

Using artificial intelligence to make interactive television more usable by people with special needs

[Aplicación de la inteligencia artificial para hacer la televisión interactiva más usable por las personas con necesidades especiales]

Luigi Ceccaroni¹, Josefa Z. Hernández², Elisa Martínez³, Paloma Martínez⁴, Xavier Verdaguer¹

¹ TMT Factory, Torre Mapfre, planta 28 A, Marina 16-18, 08005 Barcelona, Spain,

luigi@tmtfactory.com, xavier@tmtfactory.com

² Artificial Intelligence Department, Technical University of Madrid, Campus Montegancedo s/n,

28660 Boadilla del Monte (Madrid), Spain, phernan@dia.fi.upm.es

³ Departament de Comunicacions i Teoria del Senyal, Enginyeria La Salle, Pge. Bonanova, 8, 08022 Barcelona,

Spain, elisa@salleurl.edu

⁴ Computer Science Department, Universidad Carlos III de Madrid, Avda. Universidad 30,

28911 Leganés (Madrid), Spain, pmf@inf.uc3m.es

Abstract. A research project, in the framework of the Spanish National Program of Technologies of Services of the Information Society, has been undertaken, whose object was the acquisition and application of new knowledge and techniques that can turn useful to contribute to considerably improve the domain of interactive television. The project, IntegraTV-4all, has been an effort towards a new television for all and to promote the access of people with special needs to the new technologies, with a development adapted and oriented to their possibilities and necessities, which includes both graphical and natural-language interfaces. This paper characterizes the general approach of the project and the artificial-intelligence techniques used, including a dialogue management for the natural-language interface and a virtual narrator in the domain of interactive systems of digital-entertainment consumption.

[Resumen. En el marco del Programa Nacional de Tecnologías de Servicios de la Sociedad de la Información, se ha emprendido un proyecto de investigación cuyo objeto era la adquisición de nuevo conocimiento y el uso de técnicas de inteligencia artificial para contribuir a mejorar las aplicaciones en el dominio de la televisión interactiva. Los resultados del proyecto, IntegraTV-4all, constituyen una contribución hacia una nueva televisión para todos y hacia la promoción del acceso de las personas con necesidades especiales a las nuevas tecnologías. El desarrollo informático, adaptado y orientado a diferentes tipos de necesidades, incluye interfaces gráficas y en lenguaje natural. Se caracterizan las técnicas de inteligencia artificial usadas, que incluyen la gestión del diálogo para la interfaz de lenguaje natural y un narrador virtual en el dominio de los sistemas interactivos de consumo de entretenimiento digital.]

1. Introduction

A model that supports the organization and dynamic presentation of multimedia content from a combination of live television programming, pre-recorded content, Internet resources and other services at each interactive system of digital-entertainment consumption (ISDEC)¹ is needed [Chorianopoulos, 2003]. Unlike classic television, ISDECs benefit most from thinking of them in terms of bits. Once in the machine, there is no need to view programs in the order they were sent [Negroponte, 1995]. Some kind of logic, either from the user or from some other source, can be applied on the multimedia content, with several scenarios as result. A few years from now, the 300-plus channels and the pre-recorded content we have now could evolve into one: MyTV, the channel you program yourself [Rose, 2003]. According to

¹ With the term ISDEC, the following concepts and technologies are represented: standalone personal video recorder (PVR), hybrid digital-recording device, home media server (HMS), server-based PVR (offered by some video-on-demand companies), PVR-integrated set-top box (STB).

Chorianopoulos et al. [2003], neither the vision of 300-plus channels, nor the vision of a single personalized channel is suitable. They propose, and we agree, a number of personalized virtual channels offering enough options in media experiences, while simplifying the choice from vast and diversified sources of media content. We here classify both MyTV and virtual channels approaches as instances of a personal guide of television programming (PGTP) super-class.

1.1 Interactive systems of digital-entertainment consumption

ISDECs offer high computational power and large memory. They can support advanced functionalities (e.g., user modeling, personalization, and speech-recognition) and store large amounts of multimedia content [Ardissono, 2001]. TMT Factory's **IntegraTV**, the ISDEC on which the project described in this paper is based, is a service of interactive television for hotels (see Figure 1) that offers the following features: digital films, video-games, music, management of digital photos, a productivity suite, Internet access, local guide of touristic offer, and typical hotel services.



Figure 1. *IntegraTV standard version's interface*

1.2 Autonomous-agents technology

Among autonomous-agents technology and principles that can be applied in ISDECs, there are three main capabilities:

- (1) *perception*: the ability to recognize:
 - a. the surrounding environment;
 - b. which content the user is paying attention to;
 - c. what is in the hard disc;
 - d. what is on the electronic programming guide (EPG);
 - e. what is on the PGTP;
 - f. what the user is saying;
- (2) *action*: the ability to respond to perceived sensation, to change one's own state or the state of the environment; many actions are usually available; common actions include all sort of data manipulation;
- (3) *cognition*: the ability to reason, including selecting from among the actions that are possible in response to perception; reasoning is a complex process that can include the ability to experiment and learn from the effects of the actions selected; cognition includes natural-language processing and comprehension, and dialogue management.

1.3 Speech-recognition in ISDECs

Apart from TMT Factory, at least two other companies, OneVideo Technology Corporation and Agile TV, are developing speech-recognition services that will let viewers change channels with voice prompts. Users speak into a microphone placed on a remote control, a set-top box or a headset. The services can recognize verbs like *find*, *scan* and *record*, topics like *sports* and *movies*, and the names of movie stars. Agile's service, called Promptu, also recognizes about 15 regional accents, and both systems claim to

filter out extraneous noise. The brains of Promptu are at the distribution hubs run by cable companies. This means consumers need two-way communications with their television provider, something satellite providers do not have. Unlike in Promptu, in OneVideo's service, OneListener, the software is installed locally, making it possible for satellite or phone companies to offer the service, too.

With hundreds of channels and thousands of hours of video now available on most cable and satellite systems, providers are working with an array of PGTP developers. But most require remote controls packed with buttons or menu screens that can create more detours than pathways to favorite shows. Current tools are not sufficient for easy-to-find and easy-to-navigate searches. No company has committed yet to offering speech-recognition services to its customers, but several companies say they could sell or give away the services to win and retain subscribers. Customers with poor eyesight or other disabilities may also be targeted.

1.4 Interaction management in natural-language dialogues

The interaction management of the natural-language interface used in IntegraTV-4all has as main objective “to achieve flexible and coherent human-computer interaction”. With this purpose, the involved techniques are dialogue models, intentional processing and language technologies.

The approach follows the dialogue model called the *thread model* [Garcia-Serrano, 2002] that makes use of the *common ground* concept [Clark, 1996] for attaining coherent and fluent dialogues within an interaction. When two speakers converse, they cannot possibly exchange all of the information necessary to ensure that their utterances are understood as intended. Instead, speakers assume that they share some *common ground* with their hearers. Clark and Schaefer [1989] define common ground as the propositions whose truth the speaker takes for granted as part of the background of the conversation. As a conversation progresses, speakers presuppose the propositions which were conveyed in previous utterances, adding to the common ground. Thus the net effect of a conversation is to increase the amount of information that the speakers share.

The *thread model* is also based on another theory according to which discourse structure is composed of three separate but interrelated components:

- (1) the structure of the sequence of utterances (called the *linguistic structure*);
- (2) a structure of purposes (called the *intentional structure*);
- (3) the state of focus of attention (called the *attentional state*) [Grosz, 1986].

The *interaction* is conventionally considered the top level discourse unit, in spoken conversation. The *sequence* is a bloc of exchanges (see below) linked together with a high degree of semantic or pragmatic coherence and with relatively consistent participation by the speakers. A new sequence is supposed to have been initiated once a breakdown in semantic coherence, pragmatic coherence or speakers' participation causes the end of the previous sequence. In practice, deciding where one sequence ends and another begins can be problematic, but it is often possible to spot three types of sequence within an interaction:

- (1) an opening sequence which sets up the interaction (*greeting*);
- (2) one or more sequences with a transactional function;
- (3) a closing sequence (*salutation*).

The *exchange* is the minimal unit of dialogue. It is conventional to distinguish three types of exchange:

- (1) Exchanges composed of only *one intervention* (or truncated exchanges). These occur when an intervention gives rise to no reaction, either verbal or non-verbal. Example:
L1: What can I do for you?
L2: [no reaction, either verbal or extra-verbal]

- (2) Exchanges composed of *two interventions*. This is the canonical scheme of the exchange. The first intervention is known as the *initiative* and the second as the *reaction* (or *answer*). Example:
 - L1: I'd like to know at what time the alarm is set to ring.
 - L2: The alarm is programmed for eight in the morning.
- (3) Exchanges composed of *three interventions*: an initiating intervention, a reactive intervention, and an evaluative intervention. Example:
 - L1: I'd like to know at what time the alarm is set to ring.
 - L2: The alarm is programmed for eight in the morning.
 - L1: Ah! Ok.

While the exchange is the basic unit of dialogue, the *intervention* is the basic unit produced by a single speaker, i.e. it is a *monologic unit*. An intervention may contain several *speech* (or *communicative*) *acts* [Austin, 1975] [Searle, 1969]. Example:

L1: Good morning. (*Greet*) What can I do for you? (*Authorize*)

Language is used for representing the world, but above all it allows speakers to carry out actions (to greet, to ask for authorization, to give orders, to make requests, to thank...). Communicative acts are specific acts produced by language. The communicative act can be considered the minimal unit of the speech/discourse grammar.

2. IntegraTV-4all

A consortium coordinated by TMT Factory and formed, besides, by the Ramon Llull University (La Salle Engineering and Architecture; User Lab), the Technical University of Madrid (Department of Artificial Intelligence) and the Carlos III University of Madrid (Department of Computer Science) carried out a project, IntegraTV-4all, which wants to extend interactive television in new directions, through the development, and integration into the IntegraTV ISDEC, of a new module that contributes to facilitate the personal autonomy and the social integration of groups such as, primarily, people with some sensorial impairment (blindness, visual deficiencies, deafness, impaired hearing ability, limitations of speech). Potentially, nevertheless, the results of the project, which counts on the aid of the Spanish Ministry of Industry, Tourism and Trade through an important grant under the PROFIT program and on the collaboration of the ONCE foundation and the ATLAS and Fundosa Teleservicios companies, could also be useful for people with some physical or psychic impairment as well as for the elderly.



Figure 2. *IntegraTV residential version's interface*

The project was subdivided in several, sequential phases, of which the ones most related to *accessibility to audiovisual means for people with special needs* are described in this section. In phase 2, a basic service of interactive television was implemented, which lets users navigate through the menus using their voice. More precisely, the output of phase 2 is characterized by the possibility for guests to activate and

use the system by voice, in such a way that they are able to take advantage of all the services of IntegraTV and to navigate through the system without the need of visual references, given that all the options on the screen and all the texts are presented by voice. During phase 3, described in detail later, a residential version of the service is developed (see Figure 2) and conversational capabilities are added to the system, within a limited domain. Table 1 shows a comparison of the main menu's content for the hotel and residential versions of the service.

Hotel version	Residential version
Movies	Movies
Games	Games and contests
Music, radio	Music, radio and photos
Photos	Internet
Internet	Productivity suite
Productivity suite	Video chat and telephone
City guide	Shopping
Guest services	Bank and accounting
TV channels	TV channels

Table 1. *Content of the main menu of the IntegraTV service*

2.1 Target users and service options

Unlike previous versions of IntegraTV, IntegraTV-4all is not only hotel-oriented, but is also thought for domestic users. It has to work for users with sensorial impairments, but it has also to be accessible for elder people and it has to be useful for any user without impairments (design for all). Target users are classified in the following groups:

1. people with visual deficiencies (including the blind);
2. people with hearing and speaking disorders (including the hard of hearing and the deaf);
3. people with access difficulties to the TV system and the elderly;
4. people without impairments.

However, the system is not apt for all the impairments described at the same time, and the previous segmentation of users helps to activate or deactivate the following options:

1. *Control by voice*: for users with visual deficiencies (see Figure 3). Data input is made by natural language, and all the options of the menus as well as all the texts and the answers of the system are spoken.
2. *Addition of sign language and subtitles*: for people with hearing and speaking disorders (see Figure 4). Data input is made with the remote control or the keyboard, and the system always presents the information with texts, and subtitles or sign language (integration with the virtual characters described later).
3. *Interface with interactive, virtual presenter*: for people with access difficulties to the TV system and the elderly. The interactive, virtual presenter assists users during navigation and is conceived as a likeable assistant which facilitates the use of the system.

2.2 User interface and functioning of the navigation system

The main menu of IntegraTV-4all includes the following options: Television - Movies - Music - Telephone - Alarm Clock - Assistant (see Figure 5). Each menu gives access to a series of tree-like sub-menus.

In one of its configurations (option 3 above), IntegraTV-4all's interface includes two types of virtual characters, which can, at times, be integrated between them:

1. *Realistic, virtual speaker (RVS)*. With the objective of improving the levels of accessibility and usability of the system, the La Salle Engineering and Architecture group developed a system of realistic, virtual locution with expressive speech, concretely, a new output interface based on a talking head of easy personalization. In order to reach this goal, the interface is equipped with the capability to show a realistic face, generated from previously recorded images, and to provide it with a natural behavior through synthetic, expressive speech along with the corresponding face expressions.
2. *Interactive, virtual presenter (IVP)*. TMT Factory is developing a virtual presenter (see Figure 5) which contributes to the replacement of the standard menu interface and appears on the screen with all the options of each section represented by icons located around it on an ellipse and with a certain depth. The IVP gives the user suggestions and assists him, when necessary, in the use of the system. Once an order is received from the user (the selection of an option), the IVP takes the corresponding icon and locates it in the foreground, while the non-chosen icons are replaced (using an animation) by the sub-options of the new section. It is important to consider, though, that the contribution of the IVP is more in the high-quality interface and friendliness, than in the capabilities of natural-language processing, which is carried out by other modules of the system.



Figure 3. *IntegraTV-4all's interface for users with visual deficiencies*



Figure 4. *IntegraTV-4all's interface for users with hearing and speaking disorders*

An example end-user scenario for the final demonstrator is as follows (see also Figure 6):

A person with reduced or no visual ability goes to a room of a hotel, opens the door and leaves his suitcases. Ten seconds after the door has been opened, the audio-guide's "Welcome to the hotel" message begins to play. The client listens to the greeting message and the description of her room. Once the description is finished, the system informs about the available options of dialogue and then waits activated, ready to begin a conversation. The client prefers to begin arranging her stuff. After a while, she decides to begin the conversation with the system, for instance to try and program the alarm clock.

3. Conversational capabilities

Although the interaction manager can be applied to any application domain, in IntegraTV-4all it is limited to the domain of the *alarm clock*, where it has conversational capabilities. Let us consider the following scene, continuation of the previous scenario, in which *J* is in her hotel room and is prepared to go downstairs to have lunch, but, before that, she decides to configure the alarm clock for the following morning (by means of the system for recognition and management of free language).

J: Computer ...morning!

IntegraTV-4all: Good morning. What can I do for you?

J: I'd like to know at what time the alarm is set to ring.

IntegraTV-4all: The alarm is programmed for eight in the morning.

J: I want to change the time.

IntegraTV-4all: Tell me to what new time you wish to program it.

J: To a quarter past nine.

IntegraTV-4all: The alarm has been set to a quarter past nine in the morning.

J: Good, thank you very much.

IntegraTV-4all: So long.

J: So long.

The program that generates the interventions of IntegraTV-4all in answer to *J*'s queries handles the dialogue as the combination of discourse threads. At any given time during the interaction, only one of these threads is developed. Threads' main features are, among others, their unique purpose (or *intention*, as stated by Grice [1975]), their state, their context space and their *commitment* (see below). All this information belongs to the common ground and is shared between user and system (the speakers).

It is necessary to assure a certain level of accuracy for the shared conjectures held in the common ground and this is obtained through the *thread commitment*, whose basis is, again, found in the theories of the common ground and the joint action. It may happen that certain events during the thread development affect the commitment (weakening or reinforcing it); if the commitment falls too low, the system tries reinforcing it, thus assuring the progress of the interaction.

The order in which the threads are developed is not fixed, but dynamically arranged by both speakers during the interaction. However, for a thread to be developed, both participants in the interaction have to be focusing on that thread. Hence, in the frame of Grosz and Sidner's theory [1986], an attentional structure is needed to keep track of speakers' focus and thread development. The reinforcement processes, sometimes carried out by the system, by avoiding focus losses and speeding up the interaction, also affect the attentional state.

A first version of this approach has been tested in a virtual assistant for e-commerce during the *ADVISE* European project [García-Serrano, 2004], and an enhanced version was implemented within the *VIP-Advisor* Project, concerning an assistant for risk management [Hernández, 2004].

3.1 Interaction processing

Any user-intervention is formed by one or more communicative acts. When the user utters her intervention, she could be aimed to develop an existing thread or to introduce a new one (*initiative*). During the interaction analysis, every utterance found is translated into actions (communicative acts). The interaction is segmented into independent sequences and these ones are segmented into exchanges. (An example of segmentation is shown in Figure 7.) Then, a set of threads, with their ordered sets of actions, is identified. The thread model considers the dialogue as a combined activity in which the speakers introduce and share goals that are developed during the interaction. For that, both participants advance in the dialogue according to certain rules (such as, for example, that after a question it is possible and probable that an answer appears). The set of possible steps (interventions), in any given moment, is different, based on the state in which the development of a specific goal (formalized by means of a thread) is.

This form of managing the development of each thread takes the name of *dialogue games*, and can be carried out, for example, using nondeterministic finite-states automata. The system considers a series of possible states and the possible transitions among them that can be made during the conversation. Transitions among states can be caused as much by the dialogue manager as by the interventions of the user. More precisely, it is the analysis of a communicative act that makes an automaton evolve towards another state.

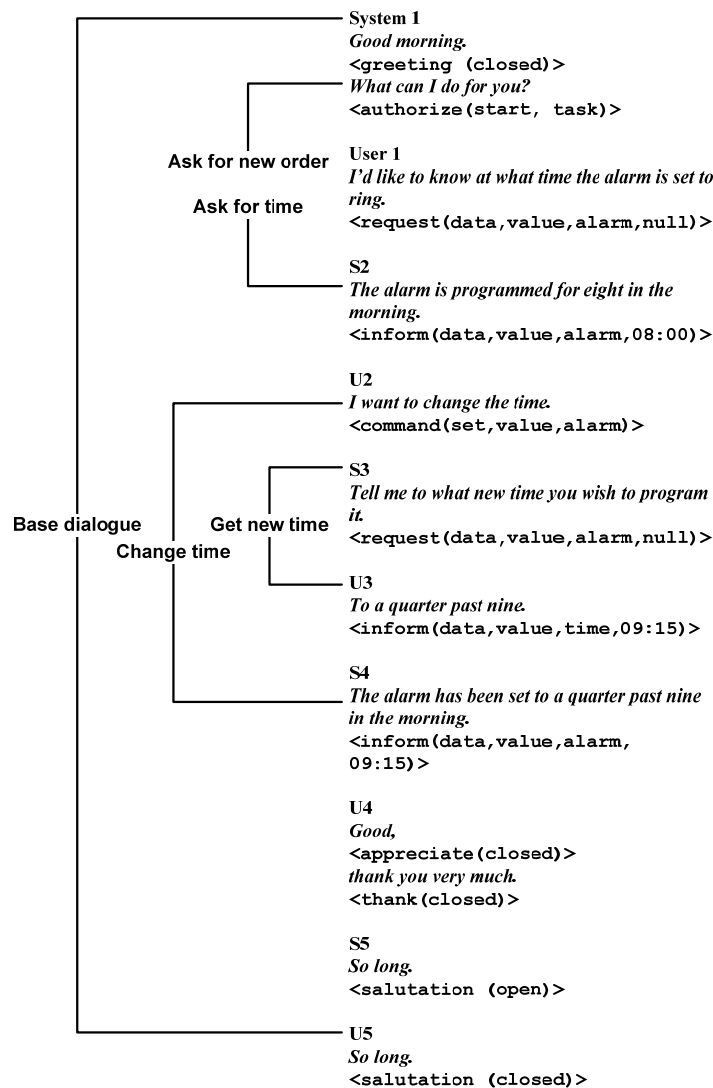


Figure 7. Example of dialogue segmentation into exchanges

With this model it is possible to maintain different threads open during a conversation. Each thread is introduced to reach a different goal. These threads are organized hierarchically, being the goal of the base thread the development of the very conversation between the IntegraTV-4all system and the user.

4. Conclusions

The IntegraTV-4all project applied advanced artificial-intelligence techniques, such as natural-language processing, to achieve adapted services of leisure and information through the television system, primarily in especially accessible hotels. These services are characterized by advanced visual and speech interfaces to facilitate the stay to guests with sensorial disabilities. These services can contribute to facilitate the personal autonomy and the social integration of groups such as, above all, people with some sensorial impairment. Potentially, nevertheless, the results of the project could also be useful for people with some physical or psychic impairment as well as for the elderly.

Acknowledgements

This work was conducted as a part of the IntegraTV-4all project. Various people at TMT Factory did a portion of the work on the prototype interface. This work was funded in part by grant FIT-350301-2004-2 from the Spanish Ministry of Industry, Tourism and Trade, through the PROFIT program. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect those of the Spanish Ministry of Industry, Tourism and Trade.

References

- [Ardissono, 2001] Ardissono, L., Portis, F., Torasso, P., Bellifemine, F., Chiarotto, A. and Difino, A. Architecture of a system for the generation of personalized Electronic Program Guides. In Proc. UM2001 Workshop on Personalization in Future TV (TV'01), Sonthofen, Germany, 2001.
- [Austin, 1975] Austin, J. L. How to do things with words. Harvard University Press; 2nd edition, 1975.
- [Chorianopoulos, 2003] Chorianopoulos, K., Lekakos, G. and Spinellis, D. The virtual channel model for personalized television. In Proc. 1st European Conference on Interactive Television, 2003.
- [Clark, 1996] Clark, H. H. Using Language. Cambridge Univ. Press, 1996.
- [Clark, 1989] Clark, H. H. and Schaefer, E. F. Contributing to Discourse. *Cognitive Science*, 13(2), 259-294, 1989.
- [Garcia-Serrano, 2002] Garcia-Serrano, A. and Calle, J. A Cognitive Architecture for the Design of an Interaction Agent. In Klusch, M., Ossowski, S. and Shehory, O. (eds.) Cooperative Information Agents, LNAI 2446. Springer, 2002.
- [García-Serrano, 2004] García-Serrano, A., Martínez, P. and Hernández, J. Z. Using AI techniques to support advanced interaction capabilities. In *Expert Systems with applications*, 26 (3): 413-426, 2004.
- [Grice, 1975] Grice, H. P. Studies in the way of words. Harvard University Press, 1975.
- [Grosz, 1986] Grosz, B. and Sidner C. Attention, Intentions, and the Structure of Discourse. In *Computational Linguistics* 12: 175-204, 1986.
- [Hernández, 2004] Hernández, J. Z. García-Serrano, A. and Calle-Gómez, J. Dialoguing with an Online Assistant in a Financial Domain: The VIP-Advisor Approach. In proceedings of AIAI 2004: 305-314, 2004.
- [Negroponte, 1995] Negroponte, N. Being digital. New York, Vintage Books, 1995.
- [Rose, 2003] Rose, F. The Fast-Forward, On-Demand, Network-Smashing Future of Television. *Wired*, 2003.
- [Searle, 1969] Searle, J. R. Speech Acts: an essay in the philosophy of language. Cambridge Univ. Press, 1969.