A Responsive Dialog System

Nigel Ward and Wataru Tsukahara[1]
University of Tokyo

Being responsive is important in dialog. In particular, back-channel feedback is essential to human conversations. Back-channel feedback is sometimes produced without thinking, in response to simple prosodic clues. A simple implementation of this behavior produces natural responses in conversation with live human subjects.

あいづちを打てる対話システム

N. ワード　塚原渉
東京大学

対話においては応答性が重要である。なかでも、あいづちは人間同士の会話にとって必要不可欠なものである。あいづちを含む会話の分析から、しばしば内容ではなく、簡単な韻律的特徴に反応してあいづちが打たれていることが分かった。この特徴を組み込んだシステムは、実際の人間との会話で自然な応答をすることができた。

1 Motivation

Modeling language as people really use it is an elusive goal. Today, thanks to advances in speech recognition, dialog-capable systems exist, but they can not yet interact naturally with humans. A broad-brush list of weaknesses of the "typical speech system of today" shows where more work is needed:

1. The information conveyed is propositional (for example, a specification of the fields of a database query);

   but for human dialog, information exchange at pragmatic and other levels is also important.

2. Priority is given to understanding and responsing accurately;

   but for human dialog, being responsive and interactive is also important.

3. The granularity of interaction is the sentence;

   but for human dialog, interaction happens frequently, in real-time, often with overlapping utterances;

4. The system can only understand words, and can only produce words;

   but for human dialog, other channels of communication, including eye contact, gestures, and prosody, are also present.

Given that these points are important to human language use, the question arises: how do we build systems with these abilities? The obvious approach is to add these abilities to a "typical" speech system. An alternative approach is to take these abilities as central, and build systems around them. This paper reports on an effort in the latter direction.

Our current work focuses on points 2 and 3 above, under the rubric of "responsiveness in dialog". (This is prerequisite to point 1, to the extent that pragmatic information is often expressed in the interaction more than in the words. This may also provide a new angle of attack on point 4, in that it opens up the possibility of multi-modal systems based on subsumption architectures (Ward 1996a).)

2 Phenomenon

Back-channel feedback is produced by one participant in response to some utterance by the other participant. Prototypical back-channel feedback: 1. encourages or allows the other to keep speaking, 2. shows attention and interest, 3, shows understanding and/or agreement. (Discussion of our exact definition appears elsewhere (Ward 1996b). In English *mm* and *uh huh* are typical back-channel feedback. In Japanese the aizuchi *un* is most typical.

3 Analysis

Many have sought for the perceptual clue that tells a participant "it's now time to produce back-channel feedback". It has often been speculated that this clue from the speaker would be prosodic, rather than involving meaning.

We looked at the prosodic environments of 900 aizuchi in natural Japanese conversation. Potential clues we considered included pitch contours, vowel lengthening or speaking rate slowdown, volume increase or decrease on final syllables, a low pitch point, and gross energy level changes (to detect when the speaker finishes speaking), as suggested in the literature (see citations in (Ward 1996b)). None of these appeared to have a strong correlation

[1] Mechano-Informatic Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113 Japan; +81-3-3182-2111 ext. 6282; fax: +81-3-3815-8356; {nigel,tsuka}@sanpo.t.u-tokyo.ac.jp; http://www.sanpo.t.u-tokyo.ac.jp/{~nigel/,~tsuka/index-j}

Upon detection of
a region of pitch less than the 30th-percentile pitch level and
continuing for at least 150ms,
coming after at least 700ms of speech,
you should produce an aizuchi 300ms later,
providing you have not done so within the preceding 1.0 seconds.

図 1: Back-Channel Feedback Rule

with whether back-channel feedback was produced or not.

However, there was one good clue: a region of low pitch. This correlation can be operationalized as the prediction rule seen in Figure 1.

It is commonly thought that silence (at the end of a speaker's turn) is a major clue for back-channel feedback. This was not the case for our data; indeed it could not be, given the swiftness of back-channel feedback and the slowness of human reaction time. It turns out that the above rule handles both back-channel feedback which was produced after the speaker paused and stopped, and that which overlapped the speaker's continued utterance.

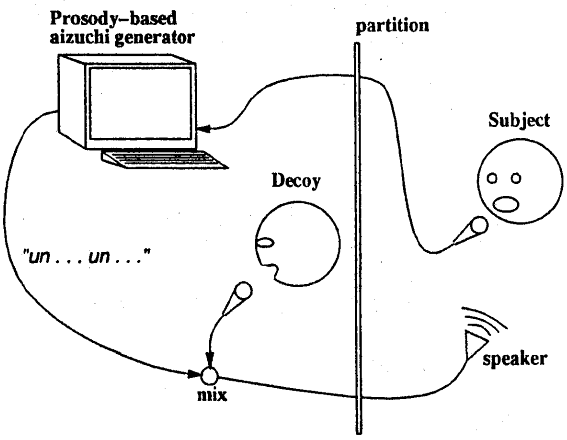

図 2: Experiment Set-up

4 Results

We tested the predictions of the above rule against the corpus of human conversations. We observed a coverage of 50% at an accuracy of 34%, over all speakers and all dialog types. For some situations performance was very good: in particular, compared to the occurrences of aizuchis produced by JH in response to KI in their 5 minute conversation, the rule correctly predicted 69% (54/78), with an accuracy of 68% (54 correct predictions / 81 total predictions).

We also have tested the performance of the rule in live conversation.

In order to get people to try to interact naturally with the system, it was necessary to fool them into thinking that they were interacting with a person. So we used a human decoy to start the conversation, and then let the system take over.

We have done experiments over the telephone and in the laboratory, with a partition so that the subject couldn't see when it the was the system that was responding (Figure 2). In both cases the system's responses seem natural.

It is interesting that even randomly produced back-channel feedback is not detected by most subjects; they still think they are talking to a human. The difference is, however, very obvious to others listening to the conversation.

Essentially the same prediction rule gave good performance for *some* English speakers in live experiments.

5 Significance

We have demonstrated a system that can keep up its end of a conversation, without doing speech recognition or understanding.

It is sometimes assumed that real-time responsiveness in spoken dialog systems is merely a question of fast algorithms and fast hardware. Here we see that that is not the whole story: there is also the human-factors issue of exactly how to time feedback. Our data suggest that this should be neither too fast nor too slow.

Our next step will be to integrate this new responsiveness with extant techniques for recognition and understanding; our aim will be to build a system that will interact truly naturally with people in a simple verbal game.

6 Acknowledgements

We thank Keikichi Hirose for the pitch tracker, Joji Habu for helping figure out how to fool subjects, and several dozen students for conversations and labeling.

References

Ward, Nigel (1996a). Reactive Responsiveness in Dialog. In *AAAI Fall Symposium on Embodied Cognition and Action*, pp. 129–133. American Association for Artificial Intelligence.

Ward, Nigel (1996b). Using Prosodic Clues to Decide When to Produce Back-channel Utterances. In *1996 International Conference on Spoken Language Processing*, pp. 1728–1731.