

Learning Portuguese with Speech Technologies

Isabel Trancoso^{1,2}, Thomas Pellegrini^{1,3}, André Silva^{1,2}, Rui Correia^{1,2}, Nuno Mamede^{1,2}, and Jorge Baptista^{1,4}

¹ INESC-ID, Lisbon Portugal

{Isabel.Trancoso,Thomas.Pellegrini,Rui.Correia,Nuno.Mamede}@inesc-id.pt
<http://call.l2f.inesc-id.pt>

² Instituto Superior Técnico, Universidade Técnica de Lisboa

³ Instituto Superior de Línguas e Administração Campus de Lisboa

⁴ Universidade do Algarve, FCHS

Abstract. This set of demos intends to illustrate different applications of speech technologies for Computer-Assisted Language Learning. Although the most typical application in this context seems to be pronunciation training, the emphasis here is on vocabulary learning, and perception. The latter is specially important for European Portuguese as a second language, which is the target of our research. The first demos are aimed at beginners level, and consist of serious games based on 3D, and speech recognition and synthesis technologies, for learning vocabulary and the use of prepositions. The second set of demos are aimed at a higher level, and use recent multimedia documents such as TV broadcast news of the preceding week as training materials. Many speech and language processing technologies are involved in this demo, such as audio segmentation, speech recognition, capitalization, punctuation, topic segmentation and indexation. The generation of the exercises using this data is fully automatic. This may be very valuable for teachers, saving them time in search for motivating materials of appropriate quality, level and topic.

Keywords: CALL, Portuguese, Serious Games

1 Introduction

In the overview of spoken language technology for education [Eskenazi, 2009], the author shows how the use of automatic speech processing for education, and especially for language education, has blossomed in many directions. This paper briefly summarizes our own contributions to this area.

The target language of our CALL (Computer-Assisted Language Learning) systems is European Portuguese (EP). Students of EP as a second language often state that their listening skills cannot cope with spontaneous speech. In fact, one well-known characteristic of EP that distinguishes it from Brazilian Portuguese in particular, is the strong use of vowel reduction and simplification of consonantal clusters, both within words and across word boundaries [Cruz-Ferreira, 2009].

Learning Portuguese with Speech Technologies

This has been the motivation for adding perception as a major focus in our exercises and games, au par with vocabulary learning and other skills.

The next section describes our 3D serious games, aimed at beginners level, which use speech recognition and synthesis technologies, for learning vocabulary and the use of prepositions. Section 3 describes our CALL applications aimed at a higher level, which use daily adapted multimedia documents such as TV broadcast news as training materials.

2 3D games

The first set of demos aims at teaching vocabulary and also the verbs and prepositions used to describe the spatial relation of objects. Exercises are solved in a game environment making use of a 3D scenario in order to further capture the student's interest, based on the Unity 3D game engine⁵.

In the first game, the player controls an avatar through first-person perspective mainly. The scenario consists of an office composed of 5 different rooms, and in each room there are several exercises to be completed. The exercises consist of asking the student to move an object in the scenario to new positions with the use of the mouse, according to a given instruction. For example: *Coloque o objecto A em cima do objecto B* (Put the object A **on top of** the object B). Answers given by the students are automatically evaluated by our game [Silva et al., 2011]. The orders are synthesized using DIXI, our in-house TTS system (Text-to-Speech) [Paulo et al., 2008], as well as the object names, in the tutorial menu. When the player does not position an object in the right place, the game describes the action that was made and the one that should have been made.

The initial feedback we received from students using this game encouraged us to develop a second version of the game in which the student is first shown a fully furnished room whose objects and positions he must memorize. The student must then go to a different room which is totally empty, and refurnish it using exactly the same pieces in the same position using voice input. The student's orders (e.g. put the sofa in the middle of the room) are automatically recognized using AUDIMUS, our in-house ASR system. If the position is correct, the animated object will appear, rotate, and drop in the desired position. A push-to-talk button is used to mark the beginning and end of each utterance.

The third version of the game is only directed at vocabulary coaching. Instead of the office environment, 4 different rooms of a house are used and the recognized vocal input is used to furnish a totally empty room, without specifying positions. Figure 1 illustrates the second version of the game.

3 Daily REAP.PT

The second set of demos involve REAP.PT - the Portuguese version of a vocabulary learning tutoring system, originally developed for American English at

⁵ <http://unity3d.com/> (last visited in November 2011)

Learning Portuguese with Speech Technologies



Fig. 1. 3D game.

Carnegie Mellon University. The system initially focused on vocabulary learning by presenting to students reading material with target vocabulary words in context [Heilman et al., 2006].

The daily REAP.PT version differs from the original system in many aspects, besides the target language: it is based on news that were published during the last 7 days, rather than on documents which were retrieved from the web at a certain time, thus ensuring that they are very recent; it includes both text news and TV news which have been automatically processed; and it allows the students to select a given segment in the text, and listen to the corresponding audio, in a karaoke style. The audio signal is either produced by the TTS system or retrieved from the original video.

In order to be able to use BN videos as learning tools, they need to be automatically segmented, transcribed and indexed. The processing pipeline consists of removing the jingles that usually start and end the news shows, segmenting the audio stream into single-speaker homogeneous speech segments, transcribing the segments automatically with our in-house automatic speech recognition (ASR) system [Neto et al., 2008]. Further modules are then applied to include punctuation, capitalization, and to segment the video into different stories and assign them multiple topic labels. The output of the BN pipeline is comprised of stories with about 300 words each on average. A further filter is applied to automatically estimate the readability level of the stories.

Figure 2 shows a screen-shot of such a page. The left-hand side of the page allows the students to navigate between the different stories of a selected topic, and to see the cooresponding video with captions. On the right-hand side, these same captions are shown with enhanced features. Words with a low confidence estimate are highlighted in red in order to warn the reader that they may be misrecognitions. Target words are hightlited in blue, as they will be the focus of the vocabulary exercises the student should complete next.

Learning Portuguese with Speech Technologies



Fig. 2. Daily REAP.PT.

3.1 Vocabulary perception

Attempting to combine the rich diversity of our BN repository with the motivating aspects of games, we developed “vocabulary perception” exercises, in which the learner transcribes a short video clip by choosing words from lists containing the correct words and some distractors. Our main objective was to give realistic speech for the learners to get used to the pronunciation of native speakers.



Fig. 3. Tick interface of the vocabulary perception game.

All the exercises are generated in a fully-automatic way. A filtering is needed to discard sentences with probably misrecognized words. A sequence of five filters