Estimation of Interest and Comprehension Level of Audience through Multi-modal Behaviors in Poster Conversations

Tatsuya Kawahara, Soichiro Hayashi, Katsuya Takanashi

School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract

We address the estimation of the interest and comprehension level of an audience in poster sessions. Compared to lecture presentations, the audience's behaviors such as gazing and backchannels are more observable in poster presentations. These multi-modal behaviors are presumably related with their interest and comprehension level. We also assume that the interest and comprehension level can be judged by particular speech acts of the audience such as questions and reactive tokens. First, we make a preliminary analysis on their correlation. Next, we investigate the relationship between the audience's behaviors and the question type. Then, we conduct prediction of questions and their type based on the multi-modal behaviors during the relevant topic segment. Experimental results show that verbal backchannels and eye-gaze patterns are good predictors to this task, and also the combination of the multi-modal features is effective.

Index Terms: multi-modal interaction, behavioral analysis, eye-gaze, backchannel

1. Introduction

Human speech communication is intrinsically bi-directional and duplex, and feedback behaviors play an important role in smooth communication [1]. Feedback behaviors of an audience are important cues in analyzing presentation-style conversations. We can guess whether the audience is attracted to the presentation by observing their feedback behaviors. This characteristic is more prominent when the audience is smaller; the audience can make not only non-verbal feedbacks such as nodding, but also verbal backchannels. Eye-gaze behaviors also becomes more observable, playing an important role in turn-taking by the audience.

We have been collecting and analyzing poster conversations, in which a researcher makes an academic presentation to a couple of persons using a poster [2]. Poster sessions have become a norm in many academic conventions including InterSpeech conferences because of the interactive characteristics. An audience can ask questions even during the presentation. By observing their reactions, particularly the quantity and quality of their questions and comments, we can guess whether the presentation is understood or liked by the audience.

In our previous work [3], we demonstrated that non-lexical kinds of verbal backchannels, referred to as reactive tokens, are a good indicator of the audience's interest level. We also investigated the relationship between the audience's turn-taking and feedback behaviors including backchannels and eye-gaze patterns [4]. The relation of turn-taking with verbal and non-verbal behaviors has been studied in other previous works, too [5, 6, 7].

The goal of this work is to estimate the interest and comprehension level of the audience based on these multi-modal behaviors. As annotation of the interest and comprehension level is apparently difficult and largely subjective, 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. Moreover, we classify questions into confirming questions and substantive questions. Prediction of these speech acts from the multi-modal behaviors is expected to approximate the estimation of the interest and comprehension level. Whereas involvement in conversations has been investigated from the viewpoint of multi-modal interactions [8], this work has a clear target on the interest and comprehension level of the poster presentation.

The multi-modal corpus and the problem setting are described in Section 2 and 3. We first give an analysis on the relationship between the audience's multi-modal behaviors and the concerned speech acts in Section 4. Then, we report experiments to predict the speech acts from the backchannel and eye-gaze behaviors, which provide estimation of the interest and comprehension level, in Section 5.

2. Multi-modal Corpus of Poster Conversations

We have recorded a number of poster conversations for multi-modal interaction analysis [2, 9]. In this study, we use ten poster sessions. 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. They 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) [10]. We also manually annotated fillers and verbal backchannels.

The recording environment 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.

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. This design is a bit different from typical posters presented in academic conferences such as InterSpeech, but makes it straightforward to assess the interest and comprehension level of the audience for each slide topic. Usually, a poster conversation proceeds with an explanation of slide topics one by one, and is followed by an overall QA and discussion phase. In the QA/discussion phase, it is difficult to annotate which topic they refer. Therefore, we focus on the conversation segments of the explanation on the slide topics.

In the ten sessions used in this study, there are 58 slide topics in total. Since two persons participated as an audience in each session, there are 116 slots (hereafter called "topic segments") for which the interest and comprehension level should be estimated.

3. Definition of Interest and Comprehension Level

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. 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 [11]. Ward [12] 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.

3.1. Annotation of Question Type

Questions are classified into two types: confirming questions and substantive questions. [1] 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". [2] 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. Substantial questions are occasionally comments even in a question form.

3.2. Relationship between Question Type and Interest & Comprehension Level

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.

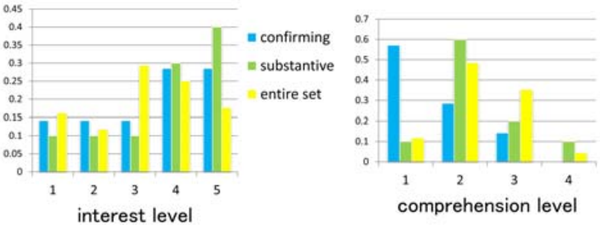

Figure 1: Distribution of interest & comprehension level according to question type

Figure 1 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 5 (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 [3], we adopt the following annotation scheme for the entire topic segments, which is used in the following sections.

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

The detection of these states would be particularly useful in reviewing the poster sessions or improving the presentations.

4. Relationship between Feedback Behaviors and Questions

Next, we investigate statistics of backchannel and eye-gaze behaviors of the audience and their relationship with questions asked by them.

4.1. Backchannels

Verbal backchannels, typically "*hai*" in Japanese and "yeah" or "okay" in English, indicate that the listener is attentive to what is being said. They also suggest the listener's interest level [14]; the listener tends to make backchannels more frequently when they are attracted. In this analysis, non-lexical reactive tokens (e.g. "wow") are excluded since the prominent part of them are used for the annotation, though their occurrence frequency is much smaller (less than 20% of all) than that of the lexical tokens (e.g. "yeah" and "okay").

Nodding is regarded as a non-verbal backchannel, and it is more frequently observed in poster conversations than in daily conversations. Our preliminary analysis showed, however, that there is not a distinct tendency in the occurrence frequency of non-verbal noddings, thus we focus on the verbal backchannel in this work.

The occurrence frequency of the verbal backchannels normalized by the presenter's utterance (sentence unit) is counted within the topic segments. The statistics are listed according to the question type in Tables 1. In the table, "entire" means the overall average computed for the entire topic segments of the data set. Since no questions were made in more than a half topic segments, the entire average is lower than the values in the other two columns. It is observed that the audience make

[1] Strömbergsson et al. [13] defined "backward questions" and "forward questions" for the similar classification.

[2] This does not mean the presenter actually answered simply by "Yes" or "No".

Table 1: Relationship of audience's backchannel (count/utterance) and questions (by type)

	confirming	substantive	entire
backchannel	0.42	0.52	0.34

Table 2: Relationship of audience's eye-gaze at the presenter (count/utterance and duration ratio) and questions (by type)

	confirming	substantive	entire
gaze occurrence	**0.38**	**1.02**	0.64
gaze duration	0.05	**0.15**	0.07

more backchannels when asking questions, especially substantive questions.

4.2. Eye-gaze at Presenter

We identify the object and the duration of the eye-gaze of all participants during the topic segments, especially prior to the audiences' questions. The target object can be either the poster or other participants. In poster conversations, unlike daily conversations, participants look at the poster in most of the time. Therefore, eye-gaze at other participants has a reason and effect. Our previous work [4] showed that eye-gaze information is related with turn-taking events; specifically, the eye-gaze by the presenter mostly controls the turn-taking.

In this work, we focus on the eye-gaze by the audience and investigate its relationship with the questions they ask. In particular, we count the eye-gaze of each person of the audience at the presenter. We measure the average occurrence count (per presenter's utterance) and the total duration (normalized per second) within the topic segments. Their statistics are listed in Table 2. We can see a significant decrease and increase when asking confirming questions and substantive questions, respectively. We reason that the audience is more focused on the poster trying to understand the content before asking confirming questions, while they want to attract the presenter's attention before asking substantive questions.

In a more detailed analysis done sentence by sentence, a gradual increase of the eye-gaze at the presenter is observed prior to substantive questions, while there is no such dynamic changes in the case of confirming questions.

The results suggest that the eye-gaze information is potentially useful for identifying the question type and also estimating the interest and comprehension level.

5. Prediction of Questions and Reactive Tokens – Estimation of Interest Level

Based on the analyses in the previous section, we conduct experiments of estimating the interest level of the audience in each topic segment. As described in Section 2, 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.

First, each of audience behaviors needs to be parameterized. We use the features described in the previous section. We compute an average count of backchannels 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.

Table 3: Prediction result of topic segments involving questions and/or reactive tokens

	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%

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 and "interesting or not" in this task. 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.

Experimental evaluations were done by the leave-one-out cross validation manner. The results with different sets of features are listed in Table 3. 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.

6. Identification of Question Type – Estimation of Comprehension Level

Next, we conduct experiments of estimating the comprehension level of the audience in each topic segment. As described in Section 2, 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.

We adopt the same features and the classifier as in the previous section. The classification results of confirming questions vs. substantive questions are listed in Table 4. 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. This is explained by a large difference in

Table 4: Identification result of confirming or substantive questions

	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%

its value among the question types as shown in Table 2. We also tried to incorporate local dynamic features computed for the two utterances prior to the questions, but did not obtain an improvement.

As the simple occurrence frequency of backchannels is not useful for this task, the syllabic or prosodic patterns of the backchannels should be investigated in the future. Strömbergsson et al. [13] investigated prosodic patterns, especially pitch patterns of the questions, according to the question type. The feature may also be useful for this task.

7. Conclusions and Future Perspective

We have investigated the relationship between the audience's feedback behaviors and speech acts such as questions and prominent reactive tokens within topic segments, by assuming that these speech acts indicate their interest and comprehension level. 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.

First, we confirmed the validity of these problem settings via a questionnaire and annotated the topic segments. Next, we made an analysis on the feedback behaviors such as backchannels and eye-gaze at the presenter, and found their typical patterns according to the type of questions. Then, we conducted experiments of estimating the interest and comprehension level via the relevant speech acts.

It is confirmed that by combining the multi-modal features, prediction accuracy of the interest level was improved from the chance rate of around 50% to over 70%. Identification of confirming questions which indicates a low comprehension level was done with an accuracy of 75%. It is possible that the audience do not ask any questions when they do not understand the content. But our primary target is those who were interested in the slide topic but had difficulty in comprehension.

We are also developing a smart posterboard [2] which can control cameras and a microphone array to record poster sessions and annotate the audience's reaction. The work presented in this paper provides high-level annotations, which will be useful in browsing the poster sessions.

Acknowledgement: This work was supported by JST CREST and JSPS Grant-in-Aid for Scientific Research.

8. References

[1] N.Ward, D.Novick, L.P.Morency, T.Kawahara, D.Heylen, and J.Edlund, editors. *Proc. Interdisciplinary Workshop on Feedback Behaviors in Dialog*, 2012.

[2] 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.

[3] 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.

[4] T.Kawahara, T.Iwatate, and K.Takanashi. Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations. In *Proc. INTERSPEECH*, 2012.

[5] N.G.Ward and Y.A.Bayyari. A case study in the identification of prosodic cues to turn-taking: Back-channeling in Arabic. In *Proc. INTERSPEECH*, pages 2018–2021, 2006.

[6] B.Xiao, V.Rozgic, A.Katsamanis, B.R.Baucom, P.G.Georgiou, and S.Narayanan. Acoustic and visual cues of turn-taking dynamics in dyadic interactions. In *Proc. INTERSPEECH*, pages 2441–2444, 2011.

[7] K.Jokinen, K.Harada, M.Nishida, and S.Yamamoto. Turn-alignment using eye-gaze and speech in conversational interaction. In *Proc. INTERSPEECH*, pages 2018–2021, 2011.

[8] C.Oertel, S.Scherer, and N.Campbell. On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. In *Proc. INTERSPEECH*, pages 1541–1545, 2011.

[9] 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.

[10] 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.

[11] 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.

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

[13] S.Strombergsson, J.Edlund, and D.House. Prosodic measurements and question types in the spontal corpus of Swedish dialogues. In *Proc. INTERSPEECH*, 2012.

[14] T.Kawahara, M.Toyokura, T.Misu, and C.Hori. Detection of feeling through back-channels in spoken dialogue. In *Proc. INTERSPEECH*, page 1696, 2008.