Task and User Adaptation based on Character Expression for Spoken Dialogue Systems

Kenta Yamamoto

Graduate School of Informatics
Kyoto University

Abstract

Many spoken dialogue systems such as smart speakers and conversational robots are used in human society. Conversational robots or agents are given various social roles such as a museum guide and an interviewer. It is important for the system to establish a relationship with the user in order to be used continuously over a long time. One of the approaches to this problem is to define a character of the spoken dialogue system. If the user feels a desirable character impression from behaviors of the system, they would be attached to the system. However, the desirable character expression should be dependent on the dialogue task and the user. The goal of this study is to make the spoken dialogue system express the appropriate character according to the dialogue task and the user. This thesis presents a character expression model for spoken dialogue systems based on the following behaviors specific to the spoken dialogue: utterance amount, backchannels, fillers, and switching pause length. The model controls these behaviors of the spoken dialogue system to give the intended character impression. This study investigates two approaches to the question of the appropriate character for the spoken dialogue system. One is task adaptation, where the system expresses the appropriate character for the task. The other is user adaptation, where the system expresses the appropriate character for the user personality.

Chapter 1 describes the background of this study, which includes spoken dialogue systems and human-robot interaction. In Chapter 2, related works on character and user adaptation of the spoken dialogue systems are reviewed. Chapter 3 describes the platform of the spoken dialogue system for an android ERICA used in this study. The robot has a physical body and generates various human-like dialogue behaviors, which are important for the character expression. In Chapter 4, the baseline character expression model using spoken dialogue behaviors is designed. It starts with data collection for training the model, in which subjects listened to the dialogue data and evaluated their character impressions of the dialogue robot. The statistical analysis showed correlation between the character impression and spoken dialogue behaviors. The baseline character representation model was constructed by using this dataset. The proposed model is given three character traits (extrovert, emotional instability, and politeness) and controls the four dialogue behaviors: utterance amount, backchannel frequency, filler frequency, and switching pause length. Subject experiments

confirmed that the proposed model expressed the characters as intended. The corpus analysis also showed that the proposed model can predict various speakers' characters in the corpus.

Chapter 5 addresses the enhancement of the character expression model using neural networks, which represent the three character traits for model training. It is necessary to collect many pairs of data on the character impressions and behaviors. Therefore, semi-supervised learning is proposed based on a Variational Auto-Encoder (VAE) which combines the limited amount of the labeled pair data with unlabeled corpus data. It is shown that the proposed model can express given characters more accurately than the baseline model with only supervised learning. Using the proposed learning method, moreover, the spoken dialogue system can express an additional dialogue behavior that is not included in the labeled data.

Chapter 6 addresses task adaptation using the character expression model. The corpus analysis shows that the tendency of characters is different among dialogue tasks. The character expression model is implemented for the tasks of job interview and laboratory guide. Subjective evaluations of the dialogue video show that expressing a character in accordance with the dialogue task by the proposed model improves the user impression of the appropriateness in formal dialogue such as job interview. In the real dialogue experiment, it is confirmed that characters designated to the task received higher evaluation scores than the baseline system.

Chapter 7 addresses user adaptation using the character expression model. The analysis of the speed dating corpus shows that the combination of the subjects' personality affects the favorable impressions. Based on the analysis, a character adaptation model that controls spoken dialogue behaviors is designed and developed. This model classifies the user personality and the system character into four classes and determines the most appropriate character of the system according to the user personality. We conducted real dialogue experiments, in which subjects talked with a robot as a laboratory guide (task-oriented dialogue) and chit-chat (non-task-oriented dialogue) in four different character conditions. It is confirmed that the extrovert character was preferred for items on the laboratory guide's skill and that the character matched to the user personality was preferred for items on how easy to talk with the robot.

This thesis is concluded in Chapter 8.

Acknowledgements

This work was accomplished at Kawahara Laboratory, Graduate School of Informatics, Kyoto University. I express my gratitude to all the people who helped me in this work. First, I would like to express my special thanks and appreciation to my supervisor Prof. Tatsuya Kawahara. His comments were essential and insightful to the progress of this work. Without his continuous support, this work would not have been completed. I also express my special thanks and appreciation to the members of my dissertation committee, Prof. Takatsune Kumada and Prof. Sadao Kurohashi for their time and valuable comments and suggestions. This thesis cannot be accomplished without Dr. Koji Inoue, Dr. Keiko Ochi, and Dr. Divesh Lala in the spoken dialogue system group of Kawahara Lab. They supported me and gave much time to meaningful discussions. Furthermore, I appreciate their cooperation to build a spoken dialogue system for the autonomous android ERICA. I also deeply thank the other members of the spoken dialogue system group in Kawahara Lab. I appreciate discussions with Mr. Haruki Kawai. I would like to thank Prof. Koichiro Yoshino, Dr. Akishige Yuguchi and Dr. Seiya Kawano, who are members of RIKEN, for their hosting my internship. I would like to thank Dr. Ritsuko Iwai at RIKEN for discussing about personality in psychology. I also express my thanks to members in the ERATO Ishiguro symbiotic human-robot interaction project. Prof. Hiroshi Ishiguro of Osaka University provided me with insightful comments. Dr. Carlos Toshinori Ishi, Dr. Takashi Minato and Dr. Kurima Sakai of ATR continuously supported me in the implementation of the spoken dialogue system for the android robot ERICA. I am also deeply grateful to Prof. Katsuya Takanashi, and Dr. Shizuka Nakamura for their kind supports. I would like to thank the current and former members of the Kawahara Lab. I am grateful for the comments and supports of Prof. Kazuyoshi Yoshii, Dr. Masato Mimura, and the other members. This work was supported by the Japan Society for the Promotion and Science (JSPS) with their financial support as a Fellowship for Young Scientists (DC1). Finally, I am truly grateful to my family for their daily and continuous support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Spoken dialogue systems are widely used in our daily lives. The goal of this study is to make the dialogue system more friendly to users. The challenges in spoken dialogue systems and the approach in this thesis are described.

1.1 Background

Spoken dialogue systems have been developed for long and many systems such as smart speakers and conversational robots are used in real world. Voice assistants on smart phones and smart speakers are often used for simple tasks such as information retrieval and music player. On the other hand, conversational robots and virtual agents are used for long conversations, such as museum guides [1], attentive listening [2], and psychological counselors [3]. To satisfy the user, the system must provide natural, non-boring interaction. In recent years, machine learning techniques have made great strides in understanding user utterances and generating responses [4, 5]. These models made it possible to generate a natural dialogue in short exchanges, such as a question-answer and chit-chat. However, they are not enough to realize the human-level of interaction. It is important not only to understand utterances but also to establish a relationship with humans and encourage the user to use the system sustainably. Various methods have been proposed for building a relationship with the user, such as showing empathy to the user [6], an acoustic-prosodic entrainment to the user [7], a mimicry of the user behaviors [8], and a self-disclosure by the system [9].

All of these elements are very important, but the user perceives the character as an overall impression of the dialog system. Character is especially easy to be perceive in systems that have the appearance of an interactive agent or robot. Characters make the dialogue system feel special and different from other systems to the user, thus making it easier to build a relationship with the user. On the other hand, the character impression is affected by many factors such as the system utterance, eye gaze, backchannels, loudness and pitch of the voice. Therefore, when proposing a new character expression model, it is necessary to avoid conflicts with existing systems. Whereas many systems control the system utterance, this thesis addresses behaviors associated with spoken dialogue.

The goal of this thesis is to construct a natural relationship between humans and spoken dialogue systems. Specifically, this thesis addresses the character expression model for spoken dialogue systems according to the dialogue task and the user personality.

1.2 Spoken Dialogue Systems

The initial studies on spoken dialogue systems were conducted about 50 years ago. The origins of the dialogue system are ELIZA [10] and SHRDLU [11]. ELIZA is a text chatting system using simple database retrieval based on keyword matching. SHRDLU understands the text commands and moves blocks in a virtual world. The basic concepts of these systems are similar to those of today's systems, but they are essentially text-based, not using speech.

Following the development of automatic speech recognition and speech synthesis, spoken dialogue systems began to be built around 1990. One of the pioneer systems is VOYAGER [12], the forerunner of today's task-oriented dialogue systems. Airline Travel Information System (ATIS) [13] is a project in which data collection and system construction were conducted for the task of retrieving flight information. These provided a basis of task-oriented dialogue systems. The development of deep learning technology [14] has dramatically improved the accuracy of speech recognition and has led to the practical application of spoken dialogue systems. Meanwhile, voice assistants in smart phones and smart speakers have been developed. These systems are used for purposes such as information retrieval and music player control and has ability to perform some simple conversation. However, it difficult to have long conversations with these systems.

Conversational robots are studied and developed based on humanoid robots [15, 16]. Interactive virtual agents were implemented as a counselor [3], a museum guide [1], and a job interviewer [17, 18]. They have the advantage to be able to utilize multi-modality. Robots and virtual agents can coordinately generate their multi-modal behaviors in accordance with user behaviors. In this context, an android ERICA has been developed. For understanding the surrounding environment and the conversational behaviors including user utterances, the system of ERICA integrates information from sensors such as microphone arrays and cameras. ERICA generates various behaviors such as facial expressions, eye gaze, head nodding, and arm gestures. Therefore, conversation with ERICA is natural through various expressions and behaviors. ERICA's appearance is designed for social tasks such as attentive listening and job interview. In this thesis, ERICA is used as the platform for a spoken dialogue system. It is expected that human-like robot is easy to build trust with the user.

1.3 Character of Spoken Dialogue Systems

In this thesis, “personality” is used as a psychological dimension for classifying users, and “character” is used as the impression that the dialog system gives to the user. Humans are known to anthropomorphize systems and perceive a character from them [19]. For example, the system which speaks positively gives an extroverted impression. Character is also important for a dialogue system to achieve human-level interaction [20]. If all conversational robots have the same look and the same speaking style, the user feels that the dialogue is very boring. On the other hand, if each dialog system has its own character and consistently speaks with the character, it is expected that users want to continue with the system and deepen their relationship with it.

Persona is the most widely-used method to achieve a character expression. This is a method of expressing individuality by specifying the personal information of the system in detail. The first dialogue system, ELIZA [10], also achieved naturalness of dialogue by using the persona of a female psychiatrist. A umber of studies have been conducted to express characters based on their age and gender [21], and to provide personas that mimic a celebrity and a animation character [22]. Paired dataset of sentences explaining the persona and dialogue examples is published [23]. On the other hand, methods have been proposed to express characters based on psychological personalities [24]. Since these methods are intended for text chat, they control the content of utterance and are dependent on the dialogue task.

This thesis addresses character expression model using spoken dialogue behaviors. Since this model allows independent control regardless the dialogue content, it can be combined with existing systems to achieve character expression.

1.4 Task and User Adaptation of Spoken Dialogue Systems

The purpose of the dialog depends on the given task of the system. Dialogues in which the system’s purpose is clear are referred to as task-oriented dialogues, while those in which the purpose is the dialogue itself are referred to as non-task-oriented dialogues.

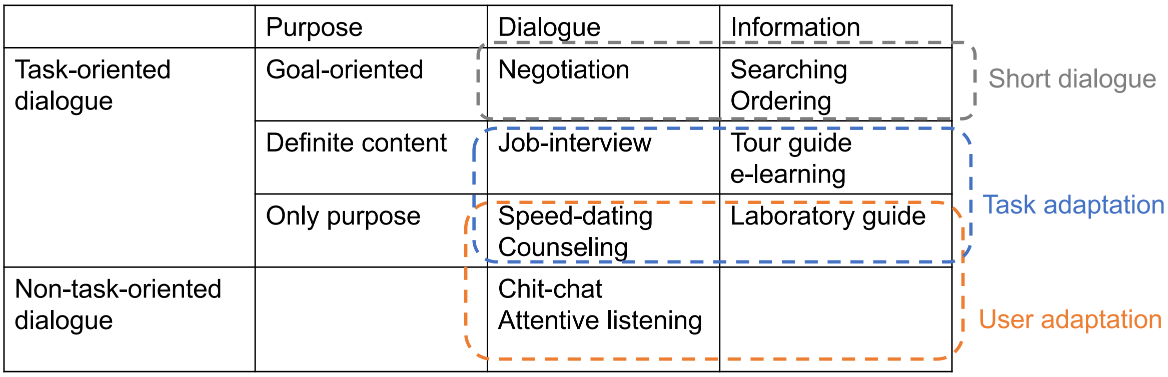
In the task-oriented dialog, tasks can be classified into three categories: goal-oriented, dialogue with definite content, and dialogue with only purpose. Goal-oriented dialogue has a clear goal, and if the desired information is exchanged between the user and the system, the goal is achieved. The main objective of a specific task is to communicate and exchange necessary information, such as searching, ordering, and reception. Therefore, it is possible to objectively judge whether the goal has been achieved. These systems are expected to complete the dialogue in the shortest time. Dialogue with definite content does not clearly indicate whether the dialogue has been accomplished or not, but the content of the dialogue is clearly defined. For example, a sightseeing guide or an e-learning agent explains the content that has been prepared in advance. Another example of the task with definite content is a job interview system with predefined questions. These tasks require dialogue that the user can clearly understand. Finally, dialogue with only purpose does not have a clear goal and a clear content. For this reason, the dialogue cannot be established simply by talking about what has been prepared in advance. Examples include speed-dating and counseling. These systems need to be designed so that they can respond to the free speech of the user.

Examples of non-task-oriented dialogue are chit-chat and attentive listening. These dialogues have neither the purpose nor the task. These dialogues are important for users to get to know each other and become attached to each other. In these dialogues, it is important to make the user feel comfortable in speaking.

Conventionally, an independent system have been built for each dialogue task or non-task. Therefore, the system also speaks in a way that was appropriate for each dialogue. However, a dialogue robot such as ERICA needs to change the way it speaks for each dialogue because it plays multiple social roles. Fig 1.1 shows the classification and adaptation method for dialogue tasks. Goal-oriented dialogues are completed with short dialogues and are not the target of this study, which aims at dialogue continuation. For dialogue with the definite content and with only the purpose, the typical dialogue is determined for each task. Therefore, it is necessary for the spoken dialogue systems to adapt its speaking style to each task. This is called task adaptation, which is one of the topic of this thesis study. Dialogue with only purpose and non-task-oriented dialogue require different approaches. As the content of these dialogues is free, the way each user speaks differs greatly from user to user. Therefore, it is necessary to realize a dialogue in which the user is comfortable in speaking. This is called a user adaptation, which is the second purpose of this thesis. In this thesis, task adaptation and user adaptation are realized through character expression.

1.5 Data Collection for Spoken Dialogue Systems

Collecting dialogue data is a very important issue to build a dialogue system. In recent years, dialogue systems using deep learning have been generating very natural dialogues [4, 5].



	Purpose	Dialogue	Information
Task-oriented dialogue	Goal-oriented	Negotiation	Searching Ordering
	Definite content	Job-interview	Tour guide e-learning
	Only purpose	Speed-dating Counseling	Laboratory guide
Non-task-oriented dialogue		Chit-chat Attentive listening	

Fig. 1.1 Task and user adaptation for dialogue tasks

These systems train their language models using a large amount of text data from the Web. The data size has been increasing year by year, and the latest models use 570GB data [5].

On the other hand, it is difficult to collect as much data for spoken dialogue as it is for text data. It is necessary to record actual conversations between people in real time. In addition, spoken dialogue requires a large number of annotations compared to text data. For example, speech transcription and time stamps, multi-modal data such as eye gaze and other annotation are needed. Behaviors such as backchannels and fillers, which do not exist in text, are observed in spoken dialogue. As a result, the data size of spoken dialogue is much smaller than that of text data. In addition, character impression, which is addressed in this study, require a new annotation because there is almost no public data available. This study proposes semi-supervised learning using a variational auto-encoder as a method to compensate for the lack of data to build a character expression model.

1.6 Approaches

This thesis addresses a character expression model for spoken dialogue systems. Figure 1.2 shows the concept of the proposed system. Character expression for spoken dialogue systems means each system speaks with consistency. It is shown that the character expression of spoken dialogue systems leads to increasing user engagement and naturalness in dialogue [19, 25, 26]. Therefore, previous studies addressed character expression models for dialogue systems [24, 22, 27, 28]. In these studies, the system expresses a character by controlling its utterances. This thesis presents a character expression model by controlling behaviors in spoken dialogue. Specifically, the system controls four dialogue behaviors: utterance amount, backchannel frequency, filler frequency, and switching pause length. The advantage of this system is that the character expression model can be added to existing spoken dialogue

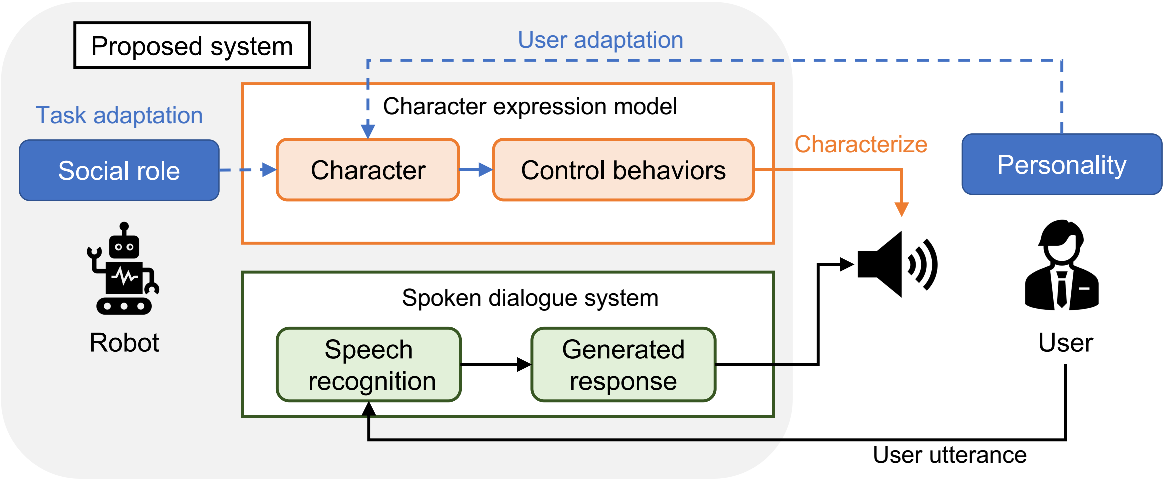


Fig. 1.2 Overview of the character expression model for spoken dialogue systems

systems. To build the model, the relationships between character impressions and behaviors are analyzed using dialogue data labeled with character impressions. Based on this result, a baseline character expression model is constructed using the logistic regression model.

Pair data of the spoken dialogue behaviors and character evaluations are required to train the character expression model. However, this manual annotation is very costly, and thus the variety and amount of behavior patterns are limited. On the other hand, various dialogue corpora are available and they contain many dialogue behaviors, but the corpus data cannot be directly used for supervised learning of the character expression model as there are no manual annotations for the character. To make the character expression model more robust and accurate, it is important to exploit both manually annotated data (supervised) and dialogue corpus data (unsupervised). To address this problem, semi-supervised learning based on a variational auto-encoder (VAE) is proposed to utilize not only manually annotated labels but also dialogue corpus data. The proposed method is designed to compensate for the data-sparseness with natural dialogue behavior data. Utilization of dialogue corpus data as unlabeled data can be applied to other expression tasks (e.g., emotion expression through dialogue behaviors) that are affected by data-sparseness due to a limited amount of training data.

It is also important what character the system should express. This thesis addresses character adaptation to determine the system character according to the dialogue task and the user. The analysis of the dialogue corpus suggests that there is a bias in the characters required for each task. The effect of task adaptation will be evaluated in job interview (definite content dialogue), laboratory guide (only purpose dialogue) and attentive listening

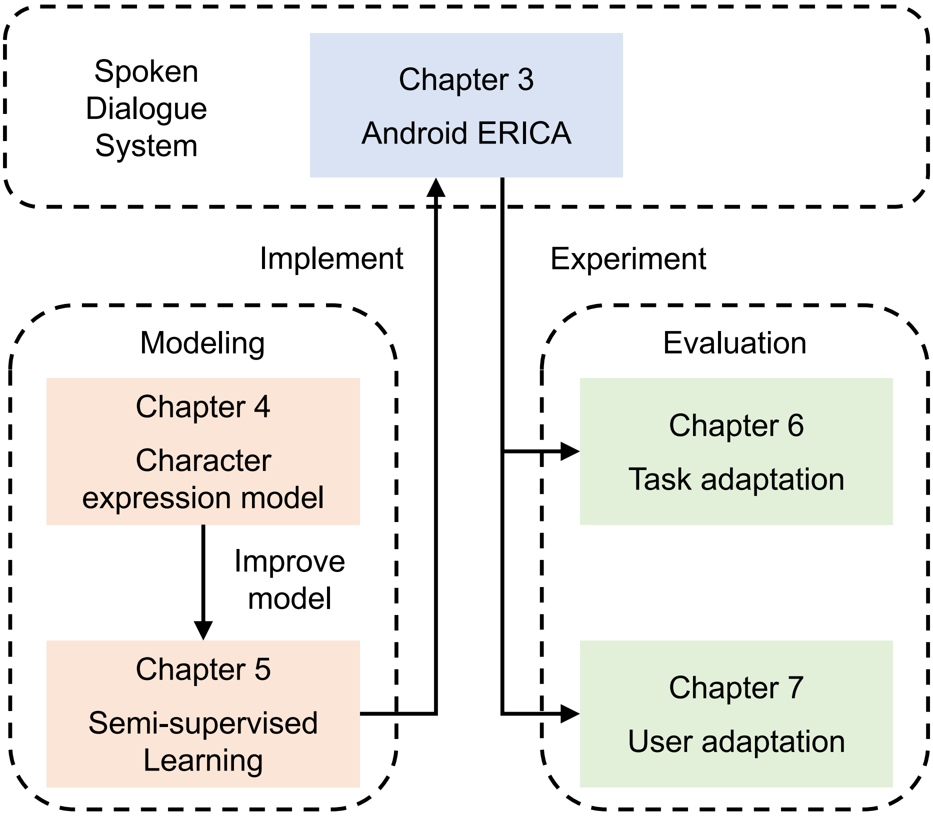


Fig. 1.3 Organization of this thesis

(non-task-oriented dialogue). It is expected that the more definite content of the dialogue, the more the task-appropriate character expression improved the impression of the task.

Previous studies analyzed the influence of the combination of the user personality and the system character on the impression of the dialogue [29, 30]. However, since each study analyzed the data based on different personality scales, the analysis results differ from one study to another. On the other hand, a large amount of data is required to model an optimal character when using generic characters. Therefore, this thesis proposes four generic character classes for character adaptation. These four classes were defined from the annotation data of character impressions on the dialogue corpus. A model that expresses four character classes is implemented in ERICA. The effect of task adaptation will be evaluated in laboratory guide (only purpose dialogue) and chit-chat (non-task-oriented dialogue).

1.7 Organization of This Thesis

Figure 1.3 summarizes the organization of this thesis. Chapter 2 addresses reviews of related works. Chapter 3 describes the architecture of ERICA and the dialogue corpus collected using

ERICA. In Chapter 4, the character expression model is presented. This chapter describes the behavior used for character expression. Chapter 5 addresses the semi-supervised training method for the character expression model. The model built in the previous chapter is improved to be more robust. In Chapter 6, task adaptation using character expression is proposed. In Chapter 7, user adaptation using character expression is proposed. Character classes for user adaptation are defined using the corpus analysis. The effect of the character expression according to the user personality is evaluated by subject experiments. Chapter 8 concludes the thesis with the future direction of the character expression of spoken dialogue systems.

Chapter 2

Literature Review

This chapter overviews the related works. At first, definitions of character in previous studies are presented. Second, user adaptation methods on dialogue systems are introduced. Finally, spoken dialogue behaviors used for the character expression model is explained.

2.1 Definition of Character in Psychology

In psychology, concepts such as character and personality have been used to explain the tendencies of human behaviors. The history of personality is as far back as Ancient Greece. The idea of classifying people into several typical personality classes was the prevailing one.

Many of the personality definition used today have been proposed since the 20th century. Eysenck posited there were two pertinent dimensions of personality such as extroversion and neuroticism [31]. These could be combined to describe four key personality types. The Myers-Briggs Type Indicator (MBTI) which classifies personality into 16 types based on Jung's theory [32]. Some researches were focused on the structure of personality and influenced by "Lexical Hypothesis" [33]. This hypothesis posits that important expressions of personality are contained in everyday speech. Allport and Odbert identified about 4,500 words in the English language that could be used to describe personality using the dictionary [34]. Based on the lexical hypothesis, Goldberg found five major personality factors via statistical analysis and named them the Big Five [35]. The five-factor model (Big Five personality scale) is one of the most widely-used and reliable scales [36]. The Big Five scale assumes the existence of five traits: extroversion, emotional instability, agreeableness, conscientiousness, and openness. These five traits have been shown to be good predictors of patterns of human behaviors [37]. Many questionnaires of Big Five scale [38, 39, 36] have been developed and are easy to use for experiments.

It is claimed that the Big Five dimensions of emotional instability, extroversion, openness, agreeableness, and conscientiousness represent the highest level in the hierarchical structure of personality. It was proposed two factors at a higher level of abstraction based on correlations between Big Five traits [40]. The first superordinate factor (alpha) is the superordinate of emotional instability, agreeableness and conscientiousness. The second factor (beta) is the superordinate of extroversion and openness.

In this study, system characters are defined based on these psychological researches.

2.2 Character in Dialogue Systems

Some research has been done on the effects of character and personality expression in dialogue systems. It is stated that personalization is important for dialogue systems to realize human-level dialogue [20, 19]. The expression of character in a dialogue system has been shown to have an impact on trust-building and task accomplishment [41, 42].

The sentence generation system PERSONAGE [24] generates response sentences according to the Big Five parameter. In addition, a profile-based method of expressing character, rather than character expression by transforming surface sentences, is also examined. The approach for defining character is using the sentences about system's information like "I like to travel.", called as *persona* [23]. In this study, corpus data set was created with the goal of giving the dialogue system a personality. The data set is composed of multi-turn dialogues between two crowd workers who are each given about five sentences of persona text containing character settings. It is a method to generate utterances with personality assigned to them using dialogue act tags and personality-related features as input to a sequence-to-sequence model using LSTM [43]. It had been studied the method to generate characterized utterances by embedding speaker information in the decoder of a sequence-to-sequence model [44]. A character expression method using role-playing dialogue data is also proposed [22]. Methods have also been proposed to maintain dialogue consistency according to persona using large-scale language models [45]. However, it is difficult to make the persona data for all dialogue systems, so they need to write persona sentences for each specific dialogue system. A method of embedding personas in system utterances using encoder and memory network was proposed [46]. 5 million personas and 700 million persona-based dialogues are used for this model training.

Recently, it is proposed methods that adapts a pre-trained model such as BERT [47] and T5 [48] to a target task by fine-tuning. A method colled prompt was proposed for adapting a pre-trained model to a target task without updating their parameters. This mehotd is a few-shot learning based on language models with manually created task descriptions and a few task examples. Prompt-tuning is a method for automatically optimizing a prompt without creating it by manual. The character expression model using prompt-tuning was proposed [49]. Such a method is expected to generate natural responses using small dataset.

Besides, these studies suppose text dialogue, which is different from spoken dialogue. In this study, the part of the Big-Five scales are used to build the character expression model.

2.3 User Adaptation

Many studies have confirmed the effectiveness of characterization according to the user. Studies about human agent interaction showed that personality affects the user's preferences and impressions of the system [29, 30]. A method for controlling the dialogue strategy based on a user model is proposed [50]. In this user model is based on three dimensions: the skill level in use of the system, the knowledge level about the target domain, and the degree of urgency.

It is shown that users are more trusting and satisfied with chat systems that are similar to their personalities [51]. Some user adaptation methods of dialogue systems were addressed in previous studies. One method accumulates the user's dialogue history and makes the user's profile [52]. This method assumed that the same user continues to use the system such as a smart speaker. In the character expression method, the system controls responses from predefined persona descriptions [53]. However, this methods can not adapt users who do not disclose their personas. Some models that obtain user persona information from the dialogue history were also studied [54–56]. Other methods were proposed using the user intent, and proficiency with the system or identifying the user preferences [57, 58]. In these methods, continuous use of the system is necessary to create the model of the user.

This study addresses a method to adapt the behavior of the system using personality to deal with first-time interaction.

2.4 Behaviors on Spoken Dialogue Systems

This section describes the behaviors required for a spoken dialogue system to achieve natural dialogue. These behaviors are observed in only spoken dialogue, not text chat dialogue.

A backchannel is a short reaction signal by the listener such as *"Yeah"* in English and *"Un"* in Japanese. It helps to express the listener's understanding and empathy to the speaker. Bachchannels are classified into six categories: Responsive interjections, Expressive interjections, Lexical reactive expressions, Repetitions, Completions, and Assessments [59]. The appropriate use of backchannels is important for the spoken dialogue systems to achieve a human-like dialogue. For this reason, many studies proposed backchannel prediction models. Prosodic [60], linguistic, and syntactic information [61, 62] are mainly used for backchannel prediction. In recent years, prediction model using a machine learning model such as recurrent neural network (RNN) have been proposed [63]. On the other hand, if the system uses the same backchannel all the time, it gives the impression that the system does not listen to the utterance. However, predicting the category of backchannel is more difficult than predicting the timing. Some studies proposed the model predicting the type of backchannels using prosodic and linguistic information [64, 65]. It is still difficult to predict the type of backchannels in real time.

A filler is a short phrase filling the silence to hold or take the conversational floor such as "*Well*" in English and "*E-*" in Japanese [66, 67]. Filler is one of the important parts for conversational robots to achieve human-like dialogue [68]. Using a filler by a conversational robot moderated the user's impression toward a long pause [69]. It was shown that the robot can smoothly acquire turns of speech by using fillers [70]. The timing of fillers depends on the language content and filler prediction models are proposed [71, 72]. As the frequency of fillers varies widely among individuals, it is difficult to predict fillers.

Turn-taking is moving a floor, which means a right to speak, from a speaker to a listener. In a question-answer system such as that implemented in smartphones and smart speakers, the start of speech is specified by either a button operation (push-to-talk) or a specific phrase (wake-word). To achieve a human-level conversation, it is essential to learn interaction patterns that resemble human's turn-taking. Common spoken dialogue systems use a time-out method. These systems speak after a period of silence (e.g. 2 seconds) has elapsed from the end of a user's utterance. If this latency is short, the system's utterance may overlap with the user's utterance. Turn-taking prediction models using speech and language features have been proposed [73, 74]. On the other hand, it has been pointed out that there are individual differences in turn-taking, and that even human prediction accuracy is about 70% [75]. This is due to the arbitrary nature of turn taking.

The rate of turn acquisition is also related to the occupancy rate of the utterance. Speakers with a high rate of utterance occupancy speak a great deal of content and actively acquire turns. On the other hand, reluctant speakers do not take many turns and do not speak much on their own. Whether speech is verbose or brief has a significant impact on the impression of dialogue. Thus, there are individual differences in the utterance amount. Methods to show favorable impressions to the user by controlling the amount of information in utterances was proposed [76].

All of the behaviors listed here vary from person to person and therefore they are also related to the impression of the character. In this study, these behaviors are used for the character expression for spoken dialogue systems.

Chapter 3

Autonomous Android ERICA

In this thesis, an autonomous android ERICA is used as a platform of the spoken dialogue system. ERICA is very similar to a human and is suitable for expressing characters [77]. This chapter describes a structure of ERICA and a dialogue corpus collected using ERICA.

3.1 Introduction

An autonomous android ERICA is developed to make her behave like a human and naturally interact with humans. ERICA is modeled as a 23 year-old woman. Her design concept is to contain both the friendliness as an android and a sense of existence as a human being. The appearance of her face and body is artificially produced in reference to characteristics of young ladies. ERICA mounts 19 active joints inside to move her face, head, shoulder, and back. It is planned to install more motors to move her arms and legs in the future. Even now, the flexibility of her face has diversity (including eyebrow, eyelids, lip, eyeballs, and tongue), which enables her to show various facial expressions. ERICA is therefore able to generate not only verbal responses but also non-verbal behaviors such as facial expression, eye-gaze, and nodding, which are used to convey a variety of her emotions.

3.2 Automatic Speech Recognition (ASR)

The microphone array captures multi-channel audio signals and identifies which direction the audio signal comes from. The sound source localization is realized by the multiple signal classification (MUSIC) method [78] with the 16-channel microphone array. This result is matched with the position of the person tracked by the depth camera (Kinect v2). If they match, the input speech is enhanced by using the delay-and-sum beamforming.

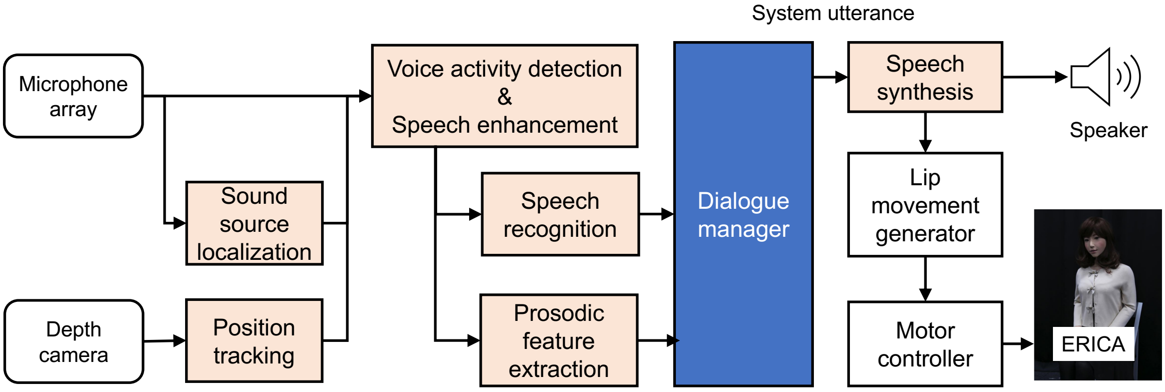


Fig. 3.1 Architecture of the spoken dialogue system on ERICA

Prosodic features such as fundamental frequency (F0) and power is calculated from the enhanced speech [79]. The automatic speech recognition (ASR) is the subword-unit-based neural network using the attention mechanism [80]. Distant speech recognition using the microphone array allows the user to initiate a dialogue without being aware of the microphone.

3.3 Text-To-Speech for ERICA

The speech of ERICA is generated by a text-to-speech engine that was developed for ERICA. It also contains many formulaic expressions, backchannels and fillers with a variety of prosodic patterns. The system allows control of speed, pitch, and volume for each word. At the same time, lip and head movements of ERICA are generated based on the prosodic features of the synthesized utterances [81, 82].

3.4 Backchannel Generation Model

Backchannel generation model determines the timing and form of backchannels from the user's utterances. A logistic regression model predicts if a backchannel would occur 500ms in the future per 100ms [83]. The input features are statistics such as mean, maximum, minimum, and range for the fundamental frequency (F0) and power. Thus, this model can continuously predict a backchannel without waiting for the end of the user utterance. The form of backchannels is selected based on the distribution of them used in the corpus described in Section 3.7. The frequency of backchannels can be adjusted by controlling the threshold for the output of this model.

3.5 Turn-taking Model

End-of-turn prediction is the problem of determining if the speaker has finished their turn. ERICA has two turn-taking models: time-out method model and prediction model. Time-out method sets a silence threshold and takes the system's turn when the length of silence from the end of the user's utterance exceeds the threshold.

The prediction model is based on Long short term memory (LSTM) model using acoustic and linguistic features on IPUs (inter-pausal units) [84]. An IPU is defined as a segment of speech which does not contain a pause greater than 200ms. Acoustic features are 40 log mel-filter bank features in 10 ms intervals. Linguistic features are word embeddings based on the transcript tokenized using Japanese tokenizer MeCab[1]. Word embeddings are conducted using Word2Vec [85] on the tokenizations with a dimension of 100. These features are extracted from the final 500ms of the IPU. These features are input into the LSTM model, which determines whether to take a turn on the IPU. When it is decided to take a turn, it is necessary to determine the swithing pause length. Real-time turn-taking is achieved by integrating a finite state model [86] that takes into account the waiting time. The prediction model enables the system to take a turn within approximately one second.

In this thesis, the time-based turn-taking model is used, since the prediction-based turn taking model is hard to control the pause length.

3.6 Dialogue Tasks

ERICA is given some social roles to realize natural and smooth dialogue with humans. Table 3.1 describes the characteristics of social roles. It is important for ERICA to perform a wide range of these roles. Attentive listening and job interview are the case where the role of listening is dominant. Laboratory guide is the opposite case where the role of speaking is dominant. Speed dating is the mixed case where the roles of listening and speaking are balanced. Dialogue data has also been collected where the operator controlled ERICA to talk with a human subject. The data are used to train dialogue models and to analyze the dialogue. The following subsections explains the system for each social role.

3.6.1 Attentive Listening

Attentive listening is to listen to the user talk carefully and give feedbacks in order to encourage the user to talk more. Attentive listening is needed for counseling systems [3]

[1] http://taku910.github.io/mecab/

Table 3.1 Daigloue task of ERICA

	Attentive listening	Job interview	Laboratory guide	Speed dating
Role	Listen	Ask	Introduce	Chat
Initiative	User	System	System	Both
Occupancy	User	System	System	Both
Content	Free	Fixed	Fixed	Free

and also casual conversation partner systems for elderly people. The attentive listening system generates both backchannels and variational listener's reactions [83]. Fig. 3.2 shows the diagram for the response generation. The following describes how each response is generated. Backchannels are generated based on acoustic features using the backchannel prediction model described in Section 3.4. A backup question is selected based on rules on the user silence. When a user is silent for a long time, the system asks a pre-prepared question to trigger a new topic. Assessments are based on the sentiment polarity (positive or negative) of the user utterance. Each word of the user utterance is compared with sentiment word dictionaries [2] [3]. If the user utterance containes positive words, the system would say a positive response such as "That is good. (いいですね．)" and "That is nice. (素敵ですね．)" On the other hand, if the user utterance containes negative words, the system would say a negative response such as "That is bad. (残念でしたね．)" and "That is hard. (大変ですね．)" Repeat is the response using a focus word extracted from the user utterance. This is expects to express understanding of the dialogue. The system extracts a focus word using the simple rule: a focus word is the latest noun or adjective in a user utterance. For example, if a user says "I ate a curry," the system response would be "A curry." If there are several continuous nouns, they are regarded as a compound word and are considered as the focus word. An elaborating question is generated using focus words to elicit more dialogue about the current topic. An elaborating question not only extends the dialogue but also expresses deeper understanding of the dialogue. The system generates a question by concatenating the focus word with interrogatives such as which, when, and what. In total, we use 11 types of interrogatives as candidates. For example, if a user says "I went to Tokyo by a train," the focus word would be "train" and the elaborating question would be "Which train?." To select the proper interrogative, the system refers to bi-gram probabilities. Generic response is a backup response used when no other response can be generated. Generic responses are "I see (そうですか)" or "I got it (なるほど)". These responses can be used for any dialogue context. The generic sentiment is another type of generic response according to the weak

[2] Japanese Dictionary of Appraisal —attitude— (JAppraisal Dictionary ver1.0): https://www.gsk.or.jp/catalog/gsk2011-c/

[3] SNOW D18: Japanese Emotional Expression Dictionary: https://www.jnlp.org/GengoHouse/snow/d18

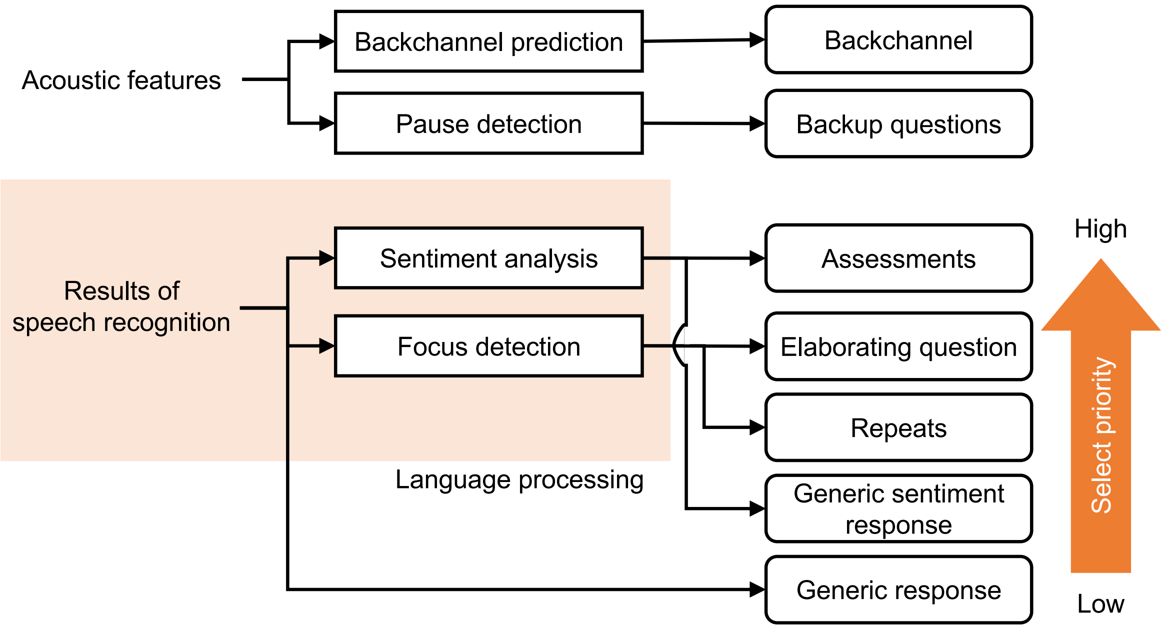


Fig. 3.2 Diagram for the response generation in the attentive listening system

sentiment of user utterances. For this response, a different sentiment dictionary [87] is used to recognized a sentiment of the user's utterance. If the user utterance is short, the system also uses a short generic response such "Yes (はい)" to avoid the barge-in.

Since each module generates each response independently, it is needed to select the suitable one. Backchannels are uttered during the user' s turn, so this module works independently from the others. Backup questions are triggered by a long pause so that this module is also independent. A priority of using the responses is shown in Fig. 3.2. The system will respond using the highest priority response which can be generated given the user' s utterance. The priority order is defined based on how likely it is to generate the response type.

3.6.2 Job Interview

The job interview dialogue is organized and system-initiative where the interviewer can control the dialogue flow. Therefore, the job interview dialogue can be implemented with a finite state transition. This system aims to dynamically generate elaborating questions according to interviewee (user) utterances. The system asks the interviewee questions on predetermined topics. In each topic, a general question (baseline question) and some elaborating questions are prepared. Depending on the user's response to the baseline questions, the system selects a elaborating question to ask. Additionally, if the user's response contains a focus word, the system asks about that word.

This study uses the system that asks prepared questions (Baseline question) as a job interview system. To reduce the influence of the dialogue content on the impression, the system asks the same content to all subjects.

3.6.3 Speed Dating

Speed dating is a dialogue between a male and a female who meet each other for the first time. They talk about topics such as their hobbies and profiles to identify if they will be able to keep the relationship in the future. Some studies were made on analyses in a speed dating dialogue corpus [88]. Subjects are expected to be more engaged in the speed dating dialogue than the case of normal chatting because there exists a clear purpose of the dialogue.

The range of conversational topics can be restricted to those that people use in the first-encounter dialogue. However, it is difficult to implement the speed dating dialogue system due to the mixed initiative where the initiative is frequently exchanged among them. A preliminary dialogue system for speed dating was implemented with the system. At first, conversational topics were defined. These topics are thought to be frequently used in the first-encounter dialogue. Then, a response table for the system is prepared to respond to user questions. The system response was prepared for each pair of topic and interrogative. For example, for the pair of hobby and what, the system response is "My hobby is reading books." A question sentence of the system was also prepared on each pair so that ERICA can make the question on the corresponding topic. Here, a dialogue system was implemented for ERICA making a question by taking the initiative when the user is silent for several seconds or that a configurable internal state of the system such as liking is higher. To handle the mix-initiative dialogue, a dialogue act classification was utilized in order to classify the user utterance into a question or others. If the user utterance contains a question, the system makes a question with the response table. Otherwise, the system generates a response towards the user statement, such as "That is nice." It is an issue of how to handle other topics where the system utterances are not prepared. If the system says the same backup utterance such as "Sorry, I do not know" too often, the user would be disengaged in the dialogue. Therefore, it is needed to generate other backup utterances that can keep user engagement.

3.6.4 Laboratory Guide

In the role of laboratory guide, the system holds the dialogue initiative and talks in most of the time. Some studies on information navigation have been made [1, 89]. The laboratory guide system is a rule-based dialogue system using finite state transitions. The predefined explanations are explained to the user. During the process, the user's level of understanding

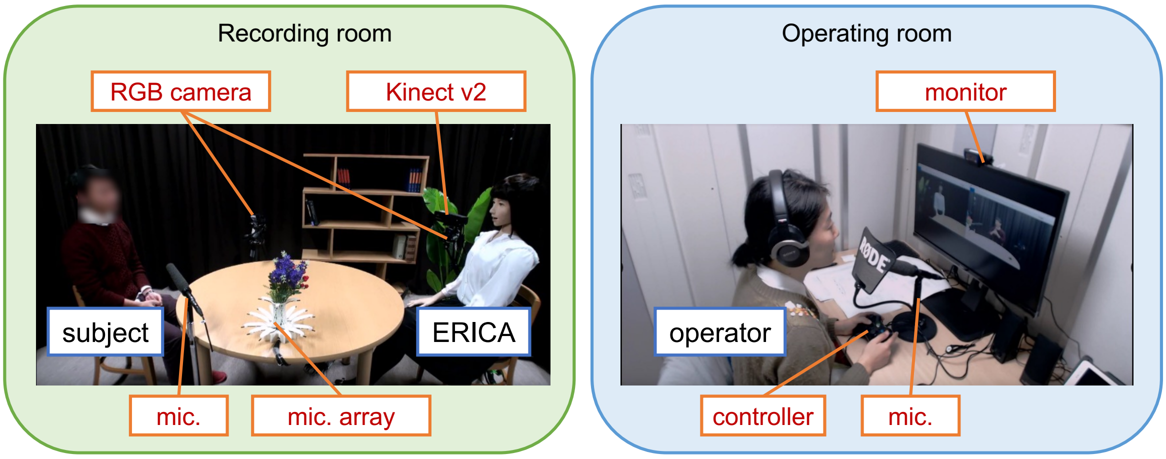


Fig. 3.3 Setting of Recording Dialogue

is checked by asking him or her questions. The system controls the content of the questions based on the user's engagement measured during the interaction [90]. This allows the user to feel that the system is explaining to him or her according to his or her level of understanding and interest.

3.7 Human-Robot Dialogue Corpus using Wizard of WOZ Method

A human-robot dialogue data have been collected in which the humanoid robot ERICA interacted with human subjects [91]. Fig. 3.3 shows a setting of the recording to dialogue. ERICA was controlled by an operator, who was in a remote room. The dialogue was one-on-one, and the subject and ERICA sat on chairs facing each other

The scenario of the dialogue are attentive listening, job interview and speed-dating according to the task of ERICA. Laboratory guide is not included in the corpus because this is one-sided dialogue, which the operator explain research topics for most of the dialogue. The participants are briefed on the dialogue in advance, and the content of the dialogue is made up in mind before the dialogue. The voice uttered by the operator was directly played with a speaker placed on ERICA in real time. When the operator spoke, the lip and head motions of ERICA were automatically generated from the prosodic information [81]. All participants were native Japanese speakers.

Chapter 4

Character Expression Model

4.1 Introduction

Humanoid robots, which naturally interact with people, have been studied and developed [15, 77, 16]. Humanoid robots have an appearance and behavior similar to those of human beings, and thus users are expected to feel a character of the robot in the dialogue. A humanoid robot is usually given a social role such as a lab guide or a counselor depending on the dialogue task. Expressing characters matching to such social roles would not only give a good impression to users but also have a good effect on the performance of the task. For example, it is easy to talk to a calm counselor.

There are several studies on character expressions for dialogue systems. PERSONAGE is a system that generates a response sentence that matches the designated character [24]. The response generation is based on the relationship between the Big Five parameters and the corresponding features of sentences. There are also some methods of expressing characters for spoken dialogue systems by linguistic patterns [92, 93]. In these studies, characters are represented by changing the style of sentences. On the other hand, in spoken dialogue, factors such as the way speaker talks also have an impact on the impression. There is a spoken dialogue system that adjusts its speaking rate to those of the user [94]. It is expected that people who talk frequently are extroverted, and those who use a lot of fillers are likely to give an impression that they are restless.

This chapter addresses a model that expresses characters by controlling behaviors of the robot in dialogue. Fig. 4.1 depicts the outline of this study. These behaviors can be controlled by adjusting parameters of the robot's utterances. To investigate the feasibility of expressing characters, we conduct an experiment and analyze the effect of controlling the dialogue behavior parameters on the impression of three character traits: extroversion,

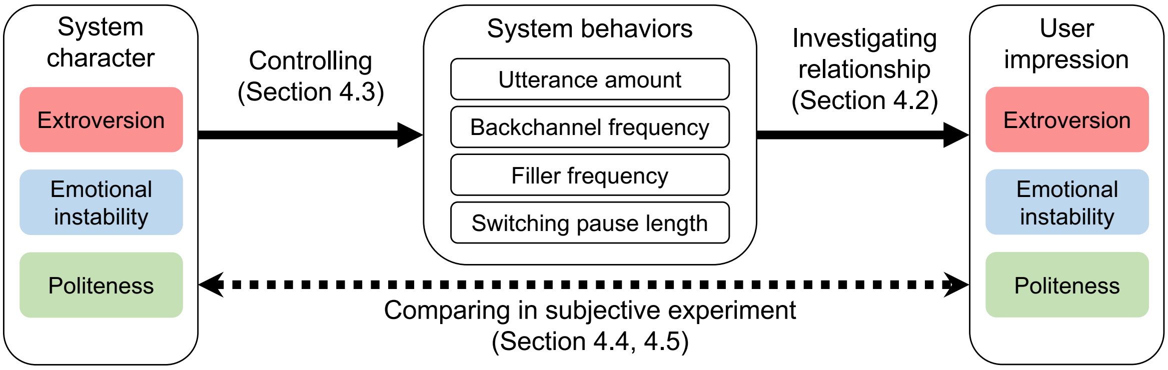


Fig. 4.1 Outline of this study

emotional instability, and politeness. Based on the analysis, we construct a model to control the dialogue behaviors based on a given character.

4.1.1 Character Expression

We define characters in consideration of the personality traits of psychology and the usage of robots in society. Then, we choose effective dialogue behaviors for the character expression.

4.1.2 Definition of Character

In psychology, several scales expressing human personality have been proposed. Among them, extroversion and emotional instability are most widely used for many character classifications. For example, Eysenck [95] expressed a personality in two dimensions using the traits of extroversion and neuroticism (emotional instability). The Big Five scale expresses a personality using five traits of extroversion, emotional instability, integrity, agreeableness, and openness [36, 39]. The presence of the top two factors of the Big Five has been confirmed [40, 96]. They are “Stability,” which ranks at the top of emotional instability, agreeableness, and conscientiousness, and “Plasticity,” which ranks at the top of extroversion and openness. It is agreed that a combination of extroversion and emotional instability can describe personality.

Personality traits in psychology are also used as characters in dialogue systems. Extroversion and emotional instability have also been used in character expression of agents in previous studies [97]. For example, an extrovert character may be preferred for a lab guide, and low emotional instability is required for a counselor. Politeness was also considered to control the impression of agents [98, 99]. Politeness is required to spoken dialogue systems to serve many social tasks such as a receptionist.

In this study, characters are defined using three traits of extroversion, emotional instability, and politeness. These three traits can be easily perceived [39] by many users and they are expected to contribute to task achievement in dialogue. Note that the correlation among these traits was partly observed [39]. We will be able to choose an appropriate character in these three traits according to the social role of the robot.

4.1.3 Dialogue Behaviors for Character Expression

Dialogue behaviors that can affect the impression of the speaker in dialogue are examined. The utterance amounts affect the impression in dialogue, for example, a person that speaks a lot seems more extroverted. Backchannels affect the impression of characters because they have various roles in dialogue such as expressing empathy and showing understanding. The frequency and type of backchannels have some effect on the impression of extroversion and emotional instability [97]. In this study, backchannel is defined as a short expression uttered by the system while the user is speaking. The backchannel frequency is defined as the number of backchannel by the robot. In this study, the candidate positions at which the robot can produce a backchannel are defined as the clause boundaries of the user's utterance. The backchannel variety is defined as the number of lexical types of the robot's backchannel in the dialogue. Fillers also have an influence on the impression of characters. Using a lot of fillers looks emotionally unstable, while the filler has been shown to be effective in improving impressions of delayed responses [69]. The filler frequency is defined as the number of filler by the robot. The switching pause, which is the length until the system responds in spoken dialogue, is an effective cue for users to recognize the system's character. It was investigated that the length of switching pause has an impact on impressions [100]. Based on the above, we adopt utterance amount, backchannel frequency, backchannel variety, filler frequency, and switching pause length as dialogue behavior features.

4.2 Analysis of Relationship between Dialogue Behavior and Character Impression

In order to construct a behavior control model, the effect of the dialogue behavior features on the character impression is investigated. In the experiments, the following hypotheses are made.

1. The extroversion is associated with the utterance amount and the backchannels frequency. A robot that talks a lot and generates many backchannels is perceived to be an extrovert.

2. The emotional instability is associated with the backchannel variety and the filler frequency. A robot that generates the same type of backchannels and many fillers is perceived to be emotionally unstable.

3. The politeness is associated with the switching pause length. A robot that waits for a long time before talking is perceived to be polite.

4.2.1 Speech Samples

Audio samples of 20 conditions are prepared for experiments. The utterances of the robot are generated using the text-to-speech software in these scenarios. The utterances of the user are made by the experimenter. These utterances are spoken in Japanese. Two scenarios are prepared with reference to a human-robot dialogue corpus recorded in a Wizard of Woz (WOZ) setting. The content of the dialogue is designed to be natural with any character and even when the dialogue behavior features (excluding latency) are adjusted.

A baseline dialogue of about one minute was created for each of the two scenarios. Backchannels and fillers appear moderately in the reference dialogue. The switching pause length is set to 0.5 seconds. In comparison dialogue used for each experimental condition, only one corresponding feature is adjusted from the baseline dialogue, and the remaining features are kept same as the baseline dialogue. In order to obtain clear results in the following experiments, we prepared the low (small, short) condition and the high (large, long) condition.

Prepared patterns of the dialogue behavior features are shown in Table 4.1. When the backchannel frequency is high, backchannels are inserted at all clause boundaries in the user utterances. When the backchannel frequency is low, all backchannels are removed. With regard to the backchannel variety, the backchannels are changed to many kinds of backchannels in a large condition, and all the backchannels are replaced with “yes" in a small condition. In the high filler frequency condition, fillers are inserted at all clause boundaries and at the sentence beginning of the utterances. In the low condition, all fillers are removed. When the switching pause is long, the turn-taking time is set to 3 seconds. When the switching pause is short, the start of the system utterance overlaps the end of the user utterance by 0.5 seconds. With regard to the utterance amount of the system, we extended the original two scenarios. Specifically, each scenario was concatenated with its following part of the dialogue in the corpus so that the length of utterances becomes twice or more. This is needed to make it easy for the subjects to recognize the difference of the utterance amount. One of the extended scenarios was system-dominant (large condition), and the other was user-dominant (small condition).

Table 4.1 Control of dialogue behavior features

Dialogue behavior features	Conditions	Details
Utterance amount	Large	Robot : 49.2 seconds, User : 25.3 seconds
	Small	Robot : 25.5 seconds, User : 38.8 seconds
Backchannel frequency	High	At all clause boundaries during user's utterance
	Low	Delete all robot's backchannels
Backchannel variety	Large	4 types
	Small	1 type
Filler frequency	High	At all clause boundaries and beginning of sentences
	Low	Delete all robot's fillers
Switching pause length	Long	3 seconds
	Short	0.5 seconds overlap

4.2.2 Experimental Procedure

Forty-six university students (28 men and 18 women, 18 - 23 years old) participated in the experiment. Each participant listened to the speech samples and responded to the questionnaire about an impression of the robot. We presented the baseline dialogue at first and the dialogue of different utterance amounts at last. The remaining conditions were randomly arranged for each experiment. Each participant listened 20 samples in total.

Participants answered questionnaires on a 7-point scale, from 1 (not at all) to 7 (completely), whether the item is true to the system. For a questionnaire of extroversion and emotional instability, we used short versions of the Big Five scales [58, 39], which is widely used in personality psychology in Japanese. It is also used for the impression evaluation of speeches [58]. Two items are added for politeness. The items we used are summarized in Table 4.2. In addition, the naturalness of the dialogue was also evaluated. Finally, we compute the average value for the corresponding items of each character trait.

4.2.3 Results

A variance analysis is performed among the three groups (Low (small, short) condition, Baseline, and High (large, long) condition) for each dialogue behavior feature. Character traits that showed significant differences ($p < .05$) between High and Low conditions by multiple comparisons in both scenarios are described below.

The analysis results for the extroversion trait are shown in Table 4.3. The higher the backchannel frequency, the lower the filler frequency, and the shorter the switching pause length are, the more the robot is considered to be extroverted. The analysis results on the emotional instability trait are shown in Table 4.4. The higher the filler frequency is and the longer the switching pause length is, the more emotionally unstable the robot is deemed. The

Table 4.2 The adjectives used in the character impression evaluation

Character traits	Items (English)	Items (Japanese)
Extroversion	Talkative	話し好き
	Quiet	無口な*
	Cheerful	陽気な
	Sociable	外向的な
Emotional instability	Melancholic	悩みがち
	Anxious	不安になりやすい
	Nervous	心配性
	Vulnerable	気苦労の多い
Politeness	Polite	丁寧な
	Gracious	礼儀正しい

(* invert scale)

analysis results on the politeness trait are shown in Table 4.5. The lower the backchannel frequency is and the longer switching pause length is, the politer the robot is deemed. The analysis results on the utterance amount are shown in Table 4.6. The robot is felt to be extroverted when the speaking time is longer, and introvert when the speaking time is shorter. No significant difference is found regarding the emotional instability. These results confirmed the hypotheses made in this chapter, except for the backchannel type.

4.2.4 Discussions

The evaluation score of naturalness except for the switching pause condition got equal to or higher than that of the baseline dialogue. Therefore, there was no problem of naturalness due to the adjustment of the features.

There were few traits where the backchannel variety had an effect. Since the number of backchannels is not so large in the one-minute dialogue, the participants may not have noticed the difference in the backchannel variety. On the other hand, the filler frequency had effects in many traits. In multiple comparisons, many character traits showed the tendency in the order of Low condition, Baseline condition, High condition. The switching pause length had a large effect on the impressions of all traits. The difference in the switching pause length is easily recognized. The utterance amount had an effect on extroversion and politeness.

The correlation coefficients between the scores of extroversion, emotional instability, and politeness were examined for all data, and a weak correlation was confirmed between extroversion and emotional instability. It is shown in Table.4.7. In addition, the results of a two-factor analysis of variance with dialogue conditions and character traits as factors showed that the F value was 64.126 and the p value was 2.01×10^{-18} among character

Variance analysis of scores based on subject evaluation

Table 4.3 Extroversion

	High		Baseline		Low			
Dialogue	(Large / Long)				(Small / Short)			Multiple comparison conditions[ii]
behavior[i]	Mean	SD	Mean	SD	Mean	SD	F-measure	
BF	5.40	1.00	4.38	1.16	4.40	1.02	35.82^{**}	**High**>**Low**, High>Baseline
BV	4.76	1.10	4.38	1.16	4.78	0.89	6.47^{**}	Large>Baseline, Small>Baseline
Fi	3.52	1.05	4.38	1.16	5.01	1.00	52.48^{**}	**Low**>**Baseline**>**High**
Sw	2.63	0.95	4.38	1.16	4.92	1.17	126.67^{**}	**Short**>**Baseline**>**Long**

$(^{*}p < .05, ^{**}p < .01)$

Table 4.4 Emotional instability

	High		Baseline		Low			
Dialogue	(Large / Long)				(Small / Short)			Multiple comparison conditions[ii]
behavior[i]	Mean	SD	Mean	SD	Mean	SD	F-measure	
BF	2.73	1.15	3.33	1.33	3.10	1.10	7.59^{**}	Baseline>High
BV	3.07	1.20	3.33	1.33	2.95	1.06	4.72	
Fi	3.88	1.58	3.33	1.33	2.79	1.17	20.59^{**}	**High**>**Baseline**>**Low**
Sw	3.90	1.48	3.33	1.33	2.77	1.14	22.17^{**}	**Long**>**Baseline**>**Short**

$(^{*}p < .05, ^{**}p < .01)$

Table 4.5 Politeness

	High		Baseline		Low			
Dialogue	(Large / Long)				(Small / Short)			Multiple comparison conditions[ii]
behavior[i]	Mean	SD	Mean	SD	Mean	SD	F-measure	
BF	4.24	1.58	4.43	1.71	5.03	1.03	9.16^{**}	**Low**>**High**, Low>Baseline
BV	5.05	1.08	4.43	1.71	5.17	1.04	11.32	Large>Baseline, Low>Baseline
Fi	4.66	1.29	4.43	1.71	4.98	1.14	5.15^{**}	Low>Baseline
Sw	4.51	1.16	4.43	1.71	2.94	1.38	46.33^{**}	**Long**>**Short**, Baseline>Short

$(^{*}p < .05, ^{**}p < .01)$

BF : Backchannel frequency,
BV : Backchannel variety, Fi: Filler frequency, Sw: Swtching pause length
i : Dialogue behavior features (**Bold letters**) showed significant difference.
ii : Multiple comparison conditions (**Bold letters**) showed significant difference ($p < 0.05$) between high and low conditions.

Table 4.6 T-test on utterance amount

Character traits	Large		Small		
	Mean	SD	Mean	SD	T ratio
Extroversion	5.74	0.82	5.03	0.84	4.99^{**}
Emotional instability	2.74	0.93	2.83	1.04	0.55
Politeness	4.76	1.22	5.91	1.04	5.69^{**}

$(^{*}p < .05, ^{**}p < .01)$

traits. There are significant differences among the three traits' rating points. Therefore, it is appropriate to control the three traits separately. However, this experiment does not assume

Table 4.7 Correlation coefficient between character traits

	Extroversion	Emotional instability
Emotional instability	-0.511	
Politeness	0.109	0.046

the independence of the character traits. These results indicate the feasibility of character expression by controlling utterance amount, backchannel frequency, fillers frequency, and switching pause length.

4.3 Character Expression Model

We construct a model that controls the dialogue behavior features based on a given character using a logistic regression model. Then, we conduct an evaluation experiment of expressing characters by the dialogue behavior control model.

4.3.1 Logistic regression model

The control model is constructed for each dialogue behavior feature. The backchannel variety is not used because it had little effect on any character traits in the experiment in Section 4.2. The model is given scores of the character traits in the 7-point scale and then outputs control values for the dialogue behaviors, modeled by logistic regressions as shown in Fig. 4.2. We used the data (features and traits) which showed a significant effect on the impression in Section 4.2 (bold items in Table 4.3, 4.4, 4.5, and 4.6) for model training. The total amount of the samples was 920. The Low condition of each dialogue behavior feature is labeled as 0 and the High condition is labeled as 1. The model learns the mapping from the evaluation score [1 - 7] of the character traits to this target label.

A cross-validation was conducted using three-quarters of each dataset for training and one-quarter for evaluation. The results of binary prediction in which the threshold is set to 0.5 are shown in Table 4.8. The results achieve F-measure scores of 0.73 to 0.91, which is reasonable for the behavior control model. The weights of the explanatory variables of the logistic regression are shown in Table 4.9.

4.3.2 Method for Controlling Dialogue Behaviors

The dialogue behavior feature is controlled using the output of the logistic regression model. To control the utterance amount, we prepare two utterance patterns: many utterances and

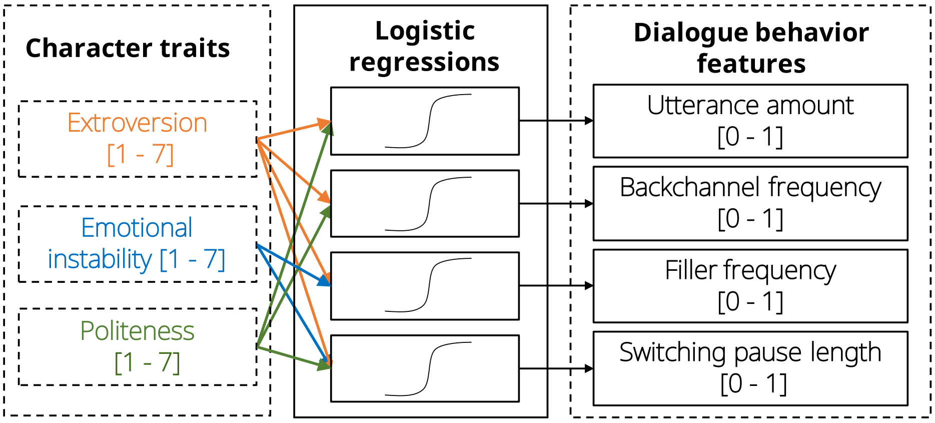


Fig. 4.2 Character expression model affecting spoken dialogue behaviors

Table 4.8 Prediction accuracy of control of each dialogue behavior feature

Dialogue behavior features	Used traits Ex	 Em	 Po	Class	Precision	Recall	F-measure
Utterance amount	✓		✓	Long	0.86	0.78	0.81
				Short	0.81	0.87	0.83
Backchannel frequency	✓		✓	High	0.74	0.73	0.73
				Low	0.75	0.74	0.74
Filler frequency	✓	✓		High	0.81	0.75	0.77
				Low	0.76	0.81	0.78
Switching pause length	✓	✓	✓	Long	0.91	0.90	0.91
				Short	0.91	0.91	0.91

Ex: Extroversion, Em: Emotional instability, Po: Politeness

fewer utterances. One of the utterance pattern is selected. Backchannels can be generated according to a statistical model [101–103]. In order to simplify the model, it is assumed that the probabilities of occurrence of a backchannel at all clause boundaries in the user's utterance are equal, and the output of the behavior control model is used for a threshold. A value sampled from a uniform distribution in the range of [0, 1] at each clause boundary is set as the occurrence probability of the backchannel, and when the probability exceeds the threshold, a backchannel is generated. Similarly, the filler frequency is controlled by adjusting the threshold of the statistical model according to the filler control model. The output of the switching pause length in the behavior control model [0, 1] is normalized in accordance with the range [-0.5, 3] of the experimental condition in Section 4.2. The value is set as the switching pause length. In the case of a negative value, it overlaps the end of the utterance of the user.

Table 4.9 Weights of explanatory variables in logistic regression models

Dialogue behaviors	Weights of explanatory variables		
	Extroversion	Emotinal instability	Politeness
Utterance amount	1.23	-	−1.16
Backchannel frequency	0.84	-	−0.62
Filler frequency	−0.78	0.30	-
Switching pause length	−1.56	0.26	0.98

Table 4.10 Pearson's product-moment correlation coefficient of a given character and the normalized impression evaluation point

Character traits	Correlation coefficient	T ratio
Extroversion	0.570	9.163**
Emotional instability	−0.004	−0.056
Politeness	0.235	3.185**

($^{*}p < .05$, $^{**}p < .01$)

Bold letters indicate the dialogue behavior features was correlated

4.4 Subjective Evaluation

We prepared dialogue samples by controlling the robot's behavior with the behavior control model. We used 16 kinds of characters based on the value of three character traits [1,7]. Eleven male university students evaluated the impression of the 16 dialogues where the generated behaviors are different. The questionnaire is same as that used in the experiment in Section 4.2. The rating points were normalized by the mean and the variance for each participant. The Pearson's product moment correlation coefficients between the ratings (Z-score) and the original values of the given character traits are shown in Table 4.10. From the results, significant correlations are confirmed with extroversion and politeness. This means that it is possible to express extroversion and politeness by using the model. There is no correlation for the emotional instability probably because the control parameter for the emotional instability is smaller than those of extroversion and politeness in the trained model.

4.5 Corpus-based Evaluation

Since the generated speech samples used in the impression evaluation were artificial data, we investigate the validity and generality of our character expression model by using a natural spoken dialogue corpus. The dialogue corpus consists of multiple dialogue tasks, and each task has corresponding suitable characters. The character expression model shown in Fig. 4.2 is applied in the backward direction (right-to-left) in order to calculate characters from spoken dialogue behaviors observed in the corpus. Finally, we examine the tendency of the identified characters for each dialogue task to confirm whether our character expression model can express characters that match each task.

4.5.1 Human-Robot Dialogue Corpus

We used a human-robot dialogue corpus where a human subject, called a subject hereafter, talked with the android ERICA [104, 91], which was remotely operated by another person, called an operator. In this corpus, three types of dialogue tasks are designed: speed dating, job interview, and attentive listening. The roles of ERICA (operators) in these tasks are a practice partner in a first-time-meeting conversation, a job interviewer, and an attentive listener to encourage a subject's talk, in the above order. In this study, we analyze ERICA's (operators') characters by our character expression model because the subjects were different people in each session so that it was difficult to reliably analyze their characteristic. The number of used dialogue sessions are 33 for speed dating, 31 for job interviews, and 19 for attentive listening. Each dialogue session lasted about 10 minutes. In each dialogue session, whereas the human subject was a different person, the operator was randomly assigned from four persons who are amateur actresses. Besides the transcription of the dialogue, we annotated the dialogue with the events and timing of backchannels [59], fillers [67], conversational turns, and dialogue act [105].

4.5.2 Analysis Method

We calculated the ERICA's (operators') characters from her spoken dialogue behaviors observed in the corpus. At first, we divided each dialogue session into two-minute segments in order to ensure a sufficient amount of data for this analysis. We also empirically confirmed that two minutes is enough duration to observe the spoken dialogue behaviors to calculate the character trait scores. For each segment sample, the corresponding character trait scores were calculated by using our character expression model as below. We first calculated feature values of the four spoken dialogue behaviors. Then, the feature values were converted to control amounts corresponding to the outputs of the logistic regression models. The amount of speech was classified into large or small (1 or 0) by using the median value of the entire corpus. The number of backchannels was normalized by the number of inter-pausal units (IPUs) [61], namely pause segmentation, of the interlocutor who is the current speaker. The number of fillers was normalized by the number of IPUs of herself. Switching pause length was linearly converted from the range of $[-0.5, 3]$ seconds to the range of $[0, 1]$. If the length was shorter than -0.5 or larger than 3 seconds, the converted value was clipped at 0 or 1, respectively. Meanwhile, we enter all possible combinations of character trait scores ($7^3 = 343$ ways) to our character expression model and then calculated the corresponding control amount of the spoken dialogue behaviors. Finally, we compared the control amounts observed from the corpus behaviors with those from each combination of character trait scores. We identified the corresponding character trait scores by the minimal Euclid distance between the control amounts.

4.5.3 Analysis Result among Dialogue Tasks

We analyzed the distribution of the estimated ERICA's (operators') characters for each dialogue task. Fig. 4.3 reports the distributions in the speed dating task. Our character expression model indicates that extroversion and politeness varied from middle to high and emotional instability was low (stable). In this dialogue task, the participants met each other for the first time, and the purpose of the dialogue is to build a relationship between them. Therefore, they should exhibit extrovert and polite behaviors. At the same time, they could show their own individual characters on their behaviors because this dialogue is not so much constrained by their participant roles. This is the reason why the distribution is varied for middle to high on extroversion and politeness.

Fig. 4.4 reports the distributions in the job interview task. Our character expression model also showed the similar tendency as in the speed dating task. Extroversion and politeness varied from middle to high. This variation can be interpreted by that the operators (interviewers) held the dialogue initiative in this dialogue so that there was more chance to control their behaviors expressing their characters. Compared to the speed dating task, extroversion relatively tended to be neutral. This can be interpreted by the style of this dialogue which is more formal. Thus, it is expected that extroversion was restricted by the style of this dialogue.

Fig. 4.5 reports the distributions in the attentive listening task. Our character expression model showed the biased distributions on extroversion and politeness. In this dialogue, the operators (attentive listeners) needed to encourage and elicit the subjects' talk. Therefore, they should behave as extrovert and polite. Moreover, the dialogue initiative was held by the subjects (story tellers) in this dialogue and the behaviors of the operators were constrained by the dialogue role (attentive listener). This is the reason why the distributions are more biased than those of other tasks.

In summary, it is shown that our character expression model can represent reasonable characters according to the scenario of each dialogue task. As a whole, the model shows that

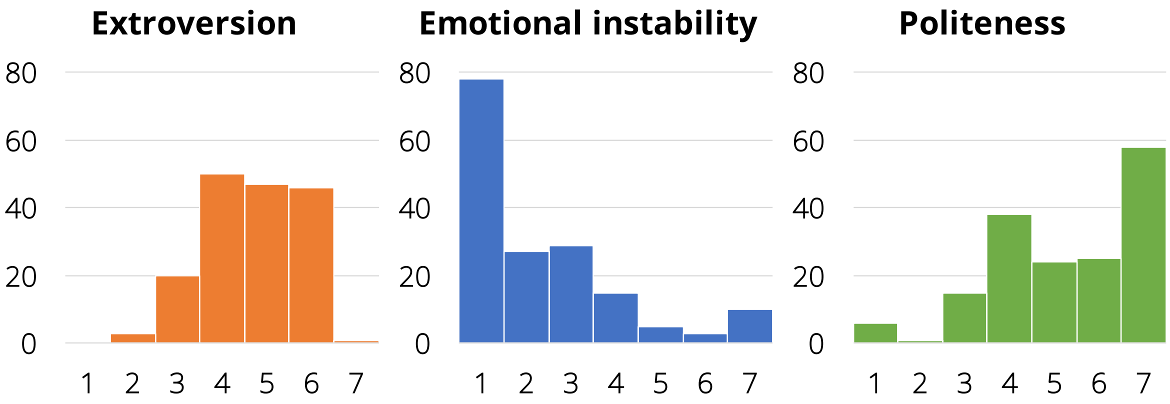


Fig. 4.3 Estimated character distributions in speed dating task

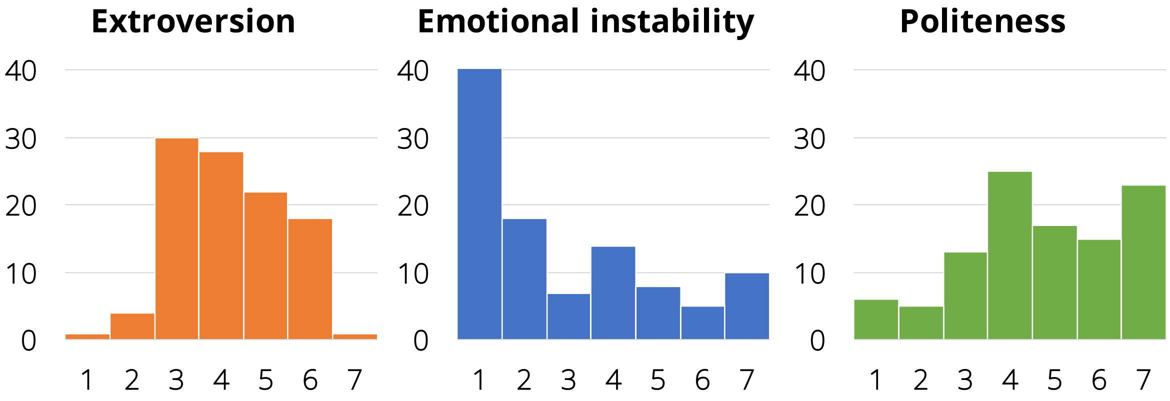


Fig. 4.4 Estimated character distributions in job interview task

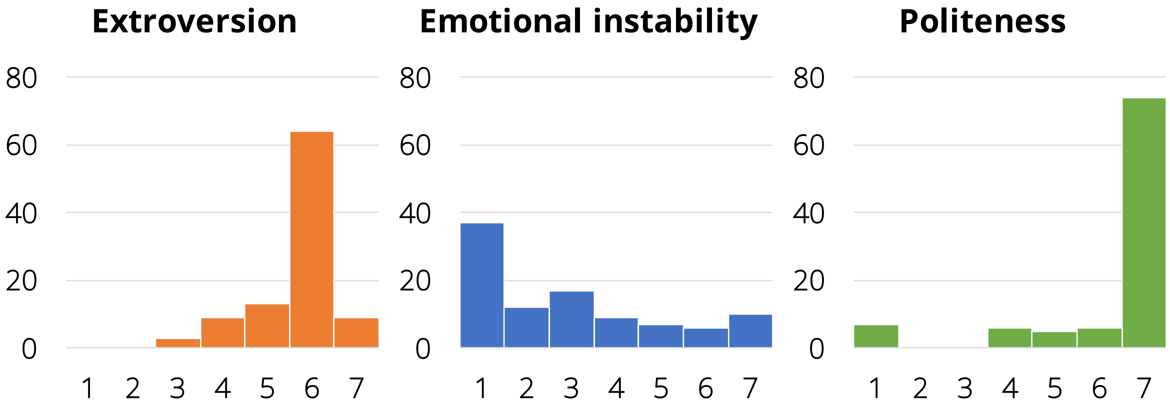


Fig. 4.5 Estimated character distributions in attentive listening task

extroversion and politeness tended to be middle or high and emotional instability was low (stable) in this corpus. These character traits are expected in dialogue where participants meet each other for the first time.

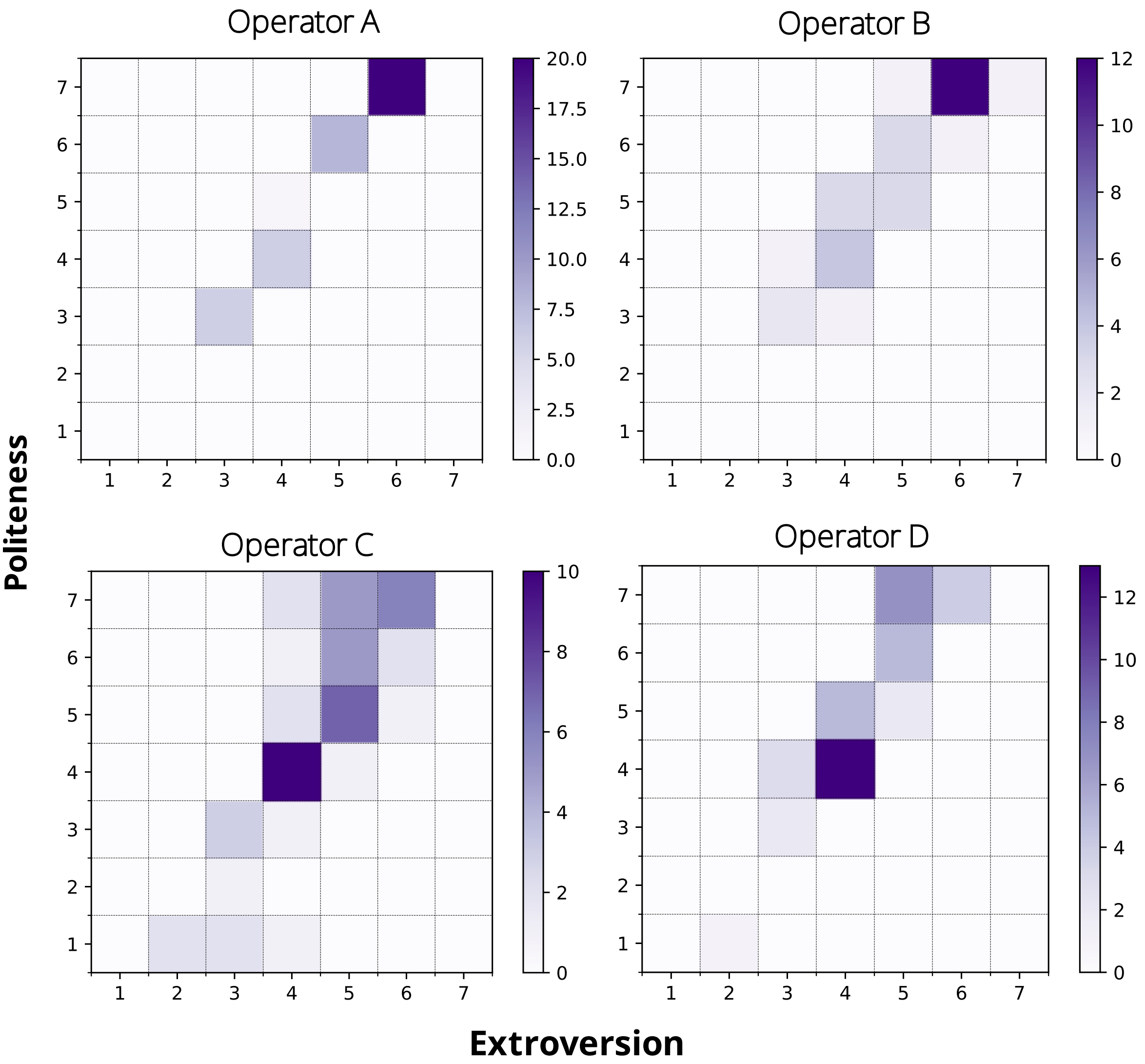


Fig. 4.6 Character distribution (extroversion and politeness) among operators in speed dating task

4.5.4 Analysis Result among Operators

Since there were four robot operators in this corpus, we further analyzed the distribution within each operator to find the individual difference among the operators in the same dialogue task. Since emotional instability was low (stable) among the entire corpus, we analyzed only extroversion and politeness in this section. We also analyzed only the speed dating task where the number of samples is the largest for each operator.

Fig. 4.6 shows a two-dimensional distribution of extroversion and politeness for each operator in the speed dating task. Our character expression model showed the different characters among the operators. The operators A and B showed high scores whereas the operators C and D showed the neutral scores. This result suggests that the operators could have their different characters in this dialogue task.

4.5.5 Subjective Evaluation

Finally, we conducted a subjective evaluation to confirm if third-party persons can perceive the calculated characters from the spoken dialogue behaviors in the corpus. We extracted 15 samples from the analysis (5 samples from each task). Note that these samples were balanced in terms of the variation of the calculated characters. The samples were taken from one operator (the operator B in Fig. 4.6) to avoid the effect of individual differences, but this difference should be examined in future work. We asked 13 subjects (3 females and 10 males, from 23 to 60 years old) to listen to each dialogue sample and then evaluate if they could agree with the calculated character. The calculated character was shown as a sentence such as “The robot was extrovert and little casual.” when the character scores were 7, 4, and 3 on extroversion, emotional instability, and politeness. Note that the adjective *little* was added to the sentence when the score was 3 or 5, and nothing was added for the scores of 1, 2, 6, and 7. If the score was 4 which was neutral, the corresponding character trait was not mentioned in the sentence. The subjects were asked to choose one from *Agree*, *Disagree*, and *Neither*. We regarded *Agree* as correct answers, but to avoid biased answering, character scores of three samples were flipped. In this case, the correct answer was *Disagree*.

Fig. 4.7 reports the ratio of correct answers. Among all samples except two samples 6 and 10, the majority of answers was the correct one. Even in the flipped character scores (sample 2 and 13), many subjects correctly answered as *Disagree*. When we regarded *Neither* as incorrect answers, the ratio of correct answers was 0.600 which was significantly higher than the chance level of 0.500 ($p = 0.004$ in the binomial test). When we did not use the *Neither* answers, the ratio became 0.770. This result demonstrates that characters represented by our model can be perceivable by humans with high accuracy.

4.6 Summary

This chapter addresses the character expression model affecting certain spoken dialogue behaviors: utterance amount, backchannel frequency, filler frequency, and switching pause length. This model are designed with the human-robot dialogue for several dialogue tasks. It is confirmed that the model represents reasonable characters for each dialogue task. Furthermore, it was also found that the model represents the individual difference of the

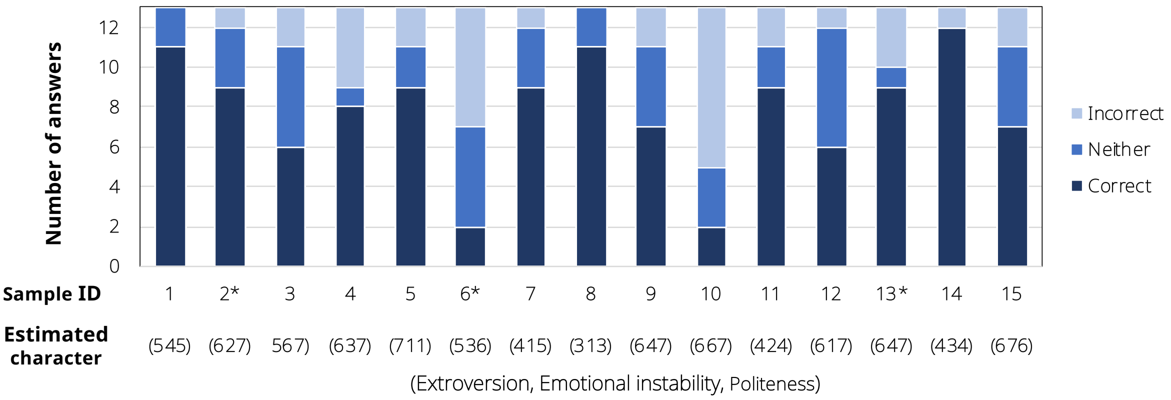


Fig. 4.7 Ratio of correct answers per sample in subjective experiment (* Character scores were flipped.)

robot operators in the same dialogue task. Finally, the subjective evaluation results showed that the subjects perceived the estimated characters from the dialogue behaviors.

Chapter 5

Semi-supervised Learning for Character Expression Model

5.1 Introduction

The character expression model that controls four spoken dialogue behaviors was proposed in Chapter 4. Pair data of the spoken dialogue behaviors and character scores are required to train the character expression model. However, this manual annotation is very costly, and thus the variety and amount of behavior patterns are limited. On the other hand, various dialogue corpora are available and they contain many dialogue behaviors, though the corpus data cannot be directly used for supervised learning of the character expression model as there are no manual annotations for the character. To make the character expression model more robust and accurate, it is important to exploit both manually annotated data (supervised) and dialogue corpus data (unsupervised).

To address this problem, semi-supervised learning based on a variational auto-encoder (VAE) is proposed to utilize not only manually annotated labels but also dialogue corpus data. The proposed method is designed to compensate for the data-sparseness with natural dialogue behavior data. Utilization of dialogue corpus data as unlabeled data can be applied to other expression tasks (e.g., emotion expression through dialogue behaviors) that are affected by data-sparseness due to a limited amount of training data.

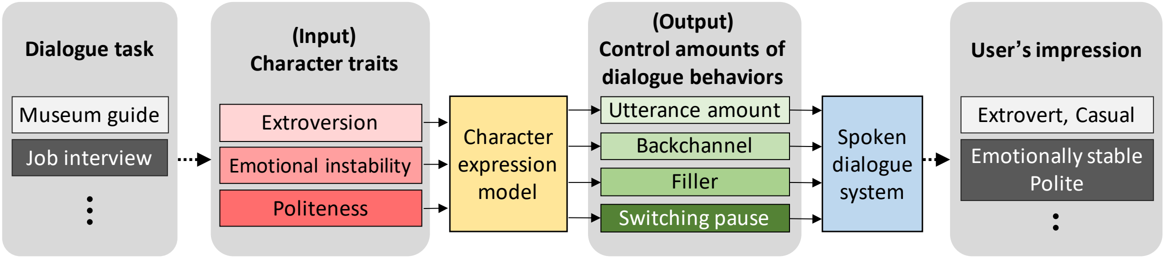


Fig. 5.1 Problem formulation of character expression

5.2 Character Traits and Spoken Dialogue Behaviors

This section addresses the problem formulation shown in Fig. 5.1. First, we define character traits used in this study as the input of the model. Then, we also describe the controlled spoken dialogue behaviors of the system.

5.2.1 Character Traits

The input of the character expression model is a set of three character traits: extroversion (extrovert vs. introvert), emotional instability (instable vs. stable), and politeness (polite vs. casual). Extroversion and emotional instability are selected from the Big Five scale [106, 107, 36]. In previous studies, the Big Five traits have been used to define the personality of dialogue systems [24, 97]. Extroversion is a major factor that determines the impression of system characters [108, 109], so we decided to include extroversion as a first priority character trait. Emotional instability is also included in our model explicitly since it would be fatal for the system in social scenarios if it is deemed as emotionally unstable. Extroversion and emotional instability have been frequently used in other psychological studies aimed at identifying personality by markers in language [110–113]. The other three traits from the Big Five are not included in this study in order to keep the model concise. Moreover, we have found a strong correlation between the other three traits and the used two traits in our preliminary study. On the other hand, politeness is adopted in the model so that the system can control its intimacy level towards dialogue partners [114]. For example, the system could behave politely in a formal situation while casually with intimate users.

5.2.2 Spoken Dialogue Behaviors

The output of the character expression model is a set of control amounts of spoken behaviors. We focus on spoken dialogue behaviors: utterance amount, backchannel frequency, filler frequency, and switching pause length. Previous studies suggested that these behaviors

	Control amount	
	0.0	1.0
Utterance amount	0% (system does not talk)	100% (user does not speak)
Backchannel	no backchannel	at all user pauses
Filler	no filler	at all system pauses
Switching pause	−0.5 sec. (overlap)	3.0 sec.

Table 5.1 Correspondence between control amount and actual behavior features

affected the impression of dialogue partners [115, 97, 69, 116–118]. Utterance amount means the ratio of utterance time between a system and a user. Backchannels are reactive tokens by listeners such as "*Yeah*" in English and "*Un*" in Japanese [60, 59]. In this study, the backchannel behavior is defined as the frequency of uttered backchannels. Fillers are short phrases filling the silence to hold (or take) the conversational floor such as "*Well*" in English and "*E-*" in Japanese [66, 67]. The filler behavior is defined as the frequency of uttered fillers. Note that the variety of forms of backchannels and fillers might also affect the impression of characters [97], but we focus on the frequencies in this study for the simplicity of the model. The switching pause length is defined as the time gap between the end of the preceding turn and the start of the following turn. Our character expression model controls these four spoken dialogue behaviors in accordance with the input of the three character traits. In Section 4.2, we artificially changed these behaviors and then investigated the effect on the impression of its character. As a result, strong relationships were confirmed between these four behaviors and the impression of the character.

Since the model output is assumed as a set of control amounts of behaviors that are normalized (e.g., from 0 to 1), their values need to be converted into actual behavior features (e.g., how many backchannels were uttered). The outputs of the proposed model are values from 0 (low or few) to 1 (high or many). The correspondence between the control amounts and the behavior feature values are summarized in Table 5.1. In this study, we use this correspondence when we create a dataset from an impression evaluation data and a dialogue corpus, which is described in the next section. We convert these control amounts to the behavior feature values by linear interpolation based on the correspondence. For example, when the control amount for backchannel is 1, the system generates backchannels at all user pauses.

5.3 Training Data

We prepared labeled and unlabeled data used for semi-supervised learning of the character expression model. The labeled data was obtained from an evaluation experiment on character impression (manual annotation) as explained in Section 4.2, and the unlabeled data was derived from a human-robot dialogue corpus as explained in Section 4.5.

5.4 Character Expression Model

A character expression model with semi-supervised learning that uses both the impression evaluation data and the corpus data is proposed. The impression evaluation data is used for supervised learning to learn the relationship between the character traits and the spoken dialogue behaviors. The dialogue corpus data is used for unsupervised learning to control the spoken dialogue behaviors naturally.

5.4.1 Network Architecture

First, we explain the architecture of the proposed model as depicted in Fig. 5.2. The model is based on a variational auto-encoder (VAE) [119] consisting of an encoder and a decoder. The encoder corresponds to a character *recognition* model, which converts the spoken dialogue behaviors to the character traits. The decoder corresponds to a character *expression* model, which controls the spoken dialogue behaviors based on the character traits. The input for the encoder is represented as a four-dimensional vector of the spoken dialogue behaviors normalized between 0 and 1. The encoder outputs a three-dimensional vector of the character traits normalized between 0 and 1 and also outputs parameters means ($\boldsymbol{\mu}$) and variances ($\boldsymbol{\sigma}$) to generate eight-dimensional latent variables (z). The latent variables are intended to capture factors other than the three character traits (e.g., dialogue task and context). The latent variable was used to suppress the overfitting. We tried several numbers for this dimension, selecting from $(2, 4, 8, 16, 32, \cdots)$. Finally, in the experiment of this paper, it is set to eight because the learning became stable when it is larger than eight, which suggests that the eight dimensions is sufficient for the current task. The input for the decoder is the three-dimensional vector of the character traits concatenated with the eight-dimensional latent variables. The decoder outputs the four-dimensional control amount of the spoken behaviors. The number of hidden layers is 3 for both the encoder and the decoder. The sigmoid function is applied to the output layer as the activation function.

The main task of this study is character expression corresponding to the decoder. When we train this VAE-based model, supervised and unsupervised learning are applied in a mixed

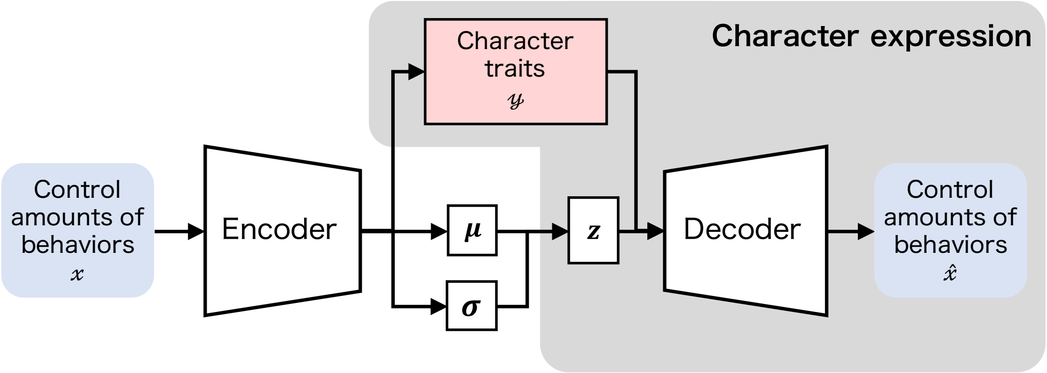


Fig. 5.2 Network architecture of the proposed model

manner in each training epoch, as depicted in Fig. 5.3. Each batch data $\mathscr{D}$ is mixed by labeled data $\mathscr{D}_l$ and unlabeled data $\mathscr{D}_u$. The labeled data $\mathscr{D}_l$ are the impression evaluation data explained and the unlabeled data $\mathscr{D}_u$ are the spoken behavior data from the dialogue corpus that does not contain any character trait data.

The spoken dialogue behavior values $\boldsymbol{x}$ and the character trait score $\boldsymbol{y}$ in the labeled data $\mathscr{D}_l$ are used to compute the encoder and decoder losses. The encoder loss is defined as

$$\mathscr{L}_{enc} = \underset{(\boldsymbol{x},\boldsymbol{y})\in\mathscr{D}_l}{\mathrm{CE}} (\mathrm{Enc}(\boldsymbol{x}),\boldsymbol{y}) \ , \tag{5.1}$$

which is the cross entropy (CE) between the outputs of the encoder $\mathrm{Enc}(\boldsymbol{x})$ and the oracle character traits $\boldsymbol{y}$ in the impression evaluation data. The decoder loss is computed using $\mathscr{D}_l$ as

$$\mathscr{L}_{dec} = \underset{(\boldsymbol{x},\boldsymbol{y})\in\mathscr{D}_l}{\mathrm{CE}} (\mathrm{Dec}(\boldsymbol{y},\boldsymbol{z}),\boldsymbol{x}) \ , \tag{5.2}$$

which is the cross entropy between the outputs of the decoder $\mathrm{Dec}(\boldsymbol{y},\boldsymbol{z})$ $(=\hat{x})$ and behaviors $\boldsymbol{x}$ in the impression evaluation data. Note that $\boldsymbol{z}$ is the latent variables following the eight-dimensional standard normal distribution $\mathscr{N}(\boldsymbol{O},\boldsymbol{I})$. When we use the decoder part (character expression) only, the latent variables $\mathbf{z}$ are randomly sampled from the standard normal distribution.

In the unsupervised learning using the unlabeled data $\mathscr{D}_u$, we use only the spoken dialogue behavior data $\boldsymbol{x}$. The whole network is fine-tuned based on the reconstruction error defined as

$$\mathscr{L}_{rec} = \underset{\boldsymbol{x}\in\mathscr{D}_u}{\mathrm{CE}} (\mathrm{Dec}(\mathrm{Enc}(\boldsymbol{x})),\boldsymbol{x}) - \mathrm{D}_{KL}[z \parallel \mathscr{N}(\boldsymbol{O},\boldsymbol{I})] \ , \tag{5.3}$$

where D_{KL} represents Kullback-Leibler divergence.

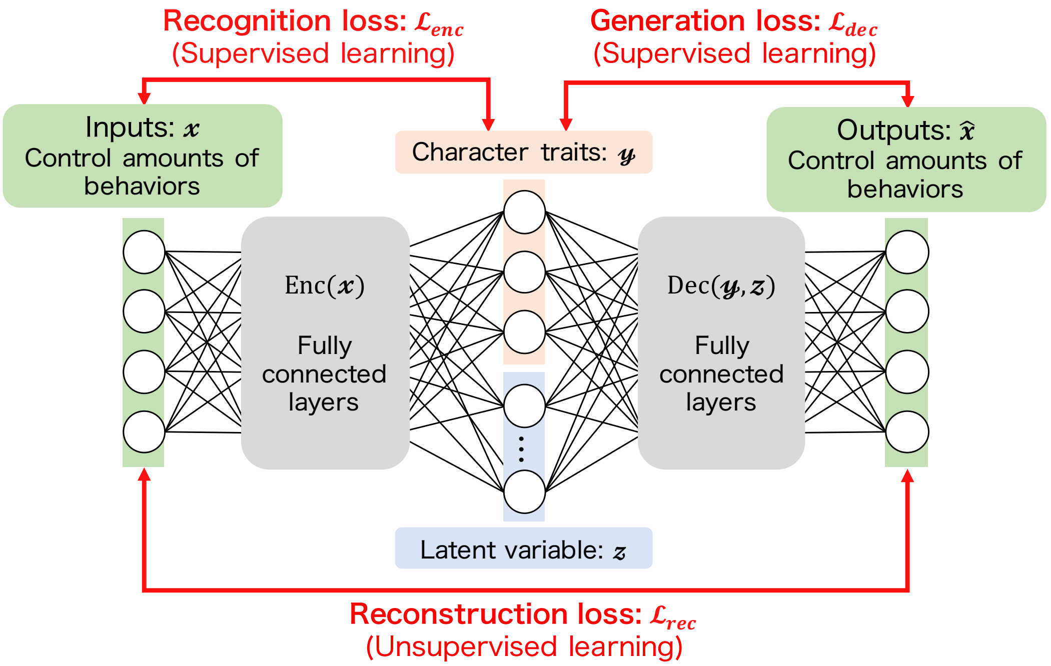

Recognition loss: $\mathcal{L}_{enc}$
(Supervised learning)
Generation loss: $\mathcal{L}_{dec}$
(Supervised learning)
Inputs: $\boldsymbol{x}$
Control amounts of behaviors
Character traits: $\boldsymbol{y}$
Outputs: $\hat{\boldsymbol{x}}$
Control amounts of behaviors
Enc($\boldsymbol{x}$)
Fully connected layers
Dec($\boldsymbol{y}, \boldsymbol{z}$)
Fully connected layers
Latent variable: $\boldsymbol{z}$
Reconstruction loss: $\mathcal{L}_{rec}$
(Unsupervised learning)

Fig. 5.3 Semi-supervised learning with character impression data and corpus data

The network parameters are optimized by using Adam to minimize the total loss defined as [1]

$$\mathscr{L} = (1-\lambda)(\mathscr{L}_{enc} + \mathscr{L}_{dec}) + \lambda \mathscr{L}_{rec} \,. \tag{5.4}$$

The weight of the loss functions (λ) adjusts the balance between the labeled and unlabeled data in the learning process. We set this weight to 0.8 to make the training process more focused on the unlabeled data. This parameter value was derived so that the training process was stable in a preliminary experiment.

5.4.2 Model Extension: Controlling Unlabeled Behavior

Another advantage of the proposed model is that it can handle unlabeled behaviors owing to the unsupervised learning. For example, we can train the mapping from the character traits to a new behavior, such as speech rate. The new behavior is not annotated with the character impression, but it could be observed in the dialogue corpus. With only a simple modification of the loss function, the proposed model can control the unlabeled behavior based on the input character traits.

To this end, the behavior data is extended to a five-dimensional vector: four dimensions for the existing spoken dialogue behaviors and one for the new behavior. In supervised

[1] In the current model, both labeled and unlabeled data are included in the same training minibatch, so the three loss functions are simultaneously optimized.

Table 5.2 Mean absolute errors (MAE) between control amounts of behavior output from the models and the oracle data (*behavior diff.* represents the difference in the level of actual behavior features in 2 min. segments that are calculated on basis of Table 5.1.)

Behavior	Reference	Baseline	Proposed	(behavior diff.)
Utterance amount	0.129	0.221	0.128*	11.16 sec.
Backchannel	0.199	0.243	0.199	2.16 times
Filler	0.113	0.326	0.113*	10.44 times
Switching pause	0.078	0.234	0.077*	0.55 sec.
Average	0.130	0.256	0.129*	

(*$p < .01$)

learning, the loss functions of the encoder and decoder cannot include those for the fifth behavior. On the other hand, in unsupervised learning, the error of the fifth behavior can be added to the loss function of reconstruction because it requires only the spoken dialogue behavior data. Therefore, the new behavior is considered in only unsupervised learning, but it is expected that the mapping between the character traits and the new behavior is learned by referring to the relationship between the four behaviors and the fifth behavior in the corpus. In other words, the fifth behavior is controlled in conjunction with the four behaviors. Since no labeled data are available for the speech rate, the supervised training process would be unstable. Therefore, the weight parameter λ (the importance of the unlabeled data) was set to 0.1 to focus on the loss of the supervised data.

5.5 Model Evaluation

We evaluated the effectiveness of the semi-supervised learning. At first, the decoder of the proposed model is evaluated to confirm that using the corpus data leads to natural behavior expression. Second, the encoder of the proposed model is also evaluated to examine whether the characters expressed by the proposed model capture the differences in each task. Finally, we investigate whether the proposed model can adequately express a new unlabeled behavior.

5.5.1 Effectiveness of Semi-supervised Learning

The proposed model was compared with a baseline model consisting of only the decoder of the proposed model, except that the latent variables were not used. The baseline model was trained with only the impression evaluation data as supervised learning. Therefore, this comparison reveals the effectiveness of semi-supervised learning.

To prepare test data, we conducted an additional impression evaluation by annotating character labels to a subset of the corpus data described in Section 4.5. First, we selected 30 audio samples from the corpus data. Note that these selected samples were not used in the model training. The data contains 10 samples for each task: attentive listening, job interview and speed dating. We asked five subjects (two females and three males) to listen to the audio samples and then evaluate the character traits of the operator. The question items were the same as for the impression evaluation explained in Section 4.5. We averaged the scores among the subjects as input character traits. The spoken dialogue behavior data was used as the oracle output. The evaluation metric was the mean absolute error (MAE) between the output of each model and the oracle data. When we calculate the control amounts using the decoder model, we input the concatenation of the character traits (3 dimensions) and latent variables (8 dimensions) sampled from a standard normal distribution to the decoder.

Table 5.2 reports the MAE of the models for each behavior and their averages. The reference is the model that always outputs the mean values of behaviors in the training data. In the evaluation data, the standard deviations of the character evaluation scores were 0.98, 0.61, and 0.39 for extroversion, emotional instability, and politeness, respectively. Thus, the variances of the input data (character traits) are so small that output (behaviors) of the proposed model is mostly similar to the average. We conducted a one-sided *t*-test between the proposed model and the baseline model. It is observed that the proposed model improved in all behaviors and significant differences were confirmed except for the backchannels. We also investigated the difference in actual behavior features (in 2 min. segments), which was calculated on the basis of Table 5.1. The proposed model controlled more accurately than the baseline by 11.16 seconds (in 2 min.), 10.44 times (in 2 min.), and 0.55 seconds for utterance amount, number of fillers, and switching pause length, respectively.

We also investigated the effectiveness of using the corpus data in semi-supervised learning by ablation study. First, we divided the labeled data of the training data into 10 parts. The amount of labeled data used was varied, as shown in Fig. 5.4. This suggests that the proposed model interpolated the sparse distribution of the labeled data by utilizing the unlabeled data. Second, the amount of unlabeled data used was varied, as shown in Fig. 5.5. It was shown that the addition of the behavior data was effective even when only a small amount of data is used.

5.5.2 Evaluation of Encoder

We also evaluate the encoder of the proposed model. In contrast to the decoder, the encoder of the proposed model can be regarded as the character recognition model. If the learning is successful, the encoder should estimate the appropriate character for an input spoken

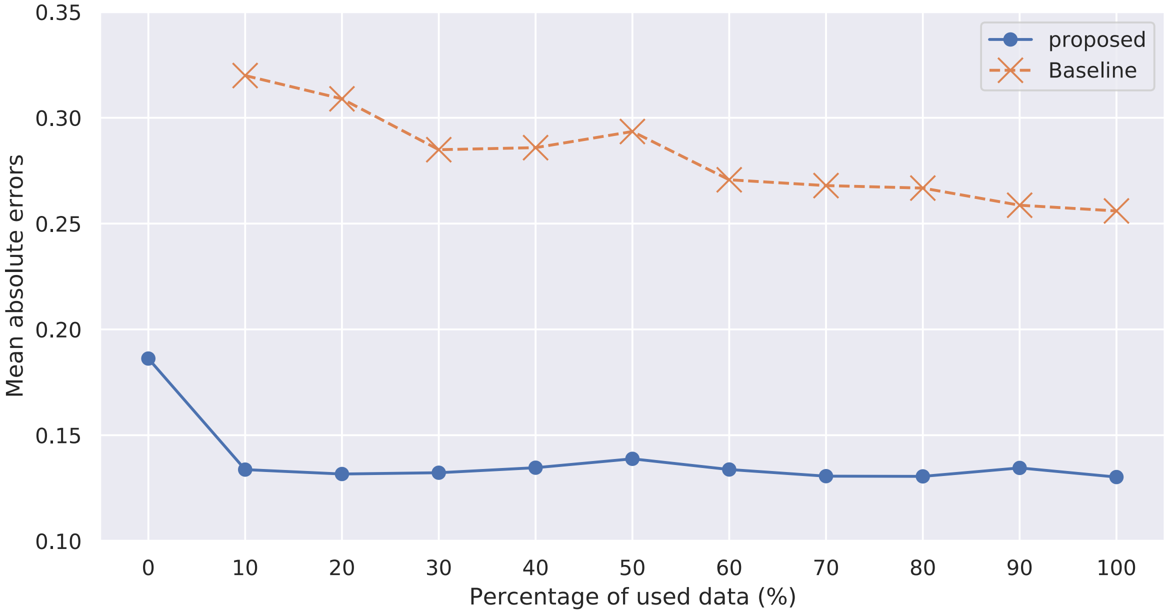


Fig. 5.4 Mean absolute errors (average of four behaviors) when the amount of labeled data used is varied

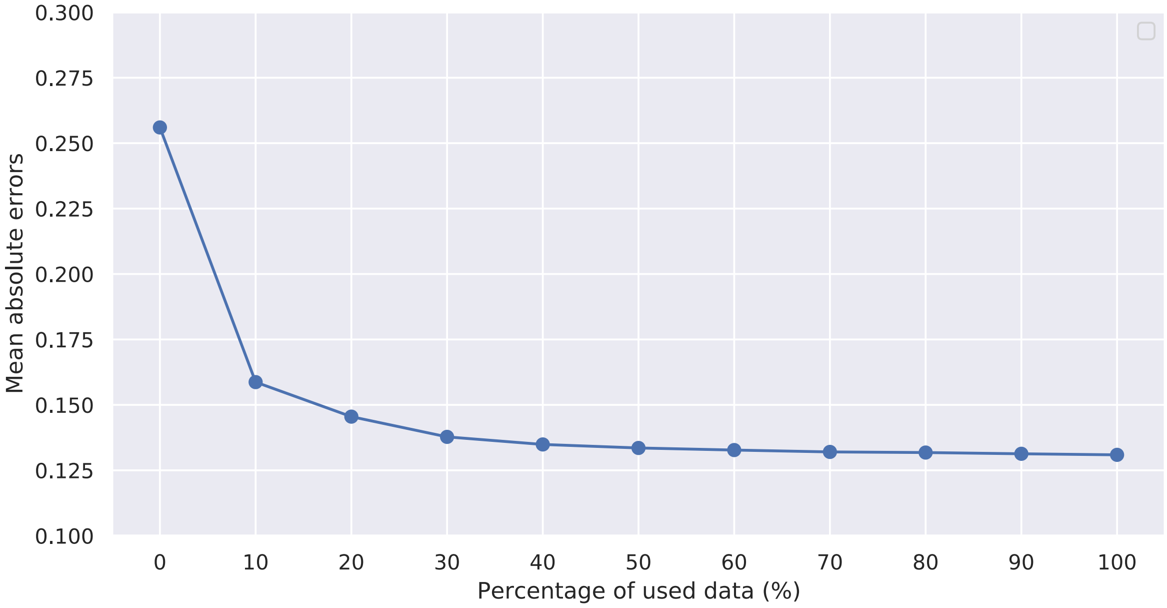


Fig. 5.5 Mean absolute errors (average of four behaviors) when the amount of unlabeled data used is varied

dialogue behavior. At first, we measured the recognition error by using the same additionally labeled data created in the previous section. We use the annotated data for evaluation of the encoder of the proposed model. In this experiment, the evaluation metric is the MAE between the output from the encoder and the baseline model. The baseline model is the encoder

Table 5.3 Mean absolute errors (MAE) between the character trait outputs from the encoder and the oracle data

Character traits	Baseline	Proposed
Extroversion	0.207	0.155 *
Emotional instability	0.128	0.183
Politeness	0.091	0.078
Average	0.142	0.138

(${}^{*}p < .01$)

Table 5.4 Example of control amounts with the proposed model when an additional unlabeled behavior (speech rate) was added (emotional instability was fixed as neutral.)

Character traits (Input)				Speech rate	
Extroversion		Politeness		(char./sec)	
0	(introvert)	0	(casual)	0.195	(5.35)
0	(introvert)	1	(polite)	0.062	(4.43)
1	(extrovert)	0	(casual)	0.526	(7.65)
1	(extrovert)	1	(polite)	0.449	(7.11)
0.5	(neutral)	0.5	(neutral)	0.339	(6.20)

model trained using only the labeled data. Table 5.3 reports the MAE for each character trait and their averages. We conducted a one-sided *t*-test between the proposed model and the baseline model. It was observed that the proposed model improved the scores for extrovert and politeness. However, in contrast to the result on the decoder, the improvement by the proposed method was limited. This may be because the reconstruction loss is calculated using the behaviors and propagates to the encoder through three character traits and 8-dimensional latent variables. The encoder learns the relationship between the four behaviors and the three character traits by the recognition loss in the supervised learning phase. Therefore, the effect of semi-supervised learning is small in the encoder evaluation.

5.5.3 Modeling of Unlabeled Behaviors

Finally, we evaluated an extension of the model by adding an unlabeled behavior as explained in Section 5.4.2. We used *speech rate* as an unlabeled behavior. Previous studies pointed out that speech rate behaviors affected the impression of extroversion [120, 121]. In this experiment, speech rate was calculated by dividing the total number of spoken characters by the total duration of the operator utterances. The calculated value was then converted to the control amount (from 0.0 to 1.0) by linear interpolation between 4.00 (min. in the corpus) and 10.94 (max. in the corpus).

We qualitatively analyzed the outputs of the model. Note that the baseline model cannot be applied to the current evaluation because this model extension is only added in an unsupervised manner. Table 5.4 reports the model outputs on the representative patterns of the character traits. The character trait patterns were combinations of extrovert/introvert and polite/casual. Emotional instability was fixed as neutral (0.5). It is observed that the more extroverted the system was, the faster it spoke. This mapping is intuitive for speech rate behaviors, which suggests that the proposed model is able to obtain the intuitive mapping for unseen (unlabeled) behaviors by referring to the relationship between unseen additional behaviors and the labeled existing behaviors.

We conducted a subjective experiment to evaluate the effect of learning speech rate. Five subjects evaluated their character impressions of the robot using the scale described in Section 4.5. Ten dialogues were sampled from speed dating task. As a result, the correlation coefficients between the speech rates and character traits are 0.26, 0.14, and -0.10 for extroversion, emotional instability, and politeness, respectively. Thus, it is shown that extroversion can be controlled by the speech rate.

5.6 Summary

The character expression model for spoken dialogue systems is enhanced by the semi-supervised learning. The model maps from the input three character traits to the output control amounts of the four spoken dialogue behaviors. The proposed character expression model is based on a VAE to utilize not only the impression evaluation data (labeled) but also the corpus data that does not contain any character labels (unlabeled). This approach allows the model to compensate for natural behavior patterns that are lacking in the impression evaluation data with natural combinations of the behaviors observed in the corpus. The experimental results showed that the proposed model expressed the target character traits through the behaviors more precisely than the baseline supervised learning.

In the proposed model, the decoder is used as a character expression model, and the encoder can be interpreted as a character recognition model, so we confirmed the encoder's ability to recognize characters. The analysis results suggested the possibility of using the encoder as a character recognition model to estimate the character that captures the task-specific tendency in the dialogue corpus.

Moreover, we also investigated the modeling of unlabeled behavior (speech rate) realized by semi-supervised learning. We confirmed that the proposed model acquired an intuitive mapping from the character traits to speech rate. This means that even if we do not have any character labels for additional behaviors, the proposed model can learn an interpretable mapping on the basis of the relationship between the additional behaviors and the existing behaviors.

Chapter 6

Task adaptation: Appropriate Character Expression for Task

6.1 Introduction

Task adaptation of spoken dialogue systems is to make systems generate behaviors appropriate in the task scenario, which leads to increasing user's satisfaction with dialogue. For tasks with definite objectives, the character required to the system should definite. In order to realize natural dialogue in these social scenarios, it is important to assign proper characters to the spoken dialogue systems.

An overview of character adaptation is shown in Fig. 6.1. It was shown by corpus analysis that the behavior and distribution of characters among each dialogue tasks. Additional subject experiments are conducted to see if the evaluation of the dialogue task can be improved by expressing appropriate characters for each task.

6.2 Corpus analysis

6.2.1 Analysis of Behavioral Tendency in Accordance with Dialogue task

We first investigate how the spoken dialogue behaviors change in accordance with the different dialogue tasks in the corpus (Section 4.5). The three dialogue tasks in the corpus are job interview, attentive listening, and speed dating, and in each task, the content of the conversation between the operator and the subject is different. For each two-minute segment,

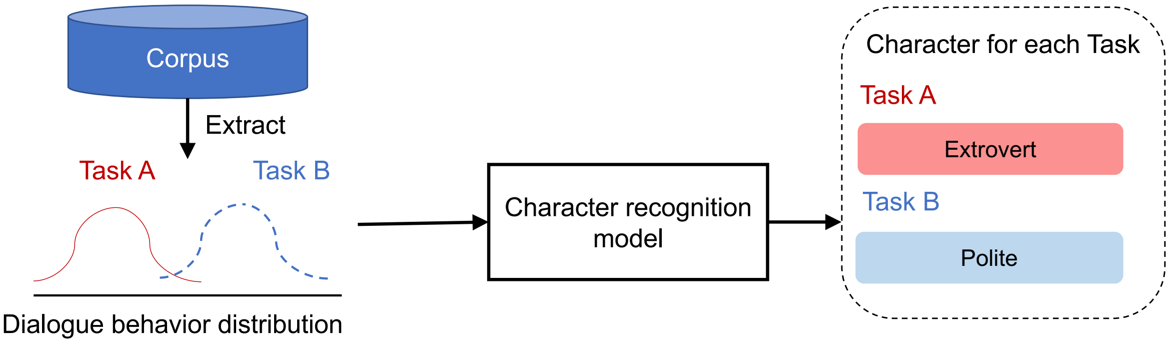


Fig. 6.1 Overview of task adaptation

the operator's spoken dialogue behaviors are normalized from 0.0 to 1.0. Note that the switching pause length was calculated as the mean value in the segment.

The distributions of the spoken dialogue behaviors in each dialogue task are reported in Fig. 6.2 (1)–(3). In the task of attentive listening (Fig. 6.2 (1)), the amount of speech and filler was smaller, backchannel was more frequent, and switching pause length was shorter. These behaviors may reflect the role of the operator as a listener, listening to the interlocutor. In job interviews (Fig. 6.2 (2)), a lot of fillers are observed, and the switching pause length is longer. This tendency is due to the fact that an interview is formal and tense dialogue. In the speed dating task (Fig. 6.2 (3)), the operators spoke many utterances with shorter switching pauses. Speed dating is a free and casual dialogue setting compared with the other tasks. These results suggest reasonable differences in the behaviors depending on the characteristics of each task. Therefore, it is important to control the spoken dialogue behaviors by the character expression model in accordance with the dialogue task.

6.2.2 Analysis using Character Recognition Model

The outputs of the character recognition model encoder (Section 5.4) which shows the characters in each dialogue task. If the distribution of the recognized characters captures the characteristics of each dialogue task, we can conclude that the proposed model can recognize the proper characters required for the specific dialogue tasks. Histograms of recognized each character traits are shown in Fig. 6.3 (1)–(3). In attentive listening, more introvert and casual characters were observed. The purpose of this task is to encourage the interlocutors to speak, so the introvert and casual character is reasonable. In job interview, more polite characters are observed, which is also reasonable because this task should be formal compared to the other tasks. In speed dating, there are more extrovert and casual characters, as the purpose of this task is to get to know each other.

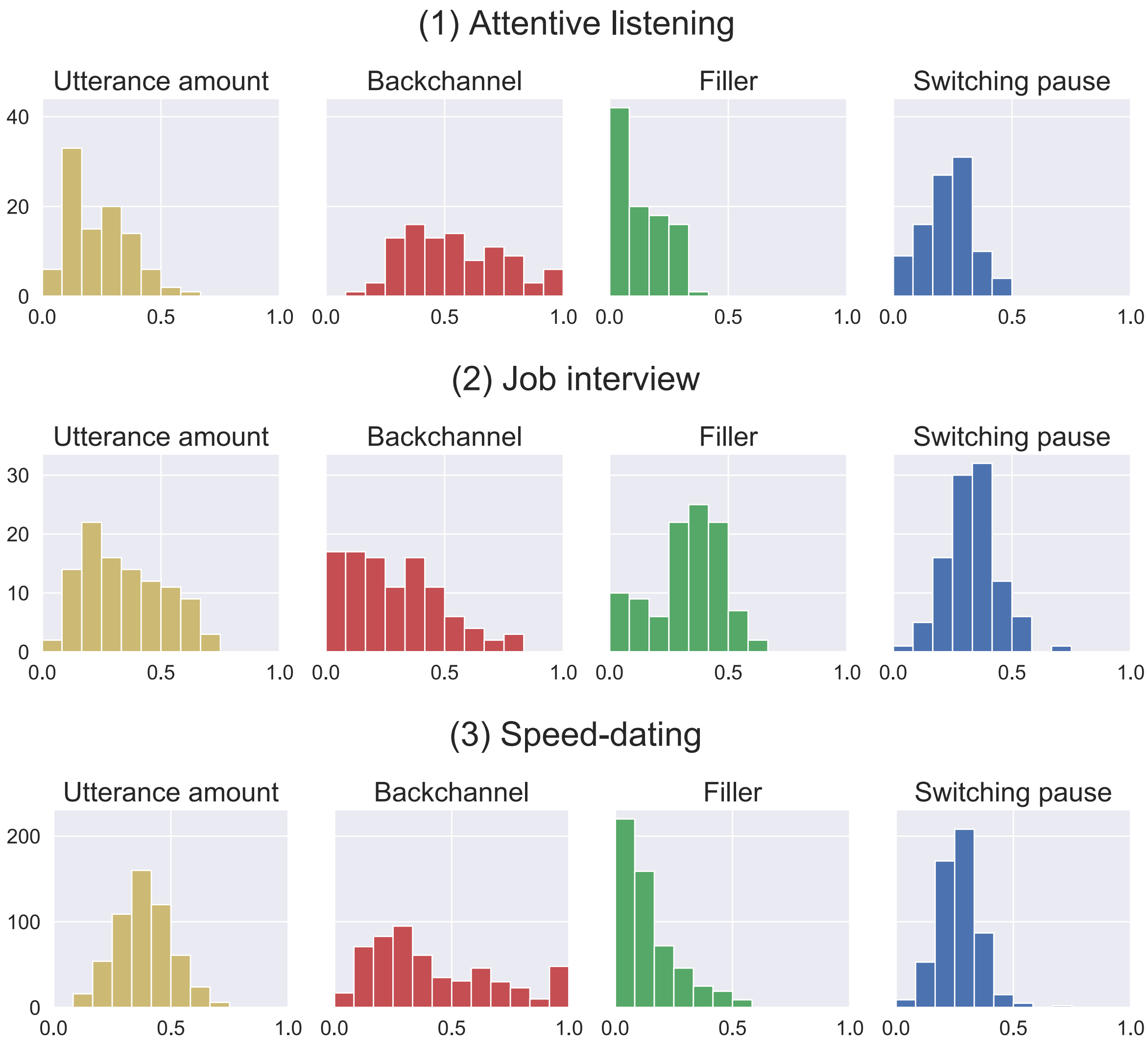


Fig. 6.2 Distribution of behaviors in each dialogue task (Histograms of each behavior)

6.3 Subjective Evaluation Using Videos

The character expression model in a spoken dialogue system has been implemented into an android ERICA and then a subjective experiment is conducted to confirm the effectiveness of the character expression in the specific dialogue task.

6.3.1 Implementation in Spoken Dialogue System

For the dialogue tasks in this experiment, we manually designed fixed dialogue flows. When we calculated the control amounts using the decoder model, the latent variable values were set by sampling from a standard normal distribution. Then, the output values were used as control amounts for the robot behaviors. The control amount values corresponded to the

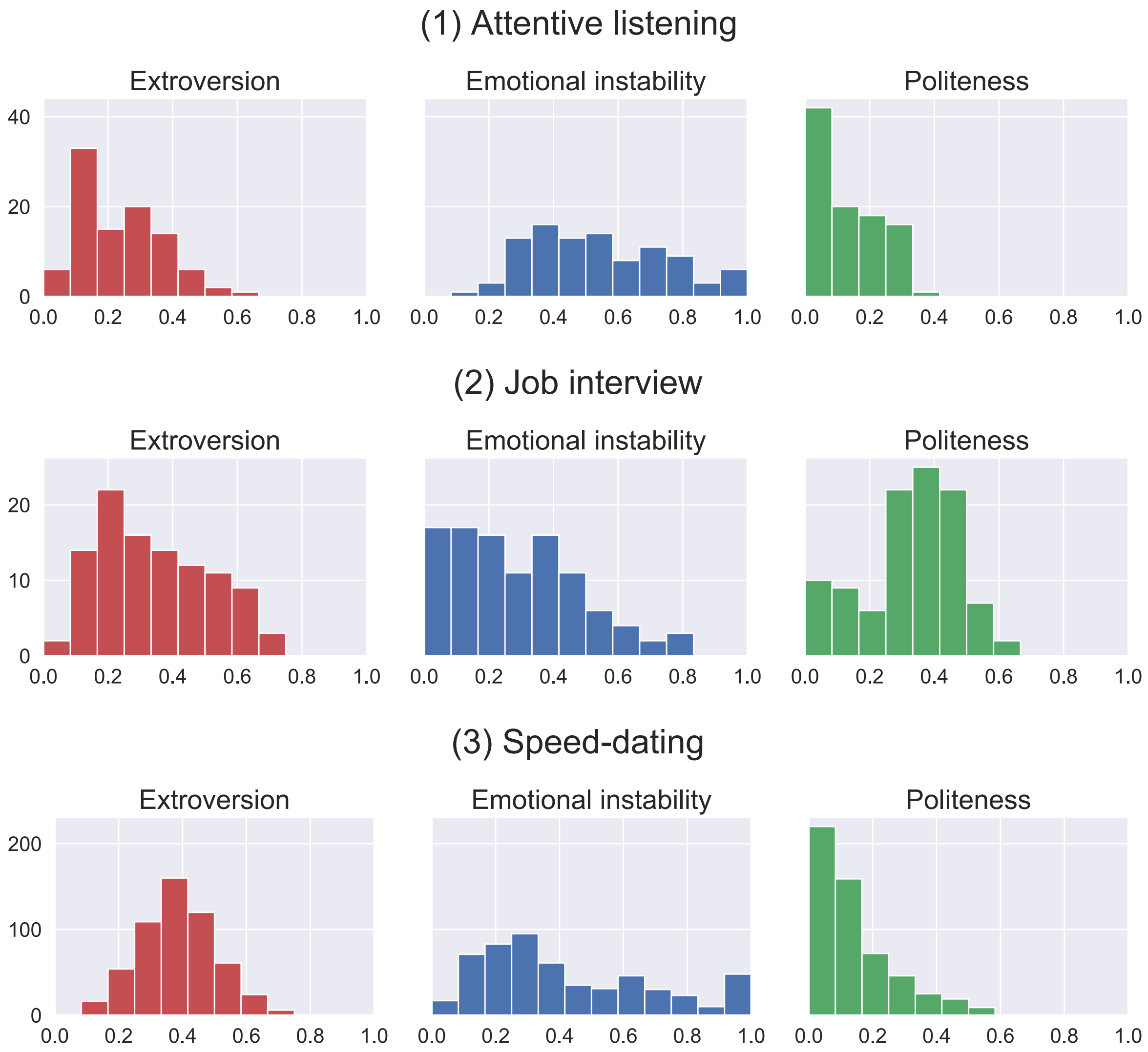


Fig. 6.3 Histograms of estimated character traits by the encoder (character recognition model) for each dialogue task

behavior setting as follows. We prepared two utterance patterns corresponding to the long and short utterance amount for the same dialogue scenario. Therefore, according to the control amount of utterance amount, the system selects one of the two-utterance patterns: long or short utterances. The system also has a function that determines the timing of backchannels every 100 milliseconds using prosodic features of the user utterance by a logistic regression model [83]. The control amount of backchannel frequency was used as a threshold of output probability in the backchannel generation module. Note that the backchannel form was fixed "*Un* (*Yeah* in English)." The control amount of filler frequency was also used as a threshold value. Fillers are inserted stochastically in the candidate positions manually designated in the system utterances. The switching pause length corresponds to the length of the pause

Table 6.1 Input character trait in each task setting

Task setting	Target characters	Input amounts of character traits		
		Extroversion	Emotional instability	Politeness
Job Interview	Polite	0.5	0.0	1.0
Laboratory guide	Extrovert, Casual	1.0	0.0	0.0
Baseline	Emotionally stable	0.5	0.0	0.5

until the robot takes a turn. The length of pause was determined by linearly normalizing the control amount to the range from 700 to $3,000$ milliseconds.

6.3.2 Experimental Setting

At first, we recorded dialogue videos where the author and ERICA talked for 8 minutes in the two different dialogue scenarios of job interview and laboratory guide. In job interview, ERICA asked questions as the interviewer, and the author responded as the interviewee. In the laboratory guide task, ERICA presented research topics for a laboratory and asked some questions to the visitor, and the author listened to the explanation and answered the questions as the visitor. For each task, we prepared two dialogue videos in two different conditions: proposed and baseline. In the proposed setting, the proposed character expression model gave a character that is regarded as appropriate for the corresponding dialogue task. In Baseline condition, ERICA spoke with an emotionally stable character. The characters represented by ERICA in each setting are summarized in Table 6.1. In the job interview, we set the character traits (polite) based on the results of the corpus analysis in Section 5.5.2. In the laboratory guide, we set the scenario where a senior student (robot) explains to a junior student. In this scenario, the dialogue should be casual to make it easy to ask questions. Therefore, we set the extrovert and casual character for the robot.

The subject experiment was conducted online with 75 university students. For each dialogue task, the subject watched the dialogue video in both the proposed and baseline setting and then made a pairwise comparison on which was more appropriate for the task. We prepared three different questions for each task as shown in Table 6.2 and questions about the character trait scores described in Section 4.2.2.

6.3.3 Experimental Results

The results are shown in Table 6.2. In job interview, two of the three question items showed significant differences (5%) in the one-tailed binomial test. For the remaining item, the

Table 6.2 Results of impression evaluation in dialogue tasks (relative evaluation)

Tasks	Items	Selection of subjects	
		Proposed	Baseline
Job interview	Which robot is settled?	62.7%**	37.3%
	Which robot listened more carefully to what you said?	53.3%	46.7%
	Which robot is better for a job interviewer?	62.7%**	37.3%
Laboratory guide	Which robot explained studies more actively?	48.0%	52.0%
	Which robot explained more clearly?	53.3%	46.7%
	Which robot is better as a laboratory guide?	52.0%	48.0%

(**$p < .05$)

proposed method was more favored, though not significant. In laboratory guide, the proposed method was more liked in two items, but there was no significant difference between the two settings. This may be because the proper character in laboratory guide depends on each subject.

In job interview, the proper character is clear because it is formal dialogue. In laboratory guide, casual presenters are desired for some people such as colleagues, while polite presentation is desired for some people such as professors vs. students. One of possible solutions is to realize an advanced character expression model that can adapt its expressed character to each user's character and individual social situation.

The results of the character impression evaluation are shown in Table 6.3. In the job interview task, the proposed method was evaluated as more polite. In the laboratory guide task, the proposed method was evaluated as more extrovert and more polite. Note that three trends are significant tendencies ($p < .10$). The intended character was recognized by the subjects with regard to most of the character traits. In the laboratory guide task, there is an inconsistency in politeness. When we increase the extroversion in the laboratory guide, the utterance amount is increased; as a result, some subjects might feel that the robot guide explained carefully and increased the impression of politeness. This is a side effect caused by extrovert traits. This suggests that we need to be careful when changing two or three traits jointly especially when their behaviors are not consistent. The results of Table 6.2 and 6.3 suggest that the task evaluation scores are higher for tasks in which the characters are correctly recognized, such as job interview.

Table 6.3 Results of character impression evaluation (7 point scales)

Character traits	Job interview		Laboratory guide	
	Proposed	Baseline	Proposed	Baseline
Extroversion	4.47	4.36	$5.21^{\dagger}$	5.06
Emotional instability	$2.33^{\dagger}$	2.96	2.45	2.55
Politeness	$5.96^{\dagger}$	5.77	$5.97^{\dagger}$	5.76

($^{\dagger} < .10$)

Table 6.4 Control amount of dialogue behavior in each condition

Task	Condition	Character			Behavior			
		Ex	Em	Po	Utt	BC	Fi	Sw
Job interview	Manual	0.00	0.00	1.00	0.14	0.09	0.58	0.48
	Corpus	0.58	0.31	0.82	0.40	0.24	0.40	0.28
	Neutral	0.50	0.00	0.50	0.34	0.20	0.05	0.20
Attentive listening	Proposed	1.00	0.00	0.00	0.73	0.98	0.09	0.08
	Corpus	0.82	2.67	0.62	0.26	0.76	0.14	0.23
	Neutral	0.50	0.00	0.50	0.34	0.20	0.05	0.20

Ex: Extroversion, Em: Emotional instability, Po: Politeness
Utt: Utterance amount, BC: Backchannel frequency
Fi: Filler frequency, Sw: Switching pause length

Table 6.5 Average behaviors in the corpus

Task	Average of features on each behavior			
	Utterance amount	Backchannel	Filler	Switching pause
Job interview	0.41	0.29	0.44	0.33
Attentive Listening	0.28	0.76	0.14	0.23

6.4 Dialogue Experiment with ERICA

The effect of character expression is examined in a dialogue experiment using the android ERICA. This experiment evaluates, when the subject talk with ERICA with the appropriate character for the dialogue task, if the subject feel that ERICA is appropriate for the task.

6.4.1 Experimental Setting

In the experiment, the system was given two social roles: job interviewer and attentive listener.

1. Job interviewer

   In this task, ERICA is interviewer and subjects are interviewee. The architecture of job interview system is described in Section 3.6.2. ERICA asks the subject pre-prepared questions, and the subject responds to the question. It is required for the interviewer to listen carefully and formally to what the subject says. Therefore, we hypothesize that an introvert and polite character would be appropriate for ERICA.

2. Attentive listener
   In this task, ERICA listens to the subject talk carefully and give feedbacks. The architecture of attentive system is described in Section 3.6.4. ERICA need to encourage subjects to talk more. Therefore, we hypothesized that an extrovert and casual character would be appropriate for ERICA.

Three character conditions are prepared for each task.

1. Manual condition
   The characters are controlled intuitively in order to represent a character appropriate for the task. An "introvert and polite" character is used for job interview. This character is determined based on the results shown in the Fig. 6.3. An "extrovert and casual" character is a condition for attentive listening. This character is determined based on the results shown in the Fig. 4.5. Therefore, it is expected that the differences in the characters are easily perceived by the subjects.

2. Corpus condition
   In this condition, the appropriate character is calculated in each task from the corpus. Using the corpus described in Section 4.5, the average values of the operator's behaviors in each task are input to the encoder of the character expression model, then the character traits are obtained as the output of the encoder. These characters are used for the appropriate character in the task.

3. Baseline condition
   In this condition, ERICA spoke with an emotionally stable character. This condition corresponds to the case where the robot does not have a character.

Table 6.4 shows the ERICA's character and the control amount of behaviors for each condition. Table 6.5 shows the average values of the operator's behaviors in the corpus. The behavior controls used in Corpus condition approximate the behavior of the operators in the corpus. It is confirmed that the character expression model can reconstruct the behavior in the corpus.

Two comparisons are conducted in the experiment.

Table 6.6 Evaluation items for each dialogue task

Task	Items
Job interview	Which robot is settled? Which robot listened more carefully? Which robot was better for a job interview?
Attentive listening	Which robot listened actively? Which robot empathized with you? Which robot was friendly? Which robot was easy to talk with? Which robot was better for a attentive listening?

1. Manual v.s. Baseline
   This is to confirm that the effect of character expression on the positive impression of the dialogue. It is expected that Manual condition is more desirable for the task than Baseline condition.

2. Manual v.s. Corpus
   This is to confirm that giving a character manually (Manual condition) has the same effect as the appropriate character in the corpus (Corpus condition).

6.4.2 Experimental Setting

In this experiment, 19 undergraduate and graduate students, 13 males and 6 females, participated. Each subject joined all six experimental conditions (2 tasks×3 conditions). Each interaction lasted approximately 5 minutes. Before starting the experiment, they were explained the two dialogue tasks and made in mind about what they would say in each dialogue for about 10 minutes. The order of the dialogue was to have the participants talk about the first task under the three conditions. After that, the second task was also performed with the same three conditions. The order of tasks and conditions was randomly rearranged.

At the end of each dialogue, the subjects evaluated questions regarding the robot's behavior and character. For the behavior evaluation, the subjects evaluated on 5-point scales from small (short) to large (long) for each the robot behavior. For character evaluation, the subjects evaluated a questionnaire shown in Table 4.2 described in Section 4.2. At the end of the experiment, the subject ranked dialogues of the three conditions on the items in Table 6.6.

Table 6.7 Average of impression evaluation of each behavior on a 5-point scale, 0: Small (Short), 1: Large (Long)

Task	Item	Manual	Baseline	t-test (p value)
Jov interview	Utterance amount	2.95	3.26	0.027*
	Backchannel frequency	2.95	3.11	0.289
	Filler frequency	4.15	3.47	0.031*
	Switching pause length	2.89	2.05	0.005**
Attentive listening	Utterance amount	2.84	2.37	0.108
	Backchannel frequency	3.79	2.89	0.004**
	Filler frequency	2.37	2.32	0.443
	Switching pause length	2.89	3.21	0.134

($^{*}p < .05$, $^{**}p < .01$)

6.4.3 Comparison 1: Manual Condition v.s. Baseline Condition

First, Manual setting and neutral conditions were compared. The results of the impression rating results for the behavior are shown in Table 6.7. In job interview, 5% significant differences of the t tests were confirmed in three behaviors of utterance amount, filler frequency, and switching pause length. In the attentive listening, a 5% significant difference of the t tests was confirmed on filler frequency.

The results of the character impression evaluation are shown in Table 6.8. In job interview, a 5% significant difference was confirmed in politeness. In attentive listening, 5% significant differences were shown in extroversion and emotional instability. This result confirms that subjects perceive differences in the behaviors and character between Manual condition and Baseline condition.

The results of relative comparisons between Manual condition and Baseline condition are shown in Table 6.9. Manual condition was evaluated as more applicable in job interview and attentive listening. In attentive listening, 5% significant differences were found in the three items of “which robot listened actively?”, “ which robot empathized with you?” and “which robot were friendly?”

These results confirmed that the subject recognized differences of behaviors and characters between Manual condition and Baseline condition. In the relative evaluation of each task, character expression were evaluated as more suitable for the task in Manual condition. The results indicated that the impression was improved by the appropriate character expression for the task.

Table 6.8 Results of character evaluation (Manual condition / Baseline condition)

Task	Item	Manual	Baseline	t-test (p value)
Job interview	Extroversion	3.64	3.89	0.253
	Emotional instability	3.47	3.22	0.171
	Politeness	5.13	4.28	0.008^{**}
Attentive listening	Extroversion	4.88	3.71	0.004^{**}
	Emotional instability	2.43	2.89	0.036^{*}
	Politeness	5.66	5.13	0.084

($^{*}p < .05$, $^{**}p < .01$)

Table 6.9 Results of evaluation tasks (Manual condition / Baseline condition)

Task	Item	Number of answering		Binomial test
		Manual	Baseline	(p value)
Job interview	Which robot is settled?	12	7	0.096
	Which robot listened more carefully?	13	6	0.052
	Which robot is better for a job interview?	12	7	0.096
Attentive listening	Which robot listened actively?	16	3	0.002^{**}
	Which robot empathized to?	15	4	0.007^{**}
	Which robot were friendly?	15	4	0.007^{**}
	Which robot was easy to talk with?	13	6	0.052
	Which robot was better for a attentive listening?	14	5	0.022^{*}

($^{*}p < .05$, $^{**}p < .01$)

6.4.4 Comparison 2: Manual Condition v.s. Corpus Condition

Then, we made a comparison between Manual condition and Corpus condition. The results of the evaluation for the behaviors are shown in Table 6.10. In job interview, there was a 5% significant difference in the t test in only utterance amount.

The results of the impression evaluation for the characters are shown in Table 6.11. In job interview, there was only 10% significant tendency of t test between Manual condition and Corpus condition in extroversion. In Manual condition, the robot is perceived by the subjects to be introvert with smaller utterance amount.

This confirmed that Manual condition was set to a more introverted character than Corpus condition in the interview (Table 6.4), and that the subject perceived the small utterance amount.

Table 6.12 shows the relative comparison between Manual condition and Corpus condition. In job interview, all items tended to be rated as more applicable in Manual condition. In attentive listening, two items “which robot listened actively?” and “which robot were friendly?” showed a significant tendency (10%). These results indicate that even when the characters were given manually, the behavior was not unnatural.

Table 6.10 Average of impression evaluation of each behavior on a 5-point scale, 0: Small (Short), 1: Large (Long)

Task	Item	Manual	Corpus	t-test (p value)
Job interview	Utterance amount	2.95	3.53	0.001**
	Backchannel frequency	2.95	3.37	0.081
	Filler frequency	4.15	4.05	0.333
	Switching pause length	2.89	3.26	0.123
Attentive listening	Utterance amount	2.84	2.58	0.165
	Backchannel frequency	3.79	3.53	0.102
	Filler frequency	2.37	2.58	0.232
	Switching pause length	2.89	2.84	0.402

($^{*}p < .05$, $^{**}p < .01$)

Table 6.11 Results of character evaluation (Manual condition / Corpus condition)

Task	Item	Manual	Corpus	t-test (p value)
Job interview	Extroversion	3.64	4.09	0.057
	Emotional instability	3.47	3.59	0.365
	Politeness	5.13	4.84	0.145
Attentive listening	Extroversion	4.88	4.84	0.443
	Emotional instability	2.43	2.54	0.252
	Politeness	5.66	5.47	0.246

($^{*}p < .05$, $^{**}p < .01$)

Table 6.12 Results of evaluation tasks (Manual condition / Corpus condition)

		Number of answering		Binomial test
Task	Item	Manual	Corpus	(p value)
Job interview	Which robot is settled?	13	6	0.096
	Which robot listened more carefully?	11	8	0.144
	Which robot is better for a job interview?	10	9	0.176
Attentive listening	Which robot listened actively?	12	7	0.096
	Which robot was empathized to?	8	11	0.144
	Which robot was friendly?	12	7	0.096
	Which robot was easy to talk with?	8	11	0.144
	Which robot was better for a attentive listening?	8	11	0.144

($^{*}p < .05$, $^{**}p < .01$)

6.5 Summary

In this chapter, the effect of character expression on task impressions was evaluated. The proposed model is implemented in ERICA that controls the behaviors appropriately. First, this system is evaluated in the subjective experiment in which the dialogue videos were viewed by third-party persons. The results showed that it is possible to realize more appropriate dialogue by expressing a given proper character in a formal job interview task. Second, this system is evaluated in the subjective experiment in which the subject talked with the system. Subjects perceive differences in the behaviors and the characters, and they felt more natural for the system expressing the appropriate character for the task. However, the best character for the task was not so obvious. In order to find the most appropriate character, we need to compare many character settings in each task, but it is left as future work.

Chapter 7

User Adaptation: Appropriate Character Expression for User Personality

7.1 Introduction

User adaptation of spoken dialogue systems is to make the systems generate behaviors appropriate to the user, which leads to increasing user satisfaction with human-robot interaction. Usually, classification of the user is conducted for user adaptation. A widely-used user classification is personality. Many studies on personality estimation [122–124] suggest that the user personality can be estimated through spoken dialogue. It was also shown that users with different personalities had different impressions of a dialogue system [29, 30]. Therefore, user adaptation based on the user personality is desired to achieve a satisfactory user experience. In this study, we propose user adaptation using a character expression, where a dialogue system expresses the character appropriate to the user personality.

It has been shown that the character expression of spoken dialogue systems leads to increasing user engagement and naturalness in dialogue [20]. Here, "personality" is used as a psychological dimension for classifying users, and "character" is used as an impression that the system gives to the user in this study. First, we define the user personality with four classes: Role model, Reserved, Self-centered and Introvert. Then, we define the system character with four classes: Role model, Reserved, Introvert and Neutral. We investigate the validity of these classifications using a speed dating corpus. Then, we conduct two subjective experiments, where the spoken dialogue system is given a social role as a laboratory guide (task-oriented dialogue) and a chit-chat (non-task-oriented dialogue), to confirm whether the combination of the user personality and the system character affects the impression of the dialogue.

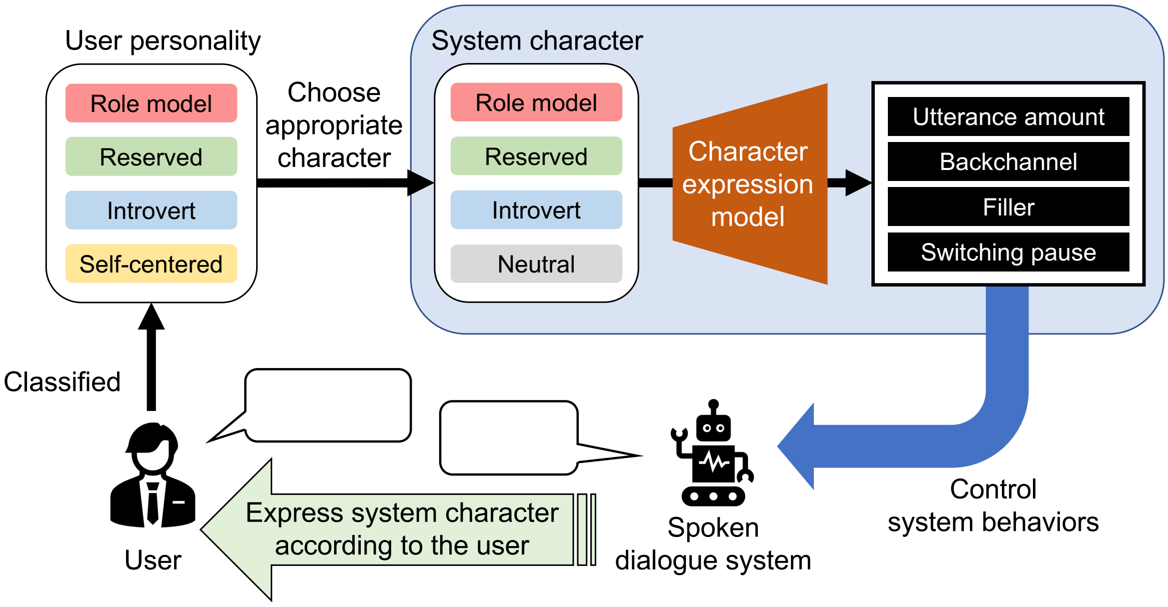


Fig. 7.1 An overview of the proposed user adaptation

In Section 7.2, we address the four classes of user personality and system character. In Section 7.3, a subjective experiment is conducted to identify the appropriate system character for the user personality in the laboratory guide task. In Section 7.4, the effect of user adaptation is evaluated in chit-chat, a non-task-oriented dialogue.

7.2 User Adaptation via Character Expression Model

An overview of the character adaptation is shown in Fig. 7.1. The system first identifies the user personality among four classes. Then, the system selects the character which is appropriate for the user personality. This classification is derived in Section 7.2.1. This is described in Section 7.2.2 We examined in Section 7.3 and 7.4. The system calculates the control amount for spoken dialogue behaviors using a character expression model according to the selected character.

7.2.1 Classification of System Character and User Personality

We define the personality and character class based on the Big Five traits [107] as summarized in Table 7.1. The Big Five traits are widely used for personality studies in psychology. We extract typical classes from the Big Five parameters and used them as user personalities and system characters. One classification was done in a previous study [125] that analyzed the

Table 7.1 Description of the Big Five traits

Traits	Typical properties		
Emotional instability (Em)	sensitive/nervous	vs.	resilient/confident
Extrovert (Ex)	outgoing/energetic	vs.	solitary/reserved
Openness (Op)	inventive/curious	vs.	consistent/cautious
Agreeableness (Ag)	friendly/compassionate	vs.	critical/rational
Conscientiousness (Co)	efficient/organized	vs.	extravagant/careless

personality ratings of over 140,000 people and then found four template clusters: Role model, Reserved, Self-centered, and Average. We apply the same methodology.

We analyzed the user personality using the speed dating dialogue corpus, where male subjects talked with the android ERICA [126], which was remotely controlled by a female operator using the Wizard of Oz (WOZ) method. They talked about getting to know each other in their first meeting. After the dialogue, each subject evaluated the item "did you have a favorable impression of the counterpart?" on a 7-point scale. However, the ratings of the subject personality and ERICA's character were not collected in this corpus.

Thus, we conducted an experiment to annotate the subject personality and the system (WOZ) character. In this experiment, 39 university students watched the video of the corpus and answered the questionnaires. The annotators answered TIPI-J [127] as their impressions of the subject and the system. TIPI-J is a Japanese translation of TIPI [38] and consists of the 10 items about personality traits on a 7-point scale. In this questionnaire, there are a couple of question items about each Big Five trait. For example, items about extroversion are "do you see yourself as extrovert and enthusiastic?" and "do you see yourself as reserved and quiet?" We prepared 195 video clips sampled from the 65 dialogues in the corpus. Each annotator evaluated approximately 20 video clips and we finally collected 778 annotations.

We normalized the evaluation score using the mean and standard deviation of each annotator's rating and clustered the Big Five scores into four classes using K-means clustering. K-means++ was used for clustering and the number of iterations was set to 300. The clustering results of the subject personality and system (WOZ) character are shown in Fig. 7.2. We name each class referring to the previous study. However, subjects clustered in the "Average" tended to be introverted. Thus, we name this class "Introvert" instead of "Average."

7.2.2 Best Combination of System Character and User Personality

Then, we analyze the effect of the combination of the WOZ character and the user personality on the impression of the dialogue. The results of the relationship between the favorable scores and the personality classes are presented in Table 7.2. We conducted a one-way

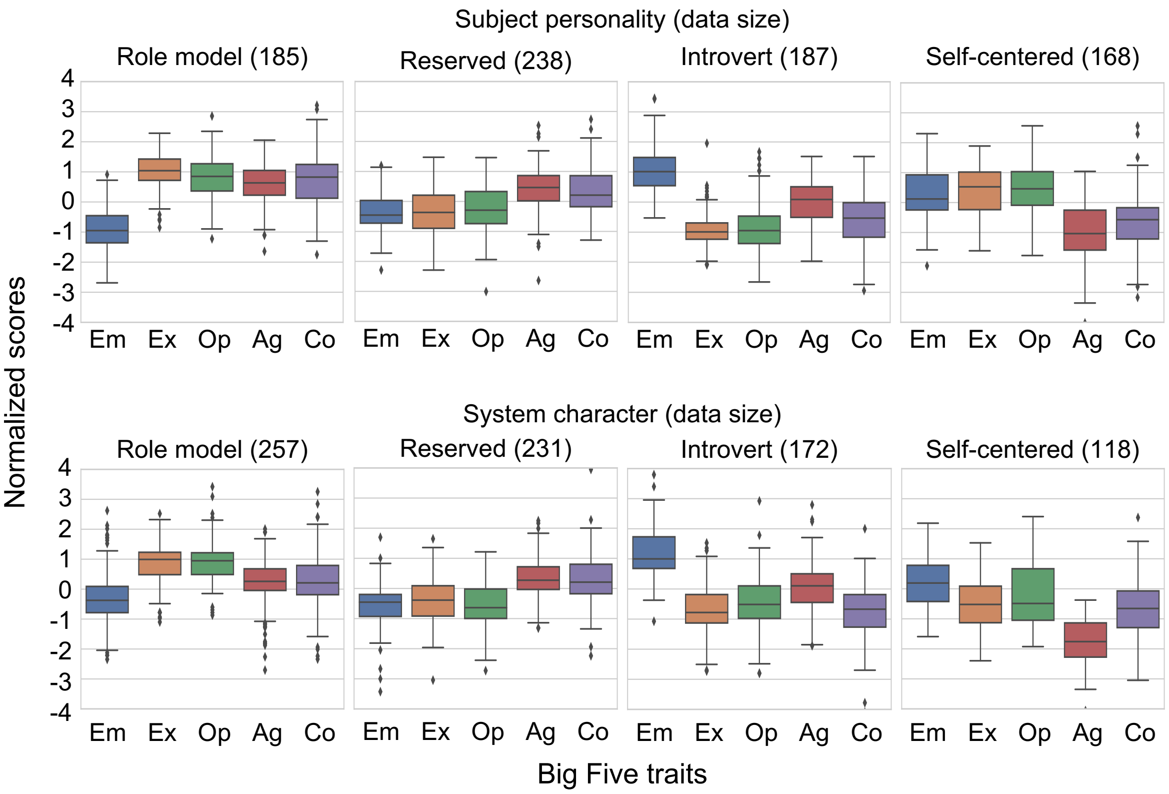


Fig. 7.2 Classification results of Big Five traits on the subject personality and the system (WOZ) character in corpus

factorial analysis of variance (ANOVA) in each subject personality class. This analysis examined whether differences in the system (WOZ) character affected users' favorable scores. The result shows that the subjects preferred different system characters depending on their personality. Highly agreeable subjects such as “Role model” and “Reserved” preferred “Introvert” system (WOZ) with less extroversion and agreeableness. Conversely, subjects with low agreeableness such as “Self-centered” and “Introvert” preferred highly agreeable systems such as “Role model” and “Reserved.” The Self-centered systems were not preferred in any case. This result suggests that a “Self-centered” character is not appropriate for the dialogue systems. Therefore, we introduce “Neutral” instead of “Self-centered” for the system (WOZ) character. The results in Table 7.2 shows that it is necessary to express different characters according to the user personality in order to make a favorable impression on the user. According to this result, we defined the user personality classes and the system character classes shown in Table 7.3.

Table 7.2 Mean of favorable scores on 7-point scale for combination between subject personality and the system character (data size)

		System character (WOZ)			
		Role model	Reserved	Introvert	Self-centered
	Role model†	4.90 (83)	4.55 (40)	**5.26** (34)	5.11 (28)
Subject	Reserved*	4.97 (76)	5.05 (77)	**5.10** (42)	5.02 (43)
personality	Introvert*	4.75 (51)	**5.03** (60)	4.94 (50)	4.92 (26)
	Self-centered†	**5.51** (47)	5.15 (54)	5.17 (46)	4.86 (32)

† $p < .10$, ∗ $p < .05$

Table 7.3 User personality and system character classes

Class name	User personality	System character	Big Five traits Em	Ex	Op	Ag	Co
Role model	○	○	▼	△	△	△	△
Reserved	○	○	▼	▼	▼	△	△
Introvert	○	○	△	▼	▼	▼	▼
Self-centered	○		△	△	△	▼	▼
Neutral		○	-	-	-	-	-

△: High, ▼: Low, -: Middle

7.2.3 Analysis between Impression of Character and Affinity: Third-party Evaluation

Then, we conduct the affinity evaluation. Affinity is calculated based on the answers of three items: "Q1: do you think they get along well," "Q2: do you think they seem easy to talk to," and "Q3: do you think they have good chemistry?" The correlations between the ratings for the three questions collected in the experiment are shown in Table 7.4. Cronbach's α coefficient [128] among the three items is 0.92. The average affinity ratings for the four personality class combinations between the subject and the operator are shown in Table 7.5. It was confirmed that the affinity scores were the highest when the operator corresponded to "Role model," regardless of which class the subject belonged to. The affinity score for Role model was significantly different from the mean score for the other classes by 5% in the t test.

When subject and operator personality are "Role model", the affinity scores are highest. Role model is characterized by high extroversion and agreeableness, and low neuroticism, indicating that active talkers fall into this class. The results suggest that the dialogue system can give an impression of high affinity when it expresses " Role model." Note that this is an

Table 7.4 Correlations between questionnaire items

	Q2	Q3
Q1	0.81	0.75
Q2	-	0.80

Table 7.5 Mean of affinity scores rated by third-party evaluators on 7-point scale for combination between system character and subject personality (the data size)

Subject	System character			
personality	Role model	Reserved	Introvert	Self-centered
Role model*	**5.26** (83)	4.62 (40)	4.48 (34)	3.82 (28)
Reserved*	**4.38** (76)	3.72 (77)	3.47 (42)	3.31 (43)
Self-centered*	**4.35** (47)	3.68 (54)	3.49 (46)	3.39 (32)
Introvert*	**3.43** (51)	3.04 (60)	2.53 (50)	2.88 (26)

*$p < .05$

impression of affinity as seen by a third party person and is not the same as an evaluation by the participants in the dialogue.

7.2.4 Analysis using Multi-modal Dialogue Corpus using a Spoken Dialogue System

We also investigate the user personality classification using another corpus. For this validation, the Hazumi corpus 1911 [129] is used. This corpus contains dialogue data between a human-operated agent and subjects. The subject evaluated his/her impression of the dialogue on 18 questions, each on an 8-point scale, in addition to his/her own rating on the Big Five scale. From 1911 dialogues in the Hazumi corpus, 27 dialogues with no missing feature data are used.

In the validation, the results of personality recognition from the agent's behavior are used as the agent's personality. To recognize personality, the character recognition model explained in Section 5.5.2 is used. The recognition results showed that the agent character was recognized as "Reserved" in all 27 dialogues. This can be attributed to the fact that the agent utterance was prepared in advance and its character was fixed. In the following analysis, it is assumed that the agent character is "Reserved".

The subject impression evaluation of the dialogues in the corpus are analyzed by subject personality class. The results of classifying the subjects into four classes based on their Big Five evaluation are shown in Fig. 7.3. The data size for each class is Role model : 3, Reserved : 10, Self centered : 7, and Introvert : 6. The results of the analysis of variance (ANOVA) in

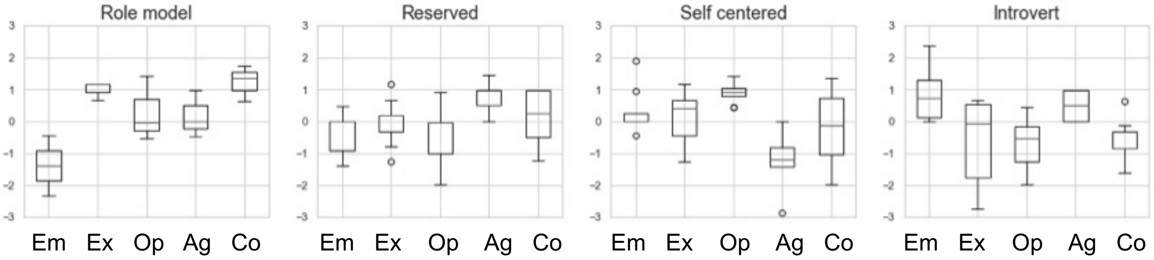


Fig. 7.3 Personality classes on Hazumi corpus

Table 7.6 Mean of impression rating in each subject personality class

Item	Personality class			
	Role	Res	Self	Int
Do the robot coordinate the conversation well?	7.3	4.8	4.0	4.8
Are you bored with the conversation?	1.7	4.2	4.4	3.7
Is the robot cooperative in the conversation?*	**7.7**	5.0	4.7	6.7
Is it a harmonious conversation?	6.0	4.3	3.4	5.0
Are you unsatisfied with the conversation?	2.0	3.9	3.9	2.5
Is the tempo of the conversation bad?	4.3	6.5	6.0	4.7
Do the conversation leave you feeling cold?	2.7	3.4	3.6	2.7
Is the conversation awkward?	4.0	3.9	4.3	2.7
Are you engaged in conversation?	6.7	3.7	3.6	4.2
Do you feel the conversation was unfocused?	3.0	3.9	4.3	2.7
Are you interested in the robot?	6.7	4.5	4.6	5.5
Are you nervous about the conversation?	1.3	4.0	4.0	3.3
Do you converse favorably?	7.7	5.5	6.1	6.8
Do you feel the conversation was active?†	**7.0**	4.1	4.0	5.8
Do you have positive conversations with each other?	7.0	4.7	5.4	6.2
Do you find the conversation boring?	1.3	3.1	3.6	4.0
Do you feel the conversation was worthwhile?	7.0	4.0	5.0	5.2
Do you feel the conversation was interminable?	4.0	5.4	5.1	4.2

Role: Role model, Res: Reserved, Self: Self centered, Int: Introvert
†$p < .10$, *$p < .05$

each item of evaluation are shown in Table 7.6. There are a 5% significant difference in "Is the robot cooperative in the conversation" and 10% significant tendencies in "Do you feel the conversation was active?". This result is consistent with the results of the compatibility analysis of Table 7.2. This suggests that the third-party person evaluates the compatibility between the subject personality and the system character based on the impression of the excitement of the dialogue.

7.3 Laboratory Guide: Task-oriented Dialogue

We first examined the effect of character adaptation in a laboratory guide, which conducts a task-oriented dialogue. We implemented the laboratory guide system on the android ERICA. In this experiment, subjects interacted with the system that expressed the four characters and evaluated the dialogue with each character.

7.3.1 Laboratory Guide System

A laboratory guide is a scenario-based dialogue system that introduces research topics to a subject. The system reads pre-defined sentences and occasionally asks the subject to ask questions. After some interaction, the system proceeds to the next topic. The average number of turns in the experiment was 12, and the dialogue duration was about 8 minutes.

The control values of the four behaviors are continuous values in the range of 0∼1 that can be directly used in the spoken dialogue system. The input of the character expression model was the mean value of each Big Five trait for each character class shown in Fig. 7.2. However, we input 0.5 to the model in the Neutral condition. The control values obtained by the character expression model for each character class are summarized in Table 7.7. The model is re-trained based on the character expression model described in Chapter 5. To train the model, the annotated data explained in Section 7.2.1 The system controls the behaviors according to the character in the manner described in Fig. 7.4. We prepared two utterance patterns corresponding to the long and short utterance amount. According to the control value of utterance amount, the system selects one of the two-utterance patterns: long or short utterances. We used the backchannel generation module [83] to control the backchannel frequency. The model determines the generation of backchannels every 100 ms by using the prosodic features of the user utterance with a logistic regression model. The control value of the backchannel frequency corresponds to the threshold of the output probability of the backchannel generation module. The control value of the filler frequency corresponds to the threshold of its probability. Fillers are inserted stochastically at the beginnings of the system utterances. The switching pause length is the length of silence before the system takes a turn. The control value of the switching pause length is linearly mapped to the switching pause length from 700 to $3,000$ ms.

7.3.2 Experimental Setting

In this experiment, 40 undergraduate and graduate students talked with the laboratory guide system with four different conditions: “Role model”,“Reserved”, “Introvert”, and

Table 7.7 Control values of dialogue behaviors in each character condition

System character	Control values of dialogue behaviors (0∼1)			
	Utterance	Backchannel	Filler	Switching pause
Role model	0.8	0.8	0.1	0.2
Reserved	0.4	0.3	0.1	0.4
Introvert	0.1	0.1	0.7	0.7
Neutral	0.5	0.3	0.2	0.3

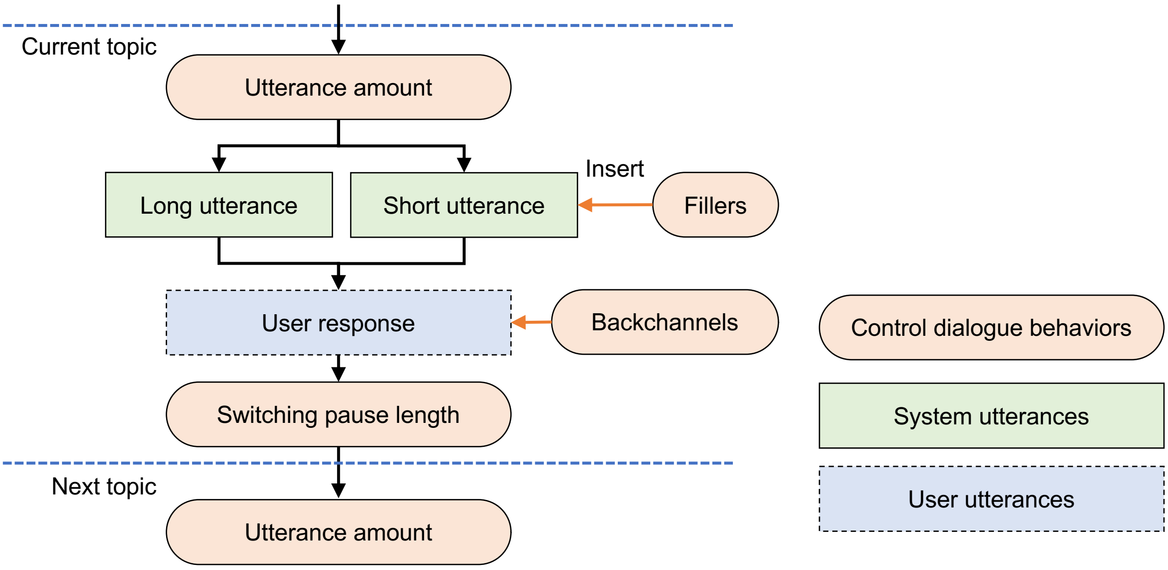


Fig. 7.4 The method of controlling dialogue behaviors for character expression

“Neutral”. In each condition, the system introduced one of four different research topics: automatic speech recognition, spoken dialogue system, acoustic signal processing, and music information processing. The system characters and dialogue topics were randomly arranged. At the beginning of the experiment, each subject answered his/her Big Five personality traits using TIPI-J [127]. At the end of the experiment, the subject answered the system character impression using TIPI-J and the questionnaire about the his/her impression with dialogue, as presented in Table 7.9. This questionnaire consists of four constructs: skill (Q1), engagement (Q2), adaptation (Q3), and naturalness (Q4). Each construct score was calculated as the average over multiple questions. “Skill” means how well the system is suited to a function as a laboratory guide. “Engagement” means how the subject is satisfied with the dialogue. “Adaptation” means how well the system adapt to the subject. “Naturalness” is used to determine whether the dialogue has broken down.

Table 7.8 Subjective evaluation scores (7-point scales) on Big Five traits for each system character condition

Character	Big Five scores: Mean (SD)				
condition	Em	Ex	Op	Ag	Co
Role model	2.80 (0.87)	4.72 (0.95)	4.37 (1.06)	4.94 (1.27)	5.02 (1.09)
Reserved	2.79 (1.02)	4.31 (1.21)	4.21 (1.10)	4.97 (1.25)	4.75 (0.97)
Introvert	4.00 (1.19)	3.63 (1.29)	4.14 (1.15)	5.06 (0.93)	4.43 (1.30)
Neutral	3.26 (1.23)	4.13 (1.24)	4.15 (1.11)	5.22 (0.94)	4.79 (1.08)

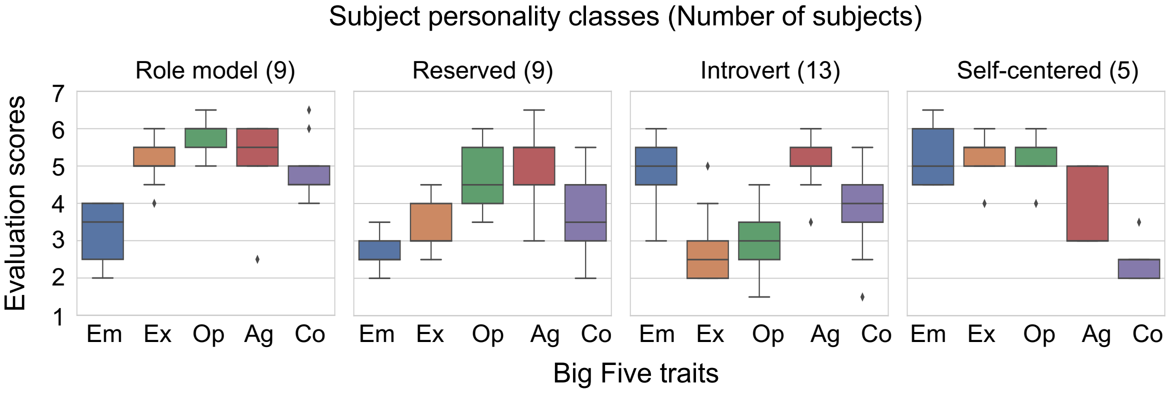


Fig. 7.5 Clustering of Big Five scores of the subjects in the laboratory guide experiment

7.3.3 Experimental Results

First, we analyze the Big Five rating for each character condition in Table 7.8. The score in each character condition was consistent with the character class tendencies in Table 7.3. This result confirms that the subjects recognized the different system character conditions. The results of the clustering of the subject personality are shown in Fig. 7.5. Based on the combination of these four personality classes and the system character conditions, we analyze the impression evaluation results.

We conducted a two-way repeated ANOVA on the conditions. The subjects' evaluation scores and the results of ANOVA are listed in Table 7.10. "System character" means the evaluation scores for each system character and "Subject personality" means the evaluation scores for each user personality. "System factor" means whether significant differences are observed among the system characters. "Subject factor" means whether significant differences are observed among the subject personalities. "Interaction effect" means whether the combinations between the system character and subject personality affect the evaluation scores. The Q4 results suggest that the character expression did not affect the naturalness. There were significant differences in the system factors of Q1. The evaluation results for each system character in Table 7.10 show that "Role model" is highly rated. Given that Q1

Table 7.9 Questionnaire used for the laboratory guide system

	Items
Q1 (Skill)	Do you think that the robot was good at explaining? Would you like the robot to explain other research topics? Would you like the robot to talk about topics other than research? Do you think the robot was a good laboratory guide?
Q2 (Engagement)	Is it easy for you to talk with the robot? Do you have a favorable impression of the robot?
Q3 (Adaptation)	Do you think that the robot adapted to you? Do you think the robot understand your personality?
Q4 (Naturalness)	Do you think that the robot spoke naturally?

Table 7.10 Mean evaluation scores (7-point scales) and the results of ANOVA in each questionnaire item

(Role: Role model, Res: Reserved, Int: Introvert, Self: Self-centered, Ne:Neutral)

										ANOVA	
Items	System character				Subject personality				System	Subject	Interaction
	Role	Res	Int	Ne	Role	Res	Int	Self	factor	factor	effects
Q1	**4.74**	4.57	3.90	4.25	**4.61**	3.91	4.36	4.56	4.71**	2.92*	2.71*
Q2	4.44	4.69	4.24	4.50	**4.68**	4.02	4.61	4.60	3.30	3.32*	2.17*
Q3	3.54	3.51	3.71	3.74	**4.08**	3.56	3.57	3.99	0.71	7.27**	1.56
Q4	4.19	4.17	3.81	4.06	4.45	4.72	3.98	4.28	$2.31^{\dagger}$	$2.54^{\dagger}$	1.65

$^{\dagger}p < .10$, * $p < .05$, ** $p < .01$

is constructed about whether the system is good as a laboratory guide, there is a suitable character for the laboratory guide. This result is reasonable for a task-oriented dialogue, since the completion of the task requires collaboration. Moreover, significant differences are observed in the subject factor of Q1, Q2, and Q3. Generally, the Role model subjects were rated highest in four personality classes.

Significant differences are also observed in the interaction effects of Q1 and Q2. This means that the combination of the system character and the subject personality influenced on the evaluation scores. Subjects' evaluation results for each combination of system character and subject personality are shown in Fig. 7.6. For example, Role model subjects prefer the Reserved system, Reserved subjects prefer the Role model system, Introvert subjects prefer the Neutral system, and Self-centered subjects prefer the Neutral system. These results show that the high agreeable subjects prefer a high agreeable character and low agreeable subjects like a neutral character. These results are not consistent with the observation in Section 7.2.1, but reasonable for the task-oriented dialogue.

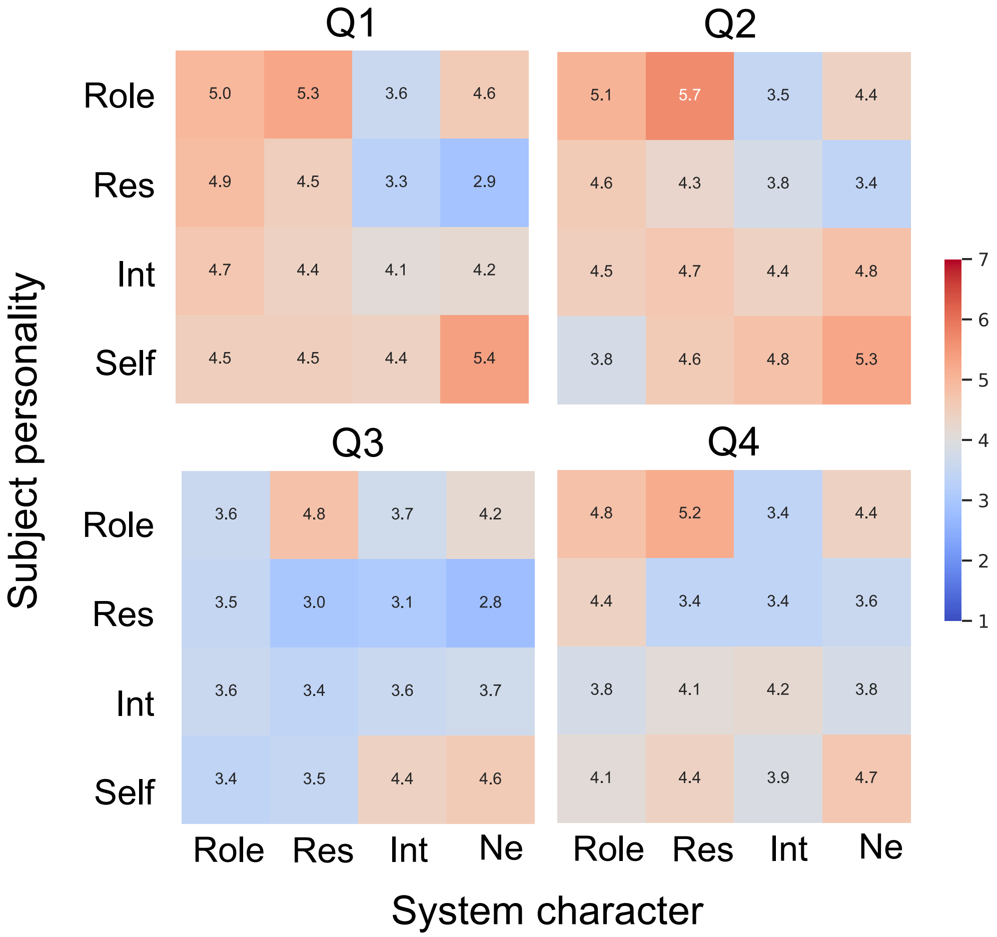


Fig. 7.6 Mean evaluation scores (7-point scales) in each combination of system character and subject personality in laboratory guide. Role: Role model, Res: Reserved, Int: Introvert, Self:Self-centered, Ne: Neutral

7.4 Chit-chat: Non-task-oriented Dialogue

We also conducted a subjective experiment using a chit-chat as non-task-oriented dialogue. We implemented a chit-chat system on ERICA. In this experiment, subjects talked with the system that expressed the four characters and talked about four topics, then they evaluated the dialogue with each character.

7.4.1 System Architecture

The chit-chat system is a combination of an example-based system and a neural generation-based system. An overview of the chit-chat system is shown in Fig. 7.7. The example-based system searches for the example most similar to the user's utterance and outputs a response. We prepared 2000 examples for one topic using a crowd-sourcing service. The system

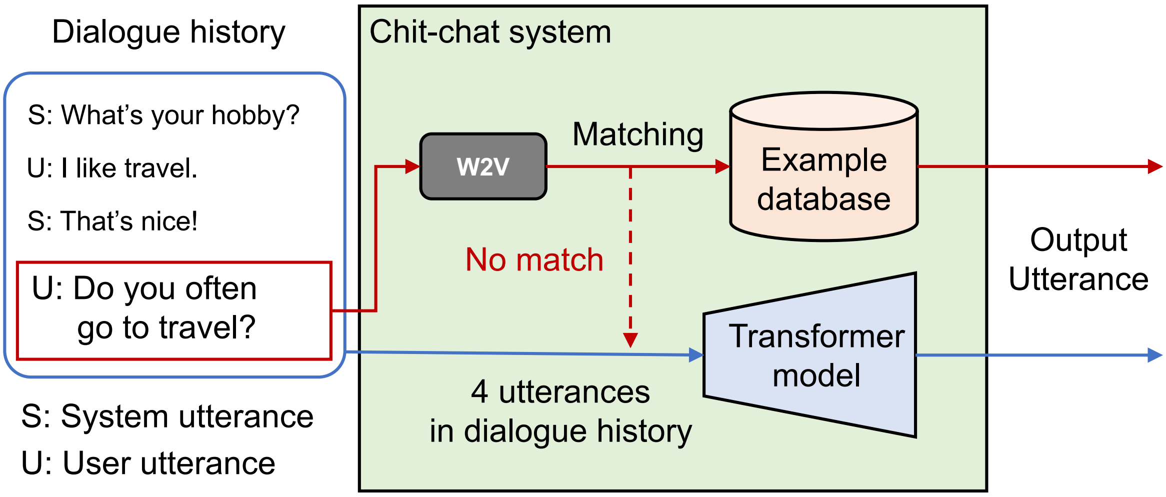


Fig. 7.7 An overview of the chit-chat system

converts the user utterance into a vector as a query. The system response is the example that has the highest degree of cosine similarity between the user utterance vector (key) and the response vector in the database (query). The query is the average vector that was computed from the content words (nouns, verbs, and adjectives) using Word2Vec[1]. The key is converted from the prepared user utterance in the database in the same manner. However, if the highest degree of similarity is lower than a threshold, the system determines that there is no matching example, and utters the generated response using the other generation model.

We prepared four items for each example in the database: expected user utterance, system response, episode, and question. The response is a reaction to the user utterance. For example, when a user asked "Do you like hot springs?", the system responds "Yes, I like hot springs." An episode is a system's self-disclosure, which follows the above response. For example, "I go to the hot spring every year." The system can ask a question to the user to continue the same topic. Therefore, the question should be related to the system episode, such as "Do you often go to hot springs?" The utterance amount is controlled by whether or not the system utters an episode.

For the neural generation model, we used the Transformer model, which was pretrained and has been fine-tuned on the JPersonaChat dataset[2]. Four utterances of the user and the system are input into the model as the dialogue history. The beam width was set to 20. The following three filters are applied to the output of the model. First, we prepared dirty word lists for removing inappropriate utterances, referring to the corpus collected from internet

[1] hottoSNS-w2v: https://github.com/hottolink/hottoSNS-w2v

[2] Japanese dialogue transformer: https://github.com/nttcslab/japanese-dialog-transformers

Table 7.11 Subjective evaluation scores (7-point scales) on Big Five traits for each system character condition

Character	Big Five scores: Mean (SD)				
condition	Em	Ex	Op	Ag	Co
Role model	3.00 (1.01)	4.71 (1.48)	4.23 (1.10)	4.67 (1.42)	4.11 (1.40)
Reserved	2.98 (1.11)	4.73 (1.32)	4.25 (1.11)	5.13 (0.99)	4.23 (1.08)
Introvert	3.71 (1.55)	3.64 (1.59)	3.70 (1.00)	5.13 (0.98)	4.16 (1.11)
Neutral	2.94 (1.10)	4.43 (1.57)	4.20 (1.10)	4.80 (1.16)	4.11 (1.44)

forum[3]. The model response candidates that contain dirty words are not output. Second, we removed questions from the generated sentences to prevent transitioning to different topics by the system utterances. Third, if the speech recognition result of the user utterance does not contain any meaningful word, then the system asks a question such as “What did you say?” The generated response is divided into a piece of sentences. Each sentence is determined stochastically whether to be included in the system’s utterance. Specifically, if the value sampled from a uniform distribution between 0 to 1 is less than the threshold value, the sentence is used. We use the value of utterance amount for this threshold. The other behaviors are controlled as described in Section 7.3.1.

7.4.2 Experimental Setting

In this experiment, 40 undergraduate and graduate students talked with the chit-chat system on four different character conditions as in Section 7.3.2. We prepared four topics: travel, movie, hometown, and student life. The characters for the dialogue topics were randomly arranged so that there was no bias in the combination. At the beginning of the experiment, each subject answered his/her Big Five personality traits using TIPI-J [127]. At the end of the experiment, the subject answered the questionnaire about three constructs: engagement (Q1), adaptation (Q2), and naturalness (Q3), as presented in Table 7.12. Each construct score was calculated as the average over multiple questions. “Engagement” means how the subject is satisfied with the dialogue. “Adaptation” means how well the system adapts to the subjects. “Naturalness” is used to determine whether the dialogue has broken down.

7.4.3 Experimental results of the chit-chat system

The Big Five rating scores in each character condition are presented in Table 7.11. This results show that the system characters are accurately perceived by the subjects. The clustering

[3] Open 2channel dialogue corpus: https://github.com/1never/open2ch-dialogue-corpus

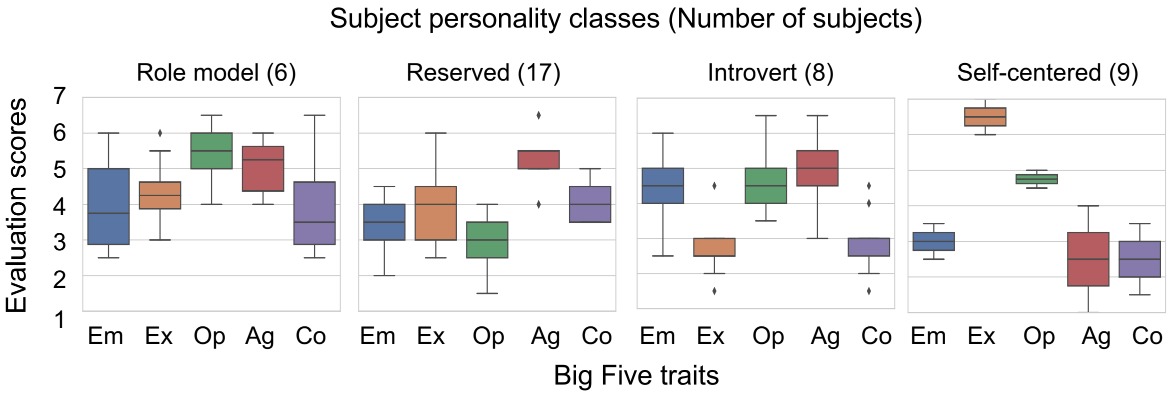


Fig. 7.8 Clustering of Big Five scores of the subjects in the chit-chat experiment

Table 7.12 Questionnaire used for the chit-chat system

Items	
Q1 (Engagement)	Do you enjoy the dialogue? Can you get along with the robot? Would you like to talk with the robot again? Is it easy for you to talk with the robot? Do you have a favorable impression of the robot?
Q2 (Adaptation)	Do you think that the robot is adapted to you? Do you think that the robot understands your personality?
Q3 (Naturalness)	Do you think that the robot spoke naturally? Do you think the robot's utterances were appropriate to the topic?

Table 7.13 Mean evaluation scores (7-point scales) and results of ANOVA in each questionnaire item

(Role: Role model, Res: Reserved, Int: Introvert, Self: Self-centered, Ne:Neutral)

	System character				Subject personality				ANOVA		
									System	Subject	Interaction
	Role	Res	Int	Ne	Role	Res	Int	Self	factor	factor	
Q1	3.56	4.10	3.72	3.69	3.93	3.69	3.89	3.75	1.65	1.10	1.04
Q2	3.33	**3.83**	3.13	3.46	3.58	3.29	3.52	3.53	2.62*	1.01	2.00*
Q3	3.80	4.19	3.91	4.03	3.81	**4.47**	3.71	3.56	1.37	5.56**	2.51*

†$p < .10$, * $p < .05$, ** $p < .01$

results of the subject personality are shown in Fig. 7.8. Based on the combination of these four personality classes and the system character conditions, we analyze the impression evaluation results.

We conducted two-way repeated ANOVA. The evaluation scores and the results of ANOVA are listed in Table 7.13. The meaning of each column in the table is the same as in Section 7.3.3. There were showed significant differences in the system factor of Q2. Subjects

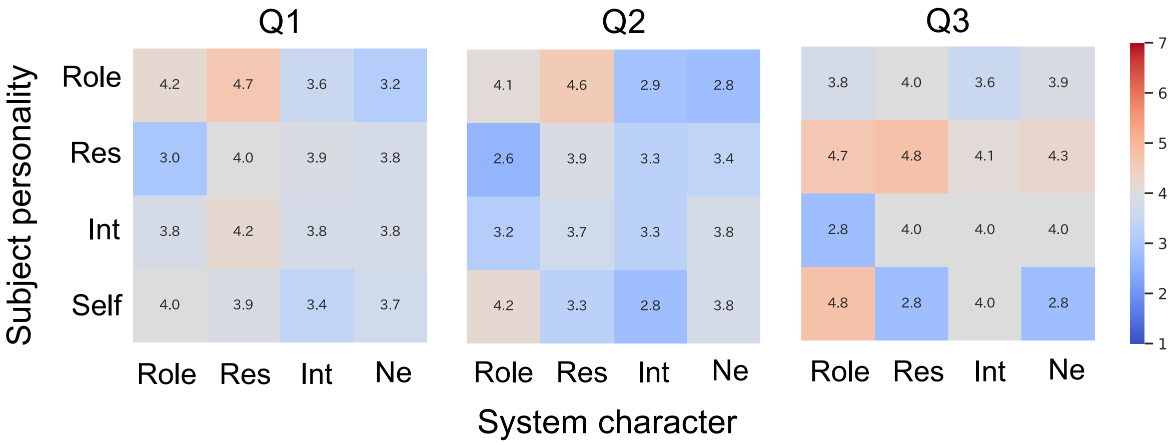


Fig. 7.9 Mean evaluation scores (7-point scales) in each combination of system character and subject personality in chit-chat. Role: Role model, Res: Reserved, Int: Introvert, Self:Self-centered, Ne: Neutral

felt that the Reserved system adapted to them. There were also significant differences in the subject factor of Q3. Reserved subjects felt that the systems spoke naturally. The result also showed significant differences in the interaction of Q2 and Q3. Subjects' evaluation results for each combination of system character and subject personality are shown in Fig. 7.9. Introvert subjects preferred the Reserved system and Self-centered subjects preferred the Role model system. These are consistent with the observation in Section 7.2.1. However, the Introvert system was not preferred by the Role model subjects and Reserved subjects. One reason for these results is that the system did not accurately express the Introvert characters.

7.5 Summary

In this study, we propose user adaptation using character expression. First, we analyzed the speed dating dialogue corpus and observed that the favorable impression depended on the combination of system character and user personality. Then, we designed the character expression model using four spoken dialogue behaviors: utterance amount, backchannel frequency, filler frequency, switching pause length. This model was implemented into a laboratory guide and a chit-chat system on the android ERICA. We conducted subjective experiments to confirm the effect of character adaptation. As a result, "Role model" characters are found to be appropriate as a laboratory guide and similar tendency to speed dating was confirmed in the chit-chat dialogue.

In future work, we will investigate the effect of character adaptation in other dialogue tasks. In addition, we will construct a real-time user adaptation model. This model will recognize user personality in the dialogue and adopt the best character for the user personality.

Chapter 8

Conclusion

This thesis addressed the character expression model for spoken dialogue systems. This chapter reviews the contributions of this thesis and addresses the future directions.

8.1 Contribution

This thesis addressed the character expression model for the spoken dialogue system,to build the relationship with the user. The goal of this study is to achieve that the spoken dialogue system expresses the appropriate character for the dialogue task and the user personality. In this study, the android ERICA was used as the spoken dialogue system because its character is easy to perceive. The character expression model controls dialogue behaviors according to input character traits: extroversion, emotional instability, and politeness. Character impression evaluation data was collected for training the model by the subjective experiment. Analysis of the collected data revealed that the speech amount, the backchannel frequency, the filler frequency, and the switching pause length were related to the impression of the character. The preliminary model using logistic regression was built and evaluated by the subjective experiment. Subject evaluated a speech sample of dialogue between a human and a robot whose behavior was controlled by the proposed model . As a result, it was confirmed that the subject recognized the characters given to the system. It was tested whether the proposed model can be used in the reverse direction to estimate characters from the speaker's behaviors. Using corpus dialogue, the speaker's characters were estimated using the proposed model from their behaviors. In subjective experiment, subjects watched a dialogue video and evaluated whether the the estimated character matched the speaker. The results showed that there was an approximate 77% agreement between the estimated character and the subject's impression of the character.

To improve the accuracy of the preliminary model, the proposed model was extended. It is costly to collect new impression evaluation data for the model expansion. To address this problem, a semi-supervised learning approach using a variational auto-encoder (VAE) was proposed. This approach utilizes both impression evaluation data (labeled data) and corpus data (unlabeled data). The encoder and decoder are trained with the labeled data and the entire model is optimized with the unlabeled data. This approach allows the model to compensate for natural behavior patterns that are lacking in the impression evaluation data with natural combinations of the behaviors observed in the corpus. The experimental result shows that the proposed model expresses the target character traits through the behaviors more precisely than the baseline model. The modeling of the unlabeled behavior (speech rate) was realized by semi-supervised learning. It was confirmed that the proposed model acquired an intuitive mapping from the character traits to the speech rate. This means that even if there are no character labels for additional behaviors, the proposed model can learn the mapping based on the relationship between the additional behaviors and existing behaviors.

Task adaptation how to use the character expression model was proposed. To verify that there is appropriate character for the task, the corpus analysis was conducted. The characters of the speakers in the dialogue corpus were estimated using the encoder of VAE as a character recognition model, and it was confirmed that the distribution of characters differed from task to task. Two subject experiments was conducted to verify the effect of task adaptation. In the first experiment, subjects watched and evaluated the dialogues videos, where ERICA talked with the user in a job interview and a laboratory guide task. As a result, it was confirmed that the evaluation score of a job interviewee's skill was enhanced by the expression of a polite character. In the next experiment, subjects talked with the ERICA expressing given characters in the job interview and attentive listening. In this experiment, there are two comparison to verify the effect of the task adaptation using the character expression. At first, manual settings of characters to match the dialogue were compared baselines without character expression. It was found that attentive listening systems with extrovert character were preferred. The second comparison was made between the best character calculated from the corpus and a manually set character. As results, it was found that there are not the large differences between the optimal character and the manual character. This result indicates that the manually set character is sufficient to express the appropriate character for the task without the need for analyzing the character from the corpus.

Finally, user adaptation using the character expression was proposed. User adaptation means that satisfaction with dialogue improves when the system express characters appropriate to the user personality. For the user adaptation, a four-class classification method for user personality and system character was proposed. As analysis results of the speed-dating dialogue corpus, it was confirmed that the favorable impression depended on the combination of the system character and the user personality. A subjective experiment result showed that there is a difference between the character desired as a laboratory guide and the character that subjects want to talk to as a dialogue partner. For example, "Role model" and "Reserved" characters are appropriate as laboratory guides. Moreover, the characters that user feels easy to talk to differ depending on their personalities. Experiments using a chit-chat system showed that it was determines whether or not the user has a positive impression of interacting with the system according to the user personality. These results obtained in this study support that switching system characters according to the user personality is effective for user adaptation.

In summary, in this thesis, the character expression model using four dialogue behaviors was proposed. Furthermore, it was also verified that the task and user adaptation using the character expression affect to improve the user's impression. Therefore, the findings of this thesis contribute to the development for spoken dialogue systems.

8.2 Future work

This study addressed a method of character expression that uses the behaviors unique to spoken dialogue. On the other hand, many character expression models have been proposed that control the content of utterance. In the future, it also consider the character expression methods that integrate these models. The proposed model uses a character definition based on the Big Five, a generic character measure. Therefore, it is believed that the proposed model can be easily integrated with other character expression models by sharing the same scale.

The character expression model was implemented on the android ERICA. Through experiments, the effectiveness of character expression by ERICA was verified. On the other hand, because ERICA has a human-like appearance, it is possible that subjects may easily perceive it's character. It is necessary to verify the effectiveness of characters in other dialogue systems such as small robots and smart speakers.

This thesis proposed two ways of using character expression: task adaptation and user adaptation. In user adaptation, prior personality assessment by the subject was required.. Therefore, it is a future challenge to build a system that can adapt to the user in real time by integrating a model for character recognition as well. It is very difficult to recognize characters using only the behaviors currently used for the character expression model. It will be considered ways that the dialog system ask personality-related questions to explicitly estimate the subject's personality.

Subjective experiments for task adaptation and user adaptation are separately conducted. For example, in the laboratory guide, the system need to express the appropriate character for the task and the user. It is not obvious whether the character should be adapted to the task or the user. To end of this problem, it is necessary to build a character adaptation model that takes into account task and user weighting. User adaptation has been based on a single interaction. In the future, the system finds the most appropriate character for the user through long-term dialogue to realize a deeper relationship with the user.

In recent years, language generation models trained on large-scale data are proposed. These models can generate natural utterances like humans. They have outperformed humans at simple tasks. These models can create utterances to express system characters in more diverse expressions than the system has ever had. Spoken dialogue systems need to improve their ability to express multi-modal behaviors as well as their ability to express utterance contents. In the future, it is expected to achieve consistent controlling utterances and behaviors by using characters as common parameters. This method allows the spoken dialogue systems to build a deep relationship with the user.

References

[1] David Traum, Priti Aggarwal, Ron Artstein, Susan Foutz, Jillian Gerten amd Athanasios Katsamanis, Anton Leuski, Dan Noren, and William Swartout. Ada and Grace: Direct interaction with museum visitors. In *ACM international conference on intelligent virtual agents (IVA)*, pages 245–251, 2012.

[2] Gary McKeown, Michel Valstar, and Maja Pantic. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[3] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jonathan Gratch, Arno Hartholt, Margot Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, and Louis Philippe Morency. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1061–1068, 2014.

[4] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 300–325, 2021.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 33:1877–1901, 2020.

[6] A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70, 2020.

[7] Rivka Levitan, Štefan Beňuš, Ramiro H. Gálvez, Agustín Gravano, Florencia Savoretti, Marian Trnka, Andreas Weise, and Julia Hirschberg. Implementing acoustic-prosodic entrainment in a conversational avatar. In *INTERSPEECH*, pages 1166–1170, 2016.

[8] Jacqueline Brixey and David Novick. *Building Rapport with Extraverted and Introverted Agents*, pages 3–13. 2017.

[9] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. "I hear you, I feel you": Encouraging deep self-disclosure through a chatbot. In *International Conference of Human-Computer Interaction (CHI)*. Association for Computing Machinery, 2020.

[10] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[11] Terry Winograd. Understanding natural language. *Cignitive pshchology*, 3(1):1–191, 1972.

[12] Victor Zue, James Glass, David Goodine, Hong Leung, Michael Phillips, Joseph Polifroni, and Stephanie Seneff. Integration of speech recognition and natural language processing in the mit voyager system. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, pages 713–716, 1991.

[13] Patti Price. Evaluation of spoken language systems: The atis domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania*, 1990.

[14] Frank Seide, Gang Li, and Dong Yu. Conversational speech transcription using context-dependent deep neural networks. In *INTERSPEECH*, 2011.

[15] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1146–1153, 1999.

[16] Hiroshi Ishiguro, Tetsuo Ono, Michita Imai, and Takayuki Kanda. Development of an interactive humanoid robot "Robovi" –an interdisciplinary approach–. In *Robotics Research*, pages 179–191. 2003.

[17] Tobias Baur, Ionut Damian, Patrick Gebhard, Kaska Porayska-Pomsta, and Elisabeth André. A job interview simulation: Social cue-based interaction with a virtual character. In *International Conference on Social Computing*, pages 220–227, 2013.

[18] Kirby Cofino, Vikram Ramanarayanan, Patrick Lange, David Pautler, David Suendermann-Oeft, and Keelan Evanini. A modular, multimodal open-source virtual interviewer dialog agent. In *International Conference on Multimodal Interaction (ICMI)*, pages 520–521, 2017.

[19] Clifford Nass, Youngme Moon, B.J.Fogg, Byron Reeves, and D.Christopher Dryer. Can computer personalities be human personalities? *Human-Computer studies*, 43:223–239, 1955.

[20] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. A survey of available corpora for building data-driven dialogue systems. *Dialogue and Discourse*, 9(1):1–49, 2018.

[21] Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Linguistic individuality transformation for spoken language. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2015.

[22] Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. Role play-based question-answering by real users for building chatbots with consistent personalities. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 264–272, 2018.

[23] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213, 2018.

[24] François Mairesse and Marilyn A. Walker. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488, 2011.

[25] Katherine Isbister and Clifford Nass. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *Human-Computer Studies*, 53(2):251–267, 2000.

[26] Maha Salem, Micheline Ziadee, and Majd Sakr. Effects of politeness and interaction context on perception and experience of HRI. In *International Conference on Social Robotics (ICSR)*, pages 531–541, 2013.

[27] Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zhu, and Quan Liu. Dually interactive matching network for personalized response selection in retrieval-based chatbots. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1845–1854, 2019.

[28] Sougata Saha, Souvik Das, and Rohini Srihari. Stylistic response generation by controlling personality traits and intent. In *Workshop on NLP for Conversational AI*, pages 197–211, 2022.

[29] Sean Andrist, Bilge Mutlu, and Adriana Tapus. Look like me: Matching robot personality via gaze to increase motivation. In *International Conference of Human-Computer Interaction (CHI)*, pages 3603–3612, 2015.

[30] Qiaoning Zhang, Connor Esterwood, X. Jessie Yang, and Lionel P. Robert Jr. An automated vehicle (AV) like me? the impact of personality similarities and differences between humans and avs. In *The Artificial Intelligence for Human-Robot Interaction (AI-HRI)*, 2019.

[31] Hans. Eysenck. *Dimensions of personality*. Oxford, 1947.

[32] David J. Pittenger. The Utility of the Myers-Briggs Type Indicator. *Review of Educational Research*, 63(4):467–488, 1993.

[33] Francis Galton. Measurement of character. *Fortnightly Review*, 36, 1884.

[34] Gordon W Allport and Henry S. Odbert. Trait-names: A psycholexical study. *Psychological Monographs*, 47, 1936.

[35] Lewis R. Goldberg. An alternative "description of personallity": the big-five factor structure. *Personality and Social Psychology*, 59(6):1216–1229, 1990.

[36] Paul T Costa and Robert R McCrae. Normal personality assessment in clinical practice: The NEO personality inventory. *Psychological Assessment*, 4(1):5–13, 1992.

[37] Daniel J. Ozer and Verónica Benet-Martínez. Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57(1):401–421, 2006.

[38] Samuel D Gosling, Peter J Rentfrow, and William B Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003.

[39] Sayuri Wada. Construction of the Big Five scales of personality trait terms and concurrent validity with NPI. *Japanese Journal of Psychology*, 67(1):61–67, 1996. in Japanese.

[40] Colin G. DeYoung, Jordan B. Peterson, and Daniel M. Higgins. Higher-order factors of the Big Five. *Personality and Individual Differences*, 33(4):533–552, 2002.

[41] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42:143–166, 2003.

[42] Heung-Yeung Shum, Xiaodong He, and Di Li. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–16, 2018.

[43] Shubhangi Tandon Sharath T.S. Stephanie Lukin Shereen Oraby, Lena Reed and Marilyn Walker. Controlling personality-based stylistic variation with neural natural language generators. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 180–190, 2018.

[44] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 994–1003, 2016.

[45] Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 167–177, 2021.

[46] Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dialogue agents. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2775–2779, 2018.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019.

[48] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer

learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[49] Tomohito Kasahara, Daisuke Kawahara, Nguyen Tung, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. Building a personalized dialogue system with prompt-tuning. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 96–105, 2022.

[50] Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. User modeling in spoken dialogue systems to generate flexible guidance. 15(1).

[51] Daniel Fernau, Stefan Hillmann, Nils Feldhus, Tim Polzehl, and Sebastian Möller. Towards personality-aware chatbots. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 135–145, 2022.

[52] Nitha Elizabeth John, Alessandra Rossi, and Silvia Rossi. Personalized human-robot interaction with a robot bartender. In *ACM Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 155–159, 2022.

[53] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Assigning personality/profile to a chatting machine for coherent conversation generation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4279–4285, 2018.

[54] Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 555–564. Association for Computing Machinery, 2021.

[55] Hongjin Qian, Zhicheng Dou, Yutao Zhu, Yueyuan Ma, and Ji-Rong Wen. Learning implicit user profile for personalized retrieval-based chatbot. In *Conference on Information and Knowledge Management (CIKM)*, pages 1467–1477, 2021.

[56] Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5808–5820, 2022.

[57] Wanling Cai and Li Chen. Predicting user intents and satisfaction with dialogue-based conversational recommendations. In *ACM Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 33–42, 2020.

[58] Takahisa Uchida, Takashi Minato, Yutaka Nakamura, Yuichiro Yoshikawa, and Hiroshi Ishiguro. Female-type android’ s drive to quickly understand a user’ s concept of preferences stimulates dialogue satisfaction: Dialogue strategies for modeling user’ s concept of preferences. *International Journal of Social Robotics*, 2021.

[59] Yasuharu Den, Nao Yoshida, Katsuya Takanashi, and Hanae Koiso. Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. In *Conference of the Oriental COCOSDA*, pages 168–173, 2011.

[60] Nigel Ward. Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, 14(1):129–182, 2006.

[61] Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and speech*, (3-4):295–321, 1998.

[62] Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *INTERSPEECH*, pages 889–892, 2005.

[63] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. Yeah, right, uh-huh: A deep learning backchannel predictor. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2016.

[64] Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. Oh, jeez! or uh-huh? a listener-aware backchannel predictor on asr transcriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8064–8068, 2020.

[65] Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. BPM_MT: Enhanced backchannel prediction model using multi-task learning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3447–3452, 2021.

[66] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the Organization of Conversational Interaction*, pages 7–55. 1978.

[67] Michiko Watanabe. *Features and Roles of Filled Pauses in Speech Communication: A corpus-based study of spontaneous speech*. Hitsuji Syobo Publishing, 2009.

[68] Matthew Marge, João Miranda, Alan W. Black, and Alexander I. Rudnicky. Towards improving the naturalness of social conversations with dialogue systems. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 91–94, 2010.

[69] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. How quickly should communication robots respond? *International Journal of Social Robotics*, 1:153–160, 2009.

[70] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *International Conference on Multimodal Interaction (ICMI)*, pages 226–234, 2019.

[71] Kengo Ohta, Masatoshi Tsuchiya, and Seiji Nakagawa. Evaluating spoken language model based on filler prediction model in speech recognition. In *INTERSPEECH*, pages 1558–1561, 2008.

[72] Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, and Akinori Ito. Filler prediction based on bidirectional lstm for generation of natural response of spoken dialog. In *IEEE Global Conference on Consumer Electronics (GCCE)*, pages 360–361, 2020.

[73] Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Ryo Ishii, Ryuichiro Higashinaka, and Yushi Aono. Neural dialogue context online end-of-turn detection. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 224–228, 2018.

[74] Shuo yiin Chang, Bo Li, Tara N. Sainath, Chao Zhang, Trevor Strohman, Qiao Liang, and Yanzhang He. Turn-taking prediction for natural conversational speech. In *INTERSPEECH*, pages 18–22, 2022.

[75] Gabriel Skantze. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 220–230, 2017.

[76] Koki Tanaka, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. End-to-end modeling for selection of utterance constructional units via system internal states. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2019.

[77] Dylan F. Glas, Takashi Minato, Carlos T. Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. ERICA: The ERATO intelligent conversational android. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 22–29, 2016.

[78] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[79] Carlos T. Ishi, Chaoran Liu, Jani Even, and Norihiro Hagita. Hearing support system using environment sensor network. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 1275–1280, 2016.

[80] Sei Ueno, Hirofumi Inaguma, Masato Mimura, and Tatsuya Kawahara. Acoustic-to-word attention-based model complemented with character-level ctc-based model. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5804–5808, 2018.

[81] Carlos T. Ishi, Chaoran Liu, Hiroshi Ishiguro, and Norihiro Hagita. Evaluation of formant-based lip motion generation in tele-operated humanoid robots. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 2377–2382, 2012.

[82] Kurima Sakai, Carlos T. Ishi, Takashi Minato, and Hiroshi Ishiguro. Online speech-driven head motion generating system and evaluation on a tele-operated robot. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 529–534, 2015.

[83] Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, 2017.

[84] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios. In *International Conference on Multimodal Interaction (ICMI)*, pages 78–86, 2018.

[85] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv*, 2013.

[86] Antoine Raux and Maxine Eskenazi. A finite-state turn-taking model for spoken dialog systems. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 629–637, 2009.

[87] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. Collecting evaluative expressions for opinion extraction. In *International Conference on Natural Language Processing (ICON)*, pages 596–605, 2005.

[88] Rajesh Ranganath, Daniel Jurafsky, and Daniel Mcfarland. It' s not you, it' s me: Detecting flirting and its misperception in speed-dates. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 334–342, 2009.

[89] Koichiro Yoshino and Tatsuya Kawahara. Conversational system for information navigation based on pomdp with user focus tracking. *Computer Speech and Language*, 34(1):275–291, 2015.

[90] Koji Inoue, Divesh Lala, Kenta Yamamoto, Katsuya Takanashi, and Tatsuya Kawahara. Engagement-based adaptive behaviors for laboratory guide in human-robot dialogue. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2021.

[91] Tatsuya Kawahara. Spoken dialogue system for a human-like conversational robot ERICA. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2018.

[92] Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. Automatic conversion of sentence-end expressions for utterance characterization of dialogue systems. In *Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 307–314, 2015.

[93] Shen Raymond, Kikuchi Hideaki, Ohta Katumi, and Mitamura Takeshi. Towards the characterization control in personalized vehicles by text-based style change. *Infomation Processing Society of Japan*, 53(4):1269–1276, 2012. in Japanese.

[94] Rivka Levitan, Stefan Benus, Ramiro H Gálvez, Agustín Gravano, Florencia Savoretti, Marian Trnka, Andreas Weise, and Julia Hirschberg. Implementing acoustic-prosodic entrainment in a conversational avatar. In *INTERSPEECH*, volume 16, pages 1166–1170, 2016.

[95] Hans Eysenck. *The biological basis of personality*. Routledge, 2017.

[96] John M. Digman. Higher-order factors of the big five predict conformity: Are there neurosis of health? *Journal of Personality and Social Psychology*, 73(6):1246–1256, 1997.

[97] Etienne D. Sevin, Sylwia Julia Hyniewska, and Catherine Pelachaud. Influence of personality traits on backchannel selection. In *ACM international conference on intelligent virtual agents (IVA)*, pages 187–193, 2010.

[98] Swati Gupta, Marilyn A Walker, and Daniela M Romano. How rude are you?: Evaluating politeness and affect in interaction. In *ACII*, pages 203–217, 2007.

[99] Ning Wang, W Lewis Johnson, Richard E Mayer, Paola Rizzo, Erin Shaw, and Heather Collins. The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies*, 66(2):98–112, 2008.

[100] Chika Nagaoka, Masashi Komori, Toshie Nakamura, and Maria Raluca Draguna. Effects of receptive listening on the congruence of speakers' response latencies in dialogues. *Psychological Reports*, 97(1):265–274, 2005. (in Japanese).

[101] Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information. In *International Conference on Automation, Robotics and Applications (ICARA)*, pages 379–384, 2004.

[102] Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel G Ward. Prediction and generation of backchannel form for attentive listening systems. In *INTERSPEECH*, pages 2890–2894, 2016.

[103] Khiet P Truong, Ronald Poppe, and Dirk Heylen. A rule-based backchannel prediction model using pitch and pause information. In *INTERSPEECH*, pages 3058–3061, 2010.

[104] Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. Talking with ERICA, an autonomous android. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 212–215, 2016.

[105] Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. Towards an iso standard for dialogue act annotation. In *International Conference on Language Resources and Evaluation (LREC)*, pages 2548–2555, 2010.

[106] John M Digman. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440, 1990.

[107] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

[108] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500, 2007.

[109] Lionel P. Robert Jr, Rasha Alahmad, Connor Esterwood, Sangmi Kim, Sangseok You, and Qiaoning Zhang. *A Review of Personality in Human-Robot Interaction*. Now Foundations and Trends, 2020.

[110] K. R. Scherer. *Scocial markers in speech*, chapter Personality markers in speech, pages 147–209. Cambridge University Press, 1979.

[111] J. B. Weaver. *Communication and Personality: Trait perspectives*, chapter Personality and self-perceptions about communication, pages 95–118. Hampton Press, 1998.

[112] A. Gill and J Overlander. Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; neuroticism is a worry. In *Annual Conference of the Cognitive Science Society*, pages 456–461, 2003.

[113] J. Oberlander and A. J. Gill. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, (42):239–270, 2006.

[114] Vasant Srinivasan and Leila Takayama. Help me please: Robot politeness strategies for soliciting help from humans. In *International Conference of Human-Computer Interaction (CHI)*, pages 4945–4955, 2016.

[115] Fabio Valente, Samuel Kim, and Petr Motlicek. Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In *INTERSPEECH*, pages 1183–1186, 2012.

[116] Laura M. Pfeifer and Timothy Bickmore. Should agents speak like, um, humans? the use of conversational fillers virtual agents. In *ACM international conference on intelligent virtual agents (IVA)*, 2009.

[117] Mingzhi Yu, Emer Gilmartin, and Diane Litman. Identifying personality traits using overlap dynamics in multiparty dialogue. In *INTERSPEECH*, pages 15–19, 2019.

[118] Katherine Metcalf, Barry-John Theobald, Garrett Weinberg, Robert Lee, Ing-Marie Jonsson, Russ Webb, and Nicholas Apostoloff. Mirroring to build trust in digital assistants. In *INTERSPEECH*, pages 4000–4004, 2019.

[119] Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. In *NIPS*, 2014.

[120] Teruhisa Uchida. Effects of the speech rate on speakers' personality-trait impressions. *Japanese Journal of Psychology*, 73(2):131–139, 2002. in Japanese.

[121] François Mairesse and Marilyn A. Walker. Automatic recognition of personality in conversation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 85–88, 2006.

[122] Onno Kampman, Elham J. Barezi, Dario Bertero, and Pascale Fung. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *ACL*, pages 606–611, 2018.

[123] Guozhen An, Sarah Ita Levitan, Julia Hirschberg, and Rivka Levitan. Deep personality recognition for deception detection. In *INTERSPEECH*, pages 421–425, 2018.

[124] Chanchal Suman, Sriparna Saha, Aditya Gupta, Saurabh Kumar Pandey, and Pushpak Bhattacharyya. A multi-modal personality prediction system. *Knowledge-Based Systems*, 236:107715, 2022.

[125] Martin Gerlach, Beatrice Farb, William Revelle, and Luís A. Nunes Amaral. A robust data-driven apprach identifies four personality types across four large data sets. *Nature Human Behaviour*, 2:753–742, 2018.

[126] Tatsuya Kawahara. Spoken dialogue system for a human-like conversational robot ERICA. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2018.

[127] Atsushi Oshio, Shingo Abe, and Pino Cutrone. Development, reliability, and validity of the japanese version of ten item personality inventory (TIPI-J). *The Japanese Journal of Personality*, 21(1):40–52, 2012.

[128] Lee J. Cronbach. Coefficient alpha and test internal structure of test. *Psychometrika*, 16:297–334, 1951.

[129] Kazunori Komatani and Shogo Okada. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, 2021.

List of Publications by the Author

Journal Articles

1. Kenta Yamamoto, Koji Inoue, Tatsuya Kawahara, “User adaptation using character expression for conversational robots”, Advanced robotics, submitted **(Chapter 7)**

2. Kenta Yamamoto, Koji Inoue, Tatsuya Kawahara, “Character expression for spoken dialogue systems with Semi-supervised learning using variational auto-encoder”, Computer Speech and Language, Vol. 79, No.101469, 2023. **(Chapter 5, Chapter 6)**

3. 山本 賢太, 井上 昂治, 中村 静, 高梨 克也, 河原 達也, “人間型ロボットのキャラクタ表現のための対話の振る舞い制御モデル”, 人工知能学会論文誌, Vol. 33, No. 5, pp. C-I37_1-9, 2018. **(Chapter 4)**

International Conferences

1. Kenta Yamamoto, Koji Inoue, Tatsuya Kawahara, “Character adaptation of spoken dialogue systems based on user personalities”, In Proc. Int’l Workshop Spoken Dialogue Systems (IWSDS), 2023. **(Chapter 7)**

2. Kenta Yamamoto, Koji Inoue, Tatsuya Kawahara, “Semi-supervised learning for character expression of spoken dialogue systems”, In Proc. INTERSPEECH, pp.4188–4192, 2020. **(Chapter 5)**

3. Kenta Yamamoto, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, Tatsuya Kawahara, “A character expression model affecting spoken dialgoue behaviors”, In Proc. Int’l Workshop Spoken Dialogue Systems (IWSDS), 2020. **(Chapter 4)**

4. Kenta Yamamoto, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, Tatsuya Kawahara, “Dialogue behavior control model for expressing a character of humanoid robots”, APSIPA ASC, pp. 1732–1737, 2018.

Domestic Conferences

1. 山本 賢太, 河野 誠也, 河原 達也, 吉野 幸一郎, “フレーズアライメントと文構造に基づくデータ拡張を用いた頑健な自然言語生成”, 研究報告自然言語処理 (NL), 2022-NL-252, 2022.

2. 山本 賢太, 井上 昂治, 河原 達也, “音声対話システムのユーザ適応に向けたパーソナリティの関係性の分析”, 人工知能学会研究会資料, SLUD-093-02, 2021.

3. 山本 賢太, 井上 昂治, 河原 達也, “VAEを用いた半教師あり学習による音声対話システムのためのキャラクタ表現”, 人工知能学会研究会資料, SLUD-C002-31, 2020.

4. 山本 賢太, 井上 昂治, 中村 静, 高梨 克也, 河原 達也,“対話のふるまいに基づくキャラクタ表現の対話コーパスにおける分析”, 人工知能学会研究会資料, SLUD-B803-08, 2019.

5. 山本 賢太, 井上 昂治, 中村 静, 高梨 克也, 河原 達也, “対話のふるまい制御によるキャラクタ表現モデルと対話コーパスによる検証”, 情報処理学会全国大会講演論文集, 2T-07, 2019.

6. 山本 賢太, 井上 昂治, Divesh Lala, 中村 静, 高梨 克也, 河原 達也, “自律型アンドロイドERICAによる傾聴対話”, 人工知能学会研究会資料, SLUD-B802-13, 2018.

7. 山本 賢太, 井上 昂治, 中村 静, 高梨 克也, 河原 達也, “自律型アンドロイドのキャラクタ表現のための対話の振る舞い制御モデルの構築と評価”, 情報処理学会全国大会講演論文集, 6Q-06, 2018.

8. 山本 賢太, 井上 昂治, 中村 静, 高梨 克也, 河原 達也, “自律型アンドロイドの対話の振る舞い制御モデルによるキャラクタ表現法の検討”, 人工知能学会研究会資料, SLUD-B508-05, 2017.

9. 山本 賢太, 井上 昂治, 中村 静, 高梨 克也, 河原 達也, “自律型アンドロイドのキャラクタ表現のための対話の振る舞い制御”, 情報処理学会全国大会講演論文集, 7M-01, 2017.