Non-humanlike Spoken Dialogue: A Design Perspective

Kotaro Funakoshi
Honda Research Institute Japan Co., Ltd.
8-1 Honcho, Wako
Saitama, Japan
funakoshi@jp.honda-ri.com

Mikio Nakano
Honda Research Institute Japan Co., Ltd.
8-1 Honcho, Wako
Saitama, Japan
nakano@jp.honda-ri.com

Kazuki Kobayashi
Shinshu University
4-17-1 Wakasato, Nagano
Nagano, Japan
kby@shinshu-u.ac.jp

Takanori Komatsu
Shinshu University
3-15-1 Tokida, Ueda
Nagano, Japan
tkomat@shinshu-u.ac.jp

Seiji Yamada
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda
Tokyo, Japan
seiji@nii.ac.jp

Abstract

We propose a non-humanlike spoken dialogue design, which consists of two elements: non-humanlike turn-taking and non-humanlike acknowledgment. Two experimental studies are reported in this paper. The first study shows that the proposed non-humanlike spoken dialogue design is effective for reducing speech collisions. It also presents pieces of evidence that show quick humanlike turn-taking is less important in spoken dialogue system design. The second study supports a hypothesis found in the first study that user preference on response timing varies depending on interaction patterns. Upon receiving these results, this paper suggests a practical design guideline for spoken dialogue systems.

1 Introduction

Speech and language are owned by humans. Therefore, spoken dialogue researchers tend to pursue a humanlike spoken dialogue. Only a few researchers positively investigate restricted (*i.e.*, non-humanlike) spoken dialogue design such as (Fernández et al., 2007).

Humanlikeness is a very important concept and sometimes it is really useful to design machines / interactions. Machines are, however, not humans. We believe humanlikenss cannot be the dominant factor, or gold-standard, for designing spoken dialogues.

Pursuing humanlikeness has at least five critical problems. (1) *Cost*: in general, humanlikeness demands powerful and highly functional hardware and software, and highly integrated systems requiring top-grade experts both for development and maintenance. All of them lead to cost overrun. (2) *Performance*: sometimes, humanlikeness forces performance to be compromised. For example, achieving quick turn-taking which humans do in daily conversations forces automatic speech recognizers, reasoners, etc. to be compromised to enable severe real-time processing. (3) *Applicability*: differences in cultures, genders, generations, situations limit the applicability of a humanlike design because it often accompanies a rigid character. For example, Shiwa et al. (2008) succeeded in improving users' impression for slow responses from a robot by using a filler but obviously use of such a filler is limited by social appropriateness. (4) *Expectancy*: humanlike systems induce too much expectancy of users that they are as intelligent as humans. It will result in disappointments (Komatsu and Yamada, 2010) and may reduce users' willingness to use systems. Keeping high willingness is quite important from the viewpoint of both research (for collecting data from users to improve systems) and business (for continuously selling systems with limited functionality). (5) *Risk*: Although it is not verified, what is called the uncanny valley (Bartneck et al., 2007) probably exists. It is commonly observed that people hate imperfect humanlike systems.

We try to avoid these problems rather than overcome them. Our position is positively exploring non-humanlike spoken dialogue design. This pa-

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 176–184,
The University of Tokyo, September 24-25, 2010.

per focuses on its two elements, *i.e.*, decelerated dialogues as non-humanlike turn-taking and an artificial subtle expression (ASE) as non-humanlike acknowledgment[1], and presents two experimental studies regarding these two elements. ASEs, defined by the authors in (Komatsu et al., 2010), are simple expressions suitable for artifacts, which intuitively notify users about artifacts' internal states while avoiding the above five problems.

In Section 2, the first study, which was previously reported in (Funakoshi et al., 2010), is summarized and shows that the proposed non-humanlike spoken dialogue design is effective for reducing speech collisions. It also presents pieces of evidence that shows quick humanlike turn-taking is less important in designing spoken dialogue systems (SDSs). In Section 3, the second study, which is newly reported in this paper, shows a tendency supporting a hypothesis found in the first study that user preference on response timing varies depending on interaction patterns. Upon receiving the results of the two experiments, a design guideline for SDSs is suggested in Section 4.

2 Study 1: Reducing Speech Collisions with an Artificial Subtle Expression in a Decelerated Dialogue

An important issue in SDSs is the management of turn-taking. Failures of turn-taking due to systems' end-of-turn misdetection cause undesired speech collisions, which harm smooth communication and degrade system usability.

There are two approaches to reducing speech collisions due to end-of-turn misdetection. The first approach is using machine learning techniques to integrate information from multiple sources for accurate end-of-turn detection in early timing. The second approach is to make a long interval after the user's speech signal ends and before the system replies simply because a longer interval means no continued speech comes. As far as the authors know, all the past work takes the first approach (*e.g.*, (Kitaoka et al., 2005; Raux and Eskenazi, 2009)) because the second approach deteriorates responsiveness of SDSs. This choice is based on the presumption that users prefer a responsive system to less responsive systems. The presumption is true in most cases if the system's performance is at human level. However, if the system's performance is below human level, high responsiveness might not be vital or even be harmful. For instance, Hirasawa et al. (1999) reported that immediate overlapping backchannels can cause users to have negative impressions. Kitaoka et al. (2005) also reported that the familiarity of an SDS with backchannels was inferior to that without backchannels due to a small portion of errors even though the overall timing and frequency of backchannels was fairly good (but did not come up to human operators). Technologies are advancing but they are still below human level. We challenge the past work that took the first approach.

Figure 1: Interface robot with an embedded LED

The second approach is simple and stable against user differences and environmental changes. Moreover, it can afford to employ more powerful but computationally expensive speech processing or to build systems on small devices with limited resources. A concern with this approach is debasement of user experience due to poor responsiveness as stated above. Another issue is speech collisions due to users' following-up utterances such as repetitions. Slow responses tend to induce such collision-eliciting speech.

This section shows the results of the experiment in which participants engaged in hotel reservation tasks with an SDS equipped with an ASE-based acknowledging method, which intuitively notified a user about the system's internal state (processing). The results suggest that the method can reduce speech collisions and provide users with positive impressions. The comparisons of evaluations between systems with a slow reply speed and a moderate reply speed suggest that users of SDSs do not care about slow replies. These results indicate that decelerating spoken dialogues is not a bad idea.

2.1 Experiment

System An SDS that can handle a hotel reservation domain was built. The system was equipped

[1]In this paper, *acknowledgment* denotes that at the level 1 of the joint action ladder (Clark, 1996), which communicates the listener's identifying the signal presented by the speaker.

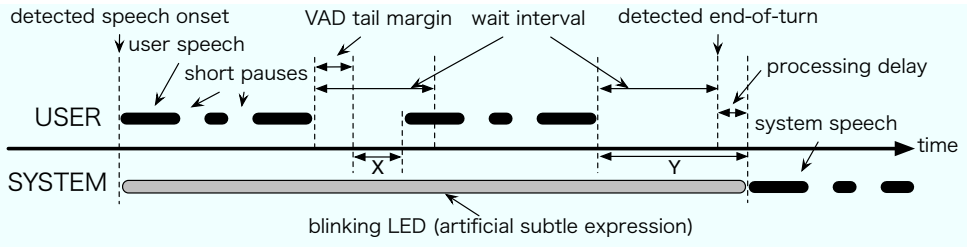

Figure 2: Behavior of the dialogue system along a timeline

with an interface robot with an LED attached to its chest (see Figure 1). Participants' utterances were recognized by an automatic speech recognizer Julius[2], and interpreted by an in-house language understander. The robot's utterances were voiced by a commercial speech synthesizer. The LCD monitor in Figure 1 was used only to show reservation details at last.

Julius output a recognition result to the system at 400 msec after an input speech signal ended, but the system awaited the next input for a fixed interval (*wait interval*, whose length is given as an experimental factor). If the system received an additional input, it awaited the next input for the same interval again. Otherwise, the system replied.

The LED started blinking at 1/30 sec even-intervals when a speech signal was detected and stopped when the system started replying. The basic function of the blinking light expression is similar to hourglass icons used in GUIs. A big difference is that basically GUIs can ignore any input while they are showing those icons, but SDSs must accept successive speech while it is blinking an LED. What we intend to do is to suppress only collision-eliciting speech such as repetitions (we call them *follow-ups*) which are negligible but difficult to be automatically distinguished from barge-ins. Barge-ins are not negligible.

Conditions and participants Two experimental factors were set-up, that is, the reply speed factor (moderate or slow reply speed) and the blinking light factor (with or without a blinking light), resulting in four conditions:

A: **slow** reply speed, **with** a blinking light,
B: **slow** reply speed, **without** a blinking light,
C: **moderate** reply speed, **with** a blinking light,
D: **moderate** reply speed, **without** a blinking light.

We randomly assigned 48 Japanese participants (mean age 30.9) to one of the four conditions.

A reply speed depends on a wait interval for which the dialogue system awaits the next input. Shiwa et al. (2008) showed that the best reply speed for a conversational robot was one second. Thus we chose 800 msec as the wait interval for the moderate reply speed because an actual reply speed was the accumulation of the wait interval and a delay for processing a user request, and 800 msec is simply twice the default length (the VAD tail margin) by which the Julius speech recognizer recognizes the end of a speech. For the slow reply speed, we chose 4 sec as the wait interval. Wait intervals include the VAD tail margin.

Figure 2 shows how the system and the LED work along with user speech. In this figure, a user utters a continuous speech with a rather long pause that is longer than the VAD tail margin but shorter than the wait interval. If the system detects the end of the user's turn and starts speaking within the interval marked with an 'X', a speech collision would occur. If the user utters a follow-up within the interval marked with a 'Y', a speech collision would occur, too. We try to suppress the former speech collision by decelerating dialogues and the latter by using a blinking light as an ASE.

Method The experiment was conducted in a room for one participant at one time. Participants entered the room and sat on a chair in front of a desk as shown in Figure 1.

The experimenter gave the participants instructions so as to reserve hotel rooms five times by talking with the robot in front of them. All of them were given the same five tasks which require them to reserve several rooms (one to three) at the same time. The meaning of the blinking light expression was not explained to them. After giving the instructions, the experimenter left the participants, and they began tasks when the robot started to talk to them. Each task was limited to up to three minutes. After finishing the tasks, the participants an-

[2]http://julius.sourceforge.jp/

swered a questionnaire. Figure 5 and Figure 6 in the appendix show one of the five task instructions, and a dialogue on that task, respectively.

2.2 Results

Reply speeds Averages of observed reply speeds were calculated from the timestamps in transcripts. They were 4.53 sec for the slow conditions and 1.42 sec for the moderate conditions.

Task completion The average number of completed tasks in the four conditions A, B, C, and D were 4.00, 3.83, 3.83, and 4.33, respectively. An ANOVA did not find any significant difference.

Speech collisions We counted speech collisions for which the SDS was responsible, that is, the cases where the robot spoke while participants were talking (*i.e.*, end-of-turn misdetections). Of course, there were speech collisions for which participants were responsible, that is, the cases where participants intentionally spoke while the robot was talking (*i.e.*, barge-ins). These speech collisions were not the targets, hence they were not included in the counts.

Speech collisions due to participants' backchannel feedbacks were not included, either. We think that it is possible to filter out such feedback because feedback utterances are usually very short and variations are small. On the other hand, as we mentioned above, it is not easy to automatically distinguish negligible speech such as repetitions from barge-ins. We want to suppress only such speech negligible but hard to distinguish from other not negligible speech.

The number of observed speech collisions in the four conditions A, B, C, and D were 5, 11, 45, and 30, respectively. First we performed an ANOVA on the number of collisions. The interaction effect was not significant ($p = 0.24$). A significant difference on the reply speed factor was found ($p < 0.005$). This result confirms that decelerating dialogues reduces collisions. The effect of the blinking light factor was not significant ($p = 0.60$).

Next we performed a Fisher's exact test (one-side) on the number of participants who had speech collisions between the two conditions of the slow reply speed (3 out of 12 for A and 8 out of 12 for B). The test found a significant difference ($p < 0.05$). This result indicates that the blinking light can reduce speech collisions by suppressing users' follow-ups in decelerated dialogues.

Impression on the dialogue and robot The participants rated 38 positive-negative adjective pairs (such as smooth vs. rough) for evaluating both the dialogue and the robot. The ratings are based on a seven-point Likert scale.

An ANOVA found a positive marginal significance ($p = 0.07$) for the blinking light in the *comfortableness* factor extracted by a factor analysis for the impression on the dialogue. In addition, an ANOVA found a positive marginal significance ($p = 0.07$) for the slow reply speed in the *modesty* factor extracted by a factor analysis for the impression on the robot. Surprisingly, no significant negative effect for the slow reply speed was found.

System evaluations The participants evaluated the SDS in two measures on a scale from 1 to 7, that is, the convenience of the system and their willingness to use the system. The greater the evaluation value is, the higher the degree of convenience or willingness.

The average scores of convenience in the four conditions A, B, C, and D were 3.50, 3.17, 3.17, and 3.92, respectively. Those of willingness were 3.58, 2.58, 2.83, and 3.42, respectively. ANOVAs did not find any significant difference among the four conditions both for the two measures.

Discussion on user preference The analysis of the questionnaire suggests that the blinking light expression gives users a comfortable impression on the dialogue. The analysis also suggests that the slow reply speed gives users a modest impression on the interface robot. Meanwhile, no negative impression with a statistical significance is found on the slow reply speed.

Although no statistically significant difference is found between the four conditions, numbers of completed tasks and convenience are strongly correlated. However, users' willingness to use the systems, which is the most important measure for systems, is inverted between condition A and D. Convenience will be primarily dominated by what degree a user's purpose (reserving rooms) is achieved, thus, it is reasonable that convenience scores correlate with the number of completed tasks. On the other hand, willingness will be dominated by not only practical usefulness but also overall usability or experience. Therefore, we can interpret that the improvements in impressions and reduction in aversive speech collisions

let condition A have the highest score for willingness. These results indicate that decelerating spoken dialogues is not a bad idea in contradiction to the common design policy in human-computer interfaces (HCIs), and they suggest to exploit merits provided by decelerating dialogues rather than pursuing quickly responding humanlike systems.

Our finding contradicts not only the common design policy in HCIs but also the design policy in human-robot interaction found by Shiwa et al. (2008), that is, *the best response timing of a communication robot is at one second*. We think this contradiction is superficial and is ascribable to the following four major differences between their study and our study.

- They adopted a within-subjects experimental design while we adopted a between-subjects design. A within-subjects design makes subjects do relative evaluations and tends to emphasis differences.
- Their question was specific in terms of response timing. Our questions were overall ratings of the system such as convenience.
- They assumed a perfect machine (Wizard-of-Oz experiment). Our system was elaborately crafted but still far from perfect.
- Our system quickly returns non-verbal responses even if verbal responses are delayed.

From these differences, we hypothesize that response timing has no significant impact on the usability of SDSs in an absolute and holistic context at least in the current state of the art spoken dialogue technology, even though users prefer a system which responds quickly to a system which responds slowly when they compare them with each other directly, given an explicit comparison metric on response timing with perfect machines.

3 Study 2: Uncovering Comfortableness of Response Timing under Different Interaction Patterns

Our conclusion in Section 2 is that SDSs do not need to quickly respond verbally as long as they quickly respond non-verbally by showing their internal states with an ASE, while many researchers try to make them verbally respond as fast as possible. Decelerating a dialogue has many practical advantages as stated above.

However, through the experiment, we have also suspected that this conclusion is not valid in some specific cases. That is, we think in some situations users feel uncomfortable with slow verbal responses primordially, and those situations are such as when users simply reply to systems' yes-no-questions or greetings. Our hypothesis is that users expect quick verbal responses (and hate slow verbal responses) only when users expect that it is not difficult for systems to understand their responses or to decide next actions. This section reports the experiment validating this hypothesis.

3.1 Experiment

To validate the hypothesis described above, we conducted a Wizard-of-Oz experiment using fixed scenarios. Participants engaged in short interactions with an interface robot and evaluated response timing of the robot. Three experimental factors were interaction patterns, response timing (wait interval), and existence of a blinking light.

Interaction patterns Five interaction patterns were setup to see the differences between situations. Each pattern consisted of three utterances. The first utterance was from the system. Upon receiving the utterance, a participant as a user of the system replied with the second utterance. Then the system responded after the given wait interval (1 sec or 4 sec) with the third utterance. Participants evaluated this interval between the second utterance and the third utterance in a measure of comfortableness.

The patterns with scenarios are shown in Figure 3. They will be referred to by abbreviations (PGG, QYQ, QNQ, PSQ, PLQ) in what follows. Note that the scenarios are originally in Japanese. Here, RequestS and RequestL mean a short request and a long request, respectively. YNQuestion and WhQuestion mean a yes-no-question and a wh-question, respectively. According to the hypothesis, we can predict that the reported comfortableness for the longer wait interval (4 sec) are worse for short and formulaic cases such as PGG and QYQ than for the long request case (*i.e.*, PLQ). In addition, we can predict that the reported comfortableness for longer intervals improves for PLQ if the robot's light blinks, while that does not improve for PGG and QYQ.

System We used the same interface robot and the LCD monitor as study 1. The experiment in this study, however, was conducted using a WOZ system.

Prompt-Greeting-Greeting (PGG)
S: Welcome to our Hotel. May I help you?
U: Hello.
S: Hello.

YNQuestion-Yes-WhQuestion (QYQ)
S: Welcome to our Hotel. Will you stay tonight?
U: Yes.
S: Can I ask your name?

YNQuestion-No-WhQuestion (QNQ)
S: Welcome to our Hotel. Will you stay tonight?
U: No.
S: How may I help you?

Prompt-RequestS-WhQuestion (PSQ)
S: Welcome to our Hotel. May I help you?
U: I would like to reserve a room from tomorrow.
S: How long will you stay?

Prompt-RequestL-WhQuestion (PLQ)
S: Welcome to our Hotel. May I help you?
U: I would like to reserve rooms with breakfast from tomorrow, one single room and one double room, non-smoking and smoking, respectively.
S: How long will you stay?

Figure 3: Interaction patterns and scenarios

First the WOZ system presents an instruction to the participant on the LCD monitor, which reveals the robot's first utterance of the given scenario (*e.g.*, "Welcome to our Hotel. May I help you?") and indicates the participant's second utterance (*e.g.*, "Hello."). Two seconds after the participant clicks the OK button on the monitor with a computer mouse, the system makes the robot utter the first utterance. Then, the participant replies, and the operator of the system end-points the end of participant's speech by clicking a button shown in another monitor for the operator in the room next to the participant's room. After the end-pointing, the system waits for the wait interval (one second or four seconds) and makes the robot utter the third utterance of the scenario. One second after, the system asks the participant to evaluate the comfortableness of the response timing of the robot's third utterance on a scale from 1 to 7 (1:very uncomfortable, 4:neutral, 7:very comfortable) on the LCD monitor.

Conditions and participants Forty participants (mean age 28.8, 20 males and 20 females) engaged in the experiment. No participant had engaged in study 1. They were randomly assigned to one of two groups (gender was balanced). The groups correspond to one of two levels of the experimental factor of the existence of a blinking light. For one group, the robot blinked its LED when it was waiting. For the other group, the robot did not blink the LED. We refer to the former group (condition) as BL (Blinking Light, n=20) and the later as NL (No Light, n=20). In summary, this experiment is within-subjects design with regard to interaction patterns and response timing and is between-subjects design with regard to the blinking light.

Method The experiment was conducted in a room for one participant at one time. Participants entered the room and sat on a chair in front of a desk as shown in Figure 1, but they did not wear headphones this time.

The experimenter gave the participants instructions so as to engage in short dialogues with the robot in front of them. They engaged in each of five scenarios shown in Figure 3 six times (three times with a 1 sec wait interval and three with 4 sec), resulting in 30 dialogues ($5 \times 3 \times 2 = 30$). The order of scenarios and intervals was randomized. The existence and meaning of the blinking light expression was not explained to them. They were not told that the system was operated by a human operator, either. After giving the instructions, the experimenter left the participants, and they practiced one time. This practice used a Prompt-RequestM-WhQuestion[3] type scenario with a wait interval of two seconds. Then, thirty dialogues were performed. Short breaks were inserted after ten dialogues. Each dialogue proceeded as explained above.

3.2 Results

End-pointing errors End-pointing was done by a fixed operator. We obtained 1,184 dialogues out of 1,200 ($= 30 \times 40$) after removing dialogues in which end-pointing failed (failures were self-reported by the operator). We sampled 30 dialogues from the 1,184 dialogues and analyzed end-pointing errors in the recorded speech data. The average error was 84.6 msec (SD=89.6).

Comfortableness This experiment was designed to grasp a preliminary sense on our hypothesis as much as possible with a limited number of participants in exchange for abandonment of use of statistical tests, because this study involved multiple factors and the interaction pattern factor was complex by itself. Therefore, in the following discussion on comfortableness, we do not refer to statistical significances.

[3]The request utterance is longer than that of RequestS and shorter than that of RequestL.

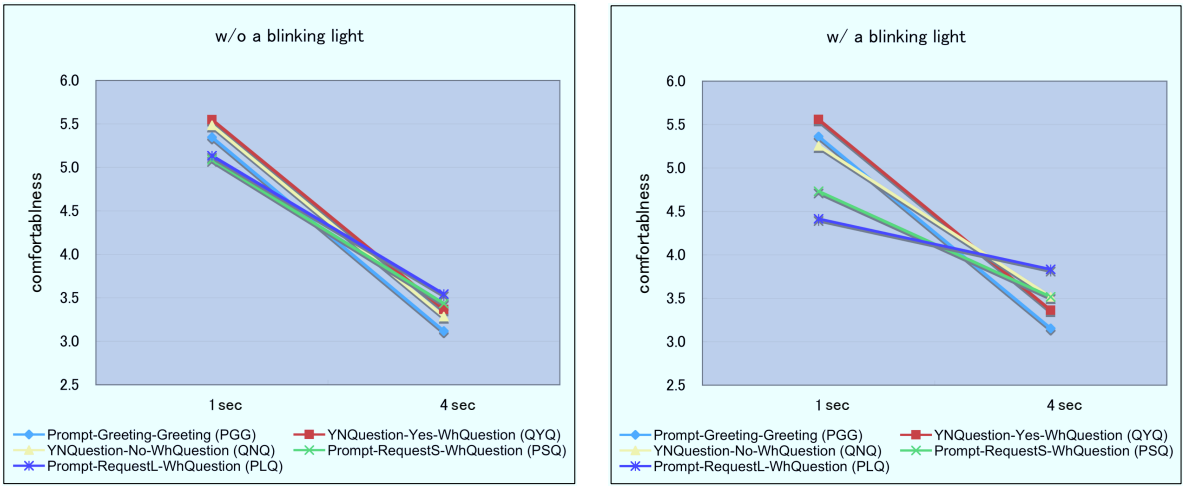

Figure 4: Comfortableness (Left: without a blinking light (NL), right: with a blinking light (BL))

Figure 4 shows regression lines obtained from the 1,184 dialogues in the two graphs for NL and BL (Detailed values are shown in Table 1). The X axes in the graphs correspond to response timing, that is, the two wait intervals of 1 sec and 4 sec. The Y axes correspond to comfortableness reported in a scale from 1 to 7. Obviously, with or without a blinking light effected comfortableness.

The results shown in the graphs support the predictions made in Section 3.1. The scores of PGG and QYQ are worse than that of PLQ at 4 sec. PGG and QYQ show no difference between NL and BL. QNQ and PSQ show differences. PLQ shows the biggest difference. In case of PLQ, the reported comfortableness at 4 sec shifted to almost the neutral position (score 4) by presenting a blinking light. This indicates that a blinking light ASE can allay the debasement of impression due to slow responses only in non-formulaic cases.

Interestingly, the blinking light expression attracted comfortableness scores to neutral both at 1 sec and at 4 sec. We can make two hypotheses on this result. One is that the blinking light expression has a negative effect which degrades comfortableness at 1 sec. The other is that the blinking light expression makes participants difficult to see differences between 1 sec and 4 sec, therefore, reported scores converge to neutral. At this stage we think that the later is more probable than the former because the scores of PGG and QYQ should be degraded at 1 sec if the former is true.

4 A Practical Design Guideline for SDSs

Summarizing the results of the experiments presented in Section 2 and Section 3, we suggest a twofold design guideline for SDSs, especially for task-oriented systems. Some interaction-oriented systems such as chatting systems are out of scope of this guideline. In what follows, first the guideline is presented and then a commentary on the guideline is described.

The guideline

(1) Never be obsessed with quick turn-taking but acknowledge users immediately

Quick turn-taking will not recompense your efforts, resources inputted, etc. Pursue it only after accomplishing all you can do without compromising performance in other elements of dialogue systems and only if it does not make system development and maintenance harder. However, quick (possibly non-verbal) acknowledgment is a requisite. You can compensate for the debasement of user experience due to slow verbal responses just by using an ASE such as a tiny blinking LED to acknowledge user speech. No instruction about the ASE is needed for users.

(2) Think of users' expectations

Users expect rather quick verbal responses to their greetings and yes-answers. ASEs will be ineffective for them. Thus it is recommended to enable your systems to quickly respond verbally to such utterances. Fortunately it is easy to anticipate such utterances. Greetings usually occur only at the beginning of dialogues or after tasks were accomplished. Yes-answers will come only after yes-no-questions. Therefore it will be able to implement an SDS that quickly responds verbally to greeting and yes-answers both without increasing development / maintenance costs and without decreasing

recognition performance, etc.

However, you should keep in mind that too quick verbal responses (0 sec interval or overlapping) may not be welcomed (Hirasawa et al., 1999; Shiwa et al., 2008). They may also induce too much expectancy in users and result in disappointments to your systems after some interactions.

Commentary on the guideline

The guideline was constructed so as to avoid the five problems pointed out in Section 1. The first point of the guideline is induced mainly from the results of study 1, and the second point is induced mainly from the results of study 2.

Although the results of study 2 indicate users prefer quick responses to slow ones as presupposed in past literature, note that the experiment in study 2 is within-subjects design with regard to the response timing factor and that within-subjects design tends to emphasis differences as discussed at the end of Section 2. The results of study 1 suggested that such an emphasized difference (*i.e.*, preference for quick responses) has no significant impact on the usability of SDSs on the whole.

5 Conclusion

This paper proposed a non-humanlike spoken dialogue design, which consists of two elements: non-humanlike turn-taking and acknowledgment. Two experimental studies were reported regarding these two elements. The first study showed that the proposed non-humanlike spoken dialogue design is effective for reducing speech collisions. This study also presented pieces of evidence that show quick humanlike turn-taking is less important in spoken dialogue system (SDS) design. The second study showed a tendency supporting a hypothesis found in the first study that user preference on response timing varies depending on interaction patterns in terms of comfortableness. Upon receiving these results, a practical design guideline for SDSs was suggested, that is, (1) never be obsessed with quick turn-taking but acknowledge users immediately and (2) think of users' expectations.

Our non-humanlike acknowledging method using an LED-based artificial subtle expression (ASE) can apply to any interfaces on wearable / handheld devices, vehicles, whatever. It is, however, difficult to directly apply it to call-centers (*i.e.*, telephone interfaces), which occupy a big portion of the deployed SDSs pie. Yet, the underlying concept: *decelerated dialogues accompanied by an ASE* will be applicable even to telephone interfaces by using an auditory ASE, which is to be explored in future work.

The guideline is supported by findings in a rather hypothetical stage. More experiments are necessary to confirm these findings. In addition, the guideline is for the current transitory period in which intelligence technologies such as automatic recognition, language processing, reasoning etc. are below human level. In that sense, the contribution of this paper might be limited. However, this period will last until a decisive paradigm shift occurs in intelligence technologies. It may come after a year, a decade, or a century.

References

C. Bartneck, T. Kanda, H. Ishiguro, and N. Hagita. 2007. Is the uncanny valley an uncanny cliff? In *Proc. RO-MAN 2007*.

H. Clark. 1996. *Using Language*. Cambridge U. P.

R. Fernández, D. Schlangen, and T. Lucht. 2007. Push-to-talk ain't always bad! comparing different interactivity settings in task-oriented dialogue. In *Proc. DECALOG 2007*.

K. Funakoshi, K. Kobayashi, M. Nakano, T. Komatsu, and S. Yamada. 2010. Reducing speech collisions by using an artificial subtle expression in a decelerated spoken dialogue. In *Proc. 2nd Intl. Symp. New Frontiers in Human-Robot Interaction*.

J. Hirasawa, M. Nakano, T. Kawabata, and K. Aikawa. 1999. Effects of system barge-in responses on user impressions. In *Proc. EUROSPEECH'99*.

N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa. 2005. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. *Journal of The Japanese Society for AI*, 20(3).

T. Komatsu and S. Yamada. 2010. Effects of adaptation gap on user's variation of impressions of artificial agents. In *Proc. WMSCI 2010*.

T. Komatsu, S. Yamada, K. Kobayashi, K. Funakoshi, and M. Nakano. 2010. Artificial subtle expressions: Intuitive notification methodology of artifacts. In *Proc. CHI 2010*.

A. Raux and M. Eskenazi. 2009. A finite-state turn-taking model for spoken dialog systems. In *Proc. NAACL-HLT 2009*.

T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita. 2008. How quickly should communication robots respond? In *Proc. HRI 2008*.

Figure 5: One of the five task instructions used in study 1

S: Welcome to Hotel Wakamatsu-Kawada. May I help you?
U: I want to stay from March 10th to 11th.
S: What kind of room would you like?
U: One non-smoking twin room and one non-smoking double room.
S: Are your reservation details correctly shown on the screen?
U: Yes. No problem.
S: Your reservation has been accepted. Thank you for using us.

Figure 6: A successful dialogue observed with the task shown in Figure 5 (translated into English)

Table 1: Detailed comfortableness scores in study 2

Interaction pattern		PGG		QYQ		QNQ		PSQ		PLQ	
Condition		NL	BL	NL	BL	NL	BL	NL	BL	NL	BL
1 sec	mean	5.34	5.36	5.55	5.56	5.48	5.25	5.09	4.73	5.13	4.41
	s.d.	1.00	1.17	1.10	1.00	1.02	1.04	1.12	1.09	1.14	1.20
	p-value	0.93		0.96		0.23		0.09		0.001	
4 sec	mean	3.12	3.16	3.37	3.36	3.28	3.52	3.43	3.52	3.54	3.83
	s.d.	0.94	1.04	0.78	0.93	0.76	0.93	0.81	0.87	0.95	0.87
	p-value	0.83		0.98		0.14		0.59		0.08	

p-values were obtained by two-sided t-tests between NL and BL. Those are shown just for reference.