A ROBUST INPUT INTERFACE IN THE SCOPE OF THE PROJECT INTERACTIVE HOME OF THE FUTURE

João P. Neto and Renato Cassaca

L^{2}F - Spoken Language Systems Laboratory
INESC ID Lisboa / IST
www.l2f.inesc-id.pt
{Joao.Neto,Renato.Cassaca}@l2f.inesc-id.pt

Abstract

This paper presents the work done in the integration of a spoken dialogue system in a new project on an *Interactive Home of the Future*. This spoken dialogue system gives access to a virtual butler that is able to control the home environment. In this system we combine automatic speech recognition, natural language understanding, speech synthesis and a visual interface based on a realistic animated face. In order to make the system available to the people using the home was necessary several modifications to the system to create a robust input interface.

1. Introduction

The concept of home automation has been followed for a long time. There are now very strong signals that the technology is available and affordable. Still some electronic manufacturers are working to create and support standards. The universities and research institutes belief in the value of creating living laboratories where they can experience, analyze and improve the technologies they are investigating. Also there are futuristic computing applications that apply ubiquitous computing technology to everyday life, and particularly at home.

The home automation concept is normally faced under different perspectives. The most common is the need to control in an easy way the home either locally or remotely. In a more intelligent way we would like that the home could anticipate our needs, learning with our own control actions under time and weather conditions. There are also some perspectives coming from a more rational management of energy and the need to enabling older adults to age in place and being supplied by appropriated medical cares.

All these perspectives must rely in a well-defined human-computer interface. However the obstacle to understanding new interfaces is high, and the control of a complete home is necessarily complex. Speech is a natural way of communication and despite all the problems will serve in the future as the most convenient interface.

This paper presents the work done in the integration of a spoken dialogue system in a new project on an *Interactive Home of the Future*. Despite this being and exposition home, with a diversity of target public, serves to us as a living laboratory to technology rehearsal.

This spoken dialogue system gives access to a virtual butler that is able to control the home environment. In this system we combine automatic speech recognition, natural language understanding, speech synthesis and a visual interface based on a realistic animated face. In order to make the system available to the people using the home was necessary several modifications to the system to create a robust input interface.

In section 2 we will present the project of the *Interactive Home of the Future* including our participation. Section 3 gives a brief description of our spoken dialogue system and in section 4 we will describe the user interaction with the system. The conclusions are presented in section 5, followed by some acknowledgements and references.

2. The project *Interactive Home of the Future*

2.1. The project

The project *Interactive Home of the Future* has born in 2002 based on the initiative from FPC (Communications Portuguese Foundation [http://www.fpc.pt]) and the Group Portugal Telecom (main telecommunications operator in Portugal).

This home was built in the second floor of FPC as a mean of exposition of new technologies of telecommunications. The main idea was to use the technologies for generating interactivity and mobility in the interaction with the house wherever the people are. This house is an advanced solution of Home-Automation integrating technologies ready available in different domains, from equipments, systems, furniture, design, infrastructure and building, and offer from telecommunications operators (cable TV, interactive TV, new telecommunications products), with some new prototypes being developed by Universities and research teams.

The idea was to show and disseminate to the generic public the present and future potentialities of the telecommunications and multimedia, create an offer of complementarities between different companies connected to the new technologies, and to contribute to the development and innovation in the area of communications and multimedia.

Professionals working in the area, under-graduated students, families and young students, are the most important groups that visit this exhibition.

2.2. Our participation in the project

The Spoken Language System Laboratory (L^{2}F) of INESC ID Lisboa was invited to participate in this project with the development of a spoken dialogue system. This system allows the house users to access through a spoken based interface to the different devices and services of the house. This system represents the concept of a virtual butler, someone that is always available to execute our requests. We combine automatic speech recognition, natural language understanding and speech synthesis. A visual interface based on a realistic animated face synchronized with the speech production process creates a good effect over the user helping in the dialogue process. The system responds to the name of "Ambrósio".

Fig. 1 - Animated face from the graphical interface of the system.

At present time our system operates only in the "Main Bedroom" controlling the devices available in that space. The devices are the ceiling lights, divided by the zones of the bed and the television, the opacity of the window glass from the bedroom to the hall, and the television. This initial situation allows us to show the real effectiveness of this system and its large applicability. In the following phase we expect to extend this control to the living room, the entrance zone and to the kitchen.

To our Laboratory this project brings the possibility to work in a living laboratory to technology rehearsal, working over real problems and observe the users problems. Also was possible a strong engagement of students, contributing to their formation and the creation of a strong relationship with the companies involved in the project.

3. Our system

3.1. Architecture of the system

Our system is based on three main blocks (see Figure 2). Two of them are responsible for the interfaces with the user and the centralized system for control of devices (based on a web server). The other block is responsible for the dialogue management.

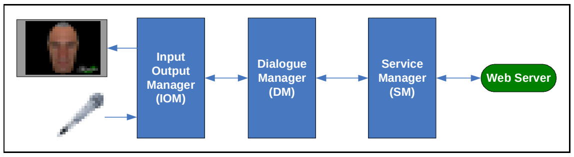

Fig. 2 - Block diagram of the system.

The user interface is based on a wireless microphone available on the bedroom and the TV set, where "Ambrósio" is visualized and answer to our requests (the TV speakers are the output point of the speech generated by our system). The control of devices has an interface based on a web server, making available the access to any device in the home.

Despite this specific application of our system we have been developing these blocks in a more generic way, in order to cope with different types of applications. We use the same system to control the home environment, to access different databases (weather information, bus information, stock market information) and to email access. Also we get access to the system from microphone, telephone, GSM, PDA and web [1].

In order to satisfy these goals both interface blocks create an independency level to the Dialogue Manager (DM). In our case the DM does not know which type of device made the request, if it was a speech request or the result of a click by the pen in a PDA application. The Input and Output Manager (IOM) creates an independency level sending the same XML format, independent of the source. The same principle of independency is applied at the Service Manager (SM). The configuration of services creates a representation that is independent of the domain [1,2].

Next we give a brief description of each of these blocks.

3.2. Input and Output Manager (IOM)

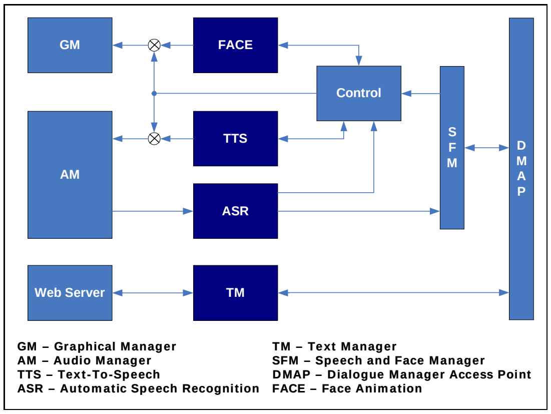

Fig. 3 - Block diagram of the Input and Output Manager.

There are 4 main blocks in this diagram: the ASR, the TTS, the FACE and the TM.

The ASR is based in *Audimus* [3], a hybrid speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of multilayer perceptrons (MLPs). This same recognizer is being used for different complexity tasks based on a common structure but with different components.

The TTS module (DIXI+) [4] is a concatenative-based synthesizer, based on the Festival framework. This framework supports several voices and two different types of unit - fixed length units (such as diphones), and variable length units. This latter data-driven approach was suitable, by adequate design of the corpus, to a limited domain of application as our present situation.

The FACE [5] is a Java 3D implementation of an animated face with a set of visemes for the Portuguese phonemes and a set of emotions and head movements.

The TM transforms the web server requests, from PDA and web, in the XML format to access the DM.

The CONTROL block receives an XML file, with text, emotions and head movement's descriptions, splitting and feeding the appropriate information to the FACE and TTS. Is responsible to the synchronization of these blocks outputs.

The SFM manages the interface between the speech and animated face blocks with the DSAP. Is a simple block that encapsulates the ASR output in a XML format and, in the other direction, responsible to send the XML file, received from the DM, to the CONTROL block.

The GM and AM are responsible to the interface with the input and output devices, both audio and graphical.

DSAP manages the communication between the SFM and TM with the DM. DM uses the hub structure of the Galaxy II.

3.3. Dialogue Manager (DM)

The DM receives the request from the IOM in a XML format and has to determine the action requested by the user and ask the SM for the execution of that action. The domains representation and the information extracted from the user request are handled through frames [6].

Each domain is internally represented by a frame, which is composed by slots and rules. Slots define domain data relationships, and rules define the system behavior. Rules are composed by operators (logical, conditional, and relational) and by functions, representing a user possible action [1].

The frame definition includes a set of rules: recognition rules, for the identification of objects from the user utterance and to specify the set of values that each slot may hold; validation rules, expressing a set of domain restrictions, avoiding invalid combination of slot values; and classification rules, used to specify the actions that must be performed when some conditions are satisfied, resulting from the valid combinations of slot values.

Our DM [6] is based on a communication hub of the Galaxy framework and in a set of main blocks: Language Interpretation (LI), Interpretation Manager (IM), Discourse Context (DC), Behavioral Agent (BA) and Generation Manager (GM). These blocks are interconnected through the Galaxy hub together with the IOM and SM.

The LI receives the user request from the IOM and transforms it in a speech act [7]. The IM receives the speech act and generates the correspondent interpretation and discourse obligation. Interpretations are frame instantiations that represent possible combinations of speech acts and the meaning associated to each object it contains. The DC manages all knowledge about the discourse, including the discourse stack, turn-taking information, and discourse obligations. The BA enables the system to be mixed-initiative, since regardless of the user request the BA has its own priorities and intentions. Whenever the system recognizes that the user is changing domains, first verifies if some previous conversation in that new domain has already taken place. The GM is responsible to communicate with the user, transforming the system intentions in natural language utterances. It receives discourse obligations from BA, transforming them into text using template files. Each domain has a unique template file, and a separate template file to produce questions that are not domain specific.

3.4. Service Manager (SM)

This block provides the DM with the necessary interface with a set of heterogeneous devices grouped by domains. A domain is a representation of a set of devices that share the same description, composed by slots and rules. As mentioned before we work simultaneous with different domains as the ones related to control the home environment, to access different databases (weather information, bus information, stock market information) and to email access.

A service represents a possible user action. The DM receives a speech act as an object, and to create an interpretation the IM needs the representation of that object in the different domains. The IM requests the information from the object to the SM. The SM searches in the available and accessible domains for that object. The SM reply back to the IM giving the object definitions. In case of no ambiguity the IM builds one interpretation and a discourse obligation for execution of the service. The IM request that execution to the SM, which responds back indicating the result of the service. In case of ambiguity the IM creates different interpretations, which requires a domain disambiguation. The BA picks that disambiguation request and sends it to the GM. The GM uses the domain independent template file to generate a disambiguation question to the user. After the user response the IM should end with only one interpretation and is able to request the service to the SM as explained before.

All the information about the devices and the domains is represented in this block, creating the DM domain independency. The SM makes available to the DM everything that is related to the domain, the representation of the domain and devices, their states and the user access restrictions. The organization of this information and the effort made to create a homogeneous representation across domains give rise to this domain independency being so easy to create and add new domains to the system.

```
<device name="Room Light">
  <frame frameType="static">
    <DESCRIPTION>
      <frame_name>domotica</frame_name>
      <db_name>domotica</db_name>
      <slot type="string" KEY="1">Action</slot>
      <slot type="string" KEY="2">Target</slot>
      <slot type="string" KEY="3">Device</slot>
      <slot type="string" KEY="4">Location</slot>
    </DESCRIPTION>
  </frame>
  <state name="Available">
    <service name="turn_on_light">
      <label>Turn on room light</label>
      <params>
        <param slotkey="1" value="turn_on"></param>
        <param slotkey="2" value="light"></param>
        <param slotkey="3" value="room"></param>
      </params>
      <execute>
        <name>connect</name>
          <success/>
          <failure/>
      </execute>
    </service>
    <service name=" turn_off_light ">
      <label> Turn off room light </label>
      <params>
        <param slotkey="1" value="trun_off"></param>
        <param slotkey="2" value="light"></param>
        <param slotkey="3" value="room"></param>
      </params>
      <execute>
        <name>disconnect</name>
          <success/>
          <failure/>
      </execute>
    </service>
  </state>
</device>
```

Fig. 4 - Definition file in XML of service "Room light"

A domain is represented by an XML structure that defines the set of available services, define the implementation code that will execute the services, define the identification of the domain objects and the templates of the system domain utterances. This domain representation is saved on XML files that are loaded by SM, creating an internal representation of the domain. In Fig. 4 we present the example of a XML representation of a service "Room light". SM could be asked for the set of services associated with the specified domains, and at a given instant it is only possible to have access to a hierarchy of services available in the actual state of the device.

4. User interaction with the system

4.1. From lab to the real world

There is always a significant mismatch from a lab environment and the real environment associated with generic public, nevertheless our efforts to create in the lab real acoustic situation and real use of the system. In the environment associated with the *Interactive Home of the Future* we found different type of users: experts, like our colleagues and students that know and are used to these kind of systems, the guides of the exposition, that start learning how to use the system, and the generic visitants, that want to try and use the system. Also there is a large spectrum of voices, covering every possible situation, male/female, adult/children. The background noise also have a large variability being normal to have in a small room groups of 30 young students. How to tailoring the system for these situations?

4.2. Modifications on the system

The use of a wireless microphone gave a large degree of freedom to the user, being appealing to the visitants to try to use the system. To cope with this situation it was necessary to have an open microphone situation. The system needs to be selective in what to send to the ASR. We requested a keyword access to the system, process being based on a DTW or MLP classifier. That means wherever we want to talk to the system we have to start by saying "Ambrósio" the name of the butler figure. From an open situation we came for a keyword filtering process that should not be user specific. At this point we were able to have some degree of control in the use of system, opening their use but not completely, since there are different voices very difficult to cope with a generic model.

In the beginning we notice that was very difficult to understand the state of the system. The only output was the animated face that was natural but neutral and only moves synchronized with the speech. We changed the behavior in order to the face reflect the state of the system. We implemented six states: silence in the input, no one is talking or the mic is closed; voice activity in the input; the name Ambrósio was called; following Ambrósio a service was requested; the system is executing the service; there is a reply from the system. Since we always get a reply from the system, despite the success or not of the service execution, this feature results more informative to the user.

With a generic use of the system was necessary to cope with the different ways of accessing the system. Despite of being only available a few set of services we found there a large number of ways to ask for turning-on the lights (at least in Portuguese). It was necessary to add to the vocabulary all the possible synonymous for the different actions and to map these new meanings to the same action in the LI module. For the ASR this implied the modification of the vocabulary, lexicon and language model. The language model resulted from an interpolation of a generic model extracted from newspaper texts and a very small model generated from sentences associated to the domain. Having always in mind the goal of task independency of the components of the spoken dialogue system, we must supply this kind of information in an XML format with the device/domain definition. The system when receives a new device should extract that information and add it to the resources available at the ASR, TTS, LI and GM.

With this representation each device definition comprise all the necessary information associated with that device/domain making the spoken dialogue system independent of the domain.

5. Conclusions

This paper presents the work done in the integration of a spoken dialogue system in a new project on an *Interactive Home of the Future*, giving access to a virtual butler that is able to control the home environment. In the development of our system we built a structure that is independent of the domains and where the necessary information associated to each device/domain is provided in XML format. There was a set of modifications in our system necessary to create a robust input interface that improved the system performance, increasing the user acceptability of a Spoken Dialogue System as a valid and useful interface.

6. Acknowledgements

This work was partially funded by FCT project POSI/PLP/41319/2001 (POSI and FEDER). There are other colleagues that contributed to this project, as Márcio Viveiros, Márcio Mourão, Sérgio Paulo, Nuno Mamede e Luís Oliveira.

7. References

[1] J. Neto, N. Mamede, R. Cassaca, L. Oliveira, "The development of a multi-purpose Spoken Dialogue System", *Proc. Eurospeech 03*, Genéve, Swiss, 2003.

[2] M. Mourão, R. Cassaca and N. Mamede, "An independent domain Dialogue System through a Service Manager", *Proc. ESTAL 2004*, Alicante, Spain, 2004.

[3] H. Meinedo, D. Caseiro, J. Neto and I. Trancoso, "AUDIMUS.MEDIA a Broadcast News speech recognition system for the European Portuguese language", *Proc. PROPOR'03*, Faro, Portugal, 2003.

[4] S. Paulo and L. Oliveira, "Multilevel Annotation Of Speech Signals Using Weighted Finite State Transducers", *Proc. 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, USA, 2002.

[5] M. Viveiros, "Cara Falante", *Graduation Thesis*, IST.

[6] P. Madeira, M. Mourão, N. Mamede, "STAR Frames - A step ahead in the design of conversational systems capable of handling multiple domains", *Proc. ICEIS*, Angers, France, 2003.

[7] D. Traum, "Speech Acts for Dialogue Agents", UMIACS, Univ. Maryland, USA, 1999.