Generation of Natural Response Timing Using Decision Tree Based on Prosodic and Linguistic Information

Masashi Takeuchi, Norihide Kitaoka and Seiichi Nakagawa

Toyohashi University of Technology, Toyohashi, 441-8580, Japan
{takeuchi,kitaoka,nakagawa}@slp.ics.tut.ac.jp

Abstract

If a dialog system can respond to the user as reasonable as a human, the interaction will be more smooth. Timing of response such as backchannels and turn-taking plays important role in such a smooth dialog as in human-human interaction. We are now developing a dialog system which can generate response timing in real time. In this paper, we introduce a response timing generator for such a dialog system. First, we analyzed conversations between two persons and extracted prosodic and linguistic information which had effects on the timing. Then we constructed a decision tree based on the features coming from the information and developed a timing generator using rules derived from the decision tree. The timing generator decides the action of the system at every 100ms in user's pause. We evaluated the timing generator by subjective and objective evaluation.

1. Introduction

In Japanese human-human dialog, well-timed responses such as '*aizuchi*' (sometimes called as '*backchannel*') and turn taking make the dialog smooth. The purpose of this study is to generate natural response timing of aizuchi and turn taking. We are developing a human-friendly spoken dialog system which can generate natural response timing during a dialog.

In this paper, we propose a method to generate the timing with decision rules derived from a decision tree based on prosodic and linguistic features. First, we investigated timing of aizuchi and turn taking in human-human dialog. From the analysis, we found that some prosodic and surface linguistic information should affect the system behavior. Then, we constructed a decision tree which selected the action of the system among waiting, making aizuchi and taking the turn at every 100ms in the user's pause. The decision tree used features derived from the information which was found to be effective in the above analysis.

The rest of the paper is organized as follows: related works are introduced in Section 2. Section 3 describes target system overview. Human-human dialogs were analized in Section 4. We introduced the method to generate response timing based on a decision tree in Section 5. Section 6 contains the experimental results and Section 7 concludes our discussions.

2. Related works

Some real time aizuchi generation systems have been developed so far. Ward[1] pointed out that low pitch region longer than 150msec in an utterance led an aizuchi and built an aizuchi generator based on this heuristic rule. Okato et al.[2] built a system to make aizuchi using models of specific pitch patterns of user's utterances. Noguchi et al.[3] proposed a method to make aizuchi using prosodic information such as changes of fundamental frequency in the end of utterances and pause. Sato et al.[4] investigate a method to detect natural turn taking timing. They adopted a decision tree based on prosodic and linguistic information with a similar technique.

We also built an aizuchi generation system using prosodic information [5]. Dialogs between human and this system were recorded and evaluated subjectively and about 70% of aizuchi generated by this system were regarded as natural. Concerning a turn taking, there are many of systems which allows users to barge in an utterance of the system [6][7], but system's response timing of aizuchi and turn taking is not considered.

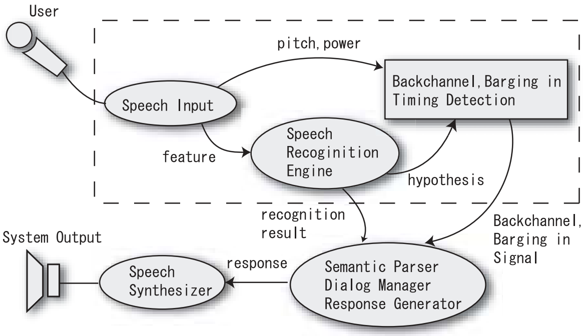


Figure 1: *Our target system*

3. Target system

Our target system is shown in Figure 1. Speech input module extracts spectral feature, pitch and power parameters. Spectral feature parameters are sent to a speech recognition engine and pitch and power are sent to a response timing generator.

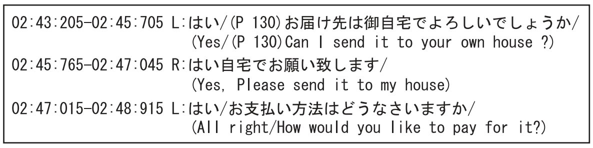
The speech recognition engine has to output intermediate hypotheses in real time[1]. Response timing generator selects the action of the system from among waiting, making aizuchi and taking the turn using the hypotheses from the speech recognition engine and pitch and power parameters from a speech input module.

4. Annotated dialog corpus

4.1. Corpus

We used annotated dialog corpus to analyze human-human dialog [11]. The corpus has 29 dialogs consistent with speech(L and R channel) and the annotation. The example of annotated dialog corpus is shown in Figure 2(P tag represents the pause and the number express the

[1] SPOJUS speech recognizer [10] developed in our laboratory satisfies this requirement.



```
02:43:205-02:45:705 L:はい/(P 130)お届け先は御自宅でよろしいでしょうか/
                      (Yes/(P 130)Can I send it to your own house ?)
02:45:765-02:47:045 R:はい自宅でお願い致します/
                      (Yes, Please send it to my house)
02:47:015-02:48:915 L:はい/お支払い方法はどうなさいますか/
                      (All right/How would you like to pay for it?)
```

Figure 2: *Annotated dialog corpus*

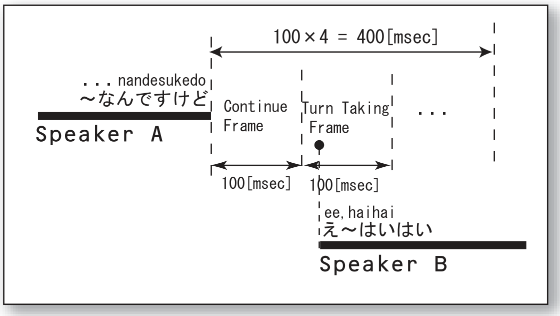


Figure 3: *Response timing analysis on a pause*

duration of the pause.). In this paper, 6 dialogs of 3 tasks were used and the total length of the dialogs was 27 minutes. These dialogs consisted of three tasks: chat, travel navigation and telephone shopping.

4.2. Analysis

4.2.1. Prosodic information

Koiso et al.[12] found a prosodic cue to decide to make aizuchi or not. There are some particular pitch and power contour patterns of the last one mora of an utterance to make the opposite speaker generate an aizuchi.

In the other side, Geluykens et al.[8] and Hirschberg[9] in term of turn taking, showed that fundamental frequency and range correlated with turn final versus turn keeping utterances in each phrase.

In Section 5, we represented the prosodic patterns with first order regression coefficients of fundamental frequency and power at the last three frames of the utterance.

Okato et.al.[2] mentioned that the duration of the utterance also relates to aizuchi. The longer the utterance is the more frequently aizuchi occurs.

4.2.2. Linguistic information

In Japanese dialog, turn taking often occurs when part-of-speech of last word in last utterance is a particle, an auxiliary verb, a verb or an interjection. In particular, turn taking is caused by particle ”ne” (chat) and ”ka” (another tasks) placed at the last of the utterance. This indicates that kinds of the last particles of utterances relate to cause turn taking. There is a growing tendency for aizuchi and turn taking topic-related phrases and keywords also lead aizuchi and turn taking. For example, in situation of telephone shopping, aizuchi often occurs just after a keyword such as product name.

5. Response timing generator

In this section, we introduce a response timing generator with a decision tree based on prosodic and linguistic information derived from the analysis in Section 4. In this study, we assume that the system can detect user's pause in real time and generates response timing in the pause.

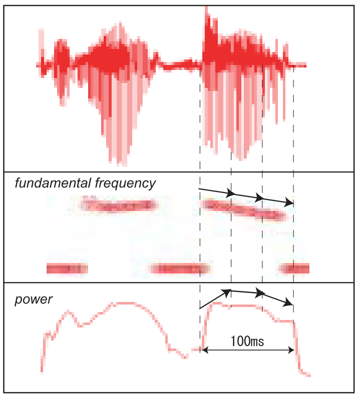


Figure 4: *Regression coefficients of fundamental frequency and power at the end of an utterance*

5.1. Construction of a decision tree for natural response timing generation

Our response timing generator first detects a pause of user's utterance and then begins to scan the pause every frame of 100 ms (Figure 3). At each frame, it classifies the frame into four classes of system behaviors: *making aizuchi*, *taking the turn*, waiting for user's successive utterance (turn keeping; *waiting(1)*) and waiting for aizuchi or turn taking (*waiting(2)*).

For this classification, we adopted a decision tree made by C4.5 learning algorithm [13]. This decision tree used following features:

1. **duration of the last utterance**
2. **part-of-speech of the last word of the last utterance**
3. **kind of the last postposition**
4. **time from the end of the previous utterance**
5. **time from the end of the last content word**
6. **duration of the content word**
7. **length between the end of the content word and the end of the utterance**
8. **fundamental frequency pattern of the end of the phrase**
9. **power pattern of the end of the phrase**

Patterns of fundamental frequency and power were described with first-order regression coefficients for fundamental frequency and power contours, respectively, in the last three regions of utterances with 50ms length and 25ms overlap as shown in Figure 4. We classified each pattern into 9 patterns such as (fall, fall, fall), (raise, fall, fall) and so on.

We prepared training data from real spoken dialog corpus with pause frames attached with tags consisting of the correct answers(aizuchi, turn taking, waiting(1), waiting(2)) and the values of the features. For example, aizuchi and turn taking generated by the decision tree are shown in Figure 5.

6. Experiment

We constructed decision rules of response timing generation derived from a decision tree trained by 3 dialogs of the corpus described in Section 4. Total length of the dialogs is 16 minutes.

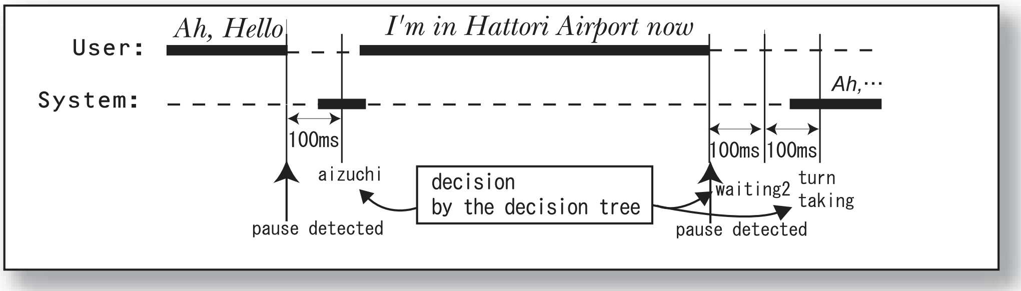


Figure 5: *Aizuchi and turn taking generated by the decision tree*

Then, we applied the decision rules to the pauses appeared in other 3 dialogs of the same corpus. Training data and test data both consist of 3 tasks described in Section 4. In this paper, we can obtain a word sequence in real time for the utterance and used the transcriptions of the data for convenience sake.

6.1. Timing reproduction experiment

We evaluated reproduction ability of the generator. All the frames discriminated by the generator are compared with real responses appeared in the corpus. We evaluated the discrimination results with *recall* rate(the rate of correctly discriminated frames among the correct appearance), *precision*(the rate of correctly discriminated frames among frames discriminated to a target class) and F-measure(harmonic average of recall and precision). Results are shown in Table 1(closed test data) and Table 2(open test data).

Table 1 shows that the response timing of the generator was similar to that of humans included in the training data. When using the generator in a dialog system, confusion of waiting(1) and waiting(2) does not matter. Thus we also show the results when treating these two waiting classes as one class.

In contrast to the closed test, Table 2 shows the lack of consistency between the system and humans who did not belong to the training set. It should be mentioned that these results did not mean the lack of the effectiveness of our method. Humans also have individual difference on this timing generation. Table 3 shows the agreement of aizuchi timings between corpus and testees. We replaced aizuchi in the corpus with pauses and testees pointed out the timings where aizuchi should be made. Then we compared the timings of the tesstees and the corpus. As shown in Table 3, the rates of agreement between human subjects were not so high. So, we cannot evaluate the generator only by this objective test and the timing by our generator may also be natural. We will confirm the fact in the next section.

6.2. Subjective evaluation

To evaluate the naturalness of the timing by our generator, we performed a subjective evaluation.

We inserted an aizuchi selected from the dialog of the same speaker who played the system at aizuchi timing point generated by our timing generator. We also made samples of turn taking, but it was almost impossible to insert an utterance which could appropriately respond to the previous user's utterance. So we picked some filled pauses as "Etto" (used as "ah" in English) to insert at the timing. Testers listened to inserted aizuchi with few preceding sentences and evaluated only the timing.

Table 1: *Results of classification of frames in pauses. (closed test)*

(a)Classification result[number]

task	in \ out	class			
		Aizuchi	Turn	Wait(1)	Wait(2)
Chat	Aizuchi	10	2	5	3
	Turn taking	0	7	2	0
	Waiting(1)	0	0	21	1
	Waiting(2)	2	0	6	17
Travel navi	Aizuchi	56	13	5	17
	Turn taking	2	63	2	9
	Waiting(1)	1	3	206	17
	Waiting(2)	22	11	1	111
Tele-phone	Aizuchi	15	9	5	1
	Turn taking	0	70	10	24
	Waiting(1)	0	1	61	0
	Waiting(2)	1	14	12	127

(b)Classification accuracy

task average	Recall	53.8%	76.0%	94.9%	75.7%
				92.8%	
	Precision	82.1%	74.1%	75.8%	78.9%
				85.4%	
	F-measure	65.0	75.0	84.3	77.2
				88.9	

Table 2: *Results of classification of frames in pauses (open test)*

(a)Classification result[number]

task	in \ out	class			
		Aizuchi	Turn	Wait(1)	Wait(2)
Chat	Aizuchi	3	0	0	0
	Turn taking	2	8	9	3
	Waiting(1)	8	0	16	13
	Waiting(2)	5	0	11	14
Travel navi	Aizuchi	2	0	3	0
	Turn taking	0	43	14	4
	Waiting(1)	4	4	82	32
	Waiting(2)	0	41	58	40
Tele-phone	Aizuchi	10	2	10	9
	Turn taking	6	35	48	38
	Waiting(1)	20	14	32	18
	Waiting(2)	5	2	71	82

(b)Classification accuracy

task average	Recall	57.4%	44.8%	49.5%	42.2%
				81.7%	
	Precision	24.8%	71.6%	38.8%	51.7%
				79.3%	
	F-measure	34.6	55.1	43.5	46.5
				80.4	

Table 3: *Agreement of the timings of aizuchi between the testees and the corpus*

testee	task	recall	precision
testee 1	chat	50.0%	47.6%
	travel navi	36.8%	35.0%
	telephone	37.5%	8.6%
testee 2	chat	20.0%	57.1%
	travel navi	5.3%	14.3%
	telephone	0.0%	0.0%
testee 3	chat	5.0%	4.8%
	travel navi	61.4%	33.3%
	telephone	25.0%	3.8%
testee 4	chat	20.0%	50.0%
	travel navi	29.8%	38.6%
	telephone	12.5%	6.7%

Table 4: *Subjective evaluation result (aizuchi)*

	1 (too early)	2	3 (good)	4	5 (too late)	6 (outlier)
Corpus	2	14	73	6	0	0
System	1	20	57	0	0	2

Table 5: *Subjective evaluation result (turn taking)*

	1 (too early)	2	3 (good)	4	5 (too late)	6 (outlier)
Corpus	0	16	62	4	1	2
System	1	18	66	2	0	3

We also compared the timing by the generator to that in the corpus. In real dialogs of the corpus, responses may have some meanings consistent with the context and the meanings may make the testees feel natural, especially in the case of turn taking. To make testees to evaluate only the timing, we also replace the real response with aizuchi and filled pause selected from other part of the dialog, as the case of the generator. We made 19 and 16 samples for aizuchi timing of humans and system, respectively, and 17 and 18 samples for turn taking of humans and systems, respectively.

Five persons listened to the data and evaluated by choosing one of the following: 1:too early, 2:early, 3:good, 4:late, 5:too late, 6:out of the question (aizuchi or turn taking should not occurs at this pause). Results are shown in Tables 4 and 5 for aizuchi and turn taking, respectively.

We cannot find significant difference between human and system, so in most cases our generator can make natural timing if prosodic and surface linguistic information is available. We can find that some samples were felt as too much unnatural even in the cases of human's timing. Contents of the response may have some correlation with naturalness of the timing.

7. Conclusions

In this paper, we tried to develop a dialog system which can generate timing of aizuchi and turn-taking in real time. Therefore, we analyzed conversations between two persons and extracted prosodic and linguistic information which had effect on the timing.

In closed test, the response timing of the generator was similar to that of the responses which humans made to in the training data. In contrast to the closed test, open test showed the lack of consistency between the system and humans who did not belong to the training. However, humans also have individual difference on this timing generation. In fact, we cannot find significant difference between human and system. So the timing made by our generator was proven to be natural.

In the future, we plan to construct a system with response in real time by combining our decision rules with speech dialog system. For the goal, we will first integrate the generator with a speech recognizer to make the timing fully automatically.

8. References

[1] Ward, N., "Prosodic features which cue back-channel responses in English and Japanese", Journal of Pragmatics 32, pp.1177-1207, 2000.

[2] Okato, Y., Kato, K., Yamamoto, M., and Itahashi, S., "Insertion of interjectory response based on prosodic information" In IEEE Workshop Interactive Voice Technology for Telecommunication Applications (IVTTA-96), pp.85-88, 1996.

[3] Noguchi, H., Den, Y., " Prosody-based detection of the context of backchannel responses", in Proc. ICSLP-98, pp.487-490, 1998.

[4] Sato, R., Higashinaka, R., Tamoto, M., Nakano, M., Aikawa, K., "Learning decision trees to determine turn-taking by spoken dialogue systems", in Proc, ICSLP-02, pp.861-864, 2002.

[5] Takeuchi, M., Kitaoka, N., Nakagawa, S., "Implementation and evaluation of an "aizuchi" generation system using prosodic information", Information Processing Society of Japan, Vol.2, pp.101-102, 2002.

[6] Hirasawa, J., Nakano, M., Kawabata, T., and Aikawa, K., "Effects of system barge-in responses on user impressions", in Proc. Eurospeech-99, Vol 3, pp.1391-1394, 1999.

[7] Kamm, C., Narayanan, S., Dutton, D., and Ritenour, R., "Evaluating spoken dialogue systems for telecommunication services", Eurospeech-97, Rhodes, Greece, pp.2203-2206, 1997.

[8] Geluykens, R., Swerts, M., "Prosodic cues to discourse boundaries in experimental dialogues", Speech Communication 15, pp.69-77, 1994.

[9] Hirschberg, J., "Communication and prosody: Functional aspects of prosody", Speech Communication 36, pp.31-43, 2002.

[10] Kai, A. and Nakagawa, S., "A Frame-Synchronous Continuous Speech Recognition Algorithm Using a Top-Down Parsing of Context-Free Grammer", in Proc, ICSLP-92, pp.257-260, 1992.

[11] SIG of Corpus-Based Research for Discourse and Dialogue, JSAI, "Constructing a Spoken Dialogue Corpus as Sharable Research Resource", Japanese Society for Artificial Intelligence, SIG-SLUD-9903-4, 1999.

[12] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y., "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs", Language and Speech, vol.41, No.3-4, pp.291-317, 1998.

[13] J. Quinlan, R., C4.5:Programs for Machine Learning, Morgan Kaufmann, 1992.