A Computer-assisted Learning Software Using Speech Synthesis and Recognition in Brazilian Portuguese

Ana Siravenha[1] *, Nelson Neto*[1]*, Valquíria Macedo*[1]*, Aldebaro Klautau*[1]

[1]Federal University of Pará (UFPA), Signal Processing Laboratory (LaPS)
www.laps.ufpa.br

Abstract:

This work presents a new version of LaPSTalk: a computer-assisted software for English language learning. All the speech interfaces (recognition and synthesis) were implemented using the Microsoft Speech Application Development (SAPI 5.1) toolkit. Also, adopting the Microsoft provided engines: Microsoft English Recognizer 5.1, Microsoft Speech Recognition Sample Engine for Brazilian Portuguese and L&H TTS, the user-system interface can be done through English and Brazilian Portuguese languages. Another feature is the custom agent, Merlin, a Microsoft Speech Agent component that provides feedback and assistance.

Key words: *CALL, speech, synthesis, recognition.*

1 Introduction

Computer-assisted instruction (CAI) is a learning process with the aid of computers. In the last 40 years, there was an exponential growth in the use of computers as providers of instructions. During this fast technological revolution, the CAI has become more refined and computers turned into a sophisticated education instrument, supporting new ways of teaching and learning. Over the years more and more systems are incorporating CAI multimodal interfaces such as sound, image, video and interactive programs that support a wide range of applications. More recently, the Computer-assisted Language Learning (CALL) has become popular.

In recent years, the performance of "personal computers" has evolved with the production of ever faster processors, a fact that encourages the economically viable implementation of speech processing algorithms in real time. The integration of high-quality audio, high-resolution images and videos with the output of text and graphics of conventional computers has generated recent multimedia systems that provide powerful "training tools" explored in CALL. The results of these efforts are prototypes with speech interfaces for training the human pronunciation in a wide variety of languages, texts reading and foreign languages teaching in pre-defined contexts [1].

An important area within the speech processing is the training of pronunciation. Intelligent tutors interact and guide students in the repetition of words and phrases, or reading sentences, in order to practice both the fluency (quality of the phonetic pronunciation), as the pitch (manipulation of the parameter [pitch]) in a certain language. These prosodics characteristics are particularly important in the language learning processing, since a incorrect prosody can block the communication between the speakers. Errors in intonation can give to the listener a misconception impression of the attitude taken by the person who is speaking, as well as a sentence without fluency makes it virtually impossible for the listener understands the dialogue context.

There is not a consensus among the linguists about the presence of teachers or tutors in the process. In [2], it is stated that pronunciation training softwares must be employed to enhance the interaction between teacher and student, where the tutor must concentrate their efforts on speech characteristics that may affect their comprehensibility, while [3], believes in replacement of "human tutors", and suggests that the students can follow their own learning rhythm. However, linguists may attend the technicians, suggesting which speech aspects of the practitioner need to be focused and stipulating limits for the shunting lines identified in the pronunciation.

A widespread style of CALL, especially in schools of languages, makes use of an high-level interface and grammar set within a context, regular situations in real life, to practice a foreign language [4]. Basically, the student must choose a response within a limited number of alternatives shown on the screen, or may be challenged to answer a question without any help from the software. These methods are called closed response and open response, respectively.

This work describes a new version of a CALL software that is under development at the Federal University of Pará, Brazil. Previous versions of this software exclusively used the English language [5]. The current version also provides support to Brazilian Portuguese and allows to building richer interfaces.

2 The Developed CALL Software

2.1 Resources

There are a large number of engines available for English, which is one of the reasons for the development of this system, however the existence of engines for others languages, including Portuguese, makes possible the implementation of a software CALL for foreigners. Therefore, changes on the source code and in the structure of the original software components are essential.

Using engines provided by Microsoft, for both Automatic Speech Recognition (ASR) and Text-To-Speech (TTS), it is possible to interact with the user through the English language. However, the familiarity with the interface by the user, was very dependent on ASR and TTS in the native language, in our case the Brazilian Portuguese language, for instructions and feedback. This situation is shared with other projects under development in Brazil (e.g., [6]). Thus, in the beginning of the project development, a multilingual structure was developed with the incorporation of the Center for Spoken Language (CSLU) toolkit.

The CSLU toolkit [7] is a platform for dialog application development. The software provides four interface levels for applications development. For example, developers can use code written on the C# language to access. However, there is no support for a generic Application Programming Interface (API) such as Microsoft Speech API (SAPI) [8] or Java Speech API (JSAPI) [9]. The approach adopted to synthesize Brazilian Portuguese texts using the CSLU toolkit did not required changes in the CSLU code. Batch files that called the CSLU TTS were used but the solution is inefficient (obviously, reading and writing from/to disk was rather slow). This strategy could be used in other situations but was abandoned in favor of a completely SAPI-based solution employing Microsoft Agents.

The personal agent, “Merlin”, created from the MSagent [10] component, provides feedback and, when necessary, assistance. Since the character Merlin is compiled to use the English TTS engine Lernout & Hauspie (L&H) TruVoice as default, it was necessary the installation of L&H TTS3000 for Portuguese language, also licensed by Microsoft. Besides, the L&H engine has proved essential for the installation of Portuguese language components. These components are libraries (DLL files) that add support for synthesis on a given language. With them it is possible to change the characters TTSModeID property.

Also, adopting the ASR engines provided by Microsoft: Microsoft English Recognizer 5.1 and Microsoft Speech Recognition Sample Engine for Brazilian Portuguese (beta version), the user can interact orally with the system via both English and Portuguese languages. The recognition is possible through activation of a grammar with restricted rules (command-and-control). The available version of Microsoft ASR engine to Brazilian Portuguese does not support dictation yet. In other words, the recognition must be guided by a grammar.

2.2 Functionalities

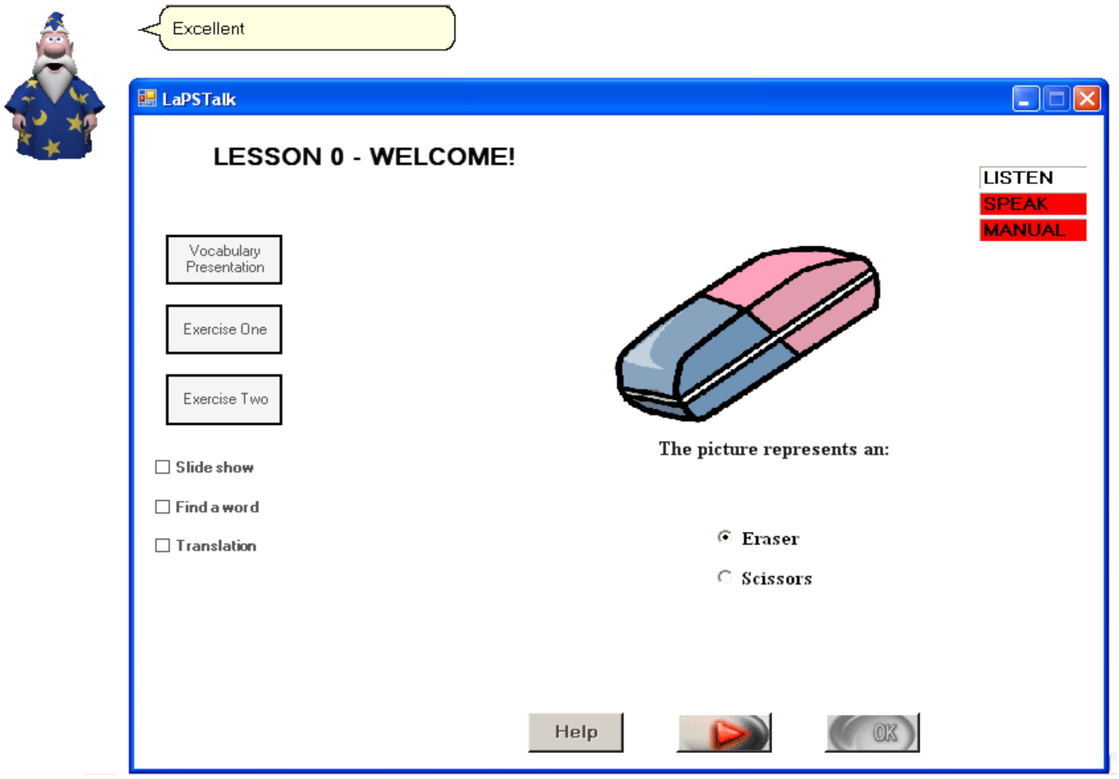
The developed software LaPSTalk, illustrated in Figure I, consists of an executable file containing ten (10) experimental modules. Each module corresponds to a lesson, which contains three (3) sections: vocabulary, exercise one and exercise two. The user is invited to respond the questioning orally, or even manually, by written, spoken and visual incentives. Besides the proposed objective exercises, there is the possibility of a prior training, where the user is asked to enrich his/her vocabulary by listening individual words or phrases synthesis, always helped by a picture, which illustrates the meaning of the word, or the action associated with the sentence. The sections are described in more details in the sequel:

- Vocabulary Presentation: stage where the user listening the words or phrases via text-to-speech, is invited to provide the transition between the sentences by hand. With the intention to make the transition more comfortable by the user, it was developed the “slide show” function. Still within the vocabulary section, there is a routine called “find a word” that, intuitively, allows the user to search for a word in the whole lesson and listen to it on its correct pronunciation. The last feature in the vocabulary section, called “translation”, enables the user to search orally a word, belonging to the lesson’s vocabulary, in Brazilian Portuguese (ASR), and get its translation in English (TTS), along with a picture to illustrate it. The engine recognition confidence is set on 0.7 and can not be adjusted by the user.

- Exercise One: tasks are proposed to test the grammar knowledge developed by the user. The software only moves forward if the user correctly answers the proposed question. If the answer is incorrect, the user is alerted by the agent, with the possibility of changing his/her answer. In case of doubt, the user can use the help option, where a synthesized sentence will help to choose the correct answer. The user can answer the questions by checking manually the desired alternative or even speaking the required answer via ASR in English.

- Exercise Two: consists on a pronunciation testing based on the vocabulary. Words are simultaneously made available so that the user can test his/her pronunciation by using automatic speech recognition in English. When the word is properly recognized, a corresponding visual illustration is shown. The engine recognition confidence level can be adjusted by the user. The available options are: easy, normal and difficult.

Intonation and rhythm plays an important role in language learning but they are not referred in this work.



LESSON 0 - WELCOME!

☐ Slide show

☐ Find a word

☐ Translation

The picture represents an:

● Eraser

○ Scissors

Figure I: One of the screens of computer-assisted software for English language learning.

3 Conclusions

This work described an on-going project that consists in implementing a CALL system with support to both English and Brazilian Portuguese. The ASR grammar used in the implemented CALL software is relatively simple (in the sense of perplexity [11]). As the recognition is always focused on contexts (limited options for response), the grammar is simple and the system is robust even if used by nonnative speakers.

The current version of the software does not control or monitor recognition errors. Informal tests showed that the determining factors for accuracy are: the training stage suggested by the engine during the installation procedure, the use of "head-mounted" microphone with a constant distance between the mouth and microphone, and the amount of environmental noise during the test (execution) stage.

We expect the future use of additional resources, like an ASR engine for Brazilian Portuguese that supports dictation (as mentioned, the current engines are restricted to using grammars), the present CALL application could be easily extended to teach Brazilian Portuguese, for example, to English speakers.

References:

[1] Ehsani F.; Knodt E.: Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New CALL Paradigm. Language Learning and Technology Vol. 2, No. 1, July, pp. 45-60, 1998.

[2] Fraser, H.: Phonetics, phonology, and the teaching of pronunciation - a new cd-rom for all esl learners and its rationale. Eighth Australian International Conference on Speech Science and Technology, pp.180-185, [S.I.], 2000.

[3] Ananthakrishnan, K. S.: Computer Aided Pronunciation System (CAPS). University of South Australia, 2003.

[4] CALL Computer-assisted Language Learning. CCLS PUBLISHING HOUSE, C1626AK11 ISBN 85-340-0602-4, 1998.

[5] Neto, N.; Siravenha, A.; Macedo, V.; Klautau, A.: A Computer-Assisted Learning Software to Help Teaching English to Brazilians. In: International Conference on Computational Processing of Portuguese Language, 2008, Aveiro. Propor 2008 - Special Session (Presentation), 2008.

[6] Velho, L.; Rodrigues, P. L.; Feijó, B.: Expressive Talking Heads: uma ferramenta de animação com fala e expressão facial sincronizadas para o desenvolvimento de aplicações interativas. Proceedings of Webmídia. SBC, 2004.

[7] http://cslu.cse.ogi.edu/toolkit/. Visited in May, 2008.

[8] http://www.microsoft.com/speech/. Visited in March, 2009.

[9] http://java.sun.com/products/java-media/speech/. Visited in May, 2008.

[10]http://www.microsoft.com/products/MsAgent/. Visited in March, 2009.

[11]Huang C.; Chen T.; Chang E.: Accent Issues in Large Vocabulary Continuous Speech Recognition. In International Journal of Speech Technology, 141-153, 2004.

Author(s):

Ana Carolina, Siravenha, MSc degree student
Federal University of Pará, Department of Electrical Engineering
Av. Perimetral 1, Guamá - Belém – Pará – Brazil – 66075-110
siravenha@ufpa.br

Nelson, Neto, PhD student
Federal University of Pará, Department of Electrical Engineering
Av. Perimetral 1, Guamá - Belém – Pará – Brazil – 66075-110
nelsonneto@ufpa.br

Valquíria, Macedo, Associate Professor
Federal University of Pará, Department of Electrical Engineering
Av. Perimetral 1, Guamá - Belém – Pará – Brazil – 66075-110
vgmacedo@ufpa.br

Aldebaro, Klautau, Adjunct Professor
Federal University of Pará, Department of Computer Science Engineering
Av. Perimetral 1, Guamá - Belém – Pará – Brazil – 66075-110
aldebaro@ufpa.br