Digital Talking Books in Multiple Languages and Varieties

Isabel Trancoso*, António Serralheiro, Céu Viana†, Diamantino Caseiro*, Isabel Mascarenhas††**

*INESC-ID/IST
Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
{isabel.trancoso, antonio.serralheiro, dcaseiro}@inesc-id.pt
**INESC-ID/Academia Militar †CLUL ††CLUL/FCSH

Abstract

This paper describes our work in digital talking book alignment, starting by our earlier efforts for the alignment of books in European Portuguese, and ending with the two challenges we are currently facing of aligning books in different varieties of Portuguese and aligning parallel books in different languages. Our alignment module proved robust enough for porting to other varieties of Portuguese, despite the fact that the assessment of the recognition module without adaptation for broadcast news had shown a very severe degradation for some of the varieties. Our work on the alignment of audio in one language with text in another is still ongoing, but informal evaluation has led us to expect that digital talking books can be a valuable tool for second language learning.

1. Introduction

Our main task in a Digital Talking Book (DTB) project[1] is the automatic alignment of the recorded speech with the written text. This paper summarizes our first efforts with the alignment of books in European Portuguese and describes the two challenges we are currently facing of aligning books in different varieties of Portuguese and aligning parallel books in different languages.

The motivation for adding this multi-variety dimension to DTBs is twofold. From a research point of view, adding different varieties of the same book allows us to verify the potential or limitations of our automatic alignment system, originally trained for European Portuguese. From a commercial point of view, it could potentially spread the audience of a given digital spoken book, since there is a common perception among speakers of Portuguese that they would very much prefer to hear recordings in their own variety and would even have difficulties in understanding a different variety, if not familiar enough with it.

The motivation for adding the multilingual dimension to DTBs is to address the question of whether digital spoken books can be used as a learning tool, by second language learners. Compared with other languages, such as English, there are relatively very few materials for learning Portuguese that include recordings with the good articulation and intonation that characterize professional speakers. So DTBs could be a very good way to get familiar with the language, for students of Portuguese. Whereas more advanced learners may profit from having a DTB with the audio and the text both in Portuguese, less advanced learners could profit from having the audio in the language they are learning (in our case, Portuguese), and the text in their first language or a language they already know (we chose English, because of its use as a lingua franca). This would potentially enable DTBs to be applied for e-learning besides their more traditional application to e-inclusion, namely for visually impaired users and, more recently, for dyslectic students.

Adding these multi-variety/multilingual dimensions to our DTB player has been fairly easy. The player was developed using a model based framework for adaptive multimodal environments (Duarte and Carriço, 2005). Besides supporting the features described in the DTB standard[2], the player introduces features complementing the synchronized presentation of text and audio, such as: addition of content related images; variable synchronization units, ranging from word to paragraph; annotation controlled navigation; definition of new reading paths; adaptation of the visual elements; behavioral adaptation reflecting user interaction, amongst others. Some of these features are visible in Figure 1.

Figure 1: The DTB player interface, presenting the toolbar (top), main text (center), table of contents (left), annotations (right) and an image (bottom left).

For the sake of testing both this interface and the automatic alignment system, we initially built a small repository with different types of book in European Portuguese: fiction, poetry, children's stories, didactic text books, etc. A similar small repository was built for Brazilian Portuguese, in the scope of a cooperation project with the

[1] RiCoBA - Rich Content Books for All

[2] www.niso.org/standards/resources/Z39-86-2002.html

Universidade Rio Grande do Sul. More recently, we extended this repository to include different versions of the same short story *O Monge Desastrado* (The Clumsy Monk, by Ana Maria Magalhães and Isabel Alçada). This short story for youths (15 min. approximately) was recorded by speakers from several varieties of Portuguese: 2 speakers of European Portuguese (EP), 2 speakers of Brazilian Portuguese (BP), and 3 speakers of African Portuguese (AF), from the countries of Angola, Mozambique and Guinea-Bissau. The same story was also manually translated to English, in order to have a pilot parallel text for our study.

This paper is structured into 3 main sections. Section 2 summarizes our initial work in the alignment of books in European Portuguese. Section 3 addresses the multi-variety problem and section 4, the multilingual one.

2. Aligning spoken books in European Portuguese

The alignment of DTBs in European Portuguese involved using our automatic speech recognition system (ASR), originally trained for broadcast news (BN) in European Portuguese, in a forced alignment mode. The BN recognizer uses hybrid acoustic models that try to combine the temporal modeling capabilities of hidden Markov models with the pattern classification capabilities of MLPs (Multi-Layer Perceptrons). The models have a topology where context-independent phone posterior probabilities are estimated by three MLPs given the acoustic parameters at each frame. The streams of probabilities are then combined using an appropriate algorithm (Meinedo and Neto, 2000). The MLPs were trained with different feature extraction methods: PLP (Perceptual Linear Prediction), Log-RASTA (log-RelAtive SpecTrAl) and MSG (Modulation SpectroGram). Each MLP classifier incorporates local acoustic context via an input window of 7 frames. The resulting network has a non-linear hidden layer with over 1000 units and 40 softmax output units (38 phones for EP plus silence and breath noises). The vocabulary includes around 57k words. For the BN development test set corpus, the out-of-vocabulary (OOV) word rate is 1.4%. The lexicon includes multiple pronunciations, totaling 66k entries.

The first challenge that we faced was to overcome the memory limitations of our original decoder (Serralheiro et al., 2003). This decoder proved unable to deal with long books, which implied splitting the text and the corresponding audio into short segments. This very tedious splitting task was avoided by adopting a totally different decoder, based on weighted finite state transducers (WFSTs) (Mohri et al., 2000). The decoder has a search space defined by a distribution-to-word transducer, constructed as $H \circ L \circ G$, where H is the HMM or phone topology, L is the lexicon and G is the language model. For alignment, G is just the sequence of words that constitute the orthographic transcription of the utterance. The decoder was extended to deal with special labels, on the input side, that are internally treated as epsilon labels, but are used to mark time transitions or boundaries. Whenever such end-of-segment labels are crossed, the time is stored in the current hypothesis. The user may choose to place those labels at the end of each phone WFST or at the end of each word WFST.

This decoder proved very robust even for aligning very long recordings. A 2-hour long book could be aligned in much less than real-time. For most of the EP spoken books, the evaluation of the aligner performance was done only informally. In fact, the visual inspection of the word labels generally guaranteed quite good results at this level. However, the alignment of some EP poetry books revealed some problems related to specific prosodic characteristics, namely in terms of larger phone durations. In order to get a more precise alignment at the phone level, we first tried alternative pronunciation rules (Trancoso et al., 2003) and later speaker adapted acoustic models.

Speaker adaptation is indeed a very straight forward procedure for digital spoken books. We have recently evaluated the phone-based alignment error before and after speaker adaptation in a small poetry corpus that by visual inspection was considered one of the worst alignment cases. The training set includes 48 minutes. The manually aligned test set includes only 2 minutes, amounting to around 580 phonetic instances. The average alignment error in this test set is less than 1 ms, without and with speaker adaptation, showing that no systematic errors are introduced. Before speaker adaptation, the average absolute error is 44.6 ms, decreasing to 22.8 ms after adaptation. 90% of the phones were correctly aligned in less or equal than 90 ms, before adaptation, and 50 ms, afterward, showing an improvement of approximately 45%. The improvements stabilized after 6 iterations. These metrics in terms of phone boundaries are not as significant to DTBs as the ones related to word boundaries. Before speaker adaptation, the average absolute error in terms of word boundaries was 58.1 ms, decreasing to 30.7 ms after adaptation.

The largest differences between the manual and automatic word boundaries occurred at strong coarticulation between consecutive words, when the phones at the boundaries are of the same type (either vowel-vowel or coronal consonant-coronal consonant). In such cases, the criterion used to set the manual boundary was to mark it at the midpoint of the transition segment (van Santen, 1992). The automatic aligner was not trained for this criterion. Other significant differences were due to very audible breath noises that were ignored in both the automatic and manual segmentation.

The alignment of didactic text books revealed the need for studying the problem of reading mathematical expressions. The DAISY Consortium[3] has recently approved MathML as a modular extension for mathematics. We are currently working in integrating this into our DTB player, and studying the implications for our aligner module. We are also addressing a related problem in the framework of another national project on lecture transcription, in which lectures of technical courses, such as Algebra, are being automatically transcribed.

[3] www.daisy.org

3. Alignment of spoken books in different varieties of Portuguese

Portuguese is spoken by more than 170 million people in virtually all continents in many different varieties. Although a detailed linguistic characterization of the different varieties is clearly out of the scope of this paper, it is worth mentioning the most striking differences. These are much more relevant in terms of phonetics and phonology than in terms of orthography and syntax.

Although there is an orthographic convention among the countries of CPLP (Community of countries of Portuguese language), there are minor differences, representing some phonetic and phonological specificities: the optional suppression of unpronounced consonants in BP (e.g. *acção* / *ação*, *excepto* / *exceto*), the optional use of hyphenation, and differences in diacritics (e.g. *tranquilo* / *tranqüilo*, accounting for the fact that *u* is pronounced as /w/, instead of deleted as in the general case involving *qui* or *que* sequences; *Jerónimos* / *Jerônimos*, accounting for the different vowel quality). Besides these differences, there are also significant ones concerning the use of prepositions, the position of clitics and the alternative use of infinitive/gerundive verb forms (e.g. *estava sempre a meterse em sarilhos* vs. *estava sempre se metendo em sarilhos* - was always getting into trouble). In order to take into account these minor differences, we have asked the first BP speaker to make whatever changes were needed in the orthography to make the reading more natural, prior to the recording. This resulted in changes in around 10 sentences.

There is common agreement that one of the most striking differences between the Brazilian and European varieties concerns vowel reduction, which is much more extreme in EP than in BP (Mateus and d'Andrade, 2000), (Barbosa and Albano, 2004). In fact, although both varieties distinguish between seven vowels in stressed position (/i e ɛ a ɔ o u/), they do not have the same reduction patterns, and quality changes are not sensitive to the same constraints.

There are also significant differences with respect to the consonantal system. One of them is the affrication of the dental plosives /t/ and /d/ before a high front vowel or glide in BP. Other (potentially troublesome) differences concern the phonetic realizations of /l r s/ in coda position.

Studies of the phonetic and phonological differences between these two varieties and the African ones are much more scarce (Gonçalves, 2001) (Mendes, 1985) (Mingas, 2000). They mention, however, that unstressed vowels in pre-tonic position are much more closer to BP than to EP. Although vowel reduction appears to be less extreme than the one observed for EP, vowel centralization and deletion are much more frequent than in BP. Regarding the consonantal system, the most striking differences in our corpus were observed for liquids: no contextual distinction is made between a simple tap and a trill, as /r/ is always pronounced as a tap; the lateral palatal (/L/) is most often pronounced as [lj] and the quality of the /l/, as well as the absence of velarization, distinguish all African varieties from both BP and EP. It is also worthwhile mentioning that the frequent absence of plural marks in many nouns and adjectives is another characteristic of African varieties, as gender and number are generally marked on pre-nominal specifiers and/or modifiers only.

Since we are using an aligner whose models were originally trained for broadcast news in European Portuguese, it is interesting to compare the performance of our recognizer when dealing with these other varieties. Table 1 presents the word error rate (WER) results obtained for a relatively small corpus of broadcast news in different varieties (results for all conditions - prepared/spontaneous speech, clean/music/noisy background, etc.). The size of the corpus in each variety is also shown in terms of number of words. This table shows that the recognizer performance degrades more or less drastically when recognizing other varieties such as the one spoken in Brazil or the ones spoken in African countries with Portuguese as official language. The degradation is much more pronounced for Brazilian Portuguese than for African Portuguese. It is also interesting to notice that for African speakers whose origin could not be guessed by the annotators, the degradation is very mild. Given the differences between the varieties that we have briefly described, the degradation in recognition performance is easily explained.

Table 1: Recognition results of broadcast news in different varieties.

Variety	Size	%WER
Portugal	19,157	20.1
Brazil	19,853	62.6
Angola	6,041	30.3
Cape-Verde	6,258	31.3
Guinea-Bissau	6,376	36.3
Mozambique	4,611	36.0
São-Tomé and Príncipe	4,356	38.6
Undistinguished African	3,690	23.1

These results reveal the inadequacy of the acoustic, lexical and language models trained for European Portuguese when dealing with other varieties. They also lead us to expect some degradation in terms of alignment precision.

Our small multi-variety corpus allowed us to access this precision. Unfortunately, as described above, we only had speakers from 5 of these varieties and the recording conditions are not exactly the same for 2 of the speakers, although all were recorded in high quality sound proof environments. Manual labeling was restricted to word boundaries and the first 2 minutes[4].

The visual inspection of the alignment results proved fairly good for all speakers. The objective results are shown in Table 2 in terms of average absolute error, and the maximum time below which 90% of the words were aligned. These results show that the degradation in recognition performance is not reflected in the aligner perfor-

[4]It would also be interesting to discuss the results of alignment boundaries at a phone level, although they are not so crucial for the purpose of spoken book alignment. However, the manual correction at this level is very time consuming and is not yet complete for all varieties.

mance. In fact, the performance of the aligner is quite robust in terms of varieties, even achieving better results than for our worst-case poetry corpus in EP.

The aligner is also robust to minor reading errors, such as the occasional word insertion or deletion, some hesitations, and the relative frequent absence of fricatives in coda position. In fact, the phenomenon of omitting the plural marks in nouns and adjectives was observed for the two first African speakers (AN1 and GB1). For GB1 this omission occurred in 50% of these plural forms. This phenomenon was not observed in the speech of journalists or government officials in our broadcast news corpus, but only in interviews with the public. For spoken books, we originally asked our (non-professional) speakers to repeat such sentences, but this did not contribute to a good naturalness.

Table 2: Alignment results of spoken books in different varieties. AV.Abs.Error (Average Absolute Error); Max90% (Maximum time below which 90% of the word boundaries were correctly aligned). Speakers: EP1 (male, European Portuguese: Lisbon), EP2 (female, European Portuguese: Lisbon), BP1 (male, Brazilian Portuguese: Rio de Janeiro), BP2 (male, Brazilian Portuguese: Rio Grande do Sul), AN1 (male, Angola), GB1 (female, Guinea-Bissau, MO1 (male, Mozambique).

Speaker	Av.Abs.Error ms	Max90% ms
EP1	11.3	32.9
EP2	18.6	37.0
BP1	16.2	31.9
BP2	16.4	38.7
AN1	20.9	42.6
GB1	22.6	45.4
MO1	18.2	43.7

4. Alignment of spoken books in different languages

We can think of at least three different ways of using DTBs for second language learning. The first one, which assumes a higher proficiency in L2 (the target language), uses audio and text in L2, and requires no changes at all to the monolingual interface. The second one uses audio in L2 and text in L1 (the original language of the student), and requires only building new DTB files (see Fig. 2). The third one, which we have not yet implemented, uses audio in L2 and a horizontally divided text screen in which the parallel texts in L1 and L2 may be simultaneously visualized.

The last two options described above assume the automatic alignment of the two parallel texts which is done in two steps: sentence alignment, followed by word alignment. In the sentence alignment step, the texts are segmented into sentences using the full stop "." as the sentence delimiter, and hand-crafted rules to detect when the full stop terminates a sentence, involving for instance, capitalization of the next word and an exception list. After

Figure 2: The DTB player interface with the translated text of Fig. 1 (L1=English).

this segmentation step, the sentences are aligned using a dynamic programming algorithm that allows alignments of 1-to-1 sentence, of 2-to-1 sentences and of 1-to-2 sentences (Gale and Church, 1993). The second step, word-to-word alignment is done using IBM-4 statistical alignments as implemented in the GIZA++ tool (Och and Ney, 2000). Alignments in both directions (source-to-target and target-to-source) are performed and combined using heuristic refinement rules. Because the books are usually too small to train reliable word alignments, they are aligned in the context of a larger (out of topic) corpus such as *europarl* (Koehn, 2005).

Sentence alignment was error free for the relatively small parallel corpus we have used in our tests. The default synchronization level for L2-learners should therefore be the sentence level. The main problem with this solution is that very often in Portuguese novels, sentences are too long, which makes the learning process more difficult. So, users should have the opportunity to switch to word level alignment. In our first experiments, we opted to show only (by highlighting) the one-to-one alignments[5] in order to avoid the typical errors caused by rare words and constructions. In our parallel corpus, only 64% of the Portuguese words are aligned on a one-to-one basis, so this solution is obviously suboptimal. One additional disadvantage is that it implies frequent inversions of the left-to-right order of highlighting. For instance, in a sequence such as *mapa do mundo*, translated as *world map*, *map* (translation of *mapa*) would be heard and highligthened before *world* (translation of *mundo*). Hence, we need to test whether the skips and the order reversal may be confusing for users. Many other alternatives are possible and worth exploring, such as highlighting groups of words, to avoid word order problems, by merging their time and position tags.

Using DTBs for second language learning opens up a wide range of possibilities. One could for instance integrate a module for automatically detecting and highlighting potentially unfamiliar words and grammar constructions, given a model of the user's reading skills. Another possibility would be to allow the users to click on words that (after lemmatization and part-of-speech tagging) could

[5] See parallel audio book demo in http://www.l2f.inesc-id.pt/projects/ipsom/pasteis_pt/monge_en.avi

direct the user to an electronically available dictionary to search for their meaning in different contexts. It would also be very interesting to test the DTB with second language learners with different levels of proficiency and try to find out at what time they start preferring the L1-audio/L1-text interface.

Given our preliminary work on L2 learning, we have only conducted informal tests with two non-native users. These tests confirmed the potential of this application, and demonstrated how some of the DTB player features, such as slowing down the narration speed may be specially relevant for language learning. We plan to conduct more formal tests in the near future, during the Summer courses for foreign students, in order to evaluate the pros and cons of different interfaces, using either error-prone word-based alignment or virtually error-free sentence alignment.

5. Conclusions and Future Work

This paper described our work in spoken book alignment, starting by our earlier efforts for the alignment of books in European Portuguese, and ending with the two challenges we are currently facing of aligning books in different varieties of Portuguese and aligning parallel books in different languages.

Our alignment module proved robust enough for porting to other varieties of Portuguese, despite the fact that the assessment of the recognition module without adaptation for broadcast news had shown a very severe degradation for some of the varieties.

Our ongoing work on improving parallel text alignment may significantly contribute to improve the interface of spoken books in different languages. Given the large number of misaligned verbal forms (namely with cliitics), adding morphological information is one of the approaches. Another one is detecting groups of words corresponding to syntactically or semantically meaningful units. Furthermore, the use of bilingual dictionaries and cognate matching, may increase alignment robustness in the presence of infrequent words.

Despite the early stage of our work in this area, informal evaluation has led us to expect that DTBs can be a valuable tool for second language learning.

Acknowledgments The authors would like to thank the 5 readers of the parallel corpus, our colleagues Carlos Duarte and Luís Carriço in the RiCoBA project, and also Julia Hirschberg, Maxine Eskenazi and Cristina Mota for many valuable suggestions. This work was partially funded by FCT projects RiCoBA (POSC/EIA/61042/2004) and WFST (POSI/PLP/47175/2002). INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

6. References

Barbosa, P. and E. Albano, 2004. Brazilian Portuguese - Illustrations of the IPA. *Journal of the Int. Phonetic Association*, 34(2).

Duarte, C. and L. Carriço, 2005. Users and usage driven adaptation of digital talking books. In *Proc. 11th International Conference on Human-Computer Interaction (HCII 2005)*. Las Vegas, Nevada.

Gale, W. and K. Church, 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Gonçalves, P., 2001. Panorama geral do Português de Moçambique. *Revue Belge de Philologie et d'Histoire*, 79:977–990.

Koehn, P., 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. Machine Translation Summit*. Phuket, Thailand.

Mateus, M. H. and E d'Andrade, 2000. *The Phonology of Portuguese*. Oxford: Oxford University Press.

Meinedo, H. and J. Neto, 2000. Combination of acoustic models in continuous speech recognition hybrid systems. In *Proc. ICSLP '2000*. Beijing, China.

Mendes, B., 1985. *Contributo para o Estudo da Língua Portuguesa em Angola*. Lisboa: Instituto de Linguística da Faculdade de Letras de Lisboa.

Mingas, A., 2000. *Interferência do Kimbundu no Português Falado em Luanda*. Luanda: Chá de Caxinde.

Mohri, M., F. Pereira, and M. Riley, 2000. Weighted finite-state transducers in speech recognition. In *ASR 2000 Workshop*.

Och, F. and H. Ney, 2000. Improved statistical aligment models. In *Proc. 38th Annual Meeting of the ACL*. Hong Kong, China.

Serralheiro, A., I. Trancoso, D. Caseiro, T. Chambel, L. Carriço, and N. Guimarães, 2003. Towards a repository of digital talking books. In *Proc. Eurospeech '2003*. Geneva, Switzerland.

Trancoso, I., D. Caseiro, C. Viana, F. Silva, and I. Mascarenhas, 2003. Pronunciation modeling using finite state transducers. In *Proc. 15th International Congress of Phonetic Sciences (ICPhS'2003)*. Barcelona, Spain.

van Santen, J., 1992. Contextual effects on vowel duration. *Speech Communication*, 11:513–546.