LEIAME: A SPEECH SYNTHESIZER SYSTEM CONTROLLED BY VOICE TO HELP THE DISABLED IN THE READING PROCESS

Giancarlo Cerutti Panosso, Marcelo Ambrosio Stefanello

Universidade Regional Integrada do Alto Uruguai e das Missões – URI, Brazil

Abstract
This paper presents the development of the LeiaMe, a speech synthesizer system driven by voice, which aims to assist visually impaired and/or motor disability users in the reading process. The idea is to replace the lack of vision of the user by their hearing capacity and, in parallel, replace motor skills by voice commands, eliminating the physical and visual contact with the computer. The developed tool makes use of mechanisms of speech synthesis and recognition, which allow the user to search for files in text format and summarize the contents of these files, resulting in an instrument of practical utility and collaborating significantly with the ideals of inclusion and accessibility.

Key words: *speech synthesizer, speech recognition, disabled, accessibility*

1. INTRODUCTION

According to statistics data from 2010, approximately 18.75% of the Brazilian population (35,774,392 people) reported having some type of visual difficulty or full disability. Parallel to this, 6.95 % (13,265,599 people) of respondents said that they had difficulties or physical disabilities (IBGE 2010). Worldwide in the same year, according to the World Health Organization, about 285 million people had some type of visual impairment, with approximately 39 million (0.58% of world population) with complete blindness and 246 million (3, 65% of the world population) with low vision (WHO 2014).

This reality has certainly reflected directly in education. The majority of the materials supplied and / or referenced in educational institutions contain mainly text, whether printed or digital. It requires from the students handling and reading skills. Consequently, students with both visual and motor limitations have difficulties using this material, or read them using your computer.

This portion of visually impaired and/or motor disability users, are those who have access to formal education or those who seek knowledge independently, require for practical use special reading tools that are beyond the conventional pattern of “visual reading” and to provide a more practical and effective acquisition of knowledge and access to information form.

In this sense was developed the LeiaMe, a speech synthesizer system driven by voice, to aid visually impaired readers and motor disability users. This application enables the user to choose files in text format and run the speech synthesis thereof. The goal is to provide access to writing without visual reading and physical contact with the computer, using only the user listening and speaking, respectively.

To support the implementation of this proposal, the activities were carried out: analysis of the operation of accessible national applications developed based on speech synthesis and recognition; literature relating to the concepts involved in the processes of speech synthesis and recognition, and literature related to technology Microsoft .NET speech development-level, used to provide the resources of speech synthesis and recognition software developed in. Finally, we carried out actions in planning and development of the LeiaMe system.

These activities are presented in this paper.

2. ANALYSIS OF SOME ACCESSIBLE NATIONAL TOOLS

This section briefly described three national tools recently developed, aimed at providing accessibility in the use of computers for users with visual or motor disabilities, and using mechanisms of synthesis or speech recognition. These tools are DOSVOX®, Motrix®, and Mecdaisy®.

After analyze them, it was found that DOSVOX and Motrix systems have a broader approach, with the purpose of enabling the use of the computer as a whole, facilitating the execution of many common tasks of everyday life. However the DOSVOX is specific for the blind users and uses only speech synthesis capabilities, while the Motrix is intended to paraplegic users, having only interaction mechanism for speech recognition. Thus, the blind user must operate the DOSVOX using keyboard while the paraplegic user must operate the Motrix visualizing the actions occur on the computer screen.

The Mecdaisy, in turn, is also directed to blind users, but with purpose speech synthesis file that contains text. However, it only supports files with specific format (Daisy), not being able to synthesize more common files like ".Txt" or ".Doc". Furthermore the user needs to use the keyboard or mouse to operate the system. On the other hand, the Mecdaisy presents interesting control features of the synthesis of the file, allowing the user to pause and resume synthesis at any time, also allows the forwarding and rewinding the paragraphs of text.

3. THE SPEECH SYNTHESIS METHOD

The speech synthesis (in English "text-to-speech" or "TTS") can be defined as the artificial generation of human speech, usually from text. Egashira (1992, p.2) quotes that "[...] voice systems from text accept as input a unlimited vocabulary text and, through extensive linguistic processing, produce a synthetic speech of the respective text".

The basic role of speech synthesis is to facilitate the interaction between human and machine, making the process faster and more natural communication. Often used in applications related to accessibility, being an alternative interface for blind users.

According to Microsoft (2011), a synthesis system based on text-to-speech is composed of two major parts: a front-end, responsible for textual analysis, and a back-end, responsible for the generation of sound itself.

The front-end parses the sequence of characters and determines the value taken by these characters, especially in the matter of punctuation symbols, numbers, among others. Also identifies statements, questions, pauses and other grammatical details.

The back-end has a rather different task: from the textual analysis by the front-end it generates the sounds corresponding to the input text. These sounds are obtained from a database with a huge amount of sound segments previously recorded.

4. THE CASE FOR SPEECH RECOGNITION

According to Deng and O'Shaughnessy (2003, p. 433), the technique of speech recognition is the process of converting an acoustic signal, obtained by a microphone (or a microphone array), a telephone, or some other receiver in a textual sequence, typically in the form of a sequence of words.

This feature has the main application serve as input when it is not possible or convenient to use your hands. It is commonly used in applications aimed at users with motor disabilities or inability to use their arms.

Similar to speech synthesis, speech recognition is also composed of a front-end and back-end. According to MICROSOFT (2011), the front-end makes the audio stream processing, isolating the sound segments that are likely part of speech and converts them into a series of numerical values that characterize the vocal sounds in the signal.

On the other hand, according to MICROSOFT (2011), the back-end is a specialized search engine that takes the output produced by front-end and performs a search through three databases: one for the acoustic model, one for the lexical part, and one for the language model.

In relation to speech synthesis, speech recognition has a very significant difference: it is a limited process. This is because it requires a grammar, in other words, a mechanism that list which terms (tokens) the application will recognize. Any command that does not obey this listing will be waived by the recognizer.

5. .NET MICROSOFT SPEECH TECHNOLOGY

The .NET Speech is a technology created and maintained by Microsoft Corporation® company that provides resources for speech synthesis and recognition, both for use in their products (operating systems, office suites etc.) and for the development of applications that using such resources.

For purposes of this work, specific funds were used for the development of applications that implement the speech synthesis and recognition, achieved through the use of System.Speech.Synthesis and System.Speech.Recognition namespaces, respectively.

According MICROSOFT (2011), the System.Speech.Synthesis namespace contains classes that allow you to initialize and configure a speech synthesis engine, creating prompts, generate speech, respond to events and modify the features of voice. In addition, you gain control over many aspects of the expression of voice, including the volume, rate of speech, emphasis, intonation and other attributes of speech.

All these features available to the developer occur thanks to the methods, properties and events that the .NET Speech classes feature. From these classes, objects that allow implementing such functionality directly in the application code can be created.

Also according to Microsoft (2011), the System.Speech.Recognition namespace provides functionality that allows to acquire and monitor the speech input, creating speech recognition grammars that produce results both literal and semantic recognition, capture event information generated by the recognition of speech and configure and manage the speech recognition engines. Just as occurs in the synthesis, the features of the recognition namespace are provided due to the set of methods, properties and events of the corresponding classes.

6. DEVELOPMENT OF THE LEIAME

This section gives the main aspects related to the development of the LeiaMe system process such as: overview, tools used in the development, general operating structure and internal architecture of the application.

6.1. Overview

The LeiaMe is a software developed to run on Windows® (Vista operating system versions or higher) and allows the user to access a file in text format (.txt) saved on your computer or media connected to it, and enable execution and control of the speech synthesis of the text contained in the selected file. This entire process is controlled only by voice commands.

The LeiaMe is based on the processes of speech synthesis and recognition for user-application interface. Therefore, it has no graphical interface and is executed in the form of window "cmd" (prompt) Windows. For their use are necessary microphone and speaker attached to the computer (or already existing on the machine).

6.2. Tools used in the development

The tools needed to develop the LeiaMe were:

• Development Environment for Visual C# 2010 Edition Expressing®, used for coding, testing and integration of all components of the application;

• C# programming language to build the source code of the application;

• DBMS SQL Server 2008 Edition Expressing® to create the database responsible for storing settings for speech synthesis;

• StartUML®, building diagrams and flowcharts of the planning steps;

• TTS Agents "Microsoft Anna" and "ScanSoft Raquel". Its function is to send the voice that performs the actual reading of the texts. The first provides the voice in English and the second in Portuguese.

All these tools were performed on a computer type notebook with Dual Core T4500 2.3GHz processor, 2GB RAM and Windows 7 32-bit operating system.

6.3. General operating structure

Initially the requirements and scenarios (use cases) provided in the LeiaMe use were defined, coming to a logical structure in flow chart form, representing the whole "script" to use the software. This structure is shown in Figure 1.

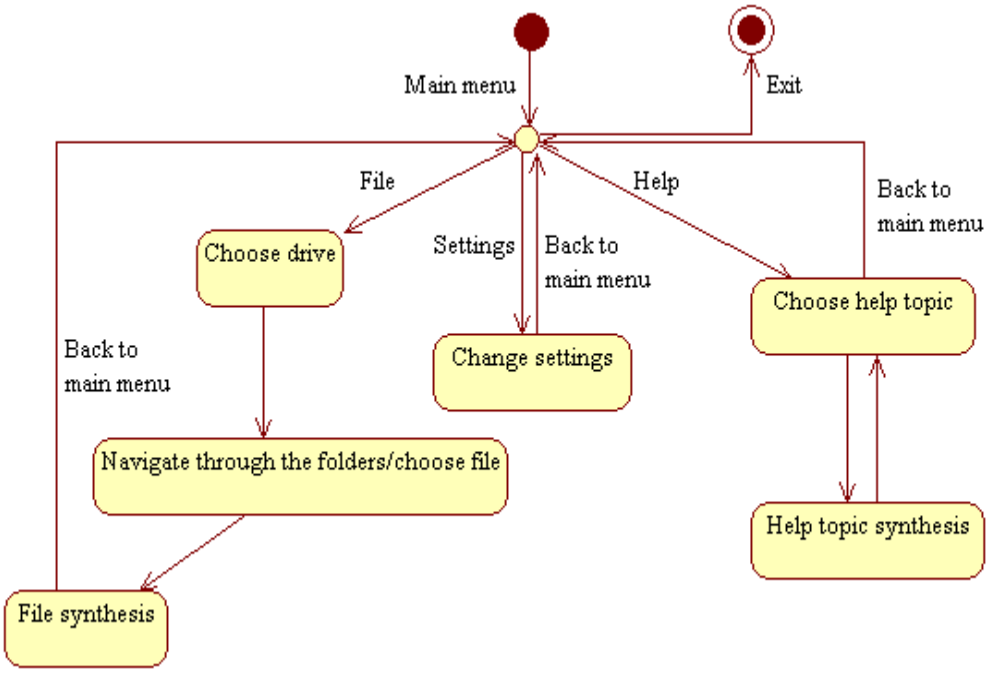


Figure 1: Structural diagram of the LeiaMe.

As can be seen in Figure 1, the application has a main menu with four options to the user: search for a file (File), change settings (Settings); listen help topics (Help) and close the application (Exit).

When choosing the "File" option, the user will be directed to select the logical drive of the computer you want to open it. This option allows files to be selected both on the computer hard drive as media connected to it, such as CDs and flash drives. After selecting and entering the selected drive, the user

can navigate through the folders on the drive to find the file you want to hear. When you find the file, simply issue the command open file. A file open time available to the user all the options control the synthesis of text, such as start, pause, stop, etc. After ending a synthesis, the user must close the file being automatically redirected to the main menu.

The "Settings" from the main menu allows the user to change one of the three properties of the synthesizer, which are: language, volume and speed. After setting a new value for one of these properties, the user is automatically redirected to the main menu.

The "Help" option allows the user to have access to help topics of the application to navigate through them and perform the synthesis of the desired topic. After synthesizing a topic being closed, the user is redirected to the list of topics. To return to the main menu, the user must launch the command "exit help".

Finally, the "Exit" option immediately terminates the execution of the application.

6.4. Internal architecture

The structure of the LeiaMe system from the point of view of the internal workings (architecture) is shown in Figure 2.

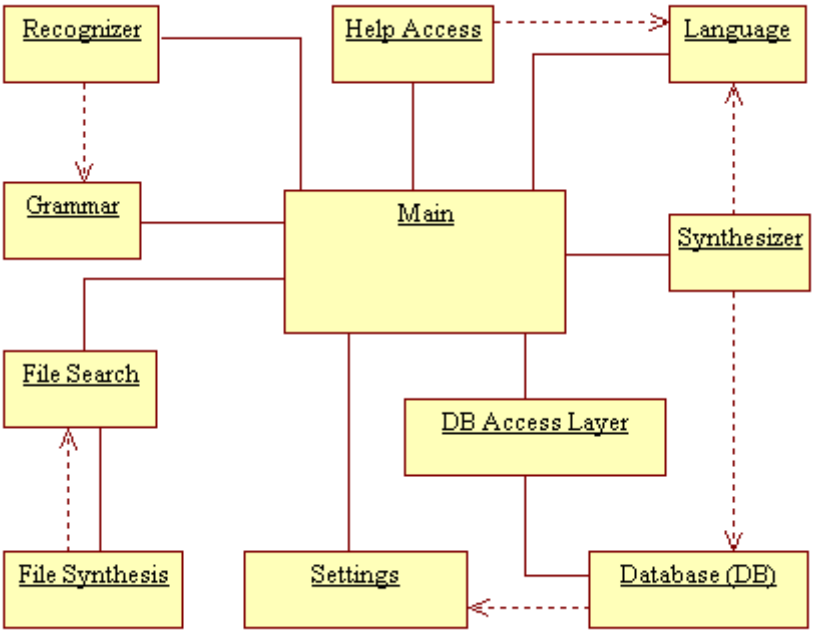


Figure 2: Internal architecture of the LeiaMe.

In Figure 2, the blocks represent both the grouping of similar features that form modules as entities involved in implementation. Solid lines represent interactions between the blocks and the dashed arrows represent dependency between blocks, leaving the dependent and pointing to what causes addiction. These modules are described below.

6.4.1 Main module

The main module is responsible for the logical grouping of the functionality of the other modules. It is the "center" of the application (main class).

6.4.2 Grammar

The grammar is the set of commands (tokens) that will be recognized and that trigger the execution of some functionality in software. It is directly connected to the main module, as this makes the combination of it with the recognizer (which is dependent on the grammar).

6.4.3 Recognizer

The recognizer is responsible for identifying voice commands issued by the user and perform tasks related to the identified module commands. Uses the event "to recognize speech" to trigger the execution of a task. It is directly connected to the main module; it is the means of communication towards user-software in all scenarios of interaction. It is also linked to grammar, having this dependency to identify only the tokens that have meaning within the context of the application.

6.4.4 Database

The database has the only purpose this application storing the configuration information of the synthesizer object (described later). This information relates to properties such as language and volume among others. It depends on the module "defines settings" to contain the record with the information needed for the synthesizer settings.

6.4.5 Access layer to the database

The function of the access to the database layer is performing any exchange of information between the application and the base, storing changes made and providing information to the configuration of the synthesizer. Parallel to this, this layer also aims to separate the application and the database in order to contribute to the integrity of the data, since it is the only channel of communication between base and main application.

6.4.6 Settings

It is responsible module for registering the settings and/or changes in the synthesizer settings. These settings are performed by the user through voice commands and are saved in the database by updating the registry that stores the information.

6.4.7 Synthesizer

The synthesizer object is responsible for performing the entire process of conversion in text-to-speech application. It aims to synthesize both the texts contained in the selected messages as feedback to the user files. It is directly connected to the main module; it is the medium of communication in the software-user direction for all interaction scenarios. Moreover, it has dependency configuration information contained in the database and is dependent on the expressions given by "language" module to the correct issue of feedback messages, corresponding to the language configured.

6.4.8 Language

The language module contains all expressions or parts of expressions that are passed to the synthesizer via the main module and arranged in such a way that they can form logical sentences to issue return messages to the user. These expressions have their versions in Portuguese, and the corresponding English language. It is connected to the main module, since this is arranging transfers and expressions to the synthesizer.

6.4.9 File search

It is the module that allows the user to select any file, in text format, to be synthesized. This module is responsible for listing the logical drives available, allowing the user to choose and enter one of them to run through the folders and select the desired file. This module is directly connected to the main module where the operation is requested to fetch a file.

6.4.10 File synthesizes

It is the module that allows the user to initiate and control the process of synthesizing the selected file in the module "File Search". It is attached to this, as it is triggered after selecting a file. Also depends on it, as to synthesize a file it must have been previously selected.

6.4.11 Help access

This module is responsible for allowing the user to browse the list of help topics and synthesize the desired topic. It is connected to the main module being fired from it. There is a relationship of dependency with the “language” module since it contains the name and content of relevant help topics available in both languages in the application.

7. CONCLUSION

The LeiaMe application presents an innovative approach to other speech synthesizers since it combines speech recognition with speech synthesis. Its operation is given entirely by voice, making your independent use of the use of input devices such as mouse and keyboard, and the same way, regardless of visual ability of the user and the help of others.

Compared to the tools studied in this work, although they are aimed at different contexts, it is concluded that the question “generalization” these applications are directed to a specific target audience and therefore more limited (Motrix for paraplegics; DOSVOX and Mecdaisy for the blind), while the LeiaMe is more generic, being aimed at users with either of two special needs.

Regarding improvements to be implemented in the LeiaMe, can highlight the inclusion of mechanisms to synthesize text files with other extensions beyond the “ .txt” as: “.doc”, “.docx”, “.rtf”, “.odt” and “.pdf”. This improvement is very desirable because it significantly increases the range of files supported by the application.

In relation to technology .NET Speech, its use in the development of the LeiaMe was very appealing because of the same resources to provide practical and effective programming, which makes the insertion process of speech synthesis and recognition software intuitive and relatively simple. As a counterpoint, we can mention the issue of technology does not yet possess mechanisms for speech recognition in the Portuguese language, which makes it necessary to grammar translation to other languages. In general, the use of this technology proved to be very satisfactory, meeting expectations and providing the resources necessary for the proper functioning of the LeiaMe.

Finally, it is concluded that all these factors and actions taken led to the successful development of the proposed system. It is believed that the LeiaMe can contribute significantly to greater inclusion and accessibility for the visually impaired and motor disability users, especially in matters of access to written information.

8. REFERENCES

Deng, L & O'Shaughnessy, D 2003, *Speech Processing - A Dynamic and Optimization - Oriented Approach*, Marcel Dekker, New York.

Egashira, F 1992, ‘Speech synthesis from text to Portuguese’. Dissertation to obtain the title of Master of Electrical Engineering, State University of Campinas, Campinas.

IBGE - Instituto Brasileiro de Geografia e Estatística 2010, *Resident population by type of disability, according to the situation of the household and age groups - Brazil - Census 2010*, viewed 15 April 2013,

<ftp://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Caracteristicas_Gerais_Religiao_Deficiencia/tab1_3.pdf>.

MICROSOFT n.d., *Speech Recognition*, MSDN Library, viewed 03 May 2013, <http://msdn.microsoft.com/en-us/library/hh361633(v=office.14).aspx>.

MICROSOFT n.d., *Speech Synthesis*, MSDN Library, viewed 28 April 2013, <http://msdn.microsoft.com/en-us/library/hh361644(v=office.14).aspx>.

WHO - World Health Organization 2010, *Global Data on Visual Impairments 2010*, viewed 15 March 2014, <http://www.who.int/blindness/GLOBALDATAFINALforweb.pdf?ua=1>.