Multimodal Technologies and Interaction

MDPI

Article

Generation of Head Movements of a Robot Using Multimodal Features of Peer Participants in Group Discussion Conversation

Hung-Hsuan Huang [1,2,3,4,*], **Seiya Kimura** [2,4], **Kazuhiro Kuwabara** [4] and **Toyoaki Nishida** [1,2,3]

[1] Faculty of Informatics, The University of Fukuchiyama, Fukuchiyama, Kyoto 620-0886, Japan
[2] Center for Advanced Intelligence Project, RIKEN, Kyoto 606-8501, Japan
[3] Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
[4] College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan
* Correspondence: hhhuang@acm.org

Received: 26 May 2019; Accepted: 15 October 2019; Published: 29 April 2020

Abstract: In recent years, companies have been seeking communication skills from their employees. Increasingly more companies have adopted group discussions during their recruitment process to evaluate the applicants' communication skills. However, the opportunity to improve communication skills in group discussions is limited because of the lack of partners. To solve this issue as a long-term goal, the aim of this study is to build an autonomous robot that can participate in group discussions, so that its users can repeatedly practice with it. This robot, therefore, has to perform humanlike behaviors with which the users can interact. In this study, the focus was on the generation of two of these behaviors regarding the head of the robot. One is directing its attention to either of the following targets: the other participants or the materials placed on the table. The second is to determine the timings of the robot's nods. These generation models are considered in three situations: when the robot is speaking, when the robot is listening, and when no participant including the robot is speaking. The research question is: whether these behaviors can be generated end-to-end from and only from the features of peer participants. This work is based on a data corpus containing 2.5 h of the discussion sessions of 10 four-person groups. Multimodal features, including the attention of other participants, voice prosody, head movements, and speech turns extracted from the corpus, were used to train support vector machine models for the generation of the two behaviors. The performances of the generation models of attentional focus were in an F-measure range between 0.4 and 0.6. The nodding model had an accuracy of approximately 0.65. Both experiments were conducted in the setting of leave-one-subject-out cross validation. To measure the perceived naturalness of the generated behaviors, a subject experiment was conducted. In the experiment, the proposed models were compared. They were based on a data-driven method with two baselines: (1) a simple statistical model based on behavior frequency and (2) raw experimental data. The evaluation was based on the observation of video clips, in which one of the subjects was replaced by a robot performing head movements in the above-mentioned three conditions. The experimental results showed that there was no significant difference from original human behaviors in the data corpus and proved the effectiveness of the proposed models.

Keywords: robot; gaze; visual focus of attention; nod; multiparty interaction; machine learning; support vector machine

1. Introduction

While companies are running projects, the communication skill of individual project members largely affects the relationship with other members and, thus, has great influence on team performance.

There has been a growing number of companies adopting group discussion in the recruitment of their employees, and, often, communication skill is regarded as even more important than professional skills. In a group-discussion-style interview, job applicants have to collaborate with each other to deliberate and produce productive results on an assigned topic. During the discussion process, their communication skill and personality are observed by the investigators of the companies. Therefore, the perception of higher communication skills may lead to the applicant's success in job hunting.

Repeated practice is considered to improve the communication skill of job applicants. However, such practice requires good partners, and finding good partners can be difficult. Research has been conducted for creating artificial partners for practicing communication—for example, an EU-funded project, TARDIS, has investigated the development of virtual agents for dyadic job interview training [1–3]. The aim of the ongoing project reported here is to develop a training environment that enables the trainees to practice group discussion with communicational robot(s) as its ultimate goal.

To build a realistic environment for effective practice, the robot must perform believable and comprehensive behaviors. Contrary to dyadic dialogs, where there are only the speaker and the addressee of the speaker's utterances as the participants, the distinction of conversational participants' roles, including speaker, addressee, and overhearer, is necessary in multiparty conversation. Otherwise, the participant cannot decide whether and how to respond to individual utterances from the other participants. In addition, because there is the potential that more interlocutors may acquire dialog turns from and transfer dialog turns to other interlocutors, managing the communication flow in a multiparty dialog is much more complex. The main difficulty emerges from the ability of the users to interact with each other, which is difficult for the robot to understand. To realize a robot that can join group discussions, functions other than those of regular conversational robot agents need to be incorporated, because the regular agents are designed only for dyadic conversation sessions. Traum [4] studied general issues in realizing multiparty human-agent interactions. Research has been conducted on handling multiparty issues—for example, the identification model of the addressee of human utterances [5–7]. Multiparty situations encounter the technical difficulty of the implementation of communicational robots. However, at the same time, such phenomena as copying and synchrony [8] between conversation partners may provide cues for robot behavior generation. Behaviors of other participants can provide cues for generating the robot's motion—for example, counterpredictive cueing experiments show that humans are reflexive to each other's gaze direction [9–11].

On the other hand, multimodal features have been shown to be effective for interpreting and predicting behaviors and intentions, such as who is talking to whom, from the features extracted from the actors themselves [5,12]. It is, therefore, of interest to examine how well the behaviors of the peer participants can be used to generate humanlike behaviors of robots. A robot's behaviors are considered to include intentional and spontaneous ones. It is supposed that intentional behaviors are more determined by the robot's internal state, while spontaneous ones are supposed to be stimulated more by the environment. During a multiparty conversation, it is natural for a participant to shift her or his attention among the potential targets (e.g., other participants or the materials about the discussion topic) or to nod from time to time. These behaviors may not necessarily have an intention (e.g., to fetch or release a speaking turn or to agree with current opinion) behind them. It is thought that, if the robot can perform such spontaneous behaviors in responding to other participants, its believability can be improved.

It is believed that the perceived attention target of someone can be approximated from her or his gaze direction. Research on human communication has shown that eye gaze is an important communication signal in face-to-face conversations. The speaker looks at the addressee to monitor her or his understanding or attitude. On the contrary, the addressee looks at the speaker to return positive feedback to the speaker [13,14]. Eye gaze also plays an important role in turn taking. In releasing a turn, the speaker looks at the next speaker at the end of the utterance [15]. Vertegaal [16] reported that gaze is a reliable predictor of addresseehood. Likewise, Takemae [17] provide evidence that the speaker's gaze indicates addresseehood and has a regulatory function in turn management. Addressee identification with low-level multimodal features has been proved to be effective [5,18]. Because of the limited degrees of freedom (DOFs) compared with a human, a robot may not be able to perform subtle gaze behaviors. In the extreme case, the robot may not have movable eyeballs. Therefore, the attention target is treated as the overall combination of head direction and gaze direction of the eyeballs of the robot. The implementation of the proposed attention model can be utilized by available physical parts of the robot. Nodding is another spontaneous and intentional behavior to convey social signals in conversation. It contributes to regulating the flow of speaking turns and showing acknowledgment, and it serves as one of the back channels [15].

The main contributions of this study are summarized as follows:

- We propose a general model to generate reactive and spontaneous head motions to show an autonomous robot's visual attention during a group discussion, where all participants play equal roles in making decisions. To improve performance, the robot's attention behaviors are distinguished in three situations: when the robot is speaking, when a participant other than the robot is speaking, and when no one is speaking. Specifically, we build one model for each situation. Although head motions of robots have been researched for years, this highly dynamic situation has not been researched sufficiently.
- The models are developed by a data-driven method using end-to-end machine learning techniques. The outputs are the timings and the target shifts (when to look at whom), while the inputs are the low-level signals and contextual features extracted only from the peer participants. Compared with previous studies, where multimodal sensory information was used only in the detection of the user(s)' state, and the robot's head motion was usually determined by heuristics [19,20], the proposed method can potentially generate fine-grained and believable motion.
- This investigation was based on a relatively large data corpus collected by the authors, which is composed of video/audio data for the group discussions of 10 four-participant groups (15 min for each group; that is, 10 h of multimodal data).
- To ensure the minimum believability of the robot, the attention models are complemented with the nodding models using the same method.
- Although the models are not designed for a specific robot, we integrated them into a popular robot, Sota, which has been developed for human-robot interaction research. We then conducted a subjective evaluation experiment based on the Sota implementation. The models are compared with two baselines (human behaviors and statistical results based ones) and are evaluated in the manner where the input stimuli are unseen to the model under evaluation to simulate a realistic situation.

This paper is organized as follows. Section 2 is a discussion of related works. After the introduction of the data corpus in Section 3, the development of the attention models is described in Section 4, and the nodding models using machine learning techniques are presented in Section 5. Section 6 provides a description of how the models can be integrated into a robot system. Section 7 gives the results of the subjective evaluation of the proposed models. Finally, Section 8 concludes this work.

2. Related Works

In the context of a group meeting, based on nonverbal features, including such features as speaking turn, voice prosody, visual activity, and visual focus of attention, Aran and Gatica-Perez [21] presented an analysis of participants' personality prediction in small groups. Okada et al. [22] developed a regression model to infer the score for communication skill using multimodal features, including linguistic and nonverbal features: voice prosody, speaking turn, and head activity. Schiavo et al. [23] presented a system that monitors the group members' nonverbal behaviors and acts as an automatic facilitator. It supports the flow of communication in a group conversation activity.

Furthermore, job interviews also have been studied in the research field of multimodal interaction. Raducanu et al. [24] made use of The Apprentice reality TV show, which features a competition for a real, highly paid corporate job. The study was carried out using nonverbal audio cues to predict the person with the highest status and to predict the candidates who are going to be fired. Muralidhar et al. [25] implemented a behavioral training framework for students with the goal of improving the perception of their hospitality by others. They also evaluated the relationship between automatically extracted nonverbal cues and various social signals in a correlation analysis.

In the domains of human-robot and human-agent interaction, to achieve natural and effective communication with humans, the realization of essential communicative behaviors and backchannels has been researched [26]. Researchers evaluated the perceived degree of agreement, affirmation, and attentiveness from synthesized acoustic and visual backchannels (head nods) of virtual agents [27,28]. These works focus on the meaningful behaviors that convey specific intentions of the agent or the robot, while the purpose of our work is more focused on spontaneous reactions that not necessarily have some intention behind. Unlike relatively more flexible virtual agents, robots inherently have less expressiveness due to the physical limitations of their actuators, in the aspects of shape, degree of freedom, smoothness, speed, and so on. Researchers have been working on the imitation of the head movements of humans for tele-operated robots [29–31]. These works do not synthesize robots' head movements in an autonomous manner but aim to replicate the head movements of the robot's human operator in real-time.

On the other hand, for autonomous robots, their behaviors have to be generated from the available information in real-time interaction. Potential information sources include the robot's own intention, perceived human partner's intention, the history of interaction, and so on. Techniques for the behavior generation of robots, including biologically inspired, data-driven, and heuristic-based ones, have been explored [32], and the present work falls into the data-driven category. There are also works about multiparty interaction in this field; however, most of them are treating the robot as having a different role (asymmetric relationship) than the human users. For example, Leite et al. [33] used two MyKeepon robots in a scripted interactive storytelling system with a group of children. Vazquez et al. investigated the effects of the robot's orientation and gaze direction in a conversation where the participant group has a brainstorming discussion about how to solve the robot's problem [34], bartender robots serving multiple customers in a bar [35–37], and a receptionist robot [38]. In the authors' work, the aim is to make the robot join the discussion and play the same role as the other peer human (or robot) participants. The dynamics of this group (robot and human) can be considered different from those of an asymmetric group. For example, in previous work in which a life-sized agent talked to two human users as a tour guide, the users spent much more time looking at the agent rather than her or his partner [5]. Therefore, dedicated behavior models are required for such a robot.

To implement robot's interaction with multiple humans, the robot must detect the states of the humans to make decisions and generate corresponding behaviors. During this process, previous works usually only handled intentional behaviors (head motion, nods, and speech). The first consideration is the robot's intention. Ishii et al. [20] developed a model that drives the robot's head in 3 dimensions, which is triggered by the speech acts of the utterances. Due to a small dataset (seven speakers engaged in interactions in a varying number of participants and different relationships), this model is a rule-based one derived from the analysis of the dataset. More general setting is the combination of the detection of partners' states based on sensory information to drive the robot's gaze and/or head orientation according to heuristics or rules. The heuristics are either derived from experimental data or from the literature [34,35,38–40]. Some researchers took a data-driven/machine learning approach to detect the state of the users— for example, whether a passing customer accepts service from a bartender robot [36] or whether a user group is willing to engage in or disengage from the interaction with the receptionist robot [38,41]. Although not for multiple users, there are also works on generating a robot's head motion with a data-driven method. Sakai et al. [37] derived the relative probabilities of the robot's head motion from a data corpus where one human user talks to two robots.

Sakai et al. [42] developed a head motion model based on the recognized speech acts of the operator of a tele-operated robot.

The present work differs from previous works not only in the conversation context (symmetric versus asymmetric roles of the participants) but also in the method. Instead of developing intentional responsive behaviors based on heuristics and detection results of the users' state, end-to-end machine learning is used for generating head motion from the features extracted from the users. This technique has the advantage of generating and updating head motions in short time spans. In that sense, the work of Stefanov et al. [43] has similarities to our work. They trained forward neural networks with end-to-end features of multiparty interaction. The dataset [44] contained sessions of three participants playing a quiz game where two player participants solve the quiz with help from the mediator. The mediator is the same through all sessions while the player participants change in each session. The learning problem is modeled as follows: the video/audio features extracted from player participants are explanatory variables, and the gaze direction or head orientation of the mediator are the target variables. The features extracted from the player participants are head orientation, gaze direction, and binary speech activity (on/off). Our work differs from that one in terms of the number of participants, the task of the experiment, and the roles of the participant; therefore, these two works cannot be compared directly in the sense of model performance. Compared with the simple and low-level features used in that study, in this work, more than 100 features were used in the learning process. In addition to low-level features, verbal and speech turn features were also included to capture the context of the decision-making group discussion (compared with the previous work where there was no specific purpose). In the aspect of group compositions, the previous study had fewer participants; one of them was the same person (the mediator), who was present throughout the dataset and played a different role than the other two participants. In comparison, our dataset contains groups with more participants, which change every time, with no bias related to their roles. Moreover, the conversation task is more open in our dataset rather than a fixed task in the previous study. Regarding machine learning problem formulation, their model is meant to predict the head motion of one specific role (the mediator), who is a single person in each session of their two datasets. On the other hand, we are trying to develop a model for generating general human behaviors from many participants. Upon these, we expect more complex interaction and larger group dynamics in our dataset. This will lead to a larger data variety and cause more difficulties in machine learning.

3. Collection of Data Corpus

To develop a realistic robot, the most intuitive source of ground truth is human-human group discussion. Therefore, an experiment was conducted to gather the data corpus for the extraction of the characteristics of human behaviors. The experimental procedures followed the ones of a previous study, using the MATRICS corpus [45], except that eye-tracker glasses were not used.

The discussion task is typically called a "survival task" [46], which is frequently used in the recruitment of Japanese companies. A list of items is shown to the applicants, and their task is to rank the items in the order of importance. One of the purposes of this task is to decide the priority of the items under the pressure that the final decision must be made within a time limit. For this task, participants' skill in logically and clearly stating the preferred item order can be observed. Another requirement for this type of task is to resolve any disagreements that might arise among participants regarding the item priority and finally reach an agreement.

The actual topic for the experiment's participants to discuss should have been easy and familiar for Japanese college students, who were the participants (and target users), to discuss. The topic, celebrity guest selection, was chosen. The participants were asked to pretend that they were the executive committee members for a college festival and were choosing the guests of the festival. The goal of the discussion task was to decide the ranked order of 15 celebrities by considering cost and audience attraction. For the first 5 min, each participant was requested to read the instructions and then decide the order of the 15 celebrities for 3 min without interacting with other members. Then, the participants engaged in a 15-min discussion to decide collaboratively the ranked order as a group.

Forty college students were recruited for the data recording of this experiment. Thirty were male, 10 were female, all were native Japanese speakers, and the discussions were in Japanese. For making sure that the participants had enough knowledge and motivation in the experimental sessions, students who had finished their job hunting or ones who were job hunting were recruited—that is, the third- or fourth-year students in college. They were divided into 10 groups, each one with four people. To prevent the occurrence of gender bias, all members in a group were of the same gender, or in equal number of the two genders. The resulting grouping was five groups that were all male and five groups that had two male and two female participants. For a situation closer to the group discussion sessions of recruitment, the combination of the participants was arranged so that they did not know each other before the experiment.

The experiment participants sat around a 1.2 × 1.2-m square table. Two video cameras, as well as a variety of sensors, were used to record all the discussion sessions. Each participant wore a headset microphone (Audio-Technica HYP-190H), which was connected to an audio digitizer (Roland Sonar X1 LE), with an accelerometer (ATR-Promotions TSND121) on her or his head. Each had a dedicated webcam (Logicool C920) to capture her or his face in a large size. Motion capture (OptiTrack Flex 3 with eight cameras) and Microsoft Kinect sensors were used to record the upper-body movements of the participants as well. The setup of the recording experiment is shown in Figure 1. As the results of the experiment, 15 min × 10 groups = 150 min of group discussion conversation was recorded in the data corpus. With the prosodic analysis of the tool Praat [47], the general statistical information of this corpus was as follows. In each session, there were, on average, 767.1 utterances (maximum 913, minimum 570, standard deviation 101.7), and the utterances had an average length at 0.898 s (maximum 10.6 s, minimum 0.2 s, standard deviation 0.868).

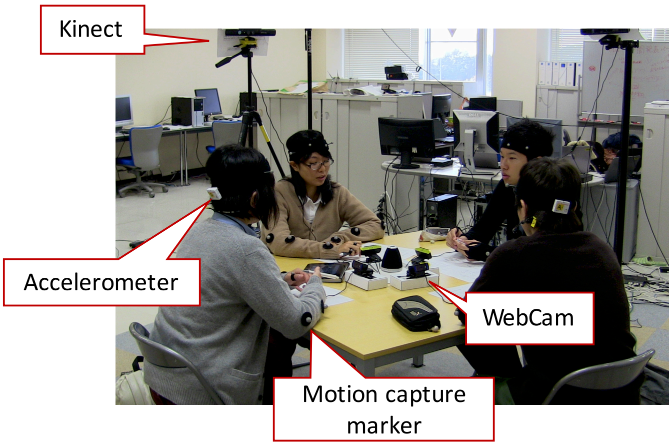

Figure 1. Setup of the data corpus recording environment.

4. Attention Model

In this particular study, the focus was on the modeling of the attention target of the robot, rather than direct interpretation of gaze direction or the head orientation of the robot. This is because the actuators of a robot typically cannot work as subtly as those of humans. They generally have many fewer DOFs than those of humans, and the motors do not drive smoothly. They may have no movable eyeballs as well. Therefore, the research issue was transformed from the imitation of detailed human movements to a simpler, abstract, and device-independent one: how to make the robot pay attention to the same target at the same time as a human would. Utilizing this attention model enabled a device-dependent behavior generator to then be used to drive the physical robots.

4.1. Definition of Attention Targets

In the experiment, which used a typical group discussion setting, the possible attention targets of the participants were considered to be one of the other participants and the discussion material (the document with the list of celebrities' names in the experiment). From the point of view of each participant, there were three other participants: the one sitting at the left-hand side, the one sitting at the opposite side of the table, and the one sitting at the right-hand side. They were called the *Left*, *Front*, and *Right* participants for the attention targets of an individual participant (or the robot to be implemented), respectively. Because the material was placed on the table and it was difficult to distinguish a gaze directed at the material itself from a gaze at the table, the attention paid to the material was called *Table*. One coder then manually coded the attention targets of all of the participants by observation of the video corpus with the annotation tool Elan [48]. The label *Away* was added when the coder could not judge the attention target in the four classes above. The annotation was on every individual participant from her or his perspective.

Table 1 shows the annotation results. The participants paid attention to the material most frequently and for the longest time. This might result from the nature of the experimental setting—the participants looked at the name list while deliberating about the celebrity candidates. This implies that they spent more time deliberating individually rather than in discussion with the others. Among the other participants, they looked at the one sitting at the opposite side of the table most frequently, and there was no obvious difference between the left- and right-hand sides. Finally, the Away class had many fewer instances than the others.

Table 1. Annotation results for the attention target of the participants of the data corpus: the columns, "avg," "max," and "min" shows the average, maximum, and minimum duration in seconds, respectively.

Attention	Instances	avg	max	min
Table	1646	18.6	350.0	0.4
Front	1071	2.6	66.7	0.2
Right	661	1.8	29.9	0.2
Left	642	2.5	17.4	0.1
Away	3	2.1	4.5	0.9

4.2. Situational Models of Attentions

The robot can participate in group discussion sessions as one of the following roles of participation: speaker, addressee, or overhearer. The attention behaviors can be considered different when a human is playing different roles. Among these, it can be difficult to distinguish between an addressee and an overhearer without verbal information. In this study, these two roles were combined, and an attention model of the robot was used for each one of the following three situations:

- Speaking model: when the robot is speaking
- Listening model: when a participant other than the robot is speaking
- Idling model: when no one is speaking

Because the table was square, every one of the four participants in a group could be treated equally. Therefore, each participant could be considered as the candidate for the robot, resulting in 600 min (15 min × 10 groups × 4 people), or 10 h of data, for the training of the robot's attention models.

The specific participant being considered for training the attention model of the robot was defined as the "focused" participant. Because the attention labels were coded from the perspective of each participant, a local-to-global transformation was required to extract the relationship among the attention targets of the participants—that is, the labels of the participants other than the focused participant were transformed relatively in the perspective of the focused participant during data extraction. In the specific example situation shown in Figure 2, the attention labels are rewritten as follows: the Left participant is paying attention the Front participant, the Front participant is paying attention to the Right participant, and the Right participant is paying attention to "Me."

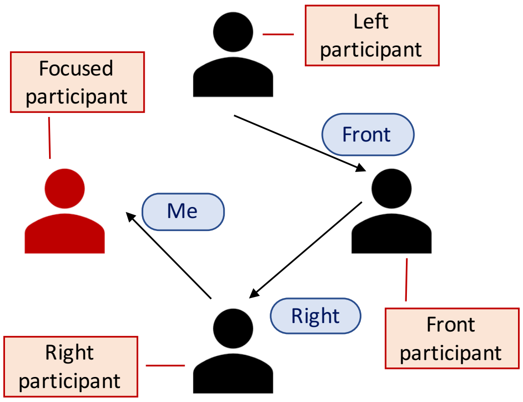

Figure 2. Conceptual diagram of the transformation of attention labels.

4.3. Multimodal Feature Extraction

To use the automatic prediction model to determine the robot's attention focus, the low-level nonverbal features that can be extracted from the behaviors of the participants directly were adopted. They were then used as the dataset to train a classification model for the attention target of the focused participant (the robot). Because of the very few instances of the Away class, this class was omitted, and the model was trained to predict the other four classes: Table, Front, Left, and Right. The extracted features were from four modalities: attention targets of the other participants, speech turn information, prosodic information of the utterances of the other participants, and the body activities of the other participants. This was because the robot was aware of its own behaviors and needed to determine its attention target according to the current situation in real time.

The multimodal low-level features that were selected were the ones that were thought to be extracted directly from the behaviors of the participants other than the robot itself at a temporal granularity of 0.1 s. This resulted in approximately 0.36 million data instances for the speaking model, listening model, and idling model in total. The distribution of the data instances is shown in Table 2. The features in five categories including low-level ones (attention, prosody, and head activity) and contextual ones (verbal and speech turns) were chosen. As discussed in the introduction section, people's attentional focus (gaze) can be affected by the others; hence, the features related to the peer participants' attention were selected. Prosodic features were selected because they present how a person is speaking; this may reveal the importance of the current utterance or the attitude of the person who is speaking. Head movement features were chosen because they may capture the nods and the overall activeness of the participant. Contextual features are considered because they may help predict how well the current speaker can attract the attention of the other participants. Numbers of parts of speech were chosen as verbal features. They were selected because of the hypothesis that the distribution of the parts of speech can capture the characteristics of a person's utterances—whether the person is providing useful information to the discussion, did not participate in the discussion actively, etc. Moreover, speech turn features may capture how well the current speaker has been contributing to the discussion up to now, which may imply the potential influence of this person. As a result, the following 122 features were selected. These features were extracted in a sliding window, symbol t denotes the window size.

- (A)ttention. This encompasses the features capturing the characteristics of the other participants' attention focus (15 features).

  - Current attention target of every participant other than the focused participant
  - Time ratio for this participant to pay attention to the focused participant since the beginning of the session
  - Time ratio for this participant to pay attention to the focused participant in the past t seconds
  - Number of changes of the attention target of this participant since the beginning of the session

- (P)rosody. This is the prosodic information while this participant is speaking. Praat was also used to compute the prosodic features of the utterances of the other participants. The distribution of pitch ($F0$) and intensity are considered (36 features).

  - Current pitch
  - Standard deviation of pitch values from the beginning of the session
  - Standard deviation of pitch values in the latest t seconds (t is the window size)
  - Average pitch values in the latest t seconds (t is the window size)
  - Difference between current pitch and the average pitch in the period from the beginning of the session to now
  - Difference between current pitch and the average pitch in the latest t seconds
  - Current intensity
  - Standard deviation of intensity values since the beginning of the session
  - Standard deviation of intensity values in the latest t seconds (t is the window size)
  - Average intensity values in the latest t seconds (t is the window size)
  - Difference between current intensity and the average intensity in the period since the beginning of the session
  - Difference between current intensity and the average intensity in the latest t seconds (t is the window size)

- (H)ead activity. This is the activity of head movements of this participant, which is measured with the three-axis accelerometer attached on the head of each participant. The number of head movements and nods in the past 5 s and so on are taken into account (30 features).

  - Amount of head activity measured in the latest 0.1 s
  - Standard deviation of activity since the beginning of the session
  - Average activity since the beginning of the session
  - Difference between the last activity value and the average since the beginning of the session
  - Average activity in the latest t seconds (t is the window size)
  - Difference between the last activity value and the average in the latest t seconds (t is the window size)

- (V)erbal features. The Japanese morpheme segmentation tool Kuromoji (https://www.atilika.com/ja/products/kuromoji.html) was used to analyze the words of utterance transcriptions and count the numbers of verbs, nouns, new nouns, existing nouns, interjections, and fillers in utterances of the participants in the past t seconds (18 features).

  - Number of new nouns
  - Number of existing nouns
  - Number of nouns
  - Number of verbs
  - Number of interjections
  - Number of fillers

- (S)peech turn. This is the feature related to speech turns. The speaking periods are identified with the phonetic analysis tool Praat (http://www.fon.hum.uva.nl/praat/)—number of utterances, ratio of speaking, last speaker, length of utterances, and so on (23 features).
  - Speaking or not
  - Total number of this participant's utterances since the beginning of the discussion
  - Duration since the beginning of the current utterance
  - Ratio of speaking periods since the beginning of the session
  - Ratio of speaking periods in the last t seconds (t is the window size)
  - Duration since the beginning of the current state (*Speaking*, *Listening*, or *Idling*)
  - Last participant who started to speak

Table 2. Distribution of data instances regarding attention targets and situations.

Attention	Speaking	Listening	Idling
Table	68,477	143,854	90,473
Front	8805	14,700	5960
Right	3660	8199	2840
Left	3550	6843	2379
Total	84,492	173,596	101,652

4.4. Automatic Prediction Model

A nonlinear support vector machine (SVM) with a Gaussian kernel was used to develop the prediction models for the three situations. SVM complexity parameter C was explored among the values 0.5, 1, 5, 10, and 15. Radial basis function kernel parameter γ was explored among the values 0.001, 0.1, 0.1, and 1. All combinations of the parameters were tested, and the best results were found with the setting where C = 10.0 and γ = 0.01. Because of the bias in the number of instances in the Table and Front classes, the number of the instances of all classes was balanced to the smallest class.

The leave-one-person-out method was used in the evaluation. That is, the data of one participant were used for testing, and the others were used for training. This procedure was repeated 40 times so that all participants' data were tested. The final results were the sum of all 40 trials. Due to the bias in the number of instances in the Table and Front classes, smaller classes were oversampled with Synthetic Minority Oversampling TEchnique (SMOTE) [49] algorithm, and the larger classes were undersampled while keeping the total weight (amount) of the dataset both in the training and testing phases. Leave-one-participant-out cross validation was used in evaluating the performances of the models.

To determine the optimal window size t, each modality in varying window sizes from 1 to 10 s of speaking, listening, and idling models was tested individually. The results are shown in Figures 3–5, respectively. The results show that there were no large differences in performance regarding the window size. The performance of the model using all modalities was always better than that using only a single modality. The activity (H) modality always performed worst. This shows that it is least representative, and this is probably because of the smaller number of features of this feature set. Table 3 shows the optimal (achieving the highest F-measure scores) window sizes regarding the feature sets and prediction models. The results show that feature set A performed best in longer window sizes in all prediction models. This may imply that the cognition of others' attention targets is based on a relatively long period (8 s) rather than an instant. These optimal window sizes are used in the following analysis.

Table 3. Optimal window sizes (in seconds) regarding feature sets and prediction models.

	A	S	P	H	V
Speaking	8	3	4	3	1
Listening	6	2	3	5	1
Idling	3	3	3	3	1

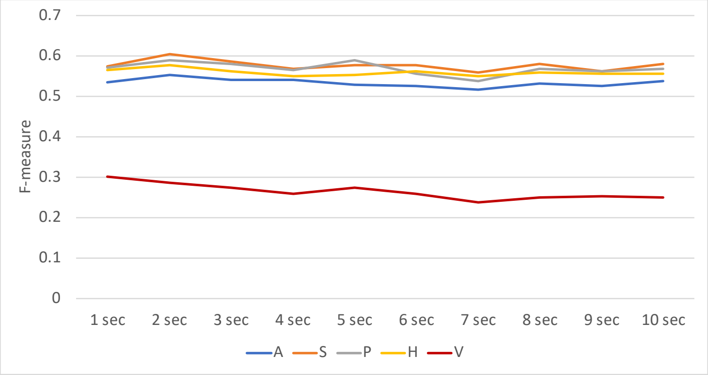

Figure 3. F-measure values (shown as the vertical axis) of each feature set in speaking model according to window size from 1 to 10 s.

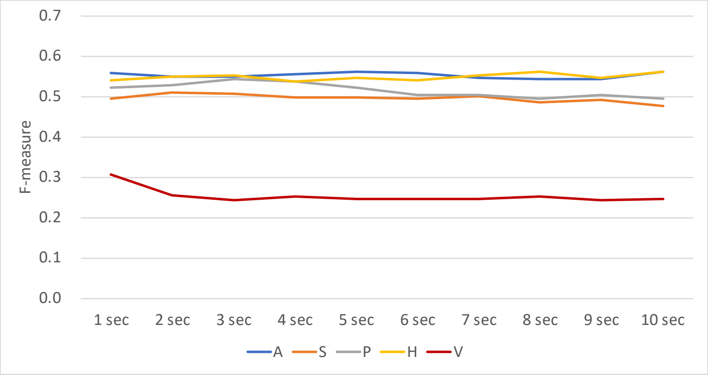

Figure 4. F-measure values (shown as the vertical axis) of each feature set in listening model according to window size from 1 to 10 s.

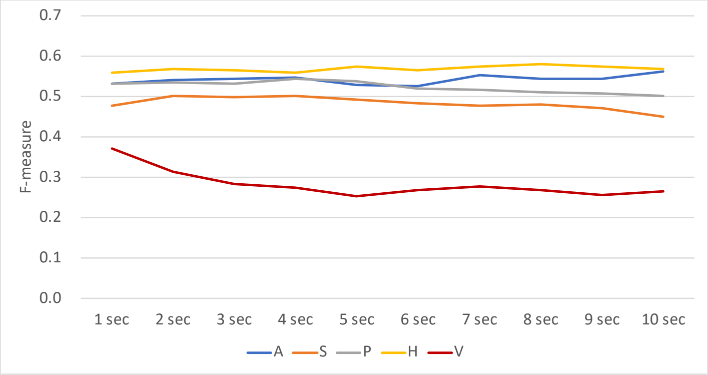

Figure 5. F-measure values (shown as the vertical axis) of each feature set in idling model according to window size from 1 to 10 s.

Table 4 shows the results regarding each attention target class with optimal window sizes. All three models were better at predicting the attention targets at side directions (Left and Right). This may be because the participants most often look forward or at the material on the table, and this means these two directions were less characteristic than the side ones. The performances of the three models are always in the order: listening > idling > speaking. This shows that the attention direction is more difficult to detect when the focused participant is speaking, and it is possible that the direction is more related to the content of the utterances. Because only nonverbal features were used in the classification, the performance in this aspect may be improved if verbal features are adopted. However, the reason why the listening model always outperformed the idling model may be because more available information was used.

Table 4. Classification results of speaking, listening, and idling models with all available feature sets and optimized window lengths.

	Attention	**Precision**	**Recall**	**F-measure**
Speaking	Table	0.423	0.515	0.464
	Front	0.411	0.393	0.402
	Right	0.501	0.425	0.460
	Left	0.541	0.522	0.532
	Overall	**0.464**	**0.460**	**0.460**
Listening	Table	0.412	0.396	0.404
	Front	0.622	0.605	0.613
	Right	0.622	0.672	0.646
	Left	0.664	0.655	0.659
	Overall	**0.580**	**0.581**	**0.580**
Idling	Table	0.471	0.610	0.532
	Front	0.556	0.514	0.534
	Right	0.452	0.385	0.416
	Left	0.655	0.570	0.610
	Overall	**0.536**	**0.529**	**0.528**

Table 5 shows the confusion matrices of the three prediction models. The speaking model generally performed worse than the other two models. Of all the attention target classes, Table was most frequent, and the Front class's recall, in particular, was low. A possible reason is that, when the participants were speaking, they did not pay attention as much to the others, so the attention targets were more random or more dependent on the contents of the speaker's utterances. Also, the participants may have changed their attention targets more dynamically and had more movements, looking at the material often, because they could not remember all the candidates. This caused it to be more difficult to distinguish the other attention targets from Table. Although the listening and idling models still mistakenly classified the attention targets as Table, the results were clearer, and this may be because the participants were more steady in these two situations. The overall tendency toward the Table class may be because this class had the largest number of instances and, consequently, had a larger variety of data.

Table 5. Confusion matrices of the proposed attention target prediction models in speaking, listening, and idling situations, respectively: the rows are actual classes and the columns are the predicted classes.

	Speaking				**Listening**				**Idling**			
	T	***F***	***R***	***L***	***T***	***F***	***R***	***L***	***T***	***F***	***R***	***L***
(T)able	51.5%	24.4%	12.7%	11.5%	39.6%	22.8%	19.7%	17.9%	61.0%	11.2%	15.4%	12.4%
(F)ront	28.0%	39.3%	19.7%	13.0%	16.4%	60.5%	13.9%	9.2%	29.5%	51.4%	10.9%	8.2%
(R)ight	25.2%	24.6%	44.3%	5.9%	19.6%	7.1%	67.2%	6.2%	21.3%	31.8%	38.5%	8.4%
(L)eft	24.2%	11.8%	11.7%	52.2%	21.2%	7.1%	6.2%	65.5%	27.4%	7.4%	8.2%	57.0%

Figure 6 depicts the overall F-measure values regarding all possible combinations of feature sets and classification models with optimal window sizes. The results showed that, not always, but usually, richer information had better results. The performances in the F-measure of the models were moderate and roughly ranged between 0.4 and 0.6. For all the models, feature sets A and P contributed most to the classification performance, while feature sets V and H were not as effective. This implies that mutual attention and speech activities attracted the participants' attention most, while the contents of utterances and head movements were not as distinguishing in various attention combinations of the participants.

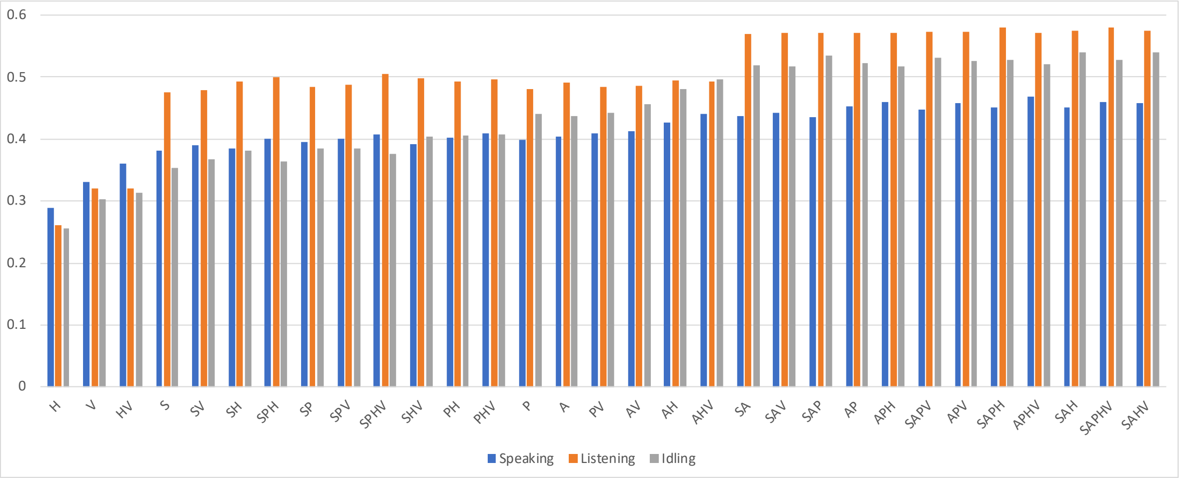

Figure 6. F-measure values (vertical axis) showing the performance of the attention focus prediction models in speaking, listening, and idling situations with all 31 combinations of feature sets: the bars are sorted in the order of the performances of listening models.

As described in Section 2, Stefanov et al. [43] is the only work that we could find that shares similar objectives at this moment. Although their work cannot be directly compared with our work in the sense of performance due to very different settings of the underlying dataset, purpose, and evaluation metrics, the same tendency in evaluation results can be found:

- The accuracy in the speaking state is lower than in the listening state. The reason is supposed to be the same: when a person is taking an intentional action (speaking), then his/her head motion is less depend on the interlocutors.
- The accuracy is higher when the features from more modalities are available.

5. Nodding Model

The second prediction model is the one that determines whether the robot should nod. Following a similar procedure as that for the attention model, the periods when the participants were nodding were manually labeled (Table 6). Nodding behavior was simplified to be only relevant to utterances; rather than the 0.1-s time slices of the attention model, the prediction timing of the nodding model was defined to be at the end of each utterance. That is, when any of the participants finished her or his utterance, this instant was treated as the prediction timings of all the four participants. Therefore, the resulting data instances share the same number of annotation labels in Table 6. In addition, for the same reason, only the models for speaking and listening situations were developed.

Table 6. Annotation results on the nods of the participants of the data corpus: the columns, "avg," "max," and "min" shows the average, maximum, and minimum durations of individual labels in seconds, respectively, and Speaking and Listening are the resulting instance numbers in corresponding models.

Nodding	Labels	Avg	Max	Min	Speaking	Listening
Yes	621	1.3	9.9	0.1	324	667
No	661	53.3	609.6	0.3	5213	15,197

As with the attention models, the five feature sets—verbal, attention, speech turn, prosody, and head activities—were extracted from the same data corpus. Instead of fixed-length windows for extracting features, in the nodding model, the features were extracted utterance by utterance. The features that could not be extracted in this way were then omitted. Tables 7 and 8 show the performance of the two-class classification model in the two situations: speaking and listening. Considering the chance level (50%) of a two-class classification problem, the performance was only moderate. The recall of "Yes" during the speaking was exceptionally low.

This may also imply that the reasons why the participant nodded were more diverse and more difficult to predict from the behaviors of other participants. The performances of all combinations of feature sets are shown in Figure 7. Unlike for the attention model, it was found that verbal features were most effective in classifying nods. Also, listening models usually performed better than those for the other situations. This may imply that people's behaviors are more determined by the others when they are listening but are not as predictable when humans are speaking. In addition, the performance was usually better when there were more modalities available during a listening situation, but it was not necessarily better in speaking situations.

Table 7. Yes/No classification results of the proposed nodding prediction models with all available feature sets.

	Nod	Precision	Recall	F-Measure
Speaking	Yes	0.715	0.487	0.579
	No	0.626	0.816	0.708
	Overall	**0.669**	**0.656**	**0.645**
Listening	Yes	0.718	0.617	0.663
	No	0.664	0.757	0.707
	Overall	**0.691**	**0.687**	**0.685**

Table 8. Confusion matrices of the proposed nodding prediction models in speaking and listening situations, respectively: the rows are actual classes and the columns are the predicted classes.

	Speaking		**Listening**	
	Yes	***No***	***Yes***	***No***
Yes	48.9%	51.1%	64.2%	35.8%
No	18.2%	81.8%	27.5%	72.5%

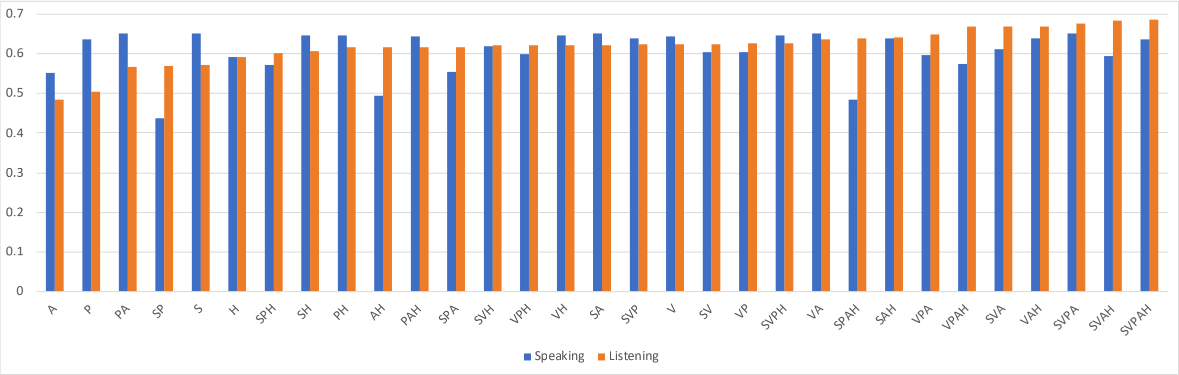

Figure 7. F-measure values (vertical axis) showing the performance of the nodding prediction models in speaking and listening situations for all 31 possible combinations of feature sets: the bars are sorted in the order of the performances of listening models.

6. Integrating the Models into a Robot

The proposed head movement models can be integrated in a multimodal framework, as shown in Figure 8. The robot engages in a group discussion task with three other participants (humans or other robots) and acquires the activities of other participants from video/audio and other sensory information. Multimodal features are extracted from these pieces of information using the preprocessing modules: the face recognizer, motion analyzer (accelerometers attached on the participants' heads), prosody analyzer, speech recognizer, and Japanese morpheme analyzer.

All available feature values were then integrated by the multimodal fuser module. It identifies the correspondence of the information coming from different sources with timestamps and generates feature vectors of input information at each prediction time point in real time. For the features to which past information is referred, data history is kept by this module. The multimodal inputs are propagated to the dialog manager (DM) module, which decides the robot's utterances, as well as other intentional behaviors of the robot. Multimodal input information is also sent to the prediction modules of nodding and attention (NP and AP, respectively). These two modules determine the timings of the robot's spontaneous head turns (changes of attention focuses) and nodding. The outputs from the DM, NP, and AP modules are then gathered in the controller (RC) module, which physically controls the robot. The RC module selects the actual actions to perform and resolve the contradictions when there are more than one module trying to move the same parts of the robot. A possible policy for the resolution is granting higher priority to intentional actions from the DM module.

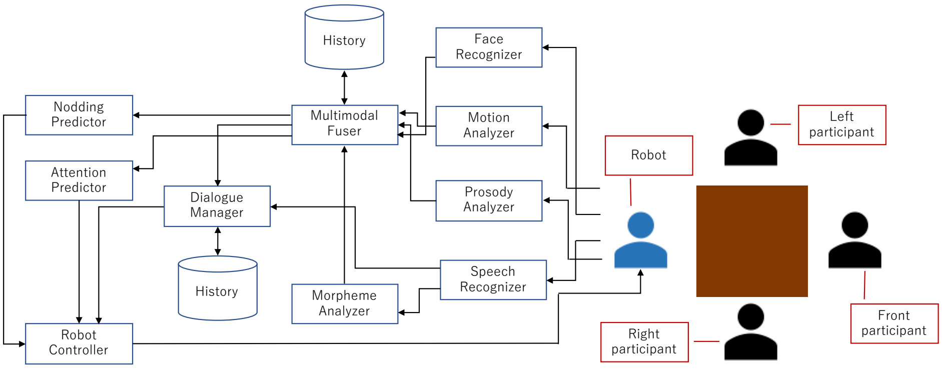

Figure 8. Proposed multimodal framework with the attention/nodding prediction modules integrated.

A robot's head is driven by motors and generally can perform neither subtle nor fast movements at the human level. However, the proposed attention model generates outputs every 0.1 s, and this provides the opportunity for fine control of the robot's movements, but the output sequence cannot be realized in all cases because of the robot's physical constraints. The models were implemented with VStone's Sota robot as an example. Sota is a desktop humanoid robot designed for communication partner applications. It has only an upper body, and its height is 28 cm. In addition to the four DOFs on its two arms, its head has three DOFs (roll-pitch-yaw), and its torso can rotate in the horizontal direction (yaw). Its head has a relatively simple structure, and there are no moving parts on its face, but the light-emitting diodes behind its eyes and mouth can be lighted.

Sota was programmed to show its attention to the Left, Right, and Front participants and Table, as well as the switching movements among these directions. The switching time between two attention targets was measured to take as long as 0.6 s. Including a steady 0.1 s to show a certain attention target for a minimum period, direction changing could not be realized within 0.7 s in the case of Sota. Therefore, a filter with a 0.7-s window was applied to the outputs of the model. Every time Sota was available to perform the next action, the model outputs in the last 0.7 s were examined to determine the next attention of Sota. Because there are seven outputs from the attention model, more than one attention can be in the sequence; the one with most instances is then selected as the joint output (Figure 9).

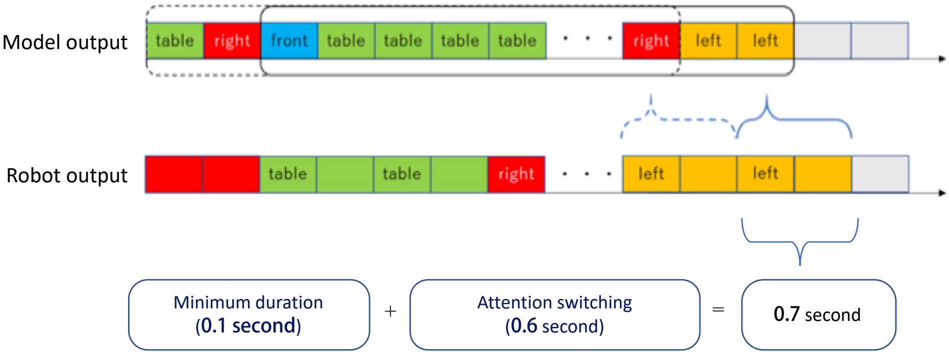

Figure 9. Conceptual diagram of the filtering from the outputs of attention model to actual robot movements.

The proposed nodding model generated an output if any of the participants finished one utterance. Unlike for the attention model, the nodding model's outputs were not in a determined period. Depending on the actual progress of the group discussion conversation, the nodding of the robot could be very short and frequent, so that it was perceived to be annoying and unnatural. Therefore, another filter was applied to the outputs of the nodding model. The average length of the nodding period of the whole data corpus, 1.3 s, was adopted as the unit of Sota's nodding movement. The concept is depicted in Figure 10. When the nodding model generates a *nod* command, Sota nods for 1.3 s. When there are several nodding outputs generated, they are concatenated until the last 1.3-s period ends.

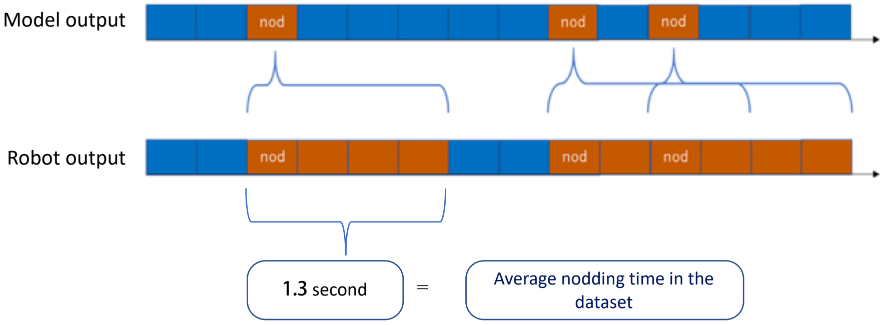

Figure 10. Conceptual diagram of the filtering from the outputs of the nodding model to actual robot movements.

Because of the control mechanism of Sota, after one movement command sent to it, the command cannot be interrupted and switched to another movement. The commands need to be sent in a one-by-one manner, and the command sequence needs to be planned beforehand. Because both the attention model and the nodding model are trying to move Sota's head, it is possible that contradictions happen. Figure 11 shows how the contradictions can be resolved between the two proposed models. Because nods have stronger semantic meanings in showing agreement, nods are treated with higher priority: at time point A, the Front output of the attention model is overwritten by the nod from nodding model. As described previously, the switching of attention targets takes 0.7 s, and Sota cannot do other actions during that period. The nod at time point B is overwritten by Right attention. Sota then starts to nod immediately after the hardware is available. Then, the last Left is overwritten by that nod.

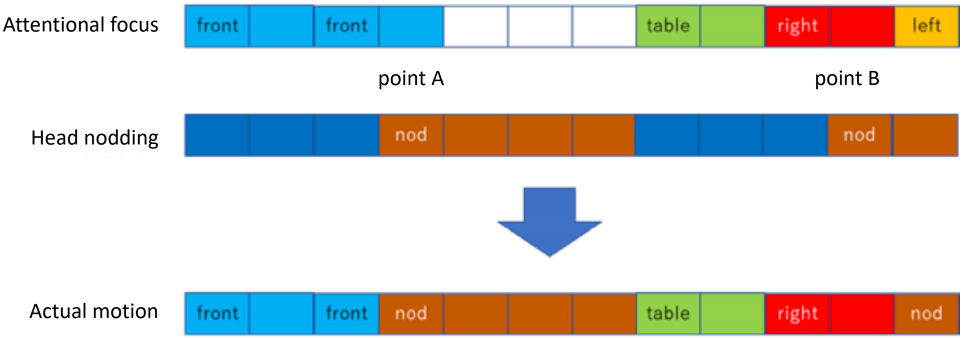

Figure 11. Conceptual diagram of the integration of the two proposed models.

Simulate the effect of the model performance on unseen inputs

7. Subjective Evaluation of the Models

In Sections 4 and 5, the proposed prediction models were evaluated in terms of accuracy. However, there is no single "correct answer" for human behaviors in a given situation. An accuracy of 100% cannot be expected, and "the higher the accuracy, the more realistic" is not necessarily true. To verify whether the head movements generated by the proposed models are perceived to be realistic, a subject evaluation is required. Therefore, a second subject experiment was conducted to evaluate the perception of predicted head movements in the aspects of naturalness and realism.

A straightforward evaluation experiment is to fully implement a robot-training system and evaluate its actual usage. Instead of that, a subject experiment based on perception from video clips was conducted. The evaluation was done with video clips because we wanted to limit the comparison on only the head movements of the robot but not on any other factors. For example, the contents of the utterances made by the WOZ operator may have heavy influences on the impression of the overall system, but it is not possible to reproduce exactly the same situation and event sequences among all of the three compared conditions. By using the current experiment setting, we can ensure that the subjects only perceive the subtle differences of behavior models, and all other factors are maintained to be exactly the same. The purpose was to conduct a fully controlled experiment on only two behaviors, not overall impression or system effectiveness. Video clips like the one shown in Figure 12 were observed and evaluated by recruited subjects. There were 10 groups by four subjects in the data corpus; hence, there were 40 video clips of individual subjects in total. Twenty of them were randomly selected and were replaced by the video of Sota, while the audio track was still the subject's original voice. Each video clip was cut from the middle part of the original experimental video data with a length of approximately 1 min. Sota's parts in the video clips were determined by the following three variations.

Model condition: Sota's head movements were determined by the proposed head movement models. The models used were created in a leave-one-subject-out manner to simulate the behavior generated for an unseen subject. That is, a dedicated model was trained with the other 39 subjects, except the one replaced by Sota.

Human condition: Sota's head movements were determined by the original labeled data of the subject.

Statistical condition: Sota's head movements were determined by statistical probabilities conducted from the data corpus (Table 9).

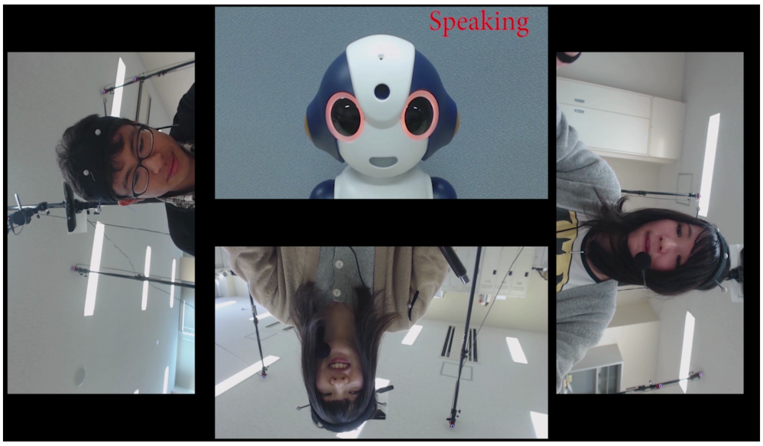

Figure 12. One scene of the video clips used in the subjective evaluation experiment.

Table 9. Probability (%) distribution of the statistical condition.

Situation	Speaking						Listening						Idling			
Model	**Attention**				**Nod**		**Attention**				**Nod**		**Attention**			
Class	T	F	L	R	Y	N	T	F	L	R	Y	N	T	F	L	R
Probability	81.0	11.4	4.2	4.4	6.2	93.8	82.8	8.5	3.9	4.8	4.4	95.6	89.0	5.8	2.5	2.8

The command sequences in all of the three conditions were passed through the filters explained in Section 6. In all, 140 subjects (84 male and 56 female, average age 39.3 with S.D. 8.8) were recruited as anonymous evaluators from a cloud sourcing service, Lancers (https://www.lancers.jp/). The composed video clips with Sota and the other three subjects shown at the same time were viewed and evaluated by the recruited evaluators. Each of them evaluated three sections of video clips uploaded to a dedicated Web site. Each section contained three variations that were listed above one session, where one of the four subjects was replaced by Sota. The sessions were arranged randomly, and the video clips in each section were ordered randomly. Therefore, each participant evaluated three sessions (nine video clips), and each session was evaluated by seven evaluators. The naturalness of Sota's head movements in reacting to the other three subjects' behavior was evaluated in the following three aspects with scores from one to ten.

Attention target and the timing to switch targets: Does Sota pay attention to appropriate targets and switches its attention to focus on the next target at appropriate timings, considering the context of the discussion?

Timing of nods: Does Sota nod at the timings that are perceived to be natural?

Overall impression: Are Sota's head movements perceived to be natural in both of the two aspects above?

The evaluators input the scores as their first impression right after watching the video clips. Table 10 summarizes the results of the subjective experiment regarding the scores of the perception of

Sota's attention, nodding, and overall head movements. The data values are the mean and standard deviation of raw scores given by the participants, and "Corr." means the correlation coefficients between the scores and the number of times of switching between the attention targets in the attention row, the correlation coefficients between the scores and number of nods in the nodding row, or both in the overall row. Values in the test column are the results of a Friedman test.

First, the human model was always evaluated to be worse. Observation showed that the human model moved less frequently than the other two conditions. This may also be because of the very simplified head behaviors. The proposed model was best at performing nodding. However, there were no significant differences among the three conditions in the attention model and overall impression. Although the statistical condition had a negative correlation with the number of movements, the model condition was evaluated in a way close to the human condition.

Table 10. Summary of the results of subjective evaluation (* : $p < 0.05$, n.s. : not significant).

Item		Model	Human	Statistical	Test
Attention	mean	4.82	4.80	4.95	n.s.
	S.D.	2.15	2.22	2.00	
	Corr.	0.58	0.52	−0.10	
Nodding	mean	5.06	4.66	4.88	*
	S.D.	2.15	2.23	1.97	
	Corr.	0.34	0.72	0.36	
Overall	mean	4.91	4.78	4.90	n.s.
	S.D.	2.07	2.18	1.89	
	Corr.	0.60	0.60	−0.21	

8. Conclusions and Future Direction

To cause the robot's head to behave in a humanlike manner in a group discussion with human users, two data-driven generation models were proposed based on the features from the other peer participants and regarding the head of the robot. One feature is the control of attentional focus, and the other is the timing of the head nod. These generation models were considered in three situations—speaking, listening, and idling—according to the robot's speaking state. The models were developed with an SVM trained on low-level multimodal features extracted from a data corpus that contained 4 people × 10 groups × 15 min = 10 h of recordings of video, audio, and sensory information. This work was based on a data corpus containing 2.5 h of the discussion sessions of 10 four-person groups. Then, the performances of these models were measured based on their accuracy in comparison with the training data and prediction models. Also, a subject experiment was performed for evaluating the perceived naturalness. The evaluation was based on the observation of video clips where one of the subjects was replaced by a robot performing head movements in the three conditions: the proposed models, statistical model, and raw human data. The experiment results showed that there was no significant difference from the original human behaviors in the data corpus.

The development of a fully autonomous robot capable to join a group discussion with human users requires implementing various functions to be realized and cannot be done in the near future. We considered that the most difficult part is real-time decision making. In the target system, this means the triggering of an appropriate and intentional action, which is supposed to improve the user's skill in the current situation. Before a fully autonomous robot or agent, a hybrid Wizard-of-Orz (WOZ) system should be practical as an intermediate stage. The decision-making part can be done by a human operator, while the spontaneous nonverbal behavior generation can be aided by the proposed generation models. Benefits from the end-to-end design, the same model should be applicable both in a fully autonomous system (Figure 8) or a WOZ system, where the dialogue manager module is replaced by a human operator.

This work was based on the hypothesis that a person's behavior is heavily affected by the other participants' behaviors in a group conversation. The results show that the accuracy of the models varied according to the conditions. Attentional focus from other participants was more predictable when the robot was in a passive state (listening or idling). This coincides with the results of the previous work done by Stefanov et al. [43]. However, nodding has fewer differences. For future work, it is planned to refine the features to improve the performance of the models further, incorporating more nonverbal information, such as postures, and more-detailed prosodic information, such as mel-frequency cepstrum coefficients (MFCC).

The proposed models depend merely on the features from the other participants but are independent of the robot's intention. This is expected to become a limitation to both the accuracy and perceived naturalness. The features reflecting the robot's intention may improve the models' performance. Adding more verbal features, such as the intention of utterances, is also planned to improve the performance of the speaking model. The relationship between the terms and who spoke them may be useful for this. For example, when the focused participant is speaking a term that was previously spoken by another participant, he or she may pay attention to that participant more. Also, other intentions, such as the taking and release of turns, can be included in the modeling.

In addition to the improvement of classification performance, implementing the framework using a a communicational robot or a virtual agent in a VR environment is planned. The robot used in our experiment was physically small. For the experiment, the participants who evaluated the video clips did not see the robot in the real world. We assume that the influence of its physical size should not be strong. However, if we use the same robot in the fully implemented system, its size may have some side influences. Therefore, we will use a human-size robot or virtual agents in a VR environment when we implement the full system. Because of the "Mona Lisa effect" mentioned in previous research [50], the users actually cannot correctly distinguish the gaze direction or the head orientation of graphical agents rendered on a 2D surface. Therefore, 2D virtual agents cannot be applied in the target application. The agent has to be perceived in a 3D space, either a physical object, i.e., a robot, or a graphical one in a virtual reality environment. The models then have to been tuned according to hardware/software implementation constraints (e.g., the rotation speed of the robot's head). The human participants may behave differently with robots than other humans, so it is also planned to investigate this aspect in a participant experiment using the implemented robot and to evaluate whether the whole system can improve the effect on the performance of the users through using the system.

Author Contributions: Conceptualization, H.-H.H.; Data curation, H.-H.H. and S.K.; Formal analysis, H.-H.H.; Funding acquisition, H.-H.H. and K.K.; Investigation, H.-H.H.; Methodology, H.-H.H.; Project administration, H.-H.H.; Resources, H.-H.H., K.K. and T.N.; Software, H.-H.H. and S.K.; Supervision, H.-H.H., K.K. and T.N.; Validation, H.-H.H.; Writing—original draft, H.-H.H.; Writing—review & editing, H.-H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partially supported by KAKENHI: Grant-in-Aid for Scientific Research (A), Grant No. 19H01120.

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

References

1. Chollet, M.; Ochs, M.; Pelachaud, C. Mining a Multimodal Corpus for Non-Verbal Signals Sequences Conveying Attitudes. In Proceedings of the 9th Edition of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, 26–31 May 2014; pp. 3417–3424.
2. Jones, H.; Chollet, M.; Ochs, M.; Sabouret, N.; Pelachaud, C. Expressing social attitudes in virtual agents for social coaching. In Proceedings of the Autonomous Agents and Multi-Agent Systems (AAMAS'14), Paris, France, 5–9 May 2014.
3. Baur, T.; Damian, I.; Gebhard, P.; Porayska-Pomsta, K.; Andre, E. A Job Interview Simulation: Social Cue-based Interaction with a Virtual Character. In Proceedings of the 2013 International Conference on Social Computing (SocialCom 2013), Washington, DC, USA, 8–14 September 2013.

4. Traum, D. Issues in Multiparty Dialogues. In Proceedings of the Advances in Agent Communication, International Workshop on Agent Communication Languages (ACL'03), Halifax, NS, Canada, 11–13 June 2003; pp. 201–211.
5. Huang, H.H.; Baba, N.; Nakano, Y. Making Virtual Conversational Agent Aware of the Addressee of Users' Utterances in Multi-user Conversation from Nonverbal Information. In Proceedings of the 13th International Conference on Multimodal Interaction (ICMI'11), Alicante, Spain, 14–18 November 2011; pp. 401–408.
6. Baba, N.; Huang, H.H.; Nakano, Y. Addressee Identification for Human-Human-Agent Multiparty Conversations in Different Proxemics. In Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye Gaze and Multimodality, Santa Monica, CA, USA, 26 October 2012.
7. Nakano, Y.; Baba, N.; Huang, H.H.; Hayashi, Y. Implementation and Evaluation of Multimodal Addressee Identification Mechanism for Multiparty Conversation Systems. In Proceedings of the 15th International Conference on Multimodal Interaction (ICMI 2013), Sydney, Australia, 9–13 December 2013.
8. Jokinen, K.; Parkson, S. Synchrony and copying in conversational interactions. In Proceedings of the 3rd Nordic Symposium on Multimodal Interaction, Helsinki, Finland, 27–28 May 2011; pp. 18–24.
9. Kingstone, A.; Friesen, C.K.; Gazzaniga, M.S. Reflexive Joing Attention depends on Lateralized Cortical Connections. *Psychol. Sci.* **2000**, *11*, 159–166. [CrossRef] [PubMed]
10. Ristic, J.; Friesen, C.K.; Kingstone, A. Are eyes special? It depends on how you look at it. *Psychon. Bull. Rev.* **2002**, *9*, 507–513. [CrossRef] [PubMed]
11. Downing, P.; Dodds, C.; Bray, D. Why does the gaze of others direct visual attention? *Vis. Cogn.* **2004**, *11*, 71–79. [CrossRef]
12. Otsuka, K. Multimodal Conversation Scene Analysis for Understanding People's Communicative Behaviors in Face-to-Face Meetings. In *Symposium on Human Interface*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 171–179.
13. Kendon, A. some functions of gaze direction in social interaction. *Acta Psychol.* **1967**, *26*, 22–63. [CrossRef]
14. Argyle, M.; Cook, M. *Gaze and Mutual Gaze*; Cambridge University Press: Cambridge, UK, 1976.
15. Duncan, S. Some Signals and Rules for Taking Speaking Turns in Conversations. *J. Personal. Psychol.* **1972**, *23*, 283–292. [CrossRef]
16. Vertegaal, R.; Slagter, R.; van der Veer, G.; Nijholt, A. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, WA, USA, 31 March–5 April 2001; pp. 301–308.
17. Takemae, Y.; Otsuka, K.; Mukawa, N. Video Cut Editing Rule Based on Participants' Gaze in Multiparty Conversation. In Proceedings of the 11th ACM International Conference on Multimedia, Berkeley, CA, USA, 2–8 November 2003.
18. Katzenmaier, M.; Stiefelhagen, R.; Schultz, T. Identifying the Addressee in Human-Human-Robot Interactions based on Head Pose and Speech. In Proceedings of the 6th international conference on Multimodal interfaces (ICM 2004), State College, PA, USA, 13–15 October 2004.
19. Gratch, J.; Okhmatovskaia, A.; Lamothe, F.; Marsella, S.; Morales, M.; van der Werf, R.; Morency, L.P. Virtual Rapport. In Proceedings of the 6th International Conference on Intelligent Virtual Agents (IVA 2006), Marina Del Rey, CA, USA, 21–23 August 2006; pp. 14–27.
20. Ishii, C.T.; Liu, C.; Ishiguro, H.; Hagita, N. Head motions during dialogue speech and nod timing control in humanoid robots. In Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010), Osaka, Japan, 2–5 March 2010; pp. 293–300.
21. Aran, O.; Gatica-Perez, D. One of a Kind: Inferring Personality Impressions in Meetings. In Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI 2013), Sydney, Australia, 9–13 December 2013.
22. Okada, S.; Nakano, Y.; Hayashi, Y.; Takase, Y.; Nitta, K. Estimating Communication Skills using Dialogue Acts and Nonverbal Features in Multiple Discussion Datasets. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016), Tokyo, Japan, 12–16 November 2016; pp. 169–176.
23. Schiavo, G.; Cappelletti, A.; Mencarini, E.; Stock, O.; Zancanaro, M. Overt or Subtle? Supporting Group Conversations with Automatically Targeted Directives. In Proceedings of the 19th international conference on Intelligent User Interfaces (IUI 2014), Haifa, Israel, 24–27 February 2014; pp. 225–234.

24. Raducanu, B.; Vitria, J.; Gatica-Perez, D. You are fired! Nonverbal role analysis in competitive meetings. In Proceedings of the 2009 IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP 2009), Taipei, Taiwan, 19–24 April 2009.
25. Muralidhar, S.; Nguyen, L.S.; Frauendorfer, D.; Odobez, J.M.; Mast, M.S.; Gatica-Perez, D. Training on the Job: Behavioral Analysis of Job Interviews in Hospitality. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016), Tokyo, Japan, 12–16 November 2016; pp. 84–91.
26. Oertel, C.; Mora, K.A.F.; Gustafson, J.; Odobez, J.M. Deciphering the Silent Participant: On the Use of Audio-Visual Cues for the Classification of Listener Categories in Group Discussions. In Proceedings of the 17th ACM on International Conference on Multimodal Interaction (ICMI 2015), Seattle, WA, USA, 9–13 November 2015; pp. 107–114.
27. Bevacqua, E.; Pammi, S.; Hyniewska, S.; Schroder, M.; Pelachaud, C. Multimodal Backchannels for Embodied Conversational Agents. In Proceedings of the 10th International Conference on Intelligent Virtual Agents (IVA 2010), Philadelphia, PA, USA, 20–22 September 2010.
28. Oertel, C.; Lopes, J.; Yu, Y.; Mora, K.A.F.; Gustafson, J.; Black, A.W.; Odobez, J.M. Towards Building an Attentive Artificial Listener: On the Perception of Attentiveness in Audio-Visual Feedback Tokens. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016), Tokyo, Japan, 12–16 November 2016; pp. 21–28.
29. Agarwal, P.; Moubayed, S.A.; Alspach, A.; Kim, J.; Carter, E.J.; Lehman, J.; Yamane, K. Imitating human movement with teleoperated robotic head. In Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2016), New York, NY, USA, 26–31 August 2016.
30. Cazzato, D.; Cimarelli, C.; Sanchez-Lopez, J.L.; Olivares-Mendez, M.A.; Voos, H. Real-Time Human Head Imitation for Humanoid Robots. In Proceedings of the 3rd International Conference on Artificial Intelligence and Virtual Reality (AIVR 2019), Singapore, 27–29 July 2019; pp. 65–69.
31. Ondras, J.; Celiktutan, O.; Sariyanidi, E.; Gunes, H. Automatic replication of teleoperator head movements and facial expressions on a humanoid robot. In Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2017), Lisbon, Portugal, 28–31 August 2017; pp. 745–750.
32. Admoni, H.; Scassellati, B. Social eye gaze in human-robot interaction: A review. *J. Hum.-Robot Interact.* **2017**, *6*, 25–63. [CrossRef]
33. Leite, I.; McCoy, M.; Lohani, M.; Ullman, D.; Salomons, N.; Stokes, C.; Rivers, S.; Scassellati, B. Emotional Storytelling in the Classroom: Individual versus Group Interaction between Children and Robots. In Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2015), Portland, OR, USA, 2–5 March 2015; pp. 75–82.
34. Vazquez, M.; Carter, E.J.; McDorman, B.; Steinfeld, J.F.A.; Hudson, S.E. Towards Robot Autonomy in Group Conversations: Understanding the Effects of Body Orientation and Gaze. In Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2017), Vienna, Austria, 6–9 March 2017; pp. 42–52.
35. Loth, S.; Huth, K.; Ruiter, J.P.D. Automatic detection of service initiation signals used in bars. *Front. Psychol.* **2013**, *4*, 557. [CrossRef] [PubMed]
36. Keizer, S.; Foster, M.E.; Wang, Z.; Lemon, O.J. Machine Learning for Social Multi-Party Human-Robot Interaction. *ACM Trans. Interact. Intell. Syst.* **2014**, *4*. [CrossRef]
37. Sakai, K.; Libera, F.D.; Yoshikawa, Y.; Ishiguro, H. Generation of Bystander Robot Actions Based on Analysis of Relative Probability of Human Actions. *J. Adv. Comput. Intell. Intell. Inform.* **2017**, *21*, 686–696. [CrossRef]
38. Bohus, D.; Horvitz, E. Managing Human-Robot Engagement with Forecasts and... Um... Hesitations. In Proceedings of the 16th International Conference on Multimodal Interaction (ICMI 2014), Istanbul, Turkey, 12–16 November 2014; pp. 2–9.
39. Sidner, C.L.; Lee, C.; Lesh, N. *Engagement Rules for Human-Robot Collaborative Interaction*; Technical Report TR2003-50; Mitsubishi Electric Research Laboratories: Cambridge, MA, USA, 2003.
40. Mutlu, B.; Kanda, T.; Forlizzi, J.; Hodgins, J.; Ishiguro, H. Conversational Gaze Mechanisms for Humanlike Robots. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2012**, *1*, 12:1–12:33. [CrossRef]
41. Bohus, D.; Horvitz, E. Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings. In Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, London, UK, 11–12 September 2009.

42. Sakai, K.; Ishi, C.T.; Minato, T.; Ishiguro, H. Online speech-driven head motion generating system and evaluation on a tele-operated robot. In Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015), Kobe, Japan, 31 August–4 September 2015.
43. Stefanov, K.; Salvi, G.; Kontogiorgos, D.; Kjellström, H.; Beskow, J. Modeling of Human Visual Attention in Multiparty Open-World Dialogues. *ACM Trans. Hum.-Robot Interact. (THRI)* **2019**, *8*, 8. [CrossRef]
44. Stefanov, K.; Beskow, J. A Multi-party Multi-modal Dataset for Focus of Visual Attention in Human-human and Human-robot Interaction. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portoroz, Slovenia, 23–28 May 2016; pp. 4440–4444.
45. Nihei, F.; Nakano, Y.I.; Hayashi, Y.; Huang, H.H.; Okada, S. Predicting Influential Statements in Group Discussions using Speech and Head Motion Information. In Proceedings of the 16th International Conference on Multimodal Interaction (ICMI 2014), Istanbul, Turkey, 12–16 November 2014; pp. 136–143.
46. Kickul, J.; Neuman, G. Emergent Leadership Behaviors: The Function of Personality and Cognitive Ability in Determining Teamwork Performance and KSAs. *J. Bus. Psychol.* **2000**, *15*, 27–51. [CrossRef]
47. Boersma, P.; Weenink, D. Praat: Doing Phonetics by Computer [Computer Software] Version 6.0.40. Available online: http://www.praat.org/ (accessed on 26 May 2019) .
48. Lausberg, H.; Sloetjes, H. Coding gestural behavior with the NEUROGES–ELAN system. *Behav. Res. Methods* **2009**, *41*, 841–849. [CrossRef] [PubMed]
49. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
50. Morikawa, O.; Maesako, T. HyperMirror: Toward Pleasant-to-use Video Mediated Communication System. In Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work (CSCW'98), Seattle, WA, USA, 14–18 November 1998; ACM Press: New York, NY, USA, 1998; pp. 149–158.