Multimodal Technologies and Interaction

MDPI

Article

Prediction of Who Will Be Next Speaker and When Using Mouth-Opening Pattern in Multi-Party Conversation †

Ryo Ishii [1,*], Kazuhiro Otsuka [2], Shiro Kumano [2], Ryuichiro Higashinaka [1,2] and Junji Tomita [1]

[1] NTT Media Intelligence Laboratories, NTT Corporation, 1-1, Hikarinooka, Yokosuka-shi, Kanagawa 239-0847, Japan; ryuichiro.higashinaka.tp@hco.ntt.co.jp (R.H.); junji.tomita.xa@hco.ntt.co.jp (J.T.)

[2] NTT Communication Science Laboratories, NTT Corporation, 3-1, Morinosato Wakamiya, Atsugi-shi, Kanagawa 243-0198, Japan; kazuhiro.ootsuka.py@hco.ntt.co.jp (K.O.); shirou.kumano.yc@hco.ntt.co.jp (S.K.)

* Correspondence: ryo.ishii.ct@hco.ntt.co.jp; Tel.: +81-46-859-2673

† This paper is an extension of the content our past conference paper.

Received: 11 September 2019; Accepted: 15 October 2019; Published: 26 October 2019

Abstract: We investigated the mouth-opening transition pattern (MOTP), which represents the change of mouth-opening degree during the end of an utterance, and used it to predict the next speaker and utterance interval between the start time of the next speaker's utterance and the end time of the current speaker's utterance in a multi-party conversation. We first collected verbal and nonverbal data that include speech and the degree of mouth opening (closed, narrow-open, wide-open) of participants that were manually annotated in four-person conversation. A key finding of the MOTP analysis is that the current speaker often keeps her mouth narrow-open during turn-keeping and starts to close it after opening it narrowly or continues to open it widely during turn-changing. The next speaker often starts to open her mouth narrowly after closing it during turn-changing. Moreover, when the current speaker starts to close her mouth after opening it narrowly in turn-keeping, the utterance interval tends to be short. In contrast, when the current speaker and the listeners open their mouths narrowly after opening them narrowly and then widely, the utterance interval tends to be long. On the basis of these results, we implemented prediction models of the next-speaker and utterance interval using MOTPs. As a multimodal-feature fusion, we also implemented models using eye-gaze behavior, which is one of the most useful items of information for prediction of next-speaker and utterance interval according to our previous study, in addition to MOTPs. The evaluation result of the models suggests that the MOTPs of the current speaker and listeners are effective for predicting the next speaker and utterance interval in multi-party conversation. Our multimodal-feature fusion model using MOTPs and eye-gaze behavior is more useful for predicting the next speaker and utterance interval than using only one or the other.

Keywords: mouth movement; multi-party conversation; next-speaker prediction; utterance interval prediction; turn-changing; eye-gaze behavior

1. Introduction

People start to have face-to-face conversations with others immediately after they are born. Face-to-face communication is one of the most important activities when people build social relationships with others. Furthermore, multi-party face-to-face conversations involving multiple persons (three or more) are very important not only for information transmission or sharing and understanding other people's intentions or emotions but also for group decision making. If a computer can understand and predict how such multifaceted conversations are carried out smoothly, it should be possible to develop a system that supports smooth communication and can dialogue with the

participants. As a basic technology for making such a system a reality, research on Automatic Meeting analysis has been actively conducted in recent years [1,2].

Turn-changing, a situation where speakers change, is a particularly important aspect of smooth communication. If the turn-changing is not successful, for example, an utterance collision will occur or unintentional silence will occur. In a two-person conversation, one person becomes a speaker and the other a listener. Because of the simple exchange that alternates between the two roles, it is easy to perform turn-changing to some extent. On the other hand, in a multi-party conversation, there are multiple listeners—in other words, multiple candidates for the next speaker—which makes it difficult to perform turn-changing. Participants cognitively predict the appropriate timing of turn changes on the basis of verbal and nonverbal cues, as well as on the appropriateness of the next speaker in their multi-party conversation and their proper utterance start timing. In addition, participants can start speaking at their own appropriate timing. If a computational model can predict the next speaker and the utterance interval between the start time of the next speaker's utterance and the end time of the current speaker's utterance, the model will be an indispensable technology for facilitating conversations between humans and between humans and conversation agents or robots. For example, this technology will help conversation agents and robots start and end utterances at the right time when they participate in a multi-party conversation. In addition, this technology may make it possible to avoid speech collisions due to time delays in remote conversational systems such as video conference systems, where video and audio transmission delays often occur. Computers may also be able to use this technology to help or encourage participants to speak at the appropriate time in face-to-face conversations.

Against this background, there is a lot of research in the engineering field that predicts the various situations related to turn-changing, such as the absence of turn-changing, the next speaker during turn-changing and the utterance interval from verbal and nonverbal information in multi-party conversation. For automatically predicting the next speaker and utterance interval in multi-party conversations, techniques using eye-gaze behavior, head movement and the respiration information of participants have been proposed [3–7]. However, it is still difficult to accurately predict the next speaker and utterance interval. We suggest that a more robust and accurate prediction model can be developed by using multimodal information other than gaze movement, head movement, and respiration. To this end, we have been searching for a new modality that can be used to predict the next speaker and the utterance interval in multiparty meetings.

In order to clarify which new modalities would be useful for such predictions, we focused on the movement of participants when they open their mouths. So far, the relationship between the turn-keeping/changing situation and mouth open information in multi-party conversation analysis has not been studied. Here, we introduce a mouth-opening transition pattern (MOTP) that indicates the transition of the degree of opening the mouth near the end of the utterance as a feature value for prediction. It is well known that the movement of mouth-opening and the start and end of an utterance are closely related [8]. In order for a person to speak, he/she has to move his/her mouth immediately before and during speaking. Then, when the utterance ends, his/her mouth will naturally close. We suggest that the movement of opening the mouth immediately after the current speaker stops speaking and just before the next speaker starts speaking can be a clue to predict who the next speaker will be. Respiration information, for example, the large breaths needed to start or continue speaking, can also be very helpful in predicting the next speaker and utterance interval [6,7]. We expect large breaths to be strongly associated with opening mouth movements. Thus, opening the mouth immediately after the current speaker stops speaking is a clue to predict the next speaker and utterance interval. On the basis of this idea, we focus on using MOTPs to predict the next speaker and utterance interval in multi-party conversation.

First, we collected data on verbal and nonverbal behaviors including utterances and the degree of mouth opening (closed, narrow-open and wide-open) from participants in a four-person conversation, which were then manually annotated. We analyzed the relationships between MOTPs

and turn-keeping/turn-changing, next speaker in turn-changing and utterance interval with the collected data. The analysis results revealed for the first time that there are various relationships between MOTPs and these behaviors. Typical results from the MOTP analysis show that the current speaker often continues to open her mouth narrowly during turn-keeping and to keep her mouth open widely and start to close her mouth after opening it narrowly during turn-changing. The next speaker often starts to open her mouth narrowly after closing it during turn-changing. Moreover, when the current speaker closes her mouth after opening it narrowly during turn-keeping, the utterance interval tends to be short. In contrast, when the current speaker and listeners open their mouths narrowly after opening them narrowly and then widely, the utterance interval tends to be long. On the basis of these results, we implemented the prediction models of next speaker and utterance interval using the MOTPs. Our evaluation of the models suggests that the current speaker's and listeners' MOTPs are effective for predicting the next speaker and utterance interval in multi-party conversations. As a multimodal-feature fusion, we implemented models using eye-gaze behavior in the form of gaze transition patterns (GTPs) in addition to MOTPs. A GTP is expressed as an n-gram, which is defined as a sequence of eye-gaze targets. Each element of a GTP contains eye-gaze target information. As eye-gaze targets, we defined a person or object classified as "current speaker", "listener", or "non-person". We considered whether mutual eye-gaze occurred and classified eye-gaze behavior when the person looked at the "current speaker" and "listener". GTPs are very useful for predicting the next speaker and utterance interval [3–5]. The evaluation results of the implemented models suggest that our multimodal-feature fusion using both MOTPs and GTPs is more useful for such prediction than non-fusion using only one or the other.

In Section 2 of this paper, we review relevant related work and highlight our new approach that utilizes mouth-opening movement. Section 3 describes the corpus data collection in multi-party conversations. Sections 4 and 5 describe our analysis of MOTP and the next speaker and utterance interval. Sections 6 describes the implementation and evaluation of the prediction models and Section 7 discusses the analysis results and multimodal-feature fusion. We conclude in Section 8 with a brief summary and mention of future work.

2. Related Work

2.1. Prediction of Next Speaker and Utterance Interval

The elucidation of the mechanism for giving and receiving conversation turns was initiated mainly in the field of sociolinguistics. Sacks et al. [9] proposed a turn-change model, arguing that turn-changing occurs only at transition-related points (TRP) near the end of an utterance. Kendon [10] analyzed conversations and discovered that verbal and nonverbal behaviors, such as eye-gaze behaviors, contributed to smooth turn-keeping and changing and the adjustment of the utterance interval.

Several studies in cognitive science have clarified the verbal and nonverbal cues for a person to perceive the presence or absence of turn-changing in two-person conversations [11,12]. For multi-party conversations, several studies have recently examined which non-verbal information of the participants is used for turn-changing. It has been shown that eye-gaze behavior [5,13–17], eye-blink [18], head movement [4,19], respiration [6,7] and hand gestures [20] are related to turn-changing.

Several studies have utilized the automatic detection model to determine whether turn-changing takes place in multi-party conversation by using speech processing techniques [14,21–25] and nonverbal behaviors such as eye-gaze behavior [14,21,22] and physical visual motion from image processing [21,22,26] near the end of a current speaker's utterance. However, these studies only estimate if the current speaker will continue speaking. For this reason, such estimation techniques are often referred to as "end-of-turn estimation".

In addition to automatic detection of turn-changing, several studies have attempted to predict the next speaker—namely, who will become the next speaker—during turn-changing and utterance intervals between the start time of the next speaker's utterance and the end time of the current speaker's utterance. A next-speaker detection model using eye-gaze behavior and head-nods has been proposed in three-person conversations with a shared poster [13]. This model deals with one poster presenter and two listeners of a conversation. All participants are paying attention to the poster for a long time and the distribution of eye-gaze behavior is unique compared to communication without a specific eye-gaze target.

We previously demonstrated which nonverbal cues are related to turn-changing, determining the next speaker during turn-changing and utterance intervals in multi-party conversation [5–7,15,19]. We also proposed a prediction model that features three processing steps to predict whether turn-keeping or turn-changing will occur, who the next speaker will be after turn-changing and the length of the utterance interval by using GTPs, which contain the n-gram information of eye-gaze targets including mutual eye-gaze information when the person looks at another person [5,15]. In previous studies by other researchers (as described above), the verbal and nonverbal behaviors right before the start time of the next speaker's utterance were used to estimate/detect the turn-keeping/changing and the next speaker during turn-changing. Compared to these, our approach has an advantage in that it can predict the next speaker and the utterance interval in advance from the information at the end time of the current speaker's utterance. In other words, our prediction technology can predict the next speaker and utterance interval about one or two seconds before the start of the next utterance. This allows conversation agents and conversation support systems to engage participants before the start of the next utterance on the basis of the prediction result of future utterance situations.

GTPs are useful feature values for predicting the next speaker and utterance interval. We also explored respiration accompanying speaking for the first time in multi-party conversation analysis and demonstrated the various strong relationships between respiration and the next speaker and utterance interval [6,7]. As a representative finding, we demonstrated that a current speaker often inhales quickly and rapidly right after the end time of his or her own utterance during turn-keeping. The listener who will become the next speaker often takes a deeper breath before speaking during turn-changing than the listeners who will not become the next speaker. The results indicated that the characteristics of the current speaker's inhalation right after her utterance and the listeners' inhalation near the end of the utterance may be useful for predicting the next speaker and utterance interval in multi-party conversation. We also implemented prediction models of next speaker and utterance interval and demonstrated that the respiration characteristics of the current speaker and listeners are useful for the prediction [6,7].

We also demonstrated that the head movement of participants extracted using a six-degrees-of-freedom head tracker near the end of an utterance has strong relationships with the next speaker and utterance interval [4,19]. Specifically, in the relationships between the head movements and the next speaker, we found that there are big differences in the amounts, amplitude and frequency of the head position and rotation movements of the current speaker near the end of the utterance between turn-keeping and changing. We also demonstrated that there are big differences in the amounts, amplitude and frequency of head position and rotation movements between the listeners who will not be the next speaker in turn-keeping and turn-changing and the listener who will be the next speaker during turn-changing [4]. We also demonstrated that the head movement of the current speaker and listeners is useful for predicting the next speaker in multi-party conversation.

Moreover, we demonstrated that there are strong relationships between the utterance interval between the start time of the next speaker's utterance and the end time of the current speaker's utterance and eye-gaze, head movement and respiration of participants during turn-keeping and changing and proposed prediction models of utterance interval using their nonverbal information [5,7,19]. The utterance interval is crucial in terms of correctly conveying intention and emotion to conversational partners and conducting a smooth conversation. The utterance interval

varies greatly depending on the utterance situation, content, intention and emotion. If the timing for starting an utterance, that is, the utterance interval, is not appropriate, there is a risk that an unintended message will be sent to the conversation partner, which may adversely affect communication [27]. In this context, if we take video conferencing systems as an example, even a short video and audio delay of about 500 ms can cause utterance collisions that will interfere with smooth turn-changing and adversely affect speech impressions [28].

As discussed above, in previous work, we clarified the relationship between the nonverbal behaviors such as eye-gaze, respiration and head movement of participants and the next speaker and utterance interval in multi-party conversation and developed prediction models of next speaker and utterance interval. However, the accuracy of the prediction models is still not sufficient. In order to build a prediction model with higher performance in the future, our research effort focusing on new modalities is very meaningful.

2.2. Mouth-Opening Movement and Speaking

It is generally known that mouth-opening movement and utterances have a very close relationship. We have to move our mouths to utter speech sounds during speaking. To develop a technology that makes use of such a relationship, there have been many studies on the utilization of lip reading in order to achieve highly accurate speech recognition in noisy environments [29].

Mouth or lip movement information is also used to improve the accuracy of active speaker detection (ASD). We briefly introduce previous research that takes such an approach. To improve the accuracy of ASD, attention is focused on utilizing visual information as well as audio information [30–32]. To focus on lip information, Cutler et al. used the image feature of the mouth region according to the audio information [33]. Haider et al. used head movements in addition to lip information in speech [34]. They also showed that the features of lip and head movement one second before the start of speech are useful for improving the function of ASD [35]. The lip information they used is the mean and standard deviation of the inner height, outer height and width of the lip. Similar to our approach, Murai et al. have shown a basic idea to deduce the next speaker from mouth movements [36]. However, they have not specifically examined which types of mouth movements are useful as feature quantities. In these studies, it was shown that lip information in the speech section and time section immediately before speech is useful for improving the performance of ASD. These previous research results [30–34,36] support the validity of our approach for predicting the next speaker and utterance interval using the mouth opening pattern at the end of an utterance.

We anticipated that opening the mouth just before the next speaker starts speaking would help predict who the next speaker would be. In recent years, the development of computer vision technology for measuring mouth movements with a single camera has progressed [37]. Thus, using these techniques, it has become practical to use mouth-opening movements to predict the next speaker and utterance interval in a conversational situation. However, there has been no research on how to analyze mouth-opening movements at the end of an utterance or how to predict the next speaker and utterance interval in a multi-party conversation. Any technology that can predict the next speaker and speech interval from mouth-opening movements will thus be of enormous practical value.

As previously discussed, participants' respiration behavior, such as taking a big breath before starting or continuing to speak, is very useful in predicting the next speaker and utterance interval [6,7]. Since deep breaths are assumed to be strongly related to the movement of opening the mouth, if we replace the movement of opening the mouth just before the start of the next utterance with respiration information, this may be useful in predicting the next speaker and utterance interval.

Our current work is the first attempt to use mouth-opening movement to predict the next speaker and utterance interval in multi-party conversation. We examine how mouth-opening movement can help with such predictions and how multimodal feature fusion using eye-gaze behavior [3,5,15] in addition to mouth-opening movement is useful. We also discuss the relationship between mouth-opening movement and respiration.

3. Corpus Data of Multi-Party Conversation

In this study, we collected data on mouth-opening movements and utterances obtained from actual multi-party conversations and used them for analysis and construction of prediction models. We recorded eight sessions of face-to-face four-person conversations. The participants were 16 individuals in their 20s and 30s who had a first-to-face relationship with each other. They were divided into four groups consisting of four people each. The four participants were seated in chairs arranged at regular intervals in a semicircle, as shown in Figure 1. We labeled the participants from left to right as Person 1, Person 2, Person 3 and Person 4. They argued and gave opinions in response to highly divisive questions such as "Do marriage and romance always go together?" and "Is it possible to strengthen social security even if taxes increase?" and needed to draw a conclusion within ten minutes. The each group performed two conversation sessions. Although the participants were given such a theme, they were allowed to speak freely.

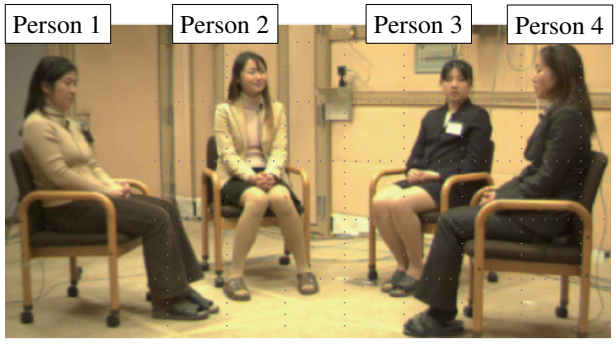

Figure 1. One scene of four-person conversation conducted for data collecting. Participants are Person 1, Person 2, Person 3 and Person 4 from the left.

We recorded the speech of each participant in the conversation with a pin microphone. We also recorded video of each participant and the group as whole with multiple cameras. The recorded audio and video were synchronized with a time resolution of 30 Hz. The total duration of the recorded conversation was 80 min. The following verbal and nonverbal information was extracted from these video and audio resources.

- Utterance: We designed the utterance unit on the basis of the inter-pausal unit (IPU) [38]. Specifically, the start and end of each utterance was denoted as an IPU. When a silence interval of 200 ms or more occurred, the utterance was separated. Therefore, if an utterance was made after a silent period of 200 ms or less, it was determined that one utterance was continued. Since this IPU can determine the start and end of an utterance using only the duration of the silent section, it is very convenient when performing real-time utterance section detection. We excluded back-channels without specific utterance content from the extracted IPUs. Later, we considered the same person's continued IPU as one utterance turn. We considered that IPU pairs by the same person in the temporally adjacent IPU pairs are turn-keeping and IPU pairs by different persons are turn-changing. The total number of IPU pairs was 904 for turn-keeping and 126 for turn-changing. Our analysis excluded overlap of utterances, that is, when the listener interrupted while the current speaker was speaking or when two or more participants spoke at the same time. Specifically, data was excluded when the IPU pair utterance interval was less than 200 ms. All told, there were 904 IPU pairs during turn-keeping and 105 IPU pairs during turn-changing.

- Degrees of mouth-opening: We define the three degrees of mouth-opening as follows.
  - Closed (*X*): The mouth is closed.
  - Narrow-open (*N*): The mouth opening is narrower than one front tooth.
  - Wide-open (*W*): The mouth opening is wider than one front tooth.
  - Unknown (*U*): A state where the opening of the mouth cannot be determined. For example, a participant has covered her mouth with her hand.

  Figure 2 shows an example of the *X*, *N* and *W*. This classification of degree of mouth-opening is a simplified version of the AU 25 (lip part), AU 26 (chin lowering) and AU 27 (mouth stretch) of the Facial Action Coding System (FACS), which is a comprehensive, anatomically based system for describing all visually discernible facial movements [39]. We focused on lip opening and ensured quantitative annotations even in non-frontal views. A skilled annotator carefully observed the person's face image and manually annotated the degree of mouth-opening for each time frame.
- Eye-gaze target: The eye-gaze target was annotated manually using bust/head and overhead views in each frame of the videos by a skilled annotator. The eye-gaze objects were the four participants (Persons 1, 2, 3 and 4) and non-persons. Three annotators labeled the eye-gaze targets. We calculated the Conger's kappa coefficient [40] to evaluate an inter-coder agreement of the three annotators. The result was .887, which is a quite high value indicating the high reliability of the annotated eye-gaze targets. We used the eye-gaze targets annotated by one annotator who had the most experience.

Figure 2. Sample images showing the degree of mouth-opening. From the left, these are images when mouth-opening is closed (*X*), narrow-open (*N*) and wide-open (*W*).

All the above verbal and nonverbal data were integrated at 30 Hz for visual display at the same time resolution using the NTT Multimodal Viewer [41]. A sample image of the viewer is shown in Figure 3. With the viewer, we can see each piece of multimodal data on the timeline and intuitively observe the data. In this example, we focus on the 16094th frame. Person 3 is speaking at this time. The mouth-openings of Persons 1, 2, 3 and 4 are *N*, *X*, *N* and *W*, respectively and the eye-gaze targets for Persons 1, 2, 3 and 4 are Person 2, non-person, non-person and Person 2, respectively.

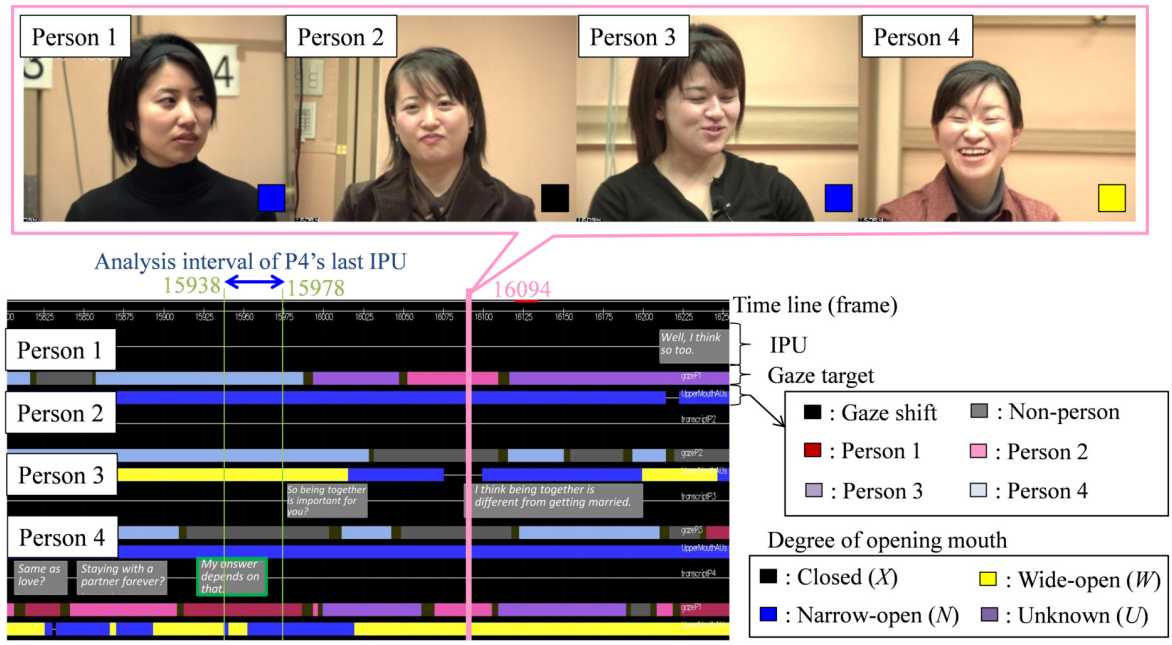

Figure 3. Sample image of displayed verbal and nonverbal data with NTT Multimodal Viewer [41]. The top shows the face image of each participant in the 16094th frame. At the bottom is the verbal and non-verbal behavior of each participant. Upper gray boxes show inter-pausal units (IPUs), middle boxes show eye-gaze targets and lower boxes show the degrees of mouth-opening for each person along the timeline.

4. Analysis of Next Speaker and MOTPs

4.1. Analysis Method

Using the collected corpus data, we analyze the relationship between the next speaker and mouth-opening movement. As analysis method, we first analyzed how the current speaker's mouth-opening movement differs quantitatively between turn-keeping and turn-changing. If the analysis results reveal a difference between the current speaker's mouth-opening movement immediately after the end of the utterance between turn-keeping and turn-changing, mouth-opening movement can be used effectively to predict turn-keeping and turn-changing in multi-party conversation. We analyzed, similarly, how the mouth-opening movement differs quantitatively between the listener who becomes the next speaker during turn-changing (hereinafter referred to as "next speaker") and the listener who does not become the next speaker during turn-keeping and turn-changing (hereinafter referred to as "listener"). If there are big differences here, they are also effective in predicting the next speaker during turn-changing, in addition to predicting the turn-keeping and turn-changing.

We assumed that the MOTP of the current speaker at the end of the IPU would be different between turn-keeping and turn-changing. For example, a current speaker might keep her mouth open to continue speaking at the end of the IPU during turn-keeping, while in contrast, she may close her mouth gradually near the end of the IPU and then keep it closed because she does not need to speak during turn-changing. Similarly, listeners who do not become the next speaker may not move their mouth very much because they do not speak during turn-keeping and turn-changing. In contrast, the next speaker may begin to open or open wider after opening narrowly because she needs to start speaking during turn-changing. In this way, we assumed that the MOTP at the end of the IPU would differ depending on if it was the beginning or the end of the utterance.

Next, we used our past research results to determine the time interval to be analyzed. Our previous studies [3–7] have demonstrated that the eye-gaze behavior, head movement and respiration that occur during the short interval between 1000 or 2000 ms before the end of a current speaker's utterance

and 200 ms after it is useful for predicting the next speaker and utterance interval. We therefore decided to analyze mouth-opening movement in a similar way: between 1000 ms before the end of the IPU and 200 ms after the end of an utterance. In analyzing the parameters of mouth-opening movement, we defined the MOTP, namely, the n-gram information and consider its elements as the degree of mouth-opening (*N*, *S* and *W*). Information on changes in the state of non-verbal behavior using n-grams has also proven useful when dealing with changes in gaze targets [4–6].

In the present work, we examine which kind of MOTP is generated using actual corpus data. For example, considering the mouth-opening movements associated with Person 4's last IPU, whose end time is the 15968th frame in Figure 3, the analysis interval of mouth-opening movements is 1200 ms (40 frames) between 1000 ms before the end of the IPU (15938th frame) and 200 ms after the end of the IPU (15978th frame). The MOTPs of Persons 1, 2, 3 and 4 are *N*, *W*, *N* and *N*-*W*-*N*, respectively.

The analysis results are presented from the next section onward.

4.2. Analysis of Current Speaker's MOTPs and Turn-keeping/Changing

We examined how often the current speaker's MOTPs occur during turn-keeping and turn-changing. The occurrence frequency of the current speaker's MOTPs during turn-keeping for 904 IPUs and turn-changing for 105 IPUs is shown in Figure 4. There are 50 kinds of the current speaker's MOTPs in all the IPUs. The occurrence probabilities of several kinds of IPUs are very small. We define the "Others" class to include 38 patterns that occurred in less than 1% of the data, as the data set is small. We determined which occurrence frequencies of the current speaker's MOTPs are significantly different between turn-keeping and turn-changing by means of a chi-squared test. The results indicate that the occurrence frequency of the current speaker's MOTPs differed significantly between the conditions during turn-keeping and turn-changing (chi-squared test result: $\chi^2 = 61.8$, $df = 11$, $p<0.01$).

Next, to verify which MOTPs differed between conditions, we conducted a residual analysis [42]. The results are shown in Figure 4, from which we deduce the following:

- The occurrence frequencies of *N*, *N*-*X*-*N* and *N*-*X*-*N*-*X*-*N* of the current speaker's MOTPs are significantly higher during turn-keeping than during turn-changing. That is, the occurrence frequency at which the current speaker continues to open her mouth narrowly (*N*), open it narrowly after closing it from opening it narrowly (*N*-*X*-*N*) and open it narrowly after closing it again from opening it narrowly (*N*-*X*-*N*-*X*-*N*) is higher during turn-keeping than during turn-changing. The *N* has the highest frequency, which is 0.47, of all the MOTPs whose occurrence frequency is significantly higher during turn-keeping than during turn-changing. There is a big difference in the occurrence frequency of *N* during turn-keeping compared to during turn-changing, which is approximately twice as high. The representative results indicate that the current speaker more often continues to open her mouth narrowly during turn-keeping than during turn-changing.
- The occurrence frequencies of *W*, *N*-*W*, *N*-*X* and *Others* of the current speaker's MOTPs are significantly higher during turn-changing than during turn-keeping. That is, the occurrence frequency at which the current speaker continues to open her mouth widely (*W*), opens widely from narrowly (*N*-*W*), closes it from opening it narrowly (*N*-*X*) and performs a mouth-opening pattern with the occurrence frequency of 1% or less (*Others*) is higher during turn-changing than during turn-keeping. The *W* and *N*-*W* have the highest frequencies, 0.14 and 0.08, respectively, of all the MOTPs excepting *Others*, whose occurrence frequency is significantly higher during turn-changing than during turn-keeping. There is a big difference in the occurrence frequency of *W* and *N*-*W* during turn-changing compared to during turn-keeping, which is approximately five times and three times as high, respectively. The representative results indicate that the current speaker more often continues to open her mouth widely or starts to close it from opening it narrowly during turn-changing than during turn-keeping.

These results suggest that a current speaker's MOTPs could be useful as a predictor of turn-keeping and turn-changing.

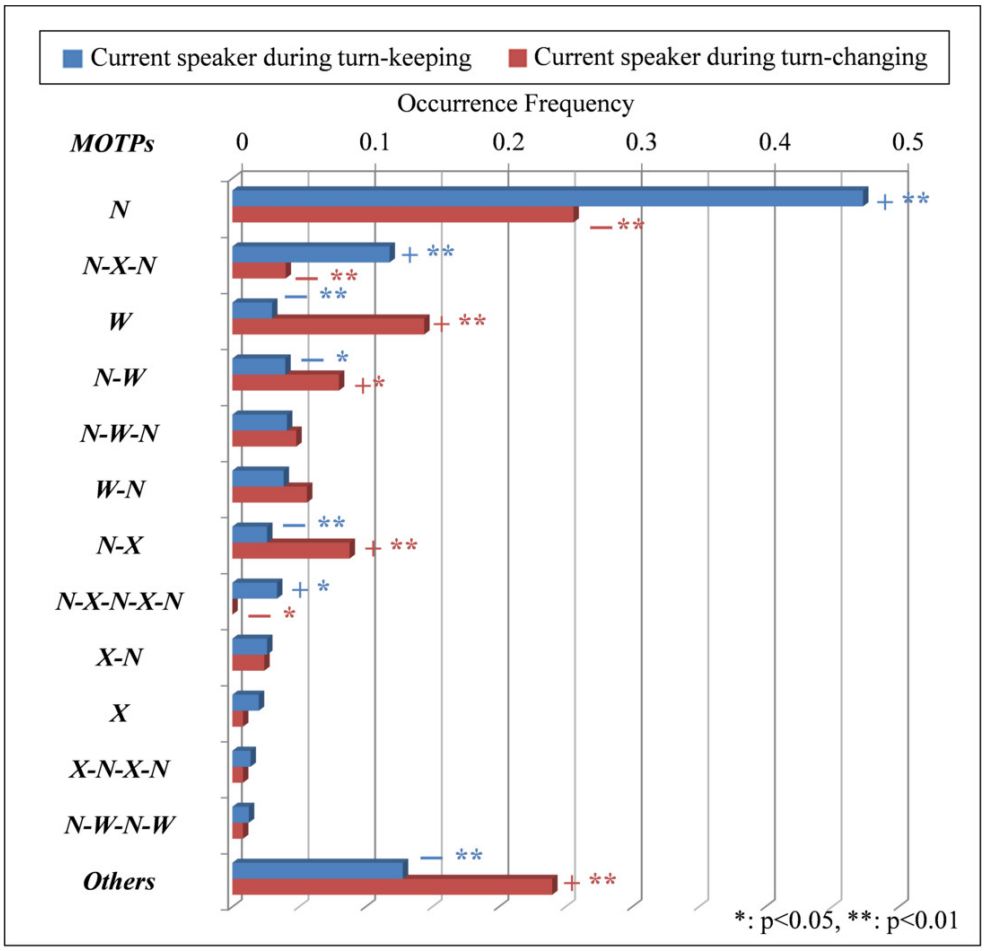

Figure 4. Occurrence frequency of current speaker's mouth opening transition patterns (MOTPs) during turn-keeping and turn-changing. Vertical axis indicates the type of MOTP. Numerical value on the horizontal axis indicates the occurrence frequency of each MOTP.

4.3. Analysis of Listeners' MOTPs and Next Speaker

We examined how often the listeners' MOTPs occur during turn-keeping and turn-changing. The occurrence frequency of the listeners' MOTPs during turn-keeping for 904 IPUs and turn-changing for 105 IPUs is shown in Figure 5. There are 55 kinds of listeners' MOTPs in all the IPUs. We define the "Others" class to include 45 patterns that occurred in less than 1% of the data, as the data set is small. We determined which occurrence frequencies of listeners' MOTPs are significantly different between turn-keeping and turn-changing by means of a chi-squared test. The results indicate that the occurrence frequency of the MOTPs differed significantly between the conditions of the listeners during turn-keeping and turn-changing and the next speaker during turn-changing (chi-squared test result: $\chi^2 = 155.8$, $df = 20$, $p<0.01$). Next, to verify which MOTPs differed between conditions, we conducted a residual analysis [42]. The results are shown in Figure 5, from which we deduce the following:

- The occurrence frequencies of *X* of the MOTPs of listeners during turn-keeping are significantly higher than those of listeners and the next speaker during turn-changing. That is, the occurrence frequency at which listeners during turn-keeping continue to close their mouths (*X*) is high. There is a big difference in the occurrence frequency of *X* of listeners in turn-keeping, 0.45, compared to listeners and the next speaker during turn-changing, which are approximately twice

and 15 times as high. In contrast, the occurrence frequencies of *X-N*, *W*, *W-N*, *N-W-N*, *X-N-X* and *Others* of the MOTPs of listeners during turn-keeping are significantly lower than those of the listeners and the next speaker during turn-changing. That is, the occurrence frequencies at which the listeners during turn-keeping start opening their mouths narrowly after closing them (*X-N*), continue to open them widely (*W*), start opening them narrowly after opening them widely (*W-N*), start opening them narrowly after opening them narrowly and then widely (*N-W-N*), start closing them after closing them and opening them narrowly (*X-N-X*) and perform a mouth-opening pattern with the occurrence frequency of 1% or less (*Others*) are low.

The representative results indicate that the listeners during turn-keeping more often continue to close their mouth than the listeners and the next speaker during turn-changing.

- The occurrence frequencies of *W*, *W-N*, *N-W-N* and *Others* of the MOTPs of listeners during turn-changing are significantly higher than those of listeners during turn-keeping. That is, the occurrence frequency at which listeners during turn-changing continue to open their mouths widely (*W*), start to open them narrowly from widely (*W-N*), start to open them narrowly after opening them narrowly and then widely (*N-W-N*) and perform a mouth-opening pattern with the occurrence frequency of 1% or less (*Others*) is higher than that of listeners during turn-keeping. The *W* has the highest frequency, about 0.12, among the MOTPs of the listeners during turn-changing. There is a big difference in the occurrence frequency of *W* of listeners in turn-changing compared to listeners during turn-keeping, which is approximately four times as high. In contrast, the occurrence frequency of *X* of the MOTPs of listeners during turn-changing is significantly lower than that of listeners during turn-keeping. That is, the occurrence frequency at which listeners during turn-changing continue to close their mouths (*X*) is lower than that of listeners during turn-keeping.

  The representative results indicate that the listeners during turn-changing more often continue to open their mouth widely than the listeners during turn-keeping.
- The occurrence frequencies of *X-N*, *W-N*, *N-W*, *X-N-X* and *Others* of the MOTPs of the next speaker during turn-changing are significantly higher than those of listeners during turn-keeping and turn-changing. That is, the occurrence frequencies when the next speaker during turn-changing starts opening her mouth narrowly after closing it (*X-N*), opening it narrowly from widely (*W-N*) and vice versa (*N-W*), starts closing it after closing it and opening it narrowly (*X-N-X*) and performs a mouth-opening pattern with the occurrence frequency of 1% or less (*Others*) are higher than those of listeners during turn-keeping and turn-changing. The *X-N* has the highest frequency, about 0.24, among the MOTPs of the next speaker during turn-changing. There is a big difference in the occurrence frequency of *W* of the next speaker in turn-changing compared to listeners during turn-keeping and turn-changing, which are approximately four times as high. In contrast, the occurrence frequency of *X* of the next speaker during turn-changing is significantly lower than that of listeners during turn-keeping and turn-changing. That is, the occurrence frequency at which the next speaker during turn-changing continues to close her mouth (*X*) is lower than that of listeners during turn-keeping and turn-changing.

  The representative results indicate that the next speaker during turn-changing more often starts to open her mouth narrowly from closing it than listeners during turn-keeping and turn-changing.

Taken together, these results suggest that listeners' MOTPs are potentially useful as predictors of the next speaker during turn-changing in addition to turn-keeping and changing.

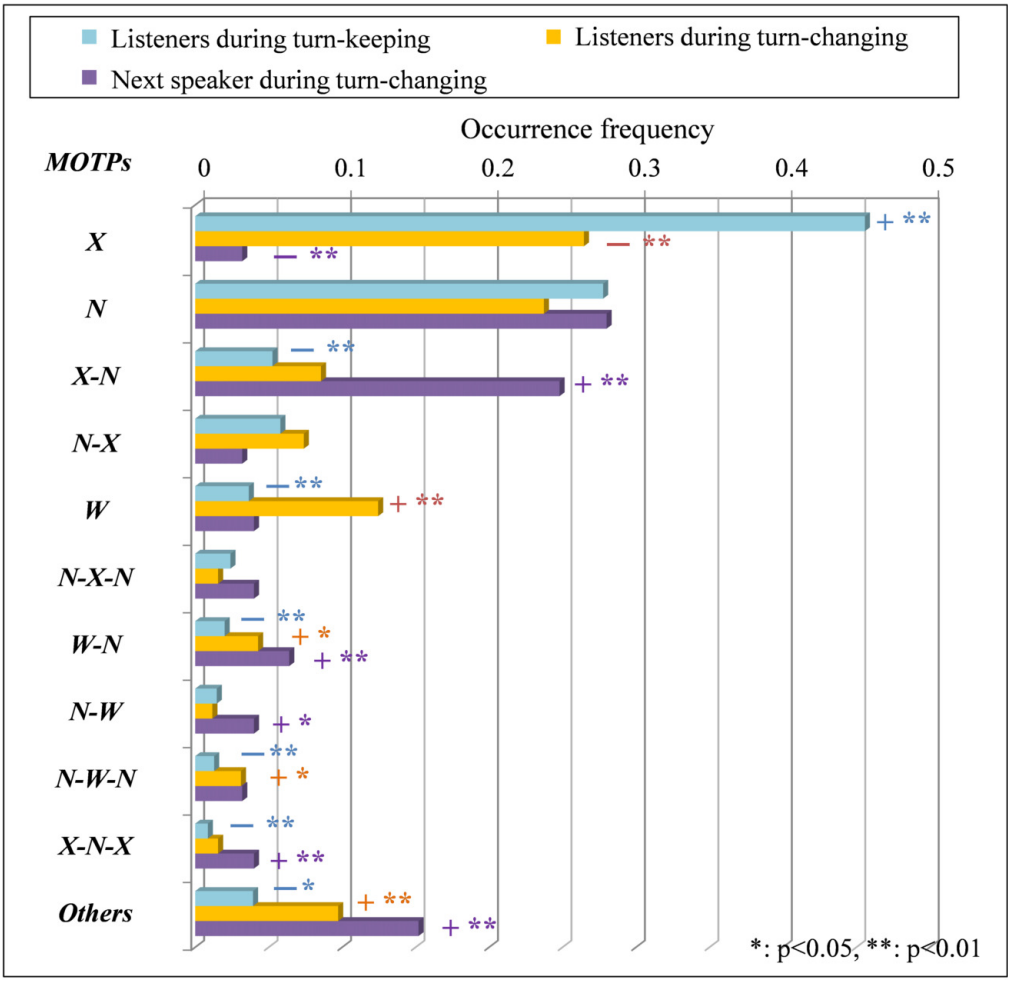

Figure 5. Occurrence frequency of MOTPs of listeners during turn-keeping and turn-changing and next speaker during turn-changing. Vertical axis indicates the type of MOTP. Numerical value on the horizontal axis indicates the occurrence frequency of each MOTP. Light blue bars show the data of listeners during turn-keeping and orange bars show the data of listeners during turn-changing. Purple bars show the data of next speaker during turn-changing.

5. Analysis of MOTPs and Utterance Interval

We analyzed the relationship between utterance interval, between the start time of the next speaker's utterance and the end time of the current speaker's utterance and the MOTPs of the current speaker and listeners. Utterance interval varies greatly depending on the difference between turn-keeping and turn-changing situations, so we analyzed the relationship between utterance interval and MOTPs during turn-keeping and turn-changing separately. Specifically, we examined the utterance interval when each MOTP of the current speaker and listeners appeared and then determined whether there were differences in the interval between the MOTPs. If there is a difference in utterance interval depending on the kinds of MOTPs, MOTPs might be useful to predict the utterance interval.

5.1. Analysis of MOTPs and Utterance Interval in Turn-keeping

Figures 6 and 7 show box plots of the utterance intervals for each MOTP of the current speaker and listeners during turn-keeping. The box shows the range from the first quartile to the third quartile and the line in the box shows the median value. The two whiskers in the boxplot respectively show the range from the minimum to the first quartile and from the third quartile to the maximum. Using one-way analysis of variance (ANOVA), we examined whether there is a difference in utterance interval among MOTPs in the current speaker and the listeners during turn-keeping. The independent variable

of ANOVA is the utterance interval. The dependent variable is the mean value of the utterance interval for each participant's MOTP. There were statistically significant differences in utterance interval according to the MOTPs of the current speaker and the listeners (ANOVA result: $F(12, 892) = 2.67$, $p < 0.01$ for current speaker in turn-keeping; $F(10, 2704) = 13.94$, $p < 0.01$ for listeners in turn-keeping). Therefore, the utterance interval differs according to the MOTP of the current speaker and that of the listeners in turn-keeping. For example, in Figure 6, when the current speaker's MOTP is *N-X*, the utterance interval is the shortest of all, with a median of 0.397 s. That is, when the current speaker opens her mouth widely from closing it, the utterance interval tends to be very short. On the other hand, when the current speaker's MOTP is *X-N-X-N*, the utterance interval is 0.588 s, which is the largest. In other words, the utterance interval tends to be largest when the current speaker opens her mouth narrowly twice, starting with the mouth closed.

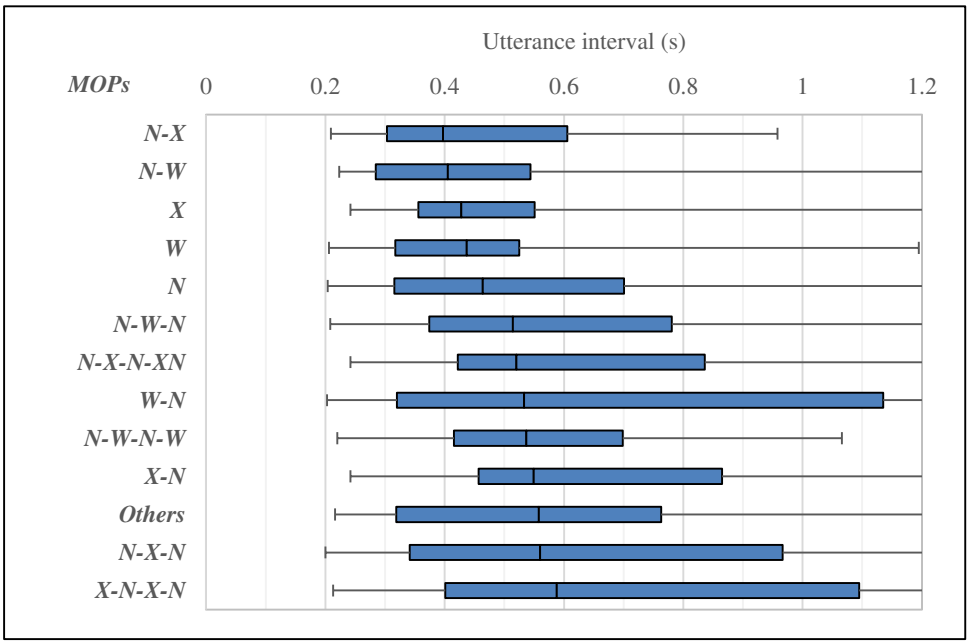

Figure 6. Box plot of utterance interval for each MOTP of current speaker during turn-keeping.

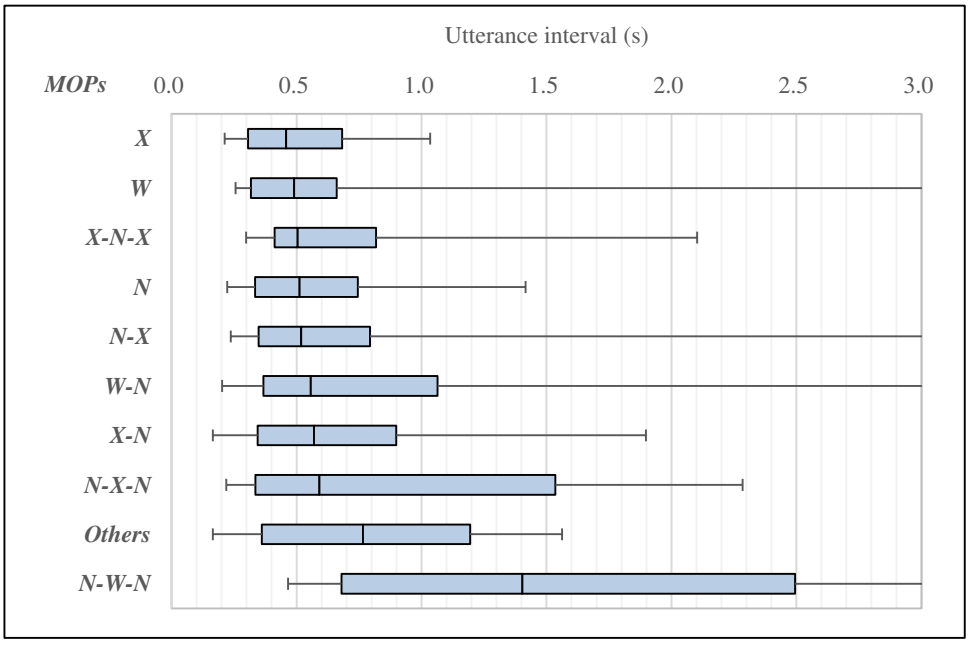

Figure 7. Box plot of utterance interval for each MOTP of listeners during turn-keeping.

In Figure 7, when the listener's MOTP is *X*, the utterance interval is the shortest, with a median of 0.437 s. That is, when the listener continues to close her mouth, the utterance interval tends to be very short. On the other hand, when the listener's MOTP is *N-W-N*, the utterance interval is 1.404 s, which is the largest. In other words, the utterance interval tends to be the largest when the listener opens her mouth narrowly, opens it widely and then opens it narrowly. Thus, during turn-keeping, the utterance interval tends to differ depending on the type of MOTPs of the current speaker and the listeners. This suggests the potential for MOTPs to predict the utterance interval effectively during turn-keeping.

5.2. Analysis of MOTPs and Utterance Interval in Turn-changing

As in the previous section, we analyzed the relationship between the utterance interval and MOTPs of the current speaker, listeners and the next speaker during turn-changing. From the results of ANOVA, there were statistically significant differences in utterance interval according to the MOTPs of only the current speaker and listeners (ANOVA result: $F(5,99) = 3.61$, $p < 0.01$ for current speaker in turn-changing; $F(7,202) = 3.35$, $p < 0.01$ for listener in turn-changing). However, there was no significantly difference in utterance interval according to the next speaker's MOTP (ANOVA result: $F(6,98) = 1.39$, $n.s$).

Figures 8 and 9 show box plots of the utterance intervals for each MOTP of the current speaker and listeners during turn-changing. We found that the utterance interval differs according to the MOTPs of the current speaker and listeners. For example, in Figure 8, when the current speaker's MOTP is *W*, the utterance interval is the shortest, with a median of 0.589 s. That is, when the current speaker opens her mouth widely, the utterance interval tends to be very short. On the other hand, when the current speaker's MOTP is *N-W-N*, the utterance interval is 1.832 s, which is the largest. In other words, the utterance interval tends to be the largest when the current speaker opens her mouth narrowly after opening it narrowly and then widely. When the listener's MOTP is *W*, as shown in Figure 9, the utterance interval is the shortest, with a median of 0.599 s. That is, when the listener opens her mouth widely, the utterance interval tends to be very short. On the other hand, when the listener's MOTP is *W-N*, the utterance interval is 2.048, which is the largest. In other words, the utterance interval tends to be the largest when the listener opens her mouth widely from opening it narrowly. Thus, during turn-changing, the utterance interval tends to differ depending on the type of MOTPs of the current speaker and the listener, suggesting that there is potential for MOTPs to predict the utterance interval effectively during turn-changing.

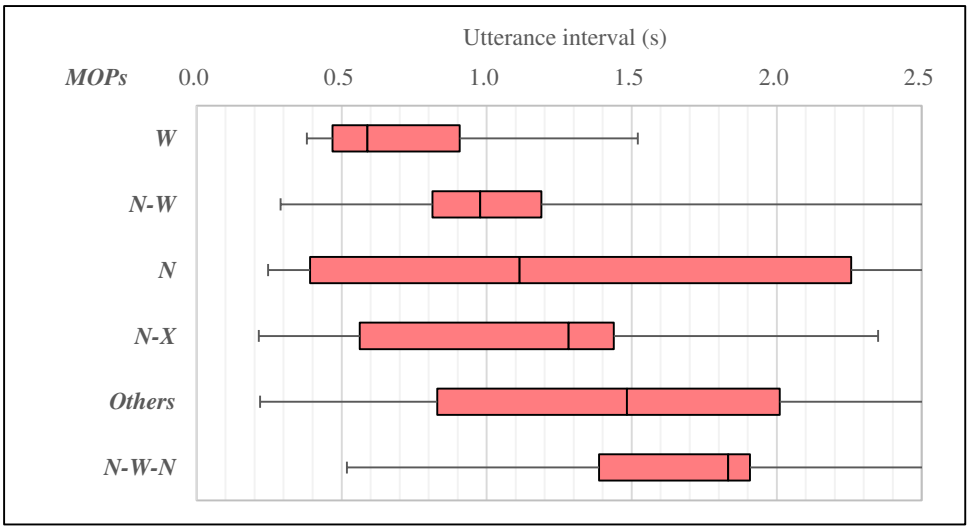

Figure 8. Box plot of utterance interval for each MOTP of current speaker during turn-changing.

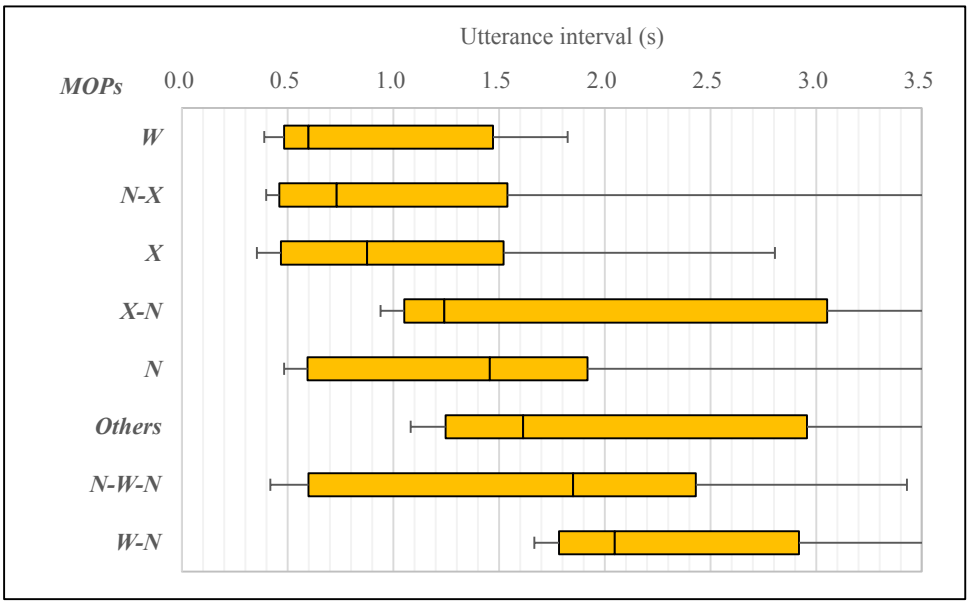

Figure 9. Box plot of utterance interval for each MOTP of listeners during turn-changing.

6. Prediction Models of Next Speaker and utterance Interval

6.1. Design Guidelines for Creating Prediction Models

The analysis results in Sections 4 and 5 suggest that the current speaker's and listeners' MOTPs are potentially useful in predicting the next speaker and utterance interval in multi-party conversations. The objective of this research was to demonstrate precisely how the MOTPs are useful for such prediction. We implemented a prediction model that predicts the next speaker and utterance interval using MOTPs extracted from 1200 ms of data between 1000 ms before the end of IPU and 200 ms after the end of IPU.

The IPU is conventionally determined solely on the basis of whether the person is speaking. Such a simple definition of utterance is suitable for real-time processing but this paper focuses on offline processing. Specifically, the part of the utterance with 200 ms of silence was used as a unit of utterance. IPU generation occurs 200 ms after the utterance actually ends. Therefore, we implemented a prediction model to predict the next speaker and utterance interval when the IPU was created, that is, 200 ms after the IPU ended in our previous studies [3–7]. In a similar manner, we created a prediction model of the next speaker and utterance interval 200 ms after the end of an IPU.

In our previous work, we implemented a three-stage prediction model to predict the next speaker and utterance interval [5,7]. In the present work, similarly, the first step predicts whether turn-keeping or turn-changing will occur. The second step predicts the next speaker during turn-changing and the third step predicts the utterance interval. By building distinct prediction models in stages, it is possible to verify in detail which stage the feature quantity used for prediction is useful for. The following sections describe our implementation and performance evaluation of each prediction model using MOTP.

6.2. Prediction Model of Turn-Keeping or Turn-Changing

First, we describe our implementation and performance evaluation of the first prediction step, which is the prediction of whether turn-keeping or turn-changing will occur. The analysis in Sections 4.2 and 4.3 suggests that the current speaker's and listeners' MOTPs may be useful in predicting turn-keeping and turn-changing. We built prediction models using a single speaker's or listeners' MOTP and evaluated their usefulness separately. We also built a prediction model that uses both the current speaker's and listeners' MOTPs and verified whether more accurate predictions could be made. For comparison, we also implemented a prediction model using GTP, which is

one of the most useful types of information for predicting turn-keeping and turn-changing [5]. We compared the performance of the prediction models using MOTP with one using GTP. In addition, we implemented prediction models using both MOTP and GTP and evaluated the effectiveness of the multimodal-feature fusion using both.

We implemented these prediction models using a support vector machine (SVM) with sequential minimal optimization (SMO) [43]. The Weka data mining tool [44] was used for this implementation and the polynomial kernel was used as the SVM function. The most useful parameters for the cost parameter (C) and hyper parameter of the kernel (γ) were determined using a grid search technique. The objective value is binary data that contain turn-keeping and turn-changing.

We implemented the following five prediction models.

- Current speaker's mouth model (SmM): This model uses the current speaker's MOTPs as feature values. Specifically, among the 13 types of MOTPs obtained in the analysis in Section 4.2, the MOTPs of the current speaker that appeared at the end of the utterance are used. The feature values are expressed as a 13-dimensional one-hot vector.
- Listeners' mouth model (LmM): This model uses the three listeners' MOTPs as feature values. Specifically, among the 11 types of MOTPs obtained in the analysis in Section 4.3, the GTP of the three listeners that appeared at the end of the utterance is used. The feature values are expressed as an 11-dimensional vector that contains information on the number of appearances of each of the 11 MOTPs.
- All-mouth model (AmM): This model uses both the current speaker's and listeners' MOTPs as feature values. The feature values are expressed as a 24-dimensional vector that combines the 13-dimensional vector of the current speaker's MOTPs and the 11-dimensional vector of the listeners' MOTPs.
- Eye-gaze model (EgM): This model uses the GTPs of the current speaker and listeners as feature values. With reference to our previous research on GTPs [5,15,16], we created a GTP that contains n-gram information of eye-gaze objects including mutual gaze information. The GTPs are created using eye-gaze targets between 1000 ms before the IPU and 200 ms after the end of IPU, the same as the MOTPs. There are 11 types of current speaker's GTPs and 14 types of listeners' GTPs. The feature values are expressed as a 25-dimensional vector that contains information on the number of appearances of each of the 25 GTPs.
- Multimodal-feature model (MuM): This model uses all the current speaker's and listeners' MOTPs and GTPs as feature values. These feature values are fused early. The feature values are expressed as a 49-dimensional vector that contains information on the number of appearances of each of the 24 MOTPs and 25 GTPs.

We used ten-fold cross-validation with 210 items of data, which include 105 items during turn-changing and 105 items obtained by random sampling from 904 data during turn-keeping to remove any bias of the amount of data between turn-keeping and turn-changing.

Table 1 lists the mean values of F-measure for each prediction model obtained by ten-fold cross-validation. The chance level (CL) indicates the theoretical value when the value of the objective variable is predicted randomly. Here, the number of data items of each binary value is the same, so the F-measure is 0.500. When we compare the CL with each model using MOTPs or GTPs, the F-measures of SmM, LmM, AmM and EgM, which are 0.696, 0.704, 0.724 and 0.760, respectively, were significantly higher than those of CL (CL vs. SmM: $t(9) = 11.8, p < 0.01$; CL vs. LmM: $t(9) = 11.3, p < 0.01$; CL vs. AmM: $t(9) = 12.9, p < 0.01$; CL vs. EgM: $t(9) = 11.7, p < 0.01$).

Table 1. Evaluation results of prediction models of turn-keeping and turn-changing. Values shown are mean of F-measures by ten-fold cross-validation.

Models	F-Measure
Chance level (CL)	0.500
Current speaker's mouth model (SmM)	0.696
Listeners' mouth model (LmM)	0.704
All-mouth model (AmM)	0.724
Eye-gaze model (EgM)	0.760
Multimodal-feature model (MuM)	0.800

The F-measure of AmM was 0.724, which is the best among the models using only MOTPs. However, there was no significant difference among SmM, LmM and AmM. These results indicate that a model using the parameters of the MOTPs of the current speaker and listeners contributes to predicting turn-keeping and turn-changing in multi-party conversations. However, the performance of the model using both the current speaker's and listeners' MOTPs did not differ from the performance when using either one or the other of them.

When we compare the models using either the MOTPs or the GTPs, the F-measure of EgM, which is 0.760, is significantly better than that of AmM, which is 0.724 ($t(9) = 1.94, p < 0.05$). This suggests that GTPs are more useful for predicting turn-keeping and turn-changing than MOTPs.

The F-measure of MuM, which uses both MOTPs and GTPs, was 0.800, which is significantly better than those of AmM and EgM (MuM vs. AmM: $t(9) = 3.24, p < 0.01$; MuM vs. EgM: t(9) = 1.94,$p <$ 0.05). These results indicate that multimodal-feature fusion using both MOTPs and GTPs is more useful for predicting the turn-keeping and turn-changing than using either the MOTPs or the GTPs individually.

6.3. Prediction of Next Speaker during Turn-Changing

As the second processing step of predicting the next speaker after predicting the turn-keeping and turn-changing, we implemented prediction models to predict the next speaker during turn-changing. As mentioned in Section 4.3, the occurrence frequencies of the MOTPs of the next speaker differ from those of the listeners during turn-changing. On the basis of these results, the listeners' MOTPs have the potential to predict the next speaker during turn-changing. Thus, we implemented the prediction models with the listeners' MOTPs as feature values and evaluated how useful the listeners' MOTPs are for predicting the next speaker during turn-changing. We also implemented a prediction model using the GTPs of the current speaker and listeners, which are known to be one of the most useful types of information for predicting the next speaker during turn-changing [6,7], to compare the performance of MOTPs and GTPs and evaluate the effectiveness of multimodal-feature fusion using both, in the same manner as in Section 6.2.

We implemented the prediction models using the SVM, which can perform multi-class classification and evaluated their performance in the same manner as in Section 6.2. The data used in the SVM contain the next speaker as objective values.

We used ten-fold cross-validation with the 105 items of turn-changing data from the corpus we collected in Section 3. Table 2 lists the mean of the F-measure of each prediction model of the next speaker during turn-changing obtained by ten-fold cross-validation. The chance level (CL) indicates the theoretical value when the value of the objective variable is predicted randomly. The CL was 0.333 because there were three next-speaker candidates during turn-changing in the four-person conversations. When we compare the models using either MOTPs or GTPS with chance level, the F-measures of LmM and EgM were 0.409 and 0.440, which are significantly better than that of CL (CL vs. LmM: $t(9) = 3.06, p < 0.01$; CL vs. EgM: $t(9) = 3.51, p < 0.01$). Comparing LmM with EgM revealed that there is no difference between them (LmM vs. EgM: $t(9) = 1.37, n.s$). These results suggest that the MOTPs and GTPs are equally useful to predict the next speaker during turn-changing.

Table 2. Evaluation results of prediction models for next speaker during turn-changing. Values shown are mean of F-measures by ten-fold cross-validation.

Models	F-Measure
Chance level (CL)	0.333
Listeners' mouth model (LmM)	0.409
Eye-gaze model (EgM)	0.440
Multimodal-feature model (MuM)	0.476

When we compare the model using both the MOTPs and GTPs with ones using only one or the other, the F-measure of MuM was 0.476, which is the highest performance of all (MuM vs. LmM: $t(9) = 2.06, p < 0.01$; MuM vs. EgM: $t(9) = 1.86, p < 0.05$). This result suggests that multimodal-feature fusion using both GTPs and MOTPs is more useful for predicting the next speaker during turn-changing than unimodal-feature processing using either MOTPs or GTPs.

6.4. Prediction of Utterance Interval

In the third step of predicting the utterance interval after predicting the next speaker, we implemented prediction models to predict the utterance interval during turn-keeping and turn-changing. As mentioned in Section 5, the utterance interval differs according to the MOTPs of the current speaker and listeners between turn-keeping and turn-changing. We implemented the prediction models of utterance interval for turn-keeping and turn-changing separately. We implemented these prediction models with the current speaker's and listeners' MOTPs as feature values and evaluated how useful MOTPs are for predicting the utterance interval during turn-keeping and turn-changing. We also implemented a prediction model using the current speaker's and listeners' GTPs to compare the performance of MOTPs and GTPs and a prediction model using both MOTPs and GTPs to evaluate the effectiveness of multimodal-feature fusion. We used these models in the same manner as in Sections 6.2 and 6.3.

We implemented these prediction models using support vector regression (SVR), which is an extension of SVM to a regression model and evaluated their performance. We implemented the two models of utterance interval during turn-keeping and turn-changing, as there is big difference in utterance interval between turn-keeping and turn-changing and the corresponding MOTPs also differ greatly. The data used in the SVR contain the utterance interval as objective values. We used the same features of LmM as those in Section 5.2.

We used ten-fold cross-validation with 904 turn-keeping and 105 turn-changing items of data from our corpus. Table 3 shows the mean absolute error between actual interval and predicted interval of each prediction model during turn-keeping and turn-changing. The error of chance level (CL) indicates the mean absolute error between the actual interval and the mean value of all actual intervals. These were 0.921 during turn-keeping and 0.711 during turn-changing. When we compare the performance of models using either the MOTPs or the GTPs with the one of chance level, the mean absolute errors of MoM and EgM, which were 0.726 and 0.711 in turn-keeping and 0.599 and 0.548 in turn-changing, are significantly better than that of CL (CL vs. MoM in turn-keeping: $t(9) = 3.32, p < 0.01$; CL vs. EgM in turn-keeping: $t(9) = 2.79, p < 0.01$; CL vs. MoM in turn-changing: $t(9) = 3.33, p < 0.01$; CL vs. EgM in turn-changing: $t(9) = 2.98, p < 0.01$). Comparing MoM with EgM reveals no significant difference between their performances. These results suggest that the MOTPs and GTPs are equally useful to predict utterance interval during turn-keeping and turn-changing.

The mean absolute errors of MuM were 0.690 in turn-keeping and 0.452 in turn-changing, which are significantly better than those of MoM and EgM (MuM vs. MoM in turn-keeping: $t(9) = 1.89$, $p < 0.01$; MuM vs. MoM in turn-changing: $t(9) = 2.25$, $p < 0.01$; MuM vs. EgM in turn-keeping: $t(9) = 1.91$, $p < 0.05$; MuM vs. EgM in turn-changing: $t(9) = 2.10$, $p < 0.05$). These results suggest that the multimodal-feature fusion using both MOTPs and GTPs is more useful for predicting the utterance interval than unimodal processing using either one or the other.

Table 3. Performance of prediction models for predicting utterance interval during turn-keeping and turn-changing. Values are mean absolute errors between actual utterance interval and predicted utterance interval by prediction models.

Models	Turn-Keeping	Turn-Changing
Chance level (CL)	0.921	0.711
Mouth model (MoM)	0.726	0.599
Eye-gaze model (EgM)	0.711	0.548
Multimodal-feature model (MuM)	0.690	0.452

7. Discussion

We demonstrated that the current speaker's and listeners' MOTPs differ depending on the next speaker and utterance interval in multi-party conversations. A key finding of our analysis is that the current speaker often continues to open her mouth narrowly during turn-keeping and starts to close it from opening it narrowly or continues to open it widely during turn-changing. The next speaker often starts to open her mouth narrowly from closing it during turn-changing. Moreover, when the current speaker starts to close her mouth after opening it narrowly in turn-keeping, the utterance interval tends to be short. In contrast, when the current speaker and listener open their mouths narrowly after opening them narrowly and then widely, the utterance interval tends to be long.

Previous studies have shown that eye-gaze [5,13–17], eye-blink [18], head movement [4,19], respiration [6,7] and hand gestures [20] are related to turn-changing. In contrast, our work here initially suggested that mouth movement plays an important role for turn-changing.

On the basis of these results, we implemented prediction models of the next speaker and utterance interval using MOTPs that are divided into three steps. The evaluation results of the prediction models suggest that the current speaker's and listeners' MOTPs are effective for predicting the turn-keeping and turn-changing, the next speaker during turn-changing and the utterance interval during turn-keeping and turn-changing in multi-party conversations.

In our previous study [5,15], the GTP was one of the most useful feature values for predicting these. As result of verifying how useful MOTPs are compared to GTPs, we conclude that GTPs are more useful to predict only the turn-keeping and turn-changing, which is the first of the three prediction steps. There is no difference when it comes to the performance of predicting the next speaker during turn-changing and the prediction of utterance interval, which are the two next prediction steps, between using either MOTPs or GTPs. We conclude that MOTPs are a very strong predictor of the next speaker and utterance interval in multi-party conversations, the same as eye-gaze behavior (GTP). In general, it is difficult to accurately estimate the eye-gaze target of people in multiparty meetings. In many cases, a wearable eye tracker is required to accurately estimate the target. Practically speaking, it is thus extremely difficult to predict the next speaker and utterance interval using the eye-gaze target, especially in the wild. In contrast, mouth-opening movement can be measured with a single camera by utilizing computer vision processing [45]. In this study, the mouth opening was manually annotated. However, considering future practicality, since it is possible to predict the next speaker and utterance interval using the mouth opening, which can be measured more easily than the line of sight, usage of mouth opening may be practical.

We also implemented models using both MOTPs and GTPs as a multimodal-feature fusion. The evaluation results of the prediction model suggest that this multimodal-feature fusion is useful for predicting the next speaker and utterance interval.

Next, we discuss a situation in which MOTPs are more useful than GTPs for predicting the next speaker in multi-party conversations. Figure 10 shows a sample of a conversational scene where turn-changing occurs from Person 1 to Person 4. In this scene, the prediction of the eye-gaze model (EgM) (implemented in Section 6) is that turn-keeping will happen next. This prediction is not correct. In contrast, the prediction results of the all-mouth model (AmM) and multimodal-feature model

(MuM) are that turn-changing will happen and the next speaker will be Person 4. These results are perfectly correct.

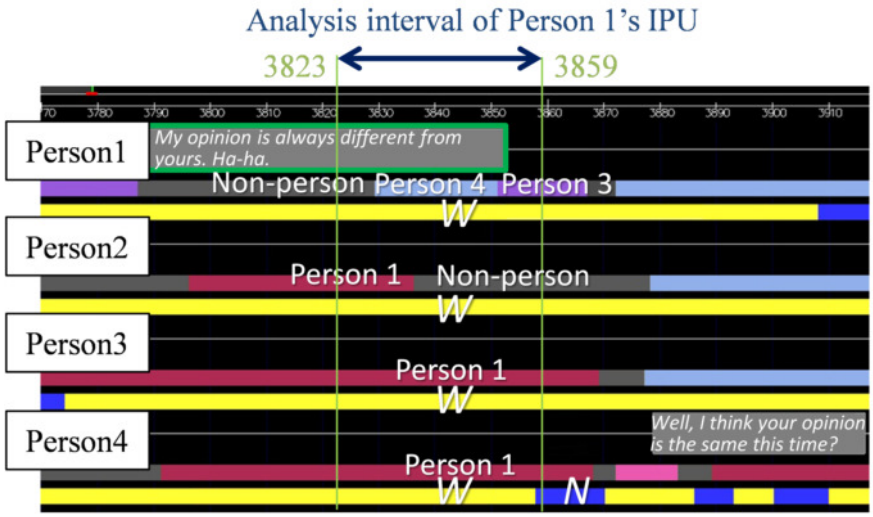

Figure 10. First sample situation where turn-changing occurs from Person 1 to Person 4. Upper gray boxes show IPUs, middle boxes show eye-gaze targets and lower boxes show the degrees of mouth-opening for each person along the timeline.

Next, we examine the differences between the prediction results by observing the feature values of each participant's MOTP and GTP. In this case, the analysis interval is from the 3823th frame, which is 1000 ms before the end of the Person1's IPU which is 3853th frame, to the 3859th frame, which is 200 ms after the end of the IPU.

First, we focus on the GTPs of Person 1, who is the current speaker. She is making eye contact with Person 3, who is a listener, after looking at a non-person and making eye contact with Person 4, who is also a listener. The GTP of Person 1 is X-L_{1M}-L_{2M}. X means looking at a non-person and L means looking at a listener. Numbers below L simply indicate different people. However, this number does not mean a specific person; rather, the numbers are dropped in the order in which the person gazes. M means making eye contact with a gaze target. Please refer to Reference [5] for detailed information. The occurrence frequency of X-L_{1M}-L_{2M} of the current speaker is very low—under 1% during both turn-keeping and turn-changing—and as such is not effective for predicting turn-keeping/turn-changing according to previous studies [5].

Next, we focus on the GTPs of Person 2, who is a listener. She is looking at a non-person after looking at Person 1, who is the current speaker. Person 2's GTP is X-S. S means looking at the current speaker. The occurrence frequency of X-S of the listener is not different between turn-keeping and turn-changing. Persons 3 and 4, who are listeners, make eye contact with Person 1, who is the current speaker. The GTPs of Persons 3 and 4 are both S_M. Their GTPs are the same and their occurrence probability is higher during turn-keeping than during turn-changing. This seems to be the reason EgM incorrectly predicted that turn-keeping will happen next.

Next, focusing on MOTPs during the 1200 ms of the analysis interval, which is same as that of the GTPs, we see that the MOTPs of Persons 1, 2, 3 and 4 are W, W, W and W-N, respectively. The first mouth-opening state of all participants is W. In this scene, all participants laugh heartily near the end of Person 1's IPU. The occurrence frequency of the current speaker's W is higher during turn-changing than during turn-keeping, as mentioned in Section 4.2. The occurrence frequency of W of the listeners during turn-changing is higher than those of the listeners during turn-keeping and the next speaker during turn-changing, as mentioned in Section 4.3. Therefore, judging from the MOTPs, Persons 2 and 3 are unlikely to be the next speaker. The occurrence frequency of W-N of the next speaker during

turn-changing is higher than those of the listeners during turn-keeping and turn-changing. Therefore, Person 4 is likely to be the next speaker.

These analysis results of the MOTPs seem to explain why AmM and MuM predicted that turn-changing will happen and Person 4 will be the next speaker. In this scene, only Person 4, who becomes the next speaker, changes her mouth-opening to narrow from wide (changes from *W* to *N*) while all persons are laughing. There are several scenes that show similar mouth-opening movements of the next speaker while laughing in our corpus. That is, *W*-*N* seems to be a typical mouth-opening movement right before the start of speaking during turn-changing after laughing.

This is just one of many instances. There are many other scenes in which combinations of participants' GTPs and MOTPs occur in turn-keeping and turn-changing. We believe that the MuM prediction model enables high-accuracy predictions by making good use of the various characteristics of the mouth-opening and gaze movements obtained in various situations of turn-keeping and turn-changing. As such, it might be effective to analyze in which situations the prediction model is able to predict the next speaker and in which it is not. Such detailed analysis is one of our future topics of research.

Finally, we discuss the relationship between mouth-opening movement and respiration, which is a useful feature value for predicting the next speaker and utterance interval [6,7], by observing the actual corpus data in detail. In References [6,7], we demonstrated that the next speaker tends to take a deep breath before starting to speak. We feel that this breathing occurs in conjunction with mouth-opening movement. Figure 11 shows a typical scene of turn-changing from Person 1 to Person 4. The analysis interval of the MOTPs is from the 13237th frame to the 13273th frame. The MOTP of Person 1, who is the current speaker and Persons 2 and 3, who are listeners, is *N*, which means narrow-open. On the other hand, the MOTP of Person 4, who will be the next speaker, is *X*-*N*. The occurrence frequency of *N* of the current speaker is higher during turn-keeping than during turn-changing (Section 4.2). There is no difference in the occurrence frequency of *N* of listeners between turn-keeping and turn-changing (Section 4.3). In contrast, the occurrence frequency of *X*-*N* of the listener who will become the next speaker during turn-changing is much higher than those of listeners who will not be the next speaker during turn-keeping and turn-changing.

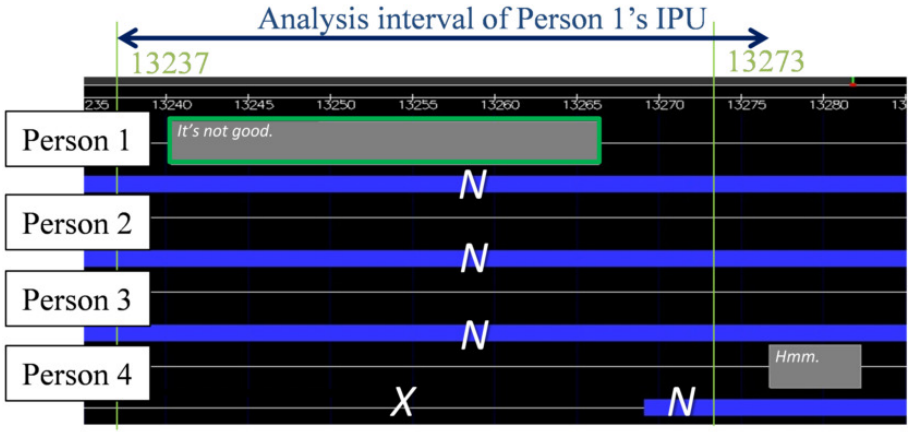

Figure 11. Second sample situation where turn-changing occurs from Person 1 to Person 4. Upper gray boxes show IPUs and lower boxes show the degrees of mouth-opening for each person along the timeline.

Therefore, the prediction result of all prediction models that use MOTPs is that turn-changing will happen next and that the next speaker will be Person 4. This prediction result is perfectly correct. *X*-*N* means that the next speaker opens her mouth narrowly after closing it. *X*-*N* happens frequently and accounts for about 25% of the MOTPs of the next speaker during turn-changing (Section 4.3). We suspect that the next speaker takes a big breath before starting to speak in those cases. In other words,

there is a possibility that the next speaker's taking a breath prior to speaking could be detected by mouth-opening movement, that is, *X*-*N*. Conventionally, a wearable measuring device is required to accurately detect respiration. Therefore, measuring the breathing of participants in a multi-party conversation has a practical problem. A technique for robustly detecting breathing by means of mouth-opening movement information would thus be very useful for predicting the next speaker in multi-party conversations. In other words, even if we do not use respiration information, we may be able to predict the next speaker and utterance interval simply by the mouth-opening movements. This verification will be the focus of our future work.

There is also the possibility that mouth-opening movement is associated with several other modality, such as phonemes, prosody and facial expressions. In the future, we plan to investigate mouth forms in detail and demonstrate the relationship between them and other modalities in multi-party conversations. We will then implement a more highly accurate model for predicting the next speaker and utterance interval. We also plan to introduce an automatic extraction technique for the degree of mouth opening that utilizes a computer vision technique [37] instead of manual annotation. With such automatic measurement, a wider variety of mouth opening parameters can be used for constructing the prediction model. It may then be possible to construct a more accurate prediction model by using these diverse and detailed feature values.

8. Conclusions and Future Work

We have demonstrated that the current speaker's and listeners' MOTPs differ depending on the next speaker and utterance interval. A key finding of our analysis is that the current speaker often continues to open her mouth narrowly in turn-keeping and starts to close it from opening it narrowly or continues to open it wider in turn-changing. The next speaker often starts to open her mouth narrowly from closing it in turn-changing. Moreover, when the current speaker starts to close her mouth after opening it narrowly in turn-keeping, the utterance interval tends to be short, the utterance interval tends to be short. In contrast, when the current speaker and listener open their mouths narrowly after opening them narrowly and then wider, the utterance interval tends to be long.

On the basis of these results, we implemented next-speaker prediction models using the MOTPs. Our evaluation of the models suggests that the current speaker's and listeners' MOTPs are effective for predicting the next speaker and utterance interval in multi-party conversations. We also implemented models using the MOTPs and eye-gaze information as a multimodal-feature fusion. The evaluation results suggest that multimodal-feature fusion using MOTPs and eye-gaze information is useful for predicting the next speaker and utterance interval. From our observation of the corpus data, we showed the possibility of multimodal fusion using mouth-opening movement and eye-gaze behavior.

In the future, we plan to explore the effectiveness of the above multimodal-feature fusion for predicting the next speaker and utterance interval by using head nods [19] and respiration [6,7].

Author Contributions: Conceptualization, R.I. and K.O.; Methodology, R.I.; Software, R.I.; Validation, R.I., K.O. and S.K.; Formal analysis, R.I.; Investigation, R.I.; Resources, R.I., K.O. and S.K.; Data curation, R.I., K.O. and S.K.; Writing—original draft preparation, R.I.; Writing—review and editing, R.I., K.O., S.K., R.H. and J.T.; Visualization, R.I. and K.O.; Supervision, K.O. and J.T.; Project administration, K.O., R.H. and J.T.; Funding acquisition, K.O. and J.T.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gatica-Perez, D. Analyzing group interactions in conversations: A review. In Proceedings of the MFI, Heidelberg, Germany, 3–6 September 2006; pp. 41–46.
2. Otsuka, K. Conversational scene analysis. *IEEE Signal Process. Mag.* **2011**, *28*, 127–131.
3. Ishii, R.; Kumano, S.; Otsuka, K. Multimodal Fusion using Respiration and Gaze for Predicting Next Speaker in Multi-Party Meetings. In Proceedings of the ICMI, Tokyo, Japan, 12–16 November 2016; pp. 99–106.

4. Ishii, R.; Kumano, S.; Otsuka, K. Predicting Next Speaker Using Head Movement in Multi-party Meetings. In Proceedings of the ICASSP, Queensland, Australia, 19–24 April 2015; pp. 2319–2323.
5. Ishii, R.; Otsuka, K.; Kumano, S.; Yamamoto, J. Predicting of Who Will Be the Next Speaker and When Using Gaze Behavior in Multiparty Meetings. *ACM TiiS* **2016**, *6*, 4.
6. Ishii, R.; Otsuka, K.; Kumano, S.; Yamato, J. Analysis of Respiration for Prediction of Who Will Be Next Speaker and When? In Proceedings of the ICMI, Istanbul, Turkey, 12–16 November 2014; pp. 18–25.
7. Ishii, R.; Otsuka, K.; Kumano, S.; Yamamoto, J. Using Respiration to Predict Who Will Speak Next and When in Multiparty Meetings. *ACM TiiS* **2016**, *6*, 20.
8. Gracco, V.L.; Lofqvist, A. Speech Motor Coordination and Control: Evidence from Lip, Jaw, and Laryngeal Movements. *J. Neurosci.* **1994**, *14*, 6585–6597.
9. Sacks, H.; Schegloff, E.A.; Jefferson, G. A simplest systematics for the organisation of turn taking for conversation. *Language* **1974**, *50*, 696–735.
10. Kendon, A. Some functions of gaze direction in social interaction. *Acta Psychol.* **1967**, *26*, 22–63.
11. Lammertink, I.; Casillas, M.; Benders, T.; Post, B.; Fikkert, P. Dutch and English toddlers' use of linguistic cues in predicting upcoming turn transitions. *Front. Psychol.* **2015**, *6*, 495.
12. Levinson, S.C. Turn-taking in human communication—Origins and implications for language processing. *Trends Cogn. Sci.* **2016**, *20*, 6–14.
13. Kawahara, T.; Iwatate, T.; Takanashii, K. Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations. In Proceedings of the INTERSPEECH, Portland, OR, USA, 9–13 September 2012.
14. Jokinen, K.; Furukawa, H.; Nishida, M.; Yamamoto, S. Gaze and turn-taking behavior in casual conversational interactions. *ACM TiiS* **2013**, *3*, 12.
15. Ishii, R.; Otsuka, K.; Kumano, S.; Matsuda, M.; Yamato, J. Predicting Next Speaker and Timing from Gaze Transition Patterns in Multi-Party Meetings. In Proceedings of the ICMI, Sydney, Australia, 9–13 December 2013; pp. 79–86.
16. Ishii, R.; Otsuka, K.; Kumano, S.; Yamato, J. Analysis and Modeling of Next Speaking Start Timing based on Gaze Behavior in Multi-party Meetings. In Proceedings of the ICASSP, Florence, Italy, 4–9 May 2014; pp. 694–698.
17. Holler, J.; Kendrick, H. Unaddressed participants' gaze in multi-person interaction: optimizing recipiency. *Front. Psychol.* **2015**, *6*, 515–535.
18. Hömke, P.; Holler, J.; Levinson, S.C. Eye blinking as addressee feedback in face-to-face conversation. *Res. Lang. Soc. Interact.* **2017**, *50*, 54–70.
19. Ishii, R.; Kumano, S.; Otsuka, K. Prediction of Next-Utterance Timing using Head Movement in Multi-Party Meetings. In Proceedings of the HAI, Bielefeld, Germany, 17–20 October 2017; pp. 181–187.
20. Holler, J.; Kendrick, K.H.; Levinson, S.C. Processing language in face-to-face conversation: Questons with gestures get faster responses. *Psychon. Bull. Rev.* **2018**, *6*, 25.
21. Chen, L.; Harper, M.P. Multimodal floor control shift detection. In Proceedings of the ICMI, Cambridge, MA, USA, 2–4 November 2009; pp. 15–22.
22. de Kok, I.; Heylen, D. Multimodal end-of-turn prediction in multi-party meetings. In Proceedings of the ICMI, Cambridge, MA, USA, 2–4 November 2009; pp. 91–98.
23. Ferrer, L.; Shriberg, E.; Stolcke, A. Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog. In Proceedings of the INTERSPEECH, Denver, CO, USA, 16–20 September 2002; Volume 3, pp. 2061–2064.
24. Laskowski, K.; Edlund, J.; Heldner, M. A single-port non-parametric model of turn-taking in multi-party conversation. In Proceedings of the ICASSP, Prague, Czech Republic, 22–27 May 2011; pp. 5600–5603.
25. Schlangen, D. From reaction to prediction: experiments with computational models of turn-taking. In Proceedings of the INTERSPEECH, Pittsburgh, PA, USA, 17–21 September 2006; pp. 17–21.
26. Dielmann, A.; Garau, G.; Bourlard, H. Floor holder detection and end of speaker turn prediction in meetings. In Proceedings of the INTERSPEECH, Makuhari, Japan, 26–30 September 2010; pp. 2306–2309.
27. Itoh, T.; Kitaoka, N.; Nishimura, R. Subjective experiments on influence of response timing in spoken dialogues. In Proceedings of the ISCA, Brighton, UK, 6–10 September 2009; pp. 1835–1838.
28. Inoue, M.; Yoroizawa, I.; Okubo, S. Human Factors Oriented Design Objectives for Video Teleconferencing Systems. *ITS* **1984**, 66–73.

29. Matthews, I.; Cootes, T.F.; Bangham, J.A.; Cox, S.; Harvey, R. Extraction of visual features for lipreading. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 198–213.
30. Chakravarty, P.; Mirzaei, S.; Tuytelaars, T. Who's speaking?: Audio-supervised classification of active speakers in video. In Proceedings of the ICMI, Seattle, WA, USA, 9–13 November 2015.
31. Chakravarty, P.; Zegers, J.; Tuytelaars, T.; hamme, H.V. Active speaker detection with audio-visual co-training. In Proceedings of the ICMI, Tokyo, Japan, 12–16 November 2016; pp. 312–316.
32. Cech, J.; Mittal, R.; Deleforge, A.; Sanchez-Riera, J.; AlamedaPineda, X.; Horaud, R. Active-speaker detection and localization with microphones and cameras embedded into a robotic head. In Proceedings of the Humanoids, Atlanta, GA, USA , 15–17 October 2013; pp. 203–210.
33. Cutler, R.; Davis, L. Look who's talking: Speaker detection using video and audio correlation. In Proceedings of the ICME, New York, NY, USA, 30 July–2 August 2000; pp. 1589–1592.
34. Haider, F.; Luz, S.; Campbell, N. Active speaker detection in human machine multiparty dialogue using visual prosody information. In Proceedings of the GlobalSIP, Washington, DC, USA, 7–9 December 2016; pp. 1207–1211.
35. Haider, F.; Luz, S.; Vogel, C.; Campbell, N. Improving Response Time of Active Speaker Detection using Visual Prosody Information Prior to Articulation. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018; pp. 1736–1740.
36. Murai, K. Speaker Predicting Apparatus, Speaker Predicting Method, and Program Product for Predicting Speaker. U.S. Patent 20070120966, 2011.
37. Cheunga, Y.; Liua, X.; You, X. A local region based approach to lip tracking. *Pattern Recognit.* **2012**, *45*, 3336–3347.
38. Koiso, H.; Horiuchi, Y.; Tutiya, S.; Ichikawa, A.; Den, Y. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Lang. Speech* **1998**, *41*, 295–321.
39. Ekman, P.; Friesen, W.V. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.
40. Conger, A. Integration and generalisation of Kappas for multiple raters. *Psychol. Bull.* **1980**, *88*, 322–328.
41. Otsuka, K.; Araki, S.; Mikami, D.; Ishizuka, K.; Fujimoto, M.; Yamato, J. Realtime meeting analysis and 3D meeting viewer based on omnidirectional multimodal sensors. In Proceedings of the ICMI, Cambridge, MA, USA, 2–4 November 2009; pp. 219–220.
42. Haberman, S.J. The analysis of residuals in cross-classified tables. *Biometrics* **1973**. *29*, 205–220.
43. Keerthi, S.S.; Shevade, S.; Bhattacharyya, C.; Murthy, K.K. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Comput.* **2001**, *13*, 637–649.
44. Bouckaert, R.R.; Frank, E.; Hall, M.A.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. WEKA–Experiences with a Java Open-Source Project. *J. Mach. Learn. Res.* **2010**, *11*, 2533–2541.
45. Amos, B.; Ludwiczuk, B.; Satyanarayanan, M. *OpenFace: A General-Purpose Face Recognition Library with Mobile Applications*; Technical Report, CMU-CS-16-118; CMU School of Computer Science: Pittsburgh, PA, USA, 2016.