Analysis of Changes in Dialogue Rhythm Due to Dialogue Acts in Task-Oriented Dialogues

Noriki Fujiwara, Toshihiko Itoh, and Kenji Araki

Graduate School of Information Science and Technology,
Hokkaido University, Sapporo, Japan
{fujiwara, t-itoh, araki}@media.eng.hokudai.ac.jp

Abstract. We consider that factors such as prosody of systems' utterances and dialogue rhythm are important to attain a natural human-machine dialogue. However, the relations between dialogue rhythm and speaker's various states in task-oriented dialogue have been not revealed. In this study, we collected task-oriented dialogues and analyzed the relations between "dialogue structures, kinds of dialogue acts (contents of utterances), *Aizuchi* (*backchannel*/*acknowledgment*), *Repeat* and interjection" and "dialogue rhythm (response timing, F0, and speech rate)".

1 Introduction

It is widely accepted that the accuracy of speech recognition and the understanding of language as well as the quality of synthesized speech are important to accomplish natural human-machine communication. The abilities of spoken-dialogue systems have recently increased exponentially. Numerous systems have been developed [1] and some of these are being used in practical applications [2]. However, the communication between humans and machines is not as natural as that between humans. Our previous study [3][4] revealed that factors such as prosody of systems' utterances and dialogue rhythm are important to attain a natural human-machine dialogue.

Kitaoka et al.[5] were interested in dialogue rhythm and developed a free-conversation spoken-dialogue system. They achieved this goal by using machine learning only on keywords and acoustic features from human-human dialogues. We were also interested in dialogue rhythm and developed a spoken-dialogue system for task-oriented dialogue. Our system has the same ability as Kitaoka's but it can also tune the acoustic features of system response to those of a user's utterances [6]. This is because the acoustic features (response timing, F0, and speech rate) of speakers' utterances in free-conversation are claimed to become synchronized with those of their partners' utterances along with increased tension in the dialogue [7]. Although the dialogue rhythm in our system did improve, it was not as smooth and natural as that of a human's. The speakers' state (rise in dialogue tension, dialogue act, and emotions) is usually claimed to influence dialogue rhythm. We believe that dialogue acts (contents of utterances) are particularly important factors in dialogue rhythm. However, our system did not use speaker's dialogue acts to attain dialogue rhythm. It is also not natural for spoken-dialogue systems to always tune the acoustic features of responses to speakers'

V. Matousek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 564–573, 2007.

utterances regardless of dialogue acts. However, there have been no studies that have investigated the relations between dialogue acts and dialogue rhythm, and there have been no studies that have investigated phenomena such as acoustic synchronicity when task-oriented dialogue is concerned. It is therefore necessary to more thoroughly investigate the rhythm of human-human dialogue to achieve a spoken-dialogue system that enables communication like that between humans.

We collected task-oriented human-human dialogue for the present study, and analyzed the relations between dialogue acts and dialogue rhythm, i.e., the response timing, F0, and speech rate.

2 Dialogue Corpus

We recorded task-oriented dialogue to analyze human-human dialogue and annotated it with dialogue-act tags and acoustic labels. The details on the process are described below.

2.1 Recording Speech Data

There was a total of 17 subjects, who were undergraduate and graduate university students. The dialogue task was a hotel reservation where one section of the dialogue was spoken by two of these subjects. The first subject played the role of a customer who made a reservation. The second subject played the role of an agent, who searched for hotels and confirmed the reservation. There are cases where the same subject played the role of the customer in one dialogue, and played the agent in another dialogue. Customers interacted according to a "situation" prepared beforehand, and adhered to its context as much as possible. We prepared seven situations. Two subjects are separated by a partition, which makes them invisible to each other. They can only communicate by speaking, without gestures or eye contact.

2.2 Acoustic Labelling

We detected the beginning and ending of utterances with speech and waveforms, and labelled each utterance "agent_start", "agent_end", "customer_start", and "customer_end" using "Wavesurfer" speech-analysis software [10]. If there was a pause that lasted longer than 300 ms, we regarded it as a border between utterances. Therefore, one dialogue act, as described in Section 2.3, often consisted of more than one utterance.

2.3 Dialogue Act Tagging

A dialogue usually consists of more than one exchange, and the structure of an exchange is usually "Initiate-Response-(Follow-up)". Initiate is a component that functions as an appeal to start a new exchange. Response is a component that functions as a reaction to initiate. Follow-up is component that functions to signal that the current exchange has finished, and is often omitted. Each component includes some dialogue acts. We consulted the literature [8][9] when we defined kinds of the dialogue acts and tagged them. The tagging procedure is described below.

1. Making transcriptions from collected dialogue data.
2. Splitting utterance (transcription) of each speaker by one dialogue act.
3. Judging and tagging the kind of dialogue act to the each splitted utterance according to a decision tree [9] proposed by Discourse Tagging Working Group in Japan.

The dialogue acts that we analyzed and discuss in this paper are described below. The numbers in parentheses denote how many times an agent or a customer used a given dialogue act in the dialogue corpus.

Initiate:

- *Wh-question* (For agents:248, For customers:184): A demand for some values or expressions as a response to a question where the speaker has not forecast his or her partner's response.
- *Request* (For agents: 138, For customers: 131): A demand is made for the listener to act, and some response indicating acceptance or rejection is needed.
- *Inform* (For agents: 50, For customers: 7): Expressing an opinion, knowledge, or facts that the speaker believes to be true.
- *Yes-No question* (For agents: 47, For customers: 35): Answers "Yes" or "No" to a question when the speaker cannot predict his or her partner's response.
- *Confirm* (For agents: 547, For customers: 55): A question is asked by a speaker who can make a prediction or has knowledge about his or her partner's response.

Response:

- *Answers* (For agents: 270, For customers: 242): Utterance that provides content to the demand in *Wh-questions*.
- *Positive* (For agents: 127, For customers: 563): An affirmative response to a *Yes-No question* and acceptance of a demand, request, or preposition.
- *Negative* (For agents: 5, For customers: 28): A negative response to a *Yes-No question* and rejection of a demand, request, or preposition.

Follow-up:

- *Understand* (For agents: 165, For customers: 168): Expressing that the goal of an exchange has been achieved after a response.

Aizuchi & Repeat:

- *Aizuchi* (*backchannel/acknowledgment*) (For agents: 295, For customers: 870): Aizuchi signifies the partner's speech has been heard or the next utterance is prompted (its function is not a definite answer but rather a lubricant to enable smoother conversation).
- *Repeat* (For agents: 187, For customers: 38): An utterance that repeats important words (keywords) included in the preceding speaker's utterances.

Here, *Aizuchi* and *Repeat* are basically included in the Follow-up. However, the contributions of these dialogue acts to dialogue rhythm are considered to differ from Follow-up and we have dealt with *Aizuchi* and *Repeat* as other dialogue acts. Fig. 1 shows an example of such a dialogue tagged with dialogue acts, and Table 1 lists the information from our dialogue corpus.

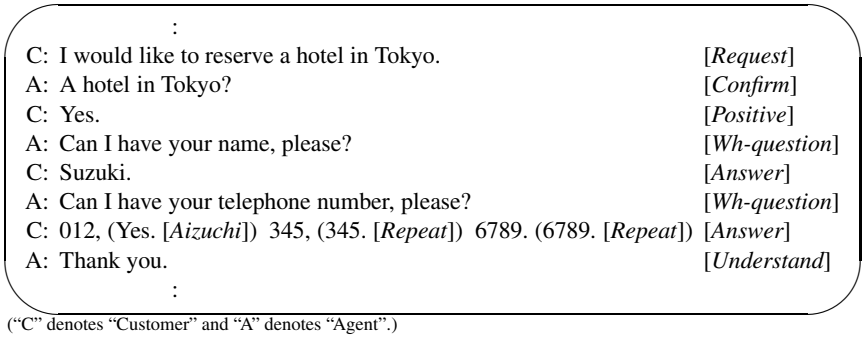

Fig. 1. Example of tagged dialogue acts

Table 1. Dialogue data

# of dialogues		50
# of subjects		17
# of utterances	agent	3844
	customer	3510
# of dialogue acts	agent	2215
	customer	2520
Ave. dialogue duration		4 min. 57 sec.

3 Analysis

We investigated the response timing, F0, and speech rate as factors contributing to dialogue acts. The response timing was calculated by subtracting the ending of a previous dialogue act from the beginning of a current one. We calculated the average log (F0), which was calculated by dividing the sum of log (F0) by the number of frames for analysis except for voiceless frames. We used "ESPS/waves+" speech analysis software [11] to estimate F0. The speech rate was calculated by dividing the number of morae by the duration of the utterance.

3.1 Analysis of Relations Between Initiate, Response, and Follow-Up

We analyzed the dialogue rhythm of utterances based on Initiate, Response, and Follow-up. Table 2 lists the averages and standard deviations for response timing, the average log (F0), and the speech rate of utterances based on Initiate, Response and Follow-up. The response timing for Response is earlier than Initiate's, and Response's speech rate is slower than Initiate's. Initiate is basically an utterance to start a new exchange and a new topic, and to dominate and manage the flow of dialogue to achieve a task. We therefore considered that Initiate needed a longer time to think about what to say than Response, and the admissible pause to commence speaking could be extended. Response is a reaction to Initiate, except in situations when information is being retrieved,

Table 2. Response timing, average log (F0), and speech rate for Initiate, Response, and Follow-up

		Initiate	Response	Follow-up
Response timing $[sec]$	Ave.	1.06	0.44	0.50
	SD	1.00	0.74	0.81
Ave. of log(F0)	Ave.	4.71	4.73	4.61
	SD	0.24	0.29	0.20
Speech rate $[mora/sec]$	Ave.	9.09	8.81	9.60
	SD	2.06	3.02	3.61

Response's thinking time was shorter than Initiate's. In addition, as utterances based on Response were affected by temporal restrictions based on the exchange structure and real time, the admissible pause to begin speaking must be short. However, as Response was more frequently included in important content than Initiate, its speech rate slowed to attract his or her partner's attention and convey content accurately. Moreover, the response timing for Follow-up is almost the same as for Response. The reason is the same as for Response which was described above. However, Follow-up's average log (F0) was lower and its speech rate was much faster than the others. The reason for this is that utterances based on Follow-up were considered to be optional utterances and the speaker thought that they were less important than the others.

3.2 Analysis Comparing Dialogue Acts

We analyzed the dialogue rhythm of utterances based on all dialogue acts (see Figs. 2, 3, and 4). First, we describe the results for turn-taking which is frequent in a dialogue and highly contributive to dialogue rhythm. In turn-taking, there is a tendency for response timing to occur early in the order of "dialogue acts belonging to Initiate", "dialogue acts belonging to Response (except *Positive*)" and "dialogue acts belonging to Follow-up". The response timing only for *Positive* occurs especially early. This is because most utterances based on *Positive* are responses to *Confirm*, which is a dialogue act to confirm information on the current task and it can easily be predicted during the process of achieving a task. Therefore, we regard utterances that are based on *Positive* to be responses that require hardly any thinking time (yes/no responses). There are therefore many overlaps in this dialogue act. There are tendencies for the response timing of the other dialogue acts, which often treat new or important information, such as *Wh-question* and *Request*, to be delayed and to only occasionally overlap. However, there are tendencies for the response timing of the dialogue act in easily predicted utterances, such as *Confirm* and *Yes-No question*, to speed up and frequently overlap.

There are significant differences in almost all combinations of dialogue acts (p <0.01) according to the results of the t-test for average log (F0). The results also show that F0 is easily affected by dialogue acts. There are basic tendencies in dialogue acts where utterances are predictable or expected by a partner, such as *Positive* and

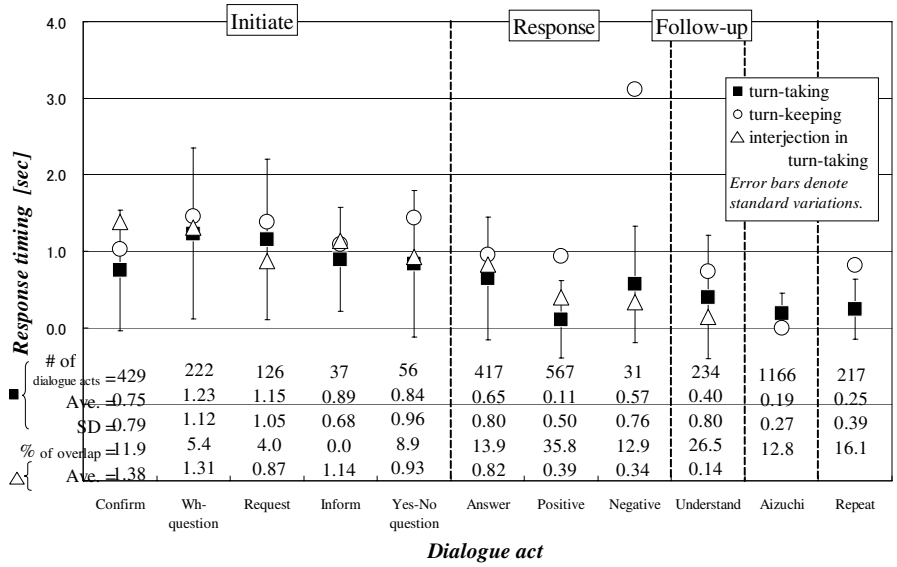

Fig. 2. Response timing for all dialogue acts

Understand, for average log (F0) to become low, and in dialogue acts where utterances are important or unexpected by a partner, for average log (F0) to become high (tendency for emphasis).

The speech rate of dialogue acts that often include new, unexpected, or important information is slow. Utterances based on the dialogue acts which modulate the dialogue rhythm are very fast. Although we thought that the thinking time for *Negative* was almost the same as or a little later than that for *Positive*, *Negative's* response timing was delayed, its average log (F0) was very high, and its speech rate was slow. We considered that this was because utterances based on *Negative* were important in the sense that they differed from the partner's expectation; therefore, the speaker emphasized them on purpose. In the case of dialogue acts belonging to Response, if "utterances which level of importance is high" were assumed to be "utterances with late response timing", there were strong correlations between response timing and F0 (r=0.84), between response timing and speech rate (r=-1.00), and between speech rate and F0 (r=-0.86). Briefly, there was a tendency by utterances which level of information was high for their response timing to be delayed, their F0 to become high and speech rate to slow; they were often emphasized. In the case of dialogue acts belonging to Initiate, there were strong correlations between response timing and F0 (r=0.97) but there were no correlations between other combinations.

Finally, in the case of relations for response timing, F0, and speech rate in each dialogue act, turn-keeping (when a partner does not take turn) appears to have the same tendencies as turn-taking. However, a close analysis reveals that the response timing is later, the average log (F0) is lower and the speech rate is faster for turn-keeping when compared with turn-taking.

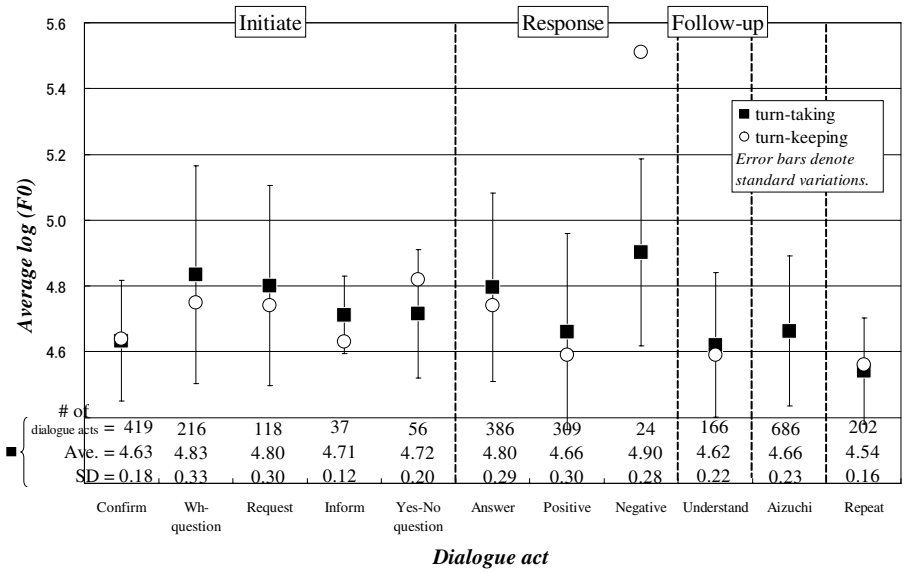

Fig. 3. Average log(F0) for all dialogue acts

3.3 Analysis Comparing Aizuchi and Repeat

The averages and standard deviations of response timing for *Aizuchi* and *Repeat* were almost the same, and their standard deviations were much smaller than those for the other dialogue acts. The reason for this is that the function to modulate dialogue rhythm is strong and there are stringent constraints about response timing, as is the case for *Aizuchi* and *Repeat*; therefore, a human may respond almost reflexively using various features as acoustic ones. The speech rate for *Repeat* is almost the same as that for *Confirm's*, which is semantically the same but *Repeat's* F0 is much lower than that for the other dialogue acts. The reason for this is that *Repeat* involves implicit confirmation by repeating the keyword(s) included in the partner's utterance, but a speaker intentionally lowers *Repeat's* F0 not to disturb his or her utterances. As above, in the case of utterances that function to modulate dialogue rhythm, if they include information to convey to a partner, such as *Repeat*, their speech rate is normal and their F0 is significantly lowered in order not to disturb partner's utterances. If one partner's utterances have no information to be conveyed (as in *Aizuchi*), their F0 is lowered and speech rate becomes very fast in order not to disturb another partner. This is how dialogue rhythm is being preserved.

3.4 Response Timing for Interjections (Filled Pause)

The average of response timing for interjections is 0.94 and its standard deviation is 0.97. We analyzed the relations between the average of response timing for all dialogue acts and the average of response timing for interjections of all dialogue acts (see

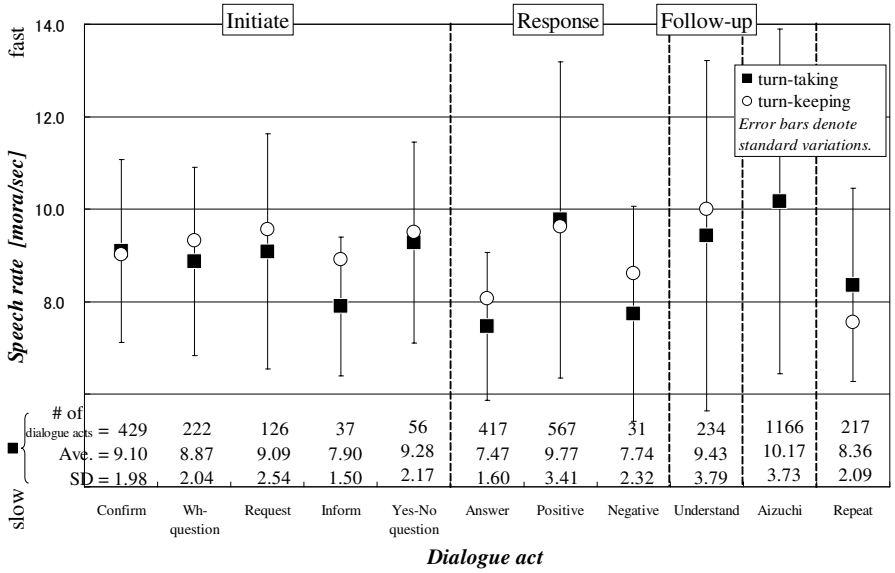

Fig. 4. Speech rate for all dialogue acts

triangles in Fig. 2), and we obtained some very interesting results. There is a strong correlation between the average of response timing for all dialogue acts and the average of response timing for interjections (r=0.73). The equation of regression line is $y = 0.92x + 0.14$, where x denotes the average of response timing for all dialogue acts and y denotes the average of an interjection's response timing for all dialogue acts. The interjection's response timing for the same thinking state, such as *Positive* and *Negative*, are almost the same, and the interjection's average of response timing for almost all dialogue acts are later than the average of response timing for all dialogue acts. We consider that there is an admissible pause for thinking time (response timing) in all dialogue acts, and the speaker utters interjection when he or she has not decided what to say yet in the admissible pause or he or she has predicted that his or her utterance will be later than the admissible pause.

3.5 Toward Accomplishment of Natural and Smooth Human-Machine Communication

From the results described in this section, we took the following into consideration in a task-oriented dialogue between two persons. We considered that there was an admissible pause (thinking time) in all dialogue acts. The pause is determined by kinds of dialogue act (including their exchange structure types) and a state of dialogue structure (turn-taking / turn-keeping). And then, the speaker utters interjection when he or she has not decided yet what to say in the admissible pause or he or she has predicted that his or her utterance will be later than the admissible pause. In the case of actual each response timing, we consider that it is determined by relations between "an admissible

pause (thinking time)" and "speaker's dialogue act, progress of determining response sentences, and level of importance for utterance contents". Therefore, it is considered that the response timing is affected by "ease of predicting the preceding partner's utterance", "difficulty of utterance contents", "dialogue act", "level of importance and novelty of utterance contents", "gap between partner's expectation and speaker's actual utterances", "time for retrieving informations" and so on. Furthermore, a decision of F0 and speech rate on the whole utterance is heavily affected by "dialogue act", "level of importance and novelty of utterance contents" and "gap between partner's expectation and speaker's actual utterances". We believe that a listener (a partner) has understood and, to some extent, modeled general response timing, F0, speech rate, and if possible, their individual averages for each speaker, the listener obtains nonlinguistic informations from the difference between the model and the actual response timing, F0, and speech rate. It is necessary to construct a model to estimate these relations using human-human dialogue in order to attain a natural and smooth dialogue rhythm in task-oriented dialogues.

4 Conclusion

In this study, we collected task-oriented dialogues and analyzed the relations between "dialogue structures, kinds of dialogue acts, *Aizuchi*, *Repeat*, and interjection" and "dialogue rhythm (response timing, F0, and speech rate)". We gained important knowledge for achieving a smooth and natural spoken-dialogue in task-oriented dialogue system.

Future work is to investigate the shift and synchronization of dialogue rhythm in a task-oriented dialogue, and the relations between dialogue rhythm and the current speaker's dialogue act taking the partner's previous dialogue act into consideration. Moreover, using the results, we plan to improve our spoken-dialogue system in order to enable smoother and more natural communication on the level comparable to human's.

References

1. Raymond, C., Esteve, Y., Bechet, F., De Mori, R., Damnati, G.: Belief confirmation in spoken dialog systems using confidence measures. In: Proc. of ASRU 2003, pp. 150–155 (2003)
2. Pouteau, X., Krahmer, E., Landsbergen, J.: Robust spoken dialogue management for driver information systems. In: Proc. of Eurospeech'97, pp. 2207–2210 (1997)
3. Yamada, S., Itoh, T., Araki, K.: Linguistic and Acoustic Features Depending on Different Situations - The Experiments Considering Speech Recognition Rate. In: Proc. of INTERSPEECH 2005, pp. 3393–3396 (2005)
4. Yamada, S., Itoh, T., Araki, K.: Is Voice Quality Enough? - Study on How the Situation and User's Awareness Influence the Utterance Features. In: Proc. of INTERSPEECH 2006, pp. 481–484 (2006)
5. Kitaoka, N., Takeuchi, M., Nishimura, R., Nakagawa, S.: Response Timing Detection Using Prosodic and Linguistic Information for Human-friendly Spoken Dialog Systems. Journal of JSAI, SP-E 20(3), 220–228 (2005)
6. Shoji, K., Takahashi, M., Ibara, S., Itoh, T., Araki, K.: Spoken Dialog System considered Rhythm and Synchronized Tendency of Conversation. (in Japanese) IPSJ SIG Tecnical Reports, SLP-61, pp. 43–48 (May 2006)

7. Nagaoka, C., Komori, M., Nakamura, T.: The interspeaker influence of the switching pauses in dialogue (in Japanese) The Japanese Journal of Ergonomics 38(6), 316–323 (2002)
8. Ichikawa, A., Araki, M., Kashioka, H., et al.: Evaluation of Annotation Schemes for Japanese Discourse. In: Proc. of ACL '99 Workshop on Towards Standards and Tools for Discourse Tagging, pp. 26–34 (1999)
9. Araki, M., Itoh, T., Kumagai, T., Ishizaki, M.: Proposal of a Standard Utterance-Unit Tagging Scheme (in Japanese) Journal of JSAI 14(2), 251–260 (1999)
10. http://www.speech.kth.se/wavesurfer/index.html
11. Software manuals of ESPS/waves+ with EnSigTM, Entropic Research Laboratory, Inc. (1997)