Reducing Speech Collisions by Using an Artificial Subtle Expression in a Decelerated Spoken Dialogue
– Should communication robots respond quickly? –

Kotaro Funakoshi [1] and **Kazuki Kobayashi** [2] and **Mikio Nakano** [3]
and **Takanori Komatsu** [4] and **Seiji Yamada** [5]

Abstract. We argue that spoken dialogue systems or communication robots do not need to quickly respond verbally as long as they quickly respond non-verbally by showing their internal states by using an artificial subtle expression. This paper describes an experiment whose results support this point. In this experiment, 48 participants engaged in reservation tasks with a spoken dialogue system coupled with an interface robot using a blinking light expression. The expression is designed as an artificial subtle expression to intuitively notify a user about a robot's internal states (such as processing) for the sake of reducing speech collisions as consequences of turn-taking failures due to end-of-turn misdetection. Speech collisions harm smooth speech communication and degrade system usability. Two experimental factors were setup: the blinking light factor (with or without a blinking light) and the reply speed factor (moderate or slow reply speed), resulting in four experimental conditions. The results suggest that the blinking light expression can reduce speech collisions and improve user impression, and surprisingly that users do not care about slow replies.

1 Introduction

An important issue in spoken dialogue systems 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. Such speech collisions lead users to stop speaking and bring about troubles in spoken dialogue systems [11] because interrupted speech is hard to recognize automatically and dialogue states after collisions are unclear. Deterioration in user impressions is also a problem.

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., [2, 8, 12, 14]) because the second approach deteriorates responsiveness of dialogue systems. 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, a user study reported that the familiarity of a spoken dialogue system with back-channel feedback was inferior to that without feedback due to a small portion of errors even though the overall timing and frequency of feedback were fairly good (but did not come up to human operators) [8]. Technologies are advancing but they are still below the human level. We challenge to the past work that took the first approach.

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 because systems' late responses tend to induce such utterances, which are not desired from the viewpoint of systems.

Taking the second approach, we showed the possibility that non-speech feedback using a blinking light based on the concept of artificial subtle expressions (see the next section) can suppress undesired utterances (repetitions) from users [4]. However, our experiment was not with an automatic spoken dialogue system but with a human-operated Wizard-of-Oz system. We showed that our method reduced repetitions (potential causes of collisions) but did not show that our method reduced collisions. We also showed that the blinking light feedback improved user impression but did not show whether the improvement in user impression was enough to compensate for the debasement of user experience due to slow replies.

This paper shows the results of the experiment in which participants engaged in hotel reservation tasks with a spoken dialogue system equipped with an artificial subtle expression-based method proposed in [4], which intuitively notified a user about the system's internal states (such as processing or busy). The results suggest that the method can reduce speech collisions and provide users with a comfortable impression and a modesty impression. The comparisons of user evaluations between systems with a slow reply speed and a moderate reply speed suggest that users of spoken dialogue systems do not care about slow replies. These results indicate that taking the second approach, decelerating spoken dialogues, is not a bad idea.

Section 2 explains artificial subtle expressions. The experiment and results are described in Section 3 and Section 4 respectively. Sec-

[1] Honda Research Institute Japan Co., Ltd., 8-1 Honcho Wako, Saitama 351-0188, Japan, email: funakoshi@jp.honda-ri.com

[2] Shinshu University, Japan

[3] Honda Research Institute Japan Co., Ltd., Japan

[4] Shinshu University, Japan

[5] National Institute of Informatics / The Graduate University for Advanced Studies, Japan

tion 5 concludes this paper.

2 Blinking Light as Artificial Subtle Expression

Although human communication is explicitly achieved through verbal utterances, non-verbal facial expressions, gaze, gestures, etc., also play an important role [6, 7]. Such non-verbal communication often influences the accuracy of utterance understanding [15].

Furthermore, researchers have reported that very small changes (called *subtle expressions*) in facial expressions and gestures might influence human communication. We believe that we can utilize such subtle expressions to make humans easily understand a robot's internal state because humans can intuitively understand subtle expressions. Some studies have been done on applying subtle expressions to human-agent interaction [1, 3, 13]. However, since they tried to enable subtle expressions on real faces and with real arms, their implementations were considerably expensive.

Equipping a robot to express its turn-taking intention by using body/eye movements as humans do may reduce speech collisions. Though, such an approach is technically difficult and uneconomical as mentioned above. Using spoken back-channel feedback (such as "well" and "uh") is another option, however, it is not an easy matter because such buck-channel feedback expressions are not arbitrarily used but require appropriate timing and situations to be used [18, 19]. Moreover, such approaches strongly restricts characters of robots, therefore, their applicabilities inevitably go low. Cultural differences also affect them.

In contrast with such human-like approaches, subtle expressions have been studied for artifacts like a robot or PC. Komatsu and Yamada [9] reported that an agent's subtle expression of simple beeping sounds with decreasing/increasing frequency enabled humans to interpret the agent's positive/negative states. Their work indicated the effectiveness of subtle expressions such as varying beeping sounds for a robot or agent. They named such subtle expressions "artificial subtle expressions (ASEs)" and defined them as expressions fulfilling the following four requirements [10].

- **Simple:** ASEs should be implemented on a single modality. It is expected that implementation cost also should be lower.
- **Complementary:** ASEs should only have a complementary role in communication and should not interfere with communication's main protocol. This means that the ASEs themselves do not have any meaning without any communication context.
- **Intuitive:** ASEs should be understood by humans who do not know about the ASEs beforehand.
- **Accurate:** ASEs should convey the specific meanings accurately. Specifically, ASEs should convey the internal states of the artifact like subtle expressions are doing.

Following their work, we proposed the use of a blinking light as a means of artificial subtle expression to intuitively notify a user about a robot's internal states (such as processing or busy) and showed that the blinking light expression potentially can reduce speech collisions, however, indirectly by showing the reduction of users' speech repetitions in a last-and-first game dialogue [4]. This paper verifies the effectiveness of the blinking light expression in a more practical task-oriented dialogue with an automatic spoken dialogue system to directly show this approach reduces collisions.

The use of a blinking light may not be the best way in terms of reducing speech collisions in comparison with other human-like approaches explained above. However, it will be easy and cheap to implement, and applicable to wider conversational agents/devices not limited to robots.

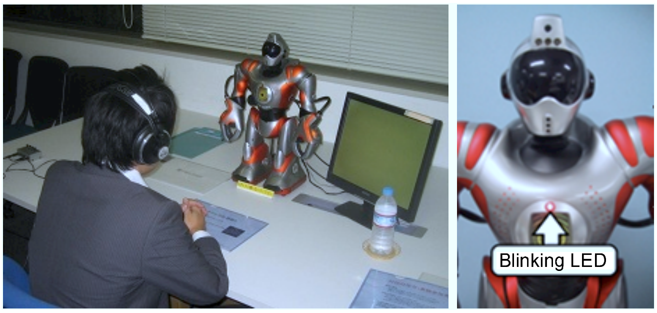


Figure 1. The LCD monitor and robot with an embedded LED

3 Experiment

We conducted an experiment in which 48 participants engaged in hotel reservation tasks with a conversational robot. The dialogue system, robot, blinking light, experimental conditions and method are explained below.

3.1 Spoken Dialogue System

A spoken dialogue system (not a voice command system) that can handle user requests in a hotel reservation domain was built. The system was equipped with an LCD monitor to show reservation information and an interface robot with an LED attached to its chest (see Figure 1). Participants' utterances were recognized by a free automatic speech recognizer Julius[6], and interpreted by using a domain-ontology-centered language understanding method [5]. The robot's utterances were voiced by a commercial speech synthesizer (NTT-IT FineVoice [17]). The LCD monitor was used to reduce the time in which the system just read up reservation details and to reduce participants' cognitive loads to catch the system's lengthy speech by showing reservation details on the monitor only when the system asked for confirmation.

The system was tuned to uplift its speech understanding performance as much as possible. For example, the language model used by Julius was built using the data collected in the same task domain in advance (4223 utterances / 140-minute speech of 47 users).

Default values were used for the parameters in Julius. Julius output a recognition result (a sequence of recognized words) to the dialogue system at 400 msec after an input speech signal ended, but the dialogue system awaited the next input for a fixed interval (we call this 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 concatenated recognition results, made an interpretation, and replied.

3.2 Robot and Blinking Light Expression

We adopted the same robot (WowWee RS-Media) and the same red LED (diameter: 4 mm) as [4]. The blinking pattern was also the same

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

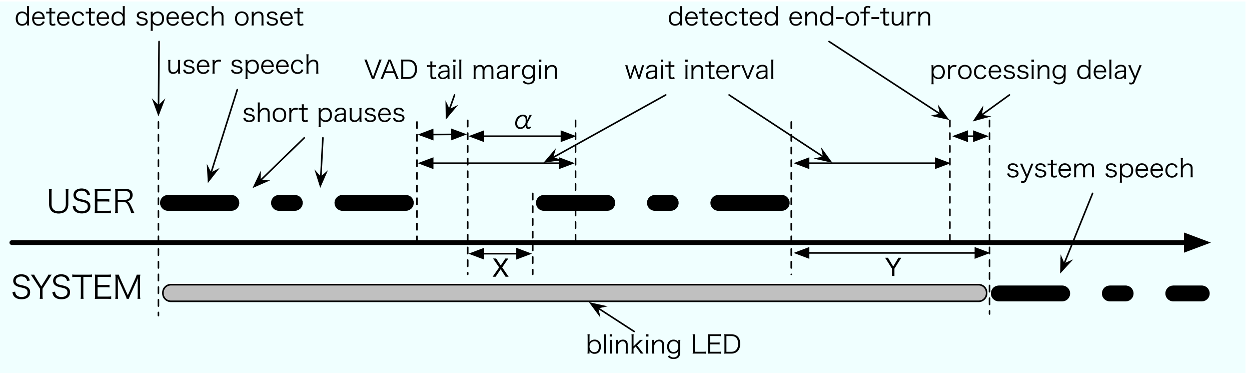


Figure 2. Behavior of the dialogue system along a timeline

(1/30 sec even-intervals). The LED started blinking 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 dialogue systems 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 and rephrasings which are negligible but difficult to automatically distinguish from *barge-ins* when speech collisions happen (we call them *follow-ups*). Barge-ins are users' intentional interruptions to systems' speech and dialogue systems must accept them.

3.3 Experimental Conditions and Participants

Two experimental factors each having two levels were setup, that is, the blinking light factor (with or without a blinking light) and the reply speed factor (moderate or slow reply speed), resulting in four experimental conditions A, B, C, and D as below.

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

We prepared a dialogue system for each group, four systems in total and randomly assigned 48 Japanese participants of 18 to 55 years old (mean age 30.9, SD = 10.2) to one of the four conditions, while genders and generations were controlled to be equally distributed among the conditions.

A reply speed depends on a wait interval for which the dialogue system awaits the next input. A user study [16] 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 4000 msec as the wait interval. Wait intervals include the VAD tail margin.

Figure 2 shows how the system and the LED works 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.

3.4 Experimental 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. They were asked to wear headphones to separate system voice and user voice. All dialogues were videotaped.

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 up to three minutes. The experimenter aborted dialogues if they could not finish a given task in three minutes. They worked on the five tasks without intermission in a fixed order. For illustration, Figure 3 shows the instruction of one of the tasks, and Figure 4 shows an example dialogue observed with this task.

After finishing the tasks, the participants answered a questionnaire about their impressions of the dialogue and robot, and other particular questions. Participants were paid for their participation.

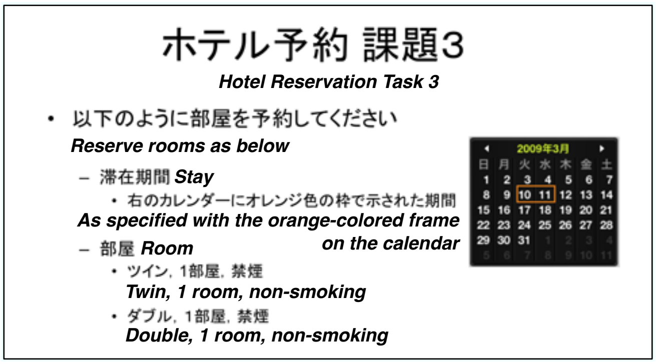
4 Results

4.1 Observed reply speeds

All dialogues were transcribed into texts with timestamps. Averages of observed reply speeds including the system's delays were calculated from those timestamps. They were 4.53 sec (SD = .75) for the slow conditions (Aand B) and 1.42 sec (SD = .43) for the moderate conditions (Cand D).

4.2 Task completion

The average numbers of completed tasks in the four conditions are shown in Table 1. A two-factor two-level ANOVA did not find any



ホテル予約 課題3

Hotel Reservation Task 3

- 以下のように部屋を予約してください
  Reserve rooms as below
  - 滞在期間 *Stay*
    - 右のカレンダーにオレンジ色の枠で示された期間
      As specified with the orange-colored frame on the calendar
  - 部屋 *Room*
    - ツイン，1部屋，禁煙
      Twin, 1 room, non-smoking
    - ダブル，1部屋，禁煙
      Double, 1 room, non-smoking

2009年3月

日	月	火	水	木	金	土
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	1	2	3	4
5	6	7	8	9	10	11

Figure 3. An example of task instructions

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 4. An observed dialogue (translated into English)

significant difference among the four conditions on the numbers of completed tasks.

Table 1. Numbers of completed tasks

reply speed	slow		moderate	
blinking light	w/	w/o	w/	w/o
condition label	A	B	C	D
mean	4.00	3.83	3.83	4.33
SD	1.28	1.34	1.27	0.98

4.3 Speech Collisions

As mentioned above, we use the term *follow-ups* to refer to all user utterances such as so called repetitions, corrections, and additions, which are potentially inducing speech collisions. In the experiment in [4], they counted repetitions and it was not difficult in their quite simple last-and-first game dialogues where each user utterance contains only one word. In our experiment, in contrast, it is hard to provide follow-ups with a concrete definition, which is necessary to count them in the dialogue data. Therefore, we do not count follow-ups but directly focus on the number of speech collisions.

We counted speech collisions for which the dialogue system 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 of this paper, hence they were not included in the counts.

Speech collisions due to participants' back-channel 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.

Speech collisions were counted both at the experiment by an experiment supporting staff and after the experiment on the dialogue transcriptions while watching videos by a research supporting staff. The two counting results were checked and merged by one of the authors. The numbers are shown in Table 2 for each participant.

First we performed a two-factor two-level ANOVA on the numbers of collisions. A significant difference between the slow reply speed (A and B) and the moderate reply speed (C and D) was found ($F_{1,44} = 4.06$, $p < 0.005$). This result confirms that making a long interval after the user's speech ends and before the system replies reduces speech collisions.

The ANOVA test did not find a significant difference on the blinking light factor. This was, however, reasonable because the effect of the blinking light suppressing troublesome follow-ups from users were expected only in the slow reply speed conditions and the numbers of collisions in the slow reply speed conditions were fairly smaller than those in the moderate reply speed conditions.

Hence, we performed a Fisher's exact test (one-side) on the numbers of participants who had speech collisions between the two conditions of the slow reply speed (A and B). The contingency table is shown in Table 3. The test found a significant difference ($p < 0.05$). This result indicates that the blinking light can reduce speech collisions by suppressing users' unnecessary follow-ups in decelerated dialogues.

Table 3. Contingency table on the numbers of participants who had collisions in the slow reply speed conditions

condition	had collisions	had no collision	total
w/ a blinking light (A)	3	9	12
w/o a blinking light (B)	8	4	12

4.4 Impression on the Dialogue

Table 4 shows the results of participants' ratings for the dialogue. The adjective pairs in the table are translated from Japanese words that we used in the questionnaire. The ratings are based on a seven-point Likert scale (1:strong agreement with a negative adjective, 4: neutral, 7: strong agreement with a positive adjective). The highest mean score for each adjective pair is underlined.

We performed a factor analysis (principal factor method) under the varimax rotation and obtained five factors from the scree plot. Table 5 shows the factor loadings under the varimax rotation.

We interpreted the factors according to the adjective pairs. The first factor was named the *likability factor*. The second factor was named the *comfortability factor*. The third factor was named the *interest factor*. The fourth and fifth factors were named *casualness factor* and *comprehensibility factor*, respectively. Their factor scores by using a regression method were calculated and compared among the conditions. Table 6 shows the factor scores for the impression of the

Table 2. Numbers of collisions by each participant

condition			participant ID in each condition												
reply speed	blinking light	label	1	2	3	4	5	6	7	8	9	10	11	12	total
slow	w/	A		1	2						2				5
	w/o	B	1			3		1		2	1	1	1	1	11
moderate	w/	C	2	5			5	6	3	3	1	14	5	1	45
	w/o	D	7		2		6	6		2		5		2	30

(All the 48 participants are different from each other. Blank slots mean zeros.)

Table 4. Rated adjective pairs for impression of the dialogue

reply speed		slow		moderate	
blinking light		w/	w/o	w/	w/o
condition label		A	B	C	D
positive	negative	mean (SD)	mean (SD)	mean (SD)	mean (SD)
casual	grave	3.42 (1.24)	3.33 (1.07)	3.50 (1.31)	4.00 (1.54)
smooth	rough	3.25 (1.29)	2.58 (1.00)	3.17 (1.59)	3.00 (1.35)
decent	indecent	4.42 (1.00)	4.08 (0.79)	4.08 (1.00)	4.42 (1.08)
exciting	dull	4.58 (0.67)	4.17 (1.03)	3.83 (1.95)	4.17 (1.11)
relaxed	tensional	3.25 (1.06)	3.25 (1.48)	3.42 (1.16)	3.75 (1.48)
easy	uneasy	3.33 (1.37)	2.42 (1.24)	2.75 (1.06)	2.58 (1.08)
warm	cold	3.25 (1.48)	3.25 (0.97)	3.67 (0.98)	3.67 (1.15)
pleasant	unpleasant	4.33 (0.89)	3.67 (1.15)	3.17 (1.03)	4.25 (0.97)
leisurely	hurried	4.75 (1.06)	4.92 (1.51)	4.83 (1.27)	4.67 (0.89)
informal	formal	3.08 (1.16)	2.67 (0.89)	2.67 (1.44)	3.17 (1.03)
light	dark	3.42 (1.00)	3.00 (0.85)	3.17 (0.94)	3.67 (0.78)
clear	confusing	4.00 (1.76)	3.58 (1.51)	4.17 (1.59)	4.50 (0.80)
likable	dislikable	4.42 (0.67)	3.75 (0.87)	4.08 (1.16)	4.33 (0.78)
good	poor	3.75 (1.22)	3.17 (1.11)	3.17 (1.19)	3.58 (1.24)
peaceful	annoying	4.33 (1.37)	3.08 (1.56)	3.83 (1.95)	3.75 (0.87)
interesting	boring	4.58 (1.24)	4.25 (1.14)	4.33 (1.67)	4.67 (1.30)
spirited	dispirited	3.42 (0.67)	2.92 (1.00)	3.00 (0.85)	3.25 (0.87)
settled	unsettled	4.50 (1.31)	3.50 (1.38)	3.83 (1.11)	3.58 (0.79)

Table 5. Results of factor analysis of impression of the dialogue (varimax rotation, factor loading matrix)

factor	item (positive)	factor loadings				
		1	2	3	4	5
1	decent	0.80	0.13	0.23	0.04	0.13
	likable	0.80	0.16	0.15	0.17	-0.23
	warm	0.58	0.01	-0.10	0.34	0.16
	peaceful	0.56	0.52	0.23	-0.25	0.15
	light	0.54	0.23	0.05	0.27	0.28
	informal	0.49	0.37	0.15	0.38	0.04
2	settled	0.35	0.86	0.06	-0.07	0.06
	easy	0.02	0.68	0.05	0.28	0.15
	good	0.13	0.62	0.37	0.42	0.14
	spirited	0.32	0.57	0.36	0.31	-0.04
	smooth	0.07	0.49	0.33	0.26	0.23
3	interesting	0.31	0.03	0.88	0.14	0.02
	exciting	-0.03	0.22	0.59	0.06	0.27
	pleasant	0.10	0.27	0.43	0.15	0.21
4	casual	0.20	0.08	0.14	0.68	-0.10
	relaxed	0.14	0.22	0.11	0.63	-0.05
5	clear	0.04	0.26	0.13	0.03	0.87
	leisurely	0.09	0.00	0.17	-0.13	0.34
contribution (%)		16.12	16.04	10.79	9.91	7.27

dialogue. A two-factor two-level ANOVA found a 10 % level significant main effect for the blinking light in the second factor's score ($F_{3,44} = 3.53, p = .07$). This result suggests that the blinking light provides users with a comfortable impression on the dialogue.

4.5 Impression on the Robot

We analyzed the impression on the robot with the same method as that on the dialogue. Table 7 shows the results of participants' ratings for the robot.

Table 7. Rated adjective pairs for impression of the robot

reply speed		slow		moderate	
blinking light		w/	w/o	w/	w/o
condition label		A	B	C	D
positive	negative	mean (SD)	mean (SD)	mean (SD)	mean (SD)
aggressive	defensive	3.67 (0.78)	3.58 (0.79)	3.92 (1.16)	4.17 (1.19)
innocent	wicked	4.08 (1.08)	3.92 (1.31)	4.17 (0.94)	4.33 (0.78)
respectful	impudent	4.75 (0.87)	4.17 (1.59)	4.42 (0.90)	5.25 (1.36)
accessible	inaccessible	3.08 (1.08)	3.42 (1.31)	3.50 (1.09)	4.00 (0.95)
pretty	provoking	4.17 (0.39)	4.17 (0.58)	4.17 (0.83)	4.08 (0.79)
tolerant	intolerant	4.17 (0.58)	3.75 (0.75)	3.92 (0.79)	4.08 (0.79)
sociable	unsociable	3.75 (1.22)	3.50 (1.17)	3.33 (1.15)	3.92 (0.90)
responsible	irresponsible	4.42 (1.08)	4.17 (1.34)	4.08 (1.24)	4.50 (1.00)
careful	careless	5.42 (1.08)	4.92 (0.90)	4.92 (1.24)	4.50 (1.31)
modest	shameless	4.08 (0.51)	4.08 (0.51)	3.83 (0.72)	3.83 (0.39)
serious	frivolous	5.00 (1.28)	4.67 (1.07)	5.00 (1.35)	4.50 (1.17)
excited	gloom	3.50 (0.52)	3.67 (1.07)	3.33 (0.78)	4.00 (1.04)
regal	servile	4.92 (1.51)	5.00 (1.21)	4.75 (0.97)	4.83 (0.72)
decent	indecent	4.25 (1.06)	3.67 (1.23)	3.58 (0.90)	4.17 (0.72)
discreet	indiscreet	4.67 (1.23)	4.33 (1.07)	3.92 (1.38)	4.33 (0.78)
friendly	unfriendly	3.75 (1.29)	3.33 (1.50)	3.17 (1.40)	4.42 (1.00)
active	inactive	3.83 (1.03)	3.58 (0.67)	3.25 (1.48)	4.17 (0.94)
confident	unconfident	4.50 (1.00)	4.25 (0.87)	4.00 (0.95)	4.67 (0.89)
patient	impatient	4.42 (0.79)	4.92 (1.31)	4.00 (1.04)	4.33 (1.07)
kind	unkind	3.67 (1.30)	4.08 (1.44)	3.33 (1.15)	4.00 (1.04)

Table 8 shows the factor loadings (principal factor method) under the varimax rotation. We interpreted the obtained factors by a factor analysis (principal factor method, varimax rotation) according to the adjective pairs. The first factor was named the *civility factor*. The second factor was named the *seriousness factor*. The third factor was named the *friendliness factor*. The fourth and fifth factors were named *aggressiveness factor* and *modesty factor*, respectively.

Table 9 shows the factor scores by using regression method for the impression of the robot. A two-factor two-level ANOVA among factor scores (regression method) found a 10 % level significant main effect for the slow reply speed in the fifth factor's score ($F_{3,44} = 3.39, p = .07$). This result suggests that the slow reply speed makes the robot look modest.

Table 6. Factor scores for impression of the dialogue

factor	reply speed	blinking light	mean	SD	ANOVA	$F_{3,44}$	p
1. likability	slow	w/o	-0.31	0.27	reply speed	0.45	0.51
		w/	0.13	0.27	blinking light	0.12	0.73
	moderate	w/o	0.21	0.27	reply speed ×blinking light	1.59	0.21
		w/	-0.03	0.27			
2. comfortability	slow	w/o	-0.25	0.26	reply speed	1.19	0.28
		w/	0.53	0.26	blinking light	3.53	**0.07**
	moderate	w/o	-0.25	0.26	reply speed ×blinking light	1.21	0.28
		w/	-0.04	0.26			
3. interest	slow	w/o	-0.03	0.28	reply speed	0.09	0.77
		w/	0.12	0.28	blinking light	0.00	0.96
	moderate	w/o	0.05	0.28	reply speed ×blinking light	0.35	0.56
		w/	-0.13	0.28			
4. casualness	slow	w/o	-0.08	0.25	reply speed	1.24	0.27
		w/	-0.20	0.25	blinking light	1.37	0.25
	moderate	w/o	0.38	0.25	reply speed ×blinking light	0.49	0.49
		w/	-0.10	0.25			
5. comprehensibility	slow	w/o	-0.18	0.27	reply speed	1.78	0.19
		w/	-0.18	0.27	blinking light	0.17	0.68
	moderate	w/o	0.29	0.27	reply speed ×blinking light	0.20	0.66
		w/	0.06	0.27			

Table 9. Factor scores for impression of the robot

factor	reply speed	blinking light	mean	SD	ANOVA	$F_{3,44}$	p
1. civility	slow	w/o	-0.19	0.28	reply speed	0.38	0.54
		w/	0.02	0.28	blinking light	0.32	0.58
	moderate	w/o	0.35	0.28	reply speed × blinking light	1.71	0.20
		w/	-0.17	0.28			
2. seriousness	slow	w/o	-0.10	0.28	reply speed	0.30	0.59
		w/	0.26	0.28	blinking light	0.85	0.36
	moderate	w/o	-0.16	0.28	reply speed × blinking light	0.12	0.73
		w/	0.00	0.28			
3. friendliness	slow	w/o	-0.11	0.26	reply speed	0.13	0.72
		w/	0.01	0.26	blinking light	1.42	0.24
	moderate	w/o	0.42	0.26	reply speed × blinking light	2.68	0.11
		w/	-0.32	0.26			
4. aggressiveness	slow	w/o	-0.39	0.28	reply speed	1.15	0.29
		w/	0.09	0.28	blinking light	1.45	0.23
	moderate	w/o	0.05	0.28	reply speed × blinking light	0.25	0.62
		w/	0.25	0.28			
5. modesty	slow	w/o	0.39	0.26	reply speed	3.39	**0.07**
		w/	0.09	0.26	blinking light	0.18	0.67
	moderate	w/o	-0.28	0.26	reply speed × blinking light	0.54	0.47
		w/	-0.20	0.26			

Table 8. Results of factor analysis of impression of the robot (varimax rotation, factor loading matrix)

factor	item (positive)	factor loadings				
		1	2	3	4	5
1	respectful	0.85	0.12	-0.02	-0.10	-0.05
	innocent	0.79	0.14	0.07	0.27	0.23
	sociable	0.72	0.02	0.08	-0.18	-0.02
	decent	0.71	0.02	0.32	0.29	0.30
	accessible	0.69	-0.16	0.37	0.02	0.22
	pretty	0.65	-0.23	0.10	0.23	0.03
	kind	0.65	0.24	0.31	-0.02	0.20
	excited	0.63	-0.43	0.33	0.00	-0.01
	tolerant	0.62	-0.01	0.05	0.19	0.01
2	serious	-0.15	0.86	-0.08	-0.18	-0.27
	careful	-0.11	0.65	0.04	-0.01	0.24
	responsible	-0.14	0.61	0.52	-0.10	0.19
	discreet	0.41	0.54	0.21	-0.05	0.45
	regal	0.17	0.47	0.07	-0.03	-0.12
3	friendly	0.54	-0.19	0.71	-0.04	0.14
	active	0.26	0.20	0.43	-0.01	-0.07
	confident	0.05	0.09	0.41	0.27	-0.22
4	patient	0.16	0.32	0.15	0.78	-0.43
	aggressive	0.34	-0.07	0.15	0.71	0.15
5	modest	0.10	-0.03	-0.06	-0.03	0.56
contribution (%)		26.15	13.02	8.55	7.60	6.08

4.6 System Evaluations

The participants evaluated the dialogue system 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 convenience scores of the four conditions in Table 10. The average willingness scores of the four conditions are shown in Table 11. The largest mean values among the four conditions are shown in bold letters. ANOVAs did not find any significant difference among the four conditions both for the two measures.

Table 10. Convenience scores

reply speed	slow		moderate	
blinking light	w/	w/o	w/	w/o
condition label	A	B	C	D
mean	3.50	3.17	3.17	**3.92**
SD	2.02	1.53	1.47	1.62

Table 11. Willingness scores

reply speed	slow		moderate	
blinking light	w/	w/o	w/	w/o
condition label	A	B	C	D
mean	**3.58**	2.58	2.83	3.42
SD	1.73	1.31	1.34	1.56

4.7 Discussion on User Preference

The analysis of the questionnaire suggests that the blinking light expression gives users a comfortable impression on the dialogue. In addition, comparing mean scores of adjective pairs and system evaluations between the conditions A and B, a tendency giving higher scores to A is observed. These support the effectiveness of the blinking light expression. However, comparing scores between C and D, a tendency giving higher scores to D is observed. This intimates that the blinking light expression includes some negative effects, too. We are planning to investigate the blinking pattern from this aspect in future work.

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 (that is, differences are small), numbers of completed tasks shown in Table 1 and convenience scores shown in Table 10 strongly correlate. However, users' willingness to use the systems, which is the most important measure for systems, is inverted between condition A and D as shown in Table 11. Convenience will be primarily dominated by to what degree a user's purpose (reserving rooms) is achieved, thus, it is reasonable that convenience scores correlate with numbers of completed tasks. On the other hand, willingness will be dominated by not only practical usefulness but also overall usability. Therefore, we can interpret that the improvements of impressions and reduction of 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 brought about by decelerating interactions rather than pursuing quickly responding human-like 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. [16], 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, that is, "Answer *your preference about the timing of the system response*." 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 dialogue systems 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 of response timing appropriateness with perfect machines.

5 Conclusion and Future work

This paper showed the results of the experiment in which forty-eight participants engaged in hotel reservation tasks with a spoken dialogue system coupled with an interface robot equipped with an LED. The experiment aimed to investigate the influences of decelerating dialogues and the effects of a blinking light expression devised as an artificial subtle expression (ASE) on reducing speech collisions and compensating the deterioration of responsiveness due to decelerating dialogues. Decelerating dialogues is the simplest way to build stable spoken dialogue systems with speech collisions reduced, and affords to employ more powerful but computationally expensive speech processing or to build systems on small devices with limited resources. The blinking light expression as an ASE is quite simple, thus it is easy and cheap to implement, and applicable to wider conversational agents/devices not limited to robots.

The two experimental factors used in the experiment were the blinking light factor (with or without a blinking light) and the reply speed factor (moderate or slow reply speed). The analysis showed that speech collisions were reduced by slowing the reply speed, and they were reduced further by using the blinking light expression with statistical significances. The analysis of a questionnaire suggested that the blinking light expression gave participants a comfortable impression, and surprisingly that users did not care about slow replies. In addition, although a statistical significance was not found, the system with the slow reply speed and a blinking light obtained the highest score for users' willingness to use the system, which is the most important measure for systems.

While our method using an LED can apply to any other interfaces on wearable/handheld devices, vehicles, whatever, it is difficult to directly apply it to call-centers (i.e., telephone interfaces), which occupy a big portion of the deployed spoken dialogue systems pie. However, the underlying framework, that is, "decelerating spoken dialog with an artificial subtle expression", will be applicable even to telephone interfaces by using an auditory artificial subtle expression which is to be explored in future work.

So far, our conclusion is that spoken dialogue systems or communication robots do not need to quickly respond verbally as long as they quickly respond non-verbally by showing their internal states by using an artificial subtle expression, while many researchers try to make them verbally respond as fast as possible. Decelerating dialogue has many practical advantages as stated above. However, through the experiment, we suspect that this conclusion is not valid in some specific cases. That is, we think in some situations users are troubled by 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. If this hypothesis is valid, and if users' expectations can be easily estimated not from unreliable information such as speech recognition results, intonations, facial motions, etc but from solid information such as the speech act type of a system's preceding utterance, we will be able to immediately achieve better speech communication systems which respond quickly only when needed. We will seek this direction in future work.

REFERENCES

[1] C. Bartneck and J. Reichenbach, 'Subtle emotional expressions of synthetic characters', *International Journal of Human-Computer Studies*, **62**(2), 179–192, (2005).

[2] L. Bell, J. Boye, and J. Gustafson, 'Real-time handling of fragmented utterances', in *Proc. NAACL-2001 workshop on Adaptation in Dialogue Systems*, (2001).

[3] S. Brave, C. Nass, and K. Hutchinson, 'Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent', *International Journal of Human-Computer Studies*, **62**(2), 161–178, (2005).

[4] Kotaro Funakoshi, Kazuki Kobayashi, Mikio Nakano, Seiji Yamada, Yasuhiko Kitamura, and Hiroshi Tsujino, 'Smoothing human-robot speech interactions by using a blinking-light as subtle expression', in *Proc. 2008 International Conference on Multimodal Interfaces (ICMI)*, pp. 293–296, (2008).

[5] Kotaro Funakoshi, Mikio Nakano, Yuji Hasegawa, and Hiroshi Tsujino, 'Semantic interpretation supplementarily using syntactic analysis', in *Proc. 2007 International Conference /Recent Advances in Natural Language Processing/ (RANLP)*, Borovets, Bulgaria, (Sep. 2007).

[6] A. Kendon, 'Some functions of gaze direction in social interaction', *Acta Psychologica*, **26**, 1–47, (1967).

[7] A. Kendon, 'Do gestures communicate?', *A Review. Research in Language and Social Interaction*, **27**(3), 175–200, (1994).

[8] Norihide Kitaoka, Masashi Takeuchi, Ryota Nishimura, and Seiichi Nakagawa, 'Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems', *Journal of The Japanese Society for Artificial Intellignece*, **20**(3), 220–228, (2005).

[9] T. Komatsu and S. Yamada, 'How do robotic agents' appearances affect people's interpretations of the agents' attitudes?', in *Proc. 25th ACM International Conference on Human Factors in Computing Systems (CHI)*, pp. 2519–2524, (2007).

[10] Takanori Komatsu, Seiji Yamada, Kazuki Kobayashi, Kotaro Funakoshi, and Mikio Nakano, 'Artificial subtle expressions: Intuitive notification methodology of artifacts', in *Proc. 28th ACM International Conference on Human Factors in Computing Systems (CHI)*, p. (to be appeared), (2010).

[11] Mikio Nakano, Yuka Nagano, Kotaro Funakoshi, Toshihiko Ito, Kenji Araki, Yuji Hasegawa, and Hiroshi Tsujino, 'Analysis of user reactions to turn-taking failures in spoken dialogue systems', in *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, (2007).

[12] Tomoko Ohsuga, Masafumi Nishida, Yasuo Horiuchi, and Akira Ichikawa, 'Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue', in *Proc. 9th European Conference on Speech Communication and Technology*, pp. 33–36, (2005).

[13] H. Prendinger, J. Mori, and M. Ishizuka, 'Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game', *International Journal of Human-Computer Studies*, **62**(2), 231–245, (2005).

[14] A. Raux and M. Eskenazi, 'Optimizing endpointing thresholds using dialogue features in a spoken dialogue system', in *Proc. 9th SIGdial Workshop on Discourse and Dialogue*, (2008).

[15] W. Rogers, 'The contribution of kinesic illustrators towards the comprehension of verbal behavior within utterances', *Human Communication Research*, **5**, 54–62, (1978).

[16] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita, 'How quickly should communication robots respond?', in *Proc. 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 153–160, (2008).

[17] S. Takano, K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima, 'A Japanese TTS system based on multi-form units and a speech modification algorithm with harmonics reconstruction', *IEEE Transactions on Speech and Audio Processing*, **9**(1), 3–10, (2001).

[18] Wataru Tsukahara and Nigel Ward, 'Evaluating responsiveness in spoken dialog systems', in *Proc. International Conference on Spoken Language Processing*, (2000).

[19] Nigel Ward, 'On the expressive competencies needed for responsive systems', in *Proc. the CHI2003 workshop on Subtle Expressivity for Characters and Robots*, (2003).