TOWARDS USABLE MULTIMODAL COMMAND LANGUAGES:
DEFINITION AND ERGONOMIC ASSESSMENT OF CONSTRAINTS ON USERS' SPONTANEOUS SPEECH AND GESTURES

S. Robbe, N. Carbonell*, C. Valot***

*CRIN-CNRS & INRIA-Lorraine, BP. 239, F54506 Vandoeuvre les Nancy Cedex
**IMASSA-CERMA, BP 73, F91223 Brétigny sur Orge Cedex
*Tel. 33 3 83 59 20 47 FAX: 33 3 83 41 30 79, E-mail: {robbe, carbo}@loria.fr
**Tel. 33 1 69 88 33 70 FAX: 33 1 69 88 33 75, E-mail: claude@cerma.fr

ABSTRACT

Within the framework of a prospective ergonomic approach, we simulated two multimodal user interfaces, in order to study the usability of constrained vs spontaneous speech in a multimodal environment. The first experiment, which served as a reference, gave subjects the opportunity to use speech and gestures freely, while subjects in the second experiment had to comply with multimodal constraints.

We first describe the experimental setup and the approach we adopted for designing the artificial command language used in the second experiment. We then present the results of our analysis of the subjects' utterances and gestures, laying emphasis on their implementation of linguistic constraints. The conclusions of the empirical assessment of the usability of this multimodal command language built from a restricted subset of natural language and simple designation gestures is associated with recommendations which may prove useful for improving the usability of oral human-computer interaction in a multimodal environment.

1 CONTEXT AND MOTIVATION

Thanks to recent research advances, speech recognizers are now capable of processing large subsets of natural language (NL) accurately. Nevertheless, spontaneous speech cannot yet be considered as a reliable substitute for artificial query/command languages, menu-driven Human-Computer Interaction (HCI) or direct manipulation, since the interpretation of linguistic reference phrases, especially anaphoric and spatio-temporal phrases, raises still numerous research issues.

An attractive solution for achieving robust quasi-"natural" HCI in the near future is to design multimodal languages that allow users to combine oral commands from a restricted subset of NL, with pointing gestures. Such languages are tractable, since multimodal spatial reference phrases (i.e. deictics associated with pointing gestures) can be reliably processed by present NL and gesture interpreters [4]. They may then supersede present forms of HCI, provided that their utility and usability (cf. [3]) is demonstrated beyond doubt.

We are currently investigating, within a prospective ergonomic research framework, utility and usability issues raised by the implementation of such languages. The main goal of the comparative empirical study reported here is to assess the effects of realistic expression constraints on the behaviours and attitudes of potential users of forthcoming multimodal interfaces integrating speech and gestures. This study addresses the following major issue:

Is it possible to define expression constraints which can both restrict users' spontaneous speech and gestures to a tractable sub-language without interfering with their activity, and be mastered easily in the course of interaction? And if so, how could such constraints be determined?

To answer these questions, we considered two experimental situations: in the first one, which served as a reference, subjects could use speech and gestures freely, whereas in the second one they had to comply with multimodal expression constraints.

While constrained oral HCI has motivated numerous experimental and empirical studies (cf. for instance, [2], [6]), less attention has been paid to the usability of speech in a multimodal HCI environment. In addition, no comparative study of constrained vs unconstrained oral interaction has been published thus far, at least to our knowledge, and the method used for defining expression constraints is original.

2 EXPERIMENTAL SETUP

Two groups, of eight subjects each, interacted during three weekly sessions (of about half an hour per subject each) with two different multimodal interfaces whose functionalities were simulated thanks to the Wizard of Oz technique (WOZ). Both groups carried out identical design tasks; but subjects in the reference group (SP) could use speech and/or 2D gestures (on a touch panel) spontaneously, while subjects in the experimental group (CS) had to comply with expression constraints.

2.1 Expression constraints

In order to obtain a tractable multimodal artificial language that would impose minimum constraints on the spontaneous expression of CS subjects, we selected an appropriate subset of the overall set of utterances and gestures used by SP subjects.

Two categories of elementary 2D gestures were allowed: pointing gestures, and simulation gestures (akin to mouse drags) for miming translations and rotations of icons on the screen. Ambiguous gestures were eliminated from the simple gestural « vocabulary » used by SP subjects.

The verbal component of the language consists in a restricted subset of NL with the following properties. Its syntax can be described by a CF grammar [1] defined on a hundred word vocabulary. Its expressiveness is equivalent to the union of the semantic interpretations of the oral commands issued by SP subjects over the three sessions. Synonymy and polysemy are excluded.

CS subjects were given a written description of this multimodal artificial language [2], and the experimenter assisted them while they performed a small set of predefined commands; this initial training stage lasted less than 10 minutes on average. On the other hand, SP subjects had no supervised initial training, but they could, before processing the first scenario, explore the capabilities of the interface in the presence of the experimenter who just answered their questions.

2.2 Application domain and subjects' tasks

Subjects had to design or modify furniture arrangements according to instructions specified in scenarios of increasing complexity. Initial furniture layouts were displayed on the screen in the form of 2D plans.

2.3 Implementation of the Wizard of Oz technique

Two human operators, hidden from the subjects, simulated the functionalities of both multimodal interfaces. One of them interpreted incoming commands and activated the corresponding software functions which were displayed on the subject's and the wizards' screens; the other one interacted verbally with subjects using a set of fifty or so pre-recorded oral messages. In addition, the CS setup included a commercial continuous speech monospeaker recognizer (Datavox) which the wizards used for interpreting subjects' utterances.

2.4 Recordings and transcripts

Subjects were videotaped throughout both experiments. Written descriptions of the recordings comprise orthographic transcripts of verbal exchanges and coarse standardized descriptions of subjects' gestures and system actions; subjects' speech and gestures were further characterized with a view to assessing the extent to which they succeeded in mastering the given set of expression constraints.

3 GLOBAL RESULTS

Results presented in sections 3 and 4 bear upon the first session only, since our present main objective (cf. section 1) is to assess the usability of artificial multimodal command languages designed according to the method presented in paragraph 2.1.

3.1 Expression constraints

Subjects in the CS group complied easily with gestural constraints: we picked out only three « incorrect » gestures in the transcripts. This result is not surprising since subjects were allowed to use a small number of simple intuitive gestures.

On the other hand, all subjects resorted to words outside the vocabulary, and six out of eight used NL structures outside the scope of the language; three subjects only resorted to incorrect (with respect to NL) syntactic structures, while five used NL words or phrases inappropriately. Subjects' reactions to linguistic and enunciation constraints are detailed in section 4.

3.2 Comparison between the CS and SP groups

An inter-group comparison suggests that subjects in the CS group benefited from the linguistic constraints with which they had to comply. Hesitations and grammatical errors are significantly [3] less frequent in their oral statements than in those from SP subjects: 13% vs 53% and 1.9% vs 23.5% respectively [4].

On the other hand, inter-group differences concerning the use of modalities are not statistically significant by reason of marked inter-individual variations (S_{CS}=67 W(8;8)=]49;87[; $p<0.05$). Detailed results of this comparative study are presented in [5].

4 SPEECH CONSTRAINTS

We describe here how CS subjects reacted to the linguistic and enunciation constraints they had to comply with. As behaviours and strategies vary greatly from one subject to another, we analyzed the CS transcripts subject by subject, with a view to defining accurate user profiles. Results are summarized in Table 1 and 2.

4.1 Inter-individual variations

Inter-individual differences affect many aspects of the verbal expression of subjects as shown in Table 1.

First, the total number (NBT) of oral (and multimodal) statements per subject ranges from 4 to 118 (cf. line 4).

Recognition rates (NRC/NBC) of the first formulations of correct commands (i.e. commands belonging to the

[1] static branching factor 5.5, dynamic branching factor 2.6

[2] We limited the description of the oral component of the language to the listing of its vocabulary and the presentation of a few commands instances; these instances were chosen so as to illustrate the structural flexibility and expressive power of the language as well as its syntactic and semantic limitations.

[3] Wilcoxon test: SCS=43 W(8;8)=]49;87[; $p<0.05$. We applied this test, since there was a significant difference between the variances for the CS and SP groups.

[4] Percentages represent numbers of relevant tokens normalized by the total number of oral and multimodal statements per group.

CF language) vary also greatly from one subject to another: from 21% of failures up to 57%.

	S1	S2	S3	S4	S5	S6	S7	S8
NBC	33	30	7	1	7	4	40	0
NRC	7	14	4	0	3	1	9	0
NBE	13	16	3	0	2	6	4	4
NBT	61	118	18	4	20	20	57	12
NTN	20	32	8	1	11	11	14	6
NRR	10	64	6	1	4	6	10	5
NRR/ NTN	0.5	2	0.75	1	0.36	0.54	0.71	0.83

Table 1. Correct and incorrect commands (CF language)

NBC: number of correct commands (∈ to the CF language)
NRC: number of unrecognized correct commands
NBE: number of incorrect commands (∉ to the CF language)
NBT: total number of statements
NTN: total number of unrecognized statements
NRR: number of repetitions/reformulations of unrecognized statements

This suggests that some brief initial training to the use of speech recognizers, together with appropriate online help, might improve significantly oral HCI.

This result also brings out the large extent of intra-speaker variability in the context of actual HCI situations. Voice interface designers should be aware of and take into account this variability, the processing of which should motivate further research efforts in the area of speech recognition.

As for the percentages of incorrect first formulations of commands, that is NBE/(NBC+NBE), they range from 9% (S7) to 60% (S6), or from 0% to 100% if subjects 4 and 8 are taken into account. This diversity points to the two following conclusions.

First, a short initial training stage (10 mn in our experiment) is not sufficient to master linguistic constraints, at least for some users; online help is obviously necessary for them. Online help may even prove beneficial to all users. It can facilitate the detection of errors, all the more so since current speech recognizers provide users with no useful information for deciding, when a recognition failure occurs, whether it should be ascribed to the inaccuracy of the recognizer or to their own incapacity to comply with some linguistic constraint. The low number of errors that subjects succeeded in correcting (5) confirms this interpretation.

Secondly, oral languages defined as restricted subsets of natural language should be used in a multimodal environment, so that users could resort to other media and modalities (gestures or typing, for instance) in cases of repeated recognition failures; in the context of our experiment, subjects often resorted to gestures in such situations. In addition, multimodality could be used by the interface manager to adapt dynamically the speech recognizer to intra-speaker variability. Online help could also take advantage of reformulations conveyed through a surrogate modality, and use such information to correct users' errors.

Finally, the strategies adopted by subjects for overcoming recognition failures can be inferred from the count of repetitions and reformulations (NRR) per subject.

Two main types of strategies, illustrated by S2 and S5 (or S6), emerge from the analysis of the values reported on line 6 in Table 1: while S2 reformulated or repeated unrecognized commands twice on average (before expressing them through gestures and multimodal commands), in such contexts S5 and S6 gave up their initial intentions (or resorted to gestural expression) in 2 instances out of 3 and 1 out of 2 respectively. Other subjects resorted to repetitions and reformulations slightly more often.

The fact that most subjects tended to switch to gestures rapidly in the context of recognition failures is a further argument for associating speech with other modalities in HCI environments.

4.2 Types of linguistic « errors »

We classified subjects' departures from the linguistic constraints imposed on them into two main categories and four sub-categories, namely:

- the use of French words (VOC) and French syntactic structures (SS) outside the command language;
- incorrect (∉ to French) syntactic structures (INC), and inappropriate use of lexical units or phrases (UA).

Results of the application of this classification to the CS subjects' oral statements over the first session are presented in Table 2.

	S1	S2	S3	S4	S5	S6	S7	S8	Total
SS	3	4	0	0	1	1	0	1	11
VOC	3	6	1	1	7	6	1	4	29
TOT1	*6*	*10*	*1*	*1*	*8*	*7*	*1*	*5*	*40*
UA	2	2	1	0	2	0	0	2	9
INC	0	0	0	0	0	1	2	1	4
TOT2	*2*	*2*	*1*	*0*	*2*	*1*	*2*	*3*	*13*

Table 2. Types of linguistic errors

SS:	NL syntactic structures outside the CF language
VOC:	words outside the CF vocabulary
TOT1:	SS + VOC
UA:	inappropriate words or phrases (in French)
INC:	incorrect syntactic structures (in French)
TOT2:	UA + INC

On the whole, despite inter-individual differences, CS subjects assimilated and complied with most of the linguistic constraints. This result demonstrates the feasibility of the approach we propose for defining tractable and usable oral or multimodal HCI languages,

(5) None of them succeeded in correcting spontaneously more than 30% of their errors.

that is the selection of appropriate restricted subsets of natural language.

The use of words outside the CF vocabulary represents the most frequent type of error in the transcripts.

The majority of words labelled VOC are more or less closely related to words belonging to the vocabulary of the CF language; the relationship is either one of synonymy or one of hyponymy. For instance, « *remplacer (replace) » [6] is synonymous with « permuter (permute) », while « *chambre (bedroom) » and « lit simple (single bed) » are hyponyms of « pièce (room) » and « *lit (bed) ». Such errors were easily detected and corrected spontaneously by subjects when the grammatical categories of both lexical units were identical. On the other hand, in the case of synonymous linguistic phrases with different structures, errors were seldom corrected (cf. for instance « *un peu (a little) » and « légèrement (slightly) »). Similarly, SS errors were less easily corrected when the infringed constraints were restrictions on grammatical categories [7].

We excluded synonyms from the CF language in order to simplify the recognition and interpretation of commands, improve the robustness of these processes, and facilitate the assimilation of linguistic constraints (by reducing the complexity of the language). But these results induce us to reconsider our a priori choices. Customization could prove a suitable trade-off: the user might tailor a standard vocabulary (without synonyms) to his liking by replacing some lexical items with words selected from predefined lists of synonyms. Thus, occurrences of words outside the CF vocabulary might be fewer.

As all subjects were French native speakers, one may be surprised by the fact that three of them made syntax errors (INC) and all of them used words in inappropriate semantic contexts. But most of these errors occurred within reformulations of unrecognized correct commands (i.e. belonging to the CF language).

Therefore, such errors might be eliminated or, at least, their number might be highly reduced, if users could be informed of the causes of recognition failures. The rate of any type of error would undoubtedly decrease significantly if the interface manager could tell users, after each recognition failure, whether this failure results from recognition limitations, enunciation faults or linguistic errors. In order to provide users with such feedback, speech recognizers should be endowed with efficient auto-diagnosis and explanation capabilities. Research efforts are necessary for achieving this objective since it raises issues which, to our knowledge, have not yet been investigated closely; solving these issues represents a stimulating scientific challenge.

5. CONCLUSION

We designed and performed two ergonomic experiments in order to assess the usability of constrained speech in a multimodal environment, and to validate the method used for defining the vocabulary and syntax of the corresponding artificial language.

Two user interfaces were simulated using the WOZ technique. In the first experiment, which served as a reference, subjects could use speech and gestures freely while, in the second one, they had to comply with multimodal constraints. Linguistic constraints were selected so as to define a restricted tractable subset of NL among the utterances of the subjects who participated in the first experiment.

On the whole, subjects who participated in the second experiment complied easily with these linguistic constraints, despite pronounced inter-individual differences as regards global numbers of utterances, command recognition rates and strategies for processing recognition failures. Nevertheless, for some subjects at least, a short preliminary training stage is not sufficient; they need online contextual help to assimilate and implement accurately the given set of linguistic constraints. In addition, error rates could be reduced significantly if users could be informed of the causes of recognition failures; in order to provide them with such feedback, research is necessary, present speech recognizers having but limited capabilities of auto-diagnosis and explanation.

REFERENCES

[1] Dauchy P., Mignot C., Valot C., *"Joint Speech and Gesture Analysis - Some Experimental Results on Multimodal Interfaces"* EUROSPEECH'93, 1315-1318, Berlin, 1993.

[2] Hauptmann A.G., McAvinney P., "Gestures with speech for graphic manipulation", *Int. Journal of Man-Machine Studies*, 38, 231-249, 1993.

[3] Nielsen J., "Usability engineering", San Diego (CA):.Academic Press, 1993.

[4] Nigay L., Coutaz J., *"A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion"*, Proc. INTERCHI'93, pp. 172-178, Amsterdam, 1993.

[5] Robbe S., Carbonell N., Dauchy P., *"Constrained vs spontaneous speech and gestures for interacting with computers: A comparative empirical study"*, Proc. INTERRACT'97, Sydney, 1997.

[6] Rudnicky A.I., Sakamoto M., Polifroni J.H., *"Spoken language interaction in a spreadsheet task"*, Proc. INTERACT'90, pp. 767-772, 1990.

Acknowledgements

We express our thanks to R. Amalberti and P. Dauchy, researchers at CERMA, for their significant contributions to the experimentation. This study was partially supported by the French Ministry of Defence (D.G.A./D.R.E.T. contract N° 95-125).

[6] Words, phrases or sentences which do not belong to the CF language are asterisked.

[7] For instance, subjects could say: « Jusqu'ici/là. (Up to here/there.) », but could not use the preposition 'jusque' in conjunction with a substantive, such as in: « *Jusqu'au mur. (Up to the wall.) ».