ANALYSIS AND MODELING OF NEXT SPEAKING START TIMING BASED ON GAZE BEHAVIOR IN MULTI-PARTY MEETINGS

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Junji Yamato

NTT Communication Science Laboratories, NTT Corporation.

ABSTRACT

To realize a conversational interface where an agent system can smoothly communicate with multiple persons, it is imperative to know how the start timing of speaking is decided. In this research, we demonstrate a relationship between gaze transition patterns and the start timing of next speaking against the end of the last speaking in multi-party meetings. Then, we construct a prediction model for the start timing using gaze transition patterns near the end of an utterance. An analysis of data collected from natural multi-party meetings reveals a strong relationship between gaze transition patterns of the speaker, next speaker, and listener and the start timing of the next speaker. On the basis of the results, we used gaze transition patterns of the speaker, next speaker, and listener and mutual gaze as variables, and devised several prediction models. A model using all features performed the best and was able to predict the start timing well.

Index Terms— Speaking timing, gaze transition pattern, prediction model, multi-party meetings, mutual gaze

1. INTRODUCTION

Face-to-face communication is one of the most basic forms of communication in daily life and group meetings are used for conveying information and making decisions. Smooth communication similar to face-to-face communication is desired in remote human-to-human and human-to-agent communication. Therefore, ways to automatically analyze multi-party meetings have been actively researched in recent years [1, 2].

Turn-taking, the situation where the speaker changes, is especially important. The participants need to predict the end of the speaker's utterance and good speaking timing. The timing of speaking varies with the situation and utterance content. It has been reported that giving back channel feedback and speaking at the right time is important for smooth and natural communication [3]. Bad timing of speaking has not only a negative effect on communication but also sends unintended messages to conversational partners. For example, only a short delay in video and audio of about 500 ms can inhibit smooth communication in remote video conferencing systems [4].

It is known that gaze behavior controls listener's response and contributes to realizing smooth turn-taking in two-person meetings. Kendon [5] reported that a speaker gazes at a listener as a "turn-yielding cue" to yield the turn to a listener at the end of an utterance. Then, the listener glances at the speaker (mutual gazing) to accept the cue and starts speaking, i.e., takes the turn. In the engineering field, several models for detecting end-of-utterance using voice information [6] and gaze behavior [7, 8, 9, 10] and the next speaker [11] in turn-taking in multi-party meetings have been proposed. These studies have found that gaze information is more effective than speech information.

In multi-party meetings, the relationship between the start timing of the next speaker's utterance and gaze behavior has not been clarified. As a relation between a listener's response and the speaker's gaze, it has been shown that the listener's response is delayed when the speaker doesn't gaze at the listener in two-person meetings [5]. In related work about detecting the timing of speaking, estimations of the timing of the listener's back channel [12] and of the timing at which turn-taking is possible have been reported [13] in two-person meetings. However, no research has tried construct a prediction model for the start timing of the next speaking against the end of the last speaking in multi-party meetings.

In this research, we focus on gaze behavior at the end of an utterance and reveal a relationship between gaze behavior and the start timing of next speaking in multi-party meetings. For the analysis, we divided participants of meetings into a speaker, next speaker, and listeners and define a gaze transition pattern (GTP) to include information about gaze shifts and mutual gaze. We analyze the effect of each GTP of a speaker, next speaker, and listener on the timing. Moreover, from the results of the analysis, we construct a prediction model for the start timing of the next speaking. Conducting the prediction model enable us to quantitatively evaluate how GTPs can clarify the mechanism that determines the timing of speaking. The prediction model will contribute to the design of an conversational agent that can speak with natural timing in multi human-to-agent communication.

2. CORPUS DATA

2.1. Definition of GTP

We introduce a method for generating GTPs. Since previous studies [5, 7, 14] have shown that gaze behavior at the end of an utterance is very deeply connected with turn-taking, we

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

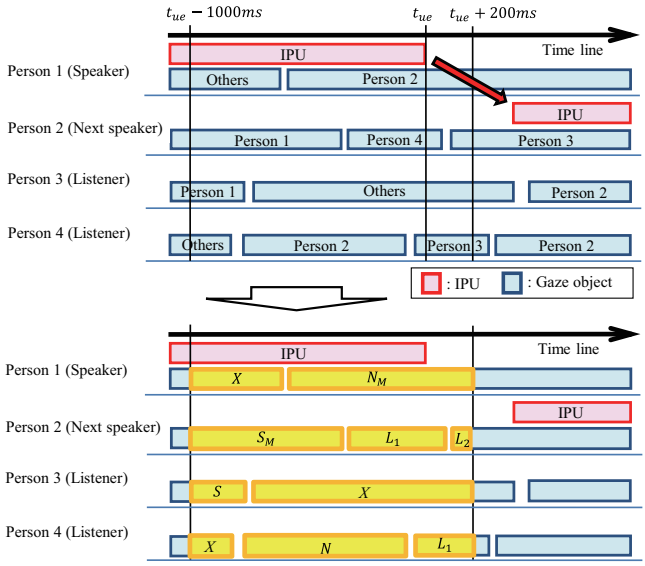

Fig. 1. Sample of GTP generation.

treat the interval between 1000 ms before the end of utterance and 200 ms after the end of utterance as the interval for analysis. In addition, Kendon [5] has demonstrated that the next speaker looks away when he/she starts to speak after having made eye contact with the current speaker at the end of the speaker's utterance in two-person meetings. Thus, it is assumed that these temporal transitions of participants' gaze behavior and mutual gaze are important for the next speaker's speaking timing. We therefore decided to focus on the mutual gaze and gaze transitions of the speaker, next speaker, and listeners, and to express them as an n-gram, which we defined as a sequence of gaze object shifts. For GTP generation, the candidate for a gaze is first classified as "speaker", "next speaker", "listener", or "others (the rest of the objects)", and labeled. We use the following gaze labels:

- S: Next speaker or listener looks at a speaker without mutual gaze (speaker doesn't look at him/her.).
- S_M: Next speaker or listener looks at a speaker with mutual gaze (speaker look at him/her.).
- L_1, L_2: Speaker, next speaker or listener looks at a listener who doesn't become the next speaker without mutual gaze. L_1 and L_2 indicate different listeners.
- L_{1M}, L_{2M}: Speaker, next speaker or listener looks at a listener who doesn't become the next speaker with mutual gaze. L_{1M} and L_{2M} indicate different listeners.
- N: Speaker or listener looks at the next speaker without mutual gaze.
- N_M: Speaker or listener looks at the next speaker with mutual gaze.
- X: Speaker, next speaker or listener looks at non-persons, such as the floor or ceiling.

As an example, Fig. 1 shows how a GTP is constructed: Person 1 finishes speaking and then person 2 starts to speak. Person 1 gazes at person 2 after he has gazed at others during the interval of analysis. When person 1 looks at person 2, person 2 looks at person 1; namely, there is mutual gaze. Therefore, person 1's GTP is X-N_M. Person 2 looks at persons 4 and 3 after making eye contact with person 1. Then, person 2's GTP is S_M-L_1-L_2. Person 3 looks at others after looking at person 1. Then, person 3's GTP is S-X. Person 4 looks at persons 2 and 3 after looking at others. Then, person 2's GTP is X-N-L_1.

2.2. Collected data in multi-party meetings

To collect a conversation corpus in multi-party meetings for the analysis of GTPs, we performed an experiment with a four-person meeting. The four participants were in their 20's and 30's, and this was the first time they had met. They faced each other and sat down. They argued and gave opinions in response to highly divisive questions like "Is marriage the same as love?" and needed to draw a conclusion within eight minutes. The corpus was created from eight recorded minutes each from the sound and video information (recorded at 30 Hz) on ten dialogs held by three groups of four different persons (12 people in total). Annotation data was created by a skilled annotator as follows:

- Gaze object: The gaze object was annotated using bust-up and overhead views from the videos with the ELAN tool [15]. The objects of gaze were the four participants (person 1, 2, 3, and 4) and others, i.e., the hall or ceiling. Conger's kappa coefficient κ [16] as an inter-coder agreement of three annotators is .887.
- Utterance: The inter-pausal units (IPUs) [17] were created after transcribing the utterances from the recorded speech. The portion of an utterance followed by more than 200 ms of silence was used as the unit of one IPU. From the created IPU, supportive responses [18] were excluded and an utterance unit continued by the same person was considered as one utterance turn. And pairs of IPUs that adjoined at the time of turn-taking were created. The numbers of the created groups of IPUs was 365. Because we analyze the GTP in the interval between 1000 ms before the end of utterance and 200 ms after the end of IPU. The data in which next speaker's IPU started before 200 ms was excluded from the 365 pairs. The remaining data was 89.0% of the whole, i.e., 325 pairs.

3. ANALYSIS

To investigate the correlation between GTPs and the speech timing of next speaker, we define the timing interval T_{int} between end time of speaker's IPU (t_{ue}) and start time of next speaker's IPU (t_{nus}) (see Fig. 2). We analyze the T_{int} by each GTP of the speaker, next speaker, and listener.

3.1. Analysis of speaker's GTPs

Boxplots of interval T_{int} by each speaker's GTP using 325 data are shown in Fig. 3. A center line in a box shows the

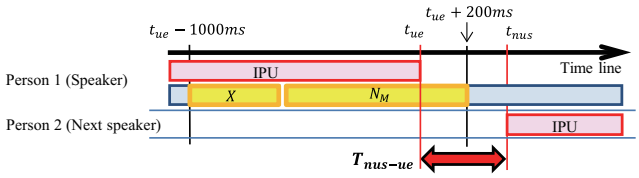

Fig. 2. Interval T_{int} between end time of speaker's IPU (t_{ue}) and start time of next speaker's IPU (t_{nus}).

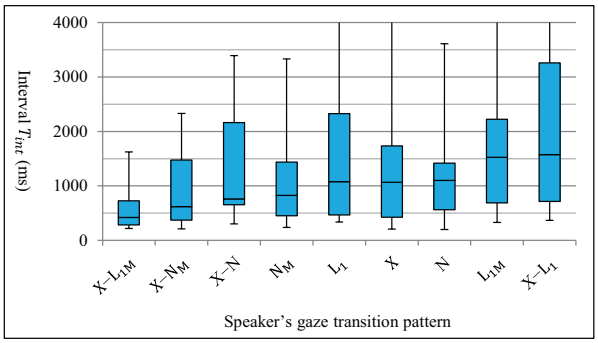

Fig. 3. Relationship between speaker's GTP and interval T_{int}.

median value. Patterns that occurred in less than 3% of the data were excluded because the number of data is small. As shown in Fig. 3, the median value and the range of the box for T_{int} differ depending on the types of pattern. X-L_{1M} has the shortest median value, 420 ms. In contrast, X-L_1 has the longest median value, 1638.5 ms. Both patterns mean that the speaker starts to look at a listener. The difference between the patterns lies in whether mutual gaze has occurred or not. This result indicates that mutual gaze is important for determining the start timing of speaking. The pattern that has the second shortest T_{int} is X-N_M. That is, when a speaker starts to make eye contact with the next speaker, the start timing of next speaker's speaking is early. In previous work [5], it has been reported that a speaker makes eye contact with a listener when smooth turn-taking occurs in two-person conversation. Namely, when these gaze behaviors occur, the next speaker starts to speak quickly.

3.2. Analysis of next speaker's GTPs

Boxplots of interval T_{int} by each next speaker's GTP using 325 data are shown in Fig. 4. Again, patterns that occurred in less than 3% of the data were excluded. Here, S and S_M, which means the next speaker continues to look at the speaker, have the shortest median values, 623 and 701.5 ms. In contrast, L_{1M}, which means the next speaker continues to look at the listener with mutual gaze, has the longest median value, 1624 ms. That is, when the next speaker continues to look at the speaker, the start timing of the next speaker's speaking is early. When the previously reported gaze behaviors mentioned above [5] occur, the next speaker starts to speak quickly. In contrast, when the next speaker doesn't look at

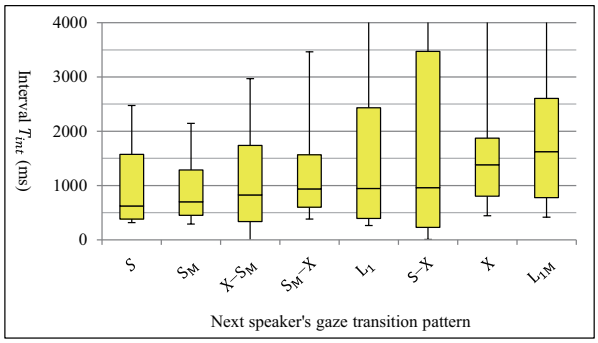

Fig. 4. Relationship between next speaker's GTP and interval T_{int}.

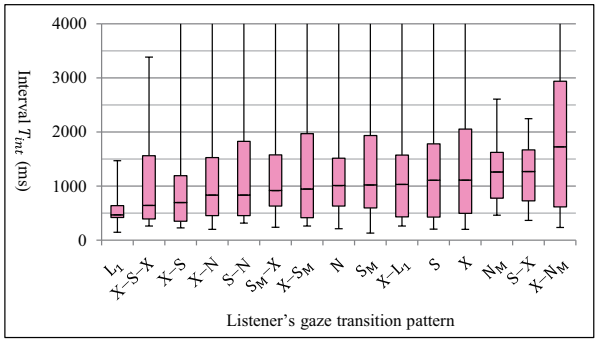

Fig. 5. Relationship between listener's GTP and interval T_{int}.

the speaker, turn-taking is not smooth and the timing of the next speaking becomes late.

3.3. Analysis of listener's GTPs

Boxplots of interval T_{int} by each listener's GTP using 650 data[1] are shown in Fig. 5. Again, patterns that occurred in less than 3% of the data were excluded. Here, L_1, which means a listener continues to look at another listener, has the shortest median value, 642 ms. In contrast, X-N_M, which means a listener starts to make eye contact with the next speaker, has the longest median value, 1726 ms. That is, when a listener starts to make eye contact with the next speaker, the start timing of the next speaker's speaking is late; when the next speaker continues to look at the speaker, the start timing of next speaker's speaking is early. This is because the next speaker looks at a listener without looking at a speaker.

Therefore, these results suggest that GTPs of the speaker, next speaker, and listener influence the start timing of speaking in turn-taking situations.

4. PREDICTING NEXT-SPEAKING TIMING

From analyses in the previous sections, we found that GTPs of the speaker, next speaker, and listener may be useful as predictors of start timing of the next speaker in multi-party

[1]There are two listeners in addition to a speaker and next speaker.

Table 1. Results of evaluation of next speaker's timing prediction model.

	Under 1000 ms	Between 1000 ms and 2000 ms	Between 2000 ms and 3000 ms	Between 3000 ms and 4000 ms	Over 4000ms	All
Baseline	1027.5 ms	353.1 ms	868.7 ms	1776.7 ms	5378.7 ms	1089.3 ms
SG	443.4 ms	361.4 ms	1094.6 ms	2072.3 ms	5656.6 ms	900.3 ms
NG	414.0 ms	394.6 ms	1163.1 ms	2080.1 ms	5626.1 ms	896.3 ms
LG	497.3 ms	394.8 ms	1323.6 ms	2298.2ms	5888.1 ms	988.7 ms
SG+NG+LG	296.9 ms	398.7 ms	976.7 ms	2011.5 ms	5471.8 ms	797.9 ms
SG+NG+LG+MG	335.7 ms	315.0 ms	793.4 ms	1883.1 ms	5101.2 ms	755.4 ms

meetings. In this section, we predict the start timing by employing SMOreg [19, 20], which implements a support vector machine for regression in Weka [21], and evaluate the accuracy of the model and the effectiveness of each feature.

The data used in SMOreg contains the start timing of next speaker as a class, and the GTPs as features. We tested the following prediction models that use the GTPs above:

- Baseline: The model outputs the average value of the interval T_{int} that is 1642.3 ms.
- SG: Uses speaker's GTP without mutual gaze. The L_{1M}, L_{2M}, and N_M labels of GTP are integrated into L_1, L_2, and N labels.
- NG: Uses next speaker's GTPs without mutual gaze. The S_M, L_{1M}, and L_{2M} labels of GTPs are integrated into S, L_1, and L_2 labels.
- LG: Uses listener's GTPs without mutual gaze. The S_M, L_{1M}, and N_M labels of GTPs are integrated into S, L_1, and N labels.
- SG+NG+LG: Uses GTPs of the speaker, next speaker and listener without mutual gaze.
- SG+NG+LG+MG: Uses GTPs of the speaker, next speaker and listener with mutual gaze. i.e., all gaze labels are used.

We employed leave-one-out with 325 data of 10 dialogs, 10-fold cross validation. Then, the error of the results predicted from the actual utterance start time was calculated. The results are shown in Table 1. Table 1 shows that the average error using the data under 1000 ms, between 1000 and 2000 ms, between 2000 and 3000 ms, between 3000 and 4000 ms, over 4000 ms, and all as test data.

As the overall result, average error of the all-features model (SG+NG+LG+MG) is 755.4 ms, which is the lowest error value among the models. This result suggests that all the features − the GTPs of the speaker, next speaker, and listener, and mutual gaze − contribute to predict the start timing of the next speaker in multi-party meetings.

In comparing SG, NG, and LG models, averages of the error are 900.3 ms in the SG model, 896.3 ms in the NS model, and 988.7 ms in the LG model. The SG and NG models provide better performance than the LG model. This suggests that the speaker's and the next speaker's gaze behavior are more important for predicting the start timing of the next speak than the listener's behavior.

The performance of SG+NG+LG is much better than the SG, NG, and LG models; specifically, the prediction performance is improved. This suggests that all participants' GTPs are strong predictors of the next speaking timing. Moreover, comparing SG+NG+LG and SG+NG+LG+MG, the averages of the error are founded to be 797.9 ms in SG+NG+LG, and 755.4 ms in SG+NG+LG+MG. This suggests that mutual gaze is useful in predicting the start timing of the next speaker. In the SG+NG+LG+MG model, the average error is fairly low, almost 300 ms: it is 335.7 ms in the test data under 1000 ms and 315 ms between 1000 and 2000 ms. However, the average error increased in the data over 2000 ms: it is 793.4 ms between 2000 and 3000 ms, 1883.1 ms between 3000 and 4000 ms, and 5101.2 ms over 4000 ms.

The interval between the end of speaking and the start time of the next speaking in smooth turn-taking situation is generally below about 2000 ms. This data is 81.8% of the 325 data of the corpus data. The SG+NG+LG+MG model can predict the start timing with high accuracy in its own way in such a smooth turn-taking situation where the interval is under 2000 ms. That the prediction model cannot predict the timing very well in data over 2000 ms is the correct result.

5. CONCLUSION

We focused on GTPs in multi-party meetings, which have not been tackled until now, and demonstrated a correlation between the start timing of next speaker and GTPs of the speaker, next speaker, and listener. On the basis of the results of the analysis, we used the variables, the GTPs of the speaker, next speaker, and listener, and mutual gaze as prediction features, and devised prediction models comprising different combinations of the features. To test which features are effective and which are not, the performance of these models was compared. As a result, it was revealed that a model using all of the features as prediction features performed the best and was able to predict the start timing of the next speaker well.

In the future work, we plan to analyze in detail the relevance of nonverbal behavior, such as head gestures, and prosody and build a more highly precise prediction model.

6. REFERENCES

[1] Kazuhiro Otsuka, "Conversational scean analysis," *IEEE Signal Processing Magazine*, vol. 28, pp. 127–131, 2011.

[2] Daniel Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: a review," *Image and Vision Computing, Special Issue on Human Behavior*, vol. 27, no. 12, pp. 1775–1787, Nov 2009.

[3] Toshihiko Itoh, Norihide Kitaoka, and Ryota Nishimura, "Subjective experiments on influence of response timing in spoken dialogues," in *Proceeding of the Interspeech*, 2009, pp. 1835–1838.

[4] Masayuki Inoue, Isamu Yoroizawa, and Sakae Okubo, "Human factors oriented design objectives for video teleconferencing systems," in *ITS*, 1984, pp. 66–73.

[5] Adam Kendon, "Some functions of gaze direction in social interaction," *ActaPsychologica*, vol. 26, pp. 22–63, 1967.

[6] Kornel Laskowski, Jens Edlund, and Mattias Heldner, "A single-port non-parametric model of turn-taking in multi-party conversation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5600–5603.

[7] Kristiina Jokinen, Kazuaki Harada, Masafumi Nishida, and Seiichi Yamamoto, "Turn-alignment using eye-gaze and speech in conversational interaction," in *INTERSPEECH*, 2011, pp. 2018–2021.

[8] Alfred Dielmann, Giulia Garau, and Herv? Bourlard, "Floor holder detection and end of speaker turn prediction in meetings," in *INTERSPEECH*, 2010, pp. 2306–2309.

[9] Lei Chen and Mary P. Harper, "Multimodal floor control shift detection," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2009, pp. 15–22.

[10] Iwan de Kok and Dirk Heylen, "Multimodal end-of-turn prediction in multi-party meetings," in *ACM International Conference on Multimodal Interaction (ICMI)*, 2009, pp. 91–98.

[11] Tatsuya Kawahara, Takuma Iwatate, and Katsuya Takanashii, "Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations," in *INTERSPEECH*, 2012, pp. 9–13.

[12] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch, "Predicting listener backchannels: A probabilistic multimodal approach," in *International Conference on Intelligent Virtual Agents (IVA)*, 2008, pp. 176–190.

[13] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch, "A multimodal end-of-turn prediction model: Learning from para social consensus sampling," in *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2011, pp. 1289–1290.

[14] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson, "A simplest systematics for the organisation of turn taking for conversation," *Language*, vol. 50, pp. 696–735, 1974.

[15] Michael Kipp, "Anvil - a generic annotation tool for multimodal dialogue," in *INTERSPEECH*, 2001, pp. 1367–1370.

[16] Anthony J. Conger, "Integration and generalization of kappas for multiple raters," *Psychol Bull*, vol. 88, pp. 322–328, 1980.

[17] Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs," in *Language and Speech*, 1998, vol. 41, pp. 295–321.

[18] Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt, "Addressee identification in face-to-face meetings," in *Conference of the European Chapter of the ACL*, 2006.

[19] Shirish Krishnaj Shevade, S. Sathiya Keerthi, Chiranjib Bhattacharyya, and K. R. K. Murthy, "Improvements to the smo algorithm for svm regression," in *IEEE Transactions on Neural Networks*, 1999.

[20] Alex J. Smola and Bernhard Scholkopf, "A tutorial on support vector regression," Tech. Rep., 1998, NeuroCOLT2 Technical Report NC2-TR-1998-030.

[21] Remco R. Bouckaert, Eibe Frank, Mark A. Hall, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, "WEKA–experiences with a java open-source project," *Journal of Machine Learning Research*, vol. 11, pp. 2533–2541, 2010.