Audio Book Creation System for Indian Languages

Krishna Tulsyan, Vandna Chaturvedi, Aradhana Vinod, Nimisha Srivastava, Ajoy Mondal, and C V Jawahar

Centre for Visual Information Technology,
International Institute of Information Technology, Hyderabad, India
krishna.tulsyan@research.iiit.ac.in,
{vandana44718,aradhana.cvit}@gmail.com, {nimisha.srivastava,
ajoy.mondal,jawahar}@iiit.ac.in

Abstract. We propose a web-based system to convert the scanned document images into digital text and there after to the corresponding audio in Daisy 3.0 format in Indian languages. The system consists of a document image segmentation module, an OCR to extract text out of document images, and a text-to-speech (TTS) module for speech synthesis. Our OCR engine employs a bi-directional LSTM based recognizer with publicly available segmentation components. Our TTS uses a Deep Voice 3.0 implementation for text to audio, in addition to the publicly available systems. The audio book creation system is validated on nearly 200 books containing 15100 pages across four Indian languages. The system is built in a modular manner, with facility for human edits, and easy to extend to many other Indian languages. The proposed system enables user to create and access audio books remotely through a simple web interface. The system archives image, OCR output, and audio of the corresponding text for further improvement in a later stage.

Keywords: OCR · TTS · audiobook · region segmentation.

1 Introduction and Related Work

Assistive technologies help the visually impaired (VI) people to access necessary information like textbooks, newspapers, magazines, and business documents. With the current rate of growth in VI population in India, it is estimated that the country will be home to 115 million VI people by the end of 2050 [1]. As of the current times, India harbors 20 percent of the estimated world VI population [1] making the need for assistive technologies imperative. There exists OCR engines, and TTSs APIs for English and many of the European languages. However, in the case of Indian languages, these are not available. Being not much commercially attractive, this area does not get to attract entrepreneurs. This demands explorations and prototyping from academic researchers. Challenges primarily stem

[1] https://www.tribuneindia.com/news/archive/nation/india-home-to-20-per-cent-of-world-s-visually-impaired-738048

from the lack of reliable OCRs and TTSs for Indian languages. India has as many as 23 languages [2] with other challenges, such as non-standardization of fonts, rendering schemes, and lack of researchers working in this domain. Commercial and reliable OCRs are still far from available for Indian languages. This is an attempt to deploy an academic OCR for a problem of practical importance.

There have been attempts in the past in creating tools and interfaces for VI in Indian languages like [6] and [4]. This paper attempts to address the need of the VI community by presenting a web-based system for robust text recognition and text-to-speech conversion. The developed system facilitates rapid prototyping and experimentation over different components like the OCR system, the text-to-speech system.

2 Our Contribution

Our objective is to develop an interactive system that can generate audio books or audio content in Indian languages. In this regard,

- we demonstrate the practical utility of Indian language OCRs adapted from our earlier work [5] for creating textual content from document images.
- we also demonstrate the Indian language TTS for speech creation in the modern deep learning frameworks based on [9] along with the publicly available TTS implementations.
- a web-based system is developed and demonstrated to convert scanned documents into audio content with the feasibility of (i) manual tagging for unique tags in Daisy (ii) manual overriding in case of failures of any of the automatic image processing or recognition modules.

3 Details of the System

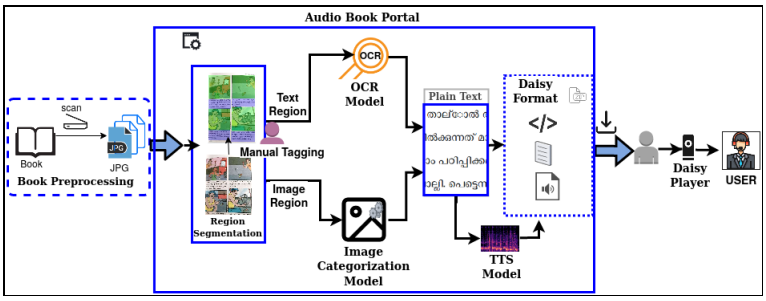

Fig. 1: shows the block diagram of our proposed system. It takes scanned pages as input and performs automatic region segmentation with manual tagging. OCR module recognizes text content and TTS module generates audio content.

The key components of our system are region segmentation, OCR, and TTS modules, as shown in Figure 1.

Region Segmentation Module: The region segmentation module extracts text regions from the document images as a precursor to the recognition module. We do it on top of Tesseract [8]. While our segmentation module is automatic, we also provide a 'user interface to correct any potential errors in the automatic region segmentation results to achieve higher text recognition accuracy and eventually, better audio of the recognized text.

OCR Module: We use bi-directional LSTM based multi-lingual OCR [8] engine to convert the extracted text segments in the document images to plain digital text contents for four Indian languages namely Hindi, English, Telugu, and Malayalam.

TTS Module: We have adopted the NYANKO-BUILD [2] implementation of Deep Voice 3 [7] trained on IndicSpeech [9] text corpus for two Indian languages (Hindi, Telugu and Malayalam) and LJSpeech [3] dataset for English corpus. TTS module takes plain text and produces audio book in DAISY format as an output.

3.1 Experiments

We have tested our system on more than 15,100 pages of the various books in four languages (statistics of the dataset is shown in Table 1). The table below highlights OCR accuracy as more than 85% for word-level for all languages. Figure 2 shows a screenshot of the live demo of our audio book portal for creating Indian language content. OCR is evaluated on a set of ground truth pages, and TTS quality is subjectively assessed.

Language	No. of Books	No. of Pages	WA	User Feedback
English	54	10966	95.88	Needs Indian English accent
Telugu	45	1075	87.71	Good quality and human accent
Hindi	45	400	86.52	Good quality
Malayalam	43	2682	85.30	Human accent

Table 1: Shows the statistics of scanned books used for the experiments, obtained text recognition accuracy, and user feedback by listening to the converted DAISY audio book. **WA:** indicates word accuracy.

4 Conclusions and Future Work

In this paper, we have presented a web-based system that converts raw scanned document pages of the book into Daisy formatted audiobook content much more

[2] https://github.com/r9y9/deepvoice3_pytorch

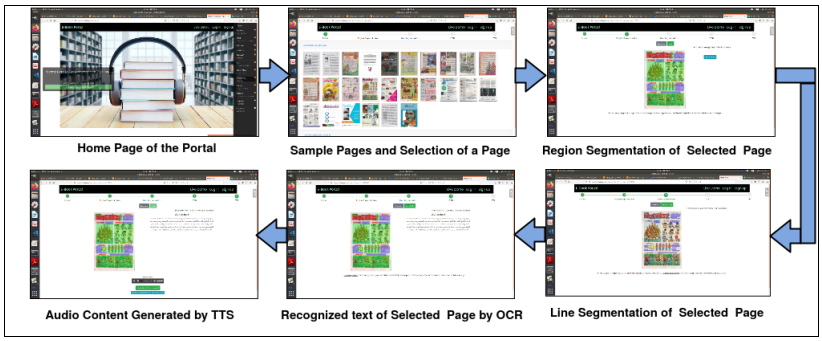

Fig. 2: shows a screenshot of the live demo of our portal for creating Indian language content.

efficiently than previously introduced systems to serve VI students and assist their teachers. The system is a practical solution that can be actively used and experimented upon by the community. The proposed work is deployed at http://bhasha.iiit.ac.in/audiobook/. In the future, we will provide a public API to communicate and enable access to the system. We will also extend the effort to create more plugin modules such as equation, table, and figure understanding modules that will enable rapid experimentation and prototyping.

References

1. Bourne, R.R., Flaxman, S.R., Braithwaite, T., Cicinelli, M.V., Das, A., Jonas, J.B., Keeffe, J., Kempen, J.H., Leasher, J., Limburg, H., et al.: Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. The Lancet Global Health (2017)
2. Chandramouli, C., General, R.: Census of India 2011. Provisional Population Totals. New Delhi: Government of India (2011)
3. Ito, K.: The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/ (2017)
4. Kanvinde, G., Gupta, S.: Accessiblenews daisy: newspapers in daisy. In: ICCWA (2011)
5. Krishnan, P., Sankaran, N., Singh, A.K., Jawahar, C.: Towards a robust OCR system for Indic scripts. In: DAS (2014)
6. Kurian, A.S., Narayan, B., Madasamy, N., Bellur, A., Krishnan, R., Kasthuri, G., Vishwanath, V.M., Prahallad, K., Murthy, H.A.: Indian language screen readers and syllable based festival text-to-speech synthesis system. In: SLPAT (2011)
7. Ping, W., Peng, K., Gibiansky, A., Arik, S.O., Kannan, A., Narang, S., Raiman, J., Miller, J.: Deep voice 3: Scaling text-to-speech with convolutional sequence learning. arXiv (2017)
8. Smith, R.: An overview of the Tesseract OCR engine. ICDAR (2007)
9. Srivastava, N., Mukhopadhyay, R., Prajwal, K., Jawahar, C.: IndicSpeech: Text-to-speech corpus for indian languages. In: LREC (2020)