Proceedings of the IJCAI'97 workshop on "Intelligent Multimodal Systems". August 24th. Nagoya, Japan. http://www.miv.t.u-tokyo.ac.jp/ijcai97-IMS/

Towards "intelligent" cooperation between modalities. The example of a system enabling multimodal interaction with a map

Jean-Claude MARTIN
LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
martin@limsi.fr

Abstract

In this paper we propose a coherent approach for studying and implementing multimodal interfaces. This approach is based on six basic "types of cooperation" between modalities: transfer, equivalence, specialization, redundancy, complementarity and concurrence. Definitions and examples of these types of cooperations are given in the paper.
We have used this approach to develop both theoretical tools (a framework, and formal notations) and software tools (a language for specifying multimodal input, and a module integrating events detected on several modalities).
These tools have been applied to the development of a prototype enabling a user to interact with a geographic map by combining speech recognition, pointing gestures with a mouse and a keyboard. We explain the underlying software architecture and give details on how the multimodal module may enable "multimodal recognition scores".
Finally, we describe what we believe "intelligent" multimodal systems should be, and how our approach based on the types of cooperation between modalities could be used in this direction.

1. Introduction

The development of multimodal systems addresses several issues [Maybury 1994]: content selection ("what to say"), modality allocation ("which modality to say it"), modality realization ("how to say that in that modality") and modality combination. Our work deals with the "modality combination" issue. A multimodal interface developer has to know how to combine modalities and why this combination may improve the interaction. Although several multimodal interfaces have already been developed [CMC 1995 ; IMMI 1995], there is still a lack of coherent theoretical and software tools.

In the first part of this paper, we propose a theoretical framework for analyzing modality combinations. The second part details two software tools based on the framework: a specification language and a multimodal module using Guided Propagation Networks. Illustrative examples are taken from a prototype enabling multimodal interrogation of a geographic map developed by [Goncalves et al. 1997].

2. Theoretical tools

A system should use multimodality only if it helps in achieving usability criteria and requirement specifications such as:

- improving recognition in a noisy (audio, visual or tactile) environment,
- enabling a fast interaction,
- being intuitive or easy to learn,
- adapting to several environments, users or user's be-haviors,
- enabling the user to easily link presented information to more global contextual knowledge,
- translating information from one modality to another modality...

These usability criteria may depend on the application to be developed. From a multimodal point of view, they can be seen as "goals of cooperation" between modalities. How can modalities cooperate and be combined to achieve each of these goals ? We propose six basic "types of cooperation" between modalities: transfer, specialization, equivalence, redundancy, complementarity and concurrency. In this section, we define each of them and give examples on how they may help in reaching usability criteria (figure 1). In our definitions, a modality is considered as a process receiving and producing chunks of information. More examples of types of cooperation can be found in [Martin et al. in press].

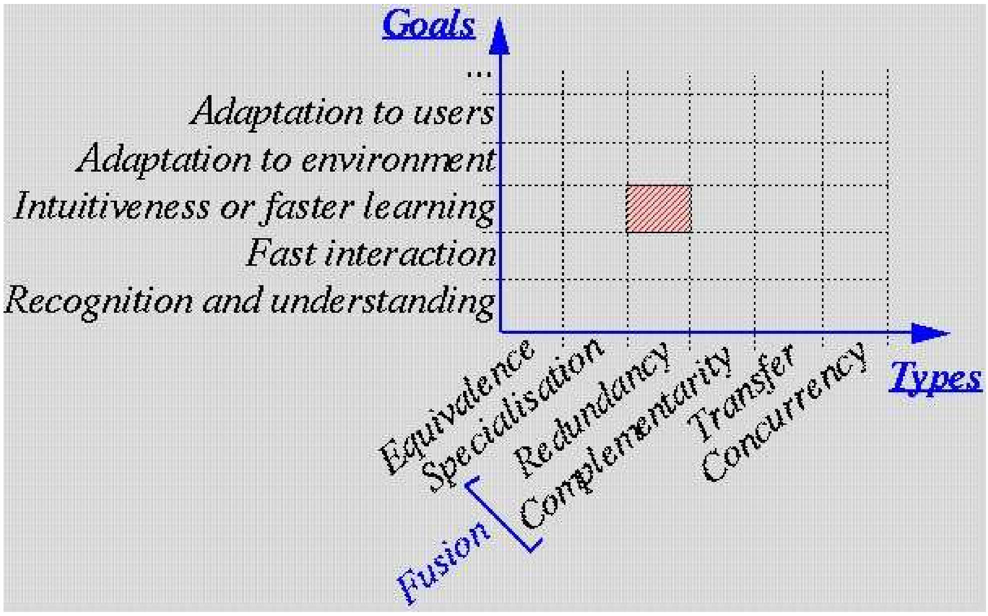


Figure 1. The framework proposed in this paper for studying and designing multimodal interfaces. Six "types of cooperation" between modalities (horizontal axis) may be involved in several "goals of cooperation (vertical axis). For instance (red box), it has been shown that with redundant displayed text and vocal output, a user learned faster how to use a graphical interface [Wang et al. 1993].

2.1. Equivalence

When several modalities cooperate by equivalence, this means that a chunk of information may be processed as an alternative, by either of them.

In COMIT, a multimodal interface that we have developed, the user can create a graphical interface (windows, buttons, scrollbars) interactively by combining speech, mouse and keyboard. For instance, the user may either utter or type "create a scrollbar" to create a new scrollbar.

The EDWARD system [Huls and Bos 1995] is applied to hierarchical file system management. It allows the user to choose at any time during the interaction the style that suits best at that moment (mouse or natural language). Experimental tests have shown that subjects tended to choose the mouse for selecting an object with a long name. Yet, when the object was difficult to locate on the screen, subjects preferred typing.

Equivalence also enables adaptation to the user by cus-tomization: the user may be allowed to select the modalities he prefers [Hare et al. 1995]. The formation of accurate mental models of a multimodal system seems dependent upon the implementation of such options over which the user has control [Sims and Hedberg 1995].

Thus, equivalence means alternative. It is clear that differences between each modality, either cognitive or technical, have to be considered.

2.2 Specialization

When modalities cooperate by specialization, this means that a specific kind of information is always processed by the same modality.

Specialization is not always absolute and may be more precisely defined: one should distinguish data-relative specialization and modality-relative specialization. In several systems, sounds are somehow specialized in errors notification (forbidden commands are signaled with a beep). On the other way, it is a modality-relative specialization if sounds are not used to convey any other type of information. It is a data-relative specialization if errors only produce sounds and no graphics or text. When there is a one-to-one relation between a set of information and a modality, we will speak of an absolute specialization.

Specialization may help the user to interpret the events produced by the computer (to link them to the global contextual knowledge). This means that the choice of a given modality adds semantic information and hence helps the interpretation process.

When a modality is specialized, it should respect the specificity of this modality including the information it is good at representing. For instance, in reference interpretation, the designation gesture aims at selecting a specific area and the verbal channel provides a frame for the interpretation of the reference: categorical information, constraints on the number of objects selected [Bellalem and Romary 1995].

In an experimental study [Bressole et al. 1995] aiming at the understanding of cooperative cognitive strategies used by air traffic controllers, non-verbal resource are revealed to be a specific vector of communication for some types of information which are not verbally expressed such as the emergency of a situation. Intuitive specialization of a modality may goes against its technical specificities. In the Wizard of Oz experiment dealing with a tourist application described in [Siroux et al. 1995], despite the low recognition rate of town names, the users did not use the tactile screen to select a town but used speech instead.

2.3. Redundancy

If several modalities cooperate by redundancy, this means that the same information is processed by these modalities.

In COMIT, if the user types "quit" on the keyboard or utters "quit", the system asks for a confirmation. But if the user both types and utters "quit", the systems interpret this redundancy to avoid a confirmation dialogue thus enabling a faster interaction by reducing the number of actions the user has to perform.

Regarding intuitiveness, redundancy has been observed in the Wizard of Oz study described in [Siroux et al. 1995]: sometimes the user selected a town both by speech and a touch on the tactile screen.

Regarding learnability of interfaces, it has been observed that a redundant multimodal output involving both visual display of a text and speech restitution of the same text enabled faster graphical interface learning [Dowell et al. 1995]. Redundancy between visual and vocal text with verbatim reinforcement was also tested in [Huls and Bos 1995] with natural language descriptions of the objects the user manipulates and the action he performs. Although speech coerced the subjects into reading the typed descriptions, the subjects made more errors and were slower than with the visual text output only.

2.4. Complementarity

When several modalities cooperate by complementarity, it means that different chunks of information are processed by each modality but have to be merged. First systems enabling the "put that there" command for the ma-nipulation of graphical objects are described in [Carbonnel 1970 ; Bolt 1980]. In COMIT, if the user wants to create a radio button, he may type its name on the keyboard and select its position with the mouse. These two chunks of information have to be merged to create the button with the right name at the right posi-tion. This complementarity may enable a faster interac-tion since the two modalities can be used simultaneously and convey shorter messages which are moreover better recognized than long messages.

In [Huls and Bos 1995], experiments have shown that the use of complementarity input such as "Is this a report ?" while pointing on a file, increases with user's experience.

Complementarity may also improve interpretation, as in [Santana and Pineda 1995] where a graphical output is sufficient for an expert but need to be completed by a textual output for novice users. An important issue con-cerning complementarity is the criterion used to merged chunks of information in different modalities. The most classical approaches are to merge them because they are temporally coincident, temporally sequential or spatially linked. Regarding intuitiveness, complementarity behavior were observed in [Siroux et al. 1995]. Two types of behavior did feature complementarity. In the "sequential" behavior, which was rare, the user would by example utter "what are the campsites at" and then select a town with the tactile screen. In the "synergistic" behavior, the user would utter "Are there any campsites here ?" and select a town with the tactile screen while pronouncing "here". Regarding the output from the computer, it was observed in the experiment described in [Hare et al. 1995] that spatial linking of related information encourages the user's awareness of causal and cognitive links. Yet, when having to retrieve complementary chunks of information from different media, users behavior tended to be biased towards sequential search avoiding synergistic use of several modalities.

Modalities cooperating by complementarity may be specialized in different types of information. In the example of a graphical editor, the name of an object may be always specified with speech while its position is specified with the mouse. But modalities cooperating by complementarity may be also be equivalent for different types of information. As a matter of fact, the user could also select an object with the mouse and its new position with speech ("in the upper right corner"). Nevertheless, the complementary use of specialized modalities gives the advantages of specialization: speech recognition is improved since the vocabulary and syntax is simpler than a complete linguistic description.

2.5. Transfer

When several modalities cooperate by transfer, this means that a chunk of information produced by a modality is used by another modality.

Transfer is commonly used in hypermedia interfaces when a mouse click provokes the display of an image. In information retrieval applications, the user may express a request in one modality (speech) and get relevant information in another modality (video) [Foote et al. 1995]. Output information may not only be retrieved but also produced from scratch. Several systems generate graphical descriptions of a scene from a linguistic description [O Nuallain and Smith 1994]. Natural language instruc-tions can also be used to create animated simulations of virtual human agents carrying out tasks [Webber 1995]. Similarly, the visual description of a scene can be used to generate a linguistic description [Jackendoff 1987] or a multimodal description [André and Rist 1995]. Let's say that all these previous examples involved transfer for a goal of translation.

Transfer may also be involved in other goals such as improving recognition: mouse click detection may be transferred to a speech modality in order to ease the recognition of predictable words (here, that...) as in the GERBAL system [Salisbury et al. 1990].

2.6. Concurrency

Finally, when several modalities cooperate by concurrency, it means that different chunks of information are processed by several modalities at the same time but must not be merged. This may enable a faster interaction since several modalities are used in parallel.

2.7. Formal notations

To define more precisely these types of cooperation, we propose logical formal notations. They aim at stating explicitly the parameters of each type of cooperation and the relation between these parameters which is subsumed by the type of cooperation. We consider the case of input modalities (human towards computer). These formal notations have helped us in defining a specification language for implementing multimodal interfaces (next section).

We define a modality as a process receiving and pro-ducing chunks of information. A modality M is formally defined by:

- E(M) the set of chunks of information received by M
- S(M) the set of chunks of information produced by M

Two modalities M1 and M2 cooperate by transfer when a chunk of information produced by M1 can be used by M2 after translation by a transfer operator tr which is a pa-rameter of the cooperation.

$$\begin{aligned}&transfer\ (M_1, M_2, tr):\\&\qquad tr(S(M_1)) \subset E(M_2)\end{aligned}$$

An input modality M cooperate by specialization with a set of input modalities Mi in the production of a set I of chunks of information if M produces I (and only I) and no modality in Mi produces I.

$$\begin{aligned}&specialisation(M, I, \{M_i\}):\\&\qquad I = S(M) \wedge \forall M_i\,,\ I \not\subset S(M_i)\end{aligned}$$

Two input modalities M1 and M2 cooperate by equiva-lence for the production of a set I of chunks of informa-tion when each element i of I can be produced either by M1 or M2. An operator eq controls which modality will be used and may take into account user's preferences, environmental features, information to be transmitted...

$$\begin{aligned}&equivalence\ (M_1, M_2, I, eq):\\&\qquad \forall\, i \in I,\ \exists\, e_1 \in E(M_1),\ \exists\, e_2 \in E(M_2),\ i = eq((M_1, e_1), (M_2, e_2))\end{aligned}$$

Two input modalities M1 and M2 cooperate by redundancy for the production of a set I of chunks of informa-tion when each element i of I can be produced by an operator re merging a couple (s1, s2) produced respec-tively by M1 and M2. The operator re will merge (s1, s2) if their redundant attribute has the same value and a criterion crit is true. A chunk of information has several attributes. For instance, a chunk of information sent by a speech recognizer has the following attributes: time of detection, label of recognized word, recognition score. The redundant attribute of two modalities plays a role in deciding whether two chunks of information produced by these modalities is redundant or complementary.

$$\begin{aligned}&redundancy\ (M_1, M_2, I, redundant\_attribute, crit):\\&\qquad \forall\, i \in I,\ \exists\, s_1 \in S(M_1),\ \exists\, s_2 \in S(M_2),\\&\qquad redundant\_attribute\ (s1) = redundant\_attribute\ \ (s2) \wedge\\&\qquad i = re(s_{1,}\, s_{2,},\ crit)\end{aligned}$$

Two input modalities M1 and M2 cooperate by complementarity for the production of a set I of chunks of in-formation when each element i of I can be produced by an operator co merging a couple (s1, s2) produced re-spectively by M1 and M2. The process co will merge (s1, s2) if their redundant attribute does not have the same value and a criterion crit is true:

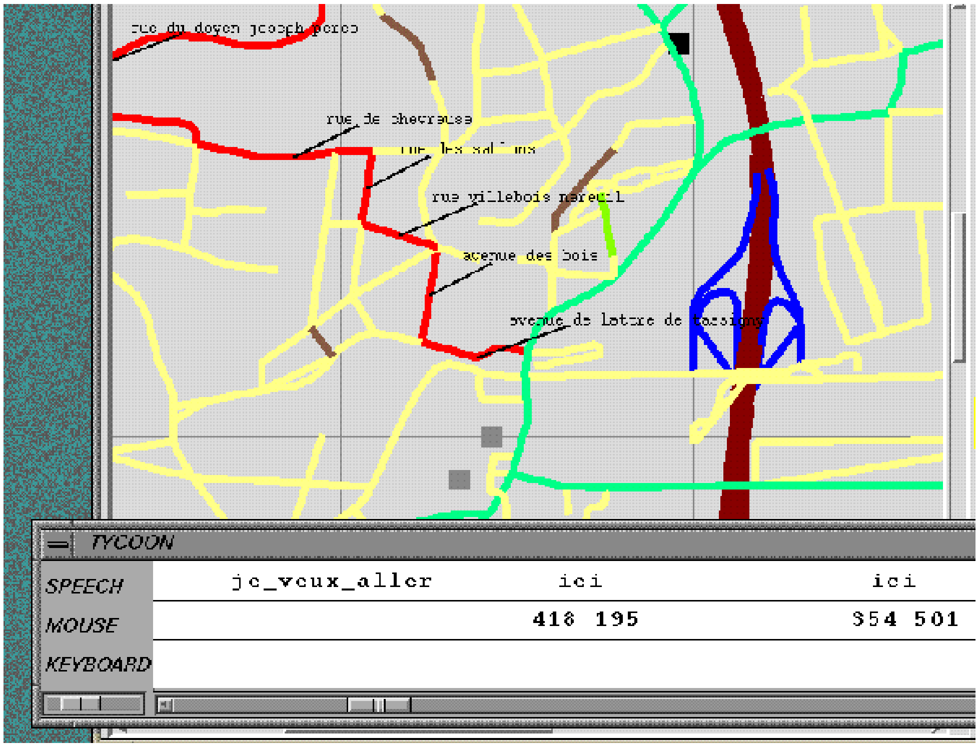
$$\begin{aligned}&\mathit{complementarity}\ (M_1, M_2, I, \mathit{redundant\_attribute}, \mathit{crit}):\\&\quad \forall i \in I,\ \exists s_1 \in S(M_1),\ \exists s_2 \in S(M_2),\\&\quad \mathit{redundant\_attribute}\ (s1) \neq \mathit{redundant\_attribute}\ \ (s2) \wedge\\&\quad i = co(s_1, s_2, \mathit{crit})\end{aligned}$$

In the next sections, we introduce a specification language based on these formal notation. This language has been used for the implementation of a multimodal prototype: CARTOON.

3. The CARTOON prototype

We have implemented CARTOON (CARTography and cOOperatioN between modalities), a multimodal interface to a cartographic application developed by [Goncalves et al. 1997] enabling the manipulation of streets, the computation of shortest itinerary... Multimodal interrogation of maps seems to be a promising application for multimodal systems [Cheyer and Julia 1995 ; Siroux et al. 1995] as more and more tourist information is available on the Internet. Figure 2 shows a screen dump during a multimodal interaction in CARTOON. A map is displayed on the screen. The user may combine speech utterances and pointing gestures with the mouse. For instance, the user may utter (translated from French) "I want to go from here to here ". Then the system computes the shortest itinerary and the streets to be taken are displayed in red. The following combinations are possible with CARTOON:

- Where is the police station ?
- Show me the hospital
- I want to go from here to the hospital
- I am in front of the police station. How can I go here ?
- What is the name of this building ?
- What is this ?
- Show me how to go from here to here



SPEECH je_veux_aller ici ici

MOUSE 418 195 354 501

KEYBOARD

Figure 2. Example of a multimodal interaction with the CARTOON prototype. The events detected on the three modalities (speech, mouse, keyboard) are displayed in the lower window as a function of time. In this case, the detected speech events were: "I_want_to_go", "here", "here". Two mouse clicks were detected. The system integrated these events as a request and displays the shortest itinerary.

In the current version, there is no linguistic analysis preliminary to the multimodal fusion. Events produced by the speech recognition system (a Vecsys Datavox) are either words ("here") or sequences of words ("I_want_to_go"). There are 38 such possible speech events. Each speech event is characterized by: the recognized word, the time of utterance and the recognition score.

The pointing gestures events are characterized by an (x, y) position and the time of detection.

The overall hardware and software architecture is described in figure 3.

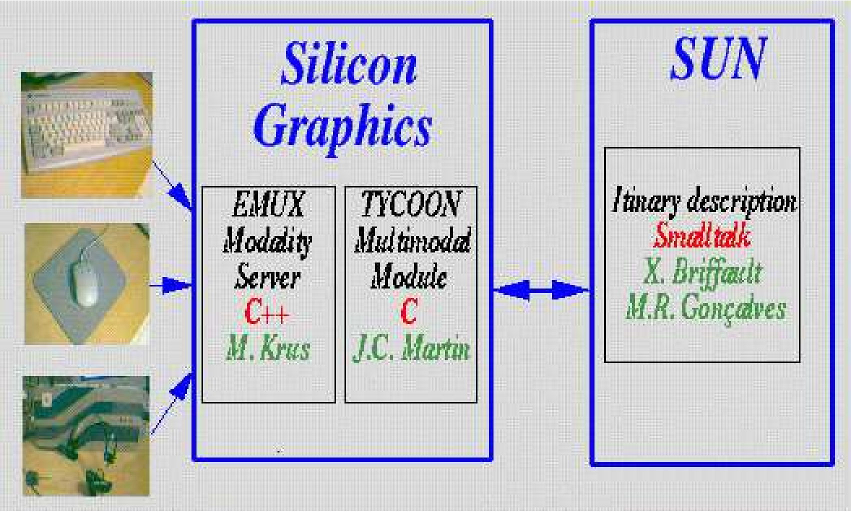


Figure 3. hardware and software architecture. Events detected on the speech, mouse and keyboard modalities (left-hand side) are time-stamped coherently by a Modality Server [Bourdot et al. 95]. The events are then integrated in our multimodal module TYCOON (in the middle) which merges them and sends messages to the cartography and itinerary application (right-hand side).

4. The specification language

The combination of modalities used in CARTOON are described in a specification language that is based on our formal notations. In this section, we explain parts of the specification file used for CARTOON.

Firstly, the modality used are specified (the objects modality is activated when one graphical object such as a building is mouse-clicked) :

```
        modality Speech Keyboard Mouse Objects
```

Then, these modalities are connected to the multimodal module:

```
        link            Speech                          Multimodal
        link            Mouse                           Multimodal
        link            Keyboard                        Multimodal
        link            Objects                         Multimodal
```

The events to be detected on each modality are also specified (38 speech items):

```
        event           Speech  where_is
                                show_me
                                I_am
                                I_want_to_go
                                ...
```

For each command of the cartographic application, the possible combination of modalities are specified. Here is the example of the command NameOf: A variable V3 is defined as the beginning of a sequence:

```
        start_sequence  Multimodal      V3
```

It is may be activated by one event among several (the word "name" typed on the keyboard or the speech items "what is the name of" or "what is that"):

```
        equivalence     Multimodal      V3
                Keyboard                name
                Speech                  what_is_the_name_of
                Speech                  what_is_that
```

This V3 variable is linked sequentially to a second vari-able V4:.

```
        complementarity_sequence        Multimodal      V3   V4
```

V4 may only be activated by a mouse event:

```
        specialization          Multimodal      V4     Mouse            *
```

V4 is bound to a parameter of an application module which is involved in the execution process:

```
        bind_application        Parameter1NameOf                V4
```

V4 is the last variable of the sequence:

```
        end_sequence            Multimodal      V4       NameOf
```

5. The multimodal module

The multimodal module used in CARTOON is based on Guided Propagation [Béroule 1985] (figure 4). Such networks comprise elementary processing units: event-detectors and multimodal units. Event detectors (square units) selectively respond to events at the moment they occur in the environment. When activated by an event, these event-detectors send a signal to the multimodal units (circle units) to which they are connected. The connections between the units are build from the specification file described in the previous section.

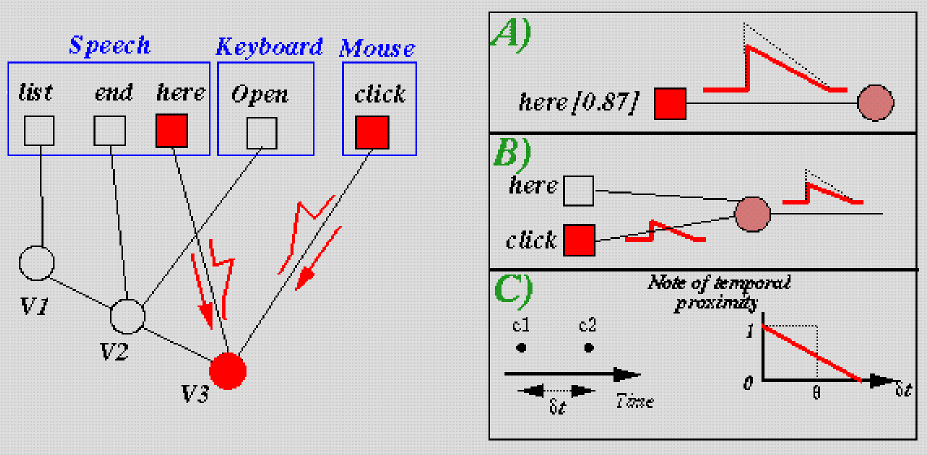


Figure 4: the multimodal module uses Guided Propagation Networks. Left-hand side: a network integrating events detected on three modalities is composed of event-detectors (square units) and multimodal units (circle units). Right-hand side: three properties of these networks enable multimodal recognition scores (see text).

The activity level of a detector at the end of a multimodal command pathway corresponds to the way an occurrence of this command matches its internal representation. This "matching score" accounts for the degree of distortions undergone by the reference multimodal command, including noisy, missing or inverse components. Initially applied to robust parsing [Westerlund et al. 1994], this feature has been adapted to multimodality [Veldman 1995]. This quantified matching score results from three properties of GPN (figure 4, right-hand side):

- A: the amplitude of the signal emitted by a speech detector is proportional to the recognition score provided by the speech recogniser
- B: a multimodal unit can be activated even if some expected events are missing (in this case, the amplitude of the signal emitted by this variable is lower than the maximum)
- C: the bigger the temporal distortion between two events, the weaker their summation (or note of temporal proximity), because of the decreasing shape of the signals.

6. Conclusion and perspectives

In this paper, we have described some theoretical and software tools that we have developed. We explained how we used them for implementing a multimodal interface to a cartography application. The main features of our work are the typology of types of cooperation that we propose and the capacity of our multimodal module to provide multimodal recognition scores.

We plan to improve the CARTOON system in the following directions:

- make user studies to test the advantages of multimodal recognition scores and to evaluate the types of cooperation that are used by the user
- develop linguistic and semantic representations (which are currently missing in our work) : we plan to connect our multimodal module to the linguistic tools developed by [Briffault et al. 1997] and test several possibilities of interaction such as early dropping of linguistic hypothesis due to multimodal results
- extend the gesture modality to circling and trajectory gestures on a tactile screen

More generally, what should be an "intelligent" multimodal system ? We propose hereafter some answers to this question. It should:

- recognize several input modalities (speech, hand and body gesture, gaze)
- generate contextual output modalities (speech, displayed text and graphics) depending on the users profile, behavior and environment
- be intuitive to use
- integrate multi-users dialogues mediated by the computer
- manipulate semantic representations
- find out dynamically the most important goal of cooperation between modalities depending on the user and environmental features

- dynamically select (these three questions have to be tackled together):
- the information to be transmitted
- the modalities to be used (and hence the media)
- the types of cooperation between modalities to be used

Acknowledgments

The author would like to thank Marie-Rose Goncalves and Xavier Briffault for the cartographic application they have developed and which is used within the CARTOON project.

References

[André and Rist 1995] André, E. and Rist, T. Generating coherent presentations employing textual and visual material. Artificial Intelligence Review, 9 (2-3), 147-165.

[Bellalem and Romary 1995] Bellalem, N. and Romary, L. Reference interpretation in a multimodal environment com-bining speech and gesture. In [IMMI 1995].

[Béroule 1985] Béroule, D. (1985). A model of Adaptative Dynamic Associative Memory for speech processing. The-sis, 31 may, Univ. Orsay. 185p. In French.

[Bolt 1980] Bolt, R.A. "Put-That-There": Voice and Gesture at The Graphics Interface. Computer Graphics 14 (3):262-270.

[Bourdot et al. 1995] Bourdot, P., Krus, M., Gherbi, R. Management of non-standard devices for multimodal user interfaces under UNIX/X11. In [CMC 1995].

[Bressole et al. 1995] Bressolle, M.C, Pavard, B., Leroux, M. The role of multimodal communication in cooperation and intention recognition: the case of air traffic control. In [CMC 1995].

[Goncalves et al. 1997] http://www.limsi.fr/Individu/goncalve/index.html
http://www.limsi.fr/Individu/xavier/index.html

[Carbonnel et al. 1970] Carbonnel, J.R. Mixed-Initiative Man-Computer Dialogues. Bolt, Beranek and Newman (BBN) Report N 1971, Cambridge, MA.

[Cheyer and Julia 1995] Cheyer, A. and Julia, L. Multimo-dal maps: an agent-based approach. In [CMC 1995].

[CMC 1995]. Proceedings of the International Conference on Cooperative Multimodal Communication (CMC'95). Bunt, H, Beun, R.J. and Borghuis, T. (Eds.). Eindhoven, may 24-26.

[Dowell et al. 1995] Dowell, J.; Shmueli, Y.; and Salter, I. Applying a cognitive model of the user to the design of a multimodal speech interface. In [IMMI 1995].

[Foote et al. 1995] Foote, J.T.; Brown, M.G.; Jones, G.J.F.; Sparck Jones, K.; and Young, S.J. Video mail retrieval by voice: towards intelligent retrieval and browsing of multi-media documents. In [IMMI 95].

[Briffault et al. 1997] http://www.limsi.fr/Individu/xavier/index.html http://www.limsi.fr/Individu/vap/index.html

[Hare et al. 1995] Hare, M.; Doubleday, A., Bennett, I.; and Ryan, M. Intelligent presentation of information retrieved from heterogeneous multimedia databases. In [IMMI 1995].

[Huls and Bos 1995] Huls, C. and Bos, E. Studies into full integration of language and action. In [CMC 1995].

[IMMI 1995] Pre-Proceedings of the First International Workshop on Intelligence and Multimodality in Multimedia Interfaces: Research and Applications. Edited by John Lee. University of Edinburgh, Scotland, July 13-14.

[Jackendoff 1987]. Jackendoff, R. On beyond zebra: the relation between linguistic and visual information. Cognition 26(2):89-114.

[Martin et al. In press] Martin, J.C., Veldman, R. and Béroule, D. Developing Multimodal Interfaces : A theoretical Framework and Guided Propagation Networks. Book following the [CMC 1995] workshop. Bunt, H. (Ed.)

[Maybury 1994] Maybury, M. Introduction. In Intelligent multimedia interfaces. AAAI Press. Cambridge Mass.

[O'Nuallain and Smith 1994] O'Nuallain, S. and Smith, A.G. An investigation into the common semantics of language and vision. Artificial Intelligence Review 8 (2-3):113-122.

[Salisbury et al. 1990] Salisbury M.W.; Hendrickson, J.H.; Lammers, T.L.; Fu, C.; and Moody, S.A. Talk and draw: bundling speech and graphics. IEEE Computer., 23(8) 59-65.

[Santana and Pineda 1995] Santana, S. and Pineda, L.A. Producing coordinated natural language and graphical ex-planations in the context of a geometric problem-solving task. In [IMMI 1995].

[Sims and Hedberg 1995] Sims, R. and Hedberg, J. Dimen-sions of learner control: a reappraisal of interactive multi-media instruction. In [IMMI 1995].

[Siroux et al. 1995] Siroux, J., Guyomard, M., Multon, F., Remondeau, C. Modeling and processing of the oral and tactile activities in the Georal tactile system. In [CMC 1995].

[Veldman 1995] Experiments on robust parsing in a multi-modal Guided Propagation Network. ERASMUS Report. LIMSI.

[Wang et al. 1993] Wang, E.; Shahnvaz, H.; Hedman, L.; Papadopoulos, K.; and Watkinson. A usability evaluation of text and speech redundant help messages on a reader inter-face. In G. Salvendy M. Smith (Eds.), Human-Computer Interaction: Software and Hardware Interfaces. pp 724-729.

[Webber 1995] Webber, B. Instructing Animated Agents: Viewing Language in Behavioural Terms. In [CMC 1995].

[Westerlund et al. 1994] Westerlund, P., Béroule, D and Roques, M. Experiments of robust parsing using a Guided Propagation Network. Proc. of the International Conf. on New Methods in Language Processing (NEMLAP), sept. 14-16, Manchester.