Multi-modal Sensing and Analysis of Poster Conversations Toward Smart Posterboard

Tatsuya Kawahara

(Kyoto University, Japan)

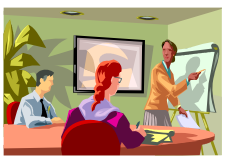
http://www.ar.media.kyoto-u.ac.jp/crest/

Directions in Dialogue Research

(Engineering Applications in mind)

- Speech-only
- Dyadic
- Human-Machine Interface

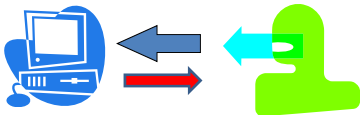
- Multi-modal
- Multi-party
- Human-Human Interaction



Human-Machine Interface

constrained speech/dialog

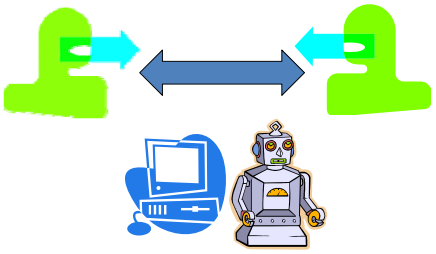
- task domain
- one sentence per one turn
- clear articulation



Human-Human Communication

natural speech/dialogue

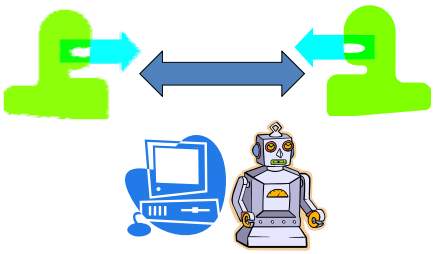
- many sentences per one turn
- backchannels



Project Overview

Problems

"Understanding" of human-human speech communication



- Speaker Diarization
- Speech-to-Text (ASR)
- Dialogue Act (?)

- Comprehension level
- Interest level

Goal (Application Scenario)

Mining human interaction patterns

- A new indexing scheme of speech archives
  - Review summary of QA
  - Portion difficult for audience to follow (→ presenter)
  - Interesting spots (→ third-party viewers)

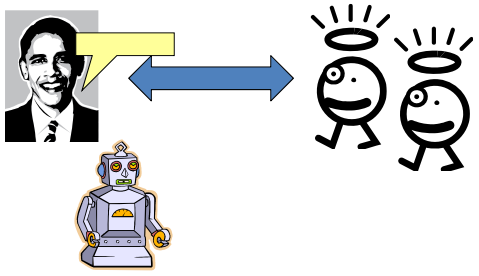
    "People would be interested in what other people were interested in."
- A model of intelligent conversational agents (future topic)

From Content-based Indexing to Interaction-based Indexing

- Content-based approach
  - try to understand & annotate content of speech…ASR+NLP
  - Actually hardly "understand"

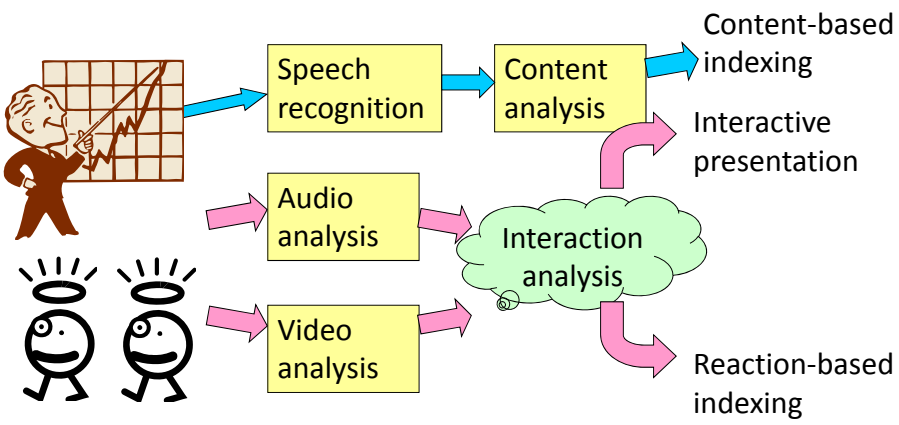
- Interaction-based approach
  - look into reaction of listeners/audience, who understand the content
  - More oriented for human cognitive process

From Content-based Approach to Interaction-based Approach

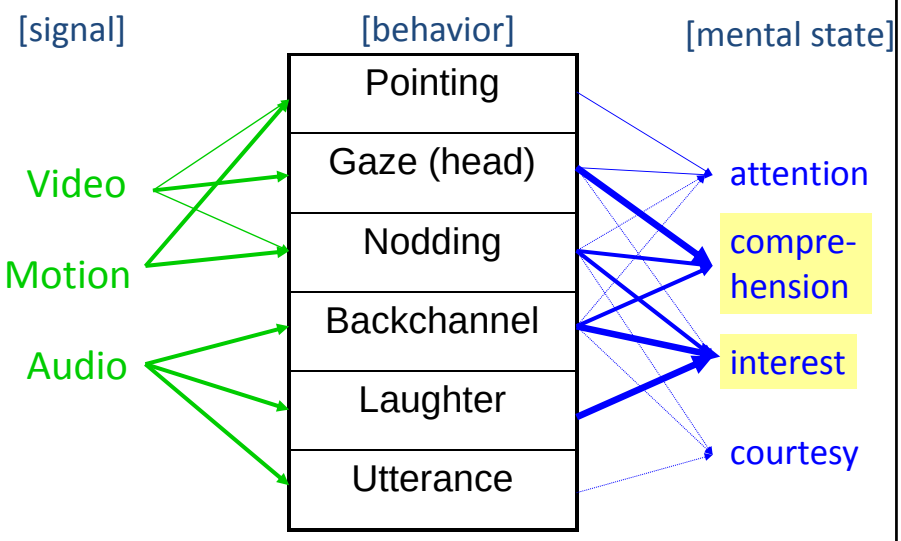


- Even if we do not understand the talk, we can see funny/important parts by observing audience's laughing/nodding
- Page rank is determined by the number of links rather than by the content

System Overview



Multi-modal Sensing & Analysis



Methodology

- Sensing devices
  - Gold-standard: special devices worn by subjects
  - Final system: distant microphones & cameras
- Milestones for high-level annotation

  "Good reactions" → "attracted"
  - Reactive tokens → interest level
  - when & who asks questions → interest level
  - kind of questions → comprehension level

Multi-modal Corpus of Poster Conversations

Why Poster Sessions?

- Norm in conferences & open-houses
- Mixture characteristics of lectures and meetings
  - One main speaker with a small audience
  - Audience can make questions/comments at any time
- Interactive
  - Real-time feedback including backchannels by audience
- Multi-modal (truly)
  - Standing & moving
- Controllable (knowledge/familiarity) and yet real

Multi-modal Sensing Environment: IMADE room

- Wire-less head-worn microphone
- Microphone array mounted on poster stand

} Audio

- 6-8 cameras installed in the room

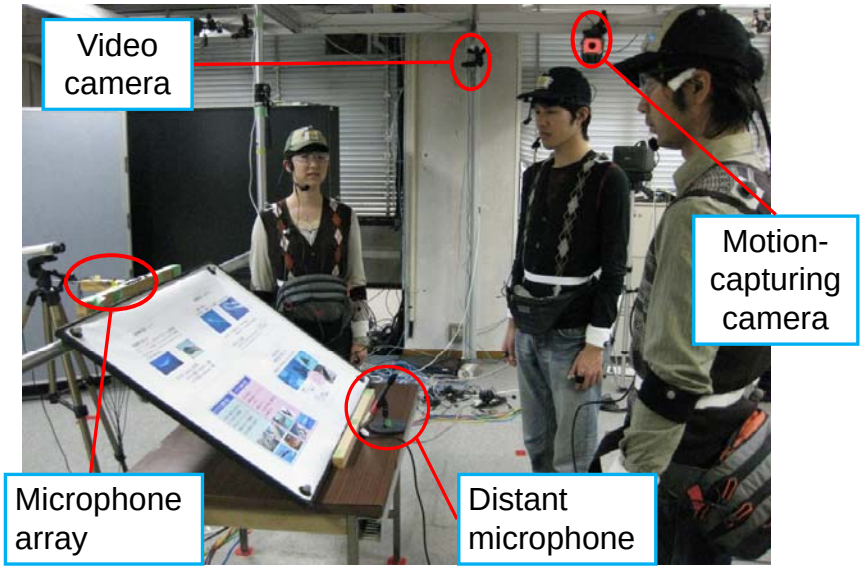
} Video

- Motion-capturing system
- Accelerometer

} Motion

- Eye-tracking recorders → Eye-gaze

Multi-modal Recording Setting



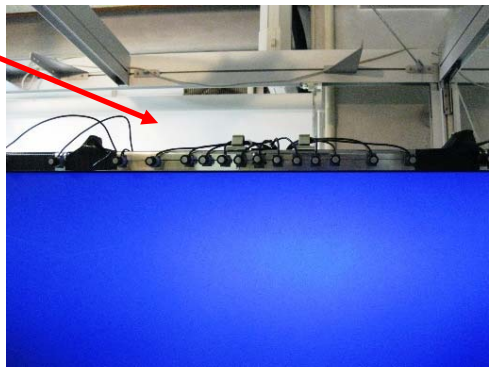
Multi-modal Recording Setting

Prototype of Smart Posterboard

65’ LCD Screen + Microphone Array + Cameras

Microphone Array mounted on LCD Posterboard

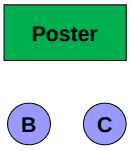
19-channel microphone array



Pre-amplifier
AD converter

Corpus of Poster Conversations

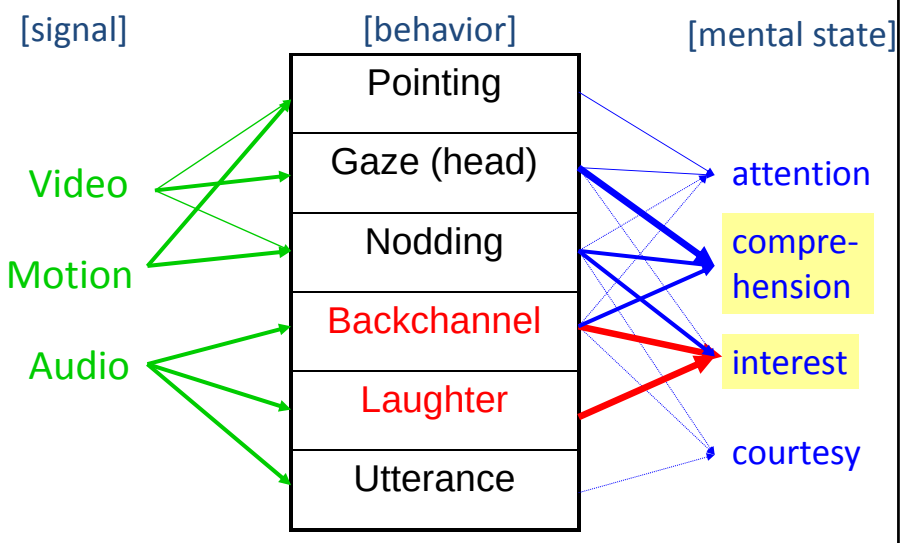
- 31 sessions recorded → 4 used in this work
  - One presenter + audience of two persons
  - Presentation of research; unfamiliar to audience
  - Each 20 min.



- Manual transcription
  - IPU, clause unit
  - Fillers, Backchannels (reactive tokens), Laughter
- Non-verbal behavior labels (almost automated!!)
  - Nodding…non-verbal backchannel ← accelerometer
  - Eye-gaze (to other person & poster) ← eye-track rec.
  - Pointing (to poster) ← motion cap.

Detection of Interest Level with Reactive Tokens of Audience

Multi-modal Sensing & Analysis



Reactive Token of Audience

- Reactive Token (*aizuchi*)
  - short verbal responses made in real time and backchannel
  - focus on non-lexical kinds (ex.) "uh-huh", "wow"
  - change syllabic & prosodic patterns, according to the state of mind [Ward2004]

- Audience's interest level

- Interesting spot ("hot spot") in the session

Prosodic Features

- For each reactive token
  - Duration
  - F0 (maximum, range)
  - power (maximum)
- Normalized for each person
  - For each feature, compute the mean
  - The mean is subtracted from feature values

Variation (SD) of Prosodic Features

- Tokens used for assessment have a large variation

			Duration SD (sec.)	F0 max SD (Hz)	F0 range SD (Hz)	Power SD (db)
Non-lexical & used for assessment	ふーん (hu:N)	114	**0.44**	22	**38**	4.3
	へー (he:)	78	**0.54**	**34**	**41**	5.4
	あー (a:)	59	0.37	**35**	**39**	**6.4**
	はあ (ha:)	55	0.24	**35**	36	**6.3**
	ああ (aa)	23	0.17	30	**38**	**6.3**
	はー (ha:)	21	**0.65**	32	30	4.8
Lexical & used for Ack.	うーん (u:N)	544	0.27	27	35	4.6
	うん (uN)	356	0.15	25	30	4.9
	はい (hai)	188	0.19	28	24	5.8
	ふん (huN)	166	0.31	25	21	4.1
	ええ (ee)	38	0.1	31	37	5.5

Relationship with Interest Level (Subjective Evaluation)

- For each token (syllable pattern) and for each prosodic feature,
  - Pick up top-10 & bottom-10 samples (largest & smallest values of the feature)
- Audio file is segmented to cover the reactive token and its preceding clause
- Five subjects listen and evaluate the audience's state of the mind
  - 12 items to be evaluated in 4 scales
  - two for interest: 興味, 関心
  - two for surprise: 驚き, 意外

Relationship with Interest Level (Subjective Evaluation)

- There are particular combinations of syllabic & prosodic patterns which express interest & surprise

Reactive token	prosody	interest	surprise
へー *he:*	duration	○	○
	F0max	○	○
	F0range	○	○
	Power	○	○
あー *a:*	duration		
	F0max	○	
	F0range		
	Power	○	
ふーん *fu:N*	**duration**	○	○
	F0max		
	F0range		
	powe		

(p<0.05)

Podspotter: Conversation Browser based on Audience's Reaction

- "Funny Spot" ← laughter
- "Interesting Spot" ← reactive token

Demo

Third-party Evaluation of Hot Spots

- Four subjects, who had not attended presentation nor listened to the content
- Listen to a sequence of utterances (max. 20sec.) which induced the laughter and/or reactive tokens
- Evaluate the spots
  - Is “Funny Spot” really funny?
  - Is “Interesting Spot” really interesting?

Third-party Evaluation of Hot Spots

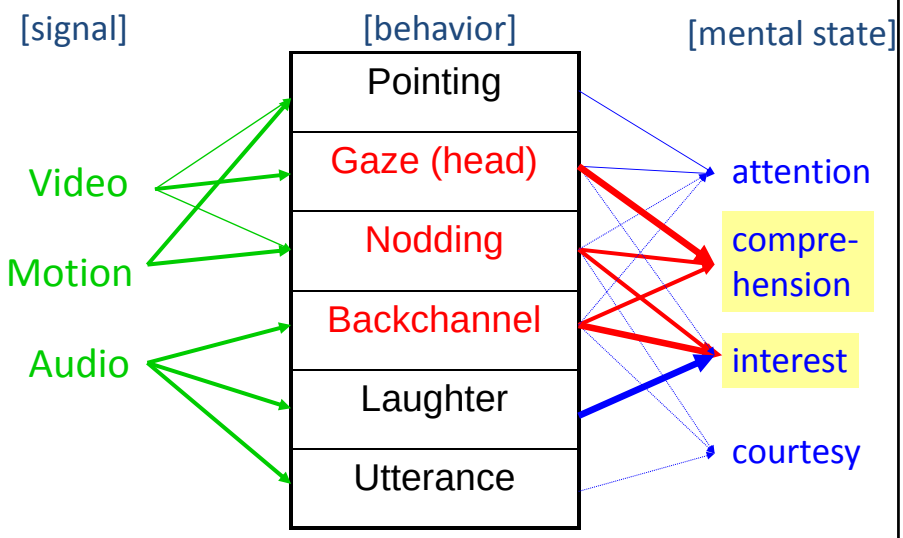
- “Funny Spot” ← laughter ?
  - Only a half are funny; 35% are NOT funny
  - Feeling funny largely depends on the person
  - Laughter was often made to relax the audience
- “Interesting Spot” ← reactive token ?
  - Over 90% are interesting and useful for the subjects

Conclusions

- Non-lexical reactive tokens with prominent prosody indicates interest level.
- The spots detected based on reactive tokens are interesting for third-party viewers.
- Laughter does not necessarily mean “funny”.

Prediction of Turn-taking with Eye-gaze and Backchannel

Multi-modal Sensing & Analysis

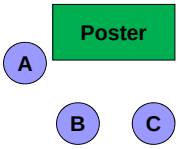


Prediction of Turn-taking by Audience

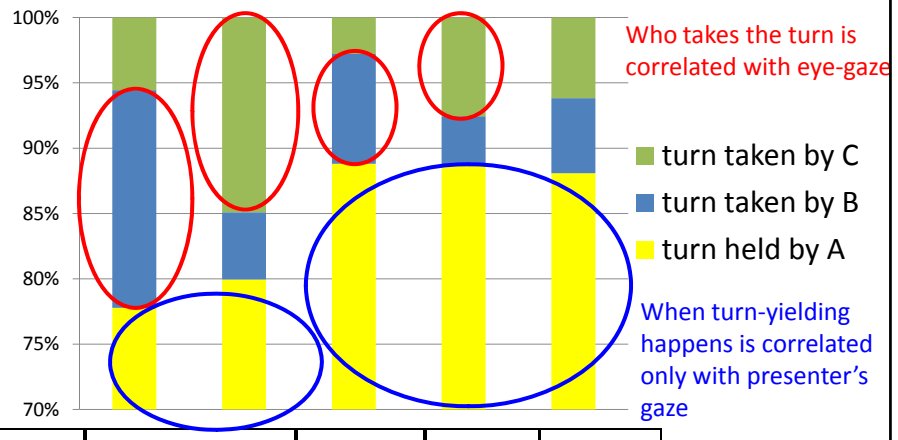
- **Questions & comments → Interest level**
  - Audience asks more & better questions when attracted more
- Automated control to beamform microphones or cameras
  - before someone in the audience actually speaks
- Intelligent conversational agent handling multiple partners
  - wait for someone to speak OR continue to speak

Prediction of Turn-taking by Audience

- When the turn is taken by (someone in) the audience
  - Detection problem (→ recall & precision)
    - ←Infrequent (~10%) compared with turn-holding by presenter
    - ←More important and informative than presenter's utterances
  - Prosody of presenter's utterance
  - Backchannel including nodding of audience
  - Eye-gaze information
- Who (in the audience) takes the turn
  - Classification problem (→ accuracy)
  - Using eye-gaze & backchannel information



Relationship between Turn-taking and Eye-gaze



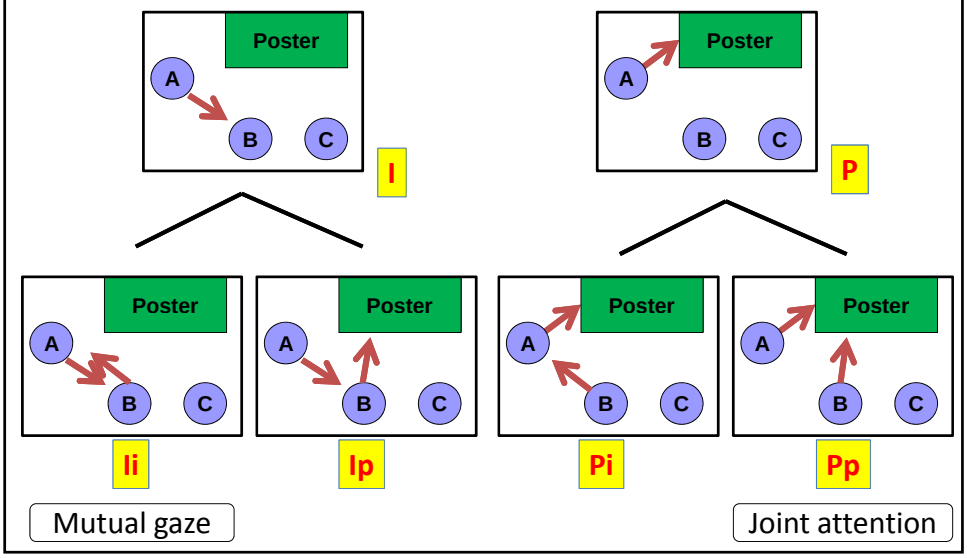
Who gazes	Presenter A		B	C	Overall average
at Who	B	C	A	A	

Relationship between Turn-taking and Eye-gaze Duration (sec.)

	turn held by presenter	turn taken by audience	
		B	C
A gazed at B	0.220	0.589	0.299
A gazed at C	0.387	0.391	0.791
B gazed at A	0.161	0.205	0.078
C gazed at A	0.308	0.215	0.355

- Presenter gazed at the person (significantly longer) before yielding the turn to him/her
- No significant difference in eye-gaze by audience

Joint Eye-gaze Event

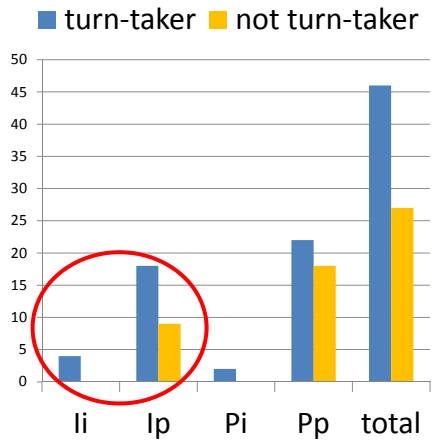


Relationship between Turn-taking and Joint Eye-gaze Events

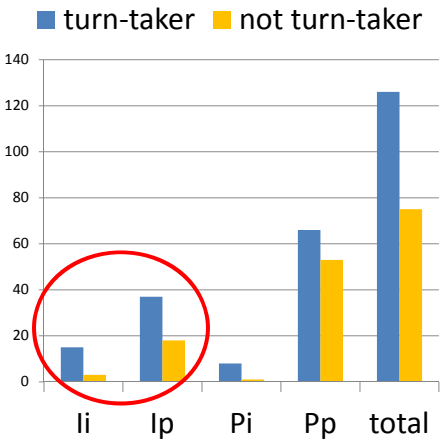
	turn held by presenter	turn taken by audience	
		Self	Other
Ii	125	17	3
Ip	320	71	26
Pi	190	11	9
Pp	2974	147	145

- Ii (mutual gaze) & Pi are not so frequent
- Ip: presenter gazes at the person before giving the turn

Relationship between Turn-taking and Backchannel (+ Eye-gaze)



Verbal backchannel



Non-verbal nodding

Features for Prediction of Turn-taking

- Prosodic features of presenter's utterance
  - F0 (mean, max, min), power (mean, max)
  - Normalized for each speaker
- Backchannel features
  - Verbal, non-verbal nodding
- Eye-gaze features
  - Object: poster (P,p) or person (I,i)
  - Joint eye-gaze event: Ii, Ip, Pi, Pp
  - Duration of above

who

when

Prediction of Speaker Change (when the turn is taken)

Feature	Recall	Precision	F-measure
Prosody	0.667	0.178	0.280
Backchannel (BC)	0.459	0.113	0.179
Eye-gaze (gaze)	0.461	0.216	0.290
Prosody + BC	0.668	0.165	0.263
Prosody + gaze	0.706	0.209	0.319
Prosody + BC + gaze	0.678	0.189	0.294

- Prosody of presenter and eye-gaze are useful, while backchannel of audience is not.

Prediction of Next Speaker (who takes the turn)

Feature	Accuracy
backchannel	52.6%
eye-gaze object/event	55.8%
eye-gaze object/event + duration	66.4%
Combination of above all	69.7%

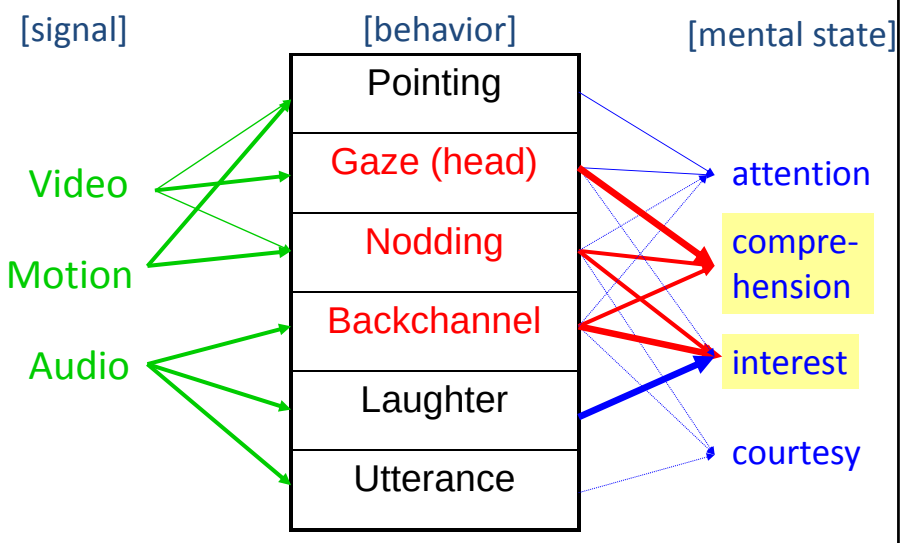
- eye-gaze and backchannel are useful, and eye-gaze duration is most effective

Conclusions

- Eye-gaze events and backchannels suggest who will make questions/comments.
  - Interest-level of the audience (?)
- Actual turn-taking by the audience happens when the presenter gazed at the person.
  - Presenter controls the turn-taking
  - Eye-gaze and backchannels may trigger this by attracting the presenter's attention (?)

Relationship between Audience's Feedback Behaviors and Question Type

Multi-modal Sensing & Analysis



Prediction of Kind of Questions asked by Audience

Questions → Comprehension & Interest level

- Confirming Questions
  - Make sure of understanding of explanation
  - Can be answered simply by "YES/NO"
- Substantive Questions
  - Asking about what was not explained
  - Can NOT be answered by "YES/NO" only; extra explanation needed

Relationship between Question Type and Backchannel

Verbal Backchannels

	Confirming	Substantive
Turn-taker	0.034	0.063
Non-turn-taker	0.041	0.038

Non-verbal Noddings

	Confirming	Substantive
Turn-taker	0.111	0.127
Non-turn-taker	0.109	0.132

Frequency (per sec.) in preceding explanation segment

Relationship between Question Type and Joint Eye-gaze Event

	Confirming	Substantive
Ii	0.053	0.015
Ip	0.116	0.081
Pi	0.060	0.035
Pp	0.657	0.818

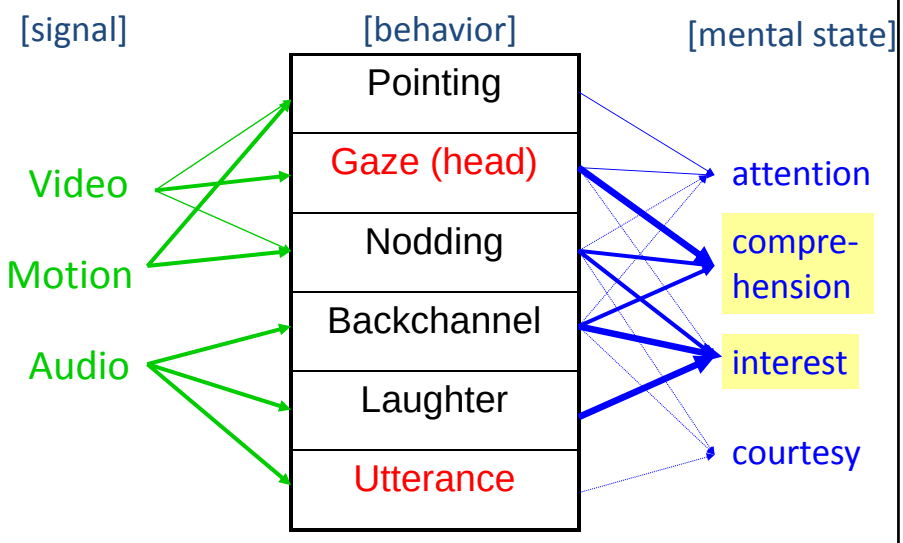
Frequency (ratio) in preceding explanation segment

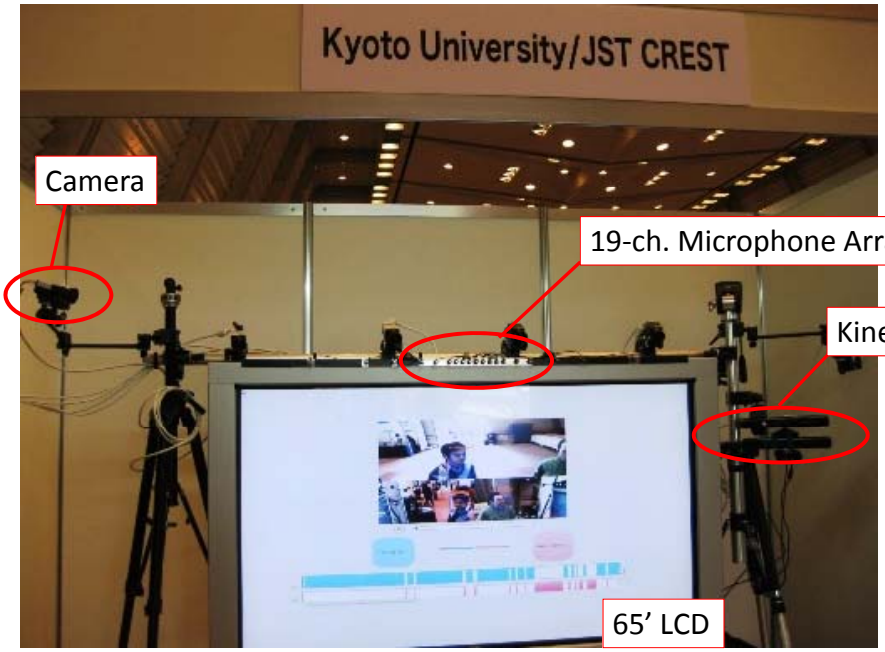
Conclusions

- Audience makes more verbal backchannels before making substantive questions while focusing on the poster.
  - Confident in understanding & shows interest (?)
- Majority of turn-taking signaled by the presenter's gazing is attributed to confirmations.
  - Grounding of understanding (?)

Smart Posterboard System

Multi-modal Sensing & Analysis



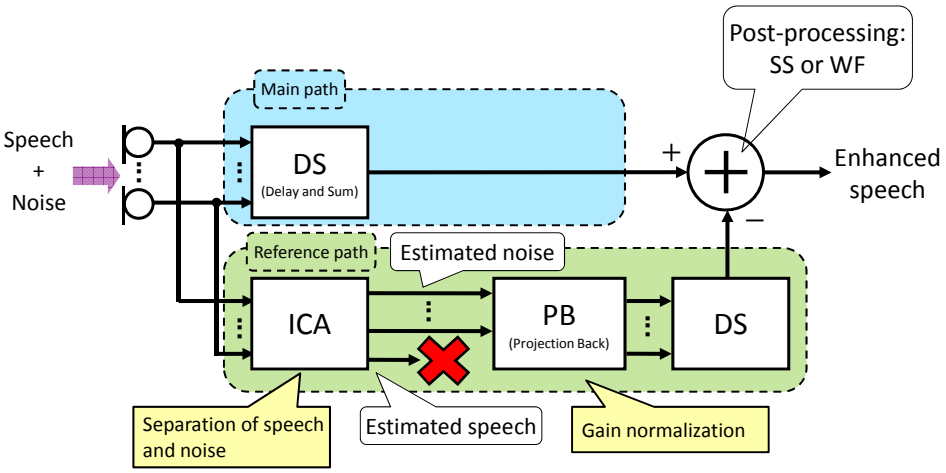


Smart Posterboard Demonstration Overview

- **Offline Diarization & Browser with 19-channel Microphone Array & 6 Cameras**
  - Speech separation & enhancement
    BSSA (Blind Spatial Subtraction Array)
  - Voice activity detection
  - Speaker localization (← video)

  } Speaker diarization
  - Gaze (head direction) detection (← video)
- **Online tracking using Kinect**
  - Speaker localization & gaze (head direction) detection
  - Speech enhancement

Speech Separation & Enhancement: Blind Spatial Subtraction Array (BSSA)



Application Scenario

- Poster session archiving + browser **Demo**
  - Interaction analysis
  - Visualization and mining
    - Review Q-A afterwards
    - Extract segments people find interesting or difficult to understand
- Automated presentation system
  - Switch slides according to interest and knowledge level of the audience
  - Answer questions

Staffs contributed to this Demo.

- Kyoto University:
  - Tony Tung, Hiromasa Yoshimoto, Randy Gomez, Soichiro Hayashi, Yuya Akita, **Tatsuya Kawahara**
- Nara Institute of Science & Technology
  - Kodai Okamoto, Yuji Onuma, Noriyoshi Kamado, Ryoichi Miyazaki, **Hiroshi Saruwatari**