TELEPHONY APPLICATIONS WITH SPEECH RECOGNITION

Nuno Borges, Rui Lopes,
Beatriz Henriques, Paulo Valente, Rui Amaral,
Emanuel Martins, Fernando Perdigão e Luís de Sá
Departamento de Engenharia Electrotécnica, Universidade de Coimbra
Instituto de Telecomunicações - Pólo de Coimbra
Polo II, Pinhal de Marrocos , 3030 COIMBRA, PORTUGAL
Tel: +351 39 796236, Fax:+351 39 796293,
Email: nuno.borges@it.uc.pt ,rui.lopes@it.uc.pt, beatriz.henriques@it.uc.pt,
paulo.valente@it.uc.pt, rui.amaral@it.uc.pt, emanuel.martins@it.uc.pt,
fernando.perdigao@it.uc.pt, luis.sa@it.uc.pt

Abstract: In this paper, we present and describe several computer telephony applications using speech recognition. These applications were developed under a research project carried out in collaboration with Portugal Telecom. Two possibilities have been explored in the developing of speech recognition applications. In the first one, speech recognition was implemented only with the help of software. In the second one, we used hardware equipped with DSP's. Both possibilities support *Word Spotting* and *Barge-in*. In addition to the telephone applications, the generic tools that have been built to develop those applications are also presented.
Keywords: Computer Telephony, Speech Recognition, *Word Spotting*, *Barge-in*. Developing Tools.

1. Introduction

Automatic speech recognition plays a major role in the developing of today's computer telephony applications. Powerful development tools for speech recognition technology allow applications to reach the market in a short time that was inconceivable just a few years ago.
Speech recognition is also a necessity. As competition grows, network operators and service providers can not afford not to build applications incorporating speech technology. This is especially true in the mobile telephony environment. Cellular phone companies want services, which differentiate them from other providers and increase their number of subscribers. Also by providing full automatic services, companies can reduce their operating costs. On the other end, services are available 24 hours a day increasing customer's satisfaction.
Portugal Telecom and Instituto de Telecomunicações run a project to explore the introduction of speech recognition technology in telephone applications. The project, referred to as SAIT, was entitled "Sistemas Avançados de Informação Telefónica". In this paper we describe some prototype applications that were developed under this project. Two implementations approaches have been considered and experimented. One implementation, referred to as PC-based, uses a PC (Personal Computer) with a telephone interface card. In this case the recognition tasks are made by software running in the PC. Since speech recognition tasks require a considerable effort from the PC, the maximum number of channels that can be simultaneously processed in one machine is limited to 2 up to 4 (Pentium 120MHz). The second approach is to equip the PC with special hardware for digital signal processing (DSP) capable of performing speech recognition. In this case, referred to as DSP-based, it is possible to handle up to 30 channels simultaneously as the computational intensive tasks involved in speech recognition algorithms are performed by dedicated DSP hardware. In the beginning of the project it was decided to construct applications in both systems, in order to understand better the advantages and limitations of each system.
The outline of the paper is as follows: in section 2 we describe the two systems used for speech recognition focusing both the algorithms and the hardware. In section 3 and 4 we describe several applications and tools that were developed: a tool for audiotext service development called "Info Maker" (3.1); a PBX-based voice mail service (3.2); a graphical tool for general application development (3.3); a phone wake-up system via web (3.4); an audiotext service called "Talking Pages" (4.1); and a commercial information service called "Tele-Balcão" (4.2). Finally in section 5 we conclude.

2. Speech recognition

2.1. PC-Based and DSP-based

In the PC-Based system, the speech recognition is made by software using continuous density Hidden Markov Models (HMM) with Linear Predictive Coding Cepstral Coefficients for acoustic signal modelling [1,2]. The used HMMs were word models for the set of digits "0" up to "9" and the "sim"/"não" words. The recogniser can also perform *Word Spotting* using a limited set of garbage models. The models were trained on a telephone speech database [3] reaching 98% recognition accuracy.
In the DSP-based system we used the ANTARES PC card from DIALOGIC® [4] which is a DSP platform supported by an open development environment oriented to commercial telephony systems. This card

basically consists on an array of four 32-bit floating-point signal processors (TMS320C31) running proprietary speech recognition software from Voice Processing Corporation [5]. This later system has less recognition accuracy that the PC-based system, around 90%, explained possibly by a reduced number of training words used during the construction of the models for the recognizer.
With this system it is possible to design speech applications with the capability to recognise digits and the words "sim"/"não". It is also possible to use *Word Spotting* and *Barge-in* (by voice or *DTMF-Dual Tone Multiple Frequency*).

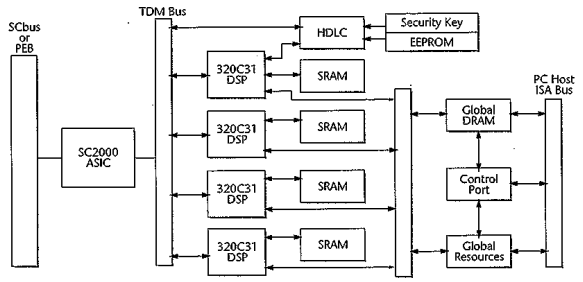

Figure 1: Antares PC-card block diagram

2.3. *Barge-in*

"Barge-In" is the ability of a system to recognise speaker input while it is playing an outgoing prompt. When a caller speaks over a prompt, the speech recogniser must be able to distinguish between the prompt itself and the speaker's voice. So, the system needs to cancel the echo of the prompts. Before attempting any recognition the characteristics of echo should be measured by sending a calibration burst signal over the telephone line. After this, the recogniser doesn't ear its own prompts allowing a faster response time to human vocal orders.

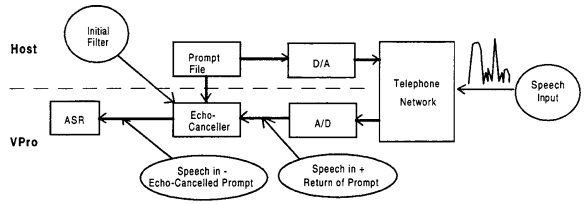

Figure 2: *Barge-in* scheme

2.4. Word-Spotting

In a word-recognition system it is desirable to allow the user answer to the questions with natural sentences like " My choice is one, please!" instead of force single word answers like "one". To give more freedom to the users, speech recognisers use the "word-spotting" technique. "Word-spotting" is the capability to detect (recognise) a given word embedded in a sentence or in noise.

3. DSP-based speech recognition

In this section we describe ITV (Interactive Voice Telephony) applications built to run in a PC equipped with an add-on card containing 4 DSPs running commercial recognition software and an add-on card interfacing with several phone lines.
As the speech recognition is made by dedicated hardware, the PC only have to control globally the system allowing the managing virtually 30 to 60 phone channels simultaneously.
All the applications made in the DSP-based system support "Barge-in" (by DTMF or voice) and "Word-spotting". For all the audio services it is generated log-like files that collect users activity. Loading this files to a database system allows the administrator see graphically the statistics of activity for each service.

3.1. Info-Maker

Info-Maker (voice **info**rmation system **maker**) is an audio-text service. Audio-text is probably one of the most common services that exists in computer based telephony. Info-Maker was built taking attention an easy voice interaction and the facility of manage the audio-text service without stopping the service itself.
Since audio-text services usually have a tree-like structure, we've constructed Info-Maker supported in the directory-tree structure of the operating system (see figure 3). Info-Maker is a MS-DOS application running nicelly in a Windows/Windows95 operating system.

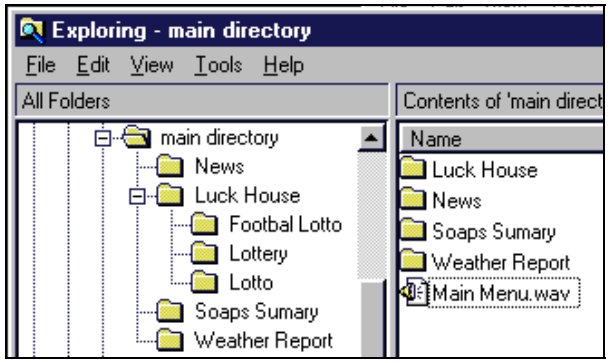

Figure 3: Tree structure example of Info-maker

Info-Maker starts from a predefined disk-directory (the main directory) and jumps from directory to directory based on the vocal commands (similar to the **`cd <directory>`**, **`cd ..`**, **`cd \`** operating system commands). Each disk directory must only have 1 file containing a voice prompt and one directory for each option stated in the prompt file. The last 2 options are always return to previous directory (like **`cd ..`**) and to the main directory (like **`cd \main_directory`**).
With this directory-tree structure info-maker allows a multiplicity of information services beginning in a single start directory.
The system can be altered while running, just adding/deleting directories of prompts to the file system.

3.2. Voice Mail

Another implemented application running in MS-DOS is Voice Mail service. The application acts like usual voice mails services but allowing also voice commands. Voice Mail retains voice messages when a subscriber can't answer the phone. The subscriber that wishes to hear his messages provides a password. After the validation of this password, he is presented with a menu allowing the management of his mail. In this menu, the subscriber can listen to all the mail available in his mailbox, record (keep) the most important messages, change the password or change the personal greeting message.
A set of maintenance files (recording errors that may occur) is available to simplify the administration tasks.

3.3. Computer Telephony Constructor

Computer Telephony Constructor (CTC) is an application constructor toolkit with a graphical environment running in Windows95. The toolkit was made in way such no programming skill is required from the application designer. The designer needs only to know how all the pieces of his application interact. Each application is build by linking telephony modules in a flowchart philosophy. To make an application consists in select common modules such as "*Detect a Call*" or "*Play a Sound File*", that are represented as icons, and drag and drop these choices into place.

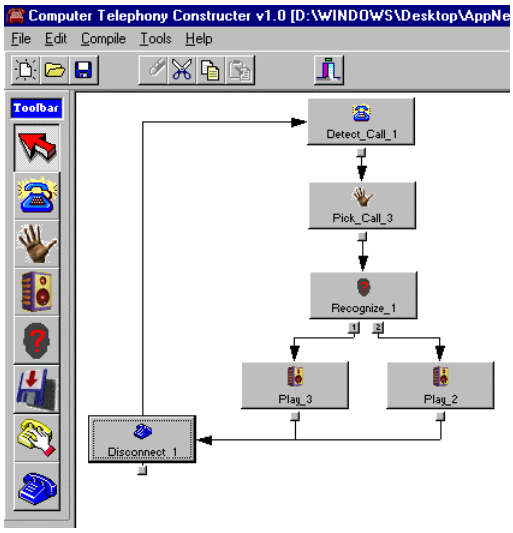

Figure 4: application example with CTC

With CTC it became possible to built the applications Info-Maker and Voice-Mail in a short period time (2 days or less), contrasting with the one month or more time that took to program with C language each application for the first time.

3.4. WebWakeUp

"WebWakeUp" is an application that implements a wake-up system, a common service provided by phone operators, but that can be activated through web hypertext pages instead of the intervention of a human operator. "WebWakeUp" is an example of the interaction between phone lines and the internet that should be expected in the near future, result of the evolution on services provided by the phone telecommunications service providers.

4. PC-Based Speech Recognition

In this section we describe ITV applications built to run in a PC running speech recognition software and equipped with an add-on card interfacing with 4 phone lines. In this case speech recognition and overall control is made only by the microprocessor of the PC, what limits the capacity of the system. In this case the number o channels that can be processed simultaneously depends of the speed of the processor. A Pentium 120 MHz can support speech recognition of about 4 phone channels simultaneously without degrading the quality of service.
The PC-Based applications support "Barge-In", by DTMF only, and "Word-spotting".

4.1. Talking Pages

This application consist of an information service by phone. It functionality mirrors the philosophy followed by the Tele-text service. That is, to each information service a different number is assigned, which a user that wants to access to a following service must know. When a user doesn't know the number of the wanted page, the system will provide a menu indicating page numbers of general services and their contents.

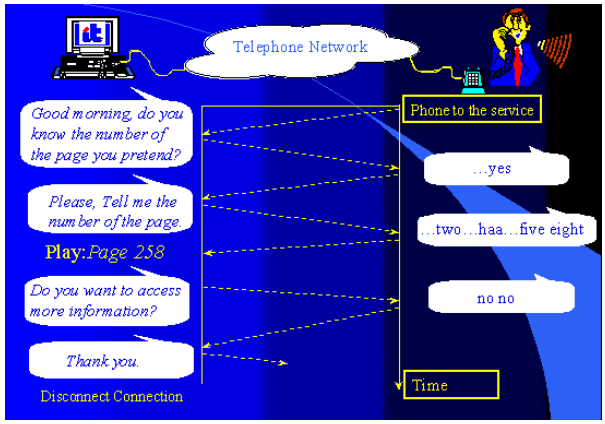

Figure 5: Example of dialog in Talking Pages

"Talking Pages" works in a philosophy similar to the Info-Maker (see section 3.1). The speech recognition runs exclusively in a PC with Windows95 with no need of an auxiliary processing board, what makes this system cheaper than that of Info-Maker, but limited to a reduced number of phone channels that the microprocessor can handle.

4.2 "Tele-Balcão"

The main purpose of "Tele-Balcão" project was the development of a system oriented to give commercial information over telephone network. It was made an effort to simplify the navigation through the information making this service intuitive and easy to understand.
How does it work? When a customer calls to "Tele-Balcão" service the system play information about several product categories to be chosen. Specified one, a menu will present all products in that class and when customer decide what product he wants to know about, a prompt file with technical and marketing information will be played.

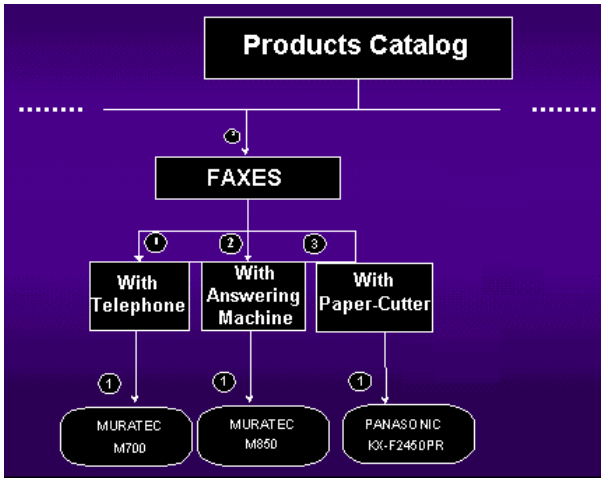

Figure 6: "Tele-Balcão" tree structure example.

5. Conclusion

We've shown in this paper several applications that we developed in the area of Automated Interactive Telephony based in speech recognition technology. We present how is possible to implement commercial applications based in computer voice telephony.
Two distinct platforms were used: PC-Based speech recognition and DSP-Based speech recognition. While PC-Based is cheaper to implement and has a good speech recognition accuracy, needs computers with high performance and have limitations in the total number of phone channels that can handle simultaneously. The DSP-Based is more expensive, but can handle a larger number of phone channels, which reduces the cost per phone channel, and doesn't require expensive computers, since speech recognition is done by dedicated DSP hardware. We can conclude that PC-Based solution is the ideal platform to small business and DSP-Based solution is the only possible choice for large systems.

6. References

[1] L. R. Rabiner and B.H.Juang, "An Introduction to Hidden Markov Models", IEEE ASSP Magazine, pp. 4-16, Jan. 86.

[2] L. R. Rabiner, "A Tutorial on HMMs and Selected Applications in Speech Recognition", Proc. IEEE, vol.77, No.2, Feb.89.

[3] R. Amaral, F. Perdigão, P. Placido, E. Sá Marta, L. Vieira de Sá, "Automatic Segmentation and Labelling of a Portuguese Telephone-Speech Digit Database", Recpad97, 9th Portuguese Conference on Pattern Recognition, 1997.

[4] DIALOGIC, "Products & Services Guide", 1997

[5] DIALOGIC, "Vpro on Antares Software Reference", 1997.