Prediction of Next-Utterance Timing using Head Movement in Multi-Party Meetings

Ryo Ishii
NTT Media Intelligence Laboratories, NTT Corporation
1-1 Hikarinooka, Yokosuka, Kanagawa, Japan
ishii.ryo@lab.ntt.co.jp

Shiro Kumano
NTT Communication Science Laboratories, NTT Corporation
3-1, Morinosato-Wakamiya, Atsugi, Kanagawa, Japan
kumano.shiro@lab.ntt.co.jp

Kazuhiro Otsuka
NTT Communication Science Laboratories, NTT Corporation
3-1, Morinosato-Wakamiya, Atsugi, Kanagawa, Japan
otsuka.kazuhiro@lab.ntt.co.jp

ABSTRACT

To build a conversational interface wherein an agent system can smoothly communicate with multiple persons, it is imperative to know how the timing of speaking is decided. In this research, we explore the head movements of participants as an easy-to-measure nonverbal behavior to predict the next-utterance timing, i.e., the interval between the end of the current speaker's utterance and the start of the next speaker's utterance, in turn-changing in multi-party meetings. First, we collected data on participants' six degree-of-freedom head movements and utterances in four-person meetings. The results of the analysis revealed that the amount of head movements of current speaker, next speaker, and listeners have a positive correlation with the utterance interval. Moreover, the degree of synchrony of the head position and posture between the current speaker and next speaker is negatively correlated with the utterance interval. On the basis of these findings, we used their head movements and the synchrony of their head movements as feature values and devised several prediction models. A model using all features performed the best and was able to predict the next-utterance timing well. Therefore, this research revealed that the participants' head movement is useful for predicting the next-utterance timing in turn-changing in multi-party meetings.

ACM Classification Keywords

H.1.2 User/Machine systems: Human information processing

Author Keywords

Head movement; Utterance interval; Turn-taking; Multi-party meetings; Synchrony

INTRODUCTION

Face-to-face communication is one of the most basic forms of communication in daily life, and meetings are held for conveying information and making decisions. In remote human-to-human and human-to-humanoid settings, smooth communication similar to that expected in face-to-face communication is desired. This expectation has prompted the development of ways to automatically analyze multi-party meetings [7, 24, 25].

Turn-changing, the situation where the speaker changes, is an especially important aspect of smooth communication. Participants in multi-party meetings need to predict the end of the speaker's utterance and who will start speaking next; they also need a strategy for good timing with respect to who will speak next. In particular, the start time of the next utterance is very important for smooth communication. The start time varies with the situation and content of the utterances. Bad timing in speaking not only has a negative effect on communication but also sends unintended messages to conversational partners [17]. For example, even a short delay in video and audio of about 500 ms can inhibit smooth communication in remote video conferencing systems [8]. If a model can be devised to predict the next speaker and the start time of the next speaker's utterance, it would lay the foundation for the development of natural conversational systems in which conversational humanoids speak with natural timing and convey their thoughts correctly and of teleconference systems that avoid utterance collisions with time delays by appraising who will speak to the participants.

Studies in sociolinguistics have reported that verbal and nonverbal behaviors, such as gaze behavior, have an important association with the next speaker and the start of the next utterance [21, 29]. In engineering, several models for detecting the end of an utterance, i.e. whether turn-keeping or turn-changing in multi-party meetings happens, have been developed using voice information [23] and gaze behavior [2, 3, 5, 18]. Moreover, for automatically detecting or predicting who becomes the next speaker in a turn change, we have proposed to use gaze behavior [9, 12, 13], head movement [10], respiration [9, 14, 16], and mouse movement [11] to predict the next speaker in multi-party meetings.

In addition to its use in next-speaker prediction, gaze behavior and respiration have been shown to have a close relationship with the start time of the next utterance, and a model

HAI '17, October 17–20, 2017, Bielefeld, Germany

DOI: https://doi.org/10.1145/3125739.3125765

Figure 1. Sample scene of multi-party meetings and coordinate system with origin at center of seated positions.

that uses only gaze behavior in multi-party meetings to predict the start time has been developed [9, 12, 13, 14, 15]. It is hoped that the associations between other nonverbal behaviors and the next speaker and the start time will be able to be demonstrated. In addition, it is hoped that more robust and more precise prediction models can be formulated by using mulitimodal information. Since it is difficult to accurately measure gaze and respiration in meetings, the predictor using them is not practical. It is important, therefore, to predict the next-utterance timing by using a more easily measured form of nonverbal behavior.

In this study, we studied the relationship between head movements and the start time of the next utterance in turn-changing in multi-party meetings and demonstrated how useful head movements are for predicting the start time of the next utterance. Several preliminary studies have investigated head movement features related to the speaking state and turn-changing. Rienks et al. reported that humans can identify the current speaker simply from the orientations of the participants' heads in multi-party meetings [27]. Duncan et al. reported that a speaker tends to turn his/her head away from their partner in turn-keeping and a listener tends to change their head orientation to grab the turn in two-person dialogs [6, 28]. Maynard reported that nodding functions as a sign of end-of-turn and turn-grabbing and approval of turn-changing [30, 31]. These studies indicate that head movement has a strong relationship with turn-changing. Head movement may relate to the start time of the next utterance in multi-party meetings. However, no research has demonstrated such a relationship. Therefore, if our study could show such a relationship, it could lead to the possibility of predicting the start time of the next utterance from the head movements of the participants in multi-party meetings. As head movements can be readily measured with a camera or depth sensor, such as Kinect, they would be very useful for constructing a system that can predict the next-utterance timing.

In this research, we collected data on participants' six degree-of-freedom head movements and utterances in four-person meetings. We demonstrated the relationship between the head movements of the participants and the start of the next utterance after the end of the previous one in turn-changing in multi-party meetings. Then, we constructed a prediction model for the next utterance timing that uses head movements near the end of an utterance. An analysis of data collected from multi-party meetings reveals that the head movements of the current speaker, next speaker, and listeners and the degree of synchronization of head movements of the current speaker and next speaker correlate significantly with the start timing of the next utterance. On the basis of these results, we used their head movements and the synchrony of their head movements as feature values and devised several prediction models. Implementing the prediction model enabled us to quantitatively evaluate how head movement can clarify the mechanism that determines the timing of speaking.

COLLECTED DATA IN MULTI-PARTY MEETINGS

We recorded four natural 12-minute four-person meetings held by four groups of four different people (16 people in total) (total of about 50 minutes) (Fig. 1). We built a multimodal corpus consisting of the following verbal and nonverbal behaviors from the recorded data.

- Utterance: A pin microphone attached to the participants' chests recorded their voices. We built the utterance unit using an inter-pausal unit (IPU) [22]. The utterance interval was manually extracted from the speech wave. The portion of an utterance followed by 200 ms of silence was used as the unit for one utterance. The supportive responses [19] from the created IPU were excluded, and an utterance unit continued by the same person was considered one utterance turn. In addition, pairs of IPUs that adjoined in time and groups of IPUs at the time of turn-keeping and turn-changing were created. Data on speech overlap situations, i.e., when a listener interrupted a speaker's utterance or two or more participants spoke simultaneously in turn-changing, were excluded from the pairs of IPUs for the analysis. There were eventually 148 IPU groups for turn-changing.
- Head movement: In order to verify the relevance between the participants' head movements and the utterance timing in detail, we used a device that can measure head movements with high accuracy. Each participant's head movements were recorded using a Polhemus FASTRAK [26]. A small receiver attached to an adjustable band on the back of the participant's head detected the three DOF position

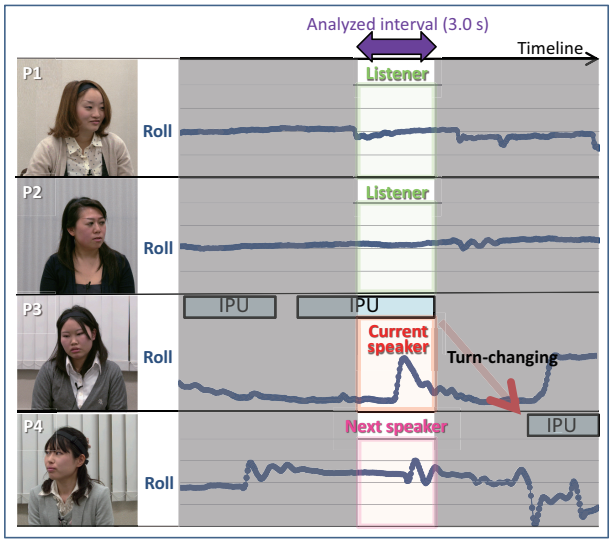

Figure 2. Sample data of roll of head posture in turn-changing. Turn-changing happens from P3 to P4.

(x, y, z) and the three DOF rotation (*yaw*, *roll*, *pitch*) at 30 Hz. The receiver's position and rotation from the sensor were treated as their head position and rotation. For the analysis, the sensor coordinate system was converted into a coordinate system with the origin located at the center of the sitting position of each participant. This coordinate system is shown in Fig. 1. The coordinates of the head position (x, y, z) and rotation (*yaw*, *roll*, *pitch*) of each participant were (0, 0, 0) and (0, 0, 0) in the coordinate system when they were sitting in the center their chair and their heads were toward the front.

All the above mentioned data were integrated at 30 Hz.

ANALYSIS OF HEAD MOVEMENT AND NEXT-UTTERANCE TIMING

Analysis method

Our previous research supposed that head movements of the current speaker, next speaker, and listeners right before the end of current speaker's utterance differ in turn-changing in multi-party meetings [10]. We analyzed the head motions of the current speaker, next speaker, and listeners near the end of an utterance separately while considering the differing characteristics of the head movements between them. Moreover, we focused on the synchrony of the head movements of the current speaker and next speaker, because the interpersonal synchrony of behaviors is very important in human communication and affects the mental state and behavior of participants [4]. It was revealed that characteristic movements of the head will appear about three seconds before the end of the utterance in turn-changing [10]. Fig. 2 shows a sample scene in which turn-changing happens from P3 to P4 and the roll data of participants. In this scene, the rolls of the speaker's and next speaker's head posture change greatly during the three seconds right before the end of an IPU. Thus, we analyzed the relationships between their head movements and the start time of the next utterance in turn-changing. We focused on head movements during the interval from three seconds before the end of an IPU to the start time of the next IPU as an analysis parameter, in the same manner as in [10]. We identified head movement waves (such as speech waveforms). The following parameters for the head position (x, y, z) and rotation (*yaw*, *roll*, *pitch*) were calculated for each wave and used in the analysis.

- *AV*: Average position or rotation.
- *MO*: Average amount of movement per second, expressed as the total amount during an interval divided by the interval length.
- *AM*: Average amplitude of movement per second, expressed as the mean amplitude of a wave during an interval.
- *FQ*: Average frequency of movement per second, expressed as the total number of waves during an interval divided by the interval length.

Table 1. *Results of correlation analysis of current speaker's head movement and start time of next utterance. Significant correlations ($p<.05$) are labeled with an asterisk (*).*

Parameters of head position	Correlation coefficient	Parameters of head rotation	Correlation coefficient
AV of x	0.053	*AV* of *yaw*	−0.035
MO of x	0.053	*MO* of *yaw*	−0.007
AM of x	0.063	*AM* of *yaw*	0.066
FQ of x	0.302*	*FQ* of *yaw*	0.167
AV of y	0.137	*AV* of *roll*	0.035
MO of y	−0.099	*MO* of *roll*	−0.198
AM of y	0.083	*AM* of *roll*	0.017
FQ of y	0.320*	*FQ* of *roll*	0.385*
AV of z	−0.029	*AV* of *pitch*	−0.045
MO of z	−0.142	*MO* of *pitch*	−0.064
AM of z	0.102	*AM* of *pitch*	−0.017
FQ of z	0.306*	*FQ* of *pitch*	0.322*

Current Speaker's Head Movement and Next-Utterance Timing

We conducted an analysis of the correlation by using Pearson's moment correlation coefficient between the *AV*, *MO*, *AM*, and *FQ* values of the x, y, and z coordinates and the *yaw*, *roll*, and *pitch* of the current speaker's head and the start time of the next speaker, i.e., the interval between the end of the current speaker's utterance and the start of the next speaker's utterance, in turn-changing with the 148 collected data. Table 1 shows the correlation coefficients. The table shows that only the *FQ* of the speaker's x, y, and z position and *roll* and *pitch* rotation coordinates have a statistically significant correlation ($p < .05$) with the start time of the next utterance.

Next Speaker's Head Movement and Next-Utterance Timing

We conducted a correlation analysis between the *AV*, *MO*, *AM*, and *FQ* values of the x, y, and z coordinates of the head position and *yaw*, *roll*, and *pitch* of the rotation of the head of the next speaker and the start time of the next utterance in turn-changing. Table 2 shows the correlation coefficients. The table shows that only the *FQ* of the z coordinate

Table 2. *Results of correlation analysis of next speaker's head movement and start time of next utterance. Significant correlations (p<.05) are labeled with an asterisk (*).*

Parameters of head position	Correlation coefficient	Parameters of head rotation	Correlation coefficient
AV of *x*	0.191	*AV* of *yaw*	0.038
MO of *x*	−0.075	*MO* of *yaw*	−0.103
AM of *x*	−0.002	*AM* of *yaw*	−0.003
FQ of *x*	0.061	*FQ* of *yaw*	0.053
AV of *y*	0.096	*AV* of *roll*	0.037
MO of *y*	−0.077	*MO* of *roll*	−0.081
AM of *y*	0.010	*AM* of *roll*	0.045
FQ of *y*	0.105	*FQ* of *roll*	0.048
AV of *z*	−0.082	*AV* of *pitch*	−0.061
MO of *z*	−0.083	*MO* of *pitch*	−0.076
AM of *z*	−0.006	*AM* of *pitch*	−0.033
FQ of *z*	0.304*	*FQ* of *pitch*	0.005

Table 3. *Results of correlation analysis of listeners' head movement and start time of next utterance. Significant correlations (p<.05) are labeled with an asterisk (*).*

Parameters of head position	Correlation coefficient	Parameters of head rotation	Correlation coefficient
AV of *x*	0.003	*AV* of *yaw*	0.013
MO of *x*	−0.010	*MO* of *yaw*	−0.084
AM of *x*	−0.031	*AM* of *yaw*	−0.086
FQ of *x*	0.294*	*FQ* of *yaw*	0.045
AV of *y*	−0.061	*AV* of *roll*	0.077
MO of *y*	−0.082	*MO* of *roll*	−0.132
AM of *y*	−0.057	*AM* of *roll*	−0.038
FQ of *y*	0.156	*FQ* of *roll*	0.028
AV of *z*	0.094	*AV* of *pitch*	−0.031
MO of *z*	−0.120	*MO* of *pitch*	−0.097
AM of *z*	0.001	*AM* of *pitch*	−0.074
FQ of *z*	0.120	*FQ* of *pitch*	0.049

of the speaker's head has a statistically significant correlation ($p < .05$) with the start time of the next utterance.

Listeners' Head Movement and Next-Utterance Timing

We conducted an analysis of the correlation between the *AV*, *MO*, *AM*, and *FQ* values of the *x*, *y*, and *z* coordinates of the positions and the *yaw*, *roll*, and *pitch* of the rotations of the listeners' heads and the start time of the next speaker in turn-changing. Table 3 shows the correlation coefficients. The table shows that only the *FQ* of the *x* coordinate of the listeners' head has a statistically significant correlation ($p < .05$) with the start time of the next utterance.

Degree of Synchronization of Head Movements of Current Speaker and Next Speaker and Next-Utterance Timing

We calculated the absolute value of the difference in each head movement parameter between the current speaker and next speaker as the degree of synchronization. We analyzed the correlation of the degree of synchronization of the*AV*, *MO*, *AM*, and *FQ* values of the *x*, *y*, and *z* positions and the *yaw*, *roll*, and *pitch* rotations of the heads of the current speaker and next speaker and the start time of the next speaker in turn-changing. Table 4 shows the results. The table shows

Table 4. *Results of correlation analysis of degree of synchronization of head movements of current speaker and next speaker and start time of next utterance. Significant correlations (p<.05) are labeled with an asterisk (*).*

Parameters of head position	Correlation coefficient	Parameters of head rotation	Correlation coefficient
AV of *x*	−0.180	*AV* of *yaw*	−0.033
MO of *x*	−0.062	*MO* of *yaw*	−0.045
AM of *x*	0.041	*AM* of *yaw*	0.044
FQ of *x*	−0.126	*FQ* of *yaw*	−0.154
AV of *y*	−0.069	*AV* of *roll*	−0.100
MO of *y*	−0.094	*MO* of *roll*	−0.083
AM of *y*	0.060	*AM* of *roll*	0.050
FQ of *y*	−0.132	*FQ* of *roll*	−0.052
AV of *z*	−0.342*	*AV* of *pitch*	−0.201*
MO of *z*	−0.095	*MO* of *pitch*	0.074
AM of *z*	0.051	*AM* of *pitch*	0.089
FQ of *z*	−0.132	*FQ* of *pitch*	−0.073

that only the *AV* of the degree of synchronization of the *z* coordinate and *pitch* of the current speaker and next speaker have a statistically significant correlation ($p < .05$) with the start time of the next utterance.

PREDICTION MODEL OF NEXT-UTTERANCE TIMING

From the analyses described in the previous sections, we found that several parameters of the head movements of the current speaker, next speaker, and listeners separately and the degree of synchronization between the current speaker and next speaker may be useful predictors of the start timing of the next speaker in multi-party meetings. In this section, we present models for predicting the start time of the next utterance using these parameters of the head movements. We constructed estimation models using SMOreg [20], which implements a support vector machine (SVM) for regression in Weka [1], and evaluated the accuracy of the models and the effectiveness of each feature. The settings of the SVM — the polynomial kernel, C (cost parameter), and γ (hyper parameter of the kernel)—were determined by using a grid search technique.

The criterion variable is the start timing of the next speaker, i.e., the duration between the end of the current speaker and the start of the next speaker. The parameters of the head movements of the speaker, next speaker, and listeners were used as feature values. We implemented six prediction models that used each head movement or all of them to evaluate the usefulness of each for predicting the next-utterance timing. The six prediction models were as follows.

- Baseline: This model outputs the average value of the interval between the end of current speaker's utterance and the start time of the next speaker's utterance, which is 1595 ms.

- CSH: This model uses the five parameters of the current speaker's head movement that showed a statistically significant correlation with the start time of the next utterance: *FQ* of the *x*, *y*, and *z* coordinates and *roll* and *pitch* of the current speaker's head.

Table 5. *Results of evaluation of next-utterance timing prediction model.*

	Under 1000 ms	Between 1000 ms and 2000 ms	Between 2000 ms and 3000 ms	Over 3000 ms	All
Baseline	1131 ms	413 ms	772 ms	3537 ms	1228 ms
CSH	805 ms	917 ms	838 ms	3420 ms	1189 ms
NSH	869 ms	823 ms	847 ms	3367 ms	1201 ms
LH	1083 ms	329 ms	1105 ms	3055 ms	1137 ms
CoH	1219 ms	320 ms	634 ms	3475 ms	1130 ms
AllH	621 ms	481 ms	1000 ms	2164 ms	856 ms

- NSH: This model uses the one parameter of the next speaker's head movement that showed a statistically significant correlation with the start time of the next utterance: the *FQ* of the *z* coordinate of the head of the next speaker.
- LH: This model uses the parameter of the listeners' head movements that had a statistically significant correlation with the start time of the next utterance. In detail, it uses the two parameters— the maximum value and minimum value of *FQ* of two listeners' *x* head position coordinates— because there are two listeners in turn-changing in a four-person meeting.
- CoH: This model uses the two parameters of the speaker's head movement that showed a statistically significant correlation with the start time of the next utterance: the *AV* of the speaker's head's *z* coordinate and *pitch*.
- AllH: This model uses all ten head movement parameters used in the CSH, NSH, LH, and CoH models.

We used four-fold cross validation in which we left out one of the four dialogs in the 148 data. We divided the test data depending on the duration, which is the criterion variable, into under 1000 ms, between 1000 and 2000 ms, between 2000 and 3000 ms, over 3000 ms. We calculated the average absolute error for each divided up set of test data and all of the data. These errors show how the prediction models' ability to predict the exact start time of the next utterance evolves over the different utterance intervals. Table 5 shows the results of the evaluation for each prediction model on each test data. The average error of the AllH model was 856 ms for all the test data, which is the lowest error among the models. This suggests that all the features—the head movements of the current speaker, next speaker, and listeners, and the degree of synchronization of head movements between the current speaker and next speaker— contribute to predicting the start timing of the next speaker in turn-changing in multi-party meetings.

Comparing the Baseline, CSH, NSH, LH, and CoH models, it can be seen that there are few differences in the averages of the error of all the test data: 1228 ms for the Baseline, 1189 ms for the CSH model, 1201 ms for the NSH model, 1137 ms for the LH model, and 1130 ms for the CoH model. The average error is fairly low in the AllH model: 621 ms for the test data under 1000 ms and 481 ms for the test data between 1000 and 2000 ms. However, the average error increased for the data over 2000 ms: 1000 ms for the test data between 2000 and 3000 ms and 2164 ms for the test data over 3000 ms.

DISCUSSION

The results of the analysis of head movements and the start time of the next utterance demonstrated that the *FQ* of the head position coordinates (*x*, *y*, and *z*) and the rotation (*roll* and *pitch*) of the current speaker are positively correlated with the start time of the next utterance. In other words, the larger the number of changes in the current speaker's head movements, including turning in various directions, nodding, or tilting, the later is the start time of the next utterance. We demonstrated that the *FQ* of the *z* head position coordinate of the next speaker is positively correlated with the start time of the next utterance. In other words, the more the next speaker's head moves up and down, the later is the start time of the next utterance. We found that the *FQ* of the *x* head position coordinate of listeners is positively correlated with the start time of the next utterance. In other words, the more the listeners' heads move right or left, the later is the start time of the next utterance. It is very interesting that listeners' head movements are related with the start time of the next speaker's utterance, even though the listeners don't participate in turn-changing. We demonstrated that the degree of synchronization of the *AVs* of the *z* coordinate and *pitch* between the current speaker and next speaker are negatively correlated with the start time of the next utterance. In other words, the more that the head height position or vertical head direction of the current speaker and next speaker match, the shorter is the start time of the next utterance. It is very interesting that the degree of synchronization of the head position and direction affect the timing of the next utterance in turn-changing. These findings concerning the relationship between head movement and the start time of the next utterance in turn-changing in multi-party meetings are new and interesting.

The evaluation of the next-utterance timing prediction models using head movements showed that the one using the head movements of the current speaker, next speaker, and listeners and the degree of synchronization of the head movements of the current speaker and next speaker is a better predictor than the ones using individual movements. This suggests that all the participants' head movements are important for predicting the timing of the next utterance. The interval between the end of one speaker speaking and the start of the next speaker in smooth turn-changing situations is generally less than about 2000 ms. This is true for 76.7% of the 146 data in the corpus. The "all-participants (AllH)" model can accurately predict the start timing in such a smooth turn-changing situation in which the interval is under 2000 ms. That the prediction model cannot predict the timing very well for data with inter-

vals over 2000 ms is a correct result. On the other hand, head movements don't affect the start time of the next utterance in non-smooth turn-changing situations.

The results of this study contribute the following insights to the development of a conversation agent system that can spontaneously perform turn-changing in multi-party meetings with humans. The first is the importance of having the conversation agent speak at an appropriate timing by predicting that timing from the movements of the persons' heads. We have already proposed a prediction model that can predict the next speaker using gaze, head movement, and respiration [9, 10, 11, 14, 16]. It is possible for a conversational agent incorporating these technologies to judge "who will be next speaker and when" from the nonverbal behavior of the participants and agents themselves. The second is the importance of controlling the motion of the conversational agent's head according to the timing of its utterances. For example, by having the agent raise or lower its head more as the timing of the utterance becomes later, the motions of the agent can appear more natural when it is the next speaker at the time of a turn change.

Now, we will clarify the subject of this research and its limitations. First, as an initial attempt at investigating the relationship between the timing of the next utterance and head movements, we focused on only head movements for three seconds. Three seconds is not necessarily the most appropriate parameter. It will be necessary to verify the effectiveness of measuring head movements over various durations. Second, we need to verify how versatile our results are in various conversation situations. The head movements at the time of a speaker change may change because of various factors such as the number of people, positional relationship, human relations, culture, conversation content, and so on. However, even if the situation changes, it is certain that head movements are important for knowing the timing of utterances. Third, for estimating utterance timings from head movements in real environments, it is necessary to verify how well the utterance timing can be estimated from the head movement depending on the sensing means. This research used FASTRAK, which can measure head movements with high accuracy. However, since the system needs to attach a magnetic sensor to the head, it is highly likely that it will be difficult to use it in a real environment. An image sensor such as Kinect would have inferior measurement accuracy compared with FASTRAK. It will be necessary to verify how useful the proposed method is when the measured data has poor accuracy.

CONCLUSIONS

We demonstrated a relationship between the head movements of participants and the start of the next utterance in turn-changing in multi-party meetings. The most noteworthy results of the analysis are that the amount of head movement of the current speaker, next speaker, and listeners has a positive correlation with the utterance interval. Moreover, the degree of synchrony of the head position and posture between the current speaker and next speaker has a negative correlation with the utterance interval. On the basis of these findings, we used head movements and the synchrony of head movements as feature values and devised several prediction models. A model using all features performed the best and was able to predict the next-utterance timing well. Therefore, we revealed that the participants' head movements are useful for predicting the next-utterance timing in turn-changing in multi-party meetings.

As head movements can be readily measured with a camera or depth sensor, such as Kinect, they would be very useful for constructing a system that can predict the next-utterance timing. We plan to create a robust and high-performance prediction model using multimodal information, such as gaze behavior, respiration, and mouse movement [9, 11, 12, 14, 15].

REFERENCES

1. Remco R. Bouckaert, Eibe Frank, Mark A. Hall, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2010. WEKA–Experiences with a Java Open-Source Project. *The Journal of Machine Learning Research* 11 (2010), 2533–2541.

2. Lei Chen and Mary P. Harper. 2009. Multimodal floor control shift detection. In *Proceedings of the International Conference on Multimodal Interaction*. 15–22.

3. Iwan de Kok and Dirk Heylen. 2009. Multimodal end-of-turn prediction in multi-party meetings. In *Proceedings of the International Conference on Multimodal Interaction*. 91–98.

4. E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. 2012. Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines. *IEEE Transactions on Affective Computing* 3, 3 (July 2012), 349–365.

5. Alfred Dielmann, Giulia Garau, and Herv? Bourlard. 2010. Floor holder detection and end of speaker turn prediction in meetings. In *Proceedings of the Annual Conference on the International Speech Communication Association*. 2306–2309.

6. S. Duncan and D. W. Fiske. 1977. Face-to-face interaction: research. *Methods and theory, Hillsdale, New Jersy: lawrence Erlbaum* (1977).

7. Daniel Gatica-Perez. 2009. Automatic Nonverbal Analysis of Social Interaction in Small Groups: a Review. *Image and Vision Computing, Special Issue on Human Behavior* 27, 12 (Nov 2009), 1775–1787.

8. Masayuki Inoue, Isamu Yoroizawa, and Sakae Okubo. 1984. Human Factors Oriented Design Objectives for Video Teleconferencing Systems. In *ITS*. 66–73.

9. Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015a. Multimodal Fusion using Respiration and Gaze for Predicting Next Speaker in Multi-Party Meetings. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. 99–106.

10. Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015b. Predicting Next Speaker Using Head Movement in Multi-party Meetings. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. 2319–2323.
11. Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2016a. Analyzing mouth-opening transition pattern for predicting next speaker in multi-party meetings. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. 209–216.
12. Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Masafumi Matsuda, and Junji Yamato. 2013. Predicting Next Speaker and Timing from Gaze Transition Patterns in Multi-Party Meetings. In *Proceedings of the International Conference on Multimodal Interaction*. 79–86.
13. Ryo Ishii, Kauhiro Otsuka, Shiro Kumano, and Junji Yamamoto. 2016b. Predicting of Who Will Be the Next Speaker and When Using Gaze Behavior in Multiparty Meetings. *The ACM Transactions on Interactive Intelligent Systems* 6, 1 (2016), 4.
14. Ryo Ishii, Kauhiro Otsuka, Shiro Kumano, and Junji Yamamoto. 2016c. Using respiration to predict who will speak next and when in multiparty meetings. *The ACM Transactions on Interactive Intelligent Systems* 6, 2 (2016), 20.
15. Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2014a. Analysis and Modeling of Next Speaking Start Timing based on Gaze Behavior in Multi-party Meetings. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. 694–698.
16. Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2014b. Analysis of Respiration for Prediction of Who Will Be Next Speaker and When? in Multi-Party Meetings. In *Proceedings of the International Conference on Multimodal Interaction*. 18–25.
17. Toshihiko Itoh, Norihide Kitaoka, and Ryota Nishimura. 2009. Subjective experiments on influence of response timing in spoken dialogues. In *Interspeech*. 1835–1838.
18. K Jokinen, K Harada, M Nishida, and S Yamamoto. 2011. Turn-alignment using eye-gaze and speech in conversational interaction. In *ISCA*. 2018–2021.
19. Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2006. Addressee identification in face-to-face meetings. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.
20. S. Sathiya Keerthi, Shirish Shevade, Chiranjib Bhattacharyya, and K.R. Krishna Murthy. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation* 13, 3 (2001), 637–649.
21. Adam Kendon. 1967. Some functions of gaze direction in social interaction. *ActaPsychologica* 26 (1967), 22–63.
22. H Koiso, Y Horiuchi, S Tutiya, A Ichikawa, and Y Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. In *Language and Speech*, Vol. 41. 295–321.
23. Kornel Laskowski, Jens Edlund, and Mattias Heldner. 2011. A single-port non-parametric model of turn-taking in multi-party conversation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. 5600–5603.
24. Hendrikus J.A. op den Akker, Daniel Gatica-Perez, and Dirk K.J. Heylen. 2012. Multi-modal analysis of small-group conversational dynamics. In *Multimodal Signal Processing*, J. Carletta & A. Popescu-Belis (Eds.) In S. Renals, H. Bourlard (Ed.). New York: Cambridge University Press, 155–169.
25. Kazuhiro Otsuka. 2011. Conversational scene analysis. *IEEE Signal Processing Magazine* 28 (2011), 127–131.
26. POLHEMUS. 2017. Fastrak. (2017). http://polhemus.com/motion-tracking/all-trackers/fastrak/.
27. Rutger Rienks, Ponald Poppe, and Dirk Heylen. 2010. Differences in head orientation behavior for speakers and listeners: An experiment in a virtual environment. *J. TAP* 7, 1(2) (2010).
28. Duncan S. and G. Niederehe. 1974. On signalling that it's your turn to speak. *J. Experimental Social Psychology* 10 (1974), 234–247.
29. Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organisation of turn taking for conversation. *Language* 50 (1974), 696–735.
30. Senko Maynard. 1987. Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation. *Journal of Pragmatics* 11 (1987), 589–606.
31. Senko Maynard. 1989. Japanese conversation: Self-contextualization through structure and interactional management. *Norwood, New Jersey: Ablex Publishing Corporation* (1989).