Coherent Back-Channel Feedback Tagging of In-Car Spoken Dialogue Corpus

Yuki Kamiya
Graduate School of
Information Science,
Nagoya University, Japan
kamiya@el.itc.nagoya-u.ac.jp

Tomohiro Ohno
Graduate School of
International Development,
Nagoya University, Japan
ohno@nagoya-u.jp

Shigeki Matsubara
Graduate School of
Information Science,
Nagoya University, Japan
matubara@nagoya-u.jp

Abstract

This paper describes the design of a back-channel feedback corpus and its evaluation, aiming at realizing in-car spoken dialogue systems with high responsiveness. We constructed our corpus by annotating the existing in-car spoken dialogue data with back-channel feedback timing information in an off-line environment. Our corpus can be practically used in developing dialogue systems which can provide verbal back-channel feedbacks. As the results of our evaluation, we confirmed that our proposed design enabled the construction of back-channel feedback corpora with high coherency and naturalness.

1 Introduction

In-car spoken dialogue processing is one of the most prevailing applications of speech technology. Until now, to realize the system which can surely achieve such tasks navigation and information retrieval, the development of speech recognition, speech understanding, dialogue control and so on has been promoted. Now, it becomes important to increase responsiveness of the system not only for the efficient achievement of the task but for increasing drivers' comfortableness in a dialogue.

One way to increase responsiveness of a system is to timely disclose system's state of understanding, by making the system show some kind of reaction during user's utterances. In human dialogues, such disclosure is performed by actions such as nods, facial expressions, gestures and back-channel feedbacks. However, since drivers do not look towards a spoken dialogue system while driving, the system has to inevitably use voice responses, that is, back-channel feedbacks. Furthermore, in the response strategy for realizing in-car dialogues in which drivers feel comfortable, it is necessary for the system to provide back-channel feedbacks during driver's utterances aggressively as well as timely.

This paper describes the design of a back-channel feedback corpus having coherency (tagging is performed by different annotators equally) and naturalness, and its evaluation, aiming at realizing in-car spoken dialogue systems with high responsiveness. Although there have been several researches on back-channel feedback timings (Cathcart et al., 2003; Maynard, 1989; Takeuchi et al., 2004; Ward and Tsukahara, 2000), in many of them, back-channel feedback timings in human dialogues were observed and analyzed by using a general spoken dialogue corpus. On the other hand, we constructed our corpus by annotating the existing in-car spoken dialogue data with back-channel feedback timing information in an off-line environment. Our corpus can be practically used in developing dialogue systems which can provide back-channel feedbacks.

In our research, the driver utterances (11,181 turns) in the CIAIR in-car spoken dialogue corpus (Kawaguchi et al., 2005) were used as the existing data. We created the Web interface for the annotation of back-channel feedbacks and constructed the corpus including 5,416 back-channel feedbacks. Experiments have shown that our proposed corpus design enabled the construction of back-channel feedback corpora with high coherency and naturalness.

2 Corpus Design

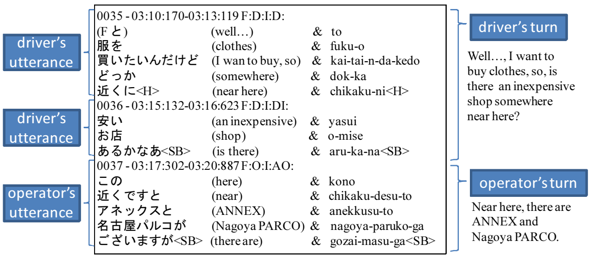
A back-channel feedback is a sign to inform a speaker that the listener received the speaker's utterances. Thus, in an in-car dialogue between a driver and a system, it is preferable that the system provides as many back-channel feedbacks as possible. However, if back-channel feedbacks are unnecessarily provided, they can not play the primary role because the driver wonders if the system really comprehends the speech.

For this reason, the timings at which the system provides back-channel feedbacks become important. Several researches investigated back-channel feedback timings in human-human dialogues (Cathcart et al., 2003; Maynard, 1989; Takeuchi et al., 2004; Ward and Tsukahara, 2000). They reported back-channel feedbacks had the following tendencies: “within or after a pause,” “after a conjunction or sentence-final particle,” and “after a clause wherein the final pitch descends.”

However, it is difficult to systematize the appropriate timings of back-channel feedbacks since their detection is intertwined in a complex way with various acoustic and linguistic factors. Although machine learning using large-scale data would be a solution to the problem, existing spoken dialogue corpora are not suitable for direct use as data, because the timings of the back-channel feedbacks lack coherency due to the influence of factors such as the psychological state of a speaker, the environment and so on.

In our research, to create more pragmatic data in which the above-mentioned problem is solved, we constructed the back-channel feedback corpus with coherency. To this end, we established the following policies for annotation:

- **Comprehensive tagging:** Back-channel feedback tags are provided for all timings which are not unnatural. In human-human dialogues, there are some cases that even if a timing is suited for providing a back-channel feedback, no back-channel feedback is not provided (Ward and Tsukahara, 2000). On the other hand, in our corpus, comprehensive tagging enables coherent tagging.
- **Off-line tagging:** Annotators tag all timings at which back-channel feedbacks can be provided after listening to the target speech one or more times. Compared with providing back-channel feedbacks in on-line environment, the off-line annotation decreases the chances of tagging wrong positions or failing in tagging back-channel feedbacks, realizing coherent tagging.
- **Discretization of tagging points:** Tagging is performed for each segment into which driver’s utterances are divided. In a normal dialogue, the listener can provide back-channel feedbacks whenever he/she wants to, but the inconsistency in the timings to give such feedbacks becomes larger in exchange for smaller restrictions. The discretization of tagging points enables not only coherent tagging but also the reduction of tagging cost.


driver’s utterance
driver’s utterance
operator’s utterance
driver’s turn
Well…, I want to buy clothes, so, is there an inexpensive shop somewhere near here?
operator’s turn
Near here, there are ANNEX and Nagoya PARCO.

```
0035 - 03:10:170-03:13:119 F:D:I:D:
(F と)          (well…)           &  to
服を            (clothes)         &  fuku-o
買いたいんだけど  (I wan to buy, so)  &  kai-tai-n-da-kedo
どっか          (somewhere)       &  dok-ka
近くに<H>        (near here)       &  chikaku-ni<H>
0036 - 03:15:132-03:16:623 F:D:I:DI:
安い            (an inexpensive)  &  yasui
お店            (shop)            &  o-mise
あるかなあ<SB>    (is there)        &  aru-ka-na<SB>
0037 - 03:17:302-03:20:887 F:O:I:AO:
この            (here)            &  kono
近くですと        (near)            &  chikaku-desu-to
アネックスと      (ANNEX)           &  anekkusu-to
名古屋パルコが    (Nagoya PARCO)    &  nagoya-paruko-ga
ございますが<SB>  (there are)       &  gozai-masu-ga<SB>
```

Figure 1: Sample of transcribed text

- **Elaboration using synthesized sound:** An annotator checks the validity of the annotation by listening to the sounds. In other words, an annotator elaborates the annotation by revising it many times by listening to the automatically created dialogue sound which includes not only driver’s voices but also sounds of back-channel feedbacks generated according to the provided timings. The back-channel feedbacks had been synthesized by using a speech synthesizer because our corpus aims to be used for implementing the system which can provide back-channel feedbacks.

3 Corpus Construction

We constructed the back-channel feedback corpus by annotating an in-car speech dialogue corpus.

3.1 CIAIR in-car spoken dialogue corpus

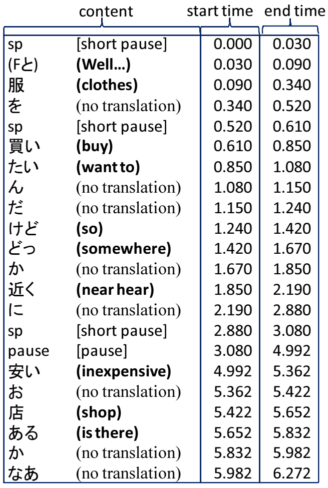
We used the CIAIR in-car spoken dialogue corpus (Kawaguchi et al., 2005) as the target of annotation. The corpus consists of the speech and transcription data of dialogues between a driver and an operator about shopping guides, driving directions, and so on. Figure 1 shows an example of the transcription. We used only the utterances of drivers in the corpus. We divided the utterances into morphemes by using the morphological analyzer Chasen[1]. In addition, each morpheme was provided start and end times estimated by using the continuous speech recognition system Julius[2].

3.2 Tagging of spoken dialogue corpus

We constructed the corpus by providing the back-channel feedback tags at the proper timings for the driver’s utterances, according to the design described in Section 2.

[1] http://chasen-legacy.sourceforge.jp

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



content		start time	end time
sp	[short pause]	0.000	0.030
(Fと)	**(Well…)**	0.030	0.090
服	**(clothes)**	0.090	0.340
を	(no translation)	0.340	0.520
sp	[short pause]	0.520	0.610
買い	**(buy)**	0.610	0.850
たい	**(want to)**	0.850	1.080
ん	(no translation)	1.080	1.150
だ	(no translation)	1.150	1.240
けど	**(so)**	1.240	1.420
どっ	**(somewhere)**	1.420	1.670
か	(no translation)	1.670	1.850
近く	**(near hear)**	1.850	2.190
に	(no translation)	2.190	2.880
sp	[short pause]	2.880	3.080
pause	[pause]	3.080	4.992
安い	**(inexpensive)**	4.992	5.362
お	(no translation)	5.362	5.422
店	**(shop)**	5.422	5.652
ある	**(is there)**	5.652	5.832
か	(no translation)	5.832	5.982
なあ	(no translation)	5.982	6.272

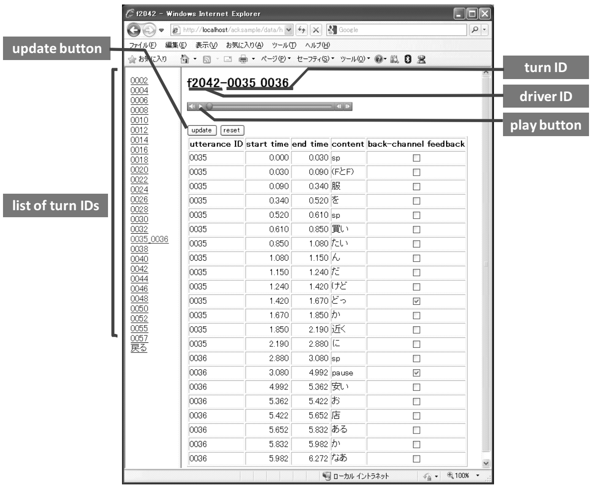
Figure 2: Sample of division of a dialogue turn into basic segments

For "comprehensive tagging," an annotator listens to each dialogue turn[3] from the start and tags a position where a back-channel feedback can be provided when the timing is found. Here, the timing of the last back-channel feedback is also used for judging whether or not the timing is unnatural.

For "off-line tagging," an annotator tags the transcribed text of each dialogue turn of drivers.

To perform "discretization of tagging points," a dialogue turn is assumed to be a sequence of morphemes or pauses (hereafter, we call them **basic segments**), which are continuously arranged on the time axis, and it is judged whether or not a back-channel feedback should be provided at each basic segment. Here, in consideration of the unequal pause durations, if the length of a pause is over 200ms, the pause is divided into the initial 200ms pause and the subsequent pause, each of which is considered as a basic segment. Figure 2 shows an example of a dialogue turn divided into basic segments.

Furthermore, for "elaboration using synthesized sound," we prepared the annotation environment where the dialogue sound including not only driver's voice but also back-channel feedbacks generated according to the provided timings is automatically created in real time for annotators to listen to. There are several types of back-channel feedbacks and in normal conversations, we choose and use appropriate back-channel feedbacks from among them according to the scene. In our study, we used the most general form "はい *hai* (yes)" for the synthesized speech since our focus was on the timing of back-channel feedbacks. The back-channel feedbacks had been created by using Hitachi's speech synthesizer "HitVoice," and one feedback was placed 50 milli-seconds after the start time of a tagged basic segment.


f2042 – Windows Internet Explorer
http://localhost/acksample/data/h
Google
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
お気に入り
ページ(P) セーフティ(S) ツール(O)
update button
turn ID
driver ID
play button
list of turn IDs
ローカル イントラネット
100%

f2042-0035 0036

0002
0004
0006
0008
0010
0012
0014
0016
0018
0020
0022
0024
0026
0028
0030
0032
0035_0036
0038
0040
0042
0044
0046
0048
0050
0052
0055
0057
戻る

update reset

utterance ID	start time	end time	content	back-channel feedback
0035	0.000	0.030	sp	□
0035	0.030	0.090	(FとF)	□
0035	0.090	0.340	服	□
0035	0.340	0.520	を	□
0035	0.520	0.610	sp	□
0035	0.610	0.850	買い	□
0035	0.850	1.080	たい	□
0035	1.080	1.150	ん	□
0035	1.150	1.240	だ	□
0035	1.240	1.420	けど	□
0035	1.420	1.670	どっ	☑
0035	1.670	1.850	か	□
0035	1.850	2.190	近く	□
0035	2.190	2.880	に	□
0036	2.880	3.080	sp	□
0036	3.080	4.992	pause	☑
0036	4.992	5.362	安い	□
0036	5.362	5.422	お	□
0036	5.422	5.652	店	□
0036	5.652	5.832	ある	□
0036	5.832	5.982	か	□
0036	5.982	6.272	なあ	□

Figure 3: Web interface for tagging

Table 1: Size of back-channel feedback corpus

drivers	346
dialogue turns	11,181
clauses	16,896
bunsetsus[4]	12,689
morpheme segments	94,030
pause segments	19,142
back-channel feedbacks	5,416

We developed a Web interface for tagging back-channel feedbacks. Figure 3 shows the Web interface. The interface displays a sequence of basic segments in a dialogue turn in table format. Annotators perform tagging by checking basic segments where a back-channel feedback can be provided.

3.3 Size of back-channel feedback corpus

Table 1 shows the size of our corpus constructed by two trained annotators. The corpus includes 5,416 back-channel feedbacks. This means that a back-channel feedback is generated at intervals of about 21 basic segments.

4 Corpus Evaluation

We conducted experiments for evaluating the tagging in the constructed corpus.

[3] A dialogue turn is defined as the interval between the time at which the driver starts to utter just after the operator finishes uttering and the time at which the driver finishes uttering just before the operator starts to utter.

[4] *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.

Table 2: Kappa values of the existing corpus

	a,c	a,d	a,b	c,d	b,c	b,d
κ	0.536	0.438	0.322	0.311	0.310	0.167

4.1 Coherency of corpus tagging

We conducted an evaluation experiment to confirm that the tagging is coherently performed in the corpus. In the experiment, two different annotators performed tagging on the same data, and then we measured the degree of the agreement between them. As the indicator, we used Cohen's kappa value (Cohen, 1960), calculated as follows:

$$\kappa = \frac{P(O) - P(E)}{1 - P(E)}$$

where $P(O)$ is the observed agreement between annotators, and $P(E)$ is the hypothetical probability of chance agreement. A subject who has a certain level of knowledge annotated 673 dialogue turns. The kappa value was 0.731 ($P(O) = 0.975, P(E) = 0.907$), and thus we can see the substantial agreement between annotators.

As the target for comparison, we used the kappa value in the existing back-channel feedback corpus (Kamiya et al., 2010). The corpus had been constructed by the way that the recorded driver's voice was replayed and 4 subjects independently produced back-channel feedbacks for the same sound. This means that the policies for tagging the existing corpus differ from those of our corpus, and are "on-line tagging," "tagging on the time axis" and "tagging without elaborating." In the exisiting corpus, 297 dialogue turns were used as driver's sound. Table 2 shows the kappa value between two among the 4 subjects. The kappa value of our corpus was higher than that between any subjects of the existing corpus, substantiating the high coherency of our corpus.

4.2 Validity of corpus tagging

In our corpus, we discretized the tagging points to enhance the coherency of tagging. However, such constraint restricts the points available for tagging and may make annotators provide tags at the unnatural timings. Therefore, we conducted a subjective experiment to evaluate the naturalness of the back-channel feedback timings. In the experiment, one subject listened to the replay of our back-channel feedback corpus and subjectively judged the naturalness of each timing. The back-channel feedback sound was generated in the same way described in Section 3.2.

In the experiment, we used 345 dialogue turns including 131 back-channel feedbacks. 98.47% of all the back-channel feedbacks were judged to be natural. Only 2 back-channel feedbacks were judged to be unnatural because the intervals between them and the back-channel feedbacks provided immediately before them were felt too short. This showed the validity of our discretization of tagging points.

5 Conclusion

This paper described the design, construction and evaluation of the back-channel feedback corpus which had the coherency of tagged back-channel feedback timings. We constructed the spoken dialogue corpus including 5,416 back-channel feedbacks in 11,181 dialogue turns. The results of our evaluation confirmed high coherency and enough naturalness of our corpus.

In the future, we will use our corpus to see to what extent the timings of back-channel feedbacks that have been annotated correlate with the cues provided by earlier researchers. Then we will develop a system which can detect back-channel feedback timings comprehensively.

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

References

N. Cathcart, J. Carletta, and E. Klein. 2003. A shallow model of backchannel continuers in spoken dialogue. In *Proc. of 10th EACL*, pages 51–58.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Y. Kamiya, T. Ohno, S. Matsubara, and H. Kashioka. 2010. Construction of back-channel utterance corpus for responsive spoken dialogue system development. In *Proc. of 7th LREC*.

N. Kawaguchi, S. Matsubara, K. Takeda, and F. Itakura. 2005. CIAIR in-car speech corpus – influence of driving status–. *IEICE Trans. on Info. and Sys.*, E88-D(3):578–582.

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

M. Takeuchi, N. Kitaoka, and S. Nakagawa. 2004. Timing detection for realtime dialog systems using prosodic and linguistic information. In *Proc. of Speech Prosody 2004*, pages 529–532.

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