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

Tatsuya Kawahara

(Kyoto University, Japan)

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

JST CREST Project (2009-2014)

- PI: **Prof. Tatsuya Kawahara (Kyoto University)**
- Kyoto University
  - Prof. Yuichi Nakamura (Video Processing)
    - Mr. Hiromasa Yoshimoto
  - Prof. Takashi Matsuyama (Computer Vision)
    - Dr. Tony Tung
  - Prof. Sadao Kurohashi (Natural Language Processing)
    - Dr. Yugo Murawaki
- Nara Institute of Science & Technology
  - Assoc. Prof. Hiroshi Saruwatari (Acoustic Processing)

Why Poster Sessions?

- Norm in conferences & open-houses
  - – But not recorded at all,
    while many lectures are now being recorded
- Interactive & multi-modal
  - – A small audience can make questions at any time
  - – Gaze and backchannels play an important role
- Long and redundant ← repeated presentations
  - → need for efficient browsing of the recordings

Smart Posterboard

[Demo@ICASSP2012]

All sensors are attached to large (65'') LCD

Goal (Application Scenario)

Modeling human interaction behaviors

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

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

Problems & Tasks

- Multi-modal signal-level sensing
  - Face detection, eye-gaze detection

    → who came to the poster
  - Speech separation, speaker diarization

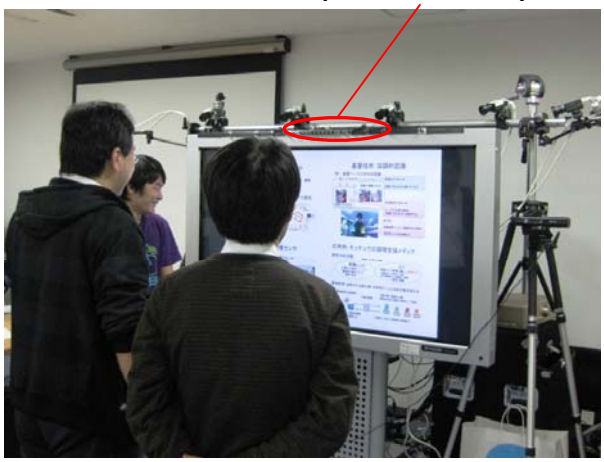
    → what they said
- High-level indexing using multi-modal behaviors
  - Interest level estimation

    → which part they were attracted
  - Comprehension level estimation

    → which part was difficult to follow

Recording of Poster Conversations with Smart Posterboard

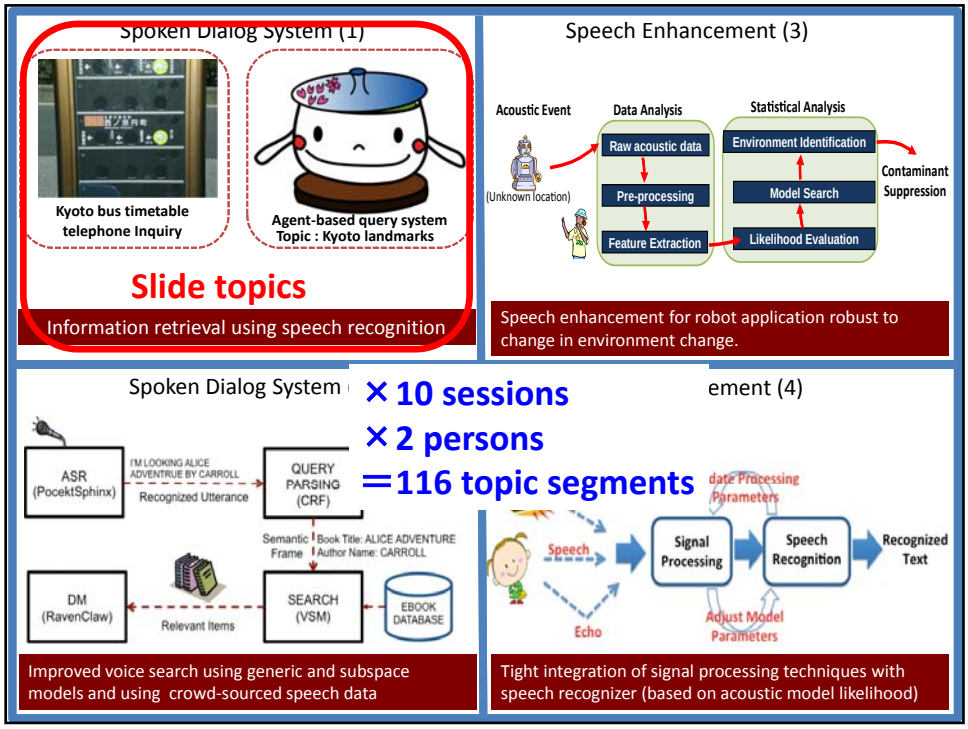
65’ LCD Screen + Microphone Array + Cameras



Setting of Poster Conversations

- Presentation of research overview
  - 4 or 8 slides of rather independent topics (=slide topics)

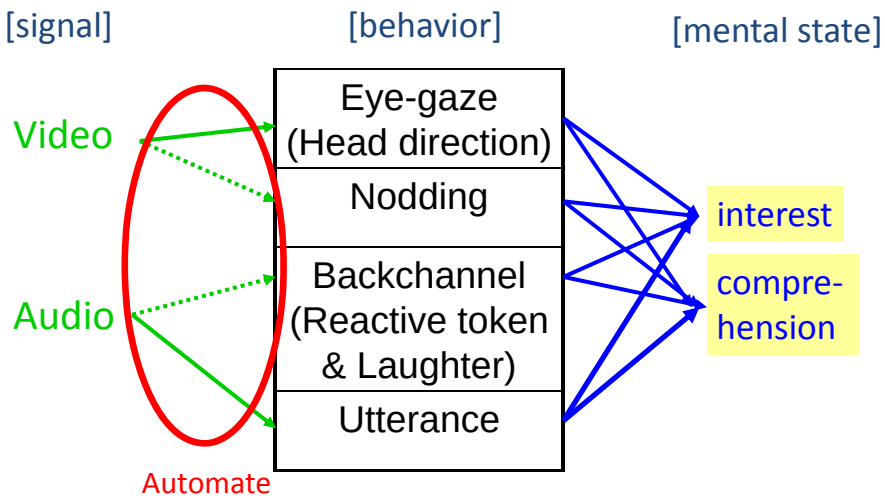
  → Easy to annotate interest & comprehension level
- Audience of two persons
  - Young researchers, who are not familiar with the presenter and the topics
- Duration: 20-30 minutes
- 10 sessions → 58 slide topics



Transcriptions & Annotations of Poster Conversations

- Manual transcription of speech
  - IPU, clause unit
  - Fillers, Backchannels (reactive tokens), Laughter
- Non-verbal behavior labels (almost automated)
  - Eye-gaze (to other person & poster)
    - ← eye-track recorder (initially for ground-truth)
    - ← Kinect sensor + head-orientation tracking
  - Nodding…non-verbal backchannel
    - ← accelerometer
    - ← Kinect sensor + head-orientation tracking

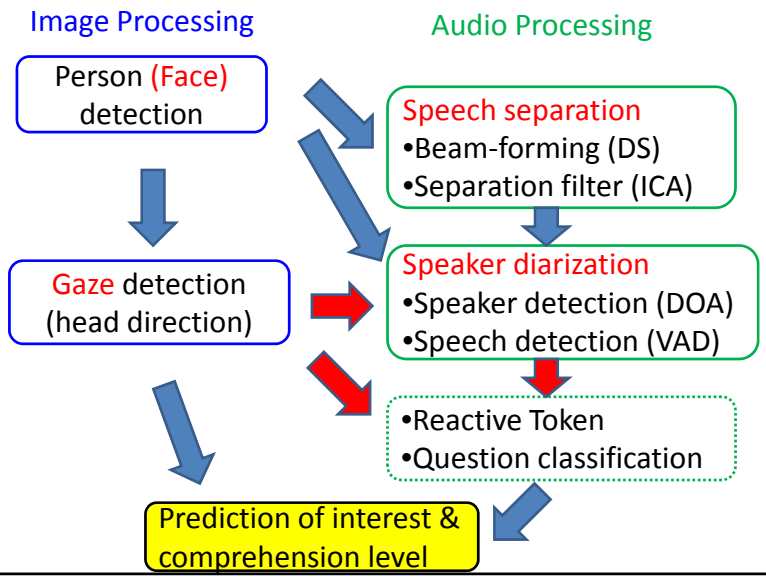
Multi-modal Sensing



Multi-modal Sensing

- Challenges in poster conversations
  - Multiple persons (+replacing)
  - Moving
  - Talking at distance (+background noise)
  - … No prior work in acoustic research!!
- All sensors are attached to posterboard
  - 19-channel microphone array, Kinect
  - [portable version] Kinect only (for 1 person)

Behavior Sensing by Multi-modal Processing



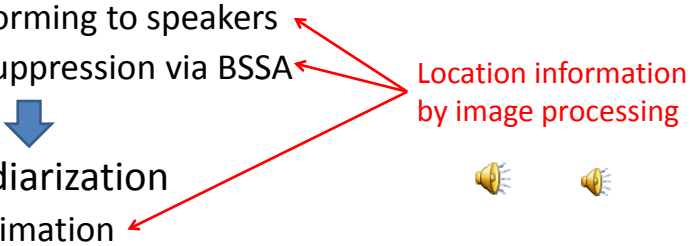
Gaze Detection

- Gaze ← Head direction tracking
  - Difference <10 degree, in poster conversations
- Procedure
  1. Face detection….color & TOF information
  2. Head model estimation…3D model
  3. Head tracking…particle filter
  4. Identification of gaze object: poster or participants
- Online & real-time processing with GPU
- Accuracy of 90%

(cf.) Nodding is also detected in this process

Speech Separation & Speaker Diarization

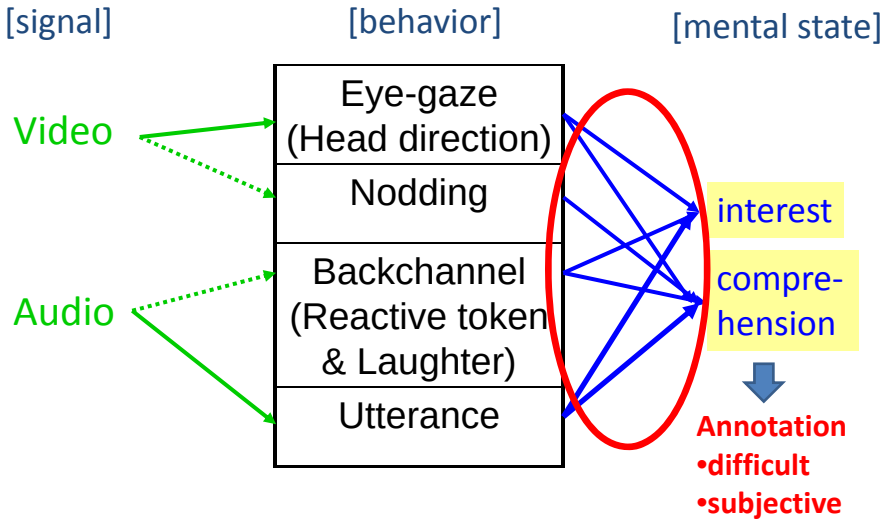
- Separation & enhancement of distant speech
  - Beam-forming to speakers
  - Noise suppression via BSSA
- Speaker diarization
  - DoA estimation
  - Voice Activity Detection on enhanced speech
  - Presenter's speech: recall & precision: 85%
  - Audience's speech: recall: 70%, precision: 85%



Detection of Reactive Tokens & Laughter

- GMM classification
- Non-lexical reactive tokens
  - 「へー」「あー」「ふーん」
  - Characteristic prosodic patterns
  - Recall: 30%, Precision 80%
    → apparent (=significant) tokens can be detected
- Laughter
  - Recall & Precision: 70%
  - Laugher is not frequent and often used for relaxing in poster conversations

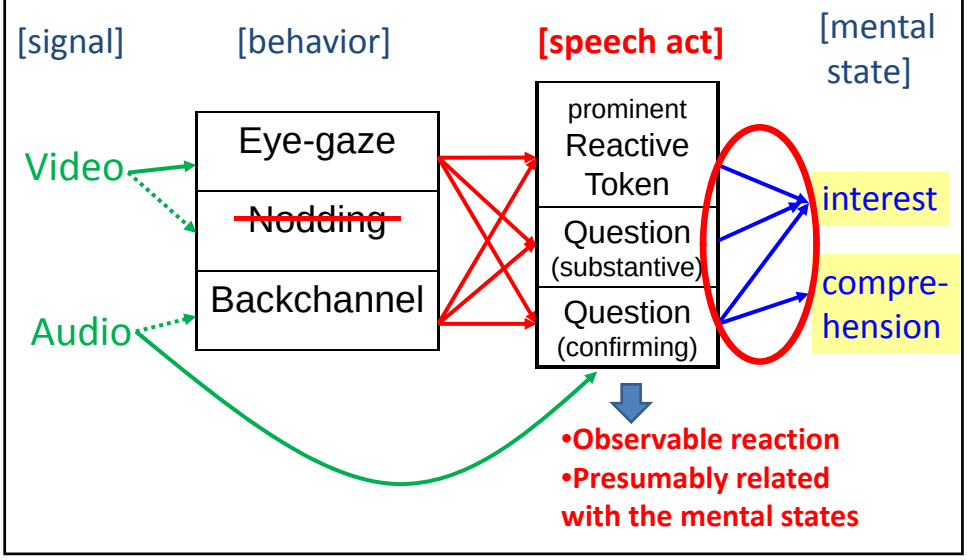
Scheme of Multi-modal Sensing & Prediction



Definition of Interest & Comprehension Levels

- "gold-standard" annotation: ask every participant to mark for each slide topic after the session
  - Not possible in a large scale
  - Subjective and may not be so reliable
- Focus on speech acts
  - Prominent reactive tokens [Kawahara IS2010&IPSJ11]
  - Questions raised by audience
    - "audience ask more questions when they are attracted."
    - Confirming questions: to make sure understanding
    - Substantive questions: asking on what was not explained

Proposed Scheme for Prediction of Mental States via Relevant Speech Acts



Relationship of Reactive Tokens and Interest Level [Kawahara IS2010]

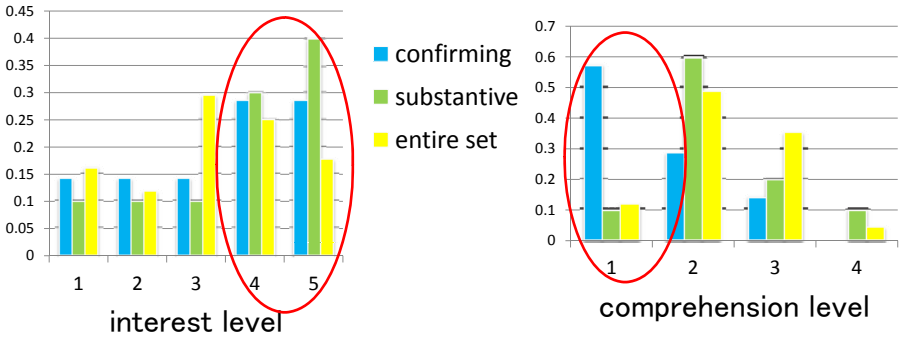
- Non-lexical
- Never used for acknowledgment ("wow")
- Prominent prosodic patterns

- Signal strong reaction

	prosody	interest
へー *he:*	duration	○
	F0max	○
	F0range	○
	Power	○
あー *a:*	duration	
	F0max	○
	F0range	
	Power	○
ふーん *fu:N*	**duration**	○
	F0max	
	F0range	
	power	

($p<0.05$)

Interest & Comprehension Level according to Question Type (4 sessions)



More questions
→ higher interest level

Confirming questions
→ low comprehension

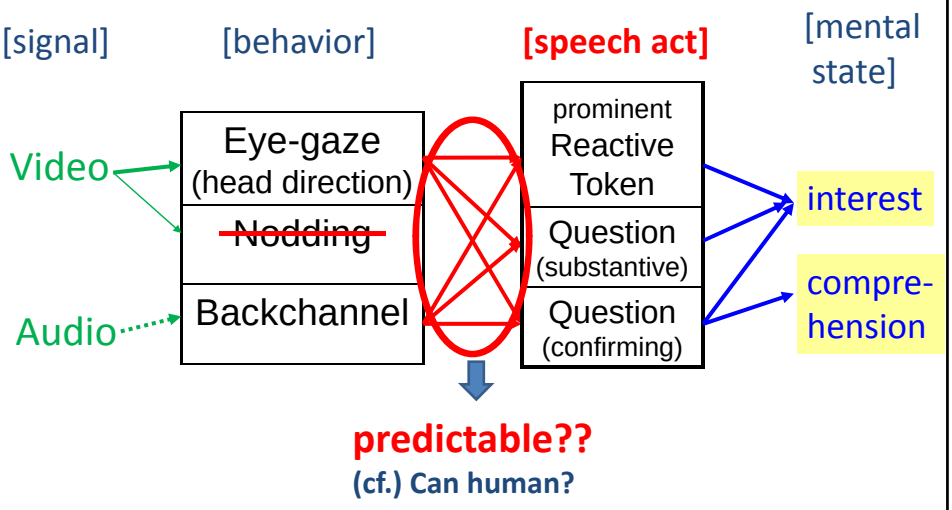
Definition of Interest & Comprehension Level

- High interest level
  - ← questions of any types
  - ← prominent reactive tokens
- Low comprehension level (in spite of interest)
  - ← confirming questions

Useful in reviewing the poster sessions

- Interesting spots (→ presenter & third-party viewers)
- Portion difficult for audience to follow (→ presenter)

Proposed Scheme for Prediction of Mental States via Relevant Speech Acts



Relationship between Backchannels and Questions

- Exclude prominent reactive tokens…less than 20%
- Majority are "*hai*" ("yeah", "okay")

- Frequency (count/utterance) in each topic segment

	Confirming	Substantive	Entire set
backchannel	0.53	0.59	0.42

- More backchannels
  → more questions, especially substantive questions

Relationship between Eye-gaze (at presenter) and Questions

- Frequency & duration of eye-gaze in each topic segment
  - In most of time, participants look at poster
  - Eye-gaze at presenter has a reason and effect

	Confirming	Substantive	Entire set
Gaze occurrence	0.38	1.02	0.64
Gaze duration	0.05	0.15	0.07

- Confirming questions ← increase in gaze at poster
  - more focused on poster, trying to understand
- Substantive questions ← increase in gaze at presenter
  - try to attract presenter's attention for taking a turn

Machine Learning for Prediction

- Features

$$F = \{f_1, f_2, f_3\} = \{\text{backchannel, gaze occurrence, gaze duration}\}$$

- Naïve Bayes classifier

$$p(c \mid F) = p(c) * \prod_i p(f_i \mid c)$$

- Estimation of $p(f_i|c)$
  - histogram quantization (3 or 4 bins)

- Circumvent estimation of model parameters
- Leave-one(session)-out cross validation using 10 sessions

Prediction of Topic Segments involving Questions and/or Reactive Tokens (=high interest)

	F-measure	accuracy
baseline (chance rate)	0.49	49.1%
(1) backchannel	0.59	55.2%
(2) gaze occurrence	0.63	61.2%
(3) gaze duration	0.65	57.8%
combination of (1)-(3)	**0.70**	**70.7%**

- Backchannel & gaze features lead to significant improvement
- Combination of both results in the best accuracy

Identification of Question Type of Confirming vs. Substantive (=comprehension level)

	accuracy
baseline (chance rate)	51.3%
(1) backchannel	56.8%
(2) gaze occurrence	75.7%
(3) gaze duration	67.6%
combination of (1)-(3)	75.7%

- All features lead to improvement
- Gaze occurrence alone achieves the best accuracy
- Need to parameterize backchannel patterns?

Summary

- Multi-modal signal-level sensing
  - "who came to the poster and what they said"
  - Combination of multi-modal information
- High-level indexing using multi-modal behaviors
  - Interest & comprehension level
  - using multi-modal features (backchannel & eye-gaze)
  - chance rate (50%) → over 70%
- Ongoing work
  - Tight integration of gaze and speech information
- Implemented on smart posterboard system
  → poster session browser