PACING SPOKEN DIRECTIONS TO SUIT THE LISTENER

Tatsuya Iwase, Nigel Ward
Mech-Info Engineering University of Tokyo, Bunkyo-ku Tokyo 113-8656
{tatsuya, nigel}@sanpo.t.u-tokyo.ac.jp*

ABSTRACT

On the basis of corpus analysis, we have made direction-giving dialog system which adjust the pace of dialog without using speech recognition. And we evaluate the naturalness of the resulting conversations by experiments. Then the system showed good performance.

And also the possibility of prosody to compensate for the weak points of speech recognition.

1 INTRODUCTION

Thanks to advances in speech recognition, it is now possible to build systems that let cooperative users can accomplish real tasks. "Cooperativeness", however, is required. Users who do not adapt to the needs of the system are likely to be disappointed, as are those who expect to be able to interact with a speech system as "naturally" as they can interact with a human interlocutor.

Two specific problems are the need to speak clearly and the need to wait for responses. Although we can expect further advances in the accuracy and speed of speech recognition to alleviate these problems, this will not eliminate them. The first problem involves the fact that speakers occasionally mumble: producing sounds which are inaudible or are not even words. Human interlocutors can however cope with this and continue the exchange regardless. The extra factors seem to be the use of context and prosody to infer the pragmatic force (or dialog act type) of the mumble; for example, classifying it as a false start, a musing, a back-channel, a snide remark or whatever. The second problem involves the fact that speakers often expect feedback while they are still talking. In natural conversation, back-channels in particular often occur before an utterance is complete. Again, the use of prosody seems to be important here.

Thus we see that a speech system whose only source of information is recognition of the words spoken will sometimes respond inappropriately or too late. For this reason, many researchers have recently turned attention to the uses of prosody, in particular for dialog act classification[3] and for inferring dialog structure and determining the timing of turn-taking and back-channel feedback[10]

Although much basic research remains to be done, it is not premature to build and experiment with systems which use prosody in these ways. One strategy is to use prosodic information to enhance a system based on word recognition[2],[4]. For example, Daly&Zue[1] studied about the method to classify English sentences into wh-questions, and others. And this research indicates that the sentences whose end shows rising pitch are yes/no-questions. We used this result to separate yes/no-questions from other sentences. An alternative research strategy is to build a system that uses only prosodic information. We have adopted this second strategy, for two reasons. Our philosophical reason is our belief that the prosodic aspects of dialog are very basic, and thus possibly appropriate as the foundation for building speech systems[9]. Our practical reason is that in such a system it is easier to evaluate the contribution of prosodic information.

Thus the goal of this research is to create a system capable of natural dialog using prosody only.

explainer:	kore wa Shinjuku doori desuna (that will be Shinjuku Avenue you know)
listener:	Shinjuku doori ... hai (Shinjuku Avenue ... okay)
explainer:	de massuguitte kondo Yotsuyayonchome sasetsu un (and go straight then at Yotsuyayonchome turn left mm)
listener:	sasetsu-ne? hai (left you say? okay)
explainer:	kore ga Gaien-nishi doori (that's Gaien-nishi Avenue)
listener:	Gaien-nishi ... doori, hai. (Gaien-nishi ... Avenue, okay.)

Figure 1: Example of corpus dialog

2 TASK

Following Schmandt[6][7], we set our system the task of conveying to the user directions for how to go from one place to another. Unlike Schmandt, who worked with English, we chose to work with Japanese. Specifically, our users were instructed to listen to the directions and take notes so they would be able to go to the place indicated. The routes used were real routes in Tokyo, involving driving on main streets only. Figure 1 shows an extract from the corpus.

3 CORPUS ANALYSIS

We gathered 10 human-human direction-giving dialogs, 27 minutes total. The explainer was given a route, marked

*http://www.sanpo.t.u-tokyo.ac.jp/ Supported in part by the Nakayama Foundation and by the Inamori Foundation.

utterance	frequency	proportion
back-channel	562 times	80.4%
mumble	63 times	9.0%
question	37 times	5.3%
irrelevant utterance	25 times	3.6%
longer back-channel	5 times	0.7%
request for delay	3 times	0.4%
request for segment	3 times	0.4%

Table 1: Frequency of listener's utterances

after listener's back-channel	
explainer's directions	81.0%
irrelevant utterances	9.3%
others	9.7%

after listener's question	
explainer's back-channel	50.7%
explainer's answer	40.6%
irrelevant utterances	4.3%
others	4.3%

Table 2: Joining tendency of utterances

on a real map, and asked to describe it to the listener. Explainer and listener were seated so as to prevent eye contact.

Our basic model for direction giving, borrowing from Psathas[5] is that the explainer has a sequence of "segments", which he utters in order, and the the listener talks between segments. Dialogs in the corpus were like this model; sometimes listeners interrupted or back-channeled in the middle of a segment. However we ignored such behavior for purposes of analysis, because it was fairly rare, and also because we did not want to add barge-in capability to our system.

Table 1 indicates that there are three main types of responses from listeners: back-channels (mostly 'un' and 'hai'), mumbles while writing, and questions. There were no requests for repetition.

So we limited our analysis to the three main response types.

Table 2 shows that listener's back-channels indicate "I see", and are generally followed by the explainer producing the next segment of the directions. It is also clear that, as expected, questions elicit responses by the explainer. In addition, we found that after listeners mumble, the explainer sometimes responds with a back-channel.

Based on this, we devised the dialog model seen in Figure 2. "Time-up" means that if the user is silent the system goes on to produce the next segment of directions. This interval is set to 2 seconds, which is the mean interval between the end of a listener's utterance and the start of the explainer's next segment. The "back-channel" after mumbles is in a dashed oval to indicate that only some mumbles are followed by back-channel feedback.

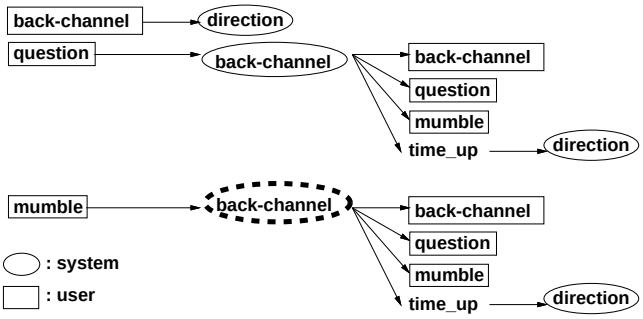

Figure 2: Dialog model

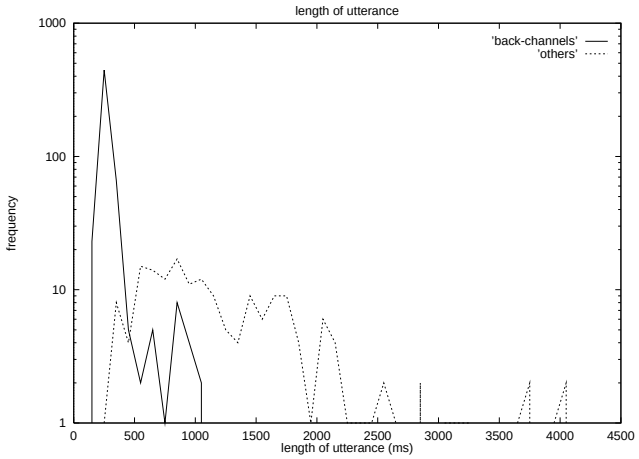

Figure 3: Utterance length of Listeners

To determine which mumbles should be followed-up with a back-channel, we looked for the "low-pitch cue"; that is, 110ms regions of low pitch were considered to be cues for back-channel feedback [8]. Defining match to a real back-channel in the corpus within 500ms as a "hit", this gave us a coverage of 58.8%, and an accuracy of 47.6%. These being higher than random prediction at the same frequency as the explainer (27%, 22%), so we judged this rule to be useful for modeling the back-channel behavior of the explainer.

Finally, we sought for prosodic criteria for distinguishing between back-channels, mumbles, and questions.

Figure 3 shows the length distribution of back-channels and other utterances. Based on this we set the threshold at 500 milliseconds; utterances shorter than this were judged to be back-channels. This rule was correct 95.1% of the time (coverage) and it detected 95.9% of the back-channels (coverage).

To classify longer utterances, we used the pitch slope of the last 200ms of each utterance, as computed by least squares data fitting. Table 3 indicates that yes/no-questions have rising pitch somewhat more often than do wh-questions or mumbles. So our system judged longer utterances with a rising pitch over the last 200ms to be yes/no-questions.

Thus there is nothing original about the prosodic features we ended up using; Figure 4 summarizes.

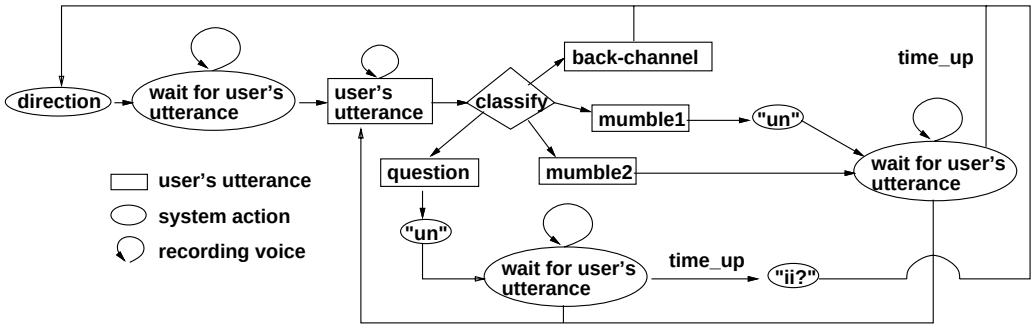

Figure 5: Direction giving system

Sentence	Proportion with rising pitch
yes/no-question	44%
wh-question	31%
mumble	25%

Table 3: Pitch slope of end of utterance

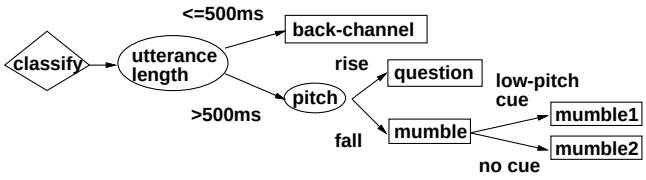

Figure 4: Classification rule

4 SYSTEM ARCHITECTURE

Figure 5 summarizes the behavior of the system. Note that the system responds to all questions with "un", meaning "uh-huh". This was the best we could do without enormously complicating the system; fortunately this response was generally appropriate. This was also followed up by "ii?", meaning "okay?", in cases where the user subsequently fell silent. Mumble 1, which has 100ms regions of low-pitch cue, always is followed by "un", and mumble 2 is never followed by "un". In Schmandt's system[6][7], one utterance of the user corresponds to one segment of the system. In our system, user can utter several times for one segment of the system. And Schmandt's system repeats previous segment as answer to user's question, but our system outputs affirmative reply.

5 EVALUATION METHOD

We contrived a method to evaluate a dialog system talking to human naturally with prosodic information. To evaluate the naturalness of dialog made by our system, we made three experiment with 10 subjects each. For evaluation, we compared a corpus of experiment conversations and questionnaires of subjects. And we introduced the concept of pacing patterns to evaluate the naturalness of dialog.

6 EXPERIMENT 1

In the first experiment, the subjects knew that the explainer was a computer system. The system gave each of the 10 subjects about a minutes worth of directions.

6.1 Analysis

First, listening to the tapes of the resulting human-computer dialogs, we found a few recurrant patterns of inappropriate pacing, as follows:

1. (missing a back-channel) When the system failed to recognize a back-channel, it didn't output the next direction, and the dialog pace became too slowed.
2. (hallucinating a back-channel) When the system misrecognizes a noise or a mumble while taking notes as a back-channel, it goes on to output the next direction, and the dialog pace was too fast.
3. (quiet subject) If subject is reticence or tense, and doesn't say anything, the system doesn't output the next direction, and the dialog stalls.

We also examined what the users had written down. 7 had written them down well enough to reach the destination, thus our task achievement rate task was 70%.

The success rate for classification of users' utterances into the three basic types was 58.9%(coverage). The extraction success rate of only back-channels was 85.3%(coverage).

6.2 User's Impressions

We also tabulated the results of user questionnaires. One key question was "What do you think of speed of direction-giving?" In dialogs where the system had failed by hallucinating back-channels, the users did indeed consider pacing to be too fast. In most other cases the users considered the pacing to be "normal". This suggests that, provided that the system correctly recognized the user's utterance types, the dialog was paced appropriately.

In response the question "was the dialog normal or strange?, on a scale from 1 to 5,", 80% answered 3 or less; thus most dialogs were not perceived to be strange.

6.3 Judges Impressions

There is often a tendency for participants to overlook infelicitous behavior by conversation partners, but third-party judges are sometimes more objective and critical. We therefore recruited 10 more subjects, different from the users of the system to listen to and evaluate the naturalness of the human-computer dialogs. We did this without first telling then that a speech system was involved. Again we found sensitivity to the failures of pacing discussed above.

After being informed that explainer in these conversations had in fact been a computer system, 90% ticked the "I was surprised to hear this" box on the questionnaire.

Finally we had them listen to a human-human direction-giving dialog, and asked for comparisons. 70% of the judges said that the human-human dialog sounded more natural than the human-computer one. Reasons given included:

- shorter turn-taking pauses
- larger vocabulary
- more false starts and fillers
- more interruptions
- the timing of back-channels was just perfect

7 EXPERIMENT 2

In experiment 1, many subjects and judges disagreed to some degree with the item "the computer system seems easy to talk with." To pursue this we did a second experiment, this time making subjects think that the explainer would be human. In this case the classification rate of subjects' utterances was 79.4%(coverage), and 86.0% of the back-channnels were detected (coverage), somewhat better than for experiment 1.

3 of the 10 dialogs broke down halfway in the "quiet subject" failure pattern, but the others were paced quite well. The questionnaire revealed that 80% of the subjects them were surprised to learn that it had been a computer system which had given them directions; thus the dialogs had been natural enough to deceive. However, in response to the question, "was the dialog normal or strange?", 50% of the subjects answered "a little strange", for the same reasons mentioned above.

In both experiments (all 20 dialogs) there were only 5 utterances which fell outside the scope of the system's dialog model, including "tsugi wa?" (meaning "and next?") and wh-questions in. In such cases the system responded inappropriately, but in each case the dialog proceeded nonetheless.

8 CONCLUSION

We found that a system which choses responses and response timing based only on the prosody of the user's utterances can give a strong impression of interacting naturally in many cases. In this we have found that Schmandt's results for English [6][7] are also true for Japanese.

In particular, in our system the pace of the dialog is naturally regulated by the user thanks to a simple mechanism: the system goes on to output the next segment of directions after the user produces an acknowledgment meaning "I see", and it gives the user more time during and after mumbles and questions.

In applications where good pacing is important and it is acceptable for responses to be appropriate only with high probability, systems which only utilize the user's prosody may be useful. Whether such a niche in fact exists, remains to be determined.

Be that as it may, we consider our results to be yet another piece of evidence for the utility of prosodic information to dialog systems.

9 REFERENCES

1. Daly, Nancy A., Zue, Victor W. "Acoustic, Perceptual, and Linguistic Analyses of Intonation Contours in Human/Machine Dialogues", *International Conference on Spoken Language Processing* :497 – 500, 1990
2. Dohsaka, Kohji., Shimazu, Akira. "A System Architecture for Spoken Utterance Production in Collaborative Dialogue", *IJCAI-97Workshop Collaboration, Cooperation and Conflict in Dialogue Systems* :25 – 31, 1997
3. Jurafsky, Daniel., Bates, Rebecca. "Automatic Detection of Discourse Structure for Speech Recognition and Understanding", *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding* :2383 – 2386, 1997
4. Kompe, R., Kuhn, T., etc. "Prosody Takes Over: Towards a prosodically guided dialog system", *Speech Communication, 15* :155 – 167, 1994
5. Psathas, George. "The Structure of Direction-giving in Interaction", *Talk and Social Structure* :195 – 216, 1991
6. Schmandt, Christopher. "Employing Voice Back Channels to Facilitate Audio Document Retrieval ", *In Proceedings of ACM Conference on Office Computing Systems* :213 – 218, 1988
7. Schmandt, Christopher. "Voice Communication with Computers", *VNR Computer Library* :199 – 204, 1994
8. Ward, Nigel., Tsukahara, Wataru. "Production of Back-Channel Feedback in Japanese may involve a Prosodically Triggered Reflex", *submitted to Language*, 1998
9. Ward, Nigel. "Responsiveness in Dialog and Prioritiesfor Language Research", *Systems and Cybernetics, Special Issue on Embodied Artificial Intelligence*, :521 – 533, 1997
10. Ward, Nigel. "Using Prosodic Clues to Decide When to Produce Back-channel Utterances" *International Conference on Spoken Language Processing 96* :1728 – 1731, 1996