

MHP Interactive Applications: Combining Visual and Speech User Interaction Modes

Vanesa Lobato

Fundación CTIC, R&D Department
Parque Científico Tecnológico Gijón
33203, Asturias, Spain
+34 984291212

vanesa.lobato@fundacionctic.org

Gloria López

Fundación CTIC, R&D Department
Parque Científico Tecnológico Gijón
33203, Asturias, Spain
+34 984291212

gloria.lopez@fundacionctic.org

Víctor M. Peláez

Fundación CTIC, R&D Department
Parque Científico Tecnológico Gijón
33203, Asturias, Spain
+34 984291212

victor.pelaez@fundacionctic.org

ABSTRACT

This paper proposes an architecture for the development of interactive digital television systems accessible via voice. The proposed architecture is applicable to any interactive digital television application developed under the MHP standard, which runs on an interactive receiver equipped with return channel via Ethernet. The user needs only a mobile device with audio I/O, capable of running a speech synthesis and recognition system which follows the W3C standards SSML and SRGS respectively. This work includes the formalization of the architecture to meet this objective and validates it through two use cases in the scopes of eHealth and Digital homes.

Categories and Subject Descriptors

H5.2 [Information interfaces and presentation]: User Interfaces - Graphical user interfaces (GUI), Input devices and strategies, Interaction styles, Voice I/O.

General Terms

Design, Human Factors

Keywords

DVB-T, Multimedia Home Platform (MHP), interactive digital television (iDTV), speech interaction, speech recognition (ASR), speech synthesis (TTS)

1. INTRODUCTION

Digital Terrestrial Television (DTT) has generated great interest related to accessibility of the audiovisual media because of multiple possibilities of DTT in terms of services and applications. Multimodal interfaces, customization and/or standardization [1] are key factors in the search for a universal design of interactive systems. This paper takes this approach, focusing on the use of standards as well as on multimodality, combining voice and graphic modes. It is important to highlight that voice interaction is an indispensable requirement for users with visual impairments, as well as one of the most natural interaction modes for the general public [2].

In 2005, the Spanish Technical Forum of Digital Television published a list of requirements for digital television accessibility [3], which includes the requirements for interactive systems and voice-based interfaces. Nevertheless, there are difficulties in

putting these requirements into practice due to a lack of standardization and hardware constraints.

Standardization is a key element to boost iDTV services as evidenced by the Multimedia Home Platform¹ (MHP) initiative of the Digital Video Broadcasting (DVB) consortium, widely accepted in Europe.

Hardware limitations are primarily due to the fact that receivers (or set-top boxes, STB) tend to have the minimum hardware features to meet their specific purposes. These features are generally insufficient for voice software requirements. Moreover, STBs rarely have an audio input, which is an essential requirement in automatic speech recognition (ASR). Aware of these limitations, in 2003 *the National Centre for Accessible Media* (NCAM) published a guide to create spoken menus for STBs and DVDs², destined to both software developers and hardware manufactures. This guide includes information not only about hardware constraints, but also about the advantages of using Text-To-Speech (TTS) technology compared to pre-recorded voices in terms of flexibility and cost.

There are several publications that address the inclusion of a voice interface in DTT applications [4] [5] but these are usually limited to the scope of electronic program guide (EPG) navigation. There are also some studies about the quality of the ASR in iTV [6], which propose different strategies for error recovery in the ASR. Some studies into accessibility are based on improving the graphical user interface, making it more suitable for people with visual impairments [7], while others deal with the humanization of the navigation experience [8] (e.g. using as remote control a mobile device). Further relevant initiatives are accessible Emplea-T [9] and IntegraTV-all [10]. Both of these address the inclusion of voice interaction in an application that has been operating in a real environment. The first one includes speech output to an iTV application, although it uses pre-recorded messages instead of TTS technology; the second one uses ASR and TTS on proprietary technology and specific hardware.

All the works described above have a design of specific prototypes for particular applications, depending in many cases on specific hardware such as a Media Centre or a specific STB.

¹ <http://www.mhp.org/>

² <http://ncam.wgbh.org/resources/talkingmenus/>

The search for a global architecture should first address the hardware limitations discussed above. One option is to manufacture a custom-built digital TV receiver, which would cover the processing needs of the system. Another alternative is to use an additional hardware device with audio I/O, connected to the digital receiver and capable of running the voice software. Because of the first option is a time-consuming approach to obtain a specific product which would probably not be compatible with the products of other manufacturers, the work presented here is based on the second alternative. This approach has advantages in terms of costs and interoperability.

The main objective to achieve in this work is the inclusion of voice interaction in interactive applications build with MHP. The resulting applications will allow people with visual disabilities to interact with MHP applications in their own homes.

2. SYSTEM ARCHITECTURE

This paper proposes a distributed architecture with two basic blocks (see Figure 1): the MHP interactive application, running on the STB; and the voice control module, running in an external and portable device with audio I/O (e.g. a PDA). Both functional blocks are connected through a network interface. Speech recognition and synthesis systems will be implemented on the mobile device, using its own audio inputs and outputs. This implies that the communication between the voice control module and the interactive application will not be used to transmit audio signals. Instead, the text for the synthesizer, the grammars that govern the process of recognition or the recognized text are the data to be transferred.

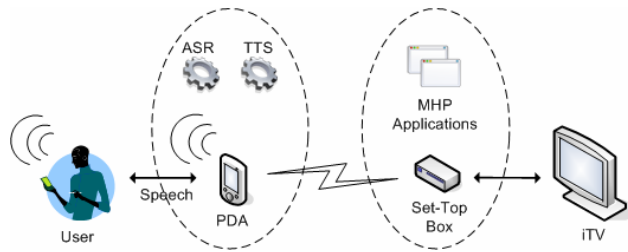


Figure 1. Global system architecture.

One of the main advantages of the proposed architecture is derived from the distribution of voice processing in a separate hardware device (e.g. a PDA). This feature makes it possible to combine a fully reusable voice module with any application that has network interface, regardless of the application scope.

Thus, three main steps have to be taken into account to design a MHP application with voice interface and based on the approach described above:

1. Development of the interactive digital television application using the MHP standard
2. Connection of the interactive application with the voice control module
3. Integration of the voice interface with the graphical interface, usual in this kind of applications, and synchronization of the two interaction modes.

2.1 Multimodal Interactive Applications

Once the architecture that specifies the hardware elements involved and their responsibilities have been defined, it is necessary to establish a mechanism to integrate and synchronize multiple modes of interaction within the interactive application for DTT. The interactive MHP applications considered here, interaction with the user based on voice and visual elements, can be defined as multimodal applications according to the classic definition of a multimodal interface: “that system which processes two or more combined input modes” [11].

Besides the application logic and the modules for the different interaction modes, the overall architectures considered in multimodal systems [11] [12] provide a number of common modules. These modules are the multimodal integrator or fusion module, the dialogue manager and the response planner or fission module. Because the expected interaction mechanisms of the system presented here are considered as alternative and individual elements, we propose a union of the fusion module, the dialogue manager and the fission module in a single module that will assume the three tasks. This simplifies the necessary architecture and makes a more efficient and suitable deployment for a MHP environment.

In addition, the proposed architecture for the interactive applications isolates the user interface from the logic necessary to integrate and sync the interaction modes (see Figure 2). The architectural pattern called Model View Presenter (MVP) has been used, as its use provides independence between the interaction modes (*views*), and also perfect synchronization through the *presenters*.

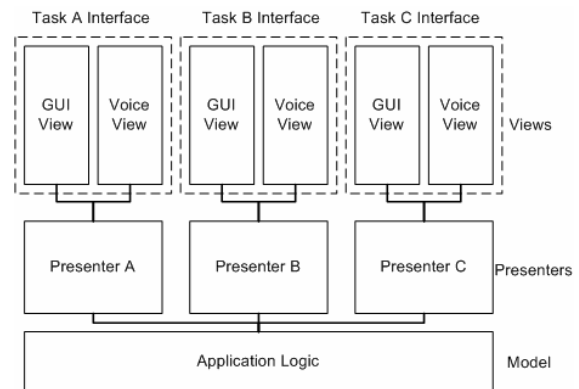


Figure 2. Interactive application architecture.

The *graphical views* (GUI) are the different interface screens, with the usual graphic components in this kind of applications, such as buttons, lists, texts, etc. The *voice views* contain the grammars for speech recognition with the commands that the user can give to the system and the messages to be synthesized. Both, graphical and speech views, translate the relevant events (button activations, reception of voice commands, etc.) into uniform actions that the presenter will trigger depending on the origin and the event type (e.g. dispatching response events to other views).

Following this pattern, events and user actions taken in different views will be transferred to the presenter, which will assure consistency and synchronization between all of the views. In this way, the presenter is fully responsible for the fusion, fission and dialogue management modules of a classic multimodal system.

2.2 Speech Control Module

In the proposed approach, speech technology lies in a common PDA which communicates with the interactive digital television application. Nevertheless, it is important to note that any hardware device with audio I/O, such a mobile phone or a tablet PC, can also be a valid option.

Speech technology has been selected with support for the standards JSGF³/SRGS⁴ and SSML⁵, to specify ASR grammars and TTS respectively. For this reason, it is possible to use any ASR-TTS engine that supports these standards.

The speech control module is based on a client-server architecture, where the client is the DTT interactive application and the server resides in the PDA. Through the exchange of information, the speech technology resident in the PDA is controlled from the STB according to the application state. This control includes updating of the active grammar in the voice recognition module (ASR), the definition of the text to synthesize and other options such as language selection in multi-language applications. On the server side, the PDA executes the operations of audio input/output associated with the ASR and TTS systems, based on information previously received from the STB. The voice commands which are identified by the ASR are sent instantly from the PDA to the STB, where they are interpreted.

Voice control on the PDA has been designed according to the method Push to Talk or 'press and talk,' because of its robustness against noise in the ASR process.

The interactive application development should be independent from the implementation details necessary to establish the communication with the synthesis and recognition services. To this end a component for MHP that encapsulates all the implementation and provides a simple interface with the necessary operations has been developed. This component is responsible for managing communications, starting the service in the PDA or the remote device, and translating the operations into the necessary network commands.

3. EXAMPLES OF USE CASES

To validate the architecture and design proposed we have taken two examples of interactive applications for DTT which verify the two basic requirements defined; that is, to be based on the MHP specification and to have an architecture based on the Model-View-Presenter pattern. The first example is an application for monitoring, recording and analyzing sleep quality in the home, developed in the scope of an European project⁶ (Figure 3). The second example focuses on the monitoring and control of digital homes (see Figure 4). In both cases, the speech interface not only

recognises the actions indicated by the user through voice commands, but also gives audio feedback to the user about ongoing actions (for example, the user might hear “*your data is being sent to your health centre*”).

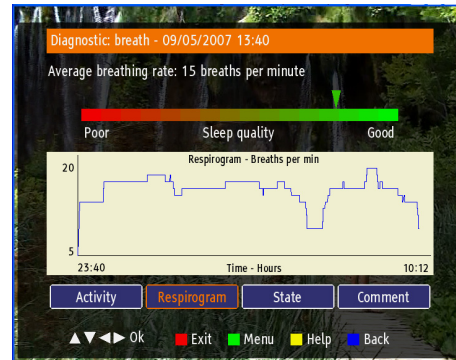


Figure 3. Application 1 (“Diagnostic results” screen).



Figure 4. Application 2 (“Digital elements” screen).

In both applications, the user can set the language in which the different elements and texts are visualized by pressing a key on the remote control unit. In addition, that functionality is available in voice-based user interaction. In this way, it is possible to change the language in which commands are recognized and texts are synthesized on-line, regardless of the speaker (user) or the characteristics of each type of voice (male/female). The languages used in the testing phase were Spanish and English.

Another useful feature in the ASR is to use grammars within semantic content, as in the SISR standard. In other words, the ASR module can return directly recognized voice commands (*Start*, *Stop*, etc.) or a semantic representation previously associated with those commands. Using grammars with a semantic content simplifies the processing of the information received from the ASR module in the interactive application. When using a grammar with multiple choices activated simultaneously in a multi-language environment (English, Spanish, etc.) there are two main options that can be considered:

1. Without semantic representation, the message from the PDA to the STB is the whole recognized chain (for example, ‘*Set Spanish language*’). In this case it is necessary to process the message in the interactive application, to first establish the type of command and then its associated value, if this exists.

³ <http://www.w3.org/TR/jsgf/>

⁴ <http://www.w3.org/TR/speech-grammar/>

⁵ <http://www.w3.org/TR/speech-synthesis/>

⁶ <http://www.estiiic.org> – SMITH Subproject

2. With semantic representation, each command is assigned a code (e.g. 1 to *Start*, 2 to *Set language*, etc.). Thus, when a voice command is recognized, the PDA sends a message consisting of a pair '*code [value]*' to the STB.

This second approach based on semantic information has been applied successfully in the developed prototypes. The initial tests carried out have demonstrated that the user can easily perform all the available functionalities by voice commands, receiving feedback via synthesized voice messages.

The results of the implementation evaluation show that the speech recognition accuracy is good for a low noise environment due to the use of the built-in PDA microphone which has short range sensitivity. Using a push-to-talk strategy to launch the speech recognition process simplifies the user interaction with the system because the user only has to touch the PDA screen to begin the process. Two limitations were found to affect the implementation of the system. Firstly, the recurrent use of the help command in order to remember the available voice commands (this is not a dialogue based system). Secondly, the complex interactions necessary to control some graphical widgets like selection lists. The latter limitation indicates that although the proposed architecture allows voice interaction to be included in existing applications, in order to improve the user experience it would be advisable to redesign some of the graphical user interfaces, or at least to consider different interaction schemes for the voice mode. This topic will be addressed in future work.

4. CONCLUSIONS AND FINAL REMARKS

In this work an architecture for the development of interactive digital television systems with speech interface, according to the MHP standard is proposed. This architecture allows voice interaction to be included in any interactive application that follows an architectural pattern similar to the Model View-Presenter and runs on a digital television receiver equipped with a return channel via Ethernet.

The hardware alternative proposed on this paper presents several advantages in comparison to others based, for instance, on the purchase of a STB built ad-hoc. In this sense, the approach presented here allows the use of any common STB which is MHP compliant to run the interactive application, and any external device with I/O audio to execute the TTS and ASR modules.

There is no doubt about the multiple advantages of multi-modal interface from the user experience point of view. In this sense, it is important to highlight the total synchronization between the different interaction modes available in the interactive applications used to validate the proposed design, both input (graphic/touch and voice), and output (graphic and voice). This synchronization means that any user interaction with the system (visual or voice based) implies an update of all the views whenever necessary.

Future work includes work on the multimodal interaction, combining all the modes simultaneously, validating the proposed architecture with more use cases and involving more users during the testing phase in order to identify new functionalities and requirements. This future work will contribute to improved versatility and potential of the global system and, of course, to a better user experience.

5. REFERENCES

- [1] Gill, J. M. and Perea, S. A. 2003. Accessible Universal Design of Interactive Digital Television. In Proceedings of the 1st European Conference on Digital Television (Brighton, UK, April 2-4, 2003). EuroITV'03.
- [2] Neto, J. and Farinazzo, V. 2008. Non-Functional Requirements to Voice User Interface on Interactive Television: an Initial Study. Proceedings of the International Conference Interfaces and Human Computer Interaction (Amsterdam, The Netherlands, July 25-27, 2008). IADIS'08.
- [3] Foro Técnico de la TV Digital. Accesibilidad en Televisión Digital para personas con discapacidad. Octubre, 2005.
- [4] Wittenburg, K., Lanning, T., Schwenke, D., Shubin, H. and Vetro, A. 2006. The Prospects for Unrestricted Speech Input for TV Content Search. In Proceedings of the working conference on Advanced Visual Interfaces (Venezia, Italy, May 23-26, 2006). AVI'06. ACM, pp. 352-359.
- [5] Ibrahim, A. and Johansson, P. Multimodal Dialogue Systems for Interactive TV Applications. In Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (Pittsburgh (PA), U.S.A., October 7-10, 2002), pp. 117-222.
- [6] Berglund, A. and Qvarfordt, P. 2003. Error Resolution Strategies for Interactive Television Speech Interfaces. In Proceedings of the IFIP TC13 International Conference on Human-Computer (Zurich, Switzerland, September 1-5, 2003). INTERACT'03. IOS Press, pp. 105-112.
- [7] Rice, M. and Fels, D. 2004. Low Vision and the Visual Interface for Interactive Television. In Proceedings of the 2nd European Conference on Interactive Television (Brighton, UK, 31 March – 2 April, 2004). EuroITV'04.
- [8] Cereijo, A., Sala, R., Ahmad, S. and Rahman, M. 2005. Beyond the remote control: Going the extra mile to enhance iTV access via Mobile devices & humanizing navigation experience for those with special needs. . In Proceedings of the 3rd European Conference on Interactive Television (Aalborg, Denmark, 30 March – 1 April, 2005). EuroITV'05.
- [9] Martín, C.B., Merchán, J.M., Jiménez, D., Menéndez, J.M. and Cisneros, G. 2007. Accesibilidad a la Televisión Digital Interactiva. In Proceedings of the 2º Congreso de Accesibilidad a los medios audiovisuales para personas con discapacidad (Granada, Spain, June 21-22, 2007). AMADIS'07. Real Patronato sobre Discapacidad, pp. 67-77.
- [10] Ceccaroni, L., Martínez, P., Hernández, J.Z. and Verdaguer, X. 2005. IntegraTV-4all: an Interactive television for all. In Proceedings of the 1st Symposium on Ubiquitous Computing and Ambient Intelligence (Granada, Spain, September 14-16, 2005). UCAmI'05. In J. Bravo, J. Alamán and T. Riesgo (Ed.).
- [11] Oviatt, S. 2002. Multimodal interfaces. Handbook of Human-computer Interaction, J. Jacko y A. Sears (Ed.), Lawrence Erlbaum, New Jersey.
- [12] Trung, H. 2006. Multimodal Dialogue Management – State of the art. Human Media Interaction Department, University of Twente.