Prediction of Turn-Taking by Combining Prosodic and Eye-Gaze Information in Poster Conversations

Tatsuya Kawahara, Takuma Iwatate, Katsuya Takanashi

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

Abstract

We investigate turn-taking behaviors in conversations in poster sessions. While the poster presenter holds most of the turns during sessions, the audience's utterances are more important and should not be missed. In this paper, therefore, prediction of turn-taking by the audience is addressed. It is classified into two sub-tasks: prediction of speaker change and prediction of the next speaker. We made analysis on eye-gaze information and its relationship with turn-taking, introducing joint eye-gaze events by the presenter and audience. We also parameterize backchannel patterns of the audience. As a result of machine learning with these features, it is found that combination of prosodic features of the presenter and the joint eye-gaze features is effective for predicting speaker change, while eye-gaze duration and backchannels preceding the speaker change are useful for predicting the next speaker among the audience.

Index Terms: multi-party interaction, turn-taking, prosody, eye-gaze

1. Introduction

Turn-taking in conversations is a natural behavior in our human activities, but it is elaborate, especially in conversations with unfamiliar people or in formal settings. Studies on turn-taking have been conventionally focused on dyadic conversations between two persons. While there are a number of studies conducting analysis on the turn-taking patterns [1, 2, 3, 4], some studies investigated a prediction mechanism for a dialogue system to take or yield turns based on machine learning [5, 6, 7, 8]. Some studies even attempt to evaluate the synchrony of dialogues [9, 10].

Recently, conversational analysis and modeling have been extended to multi-party interactions such as meetings and free conversations by more than two persons. Turn-taking in multi-party interactions is more complicated than that in the dyadic dialogue case, in which a long pause suggests yielding turns to the (only one) partner. Predicting whom the turn is yielded to or who will take the turn is significant for an intelligent conversational agent handling multiple partners [11, 12] as well as an automated system to beamform microphones or zoom in cameras on the speakers. Studies on computational modeling on turn-taking in multi-party interactions are very limited so far. Laskowski et al. [13] presented a stochastic turn-taking model based on N-gram for the ICSI meeting corpus. Jokinen et al. [14] investigated the use of eye-gaze information for predicting turn-holding or giving in three-party conversations.

In this study, we deal with turn-taking behaviors in poster sessions, which are commonly done in academic conventions including InterSpeech conferences. Conversations in poster sessions are different from those in meetings and free conversations addressed in the previous works mentioned above, 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, and thus prediction of such events is important in online applications such as automated recording control and a conversational agent. Therefore, the goal of this work is to predict turn-taking by the audience in poster conversations, and, if that happens, which person in the audience will take the turn to speak.

We approach this problem by combining multi-modal information sources. While most of the aforementioned previous studies focused on prosodic features of the current speakers, it is widely-known that eye-gaze information plays a significant role in turn-taking [15], and the works by Jokinen [14] and by Bohus [11] exploited that information in their modeling. 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. We investigate several kinds of parameterization of eye-gaze patterns including the poster object, and explore effective features related with turn-taking. Moreover, we investigate the use of backchannel information such as nodding and verbal reactions by the audience during the presenter's utterances.

In this paper, we first describe the corpus of poster sessions and its annotations in Section 2. Then, we present an analysis on individual features of eye-gaze and backchannel information in Section 3. Prediction results based on machine learning with these features are presented in Section 4. Section 5 gives discussions and conclusions.

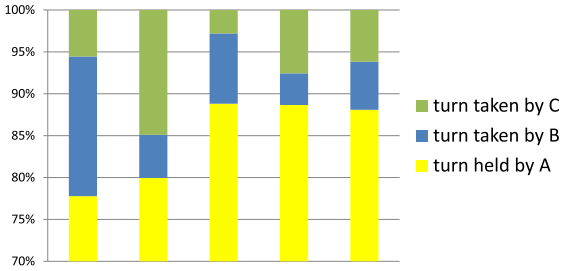
2. Multi-modal Corpus of Poster Conversations

We have recorded a number of poster sessions designed for multi-modal data collection [16]. In this study, we use four poster sessions, in which the presenters and audiences are different from each other. They are all in Japanese. In each session, one presenter (labeled as "A") had prepared a poster on his/her own academic research, and there was an audience of two persons (labeled as "B" and "C"; "B" standing closer to "A"), 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 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). We also manually annotated fillers and verbal backchannels.

Table 1: Statistics of turn-taking by audience (occurrence frequency)

	#turn held by presenter A	#turn taken by audience B	C	total
session1	845	44	50	94
session2	419	37	12	49
session3	356	17	39	56
session4	422	35	42	77
total	2042	133	143	276



Who gazes	Presenter A		B	C	Overall average
at Who	B	C	A	A	

Figure 1: Statistics of eye-gaze and its relationship with turn-taking (ratio)

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

	turn held by presenter A	turn taken by audience 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

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

who	presenter		
audience	gazes at	audience **(I)**	poster **(P)**
	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 (occurrence frequency)

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

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 an accelerometer attached with a cap. Noddings are detected with the accelerometer. 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.

The statistics on turn-taking are summarized in Table 1. In majority of utterances (IPUs) of the presenter (A), the turn was held by himself/herself. The ratio of turn-taking by the audience (either B or C) is only 11.9%. In this work, therefore, prediction of turn-taking is formulated as a detection problem rather than a classification problem. The evaluation measure should be recall and precision of turn-taking by the audience, not the classification accuracy of turn-holding and yielding by the presenter. This is consistent with the goal of the study described in Section 1.

3. Analysis on Eye-Gaze and Backchannel Features in Turn-Taking

First, we investigate statistics of eye-gaze and backchannel events and their relationship with turn-taking by the audience.

3.1. Distribution of Eye-Gaze

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. The statistics are shown in Figure 1 in relation with the turn-taking events. It is observed that the presenter was more likely to gaze at the person in the audience right before yielding the turn to him/her. We can also see that the person who takes the turn was more likely to gaze at the presenter, but the ratio of the turn-yielding by the presenter is not higher than the average over the entire data set.

We also measure the duration of the eye-gaze. It is measured 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.

3.2. Joint Eye-Gaze Events

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.

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.

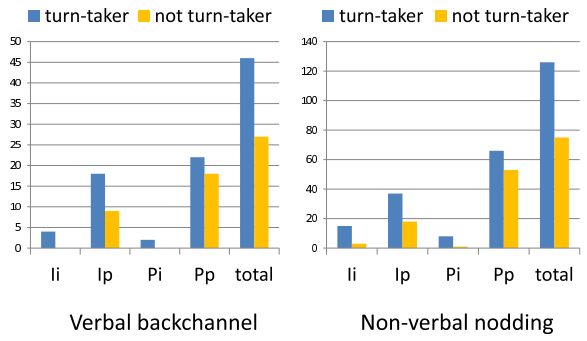


Figure 2: Statistics of backchannels and their relationship with turn-taking (occurrence frequency)

3.3. Dynamics of Eye-Gaze

In the analysis of the previous subsections, gazing information by the audience is not so clearly related with turn-taking. We hypothesize that the audience might have sent a signal to the presenter by gazing that he would like to take a turn, but turn-taking actually happens when the presenter looks back to him/her. In order to confirm this, we investigate the dynamic patterns of the eye-gaze events by a window of 2.5 seconds over 10 seconds before the end of the presenter's utterances. As a result, we observed a tendency that the frequency and duration of "Ii" and "Ip" are increasing toward the end of the utterances, while "Pi" appeared relatively longer in the segment of 5 seconds before the end of the utterances. This suggests that "Pi" is followed by "Ii" or "Ip". Then, we count bigram of the joint eye-gaze events, but the number of counts, except those with "Pp", are not large enough to derive any meaningful conclusion.

3.4. Backchannels

Verbal backchannels, typically "*hai*" in Japanese and "yeah" or "okay" in English, indicate that the listener is understanding what is being said. They also suggest the listener's interest-level [17, 18] and activate interaction. 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 2 according to the joint eye-gaze 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. Prediction of Turn-Taking by Audience

Based on the analysis in the previous section, we parameterize features for predicting 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 yielded 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.

Prediction experiments were conducted based on machine learning using the data set described in Section 2 in a cross-validation manner; one session is tested using the classifier trained with the other three sessions, and this process is repeated four times by changing the training and testing set.

4.1. Prediction of Speaker Change

For the first sub-task, prosodic features are adopted as a baseline based on the previous works (e.g. [14, 8]). Specifically, we compute F0 (mean, max, min, and range) and power (mean and max) of the presenter's utterance prior to the prediction point. Each feature is normalized by the speaker by taking the z-score; it is subtracted by the mean and then divided by the variance for the corresponding speaker.

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 as below:

1. Eye-gaze object

   For the presenter, (P) poster or (I) audience;

   For (anybody in) the audience, (p) poster, (i) presenter, or (o) other person in the audience.

2. Joint eye-gaze event: "Ii", "Ip", "Pi", "Pp"

   These can happen simultaneously for multiple persons in the audience, but we choose only one by the priority order listed above.

3. Duration of the above 1. ((I) and (i))

   A maximum is taken over persons in the audience.

4. Duration of the above 2. (except "Pp")

Note that these parameters can be extended to any number of the persons in the audience, although only two persons were present in this data set.

We tried support vector machines (SVM) and logistic regression (MaxEnt) model for machine learning, but they show comparable performance. The result with SVM is listed in Table 5. Here, we compute recall, precision and F-measure for speaker change, or turn-taking by the audience. As mentioned in Section 2, this case accounts for only 11.9% and its prediction is a very challenging task, 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 addressed in Section 1.

Among the individual features, as shown in Table 5, the prosodic features obtain the best recall while the eye-gaze features achieve the best precision and F-measure. In the table, combination of all four kinds of the eye-gaze parameterization listed above is adopted, however, using one of them is sufficient and there is not a significant difference in performance among them. Combination of the prosodic features and eye-gaze features 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.

4.2. Prediction of Next Speaker

Predicting the next speaker in a multi-party conversation (before he/she actually speaks) is also a challenging task, and has

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

Table 6: Prediction result of the next speaker

	feature	accuracy
1.	eye-gaze object	53.8%
2.	joint eye-gaze event	53.8%
	1.+2.	55.8%
3.	1.+2. + duration	66.4%
BC	backchannel	52.6%
	combination of above all (3.+BC)	**69.7%**

not been addressed in the previous work [14]. 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 which are described in the previous sub-section, but the features are computed for individual persons in the audience, instead of taking the maximum or selecting among them.

In this experiment, SVM performs slightly better than logistic regression model, thus the accuracies obtained with SVM are listed in Table 6. As there are only two persons in the audience, random selection would give an accuracy of 50%.

The simple eye-gaze features focused on the prediction point (1. and 2.) obtains an accuracy slightly better than the chance rate, but incorporating duration information (3.) significantly improves the accuracy. In this experiment, the backchannel features have some effect; as shown in Section 3.4, the person who made more backchannels is more likely to take the turn. By combining all features, the accuracy reaches almost 70%.

5. Conclusions

We have investigated para-linguistic and non-verbal patterns observed prior to turn-taking events in multi-party interactions in poster sessions, and conducted prediction experiments using these features. For prediction of speaker change or turn-taking by the audience, both prosodic features of the presenter and eye-gaze features of all participants are useful. The most relevant among the eye-gaze information is the presenter's gazing at the speaker to whom the turn is to be yielded. This is presumably affected by the characteristics of the poster session in which the presenter takes a major role in the conversation. For prediction of the next speaker, on the other hand, backchannel information by the audience is also useful as well as the eye-gaze information.

Based on the findings, we plan to design a smart posterboard which can control cameras and a microphone array to record the sessions and annotate the audience's reaction, which is critically important in poster conversations [18]. These findings will also be useful for an intelligent conversational agent that makes an autonomous presentation.

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

6. References

[1] Tomoko Ohsuga, Masafumi Nishida, Yasuo Horiuchi, and Akira Ichikawa. Investigation of the Relationship Between Turn-taking and Prosodic Features in Spontaneous Dialogue. In *Proc. INTERSPEECH*, pages 33–36, 2005.

[2] Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Analysis of Prosodic and Linguistic Cues of Phrase Finals for Turn-Taking and Dialog Acts. In *Proc. INTERSPEECH*, pages 2006–2009, 2006.

[3] Nigel G. Ward and Yaffa Al Bayyari. A Case Study in the Identification of Prosodic Cues to Turn-Taking: Back-Channeling in Arabic. In *Proc. INTERSPEECH*, pages 2018–2021, 2006.

[4] Bo Xiao, Viktor Rozgic, Athanasios Katsamanis, Brian R. Baucom, Panayiotis G. Georgiou, and Shrikanth Narayanan. Acoustic and Visual Cues of Turn-Taking Dynamics in Dyadic Interactions. In *Proc. INTERSPEECH*, pages 2441–2444, 2011.

[5] Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, and Kiyoaki Aikawa. Learning Decision Trees to Determine Turn-Taking by Spoken Dialogue Systems. In *Proc. ICSLP*, pages 861–864, 2002.

[6] David Schlangen. From Reaction to Prediction: Experiments with Computational Models of Turn-Taking. In *Proc. INTERSPEECH*, pages 2010–2013, 2006.

[7] A. Raux and M. Eskenazi. A Finite-State Turn-Taking Model for Spoken Dialog Systems. In *Proc. HLT/NAACL*, 2009.

[8] Nigel G. Ward, Olac Fuentes, and Alejandro Vega. Dialog Prediction for a General Model of Turn-Taking. In *Proc. INTERSPEECH*, pages 2662–2665, 2010.

[9] Stefan Benus. Are We 'in Sync': Turn-Taking in Collaborative Dialogues. In *Proc. INTERSPEECH*, pages 2167–2170, 2009.

[10] Nick Campbell and Stefan Scherer. Comparing Measures of Synchrony and Alignment in Dialogue Speech Timing with Respect to Turn-Taking Activity. In *Proc. INTERSPEECH*, pages 2546–2549, 2010.

[11] D. Bohus and E. Horvitz. Models for Multiparty Engagement in Open-World Dialog. In *Proc. SIGdial*, 2009.

[12] Shinya Fujie, Yoichi Matsuyama, Hikaru Taniyama, and Tetsunori Kobayashi. Conversation Robot Participating in and Activating a Group Communication. In *Proc. INTERSPEECH*, pages 264–267, 2009.

[13] Kornel Laskowski, Jens Edlund, and Mattias Heldner. A single-port non-parametric model of turn-taking in multiparty conversation. In *Proc. ICASSP*, pages 5600–5603, 2011.

[14] 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.

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

[16] 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.

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

[18] 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.