FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction
14-15 July 2018, Stockholm, Sweden

Proposal of a Multimodal Framework for Generating Robot's Spontaneous Attention Directions and Nods in Group Discussion

***Hung-Hsuan Huang[1 2 3], Seiya Kimura[3], Kazuhiro Kuwabara[3], Toyoaki Nishida[2 1]**
[1] RIKEN Center for Advanced Intelligence Project, Japan
[2] Graduate School of Informatics, Kyoto University, Japan
[3] Graduate School of Information Science and Engineering, Ritsumeikan University, Japan
*Contact: hhhuang@acm.org

Abstract

Our ongoing project is aiming to build a robot that can participate group discussion, so that its users can repeatedly practice group discussion with it. In this paper, we propose a multimodal framework to incorporate the modules to generate spontaneous head movements, shifts of attention focus and nodding of the robot. The generation models are derived from human-human group discussion data corpus with support vector classifiers. Dedicated models are developed according to conversation situations: when the robot is speaking, when the robot is listening to other participants, and when no participant is speaking. Low-level verbal and non-verbal (speech turn, prosody, face direction, and head activities) features extracted from the participants other than the focused one (the robot) are adopted in the learning process.

1 Introduction

There has been a growing number of companies that adopted group discussion in the recruitment of their employers, and often communication skill is treated even more important than professional skills. During the discussion process, the perception of higher communication skills may lead to the applicant's success in job hunting. Repeated practice is considered to improve the communication skill of job applicants. However, it requires good partners for the practices, which can be difficulty for many students. In our ongoing project, we aim to develop a training environment that allows the trainees to practice group discussion with communicational robot(s). In the domain of human-robot interaction, there are works [Vazquez *et al.*, 2017] about the interaction with multiple users, however, most of them are treating the robot as a different role (asymmetric relationship) than the human users. In our work, we focus on how to make a robot to join the discussion just like the other human (or robot) participants, and therefore dedicated behavior models are required.

In order to build a realistic environment for effective practice, the robot has to perform believable and comprehensive behaviors. Unlike dyadic conversation where there is only one communication partner for the robot, there are more than one partners in a group discussion conversation. In this context, the situated determination of the participant whom the robot should pay attention to can be considered essential in realizing robot's believeness. Due to the limited degree of freedom comparing to a human, a robot may not be able to perform subtle gaze behaviors. In the extreme case, the robot may not have movable eye balls. Therefore, we treat the attention focus as the overall combination of head direction and gaze direction of the eye balls of the robot. The implementation of the proposed attention model can be utilized by available physical parts of the robot. Nodding, is another spontaneous and intentional behavior to convey many signals in conversation. Multiparty situations leverage the technical difficulty of the implementation of communicational robot. However, in the same time, the phenomenons like copying and synchrony [Jokinen and Parkson, 2012] between conversation partners may provide the cues for robot behavior generation. Robot's behaviors can be considered to include intentional and spontaneous ones. We suppose that intentional behaviors are more determined by the behaviors themselves, while the spontaneous ones are supposed to be more stimulated by the environment.

Multimodal features have been shown effective to interpret and predict the behaviors and intentions like who is talking to whom from the features extracted from the actors themselves [Huang *et al.*, 2011; Otsuka, 2011]. In this paper, we propose the framework of robot's head movement generation with the prediction models on attention focuses and nodding which are driven by multimodal cues extracted from the conversational partners rather than the actor himself / herself.

2 Multimodal Framework

In order to generate believable behaviors, the most intuitive source of ground truth is human-human group discussion. This work is based on a data corpus of 40 college students (four people x 10 groups x 15 minutes, 10 hours in total) following the same experimental procedure of the MATRICS corpus [Nihei *et al.*, 2014]. The proposed multimodal framework is shown in Figure 1. From the point of view of a robot, it engages group discussion task with three other participants (human or other robots) and acquires the activities of other participants from video / audio and other sensory information. Multimodal features are extracted from these pieces of information by preprocessing modules, Face Recog-

10.21437/AI-MHRI.2018-4

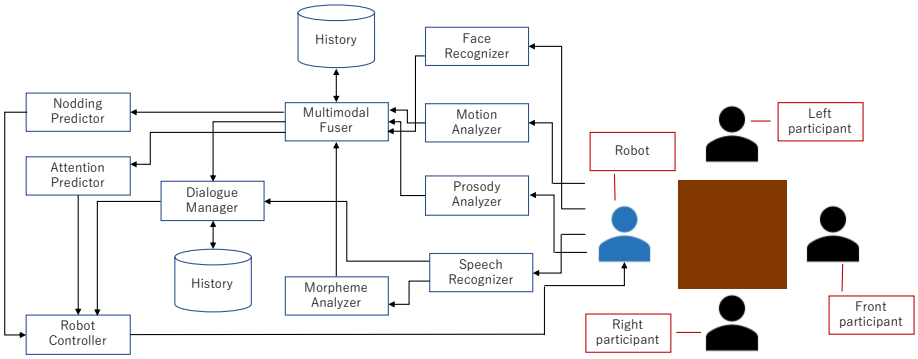

Figure 1: Proposed framework with the attention / nodding prediction modules integrated

nizer, Motion Analyzer (accelerometers attached on the participants' heads), Prosody Analyzer, Speech Recognizer, and Japanese Morpheme Analyzer. All available feature values are then integrated by the Multimodal Fuser module. It identify the correspondence of the information coming from different sources with timestamps and generates feature vectors of input information at each prediction time point in realtime. For the features which past information is referred, data history is kept by this module. The multimodal inputs are propagated to Dialogue Manager (DM) module which decides the robot's utterances as well as other intentional behaviors of the robot. Multimodal input information are also sent to the prediction modules of nodding and attention (NP and AP), respectively. These two modules determine the timings of the robot's spontaneous head turns (changes of attention focuses) and nodding. The outputs from DM, NP, and AP modules are then gathered in the controller (RC) module which physically controls the robot. The RC module selects the actual actions to perform and resolve the contradictions when there are more than one modules trying to move the same parts of the robot. A possible policy for the resolve is granting higher priority on intentional actions from the DM module.

3 Attention Model

The attention model is an extension to our previous work [Kimura *et al.*, 2017] where the model was first presented. The performance of the prediction model was further improved since then with the integration of linguistic features and the refinement of other features. This model predicts the attention focuses of the robot to four classes, the participants who sit at *Left*, *Front*, *Right* sides of the robot and *Table* where the distributed materials were placed on. Since the available information is different regarding to the situation of conversation, dedicated prediction models are built in the following three situations: *Speaking*: when the robot is speaking. *Listening*: when a participant other than the robot is speaking. *Idling*: when no one is speaking.

The multimodal low-level features selected are the ones supposed to be able to be extracted directly from the behaviors of the participants other than the robot itself at temporal granularity, 0.1 second. This resulted to 84,492 data instances for speaking model, 173,596 data instances for listening model, and 101,652 data instances for idling model. As a result, the following 122 features are selected:

- (V)erbal features. We used the Japanese morpheme segmentation tool, Kuromoji [1] to analyze the words of utterance transcriptions and count the numbers of verbs, nouns, new nouns, existing nouns, interjections, and fillers in utterances of the participants in last five seconds (18 features).
- (A)ttention: the features related to the attention focus where the other three participants are paying to. Current attention focus, time ratio in paying attention to the focused participant, frequency of changing attention focuses and so on (15 features).
- (S)peech turn: the features related to speech turns. The speaking periods are identified with the phonetic analysis tool, Praat [2]. Number of utterances, ratio of speaking, last speaker, length of utterances, and so on (23 features).
- (P)rosody: the prosodic information while this participant is speaking. Praat was also used to compute the prosodic features of the utterances of the other participants. The distribution of pitch ($F0$) and intensity are taken into account (36 features).
- (H)ead activity: activity of head movements of this participant, which is measured with the three-axis accelerometer attached on the head of each participant. The amount of head movements and nods in past five seconds and so on are taken into account (30 features).

A support vector machine (SVM) with gaussian (RBF) kernel was used to develop the prediction models for the three situations. SVM complexity parameter C was explored among the values: 1 to 10. RBF kernel parameter γ was explored among the values: $10^{-}3$ to 1. All combinations of the parameters were tested, and the best results were found with the setting where $C = 1.0$ and $\gamma = 0.01$. Because the effective duration can be modality dependent, the features are computed with sliding window in multiple lengths depending on the modalities. All combination of window size t varies

[1] https://www.atilika.com/ja/products/kuromoji.html
[2] http://www.fon.hum.uva.nl/praat/

from 1 to 10 seconds were explored to find the optimal length for each modality. Due to the bias in the number of instances in table and front classes, We oversampled smaller classes with Synthetic Minority Oversampling TEchnique (SMOTE) algorithm and under-sampled the larger classes while keeping the total weight (amount) of the dataset both in training and testing phases. Leave-one-participant-out cross validation was used in evaluating the performance of the models.

Table 1 shows the results regarding to each attention focus. For all three models, they were better in prediction the attention focuses at side directions (*left* and *right*). This may be because the participants most often look forward or the material on the table, as a result, these two directions were less characteristic than the side ones. It can also be observed that the performance of the three models are always in the order: listening >idling >speaking. This shows that attention is more affected by the other participants when the focused participant is not speaking. The possible reason is, when the participant were speaking, they pay less attention to the others, the attention focuses were more random or more dependent on the contents of the speaker's utterances. On the other hand, the reason why listening model always outperforms better than idling model may come from the fact that more available information are used. All the attention focus class are often as *Table*, especially *Front* class' recall is low. This cause it more difficult to distinguish the other attention focuses from *Table*. The overall tendency to *Table* class may due to the fact that this class has largest number of instances and consequently covered larger variety of data.

Table 1: Classification results of the proposed attention focus prediction models with all available feature sets

	Attention	Precision	Recall	F-measure
Speaking	Table	0.423	0.515	0.464
	Front	0.411	0.393	0.402
	Right	0.501	0.425	0.460
	Left	0.541	0.522	0.532
	Overall	**0.464**	**0.460**	**0.460**
Listening	Table	0.412	0.396	0.404
	Front	0.622	0.605	0.613
	Right	0.622	0.672	0.646
	Left	0.664	0.655	0.659
	Overall	**0.580**	**0.581**	**0.580**
Idling	Table	0.471	0.610	0.532
	Front	0.556	0.514	0.534
	Right	0.452	0.385	0.416
	Left	0.655	0.570	0.610
	Overall	**0.536**	**0.529**	**0.528**

Figure 2 depicts the overall F-measure values regarding to different combinations of feature sets. From the results, we found that not always but usually richer information has better results. For all the tree models, feature set A and P contribute most to the classification performance while feature set H, V, and S were not so effective. This implies that mutual attention and speech turn taking play most important roles in determining appropriate attention focus for the robot.

4 Nodding Model

The second prediction model is the one that determines whether the robot should nod or not. Follow the same procedure as the attention model, the periods when the participants were nodding are manually labeled (Table 2). Since nodding behavior is only relevant to utterances, rather than the 0.1-second time slices of the attention model, the prediction time points of nodding model is set to be the end of each utterance. That is, when anyone of the participants finished his / her utterance, this instant is treated at the prediction points of all of the four participants. Table 2 also shows the number of resulted data instances for the learning process. And for the same reason, only the models for speaking and listening situations were developed.

Similar to attention models, the five feature sets, verbal, attention, speech turn, prosody, and head activities are extracted from the same data corpus while the unavailable features were omitted. Table 3 shows the performance of the two-class model in the two situations, speaking and listening. Considering the chance level (50%) of a two-class classification problem, the performance was only moderate. The recall of *Yes* was exceptionally low. This implies that the reasons caused the participant to nod were more diverse and more difficult to predict from the behaviors of other participants. The performance of all combinations of feature sets is shown in Figure 2. Unlike attention model, it is found that verbal features were most effective in classifying nods. It can also be found that listening models usually perform than the other situations, this may imply that people's behaviors are more determined by the others when they are listening while it is not so predictable when human are speaking. In addition, the performance is usually better when there are more modalities available during listening situation but it is not necessary better in speaking situations.

Table 2: Annotation results on the nods of the participants of the data corpus. The columns, "avg", "max", and "min" shows the average, maximum, and minimum duration of individual labels in seconds, respectively. Speaking and Listening are the resulted instance numbers in corresponding models

Nodding	Labels	Avg	Max	Min	Speaking	Listening
Yes	621	1.3	9.9	0.1	324	667
No	661	53.3	609.6	0.3	5,213	15,197

Table 3: Yes / No classification results of the proposed nodding prediction models with all available feature sets

	Nod	Precision	Recall	F-measure
Speaking	Yes	0.715	0.487	0.579
	No	0.626	0.816	0.708
	Overall	**0.669**	**0.656**	**0.645**
Listening	Yes	0.718	0.617	0.663
	No	0.664	0.757	0.707
	Overall	**0.691**	**0.687**	**0.685**

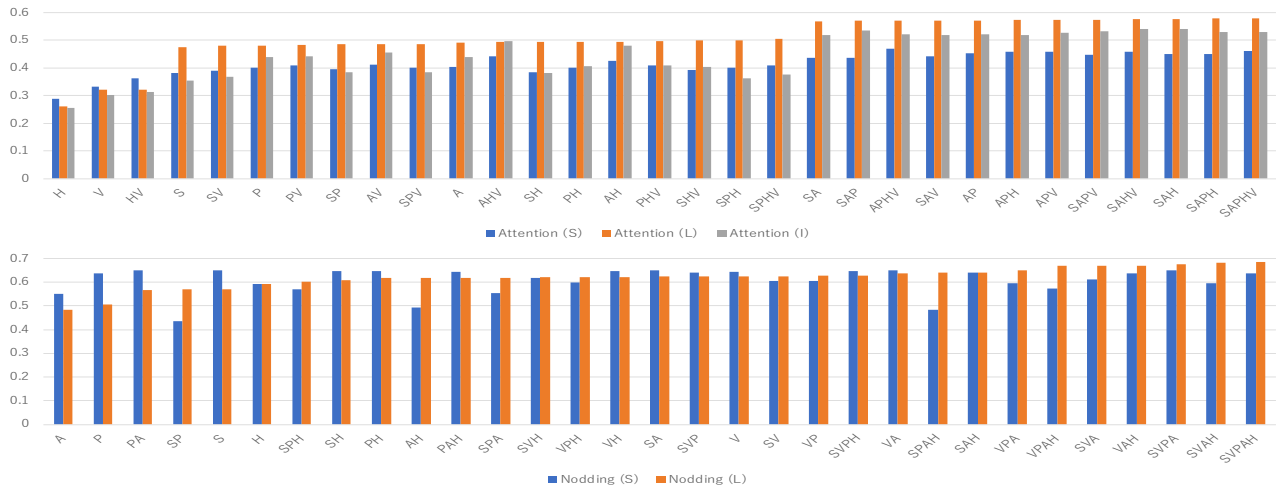

Figure 2: F-measure values (vertical axis) showing the performance of the attention focus prediction models in speaking, listening, and idling situations as well as nodding models in speaking and listening situations with all 31 combinations of feature sets. The bars are sorted in the order of listening model which has best performance

5 Conclusion and Future Direction

In order to drive the robot's head to be more life-like in group discussion sessions, we propose the models in directing the robot's attention toward the other participants in three situations: speaking, listening, and idling. The models for determining whether the robot should nod or not in the first two situations were also proposed. These models are derived from a data corpus containing 10 discussion sessions done by four-people groups. We then use low-level verbal / non-verbal features to train support vector machine models to for these two behaviors. Although the results were moderate, they showed there is tendency existing in the attention focuses of participants in group discussion.

For the future work, we would like to refine the features to further improve the performance of the models, incorporating more non-verbal information like postures and more detailed prosodic information like MFCC. Also, we would like to add more verbal features like the intention of utterances to improve the performance of speaking model. The relationship between the terms and who spoke them may be useful in this aspect. For example, when the focused participant is speaking a term which was previously spoken by another participant, he or she may pay attention to that participant more. In addition to the improvement of classification performance, we would like to implement the framework to a communicational robot or a virtual agent in a VR environment. The models then have to been tuned according to hardware / software implementation constraints (e.g. the rotation speed of the robot's head). It could be anticipated that the human participants may behave differently with agents than other humans, we will also investigate this aspect in the participant experiment using the implemented robot / agent.

References

[Huang *et al.*, 2011] Hung-Hsuan Huang, Naoya Baba, and Yukiko Nakano. Making virtual conversational agent aware of the addressee of users' utterances in multi-user conversation from nonverbal information. In *13th International Conference on Multimodal Interaction (ICMI'11)*, pages 401–408, 2011.

[Jokinen and Parkson, 2012] Kristiina Jokinen and Siiri Parkson. Synchrony and copying in conversational interactions. In *3rd Nordic Symposium on Multimodal Interaction*, pages 18–24, 2012.

[Kimura *et al.*, 2017] Seiya Kimura, Hung-Hsuan Huang, Qi Zhang, Shogo Okada, Naoki Ohta, and Kazuhiro Kuwabara. Proposal of a model to determine the attention target for an agent in group discussion with non-verbal features. In *5th International Conference on Human Agent Interaction (HAI 2017)*, pages 195–202, Bielefeld, October 2017.

[Nihei *et al.*, 2014] Fumio Nihei, Yukiko I. Nakano, Yuki Hayashi, Hung-Hsuan Huang, and Shogo Okada. Predicting influential statements in group discussions using speech and head motion information. In *16th International Conference on Multimodal Interaction (ICMI 2014)*, Istanbul, pages 136–143, November 2014.

[Otsuka, 2011] Kazuhiro Otsuka. Multimodal conversation scene analysis for understanding people ' s communicative behaviors in face-to-face meetings. In *Symposium on Human Interface 2011*, pages 171–179. Springer Berlin Heidelberg, 2011.

[Vazquez *et al.*, 2017] Marynel Vazquez, Elizabeth J. Carter, Braden McDorman, Jodi Forlizzi Aaron Steinfeld, and Scott E. Hudson. Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze. In *12th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2017)*, pages 42–52, Vienna, Austria, March 2017.