A Theoretical Framework for Multimodal User Studies

Jean-Claude Martin[1,2], Luc Julia[2], Adam Cheyer[2]

[1] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS), BP 133, 91403 Orsay Cedex, France, martin@limsi.fr
[2] SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, U.S.A.
cheyer@ai.sri.com, julia@speech.sri.com

Abstract. Researchers have conducted studies in various application areas to model human behavior during multimodal interactions with a real or simulated system. We propose a theoretical framework based on "types" and "goals" of cooperation between modalities. Using a survey of current multimodal research, we analyze several multimodal experiments and systems, examining which aspects of our framework were already considered by the researchers, and what elements the proposed framework could have added to these investigations. It is our hope that the framework will make it easier to link user observation on one side and software design on the other side.

Keywords

Multimodality, user modeling, theoretical framework

1 Introduction

In the area of multimodal system design, user studies have been conducted to explore how users combine modalities when interacting with real systems. To free the experiment from the limitations of today's technology, many experiments use a "Wizard of Oz" approach, in which the functional part of the system is simulated by a hidden human participant. Because of the complexity of such studies, experiments generally focus on one or more aspects of multimodal human-computer interaction. For instance, several studies have investigated the role of time in multimodal input interaction (i.e., are spoken deictics simultaneous with related gestures?). The methods used for analyzing the resulting multimodal corpus often ignore important information that might prove useful for the development or improvement of the multimodal system being considered.

In this paper, we define our terminology and our framework, and explain how the framework has been applied to multimodal system development. We provide a survey of ten existing multimodal experiments and discuss how they fit into our framework. By doing so, we establish a list of questions that we believe should be answered when analyzing a multimodal corpus.

2 A framework based on cooperation between modalities

We have developed a conceptual framework called TYCOON for building multimodal systems (see Martin and Béroule 93, Martin et al. 95 for more detail). This framework answers the following questions: "What is multimodality?" and "Why should we use multimodality in human-computer interaction?"

There is no real agreement on the definitions of "media" and "modality". In this paper we use our own definitions, which are as follows. A "media" is a physical device enabling the exchange of information between the user and the computer. Examples of media are keyboard, mouse, pen, microphone, screen and loudspeaker. A "modality" is a way to exploit a specific

media. Examples of modalities are typed command language, typed natural language, spoken natural language, 2D gestures with the mouse and 2D gestures with a pen. During user interactions, "chunks" of information are transmitted across several modalities from the user to the computer and vice-versa. Related chunks of information can be grouped into higher-level entities (i.e., commands). As for "multimodality", we propose the following definition: "the cooperation between several modalities in order to improve the interaction". More precisely, we have listed six basic "types" of cooperation between modalities:

- complementarity: Different chunks of information belonging to the same command are transmitted over more than one modality;
- redundancy: The same chunk of information is transmitted using more than one modality;
- equivalence: A chunk of information may be transmitted using more than one modality;
- specialization: A specific chunk of information is always transmitted using the same modality;
- concurrency: Independent chunks of information are transmitted using different modalities and overlap in time;
- transfer: A chunk of information produced by one modality is analyzed by another modality.

As our primary aim was the building of multimodal systems, our conceptual framework produced formal notations and two software tools: a specification language and a multimodal module. We have applied these tools as a configuration system for a prototype enabling multimodal interaction with a map (Martin 97).

The second dimension of TYCOON is called "goals of cooperation". This describes the requirements of a human-computer interface: the system must recognize and understand the messages sent by the user (and vice-versa), the system must be intuitive to use, interactions must be fast enough, and the system must adapt to different users and to different environmental conditions. As we will see in the next section, a type of cooperation may be involved in several goals. Regarding user studies, this dimension is very open and can be extended with any other goal to be tailored to a current user study. Symmetrically, a user study may choose to focus on only one of these "goals" of cooperation between modalities.

3 A survey on user studies for multimodal interfaces

Several multimodal user studies and Wizard of Oz experiments have already been published. To have a global view of existing user studies for multimodal interfaces, we analyzed ten of them as they were described in the literature. Table 1 summarises this survey.

4 Existing multimodal studies and TYCOON

We have compiled a list of questions based on our framework. We believe that such questions should be answered when analyzing a multimodal corpus to ease the specification of a multimodal module. Existing studies have answered some of them.

4.1 What are the monomodal features of a user's behavior?

This question is often tackled in user studies. It has an obvious impact on the development of software modules managing each modality as speech and gesture. Several studies, such as (Guyomard et al. 95, Oviatt et al. 97), propose categories of words and gestures from their corpuses. A speech-only session is often compared to a multimodal session to determine whether availability of several modalities modifies spoken behavior (i.e. a smaller set of word categories makes speech recognition easier).

	Guyomard et al. 95	Oviatt et al. 97	Denda et al. 97	Trafton et al. 97	Fais 97
application	tourist map	real estate and map update	tourist map	tourist map and hotel booking	manipulating a map
real or simulated	both	simulated	real	simulated	real
input media and modalities	speech tactile screen	speech pen (pointing, drawing, writting)	speech tactile screen (pointing)	keyboard (natural language) mouse menus	speech keyboard tactile screen
output media and modalities	speech graphics text	speech graphics	speech graphics	graphics	speech graphics text persona

	Ando et al. 94	Mignot and Carbonel 96	Huls and Bos 95	Petrelli et al. 97	Wang et al. 93
application	interior system design	interior system design	file manipulation	form filling	using a graphical interface
real or simulated	both	simulated	real	simulated	real
input media and modalities	speech tactile panel	speech tactile screen	keyboard (limited language) mouse menus	pen (pointing, drawing, writting)	mouse
output media and modalities	graphics	speech graphics	speech graphics text	graphics text	speech graphics text

Table 1. Summary of a survey on ten multimodal user studies. The five columns in the upper table deal with map-based domains, whereas the lower five columns deal with other kinds of applications.

4.2 Does the subject use complementarity? Why? Which criteria should be used by the system to merge the chunks of information?

Complementarity means that different chunks of information belonging to the same command are transmitted by the subject using more than one modality and must be merged by the system. Thus, the system must detect complementarity and must know which fusion criteria to apply. Some studies have observed subtypes of complementarity in the subject's behavior. In (Guyomard et al. 95, Siroux et al. 97), deictics were observed in examples such as "Are there any beaches in this locality?" completed by a touch on a locality. Other observed examples involved different combinations of oral utterance and tactile activities. Examples of substitution of the linguistic referent with the tactile designation (e.g., "What are the camping sites at ?" completed by a pointing gesture on a town) had a low frequency. When addressing similar questions, different experiments sometimes get different results. Some experiments observed temporal relationships between speech and gestures, such as temporal coincidence (Catinis and Caelen 95) or temporal sequence (Oviatt et al. 97); others did not observe any temporal relationship at all (Mignot and Carbonell 96). It appears that the type of microphone (phone handset, head microphone, or table microphone) and the medium used for gestures (tactile screen or pen) might modify the multimodal behavior of the subjects, including the

role of time. Regarding the goals of cooperation, complementarity may be useful in improving speech recognition (Oviatt et al. 97) or making interaction faster (Huls and Bos 95).

4.3 Does the subject use redundancy? Why? Which criteria should be used by the system to merge the chunks of information?

Redundancy means that the subject sends to the computer the same chunks of information by using two (or more) modalities. Redundancy was observed in (Guyomard et al. 95, Siroux et al. 97): a confirmative relationship for which the oral syntagma was sufficient was accompanied by a tactile designation that was redundant with the linguistic reference. On the other hand, redundancy was very rarely observed in (Oviatt et al. 97). In only 2% of the commands, speech provided duplicate but less precise information about location. Symmetrically, drawn graphics provided partially duplicated information about the type of object followed with more precise speech. In (Mignot and Carbonel 96), continuous gestures were combined with a direct-manipulation spoken style (and high redundancy between both), but discrete gestures were combined with a communication spoken style (and low redundancy between both). In (Petrelli et al. 97), when very short labels (one character) were available, users strongly adopted a redundant strategy (they referred to the object in a linguistic way and used pointing, too). There may be a reason that redundant behavior is seldom considered in user studies. Such user studies should try to find out this reason so that the system can make inferences on a subject's intentions and determine which criteria to apply to merge redundant chunks of information.

4.4 Does the subject makes use of equivalence? Why?

It may appear that in some case, the subject uses speech to get some information, and then later uses gesture to get the same type of information. We call this behavior "equivalence", which does not mean that there is no difference between the two modalities. In fact, it could be of interest to know whether there is a particular reason for the subject to switch from one modality to the other. This information is seldom analyzed in multimodal user studies. Instead, user studies provide statistics on all possible types of command and the modality with which they were expressed most often. Equivalence within a single user's behavior is ignored.

4.5 Does the subject use concurrency? Why?

Concurrency means that independent chunks of information are transmitted on different modalities and that they overlap in time. User studies can be useful in finding out how much it happens in a user's behavior. In fact, it is seldom observed in a multimodal corpus. Yet, in (Mignot and Carbonel 96), one example was observed: the user was speaking a command when he realized that he had to move a piece of furniture before executing the command, so he used gesture to move the furniture while continuing to speak.

4.6 Does the subject use specialization? Why?

Specialization means that a specific type of information is always transmitted through the same modality. For instance, what can we learn from a corpus where a subject always uses speech for commands providing the locations of buildings ("Where is the shopping center?")? It is possible to make a distinction between several subtypes of specialization. The specialization may be relative to the modality (i.e., the subject does not use speech for command other than the location of building; yet the subject also asks for the same information with a gesture modality). Symmetrically, the specialization may be relative to the type of commands. Such information can be useful to a multimodal system because it can help to make predictions by improving the recognition of the commands in which a modality is specialized.

4.7 Multimodal versus monomodal

In user studies, the rate of observed multimodal commands is often compared to the rate of monomodal commands. This may help the system to decide whether the type of cooperation brought into play is multimodal or monomodal. In (Oviatt et al. 97), 19% of the corpus is multimodal, 17.5% pen-only and 63.5% speech-only. In (Petrelli et al. 97), users with preliminary computer experience performed 84% multimodal input (with shorter written input and transferring part of the reference meaning on the pointing) and nonexperts only 30%. In (Mignot and Carbonel 96), knowledge about the type of command can be used to infer the probability of use of multimodality (58% of rotation commands are expressed multimodally). Yet, details contrasting possible subtypes of multimodal (either redundant or complementary) and monomodal behavior (equivalence, specialization, concurrency) are not always provided.

4.8 Are there strong differences between individuals?

In (Siroux et al. 97), the rate of use of the tactile screen differs a lot from one subject to another (from 2% to 95%). In (Mignot and Carbonel 96), high interindividual differences between users were observed: two subjects specialized in each modality according to the information such as adding new furniture (speech and discrete gesture) or moving furniture (continuous gesture only) and two other subjects preferred only speech. When redundant behavior is seldom observed, it is of importance to know whether only one subject used it during a session. A probabilistic model could help the system know the most likely type of cooperation between modalities for a given user and hence help in multimodal module processing. Interindividual differences regarding multimodal behavior should not be ignored during user modeling.

4.9 Does multimodal behavior change over time?

Answers to this question might be useful in providing the system with online adaptation to a user's behavior dynamics. For instance, a specialization of speech in one type of command may evolve toward equivalence between speech and gesture for this command type. In (Mignot and Carbonel 96), the percentage of multimodal observations increased between the first session and the third one.

5 Discussion

5.1 Multimodal output

Although we have focused in this paper on multimodal input to the computer, some multimodal user studies also deal with multimodal output. In (Wang et al. 93), it was observed that redundancy between speech output and text display enabled the user to shorten learning time for use of a graphical interface. In (Huls and Bos 95), the efficiency of such a redundant combination was also studied. In (Fais 97), speech output was used along with a graphical synthetic speaking face (although the content may be redundant, some low-level events like phonemes are better transmitted using visual cues than simply audio speech). Studies should also be made about how the subject reacts when the system uses equivalence: this may help the system in selecting between two "equivalent" modalities. User studies can also help in finding out whether a concurrent use of several modalities by the system is helpful and appreciated by the subjects (Bearne et al. 94). Finally, the spatial overlay of the different visual modalities like text, graphics, or pictures seems to have an impact on the way subjects integrate the chunks of information they receive (Hare et al. 95).

5.2 Using TYCOON in multimodal user studies

The distinction between complementarity and redundancy is not always mentioned in user studies. Some authors refer to "combined situations" without giving more details on the relationship between the content of the information transmitted on the different modalities. This difference is sometimes lost in statistics opposing monomodal and multimodal behaviors. In fact, as mentioned in (Mignot and Carbonel 96), the difference is not always obvious. What looks like a complementarity may in fact be seen as a redundancy when considering the pragmatic knowledge the subject has of the system. Moreover, each type of cooperation may be involved at several levels of abstraction, in either the content or the form.
Studies like (Fais 97, Oviatt et al. 97) have evaluated the influence (and transfer of information) of the combination of output modalities on the subject's behavior. Transfer means that a chunk of information produced by one modality is used by another modality. In fact, the relationship between the modality used by the subject and the modality used in response by the system can be seen as a transfer. Transfer is also studied when people use speech to describe images, transferring information from visual perception to verbal production (Gapp 94).
Finally, multimodal conversations between human (and animated agents) seem to involve several types of cooperation: facial expressions can replace sequences of words (equivalence) or accompany them (redundancy and complementarity), for example, gazing at the other person to see how she follows (transfer between gaze and speech) such as is described in (Cassell et al. 94).

We think that TYCOON could be a useful theoretical framework for analyzing multimodal corpuses. We have provided examples on how some existing studies fit into it. As we believe that there should be a particular user study for each system, aiming at open theoretical tools for multimodal user studies can be of interest. A subset of our ideas initially introduced in (Martin and Béroule 93) was renamed "Care" properties by (Coutaz and Nigay 94) and was also used in a Wizard of Oz experiment where the wizard is seen by the subject (Catinis and Caelen 95).

5.3 Conclusion

We have described how a theoretical framework that we originally used for building multimodal interfaces could be useful for evaluating multimodal user studies. We have provided a list of questions that should be considered during the analysis phase of these user studies. These questions should have an impact on what is recorded during the experiments and during the subject debriefing phase. Furthermore, these types of cooperation between modalities may also provide information about dialogue or the subject's intentions. What we intend to do in the near future is to apply the TYCOON framework to provide data analysis for the multimodal Wizard of Oz experiment described in (Cheyer et al. 1998).

References

ACL (1997) Proceedings of the workshop "Referring phenomena in a multimedia context and their computational treatment", ACL/EACL'97, July 11th, Madrid.
http://www.dfki.uni-sb.de/imedia/workshops/mm-references.html

Ando, H., Kitahara, Y., Hataoka, N. (1994) Evaluating multimodal interface using spoken language and pointing gesture on interior design system. Proceedings of International Conference on Spoken Language Processing. 567-570.

Bearne, M., Jones, S., Sapsford-Francis, J. (1994) Towards usability guidelines for multimedia systems. Proceedings of the second ACM International Conference on Multimedia (MULTIMEDIA'94), San Francisco, 15-20 October. 105-110.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M. (1994) Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. SIGGRAPH'94. Computer Graphics Proceedings, Annual Conference Series. 413-420.

Catinis, L., Caelen, J. (1995) Analyse du comportement multimodal de l'usager humain dans une tache de dessin. Actes des 7èmes Journees sur l'Ingénierie de l'Interaction Homme-Machine (IHM'95). 123-129. In French.

Cheyer, A., Julia, L., Martin, J.C. (1998) A unified framework for constructing multimodal experiments and applications. Proceedings of the Second International Conference on Cooperative Multimodal Communication (CMC'98). Tillburg, the Netherlands, 28-30 January.

Coutaz, J., Nigay, L. (1994) Les propriétés CARE dans les interfaces multimodales. Actes des 6èmes Journées sur l'Ingéniérie de l'Interaction Homme-Machine (IHM'94), Lille. 7-14. In French.

Denda, A., Itoh, T., Nakagawa, S. (1997) Evaluation of spoken dialogue system for a sightseeing guidance with multi-modal interface. In IJCAI-IMS (1997). 41-48.

Fais, L. (1997) Effects of interface on linguistic behavior in task-oriented, multimodal dialogues. In IJCAI-IMS (1997). 35-40.

Gapp, K.P. (1994) From vision to language: a cognitive approach to the computation of spatial relations in 3D space. Proceedings of the First European Conference on Cognitive Science in Industry, Luxembourg. 339-357.
http://www.dfki.uni-sb.de/vitra/index/node70.html#absb110

Guyomard, M., Le Meur D., Poignonnec, S., Siroux, J. (1995) Experimental work for the dual usage of voice and touch screen for a cartographic application. Proceedings of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems, Vigso, Denmark. 30 May - 2 June 1995. 153-156.

Hare, M., Doubleday, A., Ryan, M., Bennet, I. (1995) Intelligent presentation of information retrieved from heterogeneous multimedia databases. Pre-Proceedings of the First International Workshop on Intelligence and Multimodality in Multimedia Interfaces (IMMI-1), Edinburgh, Scotland, July 13-14.

Huls, C., Bos, E. (1995) Studies into full integration of language and action. Proceedings of the International Conference on Cooperative Multimodal Communication(CMC/95), Eindhoven. Part II: 161-174.

IJCAI-IMS (1997) Proceedings of the IJCAI-97 Workshop on "Intelligent Multimodal Systems", Nagoya, Japan, August 24.
http://www.miv.t.u-tokyo.ac.jp/ijcai97-IMS/

Martin, J.C., Béroule, D. (1993) Types et buts de coopérations entre modalités dans les interfaces multimodales. Actes des 5è Journees sur l'Ingénierie de l'Interaction Homme-Machine. 17-22. 19-20 October. Lyon. In French.

Martin, J.C., Veldman, R., Béroule, D. (1995) Towards adequate representations technologies for multimodal interfaces. Proceedings of the International Conference on Cooperative Multimodal Communication (CMC'95). Part II: 207-223.

Martin, J.C. (1997) Towards "intelligent" cooperation between modalities. The example of a system enabling multimodal interaction with a map. In IJCAI-IMS (1997). 63-69.

Mignot, C., Carbonell, N. (1996) Commandes orales et gestuelles: Une étude empirique. Techniques et Sciences Informatiques. Vol 15, No 10. 1399-1428. In French.

Oviatt, S., De Angeli, A., Kuhn, K. (1997) Integration and synchronization of input modes during multimodal human-computer interaction. In ACL (1997). 1-13.

Petrelli, D., De Angeli, A., Gerbino, W., Cassano G. (1997) Referring in multimodal systems: The importance of user expertise and system features. In ACL (1997). 14-19.

Siroux, J., Guyomard, M., Multon, F., Remondeau, C. (1997) Multimodal references in GEORAL TACTILE. In ACL (1997). 39-43.

Trafton, J. G., Wauchope, K., Stroup J. (1997) Errors and usability of natural language in a multimodal system. In IJCAI-IMS (1997). 49-53.

Wang, E., Shahnvaz, H., Hedman, L., Papadopoulos, K., and Watkinson, N. (1993) A usability evaluation of text and speech redundant help messages on a reader interface. In G. Salvendy and M. Smith (eds.). Human-Computer Interaction: Software and Hardware Interfaces. 724 -729.