THE UTILITY OF ELAPSED TIME AS A USABILITY METRIC FOR SPOKEN DIALOGUE SYSTEMS

Marilyn Walker, Julie Boland, Candace Kamm

AT&T Labs—Research
180 Park Avenue
Florham Park, NJ 07932-0971 USA

ABSTRACT

It is commonly assumed that elapsed time is an important objective metric for evaluating the performance of spoken dialogue systems. However, our studies based on the PARADISE framework consistently find that other predictors are stronger contributors to user satisfaction than elapsed time. In this paper, we show that several possible explanations for this apparently counter-intuitive finding are not feasible. Our conclusion is that users of spoken dialogue systems are as much or more attuned to *qualitative aspects* of the interaction as they are to elapsed time.

1. INTRODUCTION

Evaluation of spoken dialogue systems is currently a very active area of research. As it becomes more feasible to build spoken dialogue systems for a range of tasks, the ability to evaluate and compare them is of utmost importance. Current approaches to evaluation are based on probing different aspects of a system's performance by collecting both subjective metrics, such as User Satisfaction, as well as objective metrics, such as task success, number of repairs, or elapsed time [10, 2, 1, 7, 9, 6]; we proposed PARADISE as a general framework for spoken dialogue evaluation with the goal of deriving a predictive model of user satisfaction as a function of objective metrics for task success, dialogue quality, and dialogue efficiency [4].

A common assumption of most current work on evaluation is that the elapsed time from the beginning to the end of a dialogue is an important objective metric [7, 6] *inter alia*.[1] Evaluation proposals under discussion for DARPA Communicator are considering collecting elapsed time and task success as the only objective metrics [14]. The intuition that users of dialogue systems are attuned to elapsed time predicts that when asked to rate two systems that differ in terms of elapsed time, users will rate the system with shorter elapsed time more highly. Furthermore, it has been suggested that measures of dialogue quality, such as the number of reprompts or the number of confirmations, are so highly correlated with elapsed time, that it is redundant to measure them.

However, in previous work applying PARADISE to the evaluation of spoken dialogue systems, we consistently find that elapsed time is either not a predictor of User Satisfaction or that it is less significant than other predictors [13, 8, 5]. When comparing a system-initiative with a mixed-initiative version of the ELVIS system for voice access to email, we found that users consistently rated the system-initiative version more highly, although mean elapsed time for this version was significantly longer than for the mixed initiative version [15]. Furthermore, our modeling procedures consistently select task success and various measures of dialogue quality as the most important predictors. Our explanation for this finding is that users have poor perception of elapsed time but are highly attuned to qualitative aspects of the dialogue. This paper will test several hypotheses for alternative explanations of this finding.

The first hypothesis (H1) that we consider follows from the posited redundancy of the dialogue quality measures.

- H1: Elapsed time is highly correlated with dialogue quality measures and can replace them as an important predictor of user satisfaction.

Hypothesis H2 is that elapsed time is only an important metric for those users who completed the experimental task successfully. Otherwise, a system could optimize the elapsed time metric by simply hanging up on the user.

- H2: Elapsed time is a more significant contributor to user satisfaction for those users who complete a dialogue task.

Hypotheses H3 and H4 are that these results are based on the current state of the art for ASR, or on the way in which we measured user satisfaction.

- H3: As recognizer performance improves, elapsed time will become a more significant predictor of user satisfaction.
- H4: Elapsed time is significantly correlated with usability, even given the current state of the art, but our method for measuring user satisfaction fails to expose this correlation.

Since there are many possible ways of measuring user satisfaction, we can obviously test only a weak version of H4.

In this paper, we first describe PARADISE in more detail. We then report a PARADISE evaluation of two different spoken dialogue systems providing voice access to email. This evaluation will show that, under the assumptions under which our previous evaluations were carried out, Elapsed Time on its own is not a strong predictor of user satisfaction. We then use the experimental corpus resulting from this evaluation to test the hypotheses put forth above.

2. PARADISE

PARADISE (PARAdigm for DIalogue System Evaluation) is a general method for spoken dialogue system evaluation, based on the model of performance in Figure 1. The model proposes that the system's primary objective is to maximize user satisfaction and that task success and various costs that can be associated with the dialogue are both contributors to user satisfaction. [4, 12, 8, 5]. The application of this

[1] This metric is also called Time-to-Completion, Dialogue Duration, or Transaction Time.

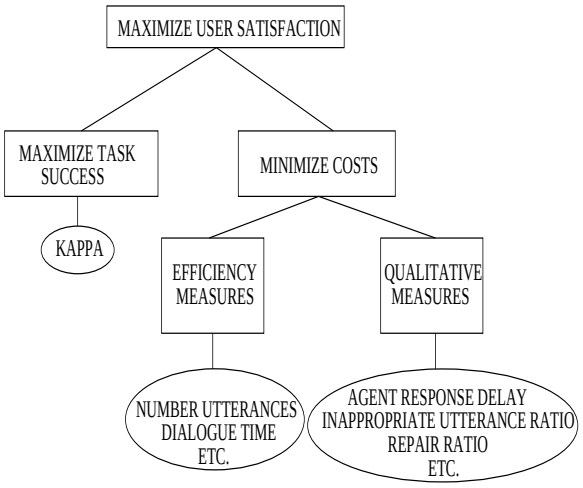

Figure 1: PARADISE's structure of objectives for spoken dialogue performance

model to a corpus of dialogues results in an objective performance function for a dialogue system that is a weighted linear combination of a task-based success metric and dialogue quality and efficiency metrics.

The motivation for PARADISE is to provide a means for (1) generating predictive models of User Satisfaction so that re-evaluation is not necessary after every change to a system; and (2) determining the contribution of various metrics to User Satisfaction.

The PARADISE performance function is derived by multivariate linear regression with user satisfaction as the dependent variable and task success, dialogue quality, and dialogue efficiency measures as independent variables. Any of the subjective and objective metrics proposed in earlier work can be used within PARADISE as either a measure of task success or a dialogue cost metric. Figure 1 illustrates some potential metrics for each of task success, dialogue efficiency and dialogue quality.

Applying PARADISE to dialogue data requires that dialogue corpora be collected via controlled experiments during which users subjectively rate their satisfaction. In addition, a number of other variables having to do with the costs of the interaction must be automatically logged by the system or hand-labelled. In the next section we describe the metrics that were logged and hand-labelled in the experiments discussed here.

3. EXPERIMENTAL SYSTEMS AND METHODS

3.1. Systems

We collected dialogue data in controlled experiments in which users completed three typical domain task scenarios in three different conversations with one of two systems:(1) ANNIE, a system that provides facilities for voice dialing, employee directory lookup and voice and email access [3]; and (2) ELVIS-SI, a system-initiative version of the ELVIS system for accessing email [15]. We also investigated the effect of a tutorial on user performance. Half of the users for each system completed a tutorial conversation with the system before starting the experimental tasks.

Each system was implemented using a general-purpose platform for phone-based spoken dialogue systems [3]. The platform consisted of a speech recognizer that supports barge-in, an audio server for both voice recordings and text-to-speech (TTS), an interface between the computer running the system and the telephone network, a module for application specific functions, and modules for specifying the application grammars and the dialogue manager.

A typical task scenario was: *You are working at home in the morning and plan to go directly to a meeting when you go into work. Kim said she would send you a message telling you where and when the meeting is. Find out* **the Meeting Time** *and* **the Meeting Place**.

To complete this task with ANNIE, users accessed their email by linearly navigating through an email folder, as in Dialogue D1:

(D1) A. Annie here.
U. Get my messages.
A. First message. This is a message from Michael about Evaluation Technical memo... (*user barges in*)
U. Next message.
A: Message 2. "This is a message from Kim about interviewing Antonio. I'd like to interview him. I could also go along to lunch. Kim." 2:16 P.M. Monday, March 23rd. What's next?
U: Next message.
A: Message 3. "This is a message from Kim about meeting today. The meeting this morning is at 10:00 in 2d 516. See you there, Kim." (*user barges in*)
U: I'm done here.

To complete this task with ELVIS-SI, users could select messages based on their content by specifying values from the sender or the subject field. Thus it was possible that users of ELVIS-SI could find the message required to complete the task above more efficiently. In ELVIS-SI, users were prompted for the selection criteria, as in Dialogue D2.

(D2) E: Hi, ELVIS here. You have 5 new and 0 unread messages in your inbox. Say Read, Summarize, or say Help for more options.
U: Read.
E: Select by Sender, Subject or in order of arrival?
U: Sender.
E: Which Sender?
U: Kim.
E: In the messages from Kim, there's 1 message about "Interviewing Antonio" and 1 message about "Meeting Today." The first message is titled, "Interviewing Antonio." (*user barges in*)
U: Next message.
A: "This is a message from Kim about meeting today. The meeting this morning is at 10:00 in 2d 516. See you there, Kim."
U: I'm done here.

In addition to differences in interface functionality, the underlying application server for the two systems was also different. In ANNIE, the email headers and messages were rendered in TTS and stored in the user's voice mailbox. In ELVIS, TTS was used to dynamically generate audio at the time that the user expressed interest in a particular email message. This meant that ELVIS-SI had a slower response time than ANNIE when playing messages. Another difference was that ANNIE's prompts were pre-recorded with a professional voice, with the dynamic content of email message headers and bodies rendered in TTS, while ELVIS used TTS for both prompts and reading email messages.

3.2. Metrics for PARADISE

The metrics used for the PARADISE analysis are a combination of task success, dialogue quality and efficiency measures extracted from (1) the recordings of the dialogues; (2) transcriptions of the recordings; (3) the system logs; and (4) a user survey collected via a web page form. Measures are described in **boldface** below.

The dialogue efficiency metrics were the Elapsed Time of the interaction (**ET**), and the number of User Turns (**Uturns**). The dialogue quality measures were Mean Recognition Score (**MRS**), and **Help%**.

Measure	ELVIS no tutorial	ELVIS tutorial	ANNIE no tutorial	ANNIE tutorial
Task Success (Comp)	.86	.86	.73	.96
User Turns	25.2	17.6	21.4	13.5
Elapsed Time (ET)	318 s	326 s	280 s	196 s
MeanRecog (MRS)	.89	.86	.67	.74
Help%	.02	.02	.14	.05
User Satisfaction	29.0	29.9	25.1	31.0

Table 1: Performance measure means per dialogue for different experimental conditions

Mean Recognition Score (**MRS**) was based on comparing the transcription with the recognizer output to calculate a concept accuracy measure for each utterance.[2] Mean concept accuracy was then calculated over the whole dialogue to produce the Mean Recognition Score for the dialogue. The **Help%** metric was derived by measuring the number of times that the system played one of its context specific help messages because it believed that the user had said *Help*. This number was then divided by the dialogue length to produce the normalized **Help%** measure.

The web page forms were used to calculate task success and User Satisfaction measures. Users reported whether they believed they had completed the task (**Comp**).[3] They also had to provide objective evidence that they had in fact completed the task by filling in a form with the information that they had acquired. To provide data on User Satisfaction, users completed the survey below.

- Was the system easy to understand in this conversation? (**TTS Performance**)
- In this conversation, did the system understand what you said? (**ASR Performance**)
- In this conversation, was it easy to find the message you wanted? (**Task Ease**)
- Was the pace of interaction with the system appropriate in this conversation? (**Interaction Pace**)
- In this conversation, did you know what you could say at each point of the dialogue? (**User Expertise**)
- How often was the system sluggish and slow to reply to you in this conversation? (**System Response**)
- Did the system work the way you expected it to in this conversation? (**Expected Behavior**)
- In this conversation, how did the system's voice interface compare to the touch-tone interface to voice mail? (**Comparable Interface**)
- From your current experience with using the system to get your email, do you think you'd use it regularly to access your mail when you are away from your desk? (**Future Use**)

The surveys had multiple choice responses ranging over values such as (*almost never, rarely, sometimes, often, almost always*). In order to produce a numerical score, each survey response was mapped into the range of 1 to 5, with 5 representing the most satisfaction. Then all the responses were summed, resulting in a **User Satisfaction** measure for each dialogue ranging from 8 to 40.

4. MODELING USER SATISFACTION WITH PARADISE

Table 1 summarizes the measures that we collected over the different experimental conditions. We performed a series of statistical tests to evaluate the statistical significance of the effects of system (ANNIE vs. ELVIS-SI) and tutorial (presence or absence) on these measures. There were no differences between ELVIS-SI and ANNIE in either overall satisfaction or task success, but there were reliable differences in some of the performance measures. ELVIS-SI had better ASR and ELVIS-SI users asked for help less often. On the other hand, ANNIE required less time and fewer turns. A tutorial resulted in fewer turns and higher satisfaction for both systems, but the effect of tutorial on satisfaction was stronger for ANNIE users. In addition, the ANNIE tutorial resulted in less time, fewer help requests, and a tendency for a higher rate of task success.

Measure	Correlation
Task Success (Comp)	.004
User Turns	.83
Mean Recog (MRS)	-.06
Help%	.09

Table 2: Correlations between Elapsed Time (ET) and other Metrics

Turning to our hypotheses, Hypothesis H1 posits that Elapsed Time is highly correlated with the dialogue quality measures, and that the dialogue quality measures can be replaced by Elapsed Time when deriving a performance model. Table 2 shows the correlation coefficients for Elapsed Time and the other objective metrics. Interestingly, the only other metric that Elapsed Time is highly correlated with is User Turns which is also an efficiency metric. Thus the first part of H1 is disconfirmed.

Since Elapsed Time is the metric of interest, and it is highly correlated with User Turns, we drop User Turns from further consideration. To see whether we can replace the dialogue quality measures with Elapsed Time, we first apply PARADISE to derive a model of User Satisfaction as a function of Comp, ET, MRS and Help%. We perform a multivariate linear regression in which User Satisfaction is the dependent variable and the other metrics are the independent variables. The only independent variables that are included in the derived models are those that are statistically significant predictors of the dependent variable, User Satisfaction. Because we normalize the factors before doing the regression, the magnitude of the coefficient directly indicates the contribution of its factor to predicting User Satisfaction [4]. The model derived from the combined ELVIS-SI and ANNIE data is given below.

$$\text{UserSat} = .24 * MRS + .29 * Comp - .14 * Help\% - .19 * ET$$

All of the independent variables are significant predictors of User Satisfaction, and the derived linear model suggests that Task Success (Comp) was the most important predictor, followed by MRS, ET and Help%. These factors together explain 34% of the variance in User Satisfaction at a statistically significant level.

Now to see whether Elapsed Time alone could be used as a predictor, we perform a regression of Elapsed Time alone against User

[2] For example, the utterance *Read my messages from Kim* contains two concepts, the *read* function, and the *sender:kim* selection criterion. If the system understood only that the user said *Read*, then concept accuracy would be .5.

[3] *Yes, No* responses are converted to *1, 0*, and null responses to the mean.

Satisfaction. The resulting model below shows that the correlation between Elapsed Time and User Satisfaction is -.23

$$\text{UserSat} = -.23 * ET$$

In contrast to the richer model given above which explains 34% of the variance in User Satisfaction, this model only explains 5% of the variance in User Satisfaction. Thus H1 is clearly disconfirmed, Elapsed Time is not redundant with dialogue quality metrics and cannot be used in place of them when deriving predictive performance models.

Data Subset	Correlation	Significance
All Dialogues	-.23	p = .006
Task Completion Dialogues	-.28	p = .002
Good ASR Dialogues	-0.01	p = .9 ns

Table 3: Correlations between Elapsed Time and User Satisfaction for Different User Groups

Hypothesis H2 posits that Elapsed Time is a more significant contributor to User Satisfaction for those users who complete a dialogue task. In this experiment, users completed the task in 113 out of the 148 dialogues (the Task Completion Dialogues shown in Table 3). H2 is supported by the correlations shown in Table 3 where there is an increase in both the significance of the correlation and the correlation coefficient for the subset of Task Completion dialogues.

Hypothesis H3 posits that as recognizer performance improves, elapsed time will become a more significant predictor of user satisfaction. This hypothesis is especially plausible since Mean Recognition Score is always a significant predictor of User Satisfaction. To explore this hypothesis we examined the 40 dialogues (out of 148) where the Mean Recognition Score was greater than 0.9. These are referred to as the Good ASR Dialogues in Table 3. H3 is disconfirmed by the correlations shown in Table 3: Elapsed Time is not significantly correlated with User Satisfaction in these dialogues.

Hypothesis H4 is that the survey that we use to measure User Satisfaction (see Section 2.2), fails to expose the relationship between Elapsed Time and User Satisfaction. Since there are many possible ways to measure User Satisfaction, we explore this hypothesis in a limited way by examining the correlation between Elapsed Time and the individual component questions of the survey. This analysis shows that the the only significant correlations are those between Elapsed Time and Task Ease (.25)[4] and Elapsed Time and Comparable Interface (.15). Since Task Ease provides the highest correlation, consider the hypothetical situation in which the Task Ease question is the only measure of User Satisfaction. The regression model below shows that the contribution of Elapsed Time to User Satisfaction is greater under these conditions.

$$\text{TaskEase} = .21 * MRS + .50 * Comp - .35 * ET - .15 * Help\%$$

However, task success (Comp) is stll a more important predictor, and dialogue quality measures such as Mean Recognition Score (MRS) and Help% still remain significant predictors. Thus even if we maximize the relationship between Elapsed Time and User Satisfaction, we find that metrics for dialogue quality and task success increase the ability of the derived models to predict User Satisfaction.

[4] And note that this correlation is only .02 larger than the correlation between Elapsed Time and User Satisfaction in Table 3 for all dialogues.

5. DISCUSSION AND FUTURE WORK

To our knowledge, this is the first detailed examination of the relationship between elapsed time and subjective metrics for dialogue system evaluation. However, in related work comparing interfaces with different modalities, Rudnicky found that elapsed time was not a good predictor of which type of interface users would choose [11].

Here, we explored several hypotheses as to the conditions under which elapsed time is an important predictor of user satisfaction. We found that elapsed time cannot be used to replace dialogue quality metrics, that the importance of elapsed time increases for dialogues where users completed the task, but that contrary to our expectations, it does not increase as recognition performance increases. We also showed that even under conditions in which the user satisfaction metric is biased towards finding a relationship with elapsed time, dialogue quality and task success metrics still remain significant predictors of user satisfaction. In future work, we hope to examine these relationships over a larger corpus of dialogues representing many different systems.

6. REFERENCES

[1] M. Danieli and E. Gerbino. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 34–39, 1995.

[2] L. Hirschman and C. Pao. The cost of errors in a spoken language system. In *Proceedings of the Third European Conference on Speech Communication and Technology*, pages 1419–1422, 1993.

[3] C. Kamm, S. Narayanan, D. Dutton, and R. Ritenour. Evaluating spoken dialog systems for telecommunication services. In *5th European Conference on Speech Technology and Communication, EUROSPEECH 97*, 1997.

[4] C. A. Kamm and M. A. Walker. Design and evaluation of spoken dialog systems. In *Proceedings of the ASRU Workshop*, 1997.

[5] C. Kamm, D. Litman, and M. A. Walker. From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP98*, 1998.

[6] L. B. Larsen. Combining objective and subjective data in evaluation of spoken dialogues. In *ESCA Workshop on Interactive Dialogue in Multi-Modal Systems*, pages 89–92, 1999.

[7] E. Levin and R. Pieraccini. A stochastic model of computer-human interaction for learning dialogue strategies. In *EUROSPEECH 97*, 1997.

[8] D. J. Litman, S. Pan, and M. A. Walker. Evaluating Response Strategies in a Web-Based Spoken Dialogue Agent. In *Proceedings of ACL/COLING 98: 36th Annual Meeting of the Association of Computational Linguistics*, pages 780–787, 1998.

[9] J. Polifroni, S. Seneff, J. Glass, and T.J. Hazen. Evaluation methodology for a telephone-based conversational system. In *Proc. First International Conference on Language Resources and Evaluation, Granada, Spain*, pages pp. 42–50, 1998.

[10] P. Price, L. Hirschman, E. Shriberg, and E. Wade. Subject-based evaluation measures for interactive spoken language systems. In *Proceedings of the DARPA Speech and NL Workshop*, pages 34–39, 1992.

[11] A. I. Rudnicky. Factors affecting choice of speech over keyboard and mouse in a simple data-retrieval task. In *EUROSPEECH93*, 1993.

[12] M. A. Walker, D. Litman, C. A. Kamm, and A. Abella. PARADISE: A general framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, ACL/EACL 97*, pages 271–280, 1997.

[13] M. Walker, D. Hindle, J. Fromer, G. Di Fabbrizio, and C. Mestel. Evaluating competing agent strategies for a voice email agent. In *Proceedings of the European Conference on Speech Communication and Technology, EUROSPEECH97*, 1997.

[14] M. Walker and L. Hirschman. Darpa communicator evaluation proposal. September,1999.

[15] M. A. Walker, J. C. Fromer, and S. Narayanan. Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics, COLING/ACL 98*, pages 1345–1352, 1998.