SOFT MACHINE: A PERSONABLE INTERFACE

John Lewis and Patrick Purcell
Architecture Machine Group
Massachusetts Institute of Technology
Cambridge MA 02139

ABSTRACT

The man-computer interface has evolved from the teletypewriter to workstations which accomplish a spatial metaphor for data management. In the experimental system described here, speech and graphics techniques are used to produce an interface in the form of a metaphorical person. Interaction takes the form of an unconstrained spoken conversation with a "graphical robot" whose animated likeness is displayed on a high-resolution computer graphics display. This system is proposed as a prototype of "casual interface" for machines which we do not use often enough to justify learning a command syntax. The realization of such systems assumes the development of limited-vocabulary speaker-independent continuous speech recognizers. The system architecture, performance, and assumptions are discussed.

KEYWORDS: computer graphics, human-machine interface, natural language interfaces, speech recognition, animation of persona.

Introduction

The extent to which metaphors of existing systems have influenced computing is suprising but undeniable. The computer interface modelled on the teletypewriter restricted the interface to a single active line of characters, resulting in line editors, line-oriented languages, and 'linear' (one event at a time) command languages.

Graphic display terminals generalize the potential "interface space" from the line of characters to a planar visual spece. Pioneering work at the Xerox Palo Alto Research Center used high resolution bit-mapped displays to create a spatial metaphor in which concurrent processes are seen as spatially separate objects located on a "desktop" display [1]. The Spatial Data Management System explored navigation through a "dataland" where the data could consist of color images, movies, and sounds, as well as iconic (quality font) representations of alphanumeric data [2].

Beyond the developments provoked by spatial metaphors, other uses of graphic displays are slower in evolving. One can speculate, for example, that color, depth representation, and iconography may better represent aspects of program structure and function such as scope, binding, and concurrent process execution.

In this work the power of the metaphor is explicitly acknowledged and a new metaphor is created. The title of this paper derives from the use of the word "soft" to describe computer interfaces whose design considers the computer user. Certainly the softest interface is another person--a programmer who can be instructed to accomplish the desired task.

Soft machine demonstrates an interface to several conventional computer programs (data-base query, electronic mail) which resembles interaction with a person. The interaction is in the form of an unconstrained spoken conversation with an electronic conversational partner (alter ego). The alter ego is an animated person-likeness which "speaks" with a high-quality speech synthesizer and "listens" with a continuous speech recognizer (Fig. 1). The animated image reflects the activity of the alter ego: the head motion of the likeness is consistent with attentive listening and the lips move with the alter ego's speech. The alter ego's conversational interests and visual character are easily personalized, suggesting 'alter ego' for this conversational partner.

Conversational Technique

The alter ego's conversational ability resides in a pattern-matching script. Patterns consist of ordered lists of words, optionally separated by wildcards (which match anything) or digits (which match a particular number of arbitrary words). Each pattern is associated with a list of responses or response procedures. Patterns are arranged in ascending order of

generality. The output of the speech recognizer is tested against the patterns until a match is found. One of the associated responses is selected and either spoken or executed (the lisp EVAL function is used to execute program fragments within the script). As an example, the user might say

"why are you taking so long?"

The corresponding output of the speech recognizer is

(? are you ? ? ?)

which matches the pattern

(1 are you *)

and provokes the response

"Do you think I am?"

The speech recognizer used in this work returns up to five words recognized from continuous speech (one or two words per sentence is typical performance) with indication of words which are heard but not recognized.

The conversational script is similar to the Eliza program [3] though less sophisticated. The purpose of soft machine's conversational ability is not to simulate conversation with another human but to complete the personal metaphor while functioning as a sophisticated means of conversational prompting and system feedback. Thus, script responses serve to indicate the system's misunderstanding in a friendly and interesting way, and to encourage the user's continued and uninhibited interaction.

A variety of statements replace the usual "?Unrecognized Command", helping to overcome the frustration which often results from the limitations of current speech recognition systems [4]. The user's inquiry or expression of frustration may result in the amusing "clearly this is an existential matter" or the potentially constructive "we were reading your mail" rather than the illegal command syndrome.

Unconstrained (amodal) conversation was a major goal of this work. The user can adress any application program at any point; the currently-active program is interrupted and its state is saved. Amodal conversation is facilitated by the limited vocabulary of speech recognizers. The vocabulary is carefully selected to consist of words which may be used in more than one application, as well as general words which support the conversation. Each word has a context list, and recognized speech is examined by a context monitor program to determine which to program it is addressed. Thus, 'next' may refer to the next piece of mail or the next database entry, or (as a non-sequiteur) become the object of a conversational remark. The context monitor uses a number of obvious heuristics to determine the applicability of the recognized words and to discard words which may be spuriously recognized or are currently inapplicable. Application programs maintain a status message which is spoken if no applicable input is recognized in the course of several conversational exchanges.

A contrived but realistic example of a conversation with Soft Machine follows (recognized words are italicized):

user: *Hello*
machine: Welcome to the machine.

u: *I* want to see the Rotch database.
m: Speak up.

u: *Show me* the *Rotch* database.
m: The Rotch slideathon is available.

u: *Show me* buildings in *Massachusetts, 1950* to *1960.*
m: Confirm state Massachusetts.

u: *OK*
m: Confirm dates 1950 to 1960.

u: *OK*
m: Initiating search
Ready.

u: Howmany?
m: What are you talking about?

u: *How* many?
m: There are 646 entries.

u: *Show* them.
m: (first slide appears)

u: *Next* slide.
m: (second slide appears)

u: Read *my* mail.
m: How do you know it is yours?

u: *Shut up.*
m: You do not like me very much.

u: Read *my mail.*
m: You have no unviewed messages.
Last message dated .... (the most recent message is read).

Real Time Animation of Electronic Persona

The text output by application programs is returned to the alter ego, who reads it to the user. The text is first massaged to translate written conventions into their verbalizations, for example, the electronic mail address "godzilla@mit-pamela" becomes "godzilla at M.I.T. pamela". The text is then output to a high-quality speech synthesizer (Prose 2000).

The speech synthesizer's audio output is simultaneously sent to speakers and analyzed by a real-time formant-tracking computer (Fig. 1). Ten formant configurations are selected.

corresponding to visually distinct lip positions. A 1024x1024x8 frame buffer stores up to eighty 128x96x8 images of lip and head movements. Currently two sequences of ten images provide "positive emotion" and "negative emotion" lip positions. The formant tracker reorigins and zooms the appropriate frame buffer images, so that the alter ego "speaks". This "lipsync" technique was developed at the Architecture Machine Group for limited bandwidth teleconferencing applications [5].

The basic lipsync technique is extended to include limited expression and head movement. When the alter ego is not speaking its head moves in a subtle but animated way characteristic of attentive listening. The remaining frame buffer images (those which are not used for lip positions) include eye and head positions. A transition matrix describes coherent random sequencing of these images.

Evaluation

Though a system such as this one would make a cumbersome interface to a text editor (for example), it is an attractive interface to a machine which one uses occasionally and does not wish to know in detail. An on-line library catalogue search facility is an example. In one library a catalogue-search terminal was installed adjacent to the card catalogue. Library users were either intimidated by the computer terminal or reluctant to learn its command syntax, and the terminal was removed for indirect access via information-desk personnel (Eisenhour library, Johns Hopkins).

Soft machine is fun to use. Novice users are facinated by its conversational ability and willingly explore the system. The combination of animation and speech techniques succeed in creating an animated persona.

Suprisingly, a limited vocabulary of several hundred words, if carefully chosen, is quite adequate to support both applications and an interesting conversational capacity. This is because the system does not need to understand all of what is said, but only to recognize words which may be meaningful to application programs, and provide a reasonable conversational response if none are found ("limited recognition").

Appropriate applications have a limited set of commands. Some information retrieval systems are appropriate both in this requirement and in being systems which one might casually encounter. A videodisk-based architectural database was interfaced to Soft Machine. Its input vocabulary of about 100 words comprising several independent keys (state, date in decades, building type) allows access to all of the 5000 entries in the database.

The speech recognizer is nevertheless the weak link in the system. The conversational interface presumes the ability to reliably identify vocabulary words embedded in speech containing a large percentage of words which are not in the vocabulary. Current continuous speech recognizers perform poorly at this task. In addition, continuous speech recognizers are usually speaker dependent, requiring retraining with each new user. It has been recognized that the parsing of natural language often requires understanding, which in turn may require human-like world knowledge and intelligence. The recent development of finite state (phonemic) probability driven recognition models argues that incremental developments in speech recognition will continue however [6], and a limited-domain speaker-independent recognizer capable of realizing an interface such as Soft Machine may become available well before the problems of language understanding are resolved.

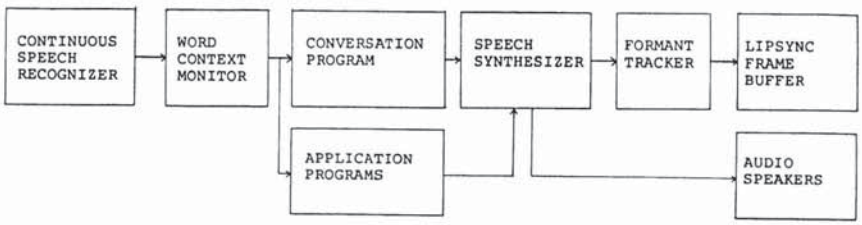

Figure 1: system flowchart

REFERENCES

1. Kay, A. and A. Goldberg, "Personal dynamic media". Computer, March 1977.

2. Bolt, R.A., Spatial Data-Management. DARPA Report. MIT Architecture Machine Group, Cambridge MA, March 1979.

3. Weizenbaum, J. "ELIZA--A computer program for the study of natural language communication between man and machine". Communications of the ACM, Vol.9 No. 1, pp. 36-45.

4. Schmandt, C. "Voice interaction: putting intelligence into the interface". Proceedings, 1982 Conference on Cybernetics and Society, IEEE, Seattle, 1982.

5. Negroponte N., "Talking heads--display techniques for persona". Unpublished paper. M.I.T. Architecture Machine Group, Cambridge Massachusetts.

6. Schwartz, R., Y. Chow, S. Roucos, M. Krasner, and J. Makhoul, "Improved hidden Markov modeling of phonemes for continuous speech recognition". 1984 International Conference on Acoustics, Speech, and Signal Processing (IEEE, 1984).