Controlling Switching Pause Using an AR Agent for Interactive CALL System

Naoto Suzuki[1], Takashi Nose[1], Yutaka Hiroi[2], and Akinori Ito[1]

[1] Graduate School of Engineering, Tohoku University, Sendai, Japan
{naoto_s@spcom.ecei,tnose@m,aito@spcom.ecei}tohoku.ac.jp
[2] Department of Robotics, Osaka Institute of Technology, Osaka, Japan
hiroi@med.oit.ac.jp

Abstract. We are developing a voice-interactive CALL (Computer-Assisted Language Learning) system to provide more opportunity for better English conversation exercise. There are several types of CALL system, we focus on a spoken dialogue system for dialogue practice. When the user makes an answer to the system's utterance, timing of making the answer utterance could be unnatural because the system usually does not make any reaction when the user keeps silence, and therefore the learner tends to take more time to make an answer to the system than that to the human counterpart. However, there is no framework to suppress the pause and practice an appropriate pause duration.

In this research, we did an experiment to investigate the effect of presence of the AR character to analyze the effect of character as a counterpart itself. In addition, we analyzed the pause between the two person's utterances (switching pause). The switching pause is related to the smoothness of its conversation. Moreover, we introduced a virtual character realized by AR (Augmented Reality) as a counterpart of the dialogue to control the switching pause. Here, we installed the character the behavior of "time pressure" to prevent the learner taking long time to consider the utterance.

To verify if the expression is effective for controlling switching pause, we designed an experiment. The experiment was conducted with or without the expression. Consequently, we found that the switching pause duration became significantly shorter when the agent made the time-pressure expression.

Keywords: Computer-assisted language learning, English learning, Spoken dialogue system, Switching pause, Augmented reality.

1 Introduction

With internationalization, learners of English conversation in the non-English-speaking world are increasing. Recently, language learning systems using a computer (Computer-Assisted Language Learning, CALL) have been used by many people [1].

There are several types of CALL systems, such as listening, grammar learning and dialogue practice [2]. In this paper, we focus on a spoken dialogue system for dialogue practice. Using the speech recognition technology, this kind of systems enable the

C. Stephanidis (Ed.): HCII 2014 Posters, Part II, CCIS 435, pp. 588–593, 2014.

learners to learn pronunciation and grammar by themselves [3,4]. Moreover, the learners can make conversation practices through making dialogues with the system. However, when the learner makes an answer to the system's utterance, timing of the answer utterance could be unnatural.

The pause between the two person's utterances (switching pause) is related to the type of the dialogue [5], and acoustic or linguistic cues [6]. Exercise of making utterances within a proper switching pause duration seems to be important to train a skill to make natural English conversation.

Speaking in the target foreign language (L2) requires more cognitive load than speaking in the speaker's mother tongue (L1) [7]. Therefore, a learner need to make more effort to react the utterance from the dialogue counterpart quickly. When talking with a computer, the system usually does not make any reaction when the user keeps silence, and therefore the learner tends to take more time (longer switching pause) to make an answer to the system than that to the human counterpart. To make the human-system dialogue more similar to the human-human one, shortening the switching pause seems to be useful for the English conversation learning.

The purpose of this study is to create an interactive CALL system that enables a learner to make a dialogue where the learner's utterances are controlled to have an appropriate pause duration.

To this end, two features were introduced to the system. First, an Augmented Reality (AR) character was introduced as a dialogue counterpart. It is known that a virtual agent makes the conversation more similar to a human-human dialogue [8]. Secondly, we made the AR character to express a "time pressure" to prevent the learner taking long time to consider the utterance.

In this paper, we report development of a spoken dialogue system with AR character, and discuss the results of the examination about investigating effects of the controlling by the expression of time pressure.

2 Related Works

For the temporal control method of user's speech, a method using entrainment in a conversation was proposed. Entrainment is a phenomenon that two (or more) persons who are making a conversation get adapted to each other at several levels such as lexical choices [9], prosodic feature [10], and even the physiological status [11]. Suzuki et al. [12] investigated how to change user's utterance and impression of the system using an eyeball CG character. They also changed speed of the system's utterance. In that study, they reported that the faster the system spoke, the slower users did. Observation from those works could be used to control properties of the learner's utterances, such as speaking speed and vocabulary.

However, we did not choose a method using entrainment, but chose the method using explicit expression of time pressure, where the AR character behaves to require the learner's quick response. Out target dialogues are supposed to have only a few utterances of the learner; therefore, it is difficult to exploit the effect of entrainment, which appears after several turns.

3 Experimental System

The spoken dialogue system is a kind of the Question-and-Answer-based systems [13], and learners are required for remembering the scenarios in advance [3]. Table. 1 shows the example scenarios. We used Julius [14] as a speech recognizer. The acoustic model was trained from the ERJ corpus [15], and the N-gram language model was trained from all sentences in the scenarios. Festival [16] was used as a speech synthesizer.

A dialogue management is simple. At first, the system waits for the learner's utterance. Then the system recognizes the learner's utterance and utter a reply sentence. To manage recognition errors and the learner's mistakes in the utterance, the system checks correspondence of recognized words and the words in the scenario without considering the word order. If ratio of the coincided words is high enough, the learner's utterance is accepted.

Table 1. An example of the dialogue scenario (Buying a hat)

System	User
Hello. May I help you?	Yes, I'm looking for a hat. Do you have one?
Yes, we do. What kind do you want?	A green one.
Like this?	Yes, like that one. Can I see it?
Yes. Here you are. Would you like to buy it?	I'm sorry. This isn't exactly what I wanted.
How about another product?	No, thank you.

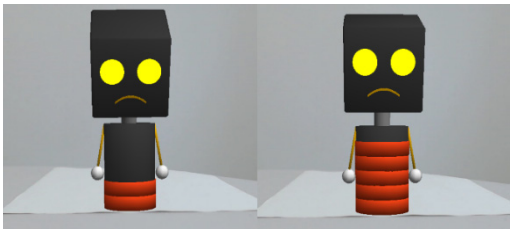

Fig. 1. The AR character with the expression of time pressure

Fig. 1 shows the AR character used in the experiment with the expression of time pressure. The expression of time pressure is to turn the body red from the bottom of the character to the top every 1 second. When the learner makes an utterance, the expression stops. At the same time, the character changes its face from frown to smile. In a condition without time pressure, the character is always smiling. In both cases, the character continues nodding to provide the learners with human-like impression [17].

4 Experiment

4.1 Experimental Conditions and Procedure

We conducted an experiment to investigate the effect of the time pressure expression. The experiment was conducted inside a soundproof chamber, and the learner's behavior was recorded by a video camera. Procedure of the experiment is as follows:

1. The participants were asked to remember two scenarios in 20 minutes as the preliminary training.
2. The participants were asked to take an examination to confirm whether the subject remembered the scenarios correctly.
3. The participants were asked to talk with the dialogue system following the scenarios.
4. The participants filled a questionnaire after the dialogue.

A subject wore a head-mounted display (HMD, SONY HMZ-T2). A Web camera was attached on the HMD. The AR character was superimposed on the image from the Web camera and displayed on the HMD. The subject looked at the AR character and talked with that character. We instructed the participants that when the character expressed the time pressure, the participant should respond to the system before the character's head turned into red. In Step 2, we confirmed that all the participants remembered the scenarios correctly.

4.2 Effectiveness of the AR Character

In this experiment, we compared the impression of the learner to the dialogue with and without the AR agent to confirm the effectiveness of the agent. We employed four undergraduate students as the participants. After finishing all dialogues, the participants answered the enquiry for choosing one of the two conditions (with or without the AR character) from six points of view: (1) Which condition did you feel easy for making dialogue? (2) Which condition did you make a dialogue more smoothly? (3) Which condition did you enjoy talking? (4) Which condition did you able to have more motivation to learn? (5) Which condition did you have more feeling of making real practice? (6) Which condition did you feel being stressed?

The results are shown in Fig. 2. As we can see, the condition with the AR character was preferred by most of the participants, which suggests the usefulness of the AR character in the context of the dialogue for English learning.

4.3 Effect of the Time-Pressure Expression

We employed ten participants (5 graduate students and 5 undergraduate students) who have studied English for around 10 years without English conversation learning.

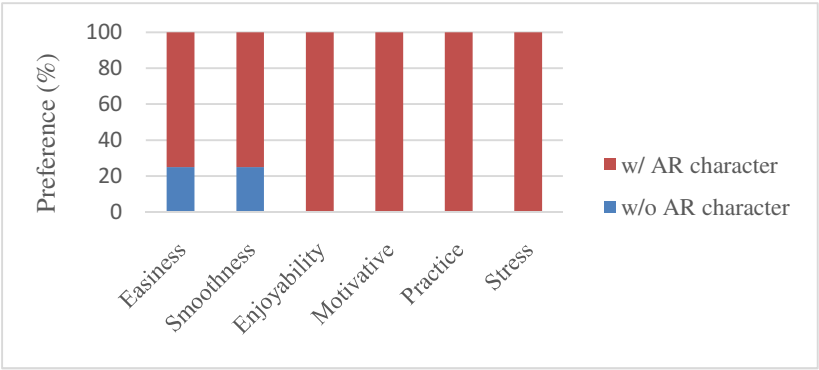

Fig. 2. Result of the preference enquiry

After the experiment, we measured the switching pauses of all the sessions from the recorded video data. Fig. 3 shows the average and standard deviation of the measured switching pauses in the two conditions. The standard deviation is indicated by the error bar. We conducted t-test and found a significant difference between the two conditions at the 1% level. As can be seen from Fig. 3, the switching pause became about 500 ms shorter by introducing the time pressure expression. This result suggested that we can control the timing of the learner's utterance by the AR character's behavior.

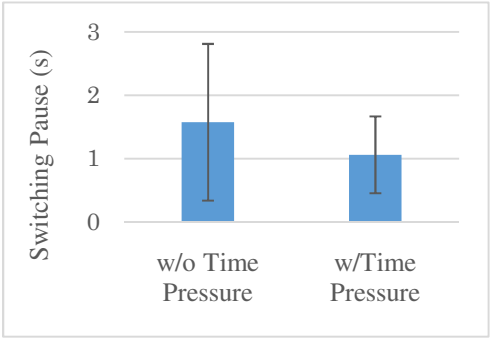

Fig. 3. Difference of the switching pause between with and without the time pressure

5 Conclusions

In this paper, we proposed a CALL system based on the spoken dialogue system and augmented reality. To control the timing of the learner's utterance when learning English conversation with AR character, we focused on the switching pause. We introduced an expression of the time pressure as a means of controlling the switching pause, and designed the experiment to investigate the effect of the expression. Consequently, we found significant difference of the switching pause duration between with and without the expression.

Acknowledgment. The present study was supported in part by a JSPS Grant-in-Aid for Challenging Exploratory Research 24652111.

References

1. Chujo, K., Nishigaki, C., Uchibori, A., Yamazaki, A.: Developing a beginning-level CALL system and its effect on college students' communicative proficiency. J. of the College of Industrial Technology, Nihon University 38, 1–16 (2005)
2. Eskenazi, M.: An overview of spoken language technology for education. Speech Communication 51(10), 832–844 (2009)
3. Kweon, O.-P., Ito, A., Suzuki, M., Makino, S.: A grammatical error detection method for dialog-based CALL system. J. of Natural Language Processing 12(4), 137–156 (2005)
4. Anzai, T., Ito, A.: Recognition of utterances with grammatical mistakes based on optimization of language model towards interactive CALL systems. In: Proc. APSIPA ASC (2012)
5. Trimboli, C., Walker, M.B.: Switching pauses in cooperative and competitive conversations. J. of Experimental Social Psychology 20(4), 297–311 (1984)
6. Miura, I.: Switching pauses in adult-adult and child-child turn takings: An initial study. J. of Psycholinguistic Research 22(3), 383–395 (1993)
7. Nation, P.: The role of the first language in foreign language learning. The Asian EFL J. 5(4) (2003)
8. Miyake, S., Ito, A.: A spoken dialogue system using virtual conversational agent with augmented reality. In: Proc. APSIPA ASC (2012)
9. Brennan, S.E.: Lexical Entrainment in Spoken Dialog. In: Proc. Int. Symp. on Spoken Dialog, pp. 41–44 (1996)
10. Levitan, R., Hirschberg, J.: Measuring Acoustic-Prosodic Entrainment with respect to Multiple Levels and Dimensions. In: Proc. Interspeech (2011)
11. Watanabe, T., Okubo, M.: Physiological analysis of entrainment in communication. J. of Information Processing 39(5), 1225–1231 (1998)
12. Suzuki, N., Kakei, K., Takeuchi, Y., Okada, M.: Effects of the speed of hummed sounds on human-computer interaction. J. of Human Interface Society 5(1), 113–122 (2003)
13. Nisimura, R., Lee, A., Saruwatari, H., Shikano, K.: Public speech-oriented guidance system with adult and child discrimination capability. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing, vol. I, pp. 433–436 (2004)
14. Lee, A., Kawahara, T., Shikano, K.: Julius — an open source real-time large vocabulary recognition engine. In: Proc. European Conf. on Speech Communication and Technology (EUROSPEECH), pp. 1691–1694 (2001)
15. Minematsu, N., Tomiyama, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M., Makino, S.: Development of English speech database read by Japanese to support CALL research. In: Proc. Int. Conf. Acoustics, pp. 557–560 (2004)
16. The Festival Speech Synthesis System, `http://www.cstr.ed.ac.uk/projects/festival/`
17. Hiroi, Y., Ito, A.: Evaluation of head size of an interactive robot using an augmented reality. In: Proc. World Automation Congress (2010)