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

Tatsuya Kawahara
Academic Center for Computing and Media Studies, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

***Abstract*—Conversations in poster sessions in academic events, referred to as poster conversations, pose interesting and challenging topics on multi-modal signal and information processing. This paper gives an overview of our project on the smart posterboard for multi-modal conversation analysis. The smart posterboard has multiple sensing devices to record poster conversations, so we can review who came to the poster and what kind of questions or comments he/she made. The conversation analysis combines speech and image processing such as head tracking, speech enhancement and speaker diarization. Moreover, we are also investigating high-level indexing of interest and comprehension level of the audience, based on their multi-modal behaviors during the conversation.**

Index Terms: multi-modal signal processing, conversation analysis, behavioral analysis

I. Introduction

Multi-modal signal and information processing has been investigated primarily for intelligent human-machine interfaces, including smart phones, KIOSK terminals, and humanoid robots. Meanwhile, speech and image processing technologies have been improved so much that their target now includes natural human-human behaviors, which are made without being aware of interface devices. In this scenario, sensing devices are installed in an ambient manner. Examples of this kind of direction include meeting capturing [1] and conversation analysis [2].

We have been conducting a project which focuses on conversations in poster sessions, hereafter referred to as poster conversations [3]. Poster sessions have become a norm in many academic conventions and open laboratories because of the flexible and interactive characteristics. In most cases, however, paper posters are still used even in the ICT areas. In some cases, digital devices such as LCD and PC projectors are used, but they do not have sensing devices. Currently, many lectures in academic events are recorded and distributed via Internet, but recording of poster sessions is never done or even tried.

Poster conversations have a mixture characteristics of lectures and meetings; typically a presenter explains his/her work to a small audience using a poster, and the audience gives feedbacks in real time by nodding and verbal backchannels, and occasionally makes questions and comments. Conversations are interactive and also multi-modal because participants are standing and moving unlike in meetings. Another good point of poster conversations is that we can easily make a setting for data collection which is controlled in terms of familiarity with topics or other participants and yet is "natural and real".

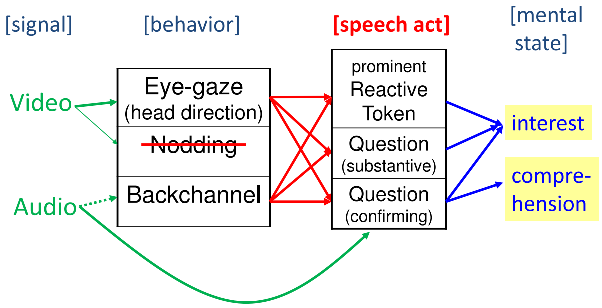

Fig. 1. Proposed scheme of multi-modal sensing and analysis

The goal of the project is signal-level sensing and high-level analysis of human interactions. Specific tasks include face detection, eye-gaze detection, speech separation, and speaker diarization. These will realize a new indexing scheme of poster session archives. For example, after a long session of poster presentation, we often want to get a short review of the question-answers and feedbacks from the audience.

We also investigate high-level indexing of which segment was attractive and/or difficult for the audience to follow. This will be useful in speech archives because people would be interested in listening to the points other people liked. However, estimation of the interest and comprehension level is apparently difficult and largely subjective. Therefore, we turn to speech acts which are observable and presumably related with these mental states. One is prominent reactive tokens signaled by the audience and the other is questions raised by them. Prediction of these speech acts from multi-modal behaviors is expected to approximate the estimation of the interest and comprehension level. The scheme is depicted in Figure 1.

In the remainder of the paper, the steps in Figure 1 are explained after a brief description of the multi-modal corpus in Section 2. Section 3 gives an overview process of audio-visual sensing of conversation participants and their speech using the ambient devices installed in the posterboard (green lines). In

Section 4, the relationship between the speech acts and mental states of the audience is analyzed to define the interest and comprehension level (blue lines). In Section 5, prediction of the concerned speech acts from the audience's multi-modal behaviors such as eye-gaze and backchannels (read lines) is addressed.

II. Multi-modal Corpus of Poster Conversations

We have recorded a number of poster conversations for multi-modal interaction analysis [3], [4]. In each session, one presenter (labeled as "A") prepared a poster on his/her own academic research, and there was an audience of two persons (labeled as "B" and "C"), standing in front of the poster and listening to the presentation. Each poster is designed to introduce research topics of the presenter to researchers or students in other fields. It consists of four or eight components (hereafter called "slide topics") of rather independent topics. The audience subjects were not familiar with the presenter and had not heard the presentation before. The duration of each session was 20-30 minutes. Some presenters made a presentation in two sessions, but to a different audience.

All speech data were segmented into IPUs (Inter-Pausal Unit) and sentence units with time and speaker labels, and transcribed according to the guideline of the Corpus of Spontaneous Japanese (CSJ) [5]. We also manually annotated fillers, laughter and verbal backchannels.

The recording environment for the "gold-standard" corpus was equipped with multi-modal sensing devices such as cameras and a motion capturing system while every participant wore an eye-tracking recorder and motion capturing markers. Eye-gaze information is derived from the eye-tracking recorder and the motion capturing system by matching the gaze vector against the position of the other participants and the poster.

III. Detection of Participants' Eye-gaze and Speech

We have designed and implemented a smart posterboard, which can record a poster session and sense human behaviors. Since it is not practical to ask every participant to wear special devices such as a head-set microphone and an eye-tracking recorder and also to set up any devices attached to a room, all sensing devices are attached to the posterboard, which is actually a 65-inch LCD screen. Specifically, the digital posterboard is equipped with a 19-channel microphone array on the top, and attached with six cameras and two Kinect sensors. An outlook of the posterboard is given in Figure 2.

Detection of participants and their multi-modal feedback behaviors such as eye-gaze and speech using these sensing devices, as shown in green lines in Figure 1, is explained in the following.

A. Eye-gaze Detection

We use the Kinect sensor to detect the participants' face and their eye-gaze. As it is difficult to detect the eye-ball with the Kinect's resolution, the eye-gaze is approximated with the head orientation. A preliminary analysis using the eye-tracking recorder showed that the difference between the actual eye-gaze and the head orientation is 10 degrees on average, but it is much smaller when the participants look at the poster. The process of the head orientation detection is as follows:

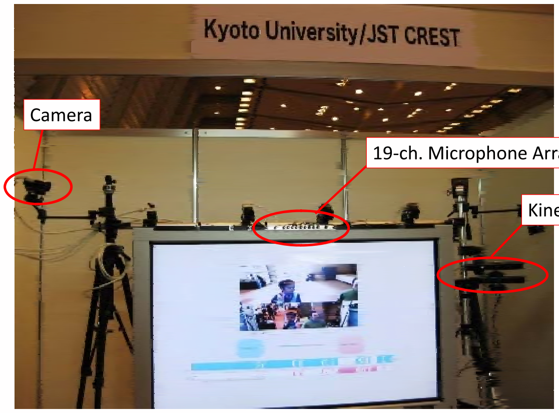

Fig. 2. Outlook of smart posterboard

1) Face detection
Haar-like features are extracted from the color and ToF (Time-of-Flight) images to detect the face of the participants. Multiple persons can be detected simultaneously even if they move around.
2) Head model estimation
For each detected participant, a three-dimensional shape of the head and colors are extracted from the ToF image and the color image, respectively. Then, a head model is defined with the polygon and texture information.
3) Head tracking
Head tracking is realized by fitting the video image into the head model. We adopt the particle filter to track the three-dimensional position of the head and its three-dimensional orientation.
4) Identification of eye-gaze object
From the six-dimensional parameters, we can compute an eye-gaze vector in the three-dimensional space. The object of the eye-gaze is determined by this vector and the position of the objects. In this study, the eye-gaze object is limited to the poster and other participants.

The entire process mentioned above can be run in real time by using a GPU for tracking each person.

B. Detection of Nodding

Nodding can be detected as a movement of the head, whose position is estimated in the above process. However, discrimination against noisy or unconscious movements is still difficult.

C. Speech Separation and Voice Activity Detection

Speech separation and enhancement are realized with the blind spatial subtraction array (BSSA), which consists of the delay-and-sum (DS) beamformer and a noise estimator based on independent component analysis (ICA) [6]. Here, the position information of the participants estimated by the image processing is used for beamforming and initialization of the ICA filter estimation. This is one of the advantages of multi-modal signal processing. While the participants move around, the filter estimation is updated online.

When we use the 19-channel microphone array, speech separation and enhancement can be performed with a high SNR, but not in real time. Using the Kinect sensor realizes real-time processing, but degrades the quality of speech.

By this process, the audio input is separated to the presenter and the audience. Although discrimination among the audience is not done, DoA (Direction of Arrival) estimation can be used for identifying the speaker among the audience. Voice activity detection (VAD) is conducted on each of the two channels by using power and spectrum information in order to make speaker diarization. We can use highly-enhanced but distorted speech for VAD, but still keeps moderately-enhanced and intelligible speech for re-playing.

D. *Detection of Reactive Tokens and Laughter*

We have investigated detection of reactive tokens and laughter, which are major reaction of the audience [7]. Reactive tokens are mostly non-lexical backchannels and indicate the listener's mental state. We focus on the particular syllabic patterns with prominent prosodic patterns which are related with the interest level [8]. Laughter can be easily detected by preparing a dedicated GMM, but it was observed that laughter is not so frequent and often used for mood relaxation in poster conversations [8].

IV. Definition of Interest and Comprehension Level through Speech Acts

In order to get a "gold-standard" annotation, it would be a natural way to ask every participant of the poster conversations on the interest and comprehension level on each slide topic after the session. However, this is not possible in a large scale and also for the previously recorded sessions. The questionnaire results may also be subjective and difficult to assess the reliability.

Therefore, we focus on observable speech acts which are closely related with the interest and comprehension level, as shown in blue lines in Figure 1. Previously, we found particular syllabic and prosodic patterns of reactive tokens ("*he:*", "*a:*", "*fu:N*" in Japanese, corresponding to "wow" in English) signal interest of the audience [9]. Ward [10] also investigated similar prosodic patterns of reactive tokens in English. We refer to them as prominent reactive tokens.

We also empirically know that questions raised by the audience signal their interest; the audience ask more questions to know more and better when they are more attracted to the presentation. Furthermore, we can judge the comprehension level by examining the kind of questions; when the audience asks something already explained, they must have a difficulty in understanding it.

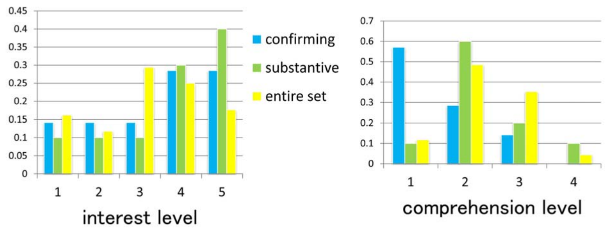

Fig. 3. Distribution of interest & comprehension level according to question type

Questions are classified into two types: confirming questions and substantive questions. The confirming questions are asked to make sure of the understanding of the current explanation, thus they can be answered simply by "Yes" or "No". The substantive questions, on the other hand, are asking about what was not explained by the presenter, thus they cannot be answered by "Yes" or "No" only; an additional explanation is needed. Substantive questions are occasionally comments even in a question form.

For four sessions collected most recently, we asked audience subjects to answer their interest and comprehension level on each slide topic after the session. Although the data size is small, we preliminarily investigate the relationship between these "gold-standard" annotations and observed questions.

Figure 3 shows distributions of the interest and comprehension level for each question type. The interest level was quantized into five levels from 1 (not interested) to 5 (very interested), and the comprehension level was marked from 1 (did not understand) to 4 (fully understood). In the graph, a majority of confirming questions (86%) indicate a low comprehension level (level 1&2). We also see a general tendency that occurrence of questions of either types is correlated with a higher interest level (level 4&5).

From these observations and the previous findings [9], we adopt the following annotation scheme for the entire corpus.

- high interest level ← questions of any types and/or prominent reactive tokens.
- low comprehension level ← confirming questions.

Note that annotation of these states is done for each topic segment and for each person in the audience (hereafter called "topic segments"). Detection of these states would be particularly useful in reviewing the poster sessions or improving the presentations.

V. Prediction of Speech Acts from Multi-modal Behaviors

In the previous section, we formulate prediction of interest and comprehension level with prediction of the relevant speech acts. Specifically, we reduce the estimation of the interest level to prediction of occurrence of questions and prominent reactive tokens, and the estimation of comprehension level to classification of the question type. The prediction and classification are done based on non-verbal feedback behaviors of the audience, without referring to the actual utterances, which usually occur at the end of topic segments. Automatic speech recognition is not realistic for distant spontaneous speech. Prediction is important to assess the mental state of the audience even if the speech acts are not done.

In this section, we investigate the relationship between the audience's multi-modal feedback behaviors and the speech acts. We conduct prediction of the speech acts as an approximate estimation of the interest and comprehension level. We used ten sessions in this study. There are 58 slide topics in total. Since two persons participated as an audience in each session, there are 116 topic segments for which the interest and comprehension level should be estimated.

First, each of audience behaviors needs to be parameterized. We focus on backchannel and eye-gaze behaviors. An average count of backchannels is computed per the presenter's utterance. Eye-gaze at the presenter is parameterized into an occurrence count per the presenter's utterance and the duration ratio within the topic segment.

Then, regarding the machine learning method for classification, we adopt a naive Bayes classifier, as the data size is not so large to estimate extra parameters such as weights of the features. For a given feature vector $F = \{f_1, \ldots, f_d\}$, a naive Bayes classification is done by

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

where c is a considered class. For computation of $p(f_i|c)$, we adopt a simple histogram quantization, in which feature values are classified into one of bins, instead of assuming a probabilistic density function. This also circumvents estimation of any model parameters. The feature bins are defined by simply splitting a histogram into 3 or 4. Then, the relative occurrence frequency in each bin is transformed into the probability form.

We set up two tasks defined in the previous ection. First, we conduct experiments of estimating the interest level of the audience in each topic segment. This problem is formulated by predicting the topic segment in which questions and/or prominent reactive tokens are made by the audience. We regard these topic segments as "interesting" to the person who made such speech acts.

Experimental evaluations were done by the leave-one-out cross validation manner. The results with different sets of features are listed in Table I. F-measure is a harmonic mean of recall and precision of "interesting" segments, though recall and precision are almost same in this experiment. Accuracy is a ratio of correct output among all 116 topic segments. The chance-rate baseline when we count all segments as "interesting" is 49.1%. Incorporation of the backchannel and eye-gaze features significantly improved the accuracy, and the combination of both features results in the best accuracy of over 70%. It turned out that the two parameterizations of the eye-gaze feature (occurrence count and duration ratio) are redundant because dropping one of them does not degrade the performance. However, we confirm the multi-modal synergetic effect of the backchannel and eye-gaze information.

TABLE I
PREDICTION RESULT OF TOPIC SEGMENTS INVOLVING QUESTIONS AND/OR REACTIVE TOKENS ("INTERESTING" LEVEL)

	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%

TABLE II
IDENTIFICATION RESULT OF CONFIRMING OR SUBSTANTIVE QUESTIONS ("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%

Next, we conduct experiments of estimating the comprehension level of the audience in each topic segment. This problem is formulated by identifying the confirming question given a question, which signal that the person does not understand the topic segment. Namely, we regard these topic segments as "low comprehension (difficult to understand)" for the person who made the confirming questions.

The classification results of confirming questions vs. substantive questions are listed in Table II. In this task, the chance-rate baseline based on the prior statistic $p(c)$ is 51.3%. All features have some effects in improving the accuracy, but the eye-gaze occurrence count alone achieves the best performance and combining it with other features does not give an additional gain.

VI. Conclusions

We have addressed multi-modal conversation analysis focused on poster sessions. Poster conversations are interactive, but often long and redundant. Therefore, simple recording of the session is not so useful.

Our primary goal is robust signal-level sensing of participants, i.e. who came to the poster, and their verbal feedbacks, i.e. what they said. This is still challenging given distant and low-resolution sensing devices, but we expect that a combination of multi-modal information sources will enhance the performance.

The next step is high-level indexing of interest and comprehension level of the audience, which are presumably useful for browsing the recording. We formulate the problem via relevant speech acts and show that reasonable performance can be achieved given non-verbal feedback behaviors of the audience.

We are also developing a poster session browser to visualize these detected events and indexes. It will be useful for assessing the effect of the processes and further improving them.

Acknowledgments

This work was supported by JST CREST. It was conducted by the members of the CREST project including Hiromasa Yoshimoto, Tony Tung, Yukoh Wakabayashi, Soichiro Hayashi, Koji Inoue, Katsuya Takanashi (Kyoto University) and Yuji Onuma, Shunsuke Nakai, Ryoichi Miyazaki, Hiroshi Saruwatari (Nara Institute of Science and Technology).

References

[1] S.Renals, T.Hain, and H.Bourlard. Recognition and understanding of meetings: The AMI and AMIDA projects. In *Proc. IEEE Workshop Automatic Speech Recognition & Understanding*, 2007.

[2] K.Ohtsuka. Conversation scene analysis. *Signal Processing Magazine*, 28(4):127–131, 2011.

[3] T.Kawahara. Multi-modal sensing and analysis of poster conversations toward smart posterboard. In *Proc. SIGdial Meeting Discourse & Dialogue*, pages 1–9 (keynote speech), 2012.

[4] T.Kawahara, H.Setoguchi, K.Takanashi, K.Ishizuka, and S.Araki. Multi-modal recording, analysis and indexing of poster sessions. In *Proc. INTERSPEECH*, pages 1622–1625, 2008.

[5] K.Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12, 2003.

[6] Y.Takahashi, T.Takatani, K.Osako, H.Saruwatari, and K.Shikano. Blind spatial subtraction array for speech enhancement in noisy environment. *IEEE Trans. Audio, Speech & Language Process.*, 17(4):650–664, 2009.

[7] K.Sumi, T.Kawahara, J.Ogata, and M.Goto. Acoustic event detection for spotting hot spots in podcasts. In *Proc. INTERSPEECH*, pages 1143–1146, 2009.

[8] T.Kawahara, K.Sumi, Z.Q.Chang, and K.Takanashi. Detection of hot spots in poster conversations based on reactive tokens of audience. In *Proc. INTERSPEECH*, pages 3042–3045, 2010.

[9] T.Kawahara, Z.Q.Chang, and K.Takanashi. Analysis on prosodic features of Japanese reactive tokens in poster conversations. In *Proc. Int'l Conf. Speech Prosody*, 2010.

[10] N.Ward. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody*, pages 325–328, 2004.