Journal title: Applied Artificial Intelligence

Article title: The Importance of Adaptivity to Provide Onboard Services: a Preliminary Evaluation on an Adaptive Tourist Information Service Onboard Vehicle

DOI: 10.1080/08839510490463442

Author: Cristina Gena and Ilaria Torre

Year: 2004

The original publication is available at www.tandfonline.com

http://www.tandfonline.com/doi/abs/10.1080/08839510490463442

@article {tandfonline:10.1080/08839510490463442,
    author = {Cristina Gena and Ilaria Torre },
    title = {The Importance of Adaptivity to Provide Onboard Services: a Preliminary Evaluation on an Adaptive Tourist Information Service Onboard Vehicle },
    journal = { Applied Artificial Intelligence },
    publisher = { Taylor & Francis },
    issn = { 0883-9514 },
    volume = {18},
    number = {6},
    year = {2004},
    pages = {549-580},
    keyword = {Computer Science},
    url = {http://dx.doi.org/10.1080/08839510490463442},
}
The importance of adaptivity to provide on-board services

A preliminary evaluation of an adaptive tourist information service on board vehicles

Cristina Gena, Ilaria Torre
Dipartimento di Informatica – Università di Torino
Corso Svizzera 185 – 10149 Torino (Italy)
Phone: +39 011 6706711
Fax: +39 011 751603
Email: {cgena, ilatorre}@di.unito.it

Abstract

In this paper we present the results of a preliminary evaluation of MastroCARONTE, an adaptive system which provides tourist information on board cars. MastroCARONTE personalizes the tourist suggestions (about hotels, restaurants and places of interests), the way they are presented and its own behavior according to a model of the user (driver) and a model of the context of interaction (e.g., time, location, driving conditions). After a short description of the role that adaptation and personalization can play in mobile systems installed on vehicles, we briefly describe the goals and architecture of MastroCARONTE, focusing on the forms of adaptation that it implements. We then describe an evaluation exercise that we performed on the system. This preliminary evaluation shows that the system can indeed provide immediately the pieces of information that are most suitable for a user, presenting them in a way which is compatible with the user capabilities and the risk of the contextual situation. We emphasize this point, since it is a must for a system to be used safely by a driver on a car.

Keywords: adaptation, personalization, user and context modeling, mobile applications on vehicles, evaluation.
1. Introduction
The goal of providing large amounts of information at the driver’s fingertips is becoming more and more important in the last decade. While ten years ago only luxury cars were equipped with complex electronic devices and information systems, today these systems are available even on small mass production vehicles. The dashboards of these cars include, and in some cases integrate, an on-board computer (providing information about the trip and the car itself), a GSM telephone, possibly with GPS (for a precise localization of the car), a navigation system (providing maps and information about roads and routes), besides more common devices such as radio, CD/DVD player, etc.

The consequence of the presence of these electronic systems, and of the fact that people are spending more and more time on cars (for commuting and working or for pleasure), was the provision of new kind of services. For example, systems for accessing news (general news or news about the weather or the traffic conditions) or information about facilities in the area where the car is located (e.g., gas stations, restaurants, hotels, shops) or about tourist locations or even systems for connecting to the Internet (mail and web services or workplace services) are being designed and some of them are already on the market. Most commercial services are based on Call Center solutions; typically they use the global positioning system (GPS), satellite technology and a hands-free, voice-activated cellular phone to link the driver and the vehicle to a Call Center, where advisors provide real-time help for different services: Emergency Services, Remote Diagnostics, Tourist Services, Roadside Assistance, Route and Parking Support, etc. (see for example the ONSTAR solution by [General Motors], ASSISTS by [BMW]).

The availability of these systems represents an opportunity but could also become a problem for the driver. On the one hand, the services may be very useful or even necessary; on the other hand, they may distract the driver causing serious dangers for active and passive safety [Green 00]. Several studies claim the importance of taking into account ergonomics and human factors, considering that the car environment is a special one, and thus, different from those traditionally studied by HCI (see the discussions in [Srinivasan 99], [Campbell et al 98]). There are even some attempts to define safety and human factors standards and regulations for these systems, as it can be seen in the reports by the Society of Automotive Engineers Intelligent Transport Systems Division SAE-ITS and in ISO reports such as [SAE 99].

Since the 80’s, moving mainly from the Artificial Intelligence community, and spreading in numerous fields, adaptation and personalization techniques have grown significantly. In the automotive environment these techniques have not been greatly exploited jet. However they could play a very important role and could
significantly contribute to solve some of the typical problems of these applications, such as the limited time for grasping all the information on the display, the limited size of the display, the impossibility of browsing and the continuously changing context around (but also inside) the car. Indeed, adaptation and personalization techniques allow a system to tailor its behavior to the user, to the context and to the circumstances in which the system is used, by varying the type and number of services that are made available, the way a service is accessed (the interface), the content of a service itself (e.g., the type and amount of information that is presented), etc.

Believing in the potentials of such techniques applied to this area, in 2000 we started the study of a framework for on-board adaptive systems and implemented a prototype, MastroCARONTE, which exploits adaptation and personalization techniques to provide tourist information to a driver (see [Console et al. 02] and [Console et al. 03] for a discussion on the framework and a detailed description of the system).

Given the complexity of adaptive systems, due also to the discretionary choices they unavoidably carry out, evaluation is considered a very important subject in the user modeling and adaptive systems community (see [Chin 01], [Chin and Crosby 02]). As recommended by the paradigm of user-centered design [Norman et al. 86], it should be performed already in the design phases to get immediate feedback from users. For such a reason we started an evaluation exercise of the first prototype of MastroCARONTE. As a specific objective of our evaluation, we aimed at showing that adaptation and personalization can contribute to the achievement of two major goals:

1. Selecting exclusively the information (service) that is most relevant for the driver in a given situation and context of interaction – car position, time of the day, user’s preferences and priorities, etc.

2. Selecting the most suitable way for presenting the information, given the driver’s characteristics and the context of interaction – context risk level, driver cognitive load, etc.

In other words, we aimed at checking whether the recommendations of the system indeed correspond to the user’s preferences and needs and whether the mode and format of the presentation is in accordance with the user’s features and contextual situation.

In order to have a confirmation of these claims, we chose one of the specific services provided by MastroCARONTE - tourist information about hotels and restaurants of Turin -, and tested the system behavior on a large group of users\(^1\). In particular, we performed two types of evaluations (then split in other sub-tasks), aimed at

\(^1\) We tested the system on 107 individuals, using a simulator that we implemented on a laptop computer.
verifying that i) recommendations were displayed in an adequate and safe way, given the current contextual condition and the user’s cognitive load and, clearly, that ii) their content matched the user’s interests, thus avoiding the user to browse alternatives in order to search the right information.

The main objective of this paper is to report the interesting results obtained from this evaluation exercise, in order to share our experience and discuss some critical points and benefits arising from the adoption of personalization techniques for on-vehicle information services. In particular, the paper is organized as follows. Section 2 provides an overview on adaptation and personalization, focusing on the opportunities and advantages for applications on mobile systems and, specifically, on-board vehicles. Section 3 sketches the framework and architecture of MastroCARONTE. Section 4 describes the evaluation exercise while Section 5 reports the results of the evaluation. Section 6 concludes the paper.

2. Adaptation and personalization: from web to mobile, to on-board applications
In this section we provide an overview on the progress of research on adaptive systems with the specific aim of positioning MastroCARONTE in the evolution of the field, from web to mobile, to on-board applications.

Just for delimiting the area, we specify that adaptation can be defined as the ability of a system to tailor its behavior with respect to a set of circumstances which usually regard the user(s) who is (are) interacting with the system, the context of interaction and possibly other parameters (see the overviews in [Brusilovsky 96, 01], [Brusilovsky et al. 98], [Kobsa et al. 01]). The first systems where simple personalization techniques had been applied were the hypertexts in the late 80’s and then the Web sites in the mid 90’s. Different aspects of a system may be involved in the adaptation process; very roughly one may distinguish between “content adaptation” (i.e., the ability to provide different services or different pieces of information or of varying their granularity/detail), “interface adaptation” (i.e., the ability to vary the mode and/or the form of the interaction), “behavior adaptation” (i.e., the ability of the system to alter its own control flow).

This ability is usually related to the fact that the system is based on intelligent agents which maintain a model of the user and a model of the context of interaction and use these models for problem solving. As we shall see, the model of the context is especially important in mobile and on-board applications.

Several and different applications of adaptive systems have been designed in the last ten years. For example, recommender and e-commerce systems aim at selecting the most appropriate product or service for a user (e.g., TELLIM [Joerding 99], SETA [Ardissono et al. 99]); tutoring systems aim at supporting and helping a student learning the concepts, with exercises that are suitable for her (e.g., ELM-ART [Weber et al. 97], AHA [De Bra et al.
information retrieval systems aim at filtering only the pieces of information that are most relevant for a user (e.g., AVANTI [Fink et al. 98], ELFI [Schwab et al. 00]), and so on.

**Mobile adaptive systems.** Mobile systems (phones, PDAs, navigation systems, etc.) constitute a new interesting and promising field of application for these technologies. Indeed, with respect to other systems, the need of adaptation is stronger and can provide more significant benefits. Consider the case of accessing an information server from a PDA or from a mobile phone. First of all, transferring information to these devices is slower than in other networks and is expensive; this suggests that transferring only the pieces of information that are relevant for the user during a specific session (or interaction) is critical. The limitation of the interface and of the browsing facilities is a further example for this need: browsing alternatives is complicated and time consuming or even impossible in some cases and very often the user has no chance of doing that; this may make the system completely useless (see for example [Smyth et al 02] who showed how adaptive navigation techniques can enhance the usability of WAP portals by personalizing the menu on the basis of the user’s historical usage, thus obtaining a 50% click-distance reduction and a corresponding 40% navigation time reduction. See also [Billsus et al. 00] who proposed the exploitation of past-interaction-based adaptation techniques to visualize large amount of information, of the Los Angeles Times newspaper, in the small display of a PDA).

Further advantages coming from applying personalization techniques to mobile systems regard the possibility of adapting the system behavior to the context of interaction, where context means at least two things: i) the adaptation to the contextual conditions: context awareness (see for example [Muller et al. 01] who proposed the use of machine learning and user modeling techniques to recognize the resource limitations of users, such as time pressure and cognitive load, on the basis of their speech) and ii) the adaptation of contents to the location: situated information and navigation supports. Several systems have been developed in this area (see for example HYPERAUDIO [Not et al. 98], a museum guide on PDA; HIPS [Opperman et al. 99], another museum guide, REAL [Wahlster et al. 02], a mobile navigation system adapting its interface to the user’s changing situation by switching between different positioning technologies such as GPS, GSM, UMTS (outdoor navigation) and Bluetooth (indoor navigation), etc. and by taking into account the user’s cognitive limitations for a given situation and the limited technical resources of the devices.

**On-board adaptive systems.** Applications on board vehicles can be considered, under some respects, a special case of those on mobile systems and are, in our view, very interesting and promising. However, this area
received limited attention and in literature there are not many examples of systems designed for on-board use (we shall discuss these systems later). The literature contains several works dealing with the provision of multiple views, accessible by different devices (for details, see [Healey et al. 02], [Herder and van Dijk 02]; see also the projects EMBASSI [Kirste 2000] and SmartKom one [Wahlster et al 01]) and it is right that every mobile system, such as a PDA or phone, might be used on a vehicle but, as several studies pointed out [Campbell et al, 1998], the vehicle special context and the safety issues impose specific design problems for on-board services.

Let us briefly discuss the reasons why it is interesting to design on-board adaptive systems. First of all the car is a context where the driver is performing a primary activity (driving) which requires the utmost attention and requires that very little attention is devoted to secondary activities such as the use of a support system. A consequence for developers of these systems is that the design of Human Machine Interfaces (HMI) must primarily take into account these issues – design guidelines [Campbell et al, 98] and recommendations for interface evaluations have been developed to this purpose [Hankey et al, 00].

A main issue, therefore, is that the user has no chance at all to interact actively with a system, e.g., to select among a set of services or to browse a space of alternatives and thus it is important that the system by its own provides the “right” type of service (information), at the “right” time and in the “right” way. This requires that the system has the ability to decide what “right” means for each specific user in each specific context of interaction. Notice that the three uses of the word “right” correspond exactly to the three main forms of adaptations singled out at the beginning of this section, “content”, “behavior” and “interface” adaptation, respectively. Thus adaptation, and specifically behavior and interface adaptation to various features of the driver (e.g., cognitive capabilities, preferences) and of the context (e.g., driving conditions, weather, time, location) can contribute to making the use of these systems more compatible with the driving task, without affecting safety. This may be achieved in different ways; for example, the system may use different communication channels (vocal rather than graphical or textual) according to the user’s features and to the driving conditions, or may activate autonomously whenever something interesting for the user has to be communicated to her but the driving conditions do not allow her to access the system. Similarly, the system should be able to delay the provision of information if the current situation seems to be critical.

It is worth noting that, compared to other environments, the vehicle technological infrastructure - including an on-board trip computer, a local net, sensors and control units - allows a system to gather and process a lot of information about the driver and the context around her (the weather and the car conditions, the number of
passengers, the speed, the date and the time of the day, and especially the location and the direction of the car). This increases enormously the number of dimensions along which it is possible to personalize the services and moreover it reduces the infrastructural complexity of the system. See, for a comparison, projects regarding pedestrian mobile applications and navigation systems (ARREAL [Wahlster et al. 02], DeepMap [Malaka et al 00] of the European Media Lab) and projects which use smart environments and which exploit techniques of detection, such as infra-red detection, bluetooth, ultrasounds, etc. (see again HYPERAUDIO [Strapparava et al. 98], HIPS [Opperman et al. 2000], IRREAL [Baus et al. 02], etc.).

Although these potentials, the public literature contains very few works about the specific subject of adaptation on-board car. Some notable examples are [Kosch et al. 01] who presents an experimental setup of an environment for ubiquitous computing in automobiles, carried on at BMW and [Fiechter et al 2000], [Thompson et al 2000] who present applications at Daimler Chrysler for personalized restaurant and route selection. Finally, some aspects concerning adaptation to driving conditions have been taken into account in Comunicar [Montanari et al 00]: the messages on the dashboard are displayed according to their priority and to the driving conditions and similarly the access to communication devices (e.g., mobile phone) is mediated by an agent which monitors the driving situation (e.g., an incoming call is delayed if the speed of the car is too high).

With respect to these on-board systems, our framework aims at exploiting all the three forms of adaptation mentioned previously: only the simultaneous adaptation toward content, interface and behavior allows a system to provide services that are: (i) useful/interesting for the user (ii) integrated (iii) safe. Notice that safety is usually related to the ability of the system to adapt the channel of interaction and the modality to provide messages, however, as we will show, adaptivity toward contents is extremely relevant as well.

In the following we will present our proposal for a framework and architecture for adaptation on-board cars, which takes into account the goals identified in this section. Then we will move to the specific prototype that is the object of the evaluation.

3. An architecture for on board adaptive systems and a prototype: MastroCARONTE
In the past two years we defined a framework and architecture for on-board adaptive systems, implementing a prototype application MastroCARONTE, for providing tourist information to the driver [Console et al 02, 03]. In the following we analyze the principles of the framework and architecture, providing concrete examples taken from MastroCARONTE.
• On board adaptation is based on *explicit models* of the user and of the context. The former includes some features that are application independent (e.g., general preferences but also cognitive characteristics such as visual capabilities, ability to capture information on a screen) and some that are application dependent (e.g., interests or propensity to spend are relevant for the tourist services domain). The context model includes many pieces of information such as the location of the car (from a GPS) as well as the driving conditions and the external conditions (e.g., type of road, traffic, speed, weather, time of the day, presence of passengers; travel information such as duration, distance from home, direction, …). These pieces of information can be obtained from sensors available on the car (e.g., speed or travel information) or inferred (e.g., the presence of passengers can be inferred from the seat belts sensors; the traffic conditions can be estimated from the position, the date and time of the day and the speed).

• The three forms of adaptation (content, interface, behavior) are equally important in our framework and the system should infer as much information as possible about the user autonomously, without requiring a lot of interaction. This means that the model must be refined or revised continuously according to the user’s behavior. We decided to base the initial model on stereotypical information (in order to provide personalized recommendation from the very first interaction with the system) and then track the user’s behavior to collect data for revising the model. In the case of MastroCARONTE the stereotypes have been derived from a psychographic research about the social and cultural behavior of the Italian population, while the possibility of learning from the user’s behavior is based on a log of the interactions with the system (specifically, the log regards the tourist facilities on which the user requested details or information about the route or which she called or she visited). Statistics derived from this log are used by a set of rules which refine the user model accordingly.

• The same system must support multiple users of a car and the same user can use the system on multiple cars. Therefore, we decided to store all information about a user (and in particular the user model) on a smart card which can be inserted in the dashboard on any car on which the system is installed.

• The system is based on a distributed architecture with a client on the car and servers accessed via a GSM or GPRS connection. This means that the exchange of information has to be minimized. In the case of MastroCARONTE, the tourist databases are located on the server but are also replicated on the car on a CD/DVD. Since in this way the information on the car might be out of date, we defined a protocol such that only the pieces
of information that are out of date are transferred from the server to the car. The tourist database contains information about hotels, restaurants and tourist locations; each item is annotated with extra information to be used for a personalized selection.

- The intelligence is distributed among several agents, some of which are on-board the car and some of which are on the server; in particular:
  1) The agent that is in charge of “behavior” adaptation is on-board and is mainly in charge of deciding if and when the system must activate autonomously and the type of service/request that is most useful to the user at each activation. MastroCARONTE activates in special conditions, according to the user’s preferences; for example it may suggest an hotel suitable for the user if it is late at night, the user is traveling for many hours and is not directed toward home.
  2) The agents for “content” adaptation are distributed. One is located on-board and is in charge of (i) contacting the server, sending relevant portions of the user and context model and (ii) once a response from the server is received, filtering out and ranking the information/services to be displayed to the user, given the full user and instantaneous context models. Another agent is located on the server and receives requests from the car, performing a first filtering and ranking of the services/information to be sent to the car.
  3) The agent for “mode of interaction and interface” adaptation is located on the car and selects the most appropriate mode of interaction, given the user model and the context and driving conditions. In MastroCARONTE the alternatives are the following: a vocal interface (to be used in situations when the driver is alone and cannot be distracted) a graphical interface, with five different styles, ranging from a very simple one in which the items in the reply are presented one at a time, to more complex ones. Also the number of extra functions (automatic connection to the phone, connection to the route planner, bookmarking) is adapted to the user preferences and context. Figure 1 contains a scheme with the main components of our architecture, instantiated on the MastroCARONTE application.

MastroCARONTE has been implemented under VxWorks, the real time operating system on which Magneti Marelli car navigation systems actually runs. In particular, it runs on a laptop simulator of this operating system. The agents and server components are implemented using the CLIPS rule-based system and the exchange of information is based on XML. The service database is a multipurpose relational one and, for the evaluation, it has been populated with tourist information about the Turin area.
It contains about 400 restaurants and 150 hotels, and also information about places of interest, which, however, have not been involved in this preliminary evaluation.

For the purpose of the evaluation exercise in the next sections, it is worth spending a few more words on the components that will be mainly involved: the user model, the interface and on the rules used by the content adaptation agent and interface agent to rank items (hotels and restaurants) and to decide the presentation format.

The USER MODEL contains several types of information. In the table below we have grouped them into categories, providing, for each one, the method used to acquire such data, in order to build the User Model.

<table>
<thead>
<tr>
<th>Category</th>
<th>Sub-category</th>
<th>Acquisition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socio-demographic features</td>
<td>E.g. age, gender, job, school level, region of residence</td>
<td>• Direct question (registration form, filled in once, when the user buys the car)</td>
</tr>
<tr>
<td>Preferences</td>
<td>Interfaces preferences (screen, sound setting, etc.)</td>
<td>• User setting</td>
</tr>
<tr>
<td></td>
<td>Preferences for specific service features (e.g. food, style of restaurants and hotels)</td>
<td>• Usage information storing</td>
</tr>
<tr>
<td></td>
<td>Familiarity with the application</td>
<td>• Inference rules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• User’s behavior tracking</td>
</tr>
<tr>
<td>Interests</td>
<td>Interest for aspects such as traveling, art, history, nature, visiting museums</td>
<td>• Initialization with Stereotype</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Updating tracking the user’s behavior</td>
</tr>
<tr>
<td>Spending Style</td>
<td>Propensity to spend</td>
<td>• Initialization with Stereotype</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Updating tracking the user’s behavior</td>
</tr>
<tr>
<td>Receptivity (ability to grasp information)</td>
<td>Cognitive capabilities</td>
<td>• Initialization with Stereotype</td>
</tr>
<tr>
<td></td>
<td>Familiarity with technology</td>
<td>• Updating tracking the user’s behavior</td>
</tr>
<tr>
<td>Cognitive status</td>
<td>Cognitive load</td>
<td>• Direct question (registration form, see above)</td>
</tr>
<tr>
<td></td>
<td>Tiredness</td>
<td>• Inferred from context features (context model*)</td>
</tr>
<tr>
<td>Diseases</td>
<td>Visual or audio diseases</td>
<td>• Direct question (registration form, see above)</td>
</tr>
</tbody>
</table>

* The context model does not require further specifications because MastroCARONTE takes into account and implements exactly the features described in the general architecture.
As one can notice from the third column, the initialization is performed off-line, e.g. at the car dealer, adding some direct questions to the typical registration form that the purchaser is asked to fill in when she buys a car (the same form is available on the web as well, e.g. for second-hand purchases). The initialization is performed once for each user by the user model initialization agent, which uses the classificatory data of the form (socio-demographic features, familiarity with technology and diseases) to predict other features of the user (services preferences, interests, propensity to spend and cognitive capabilities). In particular, interests and propensity to spend, whose values highly contribute to the recommendation of hotels and restaurants, are predicted using stereotypes about the behavior of the Italian population, while preferences and cognitive capabilities are inferred by means of a set of rules. The resulting model is stored on the user’s smart card. These initial estimates are then extended and revised by the system by performing analyses on the behavior of a specific user: MastroCARONTE tracks the user’s actions and updates her model based on statistics taken from this log of actions.

Before moving to the interface, we briefly provide some details about stereotypes, as they will be heavily involved in the analysis of the evaluation results. Stereotypes are derived from a psychographic study (Sinottica Eurisko) on the Italian population, widely used for marketing purposes. This study segments the Italian population in classes of users with homogeneous Life Style; for each class it provides probabilistic information (distribution) about socio-demographic features (e.g. age, school degree, job, geographic area, etc.) and psychographic ones (e.g. cultural level, propensity to spend, interest for traveling, shopping, body care, etc.). These distributions can be regarded as conditional probabilities of the features, given the class\(^2\) (for all the implementation details and in particular for the aspects regarding the dynamic refinement of the user model, which is a complex task, and thus it cannot be explained here, see [Console et al. 02] and [Console et al. 03]).

\(^2\) In this way, applying Bayes theorem, we can compute, for each class (stereotype) \(S\) the probability that a user belongs to \(S\) from the socio-demographic user’s features of the registration form; then we can use this distribution to predict the probabilistic values of the other life style features associated with that stereotype. In this version of MastroCARONTE we used Bayesian network to implement the stereotypes (providing a priori probabilities for the stereotypes) and to initialize the user model. In the network, classificatory inference allows to compute the a-posteriori probability of the stereotypes given the socio-demographic features; predictive inference allows to compute probability distribution for predictive features starting from those of the stereotypes.
The **INTERFACE** for presenting information to the user includes two media: audio speakers and a screen. As regards the screen we defined a number (five in the prototypes) of presentation styles. Each style defines a format (number of items to be displayed, fonts, colors) as well as the number of extra services that the user can access; these services include the possibility of calling a restaurant/hotel, the possibility of getting detailed information about it, the possibility of asking for the route to reach it, the possibility of bookmarking the item (these services are very useful for getting feedback about the user’s behavior).

The **SELECTION AND RANKING OF ITEMS** is carried out by two agents. The content agent on the server retrieves a list of items using some criteria that, using an abstract SQL, could be expressed like that:

```
SELECT identificationKey FROM (restaurants OR hotels) {depending on the type of query}
WHERE Location closeTo currentGPSposition AND Attributes includeAtLeastOneOf UserPreferences
```

The ranking of these items is performed by the content agent on the car and is based on a set of rules that make two types of evaluations. The first one provides a score to each hotel or restaurant, according to the user’s characteristics, namely, her propensity to spend (coming from the predictions in her model) and an estimation of her preferences, such as preference for a type of food or a type of place (computed with another set of rules). The second evaluation regards contextual information (time of the day, distance, type of area – metropolitan, extra urban, etc.) and is combined with the previous one to rank the selected items.

The **SELECTION OF THE INTERFACE**, medium and format, is more complex as it involves several parameters, such as the user receptivity and preferences (but also estimates of the user tiredness), the driving conditions (e.g., speed, traffic), contextual information (e.g., time of the day, weather, etc.). This is achieved by means of a set of rules which operate in steps, first deciding the medium, then the format.

In order to make the process clearer, let us consider an example of interaction. In case of a query concerning restaurants, the agent on the car sends information about the location of the car and about some of the user's features and preferences (preferences about food, propensity to spend). The agent on the server retrieves the restaurants that may be of interest, sending a list of references to the car. The agent on the car asks for up-to-date information for those restaurants whose reference is unknown. Then it ranks the list according to the complete user model and to the instantaneous context conditions. Finally, it passes the ranked list to the interface agent which decides the most suitable presentation medium and format.

In the following we show, as an example, the layouts displayed by the system with three different users/contexts:
a) a university student with high level of receptivity (young, familiar with electronic devices, no visual problems, etc.), medium/low propensity to spend, while driving at a low speed with no traffic (see Figure 2),
b) the same student, while driving at a higher speed, but with non traffic and in a straight way (see Figure 3),
c) a middle age lady with high propensity to spend (see Figure 4).

In all the case, the screenshots on the left show the ranked list of selections, displayed according to the selected format and style; the screenshots on the right show the details of a restaurant, using again the selected format and thus activating different sets of services. The screenshots are exactly as shown on the dashboard’s display of the car.
4. Evaluation of the system

In the UM community, the importance of systems evaluation has been strongly advocated (see [Chin 01], [Chin and Crosby 02], [Petrelli et al. 99]) and now it is a shared principle. The methodologies for evaluating user-adapted systems are generally borrowed from the methodologies used in HCI and by those exploited for evaluating the information selection process (mainly in information retrieval systems). They can be classified in predictive evaluation, formative evaluation, summative evaluation, etc.; for more details see [Preece et al 02] and encompass metrics such as precision and recall, evaluation of the ordering, coverage, MAE and RMSE, reversal rate and sensitivity measures, etc., for more details see [Good et al 99]. All these methodologies should be exploited during the different design phases of the system development, according to the user-centered design approach. Anyway it is important to underline that, compared to the fields from which these techniques have been borrowed, in the context of adaptive applications, the concept of the typical user of a system cannot be applied, and thus each evaluation has to be performed toward the different user profiles and in the different contexts of use [Paramithys et al. 01]. Moreover, its is also interesting to follow a layered evaluation approach which considers the different tasks of the system and which differentiates, at least, problems concerning content adaptation and interface adaptation [Brusilovsky et al 01]. In particular, as regards the automotive environment, in our opinion, an evaluation process requires considering several aspects: first of all, of course, the matching between the real users preferences and the features of the items suggested by the system (contents); second, the correct weight to external conditions, like distance, time pressure, etc., in the selection of the items and especially in their presentation (interface adaptation). Finally and most important, the analysis must evaluate if the system adapts its content, presentation and behavior in order to, and respecting the requirement of, being safe and not intrusive (see [Hankey et al, 00] for details on recommendations for interface evaluations).

Methodology of the evaluation.

With respect to the evaluation principles and techniques mentioned above, and taking into account the specific evaluation requirements for automotive environment, we carried out an evaluation of contents (recommendations) and interface, considered under the usability point of view and, most important, under the safety one. We also integrated it with a layered approach based on the three typical tasks of personalized hypermedia applications, identifiable with i) acquisition method and primary inferences, ii) representation and secondary inferences, iii) adaptation production [Kobsa et al. 01]. We performed a formative evaluation for the points i and ii and a both formative and predictive evaluations for point iii (which includes both content and interface adaptations)
a similar approach has been proposed by [Paramythis et al 01]. And we used the MAE and RMSE metrics to evaluate our tests.

Content evaluation: evaluation of items selection and ranking

MastroCARONTE selects and ranks contents (recommendations about restaurants and hotels, for the evaluation test) on the basis of the user interests and preferences and the distance of the place from the user, but it weights them taking into account also the type of area (metropolitan, suburban, etc.), the hour in the day, the hours driving, etc. In this way, if for example it is late in the night, the system ranks the retrieved hotels weighting distance higher than other features (less important in that moment). In the current version of the prototype, these weights are applied with fixed rules and do not change from a user to another, thus we could evaluate these two aspects of content selection in a separate way. The evaluation of the first one was much more complex as, in order to be significant, we needed to collect the preferences of many users and compare them with those predicted by the system. For the second one, we defined some combinations of contextual conditions (distance, kind of area, time) and evaluated, for a few subjects, if the recommendations satisfied the desires of the subjects regarding the priorities of these conditions. Given that, in the following we will focus on the first aspect.

To obtain reliable user data we chose to collect self-reported information from target users. Thus we decided to exploit a questionnaire we personally distributed to 107 users identified following a proportional layered sampling strategy, where the population is divided into layers, related to the variables that have to be estimated, and containing each one a number of individuals proportional to its distribution in the target population. We identified eight groups characterized by different age, sex, education, job, technology expertise, geographic area, etc. (that are the same descriptive data used by the system to classify each user – see the User Model sub-section of section 3). Every group identifies a potential user of the system. For instance group s1 (the 5% of our sample) is characterized by age: 26-35, sex: male, education: high school, job: autonomous workers, technology expertise: medium, etc.; while group s8 (the 23% of our sample) is characterized by age: 36-45, sex: male and female, education: high school/degree, job: manager/professionals, etc..

We asked each one of the users involved in the evaluation to fill in a questionnaire, thus collecting two sets of data:

3 We plan to improve the next version of the system with personalized weights to these features, according to the relevance for each user.
(a) information useful to the system to classify users and to generate the user model and personalized recommendations and interaction,

(b) information about the real users’ preferences about hotels and restaurants useful to calculate the distance between system predictions (recommendations) and actual users preferences. In particular, in the evaluation we compute the distance between the features of the recommended hotels (restaurants) with the actual user preference in terms of these features.

Six main topic areas were identified in the questionnaire: personal data, information about visual problems, familiarity towards computer and interfaces, food and restaurant preferences, restaurant price preferences, hotel price preferences. The final questionnaire was made up of 14 questions. The questionnaires were auto-filled by the users to avoid any possible interviewer’s interferences and gained a week after the distribution. The questionnaire was anonymous and introduced by a written presentation explaining the general research aims (collecting real and anonymous users data for a computer science research, etc). For the items concerning personal data, visual diseases, computer and interfaces, the participants were required to thick the appropriate answer from a set of alternatives. In the other questions, users had to express their level of agreement with the options concerning the given questions by choosing an item of a 5-point Likert scale.

The survey was conducted in September-October 2002 and the participants were Italian citizen living in - or in the suburbs of - the city of Genoa and Turin, in the North of Italy.

We inserted the data set labeled as (a) in a PC simulator version of the system to generate the system responses. Then we analyzed the correctness of a prediction on user’s preferences using two metrics, MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) that evaluate the distance between the system predictions and the collected users’ preferences -data set (b)-. Both user’s actual preferences and system preferences prediction were expressed by means of rate vectors, on a scale ranging from 0 to 5. Obviously, higher values mean worst predictions and thus recommendations.

Until this point, the test evaluates the recommendations provided to a user at her first interaction with the system. Indeed, data set (a) contains the same information that are requested to the driver when she buys a car. As a consequence we only consider the model generated by the stereotypes and the predictions based on this model. However, as explained in the presentation of the system, MastroCARONTE is able to learn from the observation of the user’s behavior and to update/revise her user model accordingly. Thus, in order to evaluate the dynamic adaptation of the models and consequently of the recommendations, we let the subjects of the test interact with the system for
several times (twelve). In each session the subject had to make a request for restaurant or hotel, look at the answer and then bookmark the items she was interested in (bookmarks are used by the system as a feedback from the user and indicate the actual user preferences). The system logs the session and after several ones updates the user model according to the user behavior. For the evaluation we applied the same methodology we used in the first part of the test, that is calculating the distance between the user preferences - always the data set (b) - and the system predictions generated after the revision of the user model.

Notice that, since the object of the evaluation regarded content adaptation, in all the interactions we simulated a standard and non-risky context. Two final remarks concern the number of interactions and the kind of feedback monitored. We decided for twelve sessions as a tradeoff between the requirement of fast adaptation and the possibility for the system to collect a minimum significant number of evidences. As regards feedback, in the test we considered only bookmarking because other ones such as the detection of direction and parking of the car, etc. required an on board test. Anyway, using one only type of feedback with the same certainty factor, we are confident to obtain normalized evidences, fitting the goals of our evaluation (using one only task and one only kind of feedback avoids also to take into account the experience of the user in managing the interface, which may influence her behavior and the consequent feedback).

**Interface evaluation**

As previously mentioned, in an automotive environment, the evaluation of the interface is extremely relevant as the way information is presented is directly connected to the safety problem. In particular two aspects have to be evaluated:

- the adaptation of the interface to the contextual conditions (e.g. traffic, weather, etc.) and contemporary to the user’s receptivity and cognitive load,
- the usability of the interface, which allows the user to access information in a direct way, avoiding distractions from the driving task.

As regards the first point, in order to evaluate the correspondence between the proposed layout and (i) contextual conditions and (ii) user’s cognitive capabilities, we used an approach which is similar to the one discuss above (it uses the same metrics) but differs in the source of the “gold standard” for the comparison. On the simulator we set two contexts (different for speed and traffic level) and asked two HCI-usability expert to choose the best layout option for each user in each context. We then compared the layout selected by the system with the one chosen by the experts (also in this case the distance is on a scale from 0 to 5).
As regards usability, we performed an empirical study with eight users, each one belonging to a different group, identified on the basis of her level of receptivity and familiarity with technology. We defined four common tasks each user had to do; in their whole, these tasks are planned to obliged the users to use all the main functionalities of the interface:

1. Looking for a hotel and making a call for reservation (on the PC set it just meant to press the button “call”).
2. Looking for a restaurant and asking for the map to reach its location.
3. Looking for a restaurant, adding it to the memo and asking for the itinerary to reach it from the current location.
4. Switch on the radio (which meant escaping from the tourist application and accessing the radio control from the same interface).

On the PC simulator, we initialized the user model of each subject with the information necessary for interface adaptation, and uploaded a standard, non-risky context for all of them. We recorded the performances and at the end of the test we briefly interviewed the subjects.

Since the current prototype is running on a PC set, we could only perform a usability test with users interacting with the computer interface. Anyway, as the display has the same layout as the one on board, we found that performing such test could be useful to check the general organization of the layout, the meaning of labels and buttons and the easiness of navigation. Obviously, the underlying idea is that non-clear layouts distract the driver.

5. Results of the evaluation

In an automotive environment, for all the reasons discussed in the introduction, the goal is to achieve good results of adaptation since the first interactions. Anyway, with respect to the distinction between content adaptation and interface adaptation, and their sub-divisions, we can specify some differences for the minimum requirements:

<table>
<thead>
<tr>
<th>MINIMUM REQUESTED RESULTS</th>
<th>Descriptive values</th>
<th>Metrics values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Adaptation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptation toward user interests and preferences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• at the first interaction</td>
<td>Medium-High</td>
<td>MAE &lt;=1</td>
</tr>
<tr>
<td>• after a number of interactions</td>
<td>High</td>
<td>MAE &lt;=0.7</td>
</tr>
<tr>
<td>Adaptation weighed with external features (distance, time, etc.)</td>
<td>Medium</td>
<td>Not applicable: qualitative method</td>
</tr>
<tr>
<td>Interface adaptation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptation toward context conditions</td>
<td>Very High</td>
<td>MAE &lt;=0.4</td>
</tr>
<tr>
<td>Usability</td>
<td>High</td>
<td>Not applicable: qualitative method</td>
</tr>
</tbody>
</table>

Table 1. Requirements for an effective and safe system in the automotive environment
The values in the last column have been derived from some authoritative references (see [Sarwar, 98], [Good et al, 99]) where, for recommender systems evaluation, the authors suggest that good value of MAE an RMSE should be near to 0.7, in a range of 0-5. We revised these values according to the phase of interaction and to the relevance of the task for safety. Therefore, during the first interaction, where few direct information from the user is available and the stereotypical knowledge is largely taken into account, a MAE value of 1 can be considered the higher acceptable value, while this value must decrease after several interactions with the system. On the other hand, providing messages to the driver when the contextual conditions are risky requires very strict levels of adaptation to the context and to her receptivity, such as those achieved with MAE values less than 0.4.

Content evaluation
Let us start with the results concerning content adaptation. First of all, as regards the ability of MastroCARONTE to adapt to external conditions, we observed that in all cases the system was able to select facilities close to the user’s location, giving priority to the closest ones at specific times (e.g., closest hotels at night), according with the subjects expectations. We can say that the first and simple requirement for MastroCARONTE of giving a correct weight to external conditions, like distance, in the selection of the items, is satisfied.

As regards the adaptation toward user interests and preferences, the analysis is much more complex and the obtained results on the distance between the user’s preferences and the system predictions are summarized in the following tables, divided in results at the first interaction and after a number of interactions:

<table>
<thead>
<tr>
<th>a) Results at the first interaction (initialization)</th>
<th>b) Results after twelve interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurants</td>
<td>Hotels</td>
</tr>
<tr>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>1.44</td>
<td>1.75</td>
</tr>
</tbody>
</table>

Table 2. Experimental results

Comparing the results we obtained with the requirements in Table 1, we can observe that the results for the first interaction are not satisfactory, while the results improve considerably when the number of interactions increases.

Notice that, in this test, the consideration of different groups has only the objective of presenting to subjects the real interface that the system would present them in a real interaction, given their profile.
Even if in many systems a non-perfect adaptation could be tolerated at the first interaction, while driving, it is not acceptable that the user browses alternatives looking for the preferred items and waits for the system learning her actual preferences. The next part of this section aims at showing the analysis we carried on in order to find out the causes of a distance of 1.44 and 1.87 versus a requirement of 1. We will show that the major responsibility for such result is the wrong inference on the users propensity to spend. Indeed, as we will see, if we ignored prices, we would obtain, for restaurants, a MAE of 1.05, which is quite good for the initialization phase.

Several are the ways that can be followed to investigate the cause, or causes, of the distance to finally find out where the system has to be modified. In particular we have identified two types of approaches, which could be defined, for simplicity, as backward and forward or, also, as bottom-up and top-down approach. The first one begins the investigation from the data, looking for cases, or groups of cases, with anomalous statistics. The idea is that of trying to limit the space of alternatives using knowledge about the behavior of the system to go back to the causes of the problem (in order to discover, for example, if it is a classification or personalization problem or some other localized problem). On the other hand, the second type of approach starts without information about the behavior and the performance of the system and tests each component which can produce errors of adaptation (knowledge base, production rules, etc.).

**Bottom-up approach**

In the evaluation of MastroCARONTE we used both approaches, but, in order to reduce the space of possible causes of error and thus to speed up the discovery of the problems, we started from the first approach. We disaggregated the results by considering different groupings of the users. In particular, we considered three ways to group the subjects involved in the test:

- **Sampling Groups** = classes of subjects with common socio-demographic features (we singled out eight groups labeled as s1, s2, s3, ……, s8).

- **Profile Groups** = classes of subjects with the same profile; a profile is the result of the user modeling process which allows to extend the knowledge on users, making inferences about some unknown features. It includes the features derived from the stereotypes (e.g. propensity to spend) and from a set of rules (e.g. receptivity), plus some of the socio-demographic features; in other words it contains the dimensions that are taken into account by the personalization rules (we singled out eighteen groups labeled as pf1, pf2, pf3, …… pf18)
- **Prediction Groups** = classes of subjects for which the system produced the same recommendations (we singled out five groups labeled as pr1, pr2, pr3, pr4, pr5). These groups are clearly related to the previous ones, in the sense that each prediction group includes a set of profile groups, in our case, for example we have that:

\[ pr2 = pf3 \cup pf6 \cup pf10 \]

The basic idea of this analysis was that by **comparing the behavior of the system for these different groups, it might be possible to understand for which reasons the system does not provide a good advice on some users.**

**Analysis of Prediction Groups.** As a first result of this deeper evaluation, we noticed that the recommendations changed significantly according to the different prediction groups: some groups received better recommendations than others (see Table 3). As noticed above, the five *prediction groups* cluster participants with common recommendations in spite of possible different socio-demographic features (they cross the initial sampling groups). For instance people belonging to group pr2 are 26-35 years old and have medium propensity to spend, while people belonging to group pr4 are 46-65 years old and have a high propensity to spend. Within group pr2 there are males and females, with different education levels and different jobs (employees, autonomous workers, managers) and within group pr4 there are males and females, with different education levels and different jobs (employees, managers, teachers). The results in the table suggest that either the accuracy of classification or the prediction for pr5 is much more problematic than that for the other groups; while on the contrary, for pr2 it is much better than for the other groups.

<table>
<thead>
<tr>
<th>Prediction groups</th>
<th>Restaurants</th>
<th>Hotels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>Group pr1</td>
<td>1.46</td>
<td>1.78</td>
</tr>
<tr>
<td>Group pr2</td>
<td>1.08</td>
<td>1.40</td>
</tr>
<tr>
<td>Group pr3</td>
<td>1.29</td>
<td>1.53</td>
</tr>
<tr>
<td>Group pr4</td>
<td>1.54</td>
<td>1.85</td>
</tr>
<tr>
<td>Group pr5</td>
<td>1.72</td>
<td>2.05</td>
</tr>
</tbody>
</table>

Table 3. Prediction groups’ results

Given these results, the next step was to investigate if these differences were uniform among all the *Profile groups* that belong to the same *Predictions group* or not.
**Analysis of Prediction Groups, divided for Profile Groups.** The idea was that, in case we observed some Profile groups with very high MAE with respect to the belonging Prediction group, we could hypothesize that the negative result of the Prediction group was due to one only Profile group, and thus limiting the probable causes of error. The results of this evaluation are reported in Table 4. As it can see, we have computed the MAE and RMSE distances between each Profile Group and the belonging Prediction Group, and then we calculated the Standard Deviation (SD) inside each Profile Group. We decided to exclude, from such a computation, Profile Groups with less then five subjects. The label NC is applied to groups which do not satisfy this condition.

| Predictions Groups | Profile Groups | Restaurants | | Hotels |
|-------------------|---------------|-------------| |         |
|                   |               | MAE | SD | RMSE | MAE | SD | RMSE |
| Group pr1         | Group pf1     | 1,662 | 0,408 | 1,904 | 2,133 | 0,149 | 2,315 |
| Group pr1         | Group pf2     | 1,434 | 0,292 | 1,749 | 1,990 | 0,235 | 2,238 |
| Group pr1         | Group pf4     | 1,397 | 0,212 | 1,767 | 2,067 | 0,893 | 2,408 |
| Group pr1         | Group pf5     | NC  | NC  | NC  | NC  | NC  | NC  |
| Group pr2         | Group pf3     | 0,980 | 0,164 | 1,370 | 0,900 | 0,150 | 1,139 |
| Group pr2         | Group pf6     | 0,829 | 0,136 | 1,260 | 1,029 | 0,167 | 1,302 |
| Group pr2         | Group pf10    | NC  | NC  | NC  | NC  | NC  | NC  |
| Group pr3         | Group pf7     | 0,86  | 0,212 | 1,17  | 2,05  | 0,229 | 2,19 |
| Group pr3         | Group pf8     | 0,82  | 0,10  | 1,13  | 1,45  | 0,21  | 1,64 |
| Group pr3         | Group pf9     | NC  | NC  | NC  | NC  | NC  | NC  |
| Group pr3         | Group pf11    | NC  | NC  | NC  | NC  | NC  | NC  |
| Group pr4         | Group pf13    | 1,14  | 0,27  | 1,63  | 1,94  | 0,34  | 2,13 |
| Group pr4         | Group pf14    | NC  | NC  | NC  | NC  | NC  | NC  |
| Group pr5         | Group pf17    | 1,27  | 0,21  | 1,74  | 1,97  | 0,19  | 2,22 |
| Group pr5         | Group pf18    | NC  | NC  | NC  | NC  | NC  | NC  |

**Table 4. Prediction and Profile groups’ results**

Even though this kind of statistics could not be exploited completely in this first evaluation (in the next one we plan to extend the set of test users), we will briefly describe how to use the available results. We already said that the aim
is discovering if, inside a *Prediction group* with high MAE, there is a *Profile group* that significantly deviates from the average MAE. See for example the profile pf1 of the group pr1. Its MAE for restaurant is 1,662 and the MAE of the belonging *Prediction Group* is 1,46.

- If we consider the Standard Deviation of pf1, we see that it is 0,408 while the Standard Deviation of all the subjects is 0,3. It means that the group is not very homogeneous or that it is homogeneous with a medium-low deviation. The value of the SD is important because it allows to understand if the deviation of the group is due to a single isolated case or not.

- Once we have seen that it is not an isolated case that determines the deviation, the problem becomes understanding the cause of the deviation of the Profile group, i.e. if it is for a problem of classification or of personalization. This second evaluation can be done easily, because we know exactly the values of the dimension used by the system to classify (the values that identify the Profile). For example, if for the group pf1 we know that personalization rules are right defined, then we can be reasonably sure that it is a classification problem.

- Finally it is even possible to identify the exact component in charge of the bad classification, calculating the MAE for the different features of the advice. For example, for restaurants, the features prices, kind of food, type of places, etc. are correlated with specific user model dimensions, which can be computed using the stereotypes (e.g., propensity to spend and thus price) or the production rules (type of food and of places). As an example, for pf1 we discovered a problem in the propensity to spend, inferred by stereotypes (we shall come back on this point later).

**Analysis of Sampling and Prediction Groups.** For completeness of analysis, we can now estimate the dependencies between the groups of the test and the recommendations provided by the system. Let us start by considering the MAE and RMSE of the *Sampling groups*, which are the initial 8 groups collected by the sampling strategy\(^5\). (see Table 5).

<table>
<thead>
<tr>
<th>Sampling groups</th>
<th>Restaurants</th>
<th>Hotels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>Group s1</td>
<td>1.39</td>
<td>1.67</td>
</tr>
</tbody>
</table>

\(^5\) Five subjects were not classifiable in any group.
<table>
<thead>
<tr>
<th>Group</th>
<th>s2</th>
<th>s3</th>
<th>s4</th>
<th>s5</th>
<th>s6</th>
<th>s7</th>
<th>s8</th>
<th>not-classified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1,41</td>
<td>1,36</td>
<td>1,59</td>
<td>1,18</td>
<td>1,49</td>
<td>1,22</td>
<td>1,42</td>
<td>1,50</td>
</tr>
<tr>
<td></td>
<td>1,69</td>
<td>1,71</td>
<td>1,92</td>
<td>1,55</td>
<td>1,78</td>
<td>1,51</td>
<td>1,73</td>
<td>1,81</td>
</tr>
<tr>
<td></td>
<td>1,81</td>
<td>1,00</td>
<td>2,06</td>
<td>1,31</td>
<td>2,05</td>
<td>1,85</td>
<td>1,70</td>
<td>1,93</td>
</tr>
<tr>
<td></td>
<td>2,08</td>
<td>1,18</td>
<td>2,28</td>
<td>1,58</td>
<td>2,23</td>
<td>2,06</td>
<td>1,88</td>
<td>2,11</td>
</tr>
</tbody>
</table>

Table 5. Sampling groups’ results

Also in this case there are differences between groups: for instance group s6 receives the best restaurant recommendations while group s3 receives the worst recommendations. We compared these results with those obtained for the prediction groups. An ANOVA comparing the groups results showed that the different results are due to a significant correlation between the kind of group taken into account (independent variable) and its related recommendations (dependent variable).

Predictions groups (103 subjects for 5 groups): restaurants MAE: F(4, 98)= 9,27, p<0,01; restaurants MAE no prices: F(4, 98)= 5,33, p<0,01; hotels MAE, F(4, 98)= 26,83, p<0,01.

Sampling groups (107 subjects for 9 groups): restaurants MAE: F(8, 98)= 2,74, p<0,01; restaurants MAE no prices: F(8, 98)= 1,71 p<0,01; hotels MAE, F(8, 98)=9,40, p<0,01.

In conclusion, using the bottom-up approach and decomposing the analysis on the basis of the groups described above, we could examine the relationships between the output of the main tasks of the system - classification, inference of user’s features and recommendations - and obtain relevant information for the revision of the system. The main result was the identification of problems regarding some profile groups (pf 1 in particular) and the proof that the main problem regarded the dimension of propensity to spend, which is used by the system to predict the range of price of restaurants and hotels to recommend. As the propensity to spend is a feature inferred by the system, we could get clues on the fact that the origin of problems was in the stereotypes used for inferring this feature. Thus, after the analysis we knew that the Knowledge Base should have been revised, but anyway we had not enough information for understanding the exact changes that should have been brought to the system. Moreover we were not able to know the percentage of deviation produced by such errors.
Top-down approach

Given the problems with not comparable Profile groups, and anyway in order to extend our analysis, we also followed this second approach and obtained interesting results. As noticed above, this approach is not guided by any kind of knowledge about possible errors, and thus analyzes each component of the system. We just sketch the main points.

- Test of the correctness of the KB. Our starting hypothesis was that each stereotype shares homogeneous preferences, behaviors and lifestyles. The inexistence of this correlation could be due to these factors:
  i) the stereotypes are too generic and therefore they cannot be used for a specific domain;
  ii) the studies on which the stereotypes are based are too old and do not reflect the current situation.

We could accomplish this test using the questionnaire and comparing the answers with the Eurisko classes (which provided us the knowledge for building the stereotypes of users, see section 3). The answers in the questionnaire showed that users almost always selected ranges of prices lower than those predicted by the lifestyle stereotypes. This finding was a confirmation of the results obtained from the previous analysis: since prices are directly related to the propensity to spend, the answers on the questionnaires demonstrated a bad inference in the stereotypes. On the contrary, the other dimensions (technology expertise, receptivity) derived from the stereotypical knowledge seemed to be well suited for the purposes of our application. Therefore the need of updating our Knowledge Base, concerning the supposed propensity to spend emerged clearly.

- Test of possible errors in the KB implementation. Given the proof of errors in the KB, the goal of this phase of analysis was to understand the reasons of such errors. We investigated several possible causes. First of all, changes in the propensity to spend may be caused by variations in the economic cycle. Another cause may be that the research we exploited describes the lifestyles in a qualitative way and thus we had to translate the qualitative values into probabilistic ones. Errors are frequent in processes like this, due to a misunderstanding in tuning the estimates. For instance, we found that working young people are described as having a high propensity to spend. However, the test demonstrated that their propensity to spend is not linear over the two main dimensions of analysis for this feature, so they generally prefer non-expensive restaurants even if they like to go out for dinner frequently. As a consequence, we concluded that it would be useful to split the propensity to spend in frequency and value. A confirmation of this problem comes from the MAE calculated with respect to university students, which are mainly represented in the group pf1. However Eurisko does not take into account
this distinction and we have no means to know its interpretation of the dimension. A possibility could be to define a new Knowledge Base, exploiting a domain specific survey to the target population.

- **Test of the Rules in charge of the personalization.** In the system, a set of rules associates user features (age, propensity to spend) to hotels/restaurants features (price, kind of restaurant, kind of food, etc.). The test showed the existence of some inhomogeneous features. For instance, restaurant suggestions for 25-36 years old resulted better than restaurants suggestions for 20-25 years old. Then, within the first group the suggestions are better suited for people characterized by a medium propensity to spend. Therefore, the aim of this analysis is to discover the associations that have to be revised, as in the case above described.

In conclusion, we present again the initial table, but inserting also the results evaluated after ignoring the calculation of distances regarding prices.

<table>
<thead>
<tr>
<th>Restaurants</th>
<th></th>
<th>Hotels</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>RMSE</td>
<td>MAE*</td>
</tr>
<tr>
<td>1,44</td>
<td>1,75</td>
<td>1,05</td>
</tr>
</tbody>
</table>

Table 6. Experimental results. (*) means without the prices predictions

As it can be seen, the results have significantly improved, and, for the Restaurants they are globally positive, compared to the expected minimum requirements specified at the beginning of this section. For the Hotels the result, even if improved, continues to be not satisfactory.

**Interface evaluation**

Finally, let us move to the evaluation of the personalization choices as regards the format and layout of the presentation. Following the methodology of evaluation discussed in the previous section, we will divide the results in two parts: results regarding the adaptation of the interface toward context conditions and usability of the interface.

As already described, for the first kind of evaluation we interviewed two HCI experts, asking them to suggest, for each one of the profiles, the ideal interface layout in two different driving contexts. We then compared these “gold standards” with the personalized layout selected by the system. More specifically, we decided to carry out an experts’ evaluation since this technique (also known as heuristic evaluation) has been demonstrated to be successful in the early interface design phases [Nielsen and Molich 90]. We used the following procedure (slightly different from the regular experts’ evaluation but tailored for an adaptive interfaces evaluations): the two experts
proposed the ideal interface for groups of subjects sharing the same features used by the system to generate layout recommendations (age, technology expertise, receptivity) in a given context. Then we calculated the distance (MAE and RMSE) between the systems proposals and experts’ suggestions. The results are shown in Table 7.

<table>
<thead>
<tr>
<th>Layout Context 1</th>
<th>Layout Context 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>0,18</td>
<td>0,18</td>
</tr>
<tr>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>0,09</td>
<td>0,09</td>
</tr>
</tbody>
</table>

**Table 7.** Layout prediction’ results

The closeness between the layouts proposed by the system and the HCI experts’ suggestions confirmed the appropriateness of layout adaptations choices. The next step of the evaluation will be on board, with subjects interacting with a version of MastroCARONTE running on a car.

As regards usability, the final results obtained from the tests we described in section 4, showed the all subjects, independently on their profile, had similar problems:

1. No subject used the “ESC” button (see Fig. 2-4), which turns back to the start page. Most subjects thought that this button shut down the application. They expected a button called “Home”.

2. The “Back” button (the second button on the right, see Fig. 2-4) was judged not so clear. Probably, a left-oriented arrow, as that one used in browsers, would be clearer. Moreover, it would be useful to add another up-oriented arrow to reach the previous menu (higher level).

3. The colors of the five dots (red and blue, see Fig.2) used to show the strength of the recommendation generated some confusion. Probably, the blue should be changed in a faded degree of red to mark the non-selected dots.

4. The label “Road” (“Strada”, see Fig. 3 and 4), used to show the map, generated some confusion. Probably, the label “Map” should be clearer.

5. Some subjects had problems with the meaning of the “Memo” button, used to add items to the “Memo” list, which is a kind of bookmark list. Some of them suggested to change this label in “Save” or “Bookmark”.

As it can be noticed from the list above, the main functionalities, which allow a user to access the recommendations, are sufficiently clear, in spite of some problems in the interpretation of a few buttons. But the navigation functionalities seem to create confusion to most users. Therefore, the obtained results provide clear indications about the corrections that have to be made to the layout, but especially provide a complete analysis of the interface and of
the safety problem connected to interface: without a usability test, the good results of the previous one (regarding the adaptation of the interface toward context conditions) would have carried us to state that the adapted interface and layout guarantee good level of safety. In contrast, we can now say that the fundamental achievements of adapting the interface to the context conditions may be compromised if understanding the mechanism of navigation requires a cognitive effort from the user with the risk of distracting her.

In conclusions, comparing the results of the tests to those provided in the table of the requirements, we can state that the results are encouraging, showing that indeed adaptation can provide interesting results on board-vehicles. Some tuning of the knowledge bases of the system is needed, especially to improve first consultation results and will be performed when developing the next prototype for which we plan to perform an on-vehicle evaluation. Some other analyses will have to discover the problems that still remain for Hotels.

6. Conclusions
At the beginning of the paper, we stated that the goal of this work was to present a first evaluation of MastroCARONTE, showing the advantages and the opportunities of adaptation and personalization techniques on-board vehicles. The results of the evaluation exercise demonstrate that these claims can be accomplished in practice. The best way to draw a conclusion is to consider some of the guidelines for the design of on-board HMI, noticing that most of them are accomplished by adaptation and personalization:

- Paying attention to the risk of interference with the main task of driving by causing a dual-tasks [Green, 2000].
- Choosing the modality of interaction. It may vary from hand-held remote control to voice control, from controls and buttons near the display or touch-screen interfaces to advanced HMI, exploiting the potential of multimodal interaction such as haptic (tactile and kinaesthetic) and acoustic interaction [Mariani, 02]. Multimodal interaction in particular has been found as positively affecting driver’s safety, resulting in faster reaction times and fewer errors for emergency response displays when compared with simple visual or auditory display [Liu et al, 99]. Indeed, the exploitation of non-visual interaction can reduce the driver’s overload and can be effective for attracting her attention (e.g. auditory interaction for alerts, warning). However, these kinds of interaction can be chosen only in particular situations (e.g., auditory modality is effective if output is simple) and for located actions [Summerskill et al 02].
- Readability of the content to be displayed must be taken into account. Besides the usability of controls and display, the items presented on the screen have to be easily readable and immediate. For instance, a study
[Burnett and Porter, 02] showed that the exploitation of landmarks vs. distance cues, in vehicle navigation systems, decreases the number of glances and workload was perceived to be lower.

- Tuning the interaction. At the beginning the interaction should be easier since the unexpected system responses are handled in a more problematic way and the user workload could increase with undesired effects. Then, when the user becomes more experienced, the selection of next actions to perform becomes easier for her, since she developed an interaction strategy [Jahn, et al. 02]. In MastroCARONTE the layout is selected on the basis of the user’s receptivity and familiarity with technological devices. But, as a set of rules periodically revises and updates the User Model, after a number of interactions (which changes according to the level of receptivity itself) the layout selected becomes richer in amount of information and services.

- The interface must be adequate to each specific user. For instance, the maximum quantity of information to be presented to old and young drivers should be different. Empirical studies [Labiale and Galliano, 02] showed that, by increasing the number of pictograms displayed on an in-car system, the number and the duration of visual fixation increase for both old and young people. However, older drivers require longer visual fixation than young drivers according to the explanatory hypothesis of a perceptual and cognitive slowing down of elderly subjects. Labiale and Galliano proposed no more than 9 pictograms for young drivers and no more than 6 for older drivers. Yet, the exploitation of Intelligent User Interface taking into account individual factors can modify the complexity of each in-car display. For instance, people having more familiarity towards technology and computers seem to be faster at learning new interfaces. Visual diseases can affect the amount of information to be read particularly in a visual-centered task such as driving. The speed of the car, the type of the roads, the traffic and the driving experience may change the way in which drivers manage the switching between primary and secondary task.

The last dimension (driving experience) has not been considered in MastroCARONTE (it could be added in the next release) but all the other ones are exactly the features taken in consideration by the system for loading the right stylesheet of presentation.

Acknowledgements
We are very grateful to Luca Console for his recommendations and his supervision on the work.

References


http://www.sae.org


