MODULAR PRODUCTION OF RICH DIGITAL TALKING BOOKS

Luís Carriço, Carlos Duarte, Nuno Guimarães
Dpt. of Informatic, Faculty of Sciences, University of Lisbon
Campo Grande, Edificio C5, 1749-016 Lisboa, Portugal
{*lmc,cad,nmg*}*@di.fc.ul.pt*

António Serralheiro, Isabel Trancoso
L2F INESC-ID
R Alves Redol 9
{*antonio.serralheiro,isabel.trancoso*}*@inesc-id.pt*

Key words: User Interface Generation, Model-based Tools, Accessibility, User Diversity

Abstract: This paper presents a framework for automatic production of rich Digital Talking Books (DTB). The production process converts existing audio tapes and OCR-based digitalization of text books into full-fledged, multiple synchronized, multimodal digital books. The framework deals with the content organization processes and User Interface definition. The first one identifies content units and its relations. The latter, based on abstract, yet DTB specific, pattern-based UI specifications, allows the definition of various forms of interaction and presentation, required by the diversity and constraints of targets users (e.g. visually impaired persons) and situations of use (e.g. learning). The framework also permits to balance the complexity and flexibility of the generated DTBs, in order to cope with the resources provided on the different execution platform.

1 INTRODUCTION

Audiotapes have served as an important medium, and sometimes the only alternative, for print-disabled reader's access to books. In several public libraries, as in the Portuguese National Library, a long time effort was made in speech recording of a large amount of printed material. However, the limitations of this analogue approach, even when compared with their printed counterparts are noteworthy.

Digital Talking Books (DTBs) are a logical answer. Work around these identifies requirements and has recently issued a standard specification (ANSI/NISO, 2002). Nevertheless, it does not propose specific solutions for interaction. In fact, the combination of synchronization, structural navigation and annotations management, using visual, audio, speech and standard interactions, poses ambiguity and cognitive problems that must be dealt with at the UI design level (Carriço et al., 2003a; Morley, 1998). These issues are stressed by the diversity of targeted users, their particular disabilities and perspectives. Exploring and evaluating distinct UIs for the same book, with different multimodal combinations, eventually enriched with new media contents, is therefore mandatory.

This paper describes DiTaBBu (Digital Talking Books Builder), a framework for the production of rich DTBs based on media indexing, speech alignment and multimodal interaction elements. The work was done in the context of the IPSOM project, joining the Portuguese National Library, a speech processing research group and multimedia interaction designers and engineers. The framework draws its requirements from: (1) the existence of large amounts of recorded material; (2) the DTB standard; (3) the flexibility needed for generation of exploratory and adjustable UIs; and (4) the ability to integrate new multimedia units in the production process.

In the following section, the article presents the requirements imposed by users, usage scenarios and particular project goals, referring related work. Section 3 points design decisions that had an impact on the generated DTBs' architecture and the execution platform. Next, the production framework is described focussing on its modularity. Finally, conclusions are made and future work is drawn.

2 REQUIREMENTS

The DiTaBBu framework results from requirements emerged from the diversity of DTB target users, the DTBs' usage possibilities, in terms of situational context, richness and support technology, its repurposing and reuse, and from the characteristics of the available source material.

2.1 Main Target Users

DTBs aim to provide easier access to books, for print-disabled communities. Work done with those communities, resulted in several guidelines (Daisy Consortium, 2002). The following resumes a list of recommended navigation features (NISO, 1999a): (1) support basic navigation (advancing one character, word, line, sentence, paragraph or page at a time, and jumping to specific segments); (2) fast forward and reverse, and reading at variable speeds; (3) navigation through table of contents or control file (to obtain an overview of the book material); (4) reading notes, cross-referencing, index navigation, bookmarks, highlighting, taking excerpts, searching....

DTBs recommendations also point different combinations of media, with emphasis on the audio component. Here, one should consider the limitations of audio. Alone, its one-dimensional nature can present only few items at a time. Combining visual and spoken presentation requires accurate synchronization or specific visual marking (Duarte et al., 2003). Methods for conveying structure and assist navigation, in a non-visual environment, have been researched: 3D audio (Goose and Moller, 1999), auditory icons (Gaver, 1993), multiple speakers and sound effects (James, 1997), among other techniques. For DTBs with multiple media presentations, the use of contextual information (such as containing sentence, paragraph or section), when navigation or continuous presentation occurs, was evaluated as well (Carriço et al., 2003a) - results point to the need for different contextual units (e.g. the further the navigation "jump" the bigger the required context). Most of the studies, however, are not yet conclusive. Exploring these and other techniques and comprehending the actual need of visually-impaired people must be still a subject of evaluation.

2.2 Usage and Playback Devices

The NISO Committee characterizes three types of usage and playback devices for DTBs (NISO, 1999b): (1) basic - portable with simple playing digital audio capabilities (no access to full-text and aims primarily to play continuous audio); (2) advanced - also portable but should allow to access documents randomly, with navigation possibilities, bookmark setting, etc.; and (3) computer-based - complete and sophisticated features. The Daisy Consortium expands these further, in terms of media combination (Daisy Consortium, 2002): (1) full audio with title element only; (2) plus navigation control; (3) plus partial text; (4) full audio and text; (5) full text and some audio; and (6) text and no audio. All these should be supported under the DTB umbrella, which means that same book "edition", and to some extent the same book (structure and content), could be presented and interacted in different ways, using more or less resources, and different media and mode combinations.

Consequently, the DTB production mechanism or execution platform should build on an architecture that promotes a clear separation between the books' content and user interface (UI). This will reinforce the coherence among several usage settings of the same book. Modularity is further emphasized when considering a distributed usage (e.g. access to digital libraries through remote "reading places").

The standard (ANSI/NISO, 2002) identifies a set of DTB modules (Content, Navigation, Media, Synchronization, Resources, ...), for which XML DTDs are defined. Presentation is handled with style sheets (CSS or XSL) and SMIL 2.0 (synchronization). This architecture enables different presentation designs, and the choice of web-based technology ensures the required wide dissemination. However, content and presentation are still intermixed at the same level. For example, for book's content, the media correspondence and the presentation sequence are both defined in the Synchronization file. This one-level modularity, although coping with several configurations for the same book (a DTB for each configuration), hardly embraces the intrinsic correspondence among them. It can (as a standard) be used as a final format for DTBs, but a clearer separation of content and UI is required, on DTB production frameworks and DTB architectures that provide an enhanced flexibility or adaptability (Duarte and Carriço, 2004).

The use of the DTB standard format has recently gained momentum (Dolphin Audio Publishing, 2003; Innovative Rehabilitation Technology inc., 2003; VisuAide, 2003). Nevertheless, other web-standard solutions should be envisaged, if a wider dissemination and ease of evolution is pursued. Formats fully compatible with common Web browsers, executing in general mobile devices, should definitely be available.

2.3 Reediting and Repurposing

Beyond these "problem-oriented" proposals, a rich framework for multimodal interaction opens the way for information repurposing and creative combination of elementary media (Carriço et al., 2003a). This broader view of DTBs further affects the modularity of DTB production and architecture. Along this rich DTBs construction, multimedia units must be identified and classified, and later reused in the authoring of new books or general documents. The stress, apart from the separation of UI and content, is now on the modularity of the content itself. Meta-information and classifiers must be introduced, either explicitly or (preferably) using content analysis techniques. A production framework that facilitates this authoring process is also further emphasised.

2.4 Source Material

The Portuguese National Library provides services for visually-impaired persons. It has a large amount of analogue spoken books (audio tapes) and it is also committed to build digital versions of the books - scanned, within a XML/HTML envelope. A need for its integration and the introduction of DTB functionalities, was clearly felt, particularly by the visually-impaired community.

Coherence and the huge amount of existing material require an "as automated as possible" form to produce the DTBs. Such framework should handle: (1) the expedite identification of speech excerpts that correspond to the textual units (alignment); and (2) an easy specification of different UIs and UI patterns for a book content. The first problem is generally handled by speech recognition technology. The latter is related with transformation tools and UI generation. Here, model-based approaches were adopted to handle generation of UIs for different users and devices (Paternò, 2000), the creation of UIs for multiple devices (Eisenstein et al., 2001; Ali and Pérez-Quiñones, 2002; Lin and Landay, 2002) or its adaptation to different devices (Calvary et al., 2001). This is a field where the transition to the commercial software world has not occurred, in part because of the abstraction level used in the specification. However, in the case of DTB production, with the particularities of the domain, there is not such a great emphasis on abstraction, and the generation process can be more easily adopted.

3 DESIGN OPTIONS

In view of the stated requirements, the work conducted within the IPSOM project has evolved through a series of design options presented next.

3.1 Navigation Features

The navigation functionalities are fully considered in the built DTBs, except for the variable speed-reading and the thinner (character and word-based) basic navigation support. The first requires a complete speech model in order to maintain low voice distortion and was not considered. An alternative, currently under evaluation, is the reduction or extension of sentence separation (silence, or breathing times), combined with small speed changes. The second feature strongly depends on the ability to isolate character and word sounds from the continuous speech recording. However, tests made with word-based navigation generated incomplete sounds, that users felt displeasing in evaluation studies (Duarte et al., 2003). An alternative may be the introduction of speech synthesis.

Pertaining to DTB categories all variants are supported, from full audio and text, to plain audio or plain text. Here, the DiTaBBu production framework facilitates the several versions' maintenance and provides the means for exploratory modality combinations.

3.2 Architecture and Platform

Based on DTB recommendations, Web-based technology was adopted. Several DTB formats and arrangements are possible. The architecture includes:

- An XML-based content specification without references to UI issues. Content includes text and other media (in specific formats), media anchoring points and correspondence (to text or between media), and structure. Navigation facilities, bookmarking and annotations, margin notes and other secondary content follow a similar approach, whenever possible, close to the DTB standards. No presentation or synchronization are considered at this level.
- One or more XSLT-based translation specifications enabling the creation of UIs for the content.
- The UI, including specific interaction objects when required. Presentation could use several formats and arrangements, from plain SMIL (plus CSS), to versions fully compliant with the DTB standard.

The introduction of the XSLT level permits to build different UIs, using alternative DTB formats, and still maintaining the coherence towards books' content. It also allows to balance the generation of the UI, between the production framework and the execution platform (fig. 1). The framework generates XML and XSLT documents (plus target language templates, CSS, ...) representing the DTB and the UI building rules. If the execution platform is able to process XSLT, a book following the above three-layer organization can be used. Alternatively, if performance is an issue or the platform does not support it, DiTaBBu could generate the final (one or two-layer) DTB configurations (e.g. a DTB fully compliant with the standard or a SMIL version, in any DTB category).

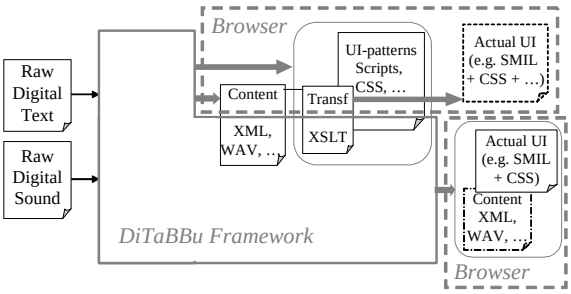

Figure 1: Balancing performance and flexibility.

Internet Explorer 6 was adopted as a base, yet powerful, execution platform. This choice enabled the use of HTML+TIME (and CSS), a representative of a SMIL 2.0 profile, and Microsoft's variant of VoiceML, for voice interaction. Both architectural arrangements are supported, since the browser processes XSLT. In the simplest form, the digital book is a (set of) HTML+TIME, CSS and media files. For voice interaction with DTBs, off-the-shelf products, recognizing Portuguese language, were initially used with very bad performance results. The Microsoft's implementation of VoiceML, provided better results, but using English as interaction language. Currently, Portuguese speech recognition software, developed within the IPSOM project's teams, is being integrated.

3.3 Automation and Initial Corpus

The initial pilot corpus was the "O Senhor Ventura" (a novel by Miguel Torga). Since the audio quality of the existing tape was very poor, both in terms of noise and diction, a clean audio stream was recorded on a soundproof booth, using a professional reader. Other books were also used (Serralheiro et al., 2003). For the moment, however, the automation process requires fairly good audio recording, making it difficult to use the original audio tapes. Refinement of the speech alignment component is currently under work, in order to circumvent audio tapes quality.

4 THE FRAMEWORK

The DiTaBBu framework generates DTBs through an automatic production process, configured by a set of specification files that allow the required flexibility. Internally the framework is decomposed into two main phases: content organization and UI-generation.

4.1 Content Organization Phase

This phase (fig. 2) is responsible for the integration of media files and auxiliary content (margin notes and indexes) with the main book content. The phase includes a set of modules for specific media processing, a structuring module and a set of linker modules. All modules generate files according to particular XML dialects. The phase result is: (1) a content description file, compliant with the DTB standard and containing the book text and other information extracted from the media files (e.g. "Part I" sounds like "part one"); (2) a set of media and linking files that establish the correspondence between content and media; (3) a set of auxiliary components (not in the figure) also linked with the content. The modules are:

- The media-processing modules generate indexable-media components, with media and media-anchoring files. These describe the media content and a set of anchor points that enable direct access to locations within the specific media (Dexter components). For example, an image-processing file could contain anchors to regions - "<region id='1' position='3:5' size='10:20'>A BARKING DOG</region>" - of an image file (e.g. a JPG).
- The structuring module (re)introduces the book structure into the main content. The result is a DTBook.DTD compliant file. The module input is the raw digital text, a set of rules to extract structure and specific structure definition if needed.
- The linker modules draw information from the above results, enrich the textual content description and structures, and convert the media-anchoring files into link specification files. Specific linker modules link auxiliary content with the main content. The process is similar to media integration except that input files may be composed of several media (e.g. margin notes with text and audio).

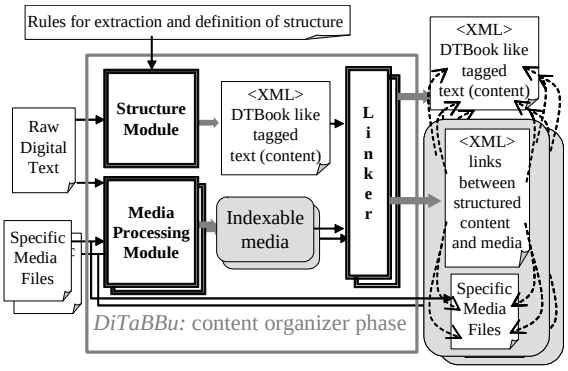

Figure 2: DiTaBBU: content organization phase.

As an example of media-processing module, consider speech alignment (fig. 3). This module is responsible for locating word limits on the speech audio signal and generating the indexable-media component. It is built of four sub modules: a striper, a characteristics extractor, a forced aligner and a tagger. The first extracts punctuation and expands abbreviations, in order to feed a stream of words, similar to the audio version, into the forced aligner.

Extractor and forced aligner form a speech recognizer. The latter is based on Weighted Finite State Transducer (WFST) (Mohri et al., 1998), as its search space is defined by a distribution-to-word transducer built outside the decoder. It can be used with and without phonological rules (Trancoso et al., 2003). The rules cover vowel devoicing, deletion and coalescence, voicing assimilation, and simplification of con-

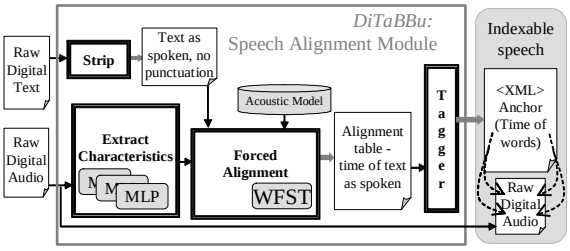

Figure 3: DiTaBBU: speech alignment module.

sonantal clusters, both within and across word boundaries. All rules are optional and compiled into WFSTs. The search space comprises a phone topology, a lexicon and a language model. The phone topology uses hybrid acoustic models, combining the temporal modelling capabilities of Hidden Markov Models (HMM) with the pattern classification capabilities of Multi-Layer Perceptrons (MLP). Three MLPs are given the acoustic parameters and the streams of probabilities are then combined using an appropriate algorithm. The language model is the sequence of words that constitute the orthographic transcription of the utterance. The lexicon is extracted from that sequence. The main advantage of this approach is that no restrictions are placed on the construction of the search space. The result is a table with the audio stream timing for each of the spoken words.

The tagger normalizes the table into a XML media-anchoring file, including a general speech description (e.g. reader, full audio time) and anchors for each aligned word.

4.2 UI-generation Phase

In the initial steps of the UI-generation phase (fig. 4, DiTaBBu presents a set of interpreter modules. Each module receives XML-based files with content, and a specification file describing the patterns and rules to be applied. Those specification files follow XML-based dialects dependent on the module. Internally it also uses XSLT code and XSLT templates that are selected and adjusted according to the specification, in order to generate the output (XML and XSLT).

Two groups of interpreters can be identified, relating to primary and secondary material. The primary material modules are the playback and navigation interpreters. They deal with the main content, not considering footnotes, margin notes and navigation auxiliaries (indexes, tables of content, ...). The respective dialects handle the visual and audio logical markup and their synchronization. For example:

- <showsync delay='2s' sunit='silence'/> means that playback will show visual synchronization marks (the visual effect is specified later in the CSS) delayed by 2 seconds and using the words between reading silences as a unit - the whole unit is marked (e.g. underlined) as narration evolves;
- <onsearch sunit='word;paragraph' based-on='paragraph,section' /> means that in result of a search, the narration (sound) should start on the word found or on the beginning of the paragraph containing that word, depending on how distant (paragraph of section) from the current reading position the searched text is.

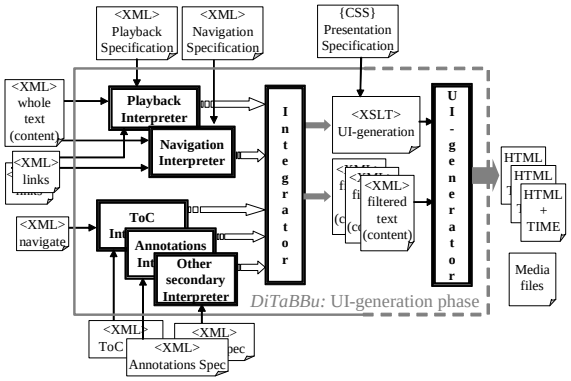

Figure 4: DiTaBBU: UI-generation phase.

The secondary modules handle auxiliary navigation structures, user annotations, etc. Apart from the specificity of their dialects (e.g. <show summary> on annotations), the synchronization rules with the primary content are also specifiable.

The remainder steps of the UI-generation phase provide: (1) the integration and filtering of the above outputs; and (2) the (optional) generation of the final presentation. The former generates a set of XSLT and XML content files that can be also interpreted by the execution platform.

5 CONCLUSIONS

This paper presented DiTaBBu, a framework for the production of DTBs. We have described the requirements and the related work, and consequent options taken for the framework design, the generated books' architecture and the execution platform. The produced DTBs provide the functionalities intended in the standards literature, including audio and text synchronization, annotations, navigation through mouse, keyboard and voice commands. Description and usability evaluation tests of the multi-faceted DTBs can be seen elsewhere (Carriço et al., 2003a; Duarte et al., 2003). These studies focused on UIs variants, generated by DiTaBBu from the same book ("O Senhor Ventura"). Different synchroniza-

tion units, visual and audio marking of navigation anchors, playback and annotation synchronization were used, as well as different forms of interaction.

The modularity of the platform's architecture enables the flexibility required for the creation of such multiple UIs for DTBs, maintaining a mostly automatic production and reinforcing coherence towards the books' contents. The use of rule based modules and templates stresses that flexibility, permitting that specification languages are maintained at a convenient high level, focussed on DTB publishing.

As ongoing work, we are conceiving tools for graphical specification of the modules configuration files. In line of hypermedia related works (Carriço et al., 2003b; Kraus and Koch, 2002) it is being defined an UML description of those specification dialects, that in turn will generate the XML specifications. Work is also being done in the integration of images: an image-processing module and image-linker modules for textual and speech-based description of such images.

REFERENCES

Ali, M. F. and Pérez-Quiñones, M. A. (2002). Using task models to generate multi-platform user interfaces while ensuring usability. In *Proceedings of Human Factors in Computing Systems: CHI 2002 Extended Abstracts*, pages 670–671, Minneapolis, MN.

ANSI/NISO (2002). Specifications for the digital talking book. http://www.niso.org/standards/resources/Z39-86-2002.html.

Calvary, G., Coutaz, J., and Thevenin, D. (2001). A unifying reference framework for the development of plastic user interfaces. In *Proceedings of Engineering for Human-Computer Interaction: EHCI 2001*, pages 173–192, Toronto, ON, Canada. Springer Verlag.

Carriço, L., Guimarães, N., Duarte, C., Chambel, T., and Simões, H. (2003a). Spoken books: Multimodal interaction and information repurposing. In *Proceedings of HCII'2003, International Conference on Human-Computer Interaction*, pages 680–684, Crete, Greece.

Carriço, L., Lopes, R., Rodrigues, M., Dias, A., and Antunes, P. (2003b). Making XML from hypermedia models. In *Proceedings of WWW/INTERNET 2003*, Algarve, Portugal.

Daisy Consortium (2002). Daisy structure guidelines. http://www.daisy.org/publications/guidelines/sg-daisy3/structguide.htm.

Dolphin Audio Publishing (2003). EaseReader - the next generation DAISY audio eBook software player. http://www.dolphinse.com/products/easereader.htm.

Duarte, C. and Carriço, L. (2004). Identifying adaptation dimensions in digital talking books. In *Proceedings of IUI'04*, Madeira, Portugal.

Duarte, C., Chambel, T., Carriço, L., Guimarães, N., and Simões, H. (2003). A multimodal interface for digital talking books. In *Proceedings of WWW/INTERNET 2003*, Algarve, Portugal.

Eisenstein, J., Vanderdonckt, J., and Puerta, A. (2001). Applying model-based techniques to the development of UIs for mobile computers. In *Proceedings of the International Conference on Intelligent User Interfaces: IUI 2001*, pages 69–76, Santa Fe, NM. ACM Press.

Gaver, W. (1993). Synthesizing auditory icons. In *Proceedings of INTERCHI'93*, pages 228–235, Amsterdam, The Netherlands.

Goose, S. and Moller, C. (1999). A 3d audio only interactive web browser: Using spatialization to convey hypermedia document structure. In *Proceedings of the 7th ACM Conference on Multimedia*, pages 363–371, Orlando, FL.

Innovative Rehabilitation Technology inc. (2003). eClipseReader. http://www.eclipsereader.com/.

James, F. (1997). Presenting htlm structure in audio: User satisfaction with audio hypertext. In *Proceedings of ICAD'97*, pages 97–103, Palo Alto, CA.

Kraus, A. and Koch, N. (2002). Generation of web applications from UML models using an XML publishing framework. In *Proceedings of the 6th World Conference on Integrated Design and Process Technology (IDPT)*.

Lin, J. and Landay, L. (2002). Damask: A tool for early-stage design and prototyping of multi-device user interfaces. In *Proceedings of the 8th International Conference on Distributed Multimedia Systems*, pages 573–580, San Francisco, CA.

Mohri, M., Riley, M., Hindle, D., Ljolje, A., and Pereira, F. (1998). Full expansion of context-dependent networks in large vocabulary speech recognition. In *Proceedings of ICASSP 98*, Seattle, Washington.

Morley, S. (1998). Digital talking books on a pc: A usability evaluation of the prototype daisy playback software. In *Proceedings of ASSETS'98*, pages 157–164, Marina Del Rey, CA.

NISO (1999a). Document navigation features list. http://www.loc.gov/nls/z3986/background/navigation.htm.

NISO (1999b). Playback device guideline. http://www.loc.gov/nls/z3986/background/features.htm.

Paternò, F. (2000). *Model-Based Design and Evaluation of Interactive Applications*. Springer Verlag.

Serralheiro, A., Trancoso, I., Caseiro, D., Chambel, T., Carriço, L., and Guimarães, N. (2003). Towards a repository of digital talking books. In *Proceedings of Eurospeech 2003*.

Trancoso, I., Caseiro, D., Viana, C., Silva, F., and Mascarenhas, I. (2003). Pronunciation modeling using finite state transducers. In *Proceedings of ICPhS'2003*, Barcelona, Spain.

VisuAide (2003). Victor reader. http://www.visuaide.com.