TITLE:

Character expression for spoken dialogue systems with semi-supervised learning using Variational Auto-Encoder

AUTHOR(S):

Yamamoto, Kenta; Inoue, Koji; Kawahara, Tatsuya

CITATION:

Yamamoto, Kenta ...[et al]. Character expression for spoken dialogue systems with semi-supervised learning using Variational Auto-Encoder. Computer Speech & Language 2023, 79: 101469.

ISSUE DATE:

2023-04

URL:

http://hdl.handle.net/2433/281689

RIGHT:

© 2022 The Authors. Published by Elsevier Ltd.; This is an open access article under the CC BY-NC-ND license.

京都大学学術情報リポジトリ

KURENAI 紅

Kyoto University Research Information Repository

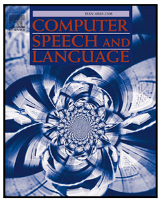
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp

Computer Speech and Language 79 (2023) 101469

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl



Character expression for spoken dialogue systems with semi-supervised learning using Variational Auto-Encoder

Kenta Yamamoto [*], Koji Inoue, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Japan

ARTICLE INFO

Keywords:
Spoken dialogue system
Character
Semi-supervised learning
Variational auto-encoder (VAE)

ABSTRACT

Character of spoken dialogue systems is important not only for giving a positive impression of the system but also for gaining rapport from users. We have proposed a character expression model for spoken dialogue systems. The model expresses three character traits (extroversion, emotional instability, and politeness) of spoken dialogue systems by controlling spoken dialogue behaviors: utterance amount, backchannel, filler, and switching pause length. One major problem in training this model is that it is costly and time-consuming to collect many pair data of character traits and behaviors. To address this problem, semi-supervised learning is proposed based on a variational auto-encoder that exploits both the limited amount of labeled pair data and unlabeled corpus data. It was confirmed that the proposed model can express given characters more accurately than a baseline model with only supervised learning. We also implemented the character expression model in a spoken dialogue system for an autonomous android robot, and then conducted a subjective experiment with 75 university students to confirm the effectiveness of the character expression for specific dialogue scenarios. The results showed that expressing a character in accordance with the dialogue task by the proposed model improves the user's impression of the appropriateness in formal dialogue such as job interview.

1. Introduction

Many spoken dialogue systems have been developed and used in smart speakers and conversational robots. In practical spoken dialogue systems, specific social roles are given to the systems, such as a psychological counselor (DeVault et al., 2014), museum guide (Traum et al., 2012), and attentive listener (McKeown et al., 2012). However, human behaviors observed during human–machine dialogues are much different from those of human–human dialogues. Our study goal is to realize a dialogue system that behaves like a human being in accordance with the social role.

It is desired for such spoken dialogue systems to express their character (e.g., extrovert) for human-like interaction (Fong et al., 2003; Serban et al., 2018). For example, museum guide systems are expected to be extroverted and intelligent, and counseling systems are expected to be introverted, emotionally stable, and agreeable. It has been found that the character expression of spoken dialogue systems led to increasing user engagement and naturalness in dialogue (Nass et al., 1955; Nass, 2000; Salem et al., 2013). Previous works investigated character expression models that controlled the linguistic patterns of system utterances (Mairesse and Walker, 2011; Ogawa et al., 2014; Miyazaki et al., 2015; Mizukami et al., 2015). However, some behaviors observed in spoken dialogue, such as speech style, affect the impression of character (Valente et al., 2012; de Sevin et al., 2010; Shiwa et al., 2009; Pfeifer and Bickmore, 2009). Therefore, spoken dialogue behaviors should be considered in addition to the content of the utterance.

[*] Corresponding author.
E-mail address: yamamoto@sap.ist.i.kyoto-u.ac.jp (K. Yamamoto).

https://doi.org/10.1016/j.csl.2022.101469
Received 10 May 2021; Received in revised form 24 November 2022; Accepted 25 November 2022
Available online 5 December 2022
0885-2308/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅
Kyoto University Research Information Repository

We proposed a character expression model that controls four spoken dialogue behaviors: utterance amount, backchannels, fillers, and switching pause length (Yamamoto et al., 2018, 2020). At first, we created pair data of the spoken dialogue behaviors and character scores in order to train a 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.

In our previous study, we collected the training data (Yamamoto et al., 2018) and proposed the character expression model (Yamamoto et al., 2020). In this paper, we enhance the model training and further analyze our character expression model in the context of dialogue tasks. Specifically, the encoder, decoder, and reconstruction loss are optimized simultaneously. Then, we implement the proposed character expression model with a spoken dialogue system of an android robot, and conduct a subjective experiment to confirm that the android robot can express the suitable character for specific dialogue tasks such as job interview and laboratory guide.

The paper is organized as follows. In Section 2, we introduce related works on definitions of character and character expression models. In Section 3, we define the character traits (input) and spoken dialogue behaviors (output) used in this study and give an analysis on the relationship between spoken dialogue behaviors and dialogue tasks in the corpus data. Training data, including both labeled and unlabeled data, are explained in Section 4. The proposed semi-supervised learning is explained in Section 5 and also evaluated in Section 6. Finally, we describe an implementation of the proposed model in a spoken dialogue system and report a subjective experiment in Section 7.

2. Related works

There have been few approaches taken in previous research to express character in dialogue. In this section, we summarize the definitions of character and previous studies on character expression in dialogue systems.

2.1. Definition of character

In psychology, concepts such as character and personality have been used to explain the tendencies of human behaviors. Various scales have been proposed to describe human personality (Eysenck, 2017; Hathaway and McKinley, 1940; Cattell, 1956). The five-factor model (Big Five personality scale) is one of the most widely-used and reliable scales (Digman, 1990; Goldberg, 1990; McCrae and John, 1992; Costa and McCrae, 1992). 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 (Ozer and Benet-Martínez, 2006).

Another approach for defining character is to describe it by sentences like “I like to travel.”, called as *persona* (Zhang et al., 2018). Recently, a large number of persona sentences together with persona-aligned dialogue has been collected, and training neural conversational models has been attempted. However, it is difficult to make the persona data tunable for users (or developers) of dialogue systems, so they need to write persona sentences for each specific dialogue task. Besides, these studies suppose text dialogue, which is different from spoken dialogue. In this study, we use the part of the Big-Five scales to realize the tunable model.

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 (Serban et al., 2018; Nass et al., 1955). The expression of character in a dialogue system has been shown to have an impact on trust-building and task accomplishment (Fong et al., 2003; Shum et al., 2018).

Other previous studies have addressed character expression models that control the linguistic patterns of system utterances (Mairesse and Walker, 2011; Miyazaki et al., 2015; Mizukami et al., 2015; Ogawa and Kikuchi, 2017). As a result, data for character expression have been collected in the form of text dialogue (Sugiyama et al., 2014; Li et al., 2016; Higashinaka et al., 2018). However, for spoken dialogue, in addition to the text style, spoken dialogue behaviors need to be considered and collected.

3. Character traits and spoken dialogue behaviors

In this section, we address the problem formulation shown in Fig. 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.

A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp

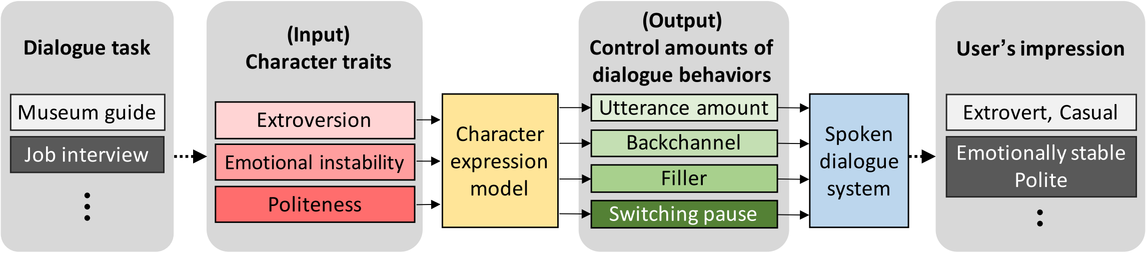


Fig. 1. Problem formulation of character expression.

3.1. Character traits used in this study

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 (Digman, 1990; McCrae and John, 1992; Costa and McCrae, 1992). In previous studies, the Big Five traits have been used to define the personality of dialogue systems (Mairesse and Walker, 2011; de Sevin et al., 2010). Extroversion is a major factor that determines the impression of system characters (Mairesse et al., 2007; Robert et al., 2020), 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 (Scherer, 1979; Weaver, 1998; Gill and Overlander, 2003; Oberlander and Gill, 2006). 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 (Srinivasan and Takayama, 2016). For example, the system could behave politely in a formal situation while casually with intimate users.

3.2. 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 affected the impression of dialogue partners (Valente et al., 2012; de Sevin et al., 2010; Shiwa et al., 2009; Pfeifer and Bickmore, 2009; Yu et al., 2019; Metcalf et al., 2019). 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 (Ward, 2006; Den et al., 2011). 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 (Sacks et al., 1978; Watanabe, 2009). 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 (de Sevin et al., 2010), but we focus on the frequencies in this study for the simplicity of the model. 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 our previous study, 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 (Yamamoto et al., 2018).

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 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.

4. 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), and the unlabeled data was derived from a human–robot dialogue corpus. We explain how we collected the data and the characteristics of each type of data.

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

Table 1
Correspondence between control amount and actual behavior features.

	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 s (overlap)	3.0 s

Table 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*[a]
	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

[a] Invert scale.

4.1. Labeled data: Character impression evaluation

To collect supervised training data, we conducted an impression evaluation of the character traits. The subjects were 46 university students (18 females and 28 males, from 18 to 23 years old). Note that this experiment was done in the Japanese language. Each subject was asked to listen to speech samples and then to evaluate his/her impression as the three character trait scores of the speaker. The evaluation was based on a 7-point scale from 1 (not at all) to 7 (completely) for ten items regarding extroversion, emotional instability, and politeness. For extroversion and emotional instability, we used eight adjectives (four for each) from a short version of the Big Five scale (Uchida, 2002; Wada, 1996). We also used two more adjectives, *polite* and *courteous*, for the third trait, politeness. The adjectives used in the evaluation are shown in Table 2.

The speech samples used in this experiment were generated as follows. In advance, we selected two dialogue scenarios from our human–robot dialogue corpus described in Section 4.2. On the basis of each scenario, we artificially generated several speech samples by controlling dialogue behaviors observed in the corpus. The robot utterances were generated by text-to-speech software. First, we generated a standard speech sample where backchannel and filler tokens appeared moderately in the original dialogue, and the switching pause length was set to 0.5 s. We adjusted each behavior one by one from the standard dialogue sample. We used these generated speech samples to compare the perceived character traits between different conditions for each dialogue behavior (e.g., high backchannel frequency vs. low backchannel frequency). Backchannels were inserted at all clause boundaries (Takanashi et al., 2003) in the user utterances in the high frequency condition, and all backchannels are removed in the low frequency condition. Fillers were inserted at all clause boundaries and at the sentence beginning of the utterances in the high frequency condition, and all fillers were removed in the low frequency condition. The switching pause length was set to 3 s in the long condition, and the start of the system utterance overlapped with the end of the user utterance by 0.5 s in the short condition. With regard to the utterance amount of the system, we prepared two different scenarios: system-dominant (large condition), and the other was user-dominant (small condition). The standard deviation of the character impression evaluation values between subjects in each trait was about 1 point on 7-point scales.

In this study, we use the character trait scores obtained through this impression evaluation as labeled data. The number of available samples was 736, and they were randomly divided into 662 samples for training and 74 samples for testing. Each sample corresponded to one where one of the subjects evaluated one of the controlled speech samples. The evaluated character trait scores are normalized from 0 to 1 by the rule explained in Table 1.

4.2. Unlabeled data: Human–robot dialogue corpus

Since it was costly to collect a large number of training labels from the above experiment, we also used a dialogue corpus as unlabeled data. We have collected a human–robot dialogue corpus where a subject talked with an android robot that was remotely controlled by a human operator (Kawahara, 2018). The voice of the human operator was directly played through the robot's speaker so that the spoken behaviors could be natural. In this corpus, there are three kinds of dialogue tasks: speed-dating, job interview,

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

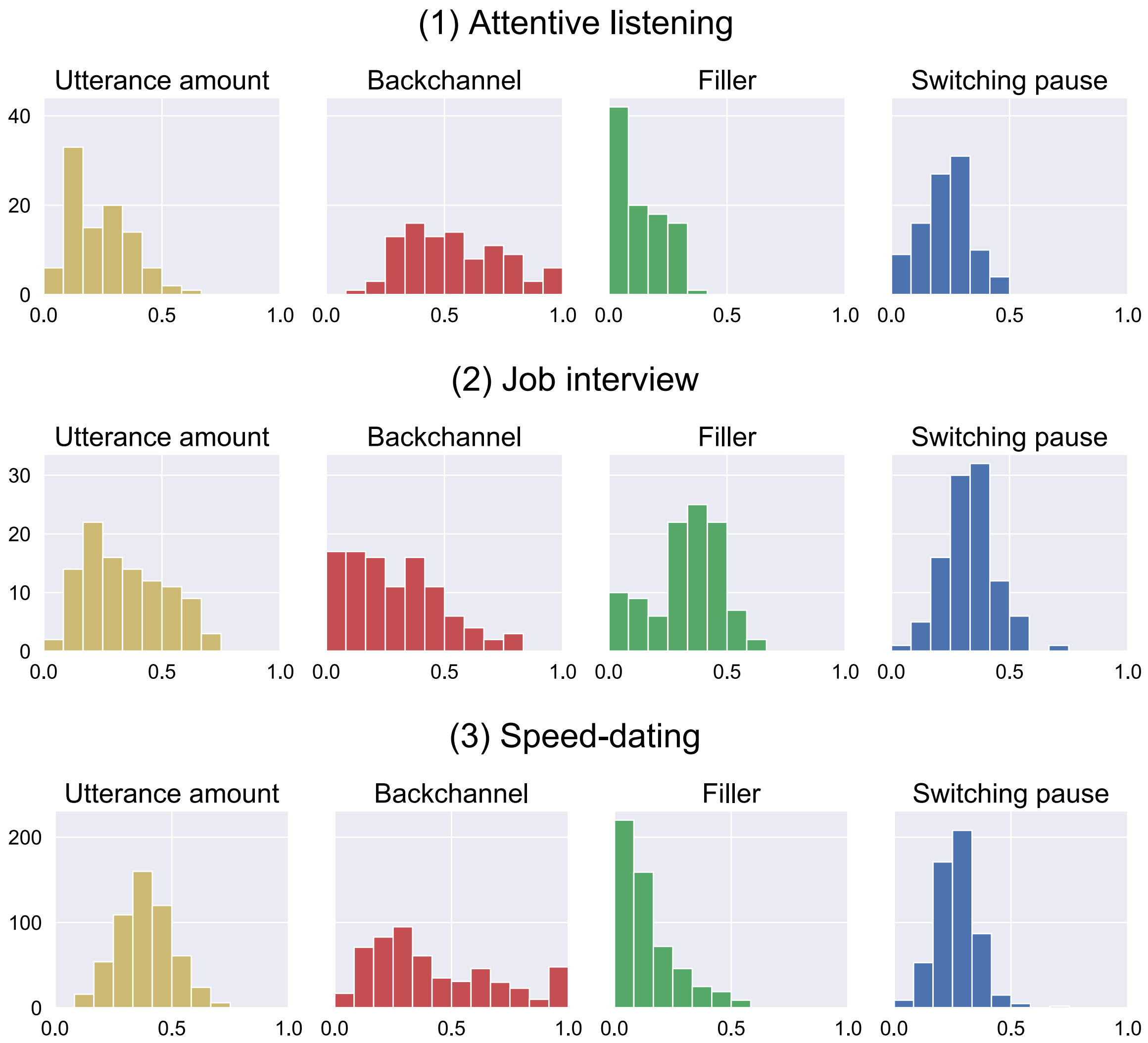


Fig. 2. Distribution of behaviors in each dialogue task (Histograms of each behavior).

and attentive listening. Each dialogue lasted about 10 min, and the number of dialogue sessions is 83, 30, and 19 for speed-dating, job interview, and attentive listening, respectively. The robot operators were four females in total, and one of them attended each session. In this study, we used the spoken dialogue behavior data of the robot operators to model the system's behavior.

We divided each dialogue session into two-minute segments. The segment length was set to 2 min that contained at least three utterance exchanges between the operator and the subject. For each segment, the four spoken dialogue behaviors were measured and also normalized to make values from 0 to 1 in the same way as we conducted on the labeled data (Table 1).

4.3. Analysis of behavioral tendency in accordance with dialogue task in corpus

The appropriateness of controlling the behavior of each task was examined by using the above dialogue corpus. We investigate how the spoken dialogue behaviors change in accordance with the different dialogue tasks in the corpus. 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 was different. For each two-minutes segment, 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. 2 (1)–(3). In the case of attentive listening (Fig. 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. 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. 2 (3)), the operators spoke many utterances with shorter switching pauses. Speed-dating

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

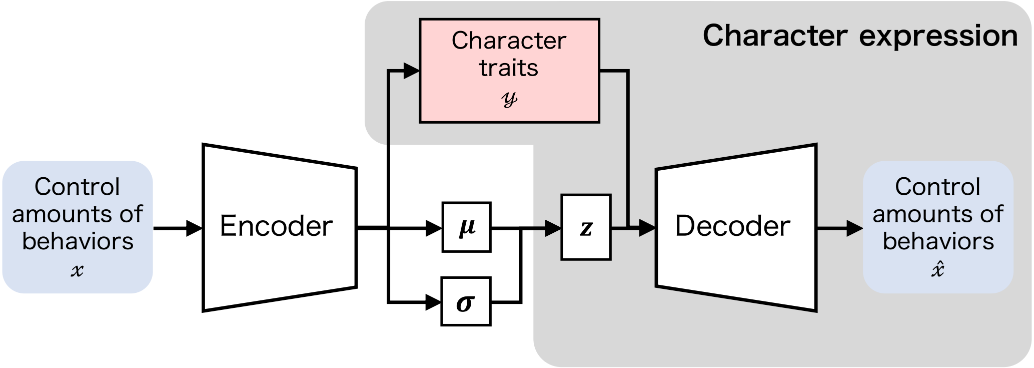


Fig. 3. Network architecture of the proposed model.

is a free and casual dialogue setting compared with the other tasks. The results so far 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. In a user experiment explained in Section 7, to confirm the effectiveness of the proposed character expression model in specific dialogue tasks, we will specify the proper character trait values according to the dialogue tasks.

5. Character expression model

We propose a character expression model with semi-supervised learning that uses both the impression evaluation data and the corpus data. 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.

5.1. Network architecture

First, we explain the architecture of the proposed model as depicted in Fig. 3. The model is based on a variational auto-encoder (VAE) (Kingma et al., 2014) 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 ($\boldsymbol{z}$). 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 $(2, 4, 8, 16, 32, \dots)$. 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 manner in each training epoch, as depicted in Fig. 4. Each batch data $\mathcal{D}$ is mixed by labeled data $\mathcal{D}_l$ and unlabeled data $\mathcal{D}_u$. The labeled data $\mathcal{D}_l$ are the impression evaluation data explained in Section 4.1. The unlabeled data $\mathcal{D}_u$ are the spoken behavior data from the dialogue corpus that does not contain any character trait data, as explained in Section 4.2.

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

$$\mathcal{L}_{enc} = \underset{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}_l}{\mathrm{CE}} (\mathrm{Enc}(\boldsymbol{x}), \boldsymbol{y}) \,, \tag{1}$$

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 $\mathcal{D}_l$ as

$$\mathcal{L}_{dec} = \underset{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}_l}{\mathrm{CE}} (\mathrm{Dec}(\boldsymbol{y}, \boldsymbol{z}), \boldsymbol{x}) \,, \tag{2}$$

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 $\mathcal{N}(\boldsymbol{O}, \boldsymbol{I})$. When we use the decoder part (character expression) only, the latent variables $\boldsymbol{z}$ are randomly sampled from the standard normal distribution.

A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp

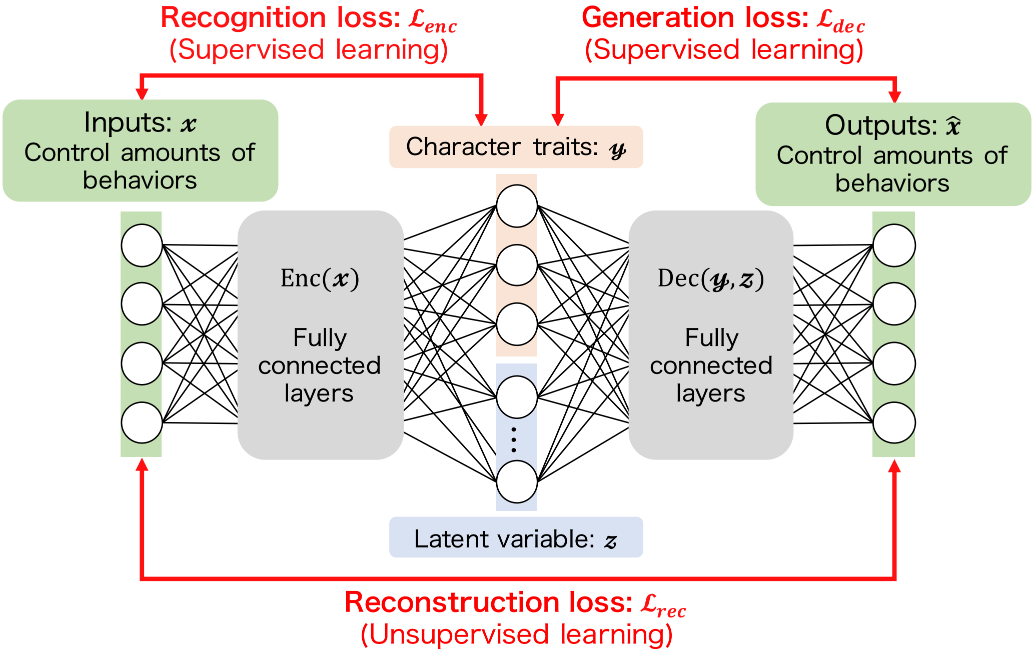


Fig. 4. Semi-supervised learning with character impression data and corpus data.

In the unsupervised learning using the unlabeled data 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

$$\mathcal{L}_{rec} = \underset{\boldsymbol{x} \in D_u}{\mathrm{CE}} \left(\mathrm{Dec}(\mathrm{Enc}(\boldsymbol{x})), \boldsymbol{x}\right) - \mathrm{D}_{KL}[\boldsymbol{z} \parallel \mathcal{N}(\boldsymbol{O}, \boldsymbol{I})] \,, \tag{3}$$

where D_{KL} represents Kullback–Leibler divergence.

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

$$\mathcal{L} = (1 - \lambda)(\mathcal{L}_{enc} + \mathcal{L}_{dec}) + \lambda \mathcal{L}_{rec} \,. \tag{4}$$

[2] Note that we empirically tune the weight parameter λ as 0.8. 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 focus on the unlabeled data. This parameter value was determined derived from that the training process was stable in a preliminary experiment.

5.2. 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 as below, 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 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.

[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.
[2] In the previous report (Yamamoto et al., 2020), we optimized separately $\mathcal{L}_{enc}$, $\mathcal{L}_{dec}$ and $\mathcal{L}_{rec}$.

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

Table 3
Mean absolute errors 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 1.)

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

[a] $< .01$.

6. 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.

6.1. *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 4.2. 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 4.1. 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 3 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 backchannel. We also investigated the difference in actual behavior features (in 2 min. segments), which was calculated on the basis of Table 1. The proposed model controlled more accurately than the baseline by 11.16 s (in 2 min.), 10.44 times (in 2 min.), and 0.55 s 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.[3] 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. 6. It was shown that the addition of the behavior data was effective even when only a small amount of data is used.

6.2. *Evaluation of encoder*

We also evaluate the encoder of the proposed model in this section. 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 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 model trained using only the labeled data. Table 4 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

[3] The result is enhanced from the previous report (Yamamoto et al., 2020) because of the improved loss function.

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

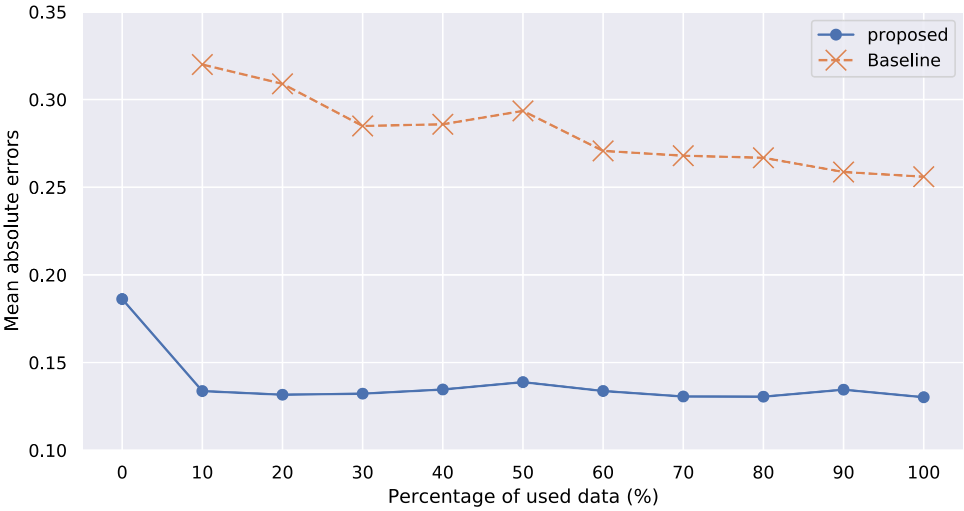


Fig. 5. Mean absolute errors (average of four behaviors) when the amount of labeled data used is varied.

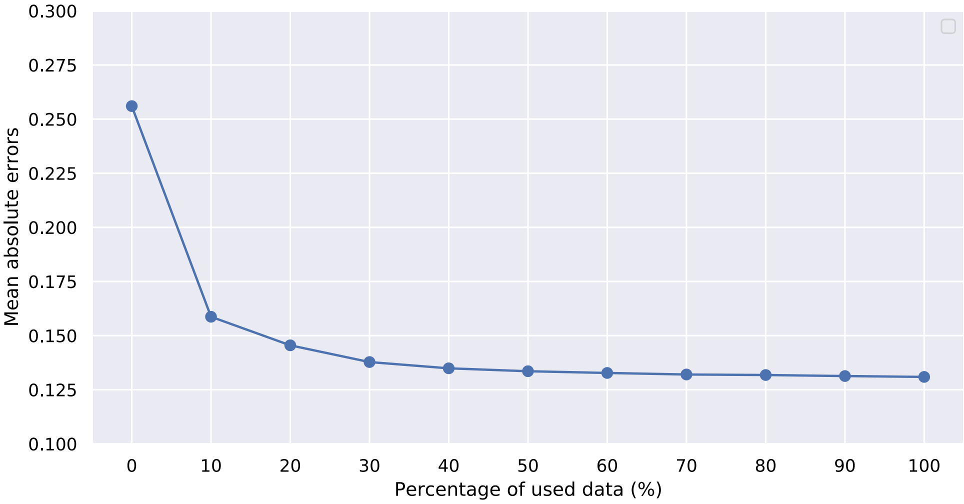


Fig. 6. Mean absolute errors (average of four behaviors) when the amount of unlabeled data used is varied.

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.

Additionally, we visualized the output of the encoder and analyzed the tendency of the characters in each dialogue task. If the distribution of recognized characters captures the characteristics of each dialogue task, we can conclude that the proposed model can recognize proper characters required in specific dialogue tasks. Histograms of recognized each character traits are shown in Fig. 7 (1)–(3). In attentive listening, more introvert and casual characters were observed. The purpose of this task for operators is to encourage the interlocutors to speak, so the introvert and casual character seems to be reasonable to make the dialogue more comfortable for talking more. In job interview, more polite characters observed, which is 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.

6.3. Modeling of unlabeled behavior

Finally, we evaluated an extension of the model by adding an unlabeled behavior as explained in Section 5.2. We used *speech rate* as an unlabeled behavior. Previous studies pointed out that speech rate behaviors affected the impression of extroversion (Uchida,

京都大学 KYOTO UNIVERSITY
A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp
京都大学学術情報リポジトリ KURENAI 紅 Kyoto University Research Information Repository

Table 4
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[a]
Emotional instability	0.128	0.183
Politeness	0.091	0.078
Average	0.142	0.138

[a] < .01.

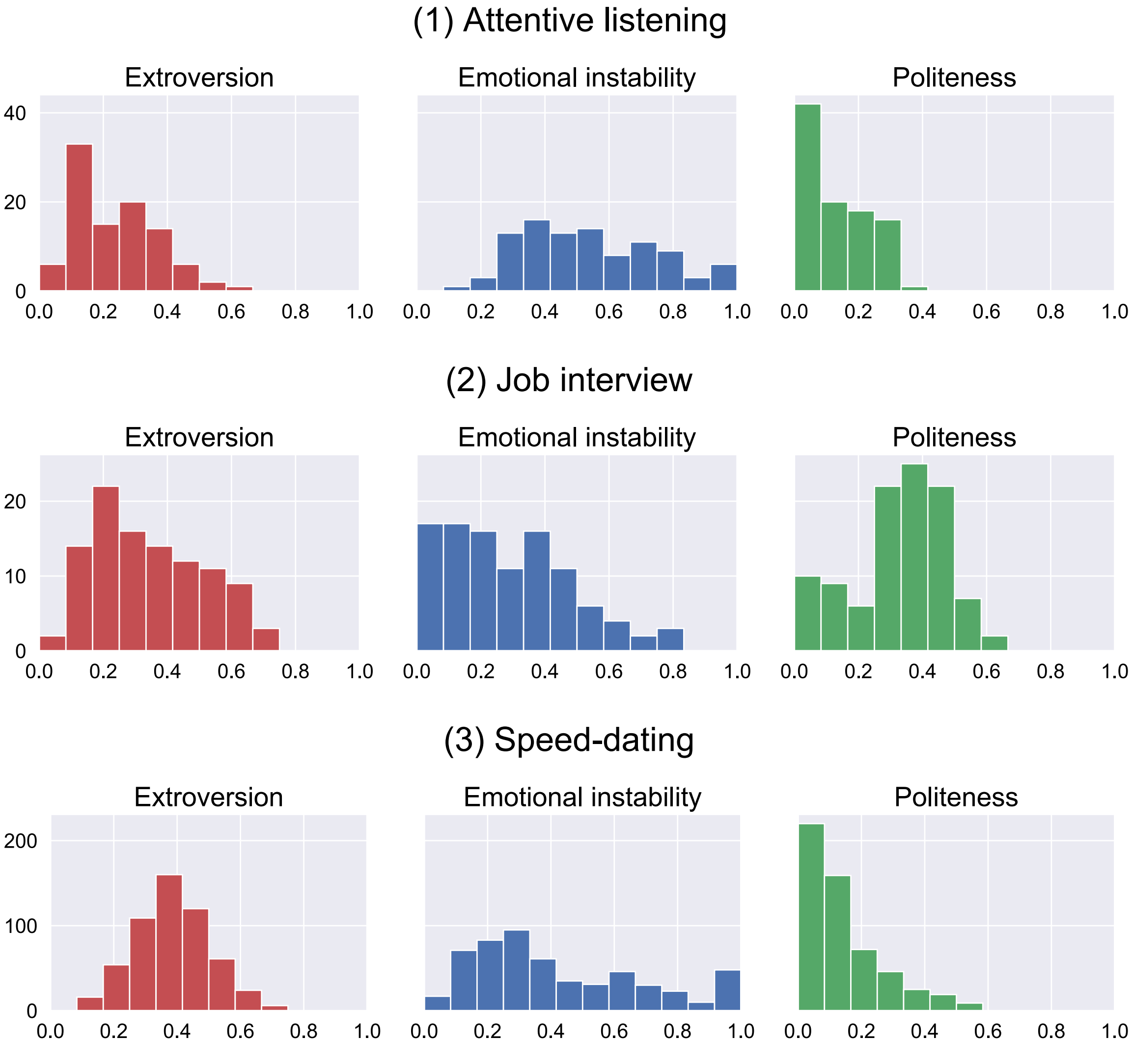


Fig. 7. Histograms of each estimated character trait by the encoder (character recognition model) in each dialogue task.

2002; Mairesse and Walker, 2006). 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 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.

A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp

Table 5
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./s)
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)

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 4.1. Ten dialogues were sampled from speed dating task introduced in Section 4.2. 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 can is shown that extroversion can be controlled by the speech rate.

7. Subjective evaluation

We implemented the character expression model in a spoken dialogue system of an autonomous android and then conducted a subjective experiment to confirm the effectiveness of character expression in specific dialogue tasks.

7.1. Implementation in spoken dialogue system

The proposed character expression model was implemented in a spoken dialogue system of an autonomous android ERICA (Kawahara, 2018). For the dialogue tasks in this experiment, we manually designed fixed dialogue flows. When we calculate the control amounts using the decoder model, the values of the latent variables were set by sampling from a standard normal distribution. Then, the output values were used as the control amounts for the behaviors of the robot. The control amount values corresponded to the 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 ms by using prosodic features of the user utterance and a logistic regression model (Lala et al., 2017). The control amount of backchannel frequency was regarded as the threshold of the 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 regarded as a probability value. Fillers are inserted stochastically in the candidate positions manually designated in the system utterances. The switching pause length corresponds to the length of silence until the robot takes a turn. The length of silence was determined by linearly normalizing the control amount to the range from 700 to 3000 ms.

7.2. Experimental setting

At first, we recorded dialogue videos where the first author and ERICA talked for 8 min in the two different dialogue scenarios of job interview and laboratory guide. In job interview, ERICA asked questions as the interviewer, and the first author responded as the interviewee. In the laboratory guide task, ERICA presented research topics of a laboratory and asked some questions to the visitor, and the first author listened to the explanation and answered questions as the visitor. For each task, we prepared two dialogue videos in two different conditions: proposed and baseline. In the proposed-method condition, the proposed character expression model was given a character that is regarded as appropriate for the corresponding dialogue task. In the baseline condition, ERICA spoke with an emotionally stable character. The characters represented by ERICA in each condition are summarized in Table 6. In the job interview, we set the character traits (polite) based on the results of the corpus analysis in Section 6.2. In the laboratory guide, we set the situation 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 seventy-five university students. For each dialogue task, the subject watched the dialogue video[4] in both the proposed and baseline conditions and then evaluated a pairwise comparison of which was more appropriate for the task. We prepared three different questions for each task as shown in Table 7 and questions about character trait scores described in Section 4.1.

[4] The videos can be seen at http://sap.ist.i.kyoto-u.ac.jp/members/yamamoto/csl

A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp

Table 6
Input character trait in each condition.

Conditions	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

Table 7
Results of impression evaluation in a dialogue task (relative evaluation).

Tasks	Items	Selection of subjects	
		Proposed	Baseline
Job interview	Which robot is settled?	47[a]	28
	Which robot listened more carefully to what you said?	40	35
	Which robot is better for a job interviewer?	47[a]	28
Laboratory guide	Which robot explained studies more actively?	36	39
	Which robot explained more clearly?	40	35
	Which robot is better as a lab guide?	39	36

[a] $p < .05$.

Table 8
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[a]	5.06
Emotional instability	2.33[a]	2.96	2.45	2.55
Politeness	5.96[a]	5.77	5.97[a]	5.76

[a] < 0.10.

7.3. Experimental results

The results are shown in Table 7. In job interview, two of the three question items showed significant differences (5%) in the one-tailed binomial test. For the remaining item, the 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 conditions. 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 people who are suitable with casual and interactive introduction such as colleagues, while polite presentation is desired for people who are in different social standings 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 8. 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 the extrovert traits. This suggests we need to be careful when changing two or three traits jointly especially when their behaviors are not consistent. The results of Table 7 and 8 suggest that the task evaluation scores are higher for tasks in which the characters are correctly recognized, such as job interview.

8. Conclusion

We have proposed a character expression model for spoken dialogue systems. 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 with semi-supervised learning 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, but the encoder can be interpreted as a character recognition model, so we verified 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.

A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp

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.

We implemented a proposed model in a spoken dialogue system that controls the behaviors appropriately. We evaluated this system 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 dialogues by expressing a given proper character in a formal job interview 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. In future work, we will construct a character expression model to take into account the correlation between character traits. In addition, to deal with tasks where the proper character depends on each user, we will extend the character expression model so that it can adapt to each user's character.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (JPMJER1401), Grant-in-Aid for Scientific Research on Innovative Areas "Communicative intelligent systems towards a human–machine symbiotic society" (JP19H05691) and Grant-in-Aid for JSPS Fellow, Japan (JP20J22284).

References

Cattell, R.B., 1956. Second-order personality factors in the questionnaire realm. J. Consult. Psychol. 20 (2), 411–418.

Costa, P.T., McCrae, R.R., 1992. Normal personality assessment in clinical practice: The NEO personality inventory. Psychol. Assess. 4 (1), 5–13.

Den, Y., Yoshida, N., Takanashi, K., Koiso, H., 2011. Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. In: Oriental COCOSDA. pp. 168–173.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Morency, L.P., 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In: AAMAS. pp. 1061–1068.

Digman, J.M., 1990. Personality structure: Emergence of the five-factor model. Annu. Rev. Psychol. 41 (1), 417–440.

Eysenck, H., 2017. The Biological Basis of Personality. Routledge.

Fong, T., Nourbakhsh, I., Dautenhahn, K., 2003. A survey of socially interactive robots. Robot. Auton. Syst. 42, 143–166.

Gill, A., Overlander, J., 2003. 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. pp. 456–461.

Goldberg, L.R., 1990. An alternative "description of personallity": The big-five factor structure. Pers. Soc. Psychol. 59 (6), 1216–1229.

Hathaway, S.R., McKinley, J.C., 1940. A multiphasic personality schedule (minnesota) : I. Construction of the schedule. J. Psychol. 10 (2), 249–254.

Higashinaka, R., Mizukami, M., Kawabata, H., Yamaguchi, E., Adachi, N., Tomita, J., 2018. Role play-based question-answering by real users for building chatbots with consistent personalities. In: SIGDIAL. pp. 264–272.

Kawahara, T., 2018. Spoken dialogue system for a human-like conversational robot ERICA. In: IWSDS.

Kingma, D.P., Rezende, D.J., Mohamed, S., Welling, M., 2014. Semi-supervised learning with deep generative models. In: NIPS.

Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K., Kawahara, T., 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In: SIGDIAL.

Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., Dolan, B., 2016. A persona-based neural conversation model. In: ACL. pp. 994–1003.

Mairesse, F., Walker, M.A., 2006. Automatic recognition of personality in conversation. In: NAACL. pp. 85–88.

Mairesse, F., Walker, M.A., 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. Comput. Linguist. 37 (3), 455–488.

Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K., 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artificial Intelligence Res. 30 (1), 457–500.

McCrae, R.R., John, O.P., 1992. An introduction to the five-factor model and its applications. J. Pers. 60 (2), 175–215.

McKeown, G., Valstar, M., Pantic, M., 2012. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans. Affect. Comput. 3 (1), 5–17.

Metcalf, K., Theobald, B.-J., Weinberg, G., Lee, R., Jonsson, I.-M., Webb, R., Apostoloff, N., 2019. Mirroring to build trust in digital assistants. In: INTERSPEECH. pp. 4000–4004.

Miyazaki, C., Hirano, T., Higashinaka, R., Makino, T., Matsuo, Y., 2015. Automatic conversion of sentence-end expressions for utterance characterization of dialogue systems. In: PACLIC. pp. 307–314.

Mizukami, M., Neubig, G., Sakti, S., Toda, T., Nakamura, S., 2015. Linguistic individuality transformation for spoken language. In: IWSDS.

Nass, K.I.A.N.D.C., 2000. Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. Hum.-Comput. Stud. 53 (2), 251–267.

Nass, C., Moon, Y., Fogg, B.J., Reeves, B., Dryer, D.C., 1955. Can computer personalities be human personalities? Hum.-Comput. Stud. 43, 223–239.

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

Ogawa, Y., Kikuchi, H., 2017. Assigning a personality to a spoken dialogue agent by behavior reporting. New Gener. Comput. 35 (1), 181–209.

Ogawa, Y., Miyazawa, K., Kikuchi, H., 2014. Assigning a personality to a spoken dialogue agent through self-disclosure of behavior. In: HAI. pp. 331–337.

A Self-archived copy in
Kyoto University Research Information Repository
https://repository.kulib.kyoto-u.ac.jp

Ozer, D.J., Benet-Martínez, V., 2006. Personality and the prediction of consequential outcomes. Annu. Rev. Psychol. 57 (1), 401–421.

Pfeifer, L.M., Bickmore, T., 2009. Should agents speak like, um, humans? The use of conversational fillers virtual agents. In: IVA. pp. 460–466.

Robert, L.P., Alahmad, R., Esterwood, C., Kim, S., You, S., Zhang, Q., 2020. A Review of Personality in Human-Robot Interaction. Now Foundations and Trends.

Sacks, H., Schegloff, E.A., Jefferson, G., 1978. A simplest systematics for the organization of turn taking for conversation. In: Studies in the Organization of Conversational Interaction. pp. 7–55.

Salem, M., Ziadee, M., Sakr, M., 2013. Effects of politeness and interaction context on perception and experience of HRI. In: ICSR. pp. 531–541.

Scherer, K.R., 1979. Scocial markers in speech. In: Scherer, K., Giles, H. (Eds.), Personality Markers in Speech. Cambridge University Press, pp. 147–209.

Serban, I.V., Lowe, R., Henderson, P., Charlin, L., Pineau, J., 2018. A survey of available corpora for building data-driven dialogue systems. Dialogue Discourse 9 (1), 1–49.

de Sevin, E., Hyniewska, S.J., Pelachaud, C., 2010. Influence of personality traits on backchannel selection. In: IVA. pp. 187–193.

Shiwa, T., Kanda, T., Imai, M., Ishiguro, H., Hagita, N., 2009. How quickly should communication robots respond? Int. J. Soc. Robot. 1, 153–160.

Shum, H.-Y., He, X., Li, D., 2018. From Eliza to XiaoIce: Challenges and opportunities with social chatbots. Front. Inf. Technol. Electron. Eng. 19 (1), 10–16.

Srinivasan, V., Takayama, L., 2016. Help me please: Robot politeness strategies for soliciting help from humans. In: CHI. pp. 4945–4955.

Sugiyama, H., Meguro, T., Higashinaka, R., Minami, Y., 2014. Large-scale collection and analysis of personal question-answer pairs for conversational agents. In: IVA. pp. 420–433.

Takanashi, K., Maruyama, T., Uchimoto, K., Isahara, H., 2003. Identification of "sentences" in spontaneous Japanese - detection and modification of clause boundaries. In: SSPR.

Traum, D., Aggarwal, P., Artstein, R., Foutz, S., amd Athanasios Katsamanis, J.G., Leuski, A., Noren, D., Swartout, W., 2012. Ada and Grace: Direct interaction with museum visitors. In: IVA. pp. 245–251.

Uchida, T., 2002. Effects of the speech rate on speakers' personality-trait impressions. Jpn. J. Psychol. 73 (2), 131–139, in Japanese.

Valente, F., Kim, S., Motlicek, P., 2012. Annotation and recognition of personality traits in spoken conversations from the AMI meetings corpus. In: INTERSPEECH. pp. 1183–1186.

Wada, S., 1996. Construction of the Big Five scales of personality trait terms and concurrent validity with NPI. Jpn. J. Psychol. 67 (1), 61–67 (in Japanese).

Ward, N., 2006. Non-lexical conversational sounds in American English. Pragmat. Cogn. 14 (1), 129–182.

Watanabe, M., 2009. Features and Roles of Filled Pauses in Speech Communication: A Corpus-Based Study of Spontaneous Speech. Hitsuji Syobo Publishing.

Weaver, J.B., 1998. Communication and personality: Trait perspectives. In: McCroksey, J.C., Daly, J.A., Martin, M.M., Beatty, M.J. (Eds.), Personality and Self-Perceptions About Communication. Hampton Press, pp. 95–118.

Yamamoto, K., Inoue, K., Kawahara, T., 2020. Semi-supervised learning for character expression of spoken dialogue systems. In: INTERSPEECH. pp. 4188–4192.

Yamamoto, K., Inoue, K., Nakamura, S., Takanashi, K., Kawahara, T., 2018. Dialogue behavior control model for expressing a character of humanoid robots. In: APSIPA ASC. pp. 1732–1737.

Yu, M., Gilmartin, E., Litman, D., 2019. Identifying personality traits using overlap dynamics in multiparty dialogue. In: INTERSPEECH. pp. 15–19.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J., 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In: ACL. pp. 2204–2213.