ASR-based exercises for listening comprehension practice in European Portuguese

Thomas Pellegrini a,∗, Rui Correia a,b,c, Isabel Trancoso a,b, Jorge Baptista a,d, Nuno Mamede a,b, Maxine Eskenazi c

a Instituto de Engenharia de Sistemas e Computadores – Investigação e Desenvolvimento, Lisbon, Portugal
b Instituto Superior Técnico, Lisbon, Portugal
c Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
d Universidade do Algarve, Faro, Portugal

Received 16 May 2012; received in revised form 22 December 2012; accepted 13 February 2013
Available online 24 February 2013

Abstract

Spoken European Portuguese (EP) is known to be difficult to understand for L2 learners, due to phenomena such as strong vowel reduction. In this paper, we present a method to automatically generate exercises aimed at improving listening comprehension skills in EP. Learners identify the words pronounced in real speech utterances. The exercises introduce two innovative aspects: using broadcast news videos for curriculum and automatically generating exercises with material updated on a daily basis. The videos are automatically transcribed by a speech recognition engine. A filtering chain, used to select appropriate sentences, was validated by a first survey comprised of both manually and automatically selected sentences. Both sets were assigned good to very good subjective quality scores. A second survey concerned the features of the exercise interface. Subjects with varying self-reported exposure to Portuguese as a second language tested several interfaces and functionalities and highlighted their preferred features. The results confirmed that the largest difficulty was the fast speech rate. All participants valued slowed-down audio and video documents, though this feature was more often used by the lowest proficiency subjects. The exercises were integrated into a Web platform where they are automatically updated daily. Though further evaluation is needed to find whether the platform affords skill acquisition, it is expected to be particularly valuable for distance learners who need opportunities to access authentic audio documents in EP.

© 2013 Elsevier Ltd. All rights reserved.

Keywords: Computer-Assisted Language Learning; Speech processing; Automatic generation of learning material; Listening comprehension

1. Introduction

European Portuguese (EP) L2 learners 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). Vowel reduction can result in vowel devoicing, voicing assimilation, deletion and coalescence.

This paper has been recommended for acceptance by M. Russell.

∗ Corresponding author at: INESC-ID, Rua Alves Redol 9, 1000-029 Lisbon, Portugal. Tel.: +351 213 100 232; fax: +351 213 145 843.
E-mail addresses: thomas.pellegrini@inesc-id.pt, thomas.pellegrini@gmail.com (T. Pellegrini), rui.correia@inesc-id.pt (R. Correia).
(when two adjacent segments are replaced by a single one which shares features of the two original ones). Common contractions can be partial or full syllable truncation and vowel coalescence. For example, in “grade de ferro”, a standard pronunciation of the two D’s would be to reduce the geminate to one phone. Hence, in our effort to develop Computer-Assisted Language Learning (CALL) tools to help Portuguese as a Second Language (PSL) learners, practicing listening comprehension is an important function. Our main objective was to provide learners with realistic speech representing the sounds and the pronunciation of EP.

The exercises consist of identifying the words pronounced in real speech utterances. Two innovative aspects were introduced: the multimedia material itself, comprised of broadcast news (BN) videos; and the automatic selection of the sentences, made possible by using our automatic speech recognition (ASR) engine to transcribe the videos and a filtering chain to select appropriate sentences. The growing interest in using games in the CALL community to motivate learners (mini-games and serious games), led us to develop exercises with game-like features such as a word puzzle interface.

This article builds on the work described in Pellegrini et al. (2011) and Correia et al. (2011). The previous studies were extended by providing performance measures for the different components of the filter chain, and by significantly enlarging the number and origin of subjects participating in the user survey. Moreover, the exercises were recently integrated into a new Web platform called Daily-REAP\(^1\) that provides written news and BN videos from a national Portuguese public TV channel. The documents are automatically updated daily and the user can access documents from the past week. Videos and their automatic transcription are presented to the user along with the corresponding listening comprehension exercises.

The paper starts with a brief overview of related work and context of this study in Sections 2 and 3, respectively. The automatic method for exercise generation is described in Section 4 along with a subjective quality evaluation performed by native speakers. The user survey that aimed at understanding the users’ preferred features is reported in the last section.

2. Related work

Literature shows that learning models have recently evolved from being teacher-centered to learner-centered. Self-directed and autonomous learning that allow the learners to work at their own pace and level are facilitated by CALL platforms available on the Web. The challenge for learning tool providers is to design engaging tools. Hoven (1999) pointed out that one essential feature of instructional design is the level of choice of tasks. In order to reach an adequate design, we carried out a user-centered study that consisted of a sequence of several exercises with different tasks (word spotting and word ordering) and different media (audio and video).

The premise of the present work is that using recent multimedia curriculum in exercises may increase the students’ motivation to learn. Videos are expected to be a motivating feature. Listening comprehension models already integrated viewing comprehension in the context of CALL multimedia systems (Hoven, 1999). Secules et al. (1992) showed how listening comprehension skills improve when French students used video-based contents. Brett (1995) showed that authentic video materials with subtitles can increase the students’ motivation.

Another interesting feature we wanted to test was the ability to slow down the speech rate. Language acquisition is stimulated by materials that can be understood through the process of meaning negotiation, either via interaction or through the help of characteristics of the input (Blau, 1990). Many studies in second language acquisition have explored modifications of the input, such as syntax simplification, restatement, rephrasing, velocity of speech manipulation, introduction of pauses, among many others (Long, 2002). In particular, slowing down the flow of speech has had a variety of results. Some studies found strong support for reduced speed (Flaherty, 1979; Kelch, 1985), while others concluded that reducing the speech rate did not account for a significant increase in learning (Smith, 2002; Blau, 1990). Blau (1990) concluded that slowing down the speech rate enhanced comprehension at the lowest levels of L2 proficiency (first-year English L2 students). As the proficiency level rose, this feature helped much less. Given these contrasting results, and the fact that speed seems to be the biggest difficulty reported by L2 learners of European Portuguese, this study explored slowed-down audio and video segments.

\(^1\) http://call.l2f.inesc-id.pt/daily-reap.pt.
Games have recently gained a strong interest in education in general and in language learning specifically. For instance, a special issue was recently devoted to serious games in *Computer & Education* (volume 56, issue 1, January 2011). These games are referred to as serious since they have an educational goal supported by entertainment features (Sørensen and Meyer, 2007). Examples of recent language learning games are Mingoville,\(^2\) for children, and Rainbow Rummy (Yoshimoto et al., 2009) and Polyglot Cubed (Grace and Castaneda, 2011) for adult learners. In Eskenazi (2009), a review of the research in spoken language technology for language learning, the author underlines the engaging role of games in CALL systems, which competitive nature has the power to hold the player’s interest. Games may target specific skills such as vocabulary (Yip and Kwan, 2006) or involve a combination of skills such as perception, grammar, and vocabulary in conversational role-playing tasks (Wik et al., 2007). Another interesting aspect of games is the possibility of gathering data without having to pay the contributors, as in the case of crowdsourcing (Skory and Eskenazi, 2011). The data can be used to enhance the content of the games in an iterative manner.

Although the interface presented in this paper is not a serious game, some gaming aspects were introduced in an effort to make the exercises more engaging.

3. Context

Our research in CALL began a few years ago in the context of a joint research program between Portuguese universities and Carnegie Mellon University. We first adapted the tutoring system developed at the Language Technologies Institute to EP. This CALL system, named REAP\(^3\) (REAder-specific lexical Practice for improved reading comprehension), teaches English as a second language, focusing on teaching vocabulary in context. It provides students with Web documents that follow specific constraints such as difficulty level, topic, and length (Heilman et al., 2006).

Our Portuguese version, REAP.PT, supports language teaching to either native or non-native speakers (Marujo et al., 2009). Target words are selected from the Portuguese Academic Word List (P-AWL) (Baptista et al., 2010), according to the student level. The current version of P-AWL contains the inflections of about 2k different lemmas, totaling 33.3k words. These words are highlighted in the selected documents, and provide the focus for vocabulary learning exercises.

REAP.PT has been extended beyond the original vocabulary exercises, namely by integrating different grammar teaching modules. The most recent innovations have involved serious games. One of them takes place in a 3D environment, aiming at teaching locative prepositions used to describe the spatial position between objects (Silva et al., 2011). Another game consists of a competitive translation game, where the learner plays against an automated agent that uses a statistical machine translation module to guess the hidden words in a sentence in the target language (Ling et al., 2011).

Current work, however, does not focus on the gaming aspects of our platform, but rather on the use of multimedia documents as learning materials. Our first effort to propose multimedia documents consisted of including a set of audio books, also known as digital talking books. However, the number of books was limited due to author copyright restrictions. An alternative was to include BN videos in the curriculum of the reading activity module. We have stored a large repository of BN videos that have been automatically transcribed on a daily basis since 2009. This corpus was a natural choice for us, making up-to-date learning materials available to the learners, with a wide choice of short stories on varied topics with the added value of videos. Instead of using past videos from this huge repository as learning materials, Daily-REAP retains only the BN shows of the past week.

The BN videos must be automatically segmented, transcribed and indexed in order to prepare and select relevant excerpts. 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, and transcribing the segments with the AUDIMUS recognition system (Neto et al., 2008). Further modules are applied to include punctuation, capitalization, and multiple topic labels, among a set of 10 possible topics: Economy, Education, Environment, Health, Justice, Meteorology, Politics, Security, Society, and Sports (Amaral et al., 2007). One topic-labeled story contains 300 words on average.

After topic segmentation and indexing, each story is fed into a classifier that automatically estimates its readability level, on a scale of 5–12, corresponding to the national Portuguese high school levels. The language level of the stories spans from 7th to 11th grade, with an average corresponding to 8th grade (Lopes et al., 2010).

\(^2\) http://www.mingoville.com (last visited January 2012).
\(^3\) http://reap.cs.cmu.edu (last visited February 2012).
After the processing pipeline and the level classification, the stories are displayed on a Web page, showing the video excerpts with their automatic transcriptions. On a side frame, the automatic transcriptions are displayed as text, showing target words (from P-AWL) in blue, exactly as in the original text version of REAP, with words transcribed with low confidence in red. The student can select any segment of this text and listen to the selection as the words are highlighted in a *karaoke* style. Fig. 1 shows the screen divided into several frames. The bottom one is for topic selection.

*Daily-REAP* also allows the students to read news in text format, which were retrieved from a single online newspaper during the previous week. Whereas this text-based version relies on text-to-speech synthesis to provide the audio of any word sequence selected by the student, the BN version allows students to listen to the real audio media. In both versions, exercises are generated on a daily basis.

4. **Automatic exercise generation**

The main idea behind the listening comprehension exercises is that in each one the student listens to a single sentence from a video or audio clip. The sentence then needs to be reconstructed by choosing words from a list that contains the correct words and some distractors. The automatic generation of such exercises entails two non-trivial steps: the selection of the sentences and the generation of distractors.

4.1. **Sentence selection**

Sentence selection is carried out by using a set of filters applied to the automatically generated transcriptions. The transcribed sentences may contain misrecognized words (word substitutions or insertions) or may be incomplete (word deletions). The audio quality also needs to be checked. For example, an outdoor report may be too noisy and is more difficult to understand when compared to indoor anchor speech. The first filter addresses these two problems by selecting sentences with high recognition confidence scores. Three additional filters control the sentence length, its syntactic completeness, and the preference for neutral declaratives ending with a full-stop. In more detail, the filter sequence is as follows:

- **ASR confidence measure — f1**

  Confidence measures (CM) at word-level estimate how reliable each hypothesized word is. Computed within a $[0, 1]$ range, CM are estimated by gathering scores and other informative variables, during ASR decoding (*Pellegrini and Trancoso, 2010*). This filter relied on CM at sentence-level, i.e. computed the average CM of the words that compose the sentence, and filtered out the sentences for which this average falls below a threshold. Several values for this threshold have been tested. The survey mentioned in Section 5.1, for instance, was performed using a threshold
of 0.85. The current value is higher (0.9), given the need for greater pruning shown in initial experiments. This value is close to the one we applied for selecting material for unsupervised retraining of acoustic models, using 1000 h of audio data (Meinedo et al., 2010).

This filter proved quite effective in filtering out sentences with a low signal-to-noise ratio (SNR). In fact, in our earlier experiments, the filter chain included a component for rejecting utterances with an SNR lower than 10 dB. The $f_1$ filter now discards most of these cases.

- **Sentence length – $f_2$**
  
  BN shows are mainly comprised of prepared speech, characterized by long sentences (17 words per sentence on average in our corpus). In Ur (1984), the author stressed the necessity of providing small units of speech for listening comprehension assessment, since in real life, discourse is typically divided into small chunks of speech. For this reason, and although different limits were initially tested (our survey, for instance, used an upper limit of 15), only sentences with a minimum of 4 words and a maximum of 10 were kept.

- **Syntactic completeness – $f_3$**
  
  Only utterances with at least one verb, and one common noun or adverb were selected. This filter guarantees the presence of content words.

- **Neutral declaratives – $f_4$**
  
  Sentence boundaries are determined by a statistical module that recovers full-stops and commas (Batista et al., 2008). In order to reject speech continuations and give preference to neutral declaratives, an additional constraint on pitch was used. In European Portuguese, neutral declaratives usually show decreasing or stable pitch pattern endings (Frota, 2000). Sentences were rejected when the pitch slope of the last voiced segment was positive, indicating a potential continuation or an interrogative. The fundamental frequency was extracted using the Snack software.\(^4\)

Fig. 2 shows the execution of the filter chain and the number of filtered sentences at each step for one month of videos (February 2011), corresponding to a total of more than 10k candidate sentences. The application of the filter chain to the set of 10k sentences resulted in drastic reductions. In total, only 292 (3%) passed all the filters. Filter $f_1$ filtered out sentences with recognition errors. The average word error rate is approximately 18% on BN data (Meinedo et al., 2010). Thus a 10-word sentence is expected to contain 2 misrecognitions on average. This fact explains the high percentage of sentences discarded by this filter (73.5%). The length filter, $f_2$, further reduced the number of sentences by 77.2%. This can be explained, as already mentioned, by the typical long length of BN sentences. Further reductions were obtained by applying $f_3$ and $f_4$, 33.0% and 29.1%, respectively.

The automatic transcriptions of the 2699 sentences kept after $f_1$ were checked manually. The associated WER was 3.9% instead of the average 18%, which confirms the good performance of this filter. The WER of the set of 290 sentences produced by the 4 filtering stages is slightly smaller: 3.1%. Thus, the filtered sentences still have some errors due to ASR. Nevertheless, the great majority of these errors concern small function words that do not affect global comprehension of the sentences. As we will show in the user evaluation section, the users have the possibility to report errors in the exercise solutions. In such a case, reported sentences could be removed from the exercise database.

The performance of filter $f_4$ is particularly interesting. The 412 input and 292 output sentences (with the corresponding audio) were checked. Interrogatives, incomplete sentences or continuations were considered as the negative class, and sentences that end or could end with a full-stop were the positive class. The resulting precision and recall were 0.879 and 0.833, respectively, achieving a 0.856 $f$-measure. The majority of the rejected segments are continuations. As it is natural to have very few direct questions in BN news, only two interrogatives were found in the sentences and detected by the filter.

---

\(^4\) [http://www.speech.kth.se/snack](http://www.speech.kth.se/snack) (last visited January 2012).
Finally, it is also important to mention that tuning the few parameters of the filters allows us to control the final number of sentences that are generated. In Daily-REAP, an average of 22 sentences per day are selected and transformed in exercises.

4.2. Distractor generation

Two techniques for generating distractors were evaluated: distractors selected according to phonetic distances (phonetic-based), and distractors gathered during the ASR decoding (ASR-based).

4.2.1. Phonetic-based distractors

This method generates a fixed number of distractors for each recognized word. As the name states, these distractors are selected by exploring the phonetic representations of each recognized word.

The candidate distractors come from P-AWL (see Section 3), which is composed of 2k different lemmas (33.3k words) that students are exposed to while using REAP:PT. For each recognized word, the two phonetically closest candidates were selected as distractors. The Levenshtein distance between the phonetic representations of each pair of words (the recognized word and the candidate distractor) was used as the phonetic distance. These representations were obtained using the leia grapheme-to-phone tool (Oliveira et al., 1991). The tool provides a “canonical” or non-reduced pronunciation of a word. For simplicity sake, the phonetic representations used for the distance algorithm are based on this canonical pronunciation. Although it would be interesting to derive distractors using the same algorithm with alternative pronunciations, this has not yet been tested. In fact, we have relied on the ASR-based distractors to take these alternative pronunciations into account (see Section 4.2.2).

Using a distance algorithm directly on top of the phonetic representations of the two words simply leads to an orthography comparison. Thus, a different weight was assigned to each substitution between a pair of phones. These weights were based on Paulo and Oliveira (2002), taking into consideration features such as voiced/unvoiced, manner and place of articulation. Table 1 shows a fragment of the weights that were used for the substitution operation. The weights for deletion and insertion are 10 and 11, respectively.

As an example of phonetic-based distractors, for the word “solução” ([sulusu]) – meaning solution – the generated distractors were “solução” ([sulusu]) – hiccup – and “seleção” ([saleshui]) – selection or team in the context of sports – as the first and second closest options, respectively.

4.2.2. ASR-based distractors

During the decoding process, several word hypotheses compete. Intrinsically, the words in competition are phonetically similar, and as such, they may be suitable distractors for the exercises.

The best hypothesis, which corresponds to the lowest score path inside a Confusion Network (CN), is the ASR transcription, and therefore is the solution of the exercises. All the other competing words generated during decoding were used as distractors. CN are compact representations of word lattices that specify the sequence of word-level confusions (Mangu and Brill, 2000). These structures are stored during the decoding process, and allow us to easily associate each best-hypothesized word with its competing words. For the best hypothesis word “contrário” ([kot/a/iu]) for instance, distractors were “contrato” ([kot/atu]), “contrária” (kot/a/iu]), and “controlar” ([kot/ula]). The interesting aspect of this method lies in the fact that it covers co-articulation effects, i.e. the distractors may correspond to multi-word common confusions. Examples are “com paixão” and “compaixão”, “se passou” and “expressou”, “para ficar” and “purificar”. For a given utterance, the number of competing words depends on two parameters used in the decoding.

<table>
<thead>
<tr>
<th>Base phone</th>
<th>Target phone</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>[i]</td>
<td>[i]</td>
<td>0</td>
</tr>
<tr>
<td>[i]</td>
<td>[e]</td>
<td>4</td>
</tr>
<tr>
<td>[i]</td>
<td>[u]</td>
<td>6</td>
</tr>
<tr>
<td>[i]</td>
<td>[e]</td>
<td>8</td>
</tr>
<tr>
<td>[i]</td>
<td>[e]</td>
<td>10</td>
</tr>
</tbody>
</table>
Fig. 3. One of the 80 sentences of the evaluation survey (“the price of the sheet increased by four cents”).

process: the beam search and the maximal number of arcs. For some utterances, however, a clear path dominates, and very few distractors can be derived by this method. This is often the case of speech spoken in a quiet studio with no disfluencies.

4.3. Evaluation

To evaluate both the quality of the filtered sentences and the quality of the distractors, we used a set of 80 sentences: 40 sentences from manual transcriptions and 40 sentences randomly chosen among the 290 automatic filtered sentences. The 40 manually transcribed sentences were manually selected, according to the same criteria used to define the filters designed for the ASR material. The filter chain was applied afterwards to this manual set (assuming CM = 1 for all words in the f1 filter). None of the sentences were discarded by the filters, showing a good adequacy between the subjective criteria and the filter chain.

Two Portuguese native speakers, with backgrounds of spoken language processing (annotator 1), and linguistics (annotator 2), answered a survey about the 80 sentences. The goals of the survey were the following:

• to test whether automatically transcribed sentences would be discriminated from manual transcriptions, therefore assessing the use of ASR material;
• in case of rejection, to identify the problems that led to the discarding of the sentence;
• and finally, to evaluate the two automatic methods of distractor generation.

4.3.1. Filter evaluation

Fig. 3 shows an example of the utterances presented to the annotators. The sentences were presented in random order, so that no information about whether the utterance was manually or automatically transcribed could be inferred. A play button allowed listening to the corresponding speech segment. The raters were asked to rate the general quality of each sentence (transcription-audio pair) on a five-point Likert scale (1 = very bad, 2 = bad, 3 = OK, 4 = good, 5 = very good). This subjective rating was expected to give an idea if an utterance and the corresponding transcription were suitable to be presented to a student in a listening comprehension task. The annotators were also asked to justify the rejection of a sentence, among a list of five problems: ASR errors, noise, syntactic problems (grammatical error), semantic problems (misrecognition that hinders the comprehension of the sentence), and other. Finally, for each sentence, annotators were asked to choose distractors among a list of randomly ordered ASR-based and phonetic-based distractors, in order to find out if they preferred one method over the other. There was no upper limit to the number of distractors that they could select.

Fig. 4 shows the number of sentences (on the Y-axis) rated on the Likert scale by the two annotators. Both rated 60.0% of the sentences with the same score, with a majority of very good ranked sentences: 82.5% and 65.0% of the 80 sentences for annotators 1 and 2, respectively. The inter-rater agreement (Cohen’s kappa coefficient) was 0.39, which can be considered as moderate.

Eight sentences contained ASR errors. Different quality ranks were assigned by the two annotators according to the type of misrecognition. Sentences with an error on a small function word were rated as very good and good by annotators 1 and 2, respectively. Errors on a content word were considered more severe, so that sentences with such
errors were rated OK by annotator 1, and bad or OK by the other rater. Only one sentence with several wrong content words was rated very bad by annotator 1.

Another annotation difference concerned the perception of noise. Annotator 1 reported only one noisy sentence, whereas annotator 2 reported 14 of them. For annotator 2, the presence of noise in both the automatic and manual sets was judged as a limitation, whereas for annotator 1, noisy examples could be interesting as examples of authentic speech.

The average scores for the automatic set were slightly lower than for the manual one: 4.32 and 4.97 for annotator 1, 4.41 and 4.70 for annotator 2, respectively. This result is mainly due to the presence of ASR errors in eight automatically transcribed sentences. The CM filter, $f_1$, was not able to filter them out, indicating that a larger CM threshold should be used. Nevertheless, the automatic score is between good and very good, thus validating the use of ASR material and the filtering chain to select the sentences.

4.3.2. Distractor evaluation

In total, more than 1k distractors were generated for the 80 sentences. ASR-based and phonetic-based distractors totaled 377 and 697, respectively. Only 3.5% of the distractors generated by the two methods were identical.

Annotator 1 selected 45.2% of the distractors in total, both ASR- and phonetic-based, fewer than annotator 2, who selected 60.0% of them. Both agreed on 57.0% of the distractors.

Annotator 1 favored ASR-based distractors, by choosing 46.1% of them versus 44.9% of phonetic-based distractors. Annotator 2 chose a slightly larger percentage of phonetic-based ones, 60.7% versus 59.0% of ASR-based distractors. These percentages do not sum up to 100.0%, since each corresponds to the ratio between the number of selected distractors of one type and the total number of proposed distractors of the same type.

Both distractor types have been chosen in similar proportions by both annotators, showing a complementarity in the two generation methods. ASR distractors have the advantage of covering interword co-articulation effects. On the other hand, when the ASR decoder is fairly confident about a hypothesis, fewer or even no ASR-based distractors are available. In this case, the sentence may be filtered out, or only phonetic-based distractors may be used.

5. User interface evaluation

Two main interfaces were implemented. Fig. 5 shows one of them, where the user is expected to watch a video clip. A slowed-down version of the normal video is also provided to help in case the speech rate is too fast. Boxes, each containing a single word, can be dragged and dropped one by one, into a sequence of empty target boxes (first row of boxes in the figure). In the figure, the user has already dropped words and obtained feedback by clicking on the check button on the left-hand side of the screen. Green check signs or red crosses indicate, for each target, if the
dropped word was correct or incorrect. Misplaced words, belonging to the correct sentence but placed in an incorrect position, are distinguished by a green background color. A score and the correct answer are given in the left-hand side of the screen. On the other hand, in the second interface, the goal changes slightly. Instead of ordering the complete set of words that form the sentence, the student is presented with a list of words that includes words that were in the audio/video clip and distractors. The student then has to spot which words in the list were present in segment that was played. It is important to notice that, in this second setup, contrary to the first, the exercise focuses only on the content words of the clip.

5.1. Experimental setup

To evaluate the system features several sets of exercises were implemented, each one with slight variations over the previous one. Volunteers completed a guided sequence of exercises and a final feedback questionnaire about their preferred functionalities and difficulties encountered.

The survey took the form of a Web session with 18 exercises, divided into 6 sets of 3 exercises. Each exercise consisted of listening to a sentence by either using audio only or video. As previously mentioned, the goal of the exercises varied along the sets: for the exercises in sets 1–3, all the words were required to be selected and ordered (first interface), whereas for the exercises in sets 4–6, only specific words, selected from P-AWL, were expected to be identified and checked from a list (second interface). The list of candidate words of all the sets included both correct words and distractors generated with the two methods explained in the previous section for sets 1–3. Sets 4–6 included the correct words and phonetic-based distractors only.

5.2. Sets 1–3

Sets 1–3 took the form of a word puzzle, in which the student first listened to a sentence, then attempted to reconstruct it by selecting words from a list of candidates including both the correct ones and some distractors.

Boxes, each labeled with a single word, can be dragged and dropped one by one, into a sequence of empty target boxes. Visual feedback is given by adding a green validation mark or a red cross for each word when the word is correct or wrong. Misplaced but correct words are marked with a green background color. A score (percentage) appears on
Table 2

<table>
<thead>
<tr>
<th></th>
<th>Set 1</th>
<th>Set 2</th>
<th>Set 3</th>
<th>Set 4</th>
<th>Set 5</th>
<th>Set 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>√</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Video (with audio)</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Slow down</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Search query</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Memorizing</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Recent material</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
</tr>
<tr>
<td>Karaoke</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>√</td>
</tr>
<tr>
<td>Manual transcriptions</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

The nine sentences of sets 1–3 were manually selected from the ALERT corpus (Amaral et al., 2007), a large set of BN shows that were manually transcribed. Hence, these sentences do not contain any ASR errors. The filter chain was applied on these sentences afterwards and none was rejected.

### 5.3. Sets 4–6

The goal of these sets was to select, among a list of words, the ones that the student heard on the clip, instead of having to order all the words to form a sentence. This group of sets also differed from the previously described sets in the way utterances were selected. Instead of being guided through a set of fixed exercises, the student had the opportunity to search for any sequence of words, and then to get BN video passages where the query appeared. The learner may be motivated by searching for a word he/she would like to hear in a real context. At the time of the survey, the words of the query could appear in the list of words to be selected, if they were in the P-AWL list. In the current version, they are excluded from the candidate words.

To enable the search feature, the so-called ASR transcript segments of BN shows dating from January 2009 through March 2011 were indexed. Transcript segments, as Richards (1983) describes, are speech segments delimited by significant pauses, and for that reason, may be viewed as an approximation to clauses. Two restrictions, corresponding to filters \( f_1 \) and \( f_2 \), were adopted to create the database of transcript segments: only segments with less than 15 words and more than 5 target words were indexed. Moreover, segments with an average CM lower than 0.85 were discarded. This filtering ended up with 90k segments, out of the 1.7 M original segments (5.3% of acceptance). Still the transcript segments may contain recognition errors, and the users were asked to report potential errors.

The free open source information retrieval library, Lucene,\(^5\) was used to index the ASR database and to retrieve the speech transcripts upon user queries. The segments could be searched using the standard search mechanism of Lucene, which takes into consideration criteria such as rarity of the terms and length of the document (Robertson and Zaragoza, 2009). If no results were returned for an exact word sequence query, the system relaxed the search using each word individually, and then retrieved the most relevant document.

Table 2 summarizes the main differences between the sets. The varying features, as shown in the first column of the table were: audio sentences (sets 1 and 2), video (sets 3–6), slowed down audio (set 2) or slowed down video in addition to the normal speed media (set 3), an engine to retrieve sentences with words queried by the user (sets 4–6), memorization of the utterance before completing the exercise (set 4), and exercises restricted to the most recent BN shows (set 6). The last set also introduced a karaoke feature, available on the correction screen, which allowed the user to watch the video with the corresponding transcription, while the words were highlighted as they were pronounced.

At the end of each exercise, the user answered two questions: “Which were the main difficulties of the exercise?” with five possible answers (none, sentence length, unknown words, speech too fast, and other with the possibility to comment for this last choice), and “Are there any errors in the answer, compared to what you have heard?”. With the latter question, we wanted to assess whether the participants were able to detect recognition errors.

---


Table 3
Number of participants according to their native language and self-reported number of years of exposure to EP.

<table>
<thead>
<tr>
<th>Native language</th>
<th># years of EP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5+</td>
</tr>
<tr>
<td>Italian</td>
<td>3 1 3 1</td>
</tr>
<tr>
<td>Spanish</td>
<td>4 4 1 2 3</td>
</tr>
<tr>
<td>French</td>
<td>2 1 1 1</td>
</tr>
<tr>
<td>Romanian</td>
<td>1</td>
</tr>
<tr>
<td>Not romance</td>
<td>3 2 2 4</td>
</tr>
<tr>
<td>Total</td>
<td>13 7 3 8 9</td>
</tr>
</tbody>
</table>

Fig. 6. Average scores obtained by the forty subjects for the six sets.

5.4. Results

Since the exercises were not intended to be used in a classroom context but rather to allow individual practice, non-native volunteers with varying EP levels were invited to take the survey. In total, 40 volunteers participated in the survey. The preliminary results with only 12 participants were reported in Correia et al. (2011).6 The average age of the 40 subjects (including the first set of participants) was 30 (σ = 7.2). They were asked to estimate the number of years they were exposed to European Portuguese. On average, this number was 3.2 years (σ = 2.3). Table 3 shows the number of participants according to their native language and to self-reported years of EP exposure. Subjects with 5 years or more of EP were grouped together with the label “5+”. The majority of the subjects (28) had a Romance language as L1. The other 12 subjects had diverse nationalities: 4 Polish, 2 Ukrainian, 2 German, 2 Serbian, 1 Lithuanian and 1 American. Scores, number of plays, queries, errors, response times, and other variables were stored in a database in order to conduct an in-depth analysis of their responses. In our previous work, the six sets of exercises were presented to the 12 subjects in the same order: from set 1 to set 6. For the new data, the order was randomized so as to avoid sequence effects. We compared the results of the 12 subjects of the first submission (no random order of the set sequence) with the results of the new 28 subjects (with a random order in the set sequence). The results were similar so we decided to merge both groups to get more reliable results.

Fig. 6 shows the average scores (and corresponding standard deviations) collected from the 40 subjects, for each of the sets. Scores from sets 1 to 3, and sets 4 to 6 cannot be directly compared since the goals, and therefore scoring

---

6 One of the original thirteen participants was excluded due to the fact that he could be considered bilingual.
methods, were different. In all sets, the maximum score was 100 but in sets 1–3, the player started with a score of 100 for each sentence, and points were subtracted when wrong words were selected, whereas in sets 4–6, the player started with a zero score and points were added when correct words were selected.

5.4.1. Score analysis

Average scores from sets 1 to 3 were 91.8, 85.3, and 93.7, respectively. Set 2 showed much lower scores than the other two sets and this may be due to the difficulty of this particular set. One sentence presented a verbal form with a clitic, another was informal speech with strong co-articulation.

Sets 5 and 6 presented very similar average scores, 82.5 and 83.9, respectively, which was expected since the only difference between both sets was the time period of the BN shows to be retrieved.

The most noticeable result was the much lower score of set 4: 66.5. This can be explained by the fact that a memorization effort was involved in this set of exercises, making this task more cognitively demanding: the users first watched the video clip, clicked next, and were redirected to another page where they could select the words that were heard.

It is interesting to compare the performance of the subjects according to their self-estimated level of EP. The subjects with 1 year of EP got the lowest average scores, 86.7 (σ = 18.0) and 81.3 (σ = 23.9) for sets 1–3 and sets 4–6, respectively, to be compared to 91.9 (σ = 16.0) and 85.4 (σ = 21.4) which are the average scores of the subjects with 5 or more EP years. The Spanish and the Italian subjects achieved the best results. In general, subjects with a Romance mother tongue (Spanish, Italian, French and Romanian) performed better than the others, with average scores of 91.1 (σ = 13.4) and 80.9 (σ = 21.4) versus 87.7 (σ = 18.6) and 74.2 (σ = 25.7), respectively, even though they reported a lower EP level than the other group: 2.5 (σ = 1.9) versus 4.1 (σ = 2.9).

The response time was stored for sets 1 and 2. It corresponds to the time interval between the first click to play the audio clip and the check button click. Fig. 7 shows the average response time in seconds as a function of the scores per sentence. The average response time was 72 s. It appears that the longer the response time, the lower the score. Spending more time on an exercise might imply that the subject is experiencing difficulties, therefore leading to lower scores. Two subjects, one with 1 year and the other with 4 years of EP contact, spent much more time than the others, about 2 min 40 s on average per sentence, and also got the two worst scores on average: 76.4% and 89.9% for sets 1–3, and 42.2% and 67.1% for sets 4–6.

5.4.2. Slow down feature

A correlation appeared between the average number of times the subjects listened to a sentence and their EP contact: participants with one EP year listened to the sentences about 8.5 times, whereas those with 5 EP years or more only did 2.9 times on average. The more experienced the subject, the less he/she listened to the sentences.
In sets 2 and 3, the participants could listen to a slowed-down version of the original audio or video sentences (audio for set 2, video for set 3). The slow-down speed ratio was empirically set to 0.7, since lower rates distorted the speech. Fig. 8 distinguishes the number of times the normal speed or the slowed-down speed utterances were played in these two sets. The least experienced participants used the slow-down feature much more often than the most experienced: 44% of the total number of plays were slow-down plays for 1 or 2 EP years subjects versus 27% for 4 and 5+EP years subjects. These findings suggest that, as Blau (1990) noted, the effect of slowing is positive at the lowest levels of proficiency.

5.4.3. Search feature

In sets 4–6, the 40 subjects made 360 queries (3 per set for each participant), of which only 194 were distinct. Most queries involved a single word (e.g. Porto, Portugal, president, football). Only 26 were comprised of several words, such as “Estados Unidos” or “Nova Iorque” (United States or New York, respectively). As expected, named entities were very popular in these queries.

5.4.4. Difficulties

The most frequently reported difficulty was the speech rate, selected in 19.8% of the exercises. This result supports the need to include a slow down feature. The sentence length factor was the second most mentioned difficulty (10.5%). The upper limit that was fixed to 15 words could better be adjusted to the subject experience. Unknown words were considered to be a difficulty only in 5.3% of the exercises. This might be explained by the fact that all exercises consist of selecting the correct words from a set of candidates, i.e. all the words are written somewhere in the screen. The task of writing down the words would significantly increase the cognitive load, introducing aspects out of the listening comprehension scope (such as spelling).

5.4.5. ASR error awareness

In order to evaluate the ASR error awareness, we conducted an experiment involving all the queries made by the participants in sets 4 and 5. A native speaker manually checked the retrieved utterances for errors. The results were compared with the answers of the non-native participants, concerning the presence of errors. The percentage of ASR errors detected by the participants is 43%. The fact that the errors very often involved very short function words, in highly reduced forms, may be responsible for this relatively low awareness. The participants also falsely identified errors in 14% of these queries. We still need to prove that additional motivation (such as bonus points) could improve ASR error awareness.
Table 4
Final questionnaire and Likert scale mean rates given by the users.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Mean score</th>
</tr>
</thead>
<tbody>
<tr>
<td>The audio and video slow down feature is useful</td>
<td>4.1</td>
</tr>
<tr>
<td>Videos make the exercises more pleasant</td>
<td>3.8</td>
</tr>
<tr>
<td>Videos allow to understand better what is said</td>
<td>3.5</td>
</tr>
<tr>
<td>I prefer to put all the words in the same order that I heard them, than tick them in an unsorted list</td>
<td>3.6</td>
</tr>
<tr>
<td>The difficulty level of the sentences is high</td>
<td>2.8</td>
</tr>
<tr>
<td>Listening/watching the audio/video at the same time that I solve the exercise is more useful that memorizing</td>
<td>3.9</td>
</tr>
<tr>
<td>It is useful to see the words being highlighted as they are being said</td>
<td>4.0</td>
</tr>
<tr>
<td>The fact that all the sentences are from real broadcast news shows is appealing</td>
<td>4.2</td>
</tr>
<tr>
<td>The fact that I can solve exercises that are based on recent news is appealing</td>
<td>4.1</td>
</tr>
<tr>
<td>The possibility to search for words is motivating</td>
<td>3.4</td>
</tr>
<tr>
<td>The scores make the exercises more appealing</td>
<td>4.0</td>
</tr>
</tbody>
</table>

5.5. Questionnaire analysis

At the end of the exercises, the students answered a feedback questionnaire. The users had to rate a set of statements according to their level of agreement, discretized on a five-point Likert scale: 1 = completely disagree, 2 = disagree, 3 = neither agree, nor disagree, 4 = agree, and 5 = completely agree. Table 4 lists those statements and the resulting mean scores. Standard deviation was about 1 for all the items.

The participants agreed that the slow-down feature was useful (4.1). This result agrees with the fast speech rate pointed out as the first difficulty. Video was judged as a positive, pleasant and useful feature (3.8). Also, with a 3.9 value, was the preference to watch the video at the same time as solving the exercise. Some users even pointed out that memorizing the utterances distracted them from the main goal. Users agreed that real and recent BN content is an additional motivation. The scores provided after each exercise were also considered to be a positive challenging feature, with a value of 4.0. The drag-and-drop interface was slightly preferred to the tick box (3.6). Finally, with an average value of 3.4, the students showed interest with respect to the search mechanism.

6. Conclusions

In this paper, we have presented listening comprehension exercises for learners of Portuguese as a second language. The exercises were recently integrated in a new Web platform called Daily-REAP\(^7\) that provides written news and broadcast news videos from the last seven days.

The video processing pipeline, the filters used to select the sentences, and the two complementary methods used to generate word distractors allow us to generate exercises automatically. A subjective evaluation validated the feasibility of this approach. The automatic set was assigned good to very good quality scores, only slightly lower than the scores of a manual set.

A survey completed by non-native subjects with various background knowledge in European Portuguese identified the users’ preferred exercise features. The participants appreciated the presence of video, especially when featuring recent news content. The possibility to search keywords to retrieve exercises was also seen as interesting, in order to listen to words in varying contexts and in real documents. Concerning the interface, the users slightly preferred to drag-and-drop words than just tick them in a list.

The study focused on the development and assessment of the exercises, but despite the very positive feedback, further evaluation needs to probe whether they aid language acquisition. The overall assessment of the whole Daily-REAP platform would also be interesting, in particular in terms of engagement. An indication of the engagement of the learners could be achieved by measuring their activity on the site over a period of time of several weeks.

---

\(^7\) http://call.12f.inesc-id.pt/daily-reap.pt.
The multimedia platform is expected to be especially valuable for distance learners who need opportunities to access authentic audio documents in EP. The daily update of the material, and the corresponding exercises (22 new sentences on average every day) may potentially increase REAP users’ motivation.

Acknowledgment

This work was partially supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under projects CMU-PT/HuMach/0053/2008 and PEst-OE/EEI/LA0021/2011.

References


