Multimodal Interaction with the Autonomous Android ERICA

Divesh Lala[*]
Kyoto University Graduate
School of Informatics, Japan
lala@sap.ist.i.kyoto-u.ac.jp

Pierrick Milhorat
Kyoto University Graduate
School of Informatics, Japan
milhorat@sap.ist.i.kyoto-u.ac.jp

Koji Inoue
Kyoto University Graduate
School of Informatics, Japan
inoue@sap.ist.i.kyoto-u.ac.jp

Tianyu Zhao
Kyoto University Graduate
School of Informatics, Japan
zhao@sap.ist.i.kyoto-u.ac.jp

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

ABSTRACT

We demonstrate an interactive conversation with an android named ERICA. In this demonstration the user can converse with ERICA on a number of topics. We demonstrate both the dialog management system and the eye gaze behavior of ERICA used for indicating attention and turn taking.

CCS Concepts

•**Human-centered computing → Human computer interaction (HCI);**

Keywords

human-robot interaction, speech, gaze behavior

1. INTRODUCTION

Realistic humanoid robots, commonly known as androids, are becoming an increasingly popular interface for human-computer interaction. As part of the ERATO project, an android named ERICA is being developed [2]. Our goal is to make ERICA behave as a human-like conversational partner. This includes not only producing appropriate speech, but also being aware of her environment and producing appropriate non-verbal behaviors such as eye gaze. These attributes require ERICA to have multimodal capabilities.

In this demonstration, users will talk with a Japanese-speaking humanoid robot named ERICA, who uses several conversational modalities. ERICA has speech recognition capabilities which allow her to converse with the user on around 30 different topics related to her personality, likes and dislikes. Another modality is the gaze of ERICA, which she uses to indicate attention towards the user. We will demonstrate these capabilities of ERICA.

2. SYSTEM OVERVIEW

A photo of the demonstration system is shown in Figure 1. ERICA is seated at a table. In front of her is a table and a space where users may converse. A spherical microphone array is placed on the table to detect user speech. Speech detected with the microphone array is sent to Julius, an automatic speech recognition system [4].

A Kinect sensor is placed beside ERICA to track the position and attention of users to control ERICA's eye gaze behavior. The system architecture is shown in Figure 2. Multimodal synchronization is implemented to accurately combine the audio and motion data. ERICA's spoken utterances are managed through a text-to-speech engine which enables her to speak natural utterances as well as backchannels and even laughs.

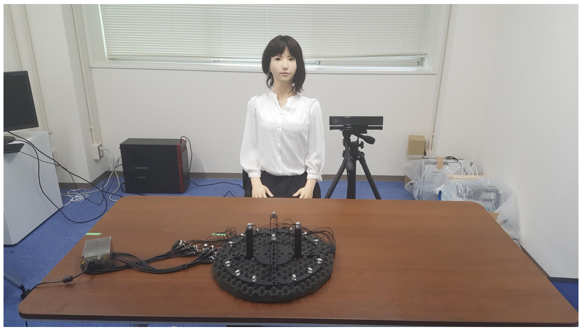

Figure 1: The setup of ERICA's demonstration system, with microphone array and Kinect sensor.

3. DIALOG MANAGEMENT

ERICA's dialog is controlled using Interaction Composer, a framework which allows us to design dialog flows for ERICA through a GUI [3]. However this framework still has

[*]International Research Fellow of the Japan Society for the Promotion of Science

ICMI'16, November 12–16, 2016, Tokyo, Japan
ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2998528

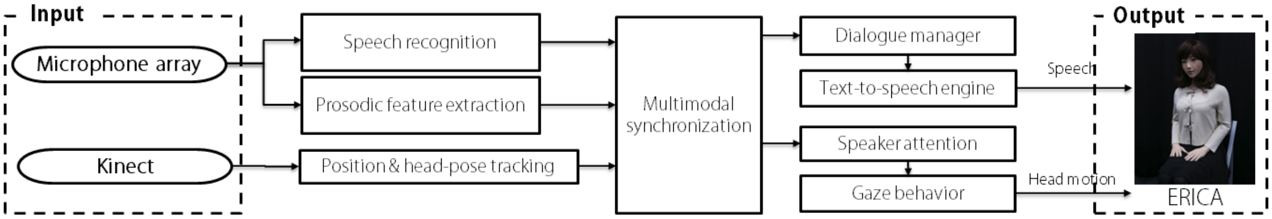

Figure 2: Generalized system architecture of ERICA.

limited functions, so we have integrated two conversational modes to allow for more natural conversation. These modes will be demonstrated to users.

The first conversational mode is where ERICA acts as the role of an attentive listener by giving appropriate feedback in the form of verbal backchannels. We extract the user's prosodic features as input and then classify whether or not a backchannel is required using a logistic regression model [5]. This mode is entered during a conversation where the user is asked by ERICA to explain something. While this explanation is occurring, ERICA will provide Japanese continuer utterances such as "Un", "Un, un" and "He-".

The statement response mode is used to repeat the user utterance as a follow-up question. It uses a model based on conditional random fields to extract the focus word [6] and then uses that focus word in the form of a question. For example, if the user says "Yesterday I went to the beach", ERICA in statement response mode will reply "Which beach?".

ERICA detects silence during a conversation and uses it to manage her behavior. For example, if ERICA asks a question and there is no reply from the user ERICA will reask the question. If ERICA asks about a topic and receives no user utterance, ERICA will suggest a random topic.

A general dialog flow of a conversation with ERICA is as follows. She begins with a short introduction then asks the user which topic they wish to talk about. We use an SVM-based model to estimate if their reply is a question or a statement. If the user replies with a question which is in the list of known topics, then ERICA will continue with a dialog flow for that topic. If the question does not match any topic, ERICA will say a short backchannel such as "E-?" to indicate she does not understand the question. If the user replies with a statement, ERICA enters statement response mode and will generate a reply using the focus word of the user's utterance. ERICA also responds to simple statements such as greetings and calling her name.

4. SPEAKER ATTENTION

One issue for multi-participant embodied conversational systems is to recognize if an utterance of a user is directed towards the embodied agent or at another party [7]. ERICA infers this by observing the head orientation of a speaker through the Kinect sensor. If a user speaks while looking at ERICA, she assumes that the utterance is for her and responds to it. If the speaker's head is turned away from ERICA she does not reply to the utterance, although she may still turn her head towards the speaker. Multiple people may interact with each other in front of ERICA, but she will reply only to utterances directed at her. We demonstrate this feature by chatting while not looking at ERICA.

5. GAZE BEHAVIOR

We will demonstrate two types of gaze behaviors of ERICA. The first indicates speaker awareness. Users are tracked in the conversational space by a Kinect sensor. ERICA will gaze towards any user from who she detects speech activity. If no speech activity is detected for a while, then ERICA will gaze at the nearest person to her.

The second behavior is to facilitate turn-taking during a conversation. The model used is based on previous research, which suggests a pattern of gaze aversion at the beginning of the speaker's turn and mutual gaze when passing the turn to their conversational partner [1]. ERICA will display gaze aversion through both slight eye and head movements, while mutual gaze is possible through the tracking of the head position of the user.

6. ACKNOWLEDGMENTS

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project

7. REFERENCES

[1] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu. Conversational gaze aversion for humanlike robots. In *Proc. 2014 ACM/IEEE Int. Conf. on Human-robot interaction*, pages 25–32. ACM, 2014.

[2] D.Glas, T.Minato, C.Ishi, T.Kawahara, and H.Ishiguro. Erica: The erato intelligent conversatioal android. In *Proc. ROMAN, 2016 (to appear)*, 2016.

[3] D. Glas, S. Satake, T. Kanda, and N. Hagita. An interaction design framework for social robots. In *Robotics: Science and Systems*, volume 7, page 89, 2012.

[4] A. Lee and T. Kawahara. Recent development of open-source speech recognition engine Julius. In *Proc. APSIPA ASC 2009*, pages 131–137, 2009.

[5] T.Kawahara, T.Yamaguchi, K.Inoue, K.Takanashi, and N.Ward. Prediction and generation of backchannel form for attentive listening systems. In *Proc. INTERSPEECH, 2016 (to appear)*, 2016.

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

[7] Z. Yu, D. Bohus, and E. Horvitz. Incremental coordination: Attention-centric speech production in a physically situated conversational agent. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 402, 2015.