Aligning and Recognizing Spoken Books in Different Varieties of Portuguese

Isabel Trancoso[1], *António Serralheiro*[2], *Céu Viana*[3], *Diamantino Caseiro*[1]

[1] L^2F INESC-ID/IST, [2] L^2F INESC-ID/Academia Militar, [3] CLUL

Lisbon, PORTUGAL

Isabel.Trancoso@inesc-id.pt

Abstract

This paper tries to present digital spoken books as a useful diagnostic tool for detecting alignment and recognition problems and for studying the porting of these technologies to different varieties of the same language - Portuguese, in our case. We summarize the main differences between European and Brazilian Portuguese (EP/BP) and describe how they affect the GtoP system. Despite the small size of our parallel spoken book corpus in the two varieties, our preliminary experiments confirmed our expectations in terms of the effectiveness of an EP-trained aligner used on BP spoken books. They also confirmed the inadequacy of an EP Broadcast News recognizer tested over literary contents, and the expected degradation in recognition scores caused by using that recognizer on a BP spoken book. Pronunciation adaptation was tested by adding variants derived by the BP GtoP system to our EP lexicon, resulting in a very small improvement in terms of recognition scores.

1. Introduction

Aligning spoken books has been the major task of our lab in a national project dealing with digital spoken books in European Portuguese and their interfaces for visually impaired users [1]. The alignment between the text and the audio files has been successfully achieved by some modifications made to our decoder based on weighted finite-state transducers. With these modifications, a 2-hour long spoken book was aligned in a single step in much less than real time.

The success of these alignment experiments made with a Broadcast News (BN) recognizer [2] led us to investigate two different research directions: the first one consisted of running recognition experiments with the same spoken books, in order to access the effect of the BN vs. literary differences in content; the second one consisted of extending our alignment and recognition experiments to digital spoken books in a different variety of Portuguese - Brazilian Portuguese. European and Brazilian varieties of Portuguese (henceforth denoted as EP and BP, respectively) have significant differences, in the written and specially in the spoken form.

This paper tries to present digital spoken books as a useful diagnostic tool for detecting alignment and recognition problems and for studying the porting of these technologies to different varieties of the same language. We shall try to answer some basic questions: (1) Can the automatic alignment system using EP-trained models be used for the BP variety? (2) What is the degradation one can expect from testing an EP recognizer trained for BN over EP spoken books? (3) What is the degradation one can expect from testing the same recognizer with BP spoken books? (4) In the absence of enough material to train speaker independent acoustic models for BP, can a BP-adapted pronunciation lexicon be of any help?

The BN recognizer that was used in all the alignment and recognition experiments 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 [3]. The MLPs were trained with different feature extraction methods: PLP (Perceptual Linear Prediction), LogRASTA (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 language model was created by interpolating a newspaper text language model built from over 400M words with a backoff trigram model using absolute discounting, based on the training set transcriptions of our BN database (45h). The perplexity is 139.5. 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, which we shall denote as EP0, includes multiple pronunciations, totaling 66k entries.

Our EP repository of aligned spoken books is already quite extensive, with fiction books, didactic text books, poetry, children's stories, etc. Because of funding limitations, our BP repository is very limited. In fact, the only parallel corpus we have in the two varieties is a short story for youths - *O Monge Desastrado* (by Ana Maria Magalhães and Isabel Alçada). The EP version was read by a professional speaker, with a Lisbon accent. The BP version was read by a young writer, with an educated *Rio de Janeiro* (RJ) accent. Both speakers took approximately 10 minutes to read the 1600-word story, although the speaker rate of the BP speaker was slightly slower (6%). The recording conditions were not exactly the same, unfortunately, but both were recorded in high quality sound proof environments. Prior to recording the text, we have asked the BP speaker to make whatever changes were needed in the orthography to make the reading more natural. This resulted in changes in around 10 sentences. Whereas our alignment experiments covered all types of spoken books, the recognition experiments reported in this paper were limited to this small parallel corpus.

Section 2 will summarize the main differences between the two varieties. Section 3 describes how these differences affected the porting of our EP Grapheme-to-Phone (GtoP) conversion module to BP. The four following sections will attempt to answer each of the questions above. The paper concludes

with a list of future problems to be addressed in this area.

2. Main differences between EP and BP

2.1. Orthographic and syntactic differences

The current orthographic convention allows for 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 meter-se em sarilhos* vs.*estava sempre se metendo em sarilhos* - was always getting into trouble).

2.2. Phonetic and phonological differences

There is common agreement that one of the most striking differences between Brazilian and European varieties concerns vowel reduction, which is much more extreme in EP than in BP [4], [5]. 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.

In pre-tonic position, /e/-/ɛ/ and /o/-/ɔ/ contrasts are neutralized in BP, and the seven-vowel system reduces to the five-vowel inventory [i e a o u], whereas in post-tonic position it reduces to [i e a u] in mid syllables and to [i ɐ u] word finally (/i e ɛ/ and /u o ɔ/ merge to [i] and [u] respectively, and /a/ is raised to [ɐ]). EP presents a single reduction pattern, as /e ɛ/ reduce to [ɨ] and /o ɔ/ to [u] when unstressed, regardless of their position relatively to stress. With few exceptions, /a/ is also raised to [ɐ] in all unstressed positions.

Mismatches between EP and BP surface forms are aggravated by the fact that in EP unstressed high vowels are often deleted and rather long consonant clusters may surface within as well as across word boundaries (e.g. *se desprezarmos* [sdʃpɾˈzaɾmʷʃ] 'if we ignore'). These clusters are not allowed in BP, as syllable nuclei are obligatorily filled (e.g. [sideʃpɾeˈzaʁmuʃ], in RJ). Accordingly, obstruent sequences with underlying empty nuclei are broken by an epenthetic vowel (e.g. *psicologia* [pisikoloˈʒiɐ] 'psychology', *afta* [ˈafitɐ] 'aphtha' in BP vs [psikluˈʒiɐ], [ˈaftɐ] in EP). Loanwords can also be treated differently (e.g. [iʒˈnɔbi] in BP vs [sˈnɔb] in EP).

On the other hand, in BP, as first shown by [6], pre-tonic vowels must also agree in height with the word stressed vowel, whereas in EP no such vowel harmony occurs (e.g. *preferência* [pɾefeˈɾẽsjɐ] 'preference' but *preferível* [pɾifiˈɾivew] 'preferable' in BP, vs [pɾifiˈɾẽsjɐ], pɾifiˈɾivɛɬ], respectively, in EP careful pronunciation). Although stressed vowels are similar in both varieties, there are some differences worth mentioning. In EP's Lisbon dialect, an additional stressed vowel ([ɐ]) may appear:

(1) [ɐ]/[a] distinguish between the 1st person plural of verbal present and past tense forms, respectively (e.g.*pensamos* [pẽˈsɐmuʃ] 'we think' / *pensámos* [pẽˈsamuʃ] 'we thought');

(2) low vowels are raised before heterosyllabic nasal consonants but they are not fully nasalized (e.g. *cara* [ˈkaɾɐ] 'face' / *cana* [ˈkɐnɐ] 'cane');

(3) front vowels centralize before palatal consonants and glides (e.g. *bandeja* [bɐ̃ˈdɐʒɐ] 'tray', *espelho* [ʃˈpɐʎu] 'mirror', *lei* [lɐj] 'law').

In BP, the two forms in (1) are homophones ([pẽˈsɐ̃mus]) and stressed vowels preceding heterosyllabic nasal consonants are strongly nasalized ([ˈkɐ̃nɐ]). As no centralization occurs, the forms in (3) may be pronounced as [bɐ̃ˈdeʒɐ], [iʃˈpeʎu] and [lej], respectively. In this case, a light diphthong may also surface (e.g. [bɐ̃ˈdejʒɐ] in BP and [bɐ̃ˈdɐjʒɐ] in EP).

This is a general phenomenon, which can also occur in both varieties with vowel /a/ as well (*caixa* ([ˈkaʃɐ] or [ˈkajʃɐ] 'box'). In EP, at least in our database, it never happens, however, with back rounded vowels. This is not the case in BP, as in the training lexicon, forms like *luz* [lujʃ] 'light' and *arroz* [aˈʁojʃ] 'rice' are attested. Contrarily to EP, this kind of variation also occurs in unstressed position and before other coronals too (e.g. *faxina* [faˈʃinɐ] or [fajˈʃinɐ] 'cleaning', *mas* [maʃ] or [majʃ] 'but' in BP vs [fɐˈsinɐ] and [mɐʃ] in EP).

With respect to the consonantal system, there are also some important phonetic differences that need to be taken into consideration. One of them is the affrication of the dental plosives /t/ and /d/ before a high front vowel or glide in BP (e.g. *fadiga* [faˈdʒigɐ] 'fatigue', *catecismo* [katʃiˈsiʃmu] 'catechism' pronounced as [fɐˈdigɐ] and [kɐˈtsiʃmu] in EP).

Other (potentially troublesome) differences concern the phonetic realizations of /l ɾ s/ in coda position. As it was already visible in some of the examples given above, while /l/ is velarized in EP, it is generally rendered as [w] in BP, forming a diphthong with the preceding vowel (e.g. *incrível* [ĩˈkɾivew] 'incredible', *sol* [ˈsɔw] 'sun' vs [ĩˈkɾivɛɬ], [ˈsɔɬ], in EP).

In the examples above, the coronal fricative in coda position was transcribed as [ʃ] or [ʒ] for both BP and EP. Those are, in fact, the most frequent realizations in RJ (82%). They are not categorical, as in this dialect [s], [z], [h] or [null] constitute other alternatives [7] [8]. By contrast, in the São Paulo's dialect (SP), the alveolar variants are predominant. The authors observe a consistent tendency for an increased palatalization in word medial position, which is in accordance with the behavior of our speaker. In EP, however, [s] and [z] are always associated with a syllable onset.

As for rhotics, EP is closer to SP as /r/ in syllable coda is realized as a tap or a flap in 62% of the cases. However, SP has retroflex realizations that do not occur in our database [7] [8]. In RJ, the velar fricative is predominant, which is common in EP. It corresponds however to 'strong-r' in EP that does not occur in this position.

3. Porting the grapheme-to-phone conversion module to BP

The above differences were taken into account in the porting of our GtoP from EP to BP. We used the same FST-based rule framework, [9] and a BP pronunciation lexicon with 19.1k entries (no inflected verbal forms) that was subdivided into training (15.3k) and test (3.8k) sets.

The number of rules derived for BP was smaller than for EP ($\sim$ 250 vs. $\sim$ 340). Over 30 EP rules are not present in the BP set because of the above mentioned adoption of a different orthography. The remaining differences can be mostly found in the extra rules for pronouncing graphemes *a, e, o* in EP, which are not counterweighted by a full implementation of vowel harmony rules in BP.

The rules resulted on a word transcription error rate of 4.3%, slightly higher than for EP (3.3%, on a different test set). The type of errors, however, was very similar. Most of the errors occurred in the transcription of graphemes *e* ([e]/[ɛ] - 24%), *o* ([o]/[ɔ] - 16%), *a* ([a]/[ɐ] - 9%, before nasal consonants and in word final position), and *x* (5%). Foreign words were responsible for 7%. The fact that we did not include any morphological analysis was responsible for 9% of the errors (e.g. *sobrecarga*, where the final *e* of prefix *sobre* was not correctly transcribed). 14% of the errors occurred in vowel sequences which were treated as raising diphthongs by the rules, but sometimes marked as hiatus in the reference lexicon.

4. Alignment of EP and BP spoken books

Our WFST-based decoder [10] 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.

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, even when the alignment involved 2-hour long recordings. The visual inspection of our small BP corpus showed very promising results, similar to those obtained with the corresponding EP one.

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 [11] 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. 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 1ms, without and with speaker adaptation, showing that no systematic errors are introduced. Before speaker adaptation, the average absolute error is 44.6ms, decreasing to 22.8ms after adaptation. 90% of the phones were correctly aligned in less or equal than 90ms, before adaptation, and 50ms, afterward, showing an improvement of approximately 45%. The improvements stabilized after 6 iterations. The alignment of the BP spoken book, also benefited from speaker adaptation, as expected, although no formal evaluation was yet conducted.

5. Recognition of EP spoken books

The recognition results obtained with the parallel EP corpus are shown in the first row of Table 1. The word error rate is far greater than the one obtained for read speech, studio recordings in Broadcast News (10.9%). The causes for this degradation may be linked with the high OOV rate of the parallel spoken book corpus (5.4%), as one OOV term can lead to between 1.6 and 2 additional errors [12], and with the very high perplexity computed over this corpus (443.9). In fact, the newspaper texts that were used to build the lexical and language models do not typically contain many verbal forms in the first or second persons, contrasting with stories such as our parallel corpus, with much dialog between the characters. It is also interesting to notice that 30% of the OOV forms are verbal forms with clitics. Although BN can be considered a very wide domain, the inadequacy of the BN recognizer for spoken books was proved by running the recognizer on other books, where similar OOV rates were found.

The relatively high deletion rate is also worth investigating. In fact, 66% of the deleted words are very short function words. In particular, the preposition *de* ([dɨ]) is very often deleted when the following word starts by a plosive sound (e.g. *nem DE comer, nem DE dormir, nem DE descansar*).

Table 1: Recognition results with EP and BP spoken books.

AUDIO	LEX	%CORR	%SUB	%DEL	%INS
EP	EP0	69.2	25.3	5.5	4.2
BP	EP0	57.1	36.5	6.4	8.5
EP	BP1	68.0	26.2	5.8	4.0
BP	BP1	59.0	34.6	6.4	7.0
EP	BP2	68.9	25.7	5.4	4.3
BP	BP2	60.5	32.9	6.6	6.9

6. Recognition of BP spoken books

The recognition results obtained with the small BP corpus are shown in the second row of Table 1. The degradation was expected, given our previous experience with BN segments in BP. The vowel reduction differences between the two varieties are patent in the substitution errors caused by OOV words as, for instance, in the word *aterrados* (frightened), pronounced as [ɐtˈʁaduʃ] in EP and [ateˈʁaduʃ] in BP, and recognized respectively as *enterrados* [ẽtɨˈʁaduʃ] 'buried' and *até ratos* [atɛ ˈʁatuʃ] 'even rats'.

This and other similar examples make us believe that a recognizer in which all the components are trained for BP may achieve better results than a similar one for EP. As a first step, in the next section, we shall adapt only the lexical component.

7. Recognition using an adapted pronunciation lexicon

In spite of being aware that adding more pronunciation variants typically introduces more substitution errors [13], and has a negative impact in terms of computational costs, we decided to investigate the use of a new pronunciation lexicon with additional pronunciation variants that would account for the BP pronunciation, but restricted to the use of the same EP-SAMPA symbols.

The transcriptions of the original lexicon (EP0) had been automatically generated and then manually corrected. Given the large number of entries, and our limited human resources, using the same procedure for BP was impossible. We therefore tried to generate BP variants using the above mentioned GtoP module only for those entries where manual correction would not be so crucial. That implied the automatic removal of entries that contained grapheme sequences characteristic of foreign words and acronyms, and entries whose orthography was likely to be different in the two variants, given the fact that the BP rules would not treat them adequately. The restricted word list for which pronunciation variants were added had $\sim 49k$ words.

The application of the BP set of rules to the 49k recognition vocabulary generated transcriptions that are very different from the corresponding EP ones generated by rule (87% different at a word level). Altogether, the new lexicon with these added variants included 108k transcriptions. We shall denote it as BP1. The third and fourth rows of table 1 show the results obtained with this lexicon. For the EP audio book, the results are slightly worse than those obtained with EP0. For the BP audio book, the results are slightly better.

Lexicon BP1 does not correspond to the same regional accent as our BP speaker. In order to take into account the most marked differences, we modified the rules affecting some graphemes (e.g. *s*, *r*, *t*, *d*), and generated transcriptions that were added to EP0. In particular, the rules for grapheme *s* became much closer to the corresponding EP rules. The new lexicon, which we shall denote as BP2, included 106k transcriptions. The fifth and sixth rows of table 1 show the results obtained with this lexicon. The results are slightly better than for BP1, reflecting the better adequacy of the new pronunciation variants. Recognition experiments using only the automatically generated BP transcriptions did not achieve good results either.

8. Conclusions and Future Work

This paper summarized the main differences between the two varieties of Portuguese and described how they affect our GtoP system. It also described several experiments with the same audio book spoken in EP and BP. Although a much larger corpus would be needed to answer our questions in a more conclusive way, our small scale experiments confirmed our expectations in terms of the effectiveness of an EP-trained aligner used on BP spoken books, and in terms of the degradation in recognition scores caused by using an EP-trained recognizer on a BP spoken book (17%, in our case). Pronunciation adaptation was tested by adding variants derived by a rule-based GtoP system with a 4.3% transcription error rate, at a word level, but yielded a very small improvement in terms of recognition scores.

Our experiments also led us to confirm the inadequacy of BN models for recognizing spoken books. Our current work on automatic TV captioning, however, will surely involve increasing the lexicon dimension and therefore decreasing the OOV rate. In this context, it will be worth exploring the possibility of separating clitics when building the lexical and language models, although our previous attempts of doing some morphological analysis have not yet brought any significant improvements [14]. We believe that the use of much larger speech and text corpora may obviously decrease the current problems, namely by using context-dependent acoustic models, but much can still be gained by studying phenomena such as vowel reduction.

This work is the first step towards porting our EP broadcast news recognizer to other varieties, and also towards the development of a variety identification module that can switch between the two varieties, thus avoiding the degradation that we currently face whenever the EP newscast includes segments by Brazilian speakers, a frequent scenario, nowadays.

Before concluding, we would like to emphasize once again the advantages of digital spoken books for spoken language research. Besides being obviously useful for data-driven prosodic modeling and unit selection in the context of text-to-speech synthesis, spoken books have been presented here as a useful diagnostic tool, not only for studying the porting of speech technologies to different varieties of the same language, but also for detecting alignment and recognition problems.

9. Acknowledgments

The authors would like to thank our colleagues Isabel Mascarenhas and Ciro Martins for their great help; Plínio Barbosa for giving us access to his unpublished manuscript, Dante Barone for sending us a phonemic pronunciation lexicon for BP, and finally Jorge Moreira and Juva Batella (the two readers of the parallel corpus). This work was partially funded by FCT projects POSI/33846/PLP/2000 and POSI/3452/PLP/2000. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

10. References

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

[2] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "Audimus.media: a broadcast news speech recognition system for the european portuguese language," in *Proc. PROPOR '2003*, Faro, Portugal, June 2003.

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

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

[5] P. Barbosa and E. Albano, "Brazilian portuguese - illustrations of the IPA," *Journal of the Int. Phonetic Association*, vol. 34, no. 2, pp. 227–232, 2004.

[6] J. M. C. Jr, *Para o Estudo da Fonêmica Portuguesa*. Rio de Janeiro: Simões, 1953, 2nd Edition.

[7] D. Callou and Y. Leite, *Iniciação à Fonética e 'Fonologia*. Rio de Janeiro: Jorge Zahar Editor, 1990.

[8] J. A. M. D. Callou and Y. Leite, "Variação e diferenciação dialectal: A pronúncia do /r/ no português do brasil," in *Gramática do Português Falado*, I. Koch, Ed. Campinas SP: Editora da Unicamp/FAPESP, 1996, vol. VI, pp. 465–493.

[9] D. Caseiro, I. Trancoso, L. Oliveira, and C. Viana, "Grapheme-to-phone using finite state transducers," in *Proc. 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, CA, USA, Sept. 2002.

[10] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," in *ASR 2000 Workshop*, Sept. 2000.

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

[12] J. Gauvain, L. Lamel, and G. Adda, "Developments in continuous speech dictation using the arpa wsj task," in *Proc. ICASSP '1995*, Detroit, USA, May 1995.

[13] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for asr: A survey of the literature," *Speech Communication*, vol. 29, pp. 225–246, 1999.

[14] C. Martins, J. Neto, and L. Almeida, "Using partial morphological analysis in language modeling estimation for large vocabulary portuguese speech recognition," in *Proc. Eurospeech '99*, Budapest, Hungary, Sept. 1999.