Predicting *Who Will Be the Next Speaker and When* in Multi-party Meetings

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato

Abstract

An understanding of the mechanisms involved in face-to-face communication will contribute to designing advanced video conferencing and dialogue systems. Turn-taking, the situation where the speaker changes, is especially important in multi-party meetings. For smooth turn-taking, the participants need to predict who will start speaking next and to consider a strategy for achieving good timing to speak next. Our aim is to clarify the kinds of behavior that contribute to smooth turn-taking and to develop a model for predicting the next speaker and the start time of the next speaker's utterance in multi-party meetings. We focus on gaze behavior and respiration near the end of the current speaker's utterance. We empirically demonstrate that gaze behavior and respiration have a relation to the next speaker and the start timing of the next utterance in multi-party meetings. A prediction model based on the results reveals that gaze behavior and respiration contribute to predicting the next speaker and the timing of the next utterance.

Keywords: turn-taking, gaze, respiration

1. Introduction

Face-to-face communication is one of the most basic forms of communication in daily life, and group meetings conducted using this kind of communication are effective for conveying information, understanding others' intentions, and making decisions. To design better communication systems that can enhance our communication beyond conversation *in loco*[*] and to develop social agents/robots that interact naturally with human conversations, it is critical to fully understand the mechanism of human 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 who will start speaking next and to consider a strategy for good timing with respect to who will speak next in multi-party meetings. If a model can predict the next speaker and the timing that the next speaker's utterance will start, the model will lay the foundation for the development of natural conversational systems in which conversational agents/robots speak with natural timing and of teleconference systems that avoid utterance collisions with time delays by apprising participants of who will speak next.

The goal of our research is to demonstrate the mechanism of turn-taking, namely what kind of information contributes to determining the next speaker and the timing of the next utterance, and to construct a prediction model that can predict who speaks next and when. To predict the next speaker and the timing of the next utterance, we developed a prediction model that has a three-step processing

* *in loco*: A Latin term meaning "in the proper place."

sequence: (I) prediction of turn-taking occurrence, (II) prediction of the next speaker in turn-taking, and (III) prediction of the timing of the next utterance. A flowchart of the model is shown in **Fig. 1**.

We focus on gaze behavior and respiration as information related to turn-taking. Gaze behavior is known to be an important cue for smooth turn-taking [3–5]. For example, the next speaker looks away when he/she starts to speak after having made eye contact with the current speaker in two-person meetings. However, previous research has only roughly demonstrated gaze behavior tendencies; it has not quantitatively revealed the relationship between gaze behavior and the next speaker and the timing of the next utterance. It is known that utterances and respiration are closely related. In order to speak, we must breathe out, and we need to take breaths to continue speaking for a long time. When starting an utterance, the next speaker inhales deeply. Moreover, a person's attitude about an utterance is frequently represented figuratively as *breathing*. For example, keeping as low a profile as possible so as not to be yielded the turn is often referred to metaphorically as *holding one's breath* or *saving one's breath*.

Thus, we focused on the detailed transitions of gaze behavior and respiration, which have not been investigated in multi-party meetings analysis. We previously examined the relationship between the transitions of gaze behavior and respiration and the next speaker and next-utterance timing and revealed that transitions in gaze behavior and respiration are useful for predicting them in multi-party meetings [6–8]. In this article, we describe how we analyzed gaze behavior and respiration in multi-party meetings to construct our prediction models.

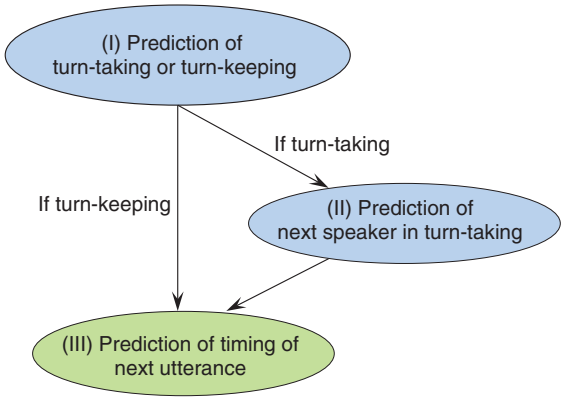


Fig. 1. Process flowchart of prediction model of next speaker and timing of next utterance.

2. Corpus of multi-party meetings

To analyze gaze and respiration, we collected a corpus of conversations in multi-party meetings. We recorded four natural (i.e., unrehearsed) meetings, such as the kind that would be held daily, conducted by four groups consisting of four different people (16 people in total) for about 12 minutes [8]. In each meeting, all four participants were in their 20s and 30s, 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 such as "Is marriage the same as love?" and were instructed to draw a conclusion within 12 minutes.

We recorded the participants' voices with a pin microphone attached to their clothing at chest level and made a video recording of the entire scene and took bust shots (head and shoulders) of each participant (recorded at 30 Hz). We recorded each participant's respiration information using a NeXus-10 MARK II biofeedback system. The respiration sensor of the system records the depth of breathing with a belt wrapped around the participant's body and outputs a value of the degree of breathing depth (called the RSP value hereafter) at 128 Hz. A high RSP value means that the person keeps taking air into the lungs. In contrast, a low RSP value means the absence of air in the lungs. We collected data during four meetings that were each about 12 minutes long (a total of about 50 minutes), and from the recorded data, we built a multimodal corpus consisting of the following verbal and nonverbal behaviors and biological information:

- Utterance: For the utterance unit, we used the inter-pausal unit (IPU) [9]. The utterance interval was extracted manually from the speech wave. The portion of an utterance followed by 200 ms of silence was used as the unit of one utterance. An utterance unit that continued by the same person was considered to be one utterance turn. Pairs of IPUs that adjoined in time, and groups of IPUs at the time of turn-keeping and turn-taking were created. There were 906 groups of IPUs created at the time of turn-keeping and 148 at the time of turn-taking.
- Gaze object: The gaze object was annotated manually using the video showing the upper body of the participants as well as overhead views from the videos by a skilled annotator. The objects of gaze were the four participants (hereafter denoted as P1, P2, P3, and P4) and other

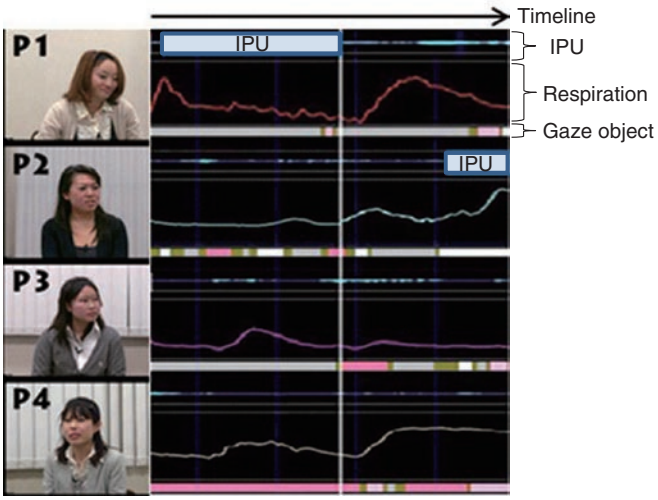


Fig. 2. Corpus data of multi-party meetings.

objects such as the floor or ceiling.

- Respiration: The RSP value measured by the Nexus MARK II has a different magnitude for each participant. To correct for these individual differences, the RSP value of each participant was normalized by the mean value μ and standard deviation δ of each participant. Specifically, the RSP value of each participant was normalized on the basis of the values of $\mu + \delta$ and $\mu - \delta$ for each participant. This enabled us to treat each participant's RSP value data for the analysis on the same scale.

All the above-mentioned data were integrated at 30 Hz for visual display using the NTT Multimodal Meeting Viewer (NTT MM-Viewer) that we developed [10] (See **Fig. 2**).

3. Prediction of next speaker and next-utterance timing based on gaze behavior

3.1 Analysis of gaze behavior

Gaze behavior is known to be an important cue for smooth turn-taking. For example, Kendon [4] demonstrated that the next speaker looks away when he/she starts to speak after having made eye contact with the current speaker in two-person meetings. Thus, it is assumed that these temporal transitions of participants' gaze behavior are important for the prediction of turn-taking situations. We therefore decided to focus on the gaze transition patterns near the end of utterances and to express them as an n-gram, which we defined as a sequence of gaze direction shifts. To generate a gaze transition pattern, we focused on the object a participant gazed at (*gazed object* hereafter) that occurs for 1200 ms during the period 1000 ms before the utterance and 200 ms after it; the candidate gazed objects were first classified as *speaker*, *listener*, or *others* (non-person objects) and labeled. At this time, the existence of mutual gaze was taken into consideration from the knowledge [4–6] that a mutual gaze takes place in two-person dialogs at the time of turn-taking. We used the following five gaze labels for the classification:

- S: Listener looks at a speaker without mutual gaze (speaker does not look at the listener.).
- S_M: Listener looks at a speaker with mutual gaze (speaker looks at the listener.).
- L_1-L_3: Speaker or listener looks at another listener without mutual gaze. L_1, L_2, and L_3 indicate different listeners.
- L_{1M}-L_{3M}: Speaker or listener looks at another listener with mutual gaze. L_{1M}, L_{2M}, and L_{3M} indicate different listeners.
- X: Speaker or listener looks at a non-person object such as the floor or ceiling.

The construction of a gaze transition pattern is shown in **Fig. 3**: P1 finishes speaking, and then P2 starts to speak. P1 gazes at P2 after he/she gazes at others during the interval of analysis. When P1 looks at P2, P2 looks at P1; namely, there is mutual gaze. Therefore, P1's gaze transition pattern is X-L_{1M}. P2 looks at P4 after making eye contact with P1. Therefore, P2's gaze transition pattern is S_M-L_1. P3 looks at others after looking at P1. P3's gaze transition

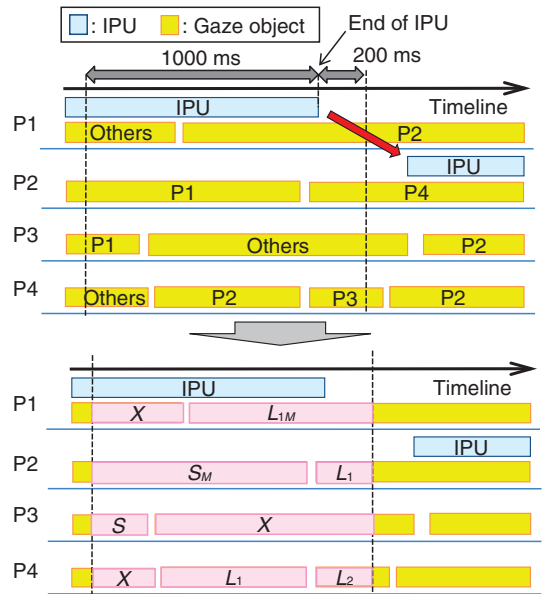


Fig. 3. Sample of gaze transition pattern generation.

pattern is therefore *S*-*X*. P4 looks at P2 and P3 after looking at others. Thus, P4's gaze transition pattern is X-L_1-L_2.

If the next speaker and the next-utterance timing differ depending on the gaze transition pattern, the gaze transition pattern may be useful for predicting them in multi-party meetings. To explore the relationship between gaze transition patterns and the next speaker and next-utterance timing, we analyzed the gaze transition pattern when turn-keeping and turn-taking occur and whether the next speaker in turn-taking and the next-utterance timing depend on the gaze transition pattern. After finishing the three analyses, we built a prediction model using gaze transition patterns.

3.2 Analysis of gaze transition pattern and turn-keeping/turn-taking

First, we analyzed how much the change in the gaze transition pattern of the speaker and listeners would differ quantitatively by turn-taking and turn-keeping. In this article, we introduce the results of analyzing the speaker's gaze transition pattern. The frequency of appearance of a speaker's gaze transition pattern under turn-taking and turn-keeping conditions using 1054 data sets is shown in **Fig. 4**. The results indicated that there were 17 gaze transition patterns. The *Others* class includes six patterns, each of which occurred in less than 1% of the data because the amount of data was small. The results of a chi-squared test showed that the frequent appearance of a gaze transition pattern differed significantly between the conditions at the time of turn-taking and turn-keeping ($\chi^2 = 87.03$, $df = 11$, $p < .01$). Next, we conducted a residual analysis to verify which gaze transition pattern differed between turn-keeping and turn-taking. The results are also shown in Fig. 4, from which we understand the following:

- A speaker's gaze transition pattern has a significantly high frequency of turn-keeping at the time of X and X-L_{1M}-X. That is, when a speaker does not look at a listener at all, or a mutual gaze with a listener is started and a gaze is stopped immediately, the frequency of turn-keeping is higher than turn-taking.
- A speaker's gaze transition pattern has a significantly high frequency of turn-taking at the time of L_{1M}, L_1, X-L_1, and L_1-X. That is, when a speaker continues to look at a listener (in spite of the presence or absence of mutual gaze), starts to gaze at a listener (not a mutual gaze), or stops looking at a listener (not a mutual gaze), the frequency of turn-taking is high.

We found that the frequency of the different gaze transition patterns for a speaker differed in turn-keeping and turn-taking. Similarly, we found that the frequency of the different gaze transition patterns for a listener differed in turn-keeping and turn-taking. Therefore, these results suggest that gaze transition patterns of the speaker and listeners are valuable information for predicting turn-keeping and turn-taking.

3.3 Analysis of gaze transition pattern and next speaker in turn-taking

Next, we analyzed the frequency of each listener's becoming the next speaker according to the speaker's and listeners' gaze transition patterns. We present here the results of analyzing the listeners' gaze transition patterns. The results of totaling who becomes the next speaker for every listener's gaze transition pattern with 148 turn-taking data sets are shown in **Fig. 5**, where the yellow bars represent the frequency that the listener herself who exhibited a gaze transition pattern (behavior) becomes the next speaker. Moreover, in a gaze transition pattern that has L_1-L_3 or L_{1M}-L_{3M}, listeners are classified into two categories: a listener who is gazed at and a listener who is not gazed at by the listener who exhibited a gaze transition pattern.

For example, in the listener's gaze transition pattern L_{1M}, the frequency of the listener exhibiting that gaze

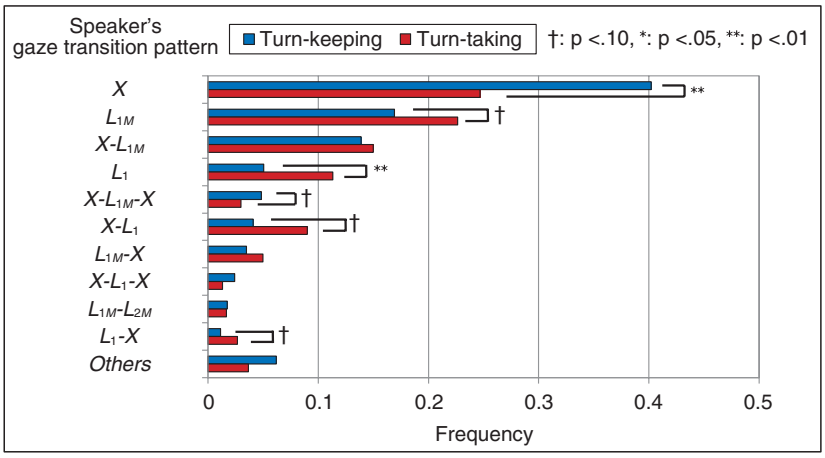


Fig. 4. Relationship between speaker's gaze transition pattern and turn-keeping/turn-taking.

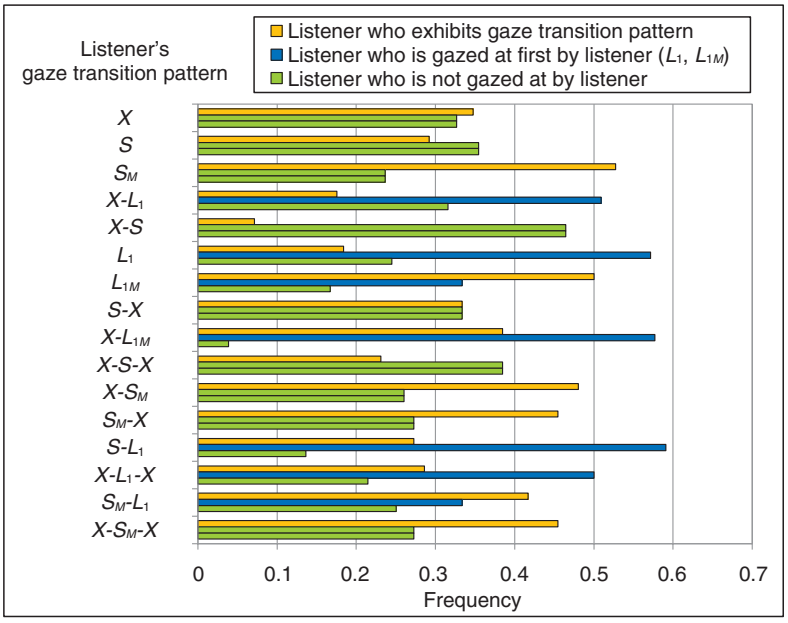


Fig. 5. Relationships between listener's gaze transition pattern and next speaker in turn-taking situation.

transition pattern who becomes the next speaker is 0.50, the frequency that the listener who is gazed at first by the listener (exhibiting the gaze pattern) is 0.33, and the frequency that another listener who is not gazed at by the listener (exhibiting the gaze pattern) becomes the next speaker is 0.17. The following becomes clear when the relationship between a listener's gaze transition pattern and the next speaker is seen in detail.

- When the listener's gaze transition pattern includes S_M, for example, S_M, X-S_M, S_M-X, S_M-L_1, or X-S_M-X, i.e., a listener makes eye contact with the speaker, the frequency that the listener becomes the next speaker is highest.
- When the listener's gaze transition pattern is X-L_1, L_1, X-L_{1M}, S-L_1, or X-L_1-X, namely, when a listener starts to look at another listener, keeps on looking at another listener without mutual gaze, looks from the speaker to another listener (without mutual gaze), or stops looking at the speaker

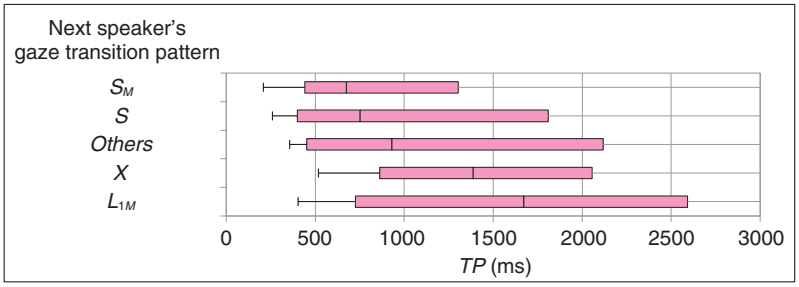


Fig. 6. Relationship between next speaker's gaze transition pattern and *TP* in turn-taking.

immediately after starting to look at the speaker (without mutual gaze), the frequency that the listener falling into pattern L_1 or L_{1M} becomes the next speaker is highest. Conversely, when the listener's gaze transition pattern is L_{1M}, i.e., a listener continues to carry out a mutual gaze with another listener, the frequency that the listener herself becomes the next speaker is highest.

We found that the frequency of each listener's becoming the next speaker in turn-taking differs depending on the listeners' gaze transition patterns. Similarly, we found that the frequency of each listener's becoming the next speaker in turn-taking differs depending on the speaker's gaze transition pattern. These results suggest that gaze transition patterns of the speaker and listeners are valuable information for predicting the next speaker in turn-taking situations.

3.4 Analysis of gaze transition pattern and next-utterance timing

An early study on this topic [4] showed that a listener's response is delayed if a speaker does not look at the listener; consequently, we think that gaze behavior is useful for predicting the timing of the next utterance. We quantitatively analyzed the correlation between the timing of the next utterance and the gaze transition pattern of the speaker and listeners. We defined timing interval *TP* as the interval between the end of the speaker's IPU and the start of the next speaker's IPU.

We analyzed the *TP* for each gaze transition pattern of the speaker and listeners in turn-keeping and of the speaker, listeners, and next speaker in turn-taking.

In this article, we introduce the results for only the next speaker's gaze transition pattern in turn-taking. Box plots of *TP* obtained for each next speaker's gaze transition pattern using 148 data sets are shown in **Fig. 6**. The *Others* class includes 38 patterns, each of which occurred in less than 5% of the data because the number of data was small. A one-way ANOVA (analysis of variance) showed that there is a significant difference in *TP* depending on the speaker's gaze transition patterns ($F(4,315) = 2.05$, $p < .10$). Here, S_M and S, which indicate that the next speaker continues to look at the current speaker, have the shortest median values, 675 and 754 ms. In contrast, L_{1M}, which means the next speaker continues to look at the listener with mutual gaze, has the longest median value, 1673 ms. That is, when the next speaker continues to look at the current speaker, the timing of the next speaker's utterance starts early. When the previously reported gaze behaviors mentioned above [4] occur, the next speaker starts to speak quickly. In contrast, when the next speaker does not look at the current speaker, turn-taking is not smooth, and the timing is late.

We found that the next-utterance timing differs depending on the next speaker's gaze transition pattern in turn-taking. Similarly, we found that the frequency of the next-utterance timing differs depending on the speaker's and listeners' gaze transition patterns in turn-keeping and turn-taking. These results suggest that gaze transition patterns of the speaker and listeners in turn-keeping and of the speaker, listeners, and next speaker in turn-taking influence the next-utterance timing situations. Therefore, it would be useful to use the gaze transition patterns to predict the next-utterance timing.

3.5 Prediction model using gaze transition pattern

The analysis results described in the previous subsections indicate that gaze transition patterns provide useful indicators of turn-keeping and turn-taking, the next speaker in turn-taking, and the timing of the next utterance in multi-party meetings. On the basis of

these results, we constructed a prediction model that features three processing steps using gaze transition patterns. The model is based on a support vector machine (SVM), in which the method is SMO (sequential minimal optimization) [11]. We also evaluated the accuracy of the model, along with the effectiveness of the gaze transition patterns. The settings of the SVM are the radial basis function (RBF) kernel.

For the turn-keeping/turn-taking prediction model, the data used in the SVM consist of the turn states of turn-taking and turn-keeping as a class and the gaze transition patterns of the speaker and three listeners as features. In this phase of the study, we employed leave-one-out with 296 data sets: 148 data sets obtained by sampling 906 data items in turn-keeping to remove the bias of the number of data, and 148 data sets in turn-taking, four-fold cross validation. We collected data from four groups; we obtained training data from three of them and test data from the remaining one. Then we calculated the average prediction accuracy. The results of the evaluation indicated an accuracy rate of 65.0% in turn-keeping and 68.2% in turn-taking. This suggests that the gaze transition patterns are useful for predicting turn-taking and turn-keeping.

For the prediction model of the next speaker in turn-taking, the data used in the SVM consist of the listener who will be the next speaker as a class and the gaze transition patterns of the speaker and three listeners as features. In this phase of the study, we employed leave-one-out with 148 data sets of four dialogs, four-fold cross validation. The results showed an average prediction accuracy rate of 61.0%. This suggests that the gaze transition patterns contribute to predicting the next speaker in turn-taking.

For the prediction model of the next-utterance timing, the data used in the SVR (SVM for regression) contain the start timing as a class and the gaze transition patterns of the speaker and three listeners in turn-keeping and of the speaker, two listeners, and next speaker in turn-taking as features. We examined two models for turn-keeping and turn-taking situations. In this phase of the study, we employed leave-one-out with 906 data sets in turn-keeping and 148 data sets in turn-taking of four dialogs, four-fold cross validation. We calculated the error in the results predicted from the actual utterance start time was calculated. As a result, the average errors were 324 ms in turn-keeping and 1281 ms in turn-taking. In a similar manner, we examined a base model that outputs the average value of interval *TP*. The average error for the base model were 469 ms in turn-keeping and 1590 ms in turn-taking, which was higher than that for our prediction model. This suggests that the gaze transition patterns contribute to predicting the timing of the next speaker's first utterance in multi-party meetings.

4. Prediction of next speaker based on respiration

Utterances and respiration are known to be closely related. In order to speak, we must breathe out, and we need to take breaths to continue speaking for a long time. When starting an utterance, the next speaker inhales deeply. Moreover, a person's attitude about an utterance is frequently represented figuratively as *breathing*. For example, as mentioned previously, when someone tries to keep as low a profile as possible so as not to be yielded the turn, it is often referred to metaphorically as *holding one's breath* or *saving one's breath*. As a first attempt to deal with respiration, we conducted a fundamental study of the relationships between respiration and the next speaker in multi-party meetings [8]. After that, we devised a prediction model of the next speaker using respiration.

We analyzed how the respiration of the speaker and listeners in turn-keeping and of the speaker, listeners, and the next speaker quantitatively differs in turn-taking. We considered that if the analysis revealed differences in the speaker's respiration between turn-taking and turn-keeping or differences in respiration between the next speaker in turn-taking and the listener in turn-taking and turn-keeping, respiration could be used as a useful indicator to predict the next speaker.

We assumed that the speaker takes a breath quickly right after the utterance to continue to speak in turn-keeping. In contrast, we assumed that the speaker doesn't take a breath quickly right after the utterance in turn-taking. Therefore, we focused on the speaker's inhalation right after the end of IPU for the analysis of speaker's inhalation. We extracted the speaker's inhalation phase just after the end of the IPU and used the following inhalation-phase parameters for the speaker in order to compare differences in inhalation in detail (see **Fig. 7**).

- *MIN*: RSP value at the start of the inhalation phase, i.e., the minimum RSP value of the inhalation phase.
- *MAX*: RSP value at the end of the inhalation phase, i.e., the maximum RSP value of the inhalation phase.
- *AMP*: Amplitude of RSP value of inhalation

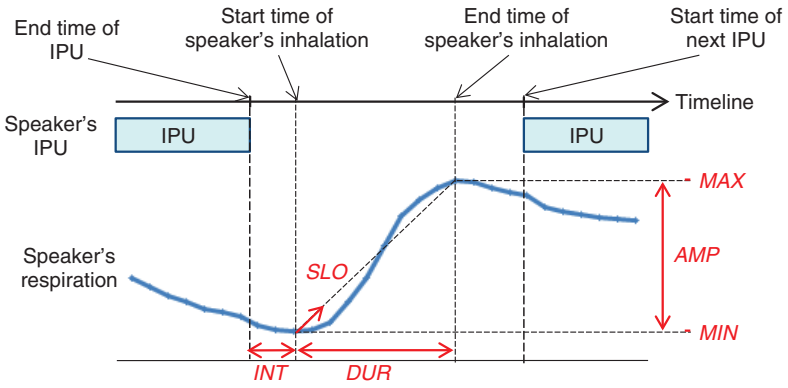


Fig. 7. Analytical parameters of inhalation right after end of IPU.

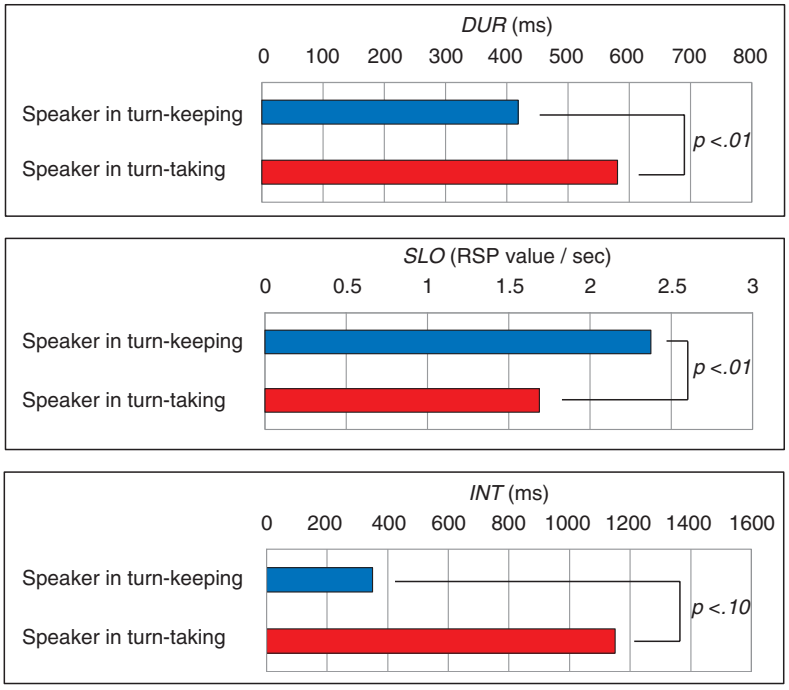


Fig. 8. *DUR*, *SLO*, and *INT* of speaker's inhalation phase right after end of IPU in turn-keeping and turn-taking.

phase.

- *DUR*: Duration of inhalation phase.
- *SLO*: Mean value of slope of RSP value per second during inhalation phase.
- *INT*: Interval between end time of speaker's IPU and start time of inhalation.

We analyzed these parameters in turn-taking and turn-keeping.

We calculated the mean value of the six parameters of the speaker in turn-taking and turn-keeping. Then, we calculated the mean value for all participants. We used a paired t-test to statistically verify whether the each parameter in turn-taking was significantly different from the same value in turn-keeping. The results suggested that there are significant differences in only *DUR*, *SLO*, and *INT* between turn-keeping and turn-taking ($t(30) = 3.08$, $p < .01$ for *DUR*; $t(30) = 2.96$, $p < .01$ for *SLO*; $t(30) = 2.04$, $p < .10$ for *INT*). These average values of *DUR*, *SLO*, and *INT* are shown in **Fig. 8**. These results reveal that the speaker

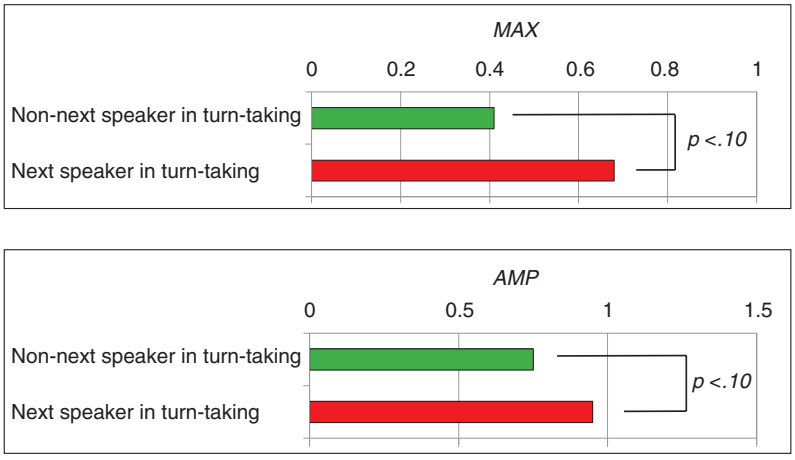


Fig. 9. *MAX* and *AMP* of inhalation phase of listeners who will not become next speaker (i.e., non-next speaker) and listener who will become next speaker (i.e., next speaker) right after end of IPU in turn-taking.

inhales more rapidly and quickly right after the end of a unit of utterance in turn-keeping.

We assumed that the next speaker takes a big breath right before starting to speak in turn-taking. In contrast, the listeners don't take a big breath. Therefore, we focused on the listeners' inhalation right before the start of next speaker's IPU for the analysis of listeners' respiration in turn-taking. We extracted the inhalation phase of the listeners. Then, we calculated the mean value of the six parameters for the listeners who will not become the next speaker (i.e., *non-next speakers*) and the listeners who will become the next speaker (i.e., *next speakers*) in turn-taking. We used a paired t-test to statistically verify whether the each parameter of the inhalation phase of the next speaker was significantly different from that of non-next speakers in turn-taking. The results suggested that there are significant differences in only *MAX*, and *AMP* between the next speaker and non-next speaker ($t(30) = 1.98$, $p < .10$ for *MAX*, $t(30) = 2.03$, $p < .10$ for *AMP*). The average values of *MAX* and *AMP* are shown in **Fig. 9**. These results reveal that the listener who will be the next speaker takes a bigger breath before speaking than listeners who will not become the next speaker in turn-taking.

The analysis results suggest that *DUR*, *SLO*, and *INT* of the speaker's inhalation right after the IPU can potentially be used to predict whether turn-keeping or turn-taking will occur, and the *MAX* and AMP of the inhalation phase of listeners can potentially be used to predict the next speaker in turn-taking.

To investigate the effectiveness of *DUR*, *SLO*, and *INT* of the speaker's inhalation right after the IPU to predict whether turn-keeping or turn-taking will occur, we constructed a prediction model based on an SVM and evaluated its performance. The SVM settings are RBF kernels. The data used in the SVM consist of the turn states of turn-taking and turn-keeping as a class and *DUR*, *SLO*, and *INT* of the speaker's inhalation as features. We employed the four-fold cross validation with 296 data sets, which includes 148 sets of data obtained by sampling 906 data items in turn-keeping in order to remove the bias of the number of data and 148 data sets in turn-taking. The accuracy rate of the prediction model was 78.7%. This suggests that parameters *DUR*, *SLO*, and *INT* of the speaker's inhalation right after the IPU contribute to predicting whether turn-keeping or turn-taking will occur.

Next, to investigate the effectiveness of *MAX* and *AMP* of the three listeners' inhalation before the next utterance in order to predict the next speaker in turn-taking, we constructed a prediction model based on the SVM as previously explained and evaluated its performance. The data used in the SVM consist of the next speaker as a class and the parameters *MAX* and *AMP* of each of the three listeners' inhalation as features. We employed four-fold cross validation with the 148 turn-taking data samples. The accuracy rate of the prediction model was 40.8%. The chance level was 33.3% because there were three next-speaker candidates in turn-taking. This suggests that the *MAX* and *AMP* parameters of the listeners' inhalation contribute to predicting the next speaker in turn-taking.

Therefore, we found that the participants' respiration is useful for predicting the next speaker in multi-party meetings.

5. Conclusion

We have developed a model for predicting the next speaker and the timing of the next speaker's utterance in multi-party meetings. As an initial attempt, we demonstrated how effective gaze behavior and respiration are in predicting the next speaker and next-utterance timing. We found from the results of analyzing gaze behavior that the next speaker and the timing of the next utterance differ depending on the gaze transition patterns of participants. The results of respiration analysis revealed that a speaker inhales more rapidly and quickly right after the end of a unit of utterance in turn-keeping than in turn-taking. The next speaker takes a bigger breath in preparing for speaking than listeners do who will not become the next speaker in turn-taking. We also constructed prediction models to evaluate how effective gaze behavior and respiration are in predicting the next speaker and the next-utterance timing. The results suggest that gaze behavior is useful for predicting the next speaker and the utterance timing, and respiration is also useful for predicting the next speaker. In the future, we plan to explore a high-performance prediction model using multimodal processing with the goal of achieving highly accurate prediction.

References

[1] K. Otsuka, "Conversation Scene Analysis," IEEE Signal Processing Magazine, Vol. 28, No. 4, pp. 127–131, 2011.

[2] D. Gatica-Perez, "Analyzing Group Interactions in Conversations: a Review," Proc. of 2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), pp. 41–46, Heidelberg, Germany, Sept. 2006.

[3] S. Duncan, "Some Signals and Rules for Taking Speaking Turns in Conversations," Journal of Personality and Social Psychology, Vol. 23, No. 2, pp. 283–292, 1972.

[4] A. Kendon, "Some Functions of Gaze Direction in Social Interaction," Acta Psychologica, Vol. 26, pp. 22–63, 1967.

[5] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn Taking for Conversation," Language, Vol. 50, No. 4, pp. 696–735, 1974.

[6] R. Ishii, K. Otsuka, S. Kumano, M. Matsuda, and J. Yamato, "Predicting Next Speaker and Timing from Gaze Transition Patterns in Multi-Party Meetings," Proc. of the 15th ACM International Conference on Multimodal Interaction (ICMI), pp. 79–86, Sydney, Australia, Dec. 2013.

[7] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Analysis and Modeling of Next Speaking Start Timing based on Gaze Behavior in Multi-party Meetings," Proc. of the 39th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 694–698, Florence, Italy, May 2014.

[8] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Analysis of Respiration for Prediction of "Who Will be Next Speaker and When" in Multi-party Meetings," Proc. of the 16th ACM International Conference on Multimodal Interaction (ICMI), pp. 18–25, Istanbul, Turkey, Nov. 2014.

[9] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An Analysis of Turn-taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogs," Language and Speech, Vol. 41, pp. 295–321, 1998.

[10] K. Otsuka, S. Araki, D. Mikami, K. Ishizuka, M. Fujimoto, and J. Yamato, "Realtime Meeting Analysis and 3D Meeting Viewer Based on Omnidirectional Multimodal Sensors," Proc. of the 11th ACM International Conference on Multimodal Interaction (ICMI), pp. 219–220, Cambridge, USA, Nov. 2009.

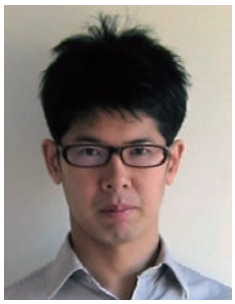
[11] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO Algorithm for SVM Classifier Design," Neural Computation, Vol. 13, No. 3, pp. 637–649, 2001.

Ryo Ishii
Research Engineer, Sensory Resonance Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.

He received the B.S. and M.S. in computer and information sciences from Tokyo University of Agriculture and Technology in 2006 and 2008, respectively, and the Ph.D. in informatics from Kyoto University in 2013. He joined NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories) in 2008. He moved to NTT Communication Science Laboratories in 2012. He was also an invited researcher at Seikei University, Tokyo, from 2011 to 2013. His current research interests include communication science, multimodal interactions, and human-computer interaction. He received the FY2014 IEICE HCG Research Award and the Association for Computing Machinery (ACM) Int. Conf. on Multimodal Interaction 2014 Outstanding Paper Award. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and JSAI (The Japanese Society of Artificial Intelligence).

Kazuhiro Otsuka
Senior Research Scientist, Supervisor, Human Information Science Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. in electrical and computer engineering from Yokohama National University, Kanagawa, in 1993 and 1995, respectively, and the Ph.D. in information science from Nagoya University, Aichi, in 2007. He joined NTT Human Interface Laboratories in 1995. He moved to NTT Communication Science Laboratories in 2001. In 2010, he was a distinguished invited researcher at Idiap Research Institute, Switzerland. His current research interests include communication science, multimodal interactions, and computer vision. He has received several awards including the Information Processing Society of Japan (IPSJ) National Convention Best Paper Award in 1998, the IAPR International Conference on Image Analysis and Processing Best Paper Award in 1999, the ACM Int. Conf. on Multimodal Interfaces 2007 Outstanding Paper Award, the Meeting on Image Recognition and Understanding (MIRU) 2009 Excellent Paper Award, the IEICE Best Paper Award 2010, the IEICE KIYASU-Zen'iti Award 2010, the MIRU2011 Interactive Session Award, the METI of Japan, Innovative Technologies Special Award (Digital Content Expo 2012), the ACM Int. Conf. on Multimodal Interaction Outstanding Paper Awards (2012 and 2014), and the IEICE Human Communication Award 2014. He is a member of the Institute of Electrical and Electronics Engineers (IEEE), IEICE, and IPSJ.



Shiro Kumano
Research Engineer, Sensory Resonance Research Group, Human Information Science Laboratory, NTT Communication Science Laboratories.

He received the Ph.D. in information science and technology from the University of Tokyo in 2009 and joined NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories) the same year. His research interests include computer vision and affective computing. He received the ACCV 2007 Honorable Mention Award. He has served as an organizing committee member of the IAPR International Conference on Machine Vision Applications. He is a member of IEEE, IEICE, and IPSJ.

Junji Yamato
Executive Manager, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E., M.E., and Ph.D. from the University of Tokyo in 1988, 1990, and 2000, respectively, and the S.M. in electrical engineering and computer science from Massachusetts Institute of Technology, USA, in 1998. His areas of expertise are computer vision, pattern recognition, human-robot interaction, and multi-party conversation analysis. He is a visiting professor at Hokkaido University and Tokyo Denki University and a lecturer at Waseda University. He is a senior member of IEEE and ACM.