PREDICTING NEXT SPEAKER BASED ON HEAD MOVEMENT IN MULTI-PARTY MEETINGS

Ryo Ishii, Shiro Kumano, Kazuhiro Otsuka

NTT Communication Science Laboratories, NTT Corporation.

ABSTRACT

We proposed a model for predicting the next speaker in multi-party meetings by focusing on the participants' head movements measured by using a six degrees-of-freedom head tracker. Results of an analysis of head movements collected from multi-party meetings revealed differences in the amounts, amplitude, and frequency of movement of the head position and rotation of the speaker near the end of an utterance in turn-keeping and turn-taking. The results also revealed the differences in the amounts of movement, amplitude, and frequency of head position movement and rotation between the listeners in turn-keeping, turn-taking, and the next speaker in turn-taking. We then built a next speaker prediction model that features two processing steps to predict whether turn-taking or turn-keeping will occur and who the next speaker will be in turn-taking. The evaluation results for the model suggest that the speaker's and listeners' head movements contribute to predicting the next speaker.

Index Terms— Head movement, next-speaker prediction, turn-taking, multi-party meetings, meeting analysis

1. INTRODUCTION

The situation in which the speaker changes (turn-taking) during conversation is especially important. The participants need to predict the end of the speaker's utterance and who will speak next and to develop a strategy for good timing with respect to who will speak next in multi-party meetings. If a model were developed that could predict next speakers and the start of their first utterance, it could lay the foundation for the development of natural conversational systems in which the conversational agents speak using natural timing. We have proposed models to predict the next speaker and the start time of the next utterance using gaze behavior and respiration in natural multi-party meetings [1, 2, 3]. For more robust and highly precise prediction, the relationship between other nonverbal behaviors and the next speaker and the start time of the next speaking and the feasibility of creating a prediction model using multimodal signal processing need to be investigated.

We are investigating the role of head movement, measured using a six degrees-of-freedom (DOF) head tracker, and a prediction model for determining the next speaker using head movement as a first attempt. Several preliminary studies on two-person dialogs have investigated the head movement feature related to turn-taking [4, 5], but no reported research has investigated the relationship between six DOF head movement measured using a head tracker and the next speaker and start timing of next speaking or the creation of a prediction model for the next speaker and start timing of next speaking in multi-party meetings.

We collected a corpus from natural multi-party meetings including the participants' six DOF head movements measured using a head tracker and utterance information. We analyzed this data to determine how the speaker's and listeners' head movements change in turn-keeping and turn-taking. The analysis results revealed differences in the amounts and amplitudes of movement of the head positions and rotations of the speaker in turn-keeping and turn-taking. They also revealed differences in the amounts, amplitude, and frequency of movement of the head position and rotation between the listeners in turn-keeping, those in turn-taking, and the next speaker in turn-taking. We then established a next speaker prediction model that features two processing steps to predict whether turn-taking or turn-keeping will occur and who will be the next speaker in turn-taking. The evaluation results for this model showed that the speaker's and listeners' head movements are effective for predicting the next speaker.

2. RELATED WORK

It is known that verbal and nonverbal behaviors, such as the gaze behavior and prosody, have an important association with the next speaker and the start time of the next utterance [4, 5, 6, 7]. Several studies have explored the idea of automatically detecting whether or not turn-taking takes place in multi-party meetings by focusing on speech processing [8, 9, 10, 11, 12, 13, 14] and visual nonverbal behaviors such as gaze behavior [8, 9, 11] and physical motion [8, 9, 15] near the end of an utterance. In addition to the prediction of turn-taking, some studies have tried to predict who will become the next speaker at the time of turn-taking and the start time of the next speaker's utterance. Kawahara et al. proposed a next-speaker detection model using prosody and gaze in three-person poster conversations [16]. We proposed a prediction model for the next speaker and the start timing of the next utterance using gaze transition patterns and res-

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

piration in multi-party meetings [1, 2, 3]. It is believed the relationship between other nonverbal behaviors and the next speaker and the start time of the next speaking and creating a prediction model that uses multimodal signal processing are important for more robust and highly precise prediction.

Several preliminary studies have investigated the head movement feature related to a speaking state and turn-taking. Rienks et al. reported that human observers can identify the current speaker just by showing only participants' head orientations in multi-party meetings [17]. 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 [4, 5]. Several studies have built models to classify whether or not turn-keeping and turn-taking occurs using broadly divided head orientations annotated by an annotator observing video of a conversation. Jokinen et al. used nine kinds of rough head postures as features for classifying turn-keeping/turn-taking [11]. deKok et al. used six kinds of head posture intentions annotated by an annotator for classifying turn-keeping/turn-taking [9]. This head posture information is coarse compared to six DOF head movements measured with a head tracker. Therefore, no research has investigated the relationship between the six DOF head movement parameters and the next speaker in multi-party meetings.

In this paper, as a first attempt to deal with the six DOF head movements, we demonstrate the relationships between them and the next speaker in multi-party meetings.

3. CORPUS OF 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) [18]. 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 for speech overlap situations, i.e., when a listener interrupted during a speaker's utterance or two or more participants spoke simultaneously in turn-changing, were excluded from the pairs of IPUs for analysis. There were eventually 906 IPU groups for turn-keeping and 148 for turn-taking.
- Head movement: Each participant's head movement was recorded using a Polhemus FASTRAK [20]. The small receiver attached to an adjustable band on the back of the participant's head detects the three DOF position (X, Y, Z) and the three DOF rotation (azimuth, elavation, roll) at 30 Hz. The receiver's position and rotation from the censor were treated as their head position and rotation. The censor coordinate system used for the analysis 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 values of the head position (X, Y, Z) and rotation (azimuth, elevation, roll) of each participant were (0, 0, 0) and (0, 0, 0) in the coordinate system when their sitting position was centered in each chair and they turned their heads toward the front.

All the above mentioned data were integrated for 30 Hz.

4. ANALYSIS OF HEAD MOVEMENT

Previous research has demonstrated that a speaker tends to turn his/her head away from their partner in turn-keeping and a listener tends to change their head rotation to grab the turn in two-person dialogs [4, 5]. We analyzed the head motions of the speaker and listeners near the end of an utterance separately while considering the differing characteristics of the head movements between them. For the speaker's head movement analysis, we analyzed how their head movement differed between turn-keeping and turn-taking. For the listeners' head movement analysis, we divided the listeners into those in turn-keeping, ones who will not become the next speaker in turn-taking (hereafter, called "listeners in turn-taking"), and the one who will be the next speaker in turn-taking (hereafter, called "next speaker in turn-taking") and analyzed how the head movements differed between the listeners in turn-keeping and turn-taking and the next speaker in turn-taking. We focused on the head movement during the interval from three seconds before the end of an IPU to the start time of the next IPU as an analysis parameter. We identified head movement waves (such as speech waveforms). The following parameters for the head position (X, Y, Z) and rotation (azimuth, elevation, roll) were calculated for each wave and used for analysis.

- MO: Average amount of movement per second, expressed as the total amount during a focused interval divided by the interval length.

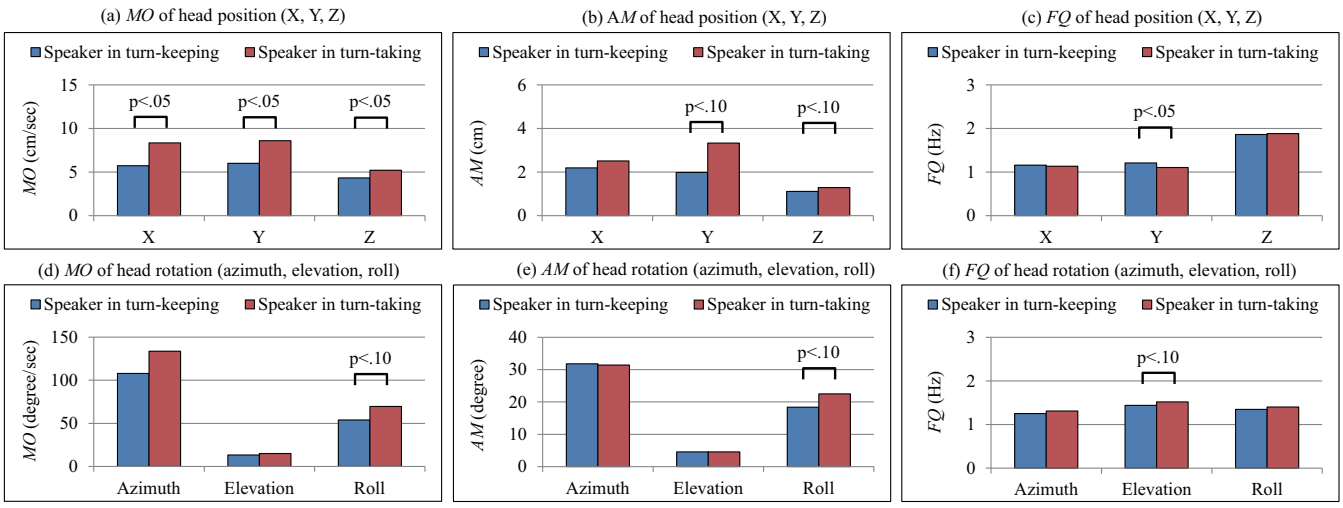

Fig. 2. Results of analysis of MO, AM, and FQ of speaker's head position (X, Y, Z) and rotation (azimuth, elevation, roll).

- AM: Average amplitude of movement per second, expressed as the mean amplitude value of a wave during a focused interval.
- FQ: Average frequency of movement per second, expressed as the total number of waves during a focused interval divided by the interval length.

4.1. Analysis of speaker's head movement

We calculated the mean MO, AM, and FQ values of the speaker's X, Y, and Z of head position and azimuth, elevation, and roll of the rotation with 906 data for turn-keeping and 148 data for turn-taking, which are shown in Fig. 2. We used an unpaired one-tailed t-test to statistically verify whether the MO, AM, and FQ of the speaker's head position and rotation in turn-taking are significantly different from those in turn-keeping. The results suggested that there is a significant difference in the MO of X, Y, Z, and roll, the AM of Y, Z, and roll, and the FQ of Y and elevation between turn-keeping and turn-taking[1]. This reveals that the MO of X, Y, Z, and roll and the AM of Y, Z, and roll are bigger in turn-taking than in turn-keeping.

4.2. Head movement of listeners

We calculated the mean MO, AM, and FQ values of the X, Y, and Z for head position and azimuth, elevation, and roll of the rotation of the listeners in turn-keeping and turn-taking and the next speaker in turn-taking, which are shown in Fig. 3. We performed a repeated one-way factorial analysis of variance to verify whether the condition of the listeners in turn-keeping and turn-taking and the next speaker in turn-taking affected the head position and rotation parameters. The results suggested there is a significant difference in the conditions for all the parameters[2]. Next, multiple comparisons using the Tukey-Kramer method were conducted to confirm the differences between each pair of conditions. The results for the MO of Y, azimuth, elevation, and roll, the AM of X, Y, Z, azimuth, elevation, and roll, and the FQ of X, Y, azimuth, elevation, and roll suggested there are significant differences between only listeners in turn-keeping and turn-taking and between listeners in turn-keeping and the next speaker in turn-taking (for the p value of multiple comparisons, see Fig. 3). The result for the FQ of Z suggested there is significant differences between only the listeners and the next speaker in turn-taking. The results for the MO of X and Z suggested there are significant differences in all the pairs of conditions. These reveal the following information.

- The MO and AM of the X, Y, Z, azimuth, elevation, and roll of listeners and the next speaker in turn-taking are larger than those of listeners in turn-keeping. In contrast, the FQ of X, Y, Z, azimuth, elevation, and roll of listeners and the next speaker in turn-taking are less than those of listeners in turn-keeping.
- The MO of X and Z of the next speaker in turn-taking is larger than those of listeners in turn-taking. In contrast, the FQ of Z of the next speaker in turn-taking is smaller than those of listeners in turn-taking.

[1] $t(1052) = 2.46$, $p < .05$ for MO of X; $t(1052) = 2.23$, $p < .05$ for MO of Y; $t(1052) = 2.19$, $p < .05$ for MO of Z; $t(1052) = 1.61$, $p < .10$ for MO of roll; $t(1052) = 4.74$, $p < .10$ for AM of Y; $t(1052) = 1.77$, $p < .10$ for AM of Z; $t(1052) = 1.82$, $p < .10$ for AM of roll; $t(1052) = 2.17$, $p < .05$ for FQ of Y; $t(1052) = 1.65$, $p < .10$ for FQ of elevation

[2] $F(2, 3159) = 16.8$, $p < .01$ for MO of X; $F(2, 3159) = 18.3$, $p < .01$ for MO of Y; $F(2, 3159) = 19.3$, $p < .01$ for MO of Z; $F(2, 3159) =??$, $p < .01$ for MO of azimuth; $F(2, 3159) = 8.5$, $p < .01$ for MO of elevation; $F(2, 3159) = 9.9$, $p < .01$ for MO of roll; $F(2, 3159) = 31.6$, $p < .01$ for AM of X; $F(2, 3159) = 19.1$, $p < .01$ for AM of Y; $F(2, 3159) = 32.2$, $p < .01$ for AM of Z; $F(2, 3159) = 6.4$, $p < .01$ for AM of azimuth; $F(2, 3159) = 15.9$, $p < .01$ for AM of elevation; $F(2, 3159) = 29.7$, $p < .01$ for AM of roll; $F(2, 3159) = 12.8$, $p < .01$ for FQ of X; $F(2, 3159) = 20.8$, $p < .01$ for FQ of Y; $F(2, 3159) = 3.8$, $p < .10$ for FQ of Z; $F(2, 3159) = 201.1$, $p < .01$ for FQ of azimuth; $F(2, 3159) = 12.6$, $p < .01$ for FQ of elevation; $F(2, 3159) = 8.3$, $p < .01$ for FQ of roll

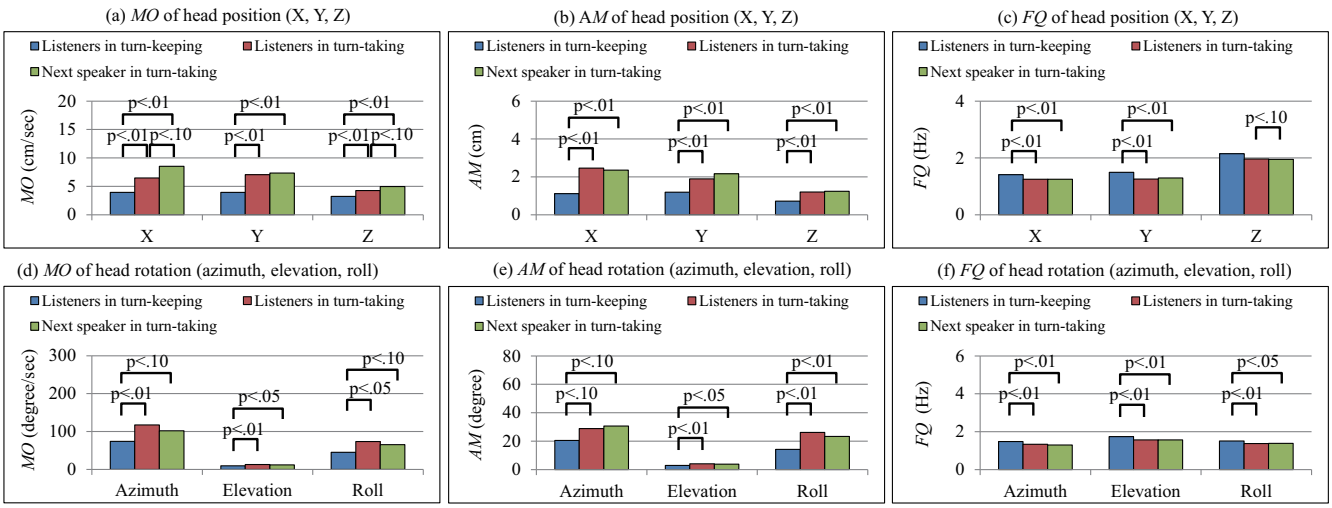

Fig. 3. Results of analysis of MO, AM, and FQ of listeners' head positions (X, Y, Z) and rotations (azimuth, elevation, roll).

5. PREDICTION MODEL

The analyses results discussed in the previous sections showed that the speaker's, next speaker's, and listeners' head movements may be useful as predictors of the next speaker in multi-party meetings. In this section, we used the speaker's and listeners' head movements as the variables and created a prediction model that feature two processing steps to predict whether turn-taking or turn-keeping will occur and who will be the next speaker in turn-taking.

5.1. Prediction of turn-keeping/turn-taking

We constructed our prediction model based on a support vector machine (SVM), in which the method used is SMO [21] implemented in the Weka data mining tool [22], and evaluated the accuracy of the model to investigate the effectiveness of the speaker's and listeners' head movements near the end of speaking for the prediction of whether turn-keeping or turn-taking occurs. The data used in the SVM contained the turn-keeping and turn-taking as a class. As the features, we used the MO of X, Y, Z, and roll, the AM of Y, Z, and roll, and the FQ of Y and elevation of the speaker and every head parameter except Z of FQ of listeners, which is different between turn-keeping and turn-taking in as described in subsections 4.1 and 4.2. We used the 10-fold cross validation of 296 data, which includes 148 data that were obtained by sampling from the 906 data in turn-keeping to remove the bias of the number of data and the 148 data in turn-taking used in the analysis in section 4. The model was 76.2% accurate. This suggests that the parameters of the speaker's and listeners' head movements near the end of speaking contribute to predicting whether turn-keeping or turn-taking occurs.

5.2. Prediction of next speaker in turn-taking

We constructed a prediction model based on the SVM as previously explained and evaluated the model's performance to investigate the effectiveness of the three listeners' head movements before the next utterance for the prediction of the next speaker in turn-taking. The data used in the SVM contained the next speaker as a class and the MO of azimuth, the FQ of roll, and the AM of Z of the three listeners, which are different between the listeners and next speaker in the turn-taking as described in subsection 4.2. We used 10-fold cross validation on the 148 turn-taking data. The prediction model was 55.2% accurate. The chance level was 33.3% because there are three next-speaker candidates in turn-taking. This suggests that the listeners' head movements contribute to predicting the next speaker in turn-taking.

6. CONCLUSION AND FUTURE WORK

Our analysis revealed that there are differences in the amount, amplitude, and frequency of of movement of the head position and rotation of the speaker between turn-keeping and turn-taking. The results also revealed differences in the amount of movement, amplitude, and frequency of the head position and rotation movement between listeners in turn-keeping, listeners in turn-taking, and the next speaker in turn-taking. On the basis of these results, we created prediction models featuring two processing steps to predict whether turn-taking or turn-keeping will occur and who will be the next speaker in turn-taking using the speaker's and listeners' head movement information. The evaluation results for the models suggest that the parameters of the speaker's and listeners' head movements near the end of speaking contribute to predicting the next speaker. As head movement can be readily measured by a camera or depth sensor, such as Kinect, the head movement is very useful for constructing a system that can predict the next speaker. In future work, we will create a prediction model for the start timing of the next speaking using head movement. Moreover, we plan to create a robust and high-performance prediction model using multimodal information, such as the gaze behavior.

7. REFERENCES

[1] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Masafumi Matsuda, and Junji Yamato, "Predicting next speaker and timing from gaze transition patterns in multi-party meetings," in *ICMI*, 2013, pp. 79–86.

[2] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato, "Analysis and modeling of next speaking start timing based on gaze behavior in multi-party meetings," in *ICASSP*, 2014, pp. 694–698.

[3] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato, "Analysis of respiration for prediction of "who will be next speaker and when?" in multi-party meetings," in *ICMI*, 2014, pp. 18–25.

[4] Starkey Duncan Jr and George Niederehe, "On signalling that it's your turn to speak," *J. Experimental Social Psychology*, vol. 10, pp. 234–247, 1974.

[5] Starkey Duncan Jr and Donald W. Fiske, *Face-to-face interaction: research, methods and theory*, Hillsdale, New Jersy: lawrence Erlbaum, 1977.

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

[7] 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.

[8] Lei Chen and Mary P. Harper, "Multimodal floor control shift detection," in *ICMI*, 2009, pp. 15–22.

[9] Iwan de Kok and Dirk Heylen, "Multimodal end-of-turn prediction in multi-party meetings," in *ICMI*, 2009, pp. 91–98.

[10] Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke, "Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody in human-computer dialog," in *ICSLP*, 2002, vol. 3, pp. 2061–2064.

[11] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto, "Gaze and turn-taking behavior in casual conversational interactions," *J. TiiS*, vol. 3, no. 2, pp. 12, 2013.

[12] Kornel Laskowski, Jens Edlund, and Mattias Heldner, "A single-port non-parametric model of turn-taking in multi-party conversation," in *ICASSP*, 2011, pp. 5600–5603.

[13] Gina-Anne Levow, "Turn-taking in mandarin dialogue: Interactions of tones and intonation," in *SIGHAN*, 2005.

[14] David Schlangen, "From reaction to prediction experiments with computational models of turn-taking," in *ISCA*, 2006, pp. 17–21.

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

[16] Tatsuya Kawahara, Takuma Iwatate, and Katsuya Takanashii, "Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations," in *ISCA*, 2012.

[17] Rutger Rienks, Ponald Poppe, and Dirk Heylen, "Differences in head orientation behavior for speakers and listeners: An experiment in a virtual environment," *J. TAP*, vol. 7, no. 1(2), 2010.

[18] 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.

[19] Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt, "Addressee identification in face-to-face meetings," in *EACL*, 2006, pp. 169–176.

[20] POLHEMUS, "Fastrak," http://polhemus.com/motion-tracking/all-trackers/fastrak/.

[21] Sathiya S. Keerthi, Shirish K. Shevade, Chiru Bhattacharyya, and K. R. K., "Improvements to platt's smo algorithm for svm classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637–649, 2001.

[22] 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," *J. Machine Learning Research*, vol. 11, pp. 2533–2541, 2010.