Spoken Dialog Training System for Customer Service Improvement

Yuta Sano*, Chee Siang Leow*, Soichiro Iida†, Takehito Utsuro†, Junichi Hoshino †,
Akio Kobayashi ‡ and Hiromitsu Nishizaki*
* University of Yamanashi, Kofu, Japan
E-mail: {sanoyuta,cheesiang_leow}@alps-lab.org, hnishi@yamanashi.ac.jp, Tel/Fax: +81-55-220-8361
† University of Tsukuba, Tsukuba, Japan
E-mail: iida.soichiro.sy@alumni.tsukuba.ac.jp, utsuro@iit.tsukuba.ac.jp, jhoshino@esys.tsukuba.ac.jp
‡ Tsukuba University of Technology, Tsukuba, Japan
E-mail: a-kobayashi@a.tsukuba-tech.ac.jp

***Abstract*—In the hospitality industry, including convenience stores and airport service counters, operational staff must be trained to serve customers satisfactorily and to avoid problems with them. This study investigates a spoken dialog training system for improving customer service by operational staff. In a conventional spoken dialog system, a system user uses the dialog system as a customer, and the dialog agent assists the system user in fulfilling his or her requirements. In our system, in contrast, the dialog agent plays the role of the customer. Consequently, the behaviors of the human and the customer are opposite to those in a traditional dialog system, and there has thus far been no research on such systems. This paper introduces a prototype of such a system that we have developed and describes a simple experiment. The results of the experiment confirm the usefulness of our system for hospitality training.**

I. Introduction

The Japanese word "Omotenashi" refers to the highest possible level of customer service, and many foreigners who visit Japan say they are impressed by Japanese hospitality. Japan provides the best customer service of any country. Even convenience stores and supermarkets, not just high-class restaurants and hotels, emphasize the spirit of hospitality. In fact, in Japan and other countries, customer satisfaction tends to be higher in stores with excellent customer service [1]. In contrast, most customers leave a shop or restaurant when they encounter bad service, and they tend to return to a shop or restaurant when they receive good service [2]. The level of customer service is very important in business.

Consequently, in the hospitality industry, including convenience stores and service counters, operational staff must be required to obtain good customer service skills, such as smooth customer service, customer satisfaction, and avoidance of problems with customers. For this reason, people newly employed in the service industry often receive customer service training.

Conventional customer service training involves learning by reading a customer service manual [3] or by having an instructor in on-the-job training (OJT) [4]. Operational staff trained using a manual, might not be able to respond suitably to situations that are not discussed in the manual. In addition, a manual cannot reproduce a real working environment. In addition, OJT involves high cost and time expenditure for acquiring instructors (trainers).

To solve these problems, this paper proposes a spoken dialog training system that does not require human instructors and can be used exclusively by trainees. A virtual agent-based interview training system [5] has been proposed for use as a computer-assisted training system. In addition, a debate training system [6] that uses automatic speech recognition and gestures has been studied as a training system similar to our customer service training system. However, even though this debate training system and our customer service training system are similar in that both systems use dialog management modules, the two systems' designs differ fundamentally. For example, our training system introduces dynamic dialog management based on scenario data for flexible response to various situations.

It can be expected that a training system based on speech interaction with a virtual agent would be very effective in improving customer service. In the hospitality industry, it is necessary to increase customer satisfaction or to solve their problems through conversation. Thus, dialog-based training is essential. The system we developed is easy to handle and can be used for training not only in assuring customer hospitality but also in handling customer complaints. The hospitality industry covers a wide range of industries, including retail outlets such as convenience stores and restaurants. A system that provides customer service training for virtual customers in many situations in a virtual space would enable trainees to train readily and improve their customer service.

Many spoken dialog systems have been studied including, for example, purpose-oriented speech dialog systems [7], such as a ticketing or reservation systems [8], [9], information providing speech dialog systems such as sightseeing guides [10], and chat speech dialog systems [11] without a fixed topic, such as Apple Siri [12]. Previous purposive and informative spoken dialog systems have included a beneficiary, i.e., a person who issues a request to the virtual agent in the dialog system. The virtual agent assists in satisfying human requirements. However, in the spoken dialog system developed in this paper, the roles of the virtual agent and the human using

the system are opposite to those of conventional speech dialog systems. In other words, the virtual agent is a "customer," and the human (trainee) converses with the virtual agent to satisfy the customer's requirements. The trainee learns the correct manner in which to serve customers by using appropriate speech in response to the various behaviors and actions of virtual customers. This reversal of the roles of virtual agent and trainee differs substantially from conventional speech dialog systems, and there has been no research on such systems.

Our goal is to develop a spoken dialog system for hospitality training. This paper reports on the development of the first prototype of such a system and experimental confirmation of its usefulness in actual customer service. Traditional spoken dialog systems have used dialog management based on the state of the dialog. Those systems use a rule-based or reinforcement learning-based agent [13], [14], such as a partially observable Markov decision process (POMDP) [15] to transition from one state to another state depending on speech input. In recent years, end-to-end (E2E) speech dialog systems have been studied [16], [17], [18], such as those that generate response sentences directly from input utterances. Yang et al. [19] also proposed an E2E-based dialog manager. Considering speech interaction for customer service training, E2E-based speech interaction control might limit variation in training situations. In this paper, we incorporate a traditional rule-based (modular-based) dialog management method. In addition, it is easier to manage dialog states in a rule-based system than in an E2E system. Since the training scenarios are expected to be prepared by humans, we think it would be easier to develop them. Although the system in this paper is a prototype, it has two characteristic features: first, it can provide customer service training in various job situations using training dialog scenarios provided in comma-separated value (CSV) format; second, it can be used by a single trainee without any human instructors.

The developed prototype system was used by subjects with practical experience, who were then asked to evaluate it after they participated in customer service training using the system. The results confirmed that customer service training using the dialog system might be useful in real work situations. The contribution of this paper is that we have built a spoken dialog training system for customer service improvement for the first time that can train customer interaction in the hospitality industry, and we have asked subjects with practical experience to confirm that the training using this system can be used in practice. The experimental results showed that our simple dialog training system could be used for practical customer service training as well.

The remainder of this paper is organized as follows. Section 2 introduces the design of the spoken dialog system, the specific dialog management flow, and the training scenario. Section 3 describes and discusses the experiment. Section 4 summarizes the paper.

II. Spoken Dialog Training System for Customer Service Improvement

A. System outline

Fig. 1 shows the configuration of the spoken dialog training system for customer service improvement. The system configuration is almost the same as that of a conventional modular-based speech dialog system. After speech recognition of the trainee's speech, morphological analysis is applied to the result. Then, the analyzed text is inputted to the dialog management module. That module updates the state of the dialog according to the content of the speech (recognized text) and sends the response of the virtual agent (customer) based on the updated status to the system control module. When that module receives a response from the dialog management module, it instructs the synthesis voice control module to play the voice of the virtual agent. In addition, it also sends the customer agent's line of dialog, the explanation of the current dialog status, and the customer service advice to the graphical user interface (GUI).

Our dialog system was developed on Unity ver.2018.3.5f1 [20], a game development platform, and the dialog management module was implemented based on "Pytransitions," [21] a finite state machine management platform implemented in Python. We used OpenJTalk [22] as our speech synthesizer.

We describe each function in detail.

B. Speech recognition

The automatic speech recognition (ASR) system uses Windows online speech-to-text API provided by Microsoft, which can be used in a Unity platform. The trainer's speech acquired from the microphone is speech-recognized, and the transcribed text is divided into words using morphological analysis and sent to the dialog management module. We used Mecab [23] as the morphological analysis engine.

C. Dialog management module

If a neural network-based response generator is used, the agent's response cannot be controlled. To evaluate customer service properly, it is desirable to use a dialog management module that can control the virtual agents' responses. Furthermore, we would like to evaluate whether the trainee can respond adequately to nasty behavior by the customer agent. Consequently, we use the simple and traditional rule-based dialog management approach in this study.

Management of the customer service flow and of the dialog status is performed using the state transitions shown in Fig. 2. The dialog state is defined in the training dialog scenario described in Section 2.4 below.

For each dialog state, keywords that are important in customer service are set in the dialog scenario file, along with the dialog state transitions according to the customer service content (speech content) of the trainee. As the dialog management module loads a training dialog scenario file, state transitions depending on the training dialog scenario are expanded in the module. At the same time, any delay in response time is suppressed by synthesizing the lines uttered

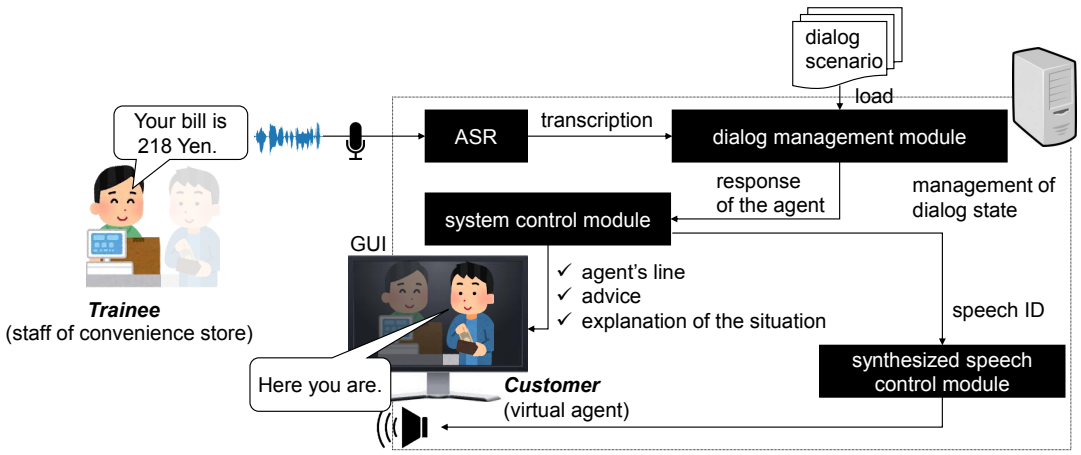

Fig. 1. Outline of the spoken dialog training system for customer service improvement.

TABLE I
A SIMPLE EXAMPLE OF TRAINING SCENARIO.

Current state	Next state	Transition trigger	Keywords	Lines of the customer	Situation description and advice to the trainee
greeting	checkout	STEP1	"Welcome"	"I'll take this."	Tell the total price.
checkout	receipt	STEP1	"218 Yen"	"Here you are."	Receive money.
checkout	correction	STEP0	"218 Yen"	"Sorry, the price is wrong."	Apologize and tell the customer the right amount.
correction	receipt	STEP2	"Sorry", "218 Yen"	"OK"	Receive money.
receipt	handover	STEP3	"Take", "218 Yen", "Exactly"	"OK"	Pass a stuff to the customer.
handover	thanks	STEP2	"Here", "Yours"	"Thanks"	Say "Thank you."
thanks	finish	STEP1	"Thank you"	—	—

by the customer agent when the scenario is loaded into the management module. For example, when the training dialog scenario data shown in Table I is loaded into the dialog management module, the state transitions shown in Fig. 2 are developed in the module.

D. Dialog scenario

A training dialog scenario must be created for each training topic (e.g., checkout at the counter of the convenience store). It consists of six items: "current state", "next state", "transition trigger", "keywords", "lines of the customer agent", and "situation description and advice to the trainee." Table I shows a simple example of a dialog scenario for customer service training at the checkout counter in a convenience store. From this scenario, the state transitions shown in Fig. 2 are generated automatically. By preparing a variety of training dialog scenario data, we can provide various customer service training options, such as complaint handling in various industries such as convenience stores and restaurants.

E. Dialog management based on scenarios

Management of the transition of the interactive state is based on the keywords contained in the speech recognition results of the trainee's voice. This is essentially the same as in a traditional spoken dialog system. The next destination of the dialog state transition is determined by how many keywords are included in the scenario. However, since our system is aimed at customer service training, keyword-based state transitions alone are not sufficient, because this would make the scenario monotonous and would degrade the effect of the training. Therefore, the scenario designer is enabled to set up transition triggers for the creation of various training patterns.

In our dialog system, we prepared seven transition triggers[1]. One is the "STEP" trigger, which is shown in Table I. This is a trigger that changes the destination of the transition depending on how many of the keywords in the scenario match. For example, if the "STEP1" trigger is set in the state transition destination, the transition to the destination will occur when the trigger matches one of the words in the keyword set. A trainee is at the "greeting" state in Table I when the virtual agent comes to the checkout counter to play, and the trainee is asked to say "Welcome" to the customer. If the trainee says "Welcome" correctly, the dialog state will move to the "checkpoint" state depending on the "STEP1" trigger.

F. Graphical User Interface

Fig. 3 shows the GUI of our spoken dialog training system for customer service improvement. The background image and the agent can be changed arbitrarily. The text of the speech recognition results after the morphological analysis is displayed in the upper left corner. This enables us to confirm

[1]These included "RAND", "RUN", "STEP", "CHOOSE", "MULTI", and "LOOP." A transition trigger was used to introduce branching to a scenario.

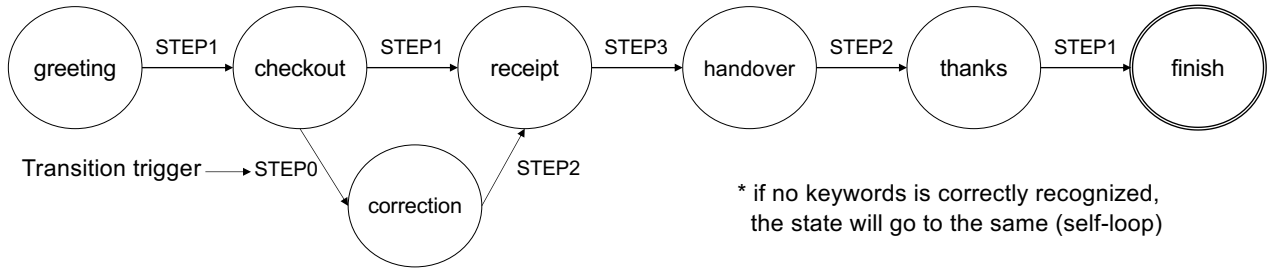

Fig. 2. Example of state transitions from the dialog scenario shown in Table I.

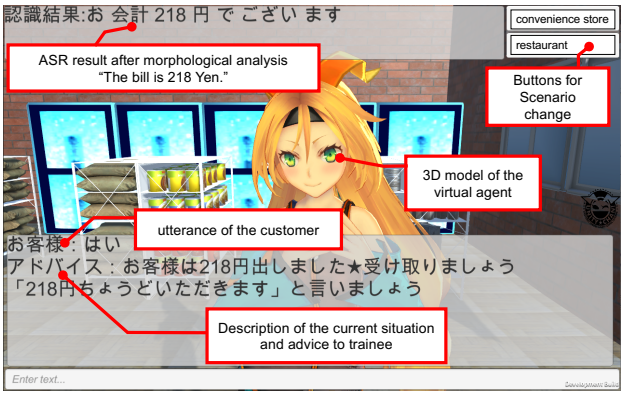

Fig. 3. Graphical user interface of the spoken dialog module.

Fig. 4. Scene of the experiment.

how the trainee's speech is heard by the virtual customer agent. The trainees' speech, "The total bill is 218 Yen," is shown in Fig. 3. The buttons in the upper right corner are used for changing the scenario. This allows the trainer to select the prepared training dialog scenario and conduct customer service training using the selected scenario.

III. Experiment

In this section, we describe an experimental test of our developed spoken dialog training system.

TABLE II
Five Training Dialog Scenarios Used in the Experiment.

Scene	Scenario
Convenience store	1. Checkout operation with guide 2. Checkout operation without guide 3. Customer's complaint for buying cigarettes
Restaurant	4. Order from a customer 5. Customer's complaint on the order

A. Experimental setup

The subjects included 19 undergraduate and graduate students. Most of the subjects had had the experience of working at a convenience store and/or restaurant. Consequently, the effectiveness of customer service training can be evaluated. Each subject wore a microphone and stood in front of a monitor, as shown in Fig. 4. The microphone used in this experiment was a head-worn vocal microphone with super-cardioid directional characteristics, which provides high sound insulation from environmental noises. This system is capable of keyboard-based text input; however, none of the subjects did use text input, but only speech input.

In this experiment, subjects evaluated our dialog training system on the seven items:

1) **Visibility of the GUI**,
2) **Ease of use of the system**,
3) **Ease of use of the interactive training**,
4) **Smoothness** of the speech interaction,
5) **Naturalness** of the training dialog,
6) **Reproducibility** of the real-world environment, and
7) **Understanding of customer service methods** when the system is used.

All of the subjects rated each evaluation item on a 5-point scale from one to five. This five-point scale is the Mean Opinion Score (MOS), where 1 represents poor and 5 represents excellent.

We prepared five training dialog scenarios to be used in the experiment, as shown in Table II. In the training dialog scenario on convenience store customer service, the subjects practiced what kind of customer service is required in any given situation at the accounting counter using three sorts of situations. On the other hand, in the dialog scenarios on restaurant service, the subjects practiced how to serve when taking an order from a customer and how to serve the food. In addition, the subjects practiced how to respond to complaints on orders from a customer.

TABLE III
SUMMARY OF THE AVERAGE RATING RESULTS FROM THE SUBJECTS.

Evaluation Items	Rating
1. Visibility of the GUI	4.3
2. Ease of use of the system	4.1
3. Ease of use of interactive training	3.4
4. Smoothness of dialog	3.0
5. Naturalness of dialog	3.6
6. Reproducibility	4.1
7. Understanding of customer service	4.3

B. Result and discussion

Table III summarizes the rating results of the seven evaluation items obtained by averaging all of the scores from the subjects. "The visibility of use of the GUI" rated very high at 4.3 because of the simplicity of the user interface. The user interface displays only the minimum necessary information. Therefore, most subjects could focus on the virtual agent playing a customer and the advice as a suggestion of customer service. "Ease of use of the system" also received a high rating of 4.1, because the customer service process could proceed only through speech interaction, which is very close to the actual environment. Based on the high evaluation of the user interface and the ease of use of the dialog system, we can conclude that our spoken dialog training system requires no additional human instructors and is easy for one person to use.

However, the subjects scored 3.4 points for ease of use of interactive training, 3.0 points for smoothness of the spoken dialog, and 3.6 points for naturalness of the training dialog, all of which resulted in a score of fewer than four points. The reason for the low usability of the interactive training is interpreted to be that the subjects often had to rephrase their words repeatedly because the keywords were incorrectly extracted owing to misrecognition by the ASR system. Although spoken dialog is very convenient for customer service training, it must also be accompanied by high speech recognition accuracy. In addition, the reason for the low evaluation of the smoothness of the dialog was that the response speed of the virtual customer and the speed of speech recognition created subtle intervals during interaction with the dialog system. This does not arise in human interaction. The ease of use and smoothness of the spoken dialog affected its naturalness.

Although some improvements are still required in our dialog system, the reproducibility of the customer service and the subjects' understandings for customer service were rated highly. In summary, the spoken dialog system for customer service training might shows promise in practice as a replacement for existing training methods.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced our prototype dialog training system for customer service improvement. Because of the need for strict dialog state management in hospitality training dialogs, we used simple rule-based dialog management. The dialog scenarios were prepared by humans to enable subjects to experience a variety of situations in the hospitality industry. The experiments revealed that the system could be used for practical training.

However, according to the experiment, our dialog system needs improvement because of the low scores on the ease of use and smoothness of the spoken dialog and the naturalness of the training dialog. In particular, improvements in the ASR system are required to improve the ease of use of speech dialogs. Currently, we are using Microsoft's ASR system, which is easy to call in the Unity environment, but we are planning use the Kaldi ASR toolkit, which can be customized with a language model and dictionary. In addition, since Kaldi runs on-premises on a local server, the speed of speech recognition can be expected to be improved, and the agent's response time can be expected to be faster [24]. Replacing the ASR system is aimed at improving the usability of the system further. In the training system, the manner of speaking [25] and politeness of wording [26] must also be assessed, and we intend to develop such assessment modules. Moreover, we will also incorporate emotional analysis based on the trainee's voice and face image as in the dialog system in [27]. We also plan to conduct a comparative experiment to evaluate what is better and what improvements are needed between actual human instruction and this system. Finally, in our system, we adopted rule-based response generation, but a deep learning-based generator can be expected to provide more natural responses [28]. The use of a deep learning-based response generator such as the one in [29] will also be considered in the future.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 20H05558.

REFERENCES

[1] K. Chopra, "Empirical study on role of customer service in delivering satisfaction at branded retail outlets in pune," *Procedia Economics and Finance*, vol. 11, pp. 239 – 246, 2014.

[2] M. Holt, "The impact of customer service on customer satisfaction," https://smallbusiness.chron.com/impact-customer-service-customer-satisfaction-2087.html, 5 2020.

[3] P. Gannon-Leary and M. McCarthy, *Customer Care: A Training Manual for Library Staff*. Chandos Publishing, 2010.

[4] P. Jain, "On-the-job training: A key to human resource development," *Library Management*, vol. 20, pp. 283–294, 08 1999.

[5] X. Jin, Y. Bian, W. Geng, Y. Chen, K. Chu, H. Hu, J. Liu, Y. Shi, and C. Yang, "Developing an agent-based virtual interview training system for college students with high shyness level," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2019, pp. 998–999.

[6] J. V. Helvert, P. V. Rosmalen, D. Börner, V. Petukhova, and J. Alexandersson, "Observing, coaching and reflecting: A multi-modal natural language-based dialogue system in a learning context," in *Proceedings of the 11th International Conference on Intelligent Environments*, 2015, pp. 220–227.

[7] C. Wu, R. Socher, and C. Xiong, "Global-to-local memory pointer networks for task-oriented dialogue," *CoRR*, vol. abs/1901.04713, 2019. [Online]. Available: http://arxiv.org/abs/1901.04713

[8] S. Kim and R. E. Banchs, "R-cube: A dialogue agent for restaurant recommendation and reservation," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, 2014, pp. 1–6.

[9] M. S. Yakoub and S. A. Selouani, "Ontology-based framework for a multi-domain spoken dialogue system," *Journal of Ambient Intelligence and Humanized Computing*, https://doi.org/10.1007/s12652-017-0625-y, 2017.

[10] H. Kashioka, T. Misu, E. Mizukami, Y. Shiga, K. Kayama, C. Hori, and H. Kawai, "Multimodal dialog system for kyoto sightseeing guide," in *Proceedings of the 3rd Annual Summit and Conference (APSIPA ASC 2011)*, 2011, pp. 1–6.
[11] J. Bang, H. Noh, Y. Kim, and G. G. Lee, "Example-based chat-oriented dialogue system with personalized long-term memory," in *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, 2015, pp. 238–243.
[12] A. Inc, "Apple siri," https://www.apple.com/siri/, 2020.
[13] H. Cuayáhuitl, S. Renals, O. Lemon, and H. Shimodaira, "Evaluation of a hierarchical reinforcement learning spoken dialogue system," *Computer Speech & Language*, vol. 24, no. 2, pp. 395 – 429, 2010.
[14] X. Li, Y.-N. Chen, L. Li, J. Gao, and A. Celikyilmaz, "Investigation of language understanding impact for reinforcement learning based dialogue systems," 2017.
[15] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.
[16] S. Constantin, Niehues, and A. Waibel, "An end-to-end goal-oriented dialog system with a generative natural language response generation," in *9th International Workshop on Spoken Dialogue System Technology*, 2019, pp. 209–219.
[17] S. Lee, Q. Zhu, R. Takanobu, Z. Zhang, Y. Zhang, X. Li, J. Li, B. Peng, X. Li, M. Huang, and J. Gao, "ConvLab: Multi-domain end-to-end dialog system platform," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Jul. 2019, pp. 64–69.
[18] N. Braunschweiler and A. Papangelis, "Comparison of an end-to-end trainable dialogue system with a modular statistical dialogue system," in *Proceedings of INTERSPEECH2018*, 2018, pp. 576–580.
[19] X. Yang, Y. Chen, D. Hakkani-Tür, P. Crook, X. Li, J. Gao, and L. Deng, "End-to-end joint learning of natural language understanding and dialogue manager," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5690–5694.
[20] U. Technologies, "Unity," https://unity.com, 2020.
[21] T. Yarkoni, "Pytransitions/transitions," https://github.com/pytransitions/transitions, 5 2020.
[22] OpenJTalk, "The Japanese TTS system "Open JTalk"," http://open-jtalk.sourceforge.net, 2020.
[23] "Mecab: Yet another part-of-speech and morphological analyzer," http://taku910.github.io/mecab/, 2020.
[24] C.-S. Leow, T. Hayakawa, H. Nishizaki, and N. Kitaoka, "Development of a low-latency and real-time automatic speech recognition system," in *2020 IEEE 9th Global Conference on Consumer Electronics*, 2020, p. Accepted.
[25] H. Deng, Y. Lin, T. Utsuro, A. Kobayashi, H. Nishizaki, and J. Hoshino, "Automatic fluency evaluation of spontaneous speech using disfluency-based features," in *Proc. of ICASSP2020*, 2020, pp. 9239–9243.
[26] S. Iida, T. Utsuro, H. Nisizaki, and J. Hoshino, "Scenario-based customer service vr training system with honorific exercise," in *Proc. of the 6th International Conference on Biomedical and Bioinformatics Engineering*, 2019, pp. 158–162.
[27] Y. Matsuyama, A. Bhardwaj, R. Zhao, O. J. Romero, S. A. Akoju, and J. Cassell, "Socially-aware animated intelligent personal assistant agent," in *Proc. of the 17th SIGDIAL*, 2016, p. 4.
[28] Q. Yang, Z. He, Z. Zhan, R. Li, Y. Lee, Y. Zhang, and C. Hu, "End-to-end personalized humorous response generation in untrimmed multi-role dialogue system," *IEEE Access*, vol. 7, pp. 94 059–94 071, 2019.
[29] A. Hatua, T. T. Nguyen, and A. H. Sung, "Dialogue generation using self-attention generative adversarial network," in *2019 IEEE International Conference on Conversational Data Knowledge Engineering (CDKE)*, 2019, pp. 33–38.