AN ALGORITHM FOR CHOOSING JAPANESE ACKNOWLEDGMENTS USING PROSODIC CUES AND CONTEXT

Wataru Tsukahara
Mech-Info Engineering, University of Tokyo, Bunkyo-ku Tokyo 113-8656 Japan
tsuka@sanpo.t.u-tokyo.ac.jp*

ABSTRACT

In human dialog a wide variety of acknowledgments are used. One function of this seems to be indicating attention, interest, and involvement to the other speaker. Based on study of Japanese memory game dialogs, we propose an algorithm for choosing among acknowledgment responses, including *hai (yes)*, *so (right)*, and *un (mm)*. The primary factors involved are aspects of the user's internal state, including confidence and liveliness, as inferred from the context and the user's prosody. We found that judges preferred algorithm-chosen responses to randomly varied responses, confirming our hypothesis that 'sensitive' and subtle choice of response may improve helpfulness and naturalness of man-machine spoken language interaction.

1 INTRODUCTION

Thanks to progress in speech recognition and understanding, high recognition accuracy can be achieved for formal speaking styles such as broadcast news reading (Young 1996); and it seems likely that eventually systems will be able to accurately recognize even spontaneous speech. This will bring us closer to the ultimate goal: spoken dialog systems which can engage in natural interaction with people. However, it will not alone be sufficient (Mane *et al.* 1996).

This need can be seen when observing interactions with current best systems which understand and retrieve information accurately and efficiently, but not responsively. They are lacking in "responsiveness" (Ward 1997) — the ability to speak at precisely appropriate times with precisely appropriate utterances. Some speech systems are responsive in ways using gesture and gaze (Thorisson 1994), backchannel feedback timing (Ward 1996), and turn-taking timing (Schmandt 1994, pp. 199-203).

However, subtle choice of what to say is one aspect of responsiveness which has not received much attention. Human participants in dialog vary their responses depending on the feelings of their interlocutors. A good example is following transcription of a human to human dialog from our corpus. In this dialog, **A** is trying to remember the names of some train stations, and **B** is trying to help him.

A: *YOYOGI HARAJUKU SHIBUYA*
B: *hai(yes) hai hai*

and here is another portion:

A: *··· HAMAMATSUCHO no tsugiwa ···*
B: *saishoni resshaga hashi-*
A: *NIHONBASHI! aha chigau? e-to SHINBASHI!*
B: *ttatoko Bu-! soso!*
(translation of above portion)
A: *··· next to HAMAMATSUCHO is···*
B: *the first train station*
A: *NIHONBASHI! (laugh) no?, uhh SHINBASHI!*
B: *No! That's it!*

In this example, **B** seems like a helpful buddy. He listens well; he does not interrupt **A** when things are going smoothly, but he gave a helpful hint when **A** is in confusion. And when **A** blurts out an answer after some struggle, B responds lively, seeming to share the moment of pleasure.

Human listeners achieve such feats of responsiveness by paying close attention to the internal state of the speaker at each moment. This is possible, we believe, by use of context and prosody. For example, emotional state such as happiness, sadness, rage, suspicion can be inferred from prosody (Morlec *et al.* 1997).

Our goal is to build a responsive spoken language system. As a first step, this paper reports on an algorithm for choice of acknowledgments, and results regarding its contribution to the helpfulness and naturalness of dialog.

2 DIALOG DATA

To study responsiveness in dialog, we wanted a task having: (1) Semantic simplicity — to simplify analysis, and to simplify problems of dialog management and speech recognition when we implement the full system, and (2) Involvement — to allow the participants to get excited and have fun interacting with each other. This led us away from some common corpus types, such as the Map Task (Anderson *et al.* 1991), because such dialogs are too complex to analyze and implement as a system. We also rejected common speech system domains, such as ticket reservations, where the recognition and understanding of content is more important than responsiveness.

*http://www.sanpo.t.u-tokyo.ac.jp/~tsuka. Thanks to Prof. Nigel Ward for direction and extensive feedback, and to Nakayama Foundation and Inamori Foundation for support.

Thus, we choose a memory game, because it is semantically trivial, but nevertheless engaging. Specifically, it starts like this: 'can you name all 29 stations of the Yamate loop line? Say them in order, and I'll give you hints if you get stuck'.

From preliminary collected 41 human-human dialogs (45 speakers), we selected the best 'tutor' who (1) did not ignore answers, (2) enjoyed dialog and (3) gave proper hints. With this 'tutor,' we constructed a corpus for analysis (6 dialogs, 30 min.). Table 1 summarizes the corpus.

gender	4 males, 2 females
social status	4 junior, 1 senior, 1 classmate
dialog length	5.0 min. (Min. 2.0 ; Max. 11.1)
number of errors	8.7 (Min. 2 ; Max. 15)

Table 1: averaged values in 6 dialogs

The dialogs contained a great variety of acknowledgments, as listed in table 2. We believe this was an important factor in encouraging the 'students' and keeping up their interest. Major acknowledgments are *hai*, *un*, *so*, duplication of these (*haihai*, *soso*, *unun*), repetition of a station name answered (denoted as <repetition> hereafter) and keeping silent. In English a similar choice appears, between *yes*, *mm*, <repetition>, *right*, *etc.*

response	frequency
hai	**82 (46%)**
hai, ok	2
hoi	1
so	**10 (6%)**
so-da	2
un	**19 (11%)**
haihai	5
soso	6
unun	1

('-' means elongation)

<repetition>	**16(9%)**
pinpo-n	10 (6%)
ye-	1
a-	2
o-	1
<digression>	3
<keep silent>	**9(5%)**
other	6
total	177

Table 2: frequency of each responses in the corpus

3 METHOD FOR MAKING DECISION RULES

We focus on the problem of choosing appropriate acknowledgments at each point, initially limiting our attention to major responses described above. We calculated: prosodic information — pause length, utterance duration, mean, variance, slope for pitch and power, and contextual information — number of hints and errors necessary. Based on simple analysis of these values, preliminary decision rules only choosing *hai*, *un* and <repetition> were proposed (table 3). Three judges listened to a dialog (4min.; from the corpus) in which responses were modified by the rules, and pointed out unnatural responses. Present decision rules are modification of these rules based on their comments.

response	condition
hai	previous is *hai* and no error
<repetition>	long delay before correct answer
un	not famous station, hints, fillers
hai	default

Table 3: preliminary rule for response choice

4 DECISION RULES

4.1 Do not produce responses

A judge pointed out that when a student answerd continuously, tutor's acknowledgment response was unnecessary. A Well-informed student answers continuously, and the tutor did not produce responses to every answers. There are six cases in the corpus and here is an example (S: student, T: tutor).

S: OKACHIMACHI-AKIHABARA-KANDA-TOKYO
T: OKACHIMACHI-AKIHABARA-KANDA-TOKYO

We decided to keep silent when a student had already started to utter next answer before the system's response.

4.2 Respond lively to lively answers

A judge pointed out that the tutor should have responded lively when a student answerd lively. From subjective observation, an utterance with high average pitch or power sounds lively. After adjusting the parameters to give a good agreement with subjective ratings of liveliness, we defined a parameter "liveliness" as follows:

$$(\text{liveliness}) \equiv \overline{f}_{0norm} + 1.5\overline{E}_{norm} \quad (1)$$

where $\overline{f}_{0norm}$ and $\overline{E}_{norm}$ are average pitch/power normalized by median pitch/power of dialog, respectively. An answer whose liveliness is over 3.5 is a "lively" answer.

There found 15 corresponding responses to "lively" answers (table 4). On the contrary to judge's opinion, liveliness (eq. 1) of acknowledgment responses in table 4 are not always high. If anything, they are salient for the use of unusual responses. In table 4, *unun*, *haihai*, *soso*, *pinpo-n*, *so-da*, *ye-* hold 2/3 of acknowledgment responses, while they are not major in the corpus (14 %). Thus, 'Lively response' could have two components: lexical item, and style of pronunciation. We decided to use *haihai*, *soso*, *unun*, *so*, *(station name)* instead of *hai*, *so*, *un*, <repetition> respectively, for lively answers.

order	'liveliness' of answer	corresponding response			
1	4.3	**soso**	9	3.6	**pinpo-n**
2	4.1	so	10	3.6	**ununun**
3	3.9	hai	11	3.5	so
4	3.9	**ye-**	12	3.5	**so-da**
5	3.9	**pinpo-n**	13	3.5	<keep silent>
6	3.8	**pinpo-n**	14	3.5	hai
7	3.7	**haihai**	15	3.5	**sososo**
8	3.7	**sososo**			

Table 4: corresponding responses for lively answers

4.3 Do not change responses

A judge pointed out that it was unnatural to change acknowledgment response type when things were going smoothly. We decided to follow this suggestion. 'Smooth'

S: SHINBASHI YURAKUCHO TOKYO
*T: un un *hai*

is defined as: (1) there is no hints or incorrect answers and (2) time required to answer is shorter than 1 sec. In the corpus, 40 out of 44 *hai* fit to this case. This is not surprising for native Japanese because *hai* is most usual acknowledgment. Furthermore, 1 out of 2 *un* and 1 out of 1 <repetition> can be predicted by this rule.

4.4 Be patient in difficulty

When the next answer is difficult to recall, it may take long time to answer. The longer is the time of filled/unfilled pause to answer, the less confident is his answer, in Japanese as in English (Brennan & Williams 1995). In the corpus, major responses for the 15 cases whose answer time are over 30 seconds are: <repetition>(5 cases), *pinpo-n*(5), *so-da*(2). Here is an example (number in parenthesis denotes unfilled pause length).

S: (7s) (1.5s)*GOTANDA*
T: suujideiuto 5 (number is 5) *GOTANDA*

We decided to use <repetition> when a student was in difficulty (time to answer is over 30 seconds). Impression of this response type is 'patience.' It will not rush the student, rather it shows an attitude to proceed together with the student. It sounds more kindly than business-like *hai*.

4.5 Praise after effort

A Tutor gives hints when a student gets stuck. A judge pointed out that the tutor should praise the student's effort when he could answer at last using hints successfully.

S: NISHINIPPORI
T: hora NIPPORI ni chikaindayo *so*
(translation of above)
S: NISHINIPPORI
T: near NIPPORI, you know *right*

In the corpus, responses to answers after one or more hints are: *un*(9 instances), <repetition>(9), *pinpo-n*(7), *so*(5), *hai*(5), *soso*(4). Although *so* is not most frequently used, more than half (9/16) of *so* and *soso* fall into this case. Thus, we decided to use *so* for this case. Different from friendly *un*, *so* and *soso* seem to express tutor's excitement for his successful hints, seemingly help praise the student in an indirect way.

4.6 Be friendly for unsure answers

Some *hai* were pointed out cold. In the corpus, they are responses to answers that (1) required no hints and (2) are not uttered confidently. Thus, we decided to use *un* for not confident answers, which were preferred by the judges. As is often pointed out, a rising intonation indicates low confidence (Brennan & Williams 1995). In the corpus, the responses to not confident answers (20 instances total) are: *hai*(12), *un*(4), *haihai*(3) and *so*(1). Here is an example for *un*.

S: NIPPORI?
T: un

Different from <repetition>, rising intonation strongly requires acknowledgments. Thus, *hai* or *un* is appropriate for this case. Also rising intonation shows low confidence, a tutor should use *un*, which sounds friendlier than hai.

4.7 Default response: *hai*

As a default response, we decided to use *hai* in case none of above rules were applicable, because *hai* is the most common acknowledgment in Japanese.

5 SUMMARY OF RULES

Rules described in previous section are summarized in table 5. Some rules has been modified from the preliminary rules in section 3. Rule 4.3 also includes other responses than *hai*. Rule 4.5, 4.6 are simplification of the preliminary rule for *un*. Rule 4.4 remains unchanged. Rule 4.1 and 4.2 are new. These rules will be applied in this order.

section	response	condition	characteristic of answer	characteristic of response
4.1	<keep silent>	rapid pace	very smooth	—
4.2	*haihai*, *soso*, *unun*	liveliness > 3.5	lively	lively
4.3	'smooth'	short answer time, no error, no hint	smooth	—
4.4	<repetition>	long answer time	difficult	patient
4.5	*so*	after hint	—	praising
4.6	*un*	rising intonation, no hints	unsure	fiendly
4.7	*hai*	other	—	default, business-like

Table 5: proposed decision rule

	hai	*so*	*un*	<repetition>	<lively>	<keep silent>	sum
	82	10	19	16	15	9	177
A	80 (98%)	0 (0%)	2 (10%)	6 (38%)	0 (0%)	0 (0%)	88 (50%)
B	54 (66%)	3 (30%)	4 (21%)	7 (43%)	5 (38%)	6 (66%)	79 (45%)

Table 6: prediction with 6 dialogs (**A**:preliminary rule, **B**:present rule)

Table 6 shows the prediction of responses in the corpus. Compared to the preliminary rules (**A**), prediction of *hai* became worse. However, it can predict other response types, which are more important for naturalness of dialog.

6 EVALUATION BY LISTENING

After some minor modification of the 'default' rule to include other than *hai*, we prepared three modified dialogs in two conditions, using pre-recorded tutor's voice[1] for responses: (1) chosen by the rule, (2) chosen randomly keeping the frequency ratio in table 2. Then five judges listened to these three dialogs in either condition and checked unnatural or not helpful responses without notification of conditions (as a whole, 7 out of 15 dialogs were "random"). Judges ranked these dialogs as to helpfulness and naturalness after listening the dialogs.

6.1 Results

Judges felt 16% (35/224) of responses as unnatural in random condition and 10 % (19 / 196) in rule-based condition. Table 7 shows judges' preference among 3 dialogs. These results suggest that judges seem to prefer rule-based responses. Table 8 shows ratio of unnatural responses

judges	'helpfulness' ranking 'best' ↔ 'worst'		
1	+	-	-
2	+	-	-
3	+	+	-
4	-	+	-
5	-	+	+
ratio of +	60%	60%	20%

+:rule
-:random

Table 7: judges' preference over 3 dialogs

rules	% unnatural
patient	26% (6/23)
default	16% (10/63)
lively	5% (1/21)
friendly	4% (1/25)

praise	3%(1/33)
smooth	2%(1/50)
keep silent	0% (0/2)

Table 8: ratio of unnatural responses for each rule

response	% unnatural rule		random	
<repetition>	26 %	(9/34)	30 %	(7/23)
'hai'	9 %	(7/75)	11 %	(15/134)
'so'	6 %	(2/35)	17 %	(6/35)
"lively"	3 %	(1/34)	13 %	(4/30)
'un'	2 %	(1/50)	20 %	(3/15)
<keep silent>	0 %	(0/2)	24 %	(4/17)

Table 9: ratio of unnatural response types

for each rule. Here <repetition> is worst, seemingly due to the incompatibility of response voice[1], which indicates importance of voice quality. For major responses, 'default' (32%, 63/196) needs further improvements, while 'smooth' (26%, 50/196) seems better. Table 9 shows ratio of unnatural response types. Differences seem little between rule and random condition for <repetition> and *hai*, because *hai* is mostly coming from 'default' rule (6/7). Other responses seem natural compared with those in random condition.

[1] Unfortunately, we had to use other person's voice for <repetition> because of difficulty of collecting voices

It is interesting that subjects sometimes judged correctly predicted responses to be unnatural (rule: 5/19, random: 8/35). This is almost certainly due to the prosodic incompatibility. This seems to show the necessity of choosing appropriate prosody among same response vocabulary.

7 CONCLUSION

We have proposed an algorithm for choice of acknowledgments in Japanese, using speaker's prosodic and contextual information to infer his confidence and liveliness. Evaluation of naturalness and helpfulness of dialog generated by this rule suggests that judges prefer rule-based responses to randomly chosen responses.

One topic needing further analysis is the role of various prosodic contours for these acknowledgment responses. Our main next step will be to actually build a quiz system which will take the role of tutor in these dialogs, and find out whether the added responsiveness makes it measurably more helpful in live dialogs with people.

In addition to the specific contributions in terms of response choice rules, we hope that our basic research strategy, of inferring subtle information on the speaker's internal state from context and prosody, will prove to be of general value for improving the usability of spoken dialog systems.

8 REFERENCES

1. Anderson, A. *et al.* The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366, 1991.
2. Brennan, S. E. & M. Williams. The Feeling of Another's Knowing: Prosody and filled pauses as cues to listeners about the metacognitive status of speakers. *J. Memory and Language*, 34:383–398, 1995.
3. Mane, A *et al.* Designing the User Interface for Speech Recognition Applications. *SIGCHI Bulletin*, 28(4):29–34, 1996.
4. Morlec, Y, G Bailly, & V Aubergé. Synthesing Attitudes with Global Rhythmic and Intonation Contours. In *EUROSPEECH 97*, pp. 219–222, 1997.
5. Schmandt, C. *Voice Communication with Computers.* Van Nostrand Reinhold, 1994.
6. Thorisson, K. Face-to-Face Communication with Computer Agents. In *Working Notes, AAAI Spring Symposium on Believable Agents*, pp. 86–90, 1994.
7. Ward, N. Using Prosodic Clues to Decide When to Produce Back-channel Utterances. In *ICSLP 96*, pp. 1728–1731, 1996.
8. Ward, N. Responsiveness in Dialog and Priorities for Language Research. *Systems and Cybernetics, Special Issue on Embodied Artificial Intelligence and Artificial Life*, 28:521–533, 1997.
9. Young, S. A Review of Large-vocaburary Continuous-speech Recognition. *IEEE Signal Processing Magazine*, pp. 45–57, 1996.