Electronic Preprint for Journal of Information Processing Vol.29

Regular Paper

Methods for Efficiently Constructing Text-dialogue-agent System using Existing Anime Characters

Ryo Ishii[1,a)] Ryuichiro Higashinaka[1] Koh Mitsuda[1] Taichi Katayama[1] Masahiro Mizukami[2]
Junji Tomita[1] Hidetoshi Kawabata[3] Emi Yamaguchi[3] Noritake Adachi[3] Yushi Aono[1]

Received: April 13, 2020, Accepted: October 6, 2020

Abstract: Starting from their early years, many persons dream of being able to chat with their favorite anime characters. To make such a dream possible, we propose an efficient method for constructing a system that enables users to text chat with existing anime characters. We tackled two research problems to generate verbal and nonverbal behaviors for a text-chat agent system utilizing an existing character. A major issue in creating verbal behavior is generating utterance text that reflects the personality of existing characters in response to any user questions. To cope with this problem we propose use of role play-based question-answering to efficiently collect high-quality paired data of user questions and system answers reflecting the personality of an anime character. We also propose a new utterance generation method that uses a neural translation model with the collected data. Rich and natural expressions of nonverbal behavior greatly enhance the appeal of agent systems. However, not all existing anime characters move as naturally and as diversely as humans. Therefore, we propose a method that can automatically generate whole-body motion from spoken text in order to give the anime characters natural, human-like movements. In addition to these movements, we try to add a small amount of characteristic movement on a rule basis to reflect personality. We created a text-dialogue agent system of a popular existing anime character using our proposed generation methods. As a result of a subjective evaluation of the implemented system, our methods for generating verbal and nonverbal behavior improved the impression of the agent's responsiveness and reflected the personality of the character. Since generating characteristic motions with a small amount of characteristic movement on the basis of heuristic rules was not effective, our proposed motion generation method which can generate the average motion of many people, is useful for generating motion for existing anime characters. Therefore, our proposed methods for generating verbal and nonverbal behaviors and the system-construction method are likely to prove a powerful tool for achieving text-dialogue agent systems for existing characters.

Keywords: text-dialogue-agent system, existing anime character, efficient construction method, utterance generation, motion generation

1. Introduction

The everyday use of robots and conversation agents by way of smartphones has led to the need for dialogue techniques that allow freely chatting with these agents and robots. In particular, in recent years, research on constructing dialogue-agent systems for entertainment and counseling has been actively conducted, and attention has been paid to the development of actual services. We are aiming to realize a dialogue-agent system that allows text chatting with conversational agents via existing anime characters that have natural movements. Many persons have had their own favorite anime characters since childhood and even as adults still wish to be able to talk with these characters. However, to date, the construction of a dialogue system that reflects the personality of a real animated character has not been realized to our knowledge. Thus, we here propose a system-construction method for making such a dream possible in a realistic way. The definition of personality in this study means a character's behavioral tendency to respond to an input stimulus, typically an utterance from a conversational partner which is one of the aspects of personality.

We worked to develop a dialogue-agent system for an existing anime character that operates on the basis of text chat as a first attempt to realize such a system. In order for the system to generate the natural behavior of a character, two elemental technologies are needed: verbal behavior (utterance) generation for responding to any user utterances, and nonverbal behavior (body motion) generation for the system's utterances. We tackled these two research problems to allow generating utterances and body motions of an existing character with the system.

Some previous studies used the Big-5 personality traits to represent the personality of conversational agents with utterance and body motion [1], [2], [3]. Although certain aspects of personality can be converted on the basis of the Big-5, such dimensions are too broad and rough when agents need to generate fine-grained answers related to their personality. Recently, a growing number of studies have been emerging that use textual descriptions of the characters in question [4], [5], [6]. Our study is in line with such recent studies since we want to generate utterances that are consistent in terms of utterance content and body motion. These

[1] NTT Media Intelligence Laboratories, NTT Corporation, Yokosuka, Kanagawa 239–08474, Japan
[2] NTT Communication Science Laboratories, NTT Corporation, "Keihanna Science City", Kyoto 619–0237, Japan
[3] DWANGO Co., Ltd., Chuo, Tokyo 104–0061, Japan
[a)] ryo.ishii.ct@hco.ntt.co.jp

prior techniques do not consider the generation of utterances and motions that reflect the impression given by an actual anime character.

In the generation of utterances, it is a major issue to be able to generate utterances in the form of text that reflect the personality of existing characters in response to any user questions. If the utterances do not properly reflect a character's personality, the user may feel discomfort or get bored quickly without feeling that they are talking to the actual character. Such utterances might also deviate from the personality that persons want and expect from an anime character. Generally, a lot of data must be collected that reflects the personality of a specific character. To generate perfectly correct utterances, appropriate utterances must be created for every utterance that the user enters. However, collecting utterance data while keeping a personality consistent is costly. In addition, appropriate answers to a variety of user questions must be generated from a limited amount of data. Therefore, it is necessary 1) to efficiently collect high-quality data that accurately reflects the individuality of a character, and 2) to use the collected data to generate appropriate answers to various user questions.

For problem 1), we propose the use of a data collection method called "role play-based question-answering" [7], in which users play the role of characters and answer questions that the users ask the characters, to efficiently collect responses to many questions that accurately reflect the personality of a particular character. For problem 2), we propose a new utterance generation method that uses a neural translation model with the collected data.

Rich and natural expressions of body motion greatly enhance the appeal of agent systems. Therefore, generating agent body movements that are more human and enriched is an essential element in building an attractive agent system. However, not all existing anime characters move as naturally and as diversely as humans. Characters can also have unique movements, for example, a specific pose that appears with a character-specific utterance. Therefore, it is also important to reproduce unique movements to realize a system that reflects a character's personality. It is known that there is a strong relationship between the content of an utterance and body motion. The motion should be suited to the content. If a skilled creator were to create a motion that satisfies the two aspects of utterance content and body motion for every utterance, the dialogue agent would probably be able to express motion that completely matches that of a character. However, in dialogue systems that generate a variety of utterances, this is not practical.

Therefore, in this context, we propose the use of a motion generation method that is comprised of two methods for introducing movements in dialogue agents of existing characters that are more human-like and natural and for introducing character-specific motions. First, we propose a method that can automatically generate whole-body motion from utterance text in order to make anime characters have human-like and natural movements. Second, in addition to these movements, we try to add a small amount of characteristic movement on a rule basis to reflect the anime character's personality.

As a target for applying these proposed utterance and motion generation methods, we construct a dialogue system with "Ayase Aragaki," a character from the light novel "Ore no Imoto ga Konna ni Kawaii Wake ga Nai" in Japanese, which means "My Little Sister Can't Be this Cute" in English. The novel is a popular light novel that has sold over five million copies in Japan and has been animated. Ayase is not the main character, but she has an interesting personality called "yandere." According to Wikipedia, this means that she is mentally unstable, and once her mental state is disturbed, she she exposes her emotions and behaves extremely violently. For this reason, she is a very popular character.

We constructed an agent text-chat dialogue system that reflects her personality by using the proposed system-construction method. We evaluated the usefulness of the implemented system from the viewpoint of whether her responses were natural and her personality could be reflected properly. As a result, both of our proposed methods for generating utterances and body motions were found to be useful for improving the users' impression of the naturalness and character-likeness of this existing anime character in the responses of the system. This suggests that our proposed system-construction method will likely contribute greatly to realizing text-dialogue-agent systems of characters.

Sections 2 and 3 describe the methods for utterance and motion generation. Section 4 describes the construction of a dialogue system, and Section 5 describes a user evaluation done using the implemented system. Finally, we discuss the results in Section 6, and we conclude this paper in Section 7.

2. Utterance Generation Method

2.1 Research Problem and Our Approach

Generally, when generating an utterance that reflects personality in a way that guarantees quality, there is a problem in that the cost is high because large-scale utterance-pair data, comprised of pairs consisting of a user's question and a system's answer, must be prepared and used manually in advance. In this research, we propose using a data collection method called "role play-based question-answering" [7] to efficiently construct high-quality utterance pairs reflecting the individuality of a character. This is a method in which multiple users participate online and ask questions to a specific character or answer a question as is, so that high-quality character-like utterance pairs can be efficiently collected. Specifically, the user has two roles. One is to ask a character a question (utterance). The user asks a character a question that they want to ask for a variety of topics. This question is notified to all users. The second role is to become a character and answer the questions. This makes it possible to efficiently collect utterance pairs of questions and answers by sharing and using the questions of the user and answering each question. In addition, the role-playing experience itself is interesting so there is no need to modify it in order to make it easier for users to participate [8]. By using this method, we thought that it would be possible to efficiently collect utterance-pair data that seems to reflect an anime character.

Our proposed utterance generation method uses a neural translation model [9], which is one of the latest machine learning methods for text generation that has been attracting attention in recent years and that extracts appropriate answers from collected data.

Fig. 1 Screenshot from site for collecting data using "role play-based question-answering."

Table 1 Results of user evaluation of experience with role play-based question-answering service.

Evaluation item	Mean score
Did you feel comfortable using the website?	4.08
Did you enjoy the role play-based question-answering?	4.53
Do you want to experience this web service again?	4.56

2.2 Data Collection Using Role-playing Question Answering

NICONICO Douga [*1] is a video steaming service that offers a channel service for fans of various characters. Use of a channel is limited to registered subscribers. In our research, a bulletin board that responds to questions in tandem with this channel service was built for an "Ayase Aragaki" channel. **Figure 1** shows a screenshot of the site. Users can freely ask Ayase Aragaki questions from a prepared text form. A user who wants to respond to Ayase Aragaki can freely answer a question. Labeling of emotions accompanying an utterance is performed simultaneously with the answering. There were eight classifications for the labeled emotions: normal, angry, fear, fun, sad, shy, surprise, and yandere.

To increase the users' motivation to participate, the website showed the ranking of users by the number of posts. In addition, a "Like" was placed next to each answer. If a user's answer seemed to be Ayase Aragaki (-like), "Like" was pressed. We devised the ranking so that the evaluation of users could be reflected in the quality of answers. In October 2017, a website was opened, and the service was operated for about 90 days. A total of 333 users participated. The collected utterance pairs exceeded 10,000 in about 20 days, and finally, 15,112 utterance pairs were obtained by the end of the service. Users voluntarily participated in this response site and were not paid. Nevertheless, the fact that we were able to collect such a large amount of data suggests that data collection using the role play-based question-answering method is useful.

A questionnaire evaluation was conducted for participating users in order to determine their satisfaction with using the question-answering site. A total of 36 users cooperated in the evaluation and responded to the items shown in **Table 1** on the basis of a five-point Likert scale (1 to 5 points). Table 1 shows the average of the user evaluation values.

Looking at the results, a high rating of 4.08 was obtained for the item "Did you feel comfortable using the website?" These ratings suggest that the service is enjoyable for users. High ratings of 4.53 and 4.56 were also obtained for the items "Did you enjoy the role play-based question-answering?" and "Do you want to experience this web service again?" These suggest that the service was attractive to the users.

Next, to evaluate the quality of the collected utterance data of Ayase, a subjective evaluation was performed by the participating users. For about 50 utterance pairs, which were selected randomly from all collected data of 15,112 utterance pairs, participating users evaluated whether they were natural and properly reflected her personality. The mean scores for naturalness and personality were 3.61 and 3.74 on a five-point Likert scale (1 to 5 points). This indicates that the quality of the response-utterance data collected through role play-based question-answering was reasonably high. However, it was surprising difficult to obtain a rating of 4.0 or more, even if the response data was created by human users. In other words, this result suggests that generating utterances that reflect a particular characterr's individuality is a difficult task even for humans.

2.3 Proposed Utterance Generation Method

We thus propose an utterance generation method that uses the collected utterance-pair data. Since the amount of data collected was not large enough to train an utterance generation model using neural networks [9], we used the approach of extracting optimal responses from the obtained utterance data. In other words, we addressed the problem of selecting a response for the most relevant utterance pair against a user's utterance. In this study, a neural translation model was used to select an appropriate utterance pair.

One of the simplest method is to use a simple application of an off-the-shelf text search engine such as LUCENE [*2] which is a popular open-source search engine for retrieval. Questions and answers are first indexed with LUCENE. We use a built-in Japanese analyzer for morphological analysis. Given an input question, the BM25 algorithm [10] is used to search for a similar question using the content words of the input question. The answers for the retrieved questions are used as the output of this method. Although simple, this method is quite competitive with other methods when there are many question-answer pairs because it is likely that we will be able to find a similar question by word matching. But, only using word-matching may not be sufficient. Therefore, we developed a more elaborate method that re-ranks the results retrieved from LUCENE. Our method focuses on recent advances in cross-lingual question answering (CLQA) [11] and neural dialogue models [9]. In addition, we matched the semantic and intention levels of the questions so that the appropriate answer candidates were ranked higher.

(1) Given question Q as input, Lucene searches the top N question-response pairs $(Q'_1, A'_1), \ldots, (Q'_N, A'_N)$ from our dataset.

(2) For Q and Q', question type determination and named entity extraction are performed, and the question type and named

[*1] http://www.nicovideo.jp/

[*2] https://lucene.apache.org/

entity (using Sekine's extended named entity [12] as the system) are extracted. To what extent a named entity asked by Q is included in A' is calculated, and this is used as the question type match score (qtype_match_score).

(3) Using a focus extraction module, the focus (noun phrase indicating a topic) is extracted from Q and Q'. If the focus of Q is included in Q', the focus score (center-word_score) is set to 1; otherwise, it is set to 0.

(4) A translation model calculates the probability that A' is generated from Q, that is, $p(A'|Q)$. We also calculate $p(Q|A')$ as the reverse translation probability. Such reverse translation has been validated in CLQA [11]. The generation probability is normalized by the number of words on the output side. Since it is difficult to integrate a probability value with other scores due to differences in range, we rank the answer candidates on the basis of this probability and calculate the translation score (translation_score; _translation_score). Specifically, when the rank of a certain answer candidate is r, the translation score is obtained as follows.

$$1.0 - (r - 1)/\text{max\_rank} \quad (1)$$

Here, max_rank is the maximum number of possible answer candidates. The translation model was learned by pretraining about 500,000 general question-response pairs and then performing fine-tuning with the utterance pairs obtained from the complete question-response. In the reverse model, the same processing was performed by exchanging the questions and responses. The translation models were created in the manner described in Ref. [13]. In the training, the OpenNMT toolkit [14] was used with the default settings.

(5) The similarity between Q and Q' is measured with a semantic similarity model. Word2vec [15] is used for this measurement. First, for each Q and Q', a word vector is obtained, an average vector is created, and cosine similarity is calculated, and this is set as a similarity score (*semantic_similarity_score*).

(6) The previous scores are added by weight and a final score is obtained.

$$\begin{aligned}&\text{score}(Q,(Q',A'))\\&\quad= w_1 * \text{search\_score}\\&\qquad+ w_2 * \text{qtypes\_match\_score}\\&\qquad+ w_3 * \text{center-word\_score}\\&\qquad+ w_4 * \text{translation\_score}\\&\qquad+ w_5 * \text{rev\_translation\_score}\\&\qquad+ w_6 * \text{semantic\_similarity\_score}\end{aligned} \quad (2)$$

Here, *search_score* is a score obtained from the ranking of search results by Lucene, and it is obtained from expression 1. Also $w_1, \ldots, w_6$ are weights, which are 1.0 in this study.

(7) On the basis of the above score, the answer candidates are ranked, and the top items are output.

The most appropriate answer sentence to a question is selectively generated with this combination of various types of language processing.

3. Motion Generation Method

3.1 Research Problem and Our Approach

Imparting appropriate body movement to agents and humanoid robots has not only been shown to improve the natural appearance but also promotes conversation. For example, actions accompanying utterances have the effect of enhancing the persuasiveness of utterances, making it easier for the other party to understand the content of the utterances [16]. Therefore, generating agent body movements that are more human and enriched is an essential element in building an attractive agent system. As mentioned above, however, not all existing anime characters move as naturally and as diversely as humans. Additionally, characters can have unique movements, for example, a special pose that appears with a character-specific utterance. Therefore, it is also important to reproduce unique movements so that the system reflects a character's personality. Since there is a strong relationship between utterance content and body motion, motion should be suited to the content. If a skilled creator were to create a motion that satisfies these two aspects for every utterance, the dialogue agent would probably be able to express motion that perfectly expresses a character. However, for dialogue systems that generate a variety of utterances this is not practical.

Therefore, for practical motion generation, we propose using a motion generation method comprised of two methods that introduce characteristic movements that are more human-like and natural and that introduce character-specific motions. First, we propose a method that can automatically generate whole-body motions from utterance text so that anime characters can make human-like and natural movements. Second, in addition, we try to add a small amount of characteristic movement on a rule basis to reflect personality.

The proposed motion generation method makes the motion of animated characters more natural and human. In a text dialogue system, linguistic information obtained from system utterances may be used as input to generate motions. In past research on motion generation using linguistic information, we mainly worked on the generation of a small number of motions using word information such as the presence or absence of nodding and limited hand gestures [17], [18], [19]. In this research, we tried to generate more comprehensive whole-body movements by using various types of linguistic information. As a specific approach, we constructed a corpus containing data on speech linguistic information and motion information obtained during human dialogue, taught the co-occurrence relationship of these using machine learning, and generated motion using speech linguistic information as input. In the next section, the construction of the corpus data, the motion generation method, and its performance are described. Then, we introduce a way to add specific body motions that reflect an anime character's personality.

3.2 Collecting Data for Motion Generation

A linguistic and non-linguistic multi-modal corpus including spoken language and accompanying body movement data was constructed for two-party dialogues. The participants in the two-way dialogue were Japanese men and women in their twenties

Table 2 List of generated labels for each motion part.

Motion part	Number of labels	List of labels
Number of nods	6	0, 1, 2, 3, 4, more than 5
Deepness of nodding	4	micro, small, medium, large
Head direction (yaw)	9	front, right-micro, right-small, right-medium, right-large, left-micro, left-small, left-medium, left-large
Head direction (roll)	9	front, right-micro, right-small, right-medium, right-large, left-micro, left-small, left-medium, left-large
Head direction (pitch)	7	front, up-micro, up-small, up-medium, up-large, up-micro, up-small, up-medium, up-large
Hand gesture	9	none, iconic, metaphoric, beat, deictic, feedback, compellation, hesitation, others
Upper body posture	7	center, forward-small, forward-medium, forward-large, backward-small, backward-medium, backward-large

through fifties who had never previously met. A total of 24 participants (12 pairs) sat facing each other. Participants sat facing each other. To collect a lot of data on various actions, such as chats, discussions, and nodding and hand gestures associated with utterances, we used dialogues in which animated content was explained. In these dialogues, each participant watched an episode (Tom & Jerry) with different content and explained the content to their conversation partner. The conversation partner was free to ask the presenter questions and to have a free conversation. For recording utterances, a directional pin microphone attached to each subject's chest was used. A video camera was used to record the overall dialogue situation and the participants. The video was recorded at 30 Hz.

The total time of the chats, discussions, and explanations was set to 10 minutes each, and in this study, the data from the first 5 minutes was used. For each pair, one chat dialogue, one discussion dialogue, and two explanation sessions were conducted. We therefore collected 20 minutes of conversation data for each pair, and we collected a total of 240 minutes of conversation data for 12 pairs.

Next, we show the acquired linguistic and non-linguistic data.

- Utterance: After manually transcribing the utterances from the voice information, the content of the utterances was confirmed and the sentences were divided. Furthermore, each sentence was divided into phrases by using a dependency analysis engine [20]. The number of divided segments was 11,877.
- Head direction: Using the face image processing tool OpenFace [21], three-dimensional face orientation information was taken from the front of the participants with a video camera. The yaw, roll, and pitch angles were obtained. Each angle was classified as micro when the angle was 10 degrees or less, small when it was 20 degrees or less, medium when it was 30 degrees or less, and large when it was 45 degrees or more.
- Nodding: Sections in the video where nodding occurred were manually labeled. Continuous nods were treated as an one nod event. In addition, the number of nodding times was classified into five stages from 1 to 5 (or more). In addition, for the depth of a nod, OpenFace was used to calculate the difference between the start of the head posture pitch at the time of the nod and the angle of the deepest rotation. The angle was classified as micro when the angle was 10 degrees or less, small when it was 20 degrees or less, medium when it was 30 degrees or less, and large when it was 45 degrees or more.
- Hand gestures: Sections in the video where hand gestures occurred were manually labeled. A series of hand-gesture motions was classified into the following four states.
  - Prep: Raise hand to make a gesture from the home position
  - Hold: Hold hand in the air (waiting time until gesture starts)
  - Stroke: Perform gesture
  - Return: Return hand to home position

  However, in this study, for simplicity, a series of actions from Prep to Return were treated as one gesture event. Furthermore, the types of hand gestures were classified into the following eight types based on the classification of hand gestures by McNeil [22].
  - Iconic: Gestures used to describe scene descriptions and actions.
  - Metaphoric: Like Iconic, this is a pictorial and graphic gesture, but the specified content is an abstract matter or concept. For example, the time flow.
  - Beat: Adjusts the tone of speech and emphasizes speech. For example, shaking or waving the hand in accordance with an utterance.
  - Deictic: A gesture that points directly toward a direction, place, or thing such as pointing.
  - Feedback: Gestures issued in synchronization with, consent to, or in response to another person's utterance. A gesture that accompanies an utterance in response to an utterance or gesture made in front of another person. In addition, gestures of the same shape performed by imitating the gestures of the other party.
  - Compellation: Gesture to call another person.
  - Hesitate: Gesture that appears at the time of hesitation.
  - Others: Gestures that are unclear but seem to have some meaning.
- Upper body posture: We observed postures when the participants were seated and there was no significant change in the seated position. For this reason, the front and back positions of the upper body were extracted on the basis of the three-dimensional position of the head. Specifically, the difference between the coordinate position in the front-back direction of the head position obtained using OpenFace and the position of the center position was obtained. From the position information, the angle of the posture change of the upper body was calculated as micro when it was 10 degrees or less, small when it was 20 degrees or less, medium when it was 30 degrees or less, and large when it was 45 degrees or more.

Table 2 shows a list of the parameters of the obtained corpus

Table 3 Performance of motion generation method and chance level. Each score shows the F-measure.

Motion part	Chance level	Proposed method
Number of nods	0.226	0.428
Deepness of nods	0.304	0.475
Face direction (yaw)	0.232	0.329
Face direction (roll)	0.297	0.397
Face direction (pitch)	0.261	0.378
Hand gesture	0.156	0.303
Upper body posture	0.183	0.311

data. In addition, ELAN [23] was used for manual annotation, and all of the above data were integrated with a time resolution of 30 Hz.

3.3 Proposed Motion Generation Method

Using the constructed corpus data, we input a word, its part of speech, a thesaurus, a word position, and the utterance action of one entire utterance as input, and we created a model that generates one action class for each clause for each of the seven actions shown in Table 2 by using the decision tree algorithm C4.5. Namely seven option labels were generated for each clause. Specifically, the language features used were as follows.

- Number of characters: Number of characters in a clause.
- Position: Position of a phrase from the beginning and end of a sentence.
- Words: Word information (bag-of-words) in phrases extracted by the morphological analysis tool Jtag [24].
- Part-of-speech: Part-of-speech information of words in a clause extracted by Jtag [24].
- Thesaurus: Thesaurus information for words in a phrase based on Japanese vocabulary.
- Utterance act: Utterance-act estimation method using word n-gram and thesaurus information [25], [26]. Utterance act extracted for each sentence (33 types).

The evaluation was performed by cross-validation done 24 times, in which the data of 23 of the 24 participants were used for learning, and the remaining data of 1 participant was used for evaluation. We evaluated how much actual human motion could be generated only from the data of others. **Table 3** shows the average of the F-value as a performance evaluation result. The chance level indicates the performance when all classes with the highest number of correct answers are output. Table 3 shows that the accuracy was significantly higher than the chance level for all generation targets (results of paired t-test: $p < .05$). The results show that the proposed method, which uses words, their parts of speech and thesaurus information, word positions, and actions performed during the entirety of speaking, obtained from spoken language is effective in generating whole-body motions as shown in Table 3.

3.4 Additional Original Motion Reflecting Character's Personality

In addition to the motion generation proposed in the previous section, motions unique to Ayase were extracted from the motions in the animation, and the four original motions shown in **Table 4** were added. These actions were selected in collaboration with Ayase's creators who have experience in creating animation.

Table 4 Additional original movements and example utterance text for triggering.

Additional original movements	*Example utterance text for triggering*
Performs roundhouse kick	*I'm gonna take you down*
Crosses her arms	*Ewww*
Raises her arms and sticks face out	*Pervert*
Points to front with her right hand	*I'll report you*

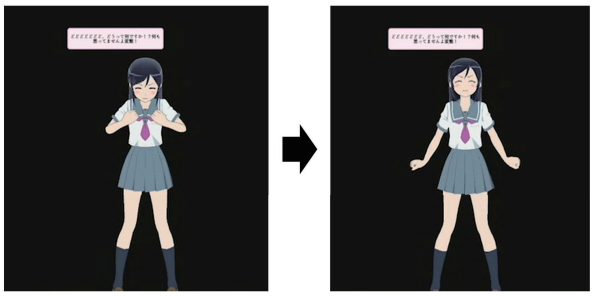

Fig. 2 Example of scene in which original motion of raising and lowering arm and protruding face is performed in accordance with text display of "Pervert!" included in system utterance.

For these original actions, words and sentences that trigger the actions were set, and when these words and sentences appear in a system utterance, the original actions take precedence over the output results of the motion generation model. **Figure 2** shows an example scene where she raises her arms and sticks her face out in accordance with the speech text "pervert." Although the number of movements was as small as four, we could not find any more distinctive movements to note, so we thought that this small number of movements was sufficient.

4. Construction of Dialogue System Reflecting Anime Character's Personality

Using the proposed methods for utterance and motion generation, we constructed a dialogue system that can respond to user utterances with utterances and motion. **Figure 3** shows a diagram of the system configuration.

The user enters input text from the chat UI. When the dialogue manager receives it, the user text is first sent to the utterance generator. After acquiring the system utterance from the utterance generation unit, the system utterance is transmitted to the motion generation unit, and the motion information in each clause of the system utterance text is obtained.

In addition to this, while this is not necessary for text dialogues, it is also possible to obtain uttered speech with the speech synthesis unit. In this system, the speech obtained from the speech synthesis unit is used to generate lip-sync motion.

The dialogue manager sends the system utterance text, motion schedules, and voices to the agent animation generator. In the agent animation part, the utterance text is displayed in a speech bubble above the character at equal time intervals from the first character, and the motion of the agent is generated in sync with the display of utterance characters according to a motion schedule. As a means for generating motion animation, a CG character was created in Unity, and animations corresponding to the motion lists in Tables 2 and 4 were generated in real time. At this time, the seven objects shown in Table 2 can operate independently,

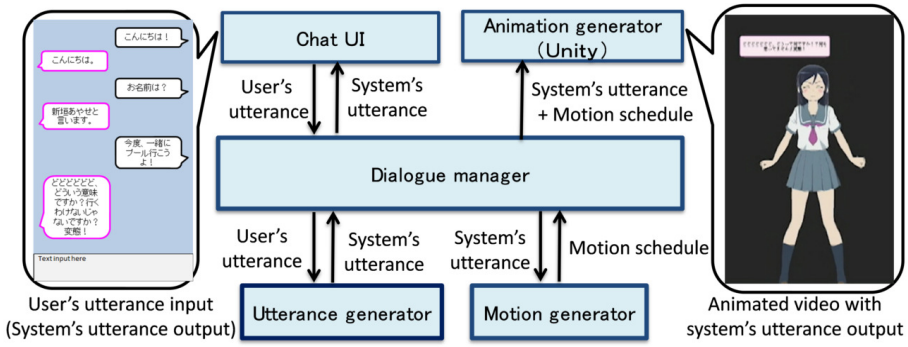

Fig. 3 Architecture of our system.

and the head motion is generated by mixing all parameters of the number of nods, deepness of head movement, and head directions (yaw, roll, pitch). When utterance text is registered as a trigger for generating the specific motions in Table 4, a specific motion is generated instead of a motion generated by our motion generation model. All motions of the agent are generated according to the timing of the utterance text display. An example of a presentation screen is shown on the right side of Fig. 3.

It is also possible to send a system utterance from the dialogue management unit to the chat UI and to present the system utterance in the chat UI in addition to the user utterance shown on the left side of Fig. 3.

5. Subjective Evaluation

5.1 Evaluation Method

The effectiveness of the proposed methods was evaluated in subject experiments using the constructed dialogue system. As an evaluation item, we evaluated the usefulness of the responses of the dialogue system with the proposed utterance and motion generation methods. The purpose of this evaluation was to focus on character reproducibility (character-like) in addition to the naturalness of the responses.

The following three conditions were set as experimental conditions for utterance generation.

- **U-AIML**: A rule-based method written in AIML, which is a general method used for utterance generation was used. Specifically, we used a large-scale AIML database that has been constructed up to 300K utterance pairs. The description of the 300k corpus can be found in Ref. [27]. In Japanese, sentence-end expressions are some of the most important elements that indicate a character [28], so these expressions were converted to expressions like those used by Ayase by using a sentence-end conversion method [29]. For example, "kawaii-desu" becomes "kawaii-desune," where "desu" is replaced with "desune," which is characteristic of Ayase. The AIML with sentence-end expressions modified for the character is appropriate for the baseline.
- **U-PROP**: An utterance is generated by using our proposed utterance generation method described in Section 2. The weight parameters w_1 to w_6 were experimentally set to 1.0 in Formula (1).
- **U-GOLD**: An utterance is generated by using the data collected with the role play-based question-answering method in Section 2.2. When multiple answers were given to a question, one was selected at random.

By comparing the U-AIML and U-PROP conditions, the usefulness of the proposed utterance generation was compared with manual utterance generation and evaluated. We also compared the U-PROP and U-GOLD conditions to evaluate how useful the proposed utterance generation method is compared with human-generated utterances.

The following four conditions were set as the experimental conditions for motion generation.

- **M-BASE**: Generates basic character movements such as for lip sync and facial expressions. For generating facial expressions, we created animations for facial expressions corresponding to the eight emotions collected with the role play-based question-answering method in Section 2.2. Facial expressions under the U-PROP and U-GOLD conditions were generated by using the collected data. For the U-AIML condition, humans annotated the emotion label for each utterance manually. The labels were used to generate facial expressions.
- **M-RAND**: In addition to lip sync and facial expressions (same as M-BASE), whole body movements were randomly generated.
- **M-PROP1**: In addition to lip sync and facial expressions (same as M-BASE), our proposed motion generation method was used on the basis of the human data described in Section 3.3.
- **M-PROP2**: In addition to lip sync and facial expressions (same as M-BASE) and using our motion generation method with human data (same as M-PROP1), a small amount of motion unique to the character was added as described in Section 3.4. This additional unique motion occurred when the utterance text triggered it as shown in Table 4. Such utterances were not necessarily included for all utterances. In this experiment, the U-PROP and U-GOLD conditions included utterances that triggered for just half of all of the utterances.

By comparing the M-BASE and M-RAND conditions, we evaluated the usefulness of motion generation for the whole body, and we compared the M-RAND and M-PROP1 conditions to evaluate the usefulness of the proposed motion generation method by learning human motion data. Also, by comparing the M-PROP1 and M-PROP2 conditions, we evaluated the usefulness of adding a small amount of unique character-specific motions in

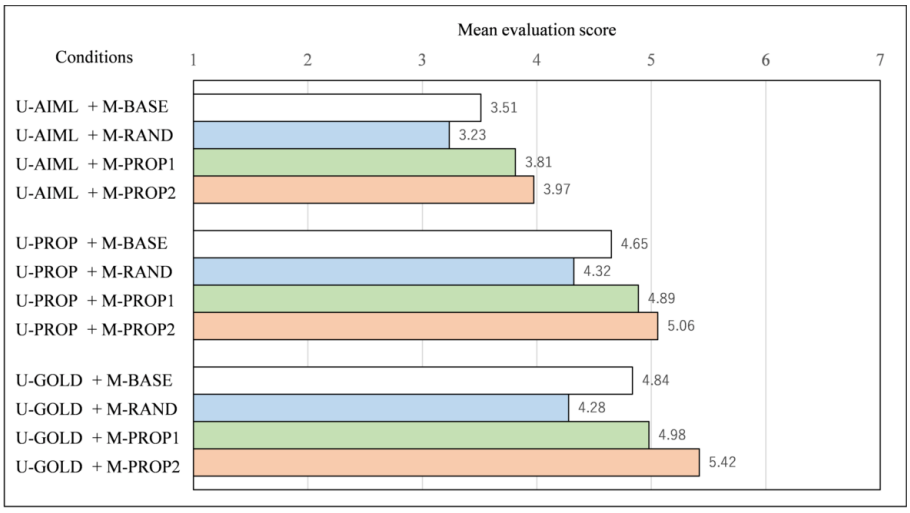

Fig. 4 Results of subjective evaluation for impression of "naturalness" of overall response.

Table 5 Items and questionnaire of subjective evaluation.

Items	*Questionnaire*
Naturalness	*Is the overall response natural?*
Character-likeness	*Does the response comprehensively reflect the personality of "Ayase Aragaki"?*

addition to the proposed generation by learning human motion data.

Twelve conditions combining these three conditions of utterance generation and four conditions of motion generation were set as experimental conditions.

As an experimental method, the same user utterance was set for comparison under each condition, and the utterance and motion of the system in response to the user utterance were evaluated. Specifically, ten question utterances were randomly extracted from the collected question-answer data. The subjects observed the user's utterance text for 3 seconds and then watched a video showing the response of the system. At each viewing, the subjects evaluated the impression of the response of the dialogue system using a seven-point Likert scale (1 to 7 points). Specific evaluation items are shown in **Table 5**. We used "naturalness" and "character-likeness" to evaluate two aspects of utterance quality in conversational agents. One aspect is whether the response is appropriate in general, or namely whether an utterance makes sense irrespective of who's speaking. The other aspect is whether the utterance is appropriate to the character in question. Ideally, we want the character-likeness to be high, but we want to maintain at least reasonable naturalness when considering the deployment of the automated agents. Note that an utterance can be rated low in terms of naturalness but high in character-likeness, or vice-versa; for example, character-likeness can be achieved by repeating utterances specific to that character (for example, clichés), and some general utterances, such as greetings, can never be uttered by particular characters.

Since all subjects knew the character of Ayase very well we think that the simple questionnaire was appropriate for the evaluation. Since videos of 10 utterances were prepared for each of the 12 conditions, video viewing and evaluation were performed 120 times. Considering the order effect, the order of videos presented to each subject was randomized. The subjects evaluated the same 10 questions. The test input questions had been removed from the training set. For U-PROP, the dataset was first split into train/dev/test. Here, the data were split so that there were no overlapping questions in the train/dev/test sets. U-PROP retrieved the responses from the train/dev sets.

5.2 Evaluation Results

Our proposed methods should be evaluated by someone who knows Ayase Aragaki well. Through an intermediary, we recruited people who had watched the anime version of "Ore no Imoto ga Konnani Kawaii Wake ga Nai," read the novel, and had a deep knowledge of Ayase Aragaki. When applying to participate in our experiment, the applicants had to fill in, through free description, what things are appealing about her and what their feelings are about her. On the basis of their description, we recruited only those people who were familiar with her as subjects. An experiment was performed with 30 subjects. The mean value of each subjective evaluation item for each subject under each experimental condition was calculated.

5.2.1 Evaluation Results for Naturalness of Response

The mean values for "naturalness" are shown in **Fig. 4**. We performed a two-dimensional analysis of variance to evaluate the effect of the factors of the utterance and motion conditions on the rating score of the overall naturalness. As a result, a simple main effect for both motion conditions was observed and no interaction effect was observed [utterance condition: $F(2, 348) = 61, 52, p < .01$, motion condition: $F(3, 348) = 10.47, p < .01$].

Since a simple main effect for the factors of the utterance conditions was observed, multiple comparisons with the Holm's method for family-wise error rate adjustment were performed to verify which utterance conditions differed under each motion condition. The results are shown in **Table 6**.

The values of the overall naturalness in the U-PROP and U-GOLD conditions were significantly higher than in the U-AIML condition under all motion conditions, M-BASE, M-RAND, M-PROP1, and M-PROP2 ($p < .01$). Therefore, this result suggests that the proposed utterance generation method and utter-

Table 6 Results of multiple comparisons for impression of "naturalness" between utterance conditions under each motion condition (left) and between motion conditions under each utterance condition (right). p-values were corrected by Holm's method for family-wise error rate adjustment. ** indicates a significance level of 1% or less ($p < .01$), * indicates a significance level of 5% or less ($p < .05$), and † indicates a significance level of 10% or less ($p < .10$).

	U-AIML vs. U-PROP	U-AIML vs. U-GOLD	U-PROP vs. U-GOLD
M-BASE	0.003e-9 **	0.001e-9 **	0.009 **
M-RAND	0.005e-9 **	0.003e-9 **	0.580
M-PROP1	0.008e-5 **	0.004e-7 **	0.309
M-PROP2	0.001e-5 **	0.007e-4 **	0.441

	M-BASE vs. M-RAND	M-BASE vs. M-PROP1	M-BASE vs. M-PROP2	M-RAND vs. M-PROP1	M-RAND vs. M-PROP2	M-PROP1 vs. M-PROP2
U-AIML	0.013 *	0.004 **	0.004 **	0.001e-2 **	0.009e-2 **	0.110
U-PROP	0.014 *	0.082 †	0.056 †	0.002 **	0.001e-1 **	0.113
U-GOLD	0.003e-2 **	0.317	0.150	0.002e-2 **	0.004 **	0.343

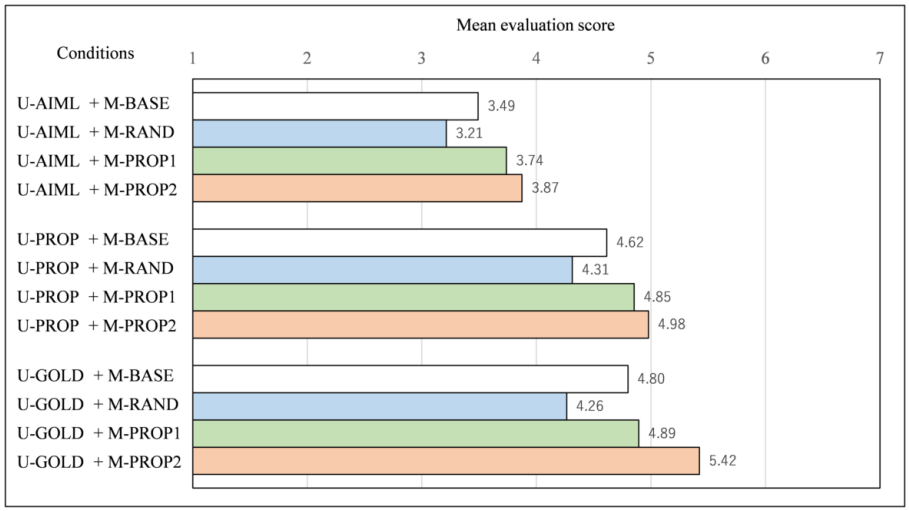

Fig. 5 Results of subjective evaluation for impression of "character-likeness" of overall response.

ances made by humans had higher rating scores in terms of overall naturalness than the general utterance generation using AIML regardless of the difference in motion condition.

Under only the M-BASE condition, the rating score of the overall naturalness in the U-GOLD condition was significantly higher than in the U-PROP condition ($p < .01$). Therefore, this result suggests that suggests that an utterance made by a human was higher in terms of score than that of our proposed utterance method only when motion generation was static (M-BASE). For the other conditions there was no difference in the rating score of the overall naturalness between the proposed utterance generation (U-PROP) and utterances made by a human (U-GOLD) when various body motions were generated (M-RAND, M-PROP1, and PROP2).

Next, since a simple main effect from motion condition factors is observable, we similarly determined which motion condition exhibited differences under each condition by making multiple comparisons. The results are shown in Table 6.

Comparing M-RAND and others, the rating score of the overall naturalness for M-RAND was lower than all other conditions under all utterance conditions ($p < .05$ for M-BASE vs M-RAND under U-AIML and U-PROP; $p < .01$ for M-BASE vs M-RAND under U-GOLD; $p < .01$ for M-RAND vs M-PROP1/M-PROP2 under all utterance conditions). This result suggests that regardless of the content of the utterance text (U-RAND), the randomized motion generation method was not appropriate than the static motion (U-BASE) and our proposed motion generation according to the utterance content (U-PROP1 and U-PROP2).

When M-BASE and M-PROP1/M-PROP2 conditions were compared, the rating scores of the overall naturalness of M-PROP1 and M-PROP2 were significantly higher than for M-BASE under U-AIML ($p < .05$). The scores for naturalness of response for M-PROP1 and M-PROP2 tended to be significantly higher than M-BASE under U-PROP ($p < .10$). However, there was no difference in the scores for naturalness of response between M-BASE and M-PROP1/M-PROP2 under U-GOLD.

These results suggest that our proposed motion generation methods (M-PROP1 and M-PROP2) are useful for improving the impression of naturalness of response better than static motion (M-BASE) when an utterance is artificially generated (U-AIML and U-PROP). However, when utterance generation is performed by a human (U-GOLD), the impression is not improved by the proposed motion generation.

Comparing M-PROP1 and M-PROP2 showed no differences between M-PROP1 and M-PROP2 under any of the utterance conditions. The results suggest that the impression of naturalness did not differ between the our generation method without the added original motion of a character (M-PROP1) and the method with the added original motion (M-PROP2).

5.2.2 Evaluation Results for Character-likeness

The mean values for "character-likeness" are shown in **Fig. 5**. The same analysis was performed for the evaluation score of the character-likeness of the overall responses, just as done for the overall naturalness score. As a result, a simple main effect of the utterance and motion conditions was observed, and no interaction effect was observed (utterance condition: $F(2, 348) = 3.92$,

Table 7 Results from multiple comparisons for impression of "character-likeness" between utterance conditions under each motion condition (left) and between motion conditions under each utterance condition (right). p-values were corrected by Holm's method for family-wise error rate adjustment. ** indicates a significance level of 1% or less ($p < .01$), * indicates a significance level of 5% or less ($p < .05$), and † indicates a significance level of 10% or less ($p < .10$).

	U-AIML vs. U-PROP	U-AIML vs. U-GOLD	U-PROP vs. U-GOLD
M-BASE	0.007e-9 **	0.002e-9 **	0.001e-1 **
M-RAND	0.002e-8 **	0.001e-10 **	0.545
M-PROP1	0.003e-5 **	0.003e-7 **	0.666
M-PROP2	0.002e-5 **	0.004e-4 **	0.204

	M-BASE vs. M-RAND	M-BASE vs. M-PROP1	M-BASE vs. M-PROP2	M-RAND vs. M-PROP1	M-RAND vs. M-PROP2	M-PROP1 vs. M-PROP2
U-AIML	0.013 *	0.004 **	0.004 **	0.001e-4 **	0.009e-2 **	0.111
U-PROP	0.018 *	0.082 †	0.076 †	0.001 **	0.001e-1 **	0.103
U-GOLD	0.003e-1 **	0.317	0.225	0.002e-1 **	0.004 **	0.229

$p < .01$, motion condition: $F(3, 72) = 3.82$, $p < .01$).

Since a simple main effect for the factors of the utterance conditions was observed, multiple comparisons with the Holm's method for family-wise error rate adjustment were performed to verify which utterance conditions differed under each motion condition. The results are shown in **Table 7**.

The values for character-likeness in the U-PROP and U-GOLD conditions were significantly higher than in the U-AIML condition under all motion conditions, M-BASE, M-RAND, M-PROP1, and M-PROP2 ($p < .01$). Therefore, the result suggests that the proposed utterance generation method and utterances made by a human had higher rating scores in terms of character-likeness than the general utterance generation using AIML regardless of the difference in motion condition.

Under only the M-BASE condition, the rating score for character-likeness in the U-GOLD condition was significantly higher than that in the U-PROP condition ($p < .01$). Therefore, this suggests that an utterance made by a human was higher in terms of score than that of our proposed utterance method only when motion generation was static (M-BASE). In other words, there was no difference in the rating scores of the character-likeness between the proposed utterance generation (U-PROP) and utterances made by a human (U-GOLD) when various body motions were generated (M-RAND, M-PROP1, and PROP2).

Since a simple main effect of the factors of the motion conditions were observed, multiple comparisons were similarly made to determine which motion condition exhibited a difference under each utterance condition. The results are shown in Table 7.

Comparing M-RAND and others, the rating score of the character-likeness for M-RAND was lower than all other conditions under all utterance conditions ($p < .05$ for M-BASE vs. M-RAND under U-AIML and U-PROP; $p < .01$ for M-BASE vs. M-RAND under U-GOLD; $p < .01$ for M-RAND vs. M-PROP1/M-PROP2 under all utterance conditions). This result suggests that regardless of the utterance text content, the randomized motion generation method (U-RAND) was not more appropriate than static motion (U-BASE) and our proposed motion generation according to the utterance content (U-PROP1 and U-PROP2).

When M-BASE and M-PROP1/M-PROP2 conditions were compared, the rating score of the character-likeness for M-PROP1 and M-PROP2 was significantly higher than that for M-BASE under U-AIML ($p < .01$). The scores for character-likeness for M-PROP1 and M-PROP2 tended to be significantly higher than M-BASE under U-PROP ($p < .10$). However, there was no difference in the scores for character-likeness between M-BASE and M-PROP1/M-PROP2 under U-GOLD.

These results suggest that our proposed motion generation method (M-PROP1 and M-PROP2) is useful for improving the impression of character-likeness to a greater extent than static motion (M-BASE) when an utterance is artificially generated (U-AIML and U-PROP). However, when utterance generation is performed by a human (U-GOLD), the impression is not improved by the proposed motion generation.

Comparing M-PROP1 and M-PROP2 showed no difference between M-PROP1 and M-PROP2 under all utterance conditions. The results suggest that the evaluation scores did not differ between the original motion being added to the basic motion by our generation model (M-PROP2) and the original motion not being added (M-PROP1).

5.3 Detailed Observation of Generated Utterance and Motion

We show examples of our generated utterances and motions that received high ratings and those that did not in the subject experiment. First, two examples of generated utterances under each utterance condition in our experiment are shown in **Table 8**. The subjective evaluation scores are mean values under the M-NONE motion condition since there was no difference in the animated whole-body motion for any of the system utterances under M-NONE. In example 1, the system utterance was "*Thank you for praising me!*", which had the highest scores (5.37 and 5.33) when the user utterance was "*I thought that Ayase was so cute when I have read the Oreimo for the first time a long time ago.*" This is one of the examples in which the proposed method was able to generate appropriate utterance text reflecting the personality of Ayase. In example 2, the system utterance was "*The ponytail looks great.*", which had the lowest scores (3.57 for naturalness and 3.43 for character-likeness) when the user utterance was "*What is your favorite hairstyle?*" Ayase is characterized by having long straight black hair so this system utterance was not appropriate. There is another character who has a ponytail hairstyle in the anime. This is only one example in which the score for the proposed method was lower than U-AIML in the experiment. In the future, we need to make improvements to prevent such inappropriate utterances from being generated.

Next, examples of generated animated body motion under each motion condition in our experiment are shown in **Fig. 6**. The subjective evaluation scores are mean values under the U-PROP utterance condition. In this example, the system utterance was "*Huh? You are a pervert! Where are you looking?*" This is a

Table 8 Examples of generated utterances in our experiment. The subjective evaluation scores are mean values under the M-NONE motion condition.

Example 1) User utterance: *"I thought that Ayase was so cute when I have read the Oreimo* for the first time a long time ago."*
(* "Oreimo" is a commonly used abbreviation for the light novel "Ore no Imoto ga Konna ni Kawaii Wake ga Nai" in Japanese.)

Utterance condition	System utterance	naturalness score	Character-likeness score
U-AIML	*"It is a pleasure to see you after such a long time."*	3.70	3.67
U-PROP	*"Thank you for praising me!"*	**5.37** (↑)	**5.33** (↑)
U-GOLD	*"You are quite a pervert to say "You are cute" to a junior high school student at night. I'm going to report you to the police!"*	5.10	5.07

Example 2) User utterance: *"What is your favorite hairstyle?"*

Utterance condition	System utterance	naturalness score	Character-likeness score
U-AIML	*"I like beautiful brown long hair. Is your hair long?"*	4.43	4.50
U-PROP	*"The ponytail looks great."*	**3.57** (↓)	**3.43** (↓)
U-GOLD	*"If I do say so myself, I like the current hairstyle because I get compliments when my hair is beautiful."*	5.33	5.43

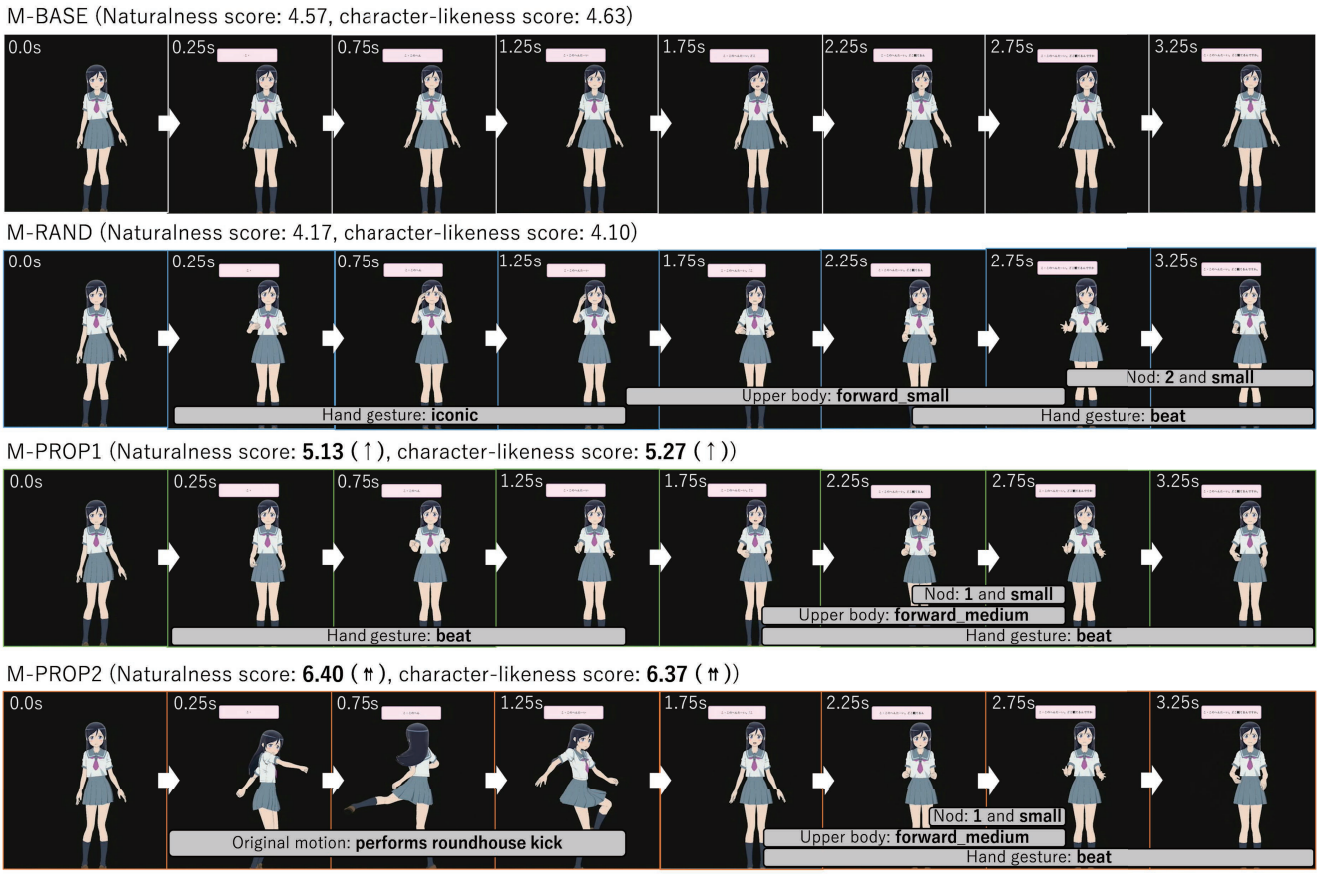

Fig. 6 Example of animated video under each motion condition. System utterance was "*Huh? You are a pervert! Where are you looking?*".

typical example where M-PROP2 had the highest scores (6.40 for naturalness and 6.37 for character-likeness) and M-PROP1 had the second highest scores (5.13 for naturalness and 5.27 for character-likeness). Using M-PROP1, an entirely human-like motion was generated. With M-PROP2, original motion, such as "performs roundhouse kick," was generated for the utterance sentence "You are a pervert!" (from about 0.25 s to 1.6 s). This combination of character-specific utterances and motions often worked very well.

Finally, examples indicating that the animated body motion under U-PROP2 was not very natural in our experiment are shown in **Fig. 7**. The system utterance was "*Wha wha wha wha wha wha, what do you mean? I don't think anything of him! Pervert!*" The naturalness and character-likeness scores were 4.43 and 4.50 under M-NONE, 3.90 and 4.07 under M-RAND, 4.53 and 4.73 under M-PROP1, and 4.60 and 4.77 under M-PROP2. The scores under M-PROP2 were almost the same as the scores under M-NONE. There are two possible reasons for this. One is that the sentence "Wha wha wha wha wha wha," which is a fairly specific colloquial sentence, was not included in our corpus data that we used to build an automatic motion generation model from utterance text. Our model generated the "beat" hand gesture mentioned above for "Wha wha wha wha wha wha," (from about 0.25 s to 1.25 s). However, it seems that this hand gesture motion

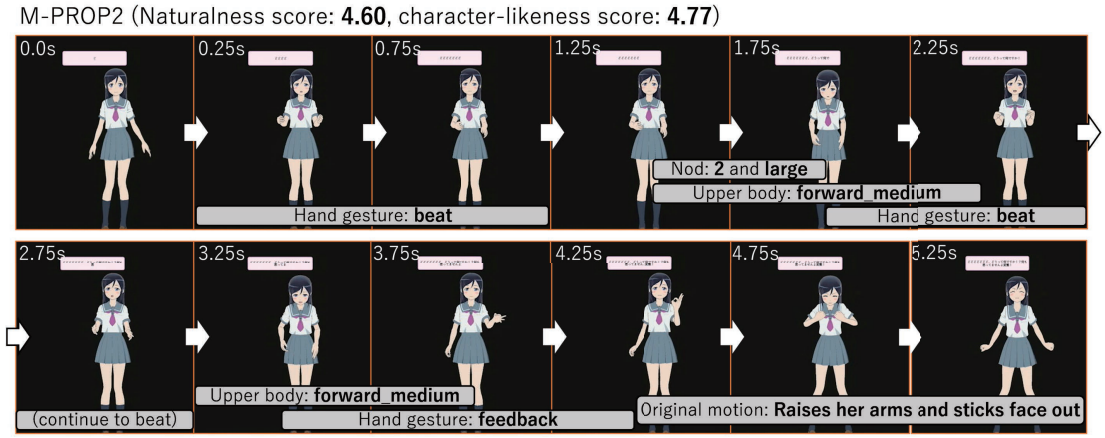

Fig. 7 Example of animated video for which naturalness and character-likeness scores are not high (4.60 and 4.77) under M-PROP2 motion condition. Naturalness and character-likeness scores under other conditions were 4.43 and 4.50 under M-NONE, 3.90 and 4.07 under M-RAND, and 4.53 and 4.73 under M-PROP1. System utterance was "*Wha wha wha wha wha wha, what do you mean? I don't think anything of him! Pervert!*".

was not suitable for this sentence. Second, it is possible that the connection between the original motion and the motion immediately before did not work well. When the last word, "pervert!", was displayed (from about 4.5 s), the original motion of "Raises her arms and sticks face out" was generated. Just before the start of this motion, the "feedback" hand gesture was generated. This caused a sudden change to a different behavior. When original motion is generated, it is important to have a smooth connection with the back and forth body motion. This requires fine-tuning but if not done properly, here is the risk of the user getting an even worse impression. In the future, we need to make improvements to prevent such inappropriate body motion from being generated.

6. Discussion

The subjective evaluation results confirmed that utterance generation and motion generation affected the impression of both naturalness and character-likeness of overall responses. In addition, the rating score of the dialogue system constructed using the proposed system-construction method (U-PROP+M-PROP1 and U-PROP+M-PROP2) was as high as 4.89 and 5.06 for naturalness and 4.85 and 4.98 for character-likeness in a rating-value range of 1 to 7.

The proposed utterance generation (U-PROP) had better naturalness and character-likeness than utterance generation using AIML (U-AIML). The proposed utterance generation method was also suggested to be more useful for generating natural responses reflecting the character's personality than the basic approach of utterance generation. The proposed motion generation method (U-PROP1 and U-PROP2) similarly improved the impression as compared with the static and random motion generation (U-BASE and U-RAND).

As a result comparing our two method of motion generation methods (M-PROP1 and M-PROP2), it was found that for the proposed motion generation method, the evaluation scores for original motion added (M-PROP2) were not higher than for original motion not added (M-PROP1) when an utterance was generated by our proposed model (U-PROP) or a human (U-GOLD). Our motion generation model (M-PROP1) can generate the average motion of many people since it was trained with the movements of 24 people. This suggests that body motion that reflects average movement can improve the impression of responses being natural and character-like without inserting an original movement. Even if an average movement is given to an anime character agent, it would be possible to sense that agent's individuality. This is a very interesting result. Of course, depending on the design and settings of the anime character, this method may not always be effective. This is because it may be better for awkward robots not to behave like a human. However, if it is appropriate for a character to move like a human, our proposed motion generation method can be an effective means of enhancing naturalness and character-likeness of responses. Detailed evaluation of effectiveness using more diverse characters is a task for future work.

Another interesting result is that M-BASE and M-PROP1/M-PROP2 under U-GOLD have the same scores for naturalness and character-likeness. When the quality of utterance generation approaches that of humans, the difference in motion quality might not significantly affect the evaluation of naturalness. Alternatively, the quality of motion generation may need to be further improved as the quality of utterance generation improves.

Our proposed utterance and generation methods allow actually achieving a dialogue system for existing anime/animation characters which has been difficult to now. We cannot claim that our proposed method creates a perfect system but we believe it proves the effectiveness of our new method.

Looking at the results in Tables 4 and 5, the overall responsiveness and character-like scores were similar. The reason for this may be that character-likeness was a very important factor as shown in the evaluation results for the naturalness of the overall responses of the existing anime character in this paper. This demonstrates the importance of generating a response that properly reflects a character's characteristics when creating a chat system for an existing character.

A future technical challenge is to generate whole-body motion

that takes into account the emotion of utterances. Since the utterance text also contains emotional information, our model can possibly generate full-body motion that takes the emotional aspects of the text into account. However, it is thought that estimating emotion from text is not sufficient. In our corpus, response sentences were given one of eight sentiment labels. We would like to build an automatic generation model for motions using these labels as input and emotion information as input.

We plan to carry out additional experiments to handle more samples and to verify in detail whether there is a mutual effect between speech and motion conditions. We will also improve the system-construction method to create a dialogue-agent system using other existing anime characters.

7. Conclusion

In this paper, we proposed a system-construction method for efficiently constructing a text chat system with animation for existing anime characters. We tackled two research problems to generate verbal and nonverbal behaviors. In the generation of verbal behavior, a major issue is how to generate utterance text that reflects the personality of existing characters in response to any user questions. For this problem, we proposed the use of the role-playing question-answering method to efficiently collect high-quality paired data of user questions and system answers that reflect the personality of an anime character. We also proposed a new utterance generation method that uses a neural translation model with the collected data. Rich and natural expressions of nonverbal behavior greatly enhance the appeal of agent systems. However, not all existing anime characters move as naturally and as diversely as humans. Therefore, we proposed a method that can automatically generate whole-body motion from spoken text so that anime characters can make human-like and natural movements. In addition to these movements, we added a small amount of characteristic movement on a rule basis to reflect personality. We created a text-dialogue agent system for a popular existing anime character by using our proposed methods for generating verbal and nonverbal behavior. As a result of a subjective evaluation of the implemented system, our methods for generating verbal and nonverbal behavior improved the impression of naturalness and character-likeness of overall responses. Since generating characteristic motions with a small amount of characteristic movement on the basis of heuristic rules was not effective, our proposed motion generation method, which can generate the average motion of many people, is useful for generating the motion of an existing anime character. Currently, we are applying our utterance and motion generation methods to characters other than Ayase Aragaki to verify their respective effectiveness [13], [30]. Therefore, our proposed methods and system-construction method are likely to contribute greatly to realizing text-dialogue-agent systems using existing characters.

References

[1] Chen, Y., Naveed, A. and Porzel, R.: Behavior and preference in minimal personality: A study on embodied conversational agents, p.49 (online), DOI: 10.1145/1891903.1891963 (2010).

[2] Mairesse, F. and Walker, M.: PERSONAGE: Personality Generation for Dialogue, *Proc. 45th Annual Meeting of the Association of Computational Linguistics*, pp.496–503, Association for Computational Linguistics (online), available from ⟨https://www.aclweb.org/anthology/P07-1063⟩ (2007).

[3] Smith, H.J. and Neff, M.: Understanding the Impact of Animated Gesture Performance on Personality Perceptions, *ACM Trans. Graph.*, Vol.36, No.4 (online), DOI: 10.1145/3072959.3073697 (2017).

[4] Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J. and Dolan, B.: A Persona-Based Neural Conversation Model, *Proc. 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp.994–1003, Association for Computational Linguistics (online), DOI: 10.18653/v1/P16-1094 (2016).

[5] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D. and Weston, J.: Personalizing Dialogue Agents: I have a dog, do you have pets too?, *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp.2204–2213, Association for Computational Linguistics (online), DOI: 10.18653/v1/P18-1205 (2018).

[6] Yu, D., Cohn, M., Yang, Y.M., Chen, C.Y., Wen, W., Zhang, J., Zhou, M., Jesse, K., Chau, A., Bhowmick, A., Iyer, S., Sreenivasulu, G., Davidson, S., Bhandare, A. and Yu, Z.: Gunrock: A Social Bot for Complex and Engaging Long Conversations, *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*): *System Demonstrations*, pp.79–84, Association for Computational Linguistics (online), DOI: 10.18653/v1/D19-3014 (2019).

[7] Higashinaka, R., Sadamitsu, K., Saito, K. and Kobayashi, N.: Question answering technology for pinpointing answers to a wide range of questions, *NTT Technical Review*, Vol.11, No.7 (2013).

[8] Ments, M.V.: *The Effective Use of Role Play: Practical Techniques for Improving Learning*, Kogan Page Publishers (1999).

[9] Vinyals, O. and Le, Q.: A neural conversational model, arXiv preprint arXiv:1506.05869 (2015).

[10] Walker, S., Robertson, S.E., Boughanem, M., Jones, G.J. and Jones, K.S.: Okapi at TREC-6 Automatic ad hoc, VLC, routing, filtering and QSDR, *Proc. TREC*, pp.125–136 (1997).

[11] Leuski, A., Patel, R., Traum, D. and Kennedy, B.: Building effective question answering characters, *Proc. SIGDIAL*, pp.18–27 (2009).

[12] Sekine, S., Sudo, K. and Nobata, C.: Extended Named Entity Hierarchy, *Proc. LREC* (2002).

[13] Higashinaka, R., Mizukami, M., Kawabata, H., Yamaguchi, E., Adachi, N. and Tomita, J.: Role play-based question-answering by real users for building chatbots with consistent personalities, *Proc. 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp.264–272, Association for Computational Linguistics (online), DOI: 10.18653/v1/W18-5031 (2018).

[14] Klein, G., Kim, Y., Deng, Y., Senellart, J. and Rush, A.: OpenNMT: Open-Source Toolkit for Neural Machine Translation, *Proc. ACL 2017, System Demonstrations*, pp.67–72, Association for Computational Linguistics (online), available from ⟨https://www.aclweb.org/anthology/P17-4012⟩ (2017).

[15] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J.: Distributed representations of words and phrases and their compositionality, *Proc. NIPS*, pp.3111–3119 (2013).

[16] Lohse, M., Rothuis, R., Gallego-Pérez, J., Karreman, D.E. and Evers, V.: Robot Gestures Make Difficult Tasks Easier: The Impact of Gestures on Perceived Workload and Task Performance, *Proc. SIGCHI Conference on Human Factors in Computing Systems*, *CHI '14*, pp.1459–1466, ACM (online), DOI: 10.1145/2556288.2557274 (2014).

[17] Ishi, C.T., Haas, J., Wilbers, F.P., Ishiguro, H. and Hagita, N.: Analysis of head motions and speech, and head motion control in an android, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.548–553 (2007).

[18] Ishi, C.T., Ishiguro, H. and Hagita, N.: Head motion during dialogue speech and nod timing control in humanoid robots, *ACM/IEEE International Conference on Human-Robot Interaction*, pp.293–300 (2010).

[19] Kadono, Y., Takase, Y. and Nakano, Y.I.: Generating Iconic Gestures Based on Graphic Data Analysis and Clustering, *The 11th ACM/IEEE International Conference on Human Robot Interaction*, *HRI '16*, pp.447–448, IEEE Press (online), available from ⟨http://dl.acm.org/citation.cfm?id=2906831.2906920⟩ (2016).

[20] Imamura, K.: Analysis of Japanese dependency analysis of semi-spoken words by series labeling, *Proc. Annual Meeting of the Association for Natural Language Processing*, pp.518–521 (2007).

[21] Schroff, F., Kalenichenko, D. and Philbin, J.: FaceNet: A Unified Embedding for Face Recognition and Clustering, *CoRR*, Vol.abs/1503.03832 (online), available from ⟨http://arxiv.org/abs/1503.03832⟩ (2015).

[22] McNeill, D.: *Hand and Mind: What Gestures Reveal About Thought*,

University Of Chicago Press (1996).
[23] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. and Sloetjes, H.: ELAN a Professional Framework for Multimodality Research, *International Conference on Language Resources and Evaluation* (2006).
[24] Fuchi, T. and Takagi, S.: Japanese Morphological Analyzer using Word Cooccurrence -JTAG, *International Conference on Computational Linguistics*, pp.409–413 (1998).
[25] Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T. and Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing, *International Conference on Computational Linguistics*, pp.928–939 (2014).
[26] Meguro, T., Higashinaka, R., Minami, Y. and Dohsaka, K.: Controlling listening-oriented dialogue using partially observable Markov decision processes, *International Conference on Computational Linguistics*, pp.761–769 (2010).
[27] Higashinaka, R., Meguro, T., Sugiyama, H., Makino, T. and Matsuo, Y.: On the difficulty of improving hand-crafted rules in chat-oriented dialogue systems, *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA*), pp.1014–1018 (2015).
[28] Kinsui, S.: *Vaacharu nihongo: Yakuwarigo no nazo* (*in Japanese*), Iwanami Shoten (2003).
[29] Miyazaki, C., Hirano, T., Higashinaka, R. and Matsuo, Y.: Towards an Entertaining Natural Language Generation System: Linguistic Peculiarities of Japanese Fictional Characters, *Proc. SIGDIAL*, pp.319–328 (2016).
[30] Ishii, R., Katayama, T., Higashinaka, R. and Tomita, J.: Generating Body Motions using Spoken Language in Dialogue, *Intelligent Virtual Agents* (*IVA'18*) (2018).

Ryo Ishii received his M.S. degree in engineering from the Tokyo University of Agriculture and Technology and joined the NTT Corporation in 2008. He received his Ph.D. degree in informatics from Kyoto University in 2013. He is currently a senior research scientist at NTT Media Intelligence Laboratories and a visiting scholar at Carnegie Mellon University. His research interests are multimodal interaction and social signal processing. He is a member of IEICE, JSAI, and HIS.

Ryuichiro Higashinaka received his B.A. in environmental information, Masters of Media and Governance, and Ph.D. from Keio University, Kanagawa in 1999, 2001, and 2008, respectively. He is currently a professor at Nagoya University. He is a visiting distinguished senior researcher at NTT Media Intelligence Laboratories. His research interests include building question-answering systems and spoken-dialogue systems. From November 2004 to March 2006, he was a visiting researcher at the University of Sheffield in the UK. He received the Prize for Science and Technology of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2016. He is a member of JSAI and ANLP.

Koh Mitsuda received his M.S. degree in engineering from the Tokyo Institute of Technology and joined the NTT Corporation in 2015. He is currently a researcher at NTT Media Intelligence Laboratories and a Ph.D. student at Tsukuba University. He is engaged in research on natural language processing and dialogue systems. He is a member of ANLP and JSAI.

Taichi Katayama received his M.S. degree in engineering from the University of Tsukuba and joined the NTT Corporation in 2011. He transferred to NTT Docomo, Inc. in 2019. He is currently an assistant manager at NTT Docomo, Inc. His research interests are natural language processing and data mining. He has worked on natural language processing and on the development of translation, FAQ, dialogue systems, and so on.

Masahiro Mizukami received his Ph.D. degree in engineering from the Nara Institute of Science and Technology and joined the NTT Corporation in 2017. He is currently a research scientist at NTT Communication Science Laboratories. His research interest is chat-oriented dialogue systems.

Junji Tomita received an M.S. degree in computer science from Keio University, Kanagawa in 1997. He joined NTT in 1997. He was a visiting scholar at the University of Washington in 2005. He worked for NTT Resonant Inc. from 2006 to 2017. He received a Ph.D. from Keio University in 2012. He is currently a senior research engineer at NTT Media Intelligence Laboratories. His research interests include natural language processing, information retrieval, and dialogue systems. He is a board member of DBSJ.

Hidetoshi Kawabata joined Rakuten, Inc. in 2007. He was a West Japan Area Manager for Rakuten Market Business in 2011. He joined Dwango Co., Ltd. in 2011. He was in charge of E-Commerce Business Development as a NicoNico Channel sales manager in 2012. He is currently a business manager for the NicoNico Channel and joined the "Narikiri AI" project in 2016. He has also belonged to the Kadokawa Digital Strategy Bureau since 2019.

Emi Yamaguchi joined Toranoana Co., Ltd. in 2004. She was a chief planner for the Customer Development Division and person responsible for the development and sales of entertainment products for women in 2005. She joined Sony Digital Entertainment Service Inc., where she launched the original comic section in 2009 and was in charge as manager of original comic production and sales in 2010. She joined Dwango Co., Ltd in 2014. She was in charge of the launch of the "Narikiri AI" project. She is currently engaged in business for the NicoNico Channel.

Noritake Adachi joined Nippon Timeshare Co., Ltd. in 2000. He then joined Dwango Co., Ltd. He was a section manager at a smartphone music site. He is currently engaged as a project manager in a conversation AI creation project called the "Narikiri AI" project (2016-present), he was involved in Artificial Intelligence Fundraising (2018), and he is a winner of the Horse Racing Programming Competition "Denno Prize" (2015-2018).

Yushi Aono received his Ph.D. degree in engineering science from Osaka University and joined the NTT Corporation in 1999. After working for NTT Cyber Space Laboratories, NTT Advanced Technology, and the NTT Research and Planning Division, he became and is currently a project manager (principal researcher) in the Cognitive Information Processing Laboratory at NTT Media Intelligence Laboratories. He is engaged in research on speech recognition and speech synthesis. He is a member of the IEICE.