Automatically Structuring Text for Audio Learning

Barbara Leporini[1], Maria Claudia Buzzi[2], Marina Buzzi[2], and Giulio Mori[2]

[1] ISTI-CNR, via Moruzzi, 1 56124 Pisa, Italy
Barbara.Leporini@isti.cnr.it
[2] IIT-CNR, via Moruzzi, 1 56124 Pisa, Italy
{Claudia.Buzzi, Marina.Buzzi, Giulio.Mori}@iit.cnr.it

Abstract. In recent years podcasting has been in great demand as a recreation and a learning tool. In this paper we describe the design and implementation of a system for automatically converting documents to structured audio. Our prototype is a Web-based service for preparing structured audio material to download on portable mp3 players. The on-line service is especially designed to aid users with special needs, such as the visually impaired. Ultimately, this would enhance comprehension for all.

Keywords: Podcasting, e-Learning, blind, mp3 files, document converting.

1 Introduction

Distance learning is increasingly used to allow a greater number of people to expand their knowledge anywhere, anytime. In addition to lessons and exercises, distance learning includes a variety of educational tools such as wikis, blogs, forums, assessment SW, games, simulations and podcasts. Although a recent phenomenon, podcasting is a fast-growing tool in the field of education [6]. A podcast is an audio or video digital-media file distributed over the Internet, that enables students and teachers to share information whenever they wish. Unlike other digital-media formats, a podcast can be syndicated, subscribed to, and downloaded automatically when new content is added. For instance, absent students can automatically receive podcasts of recorded lessons by accessing the Internet. This study was restricted to audio podcasting, which is easily used with mobile devices including portable media players.

Efficacy of podcasting has been shown by numerous studies: in fact many individuals learn better when listening to educational material than by accessing written learning objects [1], [5], [8], [10]. Advantages of audio learning material include:

- Multitasking: a student can listen to a vocalized text on an mp3 player, while performing other activities or in motion (driving, travelling or walking)[10];
- Ubiquity: podcasting enables distance learning "anytime, anywhere" beyond class limits [1]
- Emotional: Students who miss a class are less anxious if they know they can hear the lesson later [10]; furthermore, the possibility of listening to didactic material more than once reinforces their self-confidence (and obviously their knowledge)

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2009, LNCS 5616, pp. 73–82, 2009.

- Personalization: possibility of repeating the study of a concept until learned, since different individuals have different learning ability rates [10]. Furthermore, flexibility of tuning speed allows adapting the velocity to student preference: if speech is too slow/rapid, playback speed can be adjusted
- Teaching improvement/assessment: teachers, by listening to their lessons, may enrich materials and refine their ability to teach since podcasting may force them to collect and present thoughts in a more logical order [7].

With podcasting, learning paths can be tuned to the individual "rhythm" of each student, since asynchronous interaction allows one to listen to content as often as necessary. Furthermore, specific SW scripts allow podcasts to be automatically transferred from a personal computer to a mobile device. This makes podcasts quite easy to use for the blind persons. Thus, podcasting is valuable for everyone but is especially useful for people with special needs such as the blind.

Audio files can be prepared by recording the content directly (e.g. lectures, radio/TV programs, etc.). Another way of producing an audio version of a document is recording via microphone. This requires considerable time and effort.

When podcasting is used for educational purposes or for reading work documents, it is important to make content as accessible and readable as possible. In order to make educational audio files easy to use, content should be well structured because a sequential and continuous reading is not appropriate for effective learning. Converting a text file to short audio files can be useful for two reasons:

1. it is faster to retrieve specific info or navigate within a podcast (for instance, when one wants to go over a lesson again)
2. it reduces the size of a single audio file. Long podcasts may decrease attention, thus reducing comprehension [4], [10].

If the speech-contents are listened to on an mp3 player, some text-to-audio converters [text2mp3, DSpeech, etc.] allow breaking the contents down into various mp3-based files. The rules for this are related only to duration time (e.g., each file takes 5 minutes) or to a break string used to split the audio content into several files. The latter possibility is interesting but must be done manually by introducing a special "break string" within the source text. This requires user time and effort.

In the next part of this paper we discuss the design and implementation of a system for automatically structuring text to speech. Our prototype is a Web-based service for preparing audio material (a set of audio podcasts) starting from structured text instead of live recording. The on-line service is especially designed to aid users with special needs, such as the visually impaired or foreign students, but it would enhance comprehension for all, not only for the differently-abled.

The paper is arranged as follows: in Section 2 we report on related studies in this field, in Section 3 we describe a web system prototype for automatically processing text to generate a structured mp3-based audio version. Lastly, in Section 4 we report conclusions and future works.

2 Related Works

Many studies confirm podcasting as an increasing technology trend that is very useful for exchanging information and integrating/improving learning approaches with

efficiency in terms of resources consumption. For many people, listening may be more attractive and less tedious than reading [4], [10]. Podcasts are often used for additional support in teaching and learning. Some common uses in higher education are: taped lectures, guest speakers, group presentations, tutorials, exam reviews, reinforcement of key concepts, and drill or repetition [10].

Recently, to facilitate the preparation of audio materials, several tools have been proposed and developed, that transform a text document into a spoken version by using text-to-speech (TTS) technology and a voice synthesizer. A TTS system converts a text into speech [2]. The output generated can be heard immediately while the audio is being produced or can automatically be recorded in audio files. Typically, these audio files are in mp3 format and can be listened to on an mp3 portable player or on smart phones. Tools like Robobraille and vozMe are examples of this converting process. RoboBraille is an email-based service which automates translation of text documents into Braille and speech. Users submit documents (e.g., text files, Word documents, HTML pages) as email attachments. The translated results are returned to the user via email [11]. VozMe (http://vozme.com/) is an easy-to-use online service for creating mp3s from text. It only requires typing in or pasting the text and pressing a button. However audio content created is not structured (is a single file) so it is not very suitable for studying.

Some SW such as Natural Reader [9] allows one to jump to the previous or next section. In [4] authors combine sophisticated speech-processing techniques (including audio-based skimming) to create a multimedia player which allows audio-centered browsing and navigation of recorded presentations. However these systems require interaction with a PC and thus are not suitable for Mobile Learning.

3 The System

3.1 Motivation

Structured content offers many advantages for blind users. For example, a blind user cannot scan a web page visually, but if structural mark-up has been coded (headings, paragraphs, lists, etc.) it is possible to scan the page audibly since screen readers identify and announce them to the user, providing additional information. Therefore, the structured content announced by screen reader provides the user with a way to navigate a page: the user may jump between headings (e.g., using JAWS screen reader “h” command) or between paragraphs (via “p” JAWS command) rather than listening or tabbing a page from top to bottom, which can be very tedious and time consuming. In addition, a structured web page provides an overview of the content, so content is easier to read. Basically this allows a blind user to scan a page and quickly jump to key information. All these features are particularly important when reading educational content, since in order to make a learning object easy and simply to read the user needs to get an overview, skip rapidly to a chosen section, or move back and forth in the text.

This concept can be extended to the audio learning object. An audio version of a story or novel does not require structuring since reading is sequential (it flows from the beginning to the end). Instead, educational materials may be more effective if broken down into units, for easy and rapid exploration. If a document is an important

source of information, learners will not simply read it once from the beginning to the end, but will also looking up specific things, so the possibility of browsing and navigating the documents is important [6]. A table of contents offers a rapid overview of a document and the possibility of skipping from one part to another. Analogously, a vocal file should also allow users to reach the desired material rapidly.

3.2 Preliminary Questionnaire

As a preliminary stage of this study, we were interested in learning if and how blind users utilize mp3 players or podcasts. We then built a questionnaire which we sent to a set of totally blind persons contacted through the Italian Association of the Blind, to which one of the authors belongs. We sent the questionnaire via email the questionnaire to potential participants also using a mailing list to which many blind people in Italy subscribe. We received 33 responses. The sample included 9 females and 23 males, age ranging from 18-24 (12%) to over 75 years (3%), as shown in Figure 1.

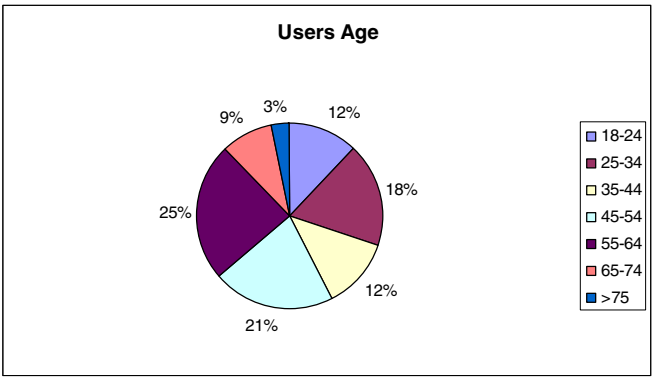

Fig. 1. Users age

The results collected by questionnaire showed that 79% (26 of 33) of the sample utilize mp3 players and 70% (23 of 33) use them habitually to hear books and other textual material.

The age distribution of mp3 player users is quite balanced, as shown in Fig. 2.. The relative percentage of mp3 diffusion decreases as age increases.

Only 52% (17 of 33) of the sample has used a text-to-audio translator: 37% (12 of 33) often and 15% sometimes.

A total of 42% (14 of 33) of the sample declared that they prefer to listen to audio files on mp3 players instead of reading text (via screen reader) and 36% (12 out of 33) believe that it could be very useful to have more structured audio material. Specifically, users ask for bookmarks, having more audio files instead a single file, being able to extract interesting parts, and simple interaction with the SW UI.

From people using text-to-audio converters we received comments on its user interface (UI): unclear parameters or unsuitable UI were indicated as uncomfortable features. This aspect is the basic concept we consider in our investigation.

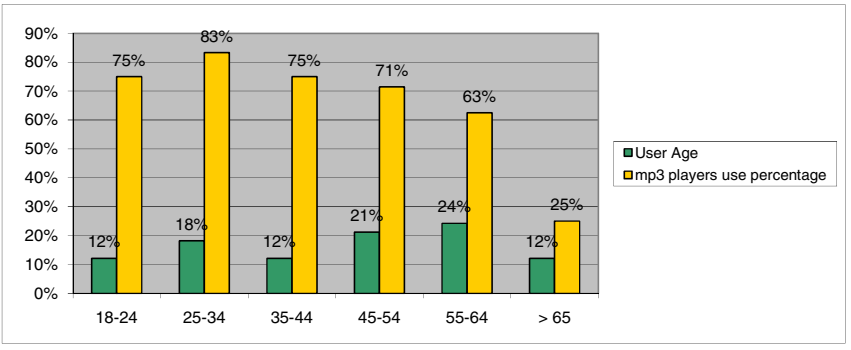

Fig. 2. Distribution of Mp3 players for age

These results motivated us to undertaken this study with two main goals:

- providing a tool for splitting a large document into logical subsections to be converted in separated audio files, also providing a document table of content, (when possible);
- developing a system that offers simple interaction via screen reader for blind users (e.g. students and teachers).

3.3 The Architecture

3.3.1 Overview

The main feature of our web-based system is the ability to structure a document in a more suitable mode for learning, producing audio files which can be listened to on a portable mp3 player. The system is conceived as a module of a Learning Management System (LMS) to handle audio learning content. This function would be useful for both teachers and students. However, the system can be used by anyone to convert structured text documents into a structured audio version.

Our prototype is a web-based service that takes a structured text file as input and produces a set of mp3 files, exploiting a SAPI (Speech Application Programming Interface) speech synthesizer for reading (or recording) document content. To obtain a set of audio files, pre-manipulation of the source text is carried out before it is transformed by the text-to-speech converting tool.

In Figure 3 the architecture of the system is shown; the server is responsible for generation of web interfaces and transformation of uploaded documents into audio files. For example: a teacher can upload a document of a lesson using the simple web interface (Fig. 4) and a student, at a different time, can connect to the server with a personal computer and download on his/her computer mp3 files containing as audio content the (text) sections of the document (Fig. 5).

For blind users additional audio files (".talk") are provided. Those files are related to the software Rockbox, an Open Source replacement firmware for portable digital audio players, that makes certain mp3 players accessible (http://www.rockbox.org). Rockbox permits to associating to an audio file (or to a folder) an additional audio file

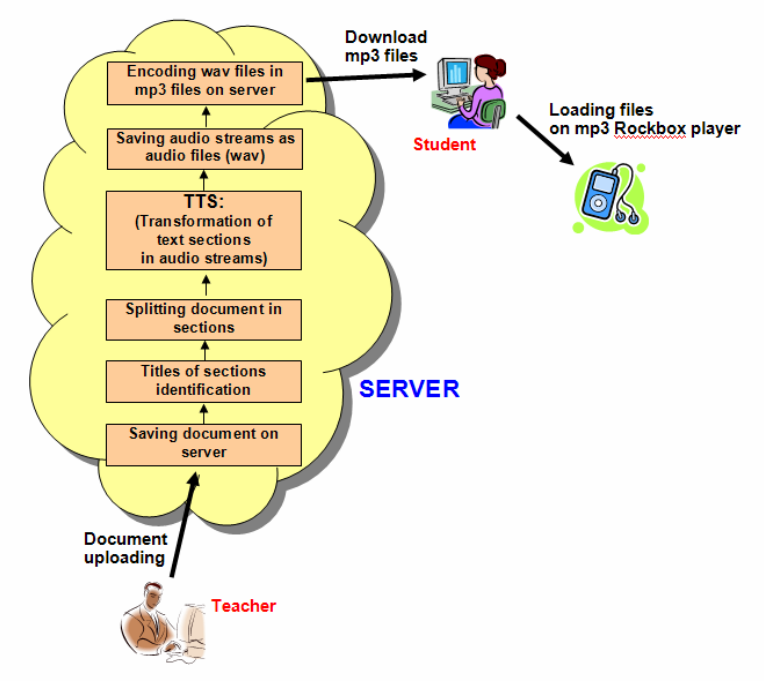

Fig. 3. System architecture - Example of use

(with the extension ".talk") containing the spoken name of the file or folder This kind of files is useful for listening to a good pronunciation of the name of the audio files or folders available on the Rockbox-based device. In fact a ".talk" file is such an audio pronunciation of file--or folder--name, necessary since Rockbox is not a screen reader and otherwise it would read the file or folder names by spelling them letter by letter.

In addition to mp3 files, our system also provides ".talk" files, so the UI of the TTS-based conversion results allows downloading normal mp3 files, or the version for Rockbox (mp3 + ".talk" files) on the PC. In the latter case, when a blind user navigates the folder, the vocal version of each file (".talk") announces the file name.

After the download, a student can transfer the mp3 files to his/her player (with a usb cable connected to his/her personal computer), ready to be listened to anytime.

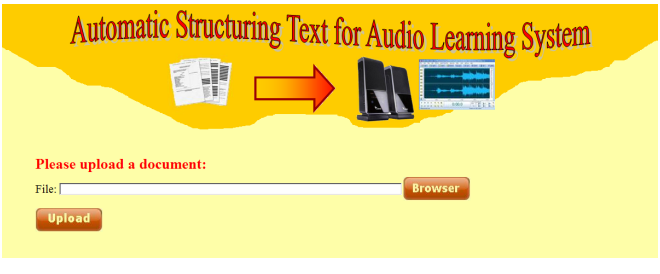

Fig. 4. Documents uploading web interface

3.3.2 Programming Environment

The system is implemented in Python (http://www.python.org/), a simple programming language which offers good performance in terms of speed computing.

We used an Apache server (http://www.apache.org) in combination with a mod_python module (http://www.modpython.org/) so that Apache could support Python language. The Win32com library of Python language offered the possibility of interacting with Windows OS and applications, without too much effort for the developers.

We use Pytts, an easy to use Python library for Text-to-Speech transformation. The results of the TTS process are audio files in wave format, so we utilize a Lame encoder to transform wave files to an mp3 version and remove wave ones, in order to save space on the server disk. To create mp3 files for Rockbox we used the RB Clip utility.

3.3.3 The Heuristics

Appropriate and suitable heuristics need to be identified and specified in order to analyze the source document for producing various audio files of its contents. Basic heuristics were implemented with regular expressions that allow easy manipulation of text. The algorithm must detect sections or other features in the source content and split the text into separate files, one for each section detected. Using the proposed prototype with plain text file, we observed that high precision is difficult to obtain since there are no clear clues for splitting the file according to logical sections. While processing Word and RTF (Rich Text Format) files, which usually contain formatting features (e.g. font type or size, heading styles, bold, italic, and so on) is easier and produces more accurate results. We are still working on improving and refining the heuristics for plain text files.

When identifying section titles (based on threshold for minimum and maximum number of chars/words, previous and next detected sections, etc.) the system filters figures and table captions, lists elements, etc. Thresholds depend on the type of document considered.

To eliminate empty or very small fragments, if the system finds sections without (or with only a few) characters after the title, put it together with the next one (see Fig. 5).

3.3.4 Conversion Process

A user can upload a document using a simple web interface and the document file is saved on the server. A cycling process analyzes the content and splits it into parts to be used as input for the audio conversion.

When the system has detected the section titles, the procedure splits the document into a number of portions corresponding to the identified titles (each section from a title until the next); each part should represent a single section. Then, the TTS library (pytts lib) transforms each section into an audio streaming residing in the computer memory and subsequently these streams are saved in wav audio files on the server. Finally, a Lame encoder transforms wav files into mp3 versions, saving space on the disk. The system presents the user with a list of the links to the files generated in a web page for downloading. The mp3 files are named with the section titles and it is possible to download one of them by clicking the related link. Through two specific buttons the zipped version of all mp3 or Rockbox files can be downloaded (see Fig. 5).

Fig. 5. Result of document conversion

At the moment, the used TTS supports English language for audio files content generation. We are planning to add an Italian language TTS in order to be able to perform user testing with blind users from the Italian Association for the Blind.

3.4 The User Interface

The system is mainly Web-based in order to allow the user to access it remotely. That means the user interface is composed of a few XHTML pages. The system is mainly structured as follows:

- An "upload page" for transferring the document to the server, shown in Figure 4;
- A "Download list page" from which the resulting audio files can be downloaded. This interface is very simple and presents a brief feedback on success or failure of the conversion. Two buttons allow downloading the zipped mp3 files or the Rockbox version: zipped mp3 and ".talk" files. Last is a link so the user can return to the upload page. Figure 5 shows the main results of a document conversion.

4 Conclusions and Future Work

In this paper we described our approach to generating podcasts from (structured) text documents. This approach is particular suitable for generating audio learning objects which need to be easily, efficiently and rapidly explored, especially when using a mobile mp3 player. Specifically we designed and developed a web-based service prototype for converting ".doc", ".rtf" and text documents into a TTS-based spoken version. The prototype currently performs very well with structured documents while some improvements are still necessary to process plain text documents. For instance, if no structure is identified, a plain text document may be reduced to flat chunks

(part1, part2, …) based on length of speech for optimal size of the generated mp3 files. However the basic idea is to provide an audio version split into several files. In this way the user is able to move forward and backward between the files and to get a rapid overview as well.

It is very important to consider the advantages of our method, which not only favours the learning process, but allows skipping the recording audio phase (so requiring less human effort) and saves the cost of hardware recording resources.

An additional benefit derives from eliminating noise (or at worst, it might be present but at an insignificant level). It is well known that normally in chaotic, noisy environments (like a classroom) live recording must consider these aspects, as well problems related to the matrix microphones positions, in order to favour a good quality signal level of speech.

Human voices obviously have a better and more natural influence on listener's comprehension, compared to a digitalized voice, but increasing good quality of recent SAPI (Speech Application Programming Interface) and speech synthesizer (such as Loquendo TTS, a vocal synthesis SW providing natural-sounding voices), attempt to reduce this gap.

Future studies will be oriented toward enriching the system by introducing other input document types (such as pdf and odt) and to improving the algorithm heuristics.

Another interesting feature to add would be the introduction of an automatic user notification service about the presence of new content audio files on the server; the user could subscribe to an RSS feed media aggregator related to a particular web page associated with a server containing formative or education podcasts of interest. This automatic notification service could be further enriched by sending an announcement to the user's mobile phone.

References

1. Aldrich, D., Bell, B., Batzel, T.: Automated Podcasting Solution Expands the Boundaries of the Classroom. In: Proceedings of the 34th annual ACM SIGUCCS conference on User services, pp. 1–4 (2006)
2. Allen, J., Sharon Hunnicutt, M., Klatt, D.: From Text to Speech: The MITalk system. Cambridge University Press, Cambridge (1987)
3. Campbell, G.: There's Something in the Air: Podcasting in Education. EDUCAUSE Review 40(6) (November/December 2005), http://connect.educause.edu/Library/EDUCAUSE+Review/TheresSomethingintheAiPo/40587
4. Cebeci, Z., Tekdal, M.: Using Podcast as Audio Learning Objects. Interdisciplinary Journal of Knowledge and Learning Objects 2 (2006)
5. Deibel, K.: Course experiences of computing students with disabilities: four case studies. In: Proceedings of the 39th SIGCSE technical symposium on Computer science education, pp. 454–458 (2008)
6. Lauer, T., Hürst, W.: Audio-based Methods for Navigating and Browsing Educational Multimedia Documents. In: Proceedings of the International Workshop on Educational Multimedia and Multimedia Education, pp. 123–124 (2007)
7. Mayer, J.: Law school innovations and Jim Milles on podcasting, November 7 (2006), http://caliopolis.classcaster.org/blog/legal_education_podcasting_project/2006/11/07/milles

8. Mermelstein, B., Tal, E.: Using Cellular Phones in Higher Education. In: Wireless and Mobile Technologies in Education (WMTE 2005) (November 2005)
9. NaturalSoft Text-toSpeech, http://www.naturalreaders.com/index.htm
10. Ormond, P.R.: Podcasting enhances learning. Journal of Computing Sciences in Colleges 24(1) (October 2008)
11. The RoboBraille Consortium. The RoboBraille email service: http://www1.robobraille.org/websites/acj/robobraille.nsf
12. W3C. Web Content Accessibility Guidelines 2.0 (December 5, 2008), http://www.w3.org/TR/WCAG20/
13. Wolff, T.B.: Podcasting Made Simple. In: Proceedings of the 34th annual ACM SIGUCCS conference on User services, pp. 413–418 (2006)