This document is published in:

Lozano, José A. et al. (eds.), 2011. Advances in Artificial Intelligence: 14th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2011, La Laguna, Spain, November 7-11, 2011. Proceedings (Lecture Notes in Computer Science, 7023), Springer, pp. 393-402.

DOI: 10.1007/978-3-642-25274-7_40

The VoiceApp System: Speech Technologies to Access the Semantic Web

David Griol, José Manuel Molina, and Víctor Corrales

Universidad Carlos III de Madrid
28911, Leganés, Spain
dgriol@inf.uc3m.es, molina@ia.uc3m.es, 100048294@alumnos.uc3m.es

Abstract. Maximizing accessibility is not always the main objective in the design of web applications, specially if it is concerned with facilitating access for disabled people. In this paper we present the *VoiceApp* multimodal dialog system, which enables to access and browse Internet by means of speech. The system consists of several modules that provide different user experiences on the web. *Voice Dictionary* allows the multimodal access to the Wikipedia encyclopedia, *Voice Pronunciations* has been developed to facilitate the learning of new languages by means of games with words and images, whereas *Voice Browser* provides a fast and effective multimodal interface to the Google web search engine. All the applications in the system can be accessed multimodally using traditional graphic user interfaces such as keyboard and mouse, and/or by means of voice commands. Thus, the results are accessible also for motor-handicapped and visually impaired users and are easier to access by any user in small hand-held devices where graphical interfaces are in some cases difficult to employ.

Keywords: Dialog Systems, Multimodality, VoiceXML, XHTML+Voice, Web Interfaces, Speech Interaction.

1 Introduction

Continuous advances in the development of information technologies and the miniaturization of devices have made it possible to access information and web services from anywhere, at anytime and almost instantaneously through wireless connections. Devices such as PDAs and smartphones are widely used today to access the web, however the contents are accessible only through web browsers, which are operated by means of traditional graphical user interfaces (GUIs). This makes it difficult to use due to the reduced size of the screen and keyboards, and also makes them less usable by motor-handicapped and visually impaired users.

Multimodal interfaces go a step beyond GUIs by adding the possibility to communicate with the devices through other interaction modes such as speech. Multimodal dialog systems [1] can be defined as computer programs designed to emulate communication capabilities of a human being including several communication modalities. The usage of these systems provides three main benefits.

Firstly, they facilitate a more natural human-computer interaction, as it is carried out by means of a conversation in natural language. Secondly, multimodal interfaces make possible the use of these applications in environments in which the use of GUI interfaces is not effective, for example, in-car environments. Finally, these systems facilitate the access to the web for people with visual or motor disabilities, allowing their integration and the elimination of barriers to Internet access [2].

In literature there are two main approaches to develop multimodal dialog systems to access web contents and services. On the one hand, some authors have developed ad-hoc solutions focused on specific tasks, as e-commerce [3], chat functionalities [4], healthcare services [5], surveys [6], or recommendation systems [7]. On the other hand, it is possible to add a speech interface to an existing web browser [8]. This approach may acquire additional complexity in the case of Information Retrieval and Question Answering systems, such as in [9]. However, these works usually emphasize on the search of documents and not on the interaction with the user.

In this paper we describe the *VoiceApp* multimodal dialog system. The system has been developed as a common ground with different web-based applications that can be easily accessed by means of a sophisticated interface which merges voice with traditional GUIs. The idea behind it is to provide an easily extensible common place to create and evaluate multimodal interfaces for web applications. All the applications in *VoiceApp* are easily interoperable so that they are very useful to evaluate the potential of voice interaction in several domains, through a variety of resources and with different users. In the current implementation of the system, the dialog systems have been developed using the XHTML+Voice (X+V) language[1]. This language combines the visual modality offered by the XHTML language and the functionalities offered by the VoiceXML language[2] for the interaction by means of speech. One of the main objectives of the system is to adequately convey to users the logical structure and semantics of content in web documents, and provide them with easy ways to select which parts of a document to listen to.

We will describe the main three applications of *VoiceApp*, although up to five applications have already been implemented. *Voice Dictionary* receives from the user the search criteria and performs a search in the Wikipedia encyclopedia, collects and processes the result of the search, and communicates it to the user by means of visual modalities and synthesized speech. This application also allows to carry out a new search or select any of the links in the result page by using speech or keyboard and mouse. *Voice Browser* is a complete speech-based web search engine. This application collects the topic that the user wants to search on the Internet, communicates this information to the Google search engine, process the resulting information, and communicates it to the user. This application also facilitates multimodal access to the links included in the result

[1] http://www.w3.org/TR/xhtml+voice/

[2] http://www.w3.org/TR/voicexml20/

of the search. Finally, *Voice Pronunciations* includes different multimedia games designed for learning foreign languages.

2 Extending Web with Voice

Hyper Text Markup Language (HTML) is the language popularly used for marking up information on the World Wide Web so that it can be displayed using commercially available browsers. However, the World Wide Web Consortium (W3C) realized that HTML was a weak mark-up language if a user wants to process web pages further, given that the use of this language to automatically infer any kind of semantic information requires the analysis of the contents of the web page. This way, the eXtensible Markup Language (XML) was developed as a solution to correct the limitation of the use of information that is available on the web by marking-up information in the web pages and also allowing developers to define their own tags with well-defined meanings that others can understand. The use of an XML-based language significantly improves the portability and interoperability of the programmable parts (including data and programs) of the service.

Many XML-based languages are currently standardized to specify various services (for instance; WSDL for Web Services, ebXML for electric commerce, or CPL for VoIP). VoiceXML is one of these significant standards, as far it makes the Internet accessible through speech, using well-defined semantics that preserves the author's intent regarding the behavior of interactions with the user and facilitating the access to the net in new devices and environments (thus making XML documents universally accessible). VoiceXML audio dialogs feature synthesized speech, digitized audio, recognition of spoken and DTMF key input (Dual-tone multi-frequency signaling), recording of spoken input, telephony, and mixed initiative conversations. The standard also enables the integration of voice services with data services using the client-server paradigm. In addition, many VoiceXML platforms are currently available for research and business use purposes (e.g., Voxeo[3]).

3 The VoiceApp Multimodal Dialog System

The *VoiceApp* system consists of a set of X+V documents. Some of them are stored from the beginning in the server of the application, while others are dynamically generated using PHP and JavaScript. This dynamic generation takes into account the information extracted from different web servers and MySQL databases in the system, and a set of users preferences and characteristics (e.g., sex, preferred language for the interaction, number of previous interactions with the system, and preferred application). Previous interactions of the users are also taken into account to adapt the system, considering users' most used application, recent topics searched using the application, or errors detected after each interaction with the system.

[3] `http://evolution.voxeo.com/`

In order to interact with the X+V documents that make up the system, a web search engine supporting speech interaction and the specifications of this language is required. There are different models for implementing this multimodal interaction on mobile devices. The fat client model employs embedded speech recognition on the specific device and allows conducting speech processing locally. The thin client model involves speech processing on a portal server and is suitable for mobile phones. The implementation of the *VoiceApp* multimodal application for both computers and mobile devices is based on the fat client model, including a multimodal browser and embedded speech recognition on the corresponding device, and a web application server in which the system is stored.

The Opera browser[4], which allows multimodal web navigation by means of speech, has been integrated for the interaction with the system using a computer. This way, users only need to connect to the application using Opera Voice in a computer with a functioning sound card and loudspeakers or headphones. Opera Voice allows the control of the Opera's interface by talking to the browser. Any ordinary browser command can be done by voice, such as refreshing a web page, navigating to and following the next link in a document, going to the next slide in an Opera Show presentation, or logging on to a password protected Web site. The voice modules that Opera downloads contain two voice types; standard, and high quality. Both of these are able to produce male, female, and child voices.

VoiceApp has also been integrated to facilitate its use by means of mobile phones and hand-held devices. In this case, the system uses the multimodal NetFront Browser v4.1[5]. NetFront supports advanced mobile voice recognition technologies based on X+V, including voice synthesis and voice recognition of mobile Internet data in voice supported web pages. Speech recognition is provided by the embedded ViaVoice speech-recognition program.

3.1 Generation of the XHTML+Voice Pages

The development of oral interfaces implemented by means of X+V implies the definition of grammars, which delimit the speech communication with the system. The <*grammar*> element is used to provide a speech or DTMF grammar that specifies a set of utterances that a user may speak to perform an action or supply information, and for a matching utterance, returns a corresponding semantic interpretation. We have defined a specific strategy to cover the widest range of search criteria in *VoiceApp* by means of the definition of speech recognition grammars in the different applications. This strategy is based on different aspects such as the dynamic generation of the grammars built from the results generated by the interaction with a specific application (e.g., to include the results of the search of a topic using the *Voice Browser*), the definition of grammars that includes complete sentences to support the naturalness of the interaction

[4] `http://www.opera.com/`

[5] `http://www.access-company.com/products/internet_appliances/netfrontinternet/`

with the system (e.g., to facilitate a more natural communication and cover more functionalities in *Voice Pronunciation*), and the use of the ICAO phonetic alphabet[6] in the cases in which spelling of the words is required in order not to restrict the contents of the search or in situations in which repetitive recognition errors are detected (e.g., in order not to delimit the topics to search using Voice Browser).

Figure 1 shows the translation between a HTML document and its equivalent X+V file. This translation is automatically carried out by means of the PHP files included in *VoiceApp*. As it can be observed, a VoiceXML application consists of one or more scripts that can call each other. A <*form*> is a basic dialog element to present information and gather user inputs, which is generally composed of several form items. The form items are subdivided into input items and control items. Variables in VoiceXML are declared by <*var*> elements, or by form items such like <*field*> with name attributes. VoiceXML has several elements to operate the control flow of the script (for example, <*if*>, <*goto*>, <*exit*>, and <*submit*>). Event handling is carried out by means of elements like <*noinput*> and <*nomatch*>.

3.2 Voice Dictionary, Voice Browser and Voice Pronunciation

As previously described, the *Voice Dictionary* application offers a single environment where users can search contents in the Wikipedia encyclopedia with the main feature that the access to the application and the results provided by the search are entirely facilitated to the user either through visual modalities or by means of speech. Once the result of an initial search is displayed on the screen and communicated to the user by means of speech, they can easily access any of the links included in the result of the search or visit the rest of applications in the system with the possibility of interrupting the system's speech in any case. This functionality is achieved by means of the dynamic generation of the corresponding grammars, in which the different links that are present in the result of a specific search are included in the dynamic X+V page automatically generated by means of a PHP script that captures the different information sources to inform the user about them (headings, text, contents, formulas, links, etc.). Figure 2 shows the initial page of the application.

Google is currently one of the most important companies for the management of information on the Internet due to its web search engine and a number of applications and services developed to access information on the net. This way, the *Voice Browser* application has been developed with the main objective of allowing the speech access to facilitate both the search and presentation of the results in the interaction with the Google search engine. The application interface receives the contents provided by the user and displays the results both visually and using synthesized speech. The application also allows the multimodal selection of any of the links included in the result of the search by numbering them and allowing using their titles as voice commands (Figure 2).

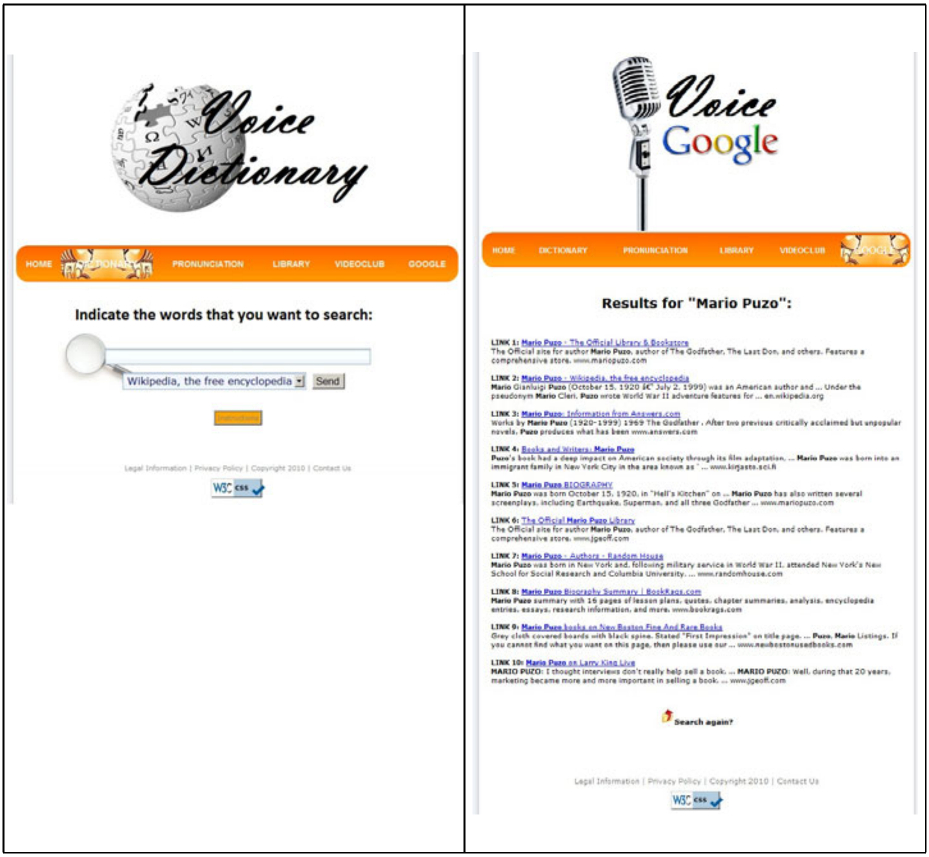
[6] International Civil Aviation Organization (ICAO) phonetic alphabet: `http://www.icao.int/icao/en/trivia/alphabet.htm`

```
% HTML document
<html>
<head>
<title>VoiceApp-Voice Browser</title>
</head>
<body>

<li>LINK 1:
<a href="http://www.beatles.com/">
<b>The Beatles</b> Find out all about
The Beatles...</li>

...

<li> LINK 10:
<a href="http://www.rarebeatles.com/">
<b>Songs, Pictures, and Stories of
The Beatles</b>
Beatles website for collectors
and fans ...</li>

</body>
</html>
```

```
% XHTML+Voice file
<?xml version="1.0" encoding="ISO-8859-1"?>
<html xmlns="http://www.w3.org/1999/xhtml"
 xmlns:vxml="http://www.w3.org/2001/vxml"
 xmlns:ev="http://www.w3.org/2001/xml-events"
 xmlns:xv="http://www.voicexml.org/2002/xhtml+voice">

<head>
    <title>VoiceApp - Voice Browser</title>
    <vxml:form id="nav">
    <vxml:block>
    To visit the links, you have to say
    "LINK" and thecorresponding number.
    </vxml:block>
    <vxml:field xv:id="app" name="app">
      <vxml:grammar src="inig.jsgf"/>
      <vxml:nomatch>
        <vxml:prompt>
        Please repeat again, I can not understand you.
        </vxml:prompt>
      </vxml:nomatch>
    </vxml:field>
    <vxml:filled mode="all">
      <vxml:prompt> Ok got them. </vxml:prompt>
      <vxml:elseif cond="app == ’home’"/>
        <assign name="window.location" expr="index"/>
      <vxml:elseif cond="app == ’link 1’"/>
        <assign name="window.location" expr="x1x"/>
     ...
      <vxml:elseif cond="app == ’link 10’"/>
        <assign name="window.location" expr="x10x"/>
      </vxml:if>
    </vxml:filled>
  </vxml:form>
    <script src="java.js" type="text/javascript"></script>
</head>

<body id="docBody" ev:event="load" ev:handler="#nav">
    <div id="cont" ev:event="click" ev:handler="#nav">
<h1>Results for: The Beatles</h1>

<li>LINK 1:<a href="http://www.beatles.com/">
<b>The Beatles</b> Find out all about
The Beatles...</li>

...

<li> LINK 10: <a href="http://www.rarebeatles.com/">
<b>Songs, Pictures, and Stories of The Beatles</b>
Beatles website for collectors and fans ...</li>

</body></html>
```

Fig. 1. Translation of a HTML document into an equivalent XHTML+Voice file

The *Voice Pronunciation* application has been developed with the main objective of implementing a web environment that facilitates second-language learning with two games that help to acquire new vocabulary and train the words pronunciation. The game *Words* shows on the screen and synthesizes orally the definition of one of the over one hundred thousand words stored in a database of


Voice Dictionary
HOME
DICTIONARY
PRONUNCIATION
LIBRARY
VIDEOCLUB
GOOGLE
Wikipedia, the free encyclopedia
Send
Instructions
Legal Information | Privacy Policy | Copyright 2010 | Contact Us
Voice Google
HOME
DICTIONARY
PRONUNCIATION
LIBRARY
VIDEOCLUB
GOOGLE
Search again?
Legal Information | Privacy Policy | Copyright 2010 | Contact Us

Indicate the words that you want to search:

Results for "Mario Puzo":

LINK 1: Mario Puzo - The Official Library & Bookstore
The Official site for author **Mario Puzo**, author of The Godfather, The Last Don, and others. Features a comprehensive store. www.mariopuzo.com

LINK 2: Mario Puzo - Wikipedia, the free encyclopedia
Mario Gianluigi **Puzo** (October 15, 1920 â€" July 2, 1999) was an American author and ... Under the pseudonym **Mario** Cleri, **Puzo** wrote World War II adventure features for ... en.wikipedia.org

LINK 3: Mario Puzo: Information from Answers.com
Works by **Mario Puzo** (1920-1999) 1969 The Godfather . After two previous critically acclaimed but unpopular novels, **Puzo** produces what has been www.answers.com

LINK 4: Books and Writers: Mario Puzo
Puzo's book had a deep impact on American society through its film adaptation. ... **Mario Puzo** was born into an immigrant family in New York City in the area known as " ... www.kirjasto.sci.fi

LINK 5: Mario Puzo BIOGRAPHY
Mario Puzo was born October 15, 1920, in "Hell's Kitchen" on ... **Mario Puzo** has also written several screenplays, including Earthquake, Superman, and all three Godfather ... www.mariopuzo.com

LINK 6: The Official Mario Puzo Library
The Official site for author **Mario Puzo**, author of The Godfather, The Last Don, and others. Features a comprehensive store. www.jgeoff.com

LINK 7: Mario Puzo - Authors - Random House
Mario Puzo was born in New York and, following military service in World War II, attended New York's New School for Social Research and Columbia University. ... www.randomhouse.com

LINK 8: Mario Puzo Biography Summary | BookRags.com
Mario Puzo summary with 16 pages of lesson plans, quotes, chapter summaries, analysis, encyclopedia entries, essays, research information, and more. www.bookrags.com

LINK 9: Mario Puzo books on New Boston Fine And Rare Books
Grey cloth covered boards with black spine. Stated "First Impression" on title page. ... **Puzo, Mario** Listings. If you cannot find what you want on this page, then please use our ... www.newbostonusedbooks.com

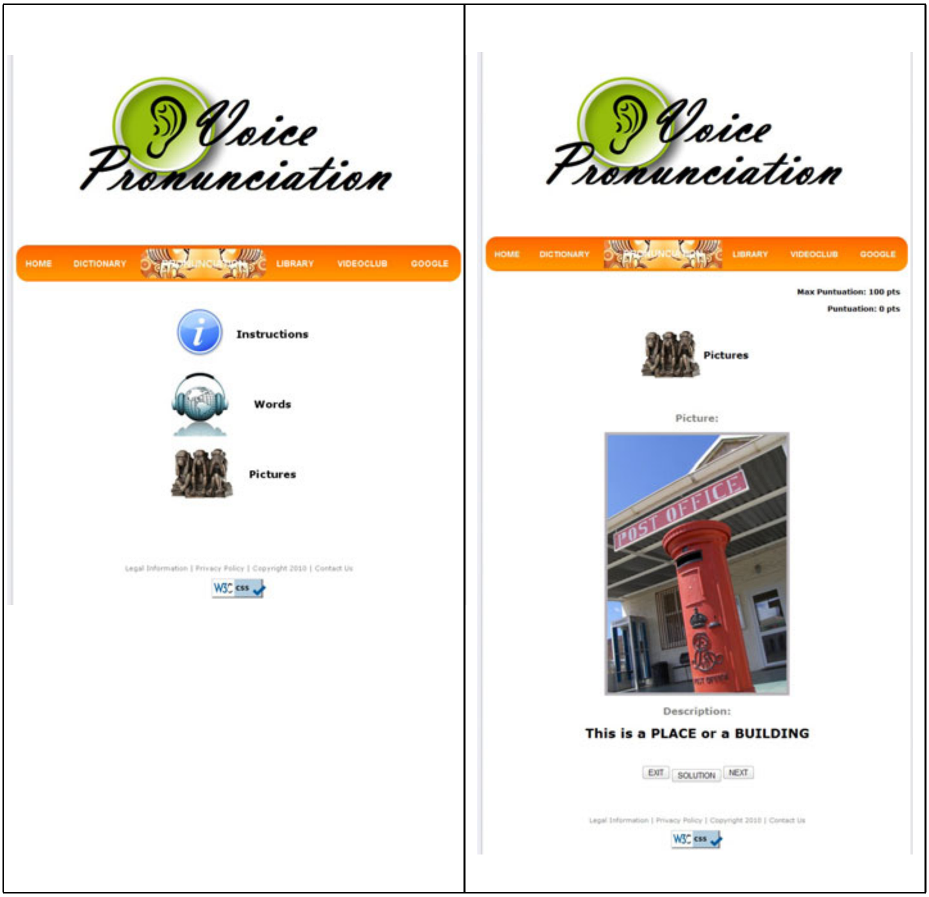
LINK 10: Mario Puzo on Larry King Live
MARIO PUZO: I thought interviews don't really help sell a book. ... **MARIO PUZO**: Well, during that 20 years, marketing became more and more important in selling a book. ... www.jgeoff.com

Fig. 2. Main page of the *Voice Dictionary* application and screen showing the result of a search using the *Voice Google* application

the application and the user must guess the word. The game *Pictures* uses images stored in a database and annotated with different difficulties, whose exact name must be correctly uttered by the user to continue in the game and increase the score (Figure 3). The specific problems and errors detected during the previous interactions of the users with this application are taken into account for the selection of the different words and images and to consequently adapt both games to the specific evolution of each user during the learning process.

4 Preliminary Evaluation

A number of tests and verifications have been carried out to maximize the functionalities and accessibility of the different applications included in the *VoiceApp* system. These tests have been very important to detect and correct programming errors and accessibility problems. One of the main identified problems was



Max Puntuation: 100 pts
Puntuation: 0 pts

Picture:

Description:
This is a PLACE or a BUILDING

Fig. 3. Main page of the *Voice Pronunciation* application and its Pictures functionality

related to the generation of inconsistencies when words with similar pronunciation were reserved to both interact with by the Opera search engine and the different applications in the system. These inconsistencies have been limited to the maximum so that the possible matches between selected words have been eliminated in the different applications.

In addition, we have completed a preliminary assessment by means of a questionnaire to measure users subjective opinion about the system. The questionnaire contained five questions: i) Q1: *Did the system correctly understand you during the interaction?*; ii) Q2: *Did you understand correctly the messages of the system?*; iii) Q3: *Was it simple to obtain the requested information? / Was it simple to play the game?*; iv) Q4: *Do you think that the interaction rate was adequate?*, v) Q5: *Was it easy to correct mistakes made by the system?*; vi) Q6: *In general terms, are you satisfied with the performance of the system?* The possible answers to the complete set questions were the same: *Never*, *Rarely*, *Sometimes*

Usually and *Always*. A numerical value between one and five was assigned for each answer (in the same order as they are shown in the questionnaire). Table 1 shows the average, maximum and minimum values obtained from the results provided by a total of 35 students and professors of our University using the different modules of the system without predefined scenarios.

Table 1. Results of the preliminary evaluation of the *VoiceApp* system (1=minimal value, 5=maximum value)

	Q1	Q2	Q3	Q4	Q5	Q6
Average value	3.6	3.8	3.2	3.7	3.2	4.3
Maximum value	4	5	5	4	4	5
Minimal value	2	3	2	3	2	3

The results of the preliminary evaluation of the *VoiceApp* system show that the users who participated in the assessment positively evaluate the facility of obtaining the requested information by interacting with the system, the appropriate interaction rate during the dialog, and overall operation of the different applications in the system. The main problems mentioned by the users include the need of improving the word error rate and achieve a better clarification of the action expected by the system at each moment of interaction. In addition, the 97% of the interactions finished achieving the objective(s) expected by the user, only the 4% of the systems turns corresponded to reprompts and the 12% to system confirmations. The error correction rate (computer as the average number of corrected errors per dialog divided by the number of corrected and uncorrected errors) was 91%.

5 Conclusions

The *VoiceApp* system has been developed as a framework for the study of the XHTML+Voice technology to develop multimodal dialog systems that improve the accessibility to information on the Internet. The programming languages XML, XHTML and VoiceXML respectively deal with the visual design of the application and allow spoken dialog with the user. This way, multimodal interaction capabilities have been integrated for both the input and output of the system. The use of additional programming languages, as PHP and JavaScript, as well as relational database management systems such as MySQL, facilitates the incorporation of adaptive features and the dynamic generation of contents for the application. Accessibility has been defined as one of the most important design requisites of the system. This way, detailed instructions, help messages and menus have been also incorporated to facilitate the interaction with the different applications in the system.

The set of applications described in this paper respectively facilitate the multimodal access for the search of contents in the Wikipedia encyclopedia, the learning of new languages by improving the words pronunciation by means of funny games, and the complete implementation of a speech-based interface to an Internet search engine.

Current research lines include the adaptation of the system for its interaction using additional languages, a more detailed assessment of each specific application, and the incorporation of new features in each one of them. Another important research line consists of the adaptation of the different applications taking into account specific user profiles considering more detailed information about their preferences and evolution.

Acknowledgements. Research funded by projects CICYT TIN 2008-06742-C02-02/TSI, CICYT TEC 2008-06732-C02-02/TEC, CAM CONTEXTS (S2009/TIC-1485), and DPS 2008-07029-C02-02.

References

1. López-Cózar, R., Araki, M.: Spoken, Multilingual and Multimodal Dialogue Systems. John Wiley & Sons (2005)
2. Beskow, J., Edlund, J., Granström, B., Gustafson, J., Skantze, G., Tobiasson, H.: The MonAMI Reminder: a spoken dialogue system for face-to-face interaction. In: Proc. of Interspeech/ICSLP, pp. 296–299 (2009)
3. Tsai, M.: The VoiceXML dialog system for the e-commerce ordering service. In: Shen, W.-m., Chao, K.-M., Lin, Z., Barthès, J.-P.A., James, A. (eds.) CSCWD 2005. LNCS, vol. 3865, pp. 95–100. Springer, Heidelberg (2006)
4. Kearns, M., Isbell, C., Singh, S., Litman, D., Howe, J.: CobotDS: A Spoken Dialogue System for Chat. In: Proc. of AAAI 2002, pp. 425–430 (2002)
5. Griol, D., McTear, M.F., Callejas, Z., López-Cózar, R., Ábalos, N., Espejo, G.: A Methodology for Learning Optimal Dialog Strategies. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 507–514. Springer, Heidelberg (2010)
6. Stent, A., Stenchikova, S., Marge, M.: Reinforcement learning of dialogue strategies with hierarchical abstract machines. In: Proc. of SLT 2006, pp. 210–213 (2006)
7. Chai, J., Horvath, V., Nicolov, N., Stys, M., Kambhatla, N., Zadrozny, W., Melville, P.: Natural language assistant: A dialog system for online product recommendation. AI Magazine 23, 63–75 (2002)
8. Vesnicer, B., Zibert, J., Dobrisek, S., Pavesic, N., Mihelic, F.: A voice-driven web browser for blind people. In: Proc. of Interspeech/ICSLP, pp. 1301–1304 (2003)
9. Mishra, T., Bangalore, S.: Qme!: a speech-based question-answering system on mobile devices. In: Proc. of HLT 2010, pp. 55–63 (2010)