Timing Detection for Realtime Dialog Systems Using Prosodic and Linguistic Information

Masashi Takeuchi, Norihide Kitaoka and Seiichi Nakagawa

Toyohashi University of Technology, Toyohashi, 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 become smoother. 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 to detect the timing based on the features coming from the information and examined the decision rules. We also applied the decision tree to a timing generator. 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 first 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 adopted an algorithm named C4.5 which could automatically construct a decision tree to detect the response timing based on the features derived from the above analysis and examined the tree to find which information was important. We finally propose a timing generation method based on the decision tree.

The rest of the paper is organized as follows: related works are introduced in Section 2. Human-human dialogs were analyzed in Section 3. We construct a decision tree in Section 4. Section 5 contains application of the tree for response timing generation and the experimental results. Section 6 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] investigated a method to detect natural turn-taking timing. They adopted a decision tree based on prosodic and linguistic information.

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

3. Annotated dialog corpus

3.1. Corpus

We used annotated dialog corpus to analyze human-human dialog [11]. The corpus has 29 dialogs consisting with speech(L and R channel) and the annotation. 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. We found that 63%, 35% and 18% of aizuchi in the dialogs of chat, travel navigation and telephone shopping, respectively, overlapped the preceding utterances, but we don't deal with the overlap in this paper.

3.2. Analysis

3.2.1. Prosodic information

Koiso et al.[12] found prosodic cues 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 the range correlated with turn final versus turn keeping utterances in each phrase.

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

3.2.2. Linguistic information

In Japanese dialog, turn-taking often occurs when part-of-speech of last word in the last utterance is a particle, an auxiliary verb, a verb or an interjection. In particular, turn-taking is caused by the particles "ne" (chat) and "ka" (another tasks) placed at the last of the utterance. This indicates that kind of the last particles of utterances is related to cause turn-taking. Topic-related phrases and some 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.

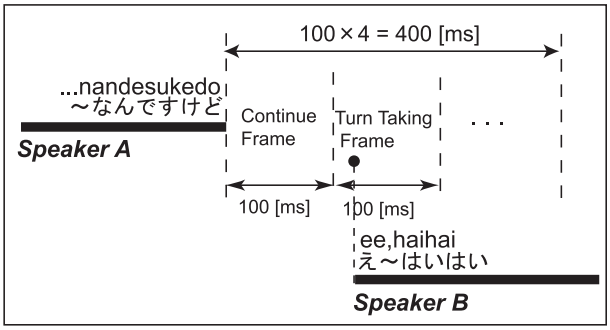

Figure 1: *Response timing analysis on a pause*

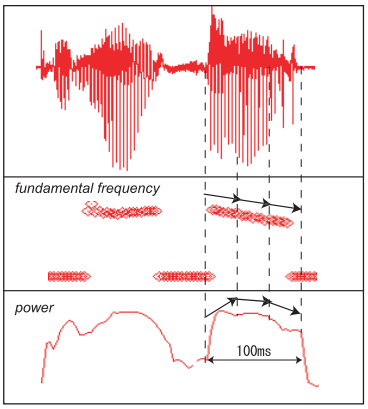

Figure 2: *Regression coefficients of fundamental frequency and power at the end of an utterance*

4. Construction of timing decision rules

4.1. Decision tree for natural response timing detection

We tried to make rules which could detect natural response timing. Pauses after user's utterances were divided to 100 ms frame and each frames was classified 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 the classification, we adopted C4.5 learning algorithm [13]. This algorithm can automatically construct a decision tree to classify the frames when given examples with the following features and the tree can be converted to a set of decision rules.

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 of the end of the phrase**
9. **power 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 2.

```
utterance_duration > 672 :
|  power1 <= 0.04 :
|  |  after_utterance_duration > 250 : continue (122.0/68.0)
|  |  after_utterance_duration <= 250 :
|  |  |  power2 > 0.04 : continue (65.0/31.0)
|  |  |  power2 <= 0.04 :
|  |  |  |  power3 <= -0.01 :
|  |  |  |  |  content_word_duration > 456 : back_channel (210.2/23.7)
|  |  |  |  |  content_word_duration <= 456 :
|  |  |  |  |  |  power3 <= -0.06 :
|  |  |  |  |  |  |  pitch3 <= -0.1 : wait (63.0/29.0)
|  |  |  |  |  |  |  pitch3 > -0.1 : change (61.0/20.0)
|  |  |  |  |  |  power3 > -0.06 :
|  |  |  |  |  |  |  pitch1 <= 0.3 : back_channel (372.3/103.9)
|  |  |  |  |  |  |  pitch1 > 0.3 : change (61.5/37.5)
```

Figure 3: *A part of the decision tree trained by all the training data*

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.

4.2. Analysis of the decision tree

Figure 3 shows a part of the decision tree made using the traing set of all the tasks (chat, travel navigation, telephone shopping).

In this decision tree we found that most of appearing features in the decision tree were prosodic features such as utterance duration, dynamics fundamental frequency and power and pause durations after the preceding utterances.

Linguistic information was hardly used. The content word duration was a major feature but the linugistic information does not play an important roll in the feature. The characteristics of this task-independent decision tree might be caused by the training data consisting of all the tasks. So we also constructed and examined a decision tree for each task.

These task-dependent decision trees were all similar to the task-independent decision tree. This assured that the prosodic information played more important roll to detect response timing than linguistic information. So we derived that aizuchi and turn-taking can be detected by only prosodic information and a small amount of surface linguistic information (such as the distinction between content words and the others). We can also imagine that humans may be able to make aizuchi not so much depending on the contents of preceding utterances.

5. Application of the decision tree for response timing generation

The decision tree constructed in Section 4 can be used a response timing generator. The timing generator first detects a pause of the user, and then classifies at every 100 ms frames of the pause to 4 classes described in Section 4.1. This procedure is illustrated in Figure 4.

5.1. Dialog system with the response timing generator

Our target system is shown in Figure 5. Speech input module extracts spectral feature, pitch & 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

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

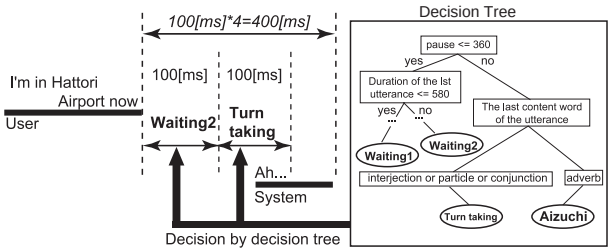

Figure 4: *Aizuchi and turn-taking generated by the decision tree*

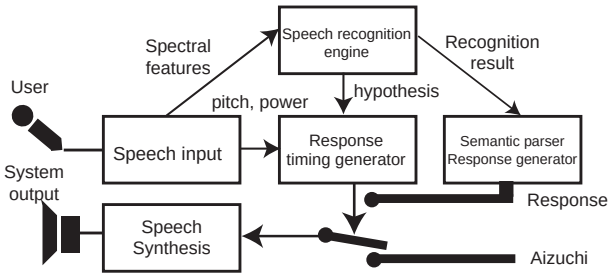

Figure 5: *Our target system*

Table 1: *Results of classification of frames in pauses*

(a)Classification result[number]

task	in \ out	class (closed / open set)			
		Aizuchi	Turn	Wait(1)	Wait(2)
Chat	Aizuchi	10 / 3	2 / 0	5 / 0	3 / 0
	Turn-taking	0 / 2	7 / 8	2 / 9	0 / 3
	Waiting(1)	0 / 8	0 / 0	21 / 16	1 / 13
	Waiting(2)	2 / 5	0 / 0	6 / 11	17 / 14
Travel navi	Aizuchi	56 / 2	13 / 0	5 / 3	17 / 0
	Turn-taking	2 / 0	63 / 43	2 / 14	9 / 4
	Waiting(1)	1 / 4	3 / 4	206 / 82	17 / 32
	Waiting(2)	22 / 0	11 / 41	1 / 58	111 / 40
Tele-phone	Aizuchi	15 / 10	9 / 2	5 / 10	1 / 9
	Turn-taking	0 / 6	70 / 35	10 / 48	24 / 38
	Waiting(1)	0 / 20	1 / 14	61 / 32	0 /18
	Waiting(2)	1 / 5	14 / 2	12 / 71	127 / 82

(b)Classification accuracy (average of all the tasks)

REC [%]	53.8 / 57.4	76.0 / 44.8	94.9 / 49.5	75.7 / 42.2
			92.8 / 81.7	
PRE [%]	82.1 / 24.8	74.1 / 71.6	75.8 / 38.8	78.9 / 51.7
			85.4 / 79.3	
F	65.0 / 34.6	75.0 / 55.1	84.3 / 43.5	77.2 / 46.5
			88.9 / 80.4	

REC:Recall PRE:Precision F:F-measure

of the system from among waiting, making aizuchi and taking the turn using the hypotheses from the speech recognition engine and pitch & power parameters from a speech input module.

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

Then, we applied the decision rules to the pauses appeared in other 3 dialogs of the same corpus. The 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 it as the transcription of the data for convenience sake.

5.2. 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/open test data).

A closed test result of 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, an open test result of Table 1 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 defectiveness of our method. Humans also have individual difference on this timing generation. Table 2 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 2, 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.

5.3. Subjective evaluation

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

We inserted an aizuchi extracted from the dialog of the same speaker who played the system (WOZ) at aizuchi timing point generated by our timing generator. We also made samples of turn-taking, but it was almost impossible to insert an alternative speaker's 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.

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. We were afraid that we could not fairly compare the generator with corpus if the system utterance had some meanings consistent with the previous user's utterance in the case of the corpus. To make testees to evaluate only the timing, we also replace the real response(i.e. the following utterance) with aizuchi and filled pause extracted from other part of the dialog, as the case of the generator.

We made 18 and 16 samples for aizuchi timing of humans and system, respectively, and 16 and 17 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 the first and second rows (Rep) of Tables 3 and 4 for aizuchi and turn-taking, respectively.

We cannot find significant difference between human (Corpus) and system, so in most cases our generator can generate 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.

Then, we also tested the real aizuchi and turn-taking to ex-

Table 2: *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 3: *Subjective evaluation result (aizuchi). "rate" is the rate of natural timing. "Rep" means aizuchi replaced by the real response.*

	1 early	2	3 good	4	5 late	6 outlier	rate[%]
Corpus(Rep)	0	12	66	9	0	3	73.3
System(Rep)	0	15	64	0	0	1	80.0
Corpus(Real)	1	13	69	6	0	6	72.6

Table 4: *Subjective evaluation result (turn-taking)*

	1 early	2	3 good	4	5 late	6 outlier	rate[%]
Corpus(Rep)	0	15	54	7	0	4	67.5
System(Rep)	3	21	55	1	0	5	64.7
Corpus(Real)	0	6	72	2	0	0	90.0

amine the effect of the replacement. For a fair comparison, the original utterances were degraded a little by adding a noise. For this experiment we used 19 and 16 samples for aizuchi and turn-taking, respectively .

The results of are shown in third rows (Real) of Tables 3 and 4 for aizuchi and turn-taking, respectively. Even if testees heard a real response, aizuchi was not always evaluated as natural. The reason is not clear so far, but it suggests the difficulty of the aizuchi timing evaluation by third parties. Naturalness of the turn-taking relies on the contents of the succeeding utterance. On the other hand the result of turn-taking in Table 4 was as expected. In the contrast to aizuchi's result, turn-taking is considered to depend on contents of following utterance than testee.

6. 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 differences on this timing generation. In fact, we cannot find a 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 a spoken dialog system. For the goal, we will first integrate the generator with a speech recognizer to make the timing fully automatically.

7. References

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

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

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

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

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

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

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

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

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

[10] Kai, A. and Nakagawa, S., 1992. "A Frame-synchronous continuous speech recognition algorithm using a top-down parsing of context-free grammar", in *Proc, ICSLP-92*, 257-260.

[11] SIG of Corpus-Based Research for Discourse and Dialogue, JSAI, 1999. "Constructing a spoken dialogue dorpus as sharable research resource", *Japanese Society for Artificial Intelligence*, SIG-SLUD-9903-4.

[12] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y., 1998. "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, 291-317.

[13] J. Quinlan, R., 1992. "C4.5:Programs for machine learning", Morgan Kaufmann.