INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Semi-supervised learning for character expression of spoken dialogue systems

Kenta Yamamoto, Koji Inoue, and Tatsuya Kawahara

School of Informatics, Kyoto University, Kyoto, Japan
{yamamoto, inoue, kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

We address character expression for spoken dialogue systems (e.g. extrovert). While conventional studies focused on controlling linguistic expressions, we focus on spoken dialogue behaviors. Specifically, the proposed model maps three character traits: extroversion, emotional instability, and politeness to four spoken dialogue behaviors: utterance amount, backchannel, filler, and switching pause length. It is costly to collect annotated data for training this kind of models. Therefore, we propose a semi-supervised learning approach to utilize not only a character impression data (labeled data) but also a corpus data (unlabeled data). Experimental results show that the proposed model expresses the target character traits through the behaviors more precisely than a baseline model that corresponds to the case of supervised learning only. Besides, we also investigate how to model unlabeled behavior (e.g. speech rate) by utilizing the advantage of semi-supervised learning.

Index Terms: Spoken dialogue system, character, personality, dialogue behavior, backchannel, filler

1. Introduction

It is desired for spoken dialogue systems, in particular, social robots and virtual agents, to express their characters (e.g. extrovert) for human-like interaction [1, 2]. In practical spoken dialogue systems, specific social roles are given such as a psychological counselor [3], a museum guide [4], and an attentive listener [5]. To make dialogue in these social scenarios more natural, it is important to assign proper characters to the systems. For example, museum guide systems are expected to be extrovert and intelligent, and counseling systems are expected to be introvert and emotionally stable. Earlier user experiments have revealed that the character expression of spoken dialogue systems led to increasing user engagement and the naturalness of the dialogue [6, 7, 8].

In our study, we focus on spoken dialogue behaviors that have not yet been studied well in character expression. Previous studies have addressed character expression models controlling the linguistic pattern of system utterances [9, 10, 11, 12]. Therefore, data for character expression has been collected as the form of text dialogue [13, 14, 15, 16]. However, in spoken dialogue, besides the above-mentioned text style, spoken dialogue behaviors should be considered. We propose a character expression model that controls four spoken dialogue behaviors: utterance amount, backchannel, filler, and switching pause length.

In our previous work [17, 18], character expression models were trained by supervised learning with manually-annotated labels obtained through impression evaluation. However, the manual annotation can be both costly and time-consuming, and therefore the variation of behavior patterns has to be limited. As a result, supervised training with the limited training data may fall in over-fitting and then lead to unnatural behavior control.

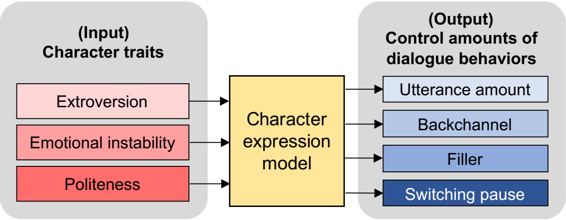

Figure 1: *Problem formulation of character expression*

In this paper, we propose a character expression model based on variational auto-encoder (VAE) [19] with semi-supervised learning to utilize not only manually-annotated labels (supervised) but also dialogue corpus data (unsupervised). It is expected that the proposed model compensates the above-mentioned data-sparseness by semi-supervised learning with natural dialogue behavior data. Another advantage of semi-supervised learning is to be able to train an expression of additional dialogue behavior (e.g. speech rate) that is not annotated at all in the labeled data but can be measured in the labeled data.

The aim of this study is to realize the character expression model, which maps target character traits to natural dialogue behaviors, by introducing semi-supervised learning. Utilization of dialogue corpus data as unlabeled data can be applied to other expression tasks (e.g. emotion expression through dialogue behaviors), which are affected by data-sparseness due to the limited training data.

In Section 2, we define character traits (input) and dialogue behaviors (output) used in this study. Training data including labeled and unlabeled ones are explained in Section 3. The proposed semi-supervised learning is explained in Section 4 and also evaluated in Section 5.

2. Character traits and spoken dialogue behaviors

At first, we address the problem setting as shown in Figure 1. The input of the character behavior model is the set of character traits. In our study, we use three character traits: extroversion (extrovert vs. introvert), emotional instability (stable vs. instable), and politeness (polite vs. casual). Extroversion and emotional instability are selected from the Big Five scale [20, 21, 22]. In previous studies, the Big Five traits have been used to define the personality of dialogue systems [9, 23]. Since using all the five traits requires a larger amount of training data and also makes a model complicated, we use the two traits: extroversion and emotional instability, in this study. Extroversion is expected to be the major factor that determines the impression on systems' characters [24]. However, if we use only extroversion in our character control model, the system could unintentionally behave as emotional instable. To avoid this, emotional instability is used in our model explicitly. In addition, politeness is also adopted in our model so that the system

http://dx.doi.org/10.21437/Interspeech.2020-2293

	Control amount	
	0.0	1.0
Utterance amount	speak not at all	speak all time
Backchannel	no backchannel	at all user pauses
Filler	no filler	at all system pauses
Switching pause	-0.5 sec. (overlap)	3.0 sec.

Table 1: *Correspondence between control amount and actual behavior features*

would be able to control its intimacy towards a dialogue partner [25]. For example, it is expected that the system behaves politely in formal situations, on the other hand, the system behaves casually with intimate (familiar) users.

The output of the character behavior model is the set of control amounts of spoken behaviors. We focus on spoken dialogue behaviors that are not observed in text-based dialogue as listed in Table 1. They are utterance amount, backchannel frequency, filler frequency, and switching pause length. Previous studies suggested that these behaviors affected the impression of dialogue partners [26, 23, 27, 28, 29, 30]. The 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 [31, 32]. In this study, the behavior of backchannel corresponds to 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 [33, 34]. The behavior of filler also corresponds to the frequency of uttered fillers. 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 behavior model controls these four spoken behaviors according to the input of the three character traits.

Since the model output is the set of control amounts of behaviors which are normalized from 0 to 1, their values need to be converted to the actual behavior features (e.g. how many backchannels uttered). We define the correspondence between them as reported in Table 1. In this study, we use this correspondence when we make the dataset from an impression evaluation data and a dialogue corpus that are described in the next section. We convert these control amounts to actual behavior values by linear interpolation based on this correspondence. For example, if the control amount of backchannel is 1, the system would generate backchannels at all user pauses.

3. Training data

We explain the labeled and unlabeled data used for semi-supervised learning for the character behavior model, respectively. The labeled data is obtained from a character impression evaluation (manual annotation), and the unlabeled data is derived from a human-robot dialogue corpus.

3.1. Labeled data: Character impression evaluation

To collect supervised training data, we conducted an experiment of impression evaluation on the character traits. In this experiment, each subject was asked to listen to speech samples and then to evaluate his/her impression on the three character traits of the speaker (7-point scale). For extroversion and emotional instability, we used 8 adjectives (4 for each) from a short version of Big Five scale [35]. We also used two adjectives, *polite* and *courteous*, for the third trait politeness. 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.

The speech samples were generated as follows. In advance, we selected two dialogue scenarios from our human-robot dialogue corpus described in Section 3.2. Based on each scenario, we artificially generated several speech samples by controlling dialogue behaviors observed in them. The robot utterances were controlled by text-to-speech software. At first, we generated a standard speech sample where backchannel and filler tokens are kept as the original dialogue and the switching pause length is set to 0.5 seconds. From the standard sample, we changed each dialogue behavior one by one. We used these generated speech samples to compare the perceived character traits between different conditions on each dialogue behavior (e.g. high backchannel frequency vs. low backchannel frequency). The detail of this sample generation (and also the analysis result) are found in our previous work [17].

In this study, we use the character trait scores obtained through this impression evaluation as labeled data. The number of available samples was 734, and they are divided into 662 samples for training and 74 samples for testing. Each sample corresponds 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.

3.2. Unlabeled data: A human-robot dialogue corpus

Since the number of training labels from the above-mentioned labeled data is limited, we also use a dialogue corpus as unlabeled data. We have collected a human-robot dialogue corpus where a subject talked with a humanoid robot that was controlled by a human operator remotely [36]. The voice of the human operator was directly played through the robot's speaker so that their spoken behaviors can be natural. In this corpus, there are three kinds of dialogue tasks: speed-dating, job interview, and attentive listening. Each dialogue lasted about 10 minutes and the numbers of dialogue sessions were 83, 30, 19 in speed-dating, job interview, and attentive listening, respectively. The robot operators were four females in total where one of them attended each session. In this study, we use the spoken behavior data of the robot operators to model the system's behavior. Since there were several robot operators and also several dialogue tasks in this corpus, it is expected that both the character and the spoken behaviors varied widely and naturally. We made manual annotation of the spoken behaviors.

We segmented each dialogue session by two minutes as one sample. This segment length was empirically determined to make enough amount of the spoken behaviors observed. For each segment, the four spoken behaviors were measured and also normalized to make the values from 0 to 1 in the same way as we conducted on the labeled data.

4. Character expression model

We propose a character behavior model trained by semi-supervised learning to utilize both the impression evaluation data and the dialogue corpus data simultaneously. Using the impression evaluation data, the proposed model acquires the relationship between the input character and the output spoken behaviors (supervised learning). Besides, using the dialogue corpus data that does not contain the character trait labels, the proposed model makes itself represent natural patterns of the behaviors (unsupervised learning).

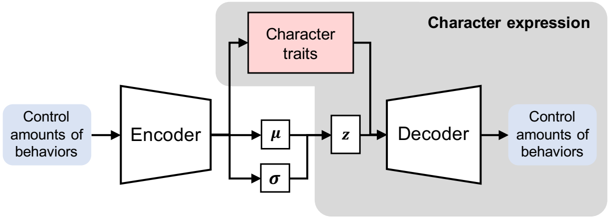

Figure 2: *Network architecture of the proposed model*

4.1. Network architecture

At first, we explain the architecture of the proposed model as depicted in Figure 2. The model is based on a variational auto-encoder (VAE) [19] consisting of an encoder and a decoder. The encoder and the decoder correspond to character recognition (behavior to character) and character representation (character to behavior), respectively. The input for the encoder is a 4-dimensional vector of the spoken behaviors normalized between 0 and 1. The encoder outputs a 3-dimensional vector of the character traits normalized between 0 and 1 and also parameters (means ($\boldsymbol{\mu}$) and variances ($\boldsymbol{\sigma}$)) to generate latent variables ($\boldsymbol{z}$). The dimension of the latent variables is set to 8, which was determined through our preliminary experiment. The latent variables are expected to capture other factors, other than the three character traits (e.g. dialogue task and context). In our preliminary experiment, we also tested an auto-encoder which does not have the latent variables and observed that the accuracy of character expression was improved by adding the latent variables. 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 4-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 as the activation function of the output layer.

The main task of this study (character expression) corresponds to the decoder. When we use the decoder part only, the 8-dimensional latent variables are randomly sampled from the standard normal distribution. When we train this VAE-based model, supervised and unsupervised learning (explained below) are applied alternately in each training epoch, as depicted in Figure 3.

4.2. Step 1: Supervised learning with character impression

In the supervised learning step, the encoder and decoder are trained separately using the impression evaluation data explained in Section 3.1. In the training of the encoder, the behavior values of each speech sample are fed as input, and the score of the character impression evaluation is predicted. The mean square error is then propagated through the encoder. Next, we train the decoder in the opposite way of the encoder. The decoder is fed the score of the character impression evaluation and then predicts the behavior values of each speech sample. The mean square error is back-propagated through the decoder.

4.3. Step 2: Unsupervised learning with dialogue corpus

In the unsupervised learning step, we use only the spoken behavior data from the dialogue corpus that does not contain a character trait data, as explained in Section 3.2. The behavior data is fed to the encoder and the output of the encoder is also fed to the decoder to predict the original input behavior data. The mean square error is then back-propagated through the whole network including the encoder and the decoder. The KL divergence is also added to the loss function so that the 8-dimensional latent variables follow the standard normal distribution.

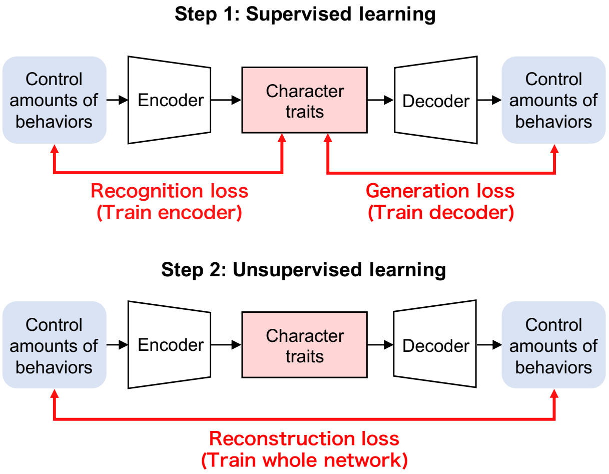
Step 1: Supervised learning
Control amounts of behaviors
Encoder
Character traits
Decoder
Control amounts of behaviors
Recognition loss (Train encoder)
Generation loss (Train decoder)
Step 2: Unsupervised learning
Control amounts of behaviors
Encoder
Character traits
Decoder
Control amounts of behaviors
Reconstruction loss (Train whole network)

Figure 3: *Semi-supervised learning with character impression data (step 1) and corpus data (step2) (Latent variables of VAE are omitted.)*

4.4. Model extension: Controlling unlabeled behavior

Another advantage of the proposed model is that it can handle unlabeled behaviors owing to 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 labeled with the character impression at all, but it can be observed in the dialogue corpus.

The behavior data (input of the encoder and output of the decoder) is extended to a 5-dimensional vector: 4 dimensions for the existing behaviors and the other is for the new behavior. In the first step (supervised learning), the training data of the fifth dimension is set to neutral (0.5), and errors are not defined. In the second step (unsupervised learning), since we use only the corpus data and consider the new behavior. The error of all the behaviors is back-propagated. In this way, it is expected that the model acquires natural parameters on the new behavior considering the relationship between the existing four behaviors and the fifth. In other words, the fifth behavior is controlled in conjunction with the four behaviors.

5. Experiment at evaluations

We evaluate the effectiveness of semi-supervised learning with a test dataset. Besides, we also investigate how much the model can handle an additional unlabeled behavior (speech rate).

5.1. Effectiveness of semi-supervised learning

The proposed model is compared with a baseline model consisting of only the decoder part of the VAE. The structure of the baseline model is the same as the decoder of the proposed model, except that the latent variables are not used. The baseline model is designed via supervised learning so it is trained with only the impression evaluation data. Therefore, this comparison reveals the effectiveness of semi-supervised learning in the current character expression task.

To conduct the evaluation of these models, we prepared

Behavior	Baseline	Proposed	(behavior diff.)
Utterance amount	0.221	0.126 *	9.67 sec.
Backchannel	0.243	0.283	1.96 times
Filler	0.326	0.137 **	9.30 times
Switching pause	0.234	0.108 **	0.44 sec.
Average	0.256	0.162 **	

(*< .05, **< .01)

Table 2: *Mean absolute errors between control amounts of behaviors output from models and the oracle data (behavior diff. represents the difference on the level of actual behavior features (in 2 min. segment) that are calculated based on Table 1.)*

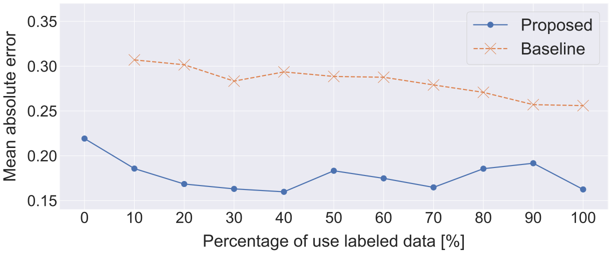

Figure 4: *Mean absolute errors (average among four behaviors) when the amount of used labeled data is varied*

a test data set using the corpus data. Specifically, we conducted another impression evaluation experiment for a subset of the natural dialogue corpus data. At first, we extracted 30 audio samples from the dialogue corpus described in Section 3.2. Note that these samples were not used in the model training. We asked 5 subjects (2 females and 3 males) to listen to the audio samples and then evaluate the character traits of the robot operator by the same items as the impression evaluation introduced in Section 3.1. We then obtained the input character data by averaging the evaluated scores among the subjects. The oracle output data corresponds to dialogue behavior data measured in each audio sample. The evaluation metric is the mean absolute error between the output of each model and the oracle data.

Table 2 reports the absolute errors of the models on each behavior and those average. We conducted a t-test between the models and confirmed that the proposed model significantly improved all the scores except for backchannel. We also confirmed the difference on actual behavior features (in 2 min. segment) that are calculated based on Table 1. The proposed model controls more accurately than the baseline by 9.67 seconds (in 2 min.), 9.30 times (in 2 min.), and 0.44 seconds for utterance amount, the number of filler, and switching pause length, respectively. We also investigated the case where the amount of used labeled data is varied, as reported in Figure 4. Whereas the error of the baseline model increases as reducing the amount of used labeled data, the proposed model relatively keeps the error until the ratio of the labeled data is about 20%. This result suggests that the proposed model interpolates the sparse distribution of the labeled data by utilizing the unlabeled data, which are natural dialogue behavior data from the corpus.

5.2. Qualitative analysis on modeling of unlabeled behavior

We also evaluated the model extension by adding an unlabeled behavior as explained in Section 4.4. In this experiment, we use speech rate as an unlabeled behavior. Previous studies pointed out that behaviors of speech rate affected the impression of extroversion [37, 38]. Here, speech rate is calculated by dividing the total number of characters of the operator utterances by the total duration of the utterances. The calculated speech rate 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).

Character traits (Input)				Speech rate	
Extroversion		Politeness		(char./sec)	
0	(introvert)	0	(casual)	0.527	(7.16)
0	(introvert)	1	(polite)	0.239	(5.43)
1	(extrovert)	0	(casual)	0.773	(8.64)
1	(extrovert)	1	(polite)	0.467	(6.80)
0.5	(neutral)	0.5	(neutral)	0.480	(6.88)

Table 3: *Example of control amounts by the proposed model when an additional unlabeled behavior (speech rate) is added (Emotional instability is fixed at stable.)*

Since we cannot apply the current task to the baseline model, we qualitatively analyzed the outputs of the proposed model. Table 3 reports the model outputs with the representative patterns of the character traits. The character trait patterns were combinations of extrovert/introvert and polite/casual. We also tried the neutral pattern (the bottom line). Emotional instability was fixed as stable. From the table, it is observed that the more extrovert, the system speaks faster. The more polite, the system speaks slowly. The result suggests that the proposed model is capable of acquiring the intuitive mapping from the unlabeled data.

6. Conclusion

We have proposed the character expression model that maps from the three character traits to the control amounts of the four spoken behaviors. The proposed model is based on variational auto-encoder with semi-supervised learning to utilize not only the impression evaluation data but also the corpus data that does not contain any character labels. This approach allows the model to compensate for natural behavior patterns that are lacking in the impression evaluation data. The experimental result shows that the proposed model expresses the target character traits through the behaviors more precisely than the baseline supervised learning.

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

We are now implementing this character expression model in the spoken dialogue system of the android ERICA. In future work, we will conduct a user experiment to confirm the effectiveness of the character expression through real dialogue.

7. Acknowledgement

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 (JP20J22284).

8. References

[1] T. Fong, I. Nourbakhsh, and K. Dautenhahn, “A survey of socially interactive robots,” *Robotics and Autonomous Systems*, vol. 42, pp. 143–166, 2003.

[2] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, “A survey of available corpora for building data-driven dialogue systems,” *Dialogue and Discourse*, vol. 9, no. 1, pp. 1–49, 2018.

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

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

[5] G. McKeown, M. Valstar, and M. Pantic, “The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent,” *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5–17, 2012.

[6] C. Nass, Y. Moon, B.J.Fogg, B. Reeves, and D. Dryer, “Can computer personalities be human personalities?” *Human-Computer studies*, vol. 43, pp. 223–239, 1955.

[7] K. Isbister and C. Nass, “Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics,” *Human-Computer Studies*, vol. 53, no. 2, pp. 251–267, 2000.

[8] M. Salem, M. Ziadee, and M. Sakr, “Effects of politeness and interaction context on perception and experience of HRI,” in *ICSR*, 2013, pp. 531–541.

[9] F. Mairesse and M. A. Walker, “Controlling user perceptions of linguistic style: Trainable generation of personality traits,” *Computational Linguistics*, vol. 37, no. 3, pp. 455–488, 2011.

[10] Y. Ogawa, K. Miyazawa, and H. Kikuchi, “Assigning a personality to a spoken dialogue agent through self-disclosure of behavior,” in *HAI*, 2014, pp. 331–337.

[11] C. Miyazaki, T. Hirano, R. Higashinaka, T. Makino, and Y. Matsuo, “Automatic conversion of sentence-end expressions for utterance characterization of dialogue systems,” in *PACLIC*, 2015, pp. 307–314.

[12] M. Mizukami, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, “Linguistic individuality transformation for spoken language,” in *IWSDS*, 2015.

[13] H. Sugiyama, T. Meguro, R. Higashinaka, and Y. Minami, “Large-scale collection and analysis of personal question-answer pairs for conversational agents,” in *IVA*, 2014, pp. 420–433.

[14] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan, “A persona-based neural conversation model,” in *ACL*, 2016, pp. 994–1003.

[15] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, “Personalizing dialogue agents: I have a dog, do you have pets too?” in *ACL*, 2018, pp. 2204–2213.

[16] R. Higashinaka, M. Mizukami, H. Kawabata, E. Yamaguchi, N. Adachi, and J. Tomita, “Role play-based question-answering by real users for building chatbots with consistent personalities,” in *SIGDIAL*, 2018, pp. 264–272.

[17] K. Yamamoto, K. Inoue, S. Nakamura, K. Takanashi, and T. Kawahara, “Dialogue behavior control model for expressing a character of humanoid robots,” in *APSIPA ASC*, 2018, pp. 1732–1737.

[18] ——, “A character expression model affecting spoken dialogue behaviors,” in *IWSDS*, 2020.

[19] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, “Semi-supervised learning with deep generative models,” *Advances in Neural Information Processing Systems*, 2014.

[20] J. M. Digman, “Personality structure: Emergence of the five-factor model,” *Annual review of psychology*, vol. 41, no. 1, pp. 417–440, 1990.

[21] R. R. McCrae and O. P. John, “An introduction to the five-factor model and its applications,” *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.

[22] P. T. Costa and R. R. McCrae, “Normal personality assessment in clinical practice: The NEO personality inventory,” *Psychological Assessment*, vol. 4, no. 1, pp. 5–13, 1992.

[23] E. D. Sevin, S. J. Hyniewska, and C. Pelachaud, “Influence of personality traits on backchannel selection,” in *IVA*, 2010, pp. 187–193.

[24] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, “Using linguistic cues for the automatic recognition of personality in conversation and text,” *Journal of Artificial Intelligence Research*, vol. 30, no. 1, pp. 457–500, 2007.

[25] V. Srinivasan and L. Takayama, “Help me please: Robot politeness strategies for soliciting help from humans,” in *CHI*, 2016, pp. 4945–4955.

[26] F. Valente, S. Kim, and P. Motlicek, “Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus,” in *INTERSPEECH*, 2012, pp. 1183–1186.

[27] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, “How quickly should communication robots respond?” *International Journal of Social Robotics*, vol. 1, pp. 153–160, 2009.

[28] L. M. Pfeifer and T. Bickmore, “Should agents speak like, um, humans? the use of conversational fillers virtual agents,” in *IVA*, 2009.

[29] M. Yu, E. Gilmartin, and D. Litman, “Identifying personality traits using overlap dynamics in multiparty dialogue,” in *INTERSPEECH*, 2019, pp. 15–19.

[30] K. Metcalf, B.-J. Theobald, G. Weinberg, R. Lee, I.-M. Jonsson, R. Webb, and N. Apostoloff, “Mirroring to build trust in digital assistants,” in *INTERSPEECH*, 2019, pp. 4000–4004.

[31] N. Ward, “Non-lexical conversational sounds in American English,” *Pragmatics & Cognition*, vol. 14, no. 1, pp. 129–182, 2006.

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

[33] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn taking for conversation,” in *Studies in the organization of conversational interaction*, 1978, pp. 7–55.

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

[35] S. Wada, “Construction of the Big Five scales of personality trait terms and concurrent validity with NPI,” *Japanese Journal of Psychology*, vol. 67, no. 1, pp. 61–67, 1996, in Japanese.

[36] T. Kawahara, “Spoken dialogue system for a human-like conversational robot ERICA,” in *IWSDS*, 2018.

[37] F. Mairesse and M. A. Walker, “Automatic recognition of personality in conversation,” in *NAACL*, 2006, pp. 85—-88.

[38] T. Uchida, “Effects of the speech rate on speakers’ personality-trait impressions,” *Japanese Journal of Psychology*, vol. 73, no. 2, pp. 131–139, 2002, in Japanese.