Expression constraints in multimodal human-computer interaction

Sandrine Robbe-Reiter*, Noëlle Carbonell*, Pierre Dauchy**

*LORIA, BP 239,F54506
Vandœuvre-lès-Nancy Cedex, France
**IMASSA-CERMA, BP 73, 91223 Brétigny-sur-Orge Cedex, France
carbo@loria.fr, pdauchy@imassa.fr

ABSTRACT

Thanks to recent scientific advances, it is now possible to design multimodal interfaces allowing the use of speech and gestures on a touchscreen. However, present speech recognizers and natural language interpreters cannot yet process spontaneous speech accurately. These limitations make it necessary to impose constraints on users' speech inputs. Thus, ergonomic studies are needed to provide user interface designers with efficient guidelines for the definition of usable speech constraints.

We evolved a method for designing oral and multimodal (speech + 2D gestures) command languages, which could be interpreted reliably by present systems, and easy to learn through human-computer interaction (HCI). The empirical study presented here contributes to assessing the usability of such artificial languages in a realistic software environment. Analyses of the multimodal protocols collected indicate that all subjects were able to assimilate rapidly the given expression constraints, mainly while executing simple interactive tasks; in addition, these constraints, which had no noticeable effect on the subjects' activities, had a limited influence on their use of modalities.

These results contribute to the validation of the method we propose for the design of tractable and usable multimodal command languages.

Keywords

Multimodal user interfaces, speech constraints, usability

CONTEXT, MOTIVATION, AND OBJECTIVES

Context and motivation

The evolution of human-computer interfaces is speeding up thanks to the development of new interaction modalities. Recent advances in speech and gesture interpretation make it possible to consider the design of input interfaces affording users spontaneous speech and gesture interaction. Such multimodal interfaces should come up to the expectations of most users, especially the general public, in-as-much as they emulate human communication.

However, some empirical results question the adequacy of human communication as a reference model for the design of HCI [1].

IUI 2000 New Orleans LA USA

Besides, spontaneous speech associated with gestures cannot yet be considered as a reliable substitute for direct manipulation, since the interpretation of linguistic reference mechanisms, such as anaphora and deixis, is still a research challenge. Therefore, designers of next generation oral or multimodal interfaces will have to define suitable speech constraints, in order to afford users reliable and usable interaction facilities. Note that "multimodal(ity)" refers to the alternate or joint use of speech and gestures, here and subsequently.

The use of speech as an input modality has motivated many empirical or experimental studies addressing a wide range of issues: the acceptability of speech input control [5, 9], the usability of speech versus mouse or keyboard [2, 4], or the integration of speech in multimodal input interfaces [6, 11]. However, only one experimental study [9] focuses on the ergonomic evaluation of various speech constraints.

Objectives

As it seems unrealistic to consider right now the development of user interfaces capable of processing spontaneous speech reliably, we have evolved a method for the design of ***"acceptable"*** oral or multimodal interaction languages. In other words, a language $\mathcal{L}$ is acceptable if:

- Present speech recognizers and interpreters are capable of processing accurately any utterance in $\mathcal{L}$;
- $\mathcal{L}$ is usable [10]; in particular, novice users can master $\mathcal{L}$ easily and rapidly, mainly in the course of HCI.

The major aim of the empirical study presented here is to assess the acceptability of a multimodal artificial command language, $\mathcal{LA}$, which we designed according to our method. The evaluation is focused on determining whether or not $\mathcal{LA}$ meets the second acceptability requirement, that is, whether:

- potential users from the general public can rapidly master $\mathcal{LA}$, while interacting with a standard graphical software package;
- speech constraints induced by $\mathcal{LA}$ will reduce neither the efficiency of their interactions nor their satisfaction.

We first present the method we propose for designing acceptable oral command languages. Then, after a brief description of the experimental setup, empirical findings are stated and discussed. Comments focus on the evolution of the subjects' behaviors in the course of the experiment.

DESIGN OF ACCEPTABLE SPEECH CONSTRAINTS

The method we propose for the design of acceptable speech constraints is based on the three following assumptions:

- H1: it has been observed that verbal exchanges between cooperating operators are limited to a restricted subset of natural language, the size of which varies according to the complexity of the task domain.
- H2: users will easily comply with speech constraints, provided that the subset of natural language (NL) defined by these constraints includes all the necessary commands for achieving efficient HCI.
- H3: in the context of H2, the elimination of ambiguity and synonymy is an acceptable speech constraint.

An efficient approach for defining such subsets of NL is to collect spontaneous speech interactions between potential users and a given application package, then to eliminate all ambiguous or synonymous words and structures from this subset. In-as-much as it satisfies hypotheses H2 and H3, the resulting command language should be easy to learn; and speech interpreters should be able to process any of its utterances reliably, provided that the complexity of the given application domain and tasks is not too high.

The automatic interpretation of utterances will be further facilitated if such a language is used in a multimodal environment, where linguistic references to objects and locations on the screen can be replaced by deictics associated with designation gestures.

EXPERIMENTAL SETUP

Overview

Eight subjects participated in this study, EC, and interacted during three weekly sessions [1] with a simulated multimodal user interface, using constrained speech and 2D gestures on a touchscreen. We organized three well spaced out sessions, in order to be able to study the possible evolution of the subjects' behaviors under the influence of practice.

Tasks

Subjects performed simple tasks relating to furniture arrangement. They had to design or modify the layout of various furnished rooms according to instructions specified in eight scenarios of increasing complexity. Initial layouts were displayed in the form of 2D plans. Graphical representations of the various pieces of furniture could be moved, using constrained speech and/or gestures. Execution of the various tasks involved: simple actions/commands (move object x, cancel or redo action), and complex ones (turn x, permute x and y, adjust position of x).

Such tasks require no specific skill or knowledge; therefore, the potential evolutions of the subjects' styles of interaction can be safely interpreted as effects of the cognitive processes involved in the assimilation of speech constraints.

Subjects

Eight volunteer subjects [2] participated in the experiment. All of them had already used software intended for the general public and were familiar with direct manipulation. Several had some programming skills, but none of them was an expert in computer science.

Simulation of the multimodal user interface

We designed a sophisticated implementation of the Wizard of Oz paradigm in order to simulate the multimodal user interface. Resorting to a software prototype would have made it impossible to compare the data collected during this study with empirical data on the spontaneous use of speech and gestures, collected earlier (see experiment EF below) using the Wizard of Oz technique out of necessity [3].

Two human operators were in charge of the simulation of the interface functionalities, without the subject's knowledge, wizards and subject working in separate rooms. One of the wizards interpreted incoming commands and activated relevant software functions; the results of his actions being displayed on both the subject's and the wizards' screens. The other one interacted verbally with the subject, using a set of pre-recorded oral messages, so that the system outputs were multimodal.

The setup also included a mono-speaker continuous speech recognizer on the market (Datavox), since the wizards could not decide in real time whether the subject's utterances complied with the specified speech constraints. They relied instead on the results of the recognition system, which were displayed both on the subject's and the wizards' screens.

Interaction language

We applied the method presented above for defining the artificial multimodal command language EC subjects had to use. It is a subset of all the commands expressed by other subjects [8], who could use speech and 2D gestures freely in a setup, EF, otherwise identical to the EC setup.

Besides, all the necessary commands were expressible easily using speech or gestures or both modalities; this extra requirement makes it possible to gain an insight into subjects' preferences regarding modalities.

The speech component of the language used in the context of EC is characterized as follows:

- its vocabulary includes about a hundred words, and its syntax can be described by a CF grammar (static and dynamic branching factors: 5.5 and 2.6, respectively);
- its semantics is the union of the meanings of all the spontaneous utterances collected during EF;
- it is free from synonymy, polysemy and ambiguity.

The gesture component comprises two types of elementary 2D gestures: designation and simulation gestures, the latter miming translations and rotations of icons on the screen. Designation gestures are not synonymous with simulation gestures, in-as-much as these types of gestures correspond to two different outlooks on HCI: communication versus manipulation [8]. Ambiguous or synonymous gestures were eliminated from both sets of gestures used by EF subjects.

Finally, the multimodal component includes all possible combinations of allowed gestures and utterances.

---

[1] Each session lasted half an hour on average, per subject.

[2] Five women and three men (from 23 to 57 years of age), all engaged in occupational activities.

[3] The accurate processing of spontaneous speech inputs being beyond the capabilities of available systems.

EC subjects were given a written description (i.e. a list of instances) of this language. The experimenter assisted them while they performed a small set of predefined commands during a short (5 to 10 min.) initial training.

Recordings and transcripts

Subjects were videotaped throughout the experiment. Written descriptions of the recordings comprise orthographic transcripts of verbal exchanges, together with standardized descriptions of subjects' gestures and system actions. Speech recognition results were also included. Utterances, gestures, system actions and speech recognition results were written down in chronological order and dated.

RESULTS AND INTERPRETATIONS

We first analyze whether HCI helped EC subjects to assimilate the given speech constraints. Then, we present and discuss results which give some insight into the influence of these constraints on the subjects' interactions with the application package, and on their use of modalities.

Assimilation of speech constraints

In the course of EC, the percentage of "incorrect" utterances [4] over the total number of utterances per session decreases from 31.1% (session 1) to 21.9% (session 3), despite a marked increase in speech recognition error rates. In addition all individual percentages decrease during the experiment, despite great inter-individual variations (from 16.3% to 75% during the first session). A qualitative analysis of incorrect utterances suggests that these inter-individual differences might stem from the diversity of individual linguistic/verbal abilities.

These results indicate that subjects' interactions with the simulated user interface helped them to learn the linguistic constraints they had to comply with. They also contribute to demonstrate the positive influence of HCI on the assimilation of linguistic constraints defined according to our design method.

Effects of speech constraints on subjects' activities

In order to elicit the possible influence of expression constraints on subjects' activities and styles of interaction, we first compared the multimodal protocols recorded during the first session of this study (EC) with those collected during the first session of our earlier empirical study (EF). There is no statistical evidence that expression constraints interfered with EC subjects' activities, and reduced the efficiency of their interactions with the application [12].

The analysis of command complexity throughout EC is also useful for assessing the possible influence of speech constraints on the efficiency of subjects' interactions with the given application package. In particular, it may provide some insight into the effects, on the subjects' cognitive workloads, of their efforts to comply with the speech constraints imposed on their spontaneous oral expression.

Complex [5] commands are sparingly used during the first session, since only 20% of the initial formulations expressed during this session are complex. The difference between simple [6] and complex commands is statistically significant ($t=2.27$; $dll=14$; $p<0.05$).

The limited use of complex commands during the first session suggests that most subjects had some difficulty in exploiting, from the start, all the functionalities of the simulated user interface.

Results relating to subsequent sessions confirm this interpretation: the percentage of complex initial formulations over the total number of initial formulations, which amounts to 18.6% during the first session, reaches 28.1% during the second one, and 34.9% during the last one. This evolution concerns all subjects; there is a statistically significant difference between the averages of the complex initial formulations expressed during the first and last sessions ($t=4.14$; $dll=14$; $p<0.01$).

Finally, most of the complex commands expressed by subjects are rotations, which cannot be formulated in terms of sequences of simple moving commands. This observation suggests that, whenever possible, many subjects preferred to split up complex commands into several simpler ones (cf. permutations) rather than resort to such commands, even at the expense of a loss of efficiency.

On the whole, these findings suggest that subjects "learned" the functionalities of the interface progressively: they first used simple, intuitive functions, then resorted to increasingly complex and powerful commands.

We observed similar behaviors in the context of our earlier empirical study, where subjects could use speech and gestures spontaneously [8]. Therefore, the evolution of command complexity during EC may be interpreted as an effect of the learning processes involved in the discovery of a new application software package, rather than as a possible negative effect of speech constraints. Thus, these constraints are unlikely to have reduced the efficiency of EC subjects' interactions with the simulated user interface.

Use of modalities

How subjects used modalities is also one of the main sources of information about their reactions to speech constraints.

Speech is the modality preferred by most subjects during the first session. 57.7% of all the commands expressed during this session are oral, 33.6% gestural, and 8.7% multimodal. But these differences are not statistically significant by reason of pronounced inter-individual variations ($t=0.64$; $dll=14$; $p<0.01$).

As for the use of multimodal commands, only one subject (S7) resorted to them frequently: 26.7% of his commands are multimodal.

During sessions 2 and 3, the use of speech decreases for 7 subjects out of 8, while the percentage of gesture commands (over the total number of commands) increases drastically: from 33.6% (session 1) to 66% (session 3). The evolutions of the use of speech and gestures between the first and last sessions are linearly correlated ($r=0.9$; $dll=5$; $p<0.01$).

As for multimodality, its use is limited during the three

[4] Utterances which do not belong to the oral component of the command language used in EC are judged incorrect.

[5] i.e. rotations, permutations and adjustments.

[6] i.e. moving/positioning (object), cancel and redo (actions).

sessions: 7.5% and 5% of the total number of commands (sessions 2 and 3 respectively).

The increasing use of gestures and the correlated decreasing use of speech may safely be interpreted as an outcome of the worsening in speech recognition rates with time. Therefore, the discussion will focus on the extensive use of monomodality in preference to multimodality.

Quantitative analyses suggest that most subjects preferred monomodality to multimodality. This finding seems to contradict one of the major results of the empirical study which S. Oviatt et al. performed on the use of speech and pen for consulting and modifying maps [11], that is, the fact that participants in their study resorted to multimodality quite often.

This difference may be ascribed mainly to the influence of expression constraints, as subjects who participated in S. Oviatt's study could use speech and gestures freely. This interpretataion is further supported by the fact that 37% of the commands expressed by EF subjects during the first session, versus 8.7% for EC subjects, were multimodal.

OVERALL CONCLUSION

One of the major goals of the empirical study presented here was to assess, in a standard HCI environment, the usability of an artificial multimodal command language, and thus to gain some insight into the efficiency of the method we used for defining this language.

This method aims at defining speech constraints which should be easy to master during interaction, and should facilitate the automatic interpretation of speech inputs.

The user interface was simulated, using the Wizard of Oz technique for processing speech and 2D gesture inputs. Subjects performed simple graphical design tasks during three weekly sessions, using the artificial interaction language we had defined.

The originality of this empirical study is twofold. Although multimodal HCI has motivated numerous studies, only one of them [8] focuses on the ergonomic evaluation of speech constraints, and none attempts to observe the evolution of subjects' behaviors with time – at least to our knowledge.

The evolution of the number of incorrect oral commands during the experiment, indicates that the speech constraints EC subjects had to comply with can be assimilated easily, mostly through interacting with the user interface.

Moreover, these constraints did not significantly reduce the efficiency of the subjects' interactions with the application package.

On the other hand, they may have affected their use of multimodality; multimodal commands being much more frequent in interaction contexts where spontaneous speech and gestures are possible. This finding suggests that speech constraints may increase users' cognitive workload, and also that multimodality might be more costly, in terms of cognitive workload, than monomodality. However, these tentative conclusions cannot be accepted without further validation. Extensive empirical or experimental research, especially ergonomic evaluation studies involving actual users in real interaction environments, are needed.

These results contribute to validating the acceptability of the artificial multimodal command language we designed. Therefore, they contribute to evaluate the adequacy of our method for designing acceptable multimodal command languages, which could prove appropriate substitutes for direct manipulation in contexts where the use of mouse and keyboard is awkward or impossible.

ACKNOWLEDGEMENTS: This work was partly supported by the French Ministry of Defense (contract DGA/DRET n° 95-125)

REFERENCES

1. Amalberti, R., Carbonell, N., and Falzon P. User representations of computer systems in human-computer interaction. *International Journal of Man-Machine Studies*. 38 (January 1993), 547-566.
2. Bekker, M.M., Van Nes, F.L., and Juola, J.F. A comparison of mouse and speech input control of a text-annotation system. *Behaviour & information Technology. 14*, 1 (1995), 14-22.
3. Coutaz, J., and Caelen, J. A taxonomy for multimedia and multimodal user interface. *Proceedings 1st ERCIM Workshop on Multimodal Human-Computer Interaction* (Lisbon, November 1991), INESC.
4. Damper, R.I., and Wood, S.D. Speech versus keying in command and control applications. *Int. Journal of Human-Computer Studies*. 42 (1995), 289-305.
5. Dillon, T.W., and Norcio, A.F. User performance and acceptance of a speech input interface in a health assessment task. *International Journal of Man-Machine Studies*, 38 (January 1993), 547-566.
6. Hauptmann, A.G., and McAvinney, P. Gestures with speech for graphic manipulation. *International Journal of Human-Computer Studies*, 47(4, 1997), 591-602.
7. Koons, D. B., Sparrell, C.J., and Thorisson, K.R. Integrating simultaneous input from speech, gaze and hand gestures. in M. Maybury (Eds.), *Intelligent Multimedia Interfaces*. MIT Press, 257-276, 1993.
8. Mignot, C, and Carbonell, N. "Natural" multimodal HCI: Experimental results on the use of spontaneous speech and hand gestures. *Proceedings 2nd ERCIM Workshop on Multimodal Human-Computer Interaction* (Nancy, November 1994), INRIA, 97-112.
9. Murray, A. G., Jones, D. M., Frankish, C.R. Dialogue design in speech-mediated data-entry: the role of syntatic constraints and feedback. *International Journal of Human-Computer Studies*. 45 (3, 1996), 263-286.
10. Nielsen, J. *Usability Engineering*. Academic Press, 1993.
11. Oviatt, S., DeAngeli, A., and Kuhn, K. Integration and synchronisation of input modes during multimodal human-computer interaction. *Proceedings CHI'97* (Atlanta, April 1997), ACM Press, 415-422.
12. Robbe, S., Carbonell, N., and Dauchy, P. Constrained vs spontaneous speech and gestures for interacting with computers: A comparative empirical study. *Proceedings INTERACT'97* (Sydney, July 1997), Chapman & Hall, 445-452.