PDF to MP3 conversion: an alternative method to reading academic papers

Dorian Miller
Department of Computer Science
University of North Carolina at Chapel Hill

Abstract:
The PDF to MP3 project provides alternative access to academic documents for the blind, weak readers and others. Using an MP3 player the person listens to a synthesized reading of the paper. The audio is divided into tracks and read in multiple voices to accommodate the listener's navigation and comprehension. The developed software performs the PDF to MP3 conversion using the Python programming language and third party components.

1 Introduction

Some people must be provided with alternative access to the power of the written word. The blind are unable to read conventional books. People with dyslexia or other reading disabilities have trouble efficiently decoding the written word. Even people without a disability could benefit from an alternative to reading material; for example, when riding the bus, instead of reading they could listen to the text, as an *audiotext*. A computer system enables a person to access audiotexts, which give similar information as from reading conventional books.

Audio versions of text must provide a literal reading and also convey the text's structure. Text structure conveys meaning. Comprehension of writing comes from understanding a series of independent consecutive points. Sections or chapters represent different parts of a document's main idea. Furthermore paragraphs are sub-points of the idea expressed in a section. In printed material, the format of section headings and whitespace between paragraphs implicitly indicates the transition between individual ideas. Similar format information must also be conveyed in audiotexts.

The organization of a document's sections is also important for how a person reads the document. Newspapers, manuals, and academic papers are examples of documents that can be accessed in random order; that is instead of reading the document front to back the reader chooses to read an individual article or section. The focus of this project is to provide audiotext versions of academic papers to convey the text and document structure. An academic paper is usually 6 to 15 pages and broken into sections like this paper. The typical reader might skim the paper front to back but then randomly access sections to learn more about the details.

The design, prototype, and evaluation of software and hardware that provides the audiotext is described in this paper. Software automatically converts the electronic text file in Portable Document Format (PDF) into MPEG-1 audio layer 3 (MP3) audio files, which can be listened to on an appropriate player. Images in the academic paper, however, are ignored. PDF is a standard format to distribute academic papers. The MP3 audio file format is chosen because users can easily afford a simple to use and versatile MP3 player. The concepts discussed, however, apply more generally to any kind of electronic document and audio format.

2 Related work

This section is an overview of different computer solutions to provide audiotexts. Each solution has its unique navigation techniques between pages, sections, paragraphs etc. and its special features.

Books on tape is the oldest form of audio books started in 1970's and continued to be provided by the Recordings for Blind and Dyslexic (RFBD) [1]. Readings of a book are stored on tapes. A special tape recorder is used to listen to the four sides of the tape; besides the two sides of a normal tape, two more sides are available by reversing the play direction. The tape recorder also provides a dial to vary the playback speed; listeners familiar with the text or just skimming it can increase the tape speed. The inconvenience of using books on tape is that one book requires a large volume of tapes and it is tedious to flip and rewind/fast-forward the tape to find the desired section. However, to assist with the navigation, beeps are used while fast-forwarding or rewinding to indicate page transitions. The listener finds relevant information with an index card accompanying the tapes, which indicates which tape, side and direction pages are on.

Digital talking book (DTB) is a standard for audio books developed by the Digital Accessible Information System (DASIY) [2] consortium and standardized in 2002 by the National Information Standards Organization (NISO). The DTB standard describes how multimedia information, such as audio files, text files and images, are composed to create an audio book. The standard is flexible and combines variable amounts of audio and corresponding text, which enables text searches. Having complete audio and text is not needed, for example, in a dictionary, which might have the complete text and audio only for pronunciations. A DTB viewer program can use the text to display word definitions to a Braille display. Also with a DTB viewer the reader can efficiently navigate between or within sections, because of the hierarchical document structure defined by the DTB standard.

Blind people rely on screen reader software, such as JAWS, to use a PC and read electronic documents. The screen reader reads all text that appears on a screen. Navigating the PC desktop and applications is possible with a series of keystrokes. Electronic documents, such as web pages, are also navigated with keystrokes, which enable moving between pages, lines, and words. The navigation, however, is limited

because there are no direct keystrokes to find the beginning of sections, paragraphs, or sentences.

The original motivation for this project was to provide access to PDF files, which until recently were not accessible with screen readers. In the meanwhile, however, Adobe has released Acrobat Reader 5.1, the standard PDF viewer, with screen reader accessibility. Regardless of Abode's recent development, the solution to listening to audiotexts on an MP3 player provides a unique and convenient access to text material.

3 Audio layout to support comprehension and navigation

To improve the listener's comprehension the layout of the audiotext supports the *active listening* [3] strategy. The active listening strategy is advocated by RFBD to improve people's comprehension when listening to books on tape. The strategy has two stages. The first stage is for the reader to gain an overview of the document by skimming the material; for example, reading the abstract, section headings, and thinking about how the sections relate to the overall idea. In the second stage, the reader reads the complete text. In the audiotext the overview material is placed before the complete text so that the listener does not have to search for the overview material. So the title, author information, abstract and section headings are heard first like the table of contents in a book. Following the overview are the sections in their entirety. With this technique the listener can gain an overview before listening to the rest of the paper.

The audiotext is divided into individual tracks to enable a listener to easily traverse the document. Each part of the overview information (title, author information, and each section heading) is in a separate track. Each of the sections in their entirety are also on individual tracks. Individual tracks are easy to access by moving back and forth between tracks. The typical paper with 10 sections will have approximately 25 tracks, which is a reasonable amount for a listener to flip through. Dividing the paper further into smaller portions, such as paragraphs, would drastically increase the track-count and thereby make it more cumbersome for the user to flip between sections.

With this technique the listener can efficiently access a desired section. First the listener can flip between the section headings (corresponding tracks) in the overview to find the section he/she wants to read. Then by means of the section number included with the section headings, the listener can browse to the desired section; the section number announced at the start of each section indicates if the desired section is before or after the current track.

The drawback to this design is that navigating within a section is complicated. A section is designated to a track which means navigation within it is by rewinding/fast-forwarding. It is a matter of hit and miss to find the beginning of paragraphs or sentences. Future work will hopefully improve the navigation within a section.

4 Annotations convey document structure

Annotating the original text of the document conveys the document structure in the audiotext, which is important for the listener's comprehension. In this work annotations take the form of spoken keywords or beeps.

Some of the tracks are annotated by spoken words with additional information to help the listeners orient themselves in the audiotext. The tracks for the title, author information, abstract, and sections start with the corresponding keywords. When the listener hears the keyword, he/she will know the approximate position in the paper and whether it is necessary to move forward or backward to reach the desired part of the document. The tracks representing the section headings are also identifiable because they do not have keywords. Listeners are most likely to listen to these short tracks consecutively so keywords would be distracting here.

Emphasizing the keyword distinguishes it from the same words appearing in the text and alerts the listener to the newly started track. Changing the characteristics of the voice that speaks them emphasizes the keywords; for example, the voice for the majority of the text may be in a female voice but the keywords will be spoken in a male voice. Hearing the emphasized "section" keyword between sections alerts the listener to a section transition and transition between major ideas.

Emphasizing printed structure by changing voice characteristics has been explored by other research and could also be applied to this application. TV Ramen uses the technique to emphasize the structure of mathematical formulas and table structures [4]; for example, when reading mathematical equations subscripts are read with a deeper voice. The voice characteristics are unique enough to be used in conjunction with the keywords and beeps and not confuse the listener.

Structural information can also be conveyed by beeps, which is an appropriate technique for experienced listeners. A short beep between paragraphs signifies the transition between them. Listeners familiar with an audiotext will understand the meaning of the beep as opposed to the novice user. Also experienced listeners might appreciate a shortening of the audiotext and a beep is likely to be shorter than a keyword. Ideally, the listener can customize the annotations as keywords or beeps.

5 Practical to use

The PDF to MP3 conversion tool will only be useful if it is convenient and practical, which is the case with MP3 technology. The tool executes in two stages. The first stage is the automatic conversion from PDF to MP3. The second stage is for the user to listen to the MP3 tracks on a MP3 player.

The conversion process is convenient because it is automated. The user provides the parameters for a customized reading voice and a PDF, which is automatically converted into MP3 files. Given the proper settings, the files are downloaded directly onto the MP3 player. Although the process is automated, it still takes several minutes for the computationally intensive conversion process and for copying files to the MP3 player.

The MP3 player's handy form factor is like a Walkman and convenient to use. The user can choose the most comfortable location to listen to a document, such as on a couch at home or outside on a grassy field. Although the user could listen to MP3's on a laptop, moving a laptop is less convenient than a MP3 player.

A MP3 player is also financially practical as it is a mainstream consumer product. The high sales volume and innovations in computer hardware are bound to further decrease the price. Besides listening to audio documents the user can use the same hardware to listen to music files, as originally intended. Using an MP3 player is more practical than the BookCourier [5] alternative. The BookCourier is a specialized text-to-speech device; the text is downloaded to the device and read. Although it has the same form factor the price is less likely to decrease because there is a limited market for it.

One function not readily available on MP3 players is changing the speed of playing the audio. The user can set the playing speed before the PDF conversion process but cannot change it while listening to it. However, certain Mp3 players are programmable, such as Archos Jukebox [6] and Neuros [7], and can be modified to provide this functionality. Furthermore using the phase vocoder the speed can be changed without changing the pitch [8]. In the frequency domain the frequencies are multiplied by the inverse of the rate change. To preserve the audio signal the phase corresponding to the original frequency is matched to the new frequency. The new audio signals time domain equivalent is played back at the new rate.

6 Design and prototype.

The PDF to MP3 conversion is feasible to implement in three stages by combining existing techniques.

The first stage, the most complex, extracts the text and structure from the electronic documents. Extracting text from electronic documents is straightforward, however, extracting structure is not. Most document formats only preserve typesetting and layout information about the text. The structure, such as title, section headings, or footnotes, is not preserved. The challenge is to infer the structure from the text style; for example, headings are likely to be bold and with a slightly larger font size than the main text. However, there is no standard so it is difficult to determine the criteria that apply in all cases. The exception is latex documents, which strictly enforce structure. The latex source, however, is not widely distributed because it is converted into other formats, such as PDF.

The second stage is to perform a text-to-speech (TTS) conversion and modify the synthesized voice to reflect the paper structure. TTS is thoroughly studied and several practical solutions exist. In TTS, the phonetics of each syllable is determined and the corresponding audio produced. Altering the pitch and speed of spoken words can change the characteristics of the voice. Ongoing efforts in the field are to make the robotic sounding synthetic voices more humanlike.
(Reference TTS research)

The final stage is to store the audio to file, which is a matter of choosing a file format. The WAV file format is the raw audio data, which a soundboard uses to reproduce the sound. The WAV data can be considerably compressed when converted into the MP3 format [9]. Depending on the audio quality and the number of channels the WAV to MP3 compression ratio ranges from 12 to 24. The audio quality can be reduced as long as the recorded synthesized speech is still comfortable to listen to.

The prototype of the described system has been implemented from existing software components. The Python [10] programming language is used to combine the different components of the system. The "pdftohtml" [11] software extracts text and font settings from a PDF and converts them into HTML. The Python HTML parser reads in the HTML file, divides the paper's components based on format, and stores the text in a data structure. The text in the data structure is processed by Microsoft's TTS [12] engine using the male voice to emphasize keywords and the female voice to read text. Microsoft Speech Application Programming Interface (SAPI) also provides the functionality to save the spoken word into WAV files. The WAV files are converted to MP3 files using LAME [13] software. The final audiotext is about 30-45 minutes long depending on paper length and reading speed.

The prototype is limited to converting papers of a fixed format, in which headings are the only bold text. The parts of the paper, such as title, abstract, and sections, are identified as being between headings. Documents with other formats could be converted but would be randomly divided into audio tracks. The listener can still listen to the paper but does not have convenient access to sections.

7 Evaluation

The purpose of the evaluation is to measure the listener's comprehension of an audiotext. The evaluation should be performed with blind readers, weak readers, and others interested in an alternative access to text.

Although a formal study was not performed, one possibility would be to perform it as follows. All participants are given an audiotext with a general subject and a chance to practice using MP3 player. The effectiveness of the audiotext is measured by the time and accuracy with which the participants complete a questionnaire about the audiotext. At the end of the experiment the participants can share their experiences of using the audiotext and suggest improvements.

It will be interesting to compare the performance between those that have used a form of audiotext and those that have not. My hypothesis is that those experienced with audiotexts will outperform those with no audiotext experience.

A possible enhancement to the experiment might improve the listeners comprehension. In addition to listening to the audiotext, the participant has an alternative access to text; for example, the sighted follow along on a paper copy and the blind use a Braille display. I hypothesize that the comprehension will improve because the listener's attention is more focused on the content. Also having two forms of the text might make the facts more memorable and therefore improve the comprehension.

8 Future Work

The audiotext navigation techniques explored in this project can be enhanced and possibly applied to electronic books, which display one-page at a time.

With programmable MP3 players it will be possible to expand the interface; buttons for play, stop, forward, backward can be replaced. The paper can be divided down to the sentence level and organized in a hierarchy of folders supported by MP3 player; for example, at the first level are sections, at the second paragraphs, and at the third sentences. Then the buttons on the MP3 player could be used to navigate the folders.

The navigation techniques used for an audiotext also apply to electronic books. Unlike a conventional book, an electronic book does not provide the same easy method of flipping pages to find the desired information. Some simple navigation techniques include flipping consecutive pages by jumping directly to a page number. Enabling access to sections can enhance the navigation.

9 References

1. web. *Recordings for Blind and Dyslexic*. in *http://www.rfbd.org/*. April 15, 2003.
2. web. *Digital Accessible Information SYstem*. in *http://www.daisy.org*. April 15, 2003.
3. video. *Video on active listening*. in *RFBD*. 1998.
4. Ramen, T. *Emacspeak --A Speech Interface*. in *CHI*. 1996.
5. web. *BookCourier*. in *http://www.ostrichsoftware.com/*. April 15, 2003.
6. web. *Archos Jukebox*. in *http://www.archos.com/*. April 15, 2003.
7. web. *Neuros*. in *http://www.neurosaudio.com/store/prod_neuros.asp*. April 15, 2003.
8. Robinson, A. *Changing the Speed of Music Without Changing the Pitch (technical discussion)*. in *http://www.seventhstring.demon.co.uk/xscribe/slowdown.html*. April 15, 2003.

9. web. *Audio & Multimedia MPEG Audio Layer-3*. in *http://www.iis.fraunhofer.de/amm/techinf/layer3/*. April 15, 2003.
10. web. *Python programming language home-page*. in *http://www.python.org/*. April 15, 2003.
11. Kruk, M. *PDF to HTML conversion tool*. in *http://pdftohtml.sourceforge.net/*. April 15, 2003.
12. web. *Microsoft Text-to-Speech research*. in *http://research.microsoft.com/srg/*. April 15, 2003.
13. web. *LAME Ain't an Mp3 Encoder (LAME)*. in *http://lame.sourceforge.net/*. April 15, 2003.