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

Tatsuya Kawahara
Kyoto University, Academic Center for Computing and Media Studies
Sakyo-ku, Kyoto 606-8501, Japan
http://www.ar.media.kyoto-u.ac.jp/crest/

Abstract

Conversations in poster sessions in academic events, referred to as poster conversations, pose interesting and challenging topics on multi-modal analysis of multi-party dialogue. This article gives an overview of our project on multi-modal sensing, analysis and "understanding" of poster conversations. We focus on the audience's feedback behaviors such as non-lexical backchannels (reactive tokens) and noddings as well as joint eye-gaze events by the presenter and the audience. We investigate whether we can predict when and who will ask what kind of questions, and also interest level of the audience. Based on these analyses, we design a smart posterboard which can sense human behaviors and annotate interactions and interest level during poster sessions.

1 Introduction

As a variety of spoken dialogue systems have been developed and deployed in the real world, the frontier of spoken dialogue research, with engineering applications in scope, has been extended from the conventional human-machine speech interface. One direction is a multi-modal interface, which includes not only graphics but also humanoid robots. Another new direction is a multi-party dialogue system that can talk with multiple persons as an assistant agent (D.Bohus and E.Horvitz, 2009) or a companion robot (S.Fujie et al., 2009). While these are extensions of the human-machine speech interface, several projects have focused on human-human interactions such as meetings (S.Renals et al., 2007) and free conversations (K.Otsuka et al., 2008; C.Oertel et al., 2011), toward ambient systems supervising the human communications.

We have been conducting a project which focuses on conversations in poster sessions, hereafter referred to as poster conversations. Poster sessions have become a norm in many academic conventions and open laboratories because of the flexible and interactive characteristics. 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 feedback in real time by nodding and verbal backchannels, and occasionally makes questions and comments. Conversations are interactive and also multimodal because people 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".

The goal of the project is signal-level sensing and high-level "understanding" of human interactions, including speaker diarization and annotation of comprehension and interest level of the audience. These will realize a new indexing scheme of speech archives. For example, after a long session of poster presentation, we often want to get a short review of the question-answers and what looked difficult for audience to follow. The research will also provide a model of intelligent conversational agents that can make autonomous presentation.

As opposed to the conventional content-based indexing approach which focuses on the presenter's

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 1–9,
Seoul, South Korea, 5-6 July 2012. ©2012 Association for Computational Linguistics

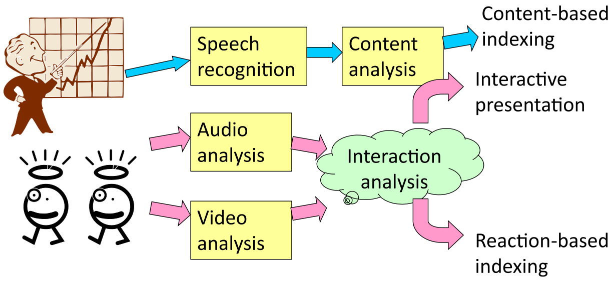

Figure 1: Overview of multi-modal interaction analysis

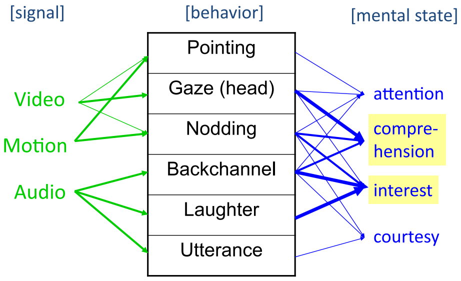

Figure 2: Flow of multi-modal sensing and analysis

speech by conducting speech recognition and natural language analysis, we adopt an interaction-oriented approach which looks into the audience's reaction. Specifically we focus on non-linguistic information such as backchannel, nodding and eye-gaze information, because we assume the audience better understands the key points of the presentation than the current machines. An overview of the proposed scheme is depicted in Figure 1.

Therefore, we set up an infrastructure for multi-modal sensing and analysis of multi-party interactions. Its process overview is shown in Figure 2. From the audio channel, we detect utterances as well as laughters and backchannels. We also detect eye-gaze, nodding, and pointing information. Special devices such as a motion-capturing system and eye-tracking recorders are used to make a "gold-standard" corpus, but only video cameras and distant microphones will be used in the practical system.

Our goal is then annotation of comprehension and interest level of the audience by combining these information sources. This annotation will be useful in speech archives because people would be interested in listening to the points other people were interested in. Since this is apparently difficult to be well-defined, however, we set up several milestones that can be formulated in objective manners and presumably related with the above-mentioned goal. They are introduced in this article after description of the sensing environment and the collected corpus in Section 2. In Section 3, annotation of interest level is addressed through detection of laughters and non-lexical kinds of backchannels, referred to as reactive tokens. In Section 4 and 5, eye-gaze and nodding information is incorporated to predict when and who in the audience will ask questions, and also what kind of questions. With these analyses, we expect that we can get clues to high-level "understanding" of the conversations, for example, whether the presentation is understood or liked by the audience.

2 Multi-modal Corpus of Poster Conversations

2.1 Recording Environment

We have designed a special environment ("IMADE Room") to record audio, video, human motion, and eye-gaze information in poster conversations (T.Kawahara et al., 2008). An array of microphones (8 to 19) has been designed to be mounted on top of the posterboard, while each participant used a wireless head-set microphone for recording voice for the "gold-standard" corpus annotation. A set of cameras (6 or 8) has also been designed to cover all participants and the poster, while a motion capturing system was used for the "gold-standard" annotation. Each participant was equipped with a dozen of motion-capturing markers as well as an eye-tracking recorder and an accelerometer, but all devices are attached with a cap or stored in a compact belt bag, so they can be naturally engaged in the conversation. An outlook of session recording is given in Figure 3.

2.2 Corpus Collection and Annotation

We have recorded a number of poster conversations (31 in total) using this environment, but for some of them, failed to collect all sensor data accurately. In the analyses of the following sections, we use four poster sessions, in which the presenters and audiences are different from each other. They are all in Japanese, although we recently recorded sessions in English as well. In each session, one presenter (labeled as "A") prepared a poster on his/her own

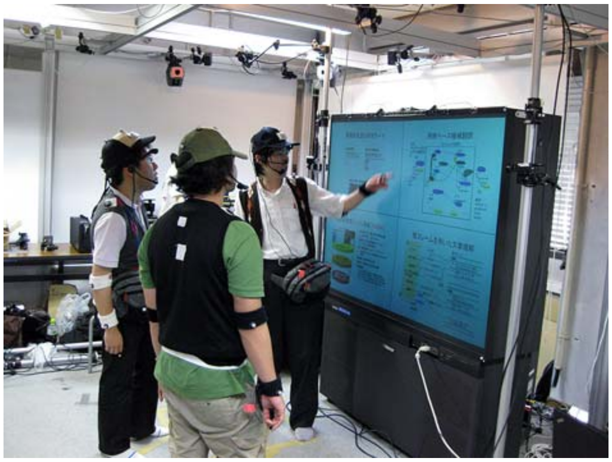

Figure 3: Outlook of poster session recording

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.

All speech data, collected via the head-set microphones, were segmented into IPUs (Inter-Pausal Unit) with time and speaker labels, and transcribed according to the guideline of the Corpus of Spontaneous Japanese (CSJ) (K.Maekawa, 2003). We also manually annotated fillers, verbal backchannels and laughters.

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. Noddings are automatically detected with the accelerometer attached with the cap.

3 Detection of Interesting Level with Reactive Tokens of Audience

We hypothesize that the audience signals their interest level with their feedback behaviors. Specifically, we focus on the audience's reactive tokens and laughters. By reactive tokens (*Aizuchi* in Japanese), we mean the listener's verbal short response, which expresses his/her state of the mind during the conversation. The prototypical lexical entries of backchannels include "*hai*" in Japanese and "yeah" or "okay" in English, but many of them are non-lexical and used only for reactive tokens, such as "*hu:n*", "*he:*" in Japanese and "wow", "uh-huh" in English. We focus on the latter kind of reactive tokens, which are not used for simple acknowledgment.

We also investigate detection of laughters and its relationship with interesting level. The detection method and performance were reported in (K.Sumi et al., 2009).

3.1 Relationship between Prosodic Patterns of Reactive Tokens and Interest Level

In this subsection, we hypothesize that the audience expresses their interest with specific syllabic and prosodic patterns. Generally, prosodic features play an important role in conveying para-linguistic and non-verbal information. In previous works (F.Yang et al., 2008; A.Gravano et al., 2007), it was reported that prosodic features are useful in identifying backchannels. Ward (N.Ward, 2004) made an analysis of pragmatic functions conveyed by the prosodic features in English non-lexical tokens.

In this study, we designed an experiment to identify the syllabic and prosodic patterns closely related with interest level. For this investigation, we select three syllabic patterns of "*hu:N*", "*he:*" and "*a:*", which are presumably related with interest level and also most frequently observed in the corpus, except lexical tokens.

We computed following prosodic features for each reactive token: duration, F0 (maximum and range) and power (maximum). The prosodic features are normalized for every person; for each feature, we compute the mean, and this mean is subtracted from the feature values.

For each syllabic kind of reactive token and for each prosodic feature, we picked up top-ten and bottom-ten samples, i.e. samples that have the largest/smallest values of the prosodic feature. For each of them, an audio segment was extracted to cover the reactive token and its preceding utterances. Then, we had five subjects to listen to the audio segments and evaluate the audience's state of the mind. We prepared twelve items to be evaluated in a scale of four ("strongly feel" to "do not feel"), among which two items are related to interest level and

Table 1: Significant combinations of syllabic and prosodic patterns of reactive tokens

		interest	surprise
hu:N	duration	*	*
	F0 max		
	F0 range		
	power		
he:	duration	*	*
	F0 max	*	*
	F0 range		*
	power	*	*
a:	duration		
	F0 max	*	
	F0 range		
	power	*	

other two items are related to surprise level [1]. Table 1 lists the results (marked by ”*”) that have a statistically significant ($p < 0.05$) difference between top-ten and bottom-ten samples. It is observed that prolonged “*hu:N*” means interest and surprise while “*a:*” with higher pitch or larger power means interest. On the other hand, “*he:*” can be emphasized in all prosodic features to express interest and surprise.

The tokens with larger power and/or a longer duration is apparently easier to detect than indistinct tokens, and they are more related with interest level. It is expected that this rather simple prosodic information is useful for indexing poster conversations.

3.2 Third-party Evaluation of Hot Spots

In this subsection, we define those segments which induced (or elicited) laughters or non-lexical reactive tokens as hot spots, [2] and investigate whether these hot spots are really funny or interesting to the third-party viewers of the poster session.

We had four subjects, who had not attended the presentation nor listened the recorded audio content. They were asked to listen to each of the segmented hot spots in the original time sequence, and to make evaluations on the questionnaire, as below.

Q1: Do you understand the reason why the reactive token/laughter occurred?

Q2: Do you find this segment interesting/funny?

Q3: Do you think this segment is necessary or useful for listening to the content?

The percentage of “yes” on Question 1 was 89% for laughters and 95% for reactive tokens, confirming that a large majority of the hot spots are appropriate.

The answers to Questions 2 and 3 are more subjective, but suggest the usefulness of the hot spots. It turned out that only a half of the spots associated with laughters are funny for the subjects (Q2), and they found 35% of the spots not funny. The result suggests that feeling funny largely depends on the person. And we should note that there are not many funny parts in poster sessions by nature.

On the other hand, more than 90% of the spots associated with reactive tokens are interesting (Q2), and useful or necessary (Q3) for the subjects. The result supports the effectiveness of the hot spots extracted based on the reaction of the audience.

4 Prediction of Turn-taking with Eye-gaze and Backchannel Information

Turn-taking is an elaborate process especially in multi-party conversations. Predicting whom the turn is yielded to or who will take the turn is significant for an intelligent conversational agent handling multiple partners (D.Bohus and E.Horvitz, 2009; S.Fujie et al., 2009) as well as an automated system to beamform microphones or zoom in cameras on the speakers. There are a number of previous studies on turn-taking behaviors in dialogue, but studies on computational modeling to predict turn-taking in multi-party interactions are very limited (K.Laskowski et al., 2011; K.Jokinen et al., 2011). Conversations in poster sessions are different from those in meetings and free conversations addressed in the previous works, in that presenters hold most of turns and thus the amount of utterances is very unbalanced. However, the segments of audiences’ questions and comments are more informative and should not be missed. Therefore, we focus on prediction of turn-taking by the audience in poster conversations, and, if that happens, which person in the audience will take the turn to speak.

[1] We used different Japanese wording for interest and for surprise to enhance the reliability of the evaluation; we adopt the result if the two matches.

[2] Wrede et al.(B.Wrede and E.Shriberg, 2003; D.Gatica-Perez et al., 2005) defined “hot spots” as the regions where two or more participants are highly involved in a meeting. Our definition is different from it.

Table 2: Duration (sec.) of eye-gaze and its relationship with turn-taking

	turn held by	turn taken by	
	presenter A	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

We also presume that turn-taking by the audience is related with their interest level because they want to know more and better when they are more attracted to the presentation.

It is widely-known that eye-gaze information plays a significant role in turn-taking (A.Kendon, 1967; B.Xiao et al., 2011; K.Jokinen et al., 2011; D.Bohus and E.Horvitz, 2009). The existence of posters, however, requires different modeling in poster conversations as the eye-gaze of the participants are focused on the posters in most of the time. This is true to other kinds of interactions using some materials such as maps and computers. Moreover, we investigate the use of backchannel information by the audience during the presenter's utterances.

4.1 Relationship between Eye-gaze and Turn-taking

We identify the object of the eye-gaze of all participants at the end of the presenter's utterances. The target object can be either the poster or other participants. Then, we measure the duration of the eye-gaze within the segment of 2.5 seconds before the end of the presenter's utterances because the majority of the IPUs are less than 2.5 seconds. It is listed in Table 2 in relation with the turn-taking events. We can see the presenter gazed at the person right before yielding the turn to him/her significantly longer than other cases. However, there is no significant difference in the duration of the eye-gaze by the audience according to the turn-taking events.

4.2 Relationship between Joint Eye-gaze Events and Turn-taking

Next, we define joint eye-gaze events by the presenter and the audience as shown in Table 3. In this table, we use notation of "audience", but actually these events are defined for each person in the audience. Thus, "Ii" means the mutual gaze by the presenter and a particular person in the audience, and "Pp" means the joint attention to the poster object.

Table 3: Definition of joint eye-gaze events by presenter and audience

who	presenter		
	gazes at	audience **(I)**	poster **(P)**
audience	presenter **(i)**	**Ii**	**Pi**
	poster **(p)**	**Ip**	**Pp**

Table 4: Statistics of joint eye-gaze events by presenter and audience in relation with turn-taking

	#turn held by presenter	#turn taken by audience		total
		(self)	(other)	
Ii	125	17	3	145
Ip	320	**71**	26	417
Pi	190	11	9	210
Pp	2974	147	145	3266

Statistics of these events at the end of the presenter's utterances are summarized in Table 4. Here, the counts of the events are summed over the two persons in the audience. They are classified according to the turn-taking events, and turn-taking by the audience is classified into two cases: the person involved in the eye-gaze event actually took the turn (self), and the other person took the turn (other). The mutual gaze ("Ii") is expected to be related with turn-taking, but its frequency is not so high. The frequency of "Pi" is not high, either. The most potentially useful event is "Ip", in which the presenter gazes at the person in the audience before giving the turn. This is consistent with the observation in the previous subsection.

4.3 Relationship between Backchannels and Turn-taking

As shown in Section 3, verbal backchannels suggest the listener's interest level. Nodding is regarded as a non-verbal backchannel, and it is more frequently observed in poster conversations than in simple spoken dialogues.

The occurrence frequencies of these events are counted within the segment of 2.5 seconds before the end of the presenter's utterances. They are shown in Figure 4 according to the joint eye-gaze

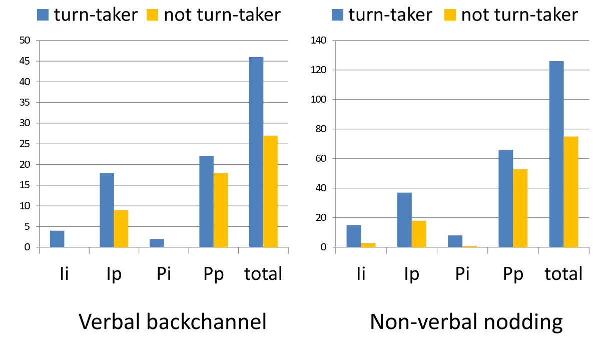

Figure 4: Statistics of backchannels and their relationship with turn-taking

Table 5: Prediction result of speaker change

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

events. It is observed that the person in the audience who takes the turn (=turn-taker) made more backchannels both in verbal and non-verbal manners, and the tendency is more apparent in the particular eye-gaze events of “Ii” and “Ip” which are closely related with the turn-taking events.

4.4 Prediction of Turn-taking by Audience

Based on the analyses in the previous subsections, we conduct an experiment to predict turn-taking by the audience. The prediction task is divided into two sub-tasks: detection of speaker change and identification of the next speaker. In the first sub-task, we predict whether the turn is given from the presenter to someone in the audience, and if that happens, then we predict who in the audience takes the turn in the second sub-task. Note that these predictions are done at every end-point of the presenter’s utterance (IPU) using the information prior to the speaker change or the utterance by the new speaker.

For the first sub-task of speaker change prediction, prosodic features are adopted as a baseline. Specifically, we compute F0 (mean, max, min, and range) and power (mean and max) of the presenter’s utterance prior to the prediction point. Backchannel features are defined by taking occurrence counts prior to the prediction point for each type (verbal backchannel and non-verbal nodding). Eye-gaze features are defined in terms of eye-gaze objects and joint eye-gaze events, as described in previous subsections, and are parameterized with occurrence counts and duration. These parameterizations, however, show no significant difference nor synergetic effect in terms of prediction performance.

SVM is adopted to predict whether speaker change happens or not by using these features. The result is summarized in Table 5. Here, we compute recall, precision and F-measure for speaker change, or turn-taking by the audience. This case accounts for only 11.9% and its prediction is very challenging, while we can easily get an accuracy of over 90% for prediction of turn-holding by the presenter. We are particularly concerned on the recall of speaker change, considering the nature of the task and application scenarios.

Among the individual features, the prosodic features obtain the best recall while the eye-gaze features achieve the best precision and F-measure. Combination of these two is effective in improving both recall and precision. On the other hand, the backchannel features get the lowest performance, and its combination with the other features is not effective, resulting in degradation of the performance.

Next, we conduct the second sub-task of speaker prediction. Predicting the next speaker in a multiparty conversation (before he/she actually speaks) is also challenging, and has not been addressed in the previous work (K.Jokinen et al., 2011). For this subtask, the prosodic features of the current speaker are not usable because it does not have information suggesting who the turn will be yielded to. Therefore, we adopt the backchannel features and eye-gaze features. Note that these features are computed for individual persons in the audience, instead of taking the maximum or selecting among them.

The result is summarized in Table 6. In this experiment, the backchannel features have some effect, and by combining them with the eye-gaze features, the accuracy reaches almost 70%.

Table 6: Prediction result of the next speaker

feature	accuracy
eye-gaze (gaze)	66.4%
backchannel (BC)	52.6%
gaze+BC	**69.7%**

5 Relationship between Feedback Behaviors and Question Type

Next, we investigate the relationship between feedback behaviors of the audience and the kind of questions they ask after they take a turn. In this work, questions are classified into 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”.[3] 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.

This annotation together with the preceding explanation segment is not so straightforward when the conversation got into the QA phase after the presenter went through an entire poster presentation. Thus, we exclude the QA phase and focus on the questions asked during the explanation phase. In this section, we analyze the behaviors during the explanation segment that precedes the question by merging all consecutive IPUs of the presenter. This is a reasonable assumption once turn-taking is predicted in the previous section. These are major differences from the analysis of the previous section.

5.1 Relationship between Backchannels and Question Type

The occurrence frequencies of verbal backchannels and non-verbal noddings, normalized by the duration of the explanation segment (seconds), are listed according to the question type in Tables 7 and 8. In these tables, statistics of the person who actually asked questions are compared with those of the person who did not. We can observe the turn-taker made significantly more verbal backchannels when asking substantive questions. On the other hand, there is no significant difference in the frequency of non-verbal noddings among the audience and among the question types.

Table 7: Frequencies (per sec.) of verbal backchannels and their relationship with question type

	confirming	substantive
turn-taker	0.034	**0.063**
non-turn-taker	0.041	0.038

Table 8: Frequencies (per sec.) of non-verbal noddings and their relationship with question type

	confirming	substantive
turn-taker	0.111	0.127
non-turn-taker	0.109	0.132

Table 9: Duration (ratio) of joint eye-gaze events and their relationship with question type

	confirming	substantive
Ii	0.053	0.015
Ip	**0.116**	0.081
Pi	0.060	0.035
Pp	0.657	**0.818**

5.2 Relationship between Eye-gaze Events and Question Type

We also investigate the relationship between eye-gaze events and the question type. Among several parameterizations introduced in the previous section, we observe a significant tendency in the duration of the joint eye-gaze events, which is normalized by the duration of the presenter’s explanation segment. It is summarized in Table 9. We can see the increase of “Ip” (and decrease of “Pp” accordingly) in confirming questions. By combining with the analysis in the previous section, we can reason the majority of turn-taking signaled by the presenter’s gazing is attributed to confirmation.

6 Smart Posterboard

We have designed and implemented a smart posterboard, which can record a poster session, sense human behaviors and annotate interactions. 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

[3] This does not mean the presenter actually answered simply by “Yes” or “No”.

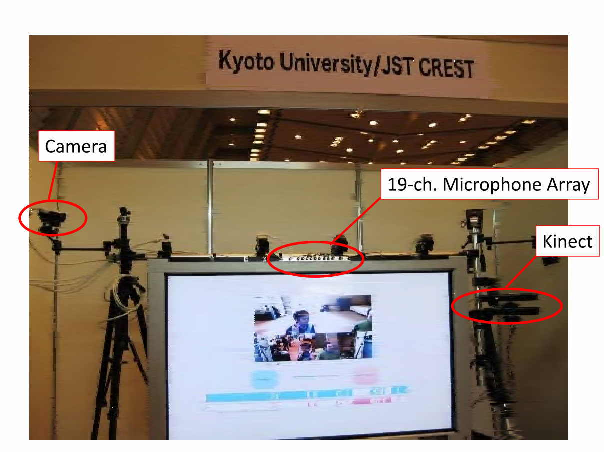

Figure 5: Outlook of smart posterboard

to the posterboard, which is actually a 65-inch LCD display. An outlook of the posterboard is given in Figure 5.

It is equipped with a 19-channel microphone array on the top, and attached with six cameras and two Kinect sensors. Speech separation and enhancement has been realized with Blind Spatial Subtraction Array (BSSA), which consists of the delay-and-sum (DS) beamformer and a noise estimator based on independent component analysis (ICA) (Y.Takahashi et al., 2009). In this step, the audio input is separated to the presenter and the audience, but discrimination among the audience is not done. Visual information should be combined to annotate persons in the audience. Voice activity detection (VAD) is conducted on each of the two channels to make speaker diarization. Localization of the persons in the audience and estimation of their head direction, which approximates their eye-gaze, are conducted using the video information captured by the six cameras.

Although high-level annotations addressed in the previous sections have not been yet implemented in the current system, the above-mentioned processing realizes a browser of poster sessions which visualizes the interaction.

The Kinect sensors are used for a portable and online version, in which speech enhancement, speaker localization and head direction estimation are performed in real time.

We made a demonstration of the system in IEEE-ICASSP 2012 as shown in Figure 5, and plan further improvements and trials in the future.

7 Conclusions

This article has given an overview of our multimodal data collection and analysis of poster conversations. Poster conversations provide us with a number of interesting topics in spoken dialogue research as they are essentially multi-modal and multi-party. By focusing on the audience's feedback behaviors and joint eye-gaze events, it is suggested that we can annotate interest level of the audience and hot spots in the session.

Nowadays, presentation using a poster is one of the common and important activities in academic and business communities. As large LCD displays become ubiquitous, its style will be more interactive. Accordingly, sensing and archiving functions introduced in the smart posterboard will be useful.

Acknowledgments

The work presented in this article was conducted jointly with Hisao Setoguchi, Zhi-Qiang Chang, Takanori Tsuchiya, Takuma Iwatate, and Katsuya Takanashi. The smart posterboard system has been developed by a number of researchers in Kyoto University and Nara Institute of Science and Technology (NAIST).

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

References

A.Gravano, S.Benus, J.Hirschberg, S.Mitchell, and I.Vovsha. 2007. Classification of discourse functions of affirmative words in spoken dialogue. In *Proc. INTERSPEECH*, pages 1613–1616.

A.Kendon. 1967. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63.

B.Wrede and E.Shriberg. 2003. Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *Proc. EUROSPEECH*, pages 2805–2808.

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

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

D.Bohus and E.Horvitz. 2009. Models for multiparty engagement in open-world dialog. In *Proc. SIGdial*.

D.Gatica-Perez, I.McCowan, D.Zhang, and S.Bengio. 2005. Detecting group interest-level in meetings. In *Proc. IEEE-ICASSP*, volume 1, pages 489–492.

F.Yang, G.Tur, and E.Shriberg. 2008. Exploiting dialog act tagging and prosodic information for action item identification. In *Proc. IEEE-ICASSP*, pages 4941–4944.

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

K.Laskowski, J.Edlund, and M.Heldner. 2011. A single-port non-parametric model of turn-taking in multi-party conversation. In *Proc. IEEE-ICASSP*, pages 5600–5603.

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

K.Otsuka, S.Araki, K.Ishizuka, M.Fujimoto, M.Heinrich, and J.Yamato. 2008. A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In *Proc. ICMI*, pages 257–262.

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

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

S.Fujie, Y.Matsuyama, H.Taniyama, and T.Kobayashi. 2009. Conversation robot participating in and activating a group communication. In *Proc. INTERSPEECH*, pages 264–267.

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

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

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