Wikiaudia: Crowd-sourcing the Production of Audio and Digital Books

Ashwini Venkatesh, Lalitha M. V., Jyothi Narayana and Kavi Mahesh

***Abstract*–Audiobooks are an indispensable aid in making books accessible to people with special needs such as visually challenged, elderly and those afflicted with problems of dyslexia and agraphia. Audiobooks are currently generated manually for the most part with consequent dependency on individuals having to invest substantial amounts of time and effort to read out an entire book. Automation in the form of text-to-speech, on the other hand, is available primarily in English and other European languages and is still inferior in quality and accuracy to human voice. In this context, this project designed and built a platform for crowd-sourcing the production of audio books in any language, particularly in several languages of India. It also provides a useful feature of generating Unicode text of the books thereby contributing to better preservation of classical and out-of-print books with the added advantage of search and browse functions which are currently not possible in "digital libraries" containing only scanned images of books.**

***Index Terms*–audiobook, crowd-sourcing, digitization, visually impaired.**

I. INTRODUCTION

Since the invention of phonography by Thomas Edison, audio books (then known as "Phonographic books") have been in use. As Wikipedia states, Audiobook was the original application envisioned by Edison which would "speak to the blind people without effort on their part."

Creation of these audio books has been a matter of research for a long time. The aim was to automate the process of manually reading books by volunteers which was time consuming. This led to technologies such as optical character recognition (OCR) and artificial voice synthesis (or text-to-speech). But this has drawbacks of being language dependent and sounding rather artificial or robotic. A vanilla one-flavour-fits-all voice might not be adequate to meet the needs of our audiences with special needs. This was in fact confirmed by our interactions with the chairman of a well-established NGO in India which works for the visually impaired.

Crowd-sourcing is a method of outsourcing the work to the crowd, where some reward is given for work that meets expected quality. Applying this model to the present problem of audiobook generation helps overcome the above mentioned difficulties.

Our goal is to create audio books and their Unicode text for all languages which can be read and written by volunteer users, with a special focus on Indian languages for which hardly any technology is available for OCR or text-to-speech synthesis. This is possible because the task can be distributed among the crowd which knows the language. Human volunteers are asked to read small parts, thereby reducing the burden on a single user by the division of work among the members of the crowd.

The design also supports generation of Unicode texts of the books. Digitizing a book by generating Unicode text is an especially relevant issue in languages such as Indian languages as this improves access to classical material and provides better preservation. It also enables enhanced searching and browsing options in the resulting documents. Additionally, the physical storage required is lesser compared to storing images.

This project has the potential for significant social impact by making books readily accessible to people with special needs. This in turn is expected to positively influence the way education is imparted to such people. As a secondary benefit, classical and out-of-print books and palm-leaf manuscripts can be better preserved by creating both digitized text and audio versions through crowd-sourcing. Finally, popular books, current research papers, newspapers and periodicals too can be converted to audio form on our platform. It may even make audiobooks less boring since several people reading it out may sound less monotonous than a single voice, be it human or robotic.

This paper gives a detailed description of Wikiaudia

Manuscript received December 03, 2014. This work is supported in part by the World Bank/Government of India research grant under the TEQIP programme (subcomponent 1.2.1) to the Centre for Knowledge Analytics and Ontological Engineering (KAnOE), http://kanoe.org at PES University, Bangalore, India.

Dr. Kavi Mahesh is the Dean of Research, Director of KAnOE and Professor of Computer Science at PES University. He can be reached at drkavimahesh@gmail.com

The first three authors are students who recently graduated from PES Institute of Technology, Bangalore, India. They can be reached at ashuven63@gmail.com, lalithamv92@gmail.com and jyothi.narayana10@gmail.com

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

- our solution for creating audio books along with Unicode text. It first describes previous work done in this field and differentiates it from the present work. This is followed by the design of the system with a description of the various functionalities. Ways of maintaining quality of the end result (in this case the final audio book and its Unicode text) are described next. In the end, the results and conclusion of the paper are outlined.

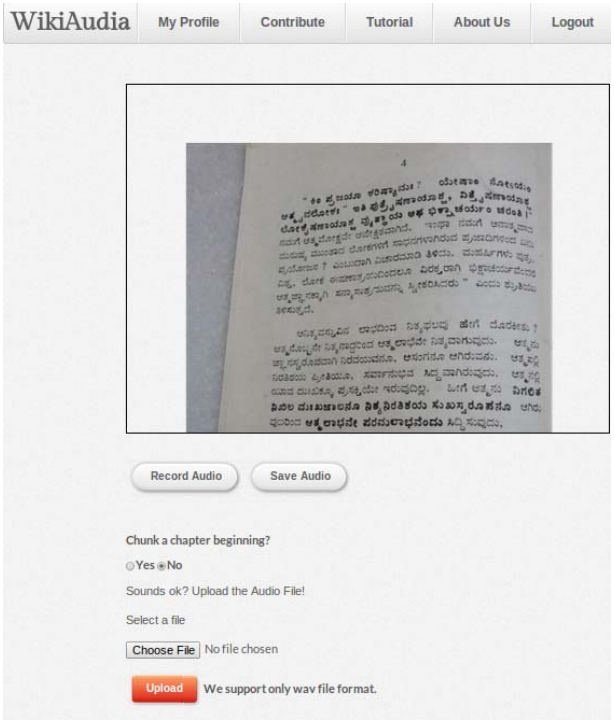

Fig 1. Screenshot of Wikiaudia Crowd-sourcing Portal.

II. WIKIAUDIA

Wikiaudia is a platform for crowd-sourcing the creation of audio books and their Unicode texts. The idea is to make use of the ability of natural human speech to produce an audiobook. Thus, we are delegating the primary task of audio generation for the audiobook to a human. This way we also collect natural human voice as opposed to a computerized version of it and also preserve the accent typical of that particular language. This is very essential in many Indian as well as other less popular languages.

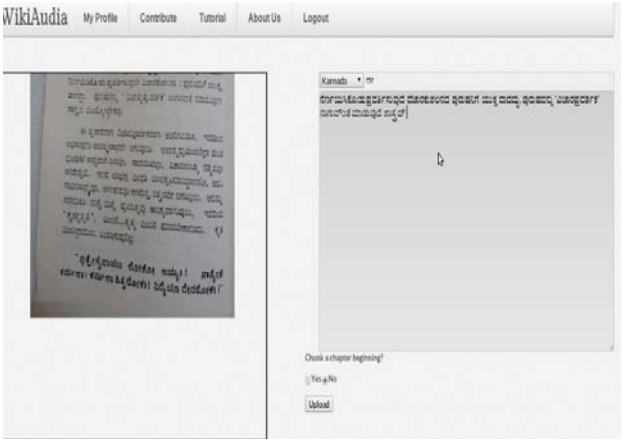

Fig 2. Screenshot of Wikiaudia showing Unicode Text being typed in alongside a Scanned Image.

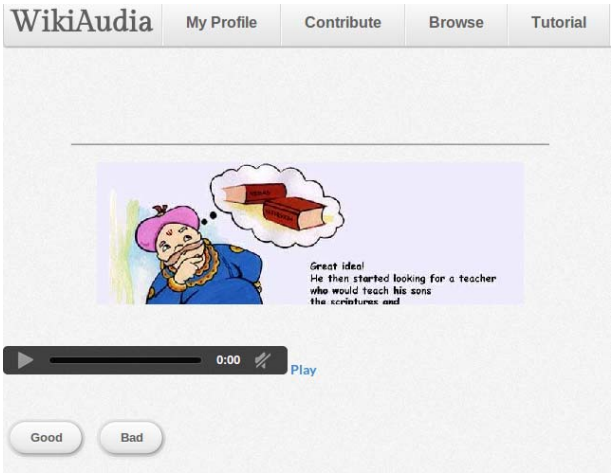

Fig 3. Wikiaudia Browser Interface for Listening to an Audio Book.

A user of the portal registers by providing his/her credentials and signing up for those languages which he/she can read and write. A user is asked to contribute only for books in these languages, which he/she can do in several ways:

- Read out a part of a book to contribute to the audiobook in making; and upload the resulting audio (see Fig. 1).
- Convert the book into Unicode text. Type the text shown in the image in the respective language using the transliteration tool provided on the portal in the browser itself (see Fig. 2).

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

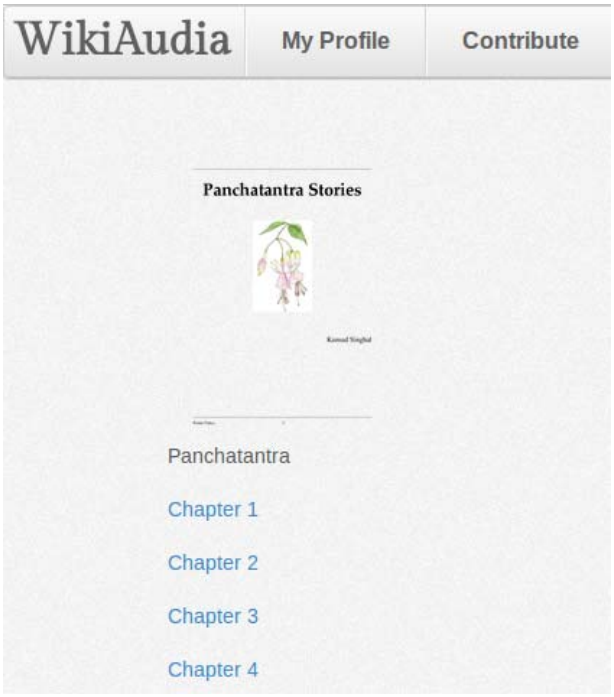

Fig 4. Chapter-wise Structure of an Audio Book in Wikiaudia.

- Request for a book to be converted to audiobook and/or Unicode text. User should upload the book into the portal in one of the supported image or text formats.
- Validate audio clips which have been uploaded by other users. The user can up-vote or down-vote the audio clip thereby validating the quality of the audio clip (see Fig 3).
- Browse and download the generated content (see Fig 4).

Wikiaudia aims at harnessing the power of the crowd by taking into consideration their interests and perspectives. The crowd here is the decision maker and these decisions are based on the needs of their communities. For example, this may imply that a certain genre of books, say classics are given importance over others, hence accelerating their production in the form of audio books. The crowd also decides what quality of audio or what accent in a particular language is acceptable.

Not only does it distribute the task among the members of a crowd, but also considerably scales down the resource requirements. Involving the crowd reduces substantially the time and effort investment of each user in generating content. In summary, Wikiaudia uses human skills that are the best in converting a textual image to both spoken and digitized textual forms, while providing an effective platform for coordinating each user's effort to eventually automatically generate usable audio books along with accurately digitized text files.

In addition, each user is also given rewards proportional to his/her contributions. Points so earned by users on completion of different tasks can be converted to certificates or coupons for on-line shopping.

Wikiaudia is live at http://54.213.4.161:8080/wa and open to anyone for contribution and use.

III. EXISTING SCENARIO AND RELATED WORK

Audio books have been predominantly generated manually. This suffers from drawbacks such as low production rate of the audiobooks since a volunteer has to devote nontrivial amounts of his/her time to complete a book. It could also become monotonous to the listener as the book is recorded in a single voice. Further, there is a higher probability of mistakes in the recording since the entire book is done by the same person. This therefore demands manual validation and editing. Wikiaudia overcomes these drawbacks through human computing and crowd-sourcing: the production rate increases, monotony is broken and even the error rate is expected to decrease.

Another probable solution would be application of optical character recognition (OCR) to images followed by automatic text-to-speech synthesis. This will not provide sufficient accuracy or an acceptable accent in the case of Indian languages. This also requires separate engines to be trained for each language. Wikiaudia overcomes this as it is independent of any language.

Other methods which are proposed in previous research work in the area of manufacturing audio books are:

1. User has to obtain an authorisation key to access a text file on a server, display the content of the same on his desktop computer, produce an audio for it and combine the audio file with sound effects. This method is only for producing personalised audio books [1].
2. An audio book is manufactured by attaching an audio device to the back cover of a book. As the user reads the book, he should also simultaneously record it. This device is needed for each book though it is suggested that the audio device can also be reused [2].

The above two methods produce personalised audio books and they do not mention anything about contributing audio books to the public domain or validating the audio recordings. As such, our target audience - primarily those who are unable to read printed or e-books by themselves - does not benefit from such approaches.

3. Another method to manufacture public-domain books using crowd-sourcing has been implemented in [3]. There are various actors playing administrative roles in this system. For example,

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

some volunteers are assigned as book coordinators to post the book and also collect all the recorded files while meta-coordinators manage the status of an audio book and also its validation. Also, in this domain the volunteer must himself download the book and navigate to an appropriate location to begin recording. In contrast, in Wikiaudia, the user will be allocated the particular content to be processed, requiring no further additional effort from his/her side to locate the content. In addition, the validation of the uploaded audio clip is also crowd-sourced. Thus our method makes the entire process of production of the audiobook crowd-based and efficient.

Traditionally, for digitizing a document, automated methods assisted by optical character recognition technology (OCR), intelligent character recognition technology (ICR) and/or natural handwriting recognition technology (NHR) have been employed. However, these methods are ineffective in cases where the quality of the content, especially digital images, is so poor that OCR/ICR or NHR technology is of little or no assistance. A bigger problem arises in the case of many languages such as those of India for which OCR technology is simply not available and is highly inaccurate where available. Other methods which are proposed include:

1. A process comprising loading a set of definitions for image types and fields, providing a document in electronic form for digitizing from a set of documents, digitizing each record of the document and validating each record of the document [4]. This method requires large amounts of software resources.
2. OCR tasks are done by users. Here, small chunks of documents are sent to the user in a mobile web-based application [5]. The chunk is as small as a word or two since it has to be accommodated in a mobile device's screen. This will take longer to complete a simple document than in our proposed method where each logical chunk is a paragraph or section. Also, our proposed method currently supports as many as 21 Indian languages with features for transliteration from standard keyboard input.

The digitization of old books is a problem that is well-solved by reCAPTCHA. The basic idea of reCAPTCHA is, what is tough on the bots is easy on humans. Though this is the same thought process behind digitization of old books, reCAPTCHA is supported by default only in eight languages and not one of them is an Indian Language [6]. Moreover, the focus and impact of both these software are quite different: while reCAPTCHA is used for providing security, Wikiaudia intends to create books particularly for Indian languages to meet the needs of special audience. Further, reCAPTCHA usually provides a single word to be read out while Wikiaudia provides the user with an image of a much larger chunk of a document to be converted to text.

IV. DESIGN

This section focuses on the design of the generation of crowd-sourced audiobooks and Unicode books. The design of Wikiaudia follows the MTV (Model-Template-View) framework. In this, the view processes the HTTP request from a client to generate an HTTP response. The view gets the data from models and passes it to the template. The template controls how the data is displayed to the user.

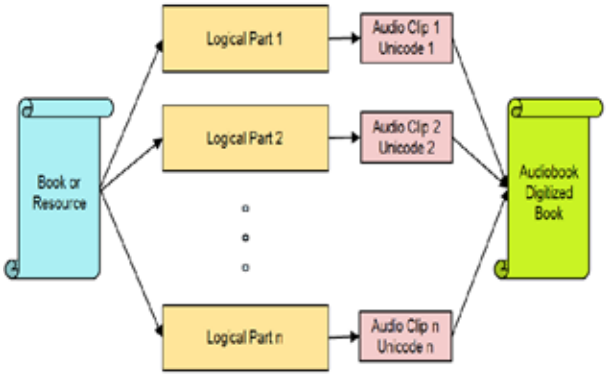

Fig 5. Process of Audio Book Generation in Wikiaudia.

Fig 6. Automatic Chunking of Images along Text Boundaries in Wikiaudia.

The book uploaded by the user is divided into a number of chunks. Each chunk is a logical division of the text. In the context of a book, a logical division is usually a paragraph or two depending on the length of the paragraph. This is done by using an image processing algorithm. This algorithm makes use of Hough lines [7] which detect straight lines in the image of a page of the book. Depending of the

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

length of the lines detected, the image is split into logical divisions which are expected to be meaningful to users – both the ones who read out and the end users who listen to the audio books (see Fig. 5 and Fig. 6). Instead, if each page of the book were to be presented to the user in its entirety, there would be discontinuity in audio across pages, that is, within the same sentence we would have different voices. Apart from reading out a chunk and uploading the audio file, for converting the book into Unicode text, since the volunteer knows how to read and write the language, he/she can use the transliteration tool on our portal to type in the language using a standard QWERTY keyboard.

Wikiaudia also uses a windowing algorithm to allocate chunks to users so that there is continuity in the voice recorded across a sequence of chunks. As a result, on an average, we can expect a minimum of 3-4 chunks of contiguous content read out or typed in by the same user.

Validation of crowd-generated audio and digital text is also crowd-sourced. Even though we can automate the validation of certain aspects of the audio clips such as by checking the loudness and noise levels in the clip, there is no way to detect whether several words in the clip have been mispronounced or the language is not clear and hence not appropriate to be present in the audiobook. This can however be done readily by using the hearing ability and judgment of human volunteers.

The final products – audio books and Unicode texts – also need indexing for navigating to particular chapters or sections. Since Wikiaudia does not make use of OCR, the system cannot figure out by itself whether a chunk marks the beginning of a chapter. Therefore, it relies on user input even for this.

Once all the chunks are processed, that is, the audio clips and/or the Unicode texts are received, the parts are concatenated together automatically. The concatenation is done chapter-wise depending on the inputs received from the user on whether a given chunk is the beginning of a chapter.

V. VALIDATION

Both the audio recording and the Unicode text of the input images of the book produced by crowd-sourcing are subjected to validation through crowd-sourcing. A user can opt to validate a book in one of his/her registered languages of choice. A set of chunks corresponding to that book is displayed to the user. On selecting one of the chunks, the input image of the chunk obtained from image processing, the corresponding audio file and the Unicode text, if available, are shown to the user which he/she can validate by either up-voting or down-voting.

Each chunk is voted by multiple users and when the votes for the chunk exceed a particular threshold, voting results are evaluated. If the ratio between up-votes to down-votes is less than a pre-set minimum acceptable value, the chunk is judged as unusable. If the number of clips which are judged to be unusable exceeds a pre-set threshold, all those clips are automatically sent for re-recording and for re-digitizing. Once these clips are recorded/digitized, the next version of the audiobook/digitized text is released for validation and use. Thus the method ensures high quality and high accuracy of the books released.

VI. RESULTS

The Wikiaudia portal has been hosted on an Amazon cloud server at http://kanoe.org/wa/ Some interesting statistics are as follows:

- 21 Languages supported for audio recording and digitizing;
- Several books uploaded including 5 in English, 2 in Sanskrit and 2 in Kannada languages;
- Sample audiobook produced fully and successfully: Panchatantra Stories, had 23 pages which were automatically divided into 69 chunks of which 66 audio clips are of good quality, 8 different users having contributed and completed the recording in 6 days.
- Each clip recorded by the user is of an average length of 2.5 minutes. To achieve this, a user may have to spend only 5-6 minutes on Wikiaudia.

We believe that Wikiaudia as a crowd-sourcing platform has the ability to contribute significantly to the cause of helping visually impaired people in getting access to books and documents in any language through crowd-sourcing wherein each volunteer user contributes by investing a small amount of time and effort.

VII. APPLICATIONS

Some important applications of Wikiaudia are:

- Mass conversion of out-of-print books into audio books: such books can be brought back to life if a user volunteers to upload a scanned copy of it.
- Translating books written in lesser known languages: this will help in decoding and understanding scripts which have been long forgotten. Say a person owns some manuscript which was written centuries ago and he really doesn't know how to read that script. For example, many palm-leaf manuscripts found in India are written in scripts such as Nandinagari and Grantha for which it is rather difficult to find person

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

who can read them today. Wikiaudia can provide the platform to connect the owner of the manuscript with potential readers to help the community decipher such manuscripts.

- Converting books into different languages: A book can be translated and digitized in other languages which in turn can be made into an audio book. For example, many books written in Tamil language were originally printed in Telugu script when printing was first introduced in India. Such books can be made available in both Telugu and Tamil through an effective use of the Wikiaudia platform.

VIII. FUTURE WORK

1. Currently, the system supports only PDF format for the input images of a book. We intend to implement support for a larger number of types of input images.
2. Mobile application for browsing the manufactured audio books can also be developed. This will help our target audience in using the audio books.
3. More accurate and efficient image processing may be needed in special cases where there are multiple-column texts or the images of pages are significantly warped or where the contrast is poor.

REFERENCES

[1] David Watson, Christopher Coombs, Sean Lewis, Raymond Clark, Andrew Clark, Scott Preston "Apparatus and method for the manufacture of audio books" US Patent Application US20070088712 A1 , 19 Apr 2007.

[2] Yi-Yang Li, "Combination of book with audio device", US Patent US5631883 A, 20 May 1997.

[3] https://librivox.org/, public domain audio books.

[4] Qingfeng Duan, "System for document digitization", US Patent US8457447 B2, Jun 4 2013.

[5] Prayag Narula, Philipp Gutheim, David Rolnitzky, Anand Kulkarni, Bjoern Hartmann, MobileWorks: A Mobile Crowdsourcing Platform for Workers at the Bottom of the Pyramid, in proc. HCOMP 2011.

[6] https://developers.google.com/recaptcha/docs/customization, alternative technique to convert to Unicode.

[7] http://wikipedia.org/wiki/Hough_transform, feature extraction technique in image analysis.

[8] http://www.samarthanam.org/, trust for disabled people in Bangalore, India.

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)