Construction of Back-Channel Utterance Corpus for Responsive Spoken Dialogue System Development

Yuki Kamiya[†], Tomohiro Ohno[‡], Shigeki Matsubara[†♮], Hideki Kashioka[♮]

[†] Graduate School of Information Science, Nagoya University
[‡] Graduate School of International Development, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
[♮] National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
kamiya@el.itc.nagoya-u.ac.jp, {ohno,matubara}@nagoya-u.jp, hideki.kashioka@nict.go.jp

Abstract

In spoken dialogues, if a spoken dialogue system does not respond at all during user's utterances, the user might feel uneasy because the user does not know whether or not the system has recognized the utterances. In particular, back-channel utterances, which the system outputs as voices such as "yeah" and "uh huh" in English have important roles for a driver in in-car speech dialogues because the driver does not look towards a listener while driving. This paper describes construction of a back-channel utterance corpus and its analysis to develop the system which can output back-channel utterances at the proper timing in the responsive in-car speech dialogue. First, we constructed the back-channel utterance corpus by integrating the back-channel utterances that four subjects provided for the driver's utterances in 60 dialogues in the CIAIR in-car speech dialogue corpus. Next, we analyzed the corpus and revealed the relation between back-channel utterance timings and information on bunsetsu, clause, pause and rate of speech. Based on the analysis, we examined the possibility of detecting back-channel utterance timings by machine learning technique. As the result of the experiment, we confirmed that our technique achieved as same detection capability as a human.

1. Introduction

With the recent advances in speech recognition technologies, a considerable number of studies have been conducted on spoken dialogue systems such as car navigation systems. Since these systems are designed to accurately respond to user's requests, their response is done after a user utterance is completely finished. However, in spoken dialogues, if a system does not respond at all during user's utterances, the user might feel uneasy because the user does not know whether or not the system has recognized the utterances.

In human-to-human dialogues, a listener informs a speaker that the listener is listening to speaker's speech not only by replaying after the speaker's utterance but also by performing actions such as back-channel utterances, laughing or nods during the utterance. In particular, back-channel utterances, which the system outputs as voices such as "yeah" and "uh huh" in English, have important roles for a driver in in-car speech dialogues because the driver does not look towards a listener while driving.

This paper describes construction of a back-channel utterance corpus and its analysis to develop the system which can output back-channel utterances at the proper timing during user's utterances in the responsive in-car speech dialogue. Until now, although there are several researches for developing the systems which output back-channel utterances (Fujie et al., 2004; Kopp et al., 2007; Ward and Tsukahara, 2000), they have not comprehensively analyzed the timings at which back-channel utterances can be provided. Our research tries to comprehensively reveal the timings at which back-channel utterances can be provided during speaker's utterances. First, we constructed the back-channel utterance corpus by recording back-channel utterances of multiple subjects and then integrating each timing. Next, we analyzed the corpus and revealed the relation between back-channel utterance timings and information on bunsetsu[1], clause, pause and rate of speech.

The rest of this paper is organized as follows. The next section explains a back-channel utterance. The construction of a back-channel utterance corpus and its analysis are reported in Section 3 and 4, respectively. Section 5 presents examination on detection of back-channel utterance timings. Section 6 concludes the paper.

2. Back-channel utterance

The definition of Japanese back-channel utterance has been discussed by various researchers. Maynard defined a Japanese back-channel utterance as "a short expression which a listener sends while a speaker is speaking" (Maynard, 1989). In addition, Horiguchi defined it as "sign which informs a speaker that a listener is listening to speaker's utterances" (Horiguchi, 1997). We can say that it is important to provide back-channel utterances at the proper timings. Because a speaker might wonder if a listener surely listens to and comprehends his/her speech if back-channel utterances are unnecessarily provided.

The proper timings at which back-channel utterances should be provided differ depending on the participants and the situations. In case of making a spoken dialogue system generate back-channel utterances, it is required to change the timing at which back-channel utterances are outputted depending on the user or situation. In our research, we think of detecting comprehensively the timings at which

[1] *Bunsetsu* is a linguistic unit in Japanese that roughly corresponds to a basic phrase in English. A bunsetsu consists of one independent word and zero or more ancillary words.

subject 1	232
subject 2	217
subject 3	65
subject 4	136

Table 1: Number of back-channel utterances

back-channel utterances can be provided, and then deciding the proper back-channel utterance timing depending on the user's preference and the situation among the detected timings. This can prevent the system from outputting back-channel utterances at the unnatural timings.

3. Construction of back-channel utterance corpus

In order to analyze comprehensively the timings at which back-channel utterances can be provided, we constructed a back-channel utterance corpus by integrating back-channel utterances which multiple subjects provided.

3.1. Recording and transcription of back-channel utterances

We recorded back-channel utterances which four subjects provided for driver's utterances (297 turns in 60 dialogues) in the CIAIR in-car speech dialogue corpus (Kawaguchi et al., 2004). We limited the utterance of the subjects to only "はい *hai* (yeah)," and ordered them to provide as many back-channel utterances as possible at non-unnatural timings. In addition, we played the subjects the sounds of driver's utterances turn-by-turn, and ordered them to provide back-channel utterances during the driver's utterances of each turn. Table 1 shows the number of back-channel utterances by each subject.

We transcribed the driver's utterances and the timings at which each subject provided back-channel utterances based on the recorded speech. Figure 1 shows an example of the transcription. Each line means a morpheme or a pause in driver's utterances. Each line contains information on the start-end time and whether or not a back-channel utterance was provided at the place. In addition, information on the part-of-speech, bunsetsu boundary and clause boundary are described in case of a morpheme. In our research, a morpheme or a pause[2] is thought of as a basic segment for analysis. By analyzing which segment a back-channel utterance was provided at, we reveal the characteristic of back-channel utterance timings. Here, information on morphological analysis, clause boundary and time of each basic segment were provided by ChaSen (Matsumoto et al., 1999), CBAP (Kashioka and Maruyama, 2004) and Julius (Lee et al., 2001), respectively.

3.2. Integration of back-channel utterances by each subject

As shown by Figure 1 and Table 1, the number and timing of back-channel utterances are different from each subject.

[2]If the length of a pause is over 200ms, the pause is divided into two basic segments: the initial 200ms pause and the left pause.

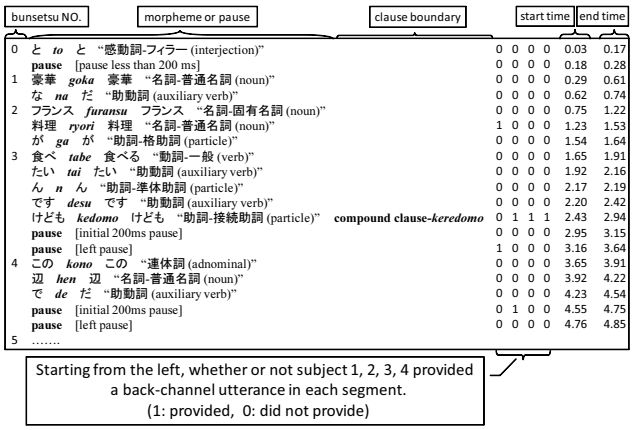

Figure 1: Sample of transcription of driver's utterances and subjects' back-channel utterances

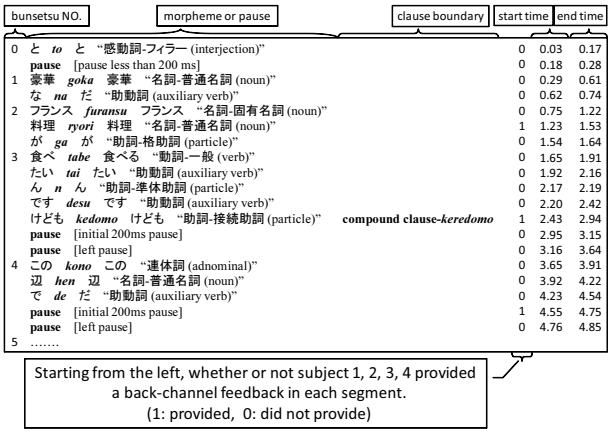

Figure 2: Sample of back-channel utterance corpus constructed by the integration

Our back-channel utterance corpus was constructed by integrating the back-channel utterances which four subjects provided. This integration was performed, by focusing on the timing at which each subject provided a back-channel utterance as voice (hereafter, **output timing**) and the timing at which each subject thought to provide the back-channel utterance in one's head (hereafter, **inspiration timing**), in the following order:

1. In case that the output timings are different but the inspiration timings are thought to be same, those back-channel utterances are eliminated except a back-channel utterance which the most subjects provided at the same output timing.

2. All back-channel utterances which were left after the above procedure are adopted as the meaningful output timing.

Figure 2 shows a sample of the back-channel utterance corpus constructed by integrating back-channel utterances of four subjects which are shown in Figure 1. In this integration, the inspiration timings of back-channel utterances

dialogue	60
turn	297
bunsetsu	1,478
morpheme	3,178
pause	756
back-channel utterance	324

Table 2: Size of back-channel utterance corpus

dialogue	50
turn	250
bunsetsu	1,227
morpheme	2,656
pause	618
back-channel utterance	265

Table 3: Size of analysis data

which were provided at "けども *kedomo* (although) (2.43-2.94 sec)" and "left pause (3.16-3.64 sec)" in Figure 1 are judged to be same and then only the back-channel utterance which was provided at "けども *kedomo* (although) (2.43-2.94sec)" is left and the other is eliminated.

Table 2 shows the size of the back-channel utterance corpus which was constructed by this means.

3.3. Evaluation of back-channel utterance corpus

To evaluate the validity of the back-channel utterance corpus constructed in the previous section, we conducted a subjective experiment. In this experiment, an evaluator listened to the mixed sounds of the driver's speech and the back-channel utterances created by using a speech synthesizer according to the corpus, and then, evaluated whether or not the timing of each back-channel utterance is unnatural. We created the sound of a listener's back-channel utterance so that the back-channel utterance was outputted at 50msec after the start time of the basic segment at which the back-channel utterance was provided. Here, we created the back-channel utterance "はい *hai* (yeah)" by using the speech synthesizer "HitVoice" produced by Hitachi, Ltd.

Among 324 back-channel utterances in the constructed corpus, 34 back-channel utterances, which occupy 10.5% of the total, were judged as unnatural. Considering that the back-channel utterances which the subjects have to hear are all the same, the availability of the constructed back-channel utterance corpus is accepted as a reasonable degree.

4. Characteristic analysis of back-channel utterance timings

In our research, we plan to adopt a statistical approach to detect the timing at which a back-channel utterance can be provided. To decide the available features in a statistical approach, we analyzed the characteristic of back-channel utterance timings by using the corpus constructed in the previous section. We focused on bunsetsu and clause as linguistic information and a pause and rate of speech as phonetical information, and investigated the relations between them and back-channel utterance timings. In the analysis, we used 50 dialogues in the back-channel utterance corpus. Table 3 shows the size of the analysis data. There were 3,274 basic segments, consisting of 2,656 morphemes and 618 pauses, in the 50 dialogues. Among them, a back-channel utterance was provided in 265 basic segments. The rate at which back-channel utterances are provided was 8.1%. Here, this numerical value means the rate at which a back-channel utterance was averagely provided in a basic segment. If the rate at which back-channel utterances are provided in the basic segments having a characteristic is higher than this numerical value 8.1%, it means back-channel utterances are easy to be provided in the basic segments.

particle	25.2%	(102/405)
interjection	8.5%	(21/246)
auxiliary verb	27.5%	(28/102)
noun	24.0%	(18/75)
conjunction	5.5%	(3/55)
adverb	9.5%	(3/42)
adjective	0.0%	(0/17)
verb	8.3%	(1/12)
suffix	9.1%	(1/11)
pronoun	0.0%	(0/10)

Table 4: POS of the final morpheme of a bunsetsu and its rate at which back-channel utterances are provided

4.1. Bunsetsu and back-channel utterance timing

Back-channel utterances are considered to be provided by a listener when the listener understood the utterances of the speaker to some extent because back-channel utterances express the contents understanding. On the other hand, a bunsetsu is the smallest meaningful unit in Japanese. Therefore, a back-channel utterance tends to be provided right after a bunsetsu is uttered.

There were 977 bunsetsus except the final bunsetsu of turns in the analysis data. Among them, a back-channel utterance was provided in the basic segments right after 178 bunsetsus. The rate at which back-channel utterances are provided was 18.2%. The rate at which back-channel utterances are provided in the basic segments right after bunsetsus was higher than the rate in all basic segments (8.1%), and thus, we confirmed that back-channel utterances tend to be provided in the basic segments right after bunsetsus.

Next, we analyzed the difference between the types of a bunsetsu. Table 4 shows the rate at which back-channel utterances are provided in the basic segments right after each type of a bunsetsu, which is classified by the POS of the final morpheme of a bunsetsu. A back-channel utterance is easy to be provided in the basic segments right after a bunsetsu of which the final morpheme is a particle, an auxiliary verb or a noun.

4.2. Clause and back-channel utterance timing

Since a clause is a more meaningful unit than a bunsetsu, a back-channel utterance is considered to tend tobe provided

interjection	5.1%	(6/118)
quotation clause	8.6%	(3/35)
compound clause -*keredomo*	67.6%	(23/34)
conjunction	7.4%	(2/27)
topicalized element -*wa*	12.0%	(3/25)
compound clause -*te*	42.9%	(6/14)
continuous clause	7.1%	(1/14)
indirect question clause	40.0%	(4/10)

Table 5: Type of clauses and its rate at which back-channel utterances are provided

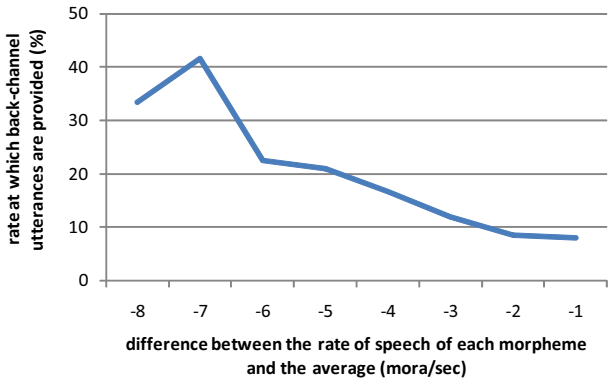

Figure 3: Rate of speech and rate at which back-channel utterances are provided

right after a clause.
There existed 341 clauses except final clauses of turns in the analysis data. Among them, a back-channel utterance was provided in the basic segments right after 66 clauses. The rate at which back-channel utterances are provided in the basic segments right after clauses was 19.4%, and higher than the rate in all basic segments (8.1%) and in the basic segments right after bunsetsus (18.2%). Therefore, we confirmed that back-channel utterances tend more strongly to be provided in the basic segments right after clauses.
Next, we analyzed the difference between the types[3] of clauses. Table 5 shows the rate at which back-channel utterances are provided in the basic segments right after each type of a clause. The rate for the basic segment right after "compound clause -*keredomo*" was high. The ease-of-provision of back-channel utterances was different between the types of clauses.

4.3. Pause and back-channel utterance timing

Since a back-channel utterance has the function to demand utterances from a partner, a back-channel utterance is considered to tend to be provided in a pause.
Among 618 pause segments, a back-channel utterance was provided in 102 pause segments. The rate at which back-channel utterances are provided in the pause segments was 16.5% and higher than the rate in all basic segments (8.1%). We confirmed that a back-channel utterance tends to be provided in the pause segments.

4.4. Rate of speech and back-channel utterance timing

Since a back-channel utterance has the function to demand utterances from a partner, a back-channel utterance is considered to tend to be provided when the rate of speech of the parter becomes slow. We measured the rate of speech (mora/second) of each morpheme in drivers' utterances and classified all morphemes into the two classes: the faster class and the slower class than the average rate of speech of a morpheme. We investigated the rate at which back-channel utterances are provided in the basic segments right after morphemes in each of the two classes. Here, the average rate of speech of a morpheme was calculated for each driver in consideration of the difference between persons. In case of a morpheme of which the rate of speech is faster than the average, the rate at which ack-channel utterances are provided in the basic segments right after the morpheme was 5.7% (99/1,746). In case of a slower morpheme, the rate at which back-channel utterances are provided in the basic segments right after the morpheme was 12.9% (165/1278). From this, we could see that if the rate of speech of a morpheme became slower than the average rate of speech, a back-channel utterance became easy to be provided in the basic segment right after the morpheme.
Furthermore, we classified all morphemes according to the difference between the rate of speech of each morpheme and the average rate of speech on a 1 mora-per-second basis, and investigated the rate at which back-channel utterances are provided in the basic segments right after morphemes in each of the classes. The result is shown in Figure 3. The slower the rate of speech of a morpheme was, the higher the rate at which back-channel utterances are provided became. We could see that a back-channel utterance tended to be provided more strongly when the rate of speech of a morpheme became slower.

5. Examination of detection of back-channel utterance timings

We examined the possibility of automatically detecting back-channel utterance timings by a machine learning technique based on the analysis described in Section 4.

5.1. Technique for detecting back-channel utterance timings

Our research has an assumption that a basic segment in a sequence of basic segments ($s_1 \cdots s_n$) of a turn is sequentially inputted. In our technique, whenever a basic segment is inputted, it is judged whether or not a back-channel utterance can be provided in the input basic segment by using Support Vector Machine (SVM). Table 6 shows the features used by our SVM in judging whether or not a back-channel utterance is provided in a basic segment s_i. These features can be obtained from $s_1 \cdots s_{i-1}$, which is a sequence of basic segments right before the input basic segment s_i. Here, our research assumes that the information of each basic segment such as morphological information, clause information and time information can be obtained after the basic

[3] In our research, we used the types of clause boundaries defined by the Clause Boundary Annotation Program (Kashioka and Maruyama, 2004).

1.	whether or not s_{i-1} is the final morpheme of a bunsetsu
2.	POS of s_{i-1} (if 1. is ture)
3.	whether or not s_{i-1} is the final morpheme of a clause
4.	type of the clause to which s_{i-1} belongs (if 3. is ture)
5.	whether or not s_{i-1} is the initial 200ms pause segment
6.	whether or not the rate of speech of s_{i-1} is slower than the average
7.	difference between the rate of speech of s_{i-1} and the average (if 6. is true)
8.	time length from the start time of the previous back-channel utterance (the start time of s_1 if there were no previous back-channel utterances) to the start time of s_i.

Table 6: Features used by our SVM

segment is inputted.

5.2. Experiment and discussion

We conducted an experiment on detecting back-channel utterance timings and examined the possibility of the detection. Among the corpus constructed in Section 3, we used 50 dialogues which were the analysis data in Section 4 as learning data, and the remaining 10 dialogues as test data. LibSVM(Chang and Lin, 2001) was used with the default option as the SVM tool. For comparison, we collected the data of back-channel utterances by another subject, which was not the subjects of the construction described in Section 3. In the evaluation, we obtained recall and precision which were respectively defined as follows:

$$precision = \frac{\#\ of\ \ correctly\ \ provided\ \ BCUs}{\#\ of\ \ provided\ \ BCUs}$$

$$recall = \frac{\#\ of\ \ correctly\ \ provided\ \ BCUs}{\#\ of\ \ BCUs\ \ in\ the\ \ correct\ \ data}$$

Here, BCUs mean back-channel utterances. If a basic segment in which a back-channel utterance was provided in the detection result matches that in the correct data, we judged that the back-channel utterance has been correctly provided.

Table 7 shows the experimental result. In the result by the subject, the f-measure was 33.8%. This indicates the difficulty of detecting the precisive back-channel utterance timings. The detection performance of our technique was slightly below that of the result by the subject.

On the other hand, in a spoken dialogue system, we think that if the detected timing has only the slight difference from the correct timing, the detected timing can be accepted. Therefore, we reevaluated a provided back-channel utterance as a correctly provided one if the difference between the detected timing and the timing of the correct data was within 200ms. Table 8 shows the reevaluation result. Our technique achieved a comparatively-high precision (70.8%). By expanding the learning data, we think back-channel utterance timings can be detected at the allowable level.

	precision	recall	f-measure
our method	54.2% (13/24)	22.0% (13/59)	31.3%
subject	63.6% (14/22)	23.7% (14/59)	33.8%

Table 7: Experimental result

	precision	recall	f-measure
our method	70.8% (17/24)	28.8% (17/59)	41.0%
subject	72.7% (16/22)	27.1% (16/59)	39.5%

Table 8: Reevaluation result

6. Conclusions

In this paper, first, we constructed the back-channel utterance corpus, which included 324 back-channel utterances, by integrating the back-channel utterances that 4 subjects provided for the utterances of the driver's 297 turns in 60 dialogues. Next, we revealed the characteristic of back-channel utterance timings, by analyzing the constructed corpus. As the result, we found the following things:

- The rate at which back-channel utterances are provided in the basic segments right after bunsetsus was 18.2%. Since this rate was higher than 8.1%, which is the average rate at which back-channel utterances are provided we can recognize the tendency that a back-channel utterance is easier to be provided in a basic segment right after a bunsetsu. In particular, a back-channel utterance is much easier to be provided in the basic segments right after bunsetsus of which the final morpheme is a particle or auxiliary.
- The rate at which back-channel utterances are provided in the basic segment right after a clause was 19.4% and higher than the average rate 8.1%. We found that a back-channel utterance is easier to be provided in the basic segment right after a clause. In particular, a back-channel utterance tends to be provided in the basic segment right after a clause labeled "compound clause-*keredomo*."
- The rate at which back-channel utterances are provided in pause segments was 16.5% and higher than the average rate 8.1%. Therefore, a back-channel utterance tends to be provided in a pause segment.
- When the rate of speech of a morpheme was slower

than the average rate of speech, the rate at which back-channel utterances are provided in the basic segments right after such a morpheme was 12.9%. A back-channel utterance tends to be provided in the basic segment right after a morpheme of which the rate of speech is slower than the average rate of speech.

As the result of the experiment on detecting back-channel utterances based on the above analysis, we confirmed that our technique achieved as same detection capability as a human.

In the future, we plan to analyze the characteristic of back-channel utterances by using the prosodic information, and then, develop the system which can comprehensively detect the back-channel utterance timings.

Acknowledgments

This research was supported in part by the Grant-in-Aid for Challenging Exploratory Research (No.21650028) of JSPS.

7. References

C. C. Chang and C. J. Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

S. Fujie, K. Fukushima, and T. Kobayashi. 2004. A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information. In *Proceedings of 2nd International Conference on Autonomous Robots and Agents (ICARA 2004)*, pages 379–384.

S. Horiguchi. 1997. *Nihongo kyoiku to kaiwa bunseki (Japanese conversation by learners and native speakers)*. Kuroshio, Tokyo. (In Japanese).

H. Kashioka and T. Maruyama. 2004. Segmentation of semantic units in Japanese monologues. In *Proceedings of International Conference on Speech and Language Technology (ICSLT 2004) and the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (Oriental-COCOSDA 2004)*, pages 87–92.

N. Kawaguchi, S. Matsubara, Y. Yamaguchi, K. Takeda, and F. Itakura. 2004. CIAIR in-car speech database. In *Proceedings of 8th International Conference on Spoken Language Processing (ICSLP 2004)*, pages 2789–2792.

S. Kopp, T. Stocksmeier, and D. Gibbon. 2007. Incremental multimodal feedback for conversational agents. *Intelligent Virtual Agents*, 4722:139–146.

A. Lee, T. Kawahara, and K. Shikano. 2001. Julius – an open source real-time large vocabulary recognition engine. In *Proceedings of 7th European Conference on Speech Communication and Technology (EUROSPEECH 2001)*, pages 1691–1694.

Y. Matsumoto, A. Kitauchi, T. Yamashita, and Y. Hirano. 1999. Japanese morphological analysis system ChaSen version 2.0 manual. In *NAIST Technical Report*, NAIST-IS-TR99009.

S. K. Maynard. 1989. *Japanese conversation : self-contextualization through structure and interactional management*. Ablex, Norwood, NJ.

N. Ward and W. Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207.