

Speech Graffiti vs. Natural Language: Assessing the User Experience

Stefanie Tomko and Roni Rosenfeld

Language Technologies Institute, School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh PA 15213
{stef, roni}@cs.cmu.edu

Abstract

Speech-based interfaces have great potential but are hampered by problems related to spoken language such as variability, noise and ambiguity. Speech Graffiti was designed to address these issues via a structured, universal interface protocol for interacting with simple machines. Since Speech Graffiti requires that users speak to the system in a certain way, we were interested in how users might respond to such a system when compared with a natural language system. We conducted a user study and found that 74% of users preferred the Speech Graffiti system to a natural language interface in the same domain. User satisfaction scores were higher for Speech Graffiti and task completion rates were roughly equal.

1 Introduction

Many problems still exist in the design of speech-based interfaces. Noisy environments and linguistic variability make interpretation of already uncertain input even more difficult, resulting in errors that must be handled effectively. What if many of these issues could be reduced by asking users to interact with speech-based systems in a structured way? Would they learn the interaction protocol? Would they prefer a more efficient yet structured interaction to one that was more natural, but perhaps less efficient?

One approach to structuring interaction is through directed-dialog systems. These generally work well for novice users, but they can be too meandering for expert users who know exactly what they want from the system. Specialized command-and-control languages give more power to the user, but they can be difficult for novices and require learning a new language for each new application encountered.

Speech Graffiti is a structured interaction protocol that is designed to be universal. Common input structures, output patterns and keywords are used for all Speech Graffiti applications, and once users are familiar with these for any single application, they can navigate

their way through any other Speech Graffiti application. Fig. 1 shows a sample Speech Graffiti dialog

User interactions with Speech Graffiti (independent of other speech interfaces) have previously been assessed in Rosenfeld et al. (2000). Here we consider a head-to-head comparison: given the chance to interact with both types of interfaces, which would people choose? What differences might be observed in performance or user satisfaction?

```
User: Theater is the Manor. Title is Tuck Everlasting.
System: THE CINEMAGIC MANOR THEATRE, TUCK EVERLASTING.
Options.
SHOW TIME, DAY, GENRE, {...}.
What are the show times?
2 MATCHES: 1:25, 5:30.
Start over.
STARTING OVER.
Area is Monroeville.
MONROEVILLE.
What are the theaters?
SHOWCASE CINEMAS PITTSBURGH EAST.
Theater is Showcase Cinemas Pittsburgh East.
SHOWCASE CINEMAS PITTSBURGH EAST.
What are the titles?
11 MATCHES, BLUE CRUSH, ABANDON, BROWN SUGAR, {...}.
Genre is comedy.
COMEDY.
Where am I?
AREA IS MONROEVILLE, THEATER IS SHOWCASE CINEMAS PITTSBURGH
EAST, WHAT ARE THE TITLES?, GENRE IS COMEDY.
What are the titles?
5 MATCHES, MY BIG FAT GREEK WEDDING, BROWN SUGAR,
JONAH - A VEGGIETALES MOVIE, {...}.
```

Figure 1. Sample Speech Graffiti interaction.

1.1 Related work

Although much research has been conducted on the design of natural language spoken dialog systems, far less research has been done on more standardized speech interfaces. Several studies have previously found that users are able to interact successfully using constrained or subset languages (e.g. Guindon & Shuldberg, 1987; Ringle & Halstead-Nussloch, 1989; Sidner & Forlines, 2002). As far as we know, no studies have been done comparing constrained, “universal” languages and natural language interfaces directly as we have done in this study. General information about the Speech Graffiti project and its motivation can be found in Rosenfeld et al. (2001).

2 Method

We conducted a within-subjects user study in which participants attempted a series of queries to a movie information database with either a Speech Graffiti interface (SG-ML) or a natural language interface (NL-ML). Participants repeated the process with the other system after completing their initial tasks and an evaluation questionnaire. System presentation order was balanced.

2.1 Participants

Twenty-three users (12 female, 11 male) accessed the systems via telephone in our lab. Most were undergraduate students from Carnegie Mellon University, resulting in a limited range of ages represented. None had any prior experience with either of the two movie systems or interfaces, and all users were native speakers of American English. About half the users had computer science and/or engineering (CSE) backgrounds, and similarly about half reported that they did computer programming “fairly often” or “very frequently.”

2.2 Training

Users learned Speech Graffiti concepts prior to use during a brief, self-paced, web-based tutorial session. Speech Graffiti training sessions were balanced between tutorials using examples from the MovieLine and tutorials using examples from a database that provided simulated flight arrival, departure, and gate information. Regardless of training domain, most users spent ten to fifteen minutes on the Speech Graffiti tutorial.

A side effect of the Speech Graffiti-specific training is that in addition to teaching users the concepts of the language, it also familiarizes users with the more general task of speaking to a computer over the phone. To balance this effect for users of the natural language system, which is otherwise intended to be a walk-up-and-use interface, participants engaged in a brief NL “familiarization session” in which they were simply instructed to call the system and try it out. To match the in-domain/out-of-domain variable used in the SG tutorials, half of the NL familiarization sessions used the NL MovieLine and half used MIT’s Jupiter natural language weather information system (Zue et al., 2000). Users typically spent about five minutes exploring the NL systems during the familiarization session.

2.3 Tasks

After having completed the training session for a specific system, each user was asked to call that system and attempt a set of tasks (e.g. “list what’s playing at the Squirrel Hill Theater,” “find out & write down what the ratings are for the movies showing at the Oaks Theater”). Participant compensation included task completion bonuses to encourage users to attempt each task in

earnest. Regardless of which system they were working with, all users were given the same eight tasks for their first interactions and a different set of eight tasks for their second system interactions.

2.4 Evaluation

After interacting with a system, each participant was asked to complete a user satisfaction questionnaire scoring 34 subjective-response items on a 7-point Likert scale. This questionnaire was based on the Subjective Assessment of Speech System Interfaces (SASSI) project (Hone & Graham, 2001), which sorts a number of subjective user satisfaction statements (such as “I always knew what to say to the system” and “the system makes few errors”) into six relevant factors: system response accuracy, habitability, cognitive demand, annoyance, likeability and speed. User satisfaction scores were calculated for each factor and overall by averaging the responses to the appropriate component statements.¹ In addition to the Likert scale items, users were also asked a few comparison questions, such as “which of the two systems did you prefer?”

For objective comparison of the two interfaces, we measured overall task completion, time- and turns-to-completion, and word- and understanding-error rates.

3 Results

3.1 Subjective assessments

Seventeen out of 23 participants preferred Speech Graffiti to the natural language interface. User assessments were significantly higher for Speech Graffiti overall and for each of the six subjective factors, as shown in Fig. 2 (REML analysis: system response accuracy $F=13.8$, $p<0.01$; likeability $F=6.8$, $p<0.02$; cognitive demand $F=5.7$, $p<0.03$; annoyance $F=4.3$, $p<0.05$; habitability $F=7.7$, $p<0.02$; speed $F=34.7$, $p<0.01$; overall $F=11.2$, $p<0.01$). All of the mean SG-ML scores except for annoyance and habitability are positive (*i.e.* > 4), while the NL-ML did not generate positive mean ratings in any category. For individual users, all those and only those who stated they preferred the NL-ML to the SG-ML gave the NL-ML higher overall subjective ratings.

Although users with CSE/programming backgrounds tended to give the SG-ML higher user satisfaction ratings than non-CSE/programming participants, the differences were not significant. Training domain likewise had no significant effect on user satisfaction.

¹ Some component statements are reversal items whose values were converted for analysis, so that high scores in all categories are considered good.

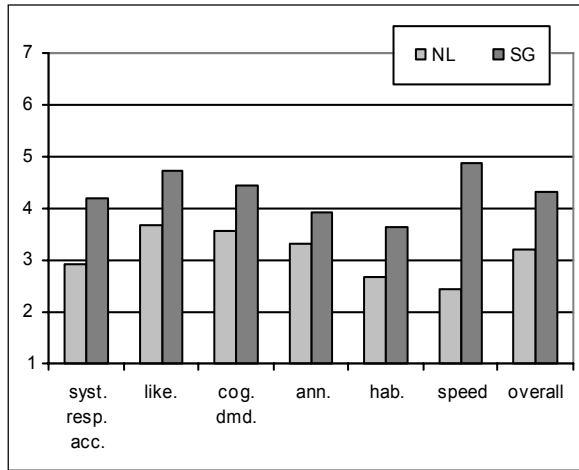


Figure 2. Mean user satisfaction for system response accuracy, likeability, cognitive demand, annoyance, habitability, speed and overall.

3.2 Objective assessments

Task completion. Task completion did not differ significantly for the two interfaces. In total, just over two thirds of the tasks were successfully completed with each system: 67.4% for the NL-ML and 67.9% for the SG-ML. The average participant completed 5.2 tasks with the NL-ML and 5.4 tasks with the SG-ML. As with user satisfaction, users with CSE or programming background generally completed more tasks in the SG-ML system than non-CSE/programming users, but again the difference was not significant. Training domain had no significant effect on task completion for either system.

To account for incomplete tasks when comparing the interfaces, we ordered the task completion measures (times or turn counts) for each system, leaving all incompletes at the end of the list as if they had been completed in “infinite time,” and compared the medians.

Time-to-completion. For *completed* tasks, the average time users spent on each SG-ML task was lower than for the NL-ML system, though not significantly: 67.9 versus 71.3 seconds. Considering incomplete tasks, the SG-ML performed better than the NL-ML, with a median time of 81.5 seconds, compared to 103 seconds.

Turns-to-completion. For *completed* tasks, the average number of turns users took for each SG-ML task was significantly higher than for the NL-ML system: 8.2 versus 3.8 ($F=26.4$, $p<0.01$). Considering incomplete tasks, the median SG-ML turns-to-completion rate was twice that of the NL-ML: 10 versus 5.

Word-error rate. The SG-ML had an overall word-error rate (WER) of 35.1%, compared to 51.2% for the NL-ML. When calculated for each user, WER ranged from 7.8% to 71.2% (mean 35.0%, median 30.0%) for the SG-ML and from 31.2% to 78.6% (mean 50.3%, median 48.9%) for the NL-ML. The six users with the

highest SG-ML WER were the same ones who preferred the NL-ML system, and four of them were also the only users in the study whose NL-ML error rate was lower than their SG-ML error rate. This suggests, not surprisingly, that WER is strongly related to user preference.

To further explore this correlation, we plotted WER against users’ overall subjective assessments of each system, with the results shown in Fig. 3. There is a significant, moderate correlation between WER and user satisfaction for Speech Graffiti ($r=-0.66$, $p<0.01$), but no similar correlation for the NL-ML system ($r=0.26$).

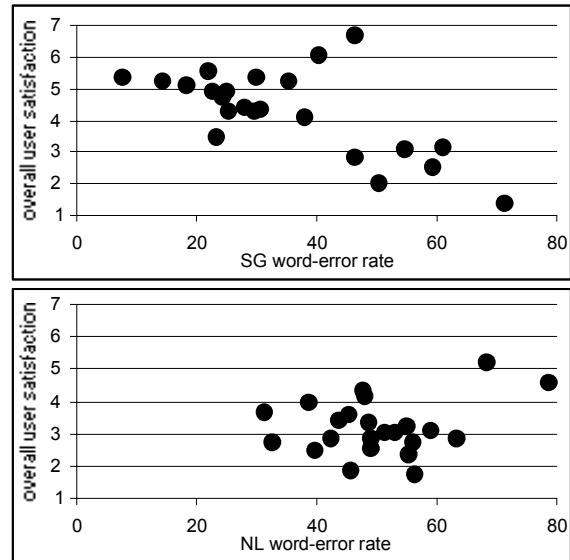


Figure 3. Word-error rate vs. overall user satisfaction for Speech Graffiti and natural language MovieLines.

Understanding error. Word-error rate may not be the most useful measure of system performance for many spoken dialogue systems. Because of grammar redundancies, systems are often able to “understand” an utterance correctly even when some individual words are misrecognized. Understanding error rate (UER) may therefore provide a more accurate picture of the error rate that a user experiences. For this analysis, we only made a preliminary attempt at assessing UER. These error rates were hand-scored, and as such represent an approximation of actual UER. For both systems, we calculated UER based on an entire user utterance rather than individual concepts in that utterance.

SG-ML UER for each user ranged from 2.9% to 65.5% (mean 26.6%, median 21.1%). The average change per user from WER to understanding-error for the SG-ML interface was -29.2% .

The NL-ML understanding-error rates differed little from the NL-ML WER rates. UER per user ranged from 31.4% to 80.0% (mean 50.7%, median 48.5%). The average change per user from NL-ML WER was $+0.8\%$.

4 Discussion

Overall, we found that Speech Graffiti performed favorably compared to the natural language interface. Speech Graffiti generated significantly higher user satisfaction scores, and task completion rates and times were similar.

The higher turns-to-completion rate for Speech Graffiti is not necessarily problematic. The phrasal nature of Speech Graffiti syntax seems to encourage users to input single phrases; we suspect that in a longitudinal study, we would find single-utterance command use in SG-ML increasing as users became more familiar with the system. Furthermore, because the SG-ML splits long output lists into smaller chunks, a user often has to explicitly issue a request to hear more items in a list, adding at least one more turn to the interaction. Thus there exists a trade-off between turn-wise efficiency and reduced cognitive load. Because of the reasonable results shown for the SG-ML in user satisfaction and completion time, we view this as a reasonable trade-off.

It is possible that if lower word-error rates can be achieved, Speech Graffiti would become unnecessary. This may be true for consistent, extremely low word-error rates, but such rates do not appear to be attainable in the near term. Furthermore, the correlations in Fig. 3 suggest that as WER decreases, users become more satisfied with the SG interface but that this is not necessarily true for the NL interface. Consider also the effect of understanding error. UER is the key to good system performance since even if the system has correctly decoded a word string, it must still match that string with the appropriate concepts in order to perform the desired action. Although WER may be reduced via improved language and acoustic models, matching input to understanding in NL systems is usually a labor-intensive and domain-specific task. In contrast, the structured nature of Speech Graffiti significantly reduces the need for such intensive concept mapping.

Future work. For Speech Graffiti, scores for habitability (represented by statements like “I always knew what to say to the system”) were typically the lowest of any of the six user satisfaction factors, suggesting that this is a prime area for further work.

In this regard, it is instructive to consider the experience of the six users who preferred the NL interface. Overall, they accounted for the six highest SG-ML word- and understanding-error rates and the six lowest SG-ML task completion rates: clearly not a positive experience. An additional measure of habitability is *grammaticality*: how often do users speak within the Speech Graffiti grammar? The six NL-ML-preferring users also had low grammaticality rates (Tomko & Rosenfeld, 2004). These users have become a motivator of future work: what can be done to make the interface work for them and others like them? (Future studies will

focus on a broader population of adults.) How can we help users who are having severe difficulties with an interface learn how to use it better and faster? To improve the habitability of Speech Graffiti, we plan to explore allowing more natural language-esque interaction while retaining an application-portable structure. We also plan to refine Speech Graffiti’s runtime help facilities in order to assist users more effectively in saying the right thing at the right time.

In addition to these core interface goals, we plan to extend the functionality of Speech Graffiti beyond information access to support the creation, deletion and modification of information in a database.

Acknowledgements

This work was supported by an NDSEG Fellowship, Pittsburgh Digital Greenhouse and Grant N66001-99-1-8905 from the Space & Naval Warfare Systems Center. The information in this publication does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

References

- Guindon, R. & Shuldberg, K. 1987. Grammatical and ungrammatical structures in user-adviser dialogues: evidence for sufficiency of restricted languages in natural language interfaces to advisory systems. In *Proc. of the Annual Meeting of the ACL*, pp. 41-44.
- Hone, K. & Graham, R. 2001. Subjective Assessment of Speech-System Interface Usability. In *Proceedings of Eurospeech*, Aalborg, Denmark.
- Ringle, M.D. & Halstead-Nussloch, R. 1989. Shaping user input: a strategy for natural language design. *Interacting with Computers* 1(3):227-244
- Rosenfeld, R., Zhu, X., Toth, A., Shriver, S., Lenzo, K. & Black, A. 2000. Towards a Universal Speech Interface. In *Proceedings of ISCLP*, Beijing, China.
- Rosenfeld, R., Olsen, D. & Rudnicky, A. 2001. Universal Speech Interfaces. *Interactions*, 8(6):34-44.
- Sidner, C. & Forlines, C. 2002. Subset Languages for Conversing with Collaborative Interface Agents. In *Proc. of ICSLP*, Denver CO, pp. 281-284.
- Tomko, S. & Rosenfeld, R. 2004. Speech Graffiti habitability: what do users really say? To appear in *Proceedings of SIGDIAL*.
- Zue, V., Seneff, S., Glass, J.R., Polifroni, J., Pao, C., Hazen, T.J. & Hetherington, L. 2000. JUPITER: A Telephone-Based Conversational Interface for Weather Information, *IEEE Transactions on Speech and Audio Processing*, 8(1): 85 –96.