Information Sharing Network targeted at Embodied Conversational Agents embedded in Interactive Kiosks

Jorge Miguel Fernandes Jorge
Instituto Superior Técnico, Lisboa, Portugal
jorge.jorge@ist.utl.pt

Abstract. Cultural Heritage symbolizes the aggregation of lifestyles practiced by our ancestors and is an important historical entity that needs to be conveyed over to future generations. Following the ideology of preserving Cultural Heritage, it becomes necessary to allow the constant evolution and update of the information present in this area. The usage of cultural multimedia applications is an important center of learning and knowledge transfer in order to encourage the visitor to discover and explore our history. Combined with the use of a virtual agent with which we can communicate through a natural language, provides a new and refreshing interactive tool. Based on these assumptions, we propose the creation of multimodal interactive kiosks with virtual agents supported by a Spoken Dialogue System. Thus, recreating all the usual structure of a Spoken Dialogue System and making it especially optimized with the introduction of the Microsoft Kinect for Windows. These agents will be contained in multimedia applications designed to Cultural Heritage and Tourism. Given that these applications contains restricted and static information we also propose the creation of an Information Representation and an Information Sharing Network between interactive kiosks, building a cooperative multiagent system, which grants access to information at various points in the network. In conclusion, with the development of a working prototype we defined a new method for creating Information Sharing Networks, whose collaborative activities can support conversational agents embedded in interactive multimedia applications for Cultural Heritage and Tourism.

Keywords: Cultural Heritage, Embodied Conversational Agents, Spoken Dialog Systems, Multimodality, Microsoft Kinect, Information Dissemination Network

1 Introduction

One of the most important facets of the history of a country is its Cultural Heritage, acting as a connection between the past and present. However, this broad knowledge has no value or utility on itself if is not made public. As such, it is necessary to create new tools to spread this knowledge and promote its dissemination.
In a time of constant globalization, tourism is the main vehicles for cultural exchange. This fact proved by the transformation of museums and monuments in tourist centers. These cultural centers are capable of boosting an entire region at socioeconomic level.

We can see by the proliferation of projects aimed at dissemination of Cultural Heritage and Promotion of Tourism. These projects draw on the most varied media content, focused primarily at the passage of historical knowledge and cultural values to visitors in a lively manner without compromising its pedagogic value.

1.1 Contributions

The subject for this dissertation comes following the opportunity that we had to participate in the development of the platform \textit{EI}^2\textit{VA}[6] that specializes in the integration of virtual agents in serious games and multimedia applications. After a careful analysis of the findings obtained in these case studies\cite{7,8}, there are areas that need improvement, particularly in interaction through natural language and in the speech recognition module\cite{6}.

With the need to change the paradigm of the recognition process presented in \textit{EI}^2\textit{VA} platform, recreate many of the functional modules of the dialogue component was a necessity, mainly through the introduction of the Microsoft Kinect and specially adapted Acoustic Models. These contributions are intended to improve the entire speech recognition process in order to achieve and aspire to a more natural and realistic interaction between the visitor and the agent communicative.

As a way to change the organizational structure of the educational information present in a multimedia applications, we created an Information Sharing Network. Here, through informative leveling and communicative entities, we sought to provide new ways to shape and deliver these same contents. Both of these contributions were, for their importance and relevance, constitute part of the \textit{EI}^2\textit{VA} integrated platform.

We conclude this work by creating a working prototype that presents all the characteristics described above, hoping to set a new methodology for creating Information Sharing Networks. Through this collaborative work, we can support the pedagogical activities of embedded conversational agents in interactive multimedia applications for Cultural Heritage.

2 State-of-the-art

2.1 Embodied Conversational Agents

Interfaces using the Virtual conversational agents are a metaphor for smoothing the interaction with the system, resorting to use of a virtual representation of a human being. This representation may be composed of a face or even representing the entire body. The use of virtual characters of this type has a special interest in areas of healthcare\cite{1}, online commerce\cite{3}, therapy\cite{2} and more recently learning\cite{9}, which has increased the work done in this area.
2.2 Spoken Dialog Systems

Aiming to interact with humans in real time, these systems should be autonomous and apparently intelligent. As such, they must possess skills such as recognizing and responding to verbal speech, manage conversational functions – handle turn taking during the dialog and feedback mechanisms for error recovery - and further capacity to express the state of discourse as well as to take a proactive stance during the dialog. Next are described the typical modules that forms the architecture of Spoken Dialog System (SDS)[4].

**Automatic Speech Recognition (ASR)** - It is the input of the spoken dialog system, as such is a component that is directly interacts with the user, specifically by using a microphone to capture audio of the user interaction.

**Spoken Language Understanding (SLU)** - Performs morphological analysis, part-of-speech tagging and syntactic analysis in order to complete its main tasks: domain classification and determine the user intent.

**Dialog Manager (DM)** - The result from natural language processing is passed to the Dialog Manager, which is the basis of this system, as it manages the remaining modules. Is also in charge, if necessary, of the access to databases or other services to obtain the information needed to present to the user.

**Natural Language Generator (NLG)** - This module is responsible for producing the answers in natural language based on the concept that receives the dialog manager.

**Text-To-Speech (TTS)** - The synthesizer module is responsible for performing audio synthesis and produce the respective phonetic information corresponding to the response generated by the natural language generator.

2.3 Information Distribution Network

With the increasing expansion of the Web applications and information access, it is important to build systems that allow information sharing in an organized and safe manner. However, most of this information is spread along many internet sites, some of difficult access. A possible solution to these problems and reduce stress associated with the update of the information is to create a distributed platform for content sharing. The base system implemented followed the foundations mentioned in the architecture described in [10], which consists of four key components: agents, tasks, organizations and an information infrastructure capable of modeling the order of supply chain of information in the network. This example created in Hong Kong with the implementation of a multiagent system targeted at tourism kiosks. Through the research of geographical information about places of tourism such as museums, shopping centers, hotels and cinemas. In this system there are four types of agents: Client Agents, Agents Name Servers, Category Group Agents and Agent Informations.

Client Agents are responsible for preparing and showing the information required by the user via a web interface. The Agent Name Server is responsible for maintaining the list of related information agents. It also acts as a channel of communication between cooperating agents in case of security restrictions. The
Category Group Agent main function is act as proxy between Agents Information and Agent Name Servers, removing load from the system with the use of a cache. Finally, the Agent Information is responsible for providing the information requested by users. This module is located at each site or organization interested in offering services.

3 Developments

3.1 Block Diagram

In this section, we present in a high-level view the main components used in the solution developed, as well as the flows of communication, illustrated through a block diagram in fig 1. The chosen structure consists of four main blocks: Application, EIFVA framework, Spoken Dialogue System and Distributed Platform.

![Block Diagram](image)

**Fig. 1.** Main functional blocks of the System

The Application Block contains the functionality to support the application and tools necessary for the interaction with the system. This block interacts directly with the user through a multimodal interface comprised of speech and touch.

The EIFVA framework provides the system environment with an animation engine, responsible for managing all graphical animation of the virtual agents. The Spoken Dialogue System then supports the action of the virtual agent. The SDS function is to handle spoken communication with the user by the management of the knowledge and control of the interaction.

Finally, the distributed platform is responsible for enabling the distribution and dissemination of information among the various kiosks. Through proper structuring of information, it becomes possible to introduce this information in the multimedia application through its process in the environment of SDS.

3.2 Spoken Dialogue Systems

Figure 2 presents the final architecture of the SDS and the flows of interaction between the constituent modules. We endeavored a great deal of our time in the development of SDS in order to mitigate the problems already discussed in the introductory section. We now give a description of developments in the various modules of the dialogue system:
Automatic Speech Recognition (ASR) - Resultant of the desired configuration, interaction in museums and monuments, acoustics and background noise affects the audio captured. This module, as previously described, is of utmost importance in a dialog system because it is the component most prone to errors that are diffused throughout the system. In order to solve this problem it is necessary to intervene in two areas: the microphone and speech recognizer engine. In pursuit of creating voice interfaces that avoid the hassle of the constant use of intrusive microphones, we opted for the use of a Microphone Array (MA) as speech signal acquisition interface. With the use of an MA is possible to benefit from beamforming, echo cancellation and noise suppression techniques. After the tests we chose the device Microsoft Kinect (MS Kinect) due to its superior quality of the captured audio. AUDIMUS[5] was selected as speech recognition engine. AUDIMUS is a hybrid speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMM) with the pattern discriminative classification capabilities of multilayer perceptron’s (MLP). The use of a MA input device required new acoustic models created specifically for this device. The system uses a different language model for each specific language. With this approach, the recognition percentage became much higher.

Spoken Language Understanding (SLU) - In order to build a module simplistic but robust, where the usual paradigm of interaction is question and answer, we opted for the use of recognition grammars rather than a complex language model. With the management of the work involved for the semantic interpreter, we adhere to the use of grammars derived from the specification SRGS - Speech Recognition Grammar Specification - standard recommended by W3C. We also introduced the use of cumulative grammars, composed by simplified versions of specific interactions with the suppression of connectors between sentences and prepositions that may cause small recognition errors.

Dialog Manager (DM) - The Dialog Manager is responsible for managing conversational state with the user. It is also accountable for managing the flow of information between the remaining modules. Receives as input the semantic representation from the action of natural language processor and determines the appropriate action to perform. In order to fulfill its functions the Dialog Manager maintains a set of relevant information about the state of the dialogue: history of
interactions performed, dialogue phase and errors occurred during the dialogue. In this module were also implemented mechanisms for incremental processing of the user interactions in order to tackle problems associated with turn taking and mechanisms of error detection and error recovery.

**Relational Knowledge Base (RKB)** - The Relational Knowledge Base is responsible for managing the system responses to the user, building on a structure-based XML files that specifies the possible outcomes for an interpella-
tion.

**Text-To-Speech (TTS)** - This component is responsible for converting the textual answer returned by Relational Knowledge Base in audio. Because this SDS being used in conjunction with a virtual agent, in addition to the audio synthesis is necessary to extract a set of fragmentary information, like phonemes and phonemes times. That Information is need by animation engine to synchronize animations with the audio representing the text response.

### 3.3 Information Sharing between Embodied Conversational Agents

Most of the information present in applications of this kind is typically immutable and of difficult access. This paradigm is easy to observe in the touristic area, where much of this information is not updated with sufficient regularity, transposing an image of disuse. In order to combat this recurrent problem, we tried to include a content distribution platform enabling the dissemination of information. This platform aims to introduce dynamism in SDS with the inclusion of information received during the execution of the system in the question answer module. To complement this distribution, we provide the system with a back office interface that allows quick modifications to information provided. In order to create a robust, but at the same time appropriate to the requirements set for the system, it was necessary to structure the information at multiple levels: N1, N2 and N3.

The level N1 is comprised of more site-specific information and also more extensive. The level N2 gathers information to spread on the network. This category of information is designed in order to allow the agent to provide a concise description of a surrounding location that contains a network node. Finally, the level N3 retains informations of the sub-networks for information sharing, events and tours available.

The topology of this network is based on a set of different entities that communicate with each other, in a star formation: the **Master Server** (MS) and the **Kiosk Server** (KS). The KS resides at each site that wants to provide or receive content, working in conjunction with a kiosk application. Stores information levels N1 - specific to the local - and N2, which is important to attract new users. The MS function is monitoring the system, manage the dissemination of information in the sub-networks based on the information stored related to sub-networks of diffusion and tours.
4 Results and Validation

Completed a brief analysis of the developments that led to the architecture described in the previous chapter, in this section is introduced an assessment of the results obtained during the evaluation of the system. The system under consideration has two architectural components of most relevance: Spoken Dialogue System and Information Distribution Platform. Given this division is also possible to find, for each of these areas, a specific methodology for validating the implementation achieved.

4.1 Development Accompaniment Group

From the beginning of the system development, especially for the implementation of Spoken Dialogue System, a group for specialized monitoring was created. This group was responsible for critically examine the progress made in this module. The steering group of the SDS had the following composition:

- 1 Specialist in Spoken Dialogue Systems
- 2 Experts Users
- 3 Non-expert users

4.2 Speech Recognition Evaluation

For testing the quality of the audio captured, we chose three devices: a headset microphone, a Andrea SuperBeam microphone and the multisensor device Microsoft Kinect for Windows. We collected the data for these assessments through a process where users were invited to use the various speech capture devices, allowing an evaluation of the recognition success rate and the resulting audio quality. For each device, users effected the dictation of 29 phrases, balanced phonetically and lexical. The main subject of this evaluation is Portuguese language.

Adapted Acoustic Model for MS Kinect: In a first step, in order to evaluate the integration of the MS Kinect with the speech recognition engine, we developed an acoustic model (AM) adapted to a particular speaker. The process of generation of this adapted acoustic model involved the dictation of approximately one hundred sentences, balanced phonetic and lexical. Due to logistics included in adapting an acoustic model to a speaker, this evaluation was performed with only one user.

Examining the Table 1, it is visible difference between the values achieved by the adapted acoustic model and the values registered in the initial tests performed with generic acoustic model for broadcast news. There has been an increase of 13% in recognition success rate compared to the best results achieved in various setups used as baseline for this evaluation. Also important to note the influence of the presence of cumulative grammars with recognition rate around 24% based on these rules.
**Generic Acoustic Model for MS Kinect:** Completed the process of evaluating an adapted acoustic model, we departed to the creation of a generic acoustic model, in order to make it independent of the speaker. Due to the logistics involved in the process, gathering a large group of people in order to make the model generic to the speaker, we opted by simplistic approach. The solution adopted was the use of a database of pre-made recordings in a chamber-deaf environment. This database had recordings of over a hundred people of various ages and sexes, allowing access to a set of more than a hundred phrases, per user, properly recorded and annotated for the training of acoustic models.

For assessing this generic acoustic model generated for the MS Kinect, we maintained the procedures for the previous tests, using again the development monitoring group and testing both the use of complete grammars and cumulative grammars.

<table>
<thead>
<tr>
<th>Recognized</th>
<th>Complete Grammar</th>
<th>Cumulative Grammar</th>
<th>Complete Grammar</th>
<th>Cumulative Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headset Generic AM</td>
<td>138</td>
<td>150</td>
<td>156</td>
<td>160</td>
</tr>
<tr>
<td>Rejected</td>
<td>36</td>
<td>24</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td>Recognition %</td>
<td>79.31%</td>
<td>83.31%</td>
<td>89.66%</td>
<td>91.95%</td>
</tr>
<tr>
<td>Avg. Value of Recognition Conf.</td>
<td>91.01%</td>
<td>93.87%</td>
<td>99.26%</td>
<td>98.28%</td>
</tr>
<tr>
<td>Cumulative Rules Recognized %</td>
<td>-</td>
<td>18.23%</td>
<td>-</td>
<td>18.39%</td>
</tr>
</tbody>
</table>

Table 2. Comparison between Headset Device with Generic AM for broadcast news and MS Kinect with the Generic AM for MS Kinect

In table 2, are evidenced comparisons between initial testing for devices headset with Generic AM for Broadcast News and the Generic AM created for MS Kinect. After a review of the values presented, it is possible to see a decrease from the values of the previous tests, a reduction of 4.6% of recognition power result of the generalization acoustic model. This loss is normal, since with the generalization of the AM a larger number of users can now use the recognition engine without compromising the reliability of the recognition process. Despite this decrease in the recognition success rate, it achieved the 89.5% using grammars containing only complete sentences, and 91.95% for the use of cumulative
grammars. These values supplant the performance of any device used in conjunction with the Generic AM for Broadcast News presented previously.

Based on values presented it is possible to validate the developments made. Starting with the choice MS Kinect as the preferred device for capturing and processing audio, together with process for creating generic and adapted acoustic models provide a significant increase in the recognition success rate. At the same time, with the use of cumulative grammars, it allows a greater flexibility and better understanding form the user’s expression as proved in previous tests.

4.3 Information Sharing Network

As stated earlier in this chapter, for the evaluation of this crucial component of the system an assessment in a more analytical view was carried out, in which we analyzed the strengths and weaknesses of the developed platform. Because this platform is still in an embryonic stage was not possible to provide a system with a real environment of use that allows full exult all of its features.

With these evaluation limitations present in mind, we will make an assessment at the technical level. In which the tactics used to address the generic problems of distributed systems such as scalability, modifiability, security and availability will be evaluated.

From a functional point of view, we focused the creation process of management interfaces to solve the most frequent tasks in this subject area. The completion of the tasks listed below can be achieved easily using Web Interfaces available to manage the information available on the network.
- Add a new point of interest to the Network;
- Modify the information present in level N2 of a point of interest;
- Broadcast an event relative to a point of interest;

5 Conclusions and Future Work

A considerable portion of the research carried out was focused in the Spoken Dialogue Systems area, primarily in the components of speech recognition and dialogue management. The developments included in these components allowed a more reliable and robust process of speech recognition in noisy environments. Part of this success, is attributed at the inclusion of Microsoft Kinect for Windows as audio capture device, in combination creation of appropriate acoustic models. At the management level, the employment of dialogue tactics to detect and mitigate the random errors enabled to strengthen and make more natural the interaction with the dialogue system. Based on these developments and proven by tests, it is possible to highlight the increasing robustness of the system in spoken dialogue interactions.

Another important issue during the development of this thesis was the creation of a platform for dissemination of information. This platform gave a new impetus to the existing system, enabling the integration of information in Spoken Dialogue System dynamically. Creating an application with virtual agents
that allows the presentation of modern and renovated content to the users. The interaction with a virtual agent allied to a constant renewal of content can direct visitors to the surrounding areas of the monuments visited. It remains so, with prospects for future development, stating in summary form the vital points that need to be improved in the current system:
- Using the visual component provided by MS Kinect for behavioral analysis of the actions and speech of the user;
- Introduction of a natural language processor and a language model working with the current semantic processor.
- Allow interaction of the Information Sharing Platform with mobile devices.

References