BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

Creating audiobooks with AI

Author: Nurtdinov Timur, Supervisor: doc. Ing. Vítězslav Beran, Ph.D.

Abstract
This system automates the entire process using large language models and neural text-to-speech, delivering a audiobook in minutes.

xnurtd00@stud.fit.vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

[Motivation Audiobook production is expensive and time-consuming. A single title requires a professional narrator, sound studio and hours of work. Modern AI presents an opportunity to automate this process , reducing production time from weeks to minutes.

[Existing solutions] Current text-to-speech systems such as ElevenLabs [1], Azure Neural TTS [2], and Google Cloud TTS [3] produce high-quality speech, but they are black-box narration tools – they accept text and return audio with a single voice, with no awareness of characters, dialogue structure, or narrative context.

No existing solution combines dialogue attribution, character profiling, voice assignment, and audio synthesis into a single end-to-end system that a user can run on an arbitrary book and interactively control the result.

[Our solution] I present system that takes an EPUB book and automatically produces a multi-speaker audiobook. An LLM pipeline attributes dialogue to characters, builds their voice profiles, and assigns a matching neural TTS voice to each one – the entire process requires no manual annotation and completes in minutes. A web interface lets the user review characters, swap voices, and adjust synthesis parameters before and after generation.

[Contributions] The key achievement is a working system validated on few novels, producing coherent character-consistent audio with no manual annotation required. The web interface goes beyond simple generation – it gives the user fine-grained control over every character, voice, and synthesis parameter, making the tool practical for real use.

2. System Overview

The system takes an EPUB file as input and produces a fully narrated multi-speaker audiobook. The processing is organized as a seven-step pipeline, illustrated in Figure 1.

2.1 Book Parsing and Dialogue Extraction

The EPUB file is parsed into chapters and paragraphs using the book's internal structure. Each paragraph is then classified as either *narration* or *dialogue*. Mixed paragraphs – where a line of speech is embedded within a narrative sentence – are automatically split into separate segments so that each segment can later be synthesized with the correct voice.

2.2 Scene Detection

The pipeline divides each chapter into scenes based on location changes. An LLM analyzes the paragraph sequence and identifies scene boundaries, assigning a short location description to each scene. These descriptions are later used to fetch ambient background sounds that play underneath the narration.

2.3 Dialogue Attribution and Character Extraction

For each dialogue segment, an LLM determines which character is speaking, using a sliding context window over the surrounding paragraphs and a running list of known characters. Once all dialogue is attributed, a second LLM pass collects all speech samples per character and builds a detailed profile for each one – including name, gender, age, personality traits, and a natural-language voice description.

2.4 Voice Assignment and Synthesis

The character profiles are matched to available voices from the selected TTS engine by an LLM that compares

voice metadata against character descriptions. Each character is assigned a specific voice along with fine-grained synthesis parameters such as stability and speaking style. The chapter is then synthesized paragraph by paragraph: each segment is sent to the TTS engine with the corresponding voice, and the resulting audio clips are merged into a single MP3 file with context-aware pauses between segments. Per-paragraph timestamps are recorded and stored, enabling the web reader to highlight the current sentence during playback.

3. User Interaction

Although the pipeline runs fully automatically, the user remains in control at every stage through a web interface.

The user can add a book to the system in two ways: by uploading an EPUB file directly, or by searching the Project Gutenberg [4] catalog from within the interface and importing a book with a single click.

After uploading an EPUB file, the user launches the pipeline and monitors its progress step by step. Once processing is complete, the user can review every extracted character and assign a voice from the available TTS voices. For each character the user can tune synthesis parameters such as stability, speaking style, and similarity, allowing fine-grained control over how each character sounds without re-running the full pipeline.

The user can also trigger ambient sound generation – the system fetches background audio clips from Freesound [5] based on the detected scene descriptions and layers them underneath the narration.

Before synthesis, the user selects a narrator style that controls how the narration text is adapted before being sent to the TTS engine.

When satisfied with the setup, the user selects which chapters to synthesize. The synthesis runs in the background and the result becomes available directly in the browser – an interactive reader that plays the audio and highlights the current paragraph in sync, with support for variable playback speed and ambient background sounds.

4. Conclusions

The presented system demonstrates that modern AI for dialogue understanding and neural TTS for speech synthesis – is capable of closing the gap between written content and professional-quality audio narration, making audiobook production accessible without the cost and effort of traditional studio production.

Acknowledgements

I would like to thank my supervisor doc. Ing. Vítězslav Beran, Ph.D. for his genuine interest in this work, his valuable guidance, and the ideas and suggestions that helped shape the final solution.

References

[1] Elevenlabs TTS. https://elevenlabs.io.

[2] Azure neural TTS. https://azure.microsoft.com/en-us/products/ai-services/text-to-speech.

[3] Google cloud TTS. https://cloud.google.com/text-to-speech.

[4] Project gutenberg. https://www.gutenberg.org.

[5] Freesound. https://freesound.org.