

# Intelligent Conversational Android ERICA Applied to Attentive Listening and Job Interview

Tatsuya Kawahara, Koji Inoue, and Divesh Lala

*Graduate school of informatics, Kyoto university, Japan*

Correspondence\*:

Tatsuya Kawahara

kawahara@i.kyoto-u.ac.jp

## 2 ABSTRACT

3 Following the success of spoken dialogue systems (SDS) in smartphone assistants and smart  
4 speakers, a number of communicative robots are developed and commercialized. Compared  
5 with the conventional SDSs designed as a human-machine interface, interaction with robots  
6 is expected to be in a closer manner to talking to a human because of the anthropomorphism  
7 and physical presence. The goal or task of dialogue may not be information retrieval, but the  
8 conversation itself. In order to realize human-level “long and deep” conversation, we have  
9 developed an intelligent conversational android ERICA. We set up several social interaction tasks  
10 for ERICA, including attentive listening, job interview, and speed dating. To allow for spontaneous,  
11 incremental multiple utterances, a robust turn-taking model is implemented based on TRP  
12 (transition-relevance place) prediction, and a variety of backchannels are generated based on  
13 time frame-wise prediction instead of IPU-based prediction. We have realized an open-domain  
14 attentive listening system with partial repeats and elaborating questions on focus words as well  
15 as assessment responses. It has been evaluated with 40 senior people, engaged in conversation  
16 of 5-7 minutes without a conversation breakdown. It was also compared against the WOZ setting.  
17 We have also realized a job interview system with a set of base questions followed by dynamic  
18 generation of elaborating questions. It has also been evaluated with student subjects, showing  
19 promising results.

20 **Keywords:** autonomous android, conversational robot, spoken dialogue system, attentive listening, job interview

## 1 INTRODUCTION

21 Spoken dialogue systems (SDS) have been deployed in many gadgets such as smartphone assistants and  
22 smart speakers in the past ten years. Millions of users now use them daily. In these systems, dialogue  
23 is regarded as a means to complete some tasks by the machine. The typical tasks include command and  
24 control of the machine or application software (e.g., music player and alarm) and simple information  
25 retrieval (e.g., weather and routing). The task goals are objective and shared by the system and users. Thus,  
26 the dialogue should be completed as soon as possible. There is a big gap from the human-human dialogue,  
27 in which task goals are not definite and the duration is flexible.

28 Conversational robots are another application of the spoken dialogue systems, but they should be designed  
29 in a different principle from “smart” devices mentioned above. We do not need a robot to ask for weather



**Figure 1.** Android ERICA

30 information and alarm. Instead, the conversational robots are expected to simply talk to or listen to humans,  
31 as humans do. An ultimate goal is to realize human-level dialogue, which is engaging and pleasant. Apart  
32 from the quality of dialogue, there is a significant difference in the style of the dialogue. In the current  
33 human-machine interface, a user is assumed to utter one sentence per one turn, which corresponds to a  
34 command or a query, to which the system will respond. On the other hand, in natural human dialogue, a  
35 participant utters many sentences per one turn, during which the counterpart makes backchannels. The  
36 difference can be analogous to half-duplex versus full-duplex in the communication channel. It is weird to  
37 nod or backchannel to a machine, but a human-looking android would help the realization of human-level  
38 dialogue.

39 With these backgrounds, we are conducting a project to develop an autonomous android ERICA (Glas  
40 et al., 2016; Inoue et al., 2016; Kawahara, 2018), who behaves and interacts just like a human, including  
41 facial look and expression, gaze and gesture, and spoken dialogue. An ultimate criterion is to pass a  
42 Total Turing Test, that is to convince people that ERICA's interaction is comparable to a human, or it is  
43 indistinguishable from a remote-operated android. We hope that through this project we can make clear  
44 what is missing or critical in natural interaction. ERICA is expected to replace some social roles currently  
45 done by a human, or used for conversation skill training in a realistic setting.

46 In this article, the major challenges and key technologies in this autonomous android are described. In  
47 particular, the systems developed for attentive listening and job interview are explained with evaluations.

## 2 ANDROID ERICA

48 Figure 1 shows a snapshot of ERICA. She does not move around, but generates elaborate facial expressions  
49 and movements of heads, eyes, and body attitudes while sitting on a chair, so that she can behave and  
50 interacts exactly like a human. We have also developed ASR (automatic speech recognition) and TTS  
51 (text-to-speech) systems as well as the human tracking system dedicated to ERICA.

### 2.1 Social interaction tasks

52 We have explored a number of social roles suited to ERICA that can take advantage of human-like  
53 presence and would realize human-like dialogue. A majority of tasks conducted by the current SDSs such  
54 as information services are not adequate; they are better suited to smartphones and smart speakers. While  
55 most conventional robots are engaged in physical tasks such as moving objects or delivering goods, recently  
56 many kinds of communicative robots are designed and introduced in public spaces. They serve as an  
57

**Table 1.** Dialogue tasks designed for ERICA

|                     | attentive listening | job interview | speed dating | lab guide |
|---------------------|---------------------|---------------|--------------|-----------|
| role of system      | listen              | ask           | all          | talk      |
| dialogue initiative | user                | system        | mixed        | system    |
| main speaker        | user                | user          | both         | system    |
| main listener       | system              | system        | both         | user      |
| turn-taking         | few                 | explicit      | complicate   | explicit  |

attendant (Fujie et al., 2009) or a receptionist (Bohus and Horvitz, 2009). They are effective for attracting people, but the resulting interaction is usually very shallow and short such as a predefined script and simple guidance. In recent years, chatting systems are also developed intensively, but the dialogue is also shallow and not so engaging. In contrast, interaction with ERICA should leverage physical presence and involve face-to-face communication. Therefore, we design “social interaction” tasks in which human-like presence matters and long deep interaction is exchanged. Here, dialogue itself is a task, and the goal of dialogue may be mutual understanding or appealing. We assign a realistic social role to ERICA, so matched users are seriously engaged beyond chatting. Specifically, we set up the following four tasks. They are compared in Table 1.

#### 2.1.1 Attentive listening

In this task, ERICA mostly listens to senior people talking about topics such as memorable travels and recent activities (Lala et al., 2017). Attentive listening is being recognized as effective for maintaining the communication ability of senior people, and many communicative robots are designed for this task. The role of ERICA is to encourage users to speak for long. In this sense, attentive listening is similar to counseling (DeVault et al., 2014).

#### 2.1.2 Job interview (practice)

While dialogue systems have been investigated for casual interviews (Kobori et al., 2016), a job interview is very important for both applicants, typically students, and companies hiring them. Each side makes a lot of preparations including rehearsal. In this setting, ERICA plays the role of interviewer by asking questions. She provides a realistic simulation, and is expected to replace a human interviewer in the future.

#### 2.1.3 Speed dating (practice)

Speed dating is widely held for giving an opportunity for people to find a partner. In this setting, two persons meet for the first time and talk freely to introduce themselves, and see if the counterpart can be a good match. There was a study (Ranganath et al., 2009) that analyzed a corpus of speed dating. In our setting, ERICA plays the role of the female participant by talking about topics such as hobbies and favorite foods. She provides a realistic simulation and gives proper feedbacks according to the dialogue.

#### 2.1.4 Lab guide

A robot is often used for introducing laboratories and museums. It is expected to attract people, and needs to keep their engagement by providing proper interaction.

### 2.2 System configuration

Figure 2 depicts the whole system configuration for android ERICA. The input devices are a 16-channel microphone array and a depth camera (Kinect v2). These sensors are placed within the interaction environment but not physically within or on the android robot. The sensors do not require any human contact or pre-dialogue calibration, so users can start and engage in dialogue naturally without awareness of any sensors (as opposed to using hand microphones, for example). We use the sensors for sound source localization, voice activity detection, and speech enhancement to obtain the segmented speech of the target

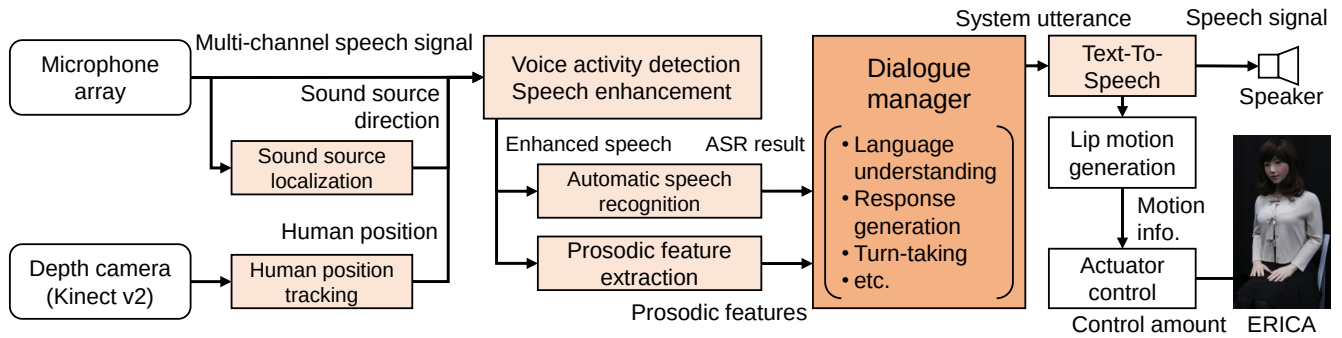


**Figure 2.** System configuration for android ERICA

94 user. The enhanced speech is then fed into the ASR (automatic speech recognition) module which uses  
 95 an acoustic-to-subword end-to-end neural network model. The ASR result is then fed into the dialogue  
 96 manager to generate a system response which will be then played by a TTS (text-to-speech) engine designed  
 97 for ERICA<sup>1</sup>.

### 3 TECHNOLOGIES FOR HUMAN-LEVEL CONVERSATIONS

98 For human-level conversation, linguistic processing and understanding is necessary but non-linguistic  
 99 behaviors are similarly important. We briefly explain several of these components for ERICA.

#### 100 3.1 Backchannel generation

101 Backchannels are short utterances uttered by listeners such as “yeah” in English and “un” in Japanese.  
 102 Listeners utter backchannels toward speakers to stimulate further talk and also to express the listener’s  
 103 attention and interest in the conversation. In order to generate backchannels, systems need to predict  
 104 the timing of backchannels, which has been tackled by many works using prosodic features (Ward and  
 105 Tsukahara, 2000; Truong et al., 2010). While Existing backchannel generation systems make a prediction  
 106 of backchanneling after the end of user utterances segmented by IPU (inter-pausal units), we implement  
 107 frame-wise continuous prediction of backchannel timing with a logistic regression that predicts if the  
 108 system should utter a backchannel within the next 500 milliseconds (Lala et al., 2017). We also proposed a  
 109 prediction model of backchannel form (type) based on both prosodic and linguistic features (Kawahara  
 110 et al., 2016).

#### 111 3.2 Flexible turn-taking using TRP

112 Flexible and robust turn-taking is a required function for natural communication between the system and  
 113 a user. End-of-turn prediction has been widely studied using linguistic and prosodic features extracted  
 114 from the user’s utterance (Skantze, 2017; Masumura et al., 2017). Existing models were trained with  
 115 actual turn-taking behaviors in dialogue corpora, but the turn-taking decision is sometimes arbitrary so  
 116 model training can become unstable. Therefore, we take into account the concept of TRPs (transition-  
 117 relevance places) (Sacks et al., 1974) where the current turn could be completed, and are independent of  
 118 the actual end-of-turn instances. We manually annotated these TRP instances in our human-robot dialogue  
 119 corpus (Hara et al., 2019). We then proposed a two-step turn-taking prediction model where the system first  
 120 detects the TRP, and if detected, predicts if the system actually takes the turn. By decomposing complex  
 121 turn-taking prediction into TRP detection and subsequent turn prediction, the accuracy of turn-taking was  
 122 improved.

<sup>1</sup> <https://voicetext.jp/news/product/151023/>

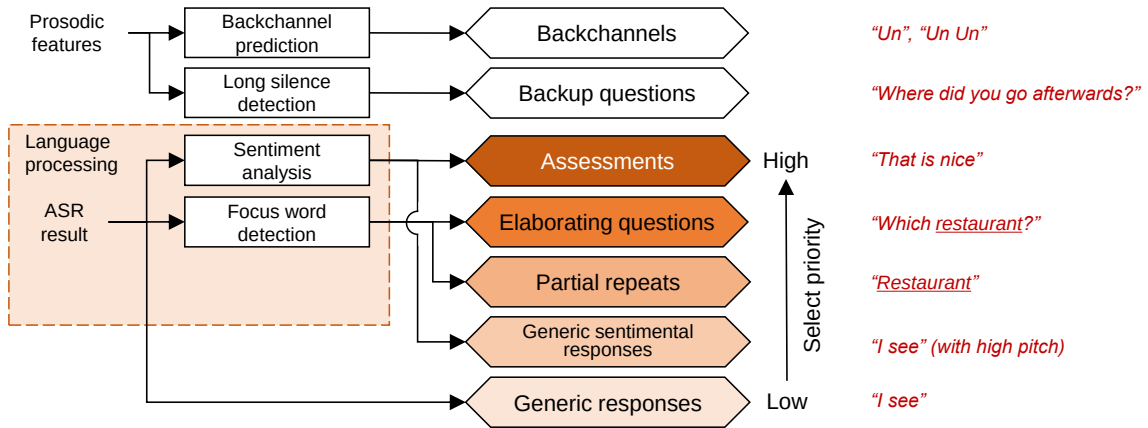


**Figure 3.** List of listener responses in our attentive listening system (Examples of generated responses are shown in right side in this figure. Underline means the focus focus word.)

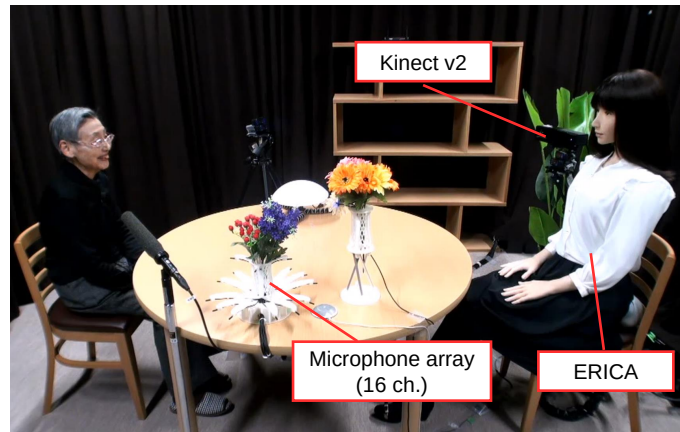


**Figure 4.** Snapshot of dialogue experiment for our attentive listening system with ERICA

## 4 ATTENTIVE LISTENING SYSTEM

We implemented an attentive listening system with ERICA (Lala et al., 2017; Inoue et al., 2020b). The aforementioned backchannel generation model is a core component in this system. Besides backchannels, the system generates a variety of listener responses: partial repeats, elaborating questions, assessments, generic sentimental responses, and generic responses, as depicted in Figure 3. Partial repeats and elaborating questions are generated based on the *focus word* of user utterances. Assessments and generic sentimental responses are based on the sentiment (positive or negative) of user utterances.

We have conducted dialogue experiments with a total of 40 senior people and confirmed that they could engage with ERICA for 5-7 minutes without a conversation breakdown. Figure 4 shows the snapshot of this experiment. We also manually evaluated each system response and found that about 60% of the system responses were acknowledged as appropriate, which means the other 40% responses can be improved. The novelty of our experiment is to compare our system with a human listener implemented in a WOZ (Wizard-of-OZ) setting with a hidden human operator tele-operating ERICA (Inoue et al., 2020b). From the subjective evaluation, our system achieved comparable scores against the WOZ setting in basic skills of attentive listening such as *encouragement to talk*, *focused on the talk*, and *actively listening*. On the other hand, there is still a gap between our system and human listeners for more sophisticated skills such as

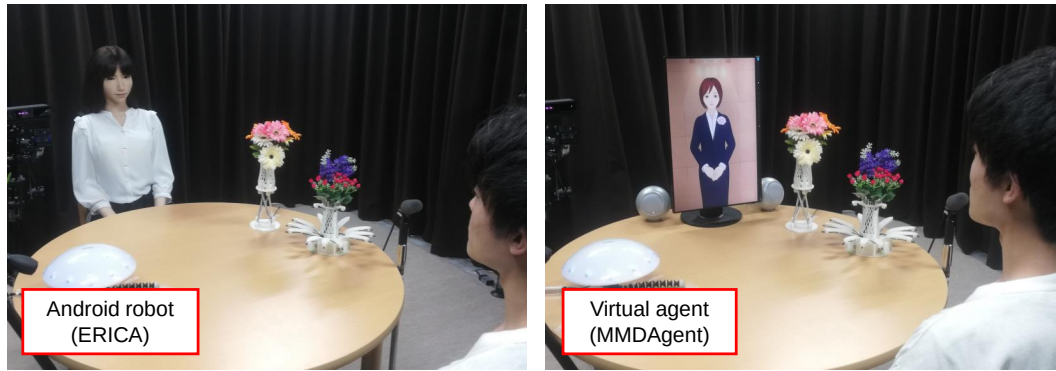


**Figure 5.** Difference of appearance between ERICA and virtual agent in our job interview dialogue experiment

138 *dialogue understanding, showing interest, and empathy towards the user.* From this experiment we could  
 139 identify the type of skills that differed between our system and human listeners. In future work, we will  
 140 improve the response generation modules to increase the ratio of appropriate responses and close the gap  
 141 with human listeners.

## 5 JOB INTERVIEW SYSTEM

142 We also implemented a job interview system where ERICA plays the role of an interviewer (Inoue  
 143 et al., 2020a). Existing job interview systems give the same pre-defined questions to any interviewees,  
 144 so interviews tend to be tedious and far from real human-human interviews. To make interviews more  
 145 realistic, our system generates follow-up questions dynamically based on initial responses from the  
 146 interviewee with two different approaches. The first approach is based on assessing the quality of the  
 147 interviewee’s utterance using a checklist of items that “*should be mentioned*” in the job interview. For  
 148 example, when the interviewee could not mention what he or she wants to contribute to the company, the  
 149 system generates a follow-up question such as “*Which part of our company do you want to contribute to?*”.  
 150 The second approach is based on keyword extraction from the interviewee’s response. For example, when  
 151 the interviewee said “*I have an experience on machine learning.*”, then the system extracts a keyword  
 152 as “*machine learning*” and then generates a follow-up question such as “*Could you explain more about*  
 153 *machine learning?*”

154 We conducted a dialogue experiment with university students to compare our system with a baseline  
 155 system that asked only pre-defined basic questions. We found that our system was significantly better  
 156 than the baseline system in the quality of the questions. It was also found that the perceived presence of  
 157 the android interviewer was enhanced by the follow-up questions (Inoue et al., 2020a). Another research  
 158 question in this job interview scenario is how much the interviewer’s appearance affects the above-reported  
 159 experimental result, so we also conducted the same experiment with a virtual agent, as shown in Figure 5.  
 160 As a result, we confirmed a similar result for the follow-up questions even with the virtual agent interviewer,  
 161 but the presence of the virtual agent interviewer was not enhanced. The presence of the interviewer is an  
 162 important factor in creating an interview system with a realistically tense atmosphere. Therefore, it is more  
 163 effective to implement the follow-up questions with android ERICA as the job interviewer than a virtual  
 164 agent.

## 6 CONCLUSIONS

Because of the pervasiveness of SNS as a communication channels and then COVID-19, importance of face-to-face communication is questioned or emphasized. The android ERICA provides a testbed for investigating what are important factors for the quality communications. The android is expected to replace some social roles under and after the COVID-19 era.

## ACKNOWLEDGMENTS

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

## REFERENCES

- Bohus, D. and Horvitz, E. (2009). Models for multiparty engagement in open-world dialog. In *SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. 225–234
- DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., et al. (2014). SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*. 1061–1068
- Fujie, S., Matsuyama, Y., Taniyama, H., and Kobayashi, T. (2009). Conversation robot participating in and activating a group communication. In *INTERSPEECH*. 264–267
- Glas, D. F., Minaot, T., Ishi, C. T., Kawahara, T., and Ishiguro, H. (2016). ERICA: The ERATO intelligent conversational android. In *International Conference on Robot and Human Interactive Communication (ROMAN)*. 22–29
- Hara, K., Inoue, K., Takanashi, K., and Kawahara, T. (2019). Turn-taking prediction based on detection of transition relevance place. In *INTERSPEECH*. 4170–4174
- Inoue, K., Hara, K., Lala, D., Yamamoto, K., Nakamura, S., Takanashi, K., et al. (2020a). Job interviewer android with elaborate follow-up question generation. In *International Conference on Multimodal Interaction (ICMI)*
- Inoue, K., Lala, D., Yamamoto, K., Nakamura, S., Takanashi, K., and Kawahara, T. (2020b). An attentive listening system with android ERICA: Comparison of autonomous and woz interactions. In *SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. 118–127
- Inoue, K., Milhorat, P., Lala, D., Zhao, T., and Kawahara, T. (2016). Talking with ERICA, an autonomous android. In *SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. 212–215
- Kawahara, T. (2018). Spoken dialogue system for a human-like conversational robot ERICA. In *International Workshop on Spoken Dialog System Technology (IWSDS)*
- Kawahara, T., Yamaguchi, T., Inoue, K., Takanashi, K., and Ward, N. G. (2016). Prediction and generation of backchannel form for attentive listening systems. In *INTERSPEECH*. 2890–2894
- Kobori, T., Nakano, M., and Nakamura, T. (2016). Small talk improves user impressions of interview dialogue systems. In *SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. 370–380
- Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K., and Kawahara, T. (2017). Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. 127–136
- Masumura, R., Asami, T., Masataki, H., Ishii, R., and Higashinaka, R. (2017). Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks. In *INTERSPEECH*. 1661–1665
- Ranganath, R., Jurafsky, D., and McFarland, D. (2009). It’s not you, it’s me: Detecting flirting and its misperception in speed-dates. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 334–342

- 205 Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of  
206 turn-taking for conversation. *Language* 50, 696–735
- 207 Skantze, G. (2017). Towards a general, continuous model of turn-taking in spoken dialogue using LSTM  
208 recurrent neural networks. In *SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. 220–230
- 209 Truong, K. P., Poppe, R., and Heylen, D. (2010). A rule-based backchannel prediction model using pitch  
210 and pause information. In *INTERSPEECH*. 3058–3061
- 211 Ward, N. G. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English  
212 and Japanese. *Journal of pragmatics* 32, 1177–1207