CantastorIA: Enhancing Audiobook Engagement through Adaptive Soundtracks and Voice Cloning

Victoria Popa*[1,2][0009−0007−0862−1820], Andrea Morelli*[1,3][1111−2222−3333−4444], Christian Di Maio*[1,4][0009−0004−7252−3836], Luca Dini*[1,5][2222−−3333−4444−5555], Cristian Cosci*[1,6][0009−0003−8483−8213], and Emanuele Fulvio Perri*[1][2222−−3333−4444−5555]

[1] University of Pisa
[2] IIT-CNR, Pisa
[3] University of Florence
[4] University of Siena
[5] ILC-CNR, Pisa
[6] University of Bologna

Keywords: Artificial Intelligence · automatic audiobook generation · voice cloning · AI Ethics · Data Privacy

1 Abstract

In recent years, the audiobook industry has grown at a tremendous rate, emerging as one of the fastest-growing audio content genres capable of attracting a wide range of listeners. Ease of use, accessibility and the ability to enhance learning opportunities are some of the advantages of audiobooks that have contributed to their increased appeal. This trend is expected to continue to grow even more in the coming years, and the market is expected to increase. By 2030 audiobooks will generate more than 21.3 percent of global book sales [1].

Research has demonstrated the many benefits of audiobooks in improving young people's access to books, increasing their enjoyment of reading, and promoting emotional intelligence [2]. Audiobooks prove to be a useful learning tool, as they improve reading accuracy and reduce emotional-behavioral problems in adolescents with dyslexia [3], and help blind students [4]. However, audiobooks also present several challenges and limitations. Among the drawbacks, we can point out that engagement and comprehension may be lower than with traditional reading methods because listeners are more prone to distraction, as it is an activity that does not involve visual memory. In addition, many listeners report abandoning an audiobook because they are dissatisfied with the narrator; this fact underscores the significant impact of narration quality on the overall listening experience. The narrative that brings the story to life has the power to either elevate the experience or diminish it by causing listeners to abandon the story itself. Therefore, creating high-level audiobooks requires a lot of effort and resources. In addition, many audiobook versions of a book may not support

* These authors contributed equally to this work

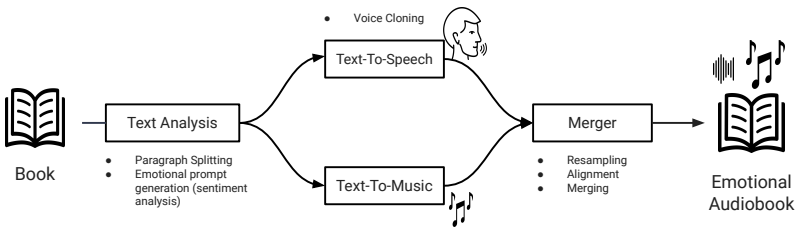

Fig. 1. Implementation Pipeline

certain languages and accents, which would limit the accessibility and variety of content offered.

We propose a model for automatically generating audiobooks with adaptive soundtracks that match the text's emotions, streamlining production and enabling multilingual output. The personalized narrative voice and the proposed background music help to increase engagement and reduce the dropout rate. To improve the concentration and attention mechanisms of the audience, including children and people with attention disorders, we propose a model with emotionally adaptive music. Furthermore, voice cloning allows personalized narration, such as enabling parents with dysphonia to read stories in their own voice.

To the best of our knowledge, existing projects do not incorporate advanced features such as voice cloning and adaptive music generation. The music produced by current audiobook projects often lacks context and variety since each emotion is usually the subject of a stand-alone track. In our model, we introduce and integrate these novel elements to enhance the automatic generation of audiobooks.

Figure 1 outlines the model's implementation pipeline, integrating text analysis, speech synthesis, and music generation for a rich audio experience. We propose "Emotional Audio-Books" that enhance traditional audiobooks with advanced NLP and audio synthesis techniques. CantastorIA aims to transcend conventional audiobook experiences by embedding intelligent soundtracks that boost engagement and satisfaction. The use of artificial intelligence for dynamic audio synthesis and emotion recognition is aimed at producing real-time music and sound effects that match the emotional content of the text. In addition, we propose voice cloning to enable personalized storytelling through increased accessibility and human connection. In the development phase it was employed sophisticated NLP, deep learning algorithms, and high-quality audio standards.

Text Analysis. The system employs advanced sentiment analysis algorithms [5] to detect emotional cues within the text. By examining the narrative, we can identify the key themes and emotional changes that shape the overall mood of the soundtrack. The analysis is then fine-tuned to capture the subtleties of literary expression and to ensure that the emotional impact of the soundtrack matches the atmosphere of the text.

Although sentiment analysis is a valuable tool for assessing emotional content, we recognize its limitations in dealing with complex emotions. Certain sentiments such as sarcasm, irony, and feelings peculiar to a particular context are just a few examples of the nuances and complexity that traditional sentiment analysis algorithms often find difficult to handle. To overcome these limitations and improve the depth of emotional understanding, we propose the integration of large language models (LLM). Recently LLMs, such as those based on transformer [6] architectures, have demonstrated significant promise in capturing the subtleties of human emotions by providing richer, more context-aware analyses. A viable option is represented by LLMs, such as ChatGPT or Gemini, as suggested by Wang et al. [7], to better detect more complex emotional nuances and context-specific emotional changes, thus helping to improve the accuracy and depth of detected emotional cues. In summary, sentiment analysis is the backbone of our text analysis methodology, and the addition of LLM can represent a progressive strategy to overcome some of its drawbacks. This hybrid approach may improve the quality and emotional fidelity of the generated soundtracks, allowing for more complex and nuanced emotional analysis.

Dynamic Soundtrack Generation. Once the emotional tone is determined, a generative audio model synthesizes a soundtrack in line with the mood of the narrative. Two examples of deep learning techniques used by this model to generate high-quality music that dynamically adapts to changes in history are transformers (Transformers) [8] and convolutional neural networks (CNNs) [9]. The purpose of integrating these soundtracks is to create a seamless audio experience that amplifies the emotional resonance and immersive quality of the audiobook without overpowering the spoken content. The latter is especially important to ensure an immersive and undisturbed listening experience.

Voice Cloning. Furthermore, voice cloning technology is incorporated by CantastorIA to add further personalization to the narration [10]. This function allows any voice to be replicated from a small sample, which can be used for audiobook narration quickly and easily. This type of technology is designed to capture the unique timbral qualities and pitch patterns of the source voice, resulting in a natural and engaging style of narration that can be customized to the user's preferences.

Evaluation Methods. Since the project has a strong impact on society, appropriate evaluation metrics must be adopted. In fact, both objective and subjective measures are used to assess the effects and effectiveness of CantastorIA. Technical measures such as spectral analysis and melacestral coefficients are commonly used in objective evaluation to judge the emotional coherence of the produced soundtracks and the naturalness and intelligibility of the cloned voices. To evaluate the sentiment analysis system performance are used benchmarks from publicly accessible datasets such as SST2. Subjective evaluation entails human-centered assessments, such as focus groups and user satisfaction surveys. Under this modality, listeners are the active party, and they are asked to score their experience in terms of liking, emotional impact, and personalization of the audiobook. Evaluating the effect of the soundtrack on listeners' over-

all experience and level of concentration is another important concern. In fact, this represents one of the main problems of audiobooks that our model aims to address. The final aim is to provide a genuinely enhanced and customized audiobook listening experience and creating opportunities for ongoing product improvements tailored to customer needs. Therefore, these evaluation techniques seek to thoroughly assess each aspect of CantastorIA by making sure that not only technical requirements are met but also that customers are engaged.

AI Ethics. In the process of developing CantastorIA, ethical aspects cannot be ignored and must be addressed appropriately. CantastorIA brings up significant ethical concerns, particularly with respect to data privacy, the use of voice cloning technology, and the social implications of AI-generated content.

Data Privacy and Consent. Regarding privacy and consent issues, we recognize the sensitivity of personal data, particularly voice data used for cloning. Hence, all voice cloning are carried out with the express consent of the user, and their data must be safeguarded using strong encryption and anonymization techniques. The use of end-to-end encryption for all data transfers and sophisticated anonymization techniques, such as differential privacy, to prevent the identification of specific individuals from data sets are two of the solutions that can be employed. Moreover, to ensure continued adherence to GDPR [11] and other data protection regulations, we will conduct periodic audits and compliance checks.

Equitable Access. An important ethical aspect concerns accessibility. We aim to make CantastorIA accessible to all users, regardless of their language, disability, or economic status. This includes developing versions for underrepresented languages and ensuring the system is affordable and accessible to economically disadvantaged users. Among the solutions identified, we can mention developing multilingual support using NLP models trained on different language datasets, by providing differentiated subsidies or pricing models to cater to users from different economic backgrounds, and by designing a user-friendly interface that includes accessibility features such as text-to-speech, voice commands, and screen compatibility.

Avoiding Bias. To prevent the perpetuation of stereotypes or offensive content, our voice and music generation algorithms will be regularly audited for biases and other ethics-related issues. Among the solutions identified, we can mention the implementation of bias detection tools to analyze and correct bias in training data and model results, the use of fairness-aware algorithms, and finally, the creation of an ethics advisory board to examine and provide guidance on content generation practices. In addition, a viable solution for identifying and resolving potential biases is to engage users in session groups and collect their feedback.

Transparency. Another key element of our ethical values is transparency. Our commitment is in ensuring transparency regarding how we make use of user data, the functioning of AI models, and the decision-making process within the system. Creating comprehensible and accessible guides and documentation that explain the use of information guidelines, AI model functionality, and decision-making

procedures represent some of the feasible options we identified. Our aim is to address these issues without neglecting aspects of privacy and the use of users' personal data. In addition, implementing explainable AI (XAI) techniques can help users understand how the models generate results, as well as providing a dashboard or portal where users can view and manage their data use consents and preferences.

Environmental and Social Sustainability. With regard to our commitment to environmental and social sustainability, we recognize the significant impact of technology on the environment and are committed to acting responsibly. In developing CantastorIA, we intend to give high priority to environmental sustainability, and coordinate our efforts with the United Nations 2030 Agenda. Using renewable energy for data centers, streamlining our algorithms and infrastructure to reduce energy consumption and carbon emissions, and periodically assessing and disclosing the environmental impact of the system are some possible approaches. Moreover, we firmly believe that AI-generated content must respect human dignity and avoid spreading unfavorable biases or stereotypes in order to ensure social sustainability. Therefore, integrating ethical standards into our content creation process is crucial, and we will continuously monitor to ensure these standards are upheld.

Human Agency and Autonomy. Another critical issue to consider is balancing human intervention and AI support. We aim to use artificial intelligence in CantastorIA to provide insights that give users more control over their end products. Users can easily edit and customize AI-created content, receive comprehensive training and support to help them make the most of AI capabilities, and have the option for the system to provide recommendations and suggestions instead of making decisions on their own.

In conclusion, we took a holistic approach to the development of CantastorIA based on a comprehensive ethical framework that includes data privacy, fair access, bias prevention, transparency, and sustainability. This multifaceted approach, combined with practical solutions for each ethical consideration, ensures the system's trustworthiness, responsible development, and positive societal impact, ultimately fostering an ethical and inclusive AI-powered audio content generation platform.

References

1. Wordsrated, https://wordsrated.com/audiobook-statistics/
2. Best, E. (2020). Audiobooks and Literacy: A Rapid Review of the Literature. A National Literacy Trust Research Report. National Literacy Trust.
3. Milani, A., Lorusso, M. L., & Molteni, M. (2010). The effects of audiobooks on the psychosocial adjustment of pre-adolescents and adolescents with dyslexia. Dyslexia, 16(1), 87-97.
4. Ozgur, A. Z., & Kiray, H. S. (2007). Evaluating Audio Books as Supported Course Materials in Distance Education: The Experiences of the Blind Learners. Online Submission, 6(4).

5. Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
7. Wang, X., Li, X., Yin, Z., Wu, Y., & Liu, J. (2023). Emotional intelligence of large language models. Journal of Pacific Rim Psychology, 17, 18344909231213958.
8. Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., ... & Zeghidour, N. (2023). Audiolm: a language modeling approach to audio generation. IEEE/ACM transactions on audio, speech, and language processing, 31, 2523-2533.
9. Kim, J., Kong, J., & Son, J. (2021, July). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In International Conference on Machine Learning (pp. 5530-5540). PMLR.
10. Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., ... & Zhao, Z. (2023, July). Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In International Conference on Machine Learning (pp. 13916-13932). PMLR.
11. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC