An Attentive Listening System with Android ERICA: Comparison of Autonomous and WOZ Interactions

Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura,
Katsuya Takanashi, and Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Japan
[inoue,lala,yamamoto,shizuka,takanashi,kawahara]
@sap.ist.i.kyoto-u.ac.jp

Abstract

We describe an attentive listening system for the autonomous android robot ERICA. The proposed system generates several types of listener responses: backchannels, repeats, elaborating questions, assessments, generic sentimental responses, and generic responses. In this paper, we report a subjective experiment with 20 elderly people. First, we evaluated each system utterance excluding backchannels and generic responses, in an offline manner. It was found that most of the system utterances were linguistically appropriate, and they elicited positive reactions from the subjects. Furthermore, 58.2% of the responses were acknowledged as being appropriate listener responses. We also compared the proposed system with a WOZ system where a human operator was operating the robot. From the subjective evaluation, the proposed system achieved comparable scores in basic skills of attentive listening such as *encouragement to talk*, *focused on the talk*, and *actively listening*. It was also found that there is still a gap between the system and the WOZ for more sophisticated skills such as *dialogue understanding*, *showing interest*, and *empathy towards the user*.

1 Introduction

In recent years, android robots have drawn much attention from researchers and the public. Their realistic appearance is their main feature, though this requires that their behaviors are also human-like. In particular, a conversational android should not only respond correctly in terms of their dialogue content, but also exhibit phenomena such as backchanneling and correct turn taking which are present in human-human conversation. Their use as an interface for natural conversation makes them an attractive prospect for research.

Since an android which can engage in free, unstructured conversation on any topic is still a long way off, we investigate a more limited task domain. In this paper we investigate attentive listening, and propose such a system for the android ERICA (Glas et al., 2016), who has been used for tasks such as job interviews (Inoue et al., 2019) and to investigate various conversational phenomena (Lala et al., 2017a, 2019). The extension of ERICA's abilities to attentive listening draws from our previous research (Inoue et al., 2016; Lala et al., 2017b; Milhorat et al., 2019; Kawahara, 2019).

In attentive listening, much of the talk is from the user. The system may interject to stimulate further conversation, but does not engage in deep discussions. The advantage of this task is that the user can theoretically talk about any topic without the system needing any deep background knowledge. Such robots are useful in areas such as elderly care, where users often desire social contact but may be isolated from family (Okubo et al., 2018; Sorbello et al., 2016; Yamazaki et al., 2012). In this case, an android which provides companionship can improve the mental and emotional well-being of the elderly.

This domain provides several technical challenges. The main requirement for attentive listening is that ERICA be seen as actively listening to the conversation. The system must be able to extract the correct topic or keyword and then generate a coherent response which can stimulate further conversation, by using a variety of responses. Furthermore, while the user speaks, ERICA should exhibit human-like listening behavior which may not necessarily be verbal. Synchronizing all these features into an autonomous system is a non-trivial task, as we wish to avoid breakdowns in the conversation.

This system draws together speech recognition, natural language processing and conversational behavior. Our goal is for ERICA to be as human-like as possible in her interactions with users. We com-

Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 85 - 94
July 1-3, 2020 ©2020 Association for Computational Linguistics

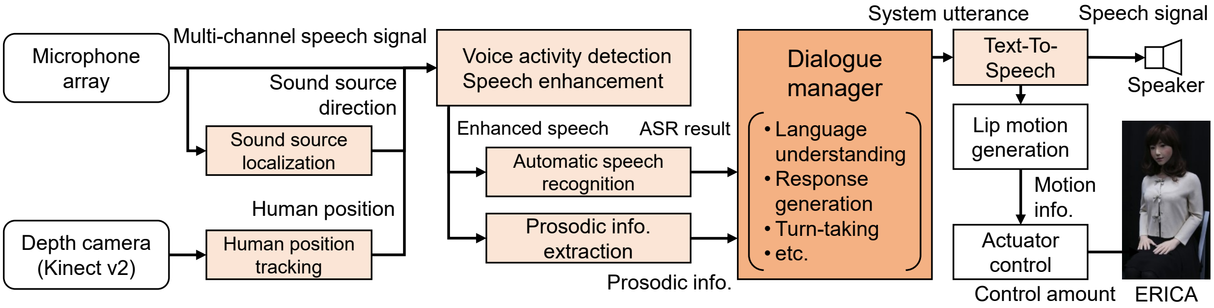

Figure 1: Architecture of a spoken dialogue system for android ERICA

pare an autonomous system to one which is controlled by a Wizard of Oz (WOZ) operator, and see how close we are to achieving human-like attentive listening.

The main contribution of this paper is a fully autonomous android attentive listener. We also report our user study which compares it to a WOZ system. The outcomes of this study will be used to guide further work in the domain of conversational androids.

2 Attentive listening system

We now describe the attentive listening system for the android robot ERICA. The whole architecture of the system is illustrated in Figure 1. First, we explain the speech processing module as the input. We then explain how to generate listener responses, followed by other necessary conversational components such as turn-taking and speech synthesis. A dialogue example can be found in Appendix A. Note that although the following processing is implemented in the Japanese language, the fundamental ideas are language-independent.

2.1 Speech processing

We use a 16-channel microphone array for automatic speech recognition (ASR) and extraction of prosodic features. Based on the multi-channel audio signals, sound source localization is conducted by multiple signal classification (MUSIC) (Ishi et al., 2016) and the direction of the audio is compared with human positions tracked by a Kinect v2 depth camera. If the sound source direction overlaps with the position of a person, enhancement of the audio is conducted and the enhanced speech is fed into an ASR system. The ASR system is implemented by an end-to-end deep neural network model (subword unit). Prosodic information including fundamental frequency (F0) and power is also extracted from the enhanced speech at 100Hz.

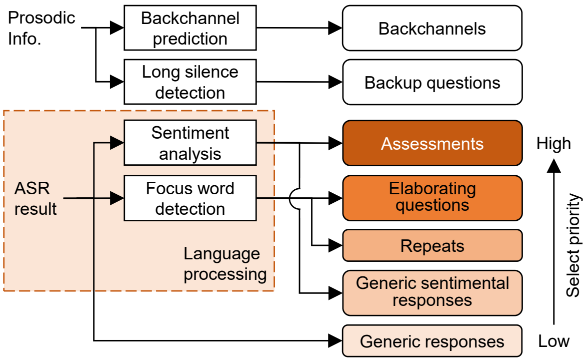

Figure 2: Architecture of listener response generation

2.2 Listener response generation

It is important for attentive listening to generate a variety of listener responses and then select an appropriate one to elicit more utterances from the user. In attentive listening, it is desirable for listener responses to express both understandings of user utterances and empathy towards users. Several attempts to implement artificial attentive listeners have been made so far (Schröder et al., 2012; DeVault et al., 2014; Han et al., 2015; Johansson et al., 2016). Our proposed attentive listening system generates backchannels, repeats, elaborating questions, assessments, generic sentimental responses, generic responses, and backup questions. Excluding backup questions, these responses do not depend on specific dialogue domains, meaning response generation is domain-independent. We now explain how the system generates each response and the selection of the final response.

Backchannels

The system generates backchannels such as "*yeah*" in English and "*un*" in Japanese. Backchannels play an important role in attentive listening in order to make users continue to talk and also to express listener attention and interest in the conversation.

There have been many works on automatic

backchannel generation, with most using prosodic features (Ward and Tsukahara, 2000; Morency et al., 2008; Ozkan et al., 2010; Truong et al., 2010; Kawahara et al., 2016). In our system, we use a logistic regression model that predicts if the system should utter a backchannel within the next 500 milliseconds. This prediction is continuously made every 100 milliseconds during the user's turn. Input features are prosodic information consisting of the statistics (e.g., mean, maximum, minimum, and range) of the F0 and power of the user's speech signal. This continuous prediction makes it possible to generate and utter backchannels during the utterances of the user, making the dialogue more smooth and natural. The backchannel form is determined based on a distribution observed in our attentive listening dialogue corpus, since continuous prediction of backchannel forms is much more difficult. In our system, the backchannels forms are "*un*", "*un un*", and "*un un un*". In Japanese, the use of many repeating backchannels represents the stronger reaction of listeners.

Repeats

For this response, the system extracts a focus word from a user utterance and repeats it. This is expected to express understanding of the dialogue. We use a simple rule to extract a focus word, defining it as the latest noun or adjective in a user utterance. For example, if a user says "*I went to Paris to visit a museum*", the system response would be "*A museum*". If there are several continuous nouns, they are regarded as a compound word and are considered as the focus word. If the ASR confidence score of the focus word is lower than a threshold, the system ignores this to avoid errors caused by ASR.

Elaborating questions

If the extracted focus word can be extended to elicit more dialogue about a topic, an elaborating question is generated. Generating the proper elaborating question not only extends the dialogue but also expresses deeper understanding of the dialogue. The system generates a question by concatenating the focus word with interrogatives such as *which*, *when*, and *what*. In total, we use 11 types of interrogatives as candidates. For example, if a user says "*I went to Paris to visit a museum*", the focus word would be "*a museum*" and the elaborating question would be "*Which museum?*". To select the proper interrogative, the system refers to bigram probabilities of all possible pairs and selects the interrogative that has the highest probability with the focus word. The probability must also be higher than a fixed threshold. If all bigram probabilities are lower than the threshold, no elaborating question is generated. Bigram probabilities are calculated in advance by using large-scale language corpora. In our case, we use the balanced corpus of contemporary written Japanese (BCCWJ)[1].

Assessments

If a user utterance contains a positive or negative sentiment, the system utterance reflects this by using an assessment response. This emotional response is expected to express empathy towards users. We first conduct sentiment analysis of a user utterance by using two kinds of Japanese sentiment dictionaries [2][3] where positive and negative words (phrases) are defined. Since sentiment responses strongly depend on the dialogue context, the dictionaries should focus on precision rather than the coverage. Therefore, we ensure that words in the dictionary are robust in terms of correctly determining the sentiment, even though the number of words is comparatively small. The system determines the sentiment of the current user utterance as positive, negative, or neutral by referring to a sentiment score of each word. If the utterance contains both positive and negative scores, the majority sentiment is used. Similar to focus word detection, if the ASR score of a word is lower than a threshold, then the corresponding sentiment score is ignored. The assessment response is selected according to the estimated sentiment of the user utterance. A positive sentiment leads to system responses such as "*That is good (いいですね)*" or "*That is nice (素敵ですね)*", and negative sentiment leads to responses such as "*That is bad (残念でしたね)*" or "*That is hard (大変ですね)*". If no sentimental words were found, this module does not output any responses.

Generic responses

The system prepares generic responses because the above-mentioned responses are not always generated. Generic responses are "*I see (そうですか)*" or "*I got it (なるほど)*". These responses can be used for any dialogue context. If the user utterance

[1] https://pj.ninjal.ac.jp/corpus_center/bccwj/

[2] https://www.gsk.or.jp/catalog/gsk2011-c/

[3] http://www.jnlp.org/SNOW/D18

is short, the system also uses a short generic response such "*Yes* (はい)" to avoid system barge-in.

Generic sentimental responses

The system also generates another type of generic response according to the sentiment of user utterances. For this response type, we use a different sentiment dictionary (Kobayashi et al., 2004) that covers a wider range of words but also expressions that might have opposing sentiments depending on the dialogue context. We designed generic sentimental responses where the surface form is the same as those of the generic responses but the prosodic pattern changes according to the estimated sentiment. By generating these responses, the system can reduce the risk of a linguistic breakdown (since they don't explicitly use an emotional linguistic response) but also express empathy towards users through prosody.

Backup questions

If a user stays silent longer than a specific amount of time (four seconds in the current system), the system generates one of several backup questions. The questions are defined in advance according to the theme of the user's talk. For example, if the theme is *traveling*, a backup question is "*Where did you go after that?*".

Response selection

Since the above-mentioned modules generate several response candidates, it is important for this attentive listening system to select the proper one among them. Backchannels are uttered during the user's turn, so this module works independently from the others. Backup questions are triggered by a longer pause so that this module is also independent. For the other response types, we designed a priority system as depicted in Figure 2. The system will respond using the highest priority response type which can be generated given the user's utterance. The priority order is based on how likely it is to generate the response type. For example, assessments use a limited dictionary so it is less likely that a user utterance will generate these kinds of responses than the other response types. On the other hand, generic responses can be used without any modeling so will inevitably be required if no other valid response type can be generated..

2.3 Turn taking

Turn-taking is an important feature of attentive listening, since we want to strike a balance between reducing barge-in from the system and allowing the system to interject during the dialogue. A simple approach in a basic spoken dialogue system is to wait until the user has been silent for a set period of time before the system can take the turn. However, this requires fine tuning and is usually inflexible.

We implement a machine learning turn-taking model that uses the ASR result as an input and supplement this with an finite-state turn-taking machine (FSTTM) as used in previous works (Raux and Eskenazi, 2009; Lala et al., 2018) to determine how much silence from the user should elapse before the turn switches to the system. Utterances with a high probability of being end-of-turn are responded to quickly, while the system will wait longer if the user says utterances such as fillers or hesitations.

2.4 Speech synthesis

The speech synthesis in the system has been designed for android ERICA [4]. Since the vocabulary of backchannels, assessments, generic responses are fixed, we recorded natural speech voices and directly play them instead of using real-time synthesis. This is also because it is still difficult to synthesize these kinds of dialogue-specific utterances with a variety of prosodic patterns using current synthesis techniques. For other responses such as repeats and elaborating questions, we can use real-time synthesis because the focus word depends on user utterances.

3 Dialogue experiment

We conducted a dialogue experiment to evaluate how the proposed system works with elderly people as subjects. We also investigated how the system compared when compared to attentive listening with a WOZ operator.

3.1 Conditions

We recruited 20 Japanese elderly people (between 70-90 years old). A snapshot of this dialogue experiment is shown in Figure 3. Subjects were asked to talk to the android robot about two topics: "*Most memorable travel experience*" and "*Delicious food you recently ate*".

We prepared two types of systems: autonomous and WOZ. The autonomous system corresponds to the proposed attentive listening system. The WOZ

[4] `https://voicetext.jp/news/product/151023/`

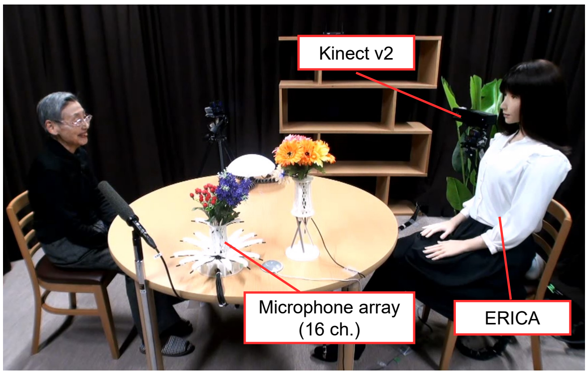

Figure 3: Snapshot of dialogue experiment

system is the case where the robot was operated by a human operator. Each subject talked with one of the systems about one dialogue topic in one condition and then did the same with the other condition. The order of the systems and topics were randomized among the subjects. After they had talked in one of the conditions, we asked them to evaluate the system individually. Note that the average word error rate (WER) of the ASR in the autonomous system was 33.8%, which suggests that the ASR with elderly people is more difficult than those with younger people. The current dialogue experiment explores what level of dialogue can be realized in this challenging situation.

The WOZ operators were two amateur actresses and each of them attended to each dialogue. The operator was asked to use the same set of listener responses as our autonomous system but also asked to properly select the timing and type of the proper response by herself. The operator spoke directly into a microphone and the voice was played via a speaker nearby ERICA, so this dialogue seemed to be natural spoken dialogue. Although the operators' voices were different from those of the speech synthesis, we asked the operators to imitate ERICA's synthesized voice as much as possible.

The dialogue time was set at seven minutes for each conversation. Our experimental trials determined this time as the longest where the autonomous system can continue with the dialogue before it becomes too repetitive. In the autonomous system, when the dialogue time passes seven minutes and the system takes the turn, the system says a fixed phrase to end the dialogue. The same rule was imposed on the WOZ operators.

Table 1: Total frequencies (per session) of each response type in the proposed system

Response type	Frequency
Backchannels	1,601 (80.1)
Repeats	90 (4.5)
Elaborating questions	16 (0.8)
Assessments	45 (2.3)
Generic sentimental responses	62 (3.1)
Generic responses	325 (16.3)
Backup questions	12 (0.6)

3.2 Evaluation on utterances of autonomous system

At first, we analyzed the distribution of response types uttered by the autonomous system. The distribution is reported in Table 1. It can be seen that all the response types could be generated in the autonomous system. As we expected, many system utterances consisted of backchannels and generic responses. On average, repeats were uttered about 4-5 times per dialogue, and elaborating questions were uttered just once, assessments were uttered about twice and generic sentimental responses were uttered about three times. This distribution will also be compared with those of the WOZ system in the later analysis.

We also evaluated each system response manually in an offline manner. In this evaluation, three criteria were considered: (1) *no error*, (2) *reaction*, and (3) *appropriate*.

The first criterion, *no error*, validates the extracted focus word. If the uttered focus word is contained in the context of user utterances and is not strange (e.g., unused words in the human dialogue), the system response was marked as accepted, otherwise rejected. The target types of responses were repeats and elaborating questions. This criterion was used to detect linguistic errors of system utterances caused by ASR or language processing errors.

The second criterion, *reaction*, focuses on the subjects' reactions after the system utterances. The reaction of the subjects to system utterances is also important for evaluation. The target types of responses were repeats, elaborating questions, and assessments. For repeats and assessments, if a subject said a positive reaction such as "*Yeah*" after a system response, the response was accepted. For elaborating questions, if a subject answered the system question the question was accepted.

Table 2: Offline evaluation on each system utterance (*No error* means the correctness of the language processing on the surface level. *Reaction* means the positive reaction of the user after the system utterance. *Appropriate* represents the appropriateness of the system utterance as effective listener responses.)

Response type	(1) No error		(2) Reaction		(3) Appropriate	
	YES	NO	YES	NO	YES	NO
Repeats	83	7	79	11	57	33
Elaborating questions	16	0	13	3	11	5
Assessments	-	-	32	13	31	14
Generic sentimental responses	-	-	-	-	25	37
Total	99 (93.4%)	7 (6.6%)	124 (82.1%)	27 (17.9%)	124 (58.2%)	89 (41.8%)

The third criterion, *appropriate*, validates appropriateness as listener responses. The target types of responses were repeats, elaborating questions, assessments, and generic sentimental responses. Since this criterion seems to be the most important but subjective, we defined the detailed criterion for each type of response as follows. For repeats, if there is another focus word that is clearly more adequate than the used one or there is no focus word in the dialogue context, the repeat response is rejected. For elaborating questions, the same criterion of repeats is firstly applied. Additionally, if the combination of the interrogative and focus word is strange or the elaborating question itself is strange, the question is rejected. For assessments, if the estimated sentiment is the opposite or the sentiment of a user utterance is neutral rather than positive or negative, the assessment response is rejected. For generic sentimental responses, if the estimated sentiment is the opposite or the sentiment of a user utterance is clearly neutral, the response is rejected. Although this criterion was expected to be the most strict, most of the utterances of the WOZ system were accepted as we observed the dialogue. Therefore, this criterion is needed to reveal the future work.

We conducted a manual evaluation of the above three criteria by two evaluators who checked their results with each other. The evaluation result is reported in Table 2. For the first criterion, *no error*, most of the responses were accepted. This means that uttered responses did not elicit linguistic errors. For the second criterion, *reaction*, about 80 percent of the system responses elicited positive reactions from the subjects. This result is good because it shows that many utterances were understood by the subjects. For the third criterion, *appropriate*, more than half of the responses were accepted. Excluding generic sentimental responses, two thirds of responses were accepted. In summary, most of the uttered responses were acceptable on the surface level and were also reacted to by the subjects. On the other hand, nearly half of them left room for improvement by considering a strict level of appropriateness for attentive listening.

We introduce some real examples on the evaluation of the third criterion. The dialogues were done in Japanese so the following is the English translation together with the original Japanese utterances (not the ASR results). E and S represent ERICA and subject utterances respectively.

The following two examples are accepted repeats.

S: This talk is about something 35 years ago.
(今から35年くらい前の話ですね。)
E: 35 years.
(35年ですか。)

S: I got on the plane, but we had to wait for one hour.
(飛行機に乗ったんですけど、その飛行機の中で一時間ほど待たせられて。)
E: One hour. (一時間ですか。)

The following example was rejected.

S: The travel to there still remains in my heart.
(そこへ行った旅行がとても心に残っています。)
E: Heart. (心ですか。)

In this case, there is no focus word in the user utterance so assessments or generic responses would be more appropriate.

The following examples are accepted elaborating questions.

S: Considering side menus, she makes delicious cakes for me.

Table 3: Average scores (standard deviations) on subjective evaluation and t-test results ($n = 20$)

	Question item	Autonomous	WOZ	p-value
(Robot behaviors)				
Q1	The words uttered by the robot were natural	5.0 (1.6)	5.9 (0.9)	.003 **
Q2	The robot responded with good timing	4.8 (1.4)	5.6 (1.3)	.022 *
Q3	The robot responded diligently	5.5 (0.7)	5.8 (1.0)	.005 **
Q4	The robot's reaction was like a human's	4.4 (1.3)	5.2 (1.3)	.008 **
Q5	The robot's reaction adequately encouraged my talk	5.0 (1.4)	5.2 (0.9)	.359
Q6	The frequency of the robot's reaction was adequate	5.1 (1.1)	5.4 (1.1)	.232
(Impression on the robot)				
Q7	I want to talk with the robot again	4.6 (1.3)	5.4 (1.5)	.005 **
Q8	The robot was easy to talk with	4.9 (1.3)	5.4 (1.2)	.116
Q9	I felt the robot is kind	4.7 (1.4)	5.6 (1.2)	.004 **
Q10	The robot listened to the talk seriously	5.6 (1.1)	6.0 (1.1)	.072 +
Q11	The robot listened to the talk with focus	5.6 (1.1)	5.7 (1.1)	.681
Q12	The robot listened to the talk actively	5.4 (1.3)	5.6 (1.1)	.385
Q13	The robot understood the talk	5.0 (1.1)	5.9 (1.4)	.002 **
Q14	The robot showed interest for the talk	5.2 (1.3)	5.8 (1.2)	.028 *
Q15	The robot showed empathy towards me	5.1 (1.4)	5.8 (1.0)	.015 *
Q16	I think the robot was being operated by a human	3.3 (1.3)	2.9 (1.1)	.286
Q17	The robot was good at taking turns	4.5 (1.1)	4.8 (1.3)	.209
(Impression on the dialogue)				
Q18	I was satisfied with the dialogue	4.6 (1.5)	5.3 (1.5)	.012 *
Q19	The exchange in the dialogue was smooth	4.6 (1.4)	5.3 (1.4)	.002 **

(+ $p < .1$, * $p < .05$, ** $p < .01$)

(副食物を考えて、おいしいケーキを作ってくれます。)
E: What cakes?
(どんなケーキですか？)

S: I stayed in Shilla hotel, a big hotel in Seoul.
(ソウルの新羅ホテルってね、大きなホテルでそこで泊まりましたけどね。)
E: What hotel?
(どんなホテルですか？)

The following is a rejected sample.

S: I put radish inside a pot.
(大根の中だけお鍋に入れて。)
E: Which pot?
(どのお鍋ですか？)

In this case, the pot seems not to be the focus word in the dialogue context and should not be elaborated upon as the next dialogue topic.

The following example is an accepted assessment.

S: I ate many dishes and visited many places, so I really enjoyed the time.
(いろんなものを食べたり見たりしてそれは楽しい時が過ごせました。)
E: That is good. (いいですね。)

The following is a rejected sample.

S: She gets angry if I do not start to eat while it is still hot.
(温かいうちに食べないと怒ります。)
E: That is bad. (残念でしたね。)

In this case, although there is actually negative word *angry*, the user was just introducing his family and the sentiment of this utterance seems to be neutral.

3.3 Comparison with WOZ system

We compared the autonomous system with the WOZ system in order to investigate how much it could match that of a human. Table 4 reports the average scores on the subjective evaluation. The question items consist of three categories: robot behavior, impression of the robot, and impression of the dialogue. The subjects evaluated each question item in the 7-point scale from 1 to 7. Overall the evaluated scores were higher than the middle point (4), meaning the autonomous system was given a positive evaluation.

Table 4: Average values (standard deviation) on the analysis of the subjects' utterances and results of t-test

Analyzed behavior	Autonomous		WOZ		p-value
Subject's utterance time / min.	38.3	(5.5)	37.5	(5.9)	.287
Number of uttered words / min.	107.5	(19.1)	112.0	(23.1)	.177
Unique number of uttered words / min.	29.0	(4.4)	32.6	(5.1)	.003 **
Number of uttered content words / min.	53.2	(9.8)	55.6	(12.3)	.220
Unique number of uttered content words / min.	23.3	(4.1)	26.3	(4.4)	.008 **

(** $p < .01$)

We conducted a paired t-test on each question item between the autonomous and WOZ systems (n=20). In the first category, significant differences were observed from Q1 to Q4, but no significant differences were observed in Q5 and Q6. This means that the subjects could perceive the difference in ERICA's utterances between the autonomous and WOZ systems. However, from Q5, there was no clear difference in encouraging the subjects' talk. From Q6, the frequency of listener responses was natural even in the autonomous system.

In the second category, significant differences were observed in questions Q7, Q9, Q13, Q14, and Q15. Interestingly, although there is no significant difference in the listening attitude (Q10, Q11, Q12), significant differences were observed in the items of dialogue understanding (Q13), showing interest (Q14), and empathy towards the user (Q15). This means that the proposed system achieved basic listening skills as well as a human operator, but there is room for improvement on sophisticated skills.

In the third category, impression on the dialogue, scores of both items had significant differences. It is expected that improving the above-mentioned items (e.g., Q13, Q14, Q15) leads to improvement on the impression of this dialogue.

We also analyzed the subjects' utterances as reported in Table 4. These measurements provide objective scores on how much the systems encouraged the subjects' talk. To count the number of words, word segmentation is required in the Japanese language so we used a public tool [5]. Content words were defined as nouns, verbs, adjectives, adverbs, and conjunctions. From our result, the numbers of uttered words and content words were not different between the autonomous and WOZ systems. Interestingly, the unique numbers of uttered words and content words were significantly different, meaning the human operators could elicit a wider variety of lexical content than the autonomous system.

Table 5: Total frequencies (per session) of each response type uttered by the WOZ operators

Response type	Frequency	
Backchannels	1,573	(78.7)
Repeats	48	(2.4)
Elaborating questions	13	(0.7)
Assessments	126	(6.3)
Generic responses	259	(13.0)
Backup questions	3	(0.2)
Others	28	(1.4)

Finally, we analyzed the distribution of listener responses in the WOZ system, as reported in Table 5. Note that generic sentimental responses are included in generic responses because it is hard to distinguish them when they are said by the WOZ operator. Compared with the case of the autonomous system reported in Table 1, assessments were used more by the human operators. Furthermore, the number of repeats was smaller in the WOZ system. This difference can be reflected in the design of the priority order of response types shown in Figure 2.

4 Conclusion

In this work, we described the implementation of an attentive listening system for the android ERICA. We discussed details of the system including how it generates various response types based on the user's utterance. Furthermore, we conducted a user study to investigate the performance of the system compared to one operated by a WOZ operator. We found that the proposed system could match the WOZ system in terms of perceived basic listening skills, but was outperformed by the human for more sophisticated skills such as displaying empathy.

Acknowledgments

This work was supported by JST ERATO Grant number JPMJER1401 and JSPS KAKENHI Grant number JP19H05691.

[5] https://taku910.github.io/mecab

References

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis P. Morency. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, pages 1061–1068.

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

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 129–133.

Koji Inoue, Kohei Hara, Divesh Lala, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2019. A job interview dialogue system with autonomous android ERICA. In *International Workshop on Spoken Dialog System Technology (IWSDS)*.

Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. 2016. Talking with ERICA, an autonomous android. In *Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 212–215.

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

Martin Johansson, Tatsuro Hori, Gabriel Skantze, Anja Höthker, and Joakim Gustafson. 2016. Making turn-taking decisions for an active listening robot for memory training. In *International Conference on Social Robotics (ICSR)*, pages 940–949.

Tatsuya Kawahara. 2019. Spoken dialogue system for a human-like conversational robot ERICA. In *International Workshop on Spoken Dialog System Technology (IWSDS)*.

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

Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. 2004. Collecting evaluative expressions for opinion extraction. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 596–605.

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

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

Divesh Lala, Koji Inoue, Pierrick Milhorat, and Tatsuya Kawahara. 2017a. Detection of social signals for recognizing engagement in human-robot interaction. In *AAAI Fall Symposium on Natural Communication for Human-Robot Collaboration*.

Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 2017b. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 127–136.

Pierrick Milhorat, Divesh Lala, Koji Inoue, Tianyu Zhao, Masanari Ishida, Katsuya Takanashi, Shizuka Nakamura, and Tatsuya Kawahara. 2019. A conversational dialogue manager for the humanoid robot ERICA. In *Advanced Social Interaction with Agents*, pages 119–131. Springer.

Louis P. Morency, Iwan De Kok, and Jonathan Gratch. 2008. Predicting listener backchannels: A probabilistic multimodal approach. In *International Conference on Intelligent Virtual Agents (IVA)*, pages 176–190.

Masataka Okubo, Hidenobu Sumioka, Soheil Keshmiri, and Hiroshi Ishiguro. 2018. Intimate touch conversation through teleoperated android: Toward enhancement of interpersonal closeness in elderly people. In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 23–28.

Derya Ozkan, Kenji Sagae, and Louis P. Morency. 2010. Latent mixture of discriminative experts for multimodal prediction modeling. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 860–868.

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

Marc Schröder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Björn Schuller, Etienne de Sevin, Michel Valstar, and Martin Wöllmer. 2012. Building autonomous sensitive artificial listeners. *IEEE Transaction on Affective Computing*, 3(2):165–183.

Rosario Sorbello, Antonio Chella, Marcello Giardina, Shuichi Nishio, and Hiroshi Ishiguro. 2016. An architecture for telenoid robot as empathic conversational android companion for elderly people. In *Intelligent Autonomous Systems 13*, pages 939–953. Springer.

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

Nigel G. Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics*, 32(8):1177–1207.

Ryuji Yamazaki, Shuichi Nishio, Hiroshi Ishiguro, Marco Nørskov, Nobu Ishiguro, and Giuseppe Balistreri. 2012. Social acceptance of a teleoperated android: field study on elderly’ s engagement with an embodied communication medium in Denmark. In *International Conference on Social Robotics (ICSR)*, pages 428–437.

A Appendix: Dialogue example

A dialogue example of the proposed system is as follows, together with original Japanese sentences. E and U represent ERICA (system) and user utterances respectively. Note that this is a written example for illustrative purposes, and backchannels are omitted in this example.

E1: Please tell me your most memorable travel.
(今までで最も印象に残っている旅行についてお話を聞かせてください。)

U1: Last year, I went to Kyoto to visit a park.
(去年、京都にある公園へ行きました。)

E2: Which park? **(elaborating question)**
(どの公園ですか？)

U2: It is *Maruyama* park.
(円山公園です。)

E3: *Maruyama* park **(repeating)**
(円山公園ですか。)

U3: There is a famous cherry blossom.
(そこには、有名な桜があります。)

E4: A famous cherry blossom **(repeating)**
(有名な桜ですか。)

U4: Then I watched the cherry blossom.
(それで、その桜を見ました。)

E5: I see. **(generic response)**
(そうなんですね。)

U5: It was really beautiful and I was moved.
(本当に綺麗で感動しました。)

E6: That is nice **(assessment)**
(いいですね。)
– long silence –

E7: Where did you go after that? **(backup question)**
(そのあとはどこかへ行きましたか？)

U7: Yes. I went shopping in a mall.
(はい。ショッピングモールへ買い物に行きました。)

E8: Where is the mall? **(elaborating question)**
(どこのショッピングモールですか？)

U8: It is in the middle of the city and it was very crowded.
(街の中心地にあるもので、とても混雑していました。)

E9: That’s hard. **(assessment)**
(大変ですね。)