# Multi-Modal Conversational Analysis of Poster Presentations Using Multiple Sensors

Hisao SETOGUCHI[†]
setoguti@ar.media.kyoto-u.ac.jp

Katsuya TAKANASHI[‡]
takanasi@ar.media.kyoto-u.ac.jp

Tatsuya KAWAHARA[†‡]
kawahara@i.kyoto-u.ac.jp

[†]Graduate School of Informatics, Kyoto University
[‡]Academic Center for Computing and Media Studies, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

## ABSTRACT

We are constructing a research environment called the "IMADE room", which can capture a variety of multi-modal human interactions. With this setting, we have designed and conducted recordings of poster sessions made by one presenter and two audiences. In addition to speech data of individual participants, gazing, nodding, and pointing behaviors are recorded through multiple sensors. This article describes the specifications of the data collection and preliminary analyses of the relationship between verbal and non-verbal behaviors.

## 1. INTRODUCTION

Currently we can easily record and archive meetings and conversations as digital archives. Many researchers have studied semi-automatic generation of indexes of audio and video materials included in these archives to enable them to be accessed more efficiently[5][8][7]. In daily conversations, people have interactions through many modalities, thus it is desirable to combine several sorts of information from these modalities to extract meta-information and create a meaningful index for archives. In this work, we investigate what sort of information should be combined and how to make use of meta-information to implement this approach.

We are developing an experimental environment where we can record audio/video, human-motion, and eye-movement, and have chosen the domain of poster presentations as our initial application domain. In addition, some verbal/non-verbal behaviors are annotated on the data to conduct preliminary analyses of these data. Specifically, we focus on the presenter's utterances and gazes and on audiences' backchannels and nodding, and investigate their correlations in terms of temporal distribution.

## 2. CLASSIFICATION OF CONVERSATIONS

Many forms of conversations have been studied based on the concreteness of their goal, information resources used, and participants' roles. They are classified as a "conversation map" shown in Table 1, in which we plot with axes of the concreteness of goals and information resources. Among these kinds of conversations, we chose poster presentations for our research. The reasons are given below:

Table 1: Conversation Map

| Goals/Resources | None | Static resources | Dynamic resources |
|---|---|---|---|
| Goal-observable | Ticket reservations | Map tasks | Cooking tutorials |
| Goal-oriented | Debates | Seminars | |
| Theme-oriented | Meetings | Poster presentations | |
| Vague | Chattering | | |

First, we can specify the themes and content of speaking in terms of the concreteness of goals in comparison with chatting where we cannot. On the other hand, we can converse more freely in presentations than in task completion, such as making ticket reservations and asking for directions.

Second, we do not always require information resources at meetings; however, at poster presentations, we at least have a poster as a static information resource, and we can actually exploit the information on it in conversation. However, in conversations on cooking, for instance, we move or make changes to the shapes of objects in the real world. In doing so, we use physical objects as dynamic information resources and proceed on the conversation. Poster sessions represent an intermediate form of conversation with limited information resources and abundant dynamic resources.

Also note that many researches have been conducted on meetings[1], seminars[3], or map tasks[2] in Table 1, but few have dealed with poster presentations, which is one reason we adopted it in this work.

In terms of contrast in participants' roles, those of the presenter and audiences are fixed in poster presentations. This differs from meetings where no participants' roles are fixed. However, different from lectures, audiences may ask questions even while the presenter is talking, and take the initiative in the conversation. Therefore, poster presentations represent an intermediate form between meetings and lectures. In addition, there is inequality in the amount of information between the presenter and audiences in poster sessions and in their respective roles. As a result, the presenter is mostly giving information through his utterances, while the audiences' behaviors mainly involve receiving the


*Workshop on Tagging, Mining and Retrieval of Human Related Activity Information* 15 November 2007 Nagoya, Japan

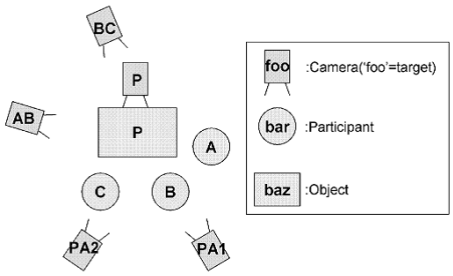


Figure 1: Camera and Poster Settings

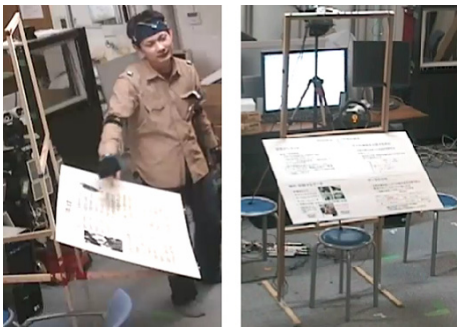

Figure 2: Specially Designed Poster Stand

information. We can make use of this characteristic of poster sessions to simplify our analysis by separating actions into presenter-specific and audience-specific behaviors.

We have recorded poster sessions conducted by one presenter with the participation of two audiences. Our aims in this setting are to detect interactions between the audiences and to analyze each audience's behavior by comparison.

# 3. ENVIRONMENT FOR MULTI-MODAL DATA COLLECTION

## 3.1 IMADE Room

We are developing a recording environment called the IMADE Room at the Faculty of Engineering of Kyoto University as a project under a Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) Grant-in-Aid for Scientific Research on Priority Areas called the "Info-plosion IT Research Platform"[6]. This environment is capable of recording audio/visual, human-motion, and physiological data to archive many kinds of multi-modal interactions.

## 3.2 Sensors used in Recording

We used many kinds of sensors to record audio, video, human motion, and eye movements. We used a wireless head-worn microphone (Shure WH-30 XLR) in order to record all participants' voices separately, and to enable them to move freely in the room at the same time. The voice signals were amplified through an ATI audio pre-amplifier and digitized at 48 kHz as 16-bit PCM format data. In addition to this, we installed an array composed of eight microphones above the poster on trial. The use of microphone array system would be more desirable if the quality of recorded and enhanced audio reaches sufficient level in the future, for it is a non-contact device of audio recording.

The IMADE room provides eight Sony built-in type cameras for the visual data recording and data can be encoded

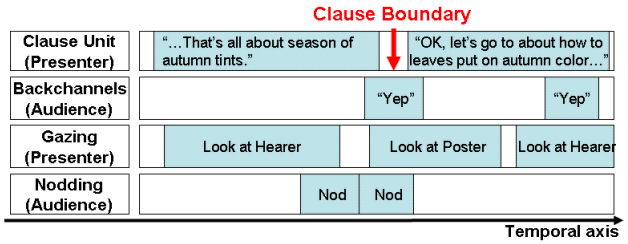


Figure 3: Example of Annotations

and stored in MPEG-2 format. It is indispensable that significant behaviors of all participants during the poster session be recorded, with at least one camera, without any of them occluded by the poster board or other participants. We kept this caution in mind and arranged the five cameras and the poster as shown in Fig. 1. Here, "P", "A", "B" and "C" indicates the poster, the presenter, and two audiences, respectively.

We also installed a motion-capturing system "PhaseSpace" to record the participants' motion. This device consists of a set of infra-red LEDs worn by the participants and the CCD sensors, embeded in the room, receiving optical signals from the LEDs to estimate the three-dimensional positions of all LEDs. The positional data were stored as a temporal data series at 30 Hz. We arranged the poster mount so that it would not occlude the LEDs and render the participants' motions unnatural. We therefore designed an easel and poster mount, as shown in Fig. 2. The poster mount was angled at 22° from the horizontal attitude.

Moreover, we used eye-tracking recorders ASL "Mobile-Eye". It is composed of goggles connected to a digital video camera (own-view camera) and an infra-red laser generator, and can trace a participant's eye movements using infra-red laser reflections and overlay the focal point estimated from his eye movements on the video recorded by the own-view camera. It recorded the wearer's eye movement data as a temporal series of changing focal points on the own-view camera's video coordinates.

# 4. RECORDING AND ANNOTATION OF CONVERSATIONS

## 4.1 Recorded Data

We recorded conversations in 11 poster sessions. However, due to system troubles, several sensor data other than audio were missing in some sessions. The duration of each session was around 20 minutes.

We had five different presenters in our experiment. Each of them had prepared a different poster on his own academic research and conducted two or three sessions. The poster had one main theme and was divided into 4 subtopics, which were arrayed in quarters on the surface of it. Two audiences in each session heard the explanation for the first time.

## 4.2 Annotations

We designed verbal and non-verbal annotations, as shown in Fig. 3. As for verbal behaviors, we transcribed units of speech segmented by more than 200ms pauses (Inter-Pausal Units: IPUs), and inserted labels of Clause Unit (CU) boundaries into the transcription with reference to the guidelines of the Corpus of Spontaneous Japanese (CSJ)[4].

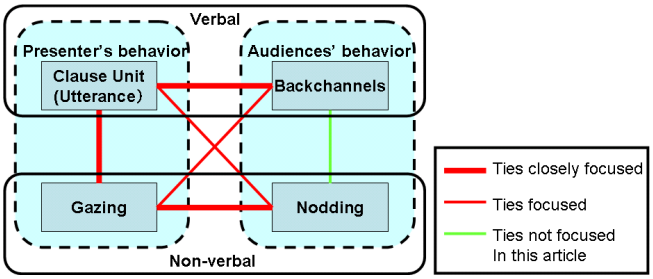


Figure 4: Ties in Behaviors in Scale of Presenter/Audience and Verbal/Non-verval

In addition, backchannels were also labeled according to our own criterion.

As for annotations of non-verbal behaviors, we labeled nodding and gazing by referring to the video data. These kinds of information are important for both developing models of non-verbal behaviors occurred in conversations and evaluating the accuracy of automatic detection using motion capture or eye-tracking data. Each record of annotations consists of at least three columns: StartTime, EndTIme, and contents (or the sort of behavior), and for verbal annotations, there is the fourth column: ParticipantID.

In this preliminary research, we annotated four kinds of behaviors mentioned above, that is, clause boundaries in presenter's utterances, gaze of a presenter, and backchannels and nodding of audiences, appeared in the first two sub-topics of a single session, whose duration was 411.196 seconds. Although there are many sorts of verbal and non-verbal behaviors in conversations, we focus on these four kinds of behaviors because we regard them as primary and characteristic behaviors by presenter or audiences.

# 5. ANALYSIS OF PARTICIPANTS' BEHAVIORS IN CONVERSATIONS

We analyze the temporal correlations between presenter's and audiences' behaviors. As shown in Fig. 4, we can classify them as to whether they are the presenter's or audiences', and whether they are verbal or non-verbal, and examine the co-occurrences of each pair of them.

It was expected that audiences display some responses to significant boundaries of the presenter's utterances or his gazes. Firstly, we examine the relationship between boundaries in the presenter's utterances and the audiences' backchannels to explore the verbal level of interactions between them (Sec. 5.1). Secondly, we examine the relationship between the presenter's gazing at the audiences and nodding of the audiences to analyze non-verbal interactions (Sec. 5.2). Finally, we investigate the effect of integration of multi-modalities, namely, that the places where utterance boundaries accompany gazes of the presenter are one of most significant points where the audiences' responsive behaviors occur frequently (Sec. 5.3).

Since the role of participants is unequal in the poster presentations, the frequencies of utterances, backchannels, and nodding are also different between presenters and audiences. Thus, the audiences' utterances and the presenter's backchannels and nodding are overwhelmingly infrequent.

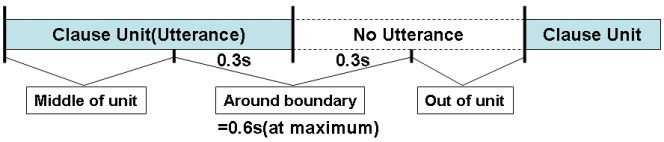


Figure 5: Definition of Durations Related with Clause Unit

Table 2: Audience B's responses to utterances

| Nodding | Frequency | Freq. per sec |
|---|---|---|
| Middle of unit | 82 | 0.24 |
| Around boundary | 16 | 0.45 |
| Backchannel | Frequency | Freq. per sec |
| Middle of unit | 43 | 0.13 |
| Around boundary | 17 | 0.48 |

Table 3: Audience C's responses to utterances

| Nodding | Frequency | Freq. per sec |
|---|---|---|
| Middle of unit | 53 | 0.16 |
| Around boundary | 11 | 0.31 |
| Backchannel | Frequency | Freq. per sec |
| Middle of unit | 42 | 0.12 |
| Around boundary | 21 | 0.59 |

## 5.1 Correlation of Presenter's Utterances with Audiences' Backchannels

We adopted the clause units explained in Sec. 4.2 as utterance units of a presenter. As shown in Fig. 5, the whole duration of the conversation can be differentiated into three kinds of periods of time: "middle of unit", "around boundary", and "out of boundary". As a general rule, we used a margin of 0.3 seconds around the end of clause units in this analysis; thus, the total duration for "around boundary" was 0.6 seconds at maximum.

We calculated the numbers of times of the audiences' behaviors and their frequency per second during "middle of unit" and "around boundary". The results are shown in Tables 2 and 3. There were 59 clause units. The sum of duration of "middle of unit" was 340.2 seconds, and that of "around boundary" was 35.4 seconds.

During "around boundary", audience B and C gave around 3.7 and 4.9 times as many backchannels as during "middle of unit", respectively. Similarly, as for nodding, during "around boundary", audience B and C nodded around twice as many compared to that during "middle of unit".

These results mean that there is a tendency for audiences to respond more often at the boundaries of utterances than in midstream. Furthermore, it is observed that the audiences use backchannels more often than nodding around the boundaries of the presenter's utterances.

## 5.2 Correlation of Presenter's Gaze with Audiences' Nodding

We conduct a similar analysis as in the previous section on the presenter's gaze and the audiences' responsive behaviors. In this analysis, we divided the entire duration of the conversation into two periods of time: duration when the presenter gazed at the audiences and duration when he did not. Tables 4 and 5 show the sum of duration when an audience is gazed at or not gazed at by the presenter, the number of audiences' responses in both periods of time, and their frequencies per second of audience B and C. The results about audience B and C are shown separately because, unlike speech, gazes cannot be directed to both audiences at the same time, but selectively to either B or C on a moment.

Table 4: Audience B's responses to gaze

| Nodding | Frequency | Freq. per sec |
|---|---|---|
| Gaze | 44 | 0.39 |
| Non-Gaze | 56 | 0.19 |
| Backchannels | Frequency | Freq. per sec |
| Gaze | 22 | 0.20 |
| Non-gaze | 40 | 0.13 |

Table 5: Audiences C's responses to gaze

| Nodding | Frequency | Freq. per sec |
|---|---|---|
| Gaze | 36 | 0.23 |
| Non-gaze | 31 | 0.12 |
| Backchannel | Frequency | Freq. per sec |
| Gaze | 38 | 0.24 |
| Non-gaze | 28 | 0.11 |

Table 6: Correlations between audiences' backchannels with presenter's utterances and gaze (mean for audiences)

| Utterances | Gaze | Duration | Frequency | Freq. per sec |
|---|---|---|---|---|
| Around boundary | On | 9.98 | 8.5 | 0.85 |
| Around boundary | Off | 22.18 | 10.5 | 0.47 |
| Middle of unit | On | 121.36 | 21.5 | 0.18 |
| Middle of unit | Off | 219.18 | 21 | 0.10 |
| Out of unit | On | 3.42 | 0 | 0.00 |
| Out of unit | Off | 35.09 | 2.5 | 0.07 |

Table 7: Correlations between audiences' nodding with presenter's utterances and gaze (mean for audiences)

| Utterances | Gaze | Duration | Frequency | Freq. per sec |
|---|---|---|---|---|
| Around boundary | On | 9.98 | 6 | 0.60 |
| Around boundary | Off | 22.18 | 7 | 0.32 |
| Middle of unit | On | 121.36 | 34.5 | 0.28 |
| Middle of unit | Off | 219.18 | 34.5 | 0.16 |
| Out of unit | On | 3.42 | 0.5 | 0.15 |
| Out of unit | Off | 35.09 | 1.5 | 0.04 |

The sum of duration when the presenter gazed at audience B and C were 112.6 seconds and 156.9 seconds, respectively. Those when the presenter did not gaze at B and C were 298.6 seconds and 254.3 seconds. We can see that the numbers of the audiences' responses per second during the presenter gazing at an audience were around 1.5 to 2.2 times as those without the presenter's gaze. This means that the presenter's gaze prompted the audiences to display some responses.

We can also find a significant difference in the number of backchannels and nods per second during the presenter gazing at audience B; however, for C, we cannot observe this tendency. Thus, we cannot conclude whether the differences in the number of backchannels and nods are a general principle or merely reflected individual differences.

### 5.3 Effect of the Integration of Multi-modalities

In Sec. 5.1 and 5.2, effects of the presenter's verbal and non-verbal behaviors were examined separately. However, it is expected that integration of the presenter's behaviors of both modalities has some special effects on the responsive behaviors of the audiences. In this section, we focus on the places where boundaries of utterances are accompanying the presenter's gaze.

As discussed in the previous sections, the presenter's utterances can be divided into three kinds of periods of time, and the period of the presenter's gazes was divided into two. Thus, by combining the utterance and gazing conditions, we can distinguish six sections for the analysis.

Tables 6 and 7 show the frequencies of the audiences' responses classified according to these periods. "Duration" in the tables indicates the sum of period of the presenter's behaviors described as a combination of "utterances" and "gazes" conditions. "Frequency" means the number of the audiences' responses and "freq. per sec" is frequency divided by duration in the same row. We take the mean of the two audiences here, for we did not find any differences in the tendencies of responses and we had too few samples.

When the presenter gazed at the audience in "around boundary", the frequency of the audiences' backchannels and nods was increased by around 1.8 to 4.7 times from that in "around boundary" without the presenter's gaze, or that in "middle of unit" with the gaze. This means that combination of utterance boundaries and presenter's gazing at the audience has a synergistic effect prompting audiences to show some responses, demonstrating the significance of integration of verbal and non-verbal modalities.

## 6. FUTURE TASKS

In this research, limited kinds and amount of annotations were utilized. However, other behaviors such as presenters' gazes, audience's pointing and other hand gestures are also important in analyzing conversations. With these kinds of data on the interactions, we can compare mutual gazes with non-mutual gazes, analyze what effects the presenter's pointing have on the audiences' behavior, and what the presenter's nodding means. We are undergoing a preliminary analysis of pointing. Furthermore, we intend to search for ways of automating annotation using data from the motion-capturing devices and eye-trackers.

## 7. ACKNOWLEDGEMENT

The authors are deeply indebted to the support by the members of Nishida & Sumi Laboratory and the members of Nakamura Laboratory of Kyoto University in developing the IMADE room. We also thank Dr. Shoko Araki and Mr. Kentaro Ishizuka of NTT Communication Science Laboratories for providing the microphone array equipments.

## 8. REFERENCES

[1] AMI Project. http://www.amiproject.org/.

[2] Anderson, A. H., M. Bader, E. G. Bard, E. H. Boyle, G. M. Doherty, S. C. Garrod, S. D. Isard, J. C. Kowtko, J. M. McAllister, J. Miller, C. F. Sotillo, H. S. Thompson and R. Weinert. The HCRC Map Task Corpus. Language and Speech, 34:351–366, 1991.

[3] CHIL - Computers In the Human Interaction Loop. http://chil.server.de/servlet/is/101/.

[4] the Corpus of Spontaneous Japanese. http://www.kokken.go.jp/katsudo/seika/corpus/public/.

[5] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio. Detecting group interest level in meetings. In Proc. Int'l Conf. Acoustics, Speech and Signal Processing, volume 1, pages 489–492, 2005.

[6] info-plosion. http://www.infoplosion.nii.ac.jp/i-explosion/ctr.php/m/IndexEng/a/Index/.

[7] ARDA VACE II:ENVIE. http://www.informadia.cs.cmu.edu/arda/vaceII.html.

[8] B. Wrede and E. Shriberg. Spotting "Hot Spots" in Meetings: Human Judgments and Prosodic Cues. In Proc. EUROSPEECH, pages 2805–2808, 2003.