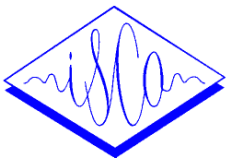

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

USER BEHAVIORS AFFECTING SPEECH RECOGNITION

Elizabeth Wade, Elizabeth Shriberg, Patti Price

SRI International
333 Ravenswood Avenue, Menlo Park, CA 94025 USA

ABSTRACT

We attempt to explain a decrease in recognition word error rate observed when users interacted over time with a spoken language system. We found no change in the language used (as measured by sentence perplexity), and only a small decrease in the number of out-of-vocabulary words. However, a behavior adversely affecting recognition, hyperarticulation, decreased over time. In addition, the acoustic match of hyperarticulated utterances to the system models also improved over time. We conclude that improvement in recognition was due to changes in speech rather than in language.

I. INTRODUCTION

Changes in the way users speak as they interact with a spoken language system over time may have consequences for recognition performance. Because humans are highly adaptive, initial recognition performance may not accurately predict later performance. System developers can benefit from considering not only initial use of a system, but also experience of a user over time. In addition, speakers interacting with a spoken language system may not exhibit the same language behavior observed in training data. Earlier, we found that recognition errors decreased as subjects interacted with the system over time [1]; the current paper more closely examines the source of this error reduction by looking at both the language and speech style of users.

Analyses were based on data collected using SRI's spoken language system (SLS), as part of a multisite collection effort [2] in which subjects solved air-travel planning scenarios. The SRI SLS combines the DECIPHER™ recognizer [3] with a robust natural-language understanding component [4], implemented in the air-travel planning domain. The system does not prompt the user for specific input; it simply accepts user-formulated queries. For example, the user might ask, "Show me flights from San Francisco to Philadelphia during the morning," to which the system should respond by displaying a table of flight information fitting those specifications.

In a previous paper [1] we reported that subject's word error rates decreased from Scenario 1 to Scenario 2. In that analysis we attempted to explain the source of this decrease; however, the addition of data in the current paper allows us to explain the phenomenon in further detail. We examine two potential causes for the decrease in error: changes in language and changes in speech style.

One possible explanation for the decrease in error is that users were changing their language to use more constructions of the types most easily recognized by the system. To test this hypothesis, we compared the perplexity of sentences in Scenarios 1 and 2 for each subject. If perplexity (essentially a measure of how unexpected a word sequence is given the system models) decreased in Scenario 2, we could conclude that subjects' behavior changed in a way that adapted to the language models of the system.

A second not contradictory hypothesis is that subjects were changing their speech style over time to better match the system's acoustic models. We coded and measured one speech style, hyperarticulation, which we had reason to believe would lead to recognition errors. If hyperarticulation was related to errors, and if the frequency of hyperarticulation decreased in Scenario 2, we could conclude that subjects' behavior changed in a way that adapted to the acoustic models of the system.

II. METHOD

2.1. Subjects

We collected speech and session logs for two scenarios from each of 24 subjects, counterbalancing for the selection and order of the scenarios they solved. The majority of subjects (17) were SRI employees recruited from an advertisement in an internal newsletter; a small number were students from a nearby university 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 involved arranging flights to two cities in three days; a second involved 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 SRI's Spoken Language System with no human in the loop. The basic characteristics of the DECIPHER™ speech recognition component are described in Murveit et al. [5,6], and the basic characteristics of the natural language understanding component are described in Jackson et al. [4]. The subjects used the real-time hardware version of the DECIPHER™ system which had a vocabulary size of 1,250 words [3,7].

SRI's SLS technology was implemented in the air travel planning domain, with which many people are familiar (see Price, [8]). The underlying database was a relational version of an 11-city subset of the *Official Airline Guide*. Recognition was based on the input of a Sennheiser HMD close-talking microphone.

The interface presented the user with a screen showing a button labeled "Click Here to Talk." A mouse click in this box caused the system to capture speech starting a 1/2 second before the click; the system automatically determined when

10.21437/ICSLP.1992-290

the speaker finished speaking based on silence duration set at a threshold of 2 seconds. Once the speech was processed, the screen displayed the words recognized, a "paraphrase" of the system's understanding of the request, and, where appropriate, a formatted table of data containing the answer to the query. When 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 and were given a short demonstration on how to use the system. Half 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 to solve the scenarios.

III. ADAPTATION

We compared Scenarios 1 and 2 for each subject to determine whether there were any changes in user behavior over time. 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 the second scenarios, $F(1,23) = 5.78$, $p < .05$. This difference was partially attributable to a lower number of total utterances in Scenario 2. In addition, we found significantly lower recognition error rates in subjects' second scenario. The mean word error rate was 20.4% for Scenario 1, but fell to 16.1% for Scenario 2, $F(1,22) = 5.60$, $p < .05$.

3.1. Language

We first hypothesized that this change in error rates might be due in part to adaptation to the language model of the recognizer. As a measure of deviation from the system's bigram language models, we used testset perplexity, which was based on the bigram probabilities of the observed word sequences. Perplexity measures the average likelihood (according to the system's models) that each word in a user's query will be followed by the next word, taking into account the base rate frequencies of the words. So a commonly phrased query like "I'd like to fly from San Francisco to Philadelphia" would have a low perplexity, since the system models would predict that each word is quite likely to follow the word that precedes it.

We confirmed the relationship between perplexity and word error in our data; there was a significant, positive average correlation between utterance word error and utterance 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 the language models of the system. For example, subjects might alter their initial language to use more common, easily recognized word sequences and to avoid rarer sequences that might tend to have more errors. However, we did not find support for this hypothesis. Perplexity decreased only slightly from 1 to 2 Scenario, with a geometric mean of 17.7 and 16.9, respectively. The magnitude of this difference was not significant given the variability both within and across subjects; perplexity within a scenario ranged from 8.8 to 38.9. The difference was nonsignificant by a Sign test, $p > .50$.

In an attempt to find converging evidence that changes in perplexity did not cause the decreased error rates, we obtained recognition results for the same sound files using a software version of the recognizer and two types of models. The bigram models were essentially the same as the original models used by the hardware, and were used as a control. The nogram models used only acoustic and word frequency information for recognition; they did not reflect any information about cross-word probabilities. Thus the recognition results from nogram models would not be improved by any user adaptation to the grammar of the original recognizer, whereas the results from the bigram models would. If the bigram results showed a greater decrease in word error than the nogram results, we could conclude that some of the decrease was due to adaptation to the language models. Figure 1 shows the word error recognition results from the two types of models. Both bigram and nogram results show essentially the same decreasing slope. This again suggests that adaptation to the language models was not a major cause of improved recognition over time.

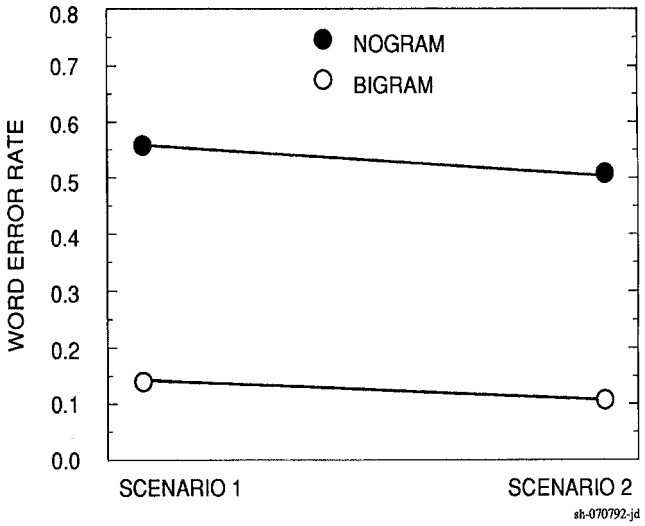

Fig. 1 - Bigram and nogram word error rates over time

We also examined whether the subjects tended to reduce their use of out-of-vocabulary words in Scenario 2. Subjects averaged 1.2 (less than 0.01%) out-of-vocabulary words in Scenario 1, as compared with 0.5 (also less than 0.01%) in Scenario 2. The number of these occurrences is so small as to be trivial; furthermore, the trend is nonsignificant, $F(1,21) = 1.74$, $p > .10$. This suggests that the use of fewer out-of-vocabulary words had little if any effect on overall recognition rates.

3.2. Speech Style

Having found no evidence for adaptation to the language models, we concluded that recognition improvement must be due to changes in user speech style. That is, as speakers became more familiar with the system, they learned to speak in ways that better matched the acoustics of the training data.

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 [9]. We predicted that a similar effect might occur for people speaking to a machine with less than perfect understanding. We noticed

that, when using an SLS as opposed to a wizard-mediated system [10], subjects tended to hyperarticulate: releasing stops, emphasizing initial word segments, pausing between words, and increasing vocal effort. Since most of the data used to train the DECIPHER™ recognizer came from wizard-mediated data collection, where recognition performance was nearly perfect, examples of "frustrated" speech were rare. For this reason, we predicted that hyperarticulation would impair recognition performance, and that perhaps the lower error rates in Scenario 2 might be due to a decrease in the frequency of hyperarticulation.

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) hyperarticulated in portions, or (3) hyperarticulated throughout the utterance. The coding was done blindly without reference to session context or recognition outcome.

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 for Scenarios 1 and 2 of the same 24 subjects. Because not enough speakers had utterances in all three categories, we combined the hyperarticulation coding of two levels for statistical purposes. For the 21 subjects who had both natural and hyperarticulate utterances, we compared recognition performance within subjects and found that the hyperarticulate utterances resulted in substantially higher word error rates, 0.25 as compared with 0.14, $F(1,20) = 15.68$, $p < .001$.

Given that hyperarticulation leads to more errors, it is possible that the overall decrease in error rates is due to a decrease in the rate of hyperarticulation. In fact, the frequency of hyperarticulated utterances decreased from an average of 46% of utterances to 30% from 1 to Scenario 2, $F(1,23) = 4.97$, $p < .05$. The decrease was more pronounced for the completely hyperarticulate utterances than for the partially hyperarticulate. As shown in Figure 2, users tended to use proportionally fewer completely hyperarticulate utterances in Scenario 2. Since this indicates a trend toward fewer hyperarticulated words within utterances, this finding may also help explain the decrease in error rate.

Converging evidence for the effect of frequency of hyperarticulation on overall recognition rates came from the experimental manipulation of instructions. Of our 24 subjects, 12 had been given instructions not to "overenunciate." Under these instructions, subjects hyperarticulated less, on 4.3 or 28.0% of all utterances as compared with 7.5 or 52.5%. This effect was reliable, $F(1,22) = 5.00$, $p < .05$. Since hyperarticulation rates decreased with instructions, we expected a comparable decrease in error rates. We compared word error rates for the two instruction groups and found that the subjects who received the instructions tended to have lower word error rates overall 0.15, as compared with 0.20; however, this effect was not significant. As Figure 3 shows, there was no interaction between instructions and session; both the instruction and no-instruction groups had similar decreases in word error rate over time. Figure 3 also shows the comparable rates of hyperarticulation. As with error rate, there was no significant interaction between instructions and hyperarticulation rate. Thus a decrease in hyperarticulation was associated with a decrease in word error both when observed over time and when manipulated by subject instructions. This finding suggests that a decrease in the rate of hyperarticulation may account for some or much of the decrease in error over time.

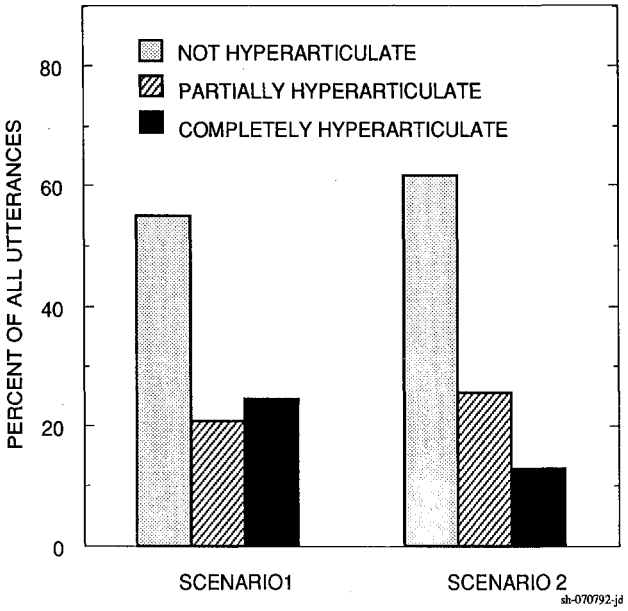

Fig. 2 - Frequency of hyperarticulate utterances over time

In addition to frequency of hyperarticulation, it is possible that the nature or degree of hyperarticulation may have changed over time. If hyperarticulated utterances themselves became more like the training data over time, this improved match might also have contributed to the reduction in error rate. Nonhyperarticulated utterances might also have become more similar to the training data. We measured the acoustic match between the utterances and the training data by running a forced alignment recognizer on the recorded sentences. Hidden Markov models associated with the sentence transcriptions were aligned to the VQ sequence produced by each sentence sound file. This procedure obtained the probability of each sentence's VQ sequence given the hidden Markov models it was aligned to.

Figure 4 shows log probabilities for hyperarticulated and nonhyperarticulated utterances in both scenarios. While the acoustic match for nonhyperarticulated utterances did not change over time, the match for hyperarticulated utterances improved sharply from the Scenario1 to Scenario 2. Because few subjects had utterances in all four categories (both hyperarticulated and nonhyperarticulated, in both scenarios), statistical tests were inappropriate. However, we observed a similar improvement in acoustic match for hyperarticulated utterances for both instruction groups, suggesting that the trend is not random. Thus, an additional factor contributing to lower recognition error rates is a change in the acoustic nature of hyperarticulated utterances over time

IV. CONCLUSION

We found that recognition word error rates decreased as users interacted with the same SLS over time. We found that this effect was due to changes in speech style rather than in adaptation to the language models of the system. We conclude that changes in one speech style, hyperarticulation, affect recognition rates in two ways. Users both decrease the rate of

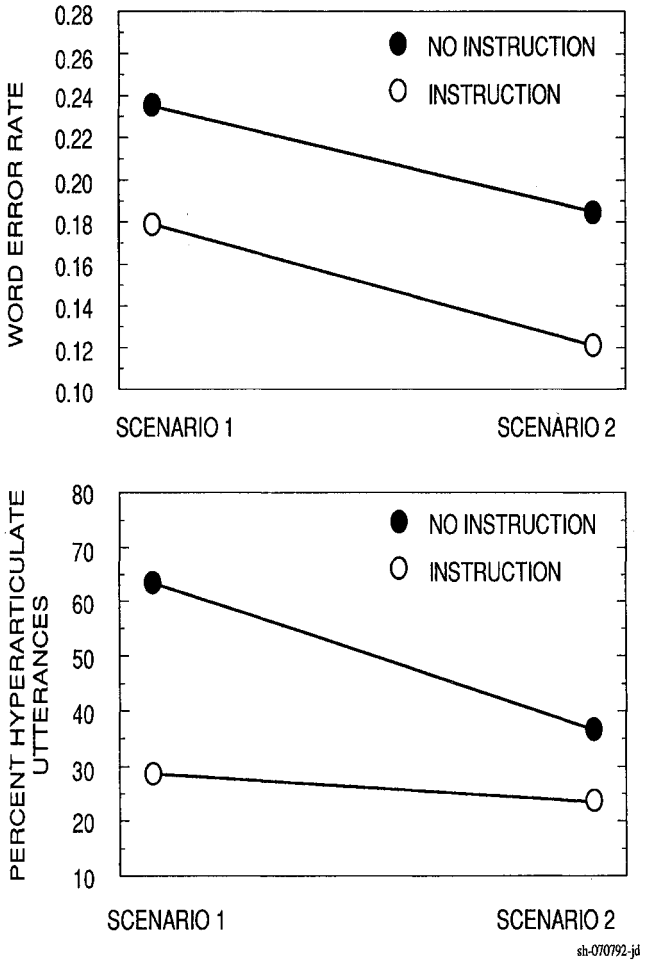

Fig. 3 - Effect of instructions to avoid overenunciation on word error and hyperarticulation rate over time

hyperarticulation and alter the way in which they hyperarticulate so as to better match the system's acoustic models. Together, these two adaptations may account for the improvement in recognition rates observed as subjects use the system over time.

ACKNOWLEDGMENTS

We gratefully acknowledge the work of Steven Tepper for system design and development, and John W. Butzburger for assistance and analyses. This research was supported by the Defense Advanced Research Projects Agency under Contract ONR N00014-90-C-0085.

REFERENCES

[1] Shriberg, E., E. Wade, P. Price, "Human-Machine Problem Solving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction," *Proc. DARPA Speech and Natural Language Workshop*, M. Marcus (ed.), Morgan Kaufmann, 1992

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

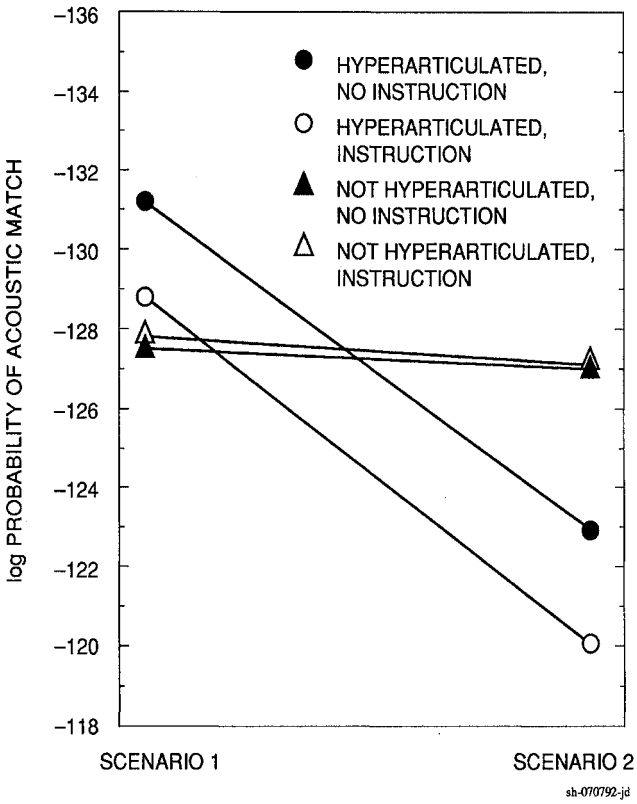

Fig. 4 - Deviation from perfect acoustic match between utterances and system models

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

[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] 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.

[6] 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.

[7] 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.

[8] Price P., "Evaluation of Spoken Language Systems: The ATIS Domain," *Proc. DARPA Speech and Language Workshop*, Morgan Kaufmann, 1990.

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

[10] Bly, B., P. Price, S. Tepper, E. Jackson, and V. Abrash, "Designing the Human Machine Interface in the ATIS Domain," *Proc. DARPA Speech and Language Workshop*, Morgan Kaufmann, 1990.