Multimodal Sensing and Recognition for Smart Posterboard

Tatsuya Kawahara
(Kyoto University)

Project Overview

Project Core Members

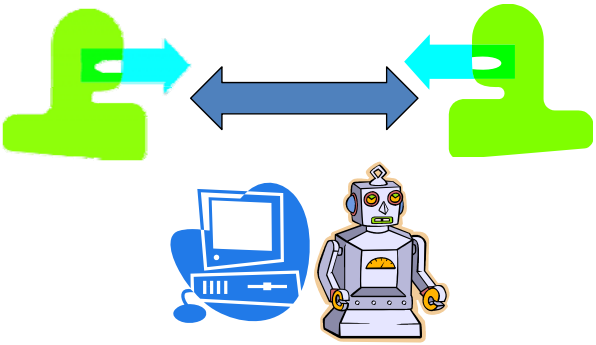
- PI : **Prof. Tatsuya Kawahara (Kyoto University)**

- Kyoto University
  - Prof. Yuichi Nakamura (Video Processing)
  - Prof. Takashi Matsuyama (Computer Vision)
  - Prof. Sadao Kurohashi (Natural Language Processing)

- Nara Institute of Science & Technology
  - Prof. Kiyohiro Shikano (Speech Processing)
  - Assoc. Prof. Hiroshi Saruwatari (Acoustic Processing)

Problems

Speech Communication



Meetings & Conversations

- Speech-to-text
  - Speech recognition
  - Captioning

- Sensing of comprehension & interest level
  - Assist comprehension
  - Presentation upon interest
  - Annotations

Goal of the Project

- Mining human interaction patterns
  (this talk)

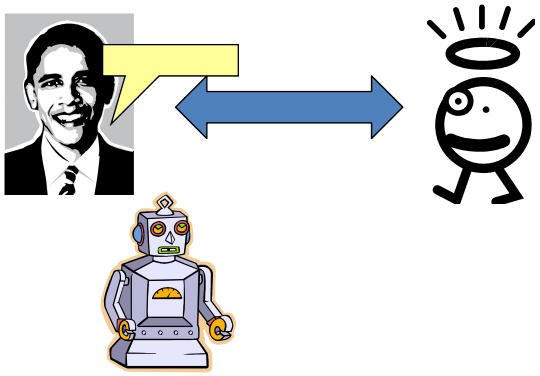
- A new indexing scheme of speech archives
  (current focus)
- A model of intelligent conversational agents
  (future topic)

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

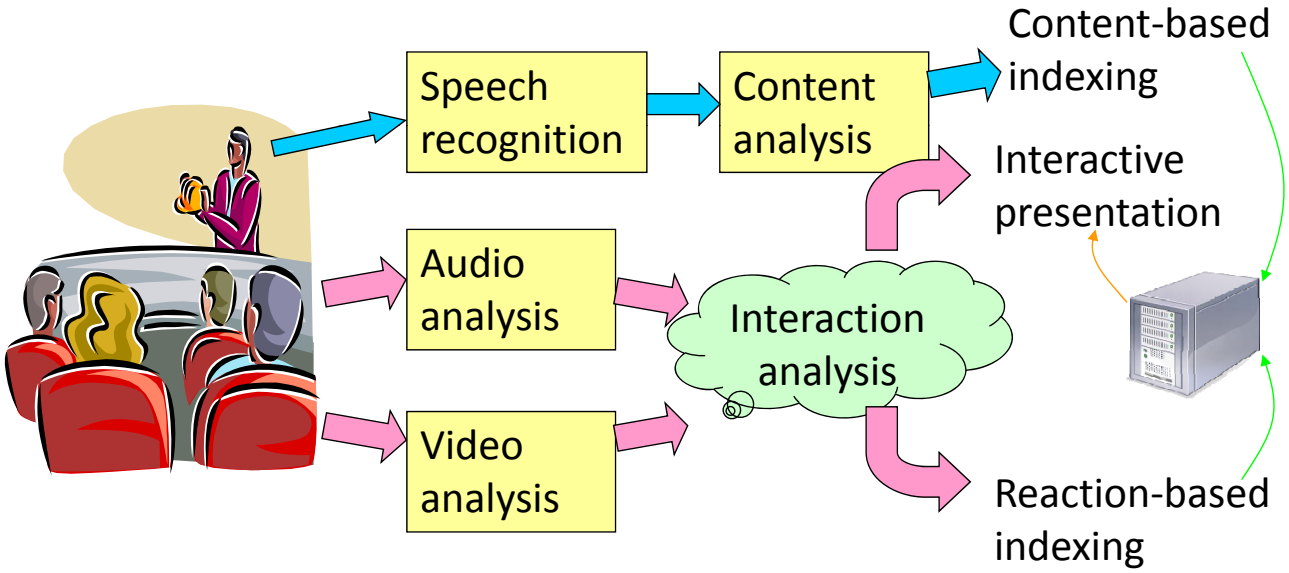


- Even if we do not understand the talk, we can see funny/important parts by observing audience's laughing/nodding
- Page rank is determined by the number of links rather than by the content

From Content-based Approach to Interaction-based Approach

	Focus	Features	Annotation
Content-based	Main speaker's utterances	lexical, prosodic …	"important"
Interaction-based	Listener's reaction	non-verbal, multi-modal	"interested"

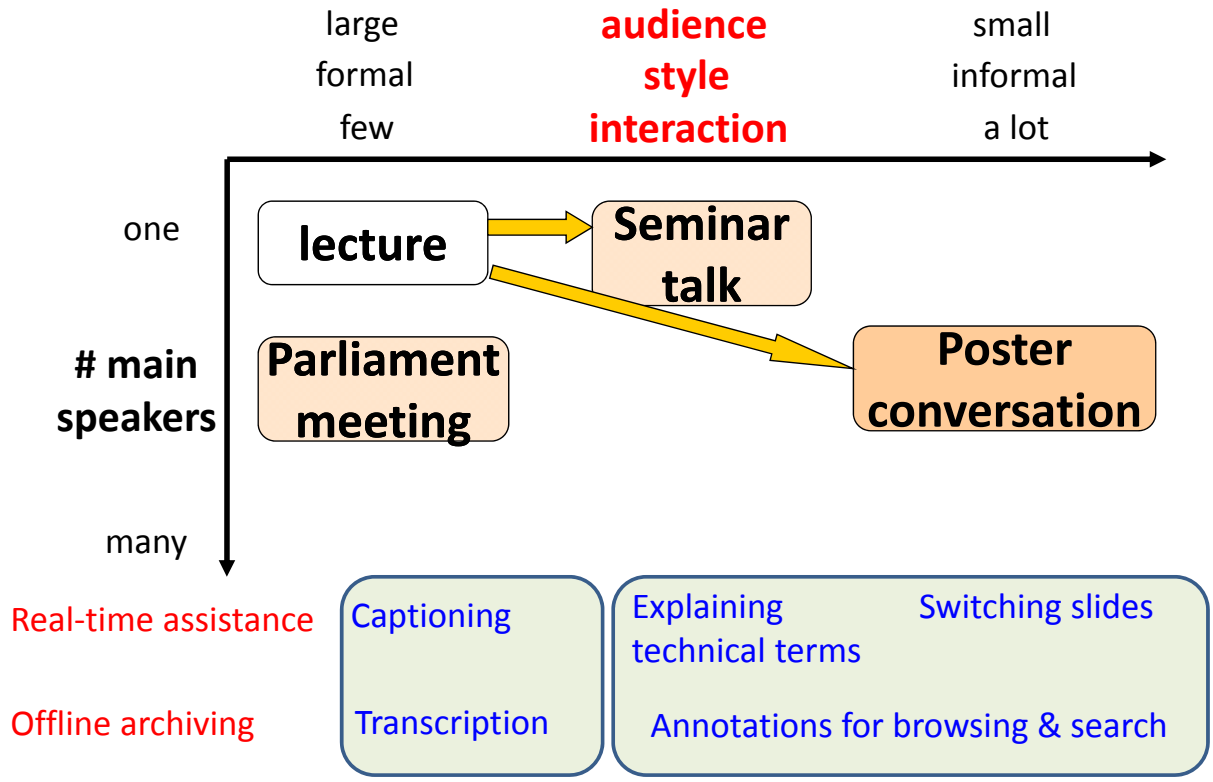
Process Overview



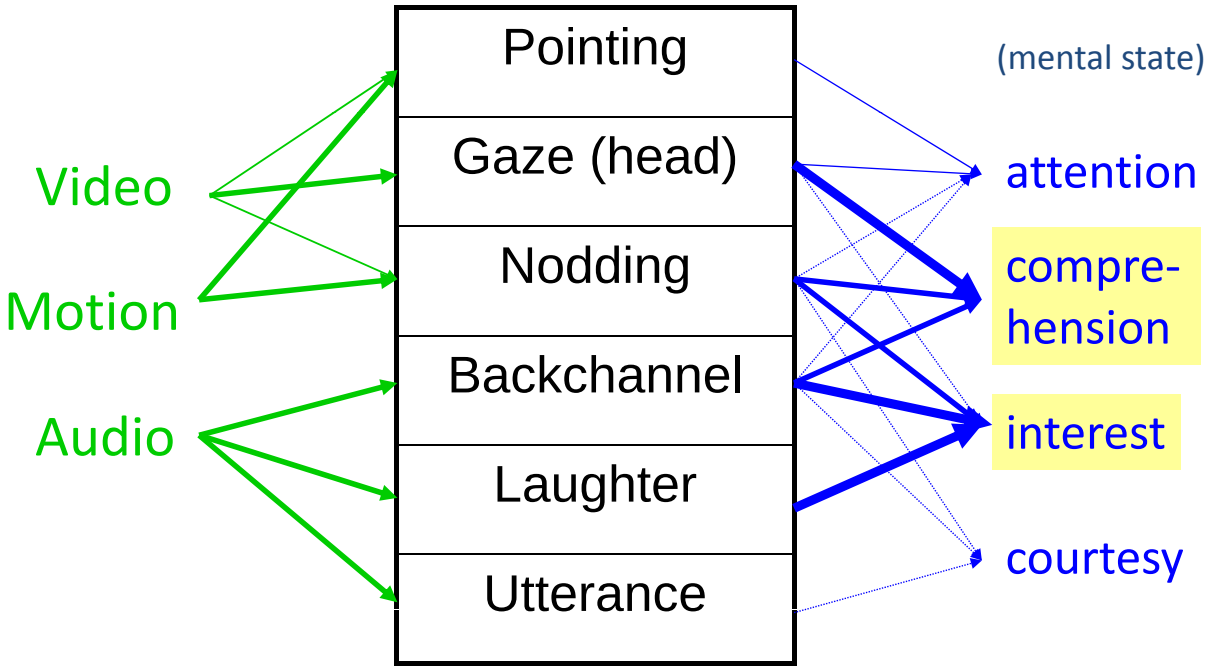
[multi-modal sensing]

[archiving & presentation support]

Targets



Multi-modal Sensing & Analysis



Multi-modal Recording of Poster Sessions

Why Poster Sessions?

- Norm in conferences & open-houses
- Mixture characteristics of lectures and meetings
  - One main speaker, with a small audience
  - Anyone of the audience can take an initiative

- Interactive
  - Real-time feedback by audience
  - including back-channels & nodding
- Multi-modal (truly)
  - Standing & moving
- Real, but controlled (knowledge/familiarity)

Multi-modal Sensing Environment: IMADE room

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

Audio

- 8 cameras installed in the room

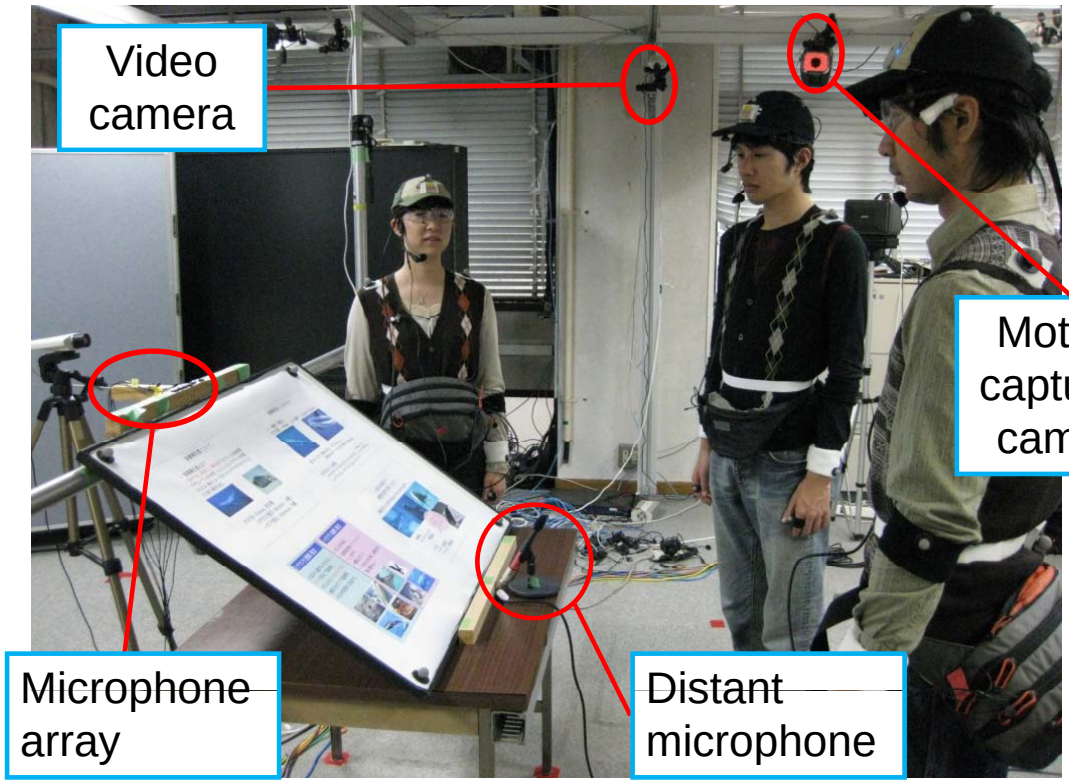
Video

- Motion-capturing system
- Accelerometer

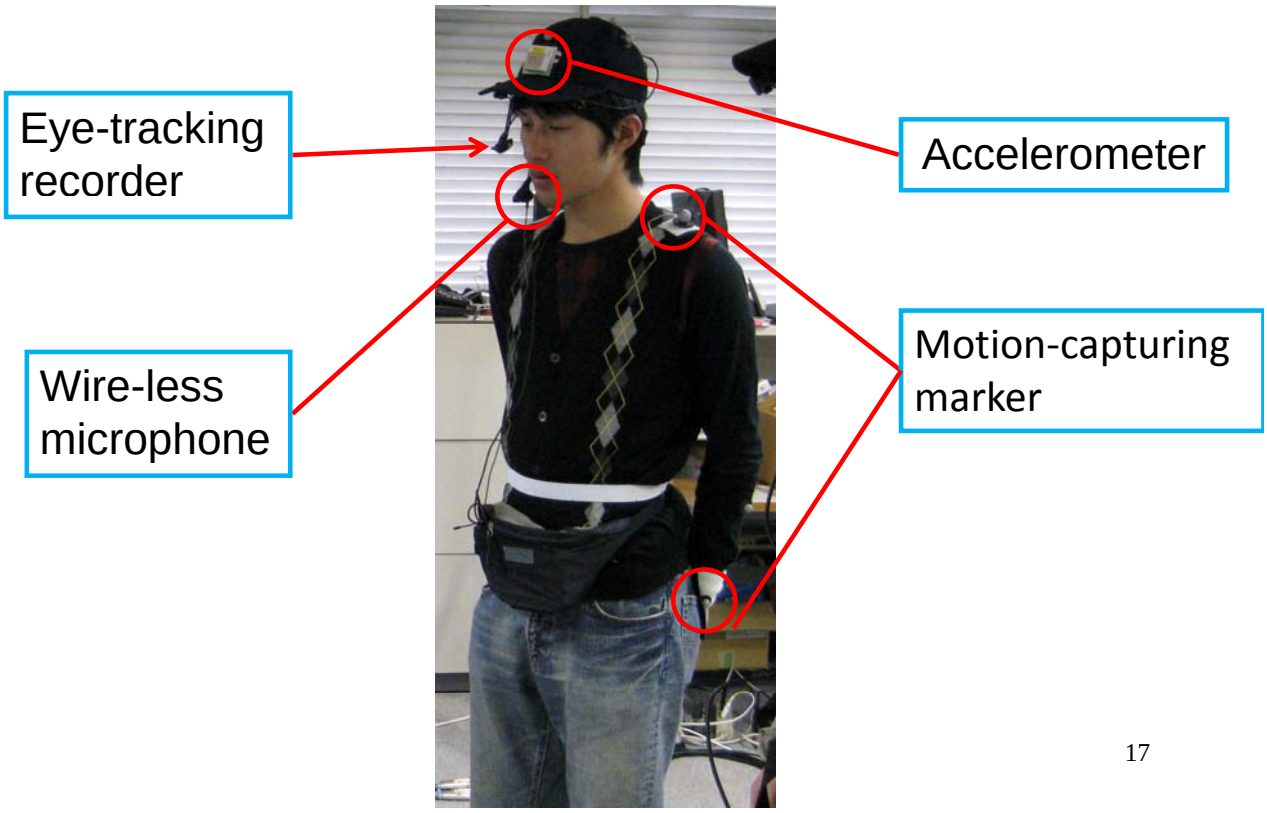
Motion

- Eye-tracking recorders → Gazing

Multi-modal Recording Setting

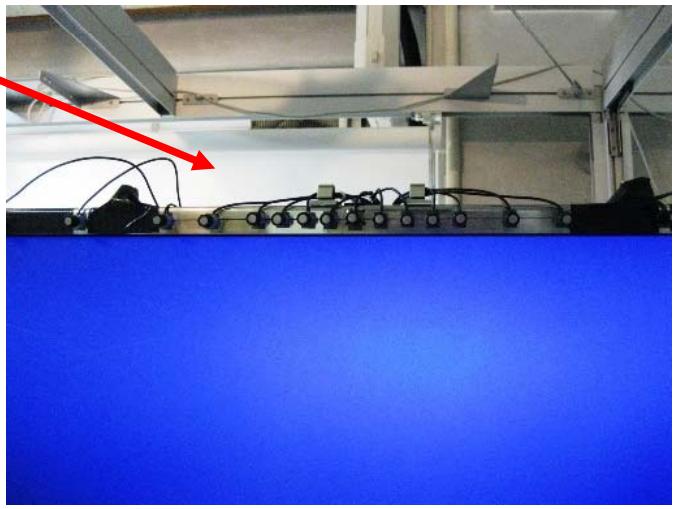


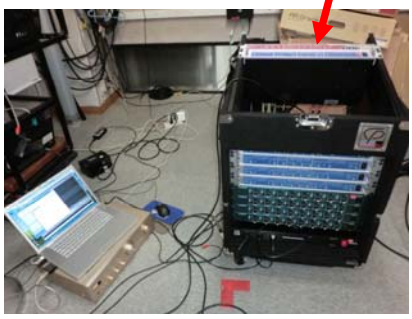
Multi-modal Recording Setting



Microphone Array settled on Posterboard

19-channel microphone array





Pre-amplifier
AD converter

Smart Posterboard

65’ LCD display + Microphone Array + Cameras

→ **Video**

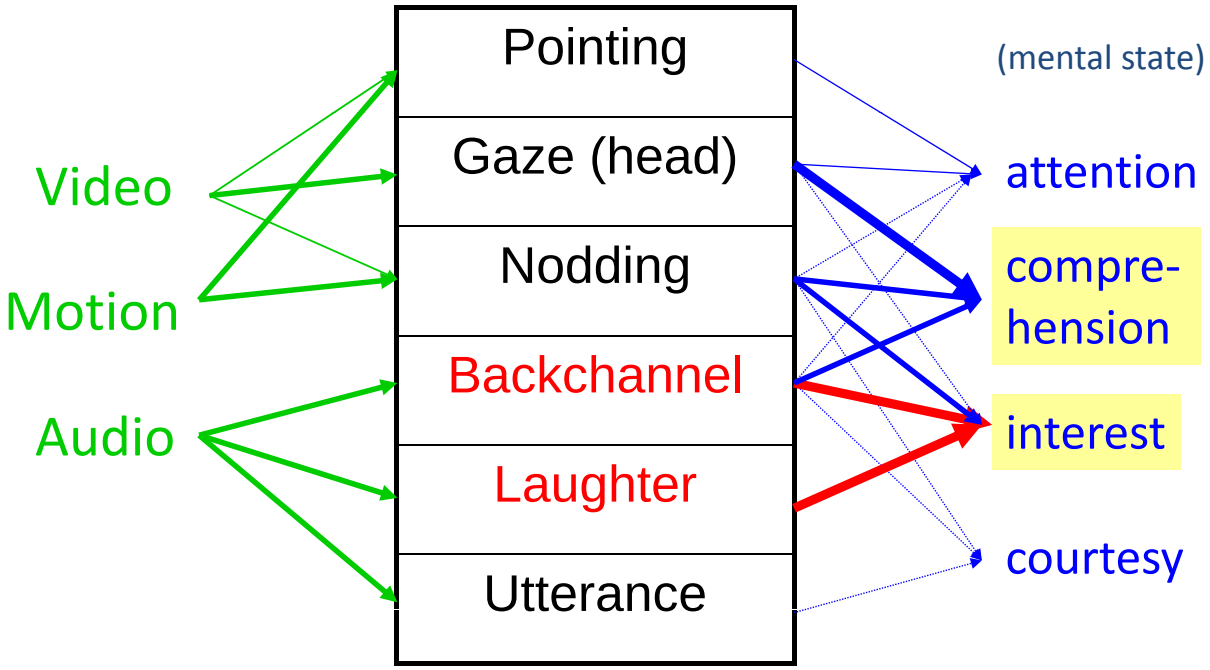
- @IMADEルーム
- 12 sessions
- Japanese, English
- annotation
  - speech
  - backchannel
  - gaze
  - nodding
  - pointing

Corpus of Poster Sessions

- 31 sessions recorded → 4 used in this work
  - One presenter + audience of two persons
  - Research presentation
  - Each 20 min.
- Manual transcription
  - IPU, clause unit
  - Reactive tokens & fillers
- Non-verbal labels (almost automated!!)
  - Nodding…non-verbal back-channel ← accelerometer
  - Gazing (to other persons & poster) ← eye-track rec.
  - Pointing (to poster) ← motion cap.

Hot Spot Detection based on Audience's Reactive Tokens

Multi-modal Sensing & Analysis



Hot Spot Detection
based on Audience's Reactive Tokens

- Hot Spots: where audience was impressed
- Reactive Tokens *(aizuchi)*
  - short verbal responses made in real-time & back-channel
  - often non-lexical (ex.) "uh-huh", "wow"
  - change syllabic & prosodic patterns, according to the state of mind (interest-level)

- Detection of audience's interest level

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

Correlation 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

Correlation 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 Spots" ← laughter
- "Interesting Spots" ← reactive tokens

Demo

Subjective Evaluation of Detected 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?

Subjective Evaluations of Detected Hot Spots

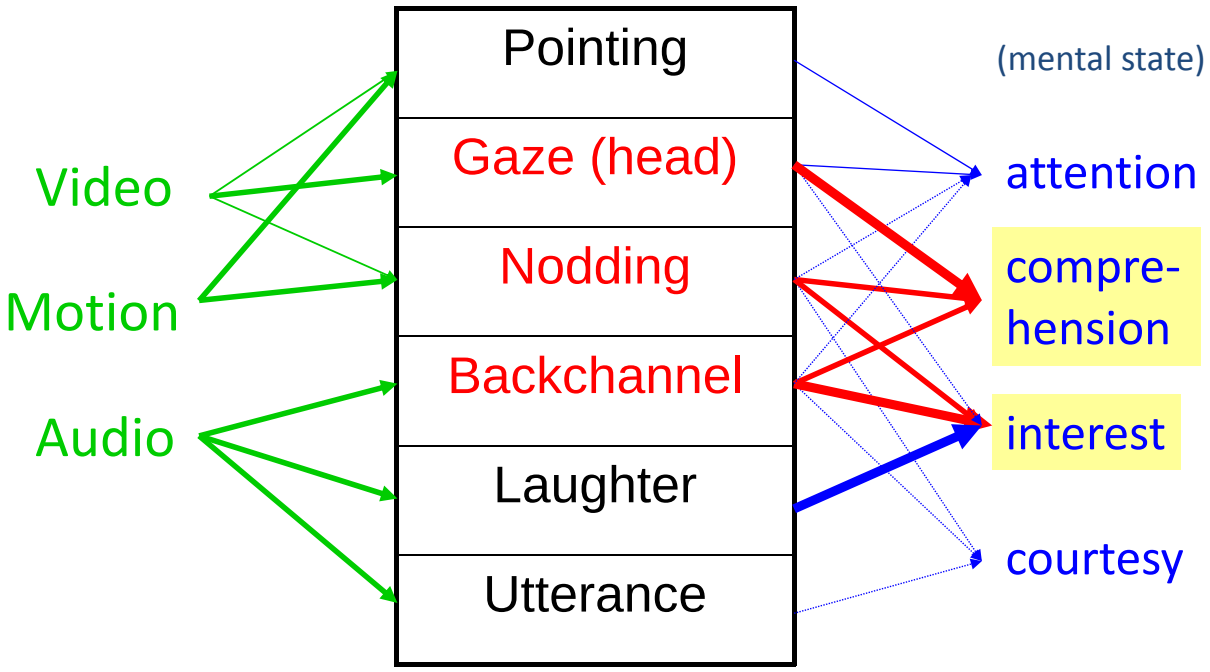
- "Funny Spots" ← laughter
  - Only a half are funny; 35% are not funny
  - Feeling funny largely depends on the person
  - Laughter was often made to relax the audience
- "Interesting Spots" ← reactive tokens
  - Over 90% are interesting and useful for the subjects

Conclusions

- Non-lexical reactive tokens with prominent prosody indicates interest level.
- Laughter does not necessarily mean "funny".

Prediction of Turn-Taking by using Eye-Gaze and Backchannel

Multi-modal Sensing & Analysis



Prediction of Turn-taking by Audience

- Questions & comments suggest comprehension & interest-level of audience

- 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

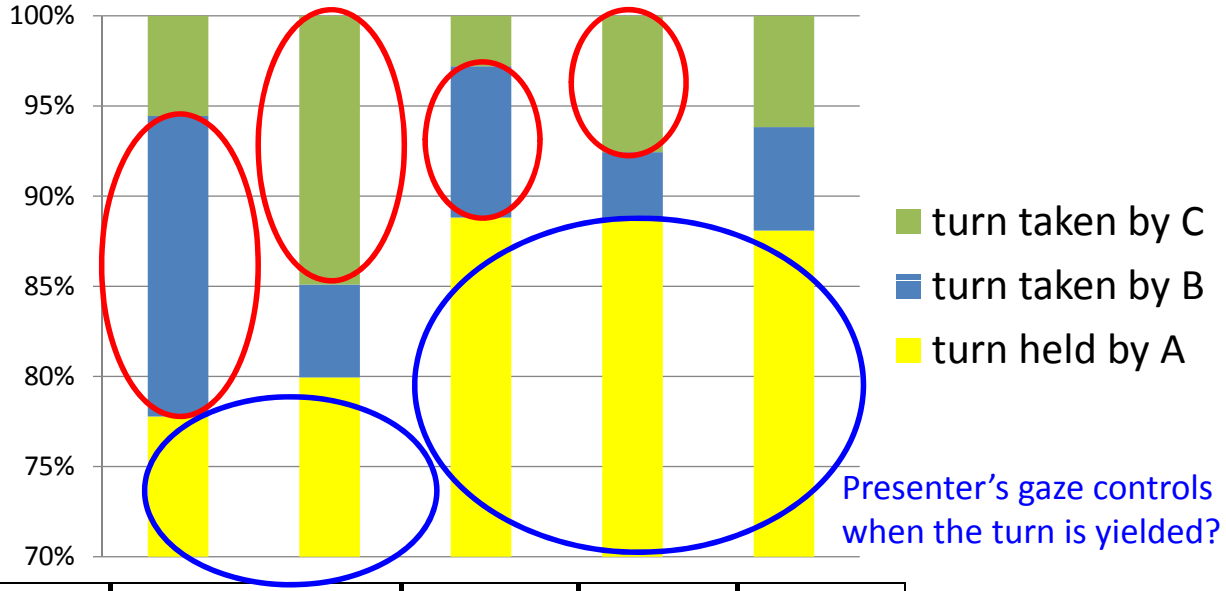
- When the turn is taken by (someone in) the audience
  - Detection problem (→ recall & precision)
  - Prosody of presenter's utterance
  - Audience's backchannel
  - Eye-gaze information
- Who (in the audience) takes the turn
  - Classification problem (→ accuracy)
  - Using gaze & backchannel information

Statistics of Turn-taking by Audience

	turn held by presenter	turn taken by audience		
		B	C	total
Session 2	845	44	50	94
Session 4	419	37	12	49
Session 5	356	17	39	56
Session 8	422	35	42	77
total	2042	133	143	276

- In majority of presenter's utterances (IPUs), turn is held
- ration of turn-taking by audience is 11.9%

Relationship between Turn-taking and Eye-gaze



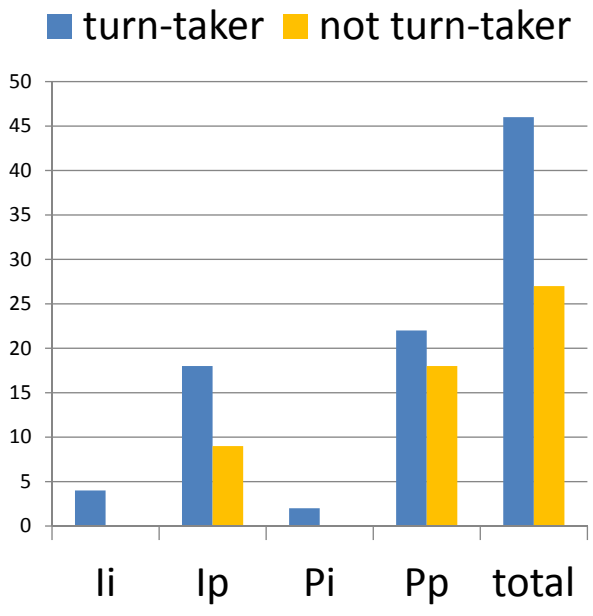
Who gazes	Presenter A		B	C	Overall average
at Who	B	C	A	A	

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

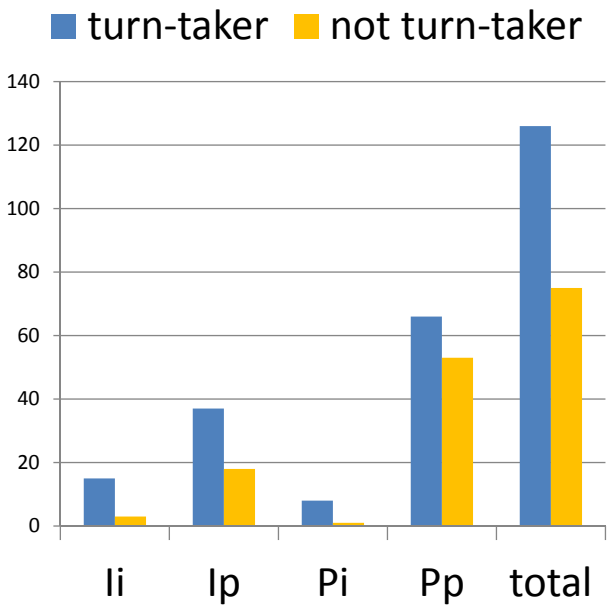
	turn held by presenter	turn taken by audience	
		B	C
A gazed at B	0.220	0.589	0.299
A gazed at C	0.387	0.391	0.791
B gazed at A	0.161	0.205	0.078
C gazed at A	0.308	0.215	0.355

- Presenter gazed at the person before yielding the turn to him/her
- Not significant difference in eye-gaze by audience

Relationship between Turn-taking and Backchannel + Eye-gaze



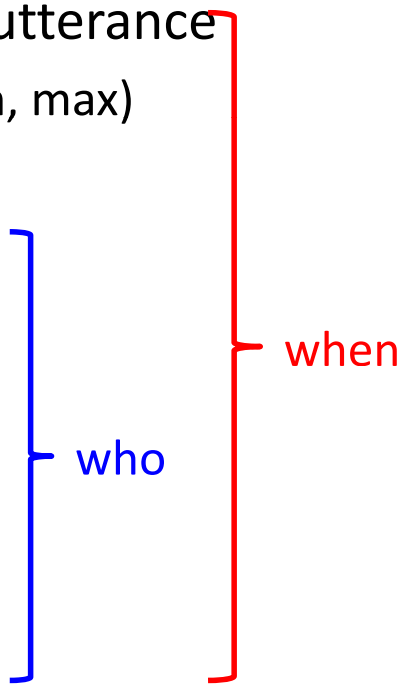
Verbal backchannel



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



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 by the 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 still controls the turn-taking (?)

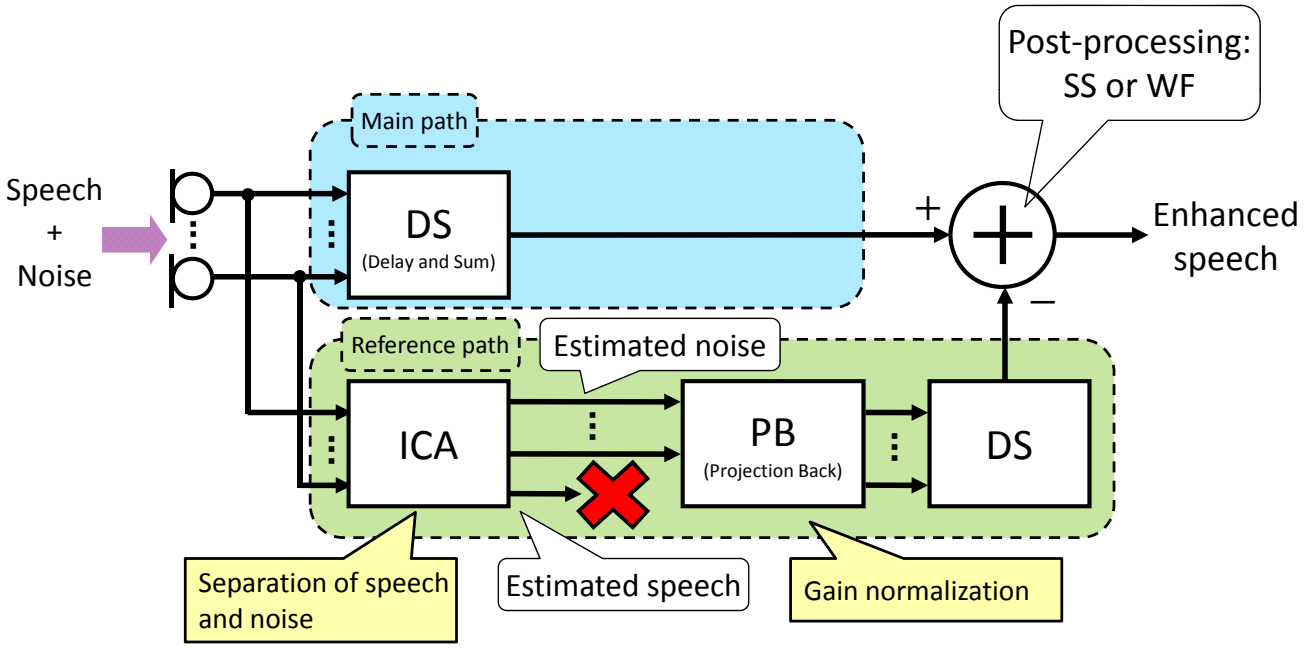
Smart Posterboard System

Smart Posterboard Demonstration Overview

- **Offline Diarization & Browser with 19-channel Microphone Array & 6 Cameras** **Demo**
  - Speech enhancement with BSSA (Blind Spatial Subtraction Array)
  - Speaker diarization based on adapted GMM
  - Speaker localization & Gaze (head direction) detection
- **Online tracking using Kinect**
  - Speaker localization & gaze (head direction) detection
  - Speech enhancement

Live Demo

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



Application Scenario

- Poster session archiving + browser
  - 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
  - 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**