Human-Machine Problem Solving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction

Elizabeth Shriberg[1,3], *Elizabeth Wade*[2,3], *Patti Price*[3]

[1]University of California at Berkeley, Department of Psychology, Berkeley, CA 94720
[2]Stanford University, Department of Psychology, Stanford, CA 94305
[3]SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94303

ABSTRACT

We have analyzed three factors affecting user satisfaction and system performance using an SLS implemented in the ATIS domain. We have found that: (1) trade-offs between speed and accuracy have different implications for user satisfaction; (2) recognition performance improves over time, at least in part because of a reduction in sentence perplexity; and (3) hyperarticulation increases recognition errors, and while instructions can reduce this behavior, they do not result in improved recognition performance. We conclude that while users may adapt to some aspects of an SLS, certain types of user behavior may require technological solutions.

1. INTRODUCTION

Data collection is a critical component of the DARPA Spoken Language Systems (SLS) program. Data are crucial not only for system training, development and evaluation, but also for analyses that can provide insight to guide future research and development. By observing users interacting with an SLS under different conditions, we can assess which issues may best be addressed by human factors and which will require technological solutions. System developers can benefit from considering not only initial use of an SLS, but also the experience of a user over time.

Systems based on current technology work best when speech and language closely resemble the training data used to develop the system. However, there is considerable variability in the degree to which the speech and language of new users match that of the training data. The current paper examines the importance of this initial match. It is possible that users whose speech does not conform to the system may be able to adapt their behavior over time (e.g., Stern and Rudnicky [11]). In order to evaluate technology in terms of the demands of the application, we need to understand the extent and the nature of such adaptation and the conditions that affect it. Although system performance can be measured in a number of ways, in this paper, we focus on (1) self-reports of user satisfaction, and (2) recognition performance. Further studies could include additional measures.

SRI has been collecting data in the air travel planning domain using a number of different systems (see Bly et al. [1]; Kowtko and Price [5]). In moving from wizard-based data collection to the use of SRI's SLS, we observed changes in user behavior that were associated with system errors. Some of these behaviors were adaptive; for example, learning to avoid out-of-vocabulary words or unusual syntax should facilitate successful interaction. Other behaviors, however, were non-adaptive and could actually impede the interaction. For example, speaking more loudly or in a hyperarticulate style may be detrimental to system performance insofar as these styles differ from those observed in training material dominated by wizard-mediated data in which system errors are minimal.

It is difficult to predict how well an SLS will need to perform in order to be acceptable to users. Both speed and accuracy are crucial to system acceptability; we have therefore collected data using versions of the system that prioritize one of these parameters at the expense of the other. The present study first addresses the issue of user satisfaction with different levels of system speed and accuracy and then focuses on an example of an adaptive behavior and another that is maladaptive. These behaviors represent a subset of potential factors influencing human-machine interaction. Because these issues are not restricted to any particular system, they should be of general interest to developers of SLS technology.

In the first study, we compared three points in the speed-accuracy space for this application: (1) an extremely slow but very accurate wizard-mediated system (described in Bly et al. [1]) with a 2-3 minute response time and a minimal error rate; (2) a software version of the DECIPHER recognizer with a response time of several times real time and a fairly low word error rate; and (3) a version of the DECIPHER recognizer implemented in special-purpose hardware using older word models, which has a very fast response time but currently has a higher word error rate.

We compared user satisfaction based on responses to a post-session questionnaire.

The second study investigated the effect of user experience on syntax and word choice. We hypothesized that one way users might adapt would be to conform to the language models constraining recognition. We therefore measured recognition performance in subjects' first and second scenarios, and compared sentence perplexities in order to determine whether any changes in recognition performance could be attributed to a change in perplexity.

The third study examined the effect of hyperarticulate speech on recognition and tested whether instructions to users could reduce this potentially maladaptive behavior. We coded each utterance for hyperarticulation and compared recognizer performance for normal and hyperarticulate utterances. We also compared rates of hyperarticulation for subjects who were either given or not given the instructions.

2. DATA COLLECTION METHODS

2.1. Subjects

Data from a total of 145 subjects were included in the analyses. Subsets of these data were chosen for inclusion in each analysis in order to counterbalance for gender and scenario. The majority of subjects were SRI employees recruited from an advertisement in an internal newsletter; a small number were students from a nearby university, employees in a local research corporation, or members of a volunteer organization. Subjects were native speakers of English, ranged in age from 22 to 71 and had varying degrees of experience with travel planning and computers.

2.2. Materials

Four different travel-planning scenarios were used. One entailed arranging flights to two cities in three days; a second entailed finding two fares for the price of a first class fare; a third required coordinating the arrival times of three flights from different cities; and a fourth involved weighing factors such as fares and meals in order to choose between two flight times. Because the task demands of the scenarios were different, we controlled for scenario in the analyses.

2.3. Apparatus

The data were collected using two versions of SRI's SLS (with no human in the loop); the first study also included data collected in a Wizard of Oz setting (Bly et al. [1]). The basic characteristics of the DECIPHER speech recognition component are described in Murveit et al.[7,9], and the basic characteristics of the natural language understanding component are described in Jackson et al. [4]. Some subjects used the real-time hardware version of the DECIPHER system (Murveit and Weintraub [8]; Weintraub et al. [12]); others used the software version of the system, which was a modified version of SRI's benchmark system (as described in the references above) tuned using the pruning threshold to improve speed at the cost of introducing a small number of recognition errors.

SRI's SLS technology was implemented in the air travel planning domain, a domain with which many people are familiar (see Price [10]). The underlying database was a relational version of an 11-city subset of the Official Airline Guide. Two DARPA/NIST standard microphones were used: the Sennheiser HMD-410 close-talking microphone and the Crown PCC-160 table-top microphone. Most data were collected with two channels; some of the early data were collected using only the Sennheiser microphone. When both microphones were used, recognition was based on the Sennheiser input.

The interface presented the user with a screen showing a large button labeled "Click Here to Talk." A mouse click on this button caused the system to capture speech starting a half second before the click; the system automatically determined when the speaker finished speaking based on silence duration set at a threshold of two seconds. The user could move to the context of previous questions via mouse clicks. Once the speech was processed, the screen displayed the recognized string of words, a "paraphrase" of the system's understanding of the request, and, where appropriate, a formatted table of data containing the answer to the query. In cases where the natural language component could not arrive at a reasonable answer, a message window appeared displaying one of a small number of error messages. A log file was automatically created, containing time-stamps marking each action by the user and by the system.

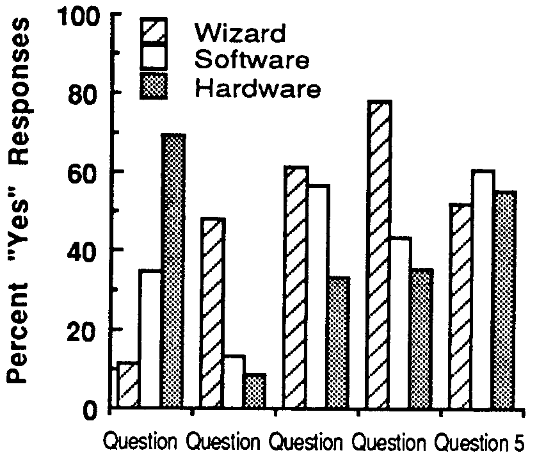
2.4. Procedure

Subjects were seated in a quiet room in front of a color monitor, and had use of a mouse and microphone(s) but no keyboard. They were given a short demonstration on how to use the system. Some of the subjects were given additional instructions explaining that, while they might have a tendency to enunciate more clearly in the face of recognition errors, they should try to speak naturally, since the system was not trained on overenunciated or separated speech. Once subjects were comfortable with the system, they were left alone in the room while they solved travel planning scenarios. After they finished as many scenarios as possible within an hour, they were asked to fill out a questionnaire and were given a choice of gift certificate for use at a local bookstore or a contribution to a charitable institution.

3. EXPERIMENTS

3.1. The Effects of Speed and Accuracy Tradeoffs on User Satisfaction

Since in general, speech understanding systems can trade accuracy for speed, we first assessed how these parameters might affect user behavior and acceptance of the system. The software version of the recognizer was slower than the hardware version (2.5 compared to 0.42 times the utterance duration), but was substantially more accurate (with a word error rate of 16.1% as compared with 24.8% on the same sound files).



1. Were the answers provided quickly enough?
2. Did the system understand your requests the first time?
3. I focused most of my attention on solving the problems, rather than trying to make the system understand me.
4. Do you think a person unfamiliar with computers could use the system easily?
5. Would you prefer this method to looking up the information in a book?

Figure 1: User Satisfaction

To assess user satisfaction, we compared questionnaire responses for 46 subjects who used the hardware, 23 who used the software, and 46 who used the earlier wizard-mediated system. Mean responses are shown in Figure 1. In general, user satisfaction with the speed of the system correlated with the response time of the system they used; when asked, "Were the answers provided quickly enough?" 69.6% of the hardware users responded "Yes." In contrast, only 34.8% of the software users and a mere 11.1% of the wizard-system users gave "Yes" responses, a significant difference from the hardware result, χ^2 (df=4) = 35.6, $p < .001$. Although hardware users were pleased with the speed of the system; they were less likely than wizard system and software users to say they focused their attention on solving the problem rather than on trying to make the system understand them (33.3% as compared with 61.4% and 56.5%, respectively), a marginally significant effect, χ^2 (df=4) = 7.8, $p < .10$.

On several other measures users found the wizard-based system preferable to either the software or the hardware. More wizard-system users said that the system usually understood them the first time (47.8% as compared with 13.0% and 8.7% for the software and hardware users, respectively), χ^2(df=4) = 22.5, $p < .001$. Overall, the wizard system users were more likely to say the system could be easily used by a person who was unfamiliar with computers (78% compared with 43.5% and 35.6% for the software and hardware, respectively) χ^2 (df=4) = 20.5, $p < .001$. However, in terms of general satisfaction, as expressed in whether the subjects said they would prefer using the system to looking the information up in a book, there was no significant difference between the groups, with 52.3%, 60.9% and 55.6% "Yes" answers for the three groups respectively.

Because the hardware system was least satisfying to users in terms of recognition accuracy, we concluded that the hardware would provide the greatest potential for user adaptation to the system. For this reason, we used the hardware system to collect data on the effects of user experience and instructions regarding hyperarticulation.

3.2. Effect of User Experience on Recognition

User experience was evaluated in a within-subjects design, counterbalanced for scenario, that compared 24 users' first and second sessions. As a global measure of adaptation, we looked at how long it took subjects to complete their two scenarios. Although subjects were not told to solve the scenarios as quickly as possible, they nevertheless took less time (10.5 compared to 13.0 minutes) to complete their second scenarios, F(1,23) = 5.78, $p < .05$. This difference was partially but not completely attributable to a lower number of total utterances in the second scenario.

The users also elicited fewer recognition errors in the second scenario. The mean word error rate was 20.4% for the first scenario but fell to 16.1% for the second, F(1,22) = 5.60, $p < .05$. However, not all users decreased their recognition error rate. There was a significant interaction between initial error rate and change in error rate from the first scenario to the second, F(1,22) = 10.98, $p < .01$. Subjects who had recognition error rates of 20% or worse in the first scenario (N=11) tended to improve recognition performance, while subjects who had better initial performance (N=13) did not (Figure 2). Subjects with initial error rates of 20% or higher went from an average of 31.3% errors down to 19.6%, while subjects with initially lower error rates showed no statistically significant change. For those subjects who did improve recognition performance, the improvement could only be due to user adaptation, since the same SLS version was used for both scenarios.

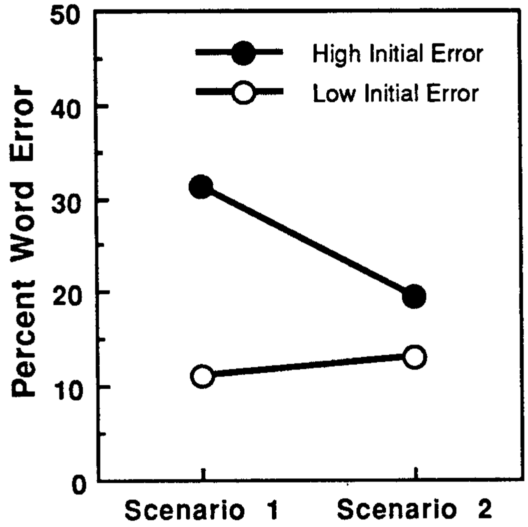


Figure 2: Recognition accuracy over time.

The improvement in recognition may be due in part to user adaptation to the language models used. As a measure of deviation from the system's language models, we used test-set perplexity, which was based on the bigram probabilities of the observed word sequences. As would be expected, there was a significant, positive average correlation between utterance word error and perplexity: mean $r = .28$, $t = 4.55$, $p < .001$. Thus, one way for subjects to improve recognition accuracy would be to change their language to conform to that of the system model. Perplexity may therefore play a role in the decrease in recognition error rates observed over time for those subjects who had an error rate of 20% or worse in their first scenario. For this group of subjects, there was a tendency to produce queries with lower sentence perplexity in the second scenario (Figure 3). Using the median as a measure of central tendency (a more stable measure due to the inherent positive skew of perplexity), we found that the average median sentence perplexity was 25.3 for the first scenario and 19.4 for the second, a reliable difference, $F(1,10) = 7.44$, $p < .05$.

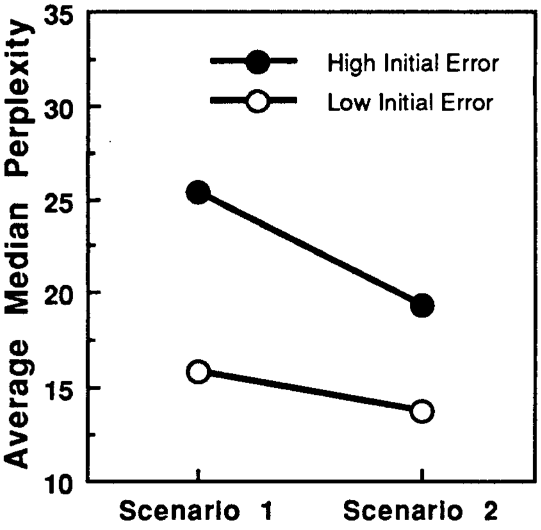


Figure 3: Median perplexity over time.

In addition to decreasing perplexity, subjects who had initial error rates of greater than 20% also tended to decrease the use of out-of-vocabulary words in the second scenario, whereas subjects who had lower error rates did not, a significant interaction, $F(1,22) = 6.10$, $p < .05$. Overall, however, the use of out-of-vocabulary words was rare.

These findings indicate that at least to some degree, subjects adapted to the language models of the system and, in doing so, managed to improve the recognizer's performance. Quite possibly, subjects were finding ways to phrase their queries that produced successful answers, and then reproducing these phrases in subsequent queries. In future work, further analyses (for example, looking at dialogue) will address this issue in greater detail.

3.3. Effect of Instructions on Speech Style

Another potential source of recognition errors arises when the speech of the user deviates from the acoustic models of the system. Since the vast majority of the data used to train the DECIPHER recognizer came from wizard-mediated data collection [6], where recognition performance was nearly perfect, examples of "frustrated" speech were rare. In human-human interaction, when an addressee (such as a foreigner) has difficulty understanding, speakers change their speech style to enunciate more clearly than usual (Ferguson [3]). We suspected that a similar effect might occur for people speaking to a machine that displayed feedback showing less than perfect understanding. We noticed that, when using an SLS as opposed to a wizard-mediated system, subjects tended to hyperarticulate: releasing stops, emphasizing initial word segments, pausing between words, and increasing vocal effort.

Although hyperarticulation is a multifaceted behavior, it was nevertheless possible to make global judgments about individual utterances. Hyperarticulation was coded for each utterance on a three-point scale by listening to the utterances. Utterances were coded as (1) clearly natural sounding, (2) strongly hyperarticulated, or (3) somewhat hyperarticulated. The coding was done blindly without reference to session context or system performance.

Using a within-subjects design, so that any differences in recognition performance could be attributed to a change in speech style, rather than speaker effects, we analyzed the speech style of 24 subjects' first scenarios (future analyses will also examine repeat scenarios). These subjects (of whom 20 were also included in the previous analysis of user experience) all used the hardware system. The subjects averaged about 10 natural sounding, 4 somewhat hyperarticulate, and 5 strongly hyperarticulate utterances each. For the 13 subjects who had at least three natural and three strongly hyperarticulated utterances, we compared recognition performance within subjects and found that the strongly hyperarticulate utterances resulted in higher word error rates, $F(1,12) = 5.19$, $p < .05$.

Hyperarticulation was reduced, however, by giving users instructions not to "overenunciate" and by explaining that the system was trained on "normal" speech. We calculated a hyperarticulation score for each subject by weighting "strongly hyperarticulated" utterances as 1, "somewhat hyperarticulated" utterances as 0.5, and "nonhyperarticulated" utterances as 0, and taking the mean weight across all utterances in the scenario. The 12 subjects who heard the instructions (the "instruction group") had lower mean hyperarticulation scores, 0.22 as compared with 0.60 for the 12 subjects who received no special instructions (the "no instruction group"), a significant difference $F(1,22) = 11.97$, $p < .01$.

Given that the instruction group had significantly fewer hyperarticulated utterances, and given that hyperarticulation is associated with lower recognition accuracy, we would expect the instruction group to have better recognition performance overall. However, although the trend was in that direction (18.1% word error for the instruction group versus 22.5% for the no-instruction group), the difference was not reliable. One possible explanation is a lack of power in the analysis, as a result of the small number of subjects and large individual differences in error rates. A second, not necessarily conflicting explanation is that the subjects given the instructions to "speak naturally" used somewhat less planned and less formal speech. We noticed that these subjects tended to have more spontaneous speech effects, such as verbal deletions, word fragments, lengthenings and filled pauses. Overall, spontaneous speech effects occurred in 15% of the 232 utterances for the instruction group, compared with 10% for the 229 utterances for the no-instruction group. Although these baseline rates are low, they may nevertheless have contributed to poorer recognition rates (see Butzberger et al. [2]). They may also be indicative of subtle speech style differences between the two groups not captured by the coding of hyperarticulation.

4. CONCLUSION

Application development can benefit from analyses of factors affecting system performance and user satisfaction. We have presented examples of ways in which the behavior and satisfaction of subjects interacting with an SLS may be affected. We have described ways in which parameters of the system itself, such as speed and accuracy, affect different aspects of user satisfaction. We have examined the effect of user experience on recognition performance and found a decrease in word error rate over repeated scenarios. Adaptation was relatively greater for those subjects who had more than 20% errors on the first scenario. The decrease in errors could be attributed at least in part to a decrease in sentence perplexity and to a reduction in the use of out-of-vocabulary words. We have also shown a significant relationship between word error rates and hyperarticulation, a speech style that occurs relatively frequently with an imperfect recognizer. We have shown that instructions not to hyperarticulate reduced this maladaptive speech style, but that instructions did not result in improved recognition performance overall.

Our studies have shown that along some dimensions, humans are flexible and can adapt in ways that improve system performance. However, hyperarticulation may be a maladaptive behavior for which a technological solution should be investigated. In particular we have found that strategies people use to try to improve normal human communication (e.g., hyperarticulation) can have the reverse effect in the context of our current models. While hyperarticulation is an "exaggerated" speech style that might improve comprehension for humans, it can cause poor recognition for automatic systems in which "exaggeration" is not adequately modeled.

Acknowledgments

We gratefully acknowledge support for this work from DARPA through the Office of Naval Research contract N00014-90-C-0085. The government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the government funding agencies. We also gratefully acknowledge Steven Tepper for software development.

REFERENCES

1. Bly, B., P. Price, S. Tepper, E. Jackson, and V. Abrash, "Designing the Human Machine Interface in the ATIS Domain," *Proc. Third DARPA Speech and Language Workshop*, pp. 136-140, Hidden Valley, PA, June 1990.

2. Butzberger, J. W., H. Murveit, E. Shriberg, P. Price, "Spontaneous Effects in Large Vocabulary Speech Recognition Applications," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

3. Ferguson, C. "Towards a Characterization of English Foreigner Talk," *Anthropological Linguistics*, 17, pp 1-14, 1975.

4. Jackson, E., D. Appelt, J. Bear, R. Moore, A. Podlozny, "A Template Matcher for Robust NL Interpretation," *Proc. DARPA Speech and Natural Language Workshop*, P. Price (ed.), Morgan Kaufmann, 1991.

5. Kowtko, J. C. and P. J. Price, "Data Collection and Analysis in the Air Travel Planning Domain," *Proc. Second DARPA Speech and Language Workshop*, pp. 119-125, Harwichport, MA, October 1989.

6. MADCOW, "Multi-Site Data Collection for a Spoken Language System," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

7. Murveit, H., J. Butzberger, and M. Weintraub,"Speech Recognition in SRI's Resource Management and ATIS Systems," *Proc. DARPA Speech and Natural Language Workshop*, P. Price (ed.), Morgan Kaufmann, 1991.

8. Murveit, H. and M. Weintraub, "Real-Time Speech Recognition System," *Proc. DARPA Speech and Natural Language Workshop*, P. Price (ed.), Morgan Kaufmann, 1991.

9. Murveit, H., J. Butzberger, and M. Weintraub, "Performance of SRI's Decipher Speech Recognition System on DARPA's ATIS Task," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.

10. Price P., "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc. Third DARPA Speech and Language Workshop*, pp. 91-95, Hidden Valley, PA , June 1990.

11. Stern, R. M. and A. I. Rudnicky, "Spoken-Language Workstations in the Office Environment," *Proc. Speech Tech' 90,* Media Dimensions, 1990.

12. Weintraub, M., G. Chen, P. Mankoski, H. Murveit, A. Stolzle, S. Narayanaswamy, R. Yu, B. Richards, M. Srivastava, J. Rabay, R. Broderson, "The SRI/UCB Real-Time Speech Recognition System," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992.