Talking with ERICA, an autonomous android

Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara
Graduate school of informatics, Kyoto University, Japan

Abstract

We demonstrate dialogues with an autonomous android ERICA, who has an appearance like a human being. Currently, ERICA plays two social roles: a laboratory guide and a counselor. It is designed to follow the protocols of human dialogue to make the user comfortable: (1) having a chat before the main talk, (2) proactively asking questions, and (3) conveying proper feedbacks. The combination of the human-like appearance and the appropriate behaviors according to her social roles allows for symbiotic human-robot interaction.

1 Introduction

Dialogue systems deployed in various devices such as smartphones and robots have been widely used to assist users in daily life. Although they can reply to users for what they are asked, their behaviors are mechanical and the primary objective of dialogue is efficiency (Wilcock and Jokinen, 2015; Skantze and Johansson, 2015). Users need to adapt their behaviors such as their utterance style for the systems, and thus the observed users' behaviors are different from those in human communication.

In the current ERATO project, an autonomous android ERICA with the appearance of human being is developed. Our goal is to make her behave like a human being and naturally interact with human beings by tightly integrating verbal and non-verbal information. For the moment, we make ERICA play a specific social role according to the conversational situation. Figure 1 illustrates some prospective social roles which could be covered by ERICA. The roles are plotted on the two axes that are in the trade-off relation: roles of speaking and listening. In the long term, ERICA is expected to replace these human beings with comparable performance.

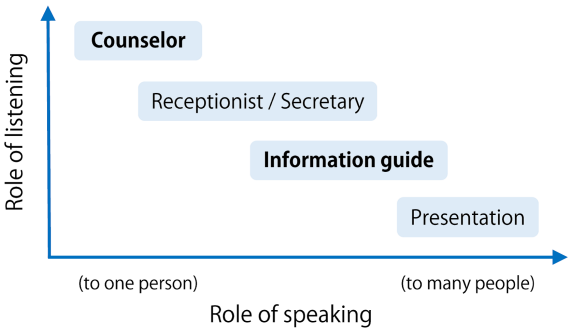

Figure 1: Social roles covered by ERICA

In this demonstration, ERICA plays two social roles: a laboratory guide and a counselor. The scenarios assume that the user meets ERICA for the first time where the user might be nervous. The highlight in the current demonstration is ERICA trying to make the user comfortable by doing the following:

1. Have a personal chat before the main talk to ease their nervousness (ice-breaking)
2. Occasionally make questions from ERICA toward the user when the user does not say anything (ERICA does not only respond to what is asked by the user)
3. Convey proper feedbacks to express that ERICA attentively listens to the user's talk and encourage the user to talk more

ERICA is enhanced by a multi-modal sensing system which consists of a microphone array and a depth camera to realize robust and smooth interaction.

Proceedings of the SIGDIAL 2016 Conference, pages 212–215,
Los Angeles, USA, 13-15 September 2016.

Figure 2: Android ERICA

2 Android ERICA

An image of ERICA is shown in Figure 2. ERICA is a 23 year-old woman. Her design concept is to contain both the friendliness as an android and a sense of existence as a human being. The appearance of her face and body is artificially produced in reference to characteristics of beautiful ladies.

ERICA mounts 19 active joints inside to move her face, head, shoulder, and back. It is planned to install more motors on her to move her arms and legs in the future. Even now, the flexibility of her face has diversity (including eyebrow, eyelids, lip, eyeballs, and tongue), which enables her to show various facial expressions. ERICA is therefore able to generate not only verbal responses but also non-verbal behaviors such as facial expression, eye-gaze, and nodding, which are used to convey a variety of her emotions.

3 Social roles played by ERICA

In this demonstration, we show the following two scenarios of different social roles played by ERICA.

3.1 Laboratory guide

In the first scenario, ERICA introduces research topics in our laboratory when a guest (user) visits there. We assume that the user meets ERICA for the first time. When people meet each other for the first time, it is common that they have a chat like a self-introduction to know each other well and ease the tension, called ice-breaking, so that they are able to establish rapport, which will result in better communication afterward. ERICA follows this protocol.

In the chatting phase, We provided 31 personal topics that ERICA and the user can discuss, such as their hometowns and hobbies, which will be useful for knowing each other. At first after a greeting, ERICA prompts the user to ask a question regarding herself. The uttered question is matched against the topic database by a language understanding module which is implemented by a two-step search, a keyword matching and a vector space model. After her reply, she occasionally makes a follow-up question which is related to the current topic. Here, we measure a pause as a cue which triggers this follow-up question. When the user replies to the follow-up question, ERICA says an assessment reply. The dialogue continues with either a new question from the user or a further follow-up question from ERICA. A dialogue example is as follows. Note that **U** and **E** correspond to utterances from the user and ERICA, respectively.

U1 What is your hobby?
E1 My hobbies are watching movies, sports, and cartoons.
(pause)
E2 Do you have the same hobbies as me? (follow-up question)
U2 Yes, I also like watching movies.
E3 Wow, I am happy to hear that. (assessment reply)

Other than questions, the user might say a statement in the chatting dialogue. To deal with this, if no topic is selected in the above matching, ERICA tries to detect a focus word from the user utterance, which is new information in the dialogue, and makes the following feedbacks using the detected focus word.

Partial repeats Simply repeat the focus word, or a phrase containing the focus word
Questions for elaboration Ask a question to elaborate the focus word
Formulaic responses Fixed phrases (e.g. “Oh really!”)
Backchannels Short responses suggesting that ERICA is listening to the user (e.g. “okay”)

The focus word detection is realized by a CRF-based classification (Yoshino and Kawahara, 2015). A dialogue example is as follows.

U1 I ate a hot dog yesterday.
E1 Hot dog? (partial repeat)
U2 Yeah, I went to a hot dog shop with my family.
E2 Where is the hot dog shop? (question for elaboration)
U3 It is near the central station.

E3 Oh really! (formulaic responses)

Once they have gone through a certain number of topics or the user says a specific key-phrase such as "Tell me about your research topics", the dialogue is switched to the laboratory guide phase. In this phase, ERICA presents several research topics, and the user can choose one of them based on his/her interest. This is designed as information navigation (Yoshino and Kawahara, 2015) and implemented by a finite state model. According to the topic selected by the user, ERICA briefly talks about the topic and asks the user if she can continue the topic in detail or not.

3.2 Counselor

Another social role played by ERICA is as a counselor of the user. In recent years, dialogue systems have been actively studied in the field of counseling and diagnoses (DeVault et al., 2014). Compared with them, ERICA can generate more realistic behaviors (not virtual) which could elicit more natural reactions from the interlocutor. The important role for counselors is to attentively listen to the user and give appropriate feedbacks to encourage the user to talk more. One of the listener's feedbacks are backchannels which are a short utterance such as "okay" and "wow." To generate appropriate backchannels, we need to predict the timing and form of the backchannel depending on the user utterance. Backchannel forms have a variety of different functions: one is to encourage the user to keep talking (called "continuer"), and the other is to show reaction to the user utterance (called "assessment") (Clancy et al., 1996).

In this demonstration, ERICA predicts the timing and form of the backchannel using prosodic information extracted from the user utterance. Here, we deal with four types of backchannels: three continuers and one assessment. Prediction of timing and the form is done by a logistic regression model trained with a corpus of counseling dialogue (Yamaguchi et al., 2016). For practical use, we recorded many backchannel voices varied in forms and levels, and choose the appropriate sample in real time. A dialogue example is as follows.

U1 It is nice weather today.
E1 *Un*. (continuer)
U2 It is the best day to play football outside.
E2 *Un, un*. (continuer, stronger than the previous one)
U3 I really like to play football. (no backchannel)
U4 I play it with my colleagues every day after work.
E3 *He-!* (assessment)

4 System

In this section, we describe a multi-modal interactive system for ERICA. Figure 3 illustrates its entire configuration. The input sensors consist of a microphone array and a depth camera. These sensors are located around ERICA, not on the android, which increases the degree of freedom of sensor arrangement.

4.1 Speech localization and recognition with microphone array

The microphone array captures multi-channel audio signals and identifies which direction the acoustic signal comes from. Here we use a 16-channel microphone array and adopt the MUltiple SIgnal Classification (MUSIC) method (Schmidt, 1986) to calculate the sound source direction. Afterwards, the input speech is enhanced by using the delay-and-sum beamforming. From the enhanced speech, we calculate prosodic information including fundamental frequency (F0) and power.

The automatic speech recognition (ASR) in ERICA is done by using the enhanced speech. Distant speech recognition elicits more natural human behavior because the user is able to use their arms and hands to show gestures. To realize the distant speech recognition, the enhanced speech is processed by a denoising auto encoder (DAE) to suppress reverberation components and signal distortion. Afterwards, the output speech signal of the DAE is decoded by an acoustic model based on a deep neural network (DNN). The DAE and DNN are trained by using multi-condition speech data so that it is robust against various types of the acoustic environment. It is also necessary for the above processes to be performed in real time.

4.2 Speaker tracking with depth camera

To realize smooth interaction, it is essential for the system to correctly identify who talks to whom and if the user is giving his/her attention toward ERICA (Yu et al., 2015). In this demonstration, we track the user's location and head orientation in the 3D space by using the Kinect v2 sensor. The user localization enables ERICA to spot if there is

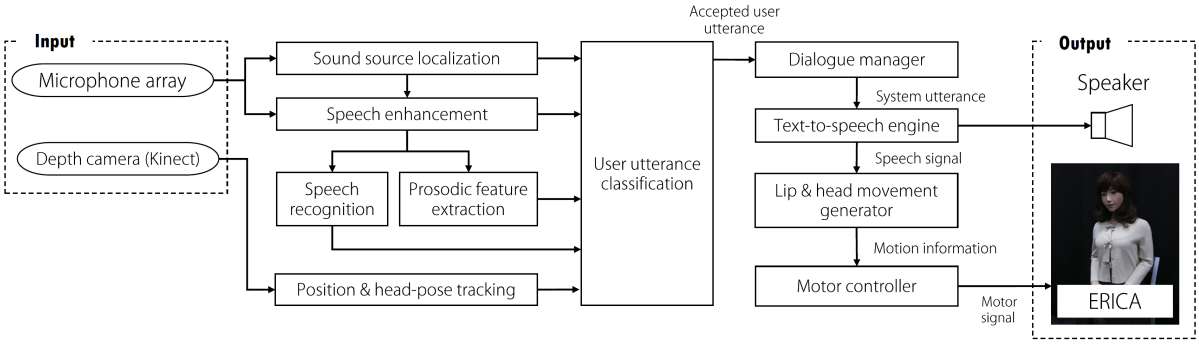

Figure 3: System architecture

a person who wants to interact with her. ERICA identifies if the user is speaking to her by the head orientation. This enables ERICA not to respond to the talking between people, for example when a person introduces ERICA to a guest standing in front of ERICA. ERICA accepts user utterances when the following are met: the user is standing in front of ERICA and looking at ERICA's face, and the sound source is coming from the direction of the user. This function is needed when we conduct a demonstration to many people such as open laboratory events.

4.3 Text-to-speech for ERICA

The speech of ERICA is generated by a text-to-speech engine developed for ERICA. It is based on the unit-selection framework from a database of many conversational-style utterances. It also contains many formulaic expressions and backchannels with a variety of prosodic patterns. At the same time, lip and head movements of ERICA are generated based on the prosodic information of the synthesized speech signals (Ishi et al., 2012; Sakai et al., 2015).

5 Conclusion

We demonstrate dialogues with ERICA who plays the two social roles. The human-like appearance of the android and the appropriate behaviors according to her social roles are combined to realize symbiotic human-robot interaction which is close to human-human interaction.

Acknowledgements

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project and JSPS KAKENHI Grant Number 15J07337.

References

P. Clancy, S. Thompson, R. Suzuki, and H. Tao. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of pragmatics*, 26(3):355–387.

D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, et al. 2014. SimSensei kiosk: A virtual human interviewer for healthcare decision support. In *Proc. Autonomous Agents and Multi-Agent Systems*, number 1, pages 1061–1068.

C. Ishi, H. Ishiguro, and N. Hagita. 2012. Evaluation of formant-based lip motion generation in tele-operated humanoid robots. In *Proc. IROS*, pages 2377–2382.

K. Sakai, C. Ishi, T. Minaot, and H. Ishiguro. 2015. Online speech-driven head motion generating system and evaluation on a tele-operated robot. In *Proc. ROMAN*, pages 529–534.

R. Schmidt. 1986. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas and Propagation*, 34(3):276–280.

G. Skantze and M. Johansson. 2015. Modelling situated human-robot interaction using IrisTK. In *Proc. SIGDIAL*, pages 165–167.

G. Wilcock and K. Jokinen. 2015. Multilingual WikiTalk: Wikipedia-based talking robots that switch languages. In *Proc. SIGDIAL*, pages 162–164.

T. Yamaguchi, K. Inoue, K. Yoshino, K. Takanashi, N. Ward, and T. Kawahara. 2016. Analysis and prediction of morphological patterns of backchannels for attentive listening agents. In *Proc. IWSDS*.

K. Yoshino and T. Kawahara. 2015. Conversational system for information navigation based on POMDP with user focus tracking. *Computer Speech and Language*, 34(1):275–291.

Z. Yu, D. Bohus, and E. Horvitz. 2015. Incremental coordination: Attention-centric speech production in a physically situated conversational agent. In *Proc. SIGDIAL*, pages 402–406.