Analysis of User Reactions to Turn-Taking Failures in Spoken Dialogue Systems

Mikio Nakano* **Yuka Nagano**** **Kotaro Funakoshi*** **Toshihiko Ito****
Kenji Araki** **Yuji Hasegawa*** **Hiroshi Tsujino***
*Honda Research Institute Japan Co. Ltd.
8-1 Honcho, Wako, Saitama 359-0188, Japan
**Graduate School of Information Science and Technology, Hokkaido University
Kita-14, Nishi-9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan
{nakano, funakoshi, yuji.hasegawa, tsujino}@jp.honda-ri.com
{calico,t-itoh,araki}@media.eng.hokudai.ac.jp

Abstract

This paper presents the results of an analysis of user reactions towards system failures in turn-taking in human-computer dialogues. When a system utterance and a user utterance start with a small time difference, the user may stop his/her utterance. In addition, when the user utterance ends soon after the overlap starts, the possibility of the utterance being discontinued is high. Based on this analysis, it is suggested that the degradation in speech recognition performance can be predicted using utterance overlapping information.

1 Introduction

Many kinds of spoken dialogue systems have been developed in the last two decades. Most previous systems employed a fixed turn-taking strategy, that is, they take a turn when the user puts a certain length of pause after his/her utterances, and they release the turn immediately when the user barges in on a system utterance. In order to improve the usability of spoken dialogue systems, the turn-taking strategy needs to be more flexible.

Thus far, there have been several approaches to this problem. Some methods try to decide when to take a turn based on not only the length of pause but also the content and prosody of the user utterance [e.g., (Sato et al., 2002; Ferrer et al., 2003; Schlangen, 2006)]. Other methods try to decide how to appropriately react to the user barge-in utterances, not just simply stopping whenever a barge-in utterance is detected [e.g., (Ström and Seneff, 2000; Rose and Kim, 2003)].

Despite these efforts, achieving appropriate turn-taking is still difficult. The features used by these methods are not always perfectly obtained. In addition, even humans cannot sometimes decide whether the system should take a turn or not (Sato et al., 2002).

Consequently, in addition to efforts towards improving turn-taking, we need to find a way to make the system cope with turn-taking errors. As a first step, we investigated how users behave when the system made mistakes in turn-taking. We have found that users tend to stop their utterances in certain situations. We expect this to be useful in avoiding misunderstanding caused by speech recognition errors of such discontinued utterances.

2 Analysis of User Reactions to Turn-Taking Failures

2.1 Dialogue Data

We analyzed two sets of human-system dialogue data using the following two different dialogue systems in Japanese. One was a car-rental reservation dialogue system in which the user could make a reservation for renting a car by specifying the date, hour, and locations for rental and return, along with the car type. The other was a video recording system in which the user could set the date, time, channel, and recording mode (long play or short play) for recording a TV program.

Both systems performed frame-based dialogue management. They employed the Julian speech rec-

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 120–123,
Antwerp, September 2007.

ognizer directed by network grammars (Kawahara et al., 2004) with its attached acoustic models. The vocabulary size for speech recognition was 225 words for the car-rental reservation system and 198 words for the video recording system. These systems also employed NTT-IT Corporation's FineVoice speech synthesizer. When collecting the data, a microphone and headphones were used. For each dialogue, the microphone input and the system output were recorded in a stereo file.

The contents of the data sets are as follows:

- Set C: (Car-rental reservation)

  Each of the 23 subjects (12 males and 11 females) engaged in 8 dialogues (total 184 dialogues). In each dialogue, users tried to make one reservation. 134 dialogues were successfully finished within 3.5 minutes, 38 failed, and 12 were aborted because of a system trouble.

- Set V: (Video recording reservation)

  This consists of 117 dialogues (9 dialogues by each of the 13 subjects (9 males and 4 females)). These subjects are different from the subjects for Set C. In each dialogue, the user tried to set the timer to record two programs. In 41 dialogues, the user successfully set up the recordings for two programs within 3 minutes. In 36 dialogues, the user set up only one of the programs. In 34 dialogues, the user could not set up the recordings, and 6 were aborted.

Both systems had variations in dialogue and turn-taking strategies so that a variety of dialogues were recorded. Thresholds for confidence scores for generating confirmation requests were changed, parameters for speech interval detection were changed, and whether the system stopped its utterances when the user barged in was changed. For each subject, different strategies were used for different dialogues. We will not explain these variations in detail since, as we will explain later, we focused on the phenomena of turn-taking failures rather than the causes of them.

After collecting data, both user and system utterances were transcribed as pronounced. Utterance segmentation was done manually based on pauses longer than 300ms, by using an annotation tool.

set \ case	(o1)	(o2)	(o3)	total
C	67	446	7	520
V	46	202	1	249

(o1) The start time of the user utterance is between the start and end times of a system utterance.
(o2) The start times of one or more system utterances are between the start and end time of the user utterance.
(o3) Both (o1) and (o2) occur.

Table 1: Frequencies of user utterances overlapping with system utterances.

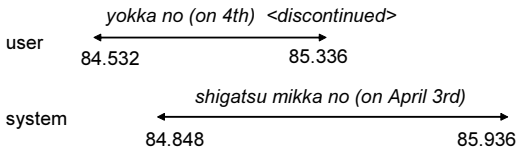

Figure 1: Example discontinuation with overlap.

The timestamps of each speech segment indicate the points in time from the start of the stereo file. Below we simply call these speech segments *utterances*. The total numbers of the user utterances and system utterances in Set C are respectively 3,364 and 5,157 and, in Set V they are 2,521 and 4,522.

2.2 Utterance Overlaps

As Raux et al. (2006) reported, there are several kinds of system turn-taking failures. The system sometimes barges in to a user utterance, and sometimes fails to take a turn. These failures are caused by several reasons, such as errors in speech interval detection, and misrecognitions of the user's intention to release a turn.

In this paper, we focus only on failures that result in overlaps between user and system utterances. We have not investigated the reason for the failure; but instead of that, we analyzed the overlapping phenomena that often occurred when the system made mistakes in turn-taking, because the goal of the analysis is not to improve turn-taking, but to find a way to recover from turn-taking failures. Table 1 shows the frequencies of user utterances overlapping system utterances.

2.3 Discontinuations

In this paper, we call utterances stopped in the middle for any reason *discontinuations*. We found that user utterances overlapping with system utterances

set	all utterances			discontinuations		
	IG	OOG	ALL	IG	OOG	ALL
C	2,662	702	3,364	9	78	87
	22.75	74.05	40.23	12.00	66.97	63.13
V	1,599	922	2,521	2	46	48
	13.08	73.89	39.69	0.00	90.43	87.39

IG means in-grammar utterances, and OOG means out-of-grammar utterances. (upper: # of utterances, lower: word error rate (%))

Table 2: Speech recognition results for all utterances and discontinuations.

d (s)	$-\infty$ – -0.4	-0.4 – -0.2	-0.2 – 0.0	0.0 – 0.2	0.2 – 0.4	0.4 – 0.6	0.6 – 1.0	1.0 – ∞
C	2/45	0/7	4/22	15/43	11/56	3/29	4/34	22/284
V	0/17	0/9	10/21	16/57	6/48	3/27	1/12	2/58

(# of discontinuations)/(# of overlapped user utterances)

Table 3: Frequency of discontinuations depending on the start time difference d.

c (s)	0.0 – 0.1	0.1 – 0.2	0.2 – 0.3	0.3 – 0.4	0.4 – 0.5	0.5 – 0.6	0.6 – 0.8	0.8 – 1.0	1.0 – ∞
C	1/50	7/44	10/67	12/66	15/52	4/36	4/75	4/45	4/85
V	1/19	4/19	9/30	13/28	2/17	3/16	2/22	0/17	4/81

(# of discontinuations)/(# of overlapped user utterances)

Table 4: Frequency of discontinuations depending on c (the length of user utterance after the overlapping starts)

are more likely to be discontinuations. Discontinuations are expected to be difficult for speech recognition mainly because they are not grammatical and include word fragments. So detecting and ignoring them would improve speech understanding. We therefore focus on analyzing discontinuations. Figure 1 shows an example of discontinuations in a car-rental reservation dialogue.

We annotated discontinuations by listening to only the user-speech channel of the stereo files. In set C, 87 utterances are discontinuations, and, in set V, 48 are discontinuations. Of these, 61 and 38 have overlaps with system utterances.

To investigate the speech recognition performance on the discontinuations, we used the same network grammar as the spoken dialogue system used in the data collection. Note that, since user speech segments are made from the timestamps in the transcriptions, they are different from those recognized at the time of data collection. As shown in Table 2, discontinuations include out-of-grammar utterances, so the word error rates are very high.[1]

2.4 Relationship between Discontinuations and Turn-Taking

One way to detect discontinuations that might be effective is to use prosodic information (Liu et al., 2003). Since prosody recognition is not yet perfect, however, it is worth exploring other methods. We therefore investigated in which turn-taking situations discontinuations are likely to exist.

Discontinuations are likely to occur when the start time of the user and system utterances are close. Table 3 shows the relationships of the frequencies of discontinuations in the overlapping user utterances depending on the start time difference d. Here, the start time difference d is defined as follows:

$$d = st(u) - st(s),$$

where $st(i)$ means the start time of utterance i, u is a user utterance and s is the first system utterance among the system utterances overlapping u. We found that people tend to stop their own utterances when d is between -0.2s to 0.4s. When d is larger than 0.4s, the user has already spoken for a while so he/she might try to finish the utterance.

Next, we investigated the end time of the overlapped user utterances, because discontinuations can be expected to occur soon after the overlapping starts. Table 4 shows the frequencies of discontinuations depending on the length of the user utterance after the overlapping starts. This is defined as c in the following formula:

$$c = \begin{cases} et(u) - st(u) & (\text{cases (o1) and (o3) in Table 1}) \\ et(u) - st(s) & (\text{case (o2) in Table 1}), \end{cases}$$

where $et(i)$ means the end time of utterance i. As we expected, when c is between 0.1s and 0.6s, the

[1] The word error rates for the out-of-grammar utterances is very high for the following reason. We transcribed the user utterances without word boundaries because it is not easy to consistently determine word boundaries for Japanese. We used a morphological analyzer to split these transcriptions into words to obtain references for computing speech recognition accuracy. This process tended to produce one-syllable out-of-vocabulary words. Therefore the references include a greater number of out-of-vocabulary words.

Set C

d (s) \ c (s)	0.0 – 0.1	0.1–0.6	0.6 – ∞
$-\infty$ – -0.2	0/0	2/12	0/40
-0.2 – 0.4	1/6	**24/62**	5/53
0.4 – ∞	0/44	22/191	7/112

Set V

d (s) \ c (s)	0.0 – 0.1	0.1–0.6	0.6 – ∞
$-\infty$ – -0.2	0/0	0/11	0/15
-0.2 – 0.4	1/2	**26/52**	5/72
0.4 – ∞	0/17	5/47	1/33

(# of discontinuations)/(# of overlapped user utterances)

Table 5: Frequency of discontinuations depending on c and d.

	Situation S			Other overlapping utterances		
set	IG	OOG	ALL	IG	OOG	ALL
C	20	42	62	285	173	458
	16.67	107.89	78.57	12.72	66.31	35.36
V	13	39	52	97	100	197
	9.52	122.73	86.15	8.44	75.06	43.14

(upper: # of utterances. lower: word error rate (%).)

Table 6: Speech recognition performance for utterances in Situation S and other cases.

user utterances are more likely to be discontinuations than other cases.

From the above analysis, the possibility that a discontinuation occurs is high when d is between -0.2s and 0.4s and c is between 0.1s and 0.6s. We call this situation, *Situation S*. Table 5 shows the frequencies of discontinuations depending on the combinations of d and c.

2.5 Predicting Speech Recognition Performance Degradation

Since discontinuations occur more frequently in Situation S than other cases, speech recognition performance would be degraded in Situation S. Table 6 shows these results. This suggests that the overlapping information can be used for predicting speech recognition performance degradation.

3 Concluding Remarks

This paper presented our preliminary analysis on user reactions to system failures in turn-taking in human-computer dialogues. We found that discontinuations are likely to occur more frequently at the overlapping utterances caused by turn-taking failure. We specified situations where user discontinuations frequently occur. It is suggested that the degradation in speech recognition performance can be predicted using utterance overlapping information. This is expected to be useful for avoiding misunderstanding.

We are planning to conduct more detailed analyses on discontinuations, such as their relationship with the subjects and the dialogue and turn-taking strategy of the system. We also plan to investigate changes in speech recognition performance when statistical language models are employed.

References

Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke. 2003. A prosody-based approach to end-of-utterance detection that does not require speech recognition. In *Proc. ICASSP-2003*.

Tatsuya Kawahara, Akinobu Lee, Kazuya Takeda, Katsunobu Itou, and Kiyohiro Shikano. 2004. Recent progress of open-source LVCSR engine Julius and Japanese model repository. In *Proc. Interspeech-2004 (ICSLP)*, pages 3069–3072.

Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. 2003. Automatic disfluency identification in conversational speech using multiple knowledge sources. In *Proc. Eurospeech-2003*, pages 957–960.

Antoine Raux, Brian Langner, Dan Bohus, Alan W. Black, and Maxine Eskenazi. 2006. Doing research in a deployed spoken dialog system: One year of let's go! public experience. In *Proc. Interspeech-2006 (ICSLP)*, pages 65–68.

R.C. Rose and Hong Kook Kim. 2003. A hybrid barge-in procedure for more reliable turn-taking in human-machine dialog systems. In *Proc. ASRU-03*, pages 198–203.

Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, and Kiyoaki Aikawa. 2002. Learning decision trees to determine turn-taking by spoken dialogue systems. In *Proc. 7th ICSLP*, pages 861–864.

David Schlangen. 2006. From reaction to prediction: Experiments with computational models of turn-taking. In *Proc. Interspeech-2006 (ICSLP)*, pages 2010–2013.

Nikko Ström and Stephanie Seneff. 2000. Intelligent barge-in in conversational systems. In *Proc. 6th ICSLP*.