Spoken Language Technologies Applied to Digital Talking Books

Isabel Trancoso[(1)]*, Carlos Duarte*[(2)]*, António Serralheiro*[(3)]*,*
Diamantino Caseiro[(1)]*, Luís Carriço*[(2)]*, Céu Viana*[(4)]

[(1)] L^{2}F INESC-ID/IST, [(2)] LaSIGE/FCUL, [(3)] L^{2}F INESC-ID/Academia Militar, [(4)] CLUL
Lisbon, Portugal

Isabel.Trancoso@inesc-id.pt

Abstract

Digital Talking Books (DTBs) offer to visually impaired users an evolution of analogue talking books that mimics the interaction possibilities of print books. This paper describes a new DTB player which tries to improve the usability and accessibility of current players, through the combination of the possibilities offered by multimodal interaction and interface adaptability, and the integration of several language processing components. Besides the potential for a greater enjoyment of the reader in general, these modifications also pave the way to the use of DTBs in different domains, from e-inclusion to e-learning applications.
Index Terms: digital talking books, Portuguese.

1. Introduction

Digital Talking Books (DTBs), as described in the ANSI/NISO z39.86 standard[1], are the digital counterpart of talking books, which have been available for many years to print-disabled readers. In order to mimic the interaction possibilities of print books, the standard defines a minimum set of features for a computer-based DTB player: no need to use visual display to operate device; variable playback speed; document accessible at fine level of detail; usable table of contents; easy skips (moving sequentially through the elements); ability to move directly to a specific target; setting and labeling bookmarks; ability to add information (highlighting and notes); presentation of visual elements in alternative formats (speech); etc.

This standard is also known as DAISY 3.0, a designation that reflects the major work done in the area by the Daisy Consortium[2]. The standard focuses on the files, their structure and content, but does not include specifications for playback devices. An auxiliary document, the Playback Device Feature List, created during the standard's development, describes the main features that playback devices should possess, but it is not normative and does not present specific implementation solutions. As a result, several DTB players developed either for previous standards or more recently for the ANSI/NISO standard are capable of DTB playback, but adopted different solutions for the presentation and interaction, and some of those solutions suffer from usability flaws that are detrimental to the reading experience. Even from the accessibility point of view, several faults can be observed, preventing the use of the playback devices by some members of the intended audience.

The DTB playback environment is most notoriously a multimodal environment. The book's textual content is presented synchronized with an audio narration, either pre-recorded by a human speaker or constructed using a text-to-speech synthesizer. When supported, speech recognition can be employed as an input mode. Alerting the reader to the presence of a bookmark can be done either visually or audibly. Several other situations can be identified where the use of multimodal interaction will benefit the reader. However, most DTB players do not take advantage of the possibilities presented by the use of multimodalities, and do not go any further than the synchronized presentation of text and narration [1]. On the other hand, the anticipation of different contexts of usage, the diverse playback environments, and the multitude of user characteristics also points to the importance of interface adaptability.

Our goals in developing a new DTB player were to improve the usability and accessibility of current players, through the combination of the possibilities offered by multimodal interaction and interface adaptability, and the integration of language processing components. Besides the potential for a greater enjoyment of the reader in general, these modifications also pave the way to the use of DTBs in different domains, from e-inclusion to e-learning.

DTBs may integrate several core language technologies and at the same time provide an ideal framework for research on these technologies. This justifies the title of the second Section of this paper which summarizes the different language technologies that may be integrated in spoken books: from basic text-to-sound alignment and speech rate modification to text-to-speech synthesis, automatic speech recognition, audio indexation, and more recently alignment of parallel texts in different languages. The third Section includes a necessarily brief summary of the architecture and main features of the DTB player. The fourth Section is devoted to two main areas of application. The e-inclusion domain has been the natural target, given the importance of DTBs for visually impaired users, but we would also like to stress their importance for a very large user group of persons with dyslexia. The other domain that we have started to explore very recently is the e-learning domain, in particular for computer aided second-language learning.

Throughout this paper, several references will be made to our current repository of aligned spoken books. For demonstration purposes, we have built a repository with different types of book in European Portuguese (EP): fiction, didactic text books, poetry, children's stories, etc. Recently, and in the scope of a cooperation agreement with the University of Rio Grande do Sul, we have extended this repository with books read in Brazilian Portuguese (BP). The availability of the same book in different varieties of Portuguese has already showed that DTBs can be useful for research on the differences between varieties of the same language [2]. This study is currently being extended to the varieties spoken in African countries with Portuguese as the official language.

[1] http://www.niso.org/standards/resources/Z39-86-2002.html
[2] http://www.daisy.org/

2. Core Language Technologies

2.1. Text-to-Speech Synthesis

Text-to-speech synthesis (TTS) can be used in DTB players for audio rendering of the text and/or alerting the user to the presence of images and annotations. The TTS system developed at L2F/CLUL for European Portuguese, known as DIXI+, is a concatenative variable length unit synthesizer [3] which has been successfully integrated in several limited domain applications. The unlimited domain version, however, is not suited for integration in a DTB player, due to its very large memory requirements (its unit inventory is based on a corpus that exceeds 5k sentences), and the fact that it does not yet have a SAPI 5.0 interface. As is the case with most current TTS systems, the limited prosodic variability that can be achieved with such an inventory is the major obstacle that must be overcome in order to achieve an enjoyable rendering of a digital talking book. These were the reasons for limiting the use of the TTS system in our DTB player to alert messages which were rendered using the Microsoft SAPI synthesizer.

Digital talking books show a great potential for research on TTS systems, both from the point of view of data-driven prosodic modeling and unit selection.

2.2. Alignment of Audio and Text

The alignment between the text and the audio files in a DTB player is trivial if the audio rendering is done through a TTS system. If, as is the case with our DTB player, the audio rendering is done by playing back a pre-recorded file, then a forced alignment stage must be performed.

The Broadcast News (BN) recognizer that was used in our alignment stage 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 [4]. 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 language model (used in section 2.3) 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 (51h). The perplexity is 139.5. The vocabulary includes around 57k words. For the BN test set, the out-of-vocabulary (OOV) word rate is 1.4%. The lexicon includes multiple pronunciations, totaling 66k entries.

The use of this recognizer is a forced aligned mode implied a modification to our decoder based on weighted finite state transducers (WFSTs) [5]. 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.

With these modifications, a 2-hour long spoken book was aligned in a single step in much less than real time. The evaluation of the aligner performance was done informally, for most DTBs. The visual inspection of the word labels generally guaranteed quite good results at this level. However, the alignment of some 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 [6] and later speaker adapted acoustic models. 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. This small scale experiment revealed the advantages of using speaker-adapted models in the forced alignment stage for DTBs.

2.3. Automatic Speech Recognition

Automatic Speech Recognition (ASR) systems can be integrated in DTB players in order to offer speech as an alternative input modality. The evaluation of a previous DTB player [7] has shown us that users adopt verbal commands for some tasks, most notably playback control. In fact, 81% considered the possibility of multimodal interaction very useful or indispensable. When verbal commands are adopted, headphones should be used whenever possible, in order to avoid the interfering effects of the audio narration.

Given the limited number of command words in this task, we have chosen to use the Microsoft SAPI recognizer. In the near future, however, we plan to use our BN recognizer for integrating indexing capabilities in our DTB player, thus allowing users to vocally search for certain words in the text.

Our previous experiments with recognition of spoken books, using a small corpus of 2 fiction books, have shown the expected out-of-domain degradation. The word error rate is far greater (30.8% for the smallest book of 12-minute duration) than the one obtained for read speech, studio recordings in BN (10.9%). The causes for this degradation may be linked with the high OOV rate of fiction books (5.4%), as one OOV term can lead to between 1.6 and 2 additional errors [8], 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 the books in our corpus, with much dialog between the characters. It is also interesting to notice that 30% of the OOV forms are verbal forms with clitics. This motivated two different lines of research that are currently being investigated: the selection of the vocabulary of a core recognizer whose language models can be interpolated with available text material of restricted size for other domains, and the treatment of clitics as separate unigrams.

The relatively high deletion rate is also worth investigating. In fact, 66% of the deleted words are very short function words. This lead us to expect that a domain-adapted recognizer can in the future be used for indexation purposes.

Just as for TTS, spoken books may be also important for research in ASR, in particular for the diagnostic of the main sources of error of large vocabulary systems in clean environments.

2.4. Alignment of Parallel Texts

The alignment of parallel text in two languages 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 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 [9]. The second step, word-to-word alignment is done using IBM-4 statistical alignments [10] as implemented in the GIZA++ tool [11]. Alignments in both directions (source-to-target and target-to-source) are performed and combined using heuristic refinement rules [12]. Because the texts are usually too small to train reliable word alignments, the texts are aligned in the context of a larger, out of topic, corpus such as the *europarl* corpus [13].

3. Main Features of the DTB Player

The developed DTB player (figure 1), besides supporting the features described in the DTB standard and accompanying documents, introduces features complementing the synchronized presentation of text and audio. These include: book's enriching with 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.

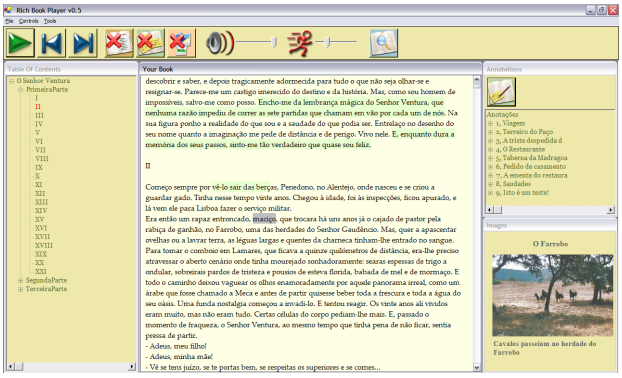

Figure 1: The DTB player interface, presenting main content, table of contents, annotations and an image.

The DTB player was developed using FAME [14], a model based framework for adaptive multimodal environments. FAME provides a sound basis for an application with the requirements of an enriched DTB player, by assisting in identifying application and user behaviors, and guiding the development of an adaptable application based on those behaviors.

FAME's general architecture for adaptive multimodal applications comprises a set of input categories, their multimodal fusion, a model-based adaptation module, the adaptive presentation layout and multimodal output fission, and the final presentation generation. The adaptation module also governs the multimodal fusion and fission procedures.

The developed DTB player allows the use of traditional input methods (mouse and keyboard) as well as the use of ASR. According to the framework, besides user inputs, application state and application generated events are also considered as inputs for the adaptation module. Thus, voice command interpretation is context dependent. For instance, a recognized *show* command can be interpreted to take different meanings according to the current context, sometimes presenting annotations, while other times images.

The adaptive output presentation employs visual elements, like text and images, and audio elements, which include voice, recorded and synthesized, and other sounds and music used to enrich the book's content. The synchronization of all these elements is fundamental for the application's success, thus being one of the most important components of the DTB player. A previous processing of the book's content generates an XML file with the content hierarchically organized from the lowest unit (word) to highest one (book). Intermediate units include sentence, paragraph, sections and chapters. Each lower unit position in the audio file is described by starting time and duration attributes. Similar attributes exist for annotations, images and sounds, informing the DTB player of their presentation time. Time based synchronization structures are then built on the fly, for each presentable component: main content, annotations, images, and other sounds and music. During the book playback, the position in the audio timeline determines the highlighted units, the presented annotations, and the exhibition of accompanying media. As the user is allowed to control the narration speed, the synchronization unit employed is adapted in run-time, since at higher reading speeds, accompanying highlighted words becomes perceptually difficult. To minimize this effect, higher reading speeds will change word synchronization, to sentence synchronization, to paragraph synchronization.

The presentation of translated material poses an additional challenge, involving another layer of synchronization between the two contents. Given languages' characteristics, it is not always possible to establish a word to word relationship between translated words, being the many words to many words relationship the only possible solution. This implies the creation of a flexible unit synchronization structure between the content in the different languages. This structure is then connected to the text-narration synchronization structures for both audio recordings, allowing for interchanging the narration controlling language.

The architecture of the DTB player also includes a TTS module, currently used with two objectives, and governed by the multimodal fission component to accomplish output commands using two modalities. The first objective is alerting the user to the presence of annotations or accompanying media. This is accomplished using both visual and audio cues. Flashing icons are accompanied by a synthesized voice alerting the user to the presence of an annotation or an image. The second goal is to offer the possibility of orally presenting written annotations. The annotation text is input to a TTS module for audio presentation. This feature is of the utmost importance for visually impaired users.

Adaptation also impacts over other aspects of the interaction with the DTB player. Visual elements can be hidden or presented (either by the user or the player's initiative) and that might cause reading disruption if it involves a new placement of the main content component. To counter this negative effect, the adaptation module rearranges the visual layout in order to minimize alterations (either in position or dimension) to this component.

The player's behavior can also be adapted according to past user behavior. For instance, if the user repeatedly chooses to hear annotations with the TTS component, the player will start presenting the annotations that way by default, and will stop the main content narration during the annotation playback, to prevent over-

lapping sounds.

4. Applications

4.1. DTBs for e-Inclusion

Visually impaired users are the main target audience for DTBs. Even though the developed DTB player has features specific for non visually impaired users, it also solves some of the usability and accessibility flaws identified in other DTB players. One of the major problems identified was the absence of alternative input modes to the keyboard. Although the keyboard can still be used for controlling the most basic operations, the more advanced ones demand a different kind of support that, for visually impaired users, has to be provided through speech recognition technologies. The developed DTB player supports these advanced operations by providing speech commands for all operations. The possibility to operate in a non-visual environment is complemented by the introduction of TTS technologies, allowing for audio presentation of annotations, as well as previously prepared image descriptions, and awareness increasing mechanisms, like cues for the presence of images and annotations. With these mechanisms the DTB player can be operated by visually impaired users, taking advantage of advanced features previously unavailable to them.

Although no tests have been conducted with users with dyslexia, we think that the adaptation capabilities will be also very important in suiting the DTB player to this community.

4.2. DTBs for e-Learning

Exploring the potential of DTBs for e-Learning was a recent trend in our group motivated by complaints from foreign students about the scarcity of available material for learning Portuguese. The possibility of selecting a word, group of words or a sentence and hearing how it sounds [15] may be important for a foreign student to learn about the segmental and prosodic characteristics of the new language. The fact that professional speakers are often used in the recordings is very advantageous from this point of view, although they complain that the use of very long sentences, so common in Portuguese fiction, makes it difficult to adapt a normal prosody.

Our future plans include conducting a formal evaluation of DTBs for e-Learning, using the versions with a single book and with parallel books whenever available. In this latter context, manually and automatically aligned parallel texts should be compared in order to evaluate the effect of alignment errors. One should also evaluate how important is to have a word based alignment versus a virtually error-free sentence alignment.

5. Conclusions

This paper described DTBs as an application area that may integrate several language processing components, from basic audio/text alignment and speech rate modification to TTS, ASR, audio indexation, and more recently alignment of parallel texts in different languages, stressing the major advantages and disadvantages of each component in this context.

The new DTB player tries to improve the usability and accessibility of current players, through the integration of these components in a multimodal adaptable interface. These features pave the way to the use of DTBs in different domains, from e-inclusion to e-learning, where we are currently designing usability tests.

6. Acknowledgments

The authors would like to thank our colleague Luís Oliveira. 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".

7. References

[1] Duarte, C. and Carriço, L., "Users and Usage Driven Adaptation of Digital Talking Books", Proc. 11th International Conference on Human-Computer Interaction (HCII 2005), Las Vegas, Nevada, USA, July 2005

[2] Trancoso, I., Serralheiro, A., Viana, C. and Caseiro, D., "Aligning and Recognizing Spoken Books in Different Varieties of Portuguese", Proc. Interspeech 2005, Lisbon, Portugal, Sept. 2005.

[3] Paulo, S. and Oliveira, L., "Reducing the Corpus-based TTS Signal Degradation Due to Speaker's Word Pronunciations", Proc. Interspeech 2005, Lisbon, Portugal, Sept. 2005.

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

[5] Mohri, M. , Pereira, F. and Riley, M., "Weighted finite-state transducers in speech recognition", Proc. ASR 2000 Workshop, Paris, France, Sept. 2000.

[6] Trancoso, I., Caseiro, D., Viana, C., Silva, F. and Mascarenhas, I. , "Pronunciation modeling using finite state transducers", Proc. 15th Int. Congress of Phonetic Sciences, Barcelona, Spain, Aug. 2003.

[7] Duarte, C. and Carriço, L., "Usability Evaluation of Digital Talking Books", Proc. Interacção 2004 - 1st National Conference in Human-Machine Interaction, Lisbon, July 2004.

[8] Gauvain, G., Lamel, L. and Adda, G., "Developments in continuous speech dictation using the ARPA WSJ Task", Proc. ICASSP 1995, Detroit, USA, May 1995.

[9] Gale W., and Church K., "A program for aligning sentences in biligual corpora", Computational Linguistics, 19:75-102, 1993.

[10] Brown P., Della Pietra V., Della Pietra S., Mercer R., "The Mathematics of Machine Translation: Parameter Estimation", Computational Linguistics, 19:263-311, 1993.

[11] Och F., Ney H., "Improved Statistical Aligment Models", Proc. 38th Annual Meeting of the ACL, 2000.

[12] Koehn P., "Noun Phrase Translation", PhD. Thesis, Univ. Southern California, 2003.

[13] Koehn P., "Europarl: A Parallel Corpus for Statistical Machine Translation", Proc. Machine Translation Summit, 2005.

[14] Duarte, C. and Carriço, L., "A Conceptual Framework for Developing Adaptive Multimodal Applications", Proc. IUI, Sydney, Australia, Jan. 2006.

[15] Li, S., Lin, H. and Chen, H., "How speech/text alignment benefits web-based learning", Proc. 13th ACM Int. Conf. on Multimedia, Singapore, Nov. 2005.