TOYOHASHI UNIVERSITY OF TECHNOLOGY

Spoken dialog and multi-modal interaction

Norihide Kitaoka

Dpt. of Computer Science and Technology,
Toyohashi Univ. of Tech., Japan

Announcement

Please answer very short questionnaire!

For ITB students

For TUT students

TOYOHASHI UNIVERSITY OF TECHNOLOGY

Norihide Kitaoka

- BS and MS from Kyoto Univ. in '92 and '94 (Prof. Doshita and Prof. Kawahara)
- DENSO CORPORATION '94-'01
- Ph. D from Toyohashi Univ. of Tech. '00 (Prof. Nakagawa)
- Toyohashi Univ. of Tech. '01-'07 (Prof. Nakagawa)
- Nagoya Univ. '07-'14 (Prof. Takeda)
- Tokushima Univ. '14-'19
- Toyohashi Univ. of Tech. '19-

Toyohashi University of Technology

TOYOHASHI UNIVERSITY OF TECHNOLOGY

- Number of Students(international)
  - Undergraduate: 1,232 (77)
  - Graduate students
    Mater course: 861 (81)
    Doctoral course: 100 (44)

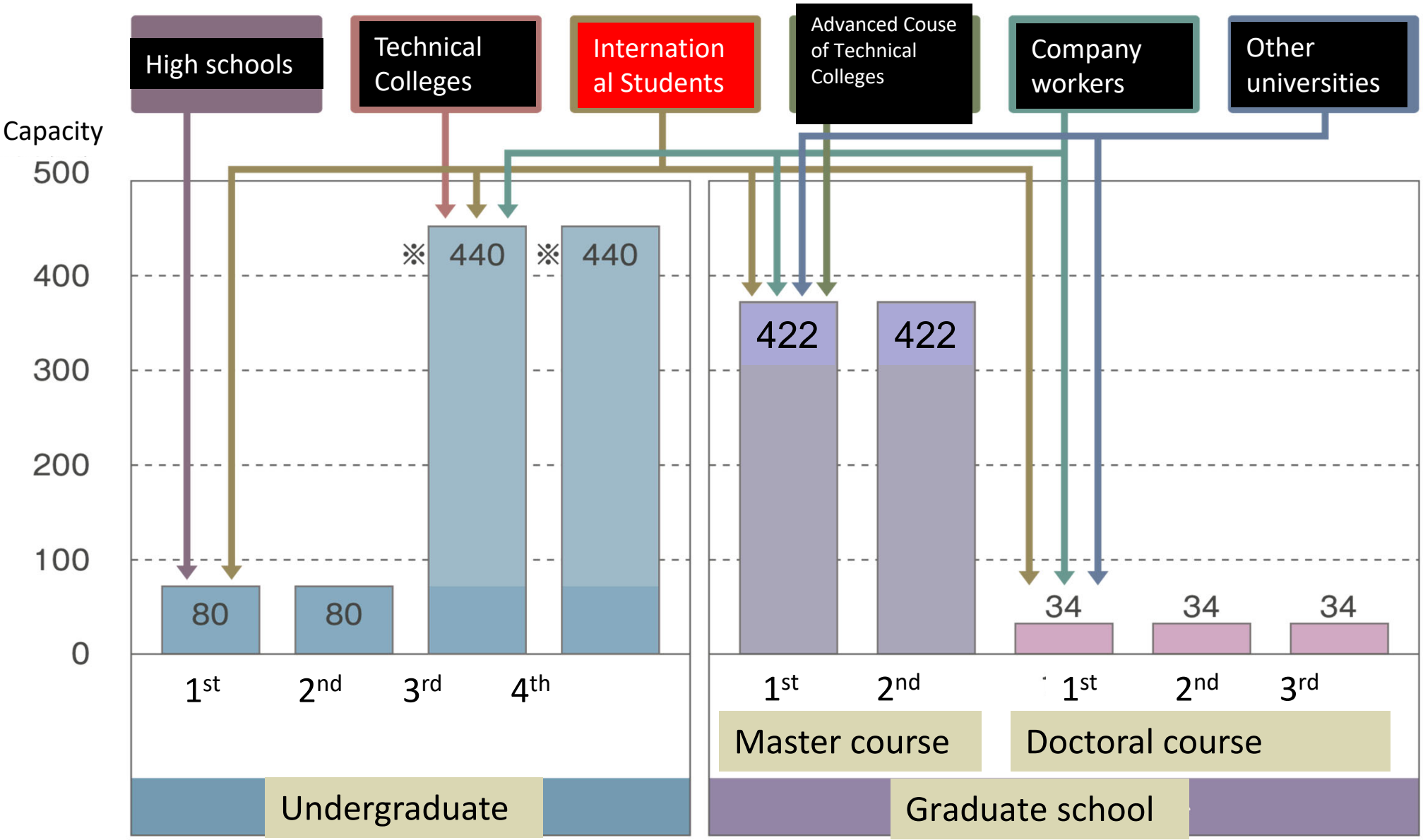
Toyohashi University of Technology

TOYOHASHI UNIVERSITY OF TECHNOLOGY

- Number of Faculty Members
  - Professors: 82
  - Associate Professors 67
  - Assistant Professors 4

President:
Prof. Akihiro Wakahara

Capacities


High schools
Technical Colleges
Internation al Students
Advanced Couse of Technical Colleges
Company workers
Other universities
Capacity
500
400
300
200
100
0
※ 440
※ 440
422
422
80
80
34
34
34
1st
2nd
3rd
4th
1st
2nd
1st
2nd
3rd
Master course
Doctoral course
Undergraduate
Graduate school

Departments

1. Mechanical Engineering
2. Electrical and Electronic Information Engineering
3. Computer Science and Engineering
4. Applied Cheistry and Life Science
5. Architecture and Civil Engineering

Institute of Liberal Arts and Sciences

Computer Science and Engineering

TOYOHASHI UNIVERSITY OF TECHNOLOGY

- Number of Students
  - Undergraduate: 296 (21)
  - Graduate students
    Mater course: 223 (15)
    Doctoral course: 26 (4)
- Number of Faculty Members
  - Professors: 15
  - Associate Professors 13
  - Assistant Professors 7

Members

Faculty members

Professor
Norihide Kitaoka

Associate Professor
Ryota Nishimura

Assistant Professor
Yukoh Wakabayashi

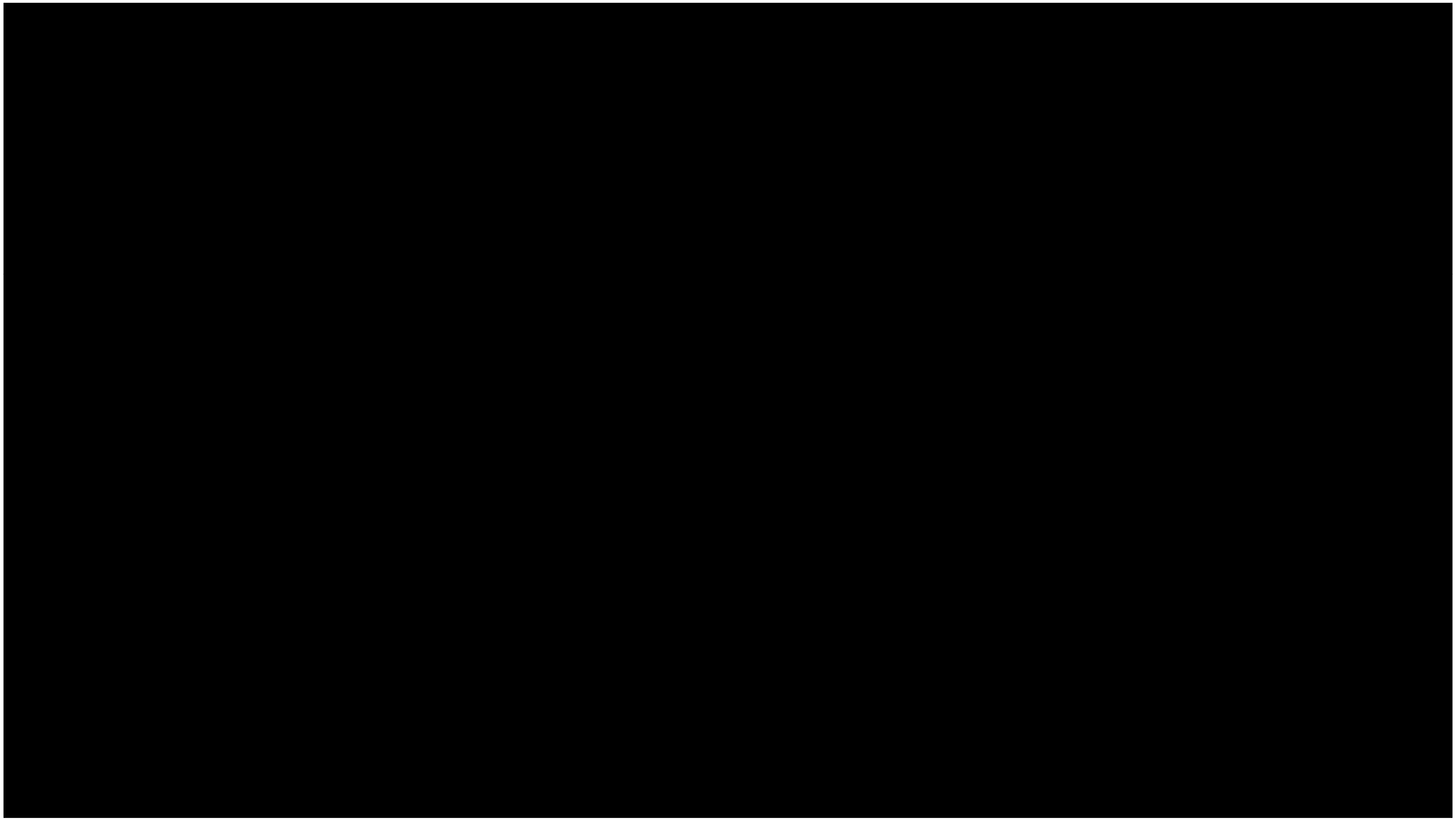
Students

D	3
M2	11
(2 International, 1 Indonesian)	
M1	10
(4 International)	
B4	11
(1 International)	

What do I do?

Speech recognition technology,
Speech Interfaces,
Spoken dialog systems,
Multimodal interaction systems

First speech recognition in iPhone



First speech recognition in iPhone (2)

First speech recognition in iPhone (2J)

What's AI ?

$x = 2$
$y = 3$
$z = 5$

$a = 5x + 2y + z$

$a = 21$

$x =$
"Who is the 35th president of USA?"

?

$a =$
"$John\ F.\ Kennedy.$"

Dialog engine

What's AI ?

$x = 2$
$y = 3$
$z = 5$

$a = 5x + 2y + z$

$a = 21$

$x =$ 

?

$a =$
「明日の天気は」

Speech recognition engine

What's AI ?

$x = 2$
$y = 3$
$z = 5$

$a = 5x + 2y + z$

$a = 21$

$x =$
「皆さんお疲れさまでした」

?

$a =$

Speech synthesis engine

Very complicated, (seems) intelligent functions!

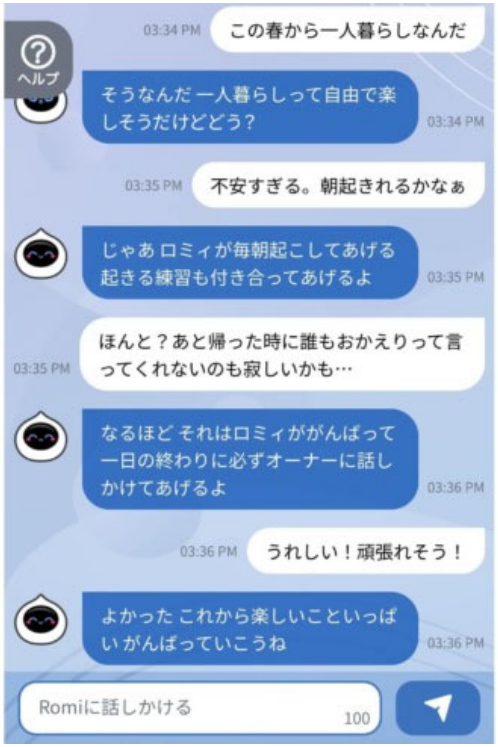
Let’s use speech synthesizer.

Let’s talk in text

開始 [redacted] ： 近年バタバタと過ごしてきたので、ゆっくりしたい、という

開始 Hiroe Kamogawa ： それはまさに「７の年」にぴったりですね。
開始 Hiroe Kamogawa ： そして
開始 Hiroe Kamogawa ： ７はあなたにとって、特に大切な数字でもありま

開始 Hiroe Kamogawa ： 「人生のメインテーマ」である軌道数も７、
開始 Hiroe Kamogawa ： 「社会に向けての自己表現」を表す表現数のも７
節目になるかもですね。
13:31:56 開始 [redacted] ： なるほど。
13:32:08 開始 Hiroe Kamogawa ： 軌道数が７ということは、
13:32:12 開始 [redacted] ： はい
13:32:35 開始 Hiroe Kamogawa ： 人生全般を通じて「穏やかな」生き方をマスターする、という風にも言えると思います。
13:32:57 開始 Hiroe Kamogawa ： 外側で起きていることを、静かに内側の自分が見つめているという意味で。
13:33:33 開始 Hiroe Kamogawa ： また「表現数」と「個人周期数」が重なる時は、いわばあなたの「本領発揮」の時、とも言えますよ！
13:33:46 開始 [redacted] ： 外側のことに、振り回されてきましたので、穏やかな生き方学びたいです。
13:34:06 開始 Hiroe Kamogawa ： ある意味、既に「ずっと学んできた」とも言えるんじゃないでしょうか。
13:34:22 開始 Hiroe Kamogawa ： そのことの「理解が深まる」一年ともなると思わ

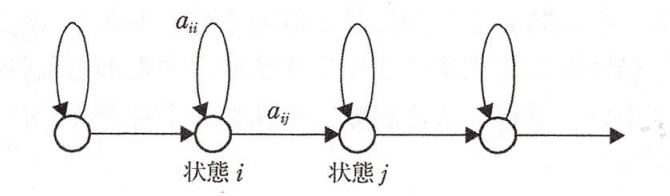


OpenAI
ChatGPT

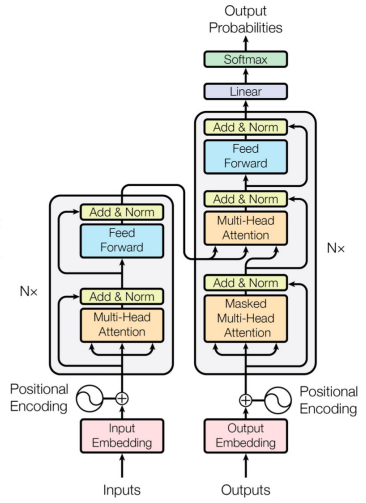
Let's use speech recognition.

My speech is recognized.

World's first speech recognizer: IBM Shoebox

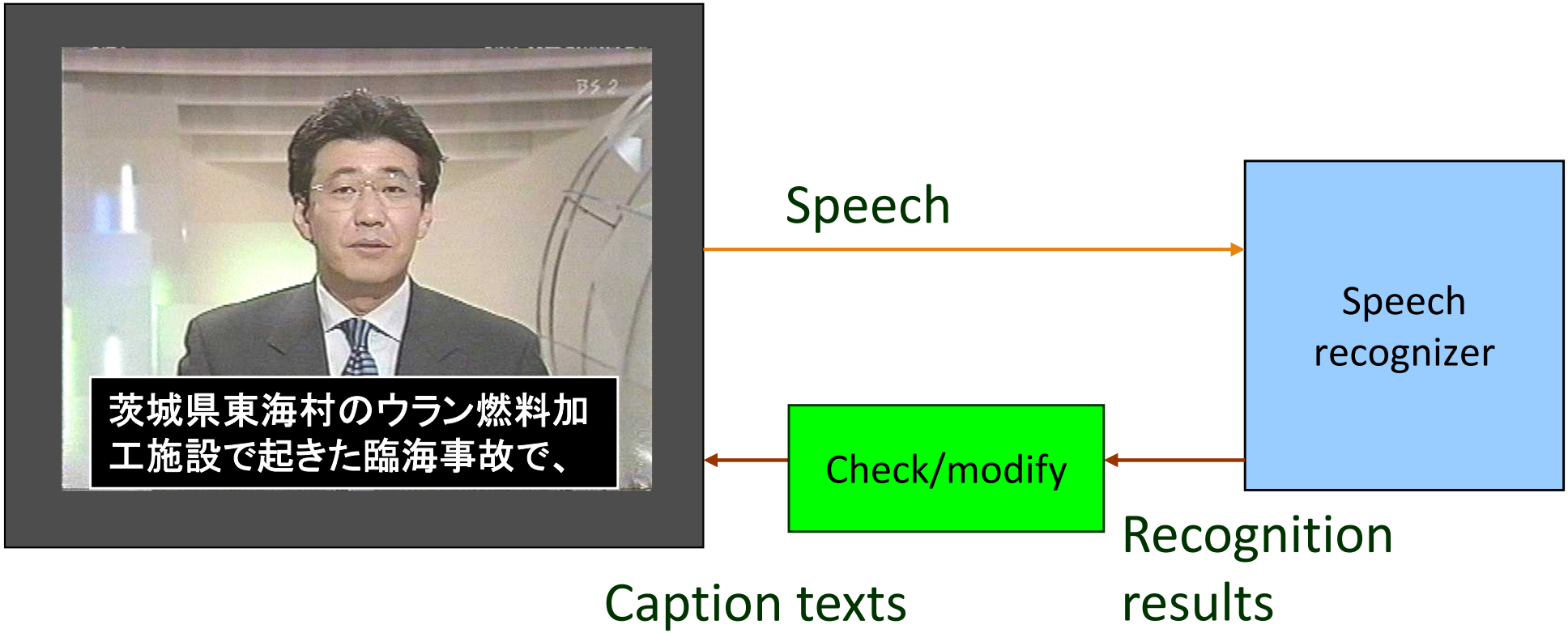


Hidden Markov model



Transformer

NHK captioning system



NHK captioning system - Respeaking -

Newscaster

Speech

Respeaker

Speech

Speech Recognizer

Speech

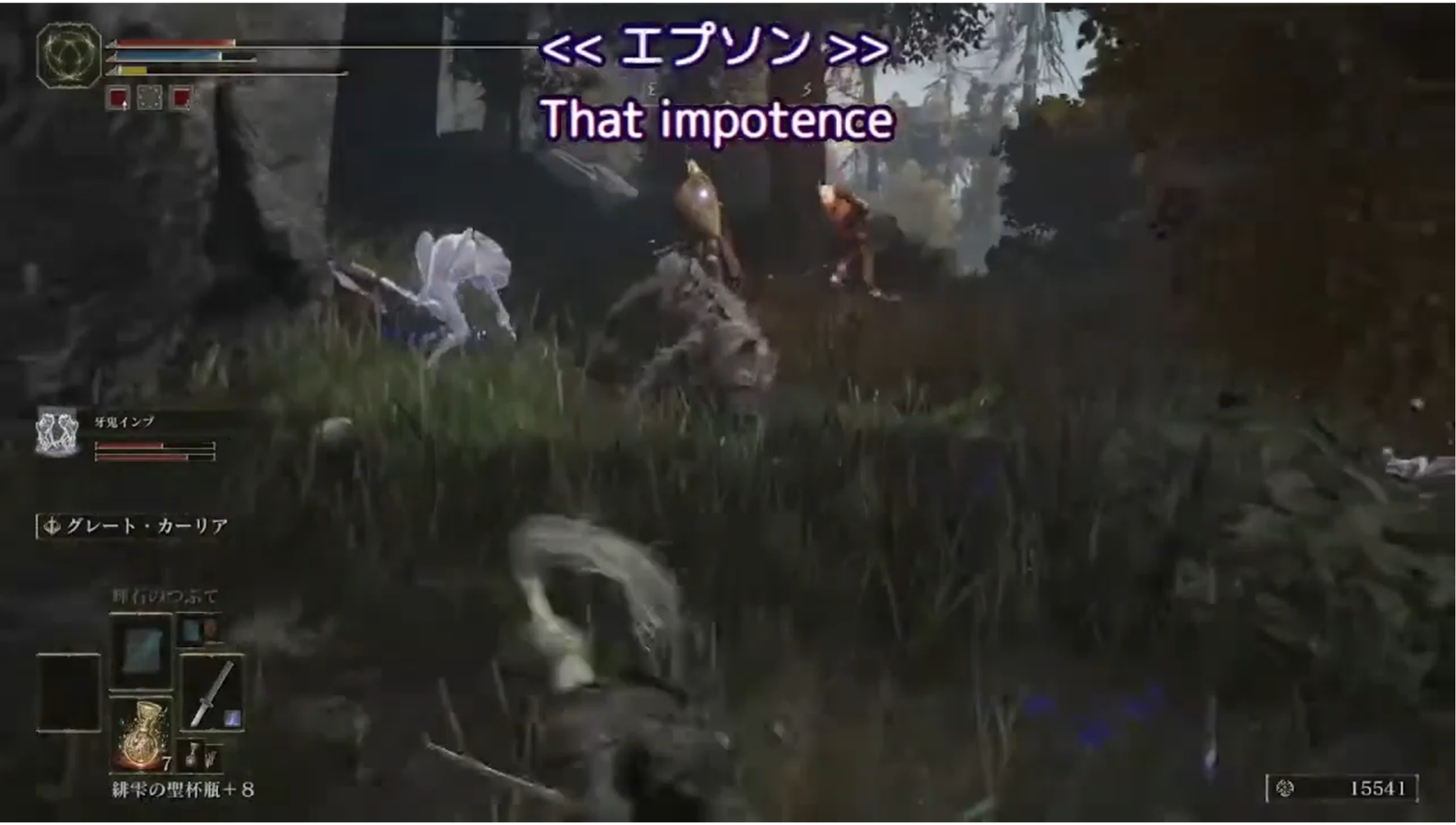
Captions

Recognition results

字幕作成システム（リスピーク方式）

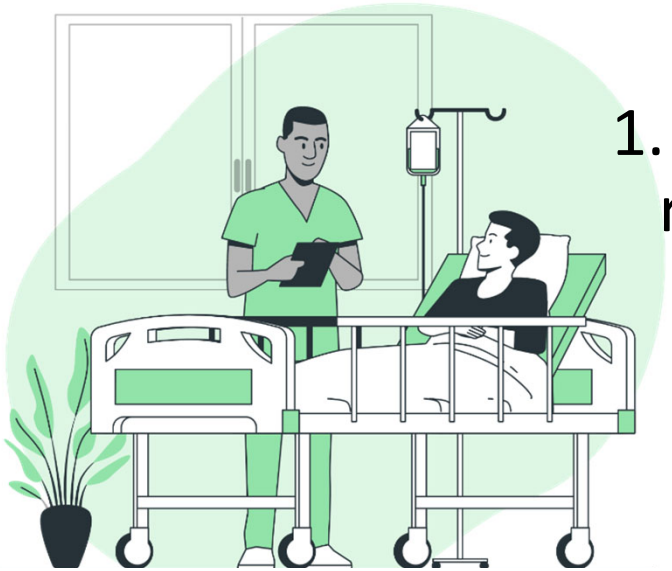
Replace!

Easily you can use captioning system
音声認識字幕ちゃん



Easily you can use captioning system
音声認識字幕ちゃん

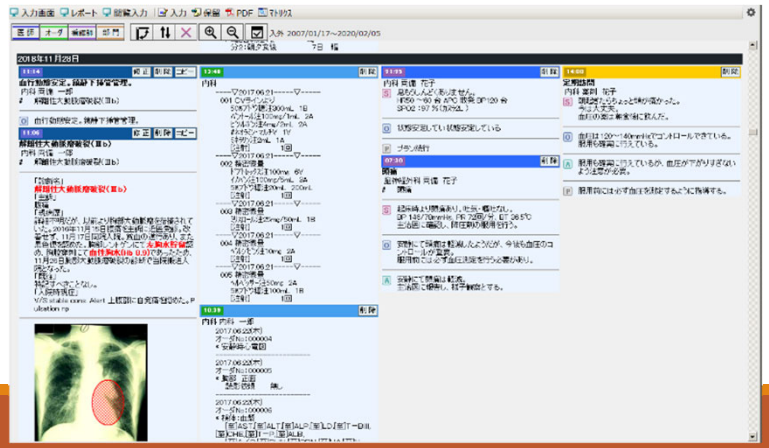
Electronic health record system using speech input



Speak the information on medical round

1. Speech recognition

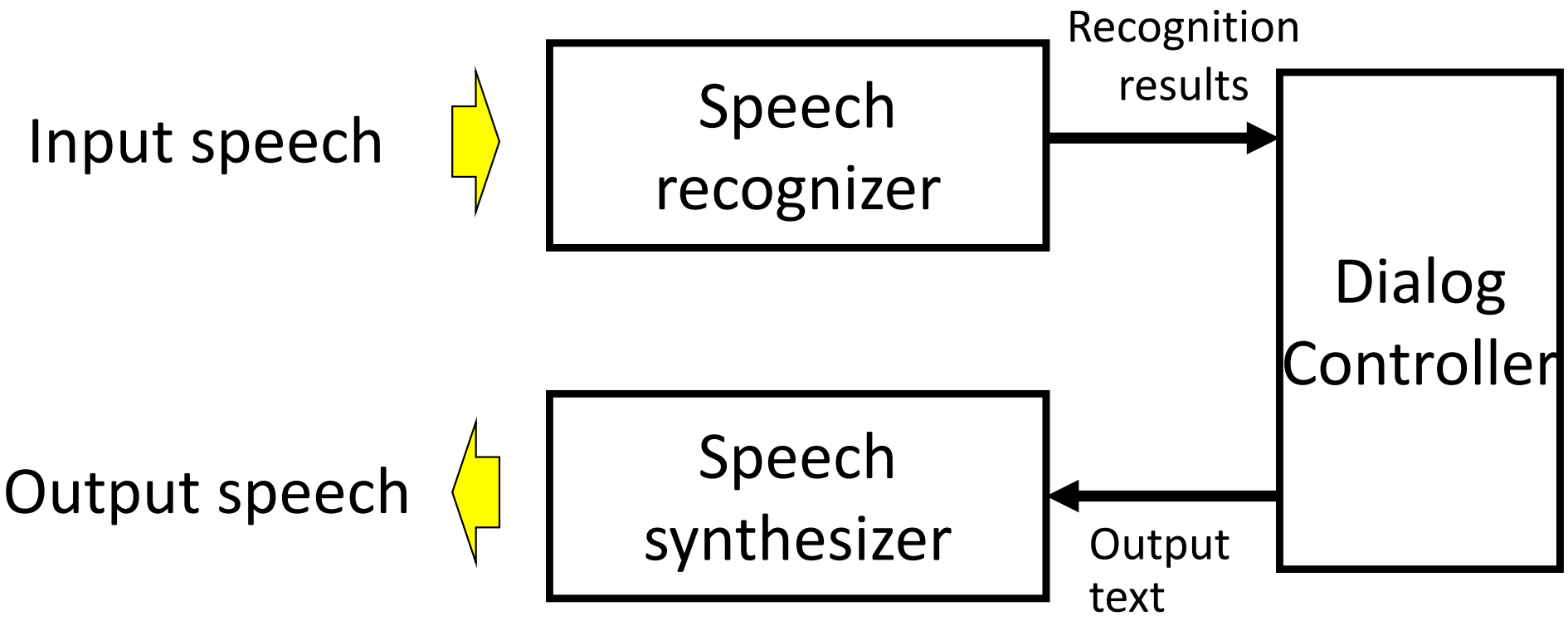
2. Structuring



3. Input to record

Spoken dialog

General structure of spoken dialog system



These are already introduced in this lecture.

⇒You can make a spoken dialog system very easily!

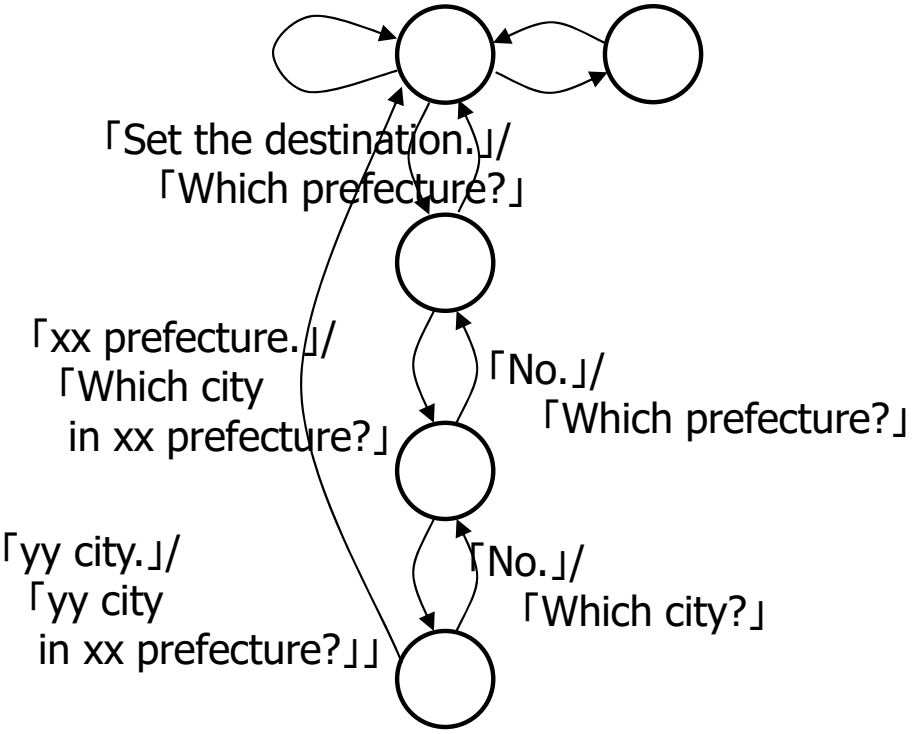
How to draw dialog flow?

Most traditional (but conventional) method:
Deterministic Finite State Automaton

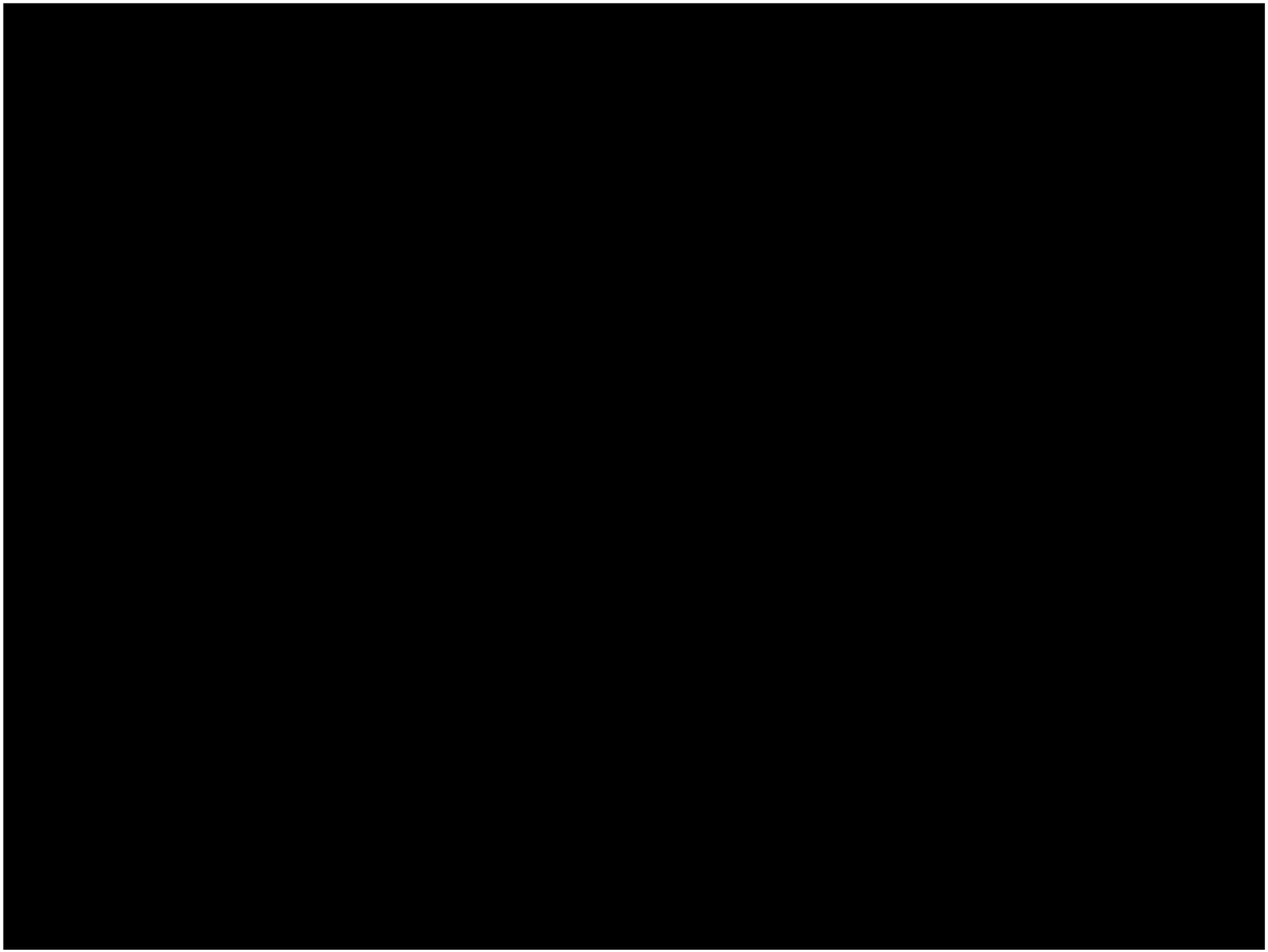
Each state describe the stage in a dialog.

Example

U:「Set the destination.」
S:「Which prefecture?」
U:「Aichi prefecture.」
S:「Which city in Aichi?」
U:「Toyohashi」
S:「Toyohashi, Aichi?」
U:「Yes.」
S:「Toyohashi Aichi is set as the destination.」

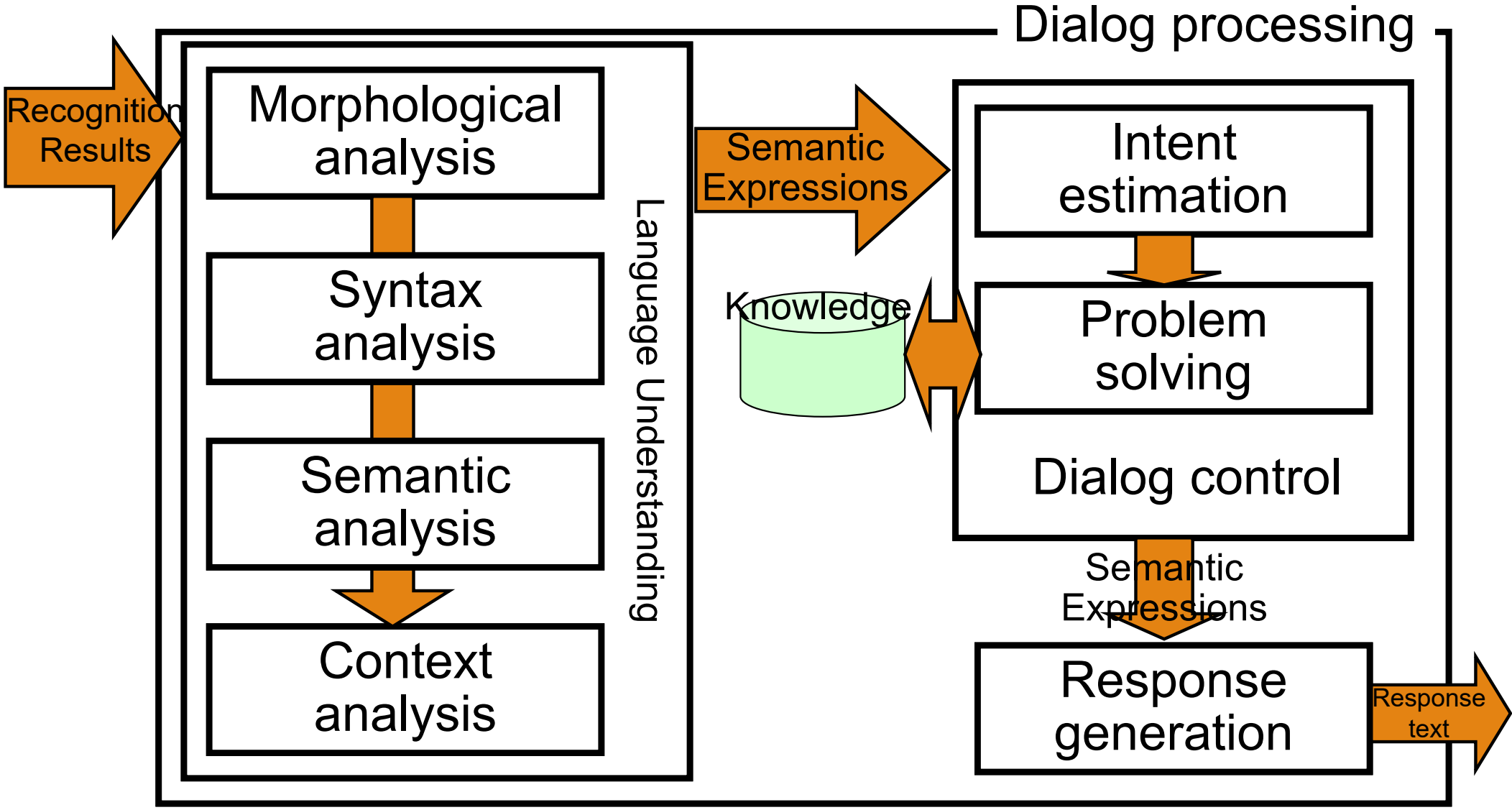


Airline Check-In Spoken Dialogue System developed at Heriot-Watt University



More complex dialog
-Mt. Fuji sightseeing information-

Classical dialog processing

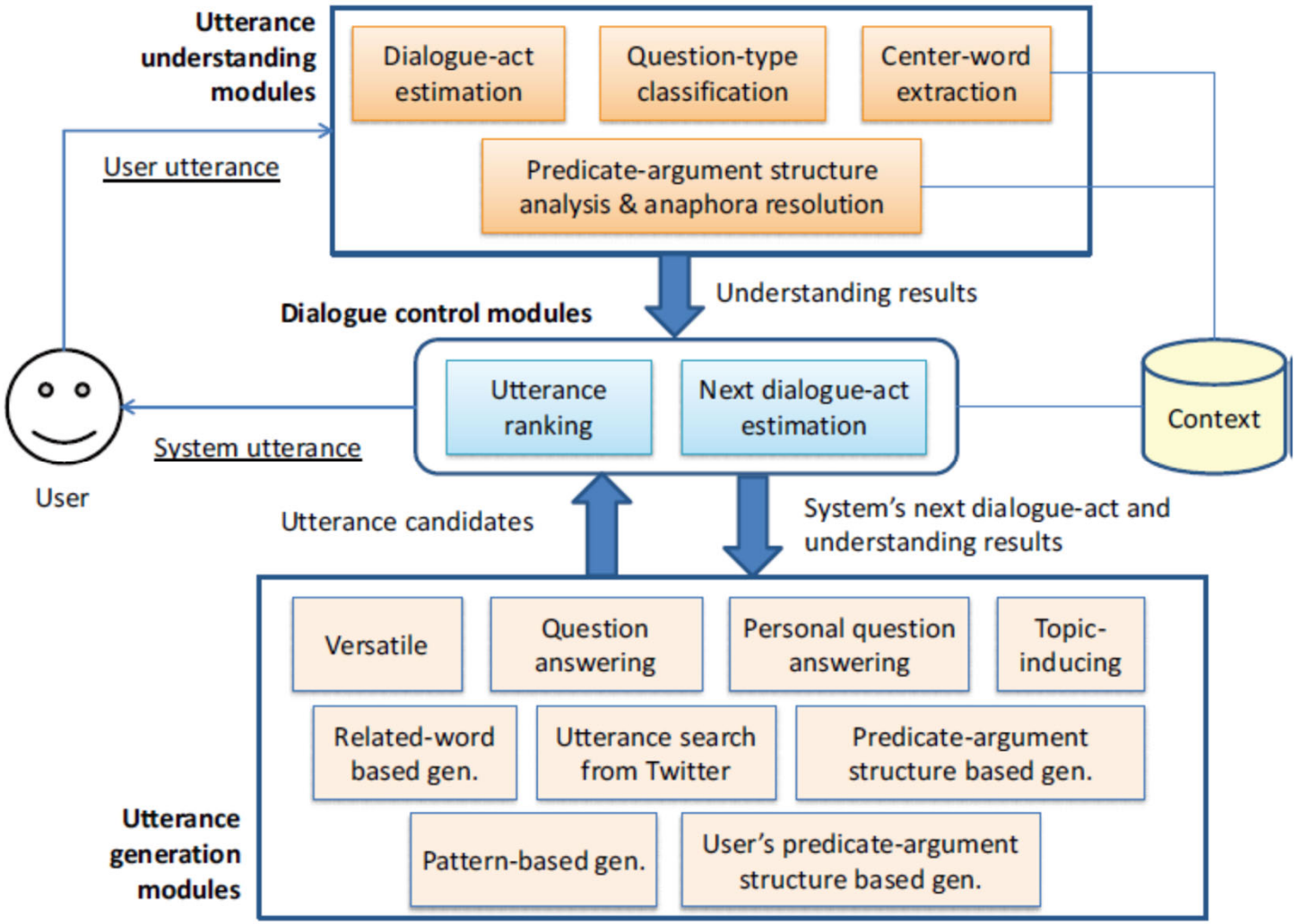


From task-oriented to “chat”

- Systems so far are used for achieving tasks.
  - Obtaining guidance
  - Setting a destination to a car navigation system
  - Hearing a music the user want to
  - Searching for a facility
- To be more familiar to machines (≒AI)
  - The system must make users feel “humanity”.
  - The system must convince the users.
  - The users want to “enjoy” the dialog itself.

Let’s make a system to talk in a joyful way.

Chat Bot API —NTT's chatbot—



API was publicly open (but now closed).

R. Higashinaka, et al. "Towards an open-domain conversational system fully based on natural language processing," Proceedings of COLING 2014, 2014

What is the future spoken dialog?

Trend in spoken dialog systems

- Intelligent dialog
  →Automaton
    →Example-based(+ α)
      →Chat (+ Task-achievement)

Dialog itself should be enjoyable/attractive.

What should we do?
More intelligence?
Witty comments?
Speaking timing?

Recent advance

DNNs/RNNs for dialog systems

- End-to-end dialog management

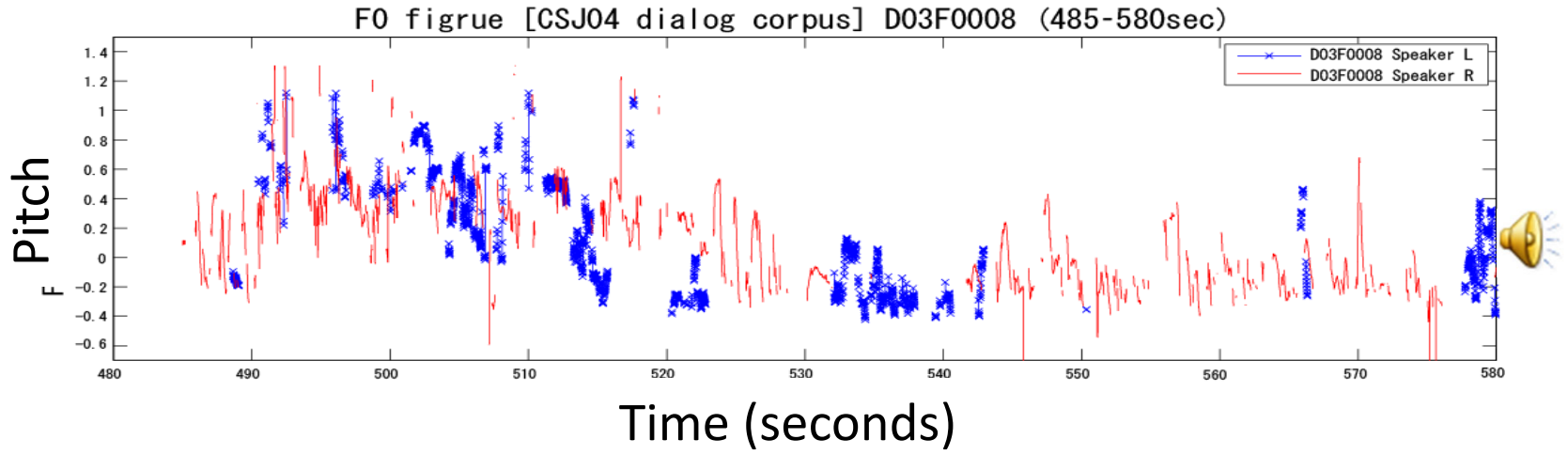
NTT's chat system (equivalent with GPT-3)

How do you feel?

Natural?

Enjoyable?

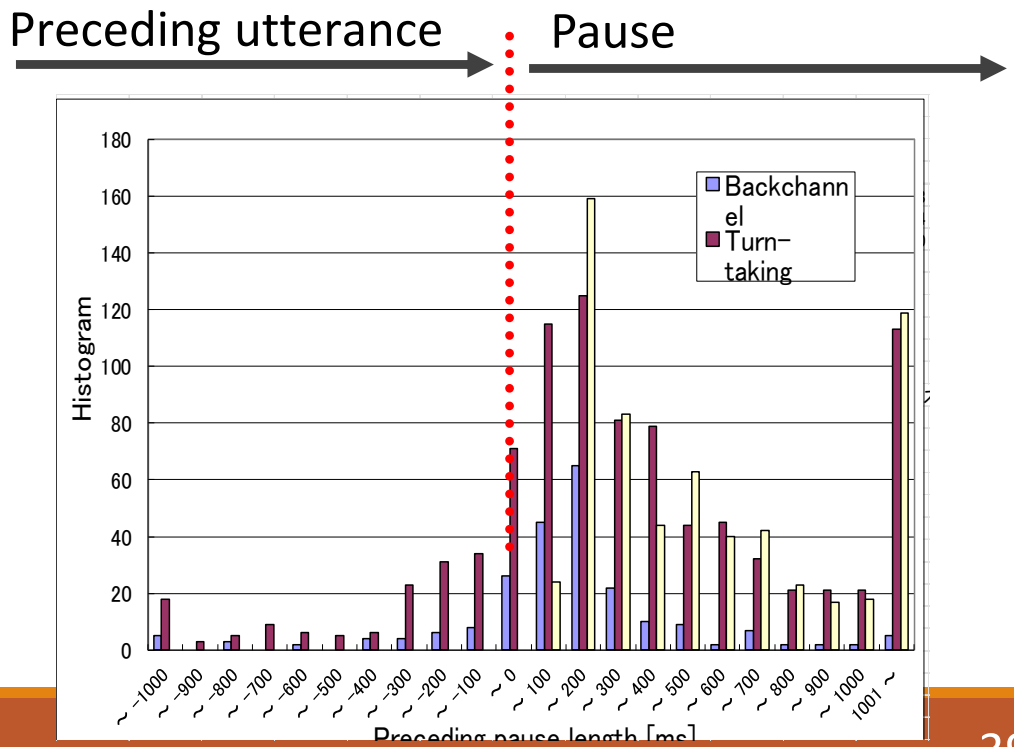
Speaking timing in dialog —human-to-human—



Human-human daialog

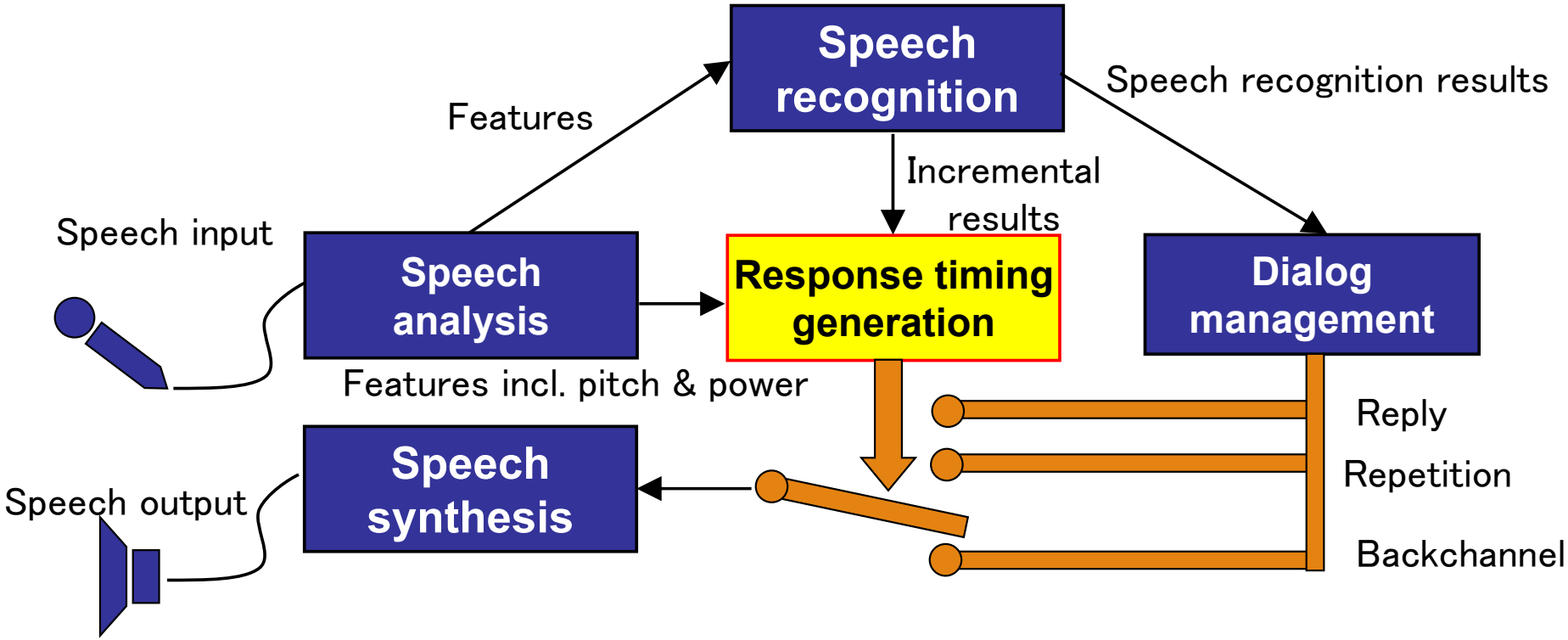
- Various length of pauses
- Sometimes overlap to the opponents' utterance

Various timings are used.



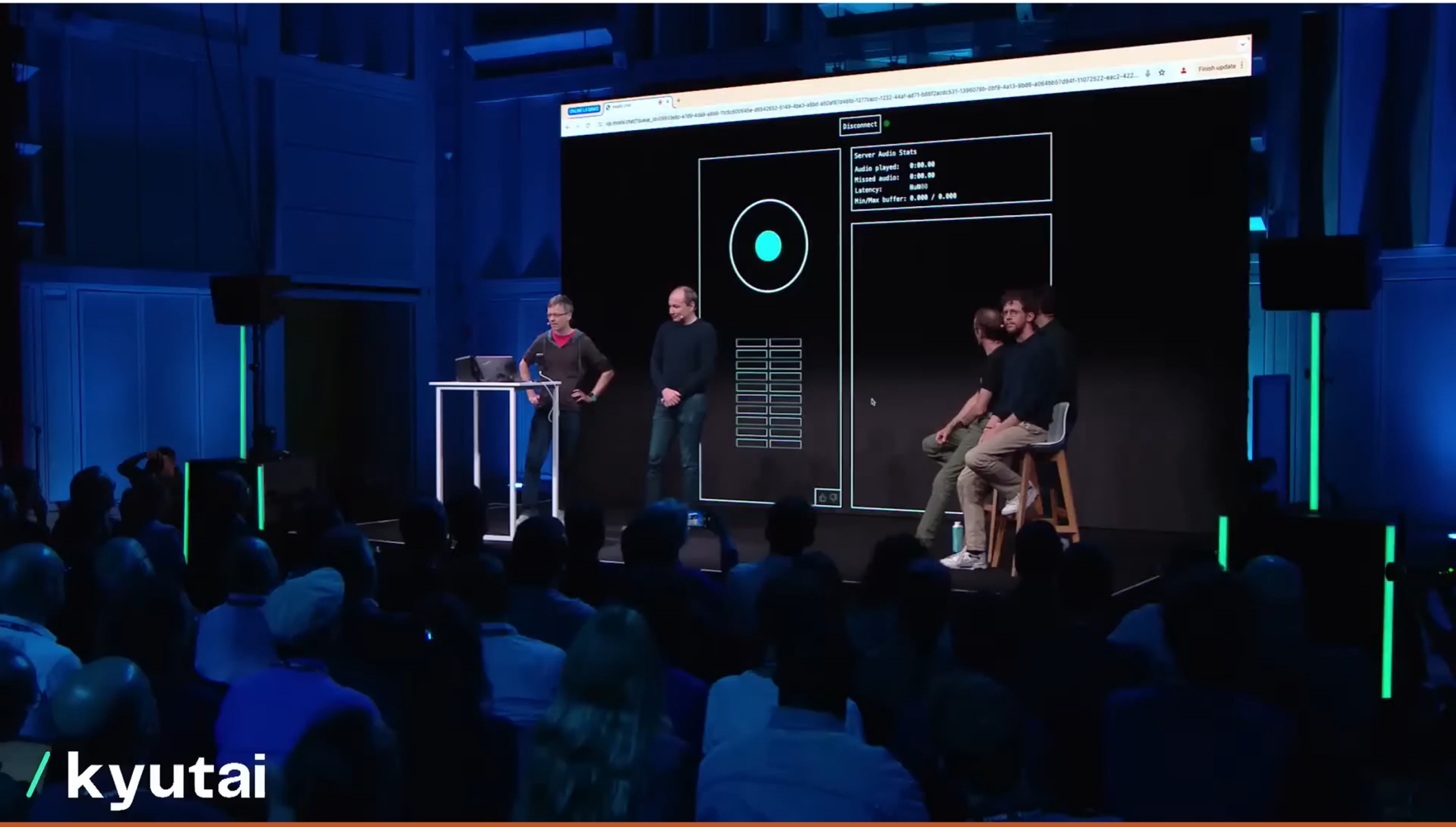
A spoken dialog system considering utterance timing

Timing generation in a spoken dialog system



State-of-the-art approach

End-to-end speech-to-speech dialog systems



TOYOHASHI
UNIVERSITY OF TECHNOLOGY

Multimodal Interface/Interaction

What is “Multimodality”?

“We can use the system only by voice” is one of solutions.

- However…

Humans communicate with each other using various methods/modalities

- Voice
- Finger pointing
- Facial expressions

Multimodal interaction

What is “Multimodality”?

Multimodal:

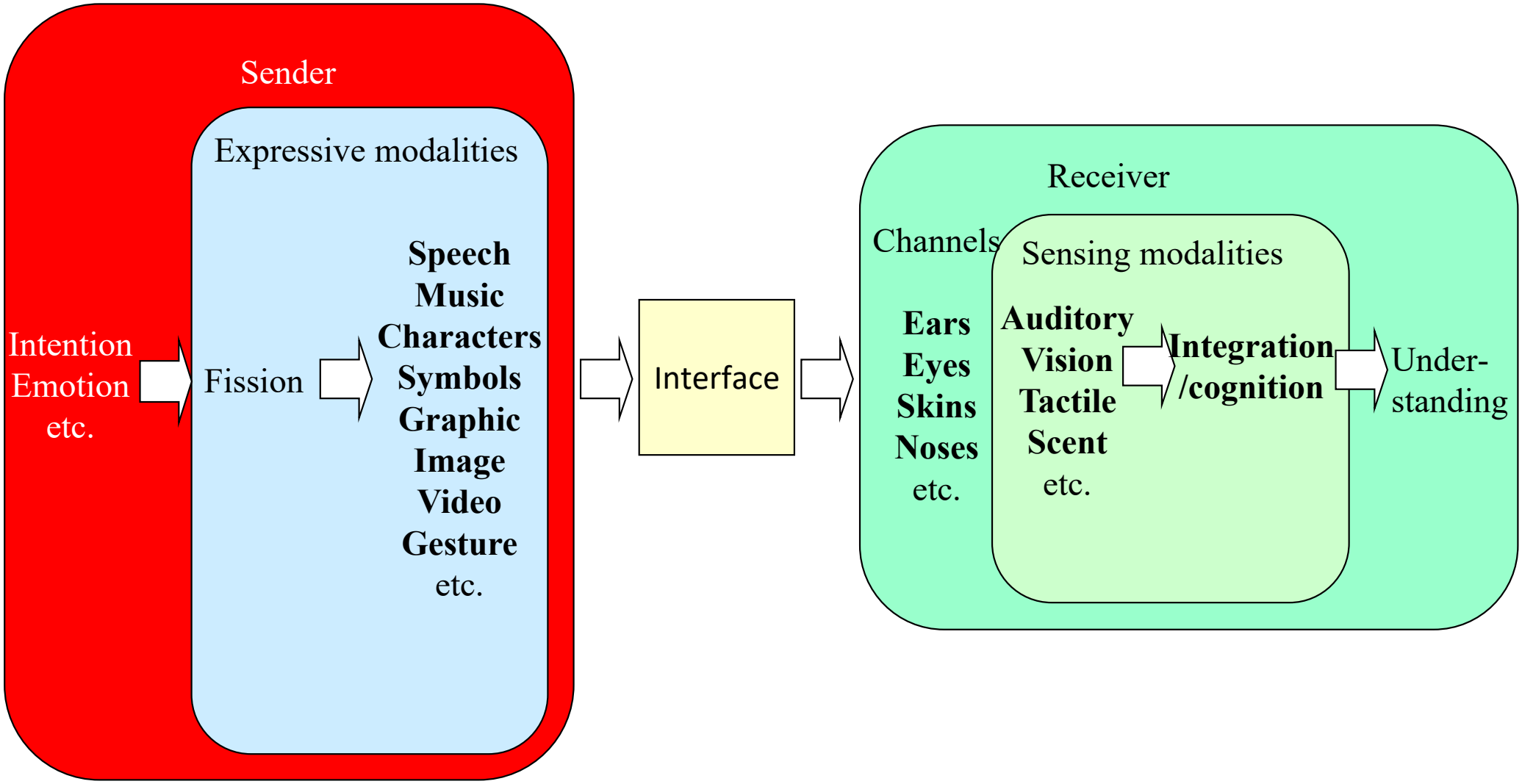
Using multiple modalities

Modality:

Humans’ perceptional aspects

- Linguistics, speech, finger pointing, …

Multimodal situation



Multimodal interaction

Multimodal Interaction: MMI

≒ Dialog using multiple modalities

- Human-human dialog
- Operation of PCs using speech and pointing
- Etc.

Modalities in human-human communication

TOYOHASHI UNIVERSITY OF TECHNOLOGY

- Verbal modalities
  - Natural language
  - Spoken language
- Non-verbal modalities
  - Gestures
  - Facial expressions
  - Gaze
  - Prosody

Gestures (Ekman '69)

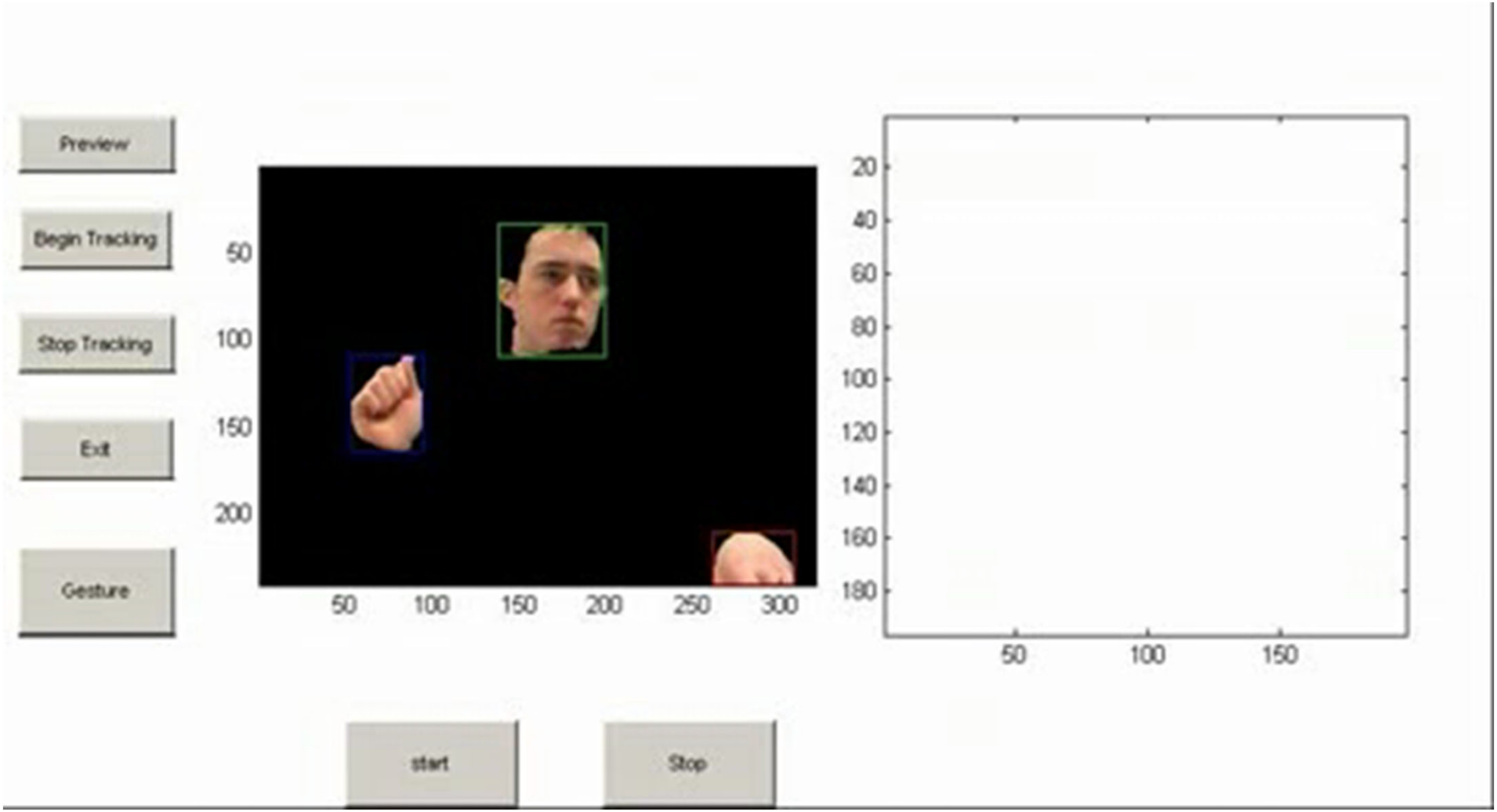
1. Emblems – generally translated directly into words
   Ex. Block sign, sign language
2. Illustrators – movements that complement verbal communication by describing or accenting or reinforcing what the speaker is saying
   Ex. Pointing to an object
3. Regulator – nonverbal messages that accompany speech to control or regulate what the speaker is saying
   Ex. nodding of the head to indicate you are listening or understanding something
4. Affected display – carry an emotional meaning or display affective states
   Ex. Big grin
5. Adaptor – forms of nonverbal communication that often occur at a low level of personal awareness
   Ex. twisting your hair, tapping your pen, scratching, tugging on your ear

Research on gestures (1)

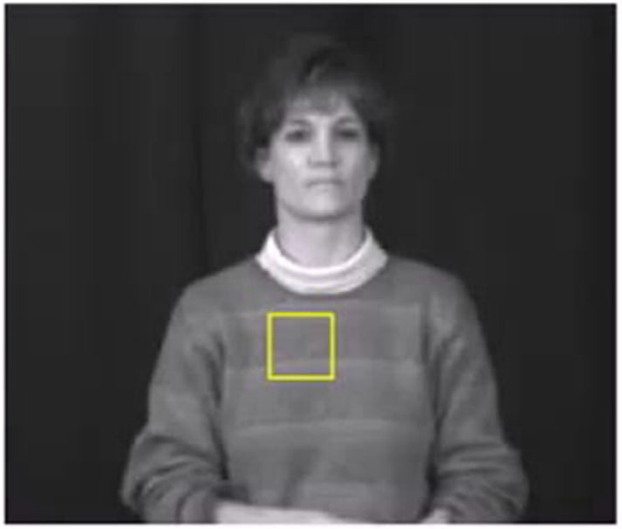
Emblems

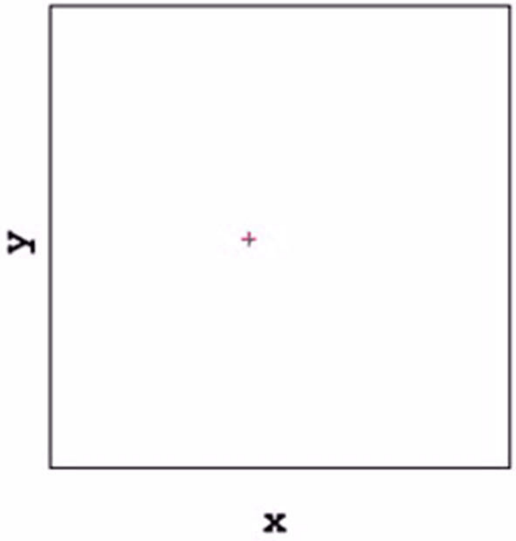
- Translated directly into words
- Sign language Recognition

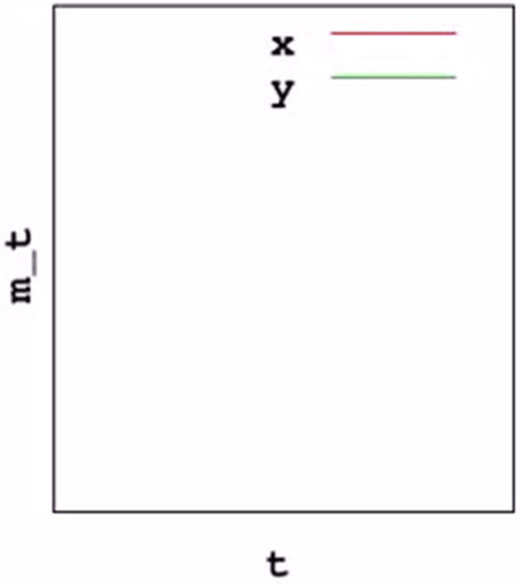
Sign Language Recognition



Sign Language Recognition







JOHN FISH WONT EAT BUT CAN EAT CHICKEN

Research on gestures (2)

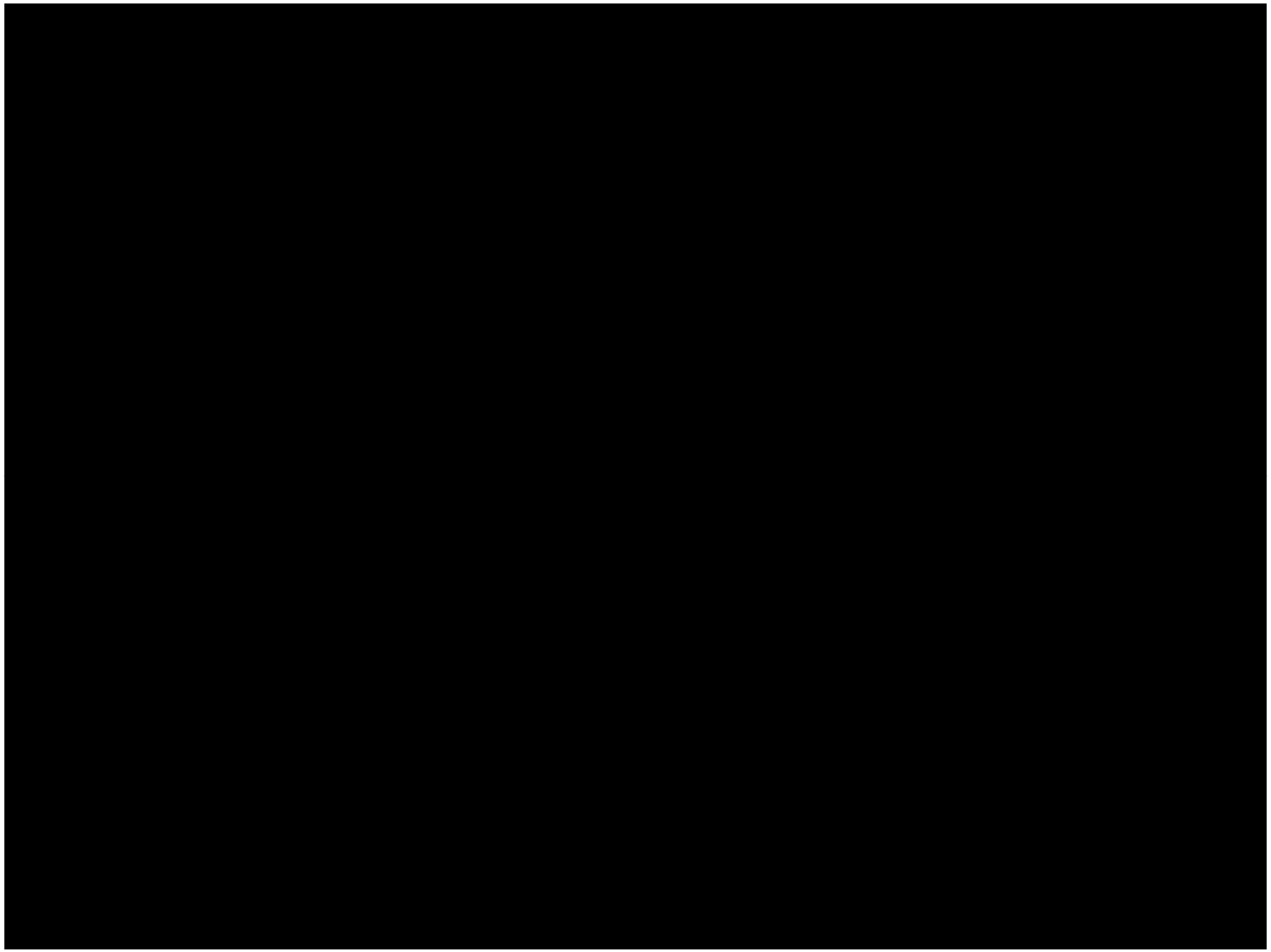
Illustrators

Movements that complement verbal communication by describing or accenting or reinforcing what the speaker is saying

Put-That-There (Bolt '80)

- First multimodal interface
- Moving the objects on the display using speech and pointing

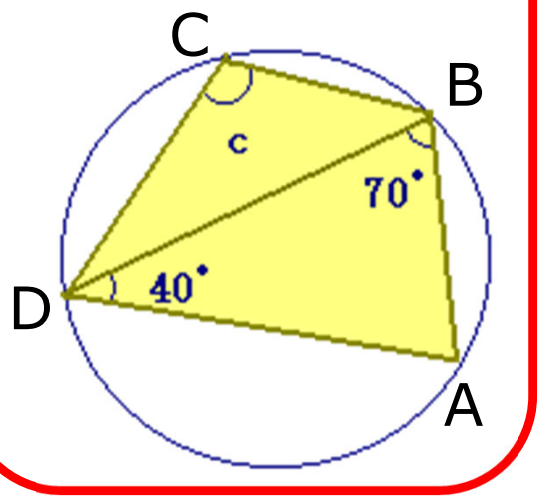
Put-that-there

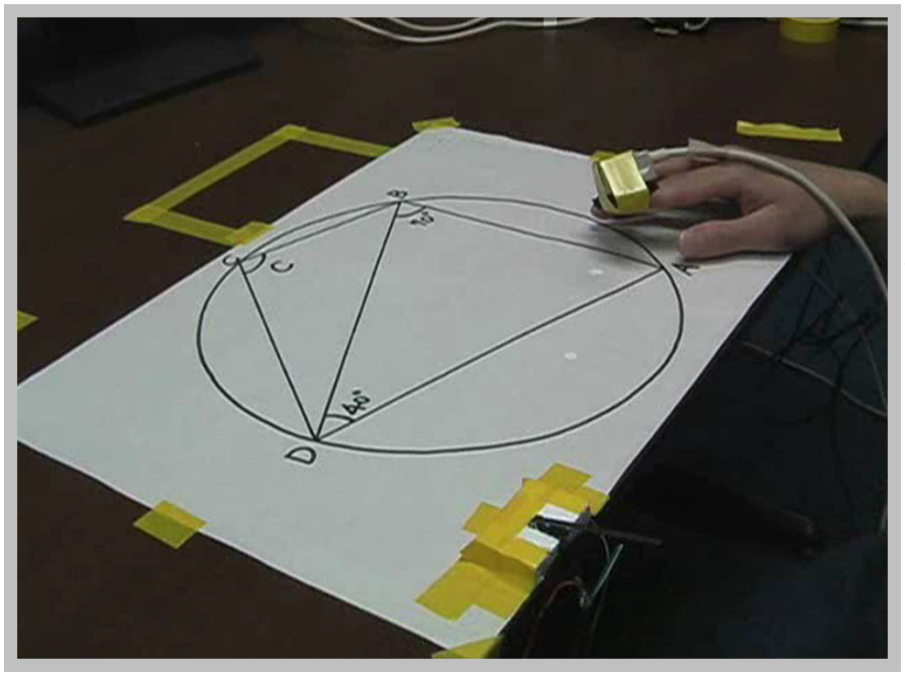


R. A. Bolt, “Put-that-there”: Voice and gesture at the graphics interface, 1980.

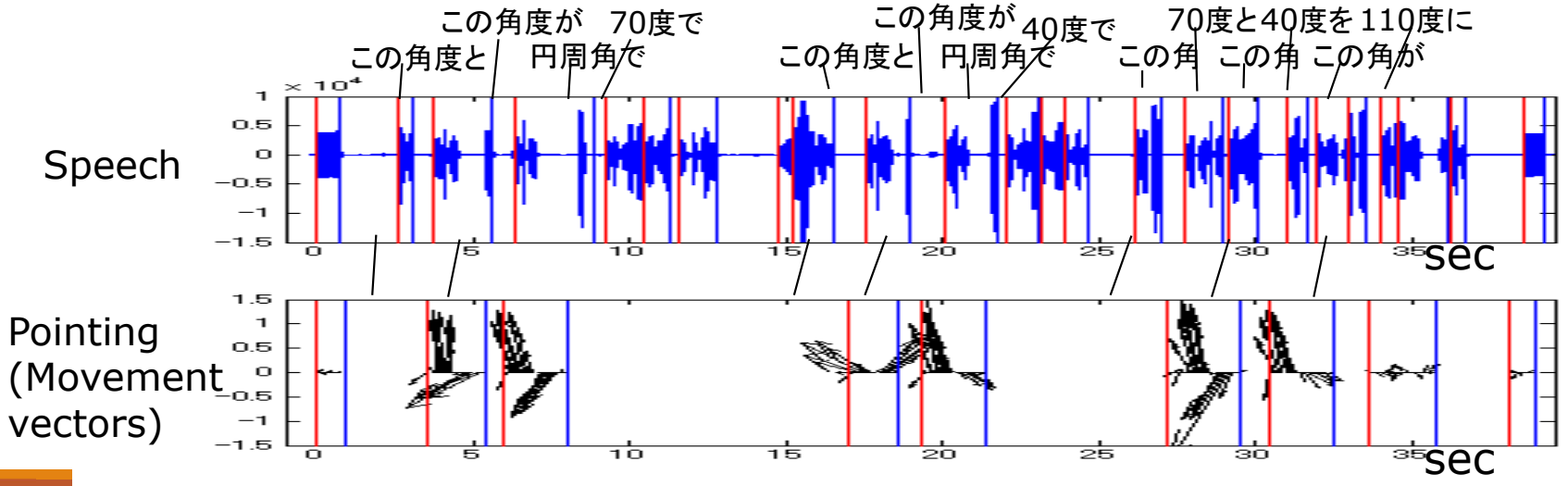
Complementary use of speech and gesture

Calculate the angle c

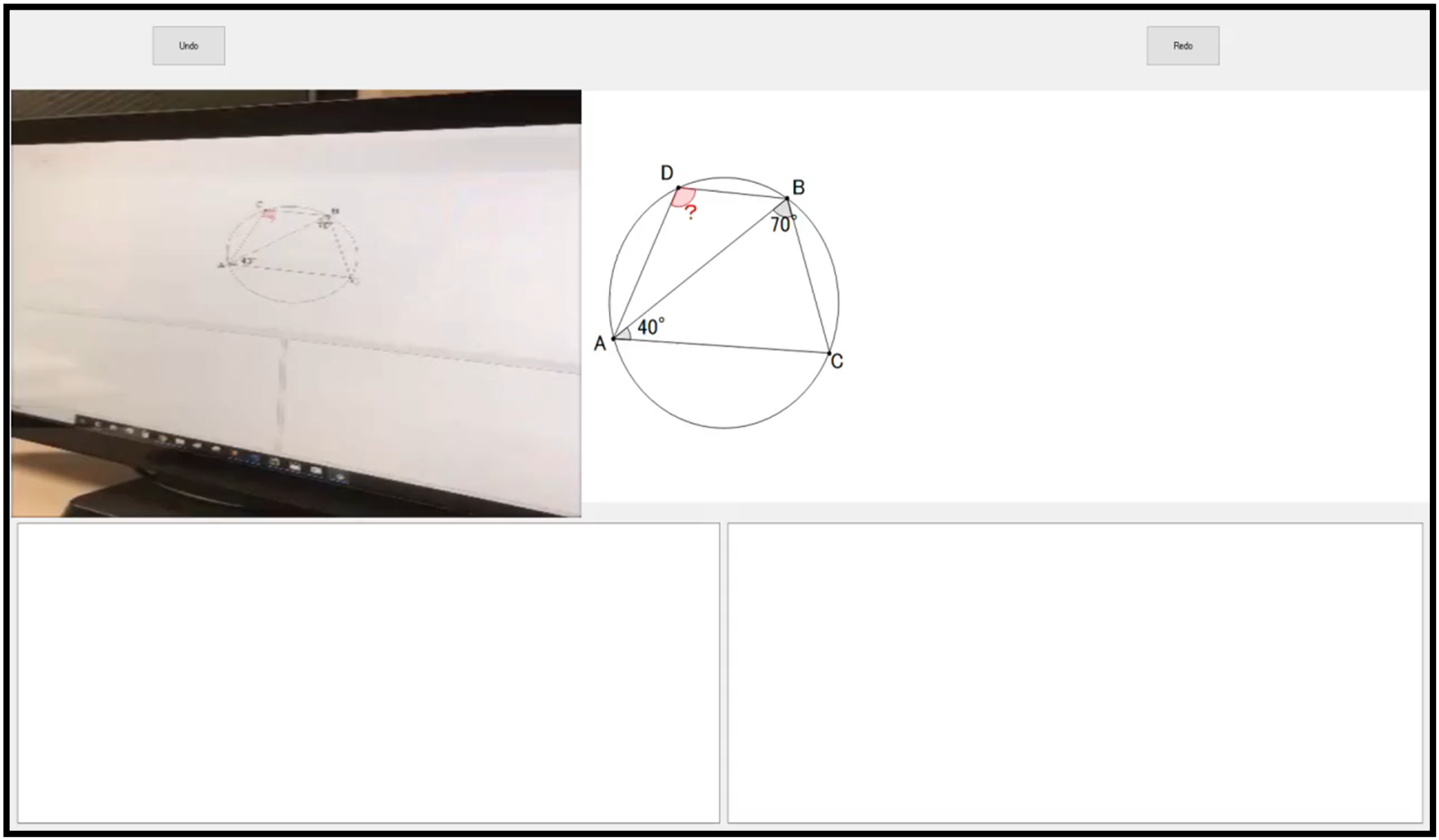




"This angle" with pointing



Complementary use of speech and gesture

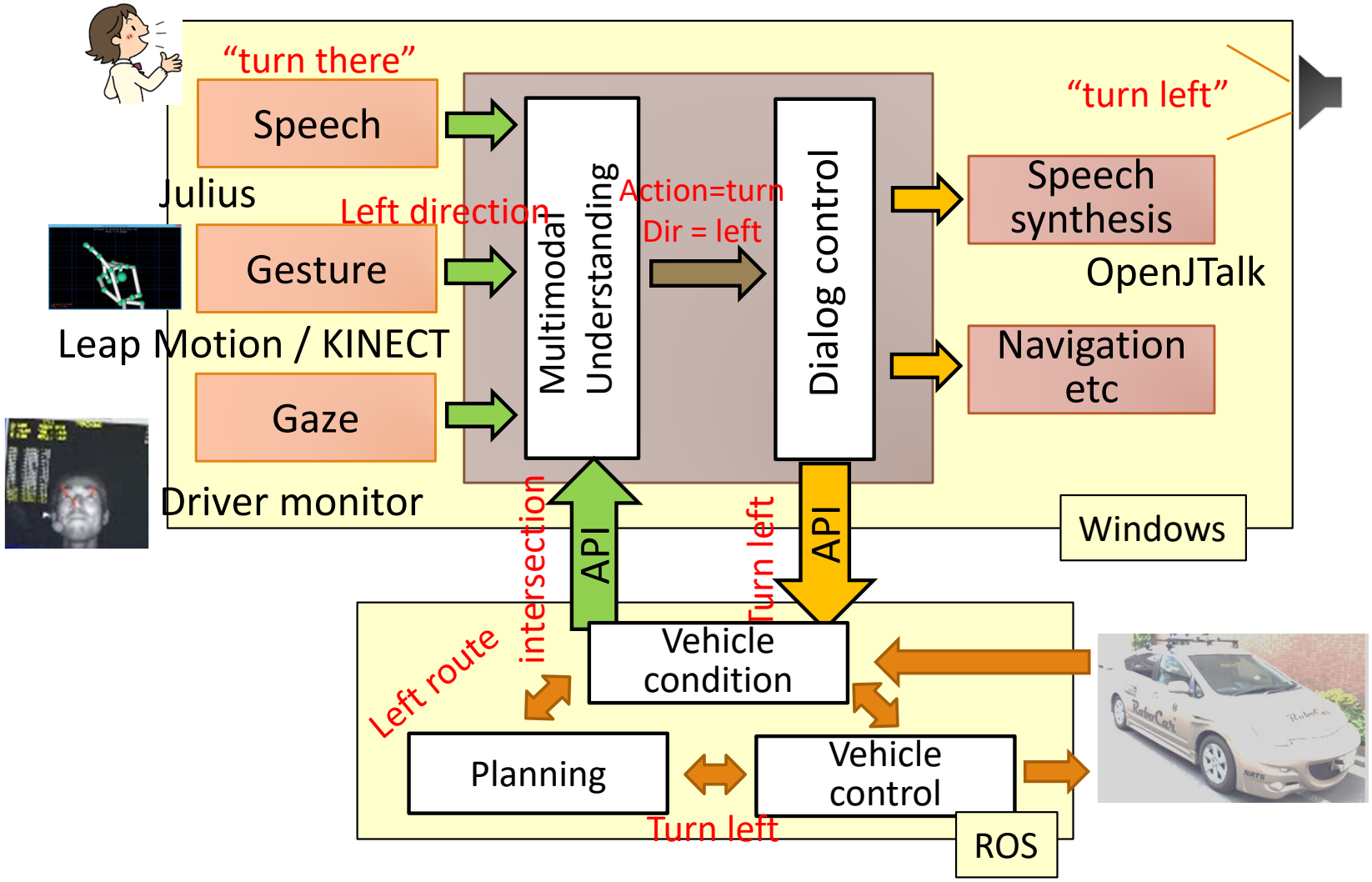


Research on gestures (3)

- Regulator
  - Nonverbal messages that accompany speech to control or regulate what the speaker is saying
    - Nodding of the head to indicate you are listening or understanding something

Multimodal interaction with autonomous vehicle

Block diagram of system



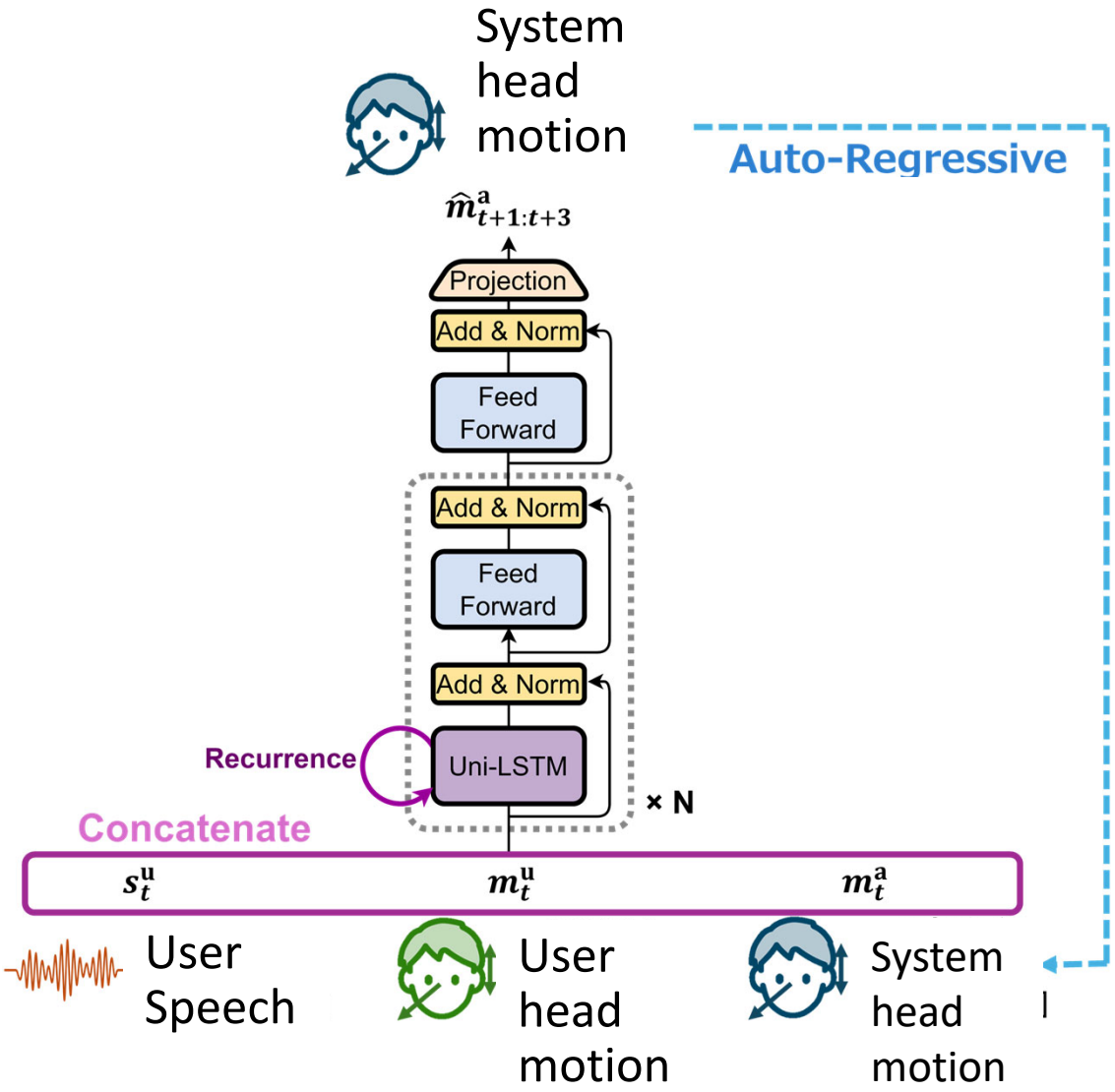
Real human-human interactions

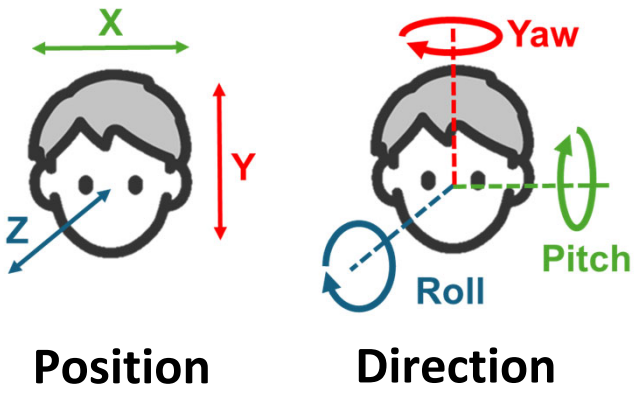
Listener responds with good timings

Speaker

Listener

Listening head generation model





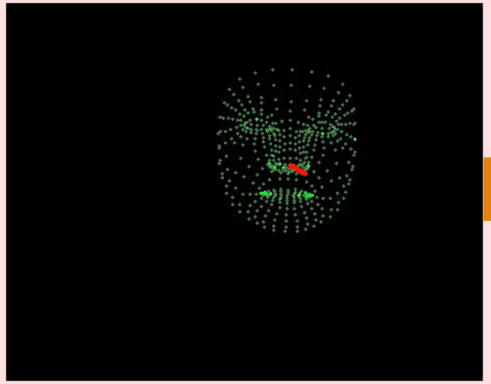
Generation results

Speaker

▲Real human

Same movement of human

Listener



▲ Generation result ▲

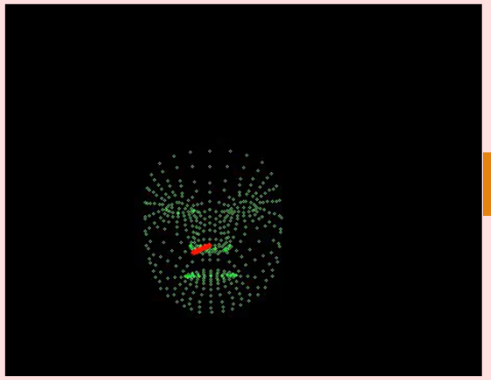
Generation results

Speaker

▲Real human

Same movement of human

Listener



▲ Generation result ▲

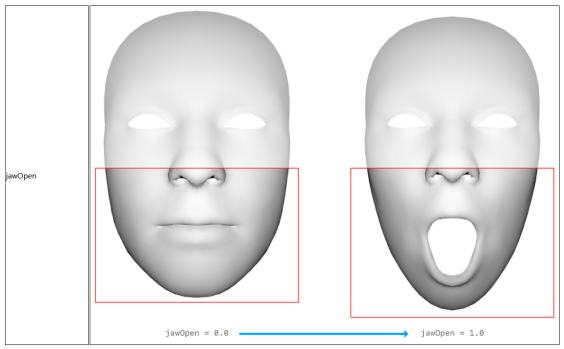
Real human-human interactions

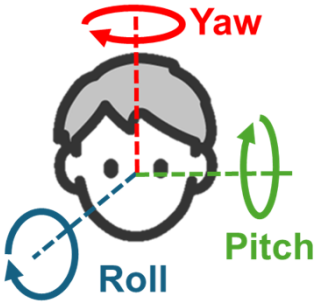
Speaker also uses head movement with good timings

Speaker

Listener

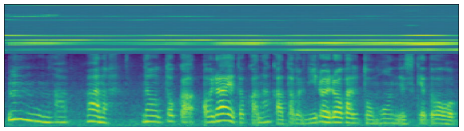
Talking head generation





Direction

Target spectrogram



Flow-prediction Network

↓Text input

あいうえお

a i u e o

Text Encoder

μ

Monotonic Alignment Search

↑Speech and movement output

Duration Predictor

Ldur

Stop gradient

Generation results

What is the future spoken dialog? (Revisiting)

Dialog itself should be enjoyable/attractive.

What should we do?
More intelligence?
Witty comments?
Speaking timing?

I don't know.

Please predict the future.

Towards Multimodal Era

Update!

TOYOHASHI UNIVERSITY OF TECHNOLOGY

Laboratory information here

Laboratory site
http://www.slp.cs.tut.ac.jp

Demonstrations