EVALUATING RESPONSIVENESS IN SPOKEN DIALOG SYSTEMS

Wataru Tsukahara[1] *and Nigel Ward*[2] *
Mech-Info Engineering, University of Tokyo, Bunkyo-ku Tokyo 113-8656 Japan

ABSTRACT

Ratings of user satisfaction, although fairly easy to elicit for today's spoken language systems, can be more elusive for systems which operate at near-human levels of performance. This problem can be alleviated by adding a 're-listening' phase before eliciting judgements: in this phase the user listens to a recording of himself interacting with the system while consulting a transcript of that interaction. This technique allows more sensitive judgements of system quality by avoiding problems arising from attention limits.

1 EVALUATING PERSONABLE SYSTEMS

Speech technology has advanced to the point where there exist systems which allow motivated users to accomplish useful tasks. However there are also potential applications where users are not so motivated and where there is no clear criterion of task-completion, including applications in education, commerce, and entertainment. In these situations it will be important for interaction with the system to itself be pleasant, otherwise the customer may hang up before making the purchase, or the student may exit the application before learning anything. Putting it more positively, systems in these domains will need to be able to motivate, persuade, and amuse users. Such skills are, of course, very valuable for human communicators: they make the difference between personable, effective salesmen and annoying drones, and between skilled private tutors and nagging nuisances. To produce spoken language systems which are not only efficient and accurate, but which also can provide such quality-of-interaction, is a long term research challenge.

One prerequisite to advances in this area is a method for measuring user satisfaction at this level of performance. For today's systems evaluation is relatively well understood: user satisfaction correlates highly with lack of gross errors, delays and infelicities (Walker *et al.* 1998). But for systems with high quality-of-interaction, where the issue is not lack of bugs but positive value, evaluation is more difficult. This paper addresses this issue.

*We thank the Nakayama Foundation, the Inamori Foundation, the International Communication Foundation, and the Japanese Ministry of Education for support.

[1]currently at Hitachi Ltd. w-tsuka@crl.hitachi.co.jp

[2]nigel@sanpo.t.u-tokyo.ac.jp,
http://www.sanpo.t.u-tokyo.ac.jp/∼nigel/

Section 2 discusses an implementation of one aspect of quality interaction, and Section 3 describes the difficulty of getting users to accurately describe their degree of satisfaction with this. Section 4 presents the re-listening evaluation method, Section 5 explains how we used it in practice, and Section 6 describes how it gave better results. Section 7 summarizes.

2 RESPONSIVENESS AND ACKNOWLEDGEMENT CHOICE

Achieving the sort of interactions envisioned above will require systems to have at least two things. The first is, of course, more accurate recognition of more spontaneous language. Second is what we have been calling 'responsiveness' (Ward 1997) — the ability to speak at precisely appropriate times with precisely appropriate utterances. In particular, we hypothesize that human participants in dialog vary their responses depending on the 'feelings' and attitudes of their interlocutors, as these change second-by-second during the course of a conversation. This is a a form of "user-adaptive behavior", but differs from that typically addressed by "user modeling" research (Rich 1999) in that it it is very swift, and in that it relies not on careful reasoning or deep domain knowledge, but rather on simple features of the context and on non-verbal cues provided by the user.

As a case study, we have been studying this in the context of a simple memory game. The game involves two participants, a 'student' and a 'tutor', and 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". Although this task is semantically very limited, it can be entertaining, and the dynamics of the interactions are often non-trivial. Here we focused on the responses of the 'tutors', as a case study in high-quality interaction.

We recorded 41 such dialogs with different participants, and from these selected one person for further study; this was a person in the tutor role who was clearly an exemplary communicator: cheerful, supportive, involved and generally making the game fun. We then elicited 5 more dialogs with this person, and studied the ways in which he gave such a good impression.

We focused attention on the various ways in which he acknowledged correct answers. There was a lot of variety (our data was in Japanese, but analogous variety exists in English, which also has many ways to acknowledge, such as

with *yes*, *right*, *uh-huh*, *mm*, *yeah* and *okay*), and this variety did not appear to be random, rather it seemed that the tutor was being 'sensitive': paying close attention to the internal state of the speaker at each moment, and chosing his acknowledgements appropriately.

We modeled this with an algorithm for acknowledgment choice based on the user's internal state. Examples of the skills which the algorithm aims to emulate include: if the user is pleased, act pleased too; don't use variety for variety's sake, but also don't be mechanically repetitive; slow down the pace of the interaction when the student is having trouble; show approval when the student gets the right answer after being stuck; and be supportive when the student is unsure. These skills are operationalized by rules which use features of the context and from the prosody of the student's utterances to guide choice of the acknowledgement, as described elsewhere (Tsukahara 1998; Tsukahara 2000a; Tsukahara 2000b).

3 USER INSENSITIVITY

As a preliminary evaluation, we judged the algorithm in three ways. First, we confirmed that it produced results similar to those seen in the corpus data. Second, we confirmed that its choices agreed with our intuitions as to what good selections would be. Third, we found that conversations synthesized (by audio cut-and-paste) using acknowledgments chosen by the algorithm sounded more natural than conversations with randomly chosen acknowledgements, in the opinions of naive judges (Tsukahara 1998).

Having achieved success by these measures, we then tried out the algorithm in a realistic context, with naive users interacting with a WOZ (Wizard of Oz) system incorporating the algorithm. In keeping with the goal of achieving responsiveness, we made this a 'hard real time' system, able to respond to the user's utterances at the same swift pace that the exemplary human tutor did. In particular, the start of each acknowledgement came no later than 360 milliseconds after the end of the user's utterance, and this allowed the dialog to continue at a cycle time of as little as 1.6 seconds from one guess to the next. As a result, users were able to get completely involved in the game of recalling as many station names as possible as fast as possible.

In these situation we found, to our surprise, that there was no systematic preference: subjects seemed insensitive to the difference between acknowledgements that were produced by the algorithm as appropriate for the situation, and acknowledgements that were produced at random.

There are several likely reasons for this:

First, in this fast-paced situation, users had no attention to spare to consider whether the system's *yeah* really should have been an *okay*, or the like; they were too busy trying to recall station names. This is in contrast to the types of systems more common today, where the pace of interaction is much slower, and users are left with free time to contemplate the prompts and responses of the system.

Second, the acknowledgements were concise and therefore somewhat subtle. In a rigid turn-taking system, producing full sentence acknowledgements like *your answer was correct, beep* or *good, at last you got it, please keep it up, beep*, inappropriate acknowledgements would be much more salient.

Third, at the end of the interaction, users probably can't remember how they felt at each moment during the interaction. Maybe they were momentarily irritated or amused or pleased by the system at various times, but after a minute or two, when asked "how did you find the naturalness of the system", those impressions have probably been forgotten.

Fourth, the interactions were short overall, lasting only a minute or two. We think that extended use, say 5 to 10 minutes, would show clearer effects: the cumulative effects of minor awkward choices would probably accumulate and create an overall impression of being hard to talk to, or conversely, the cumulative effects of consistently saying just the right thing would lead to an overall impression of high-quality. However we were not able to do extended-use experiments, primarily because longer interactions would reveal the limitations of our set-up, destroying the user's illusion of being able to talk freely and be understood.

In response to these problems, we considered several variant evaluation methods. We considered telling users to pay attention to the acknowledgements, but this would have distracted them and probably changed their behavior. We considered asking users to think aloud as they interacted with the system, perhaps pausing it after each acknowledgement, but this would have destroyed the hard real time nature of the interaction, and thereby violated the presuppositions of our choice algorithm. We considered using third-party observers, to watch and listen to the subjects interacting with the system, either live or recorded, and to judge the quality of the interaction, but it is known that third-party observers' opinions of what constitutes a good interaction do not always agree with the opinions of participants.

4 EVALUATION BY RE-LISTENING

To avoid these problems, we introduced a 're-listening' phase, after the interaction, in which the user listens to a recording of his interaction. This allows the user to devote full attention to the task of evaluating system quality, while also judging the system with reference to his private knowledge of how he felt during each moment of the interaction. During re-listening we allow the user to stop or rewind the play-back at will.

We also give the user a partial transcript of his interaction with the system, for him to follow along as he listens. This must be computer-generated in order to be available immediately, before the user's impressions can fade. For utterances in which the designer is interested in the subject's opinion, the transcript includes a tiny 7-point scale, printed above the utterance, for the user to rate the appro-

priateness of that utterance. Figure 1 shows an example of a transcript.

5 EXPERIMENT PROCEDURE

We used the above evaluation method in experiments as follows.

We wanted to get users' impressions of the acknowledgements, but wanted to do so in a natural condition. Thus we did not tell them in advance the purpose of the experiment, just asked them to "use this system". Acknowledgements are of course useless in isolation, so we built a full system to play the role of the tutor in the memory game, including not only chosing when to acknowledge and which acknowledgment to use, but chosing when to give a hint, and which hint to give. Pilot studies had revealed, not surprisingly, that users are very sensitive to mis-recognitions, such that a failure to recognize a guess as correct or not dominates the user's impression of the system as a whole, making questions of acknowledgement choice, for example, fade into insignificance. We therefore used a wizard for the speech recognition function: the wizard's only role was to determine whether the guess was correct or incorrect (even this this required some training for the wizard to get up to speed, and he still made some mistakes; data from such runs were discarded). The system used this correct/incorrect information to determine whether to produce an acknowledgment or a hint, and, if an acknowledgement, used the context and the prosody of the user's utterance to chose a suitable item.

Most users believed they were interacting with a fully automatic system, and yet their behavior was, it seemed to us, as natural as if they were talking to a human. So we were happy with this setup.

As a control condition, we used a version of the system that chose acknowledgements at random, while preserving the frequencies of each acknowledgment seen in the corpus. We considered this a fair baseline, as the best model one could imagine that did not consider the context or the user's prosody.

Each user interacted with the full system and the random system for about 90 seconds each. The order of presentation was varied at random. After interacting with each system, we asked:

1. "Which computer would you like to use?"

2. "How would you rate the overall naturalness of the acknowledgements produced by the system?" for each system.

3. Then we ran the re-listening phase. During this phase we had users rate the naturalness of each acknowledgement, on a 7 point scale.

After this listening was complete, we then gave a questionnaire regarding impressions of the system as a whole.

4. "How would you rate each system on naturalness, friendliness . . . ?" (on a 7 point scale)

5. We asked question 1 again.

Although the results of the experiments are not directly relevant to the question of how to evaluate such systems, we will note in passing that they did provide evidence for our claim that responsiveness, specifically the subtle choice of acknowledgements attuned to the user's inferred internal state, is indeed valued by users, as seen in Table 1[1]. There also appeared to be individual differences: a consistent minority preferred random acknowledgements, and from the comments it seemed that at least some of these users would have been even happier with a more formal, mechanical, style of interaction, with less variation in the acknowledgements.

Table 1: Subjects Preferences

	Algorithmic Choice	Random Choice	
First Impression	6	7	n.s.
After Re-listening	10	3	$p < 0.05$

6 VALIDATION

Although the above experiment was not designed to validate re-listening as an evaluation method, some of the data gathered does suggest that re-listening does give sensitive and reliable measurements of user satisfaction.

First, preferences were more internally consistent, in that judgements of the appropriateness of specific responses correlated better with judgements of the usability of the system as a whole. The correlation with 'kindness' preferences is one example. Logically, one would expect the user's preference for a system to correlate with his perception of the kindness (*yasashisa* in Japanese) of that system. However this correlation was weak before re-listening (r=0.4 with p=0.16 by the U-test), and of the users who stated preferences for the algorithm-based system, two actually ranked it as less kind than the random system, as seen in the top graph in Figure 2. After re-listening, however, this discrepancy disappeared, as seen in the lower graph, and the correlation between preferences and judgements of kindness was as expected (r=0.71 with p=0.014 (significant), by the U-test).

Second, after re-listening, subjects volunteered more comments regarding the appropriateness of individual items (4/13 subjects before vs. 11/13 subjects after, $p < 0.05$).

Third, the results obtained after re-listening were consistent with the results of the non-real-time evaluations.

[1]While the results here are only just significant, there is collaborating evidence: in a preliminary experiment — identical in all respects other than that the hints were produced by the wizard, not automatically — 12 of 15 users preferred the system that did algorithm-based acknowledgement choice.

```
Direction: evaluate naturalness of each acknowledgements.
Example  : X++++++ "extremely unnatural",  ++X++++ "slightly unnatural", +++X+++ "neutral"
-------------------------------------------------------------------------------------------------------
YOU    :(Wrong Answer)          Takadanobaba      Mejiro           Ikebukuro            Ootsuka
SYSTEMR:              No  (Hint)          05Yes!          06yeah             07yeah              08umm
                                          +++++++         +++++++            +++++++             +++++++
-------------------------------------------------------------------------------------------------------
```

Figure 1: Sample Transcript Fragment

7 SUMMARY

Building spoken language systems that operate at near-human levels, able to motivate, charm, persuade, and amuse users, will doubtless require lots of attention to the 'little things' in dialog, individually minor, but in aggregate determining whether users find the system to be fun to use or just tolerable. These little things can easily escape the user's conscious attention: they are certainly far less salient than recognition errors.

We have shown that more sensitive judgements of system quality, including such little things, can be obtained by evaluation with re-listening, that is, having users use the system and then listen to a recording of their interaction while consulting a transcript of that interaction, thereby avoiding problems that arise due to attention limits.

We envision that this method will be useful for establishing and quantifying the value of various kinds of responsiveness, leading to general design guidelines, and will also be useful for evaluating and tuning various systems.

8 REFERENCES

Rich, Elaine (1999). Users are Individuals: Individualizing User Models. *International Journal of Human-Computer Studies*, 51:323–338.

Tsukahara, Wataru (1998). An Algorithm for Choosing Japanese Acknowledgments Using Prosodic Cues And Context. In *International Conference on Spoken Language Processing*, pp. 691–694.

Tsukahara, Wataru (2000a). *Choice of Acknowledgements based on Prosody and Context in a Responsive Spoken Dialog System (in Japanese)*. PhD thesis, University of Tokyo, School of Engineering.

Tsukahara, Wataru (2000b). Evaluating the Effectiveness of Subtle Choices in Acknowledgements in Japanese based on Prosodic Cues and Context (in Japanese). In *Information Processing Society of Japan, Spoken Language Processing Society, 34th Meeting*, pp. 57–62.

Walker, M. A., D. J. Litman, C. A. Kamm, & A. Abella (1998). Evaluating Spoken Dialog Agents with PARADISE: Two case studies. *Computer Speech and Language*, 12:317–348.

Ward, Nigel (1997). Responsiveness in Dialog and Priorities for Language Research. *Systems and Cybernetics*, 28(6):521–533.

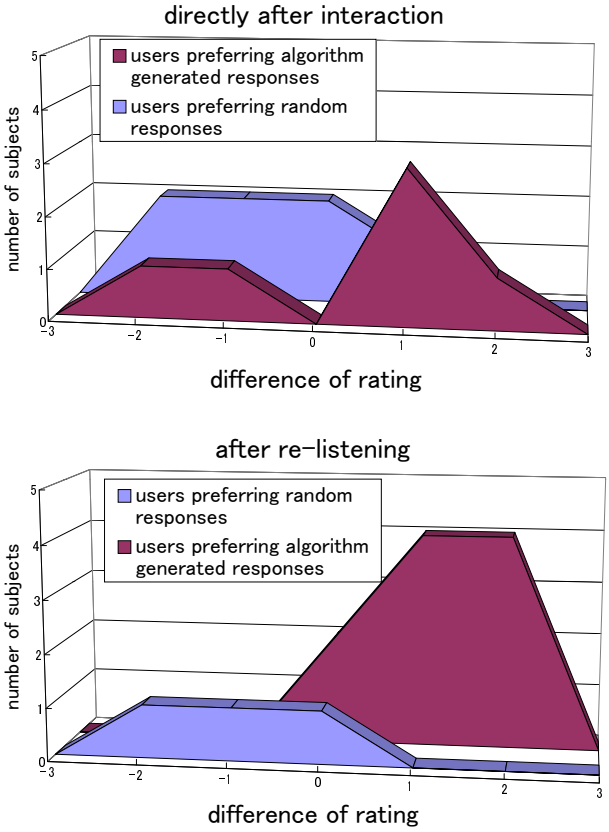

Figure 2: Histograms of the Differences Between the Rating of Kindness for the Algorithm-Based system and the Rating of Kindness of the Random System, indicating the varying ability of subjects to notice differences between the two