

Multimodal Interaction within Ambient Environments: an Exploratory Study

Abstract. In this paper we present the results of a Wizard of Oz experiment which shows that speech is a favorite modality within smart room environments for a large part of users. The experiment also shows that output modalities used by the system have an important influence on the users' input modalities for a large category of users. The experiment took place in a smart room because this kind of environment does not require any particular knowledge about computers and their use and thus allowed us to study the behavior of ordinary people including subjects who are not familiar with computers. We think that the results presented in this paper will be useful for the design of intelligent multimodal systems.

Keywords: Multimodal, Interaction, Wizard of Oz, Modality, Ambient Environment, Ordinary People.

1 Introduction

Multimodal interfaces have been extensively studied for several years. At the beginning, most of the studies addressed the input side of multimodal interfaces [1], mainly trying to solve the problem of information fusion using different approaches, to structure the design space of multimodal interfaces and to study the multimodal behavior of users. More recently, researchers have been more interested in the output side of multimodal interfaces [2]. Different issues have been addressed such as the intelligent presentation of information and the adaptation of output multimodal interfaces. However, the input side and the output side of multimodal interaction are not independent phenomena and we need to know more about the relationship that exists between the multimodal behavior of a user and the multimodal behavior of a system and how each side may affect the other side. In this paper, we focus on the influence of system output modalities on user input modalities in an ambient environment. We present a Wizard of Oz experiment which took place inside a smart room. This kind of environments allows one to interact with familiar everyday-life objects such as lights, windows, etc. Furthermore, ambient environments do not require particular knowledge about computers and thus are well adapted to study the behavior of ordinary people who do not have particular knowledge about computers.

2 Related Work

Related research work mainly addressed the study of multimodal user behavior in order to evaluate the usefulness of multimodality and to extract patterns allowing the development of more robust multimodal systems.

In [3], the authors showed how speech and gesture were combined by users. They showed that the task can influence users' choice of input modalities. Tasks without spatial components were almost always completed by users using speech only, whereas those with spatial components were performed using multimodal combinations. In [4] the authors studied the stability of multimodal patterns depending on the user's age. The patterns of children presented many similarities with those of adults even though children follow a simultaneous multimodal pattern more often. In [5], the authors showed that users switch modes under certain contexts. For instance, when recognition errors occur, users will shift from one mode to another in order to recover. In another interesting work, Oviatt & al. [6] focused their study on the speech modality. They showed how users adapt their speech signal input to converge with a text-to-speech output system. However this study is mainly concerned with how users adapt the attributes of a given modality (speech) rather than how users are influenced when choosing between different input modalities. To sum up, previous works showed that different factors may influence the users' choice of input modalities. Task, context, users' age and errors have been the most studied parameters. In this paper, we focus on system output modalities. In our study we preserve all the previous parameters from any changes and we vary only the system output modality (tasks follow the same schema, context does not change, all users are adults and error conditions are minimized).

3 Experiment

We conducted a Wizard of Oz experiment to compare the input modalities used by the subjects when the output modality of the system changes. The experiment took place inside a smart room in our lab. This smart room is a testbed for "ambient intelligent" environments, in which people can interact with assisting computers in a natural way, through various modalities. Fifteen unpaid adult volunteers, 5 males and 10 females, aged 35 years on average, served in the experiment. There were three different output modalities for the system: text, graphics (icons) and speech synthesis. For the user, three input modalities were available: speech, pointing gestures on a touch screen and button presses on a remote control. In the remainder of this paper we will refer to these three modalities, respectively by speech, touch screen and remote control (even though the touch screen and the remote control are devices, not modalities). Speech recognition and the detection of touches on the screen were simulated by the operator of the WoZ system (who was located in another room). Speech recognition was simulated because we didn't want to use an intrusive device (microphone) and we wanted to minimize recognition errors without constraining users to a limited vocabulary. We induced the users into believing that microphones were embedded into the room walls and that the screen was a touch screen. These three modalities were chosen because of the ambient context where the experiment took place. For instance, using a keyboard to switch the light on would not have been very relevant to an "assisted living" setting. For a first study, the issue of combined modalities is not considered here. Studying the influence of system output modalities on user input *combined* modalities, is planned as one of our future experiments.

Therefore we defined the experiment's tasks in a way that allows the subjects to interact using one single modality (chosen among three of them). The subjects had to perform six tasks: switch the light on, listen to music, increase the sound level, decrease the light level, stop the music and switch the light off. As said before, the studied factor is the influence of system output modalities. Any other factor such as task complexity has to be kept static. Hence, every task has the same 5-steps structure. Each task begins by a task presentation step (1). During this step and to avoid influencing the subject with a particular modality, the system introduces the task using all output modalities (Text + Graphics + Speech) in a redundant way. Then we observe which input modality the subject uses to start the task (2). Then the system asks a question (3) to get more details about the task parameters and proposes two possible answers. Now, the system uses only one modality (speech, text or graphics). Contrary to step (1), this step of the experiment aims at trying to influence the user with a particular modality. We then observe which input modality is used by the subject to answer the question (4). Finally and after getting the answer from the subject, the system performs the action (5) (switches the light on, plays music, etc.).

We predicted that the output modality used by the system would have an influence on the input modalities used by the subject. In other words, we predicted that the subject would use different input modalities depending on system output modality.

4 Results

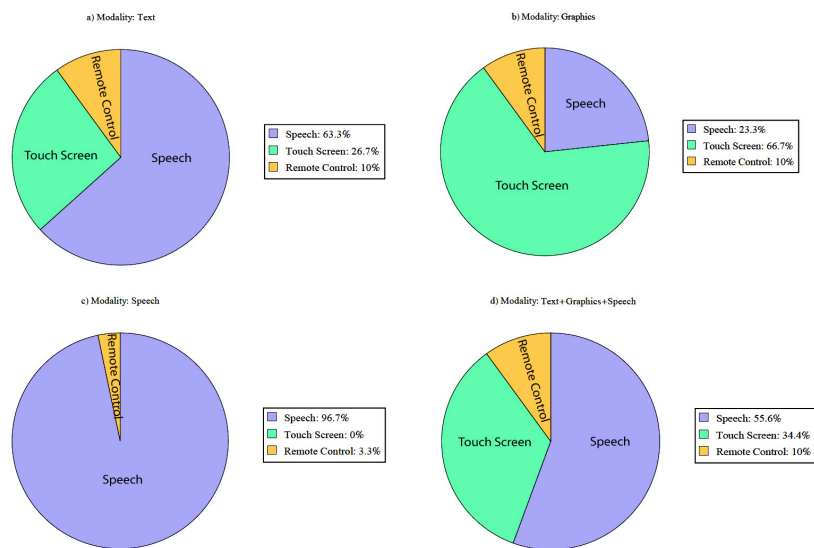


Fig. 1. Input modalities used by the subjects when the system uses Text (a), Graphics (b), Speech(c), Text + Graphics + Speech (d).

We analyzed the input modalities used by the subjects depending on system output modalities (Fig. 1). The results show that speech input is a favorite modality within smart room environments for a large part of users, except when graphics modality is used by the system. In this case, the touch screen is preferred to speech. These results

are confirmed by the answers to the questionnaires. Moreover we can observe a strong relation, on the one hand, between verbal modalities, and on the other hand between non-verbal modalities. For instance when the system uses text or speech, users tend to use speech. However, when the system uses graphics then users tend to use pointing gestures on the touch screen. Finally, we can also observe that when the 3 output modalities are used simultaneously then speech is the preferred modality but with a lower percentage than when text or speech synthesis are used alone. It seems that each modality has a kind of *influence power* and when the modalities are combined their influence powers are combined too. This is a very interesting phenomenon which needs to be investigated and confirmed in further experiments.

5 Conclusion and Future Work

We have developed a smart room and WoZ platform to study Human-System interaction within ambient environments. Our first experimental study concerned the multimodal interaction within such environments. We have presented a Wizard of Oz experiment which allowed us to study the relation between system output modalities and user input modalities. The experiment shows that the output modalities used by the system have an important influence on the user input modalities. The experiment shows also that speech is a favorite modality within smart room environments for a large part of users, except when graphics modality is used by the system. We think that these results will be useful for the designers of intelligent multimodal systems. In future work we will build another Wizard of Oz experiment where the tasks can be performed either by using single modalities or combined modalities so that we can study if the monomodal-multimodal system behavior may influence the monomodal-multimodal user behavior. This would allow us to formalize the notion of *influence power* of a modality and study the influence of combined modalities to see if there is a law which explains how these influence powers are combined.

References

1. Bolt, R. A.: Put-That-There: Voice and Gesture at the Graphics Interface. In Computer Graphics, vol. 14, n° 3, 262-270 (1980)
2. Bordegoni, M., Faconti, G., Maybury, M.T., Rist, T., Ruggieri, S., Trahanias P., Wilson, M.: A Standard Reference Model for Intelligent Multimedia Presentation Systems. In Computer Standards and Interfaces 18 (6-7), 477-496 (1997)
3. Oviatt, S.L., De Angeli, A., Kuhn, K.: Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. In Proc. of CHI '97, 415-422 (1997)
4. Xiao, B., Girand, C., and Oviatt, S.L.: Multimodal Integration Patterns in Children. In Proc. of ICSLP'2002, (2002), 629-632 (2002)
5. Oviatt, S. L., Bernard, J. Levow, G.: Linguistic adaptation during error resolution with spoken and multimodal systems, Language and Speech, Special Issue on Prosody and Conversation, vol. 41, nos. 3-4, 415-38 (1999)
6. Oviatt, S.L., Darves, C., Coulston, R.: Toward Adaptive Conversational Interfaces: Modeling Speech Convergence with Animated Personas. Transactions on Human Computer Interaction (TOCHI), Special Issue on Mobile and Adaptive Conversational Interfaces, 11(3), 300-328 (2004)