Modeling Multimodal Multitasking in a Smart House

Pilar Manchón, Carmen del Solar, Gabriel Amores, and Guillermo Pérez

***Abstract*—This paper belongs to an ongoing series of papers presented in different conferences illustrating the results obtained from the analysis of the MIMUS corpus. This corpus is the result of a number of WoZ experiments conducted at the University of Seville as part of the TALK Project. The main objective of the MIMUS corpus was to gather information about different users and their performance, preferences and usage of a multimodal multilingual natural dialogue system in the Smart Home scenario. The focus group is composed by wheel-chair-bound users. In previous papers the corpus and all relevant information related to it has been analyzed in depth. In this paper, we will focus on multimodal multitasking during the experiments, that is, modeling how users may perform more than one task in parallel. These results may help us envision the importance of discriminating complementary vs. independent simultaneous events in multimodal systems. This gains more relevance when we take into account the likelihood of the co-occurrence of these events, and the fact that humans tend to multitask when they are sufficiently comfortable with the tools they are handling.**

***Index Terms*—Multimodal corpus, HCI, multimodal experiments, multimodal entries, multimodal multitasking.**

I. Introduction

MULTITASKING is part of human nature. As a matter of fact, we all multitask to one level or another at times: sometimes we need to, sometimes we prefer to. Nowadays however, multitasking has become more of a necessity than an option. Most people are expected or pushed to handle several tasks at once to meet the requirements of a job, or of life itself. Time is a very valuable resource and if multitasking can save us some, then it is a good option.

For a long time we thought that human multitasking was performed as in computer parallel processing, that is, we thought that humans could perform several tasks in parallel in a similar fashion to the way computers can do parallel processing. Nowadays however, it seems that what the human brain does is really sequential processing, i.e., switching tasks very quickly [13]. This quick switching occurs in the Brodmann's area 10 [5, 13], which is part of the frontal lobes. This area is responsible for maintaining long-term goals and achieving them. In forthcoming sections we will see some of the cognitive aspects of human multitasking and how this should affect the design and strategies of multimodal dialogue systems.

Manuscript received November 30, 2008. Manuscript accepted for publication March 3, 2009.

Authors are with the University of Seville, Seville, Spain (e-mail: pmanchon@us.es, carsolval@alum.us.es, jgabriel@us.es, gperez@us.es)

Being '***multitasking***' such a natural and common thing to do then, it seems only logical that a multimodal dialogue system for a smart house should be enabled to handle it. Although coping with multitasking is already quite complex, handling it in a multimodal environment increases its complexity significantly, since it implies a number of ambiguous situations that the system must be able to process.

In forthcoming sections, we will analyze potential gains of enabling a multimodal system to handle human multitasking, as well as what it implies. First, we will present a quick and overall overview of the WoZ corpus. Then, some information on how the human brain handles multitasking and its impact on user modeling will be presented. Later on, we will focus on the experiments' task description for this specific issue and the results obtained from the corpus analysis. Last but not least, we will develop some conclusions and future lines of research.

II. MIMUS Corpus

Although both the WoZ platform and the corpus have been fully described in previous articles [6, 7, 8], it is important to at least outline the main characteristics and motivation behind the corpus. The MIMUS corpus is the result of a multimodal WoZ set of three experiments. The experimental design was stimulated by Oviatt's previous research [9, 10, 11, 12]. The primary objective of these experiments was to collect data in order to extend and configure an existing spoken dialogue system by adding new input and output modalities. Although data in English had been collected in previous experiments, we are not aware of any multimodal corpus in Spanish that would comply with the necessary requirements. The goal was to identify and gather information regarding:

1. **Any possible obstacles or difficulties to communicate:** potential unforeseen aspects of the interaction in this domain.
2. **Any biases that prevent naturalness:** people may not address computers the same way they address other people.
3. **A corpus of natural language in the home domain:** to generate a natural language grammar in the home domain.
4. **Modality of preference in relation to task and scenario:** what do users prefer given certain tasks in this scenario?
5. **Output modality of preference in relation to the type of information provided:** how do users prefer or need to have the information presented to them?

6. **Task completion time:** how long to perform simple tasks, and time variations depending on the modalities they use.
7. **Combination of modalities for one particular task:** how does the combination of modalities affect communication?
8. **Inter-modality timing:** what is the optimal time-frame to considered separate inputs as a unique communicative act?
9. **User evolution, learnability and change in attitude:**, do users preferences and behavioural patterns change as they learn?
10. **Pro-activity and response thresholds in multimodal environments** (Experiment 2)
11. **Relevance of scenario specific-factors/needs**
12. **Multimodal Multitasking:** multimodal input fusion and ambiguity resolution
13. **Language use:** do Oviatt's findings apply in Spanish?
14. **Personification:** do people interact differently or have a different perception if the system is personified?

Subjects had a tablet-PC where they could see the house set up, and they were surrounded by the devices they could control. They could use the pen to click/tap on the screen and/or talk through the microphone. All tasks could be fully performed using either speech or the graphical interface, as well as using them both in combination (multimodally). Subjects chose what to do.

The experiments bring some insight into the users' **speech and pen** multimodal integration patterns on a system application that controls **lights, a blind, a radio, a heater, an alarm, the main door, a security camera, and a telephone**. All interactions were recorded from different perspectives: all the graphical events were automatically logged, the subjects' facial expressions were recorded with a webcam, the full scene was recorded on a digital camera and all the audio was also recorded and later transcribed. The experiment set-up is illustrated in figure 1. For specific information regarding the corpus annotation, number of subjects, etcetera please refer to [9].

The experiments took place in a lab specially prepared to simulate a smart house, where the subjects could see the physical devices turn on/off as well as their graphical representations on the screen. Twenty-one of these subjects complied with the user profile set for the experiments (wheelchair-bound, full upper body mobility). Subjects were alone and undisturbed during the experiments.

The set consisted of:

- Two complementary experiments (1A and 1B) where naïve subjects were interacting with the wizard.
- One experiment (Experiment 2) where the naïve subjects became naïve wizards.

In chronological order, subjects received the appropriate information, filled in the survey, went through 1A, filled in the first questionnaire, went through 1B and finally filled in the second questionnaire; then the WoZ set up was disclosed and explained. At this point, they were asked to perform as wizards for other naïve subjects (these subjects were however not naïve) and trained to use the system as wizards for experiment 2. The objective of this second experiment was to compare, what they claimed they would have liked the system to do with what they would really do given the opportunity.

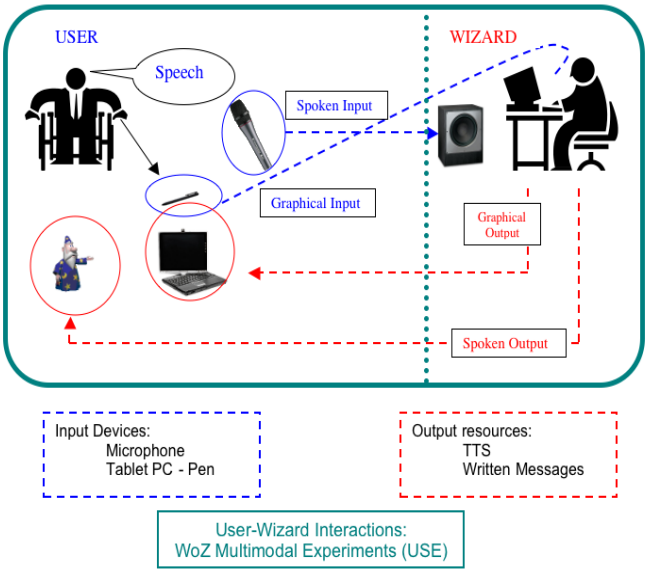

Fig. 1. The experiments set-up.

In previous articles, we have discussed some of these issues [6, 7, 8]. In this article, we will focus on the aspects related to Multimodal Multitasking in a smart house.

III. Human Multitasking

We saw in the introduction that humans do sequential processing rather than parallel processing, so when it comes to multitasking, what it turns out we do is really switching tasks very quickly. What is also a well-known fact is that multitasking is very expensive in terms of cognitive load, especially when tasks are either very demanding or when the subjects are learning or not quite familiar with them.

Multitasking becomes easier when subjects can run part of the process in auto-pilot, i.e., they make part of the process a routine. What happens in this case is that the prefrontal cortex [5] surrenders control to other brain regions: once the task is learned other areas take control [3, 4].

The ability to multitask however may vary according to gender and age: women multitask better in general, and although the reason why is still a mystery, it looks like they use more areas of the brain when they multitask. It also seems like younger adults multitask better: children and adults over 50 tend to have more difficulties [1, 13].

Another interesting issue in human multitasking is what is usually known as the Psychological Refractory Period (PRP), which is basically a delay in responding to the second stimulus in events closely together in time. The task alternation cost could range between 0.5 and 1 second [5]. The fact that humans multitask does not necessarily mean that it is the most efficient way to perform two tasks. Due to this alternation cost or delay, it turns out that multitasking may indeed result in taking more time to perform both tasks

simultaneously than it would sequentially. However, some experiments show that this PRP may not occur depending on the combination of stimuli-modality and response modality. According to the results presented in [3], there was no slowing down in the following combinations:

- Visual stimuli – Manual response
- Auditory stimuli – Vocal response

Since humans tend to minimize the cognitive effort, the obvious hypothesis would be that when multitasking, if a subject is presented with visual stimuli, his quicker or more natural response would be manual (the same interaction modality); and when presented with auditory stimuli, vocal response would logically be more frequent.

Some studies [2] also show that subjects choose the point at which they switch tasks when they have enough time to do so; moreover, they choose the point where "*the interruption would have a lesser effect on dialogue*".

These facts are important in order to understand the motivation behind this analysis, as well as the impact these considerations may have on the design and implementation of multimodal dialogue systems.

IV. Multitasking in Mimus

A. Overall Objectives and Task Description

The main objective of the experiments was to record the interactions of completely naïve subjects with the wizard. The first experiment (1A) consisted of 11 simple tasks in the house. The subjects would turn on or off different devices, make phone calls, etc. Tasks were presented as situations where the subject had to infer what to do. In the second experiment (1B), tasks were more complex than in 1A: the subject was required to performed several actions per task, either sequentially or simultaneously, having to remember relevant information and making rushed decisions. Accessibility, friendliness, usability and naturalness were all taken into account before and after the experiment. Multimodal multitasking and mixed-modality events were encouraged, although not enforced during the experiment. As in the previous experiment, tasks were posed as situations where the subjects had to decide what to do and how to do it. Here are two examples of tasks both in 1A and 1B:

Task 1: (**Experiment 1A**) "*You just got home from work. You sit and relax; now you want to read the book you have on the table, but it is too dark to read where you are.*"

Task 1: (**Experiment 1B**) "*It will be dark soon and you just heard some noise outside. You might want to have some light there and see what's going on with the outside camera*"

In Experiment 1A most tasks were simple, although at some point, an unexpected event occurs that encourages them to multitask. They are talking on the phone when the doorbell rings: they can hear it and they can see it on the screen. Since they could graphically activate the camera and open the door, subjects were able to multitask, that is, they were not forced to stop the current task or wait until it was finished to handle the new situation. In 1B however, tasks usually implied more than one action, and unexpected events occurred several times. Here are some examples:

Task 4: "*You want to make sure that light is visible through the window because your neighbor will bring you a registered letter you were waiting for.*"

Call: "*Hi, this is Charles, your boss. Listen, I got news. It is extremely important that you listen carefully because I need you to make some phone calls straight-away. You must call our best customers, John and Sally in conference call as soon as possible. I know you have their numbers in your phone directory. You must tell them that everything is going well and that the problems they heard about are already solved. Then make sure you call me back the very minute you hang up with them. Talk to you soon. Bye!*"

Task 9: "*You need that package!*"

B. User Motivation

One of the limitations of WoZ experiments is that subjects are not usually as motivated as they would in a real situation. The better we can emulate that motivation, the closer the results might be to the real thing.

In order to emulate the motivation and sense of priority or urgency the user may have in a real situation, subjects were immersed in the experiments using a role-play strategy with simple keys anybody could identify with. Having to call your mother, your boss or someone important from work is something most people can relate to. In addition to this, they were given explanations as to why they needed to do things, the system would remind them of whatever task they needed to do fast, and the messages would also convey the importance of handling everything well and at once: "*call straight-away*", "*best customers*", "*call me back the very minute you hang up*", etc.

It turned out that most subjects did get into their roles and tried their best to cope with the situations.

C. Ambiguity

As previously mentioned, multitasking in a multimodal system may be very intuitive; however, it also brings along some additional complexity that the system must be able to handle. As described in [6, 7, 8], the system allows not only for entries in different modalities, but also for multimodal entries, i.e., users may combine complementary inputs in different modalities in one single communicative act:

User: "*Turn the light on*" + [click on icon]

As illustrated in [6], these combinations may be quite complex to handle, since the timeframe within which the system ought to consider that they are potentially related is not as slim as we would like.

As illustrated in figure 2 [6], during the same experiments 91% of the clicks occurred in the [-2.6, +2.6] interval[1] in examples similar to the previously provided one. If the system

[1] 2.6 seconds before speech onset and 2.6 seconds after speech end

also allows for user multitasking, the ambiguity is unavoidable.

In previous articles [6], the importance and complexity of determining whether the user's inputs were complementary or unrelated was illustrated. Basically, in order to allow for multimodal entries, the system must be able to 'fuse' inputs coming through different channels. This fusion process may be quite complex since the number of factors to decide whether the inputs should be fused or not is significant:

1. Dialogue Moves generated,
2. Modality,
3. Inter-Input timing,
4. Dialogue Move order,
5. Existing Dialogue Moves,
6. Existing Dialogue Histories,
7. Scenario and contextual factors.

In the smart house scenario, we should consider additional factors that may not apply or be available in other scenarios:

8. User profile,
9. User routine or habits.

This decision process is obviously directly related to multitasking as well, since we are to assume that whenever two pseudo-simultaneous inputs occur and they are compatible according to all factors above, then the system must decide whether to fuse them or treat them as separate but simultaneous tasks.

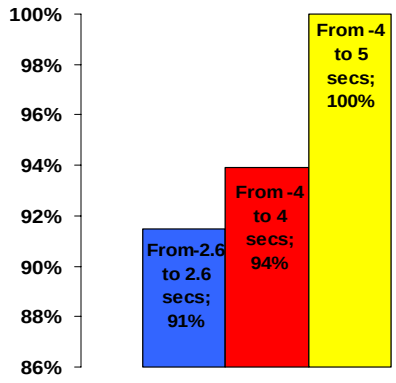

Fig. 2. Clicks before, within and after speech.

D. *Significance of Results*

Since the number of subjects is not too large (21 subjects) and the number of times they multitasked could not really be substantial under the experiments' circumstances, the analysis hereby presented must be understood as a search for trends and direction rather than for undisputable statistically significant results.

V. Results

The data analysis shows that 57% of the users will multitask at least in induced situations, where two equally urgent tasks co-occurred. Before these situations were imposed, no real need to multitask was sensed by the users.

The good news is that once users discovered they could multitask, the number of users who would do it increased. This is not surprising given the fact that multitasking occurs more often when the subjects are habituated to the tools or procedures they are using. As the users interact with the system, their confidence and familiarity grow, which makes them in turn more likely to multitask. Although the experiments did not allow for elongated conversations after the "induced situations", the fact that more multitasking occurred even when the situations were no longer as critical is noteworthy. See figure 3.

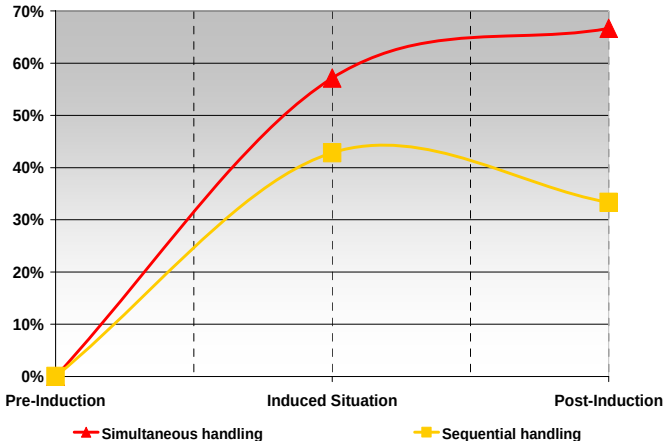

Fig. 3. Simultaneous vs. Sequential users.

Figure 4 presents the same analysis as in figure 3, although the data has been disaggregated in terms of men and women performance. It is worth noting that, as expected, women have more of a tendency to handle tasks simultaneously than men: 100% of women were multitasking simultaneously at the end of the experiment.

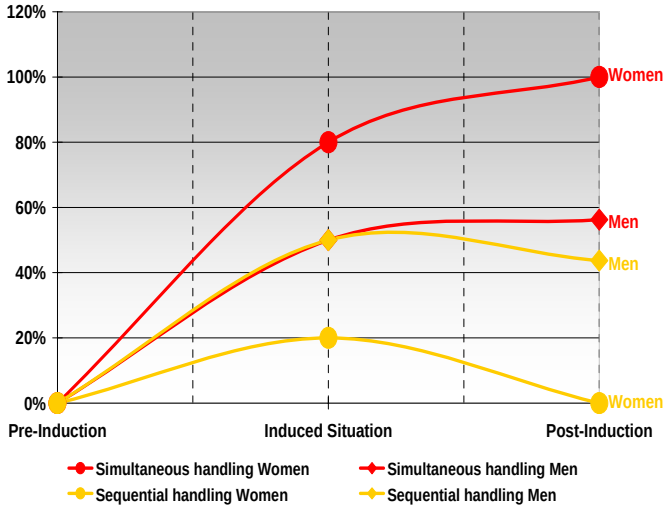

Fig. 4. Women vs. Men simultaneous handling.

Due to the observations regarding the predisposition of humans to be more or less likely to do simultaneous handling depending on their age, the graph in figure 5 has been presented. All 21 subjects ranged from 19 to 54 years old. Subjects above 30 were less inclined to do simultaneous handling than those under 30.

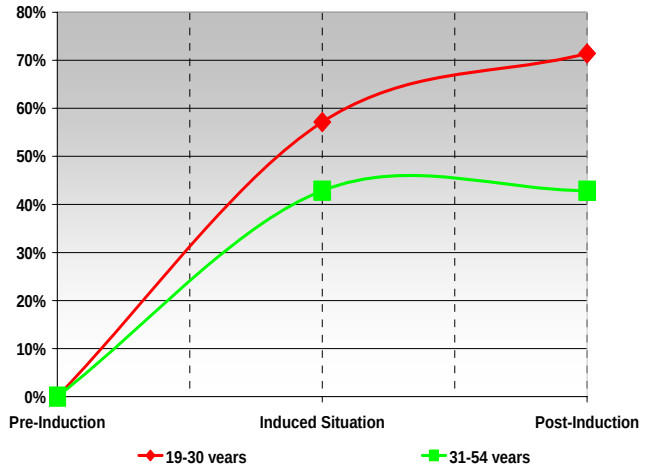

Fig. 5. Simultaneous handling in different age ranges.

The average response time in the first task after the instructions had been provided was 1.156 seconds for simultaneous handlers, ranging from 0.24 to 2.44 seconds; and 1.194 seconds for the rest of subjects, ranging from 0.40 to 1.95 seconds.

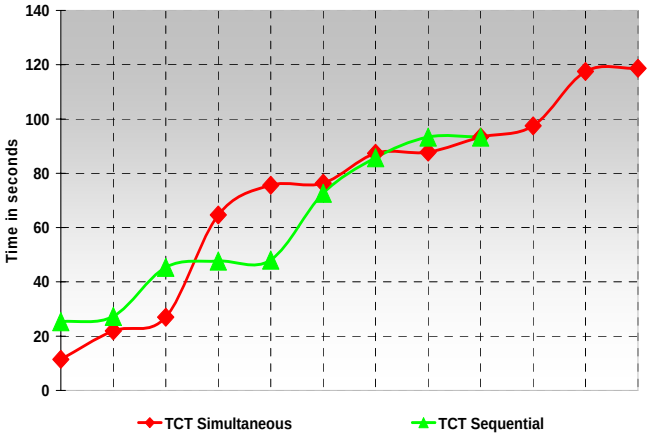

Fig. 6. Task Completion Time (TCT).

With regard to task completion time, simultaneous handlers averaged 73.26 seconds to complete both tasks, whereas sequential handlers took and average of 59.85 seconds.

As mentioned in previous sections, it was reasonable to expect some kind of cognitive effort minimization effect, so that subjects would continue to use the stimuli modality to carry on with the interaction. It only seemed reasonable as well to think that subjects might have a tendency to continue using the modality they were initially using when the second task was initiated, trying perhaps to use alternative channels for additional tasks as long as there were any available. This might also help disambiguate what dialogue or task history the next input was directed to.

The unexpected result here however is how frequently subjects switched interaction modalities. The expected result when the experiment was designed was that, given a task and a modality in use, subjects would continue to use the same modality for the current task; if a second task was presented in a different modality, then the expected behavior was that subjects would usually continue to use the first modality for the first task, and start using the new modality for the new task. The fact that visual stimuli and manual response also minimizes the cognitive effort would make this possibility seem even more plausible.

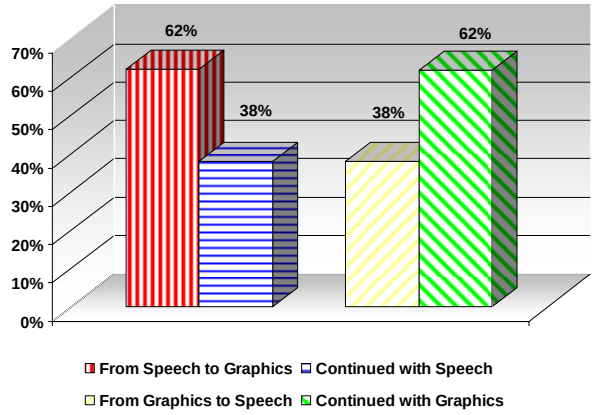

Fig. 7. Switching modalities in an ongoing task.

The first task was a phone conversation (speech); the second task consisted in checking who was on the door and opening it. Since the first task had the speech-channel occupied, the second task was presented on the screen with a graphical menu of options. However, as illustrated in figure 7, 62% of the subjects switched modalities (ended the conversation through the graphical menu) in task 1. Only 38% of the subjects switched modalities in task 2, where the subjects were initially presented with a graphical menu.

If we consider in this distribution of task switching whether the subjects chose to handle the two tasks sequentially or simultaneously, then we find that only 8% of the subjects who handled both task simultaneously did not switch modalities. About 33% of these subjects used only speech for all tasks.

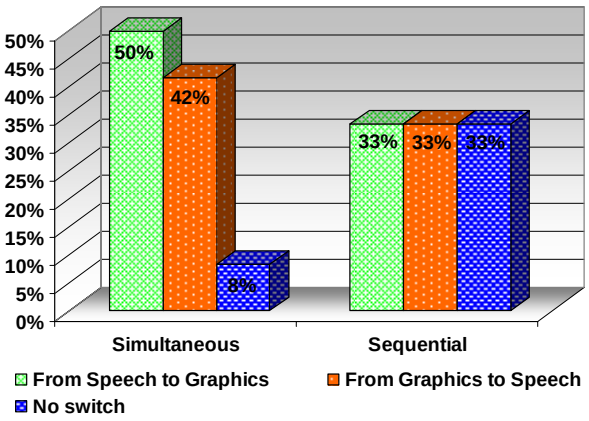

Fig. 8. Switching modalities in an ongoing task.

Only in 5% of the cases in which the subjects were multitasking, two pseudo-simultaneous[2] inputs in different modalities occurred.

[2] They are considered pseudo-simultaneous inputs when the time difference between them is in the [-2.6; +2.6] interval.

VI. Conclusions and Future Work

Although it would have been very interesting to have more data with "induced multitasking", the fact is that these results provide a good starting point to handle multitasking in a multimodal system in a smart house. As a matter of fact, in these results we could say that the relationships we did not find were almost more relevant than the ones we found: some of the cognitive factors that were initially thought to be relevant to model multitasking do not seem to be reliable or significant for that matter.

On the one hand, we can see that more than 50% of the subjects try to multitask alternating their attention to the tasks involved. As the users become more familiar with the system, the number of subjects who multitask increases, so it would be reasonable to assume that a high percentage of users will end up multitasking more frequently. There are not enough data to estimate the right percentage, but extrapolating from the tendency shown in figure 3, it would be safe to say that it should be above 60%. This evidences the need to handle multitasking in this kind of systems.

On the other hand, the results show quite unexpectedly that subjects are more likely to switch modalities than we might have initially considered under the given circumstances. Our initial impulse to bet on "modality coherence" or "cognitive effort minimization" as determining factors in the disambiguation process seems now unpromising. It is true nonetheless that the length of the experiment does not let us see whether theses factors may come into play in more advanced stages, when users are no longer learning or getting used to the system but rather optimizing their performance. In future experiments we might be able to see whether these factors may have an impact at all at any stage. For now, we can at least deduce that in early stages the disambiguation process must rely on additional factors.

If we take into account some additional data from the same experiments regarding the preferred input modality [6] (see figure 9), it may also be interesting to note that some subjects preferred speech over any other modality. As a matter of fact, there were subjects who did not once use the graphical interface. As previously mentioned, 42% of the '*simultaneous multitaskers'* switched from graphics to speech, and 33% of the same group only used speech.

What matters overall is how to model multitasking in the multimodal system, and what issues are relevant at user profile level to adapt the system's behavior to the corresponding profile.

The data available do not suffice to design a fool-proof disambiguation algorithm in multitasking. However we should consider that:

- A significant number of users would multitask given the chance if allowed to.
- Female as well as younger users are more likely to multitask then other users: the system's behavior should probably be different depending on the user gender, age and level of expertise.
- The user's modality preference may over-ride the importance of other factors: given enough time, the system may be able to assign a particular preference profile to users. If these preferences are strong, other factors may not have an impact on performance.
- At early stages at least, the principles of modality coherence and cognitive effort minimization do not help disambiguate events.

Given the high cognitive effort that multitasking implies, what may also be interesting to take into account in a human-aware system is the classification of tasks in different levels of complexity [14]. Our ability to multitask or our tendency to do so may be affected by the combined complexity of these tasks.

In the same line of thought, we may also want to take into account the time at which the tasks are being performed. In previous papers where the experiments procedures were described, it was noted that the experiments were always conducted during the approximate same timeframe. This precaution was taken in order to avoid different levels of performance that could be due to human performance variability throughout the day. By the same principle, the ability of a user to multitask may be impaired or enhanced depending on the time at which the tasks are being carried out, which may in turn have an impact on the system's adaptation strategies.

It may also be worth noting that the number of 'ambiguous' cases in which different inputs were pseudo-simultaneous and turned out to be separate tasks is quite small. Nonetheless, this may also be due to the subjects' lack of familiarity with the system. As we increase the time interval to include all 100% of the multimodal entries (see figure2), i.e., the borderline cases, and users become more proficient with the system, the number of ambiguous cases may also increase. This issue may be studied in future research.

Future work in this area also implies the implementation of human-aware strategies in a real world system. With the current system, we will be able to analyze the users' behavior. As more data are collected, we might be able to see whether there is any discrimination in terms of the users' gender or age, variability as the users become more proficient with the system, differences in performance according to time of day or task difficulty and the impact of the system's ability to adapt to different profiles.

Acknowledgment

Our thanks to the VI Frame Program TALK Project which provided the corpus, and to the GILDA Project (Spanish Ministry of Industry) within which this analysis has been conducted.

References

[1] Bush, C., "How to Multitask," *New York Times Magazine*. April 8, 2001.

[2] Heeman, P., Yang, F., Kun, A. and Shyrokov, A., "Conventions in Human-Human Multi-Threaded Dialogues: A Preliminary Study," in *Proceedings of IUI'05*, San Diego, CA, USA. January, 2005.

[3] Levy, J., Pashler, H., "Is dual-task slowing instruction dependent?" *Journal of Experimental Psychology: Human Perception and Performance*, 27, 4, pp. 862-869, 2001.

[4] Levy, J., Pashler, H., "Task prioritization in multitasking during driving: Opportunity to abort a concurrent task does not insulate braking responses from dual-task slowing," *Applied Cognitive Psychology*, 22, pp. 507-525, 2008.

[5] Pashler, H., "Task switching and multitask performance," in Monsell, S., Driver, J. (eds.). *Attention and Performance XVIII: Control of mental processes*. Cambridge, MA: MIT Press, 2000.

[6] Manchón, P., del Solar, C., Amores, G., and Pérez, G., "Multimodal Interaction Analysis in a Smart House," in *Proceedings of the 9th international Conference on Multimodal interfaces ICMI '07*, Nagoya, Aichi, Japan, November 12 - 15, 2007.

[7] Manchón, P., del Solar, C., Amores, G., and Pérez, G., "The MIMUS Corpus," in *Proc. of LREC 2006 International Workshop on Multimodal Corpora From Multimodal Behaviour Theories to Usable Models*, pp. 56-59, Genoa, Italy, 2006,

[8] Manchón P., Pérez G., and Amores G., "WOZ experiments in Multimodal Dialogue Systems," in *Proceedings of the ninth workshop on the semantics and pragmatics of dialogue*, pp. 131-135, Nancy, France. June, 2005.

[9] Oviatt, S. L., "Multimodal interactive maps: Designing for human performance," *Human-Computer Interaction*, (special issue on "Multimodal interfaces"), pp. 93-129, 1997.

[10] Oviatt, S. L., DeAngeli, A. and Kuhn, K., "Integration and synchronization of input modes during multimodal human-computer interaction," in *Proceedings of Conference on Human Factors in Computing Systems CHI '97*, 1997.

[11] Oviatt, S., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M. and Carmichael, L., "Toward a Theory of Organized Multimodal Integration Patterns During Human-Computer Interaction," in *Proc of 5th International Conference on Multimodal Interfaces, ICMI'2003*, pp. 44-51, Vancouver, British Columbia, Canada, 2003.

[12] Oviatt, S., Coulston, R., and Lunsford, R., "When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns," in *Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI 2004)*, State College, Pennsylvania, USA, October 14-15, 2004.

[13] Wallis, C., "The Multitasking Generation," *TIME Magazine*, March 19, 2006.

[14] Wild, P., Johnson, P. and Johnson, H., "Towards a Composite Modeling Approach for Multitasking," in *Proc. of Tamodia'04*, Prague, Czech Republic, November, 2004.