Enhancing Accessibility through Speech Technologies on AAL Telemedicine Services for iTV

Héctor Delgado, Aitor Rodriguez, Antoni Gurguí, Enric Martí, Javier Serrano, and Jordi Carrabina

Center for Ambient Intelligence and Accessibility of Catalonia,
Campus UAB, 08193 Bellaterra, Spain
{hector.delgado,aitor.rodriguez,antoni.guirgui,enric.marti,javier.serrano,jordi.carrabina}@uab.cat
http://centresderecerca.uab.cat/caiac

Abstract. Ambient Assisted Living Technologies are providing sustainable and affordable solutions for the independent living of senior citizens. In this scenario, telemedicine systems enhance distance patient's health care through interactive audiovisual media at home. Today TV is becoming the main connected device at home. However, Interactive TV applications must be fully adapted, particularly to the available input device: the Remote Control (RC). Despite this adaptation, some tasks are still uncomfortable due to the RC limitations. Therefore, more user-friendly input modalities are strongly desired. Spoken language allows distance hands- and eyes-free operation within the room, providing an intuitive and natural interface. This paper presents some accessibility facilities based on speech technologies for an interactive TV telemedicine service. The specific layout for TV environments, the help of an avatar and the voice navigation will enhance the user access, while the speech-based creation of medical reports reduces dramatically the time physicians need to write reports.

Keywords: telemedicine, accessibility, multimodal interaction, natural user interfaces, interactive TV, speech technologies, avatar

1 Introduction

Ambient Assisted Living (AAL) aims to produce technological support to help the aged to live independently, extending the time they stay in the environment where they are used to live. Telemedicine, understood as the deployment of telecommunication technologies to provide distance health care and health information to patients, can be extremely beneficial to provide close interactions between patients and experts. It is especially significant for people living in remote areas, people in dependant situations and even for senior citizens requiring an intensive use of medical assistance and careful monitoring [6]. Given the growth of Internet technologies, a number of Web applications to access to

different health services have been developed to be used in home environments. However, accessing to these services usually depends on the availability of a computer or a specific device to connect to the health system. The rise of connected devices in people's daily lives such as smartphones, set-top boxes (STB) for digital TV, HbbTV and tablet PCs has generated a great variety of multimedia systems with Internet capabilities giving to the end-user (i.e. the patient) the possibility to access the health system anywhere, anytime. In the home environment, in spite of the rise of other connected devices, the TV is still the most used device for the consumption of multimedia contents and it is a realistic candidate to become, in a near future, the main platform to access to specific interactive services related to education, health and home automation [13]. In this context, the interactivity model of a TV environment must be taken into account to ensure proper access conditions for all users. People usually watch TV some meters away, interacting through a Remote Controller (RC) and sometimes in company. This implies, for instance, that the components and fonts of the interactive applications must be sized large enough for a comfortable readability and the navigation must be adapted to the RC instead of a pointer on the screen. In spite of these adaptations, some tasks such as introducing user data through the RC might be considerably tedious, frustrating and inefficient. Thus, the need of other input methods becomes practically indispensable for applications with a higher degree of complexity. The use of multimodal and more natural interfaces will facilitate the user navigation and, in the case of telemedicine systems, improve some medical procedures. Spoken language has several characteristics that make it a potential interaction method between user and computer. It allows distance hands- and eyes-free operation within the room, and provides an intuitive and natural interface. The use of speech technologies as user interface can result in very beneficial interactive TV applications. The user can give orders to the system using their voice, which will be obtained by the system through Automatic Speech Recognition (ASR). On the other hand, the system provides answers by means of speech synthesis.

This work is the result of a research and development project developed in collaboration with physicians to enhance the accessibility capabilities of a traditional telemedicine system. In this paper we present an interactive application for telemedicine in digital TV enriched with accessibility features based on speech technologies: (1) a spoken user interface that provides voice navigation and (2) a speech-assisted system for the creation of medical reports. In addition, we include the use of an avatar to provide the information to the user more effectively. The rest of the paper is organized as follows: section 2 describes the presented service, the main use cases and the system architecture; section 3 details the implementation issues for the used speech technologies and the avatar; finally, in section 4 we present the conclusions of our work.

2 System Overview

This section describes the AAL service, the use cases and the system architecture.

2.1 Service and Use Case Description

The implemented service is an interactive application for a digital TV environment that aims to enhance the accessibility in health and telemedicine systems. Here we refer to accessibility in a broad sense, as a set of measures that helps to improve access to everybody in general, mainly oriented to those people unfamiliar with technology (e.g usability enhancements). The service is connected to the Info 33 health system [2], which manages the clinical information of patients throughout their life. It allows an efficient monitoring of programs for prevention and health promotion through the clinical knowledge and a universal coding for clinical interventions. This information is certified by a health professional chosen by the user providing validity and reliability of the information recorded. The service presented in this work facilitates the access to the Info 33 system from a connected TV applying some accessibility techniques successfully used in traditional Web environments, such as the multimodal user interaction, the accessible design of web-based user interfaces based on the W3C Web Content Accessibility Guidelines (WCAG) [5], and the addition of an avatar for an assisted navigation and an ASR system for the user interaction. The service has been built as a Rich Internet Application (RIA) to improve its performance and the user experience [8].

In this scenario, two main use cases were defined: 1. An end-user accessing to his health information: A user is at home watching TV when he decides to consult his medical record and see if there is any automatic notification from the system (e.g. an appointment reminder or a vaccine alert). He starts the interactive application from the provided Media Center platform and an avatar welcomes him giving spoken information about the main navigation options. The user can browse the different sections in the service (e.g. personal data, clinical data, the record of vital signs, the lab test results, the medical reports and the automatic alerts) not only with the RC supplied with the platform, but also by saying the key words -giving orders- to navigate through the interactive interface. At any time, he can request help information about a specific section of the interface and the avatar provides it. 2. The professional user (e.g. the doctor) creating medical reports in a multimodal way: A doctor sees a patient at home to scan and diagnose his condition. The doctor accesses to the Info 33 health system through the generated TV interface identified as a doctor in order to be allowed to generate a new report. Thus, he selects the report section through his voice or the RC and opens a new report to edit it. During the exploration, the physician dictates the findings, which are being written semi-automatically (with confirmation and editing options) in a fully hands-free mode into the report. At all times, the report can also be modified manually by editing it directly with a connected keyboard.

2.2 System Architecture

The above described service has been built as a RIA into a custom Set-top Box (STB) platform, which will be provided to the system end-users. This platform

connects to the Info 33 server through an IP network (e.g. Internet, IPTV networks). It contains the client application and only requests to the server the dynamic XML data related to the logged user. The application is shown on the TV through an HDMI interface and allows a multimodal user interaction through the support of different user input devices that are automatically detected by the user interface to interact consequently. Figure 1 depicts this overall architecture.

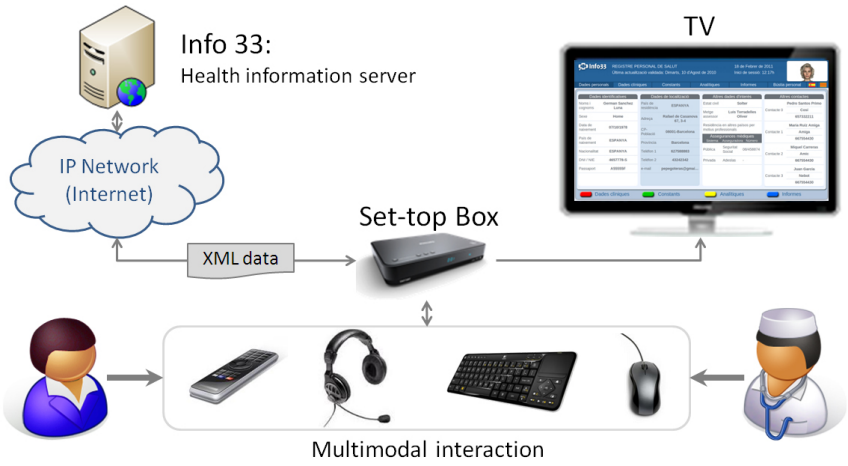

Fig. 1. Overall architecture of the presented telemedicine system.

Due to the implementation of the RIA at the client platform, the interaction with the server is limited to the XML interchange of the specific health data, leaving the rest of the tasks for the STB platform in the client side. The middleware, which includes the RIA, the ASR module, the speech synthesis module, the avatar engine, the security module and a Web browser, as well as the built-in services have been built on top of an Intel Dual-Core Atom N330 @1.6GHz and a Linux OS. The STB architecture is described in Figure 2.

3 Speech Technologies for Multimodal Interaction

In an environment of digital TV, where the availability of a keyboard as input device is not always assured (a RC is usually used), other multimodal interaction methods are strongly desired to improve user interaction and accessibility. Speech is a natural way of communication. Due to its characteristics, it can be remotely used in hands- and eyes-busy situations. Previous work has been carried out in the application of speech recognition in medical environments [9]. On one hand, spoken dialogues systems for health care and telemedicine might empower users to introduce basic information and vital sign health data in order to

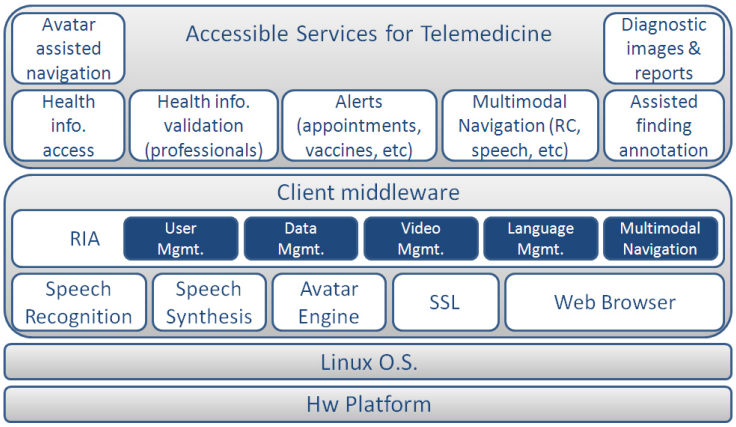

Fig. 2. The STB architecture of the presented telemedicine system.

perform self-monitoring tasks, or provide patients with useful information about appointments and other relevant issues. On the other hand, continuous speech recognition dictation systems make the creation of medical reports lighter to the professionals of medicine [12].

The project aims to offer users an intuitive and easy way of accessing and introducing medical data into the interactive TV system, by means of their voice. On the patient's side, voice commands may be used to browse the application in a natural fashion. On the physician's side, medical information can be input through their voice without the need of an external device while the patient is being explored, providing an effective and efficient method for the generation of medical reports. This way, doctors save a significant amount of time writing reports and can pay attention to the patient, whose experience and degree of comfort will be improved.

Unlike dictation systems in medical domain, the presented approach focuses only on the subset of the medical findings. Limiting the ambit of application to the fixed list of findings implies an increment in terms of accuracy, since the task is significantly simpler than modeling more complex forms of language. In this situation, the language can be modeled through finite state grammars rather than statistical N-grams. The system also provides error correction. When the user pronounces a finding, the application will notify of the recognized utterance by means of speech synthesis, thus there is no need to look at the screen to check the result. With a simple error correction method via voice commands, incorrect findings can be easily removed, as well as corrected by re-speaking. More detailed explanations about the findings list and the error correction method are given in section 3.3.

The speech-based functionality developed in this work consists of two different parts: first, a spoken user interface that provides voice navigation through the application, and second, a system for speech-assisted creation of medical re-

ports. In the subsequent subsections the training of the necessary infrastructure and development of the speech modules are explained in more detail.

3.1 ASR, Speech Synthesis and Avatar setup

The current subsection gives an overview on the preparation of the necessary tools that are used later for the development of the speech-based functionality. It comprises the ASR training, the speech synthesis and the avatar.

ASR training. The acoustic models have been trained using the SpeechCon Catalan speech corpus [11]. The corpus consists of spontaneous and read speech from 550 speakers, recorded with four microphones at different distances. Each utterance is stored in 4 independent (one per microphone) 16 bit, 16 kHz uncompressed audio files. The audio files are then parametrized into a 39-dimensional feature vector consisting on 12 cepstral coefficients plus the 0th coefficient, deltas and delta-deltas.

The acoustic models consist in a set of cross-word tied-state triphone Hidden Markov Models (HMM) derived from 40 monophones HMMs, covering the sound units of Catalan language. The models were trained according to the standard Maximum Likelihood approach. Finally, the models are refined by applying the Discriminative Training technique. The whole training process has been carried out using the HTK toolkit [14]. Further information about the training process can be found in [7].

Speech Synthesis. Text-To-Speech (TTS) is utilized to generate the system responses dynamically. It makes possible to check results in eyes-busy situations. This way, the physician does not have to advert the eyes from the patient. The Festival [3] software is used for this purpose, combined with the Festcat [4] package that contains Catalan synthetic voices for Festival.

The Avatar. To improve interaction, we have included an avatar inside the interface. Here, avatar refers to a two-dimensional video representing a person or a virtual character. The main purpose of the avatar is to inform the user about the different navigation options available for each scene. Furthermore, there are many compelling reasons to include an animated agent in the interface. On one hand, avatars have demonstrated to be an effective way to improve user's understanding of synthetic voices [10]. On the other hand, it improves user's natural perception of the interaction, as the user acts as if they were both listening and speaking to a person. At the same time, a secondary purpose is to facilitate the learning and use of the interface. From the patient's point of view, the avatar can be seen as a medical assistant that takes notes and informs the patient about their medical condition and not as intrusion between him and his physician. At the same time, patient's confidence is very important when we talk about confidential or private information, as the user needs to feel

comfortable to give or receive sensitive data and results. The avatar has been built off-line using AlterEgos [1]. This software generates the avatar's animation from a speech sound and the text speech files. For each scene, a text file with the scene presentation dialog is created. Using this file, a speech sound file is synthesized, using Festival. Then, the avatar facial animation is built, using both sound and text for lips synchronization. Afterward, the video is embedded in the scene. As the user navigates through scenes, the video is played accordingly.

3.2 Speech-based user interface

The speech-based user interface has been intended to facilitate access and browsing through the application, exploiting the naturalness of spoken communication as user input in ubiquitous systems.

The user can select the different options in the main menu through their voice. It allows a completely hands-free operation in order to navigate through the application. This method can be used by both end-users and professionals. The system implements a method for discarding the background speech based on 'universal' keywords. The system remains in 'wait' state, discarding speech input, until the universal keyword 'menu' followed by one of the possible menu entries is pronounced. Then, the best hypothesis is calculated. Next, a confidence measure is obtained to determine how likely the hypothesis is. If the confidence measure falls below an arbitrary threshold, the command is ignored and no action is performed. Otherwise, the menu entry corresponding to the hypothesis is accessed and shown in the graphic user interface.

3.3 Speech-based creation of medical reports

As said before, physicians spend a significant period of time writing medical reports. This fact makes patients feel uncomfortable, producing a feeling of impatience and inefficiency. The developed system offers the physician a way to generate the medical report while they are exploring the patient. It has important implications: doctors do not need to interrupt the exploration process, patients feel better treated, there is no dependency on other devices like a keyboard or RC, contact-time is optimized, etc.

Unlike conventional dictation systems in medial environment, the implemented module may be considered as a spoken dialogue system intended to introduce medical findings corresponding to the current appointment. The fact of considering only a subset of the medical language simplifies the task complexity noticeably. Although the system is not as flexible as a dictation system, the increment in accuracy is worth it, providing Word Error Rates (WER) next to 0%. The database of medical findings contains a list of 2144 sentences describing findings. Each one has a unique code within a whole categorization of findings.

The system consists of an ASR engine and a TTS module, as well as a simple dialogue manager that establishes the dialogue flow and deals with recognition errors and possible confirmations. Figure 3 depicts the dialogue flow. Once in new report mode, the system stays in 'wait' state until the keyword 'type' plus

a medical finding is pronounced. Then, the best hypothesis from the database of the codified medical findings and a confidence measure are estimated. If the confidence measure reaches a set threshold, the hypothesis is considered correct, the TTS module pronounces it and the finding is stored. Otherwise, the system asks for confirmation by means of the TTS module. If the proposed hypothesis is correct, the user says "yes" and the finding is stored. If incorrect, the user can say "no" and the transaction is canceled, or can correct it proceeding with the finding, in which case the flow goes back to the hypotheses computation state. In the event that an incorrect finding has been stored, the last entry can be deleted by saying 'remove'. The dialogue finished when the 'store' or 'cancel' report options are activated. In addition, the report can be edited at any time by means of the keyboard.

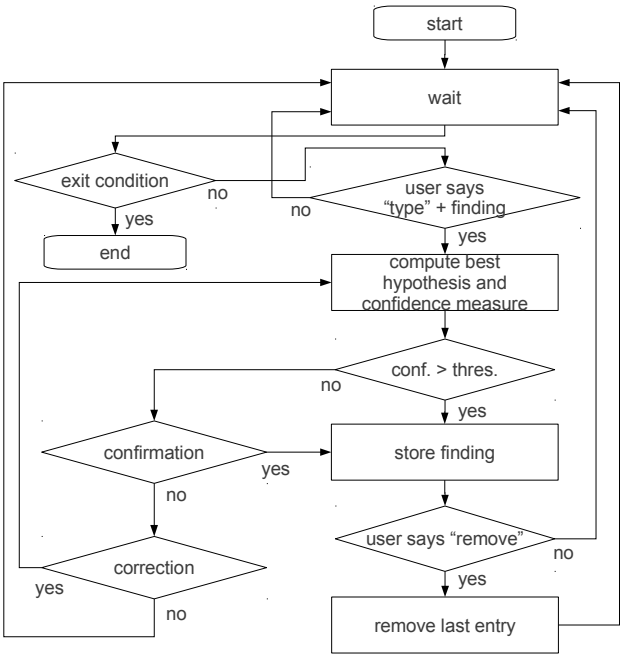

Fig. 3. Dialogue flow.

The speech recognizer has been tested in order to measure performance. A test corpus of 400 utterances from the whole set of 2144 sentences were recorded in quiet conditions. Due to the on-line constraints of the current application, the recognizer parameters have been tuned to achieve real time operation. In this situation, a word error rate of 0.7% has been obtained.

4 Conclusions

In the current scenario of connected interactive TVs, ambient intelligence and interactive applications anywhere and anytime, the accessibility issues for this kind of environments must be reviewed to ensure a proper access by everyone. This is especially important in health systems, which should be properly accessed by the largest possible number of users regardless of their age and condition. We have developed an interactive application for digital TV to access to the medical patient's record health system Info 33, in direct collaboration with physicians. We have applied speech recognition and synthesis techniques for voice navigation and medical report dictation, facilitating the data input in the TV environment. The use of a specific speech recognizer based on finite state grammars instead of dictation systems based on N-grams has been demonstrated to be effective in the task of inputting medical findings into the system. Furthermore, the presentation of an avatar as a navigation help is an effective way to improve the user empathy with the interface.

Acknowledgments. This work has been partly founded by the Spanish Ministerio de Industria, Project Reference AVANZA TSI-020302-200.

This article is supported by the Catalan Government Grant Agency Ref. 2009SGR700.

References

1. Alteregos v1.2, `http://www.alteregos.com`
2. Info 33 Health System, `http://www.mag.es`
3. Black, A., Taylor, P., Caley, R.: The Festival Speech Synthesis System (June 1999), `http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html`
4. Bonafonte, A., Aguilar, L., Esquerra, I., Oller, S., Moreno, A.: Generation of Language Resources for the Development of Speech Technologies in Catalan. In: Proceedings of the Language Resources and Evaluation Conference LREC 06. Genova, Italy (2006)
5. Caldwell, B., Cooper, M., Guarino Reid, L., Vanderheiden, G.: Web Content Accessibility Guidelines 2.0. W3C Working Draft (December 2007)
6. Cruz-Martín, E., del Árbol Pérez, L.P., Fernández González, L.C.: The teleassistance platform: an innovative technological solution in face of the ageing population problem. In: The 7th International Conference of the International Society for Gerontechnology (2008)
7. Delgado, H., Serrano, J., Carrabina, J.: Automatic Metadata Extraction from Spoken Content using Speech and Speaker Recognition Techniques. In: FALA2010. VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop. pp. 201–204. Vigo, Spain (2010)
8. Driver, M., Valdes, R., Phifer, G.: Rich Internet Applications are the next evolution of the Web. Tech. rep., Gartner (2005)
9. Jokinen, K., McTear, M.F.: Spoken Dialogue Systems. Morgan & Claypool Publishers (2009)

10. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature (December 1976)
11. Moreno, A., Febrer, A., Márquez, L.: Generation of Language Resources for the Development of Speech Technologies in Catalan. In: Proceedings of the Language Resources and Evaluation Conference LREC 06. Genova, Italy (2006)
12. Rosenthal, D.I., Chew, F., Dupuy, D.E., Kattapuram, S., Palmer, W.E., Yap, R.M., Levine, L.A.: Computer-Based Speech Recognition as a Replacement for Medical Transcription. American Journal of Roentgenology (1998)
13. Tsekleves, E., Cosmas, J., Aggoun, A., Loo, J.: Converged digital TV services: the role of middleware and future directions of interactive television. International Journal of Digital Multimedia Broadcasting (2009), http://eprints.mdx.ac.uk/7809/
14. Young, S.: ATK - An Application Toolkit for HTK, 1.4.1 ed. (June 2007)
15. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book Version 3.4. Cambridge University Press (2009)