Towards a Repository of Digital Talking Books

António Serralheiro[1], *Isabel Trancoso*[2], *Diamantino Caseiro*[2], *Teresa Chambel*[3], *Luís Carriço*[3], *Nuno Guimarães*[3]

[1] L^{2}F INESC-ID/Academia Militar, [2] L^{2}F INESC-ID/IST, [3] LASIGE/FC

Lisboa, PORTUGAL

antonio.serralheiro@inesc-id.pt

Abstract

Considerable effort has been devoted at L^2F to increase and broaden our speech and text data resources. Digital Talking Books (*DTB*), comprising both speech and text data are, as such, an invaluable asset as multimedia resources. Furthermore, those DTB have been under a speech-to-text alignment procedure, either word or phone-based, to increase their potential in research activities. This paper thus describes the motivation and the method that we used to accomplish this goal for aligning DTBs. This alignment allows specific access interfaces for persons with special needs, and also tools for easily detecting and indexing units (words, sentences, topics) in the spoken books. The alignment tool was implemented in a Weighted Finite State Transducer framework, which provides an efficient way to combine different types of knowledge sources, such as alternative pronunciation rules. With this tool, a 2-hour long spoken book was aligned in a single step in much less than real time. Last but not least, new browsing interfaces, allowing improved access and data retrieval to and from the DTBs, are described in this paper.

1. Introduction

The framework of this paper is a national project known as IPSOM, whose main goal is to improve the access to digitally stored spoken books by the visually impaired community. Spoken books have been mainly provided by the National Library (BN, *B*iblioteca Nacional) in analogue format (cassette) and have lately been under a gradual conversion process to digital format (CDROM).

To improve the usability of these spoken books, the IPSOM project aims to provide both specific access interfaces for persons with special needs, and also tools for easily detecting and indexing *u*nits (words, sentences, topics) either written or spoken. Hence, a good word-by-word synchronization between the text and its audio recording is mandatory for unit access and thus spoken book alignment at a word level is a major task of the IPSOM project. From the point of view of research in the area of speech processing, one of the most interesting aspects of the IPSOM project is the fact that indexed spoken books provide an invaluable resource for data-driven prosodic modeling and unit selection in the context of text-to-speech synthesis. This is a good motivation to perform the alignment not only on the basis of words but also of sub-word units.

Another major task is the generation of meta-information (e.g. book context, relevant ideas, ...) that associated with the audio data constitutes what is usually designated as Digital Talking Books (*DTB*) [1] [2]. By providing multimedia interfaces for access and retrieval, one can also broaden the usage of multimedia spoken books (for instance in didactic applications, etc.).

DTBs may provide the "talking" capability by means of a text-to-speech synthesizer, allowing a direct access to each text word within the book. Our automatic aligner easily provides this same word synchronized access for books read by a human voice, with all the naturalness and emotions that current synthesizers are still unable to convey, which causes them to invariably induce some fatigue to the listeners.

This paper has two main parts: the first one is devoted to speech processing issues (Section 3) and the second one deals with the *DTB* production and user interface (Section 4). Before, however, we shall describe the pilot corpus that was used in this work (Section 2).

2. Pilot Corpus

Existing spoken books at BN have been recorded by volunteers (non-professional readers) and stored in analogue tapes, that by their sequential access mode, results in an extremely slow (and error-prone) information retrieval process. This handicap could be easily overcome through their conversion to CDROM, if other problems had not been found, namely: low audio quality (multiple copies and damaged masters), and audible differences of quality through the same book (manual spectral equalization, and uncalibrated multiple recording sessions). These problems, together with the non-systematic reading of tables, figures, chapter numbers, footnotes, preface, etc., made the current material not suitable for automatic text-to-speech alignment. Consequently, it was decided to record new spoken books to serve as a *pilot corpus* for the new recording and alignment procedure.

The first book in this pilot corpus was "O Senhor Ventura", a novel by Miguel Torga. The book was read by a professional speaker in a sound-proof booth. It was recorded directly to DAT and later down-sampled to 16kHz. The digital audio file was then manually edited to remove some reading errors and extraneous noises (although breathing sounds were kept to enhance naturalness), resulting in 2h 15m of audio. The text, amounting to 137,944 words, was pre-processed to deal with abbreviations, numbers and special symbols, resulting in a lexicon with around 5k different forms. Although very intelligible, as expected from a professional speaker, the speaking rate was relatively high, averaging more than 174 words per minute.

This pilot corpus was further extended to include other literary styles such as social critic ("Manifesto Anti-Dantas" by José de Almada Negreiros) and poetry ("Um Momento de

Palavras" by David Mourão Ferreira). The former was read by a professional actor. The latter was read by its author, a well known writer, who had a program on the main TV channel on Portuguese literature. Contrarily to the first novel, these two works were already existing recordings, available on CD. These new texts amount to more than 1900 and 4000 words and their recordings to 11m and 59m, respectively. The "Manifesto Anti-Dantas" by its own nature was read with a highly stressed voice (shouted several times) and with a constantly varying speech rate (averaging 170 words per minute). On the contrary, the poems were read in a quiet way, with slow speaking rate (67 words per minute), and much intonation variability, resulting in highly intelligible speech.

3. Alignment System

The alignment system comprises a feature extraction front-end followed by a forced aligner as described in the next subsections.

3.1. Acoustic Modeling

The hybrid acoustic models used in the alignment of these spoken books were originally developed for broadcast news [3], in an effort to combine the temporal modeling capabilities of *HMMs* (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. Each *MLP* classifier incorporates local acoustic context via a multi-frame input window of 7 frames. The resulting network has a single hidden layer with 500 units and 39 output units (38 phones for European Portuguese plus silence).

3.2. Alignment

An aligner is basically a speech recognizer that keeps track of the time boundaries between words or phones. Our recognizer is based on *WFSTs* (Weighted Finite State Transducer) [4] in the sense that its search space is defined by a distribution-to-word transducer that is built outside the decoder. *WFSTs* have been successfully used in many written and spoken language applications, providing an efficient and elegant way of combining different types of knowledge sources, which makes them good candidates for alignment purposes.

The search space is 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 the sequence of words that constitute the orthographic transcription of the utterance. The main advantage is that no restrictions are placed on the construction of the search space, which means that it can easily integrate other sources of knowledge, and the network can be optimized and replaced by an optimal equivalent one. This last advantage is a disadvantage from the perspective of alignment, as there are no warranties that the output and input labels are synchronized. To solve this problem, 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*.

Our aligner can be used in two different modes: with and without phonological rules [5]. These rules may account for alternative pronunciations in a similar way of a lexicon with multiple pronunciations per word.

The way we dealt with pronunciation variation has some similarities with the one described in [6]. The variations that depend on word-level features of lexical items (such as part-of-speech) and those that are particular to specific lexical entries (such as many acronyms in Portuguese, for instance) are just included in the lexicon. The remaining variants that depend on the local immediate segmental context are modeled through rules. Some of these rules concern variations that depend on the stress and syllable position. The lexicon uses different labels for representing segments in particular positions.

The main phonological aspects that the rules are intended to cover are: vowel devoicing, deletion and coalescence, voicing assimilation, and simplification of consonantal clusters, both within words and across word boundaries. Some common contractions are also accounted for, with both partial or full syllable truncation and vowel coalescence. Vowel reduction, including quality change, devoicing and deletion, is specially important for European Portuguese, being one of the features that distinguishes it from Brazilian Portuguese and that makes it more difficult to learn for a foreign speaker. As a result of vowel deletion, rather complex consonant clusters can be formed across word boundaries.

The rules are specified using a finite-state grammar whose syntax is similar to the Backus-Naur-form augmented with regular expressions. Each rule is represented by a regular expression, and to the usual set of operators we added the operator $\rightarrow$, simple transduction, such that $(a \rightarrow b)$ means that the terminal symbol a is transformed into the terminal symbol b. The language allows the definition of non-terminal symbols (e.g. *$vowel*). All rules are optional, and are compiled into *WFSTs*.

3.3. Alignment results

Our pilot corpus allows us to do alignment tests at a word level, but not at a phone level, as required, for instance, for text-to-speech research. In order to evaluate the quality of the phone level transcriptions, we have used a fragment of the EUROM.1 corpus [7], for which we have manual phone level alignment [8].

A major advantage of the *WFST* approach is that it allowed us to align the full audio version of the book in a single step. This is specially important if we take into account that the memory limitations of our previous alignment tool imposed a maximum of 3-minute audio segments. We thus avoid the very tedious task of manually breaking-up the audio into smaller segments with their associated text.

The word segmentation of the largest book ran in 0.024 real-time (RT), requiring 200MB of RAM, excluding acoustic modeling. The phone level alignment of the book ran at 0.027 xRT when using the canonical pronunciations of the lexicon, and 0.030 xRT when using also the pronunciation rules.

An informal evaluation of the alignment procedure at word level was done using the publicly available Transcriber tool[1], which allowed us to subjectively access the good quality, by simultaneously listening and seeing on a word-by-word basis. Figure 1 illustrates the use of this tool.

The same level of performance was obtained for the alignment of the other two books. However, in the poetry case, the final post-tonic syllable of several words was incorrectly

[1] http://www.etca.fr/CTA/gip/Projets/Transcriber/

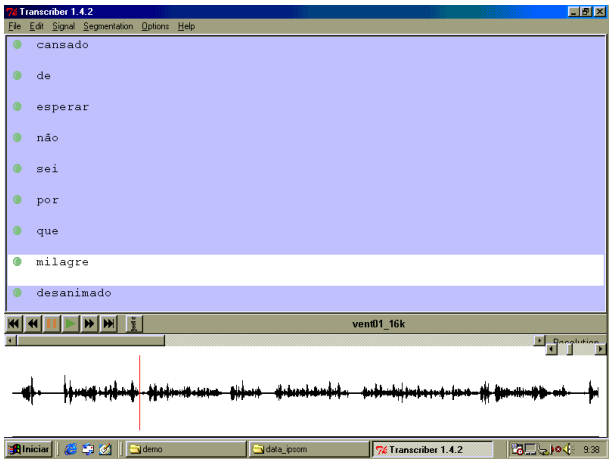

Figure 1: Illustration of the word-level alignment of spoken books.

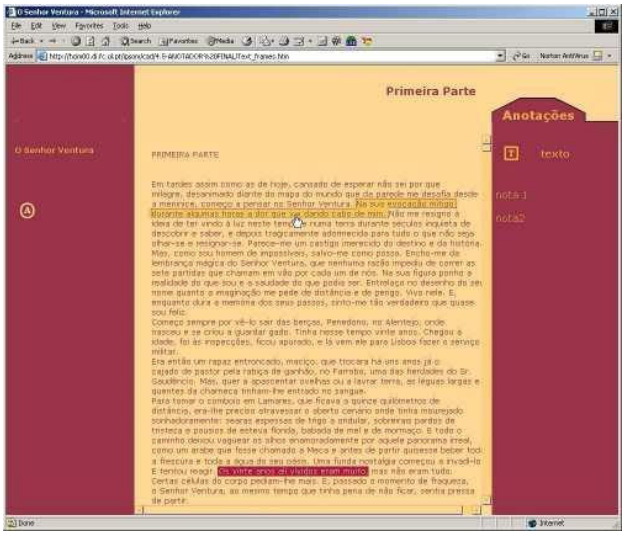

Figure 2: Prototype of a DTB.

aligned. We suspect that this was due to the rather long duration of the stressed syllable which in the observed cases was separated from the post-tonic by a considerably long silence. This is one of the aspects that we are currently investigating.

4. DTB Production System and User Interface

The DTB Production System takes the results from the alignment system and produces DTBs [9]. The production consists of three phases: (1) the standardization of contents (2) the enrichment and storage of contents and (3) the generation of the user interface (UI). The first phase results in a standard DTB specification, composed by a set of XML and audio files, that include the book content (text and audio) and structural information and synchronization at the basic level, as provided by the alignment system.

The enrichment and storage phase draws from the content of the book and builds a set of meta-information (e.g. book context, relevant ideas) that is included in the DTB structure. At the UI generation phase, it is then possible to introduce in the resulting DTB application new media components, gathered from a media repository and related to the original book (e.g. sea sounds and images, if the book is set in a maritime environment). That meta-information, along with the corresponding content, can also be stored in the media repository, as classified multimedia ”sentences”, that can be reused in the production of other books or documents.

In the last phase, the production system receives the enriched DTB specification and combines it with a set of modules that determine the UI and the implementation language of the complete DTB. Synchronization units determine the granularity levels of multi-modal information (text plus audio) that are available for interaction. Currently, word, sentence, paragraph and silence-based are supported. The ”silence-based” synchronization is extracted from the audio file, and is determined by the pauses that the reader makes in the production of the audio stream. The sentence-based and paragraph units are syntactic and automatically extracted from the text.

The presentation and interaction models determine how the synchronization is presented to the user (e.g. how it is visually marked), how the user interacts with the book (e.g. by speech, mouse/keyboard), including the targeted device (PC, Tablet, PDA), and how the application integrates annotations and other media. They provide patterns that the UI generation phase expands into the DTBs. Specifications are available for two target languages, namely HTIMEL [10] and Microsoft’s version of SMIL 2.0 [11], HTML+TIME.

4.1. Preliminary evaluation

A DTB prototype was generated from the first book of the pilot corpus. The whole production framework (i.e alignment and DTB production systems) was fed with a file from scanned text and an audio file, resulting in a DTB specification. Several variants were also obtained, on the last phase of the production system by changing the alignment units (other than words), the presentation and the interaction models.

Figure 2 depicts an example of DTB, played by a common Web Browser. The text presented is visually marked (change of color/tone) in synchrony with the audio. Annotation capabilities (editing) are also provided, appearing in the right column and also synchronized with the audio. The synchronization unit for navigation (denoted by the marked text at the top) is different from the playback unit. When the user clicks on a word, the audio stream jumps to the word at the beginning of the sentence, whereas the emphasized text during playback corresponds to the text spoken between two pauses. Search synchronization is also considered, but not shown.

Preliminary field trials have shown some discomfort in users if word-based alignment was used in the playback of DTBs. In fact, the users reported the need for more context during ”free search” (e.g. longer units). This was specially noticed if the visual counterpart of the DTB was not shown or seen (e.g. visually impaired users). As a result, the current version of the production system allows a greater design flexibility.

5. Conclusions and future work

The paper described our work on spoken books in the framework of the IPSOM project, emphasizing the problems of the actual repository and the alignment tools that were developed. We verified that, with proper recording procedures, the alignment task can be fully automated in a very fast single-step procedure, even for a 2-hour long recording. This is specially important if we take into account that the memory limitations of our previous alignment tool imposed a maximum of 3-minute audio segments. With this new tool, we avoid the very tedious process of partitioning audio and text into corresponding segments.

The use of phonological rules seems to provide reasonably good alternative pronunciations, specially accounting for vowel reduction and inter-word co-articulation phenomena. The better phone level alignment of spoken books achieved with these rules will also be crucial for our research in text-to-speech synthesis, namely for prosodic modeling and unit selection, using data-driven approaches.

The word boundaries computed using the *WFST*-based alignment allowed for the development of more sophisticated browsing tools for spoken books which can be specially important for non-fiction, technical books, for which there is a great request from the visually impaired community.

6. Acknowledgments

This work was partially funded by FCT projects POSI/ 3452/ PLP/2000 and POSI/33846/PLP/2000. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

The authors would like to thank Isabel Bahia for her kind cooperation in reading the book and our colleagues from CLUL, Céu Viana and Isabel Mascarenhas, for their help with the phonological rules and manual labeling.

7. References

[1] ANSI/NISO Z39.86 - 2002 Specifications for the Digital Talking Book, *http://www.niso.org/standards/index.html*

[2] DAISY 2.02 Spec., Formal Recommendation, February 28, 2001. *http://www.daisy.org/products/menupps.htm*

[3] H. Meinedo and N. Souto and J. Neto, "Speech Recognition of Broadcast News for the European Portuguese Language",in Proc. ASRU2001,IEEE Automatic Speech Recognition and Understanding Workshop, Madonna di Campiglio, Italy, December 2001.

[4] M. Mohri, M. Riley, D. Hindle, A. Ljolje, F. Pereira, "Full Expansion of Context-Dependent Networks in Large Vocabulary Speech Recognition", in Proc. ICASSP 98, Seattle, Washington, 1998.

[5] I. Trancoso, D. Caseiro, C. Viana, F. Silva, I. Mascarenhas, "Pronunciation modeling using finite state transducers", in Proc. ICPhS'2003, Barcelona, Spain, August 2003.

[6] T. Hazen, L. Hetherington, H. Shu, and K. Livescu, "Pronunciation Modeling Using a Finite-State Transducer Representation", in Proc. PMLA 2002 Workshop, Aspen Lodge, Estes Park, Colorado USA, September 2002.

[7] C. Ribeiro, I. Trancoso and M. Viana, *EUROM.1 Portuguese Database*, Report of ESPRIT Project 6819 SAM-A, 1993.

[8] A. Serralheiro, D. Caseiro, H. Meinedo, I. Trancoso, "Word Alignment in Digital Talking Books Using WFSTs", in Proc. ECDL'2002 - 6th European Conference on Digital Libraries Roma, Italy, September 2002, Ed. Springer-Verlag.

[9] L. Carriço, N. Guimarães, C. Duarte, T. Chambel, H. Simões, "Spoken Books: Multimodal Interaction and Information Repurposing", in Proc. 10th International Human-Computer Interaction, Crete, Greece, June, 2003.

[10] T. Chambel, N. Correia, N. Guimarães, "Hypervideo on the Web: Models and Techniques for Video Integration", International Journal of Computers & Applications, 23 (2), 2001.

[11] W3C (2001). "Synchronized Multimedia Integration Language (SMIL 2.0)", *http://www.w3.org/TR/smil20/*