
Facial expression as communication signal

Function		Facial expression
Syntactic function	Emphasize utterance	Eyebrow raise, blink
	Syntactic structure	Eyebrow raise, blink
	Change topic	Eyebrow raise, blink
Semantic function	Complement linguistic expression	Nod, shake, emblem
Conversation coordination function	Back-channel	Nod, smile
	Turn taking	gaze
emotion expression	Complement linguistic expression, reinforce semantics, Speaker's opinion/ evaluation feedback	Facial expression: happy, sad, surprise, disgust, etc

We need to make agent understand/generate these communication signals

What should be decided?

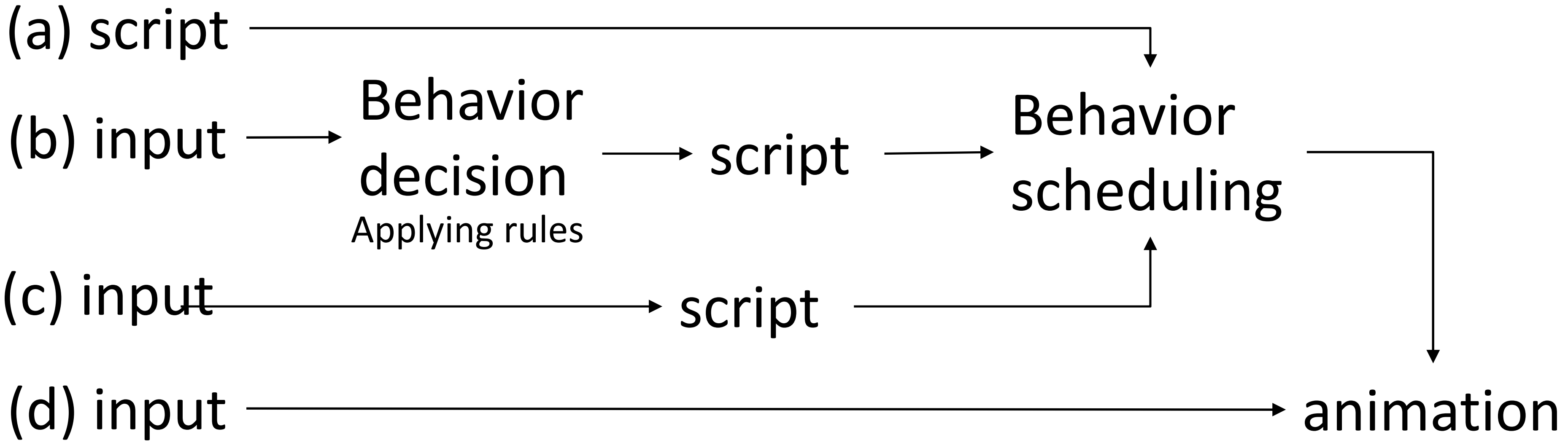
- When?
  - With what words should nonverbal behaviors be co-occurred?
- What?
  - What type of behaviors should be performed?



Approaches

- (a) Manually generated script
- (b) Rule-based
- (c) Behavior prediction
- (d) Joint position prediction

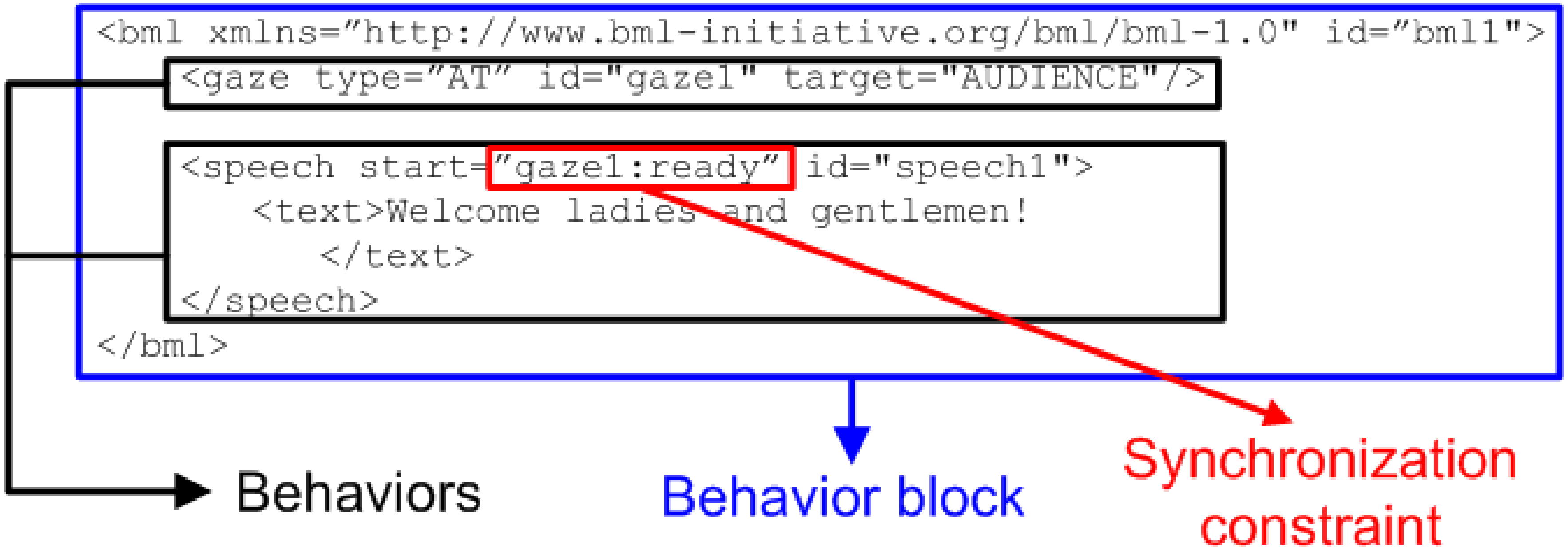
Data-driven/machine learning



Script

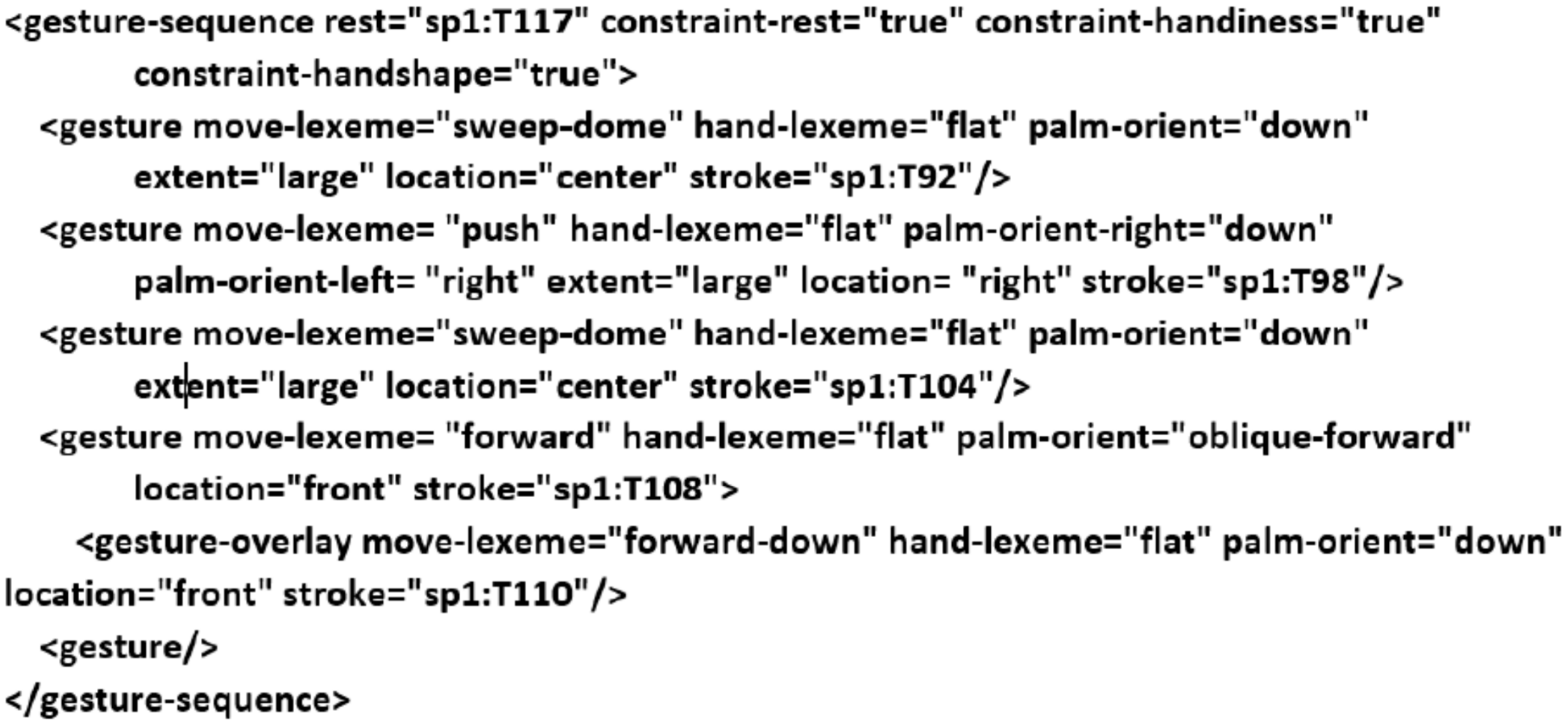
- Describe agent’s behaviors using markup language
  - text
  - gesture
  - facial expression
  - gaze
  - background image, etc
- Markup languages
  - BML, FML
  - MPML
  - (Microsoft MS agent)

BML example (1)



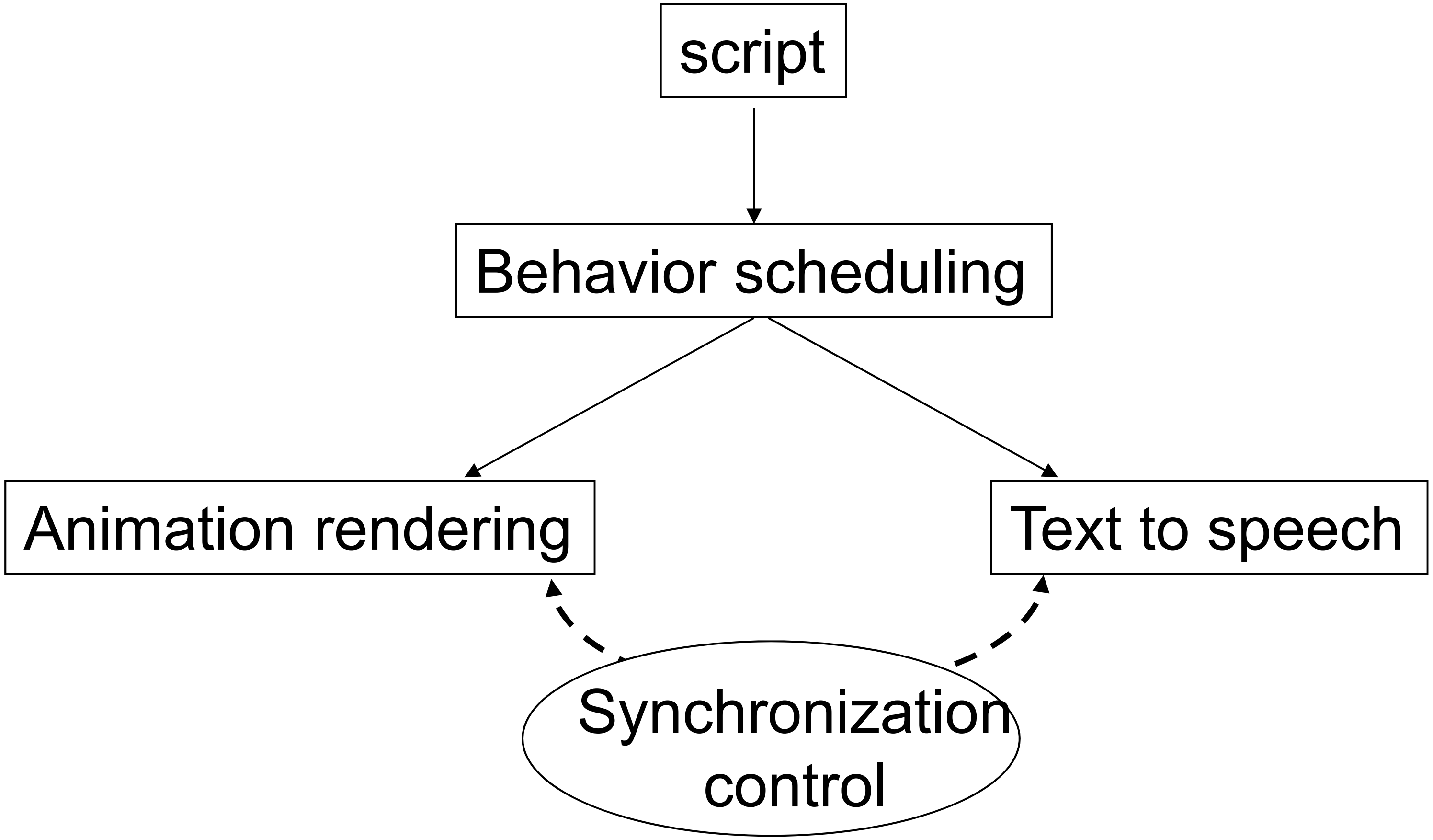
```
<bml xmlns=”http://www.bml-initiative.org/bml/bml-1.0" id=”bml1">
   <gaze type=”AT” id="gaze1" target="AUDIENCE"/>

   <speech start=”gaze1:ready” id="speech1">
      <text>Welcome ladies and gentlemen!
         </text>
   </speech>
</bml>
```

BML example (2)

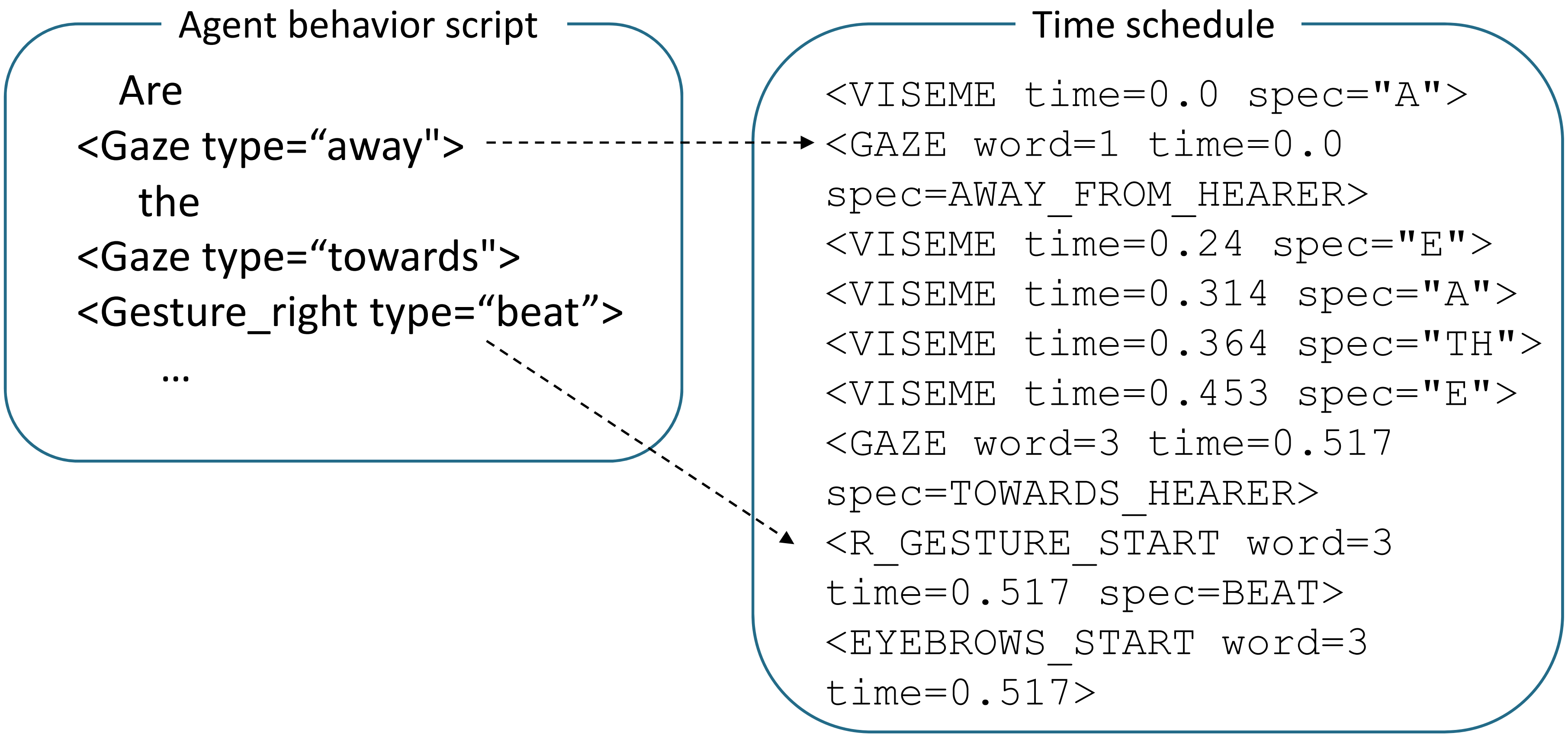


```
<gesture-sequence rest="sp1:T117" constraint-rest="true" constraint-handiness="true"
        constraint-handshape="true">
  <gesture move-lexeme="sweep-dome" hand-lexeme="flat" palm-orient="down"
        extent="large" location="center" stroke="sp1:T92"/>
  <gesture move-lexeme= "push" hand-lexeme="flat" palm-orient-right="down"
        palm-orient-left= "right" extent="large" location= "right" stroke="sp1:T98"/>
  <gesture move-lexeme="sweep-dome" hand-lexeme="flat" palm-orient="down"
        extent="large" location="center" stroke="sp1:T104"/>
  <gesture move-lexeme= "forward" hand-lexeme="flat" palm-orient="oblique-forward"
        location="front" stroke="sp1:T108">
    <gesture-overlay move-lexeme="forward-down" hand-lexeme="flat" palm-orient="down"
location="front" stroke="sp1:T110"/>
  <gesture/>
</gesture-sequence>
```

Realization process



Script to behavior schedule



Agent behavior script

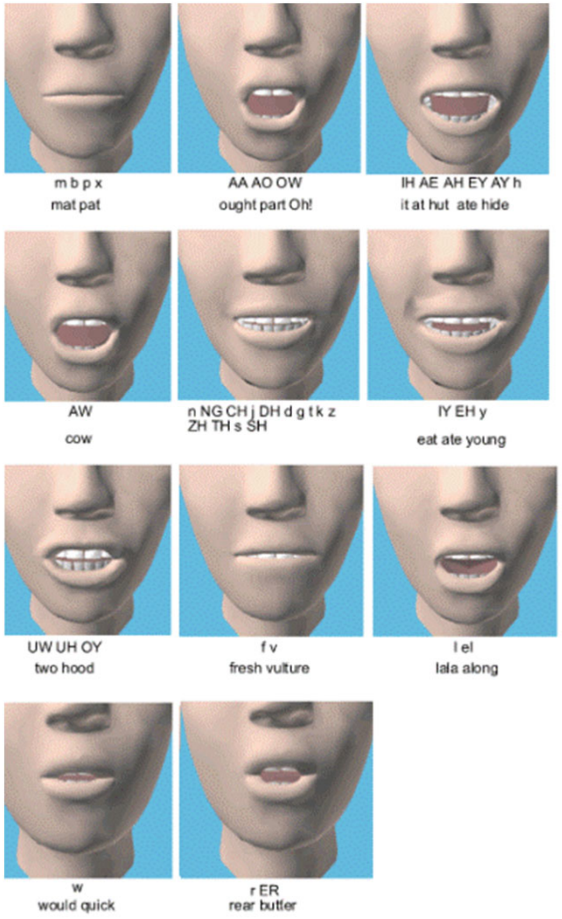
```
   Are
<Gaze type=“away">
    the
<Gaze type=“towards">
<Gesture_right type=“beat”>
     …
```

Time schedule

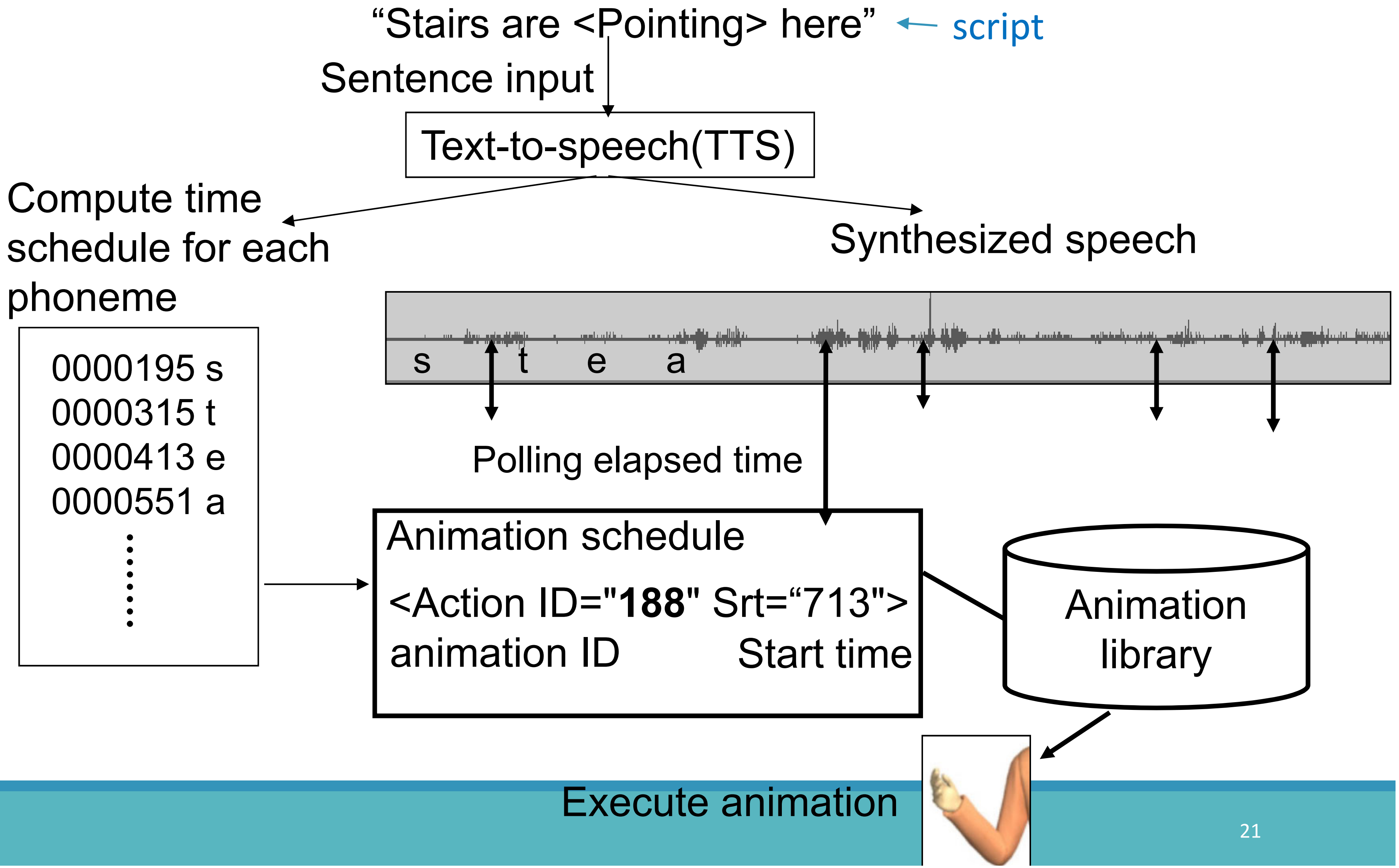
```
<VISEME time=0.0 spec="A">
<GAZE word=1 time=0.0
spec=AWAY_FROM_HEARER>
<VISEME time=0.24 spec="E">
<VISEME time=0.314 spec="A">
<VISEME time=0.364 spec="TH">
<VISEME time=0.453 spec="E">
<GAZE word=3 time=0.517
spec=TOWARDS_HEARER>
<R_GESTURE_START word=3
time=0.517 spec=BEAT>
<EYEBROWS_START word=3
time=0.517>
```

Lipsync and viseme

- Solution 1: use lipsync functionality provided by the animation engine
  - Unity
- Solution 2: implement by yourself
  - Get phoneme timing (and viseme) from TTS engine (e.g., Microsoft speech API (SAPI))
  - Execute viseme animation at the right timing by implementing a timer



Synchronization between speech and animation



"Stairs are <Pointing> here" ← script

Sentence input

Text-to-speech(TTS)

Compute time schedule for each phoneme

Synthesized speech

0000195 s
0000315 t
0000413 e
0000551 a
⋮

s t e a

Polling elapsed time

Animation schedule

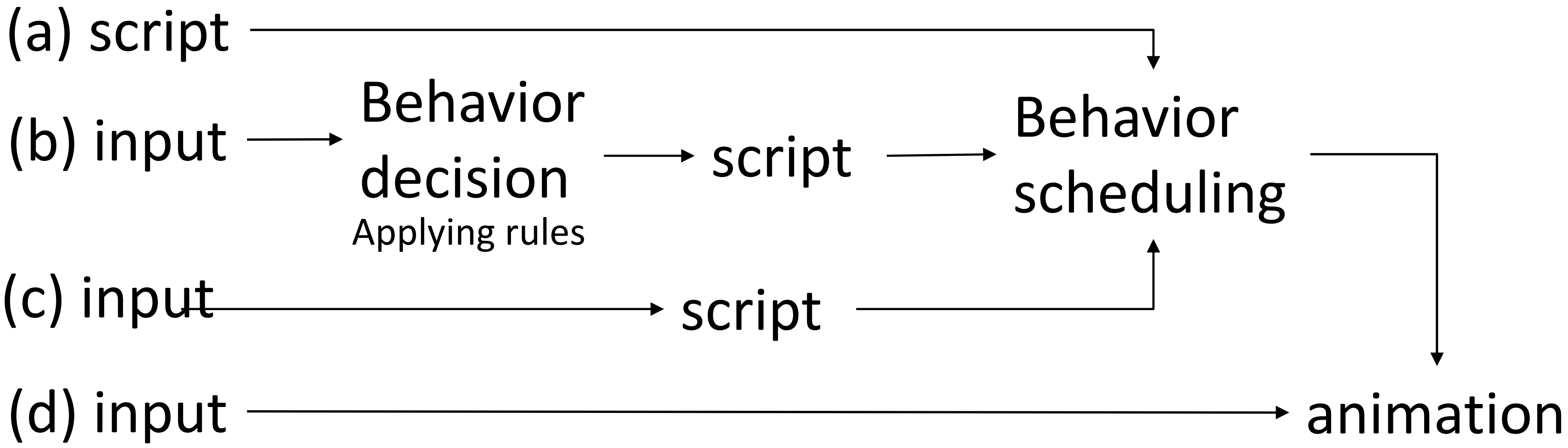
<Action ID="**188**" Srt="713">
animation ID Start time

Animation library

Execute animation

Approaches

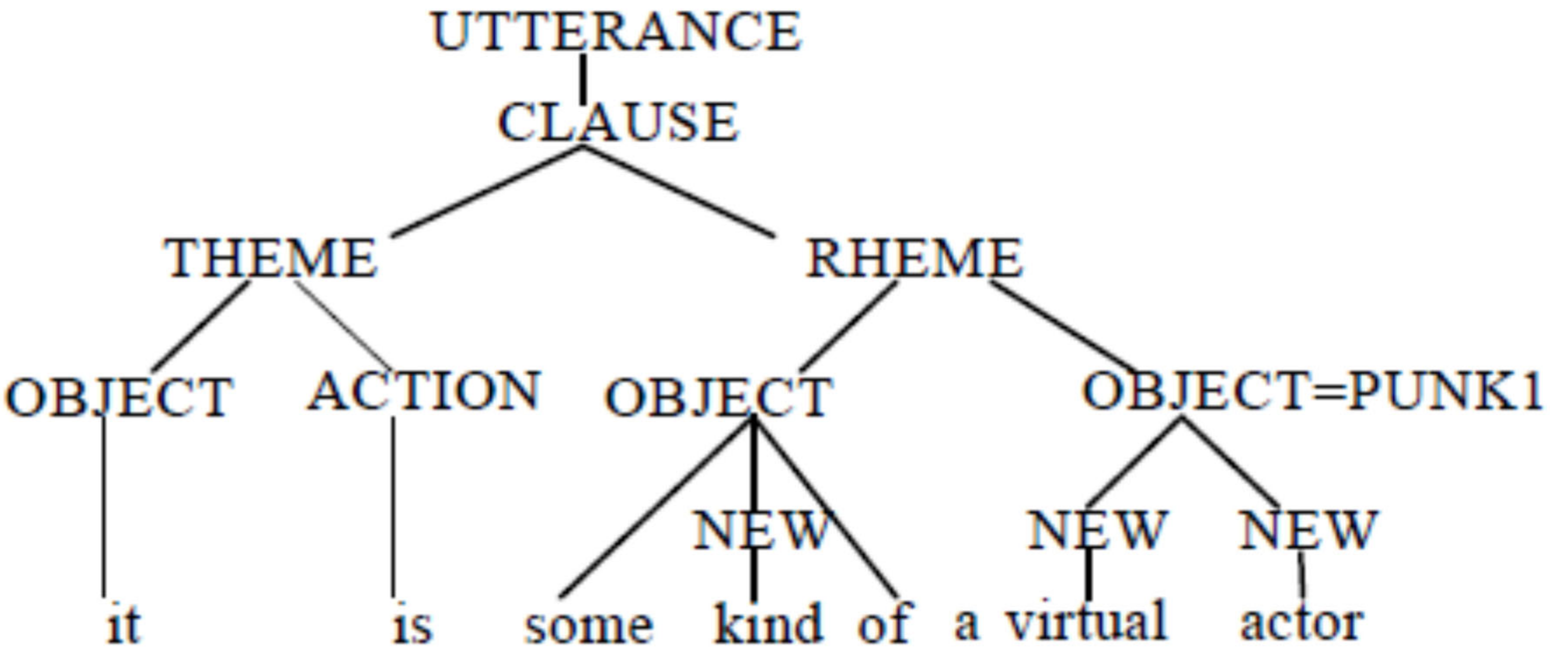
- (a) Manually generated script
- (b) Rule-based
- (c) Behavior prediction
- (d)) Joint position prediction



Rule-based approach

- BEAT (Cassell et al 2001)
- Automatically Generate agent nonverbal behaviors from text input
- Approach
  - Define gesture decision rules based on the findings in previous nonverbal communication studies
  - Analyze linguistic information in the text
  - Apply the rule to the linguistic information to generate a script
  - Produce animation synchronized with speech from the script

Example of linguistic analysis



Rule examples

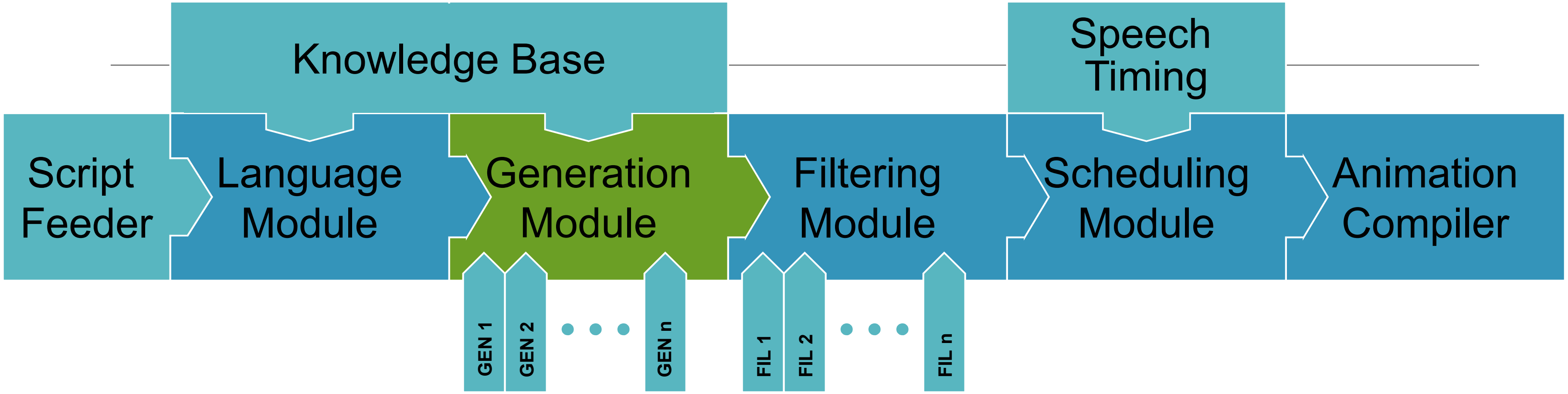
Gesture rule

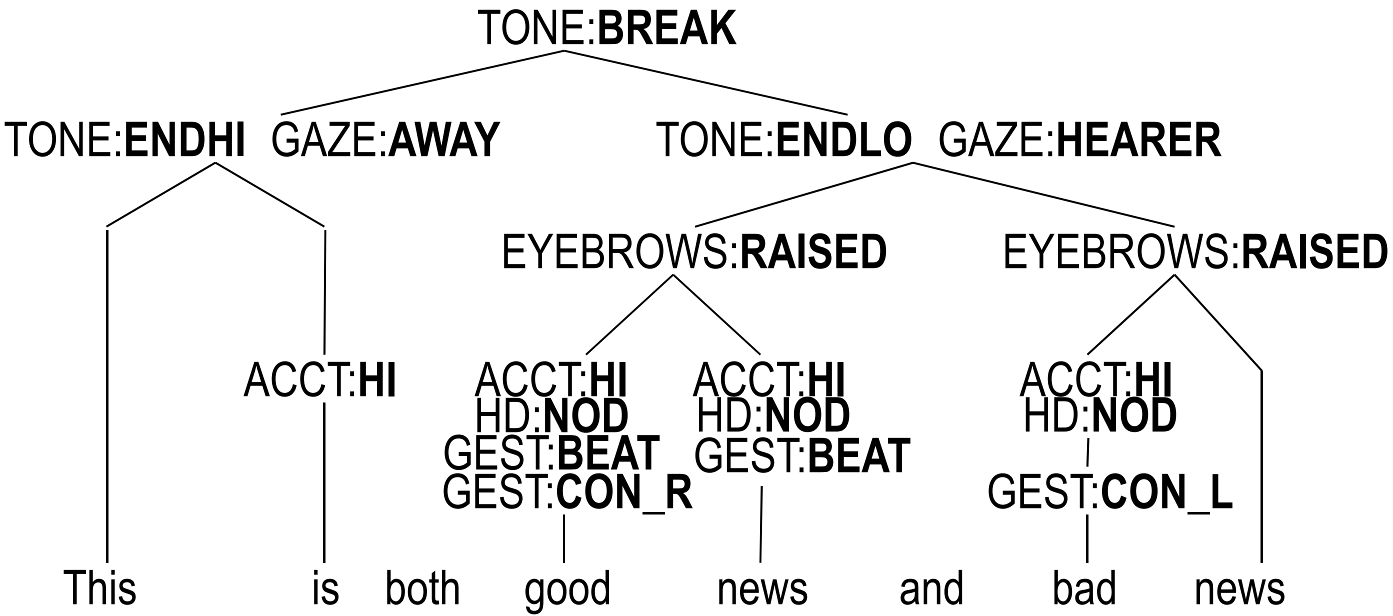
```
FOR each RHEME node in the tree
    IF the RHEME node contains at least one NEW node
    THEN Suggest a BEAT to coincide with the OBJECT phrase
```

Gaze rule

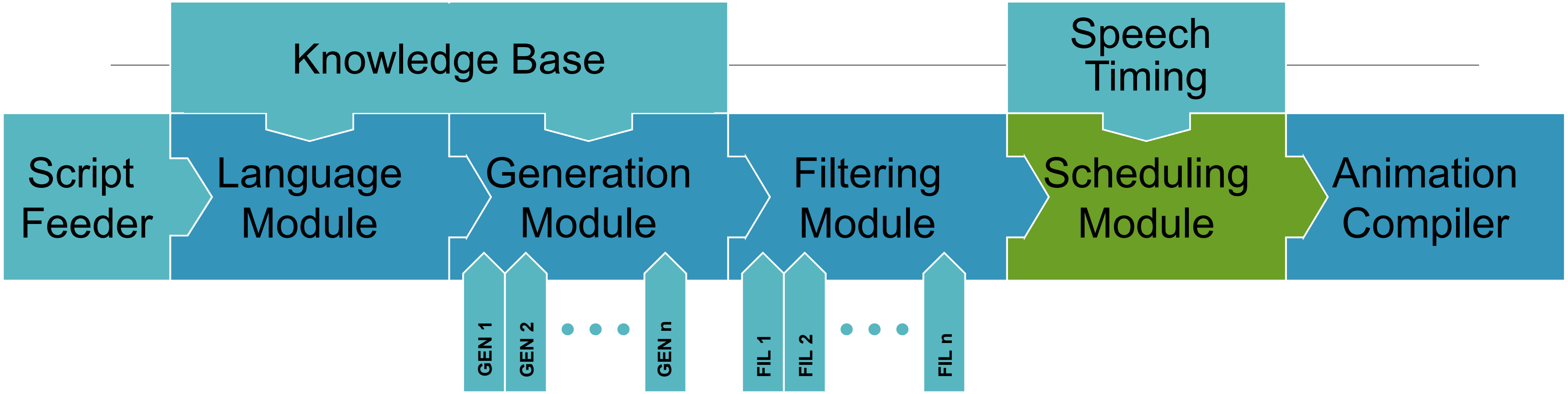
```
FOR each THEME
    IF at beginning of utterance OR 70% of the time
    Suggest Gazing AWAY from user
FOR each RHEME
    If at end of utterance OR 73% of the time
    Suggest Gazing TOWARDS the user
```

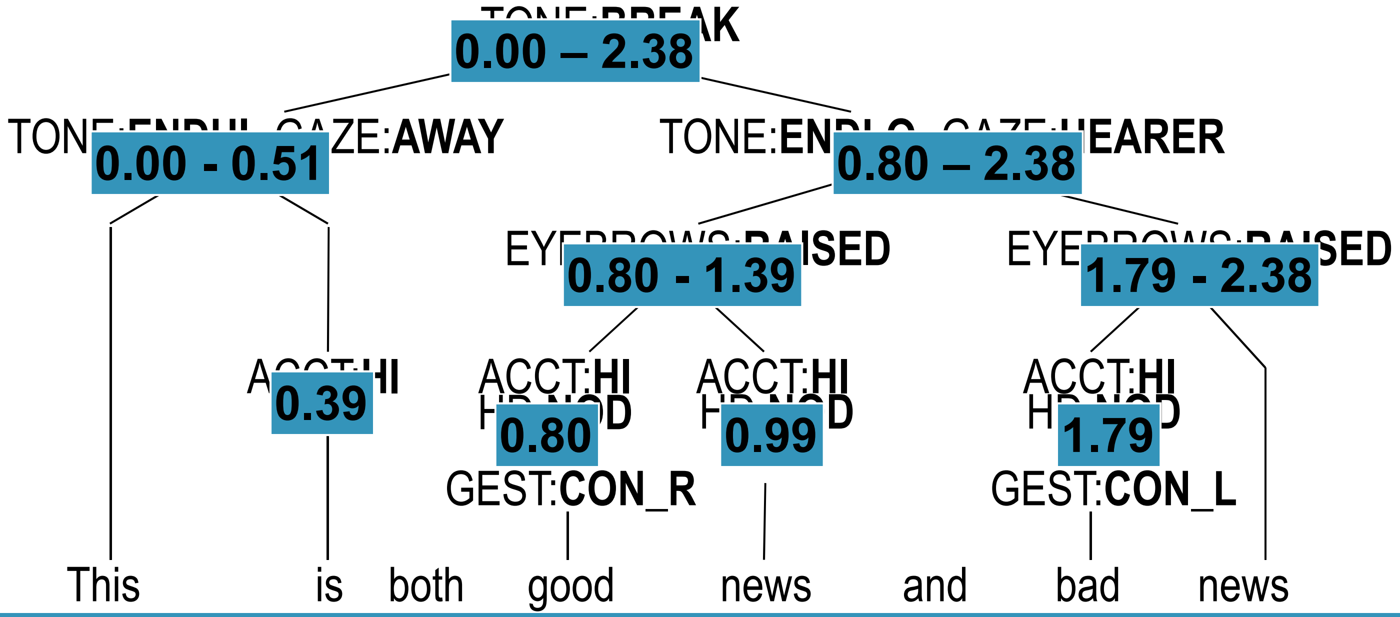
Processing: Behavior Generation



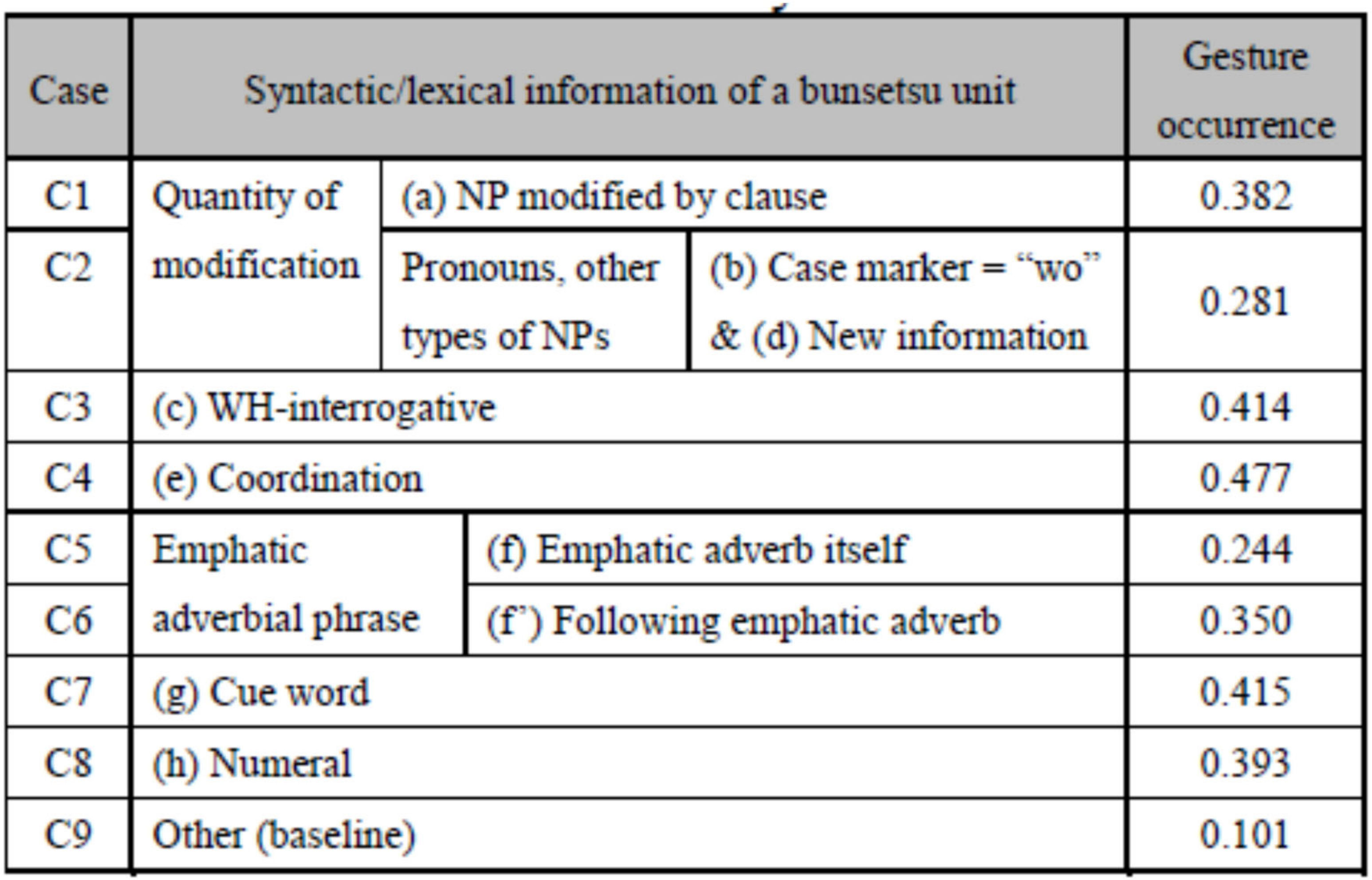

TONE:**BREAK**
TONE:**ENDHI** GAZE:**AWAY**
TONE:**ENDLO** GAZE:**HEARER**
EYEBROWS:**RAISED**
EYEBROWS:**RAISED**
ACCT:**HI**
ACCT:**HI** HD:**NOD** GEST:**BEAT** GEST:**CON_R**
ACCT:**HI** HD:**NOD** GEST:**BEAT**
ACCT:**HI** HD:**NOD** GEST:**CON_L**
This
is
both
good
news
and
bad
news

Processing: Behavior Scheduling



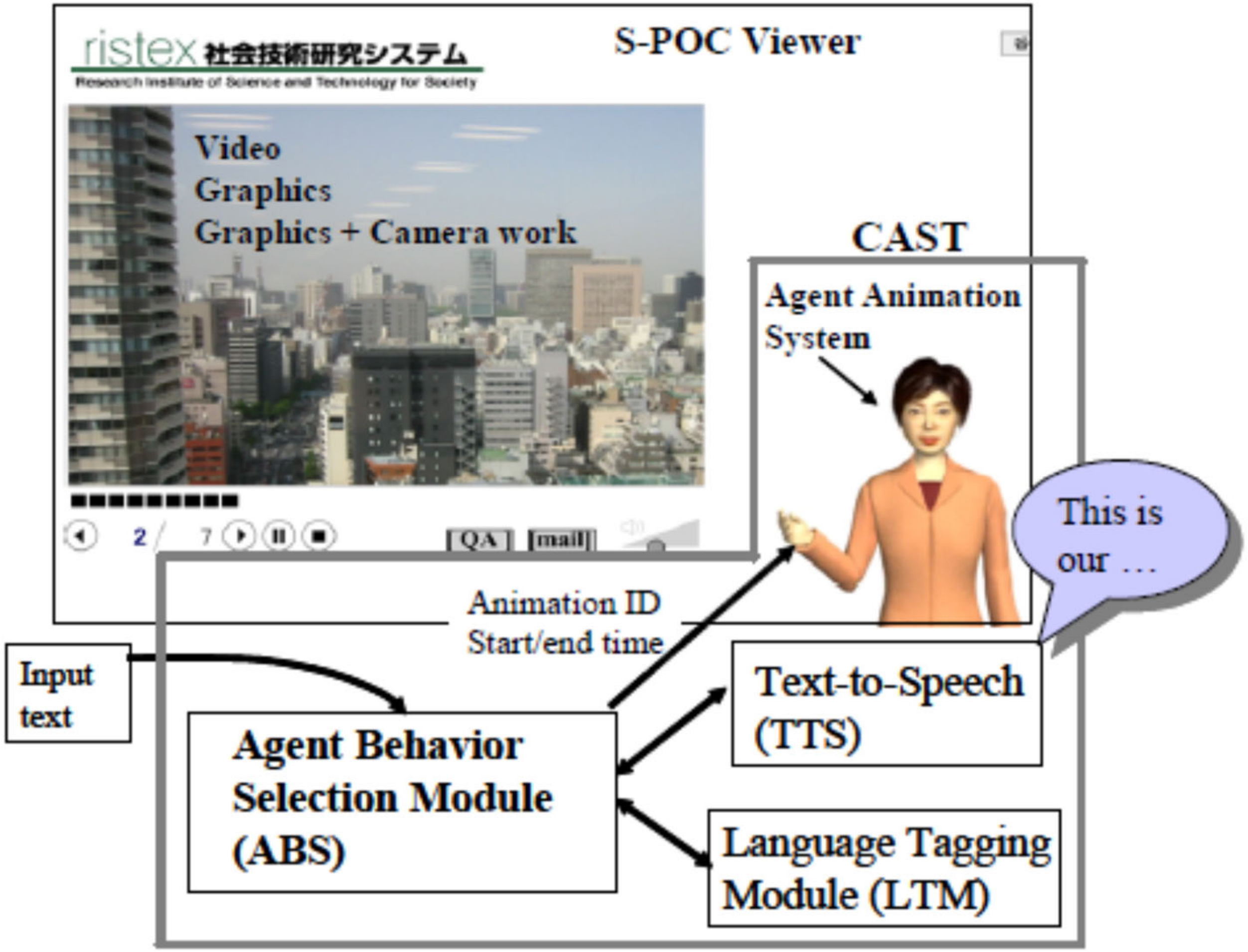


Gesture distribution in Japanese presentation



Case	Syntactic/lexical information of a bunsetsu unit			Gesture occurrence
C1	Quantity of modification	(a) NP modified by clause		0.382
C2		Pronouns, other types of NPs	(b) Case marker = “wo” & (d) New information	0.281
C3	(c) WH-interrogative			0.414
C4	(e) Coordination			0.477
C5	Emphatic adverbial phrase	(f) Emphatic adverb itself		0.244
C6		(f’) Following emphatic adverb		0.350
C7	(g) Cue word			0.415
C8	(h) Numeral			0.393
C9	Other (baseline)			0.101

Presentation agent



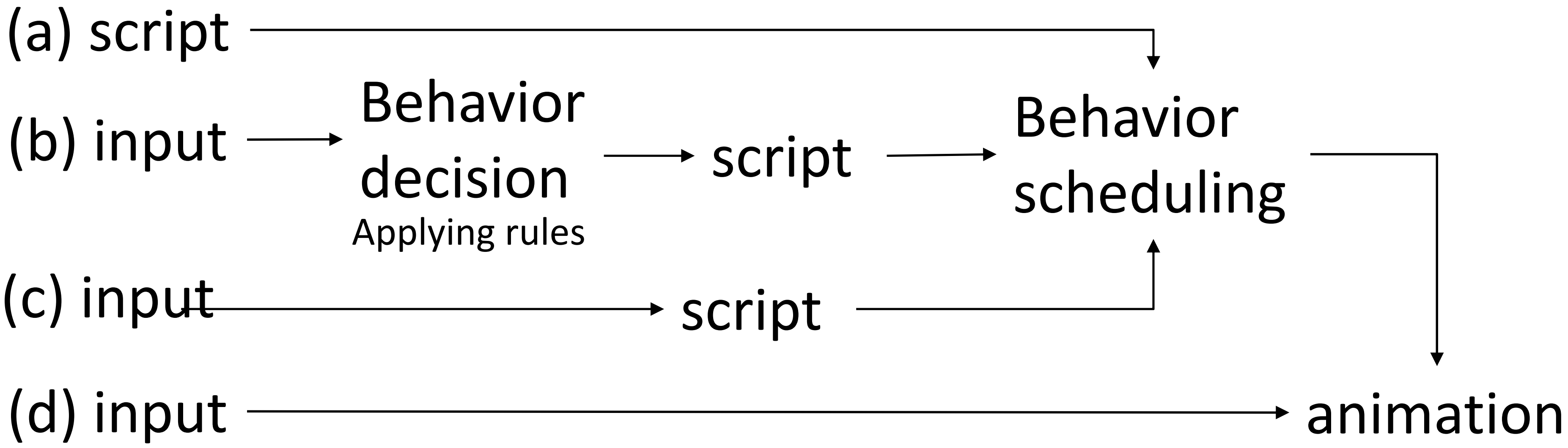
Co-articulation

- Xu et al. (2014)
- Co-articulation between gestures
  -> previous gesture affects the shape of the next gesture
- Co-articulation within gesture units
  ->when gestures go into relax, rest positions or holds?

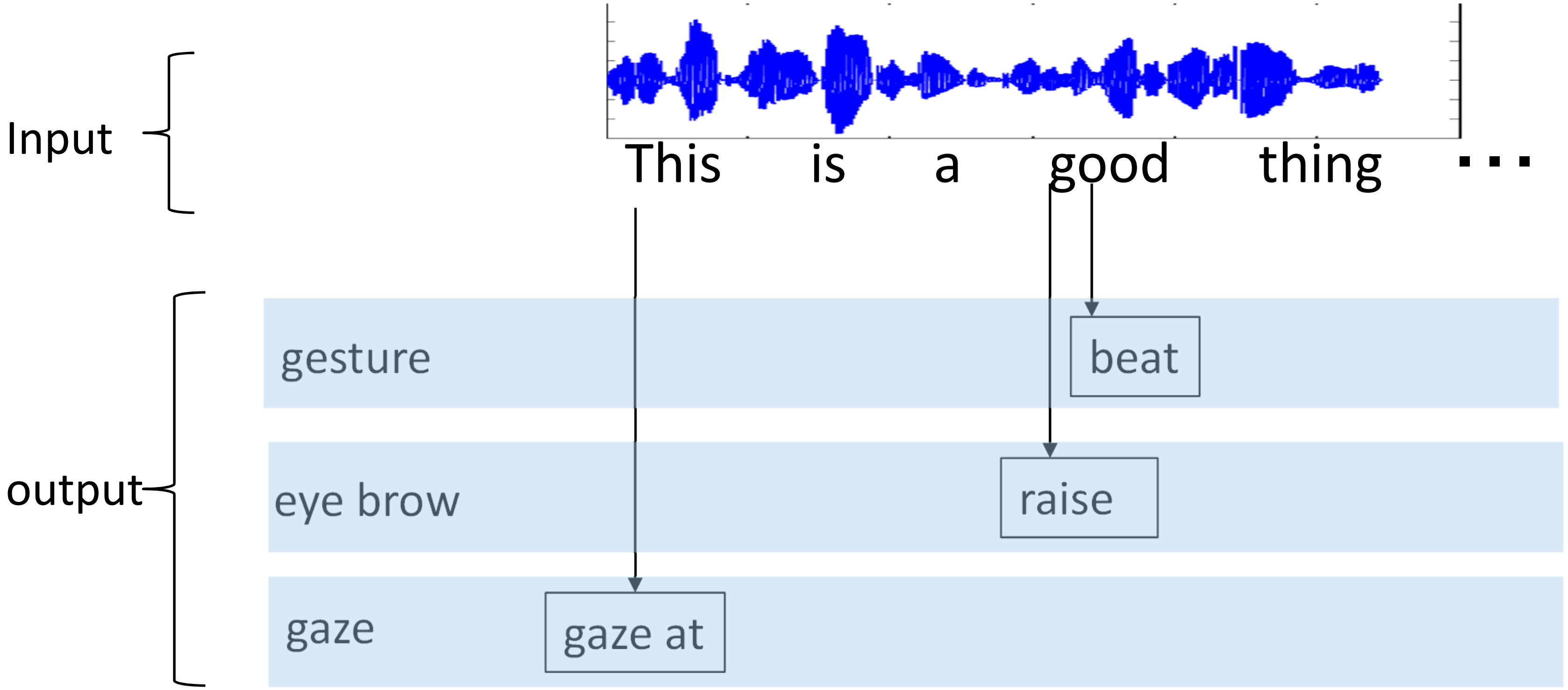
https://www.youtube.com/watch?v=A-3Ic-zCqnM&feature=youtu.be

Approaches

- (a) Manually generated script
- (b) Rule-based
- (c) Behavior prediction
- (d)) Joint position prediction



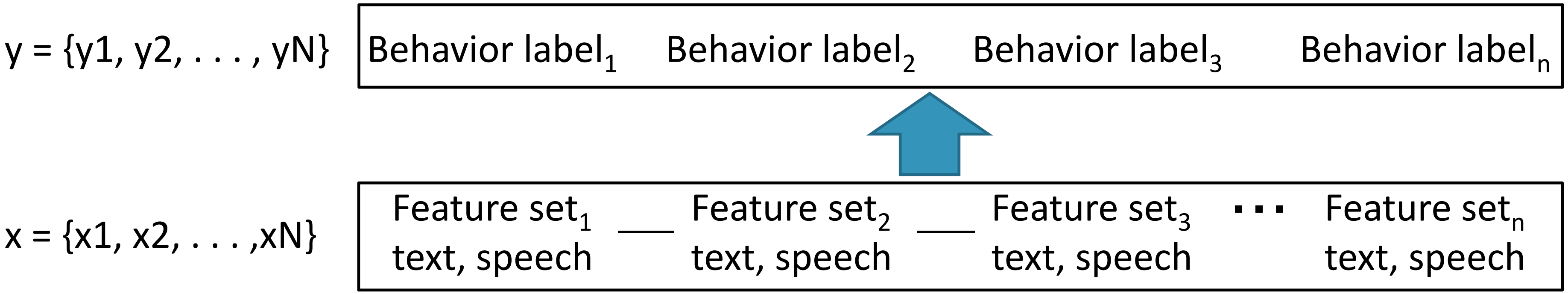
Predicting behavior labels



We want a model that predicts a sequence of behaviors

Predicting behavior labels (Cont.)

- Predict sequence of behaviors using temporal modeling, such as CRF (conditional random field).
- Prediction task
  - Input: x = {x1, x2, . . . ,xN}
    utterance transcription, part-of-speech tags, prosody features
  - Output: predict a sequence of gestural signs y = {y1, y2, . . . , yN}



y = {y1, y2, . . . , yN}

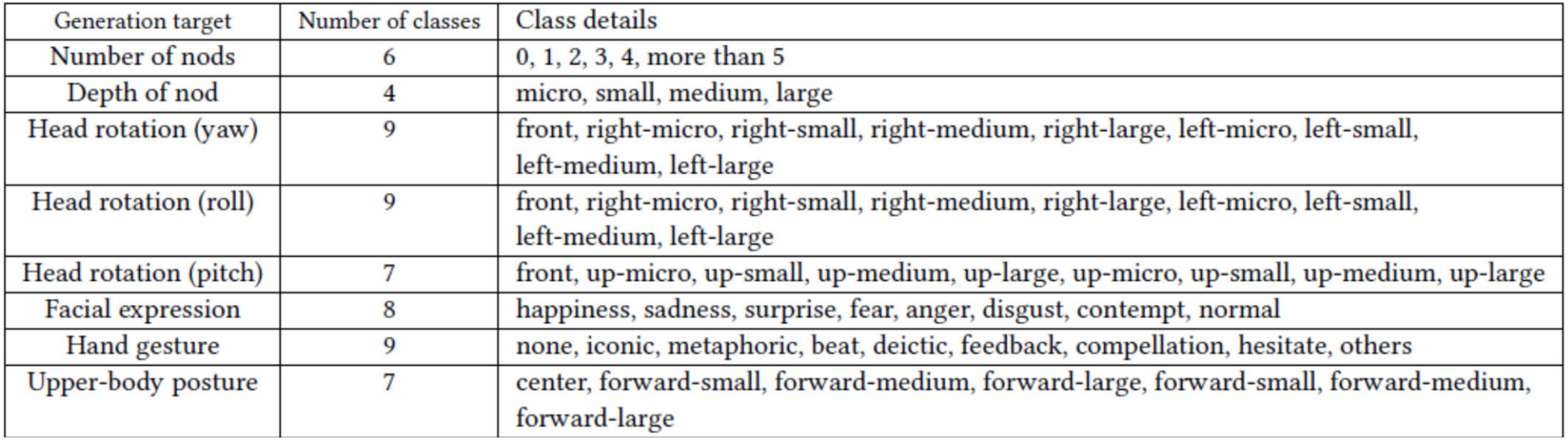
Behavior label$_1$ Behavior label$_2$ Behavior label$_3$ Behavior label$_n$

x = {x1, x2, . . . ,xN}

Feature set$_1$ text, speech — Feature set$_2$ text, speech — Feature set$_3$ text, speech ▪ ▪ ▪ Feature set$_n$ text, speech

Predicting behavior sequence (1)

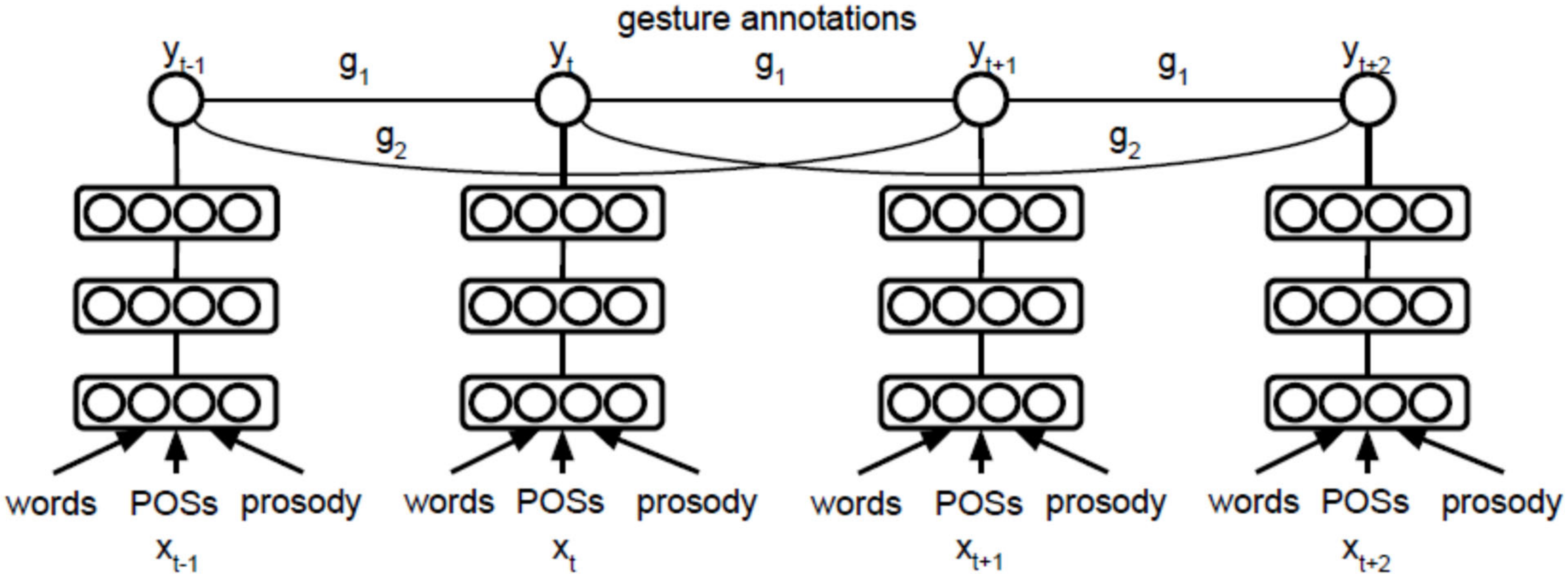
- Ishii et al (2018)
  - Input: set of features: Length of phrase , Word position, Bag of words, Dialogue act, Part of speech, synonyms
  - Output: behavior labels (choose one class for each behavior)



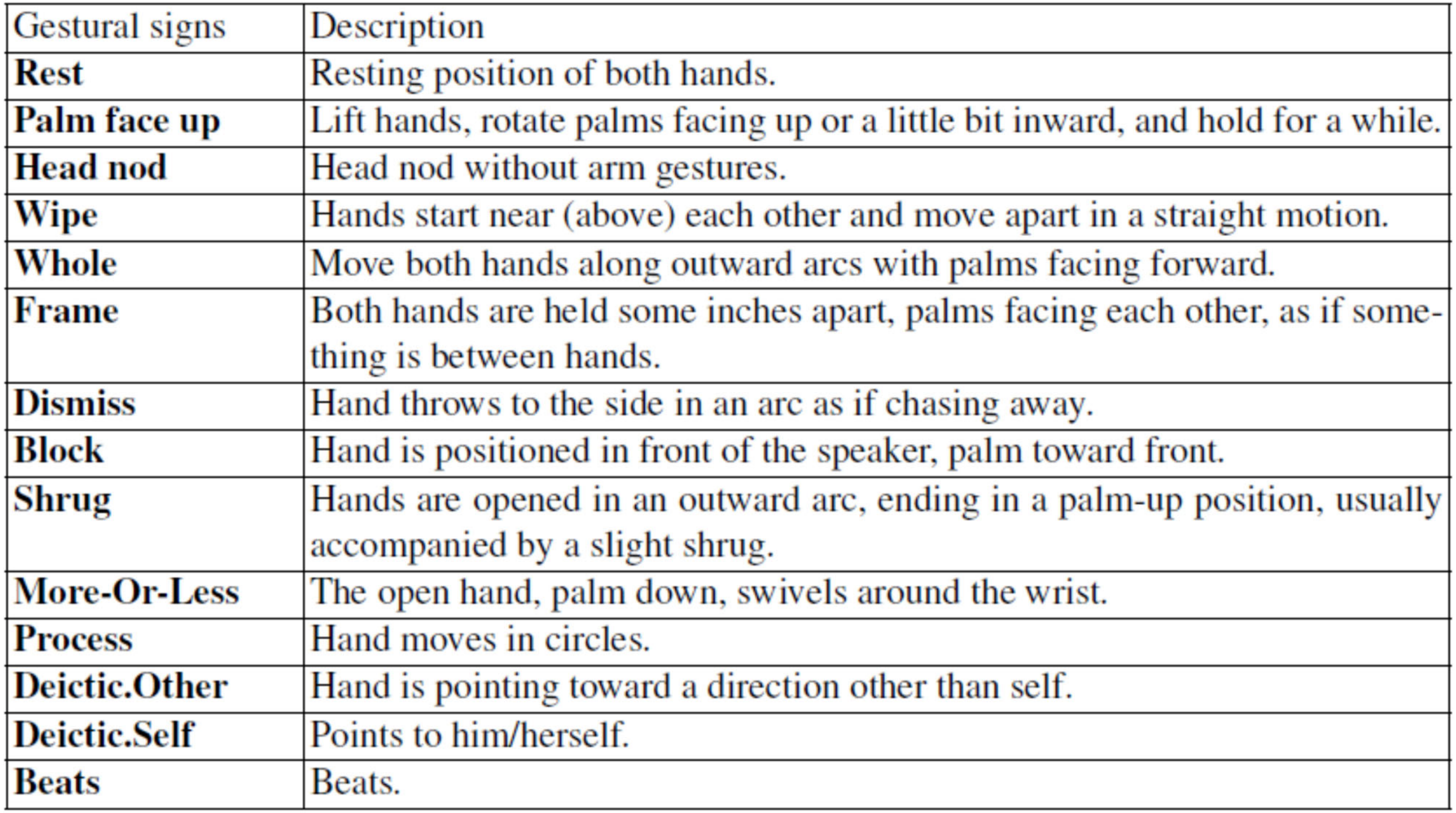
Generation target	Number of classes	Class details
Number of nods	6	0, 1, 2, 3, 4, more than 5
Depth of nod	4	micro, small, medium, large
Head rotation (yaw)	9	front, right-micro, right-small, right-medium, right-large, left-micro, left-small, left-medium, left-large
Head rotation (roll)	9	front, right-micro, right-small, right-medium, right-large, left-micro, left-small, left-medium, left-large
Head rotation (pitch)	7	front, up-micro, up-small, up-medium, up-large, up-micro, up-small, up-medium, up-large
Facial expression	8	happiness, sadness, surprise, fear, anger, disgust, contempt, normal
Hand gesture	9	none, iconic, metaphoric, beat, deictic, feedback, compellation, hesitate, others
Upper-body posture	7	center, forward-small, forward-medium, forward-large, forward-small, forward-medium, forward-large

Predicting behavior sequence (2)

- Chiu et al (2015): Deep Conditional Neural Field (DCNF)
- Predicting gesture labels by combining CRF and deep learning



Set of gesture labels for prediction

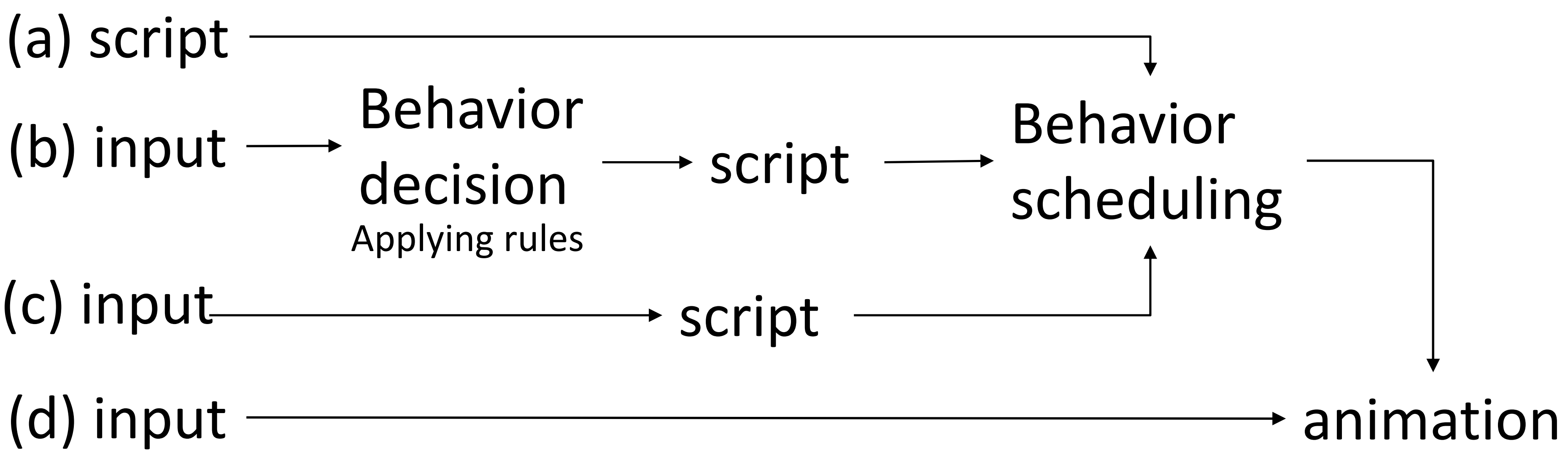


Gestural signs	Description
Rest	Resting position of both hands.
Palm face up	Lift hands, rotate palms facing up or a little bit inward, and hold for a while.
Head nod	Head nod without arm gestures.
Wipe	Hands start near (above) each other and move apart in a straight motion.
Whole	Move both hands along outward arcs with palms facing forward.
Frame	Both hands are held some inches apart, palms facing each other, as if something is between hands.
Dismiss	Hand throws to the side in an arc as if chasing away.
Block	Hand is positioned in front of the speaker, palm toward front.
Shrug	Hands are opened in an outward arc, ending in a palm-up position, usually accompanied by a slight shrug.
More-Or-Less	The open hand, palm down, swivels around the wrist.
Process	Hand moves in circles.
Deictic.Other	Hand is pointing toward a direction other than self.
Deictic.Self	Points to him/herself.
Beats	Beats.

Table 1: A formalized representation of co-verbal gestures for computational prediction.

Approaches

- (a) Manually generated script
- (b) Rule-based
- (c) Behavior prediction
- (d) Joint position prediction


(a) script

(b) input

Behavior decision

Applying rules

script

Behavior scheduling

(c) input

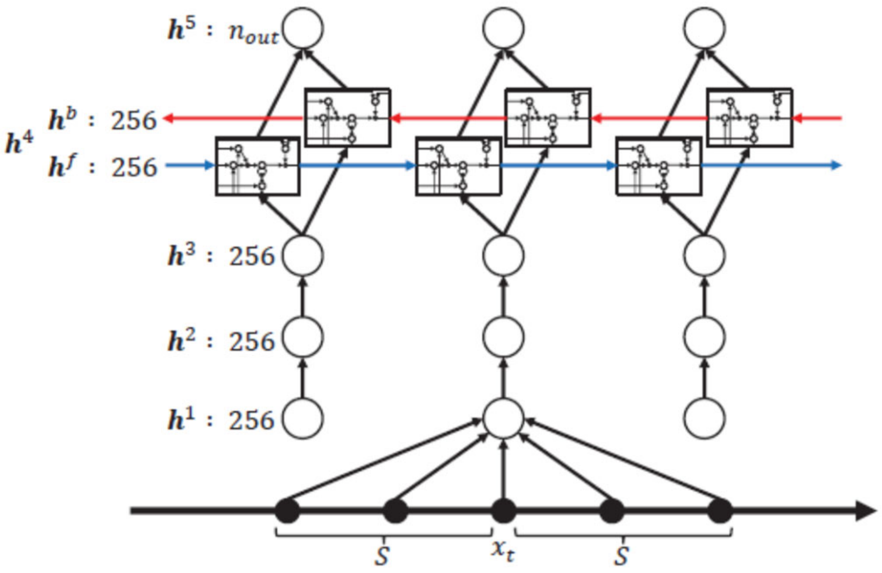
script

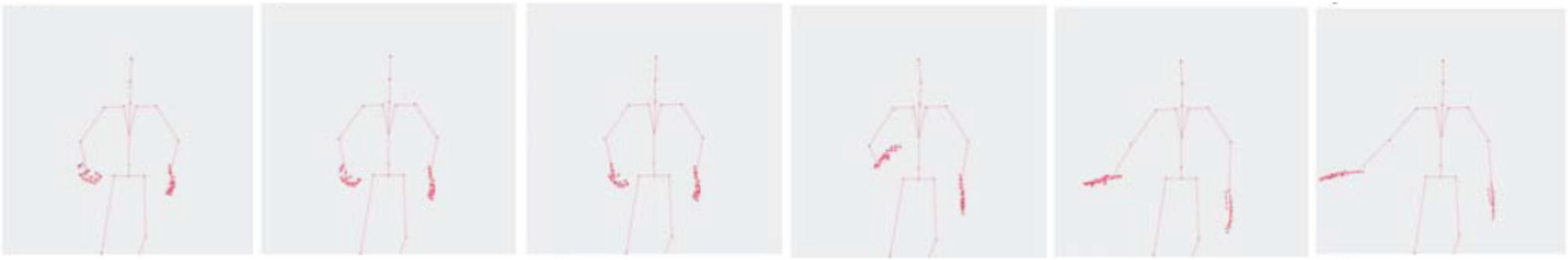
(d) input

animation

Joint position prediction approach

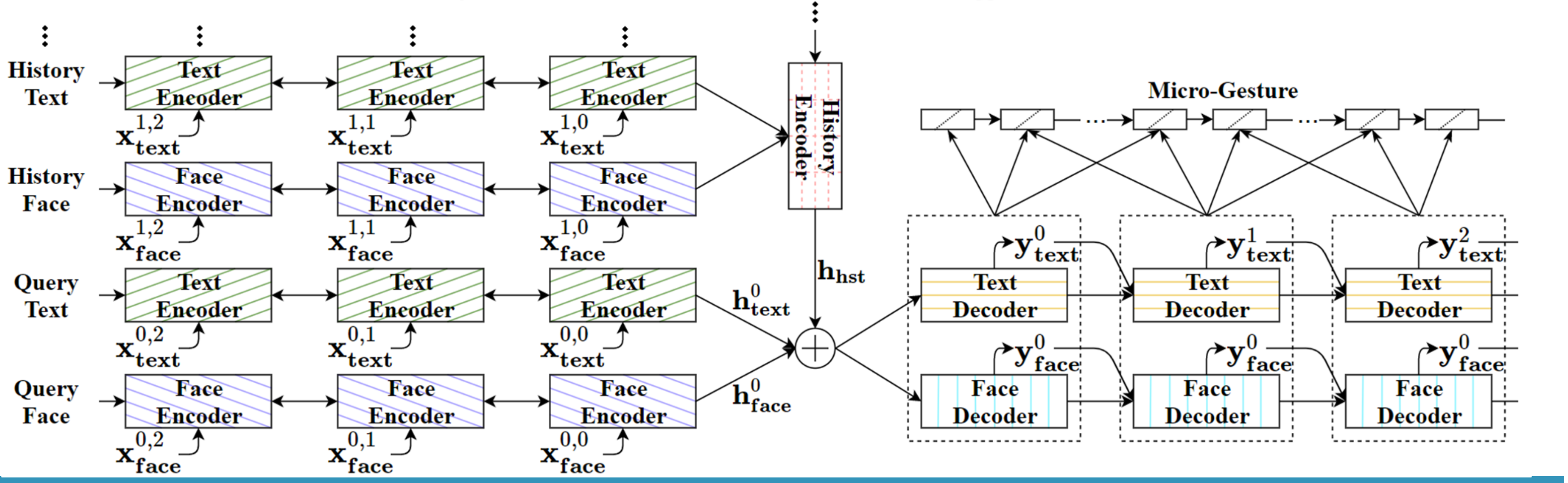
- Predicting next joint positions using LSTM
- Input: speech audio feature (MFCC)
- Output: set of joint positions





Seq2Seq multimodal dialogue system

- Chu et al. 2018
- Categories of face and head motion expression
  - Obtain 18 types of AUs and 3D head pose data from OpenFace
  - Clustering the behavior patterns using k-means(k=200) as behavior templates
- Create Seq2Seq model by combining sequence of words and sequence of behavior templates
- Send the decoder output to TTS and animation engine => chat bot!



http://www.cs.toronto.edu/face2face

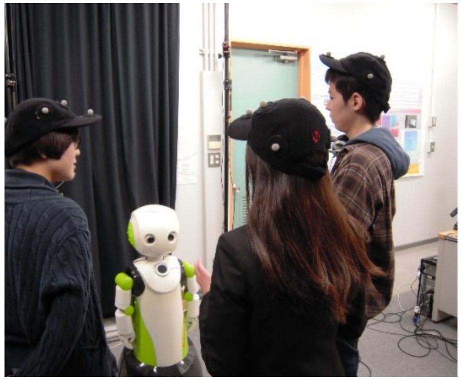
Chu, H., Li, D. and Fidler, S. (2018). A Face-to-Face Neural Conversation Model. CVPR 2018.

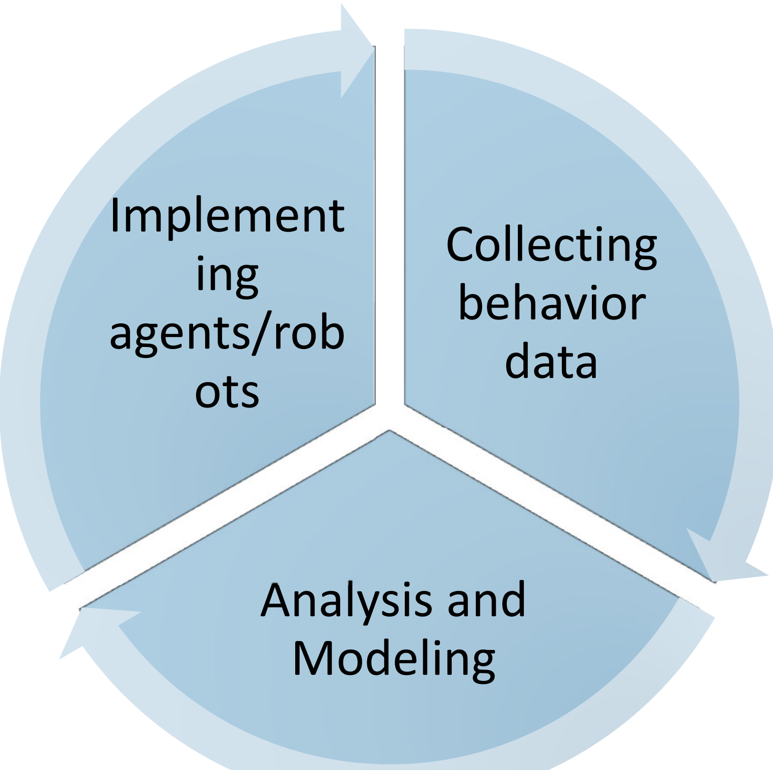
Discussions

- Wiring script is time consuming
- Defining rules need domain knowledge, and still need human effort
- Label prediction can only predict limited kinds of behaviors.
- Position prediction approach does not care about relationship between linguistic information and communication signals.

Communicating with virtual agent

Our approach

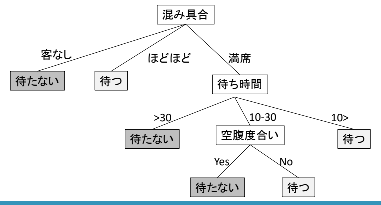


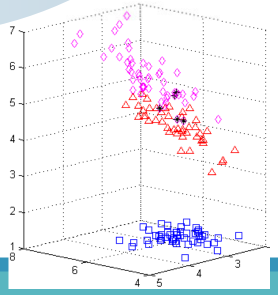


Implementing agents/robots

Collecting behavior data

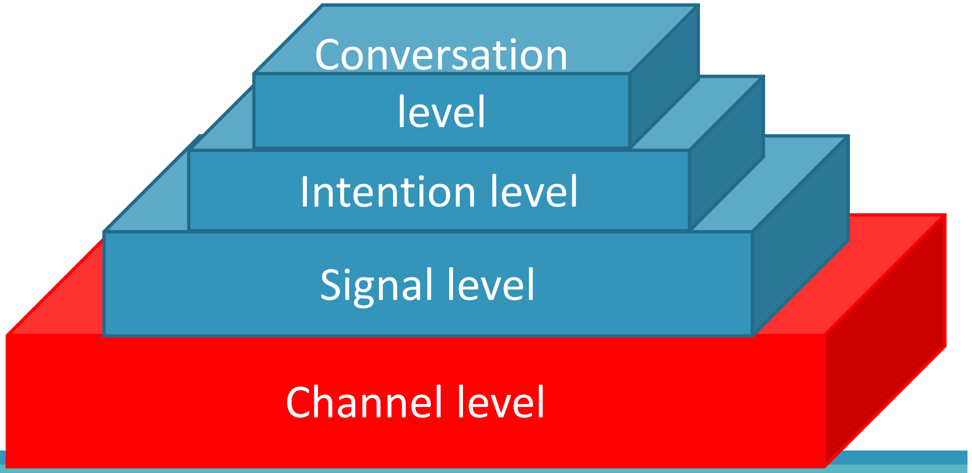
Analysis and Modeling



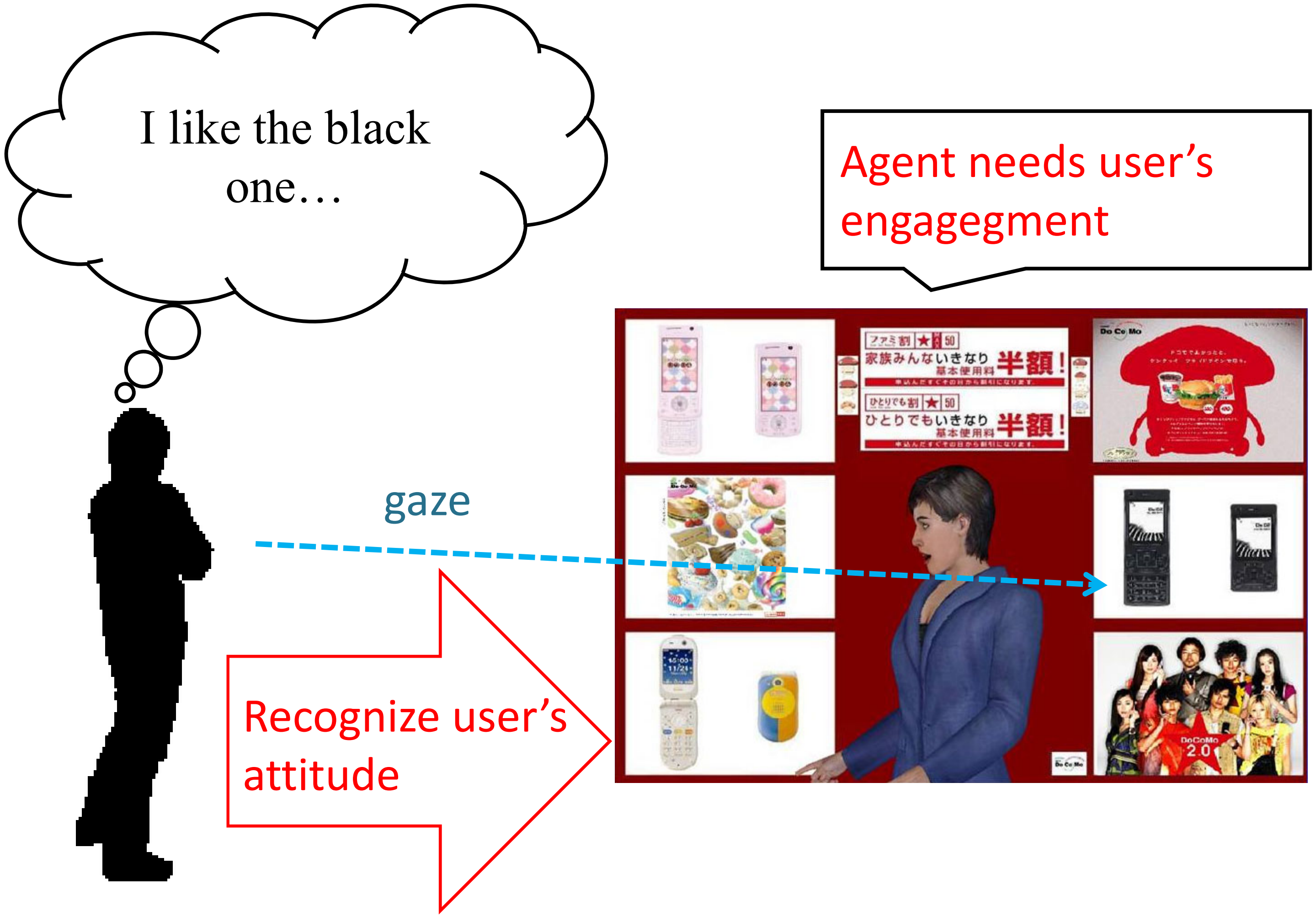


Establishing communication

- Level of conversation (Clark; 1996, Paek et al; 1999)
  - Channel level: the listener pays attention to and perceive signals from the speaker.
  - Signal level: the listener identify the signal as a communication signal
  - Intention level: the listener understand the speaker's utterance
  - Conversation level: the listener agree or disagree to work on the joint action proposed by the speaker



Engagement



I like the black one…

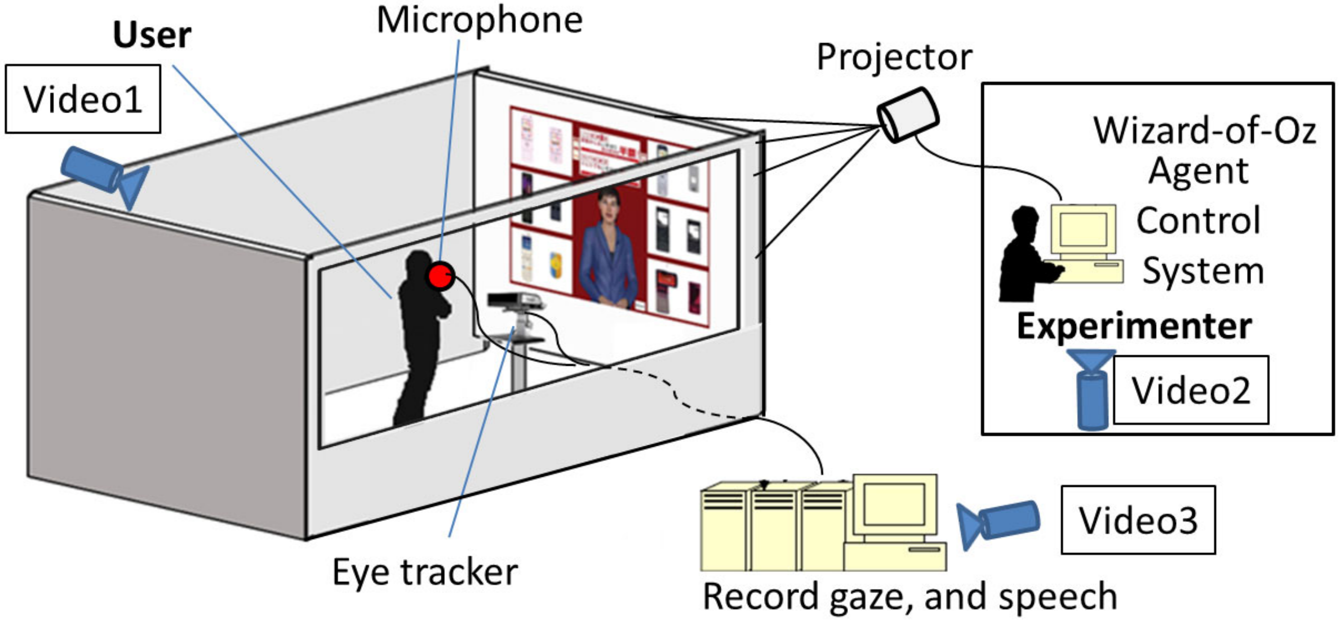
gaze

Agent needs user’s engagegment

Recognize user’s attitude

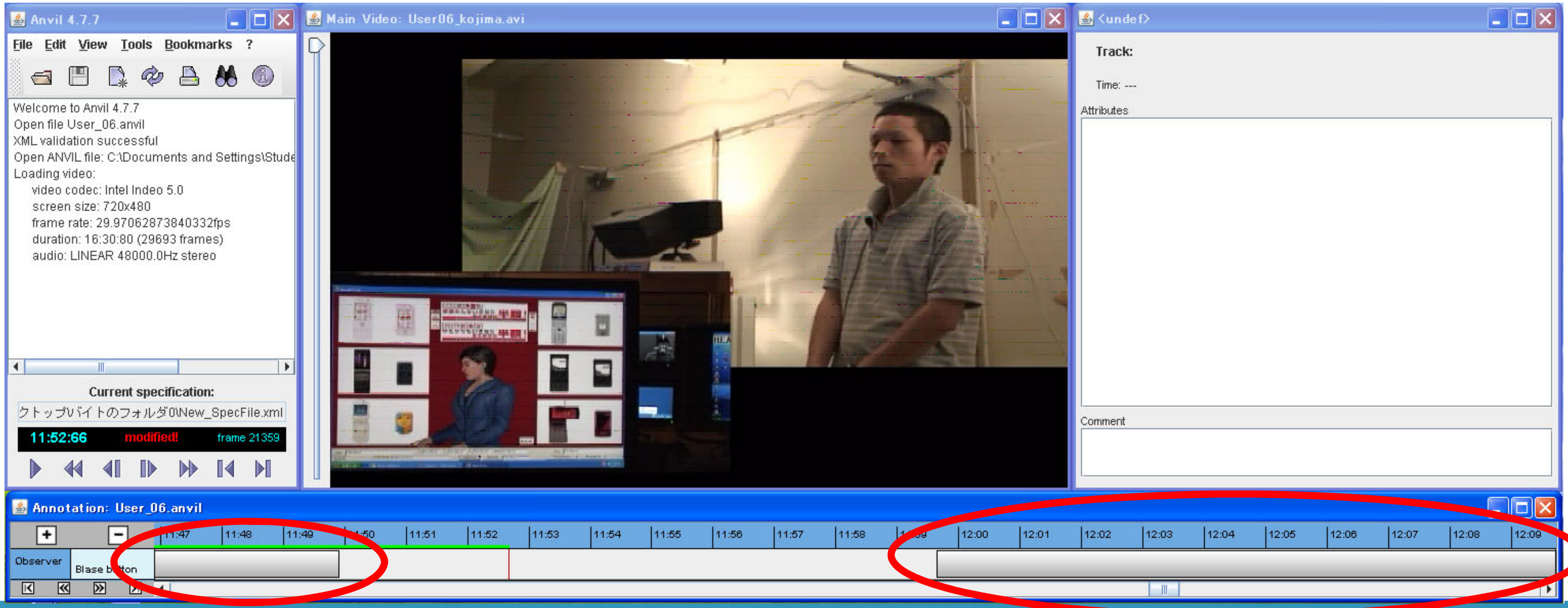
WOZ experiment for corpus collection

- Wizard-of-Oz system
  - Experimenter interpreted the subject’s utterance and typed in the response
    - Subject can ask : price and functions of new models of cell phones
- 10 dialogues
  - Average length
- Collected dat
  - Eye gaze
  - Head pose
  - Speech



Video annotation

- We recruited another 10 annotators, and asked them to watch the video and mark the time when the subject looked disengaged from the conversation.
- Disengagement score: how many people judged a given time (30fps) as disengagement



```
Welcome to Anvil 4.7.7
Open file User_06.anvil
XML validation successful
Open ANVIL file: C:\Documents and Settings\Stude
Loading video:
    video codec: Intel Indeo 5.0
    screen size: 720x480
    frame rate: 29.97062873840332fps
    duration: 16:30:80 (29693 frames)
    audio: LINEAR 48000.0Hz stereo
```

Current specification: クトップ\バイトのフォルダ0\New_SpecFile.xml

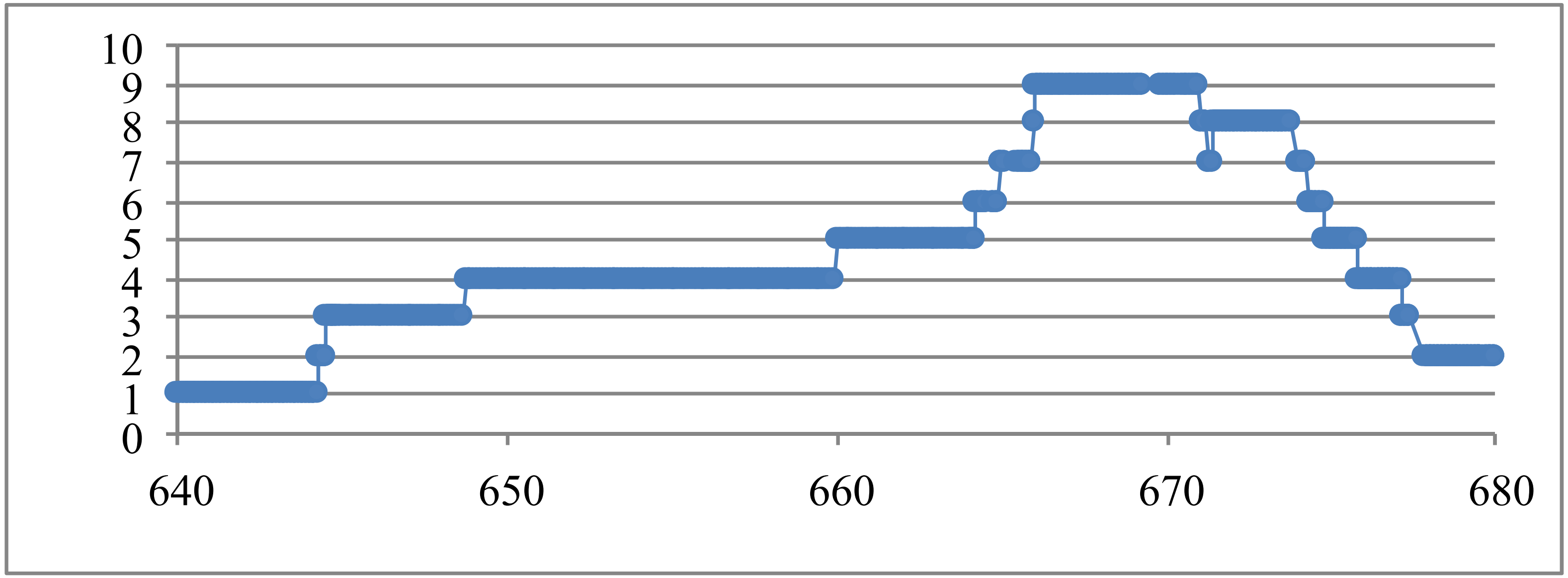
11:52:66 modified! frame 21359

Track:

Time: ---

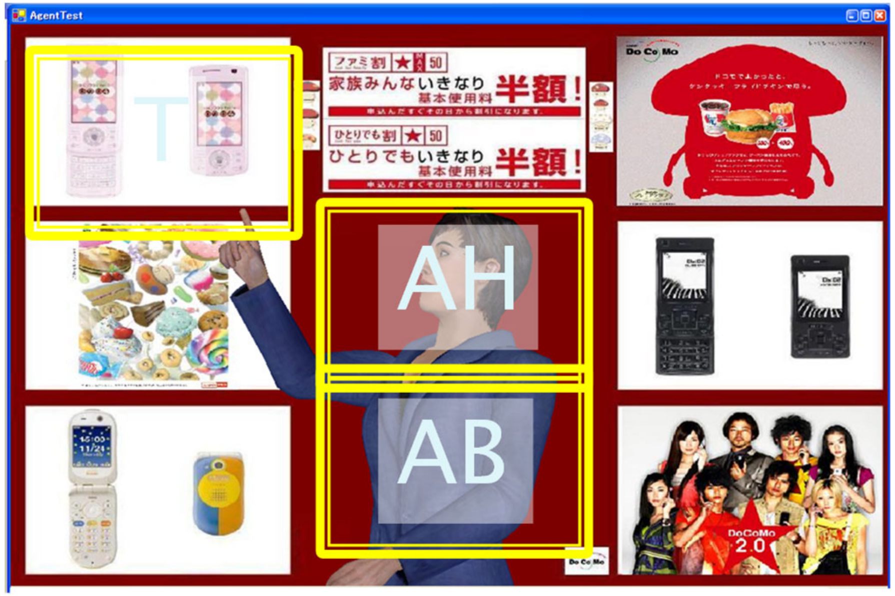
Parasocial consensus data

- When the disengagement score was 3, the score reached higher scores. The average peaks for such shifts were over 5
-> set the disengagement threshold at 3



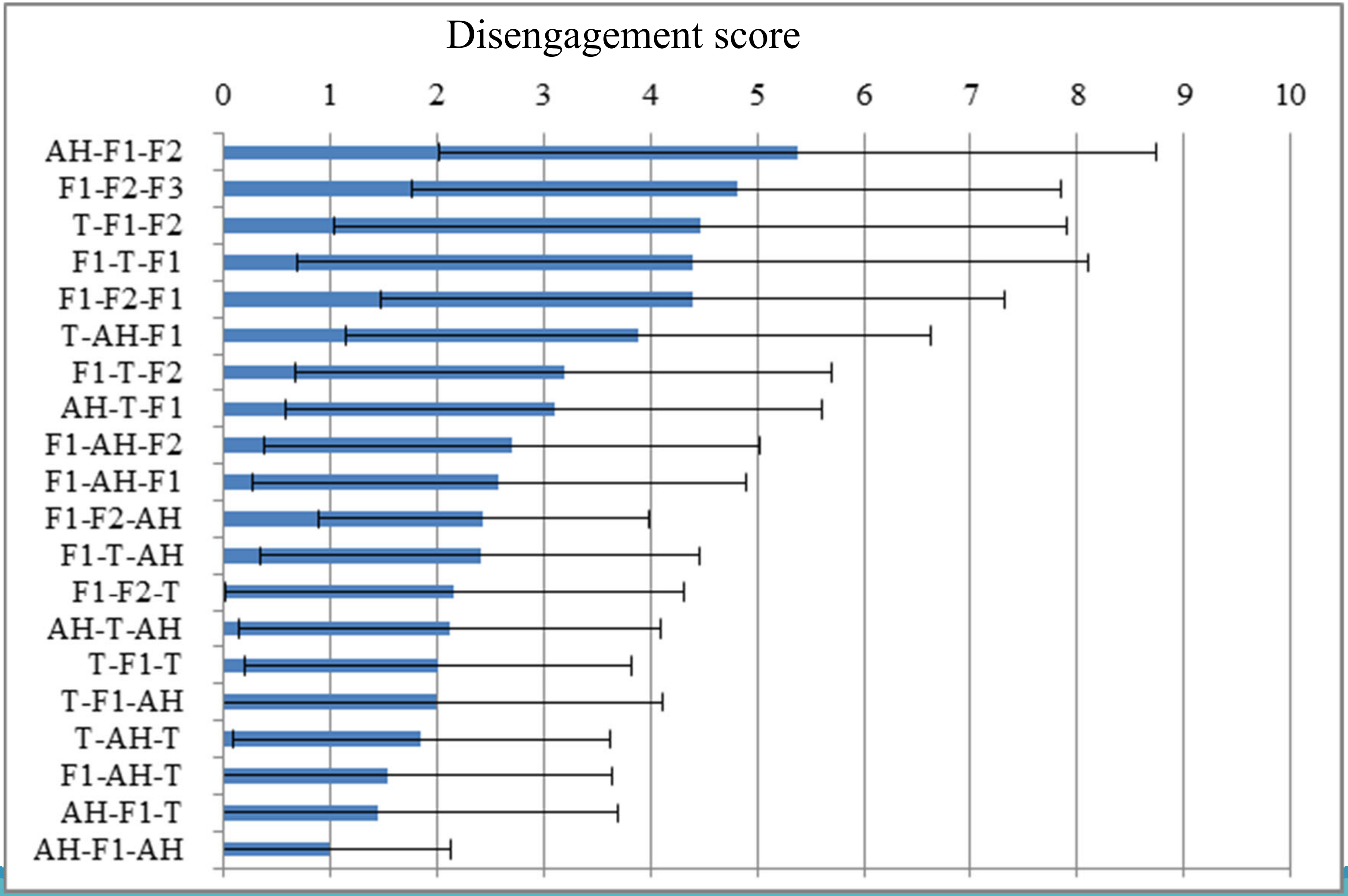
Data analysis

- Gaze 3-gram patterns
- Eye movement distance
- Pupil size
- Duration
- Head pose


AgentTest
T
AH
AB
家族みんないきなり 基本使用料 半額！
ひとりでもいきなり 基本使用料 半額！
DoCoMo 2.0

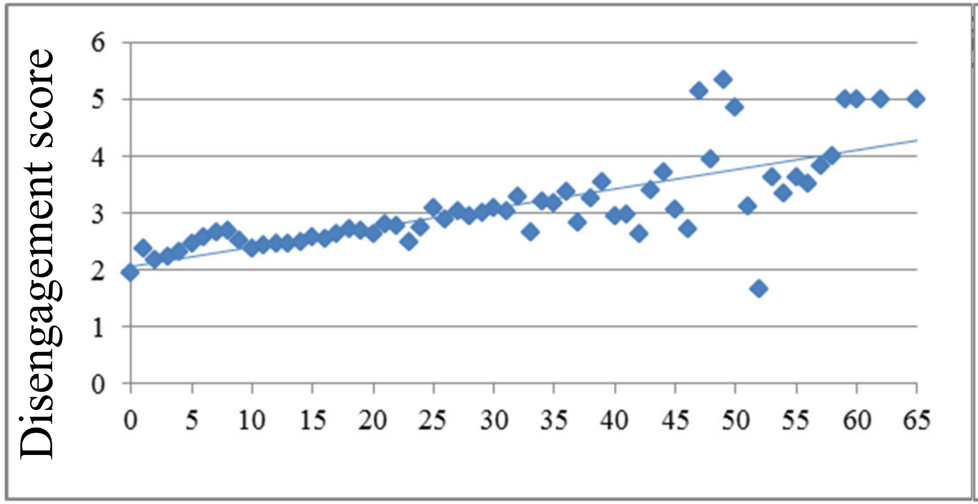
T: look at target object
A : look at agent
F : look away

Gaze patterns and disengagement

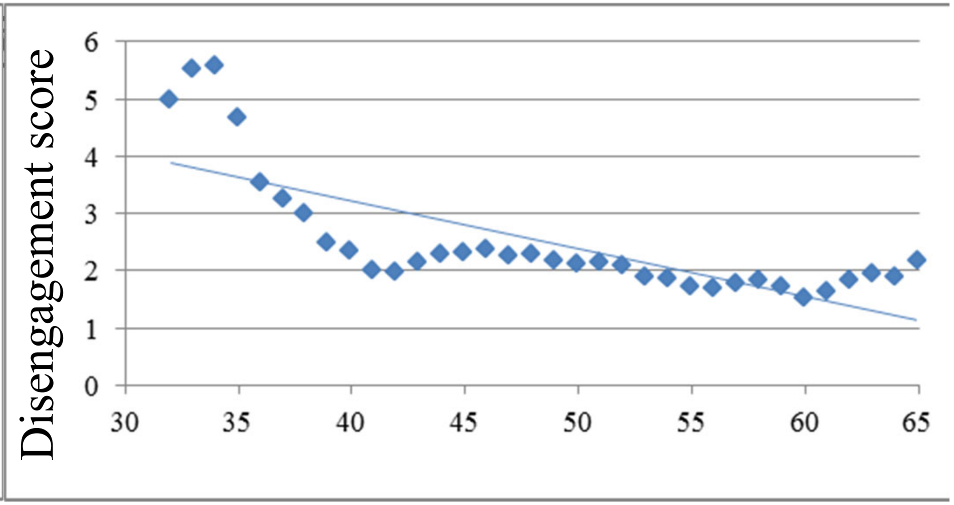


Eye move distance, pupil size

- The larger the eye movement distance is, the higher the disengagement score becomes
- The smaller the pupil size is, the higher the disengagement score becomes



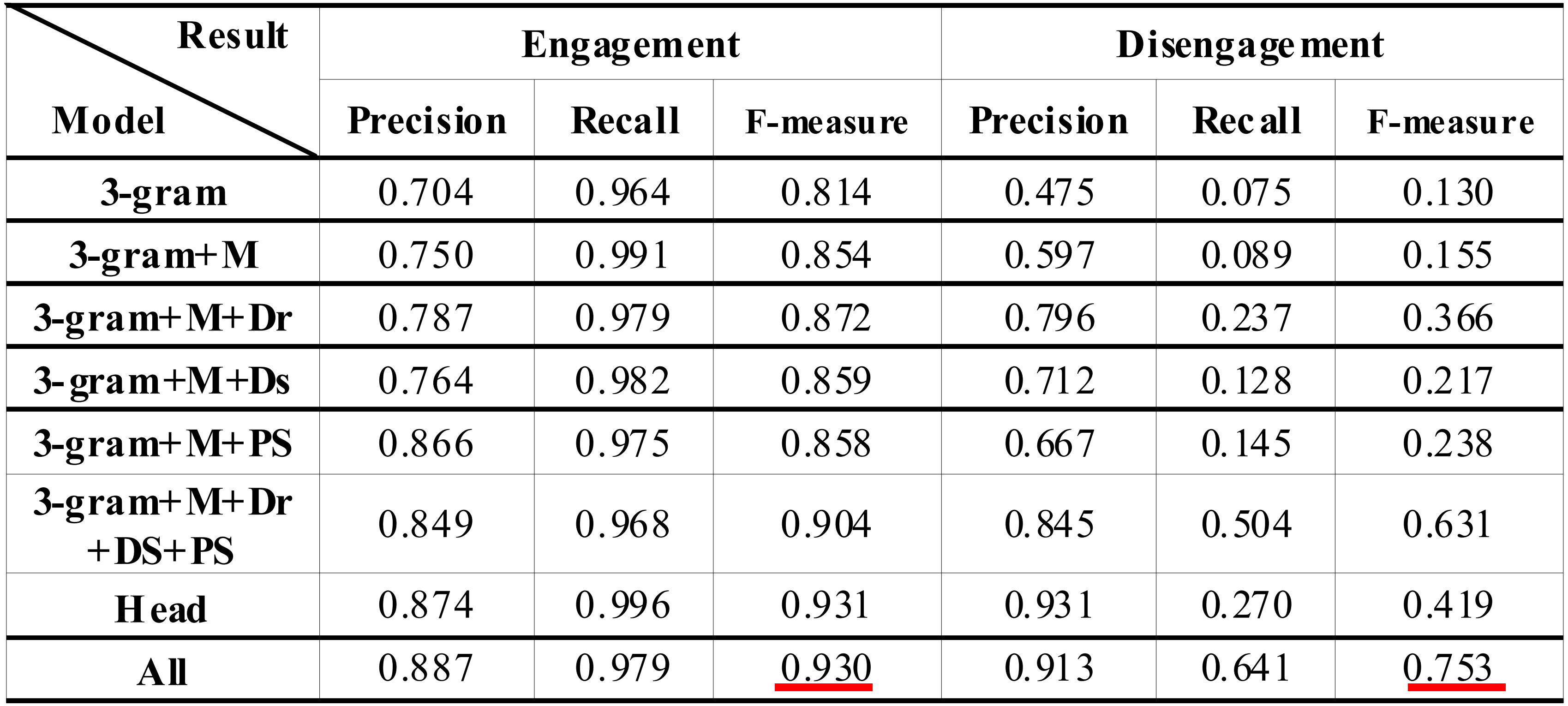
Distribution of eye movement distance (pixels)



Pupil size distribution (mm)

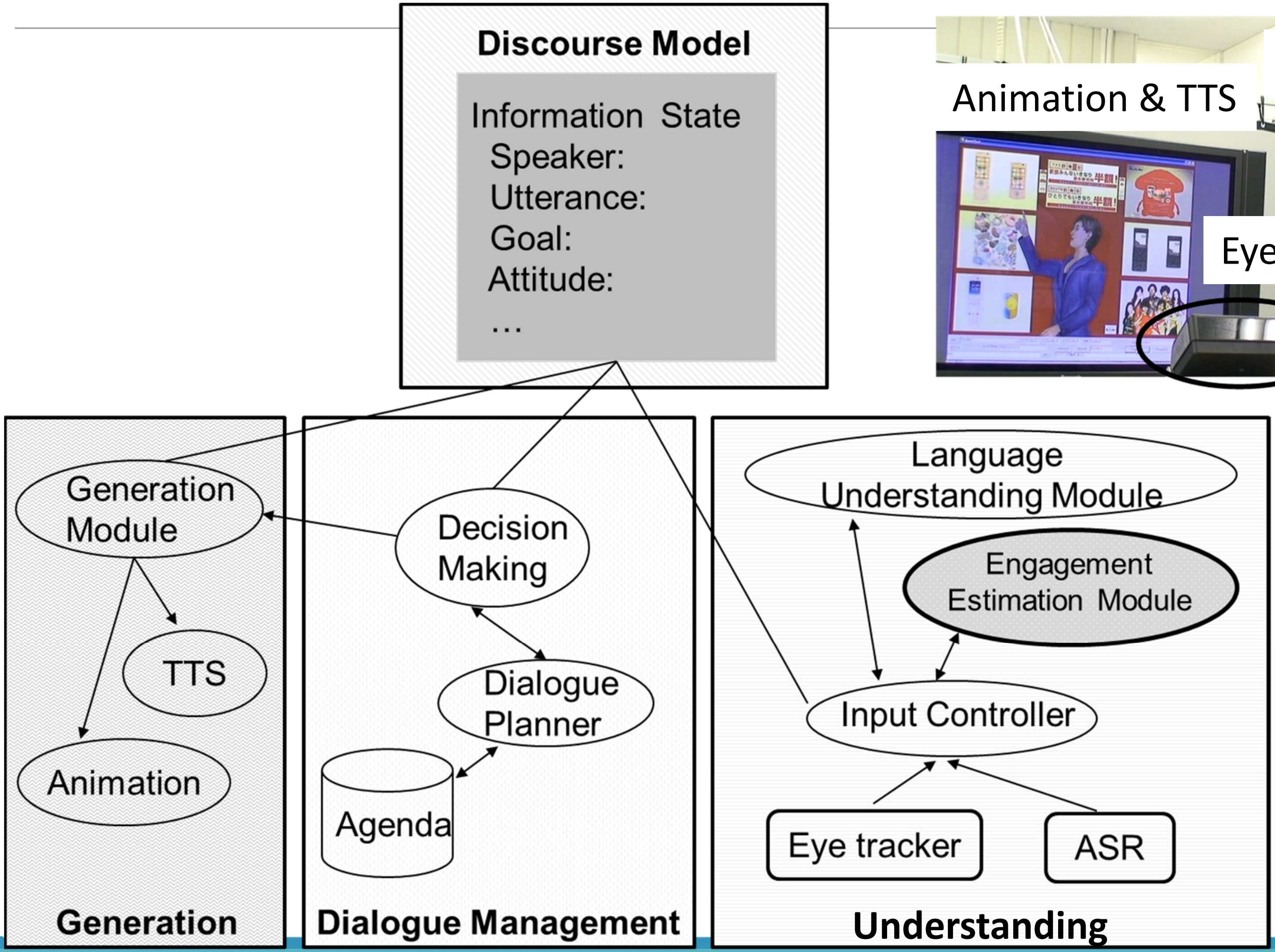
Model evaluation

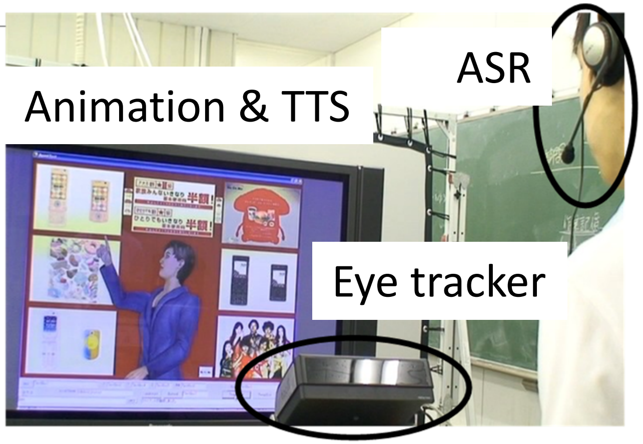
Results of SVM



Result / Model	Engagement			Disengagement		
	Precision	Recall	F-measure	Precision	Recall	F-measure
3-gram	0.704	0.964	0.814	0.475	0.075	0.130
3-gram+M	0.750	0.991	0.854	0.597	0.089	0.155
3-gram+M+Dr	0.787	0.979	0.872	0.796	0.237	0.366
3-gram+M+Ds	0.764	0.982	0.859	0.712	0.128	0.217
3-gram+M+PS	0.866	0.975	0.858	0.667	0.145	0.238
3-gram+M+Dr +DS+PS	0.849	0.968	0.904	0.845	0.504	0.631
Head	0.874	0.996	0.931	0.931	0.270	0.419
All	0.887	0.979	0.930	0.913	0.641	0.753

Implementation

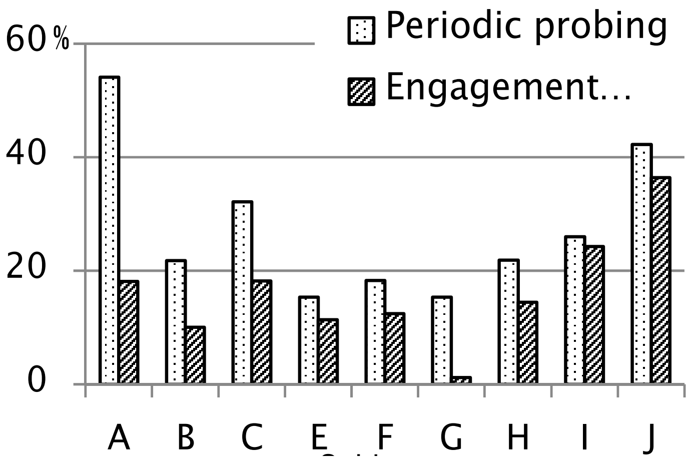




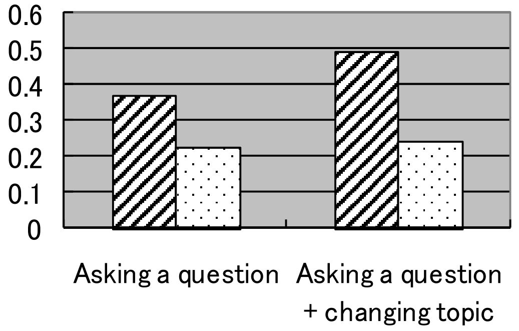
Demo video

Evaluation experiment

- Experimental conditions
  - Engage estimation : the proposed system
  - Periodic probing : probe every 10 utterances
- Subjects: 10 university students (7 male, 3 female)
- Subjective evaluation
  - Awareness of engagement, Appropriateness of behavior, Smoothness of conversation, Intelligence
- Subject's nonverbal behaviors
  - Decrease the number of disengagement status
- Subject's verbal behaviors
  - Subject asked questions and changed a topic when the agent gave the probe
- If the agent estimates the user's engagement and gives proes based on this;
  - Improve the impression to the agent
  - Decrease the user's disengagement states
  - Trigger subject's utterance



Frequency of Disengagement



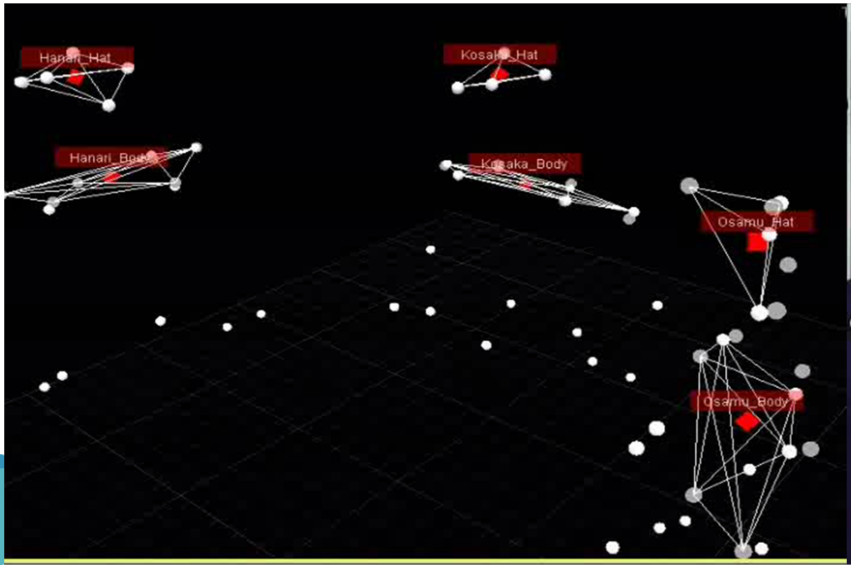
Frequency of user's verbal contribution

Dominance estimation

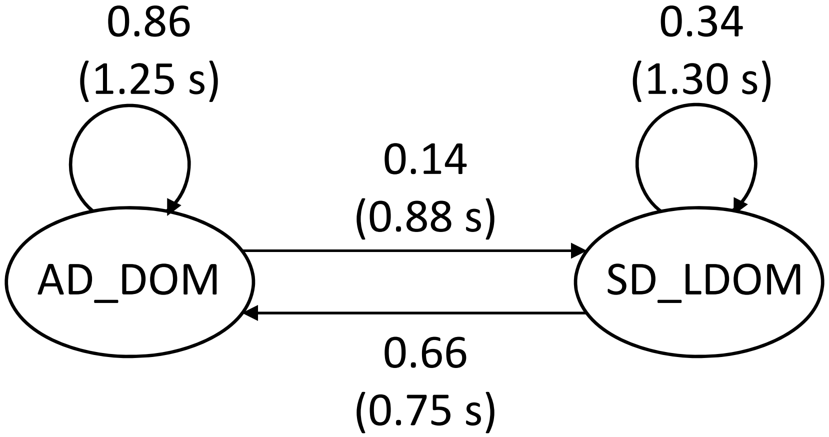
- In multiparty communication, there exists a dominant person and less dominant person.
  - Dominant participant : Leading a conversation
  - Less dominant participant : Small contribution to the conversation, fewer chances to speak
- Regression model for estimating dominance

Dominance score= (0.80) × amount of gaze at others + (0.162) × amount of mutual gaze + (0.94) × amount of speech + (0.256) × breaking a silence + (-0.25)

- Establish a robot attention model by considering dominance

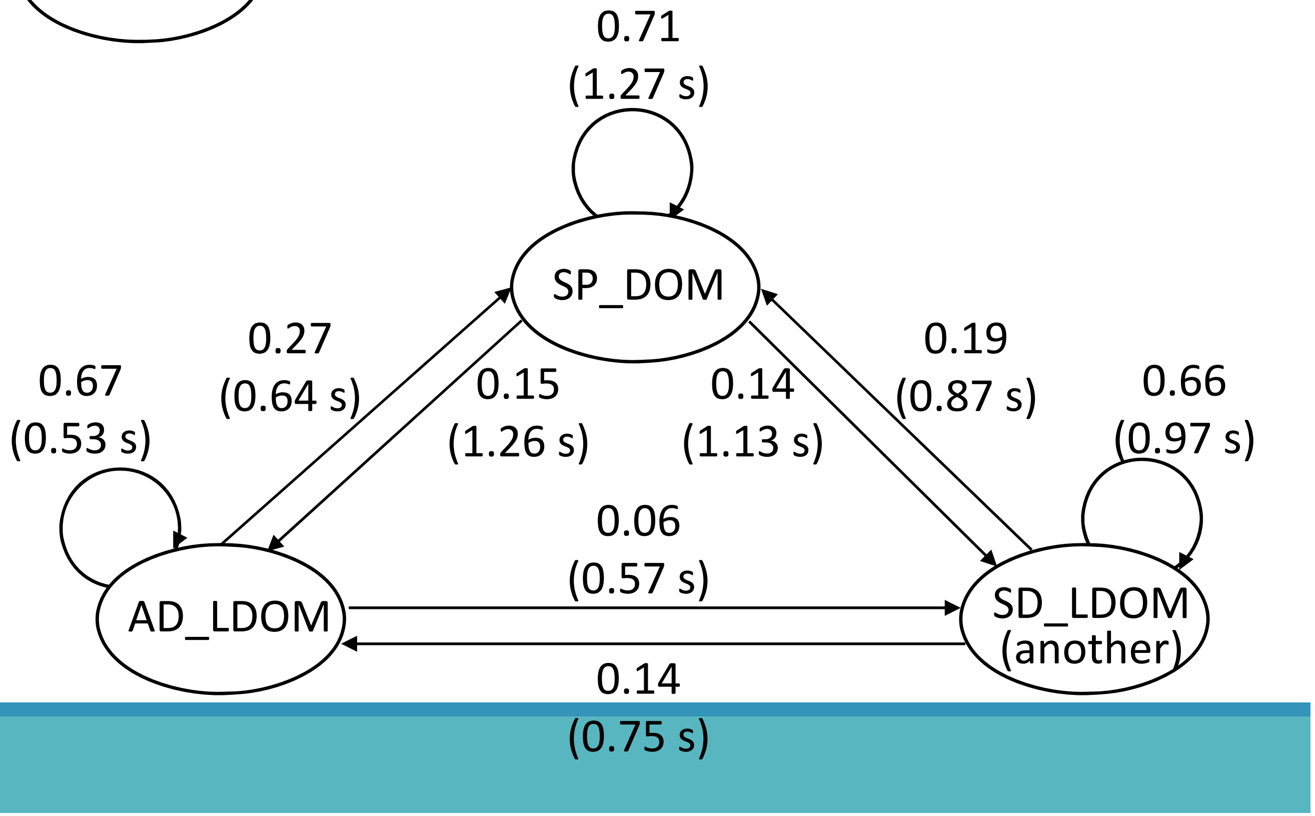


Robot head gaze model



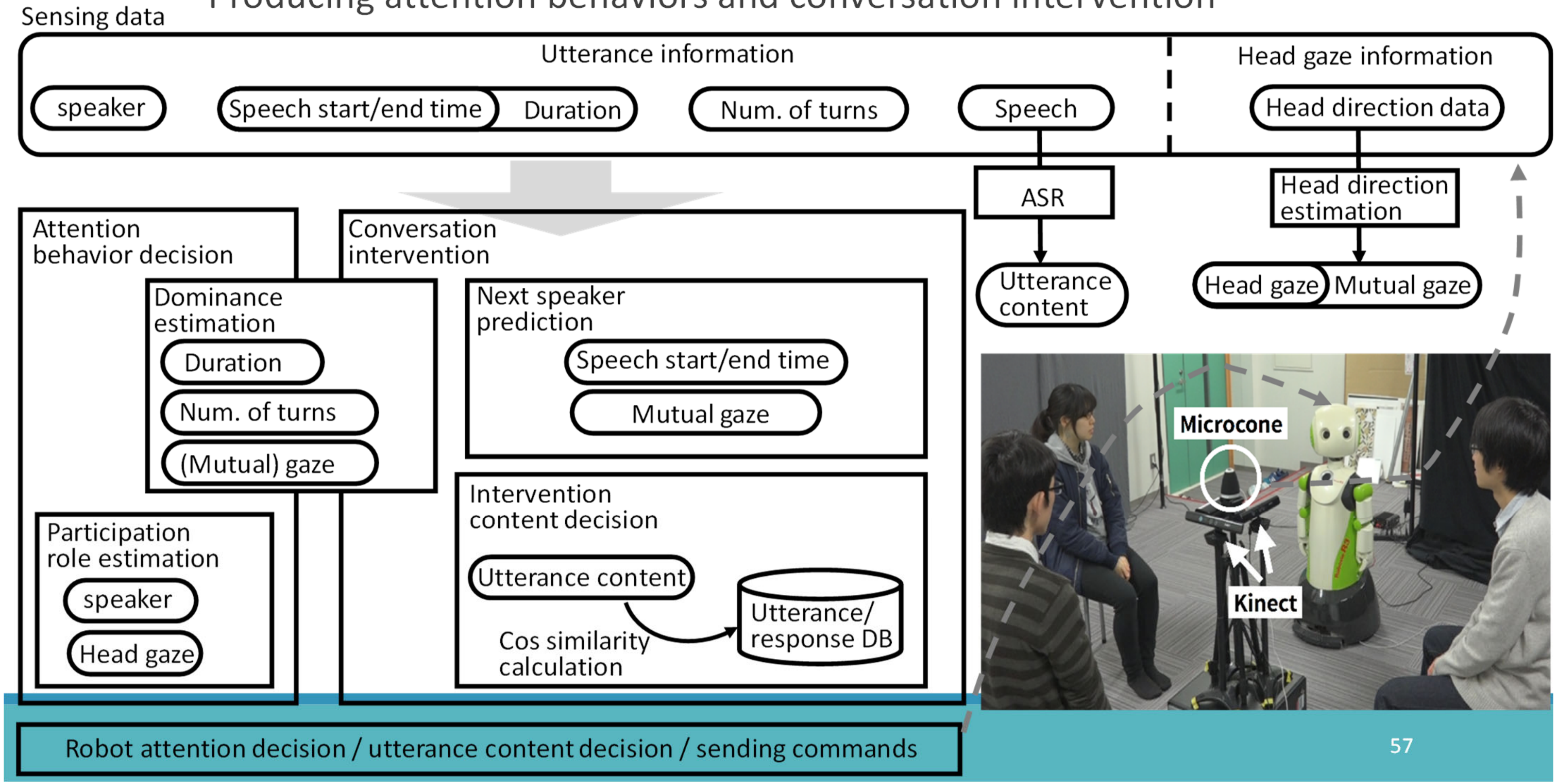
Gaze behaviors are different depending on the participation roles: speaker, addressee, side participant

Is the speaker/addressee/side participant dominant/less dominant?



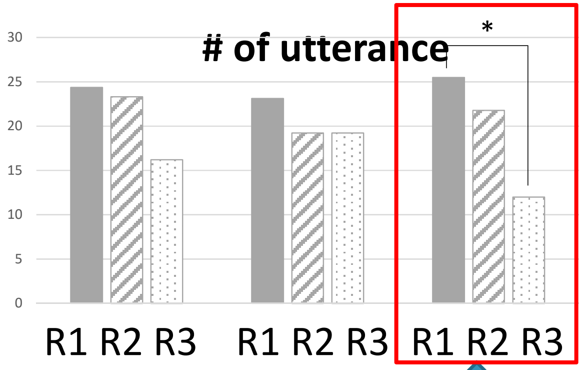
System architecture and functions

- Functions of conversation intervention robot
  - Estimating dominance and participation roles
  - Producing attention behaviors and conversation intervention

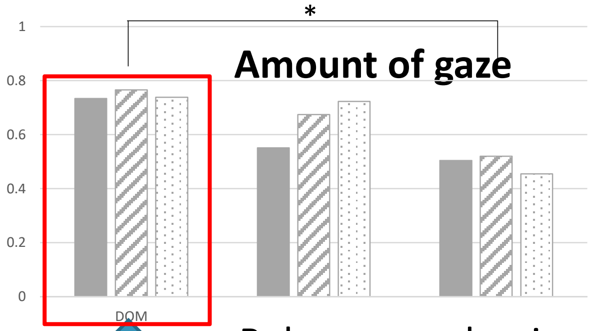


Evaluation experiment

- If a robot only looks at a speaker, the discrepancy of the amount of speech between the participants becomes larger.
- If a robot performs as a dominant participant, the amount of gaze communication of the group increases.
- Dominant person does not like a dominant robot very much.



Always look at a speaker



Behave as a dominant participant