ABSTRACT
We describe a unified design for voice interaction with simple machines; discuss the motivation for and main features of the approach, include a short sample interaction, and report the results of two preliminary experiments.

Keywords: Spoken dialog systems, speech, user interface

INTRODUCTION
For speech recognition to become widespread, users must learn how to speak to and interact with a variety of systems (information servers, handheld devices, transaction servers, household appliances, etc.). This includes knowing what vocabulary and syntax to use with each different system, as well as having some way of ascertaining a system’s capabilities and limitations.

One solution to this problem is to use unconstrained, natural language dialog interfaces, in which a system is designed to respond to open, conversational input (“When’s the first flight to New York Monday?” “Did my stocks go up?”). However, this approach can be problematic for both developers and users: a large amount of domain knowledge is required to sufficiently model possible user input, and the large vocabularies and complex grammars necessary for such systems can adversely affect speech recognition accuracy. Users may also experience problems if they overestimate a system’s knowledge and ask it questions that it is not equipped to handle.

Another approach is to use machine-driven dialogs to guide users to their goals, but this is not much of an improvement over the touch-tone menu interfaces so ubiquitous in telephone-based systems. In these systems, the user is often forced to listen to a variety of options, most of which are presumably irrelevant to their goal. Interactions are slowed by this forced iteration of options at each step, and although frequent users may be able to speed up their interactions by memorizing the appropriate sequence of keypresses, these sequences are not valid across applications, and users therefore must learn a separate interface pattern for each new system used.

ANOTHER APPROACH
We have been working on an alternate paradigm for voice interface systems, called the Universal Speech Interface (USI). With this approach, users learn a set of strategies that help them explore and use any application that is designed using the USI protocol. The core features of the USI are a small set of keywords (less than ten) and a standard structure for input and output.

The USI keywords are designed to provide standard mechanisms for interaction universals, which we derived by analyzing several applications and application categories prior to developing the USI vocabulary. These universals include help, orientation, navigation, error correction, and general system interaction.

The standard structure provides principles governing the regularities in the interaction, such as “the system will tersely paraphrase whatever part of the input it understood” and “input is always provided in phrases, each conveying a single information element.” Each application designer can choose how flexible the grammar should be for the individual phrases. This can range from a tightly proscribed format to unconstrained natural utterances. In the applications we have designed so far, in order to help improve recognition accuracy, the phrase grammar is quite stylized, as is seen in the examples below.

Our approach requires users to learn a specific interaction style. We found that the USI can be effectively taught in a five minute, one-on-one tutorial session, after which the user should be able to use any USI-compliant application. We are also experimenting with ways to effectively use more limited training time. One of our implementations is telephone-based, and includes a 90-second tutorial introduction (which advanced users can skip), which appears to do a reasonable job of telling users what they need to know in order to start using the system.

EXAMPLE
Our first USI application is a telephone-based interface to a database of current information about movies and theaters in the Pittsburgh area.

User (U) wants to know where Casablanca is playing:
1 U: Movie is Casablanca, theaters are what?
2 Movieline (M): Casablanca, two theaters: Showcase East, Waterworks Cinema.

User would like to find a comedy showing in Squirrel Hill:
local help (audio signals should often decrease the duration of output times [1,2], and, if used in place of lexical descriptions, signals in interfaces has been shown to increase response times [5,4]).

Another key feature of the USI approach is the use of non-linguistic signals to "pack the audio channel." Using audio signals in interfaces has been shown to increase response times [1,2], and, if used in place of lexical descriptions, audio signals should often decrease the duration of output messages, which can affect overall task completion times. We believe that audio signals can also help universalize applications and reinforce learning across applications. In the example above, <ellsig> (lines 4, 6, 16) is an ellipsis signal (currently implemented as three short beeps), which indicates that the list has not reached its end. <alert> (line 9) is used to warn users that something in their input was not recognized; this is currently implemented as a beep.<dadada> (line 4) is a spoken placeholder signal, indicating that the user can fill in a value here.

PRELIMINARY RESULTS
In hope of finding a preliminary validation of our approach and obtaining useful feedback for refining it, we have performed two initial experiments.

In the first experiment, 15 subjects each used the USI movieline to find the answers to five movie information retrieval tasks. Our goal in this study was to make sure that the USI approach was usable in the most basic sense: did users understand the concepts of structured, phrasal input and keywords enough to complete basic tasks? All of the subjects were indeed able to complete tasks (although one needed the aid of a USI "cheat sheet"); nearly all the users formed correct USI-style commands within their first three utterances. One of the most informative results of this study was the need for explicit confirmation. This was not included in our first version, and resulted in users often not being sure how to correct errors since they were not sure what the system had and had not understood.

Our second experiment has so far provided us with anecdotal feedback only. Five subjects used the USI movieline to find the answers to five movie information retrieval tasks. Our goal in this study was to make sure that the USI interface was usable in the most basic sense: did users understand the concepts of structured, phrasal input and keywords enough to complete basic tasks? All of the subjects were indeed able to complete tasks (although one needed the aid of a USI "cheat sheet"); nearly all the users formed correct USI-style commands within their first three utterances. One of the most informative results of this study was the need for explicit confirmation. This was not included in our first version, and resulted in users often not being sure how to correct errors since they were not sure what the system had and had not understood.

Our second experiment has so far provided us with anecdotal feedback only. Five subjects used the USI movieline to find the answers to five movie information retrieval tasks. Our goal in this study was to make sure that the USI interface was usable in the most basic sense: did users understand the concepts of structured, phrasal input and keywords enough to complete basic tasks? All of the subjects were indeed able to complete tasks (although one needed the aid of a USI "cheat sheet"); nearly all the users formed correct USI-style commands within their first three utterances. One of the most informative results of this study was the need for explicit confirmation. This was not included in our first version, and resulted in users often not being sure how to correct errors since they were not sure what the system had and had not understood.

Our second experiment has so far provided us with anecdotal feedback only. Five subjects used the USI movieline to find the answers to five movie information retrieval tasks. Our goal in this study was to make sure that the USI interface was usable in the most basic sense: did users understand the concepts of structured, phrasal input and keywords enough to complete basic tasks? All of the subjects were indeed able to complete tasks (although one needed the aid of a USI "cheat sheet"); nearly all the users formed correct USI-style commands within their first three utterances. One of the most informative results of this study was the need for explicit confirmation. This was not included in our first version, and resulted in users often not being sure how to correct errors since they were not sure what the system had and had not understood.

ACKNOWLEDGMENTS: We thank the Pittsburgh Digital Greenhouse for providing seed funding for this project.

REFERENCES
2. Leplâtre, G. and Brewster, S. Designing non-speech sounds to support navigation in mobile phone menus, in Proceedings of ICAD ’00 (Atlanta, GA, April 2000)