A VIRTUAL BUTLER CONTROLLED BY SPEECH

A. Uria, A. Ortega, M. I. Torres, A. Miguel*, V. Guijarrubia, L. Buera*,*
J. Garmendia, E. Lleida*, O. Aizpuru, A. Varona, E. Alonso, O. Saz*

Departamento de Electricidad y Electrónica
Facultad de Ciencia y Tecnología.
Universidad del País Vasco
manes@we.lc.ehu.es

* Communication Technologies Group (GTC)
I3A, University of Zaragoza, Spain
lleida@unizar.es

ABSTRACT

The aim of this work was to develop a virtual butler-service to be installed at home to control electrical appliances and to provide information about their conditions. The framework of this project lays on FAGOR Home Appliance. The overall goal was to develop a smart home where anyone, even physically handicapped people, could control the appliance with the voice. A spoken dialog system was developed allowing a spontaneous and speaker independent speech input. A Speech Recognition module is used to convert the speech input signal into text. A speech-understanding model translates the recognized utterance into a sequence of task dependent frames. The Dialog Manager generates the most suitable answer. It also controls the device activation when the required information is ready. The butler will be able to fully control and program a washing machine, a dishwasher machine and an oven, being able to show some recipes. In order to test the Speech Recognition module, a specific corpus was recorded in the FAGOR Home Appliance facilities. Some experiments were carried out with this corpus, obtaining up to a 67,31% of improvement when using speaker and environment adaptation.

1. INTRODUCTION

The framework of this project lays on FAGOR Home Appliances, which is a electrical appliance multinational in Spain: the 44% of FAGOR sales are on the international market, and 70% of these overseas sales are made in countries as competive as France, Germany and the United Kingdom.

The aim of this work is to develop a virtual butler service that would be installed at home to control electrical appliances and provide information about their conditions. The system would allow to ask for the state of each appliance, to program them or to consult a database of recipes while you are cooking. All of these tasks occur in a natural way due to a dialog that is established between the virtual butler and the user.

This work has been partially supported by the national project TIN 2005-08660-C04.

All the queries can be carried out both inside the house or by telephone outdoors. The telephone bestows mobility on the user since the system can be controlled or requested for information from anywhere.

The virtual butler enables interaction with the user. Vocal input is managed by a spoken language system, which aims to provide an effective interface between the user and the system through simple and natural dialogs. Concerning the output, the system replies to the user by generating multimodal presentations which combine spoken output, dynamical graphic displays and actions such as switching on/off or programming the different appliances.

The system accepts spontaneous speech and it is user independent, so anyone can have a conversation with the virtual butler. On the other hand, the system is independent of the platform, therefore it may be developed, deployed, and maintained on multiple operating system platforms.

This paper is organized as follows: the specific corpus for this application is shown in Section 2, the architecture of our system is presented in Section 3. This will be followed by a description of the main modules of our system, which are Speech Recognition (Section 4), Understanding (Section 5) and Dialog (Section 6), ending with the speech recognition results with the recorded corpus (Section 7) and an example to understand the system operation mode (Section 8). Finally, the conclusions and further work are presented (Section 9).

2. CORPUS

In order to complete the preliminary speech recognition results in the domotic scenario, a specific corpus was recorded in the kitchen of the FAGOR Home Appliance facilities. It is composed by 48 speakers with 125 utterances per speaker. 3 tasks were considered: control of appliances in the kitchen (90 utterances per speaker), continuous digits (15 utterances per speaker) and 20 phonetically balanced utterances per speaker. On the other hand, 8 audio channels were recorded: 3 located in the kitchen (freezer, extractor hood and washing machine), 3 placed

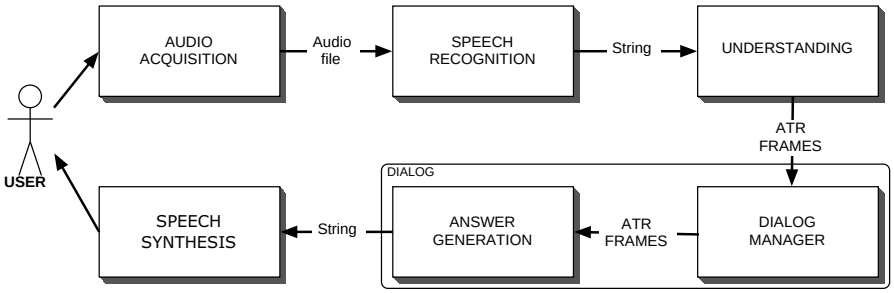

Figure 1. Architecture of the system

on the speaker, a close talk and 2 lapel microphones and finally 2 channels were recorded with a dummy (right and left ears) placed close to the speaker. In all cases, the frequency sample is 16 KHz and the audio signal is coded with 16 bits.

Three acoustic environments were considered in the recording: no appliances on (E0) with 45 dBA of typical Sound Pressure Level, SPL, extractor hood on (E1) with a 60 dBA of typical SPL, and washing machine on (E2) with a 62 dBA of typical SPL. Also, 2 speaker positions were defined: P0, in front of the washing machine, and P1, in front of the extractor hood. 15 utterances for each speaker were recorded for each position and acoustic environment.

3. ARCHITECTURE OF THE SYSTEM

The system is composed of several modules in a serial architecture, as shown in Figure 1. When the user speaks, the audio signal is captured and converted into a electrical signal by an Audio Acquisition module. This signal is decoded into a text string by the Speech Recognition module. The next step is to understand what the user expects, that is, to extract the meaning from the text string (Understanding module). The obtained information is processed by the Dialog Manager and thus an answer is generated by the Answer Generation module. This answer, in text format, is sent to the Speech Synthesizer which allows the system to ask the user for information, generating, in this way, a dialog between the human and the machine. Also other output options can be activated from the string obtained by the Answer Generation module, but in all cases this feedback pretends to work as if the user was speaking to a human butler.

It is worth to stand out that the whole system is task independent, so there is a possibility for extended plug-and-play just adapting the Answer Generation module, which is the only module that involves internal files. Furthermore this system is independent of the platform, so it may be run on multiple operating system platforms. The code of the system is Java and C so it is hardware independent application. In fact, it has been tested on Mac Os X, Linux ...

The main modules in this system, which are Speech Recognition, Understanding and Dialog, are explained in the following sections.

4. SPEECH RECOGNITION MODULE

In our system, the Speech Recognition module has a capital importance, otherwise the system will be wrong from the first stage on. This fact can generate errors in Understanding module and consequently produce mistakes in the dialog. That is, the error is propagated through all the system.

The technology of Automatic Speech Recognition systems, ASR, has evolved dramatically over decades of research. In an ASR system, we can establish two model levels which interact when searching for spoken utterances hypotheses. The first level is the acoustic level, which usually is generally assumed to follow a Hidden Markov Model, HMM, distribution, many authors have contributed to the development of these models [1, 2, 3]. The second level is the language model and referes to how words are concatenated, which can as simple as static rules of grammar, or more complex as an statistically estimated N-gram grammar [4]. This last statistical model has been selected for this work.

Another essential topic is to know how the people speak to the virtual butler. That is the reason to carry out a survey of the conversation style that people would hold on with a system. Around 250 people virtual dialog samples have been considered. A grammar has been generated with the obtained dialog samples so as to create thousands of sentences. A k-Testable in the Strict Sense Language Model (k-TSS LM) language model has been produced by inference. In this work, we make use of k-TSS language model [5] since in contrast to a probabilistic n-gram model, the former one preserves the syntactic constraints.

With regard to the acoustics models, the noises in a kitchen can be modeled and introduced in the recognition system providing a more robust system.

5. UNDERSTANDING MODULE

This module tries to get the meaning from the text string output of the Speech Recognition module looking for key words. In order to extract all this information provided by the string two concepts have been used: *Attributes* and *Frames*.

5.1. Attributes

The *Attributes* are collections of key words that have a similar meaning where each key word belonging to one and only one attribute. An example of the use of *Attributes* is shown in table 1.

Table 1. Example of an attribute

Attr_ElectricalAppliance
washing_machine
dishwasher
oven

5.2. Frames

The *Frames* are files which contain all needed information to determinate and to carry out a task. So, there is a file for each task. Each frame contains all the attributes that the system needs for the corresponding task. Apart from these attributes, each frame also contains some tags which give aid to set the task. Moreover, these frames enclose tables that help the Understanding module.

5.3. Operation mode of the Understanding module

An example is introduced in order to explain the Understanding module. The input of this module is a text string, as shown in the example 1. The Understanding module would process the input and detect the key words. In the example, the key words have been underlined to emphasize.

Example 1 *I want to turn on the oven.*

Once the key words have been detected, the Understanding module looks for the corresponding attributes. For the Example 1 the attributes are shown in the table 2.

Table 2. Attributes associated to Example 1

KEY WORD	ATTRIBUTE
turn on	Atr_Switch_on
oven	Atr_ElectricalAppliance

When the Understanding module has obtained all attributes, it checks the frames looking for the most appropriated task among all the tasks. In the example the task would be to switch on the oven.

Once the task has been determined, the Understating module completes the necessary attributes to carry out the task. In the example the Understanding module obtains another attribute: the function of the oven (grill), thanks to the tables. As a summary, the Understanding module output for the example is shown in table 3.

Table 3. Output for Understanding module to Example 1

FRAME	Switch on the Oven
ATTR : Switch_on	turn on
ATTR : ElectricalAppliance	oven
ATTR : Oven_Function	grill

6. DIALOG MODULE

From the point of view of the complexity of the dialogue, 5 levels of complexity were defined according to the used technique to represent the dialogue acts [6]. Many dialog systems have been developed during the last years covering most of this complexity levels [4, 7].

In our system, the Dialog module has two components: Dialog Manager and Answer Generation, as it can be seen in figure 1.

6.1. Dialog Manager

The operation of the Dialog Manager can be seen in Figure 2. First of all, the Dialog Manager requests a task to the Understanding module. When the task is determined, the Dialog Manager asks for the information of it to the Understanding module. With this information, an answer is generated in the Answer Generation module, which is processed by the Dialog Manager.

The answer of the user can be an affirmation, a negation or it can be a sentence without any affirmation or negation. When the speaker gives an affirmative statement, the Dialog Manager looks for more information about the task. If it obtains some information, the Answer Generation produces a new answer for the user; otherwise the Dialog Manager checks if all needed information is available. If it is not, the Answer Generation module asks about the missing information.

Otherwise, if the user gives a negative statement or the sentence does not include any affirmation or negation, the Dialog Manager looks for a new task, looking for all the needed information about the new task.

6.2. Answer Generation

The Answer Generation module produces an answer with the information provided by the Dialog Manager. The structure of this answer is defined by the task and the

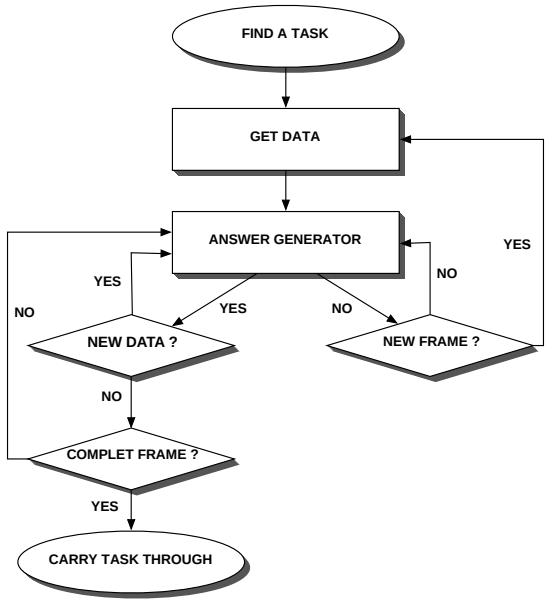

Figure 2. Way of operation of the Dialog Manager module

available information of the system, requesting for confirmation. Therefore, the Answer Generation module is task dependent.

7. SPEECH RECOGNITION RESULTS

Word acoustic models are built from a set of left and right context-dependent and context-independent units. Each unit is modelled by one-state continuous density HMMs with 16 Gaussians. In addition, two silence models for long and interword silences are considered. Each phoneme is modelled by the left contextual unit, the in-contextual unit and right contextual unit. So, for example, the word acoustic model for Spanish digit "dos" ("two") can be obtained by the concatenation of the following units: $/\#< d/\ /d/\ /d > o/\ /d < o/\ /o/\ /o > s/\ /o < s/\ /s/\ /s > \#/$, where $\#$ is the silence unit, $/ < /$ is the left context-dependent unit, $//$ is the context-independent unit and finally $/ > /$ is the right context-dependent unit. The used training corpus is composed by 7,970 utterances of the SpeechDat-Car training corpus [8] and 5,198 of the Albayzin training corpus.

The language model is a stochastic grammar, trained with the CMU Toolkit (v2.0) with 47,657 utterances generated by the virtual dialog samples. So, 363 1-grams, 1,819 2-grams and 5,379 3-grams are considered, giving a perplexity of 4.25.

Since a domotic scenario is usually used by one person, acoustic model adaptation techniques can provide a very good performance in speech recognition. In order to study it, MAP, *Maximum A Posteriori* algorithm is proposed [9], which provides better results than ML, *Maximum Likelihood*, technique when there are not enough data.

The recognition task is carried out only with the control of appliances utterances, splitting all of them into two corpus: adapting and testing, including for both of them signals of all the positions (P0 and P1) and all the acoustic environments (E0, E1 and E2). The Word Error Rate (WER) results for the different experiments are presented in Table 4, where the "baseline" column results have been obtained without any acoustic model adaptation but CMN, *Cepstral Mean Normalization* [10]. The "MAP" column refers to the WER obtained via MAP adaptation. Finally, "Improvement MAP" column presents the WER reduction concerning baseline. It can be observed that an important improvement is achieved with all different microphones.

8. AN EXAMPLE OF A DIALOG

In this Section we will show an example of a dialog. In this case, a user wants to turn on the oven, and the system requests for the needed information. The key words are underlined for clarity and the information that is obtained by the system it is shown in Table 5.

t_1 USER: I want to cook the chicken
BUTLER: Which temperature do you want to set the oven within the grill function?

t_2 USER: No, what I want to select is the solera function
BUTLER: Which temperature do you want to set the oven within the solera function?

t_3 USER: at eighty degrees
BUTLER: Do you want to set the oven at eighty degrees?

t_4 USER: yes, for two hours time
BUTLER: Do you want to set the oven for two hours time?

t_5 USER: It is ok.
BUTLER-> Do you want something else?

...

9. CONCLUDING REMARKS AND FURTHER WORK

We have introduced a virtual butler controlled by speech which allows spontaneous speech with an independent user. The goal is to control the appliances of a kitchen. The input and the output of the system is speech, although some different ways of communication can be used (dynamical graphic displays...).

Some experiments were carried out with a specific corpus recorded in a real kitchen, showing that speaker acoustic model adaptation can improve the results in speech recognition (up to 67.31% of improvement concerning baseline).

In this work the involved language has been the Spanish but it has been considered as a further work, that the

Table 4. Speech recognition results

	Baseline	MAP	Improvement MAP (%)
Close-talk	1,91	1,34	29,69
Left lapel	4,43	1,50	66,13
Right lapel	4,88	1,59	67,31
Left Dummy	72,79	29,22	59,86
Right Dummy	54,56	22,91	58,00
Washing Machine	48,95	23,50	51,99
Extractor hood	62,90	22,98	63,47
Freezer	71,73	29,24	59,23

Table 5. Frames and attributes obtained by the system for the example in Section 8

	FRAME	ATTRIBUTES
t_1	Switch on the oven	cook (atr_SwitchOnOven) chicken (atr_food) grill (atr_OvenFuction)
t_2	Switch on the oven	solera (atr_OvenFuction) no (atr_Negation)
t_3	Switch on the oven	eighty degrees (atr_OvenTemperature)
t_4	Switch on the oven	for two hours (atr_OvenTime) yes (atr_Afirmation)
t_5	Switch on the oven	It is ok (atr_Afirmation)

user could employ another languages such as Basque or English.

Another future work can be the virtual-butler to warn about problems, such as, when it is a power cut or water or gas leak. In this situation the dialog acquires a vital importance, because in these situation it is the virtual butler who takes the initiative.

10. ACKNOWLEDGEMENTS

The authors want to acknowledge FAGOR Home Appliances for the cooperation in this work.

11. BIBLIOGRAPHY

[1] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," in *Inequalities III: Proceedings of the Third Symposium on Inequalities*, Oved Shisha, Ed., University of California, Los Angeles, 1972, pp. 1–8, Academic Press.

[2] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, 1976.

[3] L. R. Rabiner, *A Tutorial on HMM and selected Applications in Speech Recognition*, chapter 6.1, pp. 267–295, Morgan Kaufmann, 1988.

[4] J. Glass, T. J. Hazen, and I. L. Hetherington, "Real-time telephone based speech recognition in the jupiter domain," in *Proc. ICASSP*, 1999.

[5] I. Torres and A. Varona, "k-tss language models in a speech recognition systems," *Computer Speech and Language*, vol. 15, no. 2, pp. 127–149, 2001.

[6] M. Dzikovska J.F. Allen, D.K. Byron, "Towards conversational human-computer interaction," *AI Magazine*, 2001.

[7] J. Glass and E. Weinstein, "Speechbuilder: Facilitating spoken dialogue systems development," in *Proc. EUROSPEECH*, 2001.

[8] Asuncion Moreno, Borge Lindberg, Christoph Draxler, Gael Richard, Khalid Choukri, Stephan Euler, and Jeffrey Allen, "Speechdat-car. a large speech database for automotive environments," in *Proceedings of LREC*. Athens, Greece, June 2000, vol. 2, pp. 895–900.

[9] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriory estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 291–298, Apr 1994.

[10] N. Hanai and R. M. Stern, "Robust speech recognition in the automobile," in *in Proc. ICSLP*, 1994.