

# Subjective Experiments on Influence of Response Timing in Spoken Dialogues

Toshihiko ITOH<sup>†</sup>, Norihide KITAOKA<sup>‡</sup> and Ryota NISHIMURA<sup>††</sup>

<sup>†</sup>Hokkaido University, <sup>‡</sup>Nagoya University and <sup>††</sup>Toyohashi University of Technology

t-itoh@media.eng.hokudai.ac.jp, kitaoka@nagoya-u.jp, nishimura@slp.ics.tut.ac.jp

## Abstract

To verify the validity of analysis results relating to dialogue rhythm from earlier studies, we produced spoken dialogues based on analysis results relating to response timing and the other spoken dialogues, and performed subjective experiments to investigate parameters such as the naturalness of the dialogue, the incongruity of the synthesized speech, and the ease of comprehension of the utterances. We used very short task-oriented four-turn dialogues using synthesized speech in Experiment 1, and approx. one-minute free-conversation dialogues in Experiment 2 using natural human speech and synthesized speech. As a result, we were able to show that a natural response timing exists for utterances, and that response timings that conform to the utterance contents are felt to be more natural, thus demonstrating the validity of the analysis results relating to dialogue rhythm.

**Index Terms:** Speech analysis, Speech communication, Interactive systems, User interface human factors, User interfaces

## 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 and some of these are being used in practical applications. However, the communication between humans and machines is not as natural as that between humans. Our previous study [1, 2] revealed that factors such as prosody of systems' utterances and dialogue rhythm (response timing, F0, and speech rate) are important to attain a natural human-machine dialogue.

We were interested in dialogue rhythm and developed spoken dialogue systems [3, 4]. We achieved a generation method of dialogue rhythm by using machine learning only on keywords and acoustic features trained from human-human dialogues. Our system can also tune the acoustic features of system response to those of a user's utterances. This is because the acoustic features of speakers' utterances in free-conversation are claimed to become synchronized with those of their partners' utterances along with increased tension in the dialogue [5]. 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 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' 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.

So, we collected task-oriented human-human dialogue, and analyzed the relations between dialogue acts and dialogue rhythm. As a result, we gained important knowledge for achieving a smooth and natural spoken dialogue in task-oriented dialogue system [6].

In this paper, to verify the validity of analysis results relating to dialogue rhythm from earlier studies, we produced spoken dialogues based on analysis results relating to response timing and performed subjective experiments.

We start by briefly discussing the analysis results obtained in previous studies, and then we discuss the subjective experiments performed in this study.

## 2. Toward accomplishment of natural and smooth human-machine communication

We briefly discuss our analysis of dialogue acts and dialogue rhythms [6] (response timing, F0 and speech rate) in order to implement natural and smooth man-machine dialogue. We analyzed the differences in dialogue rhythm with different dialogue acts, and we investigated parameters such as trends in the distribution of response timings of different roles, the effects of individual differences in dialogue rhythm, the relationship between detailed dialogue acts (utterance contents) and dialogue rhythm, and the effects of the other person's previous utterance on the current speaker's dialogue rhythm. From these results, we inferred that the dialogue rhythm in a task-oriented dialogue between two speakers is characterized as follows. First, although there are individual effects (individual differences) in the relationship between dialogue act and dialogue rhythm, these do not cause much disruption in the relationship between each dialogue act, and exhibit some degree of generality. As for the response timings, these depend on the dialogue act. If we conduct an analysis using more detailed units of utterance contents, then we find that the variance is smaller and the response timing is fixed to some degree. However, there is a range of possible response timings even in the utterance contents, and although the magnitude of this range varies significantly depending on the contents of each utterance, the start position (time) of this range is more or less the same for all utterance contents. The size of this range is determined by the contents of the utterance (including the type of utterance within the exchange structure), the state of the dialogue structure (alternating speakers or continuation by the same speaker), and whether or not there is a change of subject (i.e., whether or not the speaker is starting a new subject). Basically, the response timing range is narrower when responding to an earlier statement than when initiating a new exchange (an Initiate-Response or an Initiate-Response-Follow-up sequence). Also, at positions where there is a change

of subject, the range of response timings becomes wider as a result of rationalization relating to the task and planning the flow of the dialogue, and the response timing is liable to vary. The interval up to the end of this response timing range is mainly where the speaker decides what to say. If the speaker cannot decide what to say within this response timing range or expects to be unable to do so, an interjection is uttered. The actual final response timing is thought to be determined by parameters such as the time duration of the partner's utterance, the predictability of the partner's previous utterance, the ease with which the contents of the partner's utterance could be understood, the actual contents of the utterance made by the current speaker, the importance of the contents of this utterance, the difference between what the partner maybe predicted (expected) to be said and what the current speaker actually decided to say and the time taken by the current speaker to retrieve the necessary information. Furthermore, a decision of F0 and speech rate on the whole utterance is heavily affected by parameters such as the actual contents of the utterance made by the current speaker, the importance of the contents of this utterance and the difference between what the partner maybe predicted (expected) to be said and what the current speaker actually decided to say.

We believe that a listener has understood and, to some extent, modeled general response timing, F0, speech rate, and if possible, their individual averages for each speaker, and the listener obtains non-linguistic 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.

### 3. Subjective experiments relating to dialogue rhythm

In the previous section, we discussed our findings from the analysis of dialogue rhythms, and the elements that are necessary for creating a spoken dialogue system capable of natural human-like spoken dialogue. To verify the validity of the analysis results obtained so far, we performed subjective experiments using human test subjects. These experiments were conducted to test the following two hypotheses based on the analysis of results:

- The listener models the general response timing in dialogue acts and utterance contents.
- The listener feels that human response times conforming to the utterance contents are the most natural.

To verify these hypotheses, we performed two types of subjective experiment.

#### 3.1. Subjective Experiment 1

In this experiment, very short four-turn task-oriented dialogues were generated by synthesized speech while varying the response timing of each speaker in the dialogue. The test subjects were asked to evaluate parameters such as the overall naturalness of the dialogue, their feelings of incongruity relating to the synthesized speech quality, and their impression of each speaker.

##### 3.1.1. Speech Samples

The following table shows the dialogue contents of the speech samples prepared for these experiments, together with the dialogue acts (shown on the left beneath each utterance) and the detailed dialogue acts level (that is, dialogue acts considering utterance contents shown on the right beneath each utterance). In the dialogue example,  $\alpha$ ,  $\beta$  and  $\gamma$  indicate the pause durations between the end of the previous utterance and the start

of this utterance. The experimental speech samples were made by varying these response timings. The following 7 response timing patterns were used in the experiments. The values in brackets represent the actual response timings used in the experiments (in units of seconds).

Pause duration	Speaker	Utterance
	A:	What is your name? [Wh-question] [Name-question]
$\alpha$	B:	My name is Suzuki Taro. [Answers] [Name-answer]
$\beta$	A:	Is your name Suzuki Taro? [Confirm] [Name-confirm]
$\gamma$	B:	Yes. [Positive] [Name-conf-posit]

1-1 Average value of the response timing of each dialogue act ( $\alpha$ :0.65,  $\beta$ :0.75,  $\gamma$ :0.11)

1-2 Average value of the response timing of each utterance content ( $\alpha$ :0.21,  $\beta$ :0.47,  $\gamma$ :0.09)

1-3 Random application of the average response timing value of each utterance contents-Pattern1 ( $\alpha$ :0.47,  $\beta$ :0.09,  $\gamma$ :0.21)

1-4 Random application of the average response timing value of each utterance contents-Pattern2 ( $\alpha$ :0.09,  $\beta$ :0.21,  $\gamma$ :0.47)

1-5 Response timing fixed at 1 second ( $\alpha$ :1.00,  $\beta$ :1.00,  $\gamma$ :1.00)

1-6 Response timing fixed at 0.5 second ( $\alpha$ :0.50,  $\beta$ :0.50,  $\gamma$ :0.50)

1-7 Response timing fixed at -0.3 second ( $\alpha$ :-0.30,  $\beta$ :-0.30,  $\gamma$ :-0.30)

The utterance speech in the speech samples was produced using the FineSpeech SDK V2.1. To make the difference between the speakers clearly discernible, speaker A was synthesized with a female voice, and speaker B was synthesized with a male voice. Parameters such as the voice pitch, emotion, intonation and speech rate were all set to SDK's default values, and the synthesized speech was not subjected to any special post-processing or the like.

##### 3.1.2. Experimental Procedure

The experiments were performed by two groups of test subjects. The first group consisted of 39 male and female IT graduate students from N University, and the second group consisted of 58 male and female IT undergraduates from H University. None of the test subjects in either group had any specialist knowledge of subjects such as speech recognition, spoken language processing or speech synthesis. With both groups, the experiments were performed in an ordinary classroom capable of holding from 70 to 80 students, and no special measures were taken to reduce background noise or echoes. A speaker was centrally placed at the front of the classroom, and its orientation, volume level and the like were adjusted so that the speech could be heard by in the whole classroom. Next, we performed subjective experiments on the two groups of test subjects according to the following procedure.

1. The test subjects were given papers containing a description of the content of the experiment, the dialogue content of the speech samples, a list of impressions and a questionnaire. After the test subjects had familiarized themselves with the content of these papers, they were also given a brief verbal description of the experiments.

2. The 7 speech sample patterns were played through in random order without any evaluation.
3. The speech samples were then selected at random (the order was different from that in Step 2.) and played back three times each, and the test subjects were asked to evaluate them on a 7-point scale with regard to how natural or incongruous they sounded. Then, if possible, the test subjects were asked to provide their impressions of each speaker either by selecting words from a list, or by providing free written responses.

The first group of test subjects performed the experiment without being given any special instructions regarding the naturalness of the dialogue, while the second group were instructed to judge the naturalness of the dialogue by focusing on the utterance contents and response timings. Also, to investigate the fluctuation of evaluation responses, the test subjects in the first group were played the same 7 patterns of speech samples a second time and asked to evaluate them again. For this second evaluation, the speech samples were again presented in random order, except to ensure that each speech sample was different from the one the test subjects had just been asked to evaluate.

### 3.1.3. Experimental Results

The results of subjective experiment 1 are shown in Table 1. In this table, the “Natural” columns show the average of the 7-point evaluation scores for naturalness of the dialogues, and the “Quality” columns show the 7-point evaluation scores for naturalness of the synthesized speech.

Table 1: *The results of subjective experiment 1*

	Group 1				Group 2	
	Set 1		Set 2		Set 3	
	Natural	Quality	Natural	Quality	Natural	Quality
1-1	3.5⑤	3.5⑥	4.6③	4.5④	4.1④	4.0④
1-2	4.7②	4.4①	4.8②	4.4⑥	5.4②	4.5①
1-3	3.0⑥	3.8④	4.4④	4.7①	3.5⑥	3.7⑥
1-4	3.8④	4.0③	3.9⑤	4.5④	4.2③	4.2③
1-5	4.5③	2.8⑦	3.3⑥	4.7①	3.7⑤	3.8⑤
1-6	5.3①	4.2②	5.5①	4.7①	5.9①	4.5①
1-7	1.6⑦	3.6⑤	1.6⑦	3.7⑦	1.2⑦	3.3⑦

Interestingly, for each evaluation set and each “Natural” item, the speech samples that gained the first (1-6) and second (1-2) highest scores and the lowest (1-7) score all occurred with the same response timing pattern. This shows that the dialogue rhythm (response timing, etc.) has a strong effect on the naturalness of the dialogue, and that a suitable (natural) response timing does exist. There is also a strong correlation between the naturalness of the response timing and the naturalness of the synthesized speech, suggesting that as the response timing becomes more natural, the feeling of incongruity of the synthesized speech becomes weaker. However, the highest evaluation scores were gained by speech sample 1-6, where the response timing was fixed at 0.5 seconds, whereas speech sample 1-2 - which used response timings based on the utterance contents proposed from the results of our earlier dialogue analysis - gained a high score but only achieved second place. In the response timing patterns with the lowest evaluation scores, the speech samples were always overlapping, which shows that care must be taken with regard to overlaps even when they are within the range of human response timings. The ranking of the other response timing patterns varied between the evaluation sets, but if we investigate their overall average scores then the speech samples can be ranked as follows: 1-1 > 1-4 > 1-5 > 1-3. This shows that the use of response timings with dialogue acts were evaluated as being more natural than random response timings.

If these results are accepted at face value, then we would have to conclude that the response timings in a dialogue sound most natural and preferable when fixed at a constant timing of approximately 0.5 seconds, which is the approximately median value of the response timing range of human utterances. We therefore performed an analysis from a slightly different viewpoint. As for the overall impressions and comments regarding the speaker of each sample, speech sample 1-6 mostly created impressions such as “calm”, “rational”, “cool and collected” and “relaxed”, while speech sample 1-2 mostly created impressions such as “human”, “sincere”, “impatient” and “tense”. The test subjects also made general remarks to the effect that since the synthesized speech was spoken slowly and indifferently, the presence of gaps made it sound more natural. It is certainly the case that the speech samples used in the experiments were all synthesized with the default settings for emotion, intonation, utterance rate and the like, and that the emotion, intonation and utterance rate were not the same as if the speech contents had been spoken by real people. It is thus possible that these results were obtained because we combined human-like response timings with the machine-like emotion, intonation and utterance rate of synthesized speech. Therefore in subjective experiment 2, these points were taken into consideration.

## 3.2. Subjective Experiment 2

In this Experiment, we concentrated on the samples that obtained the highest scores for naturalness of the dialogue in subjective experiment 1 - i.e., the samples with response timings based on the utterance contents, and the samples with a fixed response timing of 0.5 seconds - and we investigated the importance of response timing to the naturalness of dialogue in real speech, and the relationship between the response timing and other dialogue rhythm factors besides the response timing, including the emotional information, intonation, pitch and speech rate of the dialogue speech.

### 3.2.1. Speech Samples

The speech samples used in this experiment were obtained from the CSJ corpus (Corpus of Spontaneous Japanese) [7] provided by the National Institute for Japanese Language. The particular spoken dialogue used in the experiment was a relatively lively extract from an informal conversation between two speakers from the CSJ corpus (D02F0025). The extract consisted of a total of 40 utterances - 22 from the L-speaker, and 18 from the R-speaker. Based on this spoken dialogue data, we produced the following four types of speech sample:

- 2-1 Utterances in the dialogue consisted of actual speech samples which were used with the unaltered response timings of the actual speech (original spoken dialogue)
- 2-2 Utterances in the dialogue consisted of actual speech samples, but all the response timings were fixed at 0.5 seconds
- 3-1 Utterances in the dialogue consisted of synthesized speech samples which were used with the unaltered response timings of the actual speech
- 3-2 Utterances in the dialogue consisted of synthesized speech samples, and all the response timings were fixed at 0.5 seconds

In the speech synthesis system, we used the Galatea Talk software [8] included in the Galatea toolkit to generate speech utterances from transcript data included with the corpus. Parameters such as the voice pitch, emotion, intonation and speech rate were all set to their default values. The original speech was a dialogue between two females, but we used a male and a female speaker to generate the synthesized speech (changing the

R-speaker to a male voice) because it is difficult to distinguish between speakers in synthesized speech. Utterances containing no utterance contents (information) such as back-channeling or filler words were used with their original timing including any overlaps.

### 3.2.2. Experimental Procedure

The experiment environment of this experiment was basically identical to that of subjective experiment 1. The test subjects were the same 58 male and female IT undergraduate students from H University that constituted the second group in subjective experiment 1. We performed subjective experiments on the group of test subjects according to the same procedure in subjective experiment 1.

### 3.2.3. Experimental Results

The results of subjective experiment 2 are shown in Tables 2 and 3. In these tables, the “Natural” column shows the 7-point evaluation scores for naturalness of the dialogue, the “Speech” columns show the scores for agreement between the prosodic features and the utterance contents, the “Intention” column shows the scores for the ease of understanding the intention and emotion of the utterances, the “Quality” column shows the scores for the naturalness of the synthesized speech, and the “Comprehension” column shows the scores for the ease of comprehension of the utterances. From the comparative tests with real speech, we found that when non-linguistic information in the uttered speech such as emotion, intonation and speech rate is conformant with the utterance contents, response timings based on the utterance contents were rated more highly than the fixed response timing of 0.5 seconds in all the evaluations. Conversely, in cases where the utterance contents are the same but the speech is synthesized, better evaluation results were obtained in all cases for dialogue speech with a fixed response timing of 0.5 seconds. The reason why a fixed response timing of 0.5 seconds was evaluated highest in subjective experiment 1 is thought to be because the best balance in synthesized speech was achieved with speech having fixed non-linguistic information such as emotion, intonation, pitch and speech rate. In other words, natural response timings alone do not outweigh other non-linguistic information and do not have sufficient importance or priority to be able to govern the overall naturalness of a dialogue. Consequently, to determine the response timing, although a natural timing can to some extent be determined from the utterance contents, to achieve dialogue that really sounds human it is necessary to determine the response timing by taking factors such as emotion in to consideration and to simultaneously produce speech whose pitch, intonation and speech rate conform to this emotion. The balance of these factors is also important. Conversely, no matter how natural the response timing is, if the non-linguistic information in the speech is insufficient then it results in feelings of incongruity and unnaturalness. In the future, it will probably be necessary to determine the response timing by taking into consideration other parameters such as the speech synthesis performance when the response timing is taken into consideration in spoken dialogue systems and the like. Also, although it seems to be difficult to fully estimate the emotion and dialogue acts from the response timing alone, it seems that this is one feature that can be used to provide an appreciable amount of information. Finally, with regard to overlaps, spoken dialogues with overlaps were also used in subjective experiment 2. But although no incongruity whatsoever is felt in samples of real speech, the evaluation results were particularly poor for overlapping parts of synthesized speech samples with regard to their naturalness, incongruity, ease of listening and so on. To implement natural human-like overlaps, it is not sufficient merely to aim for human-like utterance contents and

response timings, and it seems that the combination and balance of non-linguistic information are also important. In the future, further detailed analysis in relation to these factors will be required.

Table 2: The results of subjective experiment 2 (real)

	Natural	Speech	Intention
2-1	6.0	6.1	5.8
2-2	5.6	5.6	5.3

Table 3: The results of subjective experiment 2 (synthesized)

	Natural	Speech	Intention	Quality	Comprehension
3-1	2.4	1.5	1.5	1.7	1.8
3-2	3.4	1.8	2.0	2.1	2.6

## 4. Conclusion

To verify the validity of analysis results relating to dialogue rhythm from earlier studies, we produced spoken dialogues based on analysis results relating to response timing and performed subjective experiments. As a result, we were able to demonstrate the validity of the analysis results relating to dialogue rhythm. However, we have found that to produce dialogue it is insufficient merely to adjust the response timing, and that the acoustic characteristics of the utterances and their balance with the response timing are also important. As for the future, we are considering performing similar subjective experiments by using tools such as voice changers to produce speech where the acoustic characteristics other than the voice quality are left intact, or speech from which some of the acoustic characteristics have been removed, allowing us to analyze the factors necessary for speech to be felt as natural. We also hope to analyze the acoustic characteristics that are necessary for implementing overlaps that are felt to be natural even when using synthesized speech.

## 5. References

- [1] Yamada, S., Itoh, T., and Araki, K., “Linguistic and Acoustic Features Depending on Different Situations - The Experiments Considering Speech Recognition Rate”, Proc. of INTERSPEECH 2005, pp.3393-3396, 2005.
- [2] Yamada, S., Itoh, T. and Araki, K., “Is Voice Quality Enough? - Study on How the Situation and User’s Awareness Influence the Utterance Features”, Proc. of INTERSPEECH 2006, pp.481-484, 2006.
- [3] Kitaoka, N., Takeuchi, M., Nishimura, R., and Nakagawa, S., “Response Timing Detection Using Prosodic and Linguistic Information for Human-friendly Spoken Dialog Systems”, Journal of JSAI, Vol.20, No.3 SP-E, pp.220-228, 2005.
- [4] Shoji, K., Takahashi, M., Ibara, S., Itoh, T., and Araki, K., “Spoken Dialog System considered Rhythm and Synchronized Tendency of Conversation”, (in Japanese) IPSJ SIG Technical Reports, SLP-61, pp.43-48, May, 2006.
- [5] Nagaoka, C., Komori, M, and Nakamura, T., “The interspeaker influence of the switching pauses in dialogue”, (in Japanese) The Japanese Journal of Ergonomics, Vol.38, No.6, pp.316-323, 2002.
- [6] Noriki Fujiwara, Toshihiko Itoh, and Kenji Araki, “Analysis of Changes in Dialogue Rhythm due to Dialogue Acts in Task-oriented Dialogues”, Springer-Verlag Lecture Notes in Artificial Intelligence (LNAI) 4629, pp.564-573, 2007.
- [7] Maekawa, K., “Corpus of Spontaneous Japanese : Its Design and Evaluation”, Proc. ISCA & IEEE Workshop SSPR 2003, pp.7-12, (2003)
- [8] Galatea Project:  
<http://hil.t.u-tokyo.ac.jp/galatea/index.html>