Journal of Universal Computer Science, vol. 13, no. 2 (2007), 186-198
submitted: 31/7/06, accepted: 15/1/07, appeared: 28/2/07 © J.UCS

Entrainment in the Rate of Utterances in Speech Dialogs between Users and an Auto Response System

Takanori Komatsu
(Future University-Hakodate, Japan
komatsu@fun.ac.jp)

Koji Morikawa
(Matsushita Electric Industrial Co, Ltd., Japan
morikawa.koji@jp.panasonic.com)

Abstract: Entrainment, a physical phenomenon in which one individual's expressed information synchronizes with another's and vice versa, can be observed between two communicators who are interacting naturally. In this study, we focused on the "rate of utterances" as communicators' expressed information and then conducted an experiment to observe whether or not entrainment naturally occurs in the rate of utterances in speech dialogs between users and an auto response system. Specifically, participants were asked to read given dialog scripts with an auto response system that replied with different rates of utterances. The results revealed that 1) when the system's rate of utterances increased, the participants produced faster rates of utterances, 2) when the system's rates decreased, participants spoke at slower rates. These results suggest entrainment in the rate of utterances naturally occurs in speech dialogs between participants and an auto response system.

Key Words: entrainment, speech interface, rate of utterances
Category: H.1.3, H.5.2

1 Introduction

Various kinds of user interfaces have been developed to improve the usability of home and information appliances. In particular, developments in speech interfaces have drawn substantial interest because we usually express speech sounds without any special effort, and we can do it while engaging in other tasks (e.g., cooking, driving a car)[Nass and Brave 2005]. However, the main reason for this attention could derive from a drastic improvement in sound recognition technologies. While many studies have been carried out to increase the recognition rates in sound recognition technologies, studies for adapting technology to fluctuations in the rate of utterances have been the subject of particular focus; in general, the rate of utterances varies between individuals, and it varies even for one person[Maekawa et al. 2002]. These fluctuations in the rate of utterances drastically decrease recognition rates[Shinozaki and Furui 2003]. Specifically, displacement and omission errors were reported to occur when sound recognition systems receive a faster rate of utterance, while insert errors occur when these systems receive a slower rate[Nanjo and Kawahara 2002]. To resolve these problems, Nanjo et al.[Nanjo et al. 2001] proposed “speaking rate dependent acoustic modeling” by means of Baysian Networks, and Okuda et al.[Okuda et al. 2002] created a method that selects appropriate window lengths using likelihood criterion. These studies take an approach where speech interfaces adapt to people’s rates of utterances.

However, some studies take another approach: interfaces explicitly prompting an appropriate rate of utterances from people. For example, Jousson et al.[Joussson et al. 2004] proposed the methodology that the interface immediately apologize to users like, “Sorry, I cannot recognize your speech. So please say it again” when it fails to recognize their speech correctly, as one in the series of “Media Equation” studies[Reeves and Nass 1996]. However, this method has a notable disadvantage: these apologies would inform the users that “this interface is less than ideal,” so the users might become frustrated with having to speak the same sentences repeatedly and lose interest in talking with this interface.

For this study, we adopted the other approach, that is, an interface that implicitly induces an appropriate rate of utterances from people. We specifically focused on entrainment, a physical phenomenon in which one individual’s expressed information synchronizes with another’s and vice versa. It is said that this phenomenon can be observed between two communicators who are interacting naturally[Condon and Sander 1974][Miyake 2002]. Many researchers have developed various interactive robots that are designed to use entrainment to interact naturally with users[Ono et al. 2001][Watanabe and Okubo 1997] [Watanabe et al. 2004][Wesugi et al. 2004]. Some studies have also focused on entrainment in synchronization of speech sounds. For example, Nagaoka et al. [Nagaoka et al. 2003] reported that the speakers in cooperative speech communications shared synchronized back-channel feedback or duration of utterances with their partners. However, few studies have focused on entrainment in the rate of utterances. If we could observe this entrainment, our findings could be applied to a speech interface that can implicitly prompt users to produce the appropriate rate of utterances.

We conducted an experiment as part of a basic trial to create such a speech interface to observe whether or not the entrainment in the rate of utterances naturally occurs in speech dialogs between users and artifacts. We focused on an auto response system (i.e., a telephone ticket reservation system) as an artifact with which users interact. We present our acquired experimental results and then discuss the occurrence of the entrainment in the rate of utterances, the aspects that affect the occurrence of the entrainment, and future trials that will enable us to create a sophisticated speech interface system.

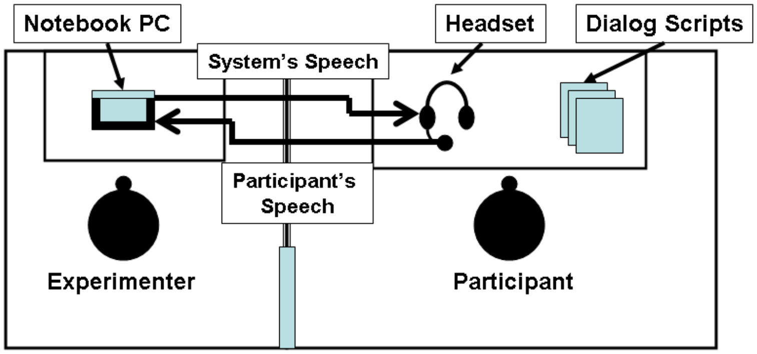

Figure 1: Experimental Setting

2 Experiment

2.1 Participants

Twenty-seven Japanese university students (16 men and 11 women; 19–24 years old) participated. A hearing test established that none of the participants had any hearing problems.

2.2 Setting and Procedure

First, an experimenter informed the participants that this experiment was to evaluate an auto response system and that the participants' task was to read given dialog scripts with this response system. We prepared three different scripts. The first script (script 1) was about ordering train tickets (Ms. A is a traveler, and Mr. B is a station officer, see Table 1); the second one (script 2) was about a typical conversation between a high school girl and boy (like the dialog skits used in language schools, Table 2), and the third (script 3) was about ordering hamburgers at a fast-food shop (Ms. A is a shop clerk, and Mr. B is a customer, Table 3). The participants were asked to read Mr. B's part, while the response system played the speeches of Ms. A.

The auto response system used in this experiment just played previously recorded speeches of Ms. A 0.5 of a second after detecting the end of the human participants speech. This "0.5 second" was fixed so that the duration of turn taking was constant throughout this experiment. This auto response system was implemented as scripting language (csh) on a linux-installed notebook PC (IBM ThinkPad T30) so that it did not have any graphical information

Ms. A: A round trip express ticket to Sapporo, please. I'd like a reserved seat.
Mr. B: When are your departure and return dates?
A: My departure is tomorrow, and my return is the day after.
B: Would you like a smoking or non-smoking seat?
A: Non-smoking please.
B: Here are your tickets. One is for your departure and the other is for your return trip.
A: Is it possible to take a different train tomorrow?
B: You can, but your seat reservation would be invalidated. Please try to catch your scheduled train.

Note: The original script was written in Japanese.

Table 1: Dialog script 1 used in this experiment

for participants. For Ms. A's speeches, we used recorded speeches of only one woman who read all three dialog scripts. We used the sound authoring software Cool Edit 2000 to prepare three different rates for Ms. A's utterances by expanding or contracting their durations without any modification of pitch values. Specifically, we prepared an 80% duration as a “high-speed” rate of utterance (H condition), 100% duration (no expansion or contraction of recorded sounds) as a “middle-speed” rate (M condition), and 120% duration as a “low-speed” rate (L condition). Note that we used Ms. A's recorded sound without any modification as the M condition rate, so the rate of utterances for all sentences was not the same throughout.

Participants were asked to wear a headset and read three dialog scripts aloud, while an experimenter went into another room to operate an auto response system (Figure 1). Before conducting the experiment, participants were asked to read these scripts silently so that they would be able to minimize their number of errors. Each participant read nine dialogs aloud with the system (3 different scripts x 3 different rates for Ms. A's speech). The order of these dialogs was counterbalanced for the participants. We observed a total of 243 dialogs (27 participants x 9 dialogs). For this paper, the rate of utterances was defined as the number of syllables in one speech divided by the speech length in [syllable/s]. These rate-of-utterance values were manually calculated from the recorded dialog sounds and scripts.

Ms. A: What time do you usually get up?
Mr. B: I get up about a half past seven on weekdays.
A: Do you have morning classes every weekday?
B: No, I listen to an English radio program.
A: You are so diligent!
B: If I have a chance, I want to study abroad. Learning English is just something that takes time.

Note: The original script is written in Japanese.

Table 2: Dialog script 2 used in this experiment

Ms. A: Can I help you?
Mr. B: I'll have a chicken burger combo.
A: What kind of soft drink would you like?
B: Coke please.
A: You can have french fries or chicken nuggets for your side order.
B: I'll have french fries.
A: All right. That comes to 504 yen.
B: I have only a 10,000 yen bill. Is that OK?

Note: The original script is written in Japanese.

Table 3: Dialog script 3 used in this experiment

2.3 Results

2.3.1 Analyzing 243 dialogs

We divided the total of 243 dialogs into three conditions (H, M, and L) and calculated the participants' average rate of utterances across 81 dialogs for each condition (regardless of the kind of dialog script). The system's average rate of utterance was 11.533 [syllable/s] in the H condition (81 dialogs), 8.830 [syllable/s] in the M condition, and 7.740 [syllable/s] in the L condition. While the participants' average rate of utterances was 9.643 [syllable/s] in the H condition, the average rate in the M condition was 9.533 [syllable/s], and the rate in the L condition was 9.188 [syllable/s] (Figure 2). Here, the results revealed significant differences in these three conditions (F(2,160)=28.43, p<.01(**)). The Newman-Keuls test revealed a significant difference between the H and

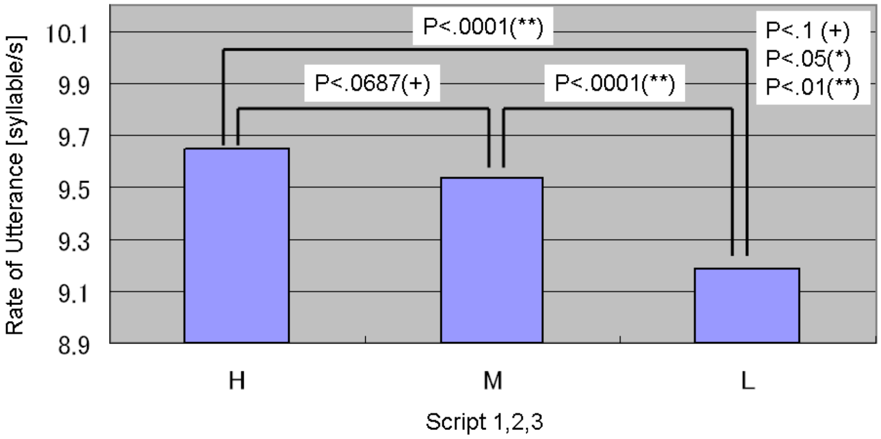

Figure 2: Participants' rate of utterances in three different conditions (H, M, and L)

L conditions (F(1,80)=45.67, p<.01 (**)) and between the M and L conditions (F(1,80)=31.34, p<.01 (**)). This test approached significance for H and M conditions (F(1,80)=3.40, p<.1(+)).

To sum up, the participants adapted their rate of utterances to follow the system's rate. Therefore, this result suggests entrainment in the rate of utterances occurs in speech dialogs between participants and the auto response system.

2.3.2 Analyzing 81 dialogs in script 1

As described earlier, we analyzed the 243 dialogs' average rate of utterances in each condition regardless of the kind of dialog scripts. Next, we observed whether or not the different scripts show different entrainment patterns.

We first focus on script 1, which is a dialog script about "buying a train ticket." The 81 dialogs performed using script 1 were divided according to the three different conditions (H, M, and L), and the average rates were calculated. The system's average rate of utterances was 11.356 [syllable/s] in the H condition, 9.562 [syllable/s] in the M condition, and 7.913 [syllable/s] in the L condition.

The participants' average rate of utterances when speaking with this system under the H condition was 9.562 [syllable/s], 9.441 [syllable/s] in the M condition, and 9.011 [syllable/s] in the L condition (Figure 3). These results revealed significant differences in speaking under these three conditions (F(2,52)=17.21,

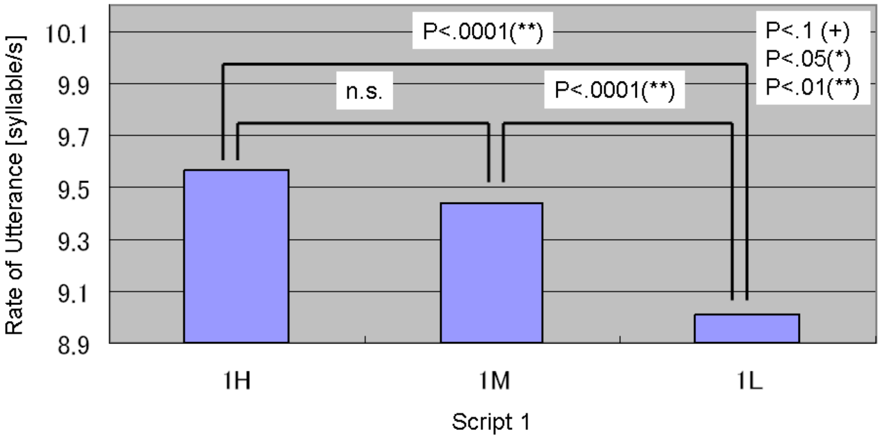

Figure 3: Participants' rate of utterances for three different conditions when performing script 1

p<.01 (**)), and the Newman-Keuls test indicated significant differences between the H and L conditions (F(1,26)=45.67, p<.01 (**)) and between the M and L conditions (F(1,26)=16.59, p<.01 (**)). However, this test revealed no significant difference between the H and M conditions (F(1,26)=1.83, n.s.).

2.3.3 Analyzing 81 dialogs of script 2

The 81 dialogs for script 2, which is a dialog script of "everyday conversation between students," were divided according to the three conditions, and the average rate of utterances for each was calculated. The system's average rate of utterance was 12.124 [syllable/s] in the H condition, 9.942 [syllable/s] in the M condition, and 7.989 [syllable/s] in the L condition.

The participants' average rates were 10.096 [syllable/s] in the H condition, 9.953 [syllable/s] in the M condition, and 9.548 [syllable/s] in the L condition (Figure 4). These results indicated significant differences between these three conditions (F(2,52)=12.45, p<.01 (**)). The Newman-Keuls test revealed significant differences between the H and L conditions (F(1,26)=21.95, p<.01 (**)) and between the M and L conditions (F(1,26)=14.38, p<.01 (**)), but revealed no significant difference between the H and M conditions (F(1,26)=1.57, n.s.).

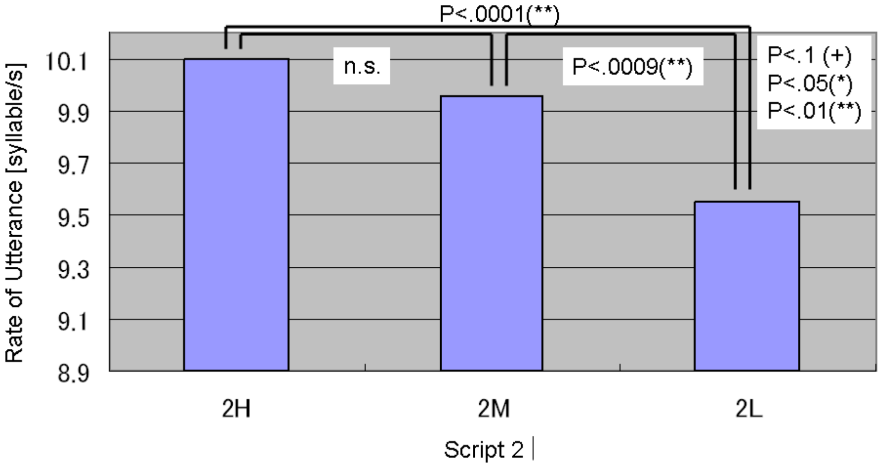

Figure 4: Participants' rate of utterances for three different conditions when performing script 2

2.3.4 Analyzing 81 dialogs in Script 3

The 81 dialogs in script 3, which is a dialog script about "ordering hamburgers at a fast food shop," were divided according to the same three conditions, and the average rate of utterances for each was calculated. The system's average rate of utterances for script 3 was 10.724 [syllable/s] in the H condition, 9.001 [syllable/s] in the M condition, and 7.060 [syllable/s] in the L condition.

The participants' average rates were 9.272 [syllable/s] in the H condition, 9.205 [syllable/s] in the M condition, and 9.007 [syllable/s] in the L condition (Figure 5). Just as with scripts 1 and 2, the results indicated significant differences between these three conditions (F(2,52)=3.48, p<.05 (*)). The Newman-Keuls test revealed significant differences between the H and L conditions (F(1,26)=4.80, p<.05 (*)) and between the M and L conditions (F(1,26)=4.38, p<.05 (*)) but revealed no significant difference between the H and M conditions (F(1,26)=.25, n.s.).

3 Discussion

3.1 Causes of the entrainment in the rate of utterances

The results of the experiment revealed that when the system's rate of utterance increased (the H condition), participants produced faster rates of utterances,

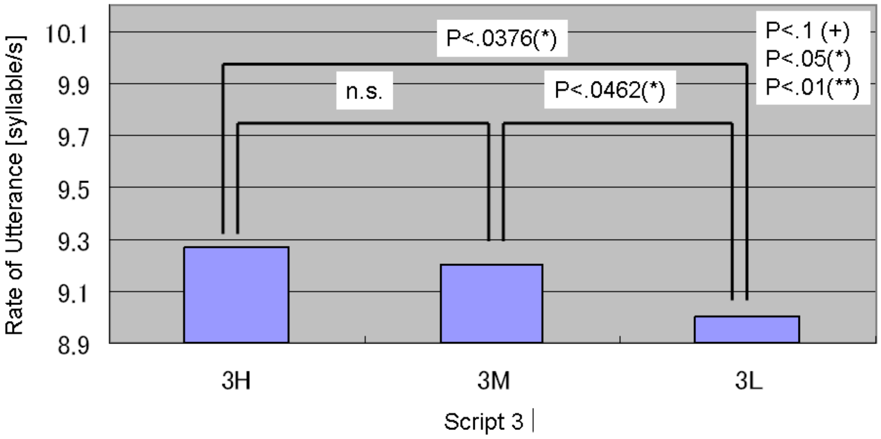

Figure 5: Participants' rate of utterances for three different conditions when performing script 3

and when the system's rate decreased, people used slower rates. This suggests that entrainment in the rate of utterances in speech dialogs occurs between participants and the auto response system.

Although significant differences were found in the rate of utterances between 1) the M and L conditions and 2) the H and L conditions, these differences were not found between the H and M conditions. The reason for this last occurrence seemed to be that Ms. A's recorded utterances (used as the M condition utterances) were originally faster. Actually, the system's (or Ms. A's) average rate of utterances in the M condition was about 9.2 [syllable/s], while the average rate of utterances in the Japanese discourse was reported to be about 8.0 [syllable/s][NIJL 2004, Murakami 2001]. Participants therefore did not have sufficient "margin" to adapt their rate of utterance to the system's faster rates, but did have enough margin to adapt to the system's slower ones.

Figure 6 shows that different dialogs have a different average rate of utterances, i.e., the average rate of utterances in the three conditions in script 2 generally had the fastest rates, and those in script 3 produced the lowest rates. The total of 243 dialogs was then divided into three scripts, and the average rate of utterance for each script was statistically analyzed. We found significant differences between the three dialogs (F(2,160)=35.40, p<.01 (**); script 2 (fastest) > script 1 > script 3 (lowest)). The reasons for these occurrences are still unclear; however, we at least have the following two possible explanations

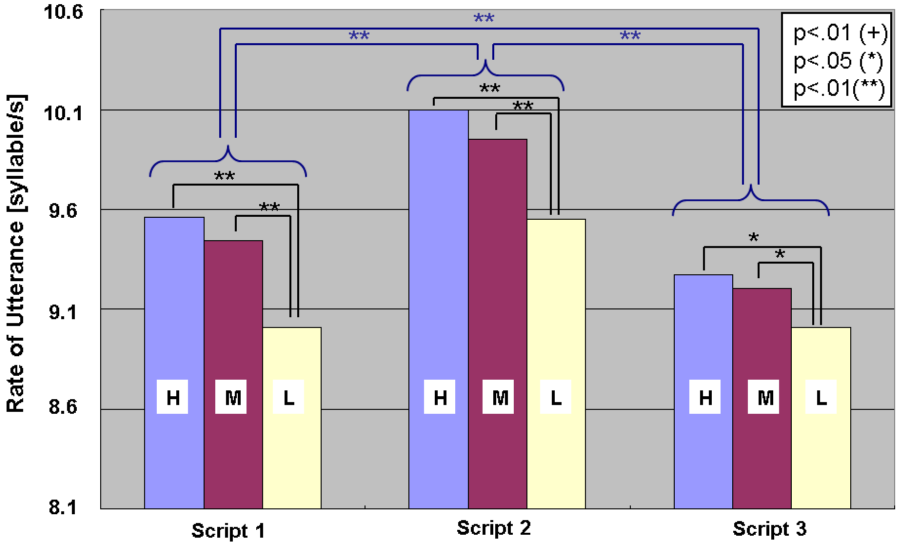

Figure 6: Participants' rate of utterances for three different conditions when performing each dialog

for them:

1. The recorded speech (Ms. A's speech) originally had different rates in each dialog, e.g., Ms. A spoke faster in script 2 and slower in script 3.

2. The effects of the "content" of each script were different, e.g., while the content of script 2 facilitated faster rates of speech, the content of script 3 facilitated lower rates.

To resolve these issues, we need to clarify this phenomenon in follow-up studies by making Ms. A's rate of utterances in each condition equal. The results of these studies will reveal the reason(s) for the different dialogs having different average rates.

The setting of this experiment was that participants were asked to read given scripts. Therefore, one could point out that this setting was completely different from natural speech observed in daily conversations. Another experiment needs to be conducted to observe more natural conversation without any restrictions being required. This will allow us to analyze the occurrence of entrainment in the rate of utterances in greater detail.

3.2 Future trials to create a speech interface system that can implicitly prompt users to achieve the appropriate rate of utterances

The results of an experiment where the appropriate rate of utterances from users can be prompted implicitly by means of entrainment in the rate of utterances would be a significant factor in creating a sophisticated speech interface system that can adapt to users' fluctuation in speech sounds. For example, when the system fails to recognize a user's speech because of fluctuations in the rate of utterances, this system could prompt the user to produce the appropriate rate by saying to them, something like this "Umm...I didn't catch your point...also there are a lot of noises in the background." In this case, the users should unconsciously change their rates to follow the system's rate due to entrainment effects but without any cognitive loads. The results of such an experiment will contribute to improving the usability of the speech interface and the recognition rates of the sound recognition system.

However, to create such a speech interface system, the following issues need to be clarified and resolved.

- What kinds of sentences should the interface speak to users? The system needs to express a sentence that implicitly prompts users to speak the same sentence again but not in an explicit way, like requesting users directly, "please speak more slowly."
- How much of a change in the rate of utterances by the users can the interface cause? In our experimental setting, we could not observe whether participants could change their rate of utterances to match the system's, particularly after the system changed its rate.

To clarify the latter issue, we are planning follow-up studies to observe how participants can adapt their rate of utterance to the system when its rate changes drastically, e.g., the system plays the H condition recorded sound and then in turn plays the L condition sound. If participants could change their rate of utterances to follow the system's drastically changing rates at every turn, we could strongly argue that entrainment naturally occurs in the rate of utterances in speech dialogs between users and an auto response system, and moreover, the results of the experiments can be undoubtedly applied to our desired speech interface system, one that can implicitly prompt users to produce the appropriate rate of utterances.

4 Conclusions

In this study, we conducted an experiment to observe whether entrainment in the rate of utterances naturally occurs in speech dialogs between users and artifacts. Based on the acquired experimental results, we discussed the occurrence of entrainment in the rate of utterances, the causes that affect this occurrence, and future trials to create a sophisticated speech interface system.

In this experiment, participants were asked to read three dialog scripts with a response system expressing three different rates of utterances. For the system's three different rates, we prepared an 80% duration of speech as a higher rate of utterance (H condition), a 100% duration (no expansion or contraction of recorded sound) as a middle rate (M condition), and a 120% duration as a lower rate (L condition). Each participant read nine dialogs aloud with an auto response system (3 different scripts x 3 different rates for system speech).

The results of these experiments revealed that when the system's rate of utterance increased (H condition), participants increased their rates of utterances, and when the system's rate decreased, people used slower rates. This suggests entrainment naturally occurs in the rate of utterances in speech dialogs between participants and the auto response system. The results of an experiment where users are prompted to produce an appropriate rate of utterances implicitly by means of entrainment in the rate of utterances would significantly assist us in creating a speech interface system that can adapt to users' fluctuations in speech sounds.

However, in this experimental setting, the rate of the system's speech sounds was faster than that in normal Japanese discourses, so participants did not show strong entrainment for this system's faster speech. Moreover, the system's speech sounds were not equalized in each condition, so the different dialogs had a different average rate of utterances. Hereafter, subsequent studies should standardize the system's rate of utterances in each condition to clarify these issues.

We are planning another study to observe how participants can adapt their rate of utterances to a system whose rates change drastically. The results of this subsequent study will clarify whether or not participants can change their rate of utterances to follow the system's drastically changing rates. If so, we will have a strong argument for the occurrence of entrainment in the rate of utterances in speech dialogs between users and an auto response system, and the results can undoubtedly be applied to our desired speech interface system, one that can implicitly prompt users to produce the appropriate rate of utterances.

References

[Condon and Sander 1974] Condon, S. W., and Sander, L. W.: Neonate movement in synchronized with adult speech: Interaction participation and language acquisition, Science, Vol. 183, 99–101 (1974).

[Joussson et al. 2004] Jousson, I, M., Nass, C., Endo, J., Reeves, B., Harris, H., and Ta, J. L.: Don't blame me, I'm only a driver: Impact of blame attribution on attitudes and attention to driving task, In Extended Abstracts of the 2004 Conference on Human Factors and Computing Systems, pp. 1219-1222 (2004).

[Maekawa et al. 2002] Maekawa, K., Koiso, H., Kikuchi, H., and Yoneyama, K.: Use of a large-scale spontaneous speech corpus in the study of linguistic variation, In Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003), pp. 643-646, (2003).

[Miyake 2002] Miyake, Y.: Co-creation system; Cognitive Processing, vol. 3, pp. 131–136 (2002).

[Murakami 2001] Murakami, J.: Study of Spontaneous Speech Recognition based on Stochastic Language Modeling, PhD Thesis of Toyohashi University of Technology (in Japanese), (2001).

[Nagaoka et al. 2003] Nagaoka, C., Komori, M., Draguna, M., Kawase, S., Yuki, M., Kataoka, T., and Nakamura, T.: Mutual Congruence of Vocal Behavior in Cooperative Dialogues: Comparison between Receptive and Assertive Dialogues (In Japanese); In Proceedings of Human Interface Symposium 2003, pp. 167–170 (2003).

[Nanjo et al. 2001] Nanjo, H., Kato, K., and Kawahara, T.: Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition. In Proceedings of EUROSPEECH 2001, pp. 2531–2534, (2001).

[Nanjo and Kawahara 2002] Nanjo, H., and Kawahara, T.: Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition. In Proceedings of IEEE-ICASSP, pp. 725–728, (2002).

[Nass and Brave 2005] Nass, C., and Brave, S.: Wired for Speech, The MIT Press (2005).

[NIJL 2004] The National Institute for Japanese Language.: the Corpus of Spontaneous Japanese, http://www2.kokken.go.jp/c̃sj/public, (2004).

[Okuda et al. 2002] Okuda, K., Kawahara, T., and Nakamura, S.: Speaking rate compensation based on likelihood criterion in acoustic model training and decoding. In Proceedings of ICSLP2002, (2002).

[Ono et al. 2001] Ono, T., Imai, M., and Ishiguro, H.: A Model of Embodied Communications with Gestures between Humans and Robots; In Proceedings of the 23rd Annual Meeting of the Cognitive Science Society (CogSci2001), pp. 732–737 (2001).

[Reeves and Nass 1996] Reeves, B., and Nass, C.: The Media Equation, Cambridge University Press, (1996).

[Shinozaki and Furui 2003] Shinozaki, T., and Furui, S.: Hidden mode HMM using Bayesian network for modeling speaking rate fluctuation, In Proceedings of Automatic Speech Recognition and Understanding, pp. 417–422 (2003).

[Watanabe and Okubo 1997] Watanabe, T., and Okubo, M.: Evaluation of the Entrainment Between a Speaker's Burst-Pause of Speech and Respiration and a Listener's Respiration in Face-to-Face Communication; In Proceedings of RO-MAN'97, pp. 392–397 (1997).

[Watanabe et al. 2004] Watanabe, T., Okubo, M., Nakashige, M., and Danbara, R.: InterActor: Speech-Driven Embodied Interactive Actor; International Journal of Human-Computer Interaction, Vol. 17(1), pp. 43–60 (2004).

[Wesugi et al. 2004] Wesugi, S., Katayama, T., and Miwa, Y.: Virtually Shared "Lazy Susan" Based on Dual Embodied Interaction Design; In Proceedings of INTERACTION 2004, pp. 263–270 (2004).