
Designing Speech Interfaces for Kiosks

Max Van Kleek
Buddhika Kottahachchi
Tyler Horton
Paul Cavallaro

EMAX@CSAIL.MIT.EDU
BUDDHIKA@CSAIL.MIT.EDU
THORTON@CSAIL.MIT.EDU
PAULC@CSAIL.MIT.EDU

MIT Computer Science and Artificial Intelligence Laboratory, Stata Center, 32 Vassar Street, Cambridge, MA 02139 USA

1. Introduction and Motivation

The Oxygen Kiosk Network (OK-Net) Project (Van Kleek, 2003) has deployed its first batch of interactive touchscreen kiosks within the MIT CSAIL laboratory. These kiosks are strategically placed in locations where people tend to congregate, such as lounges and elevator lobbies. The kiosks currently run an application known as SKINNI (Kleek et al., 2004), the Smart Kiosk Information Navigation and Note-posting Interface, which provides information about lab and MIT-wide current events, a basic directory, and also map navigation information useful to both visitors and lab members alike. The oversized widgets and simple, single touch manipulation of SKINNI's GUI was designed for fingertip navigation on touchscreen displays. At present, the kiosks are also equipped with a small keyboard to facilitate text input, for searching or posting new announcements.

While the kiosks that are currently deployed have seen sustained usage, informal discussions with kiosk users have led us to believe that the touchscreen interface is far from ideal for many purposes. In particular, when a user wishes to look up a specific piece of information, such as looking someone up in the CSAIL directory, or finding directions to a specific room within the laboratory, using the SKINNI GUI via touchscreen displays feels tedious. It requires an elaborate sequence of button taps or a combination of button taps and keyboard entry to narrow down a search, or even a brute-force visual scan of a long list of items. Even after several GUI design iterations we have been unable to come up with a better layout that makes GUI navigation more efficient, without increasing the initial learning curve. Therefore, we decided to augment SKINNI, using natural language via speech with the hope that this would improve the overall user experience, through a more direct and natural query interface. This paper describes our design process, implementation, and concludes with an evaluation done as an informal user study.

Therefore, we decided to augment SKINNI, using natural language via speech with the hope that this would improve the overall user experience, through a more direct and natural query interface. This paper briefly examines related work, describes our design process, implementation,

and concludes with an evaluation done as an informal user study.

1.1 Related Work

Gauvain et al.'s speech recognition work with the ESPRIT MASK project, a public information kiosk designed to provide passengers with timetables and other information at railway stations in Paris, has strong similarities to our own work (Gauvain et al., 1996). They identified many of the challenges with designing a speech recognizer for public use in an open space. Issues such as robustness, speaker independence, and signal capture/noise compensation were taken into account in the custom speech recognizer they designed for this purpose. Their focus was primarily on improving recognition accuracy, and therefore other aspects of their system were not described in detail. Christian et al.'s "Smart Kiosk" project, which was partially an inspiration for the OK-net project, had an embodied conversational user interface agent capable of speech. After an initial trial, they discovered that in order for speech interaction to be robust and usable, they had to carefully craft a domain-specific grammar. As an example, they describe how they chose their grammar for a game of speech-driven blackjack (Christian & Avery, 2000). The work done in both these projects has influenced our own design and implementation.

2. Design considerations

Many of the original design requirements of the SKINNI GUI interface apply to the speech interface as well. The interface must have a low usability threshold; ie. it must be easily learnable for new users. In addition, the interface must accommodate users of diverse ages, nationalities and computer experience.

2.1 Usability of speech interfaces

Speech interfaces pose a variety of additional usability challenges. Since speech interfaces are still relatively new and uncommon, users are likely to be inherently less familiar with speech interfaces than with GUIs, and may have widely different expectations of their interaction capabil-

ities. Furthermore, since our speech system (described later) is limited to pre-specified forms of queries, rather than unconstrained spoken interaction, it becomes a challenge for the application designer to ensure that the system can recognize the most common forms of queries the user might ask. Finally, the user interface must handle user misrecognition, since it is likely to occur more often than with a GUI interface.

Acoustic requirements must also be considered when designing speech interfaces for public kiosks. Users are unlikely to want to put on a headset microphone for short interactions, and far field microphones generally capture too much ambient noise. Directed noise-canceling linear array microphones provide a good compromise at a distance of 2-3 feet; however, they require users' mouths to be positioned within a constrained physical volume near the center of the array. This may pose a barrier to making the system accessible to users of various heights or users in wheelchairs.

3. Current Implementation

For our implementation, we opted to use the in-house conversational speech engine, Galaxy (Polifroni & Seneff, 2000). Galaxy is attractive for its speaker-independent speech recognition engine that is resilient to outside noise, speech disfluencies and accents, as well as for its advanced dialogue and discourse capabilities.

At present, we have outfitted two prototype kiosks for speech, with speakers and active noise-canceling linear array microphones. A sound daemon running on these machines performs speech detection and capture, and transmits speech waveforms to a separate machine designated as the Galaxy Hub. The hub coordinates recognition and natural language processing, and passes a parsed utterance back to a SKINNI-speech component known as the *back-end script*. This script interprets the parsed utterance, sends actions to the SKINNI GUI, computes a verbal response, and updates the speech feedback UI. The response to be verbalized is sent back to the Hub, which passes it to the appropriate text-to-speech (TTS) engine, and sends synthesized waveforms back to the kiosk, where they are played by the sound daemon. The Galaxy domain, consisting of recognition grammars, hub and startup scripts, was created via SpeechBuilder (Weinstein, 2001), a web-based tool for building Galaxy domains.

3.1 Designing the Domain Grammar

The most challenging aspect of building the system was designing the recognition grammar that Galaxy uses to interpret spoken utterances. We needed to determine all the basic ways people would phrase their queries. Due to the large number of different phrasings possible for even the simplest queries, we limited our initial speech capabilities to only directory field queries (ie. office, telephone number, email address) and ?show me room? queries for the

map. To determine the most common phrasings of these queries, we initially built up what we thought were the most common forms, and then asked 10 lab members how they would phrase the queries. We then extracted common forms from their phrasings and added them to the grammar. In the future, we hope to design a system that can incrementally learn new phrasings for queries.

3.2 Speech state feedback GUI

Initial impressions from using speech on the kiosks led us to realize the importance of conveying immediate feedback about the state of the speech system to the user. Particularly due to the variable delays incurred in processing the speech, parsing the result, and network transmission, without feedback users were left unsure whether the system heard the speech, encountered an error, or was still in the midst of processing it. In response, we designed a GUI that displays immediate feedback about whether speech was enabled (via a red/green microphone icon), when speech was detected, when it was being processed, output from the recognizer, and, finally, a textual rendering of the spoken response from the back-end script. We have found that this allows the user to easily determine when the kiosk is listening, why or how an error occurred, and how to proceed. This is particularly helpful when new users are familiarizing themselves with the speech interface since it allows them to quickly grasp the systems' capabilities. We believe this will help reduce the number of users who give up on the speech interface too early if they initially experience problems efficiently interacting with the kiosks.

4. Prototype Evaluation

To gauge the usability of our system, we performed an informal user study with 10 lab members. Although the study was too small to be statistically significant, and not a complete representative sample of the potential user population (ie., we had no outside visitors), this study assisted us in identifying design problems to fix prior to deployment. Our study asked subjects to look up the telephone number of eighteen different people in lab (selected randomly) using the kiosks. We asked our subjects to use the speech interface for the first six names, the touchscreen interface for the second six names, and their own preferred interface for the final six names.

As indicated in Table 1, our study yielded mixed results. As we wished to evaluate how well first time users would perform, subjects were not told what forms of queries the kiosk could recognize. This is reflected by our recognition rate for the first two trials being particularly low at 50% across users. However, they improved to 72% thereafter. Misrecognitions caused severe time penalties as a result of requiring an exhaustive search by the speech recognizer, combined with the time required to generate a spoken error response. A combination of the high error-rate and misrecognition penalty yielded a longer average time

	Unrecognized Speech(secs)	Recognized Speech(secs)	Overall Speech(secs)	Touch (secs)
Min	10	3	3	4
Max	25	9	25	19
Mean	16.78	5.22	9.11	7.33
S.D.	5.13	0.92	6.24	3.18

Table 1. Study Results: The times indicate the interval between when the subject hears the name that needs to be queried and when the subject extracts the corresponding telephone number from the screen.

for speech over touch. However, when the recognition was successful, times were consistently shorter with low variance. Also, it was encouraging that 8 out of our 10 subjects preferred using the speech interface when given a choice between the two.

5. Current Progress and Future Work

We are currently working on improving the speech domain, as well as updating our voice models to improve recognition accuracy. With help from the Spoken Language Systems group, we are also starting to build new voice models based on raw acoustic data from kiosks rather than models built from telephone data. Furthermore, we are working to extend our existing speech domain such that all functionality exposed via our touchscreen interface is also exposed via the speech interface. Conversation support, with dialogue and discourse management, will follow. We would also like to support multiple languages. We are optimistic that our efforts will render speech an efficient and natural interaction modality for kiosks.

References

- Christian, A. D., & Avery, B. L. (2000). Speak out and annoy someone: experience with intelligent kiosks. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 313–320). The Hague, The Netherlands: ACM Press.
- Gauvain, J., Gangolf, J., & Lamel, L. (1996). Speech recognition for an information kiosk. *Proc. ICSLP '96* (pp. 849–852). Philadelphia, PA.
- Kleek, M. V., Horton, T., & Boyle, E. (2004). Skinni: Connecting coworkers using information kiosks in the workplace. http://www.ai.mit.edu/emax/papers/CHI-Skinni_v6.pdf.
- Polifroni, J., & Seneff, S. (2000). Galaxy-II as an architecture for spoken dialogue evaluation. *Second International Conference on Language Resources and Evaluation*. Greece.
- Van Kleek, M. (2003). Intelligent environments for informal public spaces: the Ki/o Kiosk Platform. M.Eng. Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Weinstein, E. (2001). Speechbuilder: Facilitating spoken dialogue system development. Master's thesis, MIT.