Interacting with Computers 24 (2012) 482–498

Contents lists available at SciVerse ScienceDirect

Interacting with Computers

journal homepage: www.elsevier.com/locate/intcom

Towards building intelligent speech interfaces through the use of more flexible, robust and natural dialogue management solutions ☆

Fernando Fernández-Martínez *, J. Ferreiros, J.M. Lucas-Cuesta, J.M. Montero-Martínez, R. San-Segundo, R. Córdoba

Grupo de Tecnología del Habla, GTH (Speech Technology Group), Universidad Politécnica de Madrid, UPM, Ciudad Universitaria s/n, 28040 Madrid, Spain

ARTICLE INFO

Article history:
Received 22 October 2010
Received in revised form 26 June 2012
Accepted 20 September 2012
Available online 22 October 2012

Keywords:
Spoken dialogue systems
Mixed initiative
Bayesian Networks
Contextual information
Usability
Electronic devices control

ABSTRACT

In this paper a Bayesian Networks-based solution for dialogue modelling is presented. This solution is combined with carefully designed contextual information handling strategies. With the purpose of validating these solutions, and introducing a spoken dialogue system for controlling a Hi-Fi audio system as the selected prototype, a real-user evaluation has been conducted. Two different versions of the prototype are compared. Each version corresponds to a different implementation of the algorithm for the management of the actuation order, the algorithm for deciding the proper order to carry out the actions required by the user. The evaluation is carried out in terms of a battery of both subjective and objective metrics collected from speakers interacting with the Hi-Fi audio box through predefined scenarios. Defined metrics have been specifically adapted to measure: first, the usefulness and the actual relevance of the proposed solutions, and, secondly, their joint performance through their intelligent combination mainly measured as the level achieved with regard to the user satisfaction. A thorough and comprehensive study of the main differences between both approaches is presented. Two-way analysis of variance (ANOVA) tests are also included to measure the effects of both: the system used and the type of scenario factors, simultaneously. Finally, the effect of bringing this flexibility, robustness and naturalness into our home dialogue system is also analyzed through the results obtained. These results show that the intelligence of our speech interface has been well perceived, highlighting its excellent ease of use and its good acceptance by users, therefore validating the approached dialogue management solutions and demonstrating that a more natural, flexible and robust dialogue is possible thanks to them.

1. Introduction

Speech is the most widely used natural means of communication between people. Speech also is of increasing importance as a user–machine interface. As a result of the knowledge and the experience accumulated during almost half a century of research in the field of speech technology, the time has now come to design automated dialogue systems that make use of the communicative aspects of speech. In particular, it is essential to incorporate into the design of these systems some ideas related to the concept of *ambient intelligence* (AmI) (Augusto, 2007; Aarts and de, 2009), for providing intelligent interfaces that are able to conduct a natural dialogue, including negotiations in order to achieve the goals required by users.

A dialogue system can be seen as a computer application that enables interaction and communication between users and machines as naturally as possible. Besides the typical recognition and text-to-speech conversion modules and other components, dialogue systems usually contain a module called dialogue manager (DM). This module is responsible for a dual task: to interpret the intention of the user and to decide how to continue the dialogue.

To provide users successfully with answers resembling a human–human interaction as much as possible, we believe that the design of a dialogue system should be approached from both a theoretical and practical point of view. Thus, we must pay attention not only to dialogue management and modelling, but also to the enhancement of these models with knowledge about the specific tasks of the dialogue and the application domain (i.e. task and domain models). Thus, it is feasible to develop procedures that support the user–machine interaction with useful elements of

☆ This paper has been recommended for acceptance by D. Murray.

* Corresponding author. Tel.: +34 91 549 57 00x4228; fax: +34 91 336 73 23.
E-mail addresses: ffm@die.upm.es (F. Fernández-Martínez), jfl@die.upm.es (J. Ferreiros), juanmak@die.upm.es (J.M. Lucas-Cuesta), juancho@die.upm.es (J.M. Montero-Martínez), lapiz@die.upm.es (R. San-Segundo), cordoba@die.upm.es (R. Córdoba).

0953-5438/$ - see front matter © 2012 British Informatics Society Limited. All rights reserved.
http://dx.doi.org/10.1016/j.intcom.2012.09.003

communication for carrying out a collaborative and cooperative dialogue.

Although the interest in ambient intelligence in the domain of home dialogue systems is growing significantly (Berton et al., 2006), the benefits that this intelligence might bring are not often demonstrated or clearly identified (de Ruyter et al., 2005).

In this work we are presenting the evaluations that we have conducted to examine the effects of our dialogue management solutions, but more specifically to address the following research questions:

- Will the level of flexibility (i.e. absence of rules or restrictions that might restrict the dialogue in any way), robustness (i.e. ability to recover missing information and to handle errors when the user input has ASR and SLU errors occurred by noises or unexpected inputs) and naturalness (i.e. ability to negotiate with the user in achieving the dialogue goals similarly to the way a human would help) achieved in the home dialogue system be perceived (e.g. by means of a good user satisfaction rate)?
- What is the effect of bringing this intelligence into a home dialogue system on the perception of the ease of use of the interactive systems in the environment?
- Will the acceptance of home dialogue systems increase if the proposed solutions are implemented in these systems?

Finding performance figures from real-world applications that can be extrapolated to other systems or be accepted worldwide is a really complicated task, as all of them are directly related to a specific dialogue system. Nonetheless, there is a general agreement on *usability* as the most important performance figure (Schulz and Donker, 2006; Turunen et al., 2006; Raux et al., 2005; Walker et al., 2000), even more than others widely used such as *naturalness* or *flexibility*.

Several usability guidelines that should be taken into account in the design of dialogue systems and their evaluation, especially for multi-modal systems, have been reviewed in Dybkjaer et al. (2004). Therefore, besides quality and efficiency metrics, automatically logged or computed, subjective tests have also been carried out in order to assess the impact of the capabilities of the system on user satisfaction and to get a valuable insight into the shortcomings and advantages of the proposed solutions.

The paper is organized as follows: first, the home dialogue system used to answer the aforementioned research questions is described. A couple of subsections, 2.6 and 2.7, introduce the two different versions of our developed prototype, HIFI-AV1 and HIFI-AV2, discussing alternative approaches for the management of the actuation order. The following section describes the experimental framework used to evaluate the performance of the proposed solutions. In the sections which follow, we successively present and discuss the results obtained for both versions of our system (i.e. HIFI-AV1 vs. HIFI-AV2). Finally, the paper concludes by highlighting some conclusions specifically addressing the aforementioned research questions. Some possible future lines of research are also proposed.

2. The dialogue management solution

The major advantage of classic knowledge-based dialogue management solutions (Bui, 2006; Lee et al., 2010) like: finite state automata or FSMs, script based systems or dialogue plans, etc., is the simplicity. They are suitable for simple dialogue systems with well-structured task. However, these approaches lack of flexibility, naturalness, and applicability to other domains.

As an alternative to these we are presenting a dialogue solution based on Bayesian Networks (BNs), that allows a greater flexibility and naturalness by appropriately defining dialogue as the interaction with an inference system (Meng et al., 2003).

This solution can be classified as a data-driven dialogue management approach (Lee et al., 2010) that, although requires time consuming data annotation, enables training to be done automatically and requiring little human supervision.

The framework applies statistically data-driven and theoretically principled dialogue modelling to dynamically allow changes to the dialogue strategy. Stochastic dialogue modelling using reinforcement learning (RL) based on Markov decision processes (MDPs) (Levin and Pieraccini, 1997) or partially observable MDPs (POMDPs) (Williams and Young, 2007) are another alternative approaches within this framework.

2.1. A spoken dialogue interface for a Hi-Fi audio system

The conversational interface that we are presenting (Fernández-Martínez et al., 2005) was included as part of the EDECAN project.[1] It allows users to control a Hi-Fi system from natural language sentences, differentially to other typical control systems based on simple commands. Thus, users can feel free to give several complex commands from a single sentence. Moreover, they neither have to memorize any command list nor use specific vocabulary cum syntax in order to control the system successfully.

The Hi-Fi audio system we are controlling is a commercial system made up of a compact disc player (with a charger of three discs), two tapes deck and a radio receiver. This system can be controlled by an infra-red (IR) remote control. Instead, users are going to control the Hi-Fi system using a microphone. Our interface translates the speech into IR commands in order to carry out different operations or actions on the system. This translation is made so that the appropriate IR commands are sent according to the intention of the user.

2.2. The spoken dialogue system

A dialogue can be defined as the verbal interaction that the user has with the system with the purpose of achieving some goals related to the control of the Hifi equipment. This interaction takes place on a turn basis (a dialogue turn can be defined as one user input action and the corresponding system output). Its length, typically measured either in terms of time or simply as the number of turns, basically depends on the situation. Particularly, we assume a new dialogue to begin as soon as the user starts addressing the system with whatever intention. Then we assume that dialogue to be finished as soon as the system manages to satisfy every goal that may have been positively identified (and hopefully requested by the user during the dialogue) or just as soon as the user decides to abandon it (e.g. by using a "cancellation" voice command).

Fig. 1 shows a block diagram of our conversational interface. The system consists of an automatic speech recognition module (ASR), which translates the audio signal into a text hypothesis of what the user has said; a language understanding module (NLU), that extracts the semantics of the user's utterance; the dialogue manager (DM), which makes use of the extracted semantic information, together with the information available at the context manager module, to determine the actions on the system that the user wants to fulfil, and to provide the user with feedback regarding the current dialogue turn; the context manager (CM), which holds the information of not only the ongoing dialogue but also of the past ones between the same user and the system;

[1] EDECAN Project Web page: http://www.edecan.es

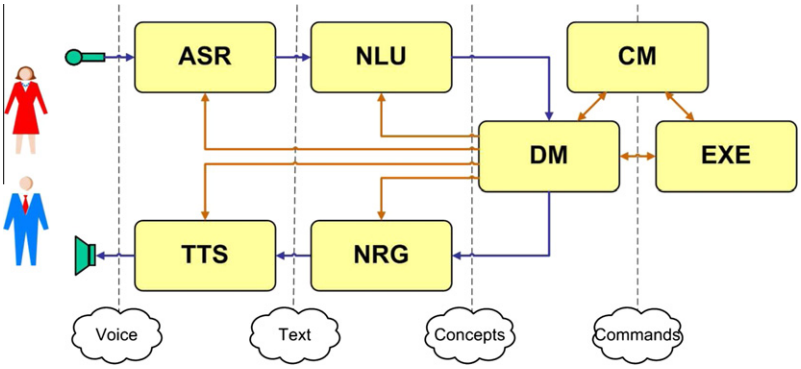

Fig. 1. Block diagram of the spoken dialogue system.

an execution or actuation module (EXE), that translates the actions into IR commands; the natural response generation module (NRG), which makes use of the semantic information provided by the dialogue manager to generate a text output, and a text-to-speech module (TTS), that synthesizes the message to the user.

In the next section we briefly outline the main advantages of our BN-based approach for dialogue modelling. A more detailed description of our system, its architecture and the implemented dialogue strategies can be found in Fernández-Martínez et al. (2005).

2.3. Dialogue management based on Bayesian Networks (BNs)

Users address the system in order to carry out some actions. First, the speech recognizer provides recognition results by using context-dependent HMM models trained with the SpeechDat database (Moreno, 1997). The recognition dictionary consisted of approximately 600 words.

Then each recognized sentence is semantically tagged by the NLU module. This tagging is done according to a concept dictionary and a set of context-dependent rules which have been previously defined by an expert trying to cover all the relevant semantic categories in the domain.

The resulting concepts can be grouped into: *actions* to be performed on the system (e.g. to play), *parameters* that can be configured in the system (e.g. the volume), and their corresponding *values* (e.g. a number). In summary, there are a total of 70 different concepts.

Table 1 shows an example of the semantic parsing of a possible sentence. This sentence has been also tagged with its corresponding dialogue goals (i.e. specific actions on the Hi-fi system) according to the user's intention. A set of 15 goals has been defined according to the available functionality.

The first task of the dialogue manager (DM) module is to identify the intention (i.e. dialogue goals) of the user considering the last utterance together with the dialogue context. Then, according to the inferred goals the DM has to make a decision regarding how the dialogue should continue. Both tasks can be accomplished using BNs.

A BN is a directed acyclic graph, DAG, with nodes and arcs where the direction of the arcs represents the probabilistic dependency between two nodes. The arrows of the acyclic graph are drawn from cause to effect (e.g. C_1 depends on G_1 in Fig. 2). A conditional probability table, CPT, quantifies these dependencies for each network node (e.g. $P(C_1|G_1)$ for the C_1 in Fig. 2).

2.3.1. Advantages of the BN-based approach

Regarding the application of the BNs to dialogue modelling and management we can highlight:

- The BN-based inference system enables a better identification of the dialogue goals according to the intention of the user (i.e. actions or activities that the user may request the system to perform) from the available semantic information (i.e. extracted concepts by semantic parsing) (Su and Zhang, 2006; Jing et al., 2008). This procedure is commonly known as the *Forward Inference process* (Meng et al., 2003).
- The BNs can be automatically obtained from training data. Automatic learning algorithms favour portability and scalability across domains. New systems can be developed at only the cost of collecting new data for moving to a new domain, which requires less time and effort than the knowledge-based approaches (i.e. when the designers develop a new application for a different domain, the entire design process must be restarted from the beginning). In addition, as training data only involves semantic information (i.e. concepts and goals), this also allows the design of the inference system with the highest possible degree of independence of the language used.
- The BNs allow a simple way of incorporating human knowledge into the models, for example by changing the BN topology by hand or by refining the dependencies (i.e. conditional probabilities) between the nodes of the network. This can also be a solution when not enough data is available for training (traditional example-based DMs, like RL or BN based DMs, may require a

Table 1
Example of the inference process of the dialogue goals from the available concepts.

U: "Play the third track from the first cd and raise the volume."	
Concepts	Dialogue Goals
STATE_ACTION=[play]	
TRACK_VALUE=[3]	
TRACK_PARAM=[track]	"device selection"
DISC_VALUE=[1]	"playing parameters definition"
DEVICE_VALUE=[cd]	"source state modification"
DISC_PARAM=[cd]	"volume adjustment"
VOLUME_ACTION=[+]	
VOLUME_PARAM=[volume]	

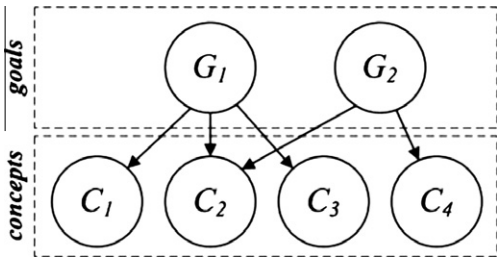

Fig. 2. Example of a BN model for dialogue management.

large number of dialogue corpora to learn an optimal policy, for instance, because of a very large state and/or policy spaces). However, all the BN models that were used during the evaluation, meaning both the directed acyclic graphs (DAGs) modelling the most significant dependencies between concepts and goals and the conditional probability tables (CPTs) that quantify these dependencies, were estimated from training data.

- BNs allow an *analysis of congruence* to be conducted between the goals assumed by the system to have been requested by the user, and all data collected during the interaction. Based on this analysis, the system can determine the flow of interaction and react according to the semantics of the application domain (e.g. performing the required tasks or asking the user for additional information if needed). In particular, it is possible to detect automatically which concepts are needed (available or not), erroneous or optional with regard to the inferred goals (through a so called *Backward Inference process* (Meng et al., 2003)). As it will be detailed in next section, the DM makes the decision on how to continue the dialogue using all the available information (including some domain-specific prior knowledge, thus turning our approach into a hybrid approach that relies on both dialogue examples and a prior knowledge to improve the robustness of the system). Thus, the dialogue could go toward the generation of messages to request the missing items, clarify the erroneous ones and ignore the optional ones respectively. This is useful for avoiding unnecessarily long dialogues and facilitates the achievement of the dialogue goals efficiently (Fernández-Martínez et al., 2008).
- The BNs enable a true mixed initiative dialogue modelling. Flexibility is probably the main asset of the proposed solution, and the most significant difference with regard to conventional approaches. Typical knowledge-based systems are usually confined to both highly structured tasks and system initiative dialogues, where a restricted and regularized language set can be expected. By using a BN-based approach instead, the user is not constrained to any predetermined goal or data sequence. Thus, the BNs provide a mixed initiative dialogue modelling in which the user is free to choose at any time the goals to be accomplished by the system. This flexibility is twofold, since it not only allows the user to decide the goals at the beginning of interaction, but also lets him/her jump to other goals without having completed the previous ones. Moreover, the user can respond with more data than those requested in a query, or even respond to a fact not asked by the system with regard to the inferred dialogue goals. However, the dialogue may turn to be system initiative, for instance, when the system solves the need to request the user a particular information item (i.e. missing concepts). To avoid sudden changes in the interpretation (which could produce disorientation or confusion in the user) the DM must integrate all available information into the decision making process of how to continue the dialogue.
- Thanks to the negotiation process between the users and the system, based on the FI and BI procedures, the system is capable of responding to complex issues (e.g. when the users provide inaccurate or insufficient information to meet the required dialogue goals) and to assist or guide the users toward the achievement of their dialogue goals by driving the dialogue in an efficient manner, minimizing the number of questions or queries and making maximum use of available information in the context of dialogue. Unlike the BN-based approach, knowledge-based approaches generally uses finite-state automata which often involve handcrafted rules. Hand-crafting rules in advance is difficult (it requires application developers who have domain-specific knowledge), and its flow is inflexible. For example, if the users provide more information that was requested by the system's question (over-informative), then the system cannot manage the dialogue flow basically because it was not designed in such a case. The agenda-based approach (Bohus and Rudnicky, 2009) is an extended version of typical knowledge-based systems that provides powerful representations for segmenting large tasks into smaller and more easily handled subtasks. However, the design process is still time-consuming and expensive mainly because of the need of human experts to design the knowledge sources (e.g. hierarchical task structure and plan recipes).

2.3.2. Forward Inference

As can be seen in Fig. 2, BNs can be adopted to model the existing causal relationship between the goals and the concepts (Fernández-Martínez et al., 2005; Meng et al., 2003). Typically, both of them are assumed to be binary (Meng et al., 1999) (i.e. a concept is true or *present* only when it is observed in the sentence). Thus, from the whole set of available evidence, e.g. $E = \{C_1 = 0, C_2 = 1, \ldots, C_M = 1\}$ for M defined concepts, a posterior probability $P(G_i = 1|E)$ can be obtained for each goal as in Eq. 1. This equation simply applies Bayes' Theorem assuming marginal and conditional independence, which is equivalent to a naive Bayes formulation (M is the number of input evidences). The computation of the posteriors $P(G_i = 1|E)$ for the N defined goals is known as the *Forward Inference* technique, FI (Meng et al., 2003; Huang and Darwiche, 1996).

$$P(G_i = 1|E) = P(G_i = 1)\prod_{j=1}^{M}\frac{P(C_j = c_j|G_i = 1)}{P(C_j = c_j)} \tag{1}$$

where $E = \{C_1 = c_1, C_2 = c_2, \ldots, C_M = c_M\}$ and $c_j \in \{0, 1\}$

Subsequently, a decision is made for each goal on the comparison of the posterior with a defined threshold, θ. As a result of this comparison, one goal is *active* or *present* if the corresponding posterior is over the threshold; otherwise the goal is *absent*.

For instance, assuming that the G_1 node in the BN depicted in Fig. 2 corresponds to the "volume adjustment" goal, then, the FI process would allow us, among others, to compute the posterior probability $P(G_1 = 1|E)$, where $E = \{C_1 = 1, C_2 = 1, C_3 = 0\}$ according to Table 2.

A possible result for this example could be $P(G_1 = 1|E) = 0.95$ so that, as the posterior is higher than the θ threshold (for simplicity, it may be set to 0.5 since $P(G_i = 1|E) + P(G_i = 0|E) = 1$), we would assume the goal to be present or active (i.e. we assume that the user intends to modify the volume settings).

The user typically refers to several goals simultaneously so multiple goal scenarios are considered. Changes in the activation of a particular goal are also allowed depending on the available evidence (which may vary according to the intention of the user and the evolving dialogue history).

An off-line evaluation (Fernández-Martínez et al., 2009) has been carried out from a set of 463 individual control sentences (without any dialogue). FI results showed a 92.29% F-measure regarding goal identification.

2.3.3. Backward Inference

After the FI process, and assuming the inferred results (i.e. those goals which were decided to be present, $G_i = 1$) as new evidence, Bayesian inference can be applied again but this time aimed at

Table 2
Example of the FI process for the "volume adjustment" goal.

Concepts	BN node	Observed?	Evidence
VOLUME_ACTION	C_1	Yes	$C_1 = 1$
VOLUME_PARAM	C_2	Yes	$C_2 = 1$
VOLUME_VALUE	C_3	No	$C_3 = 0$

Table 3
Concept analysis used to drive the dialogue.

	$P(C_j = 1\|E^*) < \theta$	$P(C_j = 1\|E^*) \geqslant \theta$
C_j absent ($C_j = 0$)	C_j **unnecessary** (No action)	C_j **missing** (Prompt to request C_j)
C_j present ($C_j = 1$)	C_j **wrong** (Prompt to clarify or notify about C_j)	C_j **required** (C_j is stored in the dialogue history)

the estimation of $P(C_j = 1|E^*)$, the probability that each concept should be present where E^* refers to the updated set of evidences (i.e. E also including goal evidence obtained through the FI process but removing the evidence corresponding to the target concept, C_j). This process is known as the *Backward Inference*, BI, technique (Meng et al., 2003).

By making a similar binary decision on the value of $P(C_j = 1|E^*)$, it is possible to check whether that concept should be present (i.e. $P(C_j = 1|E^*) > \theta$) or not.

2.3.4. Concept analysis

The BI result can be compared with the actual occurrence of the concept enabling the classification presented in Table 3.

As a result of this analysis (Meng et al., 2003) every concept can be properly classified allowing the DM to carry out a suitable action (a possible dialogue proceeding strategy has been suggested below each result). For example, the system can control the dialogue prompting about the *missing* concepts as in the following example.

This example is basically the same that we presented earlier for the FI process but this time assuming that no value nor action have been referred by the user (e.g. we assume the user turn to be like: *"Change the volume"*).

Assuming that the "volume adjustment" goal has been positively identified, we now include $G_1 = 1$ as new evidence and perform the BI process which could result in the concept posteriors presented in Table 4.

According to these results, the network clearly points to the value (i.e. VOLUME_VALUE) as a concept which is expected to be present under the available evidence (i.e. the goal and the parameter are present but the action is not). As it is actually absent then it is classified as missing leading the system to a request turn (e.g. *"What would you like to do with the volume, raise it or lower it?"*).

An action (i.e. VOLUME_ACTION) may also have been a good candidate as well, though not for the defined threshold. In this case, its posterior is computed assuming the value to be absent. Therefore, under similar conditions, the action is shown to be less likely than the value (i.e. 0.45 and 0.90 respectively), however this basically depends on the available training data and the learnt CPTs. Since the occurrence and the BI result match (i.e. both absent), the system is not expected to do anything in this regard (i.e. the concept is *unnecessary*).

Finally, as could be expected, the parameter posterior is clearly aligned with the fact that it is so often referred when trying to change the volume. It is indeed present, so it is simply regarded as *required* and directly stored in the history of the ongoing dialogue.

The accuracy of this analysis, as well as a correct identification of the corresponding dialogue goals, is of vital importance to ensure the appropriate behaviour of the SDS. For instance, a possible misclassification may occur when considering a *required* concept as *wrong*. In that case, the system would probably try to correct or clarify a concept that is not actually erroneous but needed to satisfy the inferred goals. From this revealing example it is clear that the resulting misbehaviour from a wrong concept classification may have a negative impact on dialogue regarding consistency, naturalness and success.

The BI process and the derived concept classification showed an 81.00% F-measure performance for the off-line evaluation presented in Fernández-Martínez et al. (2009).

2.4. The use of contextual information

The DM is also provided with a set of contextual information handling strategies. Regarding the benefits of applying those strategies for dialogue management we emphasize:

- Systems usually have to deal with situations in which users omit certain information. Sometimes that information is essential for the proper outcome of the dialogue. The proposed solution allows, through the negotiation process based on the inference procedure, omitted information (i.e. missing concepts) to be obtained.
- This solution has also the ability to recover the remaining information from the dialogue context quickly. Several dialogue strategies that benefit from contextual information have been designed and implemented.
  Thus, the robustness of the dialogue system is improved since all the responses are produced consistently with the context of the ongoing dialogue. These strategies are based on:
  - the available confidence measures (both from the speech recognition and the language understanding modules),
  - the history of the ongoing dialogue (∼short term history, i.e. the dialogue concepts referred to so far during the ongoing dialogue),
  - the history of dialogue (∼long term history, i.e. the dialogue concepts referred to so far during past dialogues),
  - the status of the system (i.e. the current values of the different functionalities of the system: CD, radio, volume, and so on),
  - the task model (e.g. a semantic frame containing all the information needed to meet a specific dialogue goal),
  - and the application domain model (e.g. information on the number of tracks of a particular CD).

As a result of the designed strategies, the system is able to deal with dialogue phenomena such as *anaphora* (i.e. elements that refer to other previous parts of the dialogue) and *ellipsis*(i.e. omission of certain essential elements of the dialogue that may be derived from given context).

Table 5 shows a possible dialogue as an example of the usefulness of the dialogue context. Typically, a particular parameter is omitted immediately in those subsequent user's commands which have the aim of assigning a new value to that parameter. Based on this assumption it is possible to check the dialogue history from

Table 4
Example of the BI process for the "volume adjustment" goal.

Concepts	BN node	Occurrence (evidence)	BI result $P(C_j = 1\|E^*)$	Classification
VOLUME_ACTION	C_1	Absent ($C_1 = 0$)	Absent ($0.45 < \theta$)	*Unnecessary*
VOLUME_PARAM	C_2	Present ($C_2 = 1$)	Present ($0.95 > \theta$)	*Required*
VOLUME_VALUE	C_3	Absent ($C_3 = 0$)	Absent ($0.90 > \theta$)	*Missing*

Table 5
Concept recovery using the dialogue history (DH).

Turn (U: user; S: system)	Details
...	
U: *"Play track number two"*	
S: *"Track number two is now playing"*	
U: *"Play number three"*	The user omits the "track" parameter info (it is a missing concept according to Table 3 analysis!)
S: *"Playing track number three"*	According to the specified value, both a "track" or a "disc" are suitable. The system disambiguates the correct one just by checking the DH from more recent to older entries and retrieving the newest one
U: *"Five"*	Referring to the "track" parameter once again
S: *"Track number five selected"*	Once again the system elicits the correct parameter
...	

more recent to older entries in order to extract, if possible, the closest suitable parameter, e.g. the user instantiates the "track" parameter for a particular disc: "Play track number two", subsequently he or she simply says: "Play number three". In the last utterance the parameter has been omitted (i.e. missing) but it can be perfectly elicited from the dialogue history applying the described procedure. If we find any, this could be included as a new evidence before applying inference. On the other hand, if we do not find any suitable parameter, we would have to expect the corresponding goal to be active although the parameter is omitted. More details regarding the dialogue strategy and the use of contextual information can be checked in Fernández-Martínez et al. (2005).

2.5. The dynamic response of the system

As a dynamic feature of the behaviour of the system, attenuation mechanisms have been introduced that lower the relevance or the latency of information stored in past phases of the evolution of dialogue. This mechanism is biologically inspired and relies on Bain's preliminary theoretical base for contemporary neural networks (Bain, 1894). According to Bain's theory, as activities were repeated, the connections between neurons strengthened thus leading to the formation of memory.

Every time the system manages to execute an action, all the related useful concepts (i.e. required concepts) are stored in the dialogue history with maximum relevance (i.e. 1.0). After being stored, and as a result of the attenuation suffered after each dialogue turn, the relevance of these elements can evolve to a level below a predefined threshold, so that they finally disappear definitively from the dialogue history. For the experimental approach, we have assumed this relevance to be subject to exponential decay with a mean lifetime of half a minute as the dialogues were not expected to be very lengthy (the threshold was set to $1/e \approx 0.37$). Because of this mechanism, it is possible to keep the dialogue history permanently updated by assigning higher weight to more recent information, and lower weight to older information.

Another immediate use of this mechanism is that automatically, and without any clarification process, both erroneous and spurious elements (i.e. dialogue concepts) could be simply discarded from dialogue if these elements are no longer referenced by the user. We have included an example of a possible dialogue showing this feature in Table 6.

2.6. Baseline approach

In Fernández-Martínez et al. (2008) we presented the baseline results corresponding to the evaluation approached for the first version of our system prototype (i.e. HIFI-AV1) with real users. We would like to highlight two of the most important results derived from them:

- First, experience proved to be a key factor regarding the dialogue performance.
- Second, the designed strategies for the use of contextual information were validated by measuring their true significance as an significant reduction of the system requests (thus resulting in more fluent and efficient dialogues).

2.6.1. The value of experience

The successful interaction between a user and a spoken dialogue system is significantly conditioned by the learning process that the user experiences when addressing the system. Most of the problems related to the user–system interaction (e.g. turn-taking issues) tend to disappear (or at least become less significant) as the users adapt their behaviour to the limitations of the system. In other words, the user–system interaction improves as the user learns how to address the system.

In the particular case of HIFI-AV1, this learning (or experience) factor was evident through the different types of scenarios evaluated. In fact, despite what we would have expected, a worse performance for free scenarios (i.e. a higher degree of initiative was allowed to the user and, therefore, much more open and complex expressions were expected to take place, thus making their corresponding recognition and understanding more difficult), the measured performance for these scenarios, from an objective point of view (Fernández-Martínez et al., 2008), was as good as for the basic and the advanced ones. The main explanation for this result was found in the greater experience accumulated by users at the time of their evaluation (i.e. free scenarios were the last). As the learning stage proceeded, users were able to exploit the acquired experience leading to more fluent and efficient dialogues (i.e. just before facing the free scenarios, every user accumulated about 80% of the total experience due to the whole evaluation process), for example by reusing those expressions or ways of addressing the system which proved to be useful.

Table 6
Dialogue example of the attenuation procedure.

Turn (U: user; S: system)	Details
U: *"Volume"*	The user does not specify any "volume" value
S: *"What do you want to do with the volume?"*	The system identifies the "volume" value as "missing"
U: *"Play track number five"*	The user is not interested in modifying the volume
S: *"Track number five is playing, would you like to do something with the volume?"*	Though decreasing, the remnant evidence level of the "volume" parameter is still significant enough (corresponding goal still positively inferred!)
U: *"Play track number seven"*	New evidence decrease
S: *"Track number seven playing, what would you like me to do with the volume?"*	Still trying
U: *"Track number nine"*	Evidence of "volume" parameter falls below the threshold and the system removes it from the memory
S: *"Track number nine now playing"*	The system stops prompting the user about the volume (i.e. only "track selection" goal is active)

2.6.2. The value of contextual information

In connection with the evaluation of our first prototype (i.e. HIFI-AV1) with real users we also made an attempt to measure the true significance of the proposed contextual information handling strategies. For this purpose, we measured the *percentage of contextual turns* as the fraction of dialogue turns in which some of the strategies are successfully applied.

Logically, any piece of information that is essential for the resolution of a dialogue but cannot be recovered from the dialogue context must be requested from the user. Therefore, and in connection with the aforementioned metric, we also measured the *percentage of system requests* which has to be limited by the contextual capabilities of the system.

The results for both metrics endorsed the valuable role of these strategies regarding the dialogue management. Specifically, we concluded that more than half of the turns relied on this type of information (Fernández-Martínez et al., 2008). In other words, without the contextual capabilities provided, the number of system requests would have increased considerably (i.e. would have doubled at least). This result is particularly important in terms of dialogue efficiency and fluency.

2.7. The new approach: refining the actuation strategy

As part of every dialogue strategy, it is important to define an actuation algorithm for deciding the proper order to carry out the actions corresponding to those positively inferred goals for a certain turn of the user (i.e. execution order).

Obviously, every information item needed must be available in the dialogue context for any particular goal to be satisfied (i.e. to carry out its corresponding actions). Hence, before going into details regarding the designed algorithms and as part of the actuation problem, first it is important to summarise every different state in which any goal may be in accordance with the evolving dialogue:

1. *Inactive*: a goal which has been negatively inferred from all the available information after the last user turn (i.e. absent).
2. *Active*: a goal which has been positively inferred (i.e. present).
3. *Complete*: an active goal which is ready to be satisfied as every information item required to fully achieve it is available.
4. *Executed*: an active and complete goal which has already been actuated.

Every goal follows the aforementioned state sequence during the dialogue but, as can be easily deduced, just one user turn may be enough for a goal to be successively activated, completed and finally executed.

Fig. 3 presents an example showing the relevance of properly deciding the order of the required actions with regard to the inferred goals and hence the user's intention.

Let's assume that the user has already addressed the system in an attempt to fulfil those five different goals. The sentence is self-contained as it includes every information item needed to fulfil all of these goals (i.e. all the inferred goals can be regarded as completed), however it is strictly necessary to carry out the corresponding actions in an appropriate order. In fact, despite the order in which every goal is roughly referred to in the sentence (e.g. the "Equipment status selection" goal could be regarded as the first one, since its related keywords are placed at the very beginning of the sentence, whereas the "Audio source selection" could be considered the last one for similar reasons), if we try to sort them according to a suitable execution order then we should do it as labelled in the figure.

For instance, if we want the playing action to take effect on the desired source, it is absolutely necessary to have previously selected that specific source in the Hi-Fi equipment, otherwise no subsequent action would be effective (or simply imagine what would happen if we just decide to switch on the equipment at the end...It would be the one and only action that would be well implemented!).

2.7.1. The HIFI-AV1 actuation algorithm: looking at priorities

It is clear that the equipment itself encodes a priority scheme among the defined goals. Therefore, in order to ensure a suitable response of the system, our first version of the algorithm was basically aimed at preserving these goal priorities. That algorithm could be briefly summarised as follows:

1. First, we looked at every active goal.
2. Then we re-ordered them according to their priority (i.e. from the highest to the lowest).
3. And finally, we go through the list of active goals checking whether each one was completed. If so, we just carried out the corresponding action and continued with the next goal in the list. If not, we simply stopped the actuation process, possibly leaving some other completed goals without their being carried out, and proceed to request the missing elements which are required to complete the goal that we are checking from the list.

After the evaluation of the HIFI-AV1 system, we conducted a thorough and comprehensive analysis of this system actuation algorithm based on both objective (i.e. quality and efficiency metrics) and subjective (i.e. questionnaires) results.

As a result of the study we realized that this algorithm had two main drawbacks that are now described in the subsequent sections.

2.7.2. The priority problem: position matters

The first important problem we detected was that priority information is not enough. Consider the following example: "*Lower the volume and switch off the equipment.*" In this case, assuming that the priority of the latter goal is bigger than the former's one, we would first switch off the equipment and then try to lower the volume, which of course would not make any sense.

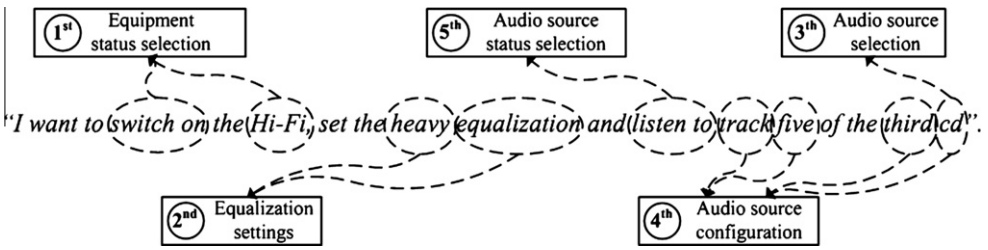

Fig. 3. Example showing the relevance of the actuation order.

Table 7
Examples of dialogues where "blocking" occurs: Hifi-AV1 solution.

Turn (U: user; S: system)	Details
...	–
U: *"I want to switch on the system, raise the equalization volume and listen to track five on third cd"*	The sentence does not contain a reference to any equalization value (e.g. the user deliberately does not specify any or just because of a possible recognition error)
S: *"What equalization would you like to set?"*	Dialogue gets blocked (one inferred goal cannot be completed yet!)
U: *"I want to switch on the equipment first."*	The user wants to power on the equipment before carrying out any other action
S: *"Which equalizer value would you like?"*	Everything remains blocked
U: *"Please, I want you to switch on the system."*	User insists, annoyed. Still no reference to any equalization value
S: *"How would you like to setup the equalizer?"*	Everything's blocked. System keeps trying to solve the equalization setting first
U: *"Ok, heavy."*	User provides an equalization value
S: *"I've switched on the system and raised the volume. I've set the heavy equalization. The cd is selected. Track five on cd three starts playing..."*	Blocking disappears
...	–

Table 8
Examples of dialogues where "blocking" occurs: Hifi-AV1 + position info.

Turn (U: user; S: system)	Details
...	–
U: *"I want to switch on the system, raise the equalization volume and listen to track five on third cd"*	Same user turn than in the previous example (Table 7)
S: *"I've switched on the system and raised the volume. What equalization would you like to set?"*	Dialogue gets blocked though, at least, two goals are executed (position information helps!)
U: *"Play track number five of the cd, please."*	Still no reference to any equalization value
S: *"The cd is selected. Which equalizer value would you like?"*	The list of active goals is updated and reordered: "audio source selection" is now the first complete goal ("equalization setting" has the same priority but worse positioned!). One more goal executed though some blocking remains (e.g. playing has not yet started!)
U: *"Heavy."*	User finally provides a value
S: *"I've set the heavy equalization. Track five on cd three starts playing..."*	Blocking disappears
...	–

Trying to figure out how to solve this problem we decided to:

- First, allow the possibility of assigning two or more different goals the same priority (e.g. "volume selection" and "equipment status selection" may be regarded as of equal priority since both are related to the same main device: the Hi-Fi's amplifier).
- And next, decide any possible tie using the position in the sentence information for each goal (estimated as the average of the positions in the sentence of all the items related to a specific goal).

We will go into more details regarding this problem and the suggested solution through the dialogue examples presented in Tables 7 and 8.

2.7.3. The blocking problem: inferred, completed but not executed

The second and, maybe, more important problem we detected was the possible blocking of some completed goals.

Tables 7 and 8 presents a pair of dialogue examples to clarify both, the priority and the blocking problems. To better understand the impact of blockings on the dialogue, we have decided to start in both examples from the same user turn and analyse the evolution that each dialogue will follow for two different solutions: the solution used by the HIFI-AV1 system and the new possible solution, described in the previous section, introducing the use of the position information.

In Table 7, we have a possible evolution of the dialogue just using priority information to decide the actuation order, while in Table 8 we are also including position in the sentence information. In both cases the actuation only takes place for those goals that have already been completed.

In the first example, the user is not able to carry out any action, so everything remains blocked, until he finally provides an equalization value. On the other hand, in the second example the blocking is somehow smoothed thanks to the position information.

According to the dynamics of the system, presented in Section 2.5, this blocking may disappear with time, although requiring the user to not make any reference to anything related to this blocking during a certain period of time can be considered as absolutely excessive. In the next section we are presenting a direct and efficient solution to the blocking problem that does not rely on this attenuation mechanism.

2.7.4. The new HIFI-AV2 actuation algorithm

As a result of the aforementioned study, we finally defined a procedure that ensures the proper implementation sequence for those actions by combining the prevalence relationships between the corresponding goals (i.e. priority information), and the order in which they appear in the sentence (i.e. position information).

In that respect, and to achieve the highest possible naturalness, the actuation for a specific active goal should take place as soon as possible, that is, as soon as the goal is completed (i.e. ready to be carried out).

On the other hand, to ensure maximum flexibility it is essential to allow these actions even if dialogue has not been completely resolved (i.e. a goal remains active but incomplete in the sentence because of errors and/or ellipsis). Therefore, the system actuation should focus exclusively on those goals that are active and complete (and not only active as in the first version algorithm).

Finally, in order to achieve the maximum possible robustness it is absolutely necessary to act coherently and orderly with respect to the rules or restrictions on the application domain (i.e. the scheme of priorities considered). Table 9 shows the evolution for the same example but this time applying the new actuation algorithm (it can be seen that the blocking problem has absolutely disappeared).

Table 9
Same example but applying HIFI-AV2 solution (no "blocking" occurs).

Turn (U: user; S: system)	Details
...	-
U: *"I want to switch on the system, raise the equalization volume and listen to track five on third cd"*	Same user turn than in previous examples (Tables 7 and 8)
S: *"I've switched on the system and raised the volume. The cd is selected. Track five on cd three is playing. What equalization would you like to set?"*	No blocking at all!!! (priority and position are combined to define the proper implementation sequence of every active and complete goal)
U: *"Heavy."*	The user provides an equalization value
S: *"I've set the heavy ecualisation. The cd was already playing."*	Dialogue is solved in less number of turns than with any of the previous solutions
...	–

3. Evaluation

In recent years several works have defined different metrics to assess the performance of HCI (Human–Computer Interaction) systems (Gibbon et al., 1997; Mertins and Moore, 2000; Callejas and López-Cózar, 2008), both objective (i.e. based on measures obtained by the system itself, such as word error rate, turn correction ratio or implicit recovery (Danieli and Gerbino, 1995)) and subjective (usually based on questionnaires presented to the evaluators at the end of the interactions, asking questions related to their perception on the performance of the system, or their satisfaction (Hirschman et al., 1990)).

In this paper we are presenting a holistic evaluation of the system based on the outcomes of users' interactions. With that purpose, we have recruited real users interacting with the system through a set of predefined scenarios. We have designed a twofold evaluation process aimed at assessing both objective and subjective features. The objective part was based on automatically collected metrics related to dialogue quality and efficiency. On the other hand, the subjective part was rather based on the assessment of different system features such as quickness or robustness. It was articulated through a user questionnaire. Both the objective and subjective metrics that we have used to perform the evaluation of our system are mostly derived from the PARADISE framework (Walker et al., 1997, 2000; Möller et al., 2007).

3.1. Evaluation scenarios

A set of 15 dialogue goals were defined covering the typical functionality available in commercial Hi-Fi systems (e.g. playing, recording, radio, volume, disc, track, or tape selection, etc.). From this goal set, different types of scenario were designed according to different initiative styles and task complexity levels. The whole set of defined scenarios added up to a total of 45, that can be grouped into the following categories:

- *Basic* (strongly guided tasks aimed at demonstrating mandatory functionality): 23 in total (see Table 14). The user has to try to fulfil just one dialogue goal (e.g. "*The user should try to stop the current disc playing*"). The dialogue context (i.e. the dialogue history and the system status) is prepared according to the targeted goal.
- *Advanced* (less guided but more complex scenarios): 19 in total (see Table 15). On the one hand this type of scenario is aimed at demonstrating the flexibility, robustness, and adaptation capabilities of the system. On the other hand, users have to try to achieve multiple dialogue goals (e.g. "*The user should try to play a particular track without referring to the specific disc the track belongs to*"). Similarly to the "basic" case, the dialogue context is prepared according to the targeted goals.
- *Free* (absolutely absence of guidance): 3 in total (see Table 16). This time the user is absolutely free to decide what to do with the system. However, and to ensure a balanced coverage of the available functionalities, we suggested that the users mainly focus on a particular device (i.e. one of the three different devices that the system is equipped with: cd, cassette or radio) thus resulting in three different scenarios. In any case, this difference was not that significant since every user tried all of them. Unlike the other two, the starting dialogue context is always set to a default state (i.e. empty history and system switched off).

3.2. Data collection

We have evaluated the two versions of our prototype, which we have called HIFI-AV1 and HIFI-AV2 respectively, corresponding to the implementation of the two different algorithms for the management of the actuation order presented in Section 2.7.

The two versions of the system have been tested by both students and members of the faculty and research staff (researchers on speech technology) from different Spanish universities on behalf of the EDECAN project. A total of 15 speakers tried the first version of our prototype whereas a different group, made up of 17 new speakers (a completely different group of participants; i.e. each participant only tested one system), was later recruited for the second evaluation thus targeting the new version of the system.

Each participant was required to complete 10 dialogues or scenarios according to the following distribution: 3 basic, 6 advanced and 1 free scenarios. Thus, a total of 150 dialogues were collected for HIFI-AV1 and 170 for HIFI-AV2.

User–system interaction took place in a specially prepared living room equipped with the Hi-Fi system where users promptly received a brief description of the tasks they were requested to accomplish for each scenario.

All the participants were classified either as *novice* or *expert* according to their previous experience on interacting with spoken dialogue systems. Every user answered on a 1–5 scale, being 5 the highest experience, so that users that rated their experience as 3 or lower were automatically regarded as novices. Regretfully, the group recruited for evaluating HIFI-AV2 did not have any novices. Therefore, and in order to compare both versions of the system fairly, we decided to remove all the novices from the HIFI-AV1 evaluation set for comparisons between both approaches (i.e. only the 6 HIFI-AV1 experts and their corresponding 60 scenarios were considered when comparing with HIFI-AV2). Table 10 shows the user distribution for each evaluation.

Both groups (i.e. HIFI-AV1 and HIFI-AV2 experts) are formed by people doing some research on speech technology, thus we expect

Table 10
HIFI-AV1 vs. HIFI-AV2: user and experience level distribution (expertise score in 1–5 scale, being 5 the highest).

	HIFI-AV1	HIFI-AV2
Number of experts	6	17
User expertise (mean)	4.5	4.29

our analysis of the results to be more focused specifically on the advantages of the new actuation algorithm rather than on different experience levels or learning skills.

Nonetheless, there is an important difference between both groups that should be remarked. Particularly, all the recruited speakers for the HIFI-AV1 evaluation were from Madrid whereas those recruited for the HIFI-AV2 one were from different Spanish regions, namely: Madrid, Aragón, Valencia, and Basque Country.

According to the analysis presented in Fernández-Martínez et al. (2010) for the HIFI-AV2 evaluation process, the different accents and regions of origin of the speakers clearly affected the speech recognition performance. This analysis showed that the word error rate (WER) was significantly better for the group of speakers from Madrid (e.g. WER almost tripled for the group of speakers from Aragón). In this regard, it is important to remark that the same speech recognizer was used in both evaluations and that it was not adapted to any specific dialect. Therefore, although no WER result can be reported for HIFI-AV1 (none of the HIFI-AV1 evaluation sentences were transcribed), a better speech recognition accuracy could be expected in this case.

This aspect (i.e. the HIFI-AV2 evaluation was performed with a group of users for which the speech recognizer performance was worse) should be considered when comparing both versions of the system.

3.3. Collected metrics

Data labelling through manual transcription is a costly and time consuming work that has not been done yet. Instead of that, a combination of dialogue quality and efficiency measures have been automatically logged or computed (Walker et al., 2000). Some of the considered metrics have been expressed as the percentage of turns where the specific event takes place.

3.3.1. Dialogue quality metrics

By tackling the blocking problem, we are particularly addressing dialogue efficiency as a major measure of the dialogue performance for benchmarking purposes between our two prototypes (i.e. *turn efficiency*, as introduced later on). Therefore, among the different quality metrics that were collected, we emphasize:

- the *recognition and understanding rejections*, that happen when either the recognition or the understanding confidence value is below a predefined threshold,
- or the *out of domain turns*, when no goal is positively inferred for a particular sentence.

Table 11
Survey to be filled by every user.

	Survey
1	What is you level of experience using speech interfaces?
2	Did the system understand what you said?
3	Did the system carry out the actions you requested?
4	Was the system's vocabulary and the available phraseology acceptable?
5	Did the system respond quickly enough?
6	Was the feedback information provided by the system easy to understand?
7	Was the system able to act coherently with dialogue context (e.g. system's status, previously executed actions, etc.)?
8	Was the system easy to use?
9	Did the system work the way you expected?
10	Was the available functionality acceptable?
11	How would you rate the system overall?
12	Would you use the system regularly instead of the IrDA remote control?

These metrics have the most important impact on dialogue efficiency since each of them involves dialogue turns that do not result in any performed action (i.e. null efficiency).

3.3.2. Dialogue efficiency metrics

Among the efficiency metrics proposed, we would like to highlight:

- the *contextual or context-dependent turns*, turns that rely on the contextual information resources for their disambiguation (i.e. implicitly inferred information). It can be estimated as the percentage of turns in which some of the contextual information handling strategies are applied successfully.
- the *system requests*, when the system decides to request some information element from the user. It can be estimated as the percentage of turns where the system requests the user for some missing or deliberately omitted information.
- and the *turn efficiency*, which can be regarded as the number of actions that are executed per turn.

3.4. User satisfaction questionnaires

In order to obtain subjective ratings of the system we conducted user satisfaction questionnaires. First, we requested users to rate the task or scenario success after each scenario. Finally, after the evaluation, users also filled out forms rating typical spoken dialogue system features on a 1–5 scale (i.e. 1 – very poor, 2 – poor, 3 – fair, 4 – satisfactory, or 5 – highly satisfactory). Among the assessed features (see Table 11) we could remark ASR and TTS performance, task ease, system response, etc.

4. Results

4.1. Objective evaluation

In this section we compare both versions of our system (i.e. HIFI-AV1 vs. HIFI-AV2) with regard to the dialogue efficiency metrics that were presented in Section 3.3.2. In this regard, a two-way ANOVA was performed to determine if there is a difference in each reported metric for the tests with different systems or scenario types. In the presence of a significant difference, multiple comparisons were performed using the Tukey procedure at the $\alpha = 0.05$ significance level. The relative improvement between systems has been highlighted in every figure of this section by means of a number inside a box depicted for each pair of columns; error bars indicating one standard deviation of uncertainty have also been included.

4.1.1. Scenario length

Table 12 shows the average dialogue or scenario length (measured as the number of turns) for every type of scenario and system (standard deviations are also indicated in parentheses). The main effect for the scenario type was significant ($F(2, 224) = 58.152$ and $p < 0.001$). Multiple comparisons using the Tukey's test (at

Table 12
HIFI-AV1 vs. HIFI-AV2: dialogue length distribution.

System	Scenario type			
	Basic	Advanced	Free	ALL
HIFI-AV1	5.00 (4.243)	7.40 (6.050)	16.33 (8.869)	7.57 (6.657)
HIFI-AV2	4.00 (3.458)	6.85 (6.556)	18.94 (8.437)	7.21 (7.280)
% (AV2-AV1) diff.	−20.00	−7.39	15.97	−4.85

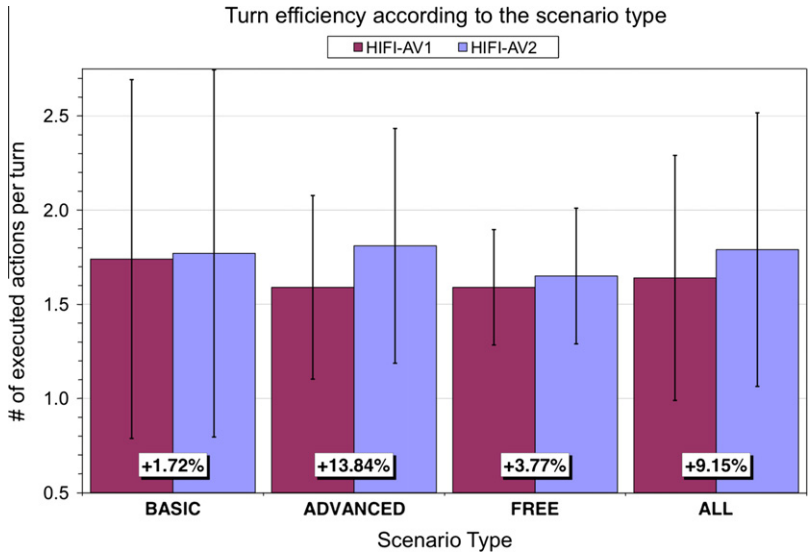

Fig. 4. Detail of the turn efficiency results.

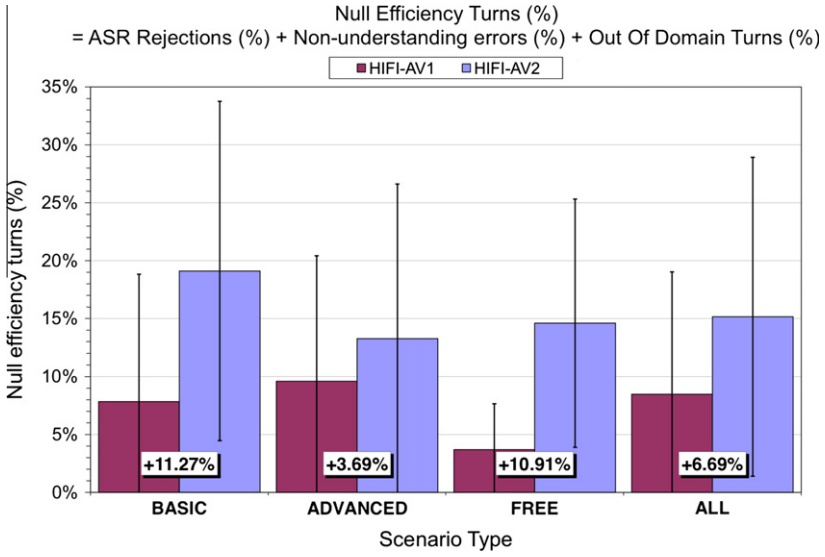

Fig. 5. Hifi-AV1 vs. Hifi-AV2: null efficiency turns (larger numbers indicate worse performance).

the $\alpha = 0.05$ significance level) indicated that advanced scenarios were longer than basic (i.e. $p = 0.001$) and that free scenarios were longer than advanced (i.e. $p < 0.001$). On the other hand, main effect for the system used was not significant ($F(1,224) = 0.167$ and $p = 0.683$).

4.1.2. Turn efficiency

Fig. 4 shows an individual analysis of the *Turn efficiency* metric for both systems across all scenario types. The two-way ANOVA analysis showed that none of the main effects were significant. We respectively obtained $F(1,224) = 1.085$ and $p = 0.298$ for the system used, and $F(2,224) = 0.483$ and $p = 0.617$ for the scenario type.

For a better understanding of this result, we have also included Fig. 5. In that figure we have presented what we have defined as the percentage of *null efficiency turns*. A null efficiency turn can be regarded as a turn that does not result in any performed action, thus limiting the turn efficiency of the corresponding scenario. Hence, a null efficiency turn could be motivated by either:

- a *recognition rejection*, a turn with a low ASR confidence result,
- a *non-understanding error*, a turn for which no valid NLU result is obtained in spite of a valid recognition result,
- or an *out of domain turn*, a turn for which no goal is positively inferred.

Table 13 presents the individual contribution of each type of error for all the scenario types and for both systems. The two-way ANOVA analysis performed for the percentage of null efficiency turns showed that the main effect for the scenario type was not significant, $F(2,224) = 0.859$ and $p = 0.425$. On the contrary, main effect for the system used was significant, $F(1,224) = 11.806$ and $p = 0.001$, thus indicating that the second system suffered a higher percentage of null efficiency turns than the first one.

Null efficiency turns measured were clearly favourable to the HIFI-AV1 system. However, none of the previously mentioned errors can be attributed to the existing differences between systems regarding the dialogue management strategy, but mainly to the greater variety of speech accents that we found among HIFI-AV2

Table 13
HIFI-AV1 vs. HIFI-AV2: source error distribution for null efficiency turns.

Metric	Scenario type							
	Basic		Advanced		Free		ALL	
	AV1	AV2	AV1	AV2	AV1	AV2	AV1	AV2
% ASR rejections	5.71	16.64	8.11	11.39	3.70	13.19	6.95	13.14
% NLU rejections	0.83	0.72	0.69	0.46	0.00	0.55	0.66	0.55
% Out of domain turns	1.29	1.75	0.79	1.43	0.00	0.87	0.86	1.47
% Null efficiency turns	7.84	19.11	9.59	13.28	3.70	14.60	8.47	15.16
% (AV1-AV2) diff.	11.27		3.69		10.91		6.69	

speakers (Table 13 points to ASR rejections as the most important contribution to null efficiency turns; (Fernández-Martınez et al., 2010) presents a detailed analysis of the corresponding recognition results per dialectal region).

As can be derived from Figs. 4 and 5, the gap between systems is reduced, particularly for basic and free scenarios, because of a worse figure for HIFI-AV2 regarding null efficiency turns (for advanced the difference is just about 3.7%, as can be checked in Fig. 5, so that the relative improvement in turn efficiency reaches its top, almost 14% as reported in Fig. 4). This effect must be taken into account for a fair comparison between systems.

Particularly, although HIFI-AV2 was around 9% better on average than the first solution, this margin could have been even much better with a similar amount of null efficiency turns (or without any, as hypothetically suggested in the next subsection).

4.1.3. Turn efficiency without null efficiency turns

To better understand the real performance of the new actuation algorithm, we have included Fig. 6 where we have corrected the turn efficiencies reported in Fig. 4 as if there was not any null efficiency turn (i.e. assuming the same amount of executed actions and discounting the amount of null efficiency turns to the overall dialogue length, thus compensating their effect on the estimated turn efficiency).

For example, HIFI-AV2 got a 1.77 turn efficiency and around 19% of null efficiency turns for basic scenarios; by assuming that none of them would ever have happened (which could be considered as an optimal upper bound for efficiency where no recognition, understanding nor inference problems take place), we would improve the turn efficiency up to 2.19. Of course, we have not measured how null efficiency turns could affect to subsequent turns in the dialogue. Therefore, this should be regarded just as an estimation of how the results could look like. Nonetheless, this figure enables a reasonable estimation of the turn efficiency improvement (without the above mentioned errors) across systems, which is reasonably constant throughout the different types of scenario (i.e. around 17% as for the overall case).

A two-way ANOVA analysis was also performed over this corrected metric. Main effect for the system used was significant, $F(1,224) = 6.881$ and $p = 0.009$. On the other hand, main effect for the scenario type was not, $F(2,224) = 1.302$ and $p = 0.273$.

4.1.4. Contextuality and system requests

We are also comparing both systems with regard to contextuality. In Fig. 7 we are presenting both the percentage of contextual turns and the percentage of turns resulting in a system request.

The two-way ANOVA analysis for contextuality showed that the main effect for the scenario type was significant, $F(2,224) = 8.015$ and $p < 0.001$. Multiple comparisons using the Tukey's test (at the $\alpha = 0.05$ significance level) indicated that advanced scenarios had a higher contextuality than basic (i.e. $p < 0.001$). On the other hand, main effect for the system used was not significant, $F(1,224) = 0.078$ and $p = 0.780$.

Regarding the percentage of system requests none of the main effects were significant. We respectively obtained $F(1,224) = 0.159$ and $p = 0.690$ for the system used, and $F(1,224) = 1.447$ and $p = 0.237$ for the scenario type.

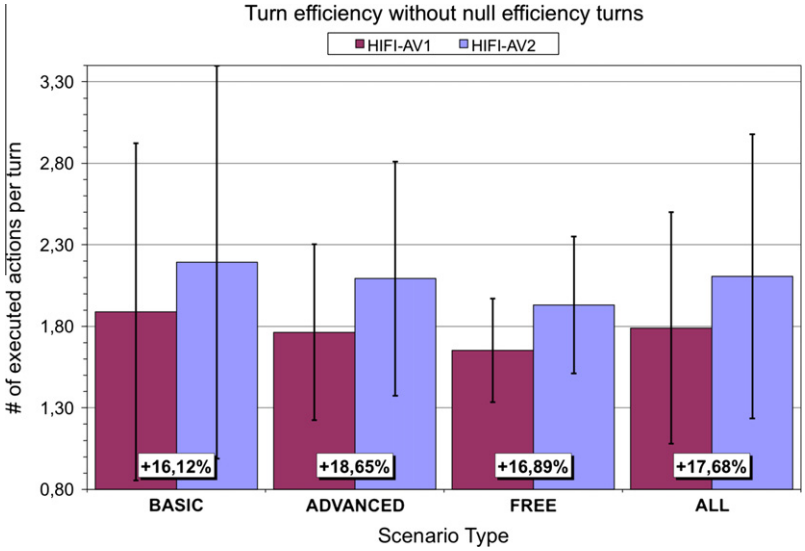

Fig. 6. Hifi-AV1 vs. Hifi-AV2: corrected turn efficiency without null efficiency turns.

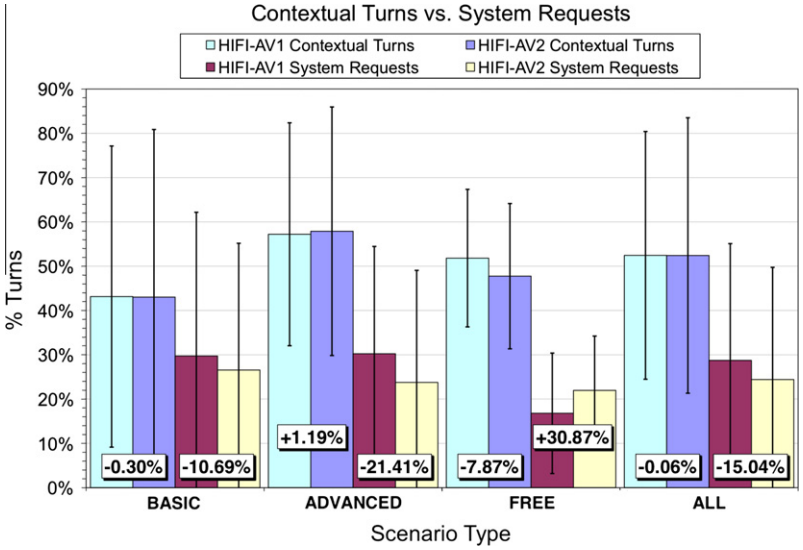

Fig. 7. Hifi-AV1 vs. Hifi-AV2: contextual turns and system requests.

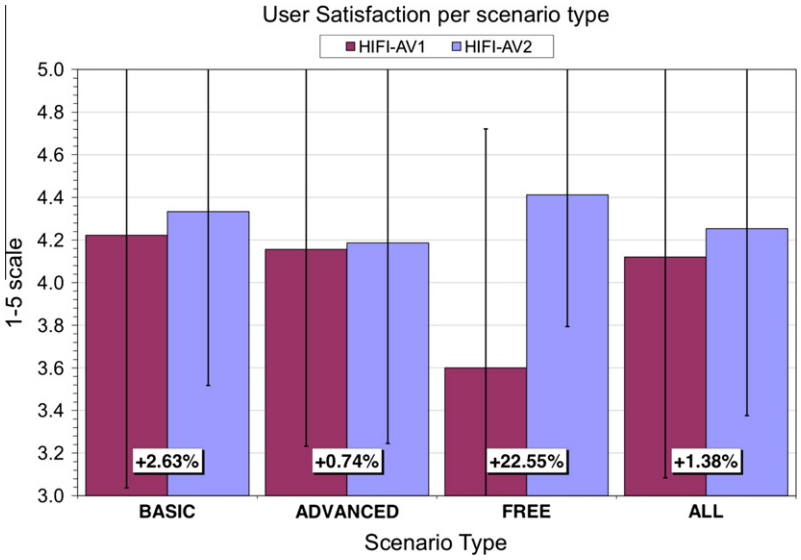

Fig. 8. Hifi-AV1 vs. Hifi-AV2: scenario success assessment.

4.2. Subjective evaluation

The subjective evaluation is based on the analysis of both a questionnaire filled out by every participant just after the evaluation, and the user satisfaction rates obtained for each scenario (the users were simply asked in this regard at the end of every scenario).

4.2.1. User satisfaction

In order to have a running estimate of the task completion rate rather than manually label each scenario, each user was asked about the level of success achieved for each scenario. Fig. 8 shows the average user satisfaction measured for each type of scenario as part of the subjective evaluation carried out for both systems.

The two-way ANOVA analysis showed that the main effect for the scenario type was not significant, $F(2,224) = 1.041$ and $p = 0.354$. On the other hand, main effect for the system used was significant, $F(1,224) = 5.339$ and $p = 0.022$, where it is seen that HIFI-AV2 system produced higher user satisfaction than HIFI-AV1.

4.2.2. Questionnaires

To conclude with the evaluation, we will also analyse the subjective ratings of both systems obtained through the user satisfaction questionnaires conducted. Fig. 9 summarises the corresponding results for both evaluations.

In this case, a two-sample Student's *t*-test assuming equal variances using a pooled estimate of the variance was performed to test, for each question included in the questionnaire, the hypothesis that the resulting mean ratings of the users for the two evaluated systems were equal.

The mean ratings from users of the two different systems were not significantly different for any of the questions. Only the difference observed for the question about the system actuation (i.e. second feature in Fig. 9, "*Did the system carry out the actions you requested?*", as previously presented in Table 11, question 3) could be regarded as marginally significant, $t(21) = 1.765, p = 0.088$.

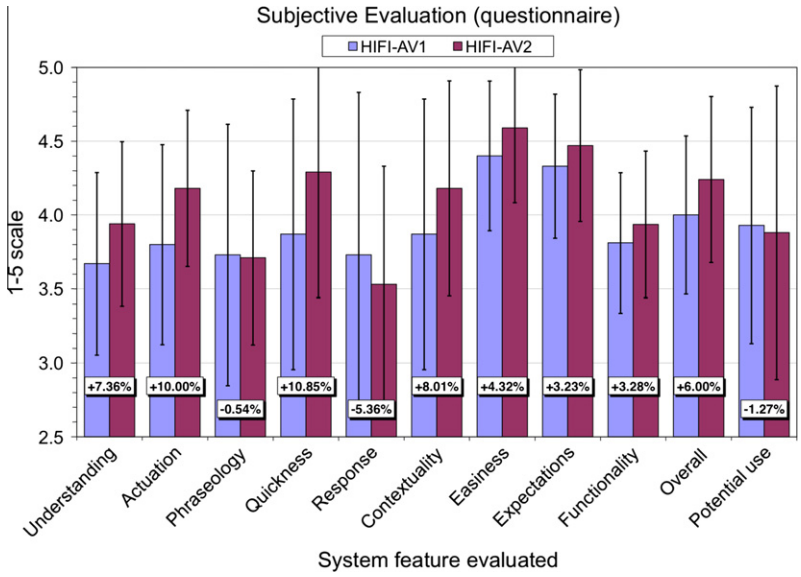

Fig. 9. Hifi-AV1 vs. Hifi-AV2: questionnaires results.

This difference was favourable to the HIFI-AV2 system that used the new actuation algorithm.

5. Discussion

5.1. Dialogue efficiency

We have compared both systems in terms of dialogue efficiency measured as the *turn efficiency without null efficiency turns*, the corrected version of the turn efficiency results presented in Fig. 6. As it can be derived from these results, when data across all scenarios are combined, the HIFI-AV2 system performs significantly better than the baseline. The absence of blockings has finally resulted in almost 18% of overall relative improvement.

In addition, we would like to point out that, on average, we measured roughly 2 goals satisfied per turn (meaning that two different actions could be accomplished in just one turn). Actually this was not our purpose but it can be considered a good result, particularly considering that we did not force or encourage the users in any sense to try solve the scenarios in as few utterances as possible.

5.2. Using contextual information

Benefits obtained through the application of contextual information handling strategies can be regarded as highly significant in both cases (more than half of the turns, roughly 55% on average, as can be observed in Fig. 7, relied on contextual information resources).

In connection with these contextual capabilities, only a quarter of the turns involved a request from the system (around a 25% on average for both evaluations; as explained in Section 2.6.2 this number would be expected to increase, at least, up to the aforementioned 55% without contextual resources).

One of the most interesting results that can be derived from Fig. 7 is that contextuality has proved to be significantly higher for advanced and free than for basic scenarios regardless of the system used. This result relies on the update of the dialogue history with the concepts that the user refers to during the interaction. As a result, the longer the dialogues, the higher the relevance of the contextual information (both advanced and free showed to be longer than basic scenarios, as presented in Table 12).

5.3. Impact on user satisfaction

The most important result that can be derived from Fig. 8, reporting on the average user satisfaction per scenario, is that the overall performance (combining all the scenario data) of our blocking-free system is significantly better than the first one.

The average rating is between 4.19 and 4.41, always above 4 which could be considered quite a good result (scale is 1–5, 5 being the best). However, the most interesting result of the individualized analysis conducted for each type of scenario is the one that we obtained for the free ones.

Free scenarios lack of any restriction on what could be done with the system so that the initiative of the user reaches its peak in this case. In particular, resulting from the absence of a specific purpose, users tended to explore the available functionality more. This freedom could favour situations in which the user may attempt to do something that actually is not allowed in the system, presumably resulting in more errors, and therefore in a worse valuation.

Nonetheless, in spite of both the higher complexity and the greater number of null efficiency turns observed for the free scenarios in the HIFI-AV2 evaluation (i.e. almost 11% more as can be seen in Fig. 5), this type of scenario was the one that received the best rating from users, which was about 22% better compared to our first prototype.

5.3.1. The questionnaire: summary of findings

As it was previously mentioned, only one of the observed differences for the questionnaire results was marginally significant. Fortunately, this was probably the most interesting one: the specific question about the system actuation (i.e. "Actuation" column-pair in Fig. 9, corresponding to question number 3 in Table 11).

This result shows that users have perceived the better turn efficiency that the new actuation algorithm has achieved by eliminating the blockings.

Table 14
Basic scenarios.

#	Description
1	The user should try to power on and off the Hifi equipment
2	The user should try to select any audio source of the Hifi equipment
3	The user should try to select any equalization of the Hifi equipment
4	The user should try to change the volume of the Hifi equipment
5	The user should try to turn on and off the mute setting of the Hifi equipment
6	The user should try to select any disc and any track
7	The user should try to start playing the cd
8	The user should try to pause and resume the cd
9	The user should try to change the playing mode of the cd
10	The user should try to stop playing the cd
11	The user should try to fast forward the cd
12	The user should try to add a new entry in the cd playlist
13	The user should try to remove any entry from the cd playlist
14	The user should try to start playing the cd playlist
15	The user should try to select any tape
16	The user should try to start playing the cassette
17	The user should try to pause and resume the cassette
18	The user should try to stop playing the cassette
19	The user should try to rewind the cassette
20	The user should try to fast forward the cassette
21	The user should try to select any radio station
22	The user should try to turn on and off the stereo mode
23	The user should try to start recording the selected audio source

5.4. Study limitations

The results that have been presented in the paper were obtained by targeting only the expert users. Regretfully, we did not welcome any novice for the second evaluation (i.e. HIFI-AV2) so that, to ensure the validity of the results, we decided to get HIFI-AV1 novices out of the comparison. However, it is still interesting and convenient to test the new version of the system with novice users (who will usually have more difficulties when interacting with the system).

Although we have measured several reliable indicators showing that the second version of the system is better than the first one, a reasonable argument to also expect a better performance for novice users, an evaluation with novices would allow us to not only confirm this, but also to measure important side effects like whether learning takes place for them at an earlier stage thanks to the improvements introduced.

Furthermore, the different populations that evaluated each of the systems (i.e. HIFI-AV1 and HIFI-AV2 experts, respectively), were discovered to differ in a significant way that impacted system performance (i.e. greater variety of dialects was found among HIFI-AV2 speakers). However, the reported results may prove to be even more valuable if we take into account that the HIFI-AV2 evaluation was performed with a group of users for which the speech recognizer performance was significantly worse (i.e. despite the worse recognition accuracy, the overall performance of the HIFI-AV2 system was better).

6. Conclusions

The intention of this work was to validate the approached dialogue management solutions, aimed at the design of better devices and intelligent interfaces that fully integrate features that improve all the aspects of the interaction with the end user, through their evaluation with real users. With this purpose in mind, a first prototype was developed and evaluated. As part of the evaluation process, we detected some features regarding the interaction of the users with the prototype that could certainly be improved: the problem of potential blockings. Therefore, a new and better actuation algorithm was then suggested proceeding to a new prototype and its corresponding evaluation.

As an immediate consequence of the elimination of blockings by the new algorithm:

- Objectively, the dialogue performance (i.e. turn efficiency after discounting the null efficiency turns) improves. We have reached roughly the number of two goals identified, completed and executed per turn. This is a good outcome, especially bearing in mind that users were not given any specification regarding the number of turns in which they had to try to overcome the different scenarios. Therefore, we can conclude that the possibilities of the system in this regard have not been fully exploited yet.

Table 15
Advanced scenarios.

#	Description
24	The user should try to power on the Hifi, change the volume and select any equalization
25	The user should try to power on the Hifi, change the equalization and set a new volume
26	The user should try to power on the Hifi, change the volume, select the cd as the audio source and start playing any track of any disc
27	The user should try to lower the volume, stop the cd and power off the Hifi
28	The user should try to select the radio as the audio source, turn the stereo on and select any radio station
29	The user should try to select the radio as the audio source, turn the stereo off and select any band
30	The user should try to select the cd as the audio source and start playing any track
31	The user should try to select the cd as the audio source and start playing any disc
32	The user should try to start playing a new track without explicitly referring to the track parameter (e.g. "play number six")
33	The user should try to start playing a new disc without explicitly referring to the disc parameter (e.g. "play number two")
34	The user should try to change the stereo mode and successively change the volume without explicitly referring to the volume parameter (e.g. "Lower")
35	The user should try to start recording without explicitly referring to any audio source
36	The user should try to start playing a track without explicitly referring to any disc
37	The user should try to select the radio as the audio source by specifying a band without explicitly referring to the desired radio station
38	The user should try to select the cd as the audio source, start playing a track of a non-existent disc, power off and then try any other action except powering on the Hifi
39	The user should try to select the radio as the audio source, select any band and any radio station, and then try several consecutive changes of station without explicitly referring to any band
40	The user should try to select the cassette as the audio source by selecting a tape, lower the volume, and then successively play, stop, and rewind the selected tape without making any explicit reference to that tape
41	The user should try to select a track and, "mistakenly", refer to the volume. Then the user should try to continue by selecting 3 new tracks while ignoring the questions that the system could make about the volume
42	The user should try to select the cd as the audio source and start playing any disc and any track. Next the user should try to select the next track and listen to it to the end. After that, the user should try to start playing another track without making any explicit reference to the track parameter (e.g. "Next")

- Subjectively, and equally important, this increased efficiency has led to greater agility and flexibility of the dialogue which, in turn, has improved the system's response. This improvement has been assessed very positively by users as can be deduced from the results corresponding to both: the collected per scenario satisfaction ratings and the "Actuation" assessment made by users through questionnaires. In particular, the individualized analysis conducted for each type of scenario puts the free scenarios as the highest-rated throughout the assessment process. This is undoubtedly a result of particular importance because the complexity of the free scenarios is maximum. These are scenarios without any restriction in which the initiative of the user reaches its top and, indeed, the nearest scenarios to the actual use of the interface.

A thorough comparison between the evaluated approaches has been completed throughout the paper. However, despite the differences resulting from different actuation algorithms, a more general discussion on the performance of the approached dialogue management solutions, and in particular focusing on their acceptance by users, is still needed.

In general, user satisfaction in relation to a particular system depends crucially on its *usability* and *functionality*. In this way we could conclude that, in order to be *useful*, a system must be *usable* first (i.e. providing services for which it is efficiently designed) and also *functional* (i.e. the services provided are of interest to users).

One of the keys for the usability of a system, and by extension for its usefulness, is its ease of use. This is the reason why we consider specially significant that this feature has been the best appreciated by users for both evaluated systems. The greater or lesser ease of use that a system is able to offer (thanks to the certainly sophisticated technologies behind the scenes), definitely conditions the final acceptance by users (in the same way as that offered functionality). According to ISO 9241 standard, the usability of a human-made object relies on both its ease of use and learnability (Part 10: Dialogue principles). We should not forget that the learning barrier has to be crossed before any effective use can take place. In this regard, the excellent ease of use of our system is also well complemented by a nice ease of learning as we already presented in Section 2.6.1.

In order to get a fluent and efficient dialogue, the user–system interaction should be: natural, flexible and robust. It is difficult to attribute each of the above features to a single aspect of the various dialogue solutions proposed. Rather, it is thanks to the synergy of these solutions, to the joint operation of all of them, how those characteristics become true. However, in relation to the excellent level of naturalness reached, we could emphasize the following reasons:

- First of all, this work focuses on speech interfaces, that is, systems based on a spoken interaction, and in that sense we must remember that speech is the most natural means of communication between humans.
- Secondly, as these interfaces are based on natural language, users can feel completely free to use any expression in order to carry out the required actions without any need to memorize either a special vocabulary or a predefined list of specific commands.
- Finally, the kind of proposed interface has the ability to negotiate with the user in achieving the dialogue goals similarly to the way a human would help, assisting the user at all times, solving his possible ignorance on how to proceed in order to fulfil those goals, properly analysing all the information provided by him and resolving any deficiencies in its content.

In relation to the requirement of the best possible flexibility, we must emphasize that the proposed dialogue manager is characterized by the absence of rules or restrictions that might restrict the dialogue in any way, resulting in greater ease of use, along with a greater naturalness. In that sense, the freedom granted to the user regarding the specification of the dialogue goals and the information provided for achieving them, is the highest. In relation to the information provided by the user, he/she could facilitate, where necessary, more information than is strictly necessary for achieving his/her dialogue goals. Furthermore, the users are not obliged to provide complete information so they can deliberately omit, if they wish, part of it without much problem.

Of course, for the latter to be possible, it implies that the system has the ability to recover that missing information. The third requirement of robustness is achieved through the use of the contextual information available. Robustness is twofold as it allows both the disambiguation of the information deliberately omitted by the user, and the recovery of lost or erroneous information during the dialogue.

In short, a more natural, flexible and robust dialogue is possible thanks to the solution for dialogue modelling based on BNs that has been suggested, and also thanks to the contextual information handling strategies. This is supported by a good user satisfaction rate, and even more by the results corresponding to the metrics that were automatically collected, which have shown the usefulness and benefits provided by the proposed solutions.

7. Future work

According to the ratings obtained from the questionnaires, it is clear that there is much room for improvement at several levels. For example, in the response generation module if we look at the least appreciated feature: the feedback provided by the system (i.e. "Response" column-pair in Fig. 9; question number 6 in Table 11).

This result was strongly influenced by the fact that the first practical application of this system was to aid people with disabilities (Ferreiros et al., 1998; Ferreiros et al., 2000). That application was mainly aimed at blind people so that a more detailed information regarding the interaction with the system (maybe more than strictly required by non-handicapped users) was proven to be particularly useful for those target users. In this regard, it is important to mention that, although system's prompts (i.e. feedback on what the system is doing or trying to do at a given point) were shortened as much as possible by removing non-critical information, the results showed that users found the listening prompts a little annoying and too verbose to be duly assimilated.

The feedback received by the users of the system is critical, especially at the beginning of the interaction. Subsequently, such information must be tailored to the degree of experience that the user gradually acquires (or even to the level of disability if the users are disabled people as previously mentioned). Certainly, users tend to need significantly less feedback as they become more

Table 16
Free scenarios.

#	Description
43	The user is absolutely free to decide what to do with the system but, please, use the cd once at least
44	The user is absolutely free to decide what to do with the system but, please, use the cassette once at least
45	The user is absolutely free to decide what to do with the system but, please, use the radio once at least

familiar with the system. This process of adaptation is absolutely necessary to avoid the most part of the response of the system becoming just redundant and/or inappropriate. The valuation for the response of the system, result obtained from a system subjective assessment questionnaire, points clearly to the need to incorporate user profiling capabilities into the system in order to adjust the behaviour and response of the system to different skill, experience or disability levels.

Acknowledgements

This work has been supported by MA2VICMR (S2009 TIC 1542–111287), ROBONAUTA (MEC ref: DPI2007–66846-c02–02) and SD-TEAM (MEC ref:TIN2008–06856-C05–03) projects. The authors thank the reviewers for their helpful comments to improve the manuscript. The authors would also like to thank the members of the EDECAN project consortium (TIN2005–08660-C04) for their participation in the evaluations.

References

Aarts, E., de Ruyter, B., 2009. New research perspectives on ambient intelligence. Journal of Ambient Intelligence and Smart Environments 1 (1), 5–14.

Augusto, J.C., 2007. Ambient intelligence: the confluence of ubiquitous/pervasive computing and artificial intelligence. Intelligent Computing Everywhere, pp. 213–234.

Bain, A., 1894. Mind and Body: The Theories of Their Relation. International Scientific Series. D. Appleton & Company.

Berton, A., Bühler, D., Minker, W., 2006. Smartkom-mobile car: user interaction with mobile services in a car environment. SmartKom: Foundations of Multi-Modal Dialogue Systems, Cognitive Technologies, pp. 523–541.

Bohus, D., Rudnicky, A.I., 2009. The Ravenclaw dialog management framework: architecture and systems. Computer Speech and Language 23 (3), 332–361.

Bui, T.H., 2006. Multimodal Dialogue Management – State of the Art. Technical Report TR-CTIT-06-01, Centre for Telematics and Information Technology University of Twente, Enschede.

Callejas, Z., López-Cózar, R., 2008. Relations between de-facto criteria in the evaluation of a spoken dialogue system. Speech Communication 50, 646–665.

Danieli, M., Gerbino, E., 1995. Metrics for evaluating dialogue strategies in a spoken language system. In: Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse, Interpretation and Generation. pp. 34–39.

de Ruyter, B., Saini, P., Markopoulos, P., van Breemen, A., 2005. Assessing the effects of building social intelligence in a robotic interface for the home. Interacting with Computers 17 (5), 522–541, http://dx.doi.org/10.1016/j.intcom.2005.03.003.

Dybkjaer, L., Bernsen, N., Minker, W., 2004. Evaluation and usability of multimodal spoken language dialogue systems. Speech Communication 43, 33–54.

Fernández-Martínez, F., Blázquez, J., Ferreiros, J., Barra, R., Macías-Guarasa, J., Lucas-Cuesta, J., 2008. Evaluation of a spoken dialogue system for controlling a hifi audio system. In: Proceedings IEEE Workshop on Spoken Language Technology (SLT08), Goa, India, pp. 137–140.

Fernández-Martínez, F., Ferreiros, J., Cordoba, R., Montero, J.M., San-Segundo, R., Pardo, J.M., 2009. A Bayesian networks approach for dialog modeling: the fusion BN. In: ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing.. IEEE Computer Society, Washington, DC, USA, pp. 4789–4792.

Fernández-Martínez, F., Ferreiros, J., Sama, V., Montero, J., San-Segundo, R., Macías-Guarasa, J., 2005. Speech interface for controlling a hi-fi audio system based on a bayesian belief networks approach for dialog modeling. In: Proceedings Interspeech, Lisbon, Portugal, pp. 3421–3424.

Fernández-Martínez, F., Lucas-Cuesta, J.M., Chicote, R.B., Ferreiros, J., Macías-Guarasa, J., 2010. Hifi-av: an audio–visual corpus for spoken language human–machine dialogue research in spanish. In: Chair, N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (Eds.), Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta, pp. 2974–2980 (May).

Ferreiros, J., Colás, J., Macías-Guarasa, J., de Córdoba, R., Pardo, J., 2000. Control de un equipo de alta fidelidad usando frases habladas de manera natural. In: Congreso Iberoamericano IBERDISCAP 2000, Madrid, Spain, pp. 187–190 (October).

Ferreiros, J., Colás, J., Macías-Guarasa, J., Ruiz, A., Pardo, J., 1998. Controlling a Hifi with a continuous speech understanding system. In: Proceedings of 5th International Conference on Spoken Language Processing, ICSLP98, Sidney, Australia, pp. 2871–2874 (December).

Gibbon, D., Moore, R., Winski, R., 1997. Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin, Germany.

Hirschman, L., Dahl, D., McKay, D., Norton, L., Linebarger, M., 1990. Beyond class A: a proposal for automatic evaluation of discourse. In: Proceedings of the DARPA Workshop on Speech and Natural Language, pp. 103–113.

Huang, C., Darwiche, A., 1996. Inference in belief networks: a procedural guide. International Journal of Approximate Reasoning 15 (3), 225–263.

Jing, Y., Pavlović, V., Rehg, J.M., 2008. Boosted Bayesian network classifiers. Machine Learning 73, 155–184.

Lee, C., Jung, S., Kim, K., Lee, D., Lee, G.G., 2010. Recent approaches to dialog management for spoken dialog systems. JCSE 4 (1), 1–22.

Levin, E., Pieraccini, R., 1997. A stochastic model of computer–human interaction for learning dialogue strategies. In: EUROSPEECH, pp. 1883–1886.

Meng, H.M., Lam, W., Wai, C., 1999. To believe is to understand. In: Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99), Budapest, Hungary, pp. 2015–2018 (September).

Meng, H.M., Wai, C., Pieraccini, R., 2003. The use of belief networks for mixed-initiative dialog modeling. IEEE Transactions on Speech and Audio Processing 11 (6), 757–773.

Mertins, I., Moore, R., 2000. Handbook of Multimodal and Spoken Dialogue Systems: Resources. Terminology and Product Evaluation. Kluwer Academic Publishers., Norwell, MA, USA.

Möller, S., Smeele, P., Boland, H., Krebber, J., 2007. Evaluating spoken dialogue systems according to de-facto standards: a case study. Computer, Speech and Language 21, 26–53.

Moreno, A., 1997. Speechdat Spanish Database for Fixed Telephone Networks. Corpus Design Technical Report SpeechDat Project LE2-4001, Universidad Politécnica de Cataluna.

Raux, A., Langner, B., Bohus, D., Black, A.W., Eskenazi, M., 2005. Let's go public! taking a spoken dialog system to the real world. In: Proc. of Interspeech 2005, Lisbon, Portugal, pp. 885–888.

Schulz, S., Donker, H., 2006. An user-centered development of an intuitive dialog control for speech-controlled music selection in cars. In: Proceedings Interspeech 2006 (ICSLP), Pittsburgh, USA, pp. 61–64.

Su, J., Zhang, H., 2006. Full Bayesian network classifiers. In: Proceedings of the 23rd International Conference on Machine Learning. ICML '06.. ACM, New York, NY, USA, pp. 897–904.

Turunen, M., Hakulinen, J., Kainulainen, A., 2006. Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: Similarities and differences. In: Proceedings Interspeech 2006 (ICSLP), Pittsburgh, USA, pp. 1057–1060.

Walker, M.A., Kamm, C.A., Litman, D.J., 2000. Towards developing general models of usability with paradise. Natural Language Engineering, Special Issue on Best Practice in Spoken Dialogue Systems 6, 363–377.

Walker, M.A., Litman, D.J., Kamm, C.A., Kamm, A.A., Abella, A., 1997. Paradise: a framework for evaluating spoken dialogue agents. In: Proceedings of the ACL/EACL 35th Meeting of the Association for Computational Linguistics. pp. 271–280.

Williams, J.D., Young, S., 2007. Partially observable markov decision processes for spoken dialog systems. Computer Speech and Language 21 (2), 393–422.