COOPERATIVE DIALOGUE AND MULTIMODAL INTERFACES

M.D. Wilson, G.A. Ringland, G. Wickler

SERC Rutherford Appleton Laboratory

Abstract

The MMI^2 ("Multi Modal Interaction with Man Machine Interfaces for knowledge based systems") demonstrator supports user interaction through English, French or Spanish natural languages, command language, by direct manipulation on a graphical display, through design gestures (as in PDAs) and non-verbal audio. The system is outlined within the MSM framework of Nigay and Coutaz (1993).

The purpose in developing the system was to demonstrate co-operative dialogue. Having developed the system several remaining problems are considered: 1) Is the interaction really multimodal, 2) How should any advantages of multimodal interaction be evaluated, 3) Can the apparent limitations in the system be overcome in the foreseeable future?

INTRODUCTION

In 1987 we wrote a proposal to the CEC for funding to develop a co-operative multimodal interface to knowledge based systems. In October 1991 the first prototype of the MMI^2 interface to support co-operative dialogue with a KBS in a computer network design task was demonstrated, in January 1993 a second system for monitoring and managing computer networks in real time was demonstrated. The design method was to collect examples of co-operative dialogue from the literature and from Wizard of Oz studies of network design, to develop a demonstration script including these examples and then to implement the system to be able to demonstrate that it could perform these examples (Wilson, 1991). Although the design method drew upon theory which promoted some generalisation from the examples, the collection of the requirements which such an interface should meet was very data and example based.

A major interface mechanism which we chose to aid in cooperative dialogue was multimodal interaction. This paper tries to place the advantages provided by multimodal interaction in the context of the limitations forced on a design by individual modes, by dialogue control and available data.

The work reported in this paper was partly funded by the CEC through Esprit project 2474, MMI^2. The partners in the project are BIM (Belgium), Intelligent Software Solutions S.A. (Spain), ADR Centre de Recherche en Informatique applique aux Sciences Sociales (France), Ecole des Mines de Saint-Etienne (France), Insitut National de Recherche en Informatique et en Automatique (France), SERC Rutherford Appleton Laboratory (UK), University of Leeds (UK).

OVERVIEW OF THE MMI^2 SYSTEM.

The integration of modes in MMI^2 was governed by four principles:

1) **there is a meaning representation formalism, common to all modes, which is used as a vehicle for internal communication of the semantic content of interactions inside the interface and also used as a support for semantic and pragmatic reasoning.**

2) **mode integration should mainly be achieved by an integrated management of a single generalised discourse context.**

3) **there are different model theories for the evaluation of symbols in the meaning representation formalism for the application domain, interface domain and user domain.**

4) **The effect of formally evaluating communication actions against a domain can cause side effects in the other domain.**

To produce co-operative dialogue it is still necessary to include a fifth principle:

5) **Informal processing of dialogue may be performed for user utterances (by the informal domain expert) and system output (by the communication planning expert).**

The architecture resulting from the application of the five principles is shown in figure 1 (Binot et al., 1990). In this, user interaction takes place through one of the presentation layer modes. These will produce, and take as input packets of information in the Common Meaning Representation (CMR) which describe the content of a communication action (by the user or the system) in a logical form, along with the mode used, the time the packet was created, and the force of the utterance (Doe et al., 1992). All operations in the dialogue management layer take place using this representation language.

The CMR is first passed to the dialogue controller which performs some presupposition checking and resolves references in conjunction with the Context Expert. It will be passed to the User Model which will derive information about the user's preferences, knowledge and misconceptions from it. It will then be passed to the informal domain expert to determine if there are any pragmatic problems with the communication action. After this it will be formally evaluated against the Formal Domain Expert or the Interface Expert to perform the command or find the answer to a question. The Formal Domain Expert will interact with the underlying knowledge base which knows about network design. The interface expert interacts with the interface itself and knows about interface objects (e.g. bar-graphs, windows, command details).The answer to the question or response to the command resulting from formal evaluation will then be passed to the communication planner which will produce a reply to the user. Similarly, if the informal domain expert identifies any pragmatic problems these will also be passed to the communication planner to form the system's next move in the dialogue with the user. The system output is created as CMR packets which will be passed to the dialogue controller and then out to the appropriate mode (named in each CMR packet) for presentation to the user.

Figure 1: Architecture of the MMI^2 System.

IS MMI^2 A MULTIMODAL SYSTEM?

Nigay and Coutaz (1993) present the MSM framework which differentiates multi-modal from multi-media systems. MMI^2 uses a single modality by their definition since user actions are through the hands (typing command language or natural language, moving a mouse for direct manipulation or drawing gestures), and all system actions are through the screen in the form of text and graphics (except minimal use of non-speech audio for system output). Since it uses text and graphics MMI^2 uses multiple modes in their terminology. It would be trivial to add a speech interface to the command language to classify the system as multi-modal as well.

It is quite possible to introduce a single keyword speech interface onto a command language interface. This allows the introduction of an extra mode and is a technology being actively pursued in Personal Digital Assistant devices by Psion and Apple. This technology uses a very small vocabulary and passes very simple messages back to the application at the level of XEvents. Similarly, it is equally possible to implement a speech to text (dictaphone) interface into such devices where large vocabularies are used, but there is no “meaning” to the text. It is merely transcribed into a word processor. Such interfaces exist with about a 97% accuracy which is sufficient to support the cleaning of the document later through a keyboard. The first example involves meaning, but low volume, and the second involves large volume and no meaning. For a co-operative multimodal interface we wish to combine complexity (intermediate volume) and meaning to allow constrained natural language dialogues which support complex quantification in commands and queries (e.g. Table 1A), and complex reference resolution (e.g. Table 1B). To support such interactions it is necessary to employ a more complex meaning representation language than XEvents which explicitly represents the meaning of each utterance at an abstract level as suggested in the criteria of Nigay and Coutaz (1993).

However, what is the appropriate level of abstraction? Clearly, events at the XEvents level must be used to capture characters typed at the keyboard, mouse events etc. Even single mouse selections on menus are not in themselves meaningful and must be mapped to the single menu item selected. Characters typed, direct manipulations and components of hand drawn gestures (as in PDAs) only reach the level of meaning when they are composed into larger units. These processes of mapping and composition are performed within the mode layer of MMI^2. The Common Meaning Representation (CMR) which passes from the mode to the dialogue management layer describes the content of a communication action (by the user or the system) in a logical form, along with the force of the utterance (see table 2 for an example). Because this logical form is a meaning bearing level of abstraction, MMI^2 passes the Nigay and Coutaz (1993) second criterion for multimodal systems: that multiple levels of abstraction are represented including meaning. However, is this sufficient? The CMR component for the force of an utterance includes only six classes: imperative, declarative, wh-interrogative, polar-interrogative, phrase (to cover some special cases including ‘greeting’, ‘closing’, and exclamation) or unknown. These convey the pragmatic component of the meaning in classes which the natural language parsers used could determine from the surface form. These classes can also be realised for the graphical interface and system output. Higher level abstractions are used within the ‘communication planning expert’ and ‘informal domain expert’ corresponding to rhetorical acts, although these are not explicit in the CMR (consistent with Maybury, 1991). However, these modules reside within the ‘Dialogue Management Layer’ of the system, and not in the Mode Layer. The purpose of the dialogue management is to provide further co-operative interface features beyond multimodality itself. The purpose of the CMR representation is to support these dialogue management functions. Is it therefore the purpose of the abstract representation of meaning in the Nigay and Coutaz (1993) to support multi-modality, or to allow for the support of undefined dialogue management facilities to provide more co-operative dialogue (e.g. Table 1E & 1F)?

The fourth criterion is that user information from different modes should be fusible together (e.g. Table 1C), and correspondingly system information should be “fissionable” between modes, in that a single rhetorical statement can be used to construct system output in multiple modes (e.g. Table 1D) (see Chappel & Wilson, 1993; Feiner & McKeown, 1991).

The third criterion suggested by Nigay and Coutaz (1993) is that of parallel use of modalities: multimodal systems should allow the user to employ multiple modalities in parallel. It is quite possible in MMI^2 to use graphics and natural language in parallel since reference from physical actions in one mode can be made in another.

Example 1C illustrates a complexity in the notion of parallel use of modes. Clearly there is an explicit reference in the word ‘this’ to an object which could be anaphoric

Table 1A: An example of complex quantification in commands and queries in an MMI^2 dialogue:

User: Which machines do not have disks?
System: cmr1 cmr4
User: Add a small disk to every machine that does not have a disk.
System: OK
User: Which machines do not have disks?
System: None
User: Does every machine have a disk?
System: Yes, cmr1, cmr2, cmr3, cmr4.
User: Does every machine have a small disk?
System: No.

Table 1B: Complex reference resolution in an MMI^2 dialogue

User: Add a 375Mb Disk to the server.
System: OK.
User: Add □□□□□□to cmr3.□
System: OK
User: What is the cost of □□□□?□□□
System: 1909 Sterling
User: What is the cost of □□□□□?□□□□
System: 4114 Sterling
User: What is the type of □□□□?□□□
System: 375 MB Disk.

Table 1C: Example of ostensive user deixis in MMI^2 where user inputs in different modes are fused into a single input.

User: Is using thin cable possible in <mouse select> □□shaft? <terminate>

Table 1D: Example of ostensive system deixis in MMI^2 where system output is divided between two modes.

System: What is the type of <system highlight cable> this cable?

Table 1E: Example of spatial reasoning in an unambiguous user frame of referance.

User: What is the type of the box □□□□□□□the server ?□□

System: Retix_2265.

Table 1F: Example of graphical reasoning in an ambiguous frame of reference?

User: Move □□□□□□□□□□to room2.□□□

System: Is the request to move machine34 or machine35? machine34 is represented by a pruple icon on the screen. machine35 is a Silicon Graphics machine and is purple in the world.

Table 1: Example interactions with the MMI^2 system.

```
(1) CMR(
[
 CMR_act_analysis(
      u_type(phrase([var(x1)]),none),
      [
       CMR_exp(
            [],
            identity(var(x1),const(cmr_Shaft0)),
            nil)],
      nil)],
ok,
Graphics,
time(56,53,23,11,6,1991))

(2) CMR(
[
 CMR_act_analysis(
            u_type(polar,question_mark),
            [
             CMR_exp(
                        [
                         anno(x1,[name(using-thin-cable),singular,definite]),
                         anno(x2,[singular,definite,neuter])],
                        description(desc(E,x1,USING,
                                    identity(var(x1),const('using-thin-cable'))),
                        description(desc(E,x2,SHAFT,true),
                        description(desc(E,x3,IS_POSSIBLE,true),
                        conj(
                                    [
                                    atom(ARG1,[var(x3),var(x1)]),
                                    atom(ARG2,[var(x3),var(x2)])])))),
                        nil)],
            nil)]
ok,
English,
time(56,53,23,11,6,1991))

(3) CMR(
[
 CMR_act_analysis(
            u_type(polar,question_mark),
            [
             CMR_exp(
                        [
                         anno(x1,[name(using-thin-cable),singular,definite]),
                         anno(x2,[singular,definite,neuter])],
                        description(desc(E,x1,USING,
                                    identity(var(x1),const('using-thin-cable'))),
                        description(desc(E,x2,SHAFT,( identity(var(x2),const(cmr_Shaft0)))),
                        description(desc(E,x3,IS_POSSIBLE,true),
                        conj(
                                    [
                                    atom(ARG1,[var(x3),var(x1)]),
                                    atom(ARG2,[var(x3),var(x2)])])))),
                        nil)],
            nil)]
ok,
English,
time(56,53,23,11,6,1991))
```

Table 2: Three CMR examples resulting from a graphical selection of a shaft (1) , the text utterance "Is using thin cable possible in this shaft ?" (2) , and their fusion (3).

in the textual discourse/dialogue or deictic to an object outside of the text. MMI^2 uses a common context space for referents from all modes, so this distinction is handled within the reference resolution process by prioritisation rules. The complexity arises in that the mouse selection can occur at any time after a previous utterance and before the explicit termination of this user turn. That is, the <mouse select> event could occur before the "Is" word, and therefore before the text mode is started, or after the "?" is typed, and therefore after the text mode has completed. The reason for this is that graphical events are passed from the mode to the dialogue management layer as they occur (e.g. mouse clicks), whereas text events are passed to dialogue management when they are terminated (Ben Amara et al.,1991). As long as the graphics event occurs before the text event it terminated, the graphical event can be classified as having occurred before the text event. Therefore the reference resolution process is one of anaphora rather than cataphora (looking back through a reference space, rather than waiting for a reference). This allows users considerable freedom in how they use modes independently, and interactively. However, it does not force the user to make the interaction of modes synergistic rather than alternate from the system perspective. Although since it permits synergistic interaction MMI^2 passes this criterion for multimodality, once again it is the purpose of providing these dialogue functionalities which has given rise to the realisation of multimodality, rather than multimodality □□□.□□□

A second issue that arises from this example is that of where the combination of modes should occur. In MMI^2 the message from the graphical mode and that from the textual mode are combined in the dialogue management layer and not in the mode layer. Table 3 shows examples of the three CMR messages resulting from the example in Table 1C, using the mouse selection of an object shown in the graphics mode (a shaft), the typing of an English natural language utterance, and the resulting combined representation after the "Dialogue Controller" has resolved the reference. This anaphora resolution is one of the functionalities provided by the dialogue management software and it seems inappropriate to try to move it up into the mode layer since references must be resolved across modes, and the boundary of the mode layer passes information relating to individual modes only.

Following the four criteria of Nigay and Coutaz (1993) MMI^2 appears to be a multimodal system. However, the reason it is multimodal it to support co-operative dialogue functionalities, not because multimodality had advantages in its own rights.

HOW SHOULD THE ADVANTAGES OF MULTIMODALITY BE EVALUATED?

In systems developed for the differently abled, or where one of the user's modalities is overloaded by other task demands there can be explicit advantages to multiple modalities by themselves. For systems where this does not apply the advantages of multimodality derive from the co-operative dialogue which it supports How can these co-operative dialogue features be defined and then evaluated?

We have unsuccessfully attempted to list the dialogue management functionalities which we are considering. From published natural language studies, our Wizard of Oz studies (Falzon, 1991) and existing user interfaces we derived a corpus of dialogue functionalities. Unfortunately these do not use the same categorisations or terminology, so are not easily stated in the form of requirements. There are also a very large number of these applying to individual modes and the dialogue management system. We attempted to classify these in terms of the sets illustrated in Figure 2. The largest circle shows the set of functionalities available in human-human dialogues. Within this is the smaller set of functionalities which would be available in an ideal intelligent user interface. Within this are the sets of minimal functionality to meet user's intentions and to provide an acceptable dialogue capability. Within this space the MMI^2 system is then optimistically represented.

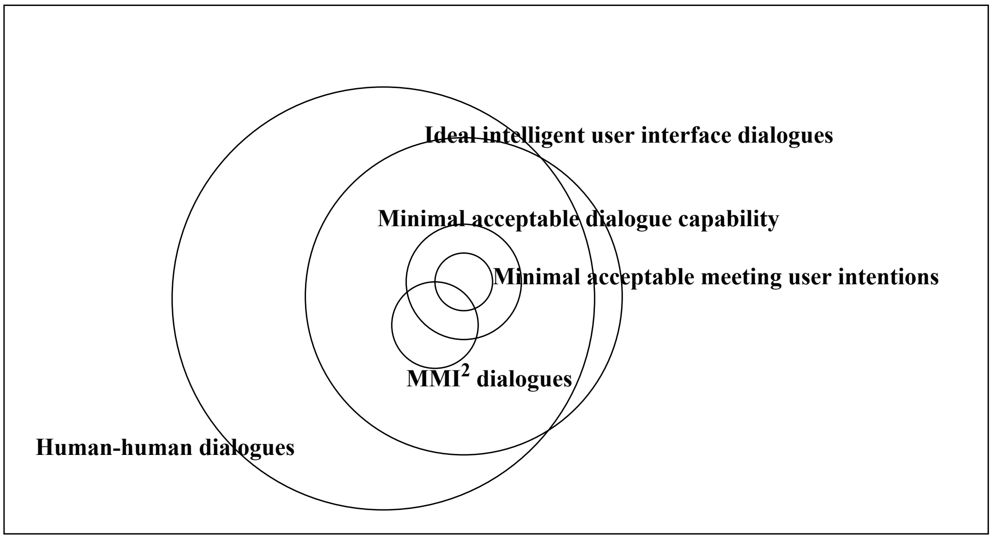
Ideal intelligent user interface dialogues
Minimal acceptable dialogue capability
Minimal acceptable meeting user intentions
MMI^2 dialogues
Human-human dialogues

Figure 2: Sets of functionalities in dialogue systems

There are many studies on human-human dialogue as a model for human-computer dialogue. These generally argue that complete human natural language and dialogue facilities are not required by human-computer systems. Indeed attempts to introduce some of them have the consequence that users overestimate other functionalities provided within a natural subset with resulting errors in interaction.

General dialogue functionalities include the ability to embed subdialogues within a dialogue, the existence of openings, bodies and closings to dialogues, the existence of clarification and tutorial subdialogues, user or system initiated subdialogues, argumentation (rhetorical) structure in system discourse, system discourse tailored to user knowledge, and the ability to ask questions, inform the system of true information and issue commands. The set of dialogue functionalities which are specifically supported by multimodality include, the expression of semantics in modes compatible with users mental models, the fission of system discourse for effective communication, common expression of some operations across modes, and common reference to objects between modes. However the examples in Table 1E and 1F show how even these break down on occasions when the attribute used for reference becomes ambiguous between the "real world" and the representation form in one mode.

The alternative to attempting to define multimodal requirements is to perform evaluation by users on the system. The problem here is that the design is not robust and incorporates many problems of entrainment and habituation within and between modes. The system was designed without a clear model of dialogue functionalities and does not present users with a clear mental model of what is possible.

An obvious example of this is found in the difference between the complexity of text generated by the system and by the complexity of natural language syntax understood as input. The system can generate complex multi-sentential output with multi-phrase sentences using conjunctions and clause structures. the mechanism is a fairly simple rhetorical planning system feeding a canned text based generator. Unfortunately, input is limited to single clause sentences. Users see what can be produced and try to input language of the same complexity. We have attempted to divide the text input window from the explanation output system as though they were different modes in order to overcome this. Although this reduces the problem, it does not eradicate it. Theoretically though, is it valid to consider natural language input and simple generated output as a separate mode from natural language output based on canned text? We need to introduce some distinction in the user's mental model to overcome the problem of entrainment on non-habitable subsets of functionality.

A second example also arises from the limitations of natural language input at the level of spelling errors. Users soon stop using natural language input of any length since they do not type perfectly correct sentences (more so in French and Spanish than in English) and there is neither an intelligent automatic spelling checker to correct their input, nor a clear error detection mechanism to provide appropriate feedback. Again, the practical solution adopted for this was to provide complex text editing facilities at the user interface to allow easy correction of errors by users if they identify the cause themselves. However, this loads the user further with the burden of graphically editing natural language input (thereby mixing the mental model of modes more) and the burden of identifying the cause of errors (which relies on a clear mental model of the mode's operation which does not exist).

These two examples are based on the natural language mode. Other examples can be drawn from the confusion of the functionality of direct manipulation modes with that of freehand gesture modes, and the structure of command language terms with menu selections. These limitations in a clear mental model of the functionality of dialogue, the functionality of each mode and how modes interact makes practical user evaluation impossible, and indicates that further work is required before such systems can be released as products.

HOW SHOULD THE LIMITATIONS BE OVERCOME?

Multimodality supports dialogue functionalities. If users are to approach multimodal systems they must have clear mental models of each mode and of the underlying dialogue capabilities.

Intellectually uninteresting problems such as the consequences of spelling errors outweigh the advantages provided by natural language input. These problems must be overcome before a practical natural language mode of any complexity is usable.

Although we have failed to characterise the functionalities of dialogue management as requirements, this is possible and is necessary to develop a clear mental model for the user of the dialogue management.

The interaction of modes and the representation of different objects in different modes need to be defined to enable the establishment of clear mental models of each mode.

This paper has focused on the limitations in the design of the MMI^2 demonstrator and their implications for multimodal interfaces. This should not be taken as a judgement that MMI^2 demonstrator itself has failed since it illustrates an architecture for multimodal systems, a meaning representation language which can express the content of several modes, a dialogue management system which fulfils the requirements placed on it, and many other advances within individual modes and modules of the dialogue management system. The purpose of the paper is to raise problems which must be addressed to further the developed of multimodal systems rather than promote the achievements of one project.

REFERENCES:

Binot, J-L., Falzon, P., Perez, R., Peroche, B., Sheehy, N., Rouault, J. and Wilson, M.D. (1990). Architecture of a multimodal dialogue interface for knowledge-based systems. In Proceedings of Esprit '90 Conference, 412-433. Kluwer Academic Publishers: Dordrecht.

Ben Amara, H., Peroche, B., Chappel, H., Wilson, M.D. (1991) Graphical Interaction in a Multi-Modal Interface. In Proceedings of Esprit '91 Conference, pgs 303-321, Kluwer Academic Publishers: Dordrecht.

H Chappel & M D Wilson (1993) Knowledge-Based Design of Graphical Responses Proceedings of the ACM International Workshop on Intelligent User Interfaces, 29-36, Orlando, Florida, January 1993, ACM Press.

Doe, G.J., Ringland, G.A., Wilson, M.D. (1992) A Meaning Representation Language for Co-operative Dialogue. Proceedings of the ERCIM Workshop on Experimental and Theoretical studies in Knowledge Representation, 33-40, Pisa, Italy, May 1992.

Falzon, P. (1991), Analysis of multimodal dialogue for domain, explanation and argumentation knowledge, in M.M. Taylor, F. Neel & D.G. Bouwhuis (Eds.),Proceedings of the second Vencona Workshop on Multi-Modal Dialogue, September 1991.

Feiner, S. & McKeown, K. (1991) Automating the Generation of Coordinated Multimedia Explanations. IEEE Computer 24(10), 33-41.

Maybury, M (1991) Planning Multimedia Explanations using Communicative Acts. In M.M. Taylor, F. Neel & D.G. Bouwhuis (Eds.),Proceedings of the second Vencona Workshop on Multi-Modal Dialogue, 1991.

Nigay, L. and Coutaz, J. (1993) A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion. In INTERCHI' 93 172-178, ACM Press: New York.

Wilson, M.D. (1991) The First MMI^2 Demonstrator: a Multi-modal Interface for Man-Machine Interaction with Knowledge Based Systems. Rutherford Appleton Laboratory Report, RAL-91-093, Dec. 1991, pgs 102.