Spoken Dialogue System for a Human-like Conversational Robot ERICA

Tatsuya Kawahara

Abstract This article gives an overview of our symbiotic human-robot interaction project, which aims at an autonomous android who behaves and interacts just like a human. A conversational android ERICA is designed to conduct several social roles focused on spoken dialogue, such as attentive listening (similar to counseling) and job interview. Design principles in developing these spoken dialogue systems are described, in particular focused on the attentive listening system. Generation of backchannels, fillers and laughter is also addressed to make human-like conversation behaviors.

1 Introduction

In the past decade, spoken dialogue systems (SDSs) have become prevailing in smartphones, car navigation systems, and smart speakers. They provide information services on many domains such as weathers and public transportation as well as some chatting function. They are useful, but the dialogue itself is very constrained in that users need to think over what the system can, and then utter one simple sentence with clear articulation before getting a response. Apparently, there is a big gap from human dialogue such as those provided by a tourist guide and a hotel concierge.

A majority of the current SDSs assume one sentence per one turn, and they respond only when the users ask. In human-human dialogue, on the other hand, we utter many sentences per one turn while the listeners make backchannels occasionally. This kind of human-like dialogue manner is necessary for humanoid robots which will be engaged in social or household roles, because humans naturally expect the humanoid robots to show human-like behaviors. We also expect the hu-

Tatsuya Kawahara
School of Informatics, Kyoto University, Japan, e-mail: kawahara@i.kyoto-u.ac.jp

manoid robots with the human-like conversation capability will be used in a variety of domains.

Above all, most current SDSs regard dialogue as a means to conduct some tasks by machines such as operations and information retrieval. In these tasks, an objective goal is defined and should be completed as soon as possible. Thus, their design principle and evaluation criteria are to make the dialogue as efficient as possible. For humans, on the other hand, a dialogue itself can be a task, for example, explanation, persuasion and consulting. Note that the goals of these kinds of dialogue are not necessarily definite, and thus interaction and redundancy are essential. We expect humanoid robots will be engaged in some of these real communication services in the future symbiotic society.

This is the motivation of our Symbiotic Human-Robot Interaction (HRI) project sponsored by the JST ERATO program, which started in 2016. The major focus is placed on the development of an autonomous android ERICA with human-like verbal and non-verbal interaction capability [1, 2, 3]. It is the major distinction of this project that places a priority on the human-like interaction rather than pragmatic applications. Another distinction of the project is the real collaboration between the (author's) spoken dialogue research laboratory and (Prof. Ishiguro's) robotics research laboratories.

2 ERICA Project

2.1 Research Goal

The goal of the project is an autonomous android ERICA, who looks, behaves and interacts exactly like a human. This involves not only spoken dialogue but also eye gaze, head movement, and gestures. A snapshot of ERICA is shown in Fig. 1.

Our ultimate goal is to pass a "Total Turing Test", convincing people it is comparable to a human. It would not be easy to make it entirely indistinguishable from a human in the foreseeable future, but our goal is to make the interaction with ERICA is as engaging as that with a human. In terms of spoken dialogue, a Turing Test can be conducted by the comparison against the remote-operated android or WoZ setting.

As this is still challenging in an open domain, we set up several social tasks designated for ERICA, as listed in Fig. 2. We believe these are promising applications of humanoid robots. On the other hand, we hope the challenge of the Total Turing Test will reveal what is missing in the current interaction and critical for natural interaction.

Fig. 1 A snapshot of ERICA

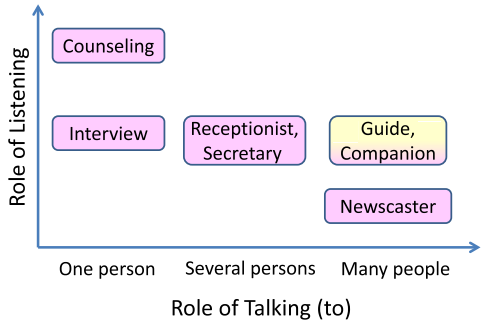

Fig. 2 Social roles of ERICA

2.2 Social Interaction Tasks of ERICA

We have explored a number of social roles suited to the android ERICA that can take advantage of human-like presence and would realize human-like dialogue. A majority of tasks conducted by the current SDSs such as information services are not adequate; they are better suited to smartphones and smart speakers. While most conventional robots are engaged in physical tasks such as moving objects, recently many kinds of communication robots are designed and introduced in public spaces. They serve as a companion [4] or a receptionist [5]. They are effective for attracting people, but the resulting interaction is usually very shallow and short such as greeting and short chatting. In recent years, chatting systems are also developed intensively, but most of the generated responses are boring or non-sense.

In contrast with these tasks, we design "social interaction" tasks in which human-like presence matters and deep and long interaction is exchanged. Here, dialogue

Table 1 Comparison of three social interaction tasks

	attentive listening	job interview	speed dating
dialogue initiative	user	system	both (mixed)
utterance mostly by	user	user	both
backchannel by	system	system	both
turn-switching	rare	clear	complex

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 will be seriously engaged beyond chatting. Specifically, we set up the following three tasks. Note that these are all dyadic style, and face-to-face interaction including non-verbal communication is important.

1. Attentive listening
   In this task, ERICA mostly listens to senior people talking about topics such as memorable travels and recent activities [6]. Attentive listening is being recognized as effective for maintaining the communication ability of senior people, and many communication 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 [7].
2. Job interview (practice)
   While dialogue systems have been investigated for casual interviews [8], 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.
3. Speed dating (practice)
   Speed dating is widely held for giving an opportunity for people to find a partner. In Japan, there was a convention of arranged meeting for marriage, named *Omiai*, which is set up for one couple. In these settings, 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 [9] 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.

While these three tasks share the key characteristics that face-to-face interaction is important, they are different in the nature of dialogue as listed in Table 1.

2.3 Dialogue Data Collection

We have set up an environment for multi-modal recording of dialogue sessions of ERICA and subjects in the Wizard of Oz (WoZ) setting. We recruited four female

actresses as an operator of ERICA. They not only respond to the user utterances but also operate ERICA's motion and eye gaze using haptic devices.

A subject sits in front of ERICA over a round table, and is engaged in dialogue for about 10 minutes. Microphone arrays, cameras, and Kinect sensors are set up on and around the table.

As of April 2018, we have collected 19 sessions of attentive listening, 30 sessions of job interview, and 33 sessions of speed dating.

2.4 Research Issues

There are many research topics involved in this project.

1. Robust automatic speech recognition (ASR)
   Humanoid robots need to deal with distant and conversational speech. This calls for integration of front-end microphone-array processing with back-end acoustic and language models.
   We note that when people speak without a microphone, the speaking style becomes so casual that it is not easy to detect utterance units. This is a serious problem in human-robot interaction, but circumvented in smartphones by using the push-to-talk interface and in smart speakers by forcing a magic word.
2. Flexible dialogue
   Humanoid robots need to deal with conversations without definite goals. This requires language understanding and generation without well-defined semantic slots. Moreover, natural turn-taking and backchanneling capability are essential for human-like interactions.
3. Natural speech synthesis
   Speech synthesis should be designed for the conversational style rather than text-reading applications, which are conventional targets of text-to-speech (TTS). Moreover, a variety of non-lexical utterances such as backchannels, fillers and laughter are needed with a variety of prosody.

Moreover, latency is critical for human-like conversation. In our corpus analysis of the WoZ dialogue, we find the average turn-switch interval is approximately 500 msec, and 700 msec would be too late for smooth conversation. Cloud-based ASR and TTS services can hardly meet this requirement. The ASR system has been based on Julius[1], but recently we have developed an acoustic-to-word end-to-end ASR system, which realizes a real-time factor of 0.03 [10]. The speech synthesis is based on VoiceText[2], but enhanced with non-lexical tokens mentioned above. All downstream natural language processing (NLP) modules and motion generation modules are tuned to run within 200 msec.

Hereafter, this article focuses on the flexible dialogue designed for the social interaction tasks mentioned in the previous section.

[1] `http://julius.osdn.jp`

[2] `http://voicetext.jp`

3 Spoken Dialogue System Design

There are several different approaches to spoken dialogue design, which are described as follows.

- State-transition flow
  A flow of dialogue is hand-crafted as a finite state machine (FSM). This approach has been widely adopted in a limited task domain with a definite task goal, such as filling forms and making transactions. It can also be extended to scenario-based dialogue such as tourist guide. Although it allows for deep interaction, it works only in narrow domains and cannot cope beyond the prepared scenario.
- Question-Answering
  The question-answering (QA) technology is widely used in smartphone assistants and smart speakers. It handles a wide variety of user queries such as weathers and news, and searches the relevant database or Internet to generate an answer. This approach provides wide coverage, but only short interaction. Moreover, it cannot cope beyond the prepared database; in this case, the system simply outputs web search results.
- Statement-Response
  The statement-response function is incorporated in ChatBot and smartphone assistants. There are several approaches to realize this function. One is to prepare a set of statement-response pairs, which can generate relevant responses to a limited number of patterns. Alternatively, the system tries to generate a sentence based on the current focus words, or simply outputs formulaic responses such as "Okay".

The above-mentioned approaches have different advantages and disadvantages, and a combination of them would make a complex system, which a humanoid robot should be equipped with. Fig. 3 depicts a hybrid system design based on these modules. For example, a lab guide system can be composed of a hand-crafted flow and some question-answer function. For this architecture, dialogue act recognition and domain recognition are needed to classify user inputs into the appropriate module. The focus word is also detected and tracked to generate relevant responses.

Moreover, we incorporate a backchanneling function as a separate module in order to generate a human-like behavior.

4 Attentive Listening System

A major demand for communication robots is to listen to users talking. Talking about troubles occasionally clarifies and solves them. Talking by remembering is important for maintaining the communication ability of senior people. People can talk to pets and dolls instead of a human. That is the reason why many communication robots are designed for attentive listening.

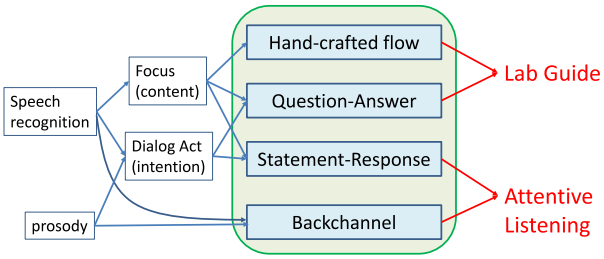

Fig. 3 Hybrid architecture of spoken dialog systems

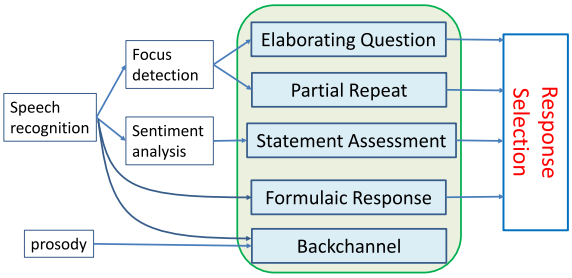

Fig. 4 Architecture of the attentive listening system

The system should encourage users to talk more smoothly. It needs to respond to any inputs, but does not require a large knowledge base. It is also important to show empathy and entrainment. Therefore, natural and smooth backchanneling is critically important, which is described in the next section.

Based on the techniques of counseling and attentive listening, we have designed an attentive listening system [6], as shown in Fig. 4.

The statement-response function is realized with the four modules.

- Elaborating questions
  Questions are generated based on a focus word. The system tries to combine it with WH phrases (i.e. which, where..) to generate a set of possible questions. The most plausible question is selected based on the N-gram probability [6] computed with a large corpus. For example, given an input “I went to a conference”, the system asks “Which conference?”
- Partial repeat
  If none of the elaborating questions is generated, the system can simply respond with the focus word. For example, given an input “I went to Okinawa”, the system says “Okinawa?” This simple response shows understanding of the user utterance

and encourages the user to talk more. Actually, an effective dialogue strategy is to first output this partial repeat and then generate an elaborating question if the user does not continue to talk.

- Statement assessment
  Sentiment analysis is used to generate an assessment of the user utterances. We prepared specific responses for sentiment categories, for example, "That's nice" for positive and objective facts. These responses should be generated with confidence after listening in a sufficient amount of time.
- Formulaic response
  If none of these above can be generated, the system outputs formulaic responses such as "I see." and "Really?" They have a similar function to backchannels.

Selection among these responses is not so simple. There are many possible responses to an input, and there is no ground truth because responses other than the one observed in the corpus can be acceptable. Therefore, we formulate as the validation problem instead of the selection problem. Namely, we design machine learning to judge if a generated response is acceptable given the linguistic and dialogue context. But this requires extra annotation of all possible responses which are not observed in the corpus. We may turn to crowd-sourcing, but this is the problem to be solved for large-scale training.

Currently, we have made evaluations with the 19 dialogue sessions of attentive listening to senior subjects, and achieved a recall of 70% and a precision of 73% for valid responses on average. This is much better than the case randomly generating according to the distribution in the corpus and the result of training with the corpus occurrence only.

During the corpus collection in the WoZ setting, almost all senior subjects believed that they were talking to ERICA, without awareness of the operator. This means that we can pass a Turing Test if we can generate the same (appropriate) responses for the corpus. But we could achieve 70% at this moment.

We have also conducted trials of the system by asking senior people to talk to the autonomous ERICA. In almost all cases, they can be engaged in dialogue lasting five minutes without serious troubles. Although a quantitative evaluation is yet to be done, ERICA occasionally made occasionally appropriate responses, while she only backchanneled most of the time in some sessions.

5 Generation of Backchannels, Fillers and Laughter

Another factor for human-like conversation is generation of non-lexical utterances. They include backchannels, fillers and laughter.

5.1 Backchannel Generation

Backchannels provide feedback for smooth communication, by indicating the listener is listening, understanding, and agreeing to the speaker. Some types of backchannels are used to express the listener's reactions such as surprise, interest and empathy. Moreover, a series of backchannels produces a sense of rhythm and feelings of synchrony, contingency and rapport [11, 12].

There are three factors in generating backchannels, which are timing (when), lexical form (what), and prosody (how). A majority of the previous works addressed the timing of backchannels [13, 14], but many conventional systems used the same recorded pattern of backchannels, giving a monotonous impression to users.

We investigated the prosody of backchannels, and found their power is correlated with the preceding user utterances [15]. We have also explored for generation of a variety of backchannel forms depending on the dialogue context [16], based on machine learning using linguistic and prosodic features [17, 18, 19]. However, the generation of backchannels and choice of their form are arbitrary, and the evaluation with the corpus observation is not meaningful. Thus, we augment the annotation by adding acceptable forms of backchannels for each occurrence point. Similar to the response selection in the attentive listening system, we formulate as the validation problem. The prediction accuracy on average including the not-to-generate case is 64.3%. We also conducted a subjective evaluation by preparing audio files of generated backchannels. The proposed method obtained much higher ratings than the random generation, and to our surprise, it is almost comparable to the original counselor's choice when we use the same voice. This result suggests that we can pass a Turing Test in terms of backchannel form generation. But we need to tune the prosody to be more natural and of variation.

5.2 Filler Generation

Conventionally, fillers have been regarded as redundant and thus those to be removed. On the other hand, an utterance of a long sentence without any fillers generated by TTS is not natural [20]. In fact, humans use fillers, not only due to disfluency, but also for showing politeness and for attracting attention.

Fillers are also used for smooth turn-taking, namely either holding the current turn or taking a turn. We are investigating a method to generate fillers at the beginning of the system utterances to indicate an intention of turn-taking or turn-holding just like human conversation [21]. They can be effective for avoiding speech collision, because the collision of fillers with the user utterances is not so harmful and TTS usually cannot cancel the speech output once generated.

5.3 Laughter Generation

Laughter is also an important component in human-human dialogue. It is used for ice-breaking and socializing each other. Recently, there are several studies to investigate laughter for humanoid robots [22]. Analysis of our corpus shows that a large majority of laughter is (1) speech laughter rather than stand-alone laughter, and (2) breath laughter rather than obvious laughter. These suggest that laughter is generated not because the subjects feel funny. On the other hand, we observed many negative laughter samples, which follow negative sentences annotated by the sentiment analysis.

Similar to backchannels, there are three factors for laughter generation, but none of them is not an easy problem and has not been seriously investigated. Moreover, inappropriate laughter would make a very bad impression compared with backchannels.

6 Conclusions and Future Directions

This article has introduced our ERICA project, which aims at human-like conversation capability. We have examined realistic social roles, which would bring users into realistic situated dialogue. They are attentive listening, job interview, and speed dating. We are collecting realistic dialogue data, and designing and implementing systems for these three tasks. Then, we try to evaluate the system by comparing its outputs against the corpus. However, the corpus may not be the ground truth in this kind of social interaction tasks, and thus we augment the annotation.

There are other issues which are not addressed in this article. One is flexible turn-taking [23]. We note again that the corpus does not provide the ground truth for this problem. Modeling non-verbal information such as valence and affect is also needed. Character modeling may be also useful for making the system human-like.

We focus on recognition of user engagement [24], which indicates a positive/negative attribute to keep the current dialogue. This is closely related to the performance of the system. The ultimate dialogue with the humanoid robot should be as engaging as human-human dialogue. Therefore, comparable performance to human-like interaction experience should be measured by the engagement level, which will be used for a Total Turing Test.

Acknowledgements This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (Grant Number JPMJER1401), Japan.

References

1. D.F.Glas, T.Minato, C.T.Ishi, T.Kawahara, and H.Ishiguro. ERICA: The ERATO Intelligent Conversational Android. In *Proc. RO-MAN*, pages 22–29, 2016.
2. K.Inoue, P.Milhorat, D.Lala, T.Zhao, and T.Kawahara. Talking with ERICA, an autonomous android. In *Proc. SIGdial Meeting Discourse & Dialogue*, volume Demo. Paper, pages 212–215, 2016.
3. P.Milhorat, D.Lala, K.Inoue, Z.Tianyu, M.Ishida, K.Takanashi, S.Nakamura, and T.Kawahara. A conversational dialogue manager for the humanoid robot ERICA. In *Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS)*, 2017.
4. S.Fujie, Y.Matsuyama, H.Taniyama, and T.Kobayashi. Conversation robot participating in and activating a group communication. In *Proc. InterSpeech*, pages 264–267, 2009.
5. D.Bohus and E.Horvitz. Models for multiparty engagement in open-world dialog. In *Proc. SIGdial*, 2009.
6. D.Lala, P.Milhorat, K.Inoue, M.Ishida, K.Takanashi, and T.Kawahara. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proc. SIGdial Meeting Discourse & Dialogue*, pages 127–136, 2017.
7. 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, Y.Xu, A.Rizzo, and L-P.Morency. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proc. AAMAS*, 2014.
8. T.Kobori, M.Nakano, and T.Nakamura. Small talk improves user impressions of interview dialogue systems. In *Proc. SIGDial*, pages 370–380, 2016.
9. R.Ranganath, D.Jurafsky, and D.McFarland. It's not you, it's me: Detecting flirting and its misperception in speed-dates. In *Proc. EMNLP*, 2009.
10. S.Ueno, H.Inaguma, M.Mimura, and T.Kawahara. Acoustic-to-word attention-based model complemented with character-level CTC-based model. In *Proc. IEEE-ICASSP*, pages 5804–5808, 2018.
11. R.Levitan and J.Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proc. InterSpeech*, pages 3081–3085, 2011.
12. B.Xiao, P.G.Georgiou, Z.E.Imel, D.Atkins, and S.Narayanan. Modeling therapist empathy and vocal entrainment in drug addiction counseling. In *Proc. InterSpeech*, pages 2861–2864, 2013.
13. N.Kitaoka, M.Takeuchi, R.Nishimura, and S.Nakagawa. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. *J. Japanese Society for Artificial Intelligence*, 20(3):220–228, 2005.
14. D.Ozkan and L.-P.Morency. Modeling wisdom of crowds using latent mixture of discriminative experts. In *Proc. ACL/HLT*, 2011.
15. T.Kawahara, M.Uesato, K.Yoshino, and K.Takanashi. Toward adaptive generation of backchannels for attentive listening agents. In *Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS)*, 2015.
16. T.Kawahara, T.Yamaguchi, K.Inoue, K.Takanashi, and N.Ward. Prediction and generation of backchannel form for attentive listening systems. In *Proc. INTERSPEECH*, pages 2890–2894, 2016.
17. N.Ward. Using prosodic clues to decide when to produce back-channel utterances. In *Proc. ICSLP*, pages 1728–1731, 1996.
18. N.Ward and W.Tsukahara. Prosodic features which cue back-channel responses in English and Japanese. *J. Pragmatics*, 32(8):1177–1207, 2000.
19. H.Koiso, Y.Horiuchi, S.Tutiya, A.Ichikawa, and Y.Den. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language & Speech*, 41(3-4):295–321, 1998.
20. S.Andersson, K.Georgila, D.Traum, M.Aylett, and R.A.J.Clark. Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In *Proc. Speech Prosody*, 2010.

21. R.Nakanishi, K.Inoue, S.Nakamura, K.Takanashi, and T.Kawahara. Generating fillers based on dialog act pairs for smooth turn-taking by humanoid robot. In *Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS)*, 2018.
22. B.B.Turker, Z.Bucinca, E.Erzin, Y.Yemez, and M.Sezgin. Analysis of engagement and user experience with a laughter responsive social robot. In *Proc. InterSpeech*, pages 844–848, 2017.
23. G.Skantze, A.Hjalmarsson, and C.Oertel. Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Communication*, 65:50–66, 2014.
24. K.Inoue, D.Lala, K.Takanashi, and T.Kawahara. Latent character model for engagement recognition based on multimodal behaviors. In *Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS)*, 2018.