

A Spoken Dialog System for Chat-like Conversations Considering Response Timing

Ryota NISHIMURA¹, Norihide KITAOKA², and Seiichi NAKAGAWA¹

¹ Department of Information and Computer Sciences, Toyohashi University of Technology, Japan

{nishimura, nakagawa}@slp.ics.tut.ac.jp

² Graduate School of Information Science, Nagoya University, Japan

kitaoka@sp.m.is.nagoya-u.ac.jp

Abstract. If a dialog system can respond to a user as naturally as a human, the interaction will be smoother. In this research, we aim to develop a dialog system by emulating the human behavior in a chat-like dialog. In this paper, we developed a dialog system which could generate chat-like responses and their timing using a decision tree. The system could perform “collaborative completion,” “*aizuchi*” (back-channel) and so on. The decision tree utilized the pitch and the power contours of user’s utterance, recognition hypotheses, and response preparation status of the response generator, at every time segment as features to generate response timing.

1 Introduction

Recently, interfaces using automatic speech recognition (ASR) have been developed. In traditional systems, however, there was no reaction to the user during a user utterance, so, the user could not know whether or not the system was hearing the utterance. Therefore, a spoken dialog system gave a *stiff* impression.

In Japanese human-human dialog, well-timed responses such as ‘*aizuchi*’ (sometimes called ‘back-channel’) and turn-taking make for smooth dialog.

The purpose of this study is to generate a natural response including *aizuchi*, collaborative completions, and turn taking considering response timing. To generate the response timing, we use a decision tree with features related to prosodic information and surface linguistic information. Using this timing generation method, we have been developing a human-friendly spoken dialog system [1]. One of our system’s goals is to become very familiar to users so that humans will chat with it.

2 Previous literature on for chat-like conversation

The properties of *aizuchi* and turn-taking has been studied [2–5]. The results indicate that pitch (F0) and power are mainly related to generating *aizuchi* and turn-taking. Some real-time *aizuchi* generation systems developed so far [6–8] use pitch (i.e., inverse of fundamental frequency (F0)) and pause duration as

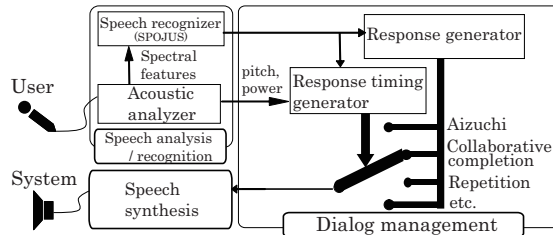


Fig. 1. Schematic diagram of dialog system.

features. Some natural turn-taking timing detection systems have also been developed [9–11]. Fujie et al. [12], for example, used prosody information, especially F0 and power of the utterance, in order to determine the appropriate timing of the feedback. They use a finite state transducer-based speech recognizer to determine the sentence of the feedback earlier than the end of the utterance. These previous studies dealt with an individual kind of response.

In this paper, we propose a unified approach for generation of various kinds of responses including *aizuchi*, and collaborative completions, considering their timing, mainly based on prosodic information. We previously proposed a method to generate *aizuchi* and turn-taking timing [1], but the approach needed pause detection and thus could not deal with overlapping responses. The new system proposed here is not pause detection-driven, but analyzes user utterances continuously even while the user is speaking. This will enable the system to deal with not only the overlapping *aizuchi* and turn-taking, but also the other responses such as collaborative completion.

3 Dialog system

To make spoken dialog systems comply with the above phenomena, we designed the novel system architecture shown in Figure 1. In this section, we introduce an overview of a developed system in the weather information domain.

3.1 Speech analysis and recognition

The speech recognizer SPOJUS [13] recognizes a user’s input. SPOJUS outputs intermediate hypotheses in real-time. We used a vocabulary of 300 words including city names, dates, types of weather, fillers etc., with word class information. Simultaneously, the system analyzes the input to extract prosodic information, such as pitch (F0) and power, using a prosodic analyzer [12, 14].

3.2 Response generator

The response generator prepares response sentences using an ELIZA-like method [15] with slot-based history management in addition to recognition hypotheses. Thus, the response generator also serves as a simple dialog manager. A template set of responses is prepared for each dialog act; *aizuchi*, collaborative completions, repetition and other ordinary responses. These templates are used in parallel, so multiple patterns of responses are generated simultaneously.

Our current system deals with ‘*aizuchi*’, ‘collaborative completions’, ‘repetition’ and other ordinary responses. Thus, four patterns of response sentences are prepared in parallel. Even while the user is speaking, the response generator continuously updates the responses using the intermediate hypotheses generated by the speech recognizer. Default sentences are also prepared and randomly selected to respond even if no appropriate sentences are prepared by the templates. During a dialog, not only the keywords included in the user utterances but also the current status of the weather extracted from a web site (<http://www.imocwx.com/>) are kept in the slots and used for response generation.

3.3 Response timing generator

The response generator only constructs response sentences. To output the sentence, the response timing generator selects an appropriate sentence with the appropriate timing. This timing generator makes a decision to respond or not and which response the system should make using a decision tree. Details are presented in Section 4.

3.4 Speech synthesizer

To output responses by speech, we use the recorded human voice or speech synthesizer voice. GalateaTalk [16] is used for the speech synthesizer, which can control speaker type, voice tone, speech rate, etc.

4 Response timing generation

4.1 Features for timing generation

According to [2] and [5], the contour patterns of pitch and power are related to the timing generation. For example, when pitch and/or power contours of the mora at the end of an utterance follow some proper patterns, the conversational partner’s *aizuchi* or turn-taking is triggered. Thus, the first-order regression coefficients of pitch and power sequences in the last three regions of utterances obtained using 55-ms length sliding window with 30-ms overlap (total length is 105 ms) as shown in Figure 2 are used. The longer region also includes the information that triggers responses, so pitch/power contours in the last 500 ms are also used. To describe such patterns, we adopted first-order regression coefficients for 100 ms-length segments with no overlaps. The coefficients of the five continuous segments express the pattern. Such coefficients can be calculated with very small computational cost, and thus the calculation can be done in real time.

‘Repetition’ and ‘collaborative completion’ occur when a keyword of the conversation topic is input by the user [17]. When the speaker is afraid the hearer cannot catch up with him/her (imagine that the speaker tells the hearer a telephone number), the speaker divides an utterance into some ‘fragments’. In such

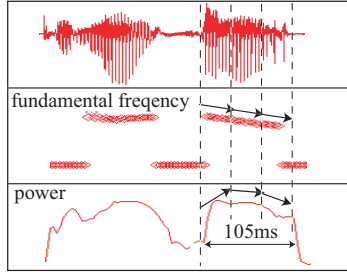


Fig. 2. Regression coefficients of fundamental frequency and power at the end of an utterance.

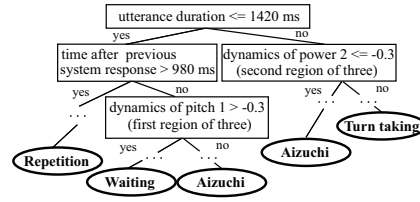


Fig. 3. Part of the decision tree.

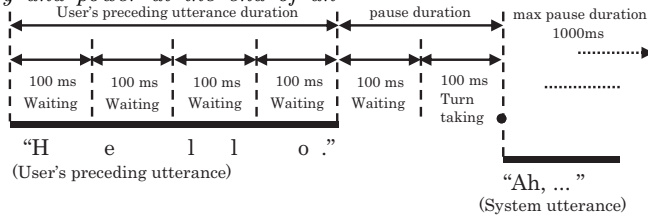


Fig. 4. Response timing generated by the decision tree

cases, the hearer often uses ‘repetitions’ of the fragments to indicate his/her understanding. To imitate this behavior, the timing generator should detect keywords in user utterances. In the recognition results (or intermediate hypotheses), the attribute of the word is attached to each word, and this information is useful to detect keywords. The task of our system is weather information, so attributes of the keywords include a place-name, date, weather in a topical place, etc. We used the attribute of the last word of hypotheses (or intermediate hypotheses) as a feature.

The following features are used in consideration of the above.

- Duration from the start time of the user’s preceding utterance
- Elapsed time from the end of the previous user utterance
- Elapsed time from the end of the previous system utterance
- Pitch/power contour of the last 100 ms (consisting of three values)
- Pitch/power contour of the last 500 ms (consisting of five values)
- Attribute of the last word of the last recognition results (or current intermediate hypotheses)

4.2 Response timing generation using decision tree

Previously, we proposed a decision tree-based timing generator [1], but it can only treat the response at after the end of user utterances. We modified this method to enable it to generate overlapping responses by scanning every segment whenever the user speaks. A part of the decision tree is shown in Figure 3.

The response timing generator decides response timing as well as the selection of a response sentence from responses prepared by the response generator, using



Fig. 5. Example of a dialog between the system and a user.

a decision tree based on the features introduced in Section 4.1. The information on whether or not the response contents have been prepared by the response generator is also used as a feature. Features are input to the decision tree every 100 ms. The decision tree selects a dialog act, which the system should do at that moment, from *aizuchi*, collaborative completion, repetition, ordinary response, and *wait*, as illustrated in Figure 4. “*Wait*” means not to output any response. The frequency of the responses except *aizuchi* and repetition is limited to one for one user utterance.

The RWC corpus [18] is used to train the decision tree for *aizuchi*, turn-taking and *wait*. RWC has 48 conversations of about 10-minute durations each for a total of 6.5 hours. It consists of 16,399 utterances. The conversation tasks are ‘car sales’ and ‘overseas trip planning’. The speaker on one side is a professional salesperson, and the questioner / customer on the other side is one of 12 men and women. C4.5 is used for machine learning.

As for the other phenomena; repetitions and collaborative completions, there were not enough training data in the corpus. So we added some rules manually referring to [2, 5, 8]. For example, ‘repetition’ occurs when two seconds or more have elapsed from the latest response of the system, and when the last word in the recognition hypothesis is a city name.

The system has some exceptional rules to continue the dialog; for example, the system prompts the user to say something after a long pause (6 seconds in our system). And, when a pause of over 1000 ms occurs after the last user utterance, the system responds to the user without depending on the tree.

5 Example of dialog with the system

An example of dialog with the system is shown in Figure 5. The top shows the user utterances, and the bottom the system responses.

In Figure 5, first the system prompted a start-up utterance. Then, the user said “Hello” and the system also said “Aha, Hello.” Next, the system said today’s weather to lead the user to the topic of weather. The system obtained the place where the user was (default value) and the weather around there, and kept the information in slots. With the next user utterance “Recently, it often rains, doesn’t it?”, the system’s **collaborative completion** “rains, doesn’t it.” was **overlapped**. The system detected the keywords/key phrases “saikin (recently)” and “ame (rain)”, and knew that it had been raining. So, the system predicted

that the user would say phrases that meant “it *often* rains” and tried to synchronize to the user with “ooi (many)”. The system has some response templates for collaborative completion and activates one of them if the user utterance and the current slot information meet a certain condition written as a decision rule. With the next user utterance “How about the weather in Hamamatsu?”, the system detected a keyword “Hamamatsu (city name)” and responded immediately by the way of **repetition**. Then the system replied regarding the weather in Hamamatsu; “It always rains.” This dialog contained some chat-like dialog-specific phenomena such as *aizuchi*, repetition, and collaborative completion. Such phenomena often occur in human-human dialogs when the dialogs warm up.

As shown above, our proposed system could work with many kinds of phenomena appearing in natural human-human spoken dialog including overlapping utterances when given appropriate rules, templates and parameters.

6 Experiments and results

6.1 Evaluation of timing generation

We subjectively evaluated the naturalness of the timing generated by the generator. Here, only *aizuchi* and turn-taking are evaluated, because phenomena with few occurrences such as repetition and collaborative completion have not been sufficiently investigated so far.

To evaluate the timing of the generator, we prepared samples of *aizuchi* and turn-taking whose timing is generated by the decision tree.

We inserted an *aizuchi* extracted from a side of a dialog at the *aizuchi* timing point generated by our timing generator. We also made a samples of turn-taking. Thus, we picked some filled pauses such as “Ettodesune” (“Well ... let’s see” in English), which is often employed at the beginning of an utterance, to insert at the time of system formation. Subjects listened to the inserted *aizuchi* with one preceding sentence and evaluated only the timing.

We compared the timing by the generator to that in the corpus. In real dialogs of the corpus, responses may have some meaning consistent with the dialog context and the meaning may make subjects feel natural, especially in the case of turn-taking. To make subjects evaluate only the timing, we also replace the *aizuchi* or filler pauses of the real response with *aizuchi* or a filled pause extracted from other parts of the dialog, as in the case of the generator. The number of samples is 20 for each phenomena. The five subjects heard these sample voices and answered questionnaires (1: too early; 2: early; 3: good; 4: late; 5: too late; and 0: outlier).

The results are shown in Table 1. The “naturalness” in the table indicates the rate of “good.” In the table, the naturalness of the decision tree timing is comparable to the naturalness of the corpus (that is, human-human dialog) timing.

Table 1. Evaluation of response timing by subjective evaluation

	Timing	too early	early	good	late	too late	outlier	naturalness(%)
aizuchi	decision tree	0	6	61	20	11	2	61.0
	corpus	14	26	58	2	0	0	58.0
turn-taking	decision tree	9	26	53	9	2	1	53.0
	corpus	7	31	51	10	0	1	51.0

6.2 Evaluation of dialog system

The subjects used and evaluated the dialog system with the timing generator. There are four kinds of systems: combinations of using / not using overlap response and using recorded / synthesized voice. After using them, the subjects answered a questionnaire. By comparing recorded voices with synthesized ones, we reveal whether or not there is a difference in the evaluation of timing according to a difference in voice quality. We required to subjects so that subjects focused on the evaluation of “timing” and “overlapping”.

According to the results of the questionnaire, two of five subjects prefer the system using *overlap* (including “bargе-in”). One said that he could confirm that the system listened to his utterance. As for familiarity with the system, four of five subjects felt the system using *overlap* was ‘very good’ or ‘good’ on the basis of a five-grade evaluation from ‘very good’ to ‘very bad’. However, an irrelevant response caused by immature speech recognition and dialog management made subjects feel uncomfortable. When listening to the real responses in the corpus, the subjects felt good for 74% and 80% of *aizuchi* and turn-taking, respectively. Compared with Table 1, this reveals that the contents affect the naturalness of timing. As for speech quality, four of five subjects prefer the recorded human voice. The evaluation of the recording voice is better, even though the timing generator has the same performance. This means that the difference in the naturalness of voice quality influences the evaluation of timing. In fact, the evaluation of timing only is difficult, and the voice quality is also unconsciously related to the evaluation.

7 Conclusions

In this paper, we developed a dialog system utilizing real-time response generation and response timing generation, to perform a chat-like friendly conversation. The naturalness of the decision tree-based timing generator was comparable to humans, and the behavior of the dialog system gives a user-friendly impression.

In the future, we will train the decision tree using the dialogs between human and the system. We will also adopt prosodic synchrony to make the system response more natural.

Acknowledgments

The prosodic analyzer used in our system was designed by Dr. Masataka Goto at the National Institute of Advanced Industrial Science and Technology (AIST), and implemented by Dr. Shinya Fujie at Waseda University.

References

1. Takeuchi, M., Kitaoka, N., Nakagawa, S.: Timing detection for realtime dialog systems using prosodic and linguistic information. *Speech Prosody 2004* (2004) 529–532
2. Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., , Den, Y.: An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech* **41**(3-4) (1998) 291–317
3. Geluykens, R., Swerts, M.: Prosodic cues to discourse boundaries in experimental dialogues. *Speech Communication*, 15 (1994) 69–77
4. Hirschberg, J.: Communication and prosody: functional aspects of prosody. *Speech Communication*, 36 (2002) 31–43
5. Ohsuga, T., Nishida, M., Horiuchi, Y., Ichikawa, A.: Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue. *Proceedings of Eurospeech2005* (2005) 33–36
6. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32 (2000) 1177–1207
7. Okato, Y., Kato, K., Yamamoto, M., Itahashi, S.: Insertion of interjectory response based on prosodic information. *IEEE Workshop Interactive Voice Technology for Telecommunication Applications (IVTTA-96)* (1996) 85–88
8. Noguchi, H., Den, Y.: Prosody-based detection of the context of backchannel responses. *Proceedings of ICSLP-98* (1998) 487–490
9. Sato, R., Higashinaka, R., Tamoto, M., Nakano, M., Aikawa, K.: Learning decision tree to determine turn-taking by spoken dialogue systems. *ICSLP-02* (2002) 861–864
10. Hirasawa, J., Nakano, M., Kawabata, T., Aikawa, K.: Effects of system barge-in responses on user impressions. *EUROSPEECH-99* **3** (1999) 1391–1394
11. Kamm, C., Narayanan, S., Dutton, D., Ritenour, R.: Evaluating spoken dialogue systems for telecommunication services. *Eurospeech-97* (1997) 2203–2206
12. Fujie, S., Fukushima, K., Kobayashi, T.: Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. *Interspeech-05* (2005) 889–892
13. Kai, A., Nakagawa, S.: A frame-synchronous continuous speech recognition algorithm using a top-down parsing of context-free grammar. 257-260 (1992)
14. Goto, M., Itou, K., Hayamizu, S.: A real-time filled pause detection system for spontaneous speech recognition. *Eurospeech-99* (1999) 227–230
15. Weizenbaum, J.: ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9 (1) (1965) 36–45
16. Kawamoto, S., Shimodaira, H., Nitta, T., Nishimoto, T., Nakamura, S., Itou, K., Morishima, S., Yotsukura, T., Kai, A., Lee, A., Yamashita, Y., Kobayashi, T., Tokuda, K., Hirose, K., Minematsu, N., Yamada, A., Den, Y., Utsuro, T., Sagayama, S.: Open-source software for developing anthropomorphic spoken dialog agent. *Proc. of PRICAI-02, International Workshop on Lifelike Animated Agents* (2002) 64–69
17. Ishizaki, M., Den, Y.: Danwa to taiwa. Tokyo Daigaku Shuppankai, (in Japanese) (2001)
18. Tanaka, K., Hayamizu, S., Yamasita, Y., Shikano, K., Itahashi, S., Oka, R.: Design and data collection for a spoken dialogue database in the real world computing program. *Proc. ASA-ASJ Third Joint Meeting* (1996) 1027–1030