# Shaping User Input in Speech Graffiti: a First Pass

**Stefanie Tomko**
Language Technologies Institute
Carnegie Mellon University.
5000 Forbes Ave.
Pittsburgh, PA 15213 USA
stef@cs.cmu.edu

**Roni Rosenfeld**
Language Technologies Institute
Carnegie Mellon University.
5000 Forbes Ave.
Pittsburgh, PA 15213 USA
roni@cmu.edu


*CHI 2006,* April 22–27, 2006, Montréal, Québec, Canada.
ACM 1-59593-298-4/06/0004.

## Abstract
Speech Graffiti is a standardized interaction protocol for spoken dialog systems designed to address some common difficulties with ASR. We have proposed a strategy of *shaping* to help users adapt their interaction to match what the system understands best, thereby reducing the chance for misunderstandings and improving interaction efficiency. In this paper we report on an evaluation of our initial implementation of shaping in Speech Graffiti, noting that our baseline strategy was not as powerful as expected, and discussing proposed changes to improve its effectiveness.

## Keywords
Spoken dialog systems, speech recognition, user interfaces

## ACM Classification Keywords
H.5.2 User Interfaces – Voice I/O.

## Introduction
The Speech Graffiti project is an attempt to address several of the current issues in automatic speech recognition (ASR) technology for human-computer interaction. Although ASR offers the promise of simple, direct access to information, factors such as environmental noise, non-native speech patterns, or friendly

yet "talkative" conversational agents can make communication with spoken dialog systems inefficient. Often, inefficiencies in human-computer speech interaction are a result of the user having spoken beyond the bounds of what the computer understands. This situation leads to misunderstandings on the parts of both the user and the system; recovering from such events can add extra time to the overall interaction. Furthermore, even when users know what to say to a system, the system often requires them to listen to a series of long prompts and menus before reaching their goal.

We have designed Speech Graffiti to be a protocol for structured, more efficient speech interaction with information access systems [10].[1] By asking users to speak in a certain way, using specific vocabulary and syntactic structures, we reduce ambiguity from both the user and the system perspectives. Speech Graffiti is a user-initiative system, meaning that it does not prompt the user for responses. Instead, it includes keywords that allow users to discover the functions of the system and the things that they can say to it. This format provides the ability to access information more directly, rather than having to navigate through layers of menus. Fig. 1 provides an example of an interaction with a Speech Graffiti movie information system.

Our previous work analyzed user interactions with the system and found that when users successfully learn and understand Speech Graffiti, they have more efficient and satisfying interactions compared to a natural language interaction (*i.e.,* one in which users can

[1] We have also investigated its use for controlling household devices [3].

**User: Theater is *Showcase North Theater***
System: Showcase Cinemas Pittsburgh North
**Genre is *drama***
Drama
**What movies are playing?**
{confsig} [an error beep, since previous utterance is not in grammar]
**WHERE WAS I?**
Theater is Showcase Cinemas Pittsburgh North, genre is drama
**OPTIONS**
You can specify or ask about title, show time, rating,
{ellsig} [a 3-beep list continuation signal]
***What* is title?**
2 matches: Dark Water, War of the Worlds
**START OVER**
Starting over
**Theater is *Northway Mall Cinemas Eight***
Northway Mall Cinemas 8
***What* is address?**
1 match: 8000 McKnight Road in Pittsburgh

**figure 1:** Example non-shaping Speech Graffiti dialog. User utterances in ALL CAPS represent universal keywords; underlined and *italicized* user input shows standard SLOT-NAME + *VALUE* input syntax.

issue queries in a less structured format, as in "what's playing at the Manor Theater tonight?") [9]. However, some users (6 of 23 in [9]) were not able to learn Speech Graffiti well enough to use it proficiently, resulting in very frustrating interactions. Our current work focuses on improving the habitability of Speech Graffiti to make it more usable for more users.

As a step towards improving the habitability of Speech Graffiti, we have implemented a strategy of *shaping*, after the cognitive psychology concept of successive conditioning of new responses [2]. The goal of shaping in Speech Graffiti is to help users learn the interaction style over time, *while using the system*.

One aspect of our original plan for Speech Graffiti was that it should be learnable via a brief, 5- to 15-minute

tutorial. Because the interaction style is standardized across applications, the initial cost of learning it should be amortized when one uses multiple Speech Graffiti applications. In [9] however, we found that some users went through the tutorial session, declared that they understood the system, and then promptly forgot what to say once the interaction began. Shaping should help such users get back on track. Ultimately, successful shaping could even preclude the need for a pre-use tutorial altogether, which is in keeping with our overall goal of increased interaction efficiency.

## Related work

Various studies and researchers have suggested that restricted languages are indeed a reasonable approach to interaction with computers and that such input is not necessarily unnatural [5,8]. For instance, [7] suggests that using a small, well-defined language may actually make interactions easier for novices, since it clarifies what is and what is not accepted by the system. The strategy of shaping is based on the phenomena of lexical entrainment and other speaker adaptations that are well-documented in both human-human and human-computer communication [1,6,12].

## Shaping

Our baseline attempt at shaping user input has two components: an expanded grammar and shaping confirmation.

*Expanded grammar*

The expanded grammar is designed to allow more natural language interaction compared to the target (as in fig. 1) Speech Graffiti grammar. Currently, the expanded grammar is hand-crafted, and its content is informed by the natural language queries made by users in our prior studies. The only structural limitation in this grammar is that utterances must map linearly to a target grammar equivalent (fig. 2).

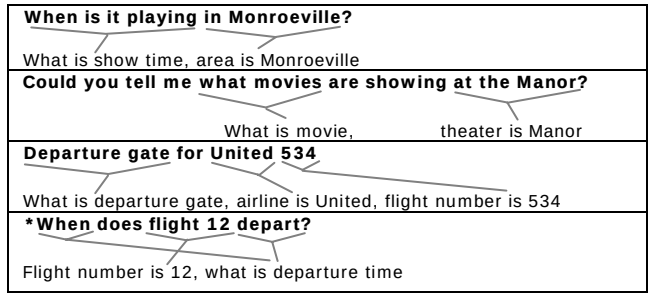


**figure 2:** Sample expanded grammar utterances and target grammar equivalents. The final example is not allowed by the expanded grammar, since it does not map in a strictly linear manner to a target language input.

A natural question here is why, since we are allowing natural language into the system here, we do not simply create a natural language system. Why, at this point, make the effort of shaping users towards the target language? The manual effort (including corpus collection) required to create the expanded grammar provides one source of motivation here. One can also imagine the existence of both shaping and non-shaping Speech Graffiti applications, such that non-shaping versions are targeted at users who are already proficient in the interaction style. This would allow developers to take advantage of the simple application generation procedure that Speech Graffiti's standard structure provides [11]. Another motivation is that, once a user has learned Speech Graffiti, the standard Speech Graffiti interactions tend to be less-error prone compared to natural language ones, resulting in more efficient interactions.

*Shaping confirmation*
Shaping confirmation is designed to explore the effectiveness of *implicit shaping*. In non-shaping Speech Graffiti, the system responds to each user input with a brief, value-only confirmation, as in the first two system responses in fig. 1. This allows the user to notice if an ASR error has occurred and to either correct or continue the interaction as appropriate. For our initial shaping version of Speech Graffiti, user input was instead confirmed with the full, target *SLOT-NAME* + *VALUE* Speech Graffiti form of the input, intended to shape user input implicitly via syntactic priming.

## Evaluation

To evaluate the initial shaping strategy, we conducted a user study comparing the initial shaping strategy to the non-shaping system. This evaluation had two goals:

- to determine the effectiveness of simple, implicit shaping on user input and interaction efficiency; and
- to collect a corpus of interactions for informing more advanced shaping strategies.

*Participants*
15 male and 14 female adults participated in the study. All were native speakers of American English and were new to the Speech Graffiti interface. In our prior study, we found that users with computer programming experience were much more likely to succeed with Speech Graffiti; because the current work was based on the idea that we should improve habitability for those types of users who had difficulties in [9], it was important for us that no current participants had significant programming experience. Participants were paid a flat rate for their participation plus a bonus based on how many tasks they completed successfully during their interactions.

*Setup*
A between-subjects experiment was designed in which participants were randomly assigned to one of three conditions: non-shaping+tutorial, shaping+tutorial, or shaping+no_tutorial. Participants in the tutorial groups were given a self-guided PowerPoint presentation covering the input structure of Speech Graffiti, its confirmation strategy, list navigation, error correction, and a few general tips. The tutorials for both the shaping and non-shaping conditions were identical with the exception of the confirmation strategies presented. Short audio examples were included for participants to listen to as they worked through the tutorial. Tutorial group participants had five minutes to work on the tutorial.

*Tasks and Assessment*
The application used in the study was the Speech Graffiti MovieLine, which provides information about theaters and movies showing in the Pittsburgh area. Participants were asked to complete a series of 15 information retrieval tasks; these were presented to users on a sheet of paper in a format designed to encourage them to use their own words when speaking to the system. Users were asked to work through the tasks in order, writing down the answers for each. Participants were given forty minutes to work through the set of tasks. At the end of the task session, users completed an evaluation questionnaire comprising 36 statements (*e.g.* "I always knew what to say to the system;" "The system was too inflexible") to be evaluated on a 7-point Likert scale [4].

## Results and Discussion

In nearly all aspects, the initial shaping strategy did not result in an improvement over the original Speech Graffiti system. Users completed an average of 8.1 tasks with the non-shaping system and 10.6 tasks with the shaping version (combined tutorial and non-tutorial), $X^2(1,29) = 11.48$, $p = 0.40$. For completed tasks, the differences in the mean time and number of turns to complete each task for the three groups were also not statistically significant (time: $F(2,29) = 0.13$, $p = 0.88$; turns: $F(2,29) = 1.33$, $p = 0.28$). Finally, mean user satisfaction scores did not differ significantly between the three conditions ($F(2,29) = 0.31$, $p = 0.73$).

From the perspective of shaping user input, our key interest in this research was how often users said something that was Speech-Graffiti-grammatical; that is, how often they speak within the target grammar. Overall, we did not find a significant difference in grammaticality between the two general conditions: non-shaping mean, 64.6%; shaping mean, 63.8% ($t(15) = 0.13$, $p = 0.90$).

Based on these results, we conclude that the simple, implicit shaping feedback is not strong enough to significantly shape user input to match the target grammar. It appears that in the shaping condition, the fact that users could retrieve the desired information using non-target-language input was a more powerful influence on *not* shaping than the feedback was *for* shaping.

Working towards a stronger model of shaping, we next analyzed user interactions in the shaping groups to assess in what ways these interactions were problematic. We identified the following nine issues, listed in rough order of frequency among users:

1. Persistent use of natural language query formats.
2. Not using START OVER when needed. (Since the query context is not automatically cleared after a query, this can lead to unwanted constraints persisting in future queries.)
3. Persistent use of *SLOT-NAME*-only query formats (*e.g.,* "theater").
4. Confusion about the semantics of "location."
5. Using utterances that are too long.
6. Using NEXT (which retrieves the next single item from a list) instead of MORE (which retrieves the next three items).
7. Persistent use of *VALUE*-only specification formats (*e.g.,* "drama").
8. Pacing issues (*e.g.*, not waiting for confirmation, or pausing mid-utterance).
9. Using "*SLOT-NAME* equal *VALUE*" specification format.

For each of the items noted above, we have designed a three-part strategy, intended to prevent the issue from occurring in the first place, to recognize it if it does occur, and to provide appropriate, shaping feedback to the user if it is recognized.

Issues 1, 3, 7 and 9 are directly related to the problem of shaping. In these situations, we propose to give the user feedback that is more explicitly shaping towards the target language, such as

User: **Could you tell me what movies are showing at the Manor?**
System: I think you meant: "theater is Manor theater, what are the movies?"

Issues 1 and 3 also point to an underlying design issue. Generally, participants were much better at formulating target language constraint specifications than they were with queries. The specification syntax is quite straightforward—"*SLOT-NAME* is *VALUE*"—and we intended the query structure to be straightforward as well: "what is *SLOT-NAME*?" However, it appears that to novice users, the query structure is actually understood more vaguely, as something like "*ask a question about SLOT*." We propose actually making the query format *more* structured, in the form of "list *SLOT-NAME*," which we hypothesize will be better assimilated by the user.

Issues 2, 5, 6 and 8 are not related to the issue of shaping, but are critical to interaction efficiency. Such issues will be addressed in future iterations with suggestion prompts that are triggered when such situations are recognized. Issue 4 is a somewhat domain-specific problem, and this term (a synonym for "area," as in "area is North Hills") will be dropped from future versions.

## Work in Progress

We are currently implementing the revised shaping strategies in preparation for a second user study, although some of the updated strategies may undergo intermediate testing in Wizard-of-Oz studies. For the second user study, we also plan to add an additional domain (such as air travel or restaurant information) in order to determine how appropriate the shaping strategies are across domains.

## References

[1] Brennan, S.E. Lexical entrainment in spontaneous dialog. In *Proc. ISSD 1996*, 41-44.

[2] Domjan, M. *The Essentials of Conditioning and Learning*. Thomson Wadsworth, Belmont CA, USA 2005.

[3] Harris, T.K. and Rosenfeld, R. A Universal Speech Interface for Appliances. In *Proc. ICSLP 2004*.

[4] Hone, K. and Graham, R. Subjective Assessment of Speech-System Interface Usability. In *Proc. Eurospeech 2001*.

[5] Jackson, M.D. Constrained Languages Need Not Constrain Person/Computer Interaction. *SIGCHI Bulletin 15*, 2-3 (1983), 18-22.

[6] Matarazzo, J.D., Weitman, M., Saslow, G. and Weins, A.N. Interviewer influence on duration of interviewee speech. *Journal of Verbal Learning and Verbal Behavior 1* (1963), 451-458.

[7] Shneiderman, B. Natural Vs. Precise Concise Languages for Human Operation of Computers: Research Issues and Experimental Approaches. In *Proc. ACL 1980*, 139-141.

[8] Sidner, C. and Forlines, C. Subset Languages for Conversing with Collaborative Interface Agents. In *Proc. ICSLP 2002*, 281-284.

[9] Tomko, S. Speech Graffiti: Assessing the User Experience. CMU LTI Tech Report CMU-LTI-04-185 (2004). www.cs.cmu.edu/~stef/papers/mthesis.ps

[10] Tomko, S., Harris, T.K., Toth, A., Sanders, J., Rudnicky, A. and Rosenfeld, R. Towards Efficient Human Machine Speech Communication: The Speech Graffiti Project. *ACM Transactions on Speech and Language Processing 2*, 1 (2005).

[11] Toth, A., Harris, T., Sanders, J., Shriver, S., and Rosenfeld, R. Towards Every-Citizen's Speech Interface: An Application Generator for Speech Interfaces to Databases. In *Proc. ICSLP 2002*, 1497-1500.

[12] Zoltan-Ford, E. How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies 34* (1991), 527-547.