# SONUS TEXERE! AUTOMATED DENSE SOUNDTRACK CONSTRUCTION FOR BOOKS USING MOVIE ADAPTATIONS

**Jaidev Shriram** **Makarand Tapaswi*** **Vinoo Alluri***
International Institute of Information Technology, Hyderabad
auto-book-soundtrack.github.io

## ABSTRACT

Reading, much like music listening, is an immersive experience that transports readers while taking them on an emotional journey. Listening to complementary music has the potential to amplify the reading experience, especially when the music is stylistically cohesive and emotionally relevant. In this paper, we propose the first fully automatic method to build a dense soundtrack for books, which can play high-quality instrumental music for the entirety of the reading duration. Our work employs a unique text processing and music weaving pipeline that determines the context and emotional composition of scenes in a chapter. This allows our method to identify and play relevant excerpts from the soundtrack of the book's movie adaptation. By relying on the movie composer's craftsmanship, our book soundtracks include expert-made motifs and other scene-specific musical characteristics. We validate the design decisions of our approach through a perceptual study. Our readers note that the book soundtrack greatly enhanced their reading experience, due to high immersiveness granted via uninterrupted and style-consistent music, and a heightened emotional state attained via high precision emotion and scene context recognition.

## 1. INTRODUCTION

In 1975, a short repetitive piece of orchestral music gained notoriety for inducing fear in its listeners, putting them in a suspenseful state of impending doom. Today, it still induces the same reaction, right when the iconic shark of the movie *Jaws* appears on screen. Movie composers, such as John Williams who created the *Jaws* theme, have long been aware of the impact music has on its listeners. Well-placed music can accentuate scenes to raise the emotional stakes or foreshadow upcoming events; ill-suited music can even suggest that something evil is afoot [1]. When taken as a whole, the musical imagery afforded by such soundtracks greatly complements the movie watching experience. In this paper, we attempt to answer whether we can create a similar experience for books (*cf*. Fig. 1).



**Figure 1**. We aim to transport readers to a musical universe by building a dense and coherent book soundtrack that is borrowed from movie adaptations.

Not unlike movies, reading literature can be an incredibly transportive process that puts one in a meditative state, while actively engaging their imagination and mind [2]. Apart from the lack of visuals, books share many similarities from a music composer's perspective: they have recurring themes, characters (allowing for unique *leitmotifs* [3]), and even long-drawn emotional and narrative arcs that can be teased and reinforced musically. Still, it is uncommon to see soundtracks that are tailored for books; it is up to the reader to curate their own playlist and craft their experience. Inconvenience aside, this requires the reader to preempt the type of music expected for a book's chapter and switch songs at relevant plot points. This is simply infeasible. In this work, we resolve these issues by proposing the first automatic system to build a dense soundtrack that plays throughout the entire reading duration of a book.

Specifically, we focus on building a soundtrack for books with movie adaptations. This allows us to draw on the corresponding movie soundtrack and take advantage of the composer/director's musical intents and instincts for the *same story*. However, adapting the soundtrack is far from trivial due to missing alignment between book parts and music segments. Our approach resolves this by finding scenes in the movie adaptation that match parts of the book and searching for music that plays in that movie scene. This results in high quality, narrative specific matches that amplify the reading experience. We conduct our experiments on *Harry Potter and the Philosopher's Stone*, cho-

sen due to its popularity and Academy Award nominated soundtrack. We evaluate the quality of our soundtrack by asking participants to read two music-accompanied chapters of the book followed by a semi-structured interview.

Our contributions can be summarised as follows: (i) We propose the first fully automatic approach for constructing book-long soundtracks. (ii) Our pipeline matches the story in the book and movie to lift musical cues from the movie adaptation. Gaps in this alignment are filled through an emotion-driven music retrieval system. (iii) A perceptual study of the soundtrack validates our proposed approach and provides further insights into soundtracking for books.

## 2. RELATED WORK

**Soundtracking for books.** Methods for constructing book (or story) soundtracks can be divided into two: generative or retrieval based. Generative approaches typically parse the text for concepts and provide that as input to a music generation pipeline. Topic extraction followed by sentiment analysis [4] or density estimation of emotion-related words [5] are used to generate melodies. On the other hand, retrieval based approaches, mine text for semantic concepts that are used to retrieve ambient music (sometimes with pitch correction) [6] or apply similar ideas to Twitter texts [7]. The general idea of cross-modal text-to-audio retrieval [8] is adapted for tag-based music [9] or characterises documents and music as a distribution over emotions [10]. Recently, Won *et al.* [11] propose a joint emotion-driven embedding space for story sentences and music enabling cross-modal retrieval. However, understanding emotions conveyed in a book depends on the larger narrative requiring a context of more than a few sentences. Thus, while [11] could be used to construct dense book soundtracks in theory, the potential switching of music and resulting change in emotion every few sentences may not result in a good user experience.

Unlike above approaches, we focus on retrieving music from the soundtrack of a book's movie adaptation. Our approach assembles coherent music for books by considering large chunks of the book chapter, aligning them with movie scenes, and finding matched music pairs. We fill in the gaps (only unmatched chapter segments) by relying on emotion-based matching. To the best of our knowledge, we are the first to perform narrative-specific music retrieval and *weave* a soundtrack for the entire book.

**Multi-modal music recommendation** systems use cues such as user location, time, and environmental information to play music in day-to-day life [12] or even traffic conditions for music in cars [13]. Perhaps closest to our work, PICASSO is a fine example of a ranking model trained on pairs of matching music and movie-clips (images, subtitles) that is used to provide music recommendations for image slideshows or audio books [14]. Instead, our work focuses on producing a *dense* book soundtrack by aligning narrative components of the book with the movie adaptation. This results in high precision narrative matches and ensures a coherent soundtrack adapted from the same story.

## 3. ASSEMBLING SOUNDTRACKS FOR BOOKS

A large book may take several (variable) hours for a reader to complete. Within the book, there are multiple chapters, each having sections indicated by the theme, emotion, or location. An ideal soundtrack should respect these scene boundaries and change accordingly. On the music front, while different tracks from the soundtrack are typically homogeneous, they are rarely played in their entirety in the movie at a single stretch. A track may include the score for the setup, conflict, climactic moment, and resolution for a certain storyline, which composers will selectively play at opportune moments to maximise emotional payoffs.

We use the movie adaptation as an intermediary between the two that links plot points from the book with snippets of music from the soundtrack album. We first divide each modality - the book, music soundtrack, and the movie into smaller segments (Sec. 3.1). Then, we obtain clues about the soundtrack associated with each plot point in the book by aligning both the book and soundtrack to the movie (Sec. 3.2). Finally, we weave together the music for the book by combining movie-based and emotion-based matches (Sec. 3.3). Fig. 2 illustrates this overall flow.

### 3.1 Segmenting the Book, Music, and Movie

We discuss methods for segmenting the book chapters $\mathcal{B}_i$ in a book $\mathcal{B} = [\mathcal{B}_1, \ldots, \mathcal{B}_L]$; identifying cohesive musical segments within each track $M_j$ of the album $\mathcal{M} = [M_1, \ldots, M_P]$; and identifying scene boundaries for a movie $\mathcal{V} = [V_1, \ldots, V_Q]$ that facilitates the alignment.

**Book narrative segmentation.** We divide the text in a book chapter based on narrative-relevant factors such as theme, location, activity composition, character constellation, or even time [15]. Due to the absence of large datasets for this task, we adopt an unsupervised approach, temporally weighted hierarchical clustering (TW-FINCH) [16], recently shown to be successful on video activity segmentation. For a chapter $\mathcal{B}_i$, we encode each paragraph using a pretrained language model $\phi_{LM}(\cdot)$ (specifically MPNet [17] that performs well for sentence embeddings and semantic search [18]), and cluster semantically similar and spatially close paragraphs to produce disjoint partitions,

$$\{\mathcal{B}_i^p\}_{p=1}^{K_i} = \text{TW-FINCH}\left(\phi_{LM}(\mathcal{B}_i)\right) . \quad (1)$$

We align individual segments $\mathcal{B}_i^p$ with music segments.

We also considered a few baselines but ignored them due to inferior results. TextTiling [19] tended to uniformly partition chapters while TopicTiling [20] often resulted in over-segmentation. In contrast, our approach yielded segments that mostly respected narrative shifts.

**Keystrength based music segmentation.** As mentioned earlier, the entire track $M_j$ from an album is (almost) never played in a portion of the movie. Hence, similar to book chapters, we focus on creating homogeneous emotionally-cohesive track segments that can be played continuously.

A key feature that music segments are required to reflect are emotions that the director/composer intends to convey. Since we expect homogeneity in music segment

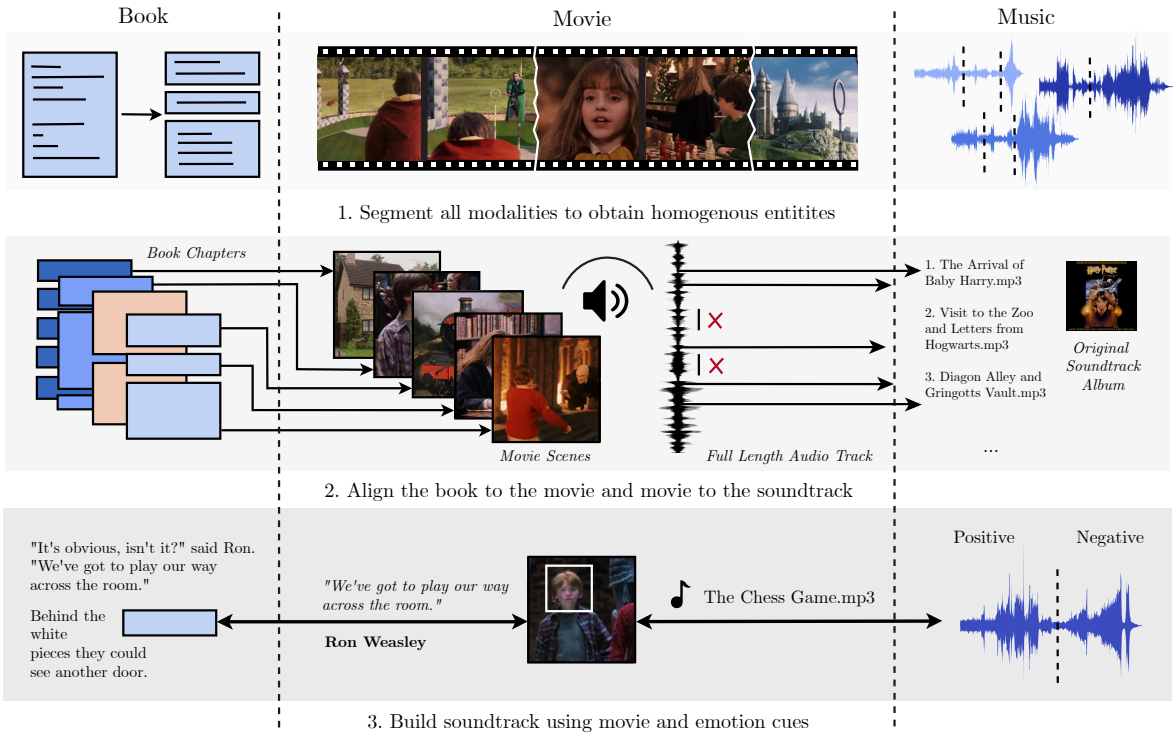


**Figure 2. Overview of approach.** We present a novel technique to build a musically rich book-length soundtrack. We accomplish this by first segmenting the book, its adaptation, and music into smaller homogenous chunks. These segments are then matched with the movie acting as an intermediary for text and music, thereby producing the final soundtrack.

emotions, *keystrength* [21] is a good attribute as it captures the tonal properties of music, and tonal changes likely suggest emotional shifts (*e.g.* major to minor mode suggesting a potential shift from positive to negative valence [22]). *Keystrength* as extracted via the MIRToolbox [23] gives us a probability for each possible key resulting in a 24 dimensional vector (12 major and 12 minor keys). We use a window size of 10 seconds with an 85% overlap, that helps ignore minor local variations in the track while retaining larger shifts. We then compute a self-similarity matrix [24] of the time-varying keystrength vector which is then used to calculate the novelty curve, with a kernel size of 64 [24]. The peaks in the novelty curve determine points of change. Finally, we use those peaks as our segment boundaries to produce $L_j$ music segments $\{M_j^c\}_{c=1}^{L_j}$ (see Fig. 3).

**Movie scene detection.** As a third step, we segment the movie into narrative-coherent scenes $[V_1, \ldots, V_Q]$ using the approach described in [25]. First, the movie is divided into shots (consecutive video frames from the same camera). Then, a dynamic programming algorithm is used to find scene boundaries (a scene consists of multiple shots) so as to maximise intra-scene similarity.

Summarising, we denote $\mathcal{B}_i^p$ as the segment of book chapter $\mathcal{B}_i$; $M_j^c$ as the music segment from track $M_j$; and $V_q$ as the $q^{\text{th}}$ scene from the movie $\mathcal{V}$.

### 3.2 Aligning the Book, Movie, and Music

We first align the book with the movie adaptation using a new two-stage coarse-to-fine alignment scheme. Then, we also align the movie audio to the soundtrack album.

**Chapter-scene coarse alignment.** We first assign a set of scenes from the movie to each book chapter. We use an approach similar to [26] where pairwise similarities are computed between a book chapter $\mathcal{B}_i$ and a video scene $V_q$ via character histograms and matched dialogues. The chapter-scene relationship is then encoded as a graph (each node represents a chapter-scene pair), with edge weights representing similarity scores. Calculating the shortest path over this graph provides the alignment between all chapters and scenes. Additional details are provided in the Appendix.

**Paragraph-scene refinement.** The coarse alignment cannot be used directly for soundtracking as a chapter $\mathcal{B}_i$ contains distinct segments $\mathcal{B}_i^p$ that likely need different music. Thus, in addition to sparse dialogue matches, we compute similarities between sentences in the chapter segment $\mathcal{B}_i^p$ and frames of the video scene $V_q$ using a pretrained vision-language model (CLIP [27]). To improve the quality of CLIP matching scores, we prune dialog and mundane sentences from the chapter segment using a TF·IDF based scoring system. This emphasises relatively rare characters and objects that are likely to give stronger matches than commonly occurring ones. We also retain sentences with a high *concreteness* index [28] that measures how likely a word can be seen or experienced (in contrast to *abstract* words). Finally, we take the top remaining sentences, encode them with CLIP, and calculate cosine similarity with all CLIP encoded video shot frames in the chapter's scenes (see Fig. 4). Scenes with a score higher than $\theta$ are assigned to the text segment using a mapping function,

$$A(\mathcal{B}_i^p) = \{V_q : \text{CLIP}(\mathcal{B}_i^p, V_q) > \theta\}\,. \tag{2}$$

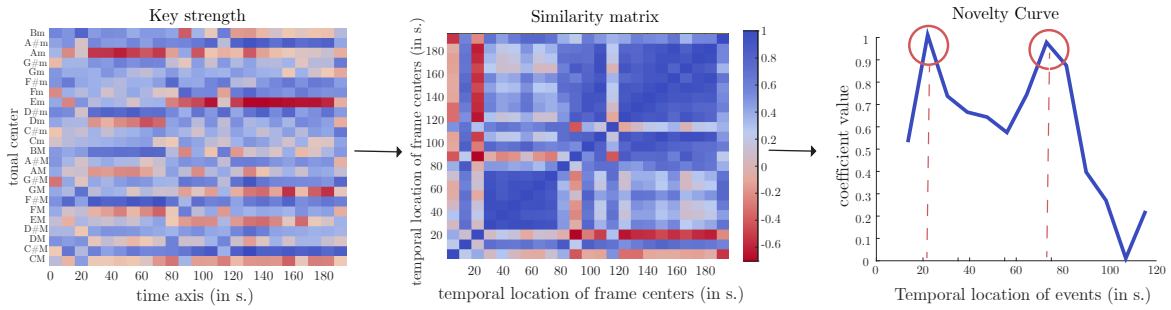


**Figure 3**. **Music segmentation pipeline.** We segment all tracks from the soundtrack to ensure a cohesive listening experience. We extract keystrength [21] that captures tonal properties of a soundtrack (left), compute the self-similarity matrix (center), and use that to calculate the novelty curve [24]. The peaks of this curve are used to segment the track.

**Aligning the movie audio track with the soundtrack album.** We identify the music in the movie by passing its audio track through Shazam [1], a popular commercial audio-fingerprint based search application [29] using its free public API. This results in high-precision estimates for the music played every few seconds. While any audio fingerprinting approach would perhaps work, we found that Shazam was quite accurate in the presence of background noise and dialogue, which is pervasive in movies.

### 3.3 Weaving the Book Soundtrack

Post alignment with the movie, book chapter segments $\mathcal{B}_i^p$ belong to one of two sets: (i) those associated with at least one movie scene $\mathcal{S} = \{\mathcal{B}_i^p : |A(\mathcal{B}_i^p)| \geq 1\}$; and (ii) those without, $\bar{\mathcal{S}} = \{\mathcal{B}_i^p : A(\mathcal{B}_i^p) = \varnothing\}$. Recall, $A(\mathcal{B}_i^p)$ is the set of aligned video scenes to a book segment (*cf*. Eq. 2).

**Extracting emotion labels.** For all text segments $\mathcal{B}_i^p$, we classify each paragraph into positive, neutral, or negative using a BERT-based emotion classifier $\phi_{\text{BERT}}$, trained on Reddit comments [30]. A majority vote across paragraphs is used to assign the emotion label to the chapter segment $E_{\text{book}}(\mathcal{B}_i^p) = \text{mode}(\phi_{\text{BERT}}(\mathcal{B}_i^p))$. For a music segment, similar to [5], we encode it's emotion as valence, $E_{\text{music}}(M_j^c)$, based on the mode of the song (major or minor). This is based on literature that indicates that tracks in minor tend to be associated with negative emotions and tracks in major with positive emotions [22]. We also tried approaches that predict emotion from audio [11, 31], but they didn't work as well for our application.

**Importing music snippets from the video scene.** For every text segment $\mathcal{B}_i^p \in \mathcal{S}$, we extract the movie timestamps corresponding to the matched dialogues or CLIP-based frame-sentence pairs. We use the audio-search to identify the overall track $M_j$ being played at any of the above timestamps in the movie (in a small neighbourhood). A specific music segment $M_j^c$ is chosen by matching emotion predictions *i.e.* $E_{\text{book}}(\mathcal{B}_i^p) = E_{\text{music}}(M_j^c)$ (we pick one randomly if there are several segments). While we can pick emotion-matching music segments without the book-movie alignment, it will likely result in spurious matches.

**Emotion-based retrieval.** For the chapter segments that are not aligned with any video scene, $\bar{\mathcal{S}}$, and those in $\mathcal{S}$ that did not find an emotion compatible soundtrack, we assign a random music segment among the set of emotionally compatible compositions. Note that while we can pick any music segment (even from different movies), we restrict to the soundtracks for this movie maintaining the composer's stylistic coherence.

[1] https://www.shazam.com/

## 4. RESULTS AND EVALUATION

Our approach is designed to be applicable to books that have mostly faithful movie adaptations in terms of few matching dialogues, a relatively similar plot, and at least some matching characters. We evaluate it on the first book and movie pair in the *Harry Potter* series, *Harry Potter and the Philosopher's Stone*.

### 4.1 Harry Potter: A Case Study

Our unsupervised text segmentation approach splits 17 chapters into 87 segments, with an average of 5.11 segments per chapter. Assuming a reading speed of 250 words per minute, each segment requires a ~4 minute track, for a total soundtrack duration of ~6 hours. Music segmentation

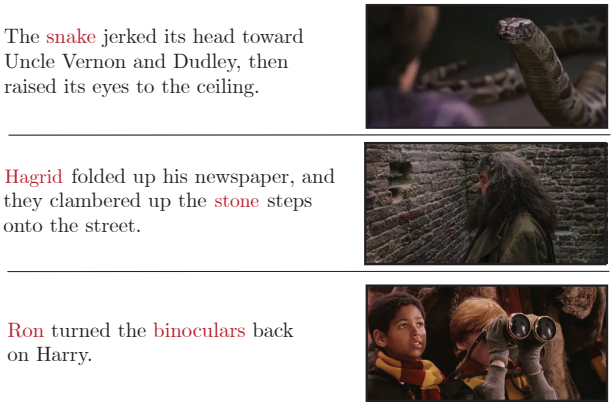


**Figure 4**. **CLIP-based paragraph-scene alignment.** Aligned examples of automatically selected visual sentences and their video frames in the scene using the vision-language model CLIP [27] in a zero-shot setting. Text highlights are words that we think that caused the match.

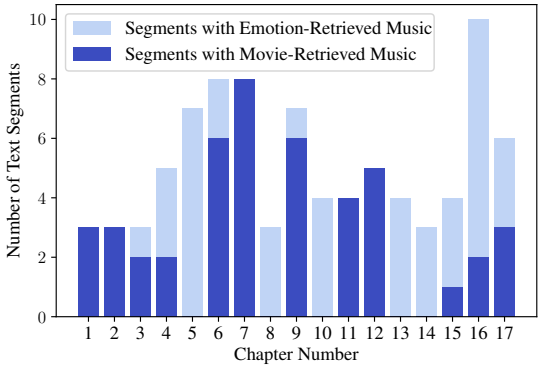


**Figure 5. Number of chapter segments with movie- vs. emotion-retrieved music.** Several chapters in the book are well represented in the movie and are predominantly soundtracked using movie cues. For others, we use an emotion-based retrieval method to build a soundtrack.

of 19 tracks from the soundtrack album results in 47 audio chunks with an average duration of ~52 seconds. For the movie, we obtain 120 scenes using the scene detection algorithm. Computing the book-movie alignment results in a strong chapter-scene alignment, with 82% accuracy, calculated as the percentage of movie shots correctly assigned to a chapter [26].

For the fine-grained alignment, 34 chapter segments have dialogue matches, while the CLIP based refinement results in a 130% improvement in the number of chapter segments associated with the movie; only 9 segments remain unmatched. After performing the final book-music alignment, 45 chapter segments have music imported from a matching video scene through the book-movie alignment. The remainder 42 segments are soundtracked using music segments fetched via emotion matching. See Fig. 5 for a chapter-wise distribution.

### 4.2 Experiment Design

As the primary goal of our method is to improve the reading experience, we obtain user feedback by setting up semi-structured interviews with 10 individuals, who each read two chapters of the *Harry Potter* book. We randomly picked one chapter for each person while the final book chapter was read by all. We chose the final chapter as it represents all aspects of our method: it includes music retrieved via movie cues as well as pure emotion based retrieval, apart from being a key chapter in terms of the plot. This allows us to fairly evaluate the effectiveness of our soundtrack at a book level as reading these chapters takes around an hour for each participant.

All participants are asked to read these chapters consecutively and answer a series of questions about several aspects of the reading experience. No user data is collected through this process and user consent is collected prior to recording the interviews. The study was also authorised by the author's institutional ethics committee.

**Reading application.** A bottleneck with playlist based approaches for soundtracking books is that they require user-input to transition at appropriate instants. We resolve this by creating an application that plays music in the background of the displayed text. Our application loops the music segment infinitely and cross-fades to the next as the reader moves on to the next text segment. This ensures complete immersiveness and lets participants with *varying* reading speeds truly enjoy the soundtrack. For reviewers to have a first-hand experience, we are packaging this application for a few chapters in the supplementary material.

**Participant information.** A total of 10 individuals (18-22 age range, 6 male, 4 female) volunteered for the study. Each participant had previously read the *Harry Potter* book without music and happened to be familiar with the movie as well. We also provided each participant with a small monetary reward as compensation for their time.

### 4.3 Findings

**The soundtrack improved immersiveness.** All participants reported that the soundtrack improved the immersiveness of the reading experience. This is remarkable as all participants noted that they typically do not listen to music while reading. Some participants were more receptive to the soundtrack than others and spoke glowingly about the movie-like immersion afforded by the music. These participants could recognise that the music played was similar to the score at the relevant movie scene. Others, who were positive but less movie-inclined, focused more on how the music *"set the environment"* (P4). Phrases such as *"set the mood"*, *"intensified the emotion"*, *"fit the vibe"*, and *"enhanced the experience"* were common among such participants.

> *"The music provided insight into the tone of the chapter and [...] beyond imagination, provided a soundtrack to what was read"* (P1)

Surprisingly, some participants reported that the first section of the random chapter that they read first threw them off initially, suggesting that there is an adjustment period to this experience. These participants reported that they were very comfortable with later sections (*"after the first 10-12 lines"* - P2) once they had gotten into a reading flow.

**The soundtrack helped visualise the book.** Some participants could recognise that the music played was similar to the score at the relevant movie scene. These participants were especially appreciative of the soundtrack and stated that it helped *"visualise the book, with movie like visuals"* (P1). Such participants typically appreciated the music's cohesiveness and were likely referring to the pleasing soundscape laid out by our soundtrack. One participant explicitly pointed out that the soundtrack helped visualise character interactions and that without, it would have just been *"two characters talking"*.

Many were also receptive to the recurring motifs and themes that appeared throughout, such as the central *Harry Potter* theme and stated that it helped imagine the fictional

world. Only one participant, PX, expressed complete disagreement with the music played in a text segment, for ironically the same reason, stating that the signature *Harry Potter* motif distracted them. Barring this, the same person spoke warmly about the remaining soundtrack, suggesting that the overtness of the motif, which permeates culture today, may be subject to individualistic preferences.

**Music helped focus.**

> *"I get distracted when reading so it helped me focus on certain parts"* (P5)

When describing the immersiveness of the reading experience, three participants specifically pointed out that the music helped them read continuously, without distractions that are typically present when reading in the absence of a soundtrack. It should be noted that few participants who described an aversion for listening to music during typical recreational reading specifically pointed out that they avoid pop music, suggesting that instrumental music, in general, may be better suited for reading purposes.

**Moderate repetition is not a concern.** Most participants noticed repetitive music in some text segments that they read but only brought it up when explicitly asked. One participant suggested playing a different, similarly composed track instead of repeating the music, but in agreement with the rest of the pool, felt that repetition did not affect the experience. These participants also said that the repetition wasn't glaringly obvious and that it did not distract or take away from the immersion. It should be noted that repetition refers to the same track being played on a cross-fade loop while a person reads a text segment, due to variable reading speeds. Despite this, it is noteworthy that our musical segments are homogeneous enough to be played repeatedly and that the text segments are narratively cohesive to warrant such repetition. When specifically asked about the diversity of the music played, all participants expressed positive opinions and noted that the tracks kept changing as and when required.

**Narrative transitions mirrored in music.** Many participants engaged in a conversation with the authors post-interview about how the soundtrack was built. All such participants were startled by the fact that the entire process was completely automatic. Their surprise was primarily due to the fact that the music was automatically matched with relevant areas in text and that it transitioned at appropriate narrative points.

> *"It actually made the experience better as the transition put you in the mood for the expected emotion - from melancholy to sad."* (P9)

When asked specifically about these narrative transitions during the interview, all participants agreed that it was emblematic of emotional or narrative shifts. Since the last chapter was common across all participants, there was strong consensus about the narrative shifts here especially, with participants noting that the music increased in tension as the final hero-villain clash developed and that it eased into more mellow, tender music once it was resolved.

> *"When the tension built in the plot, the music transitioned to match it."* (P1)

Some participants explicitly appreciated the foreshadowing made possible by the music, as suspenseful music at the start of a segment would precede a similar narrative plotline. Others simply elaborated on their earlier comments about the immersiveness and re-emphasised the high congruency between text segments and the matched music.

**Low-arousal music preferred.** We asked participants to describe the emotional suitability of the music, specifically, and in line with prior answers, and received a favourable response. The few critiques that were present revolved around the energy/arousal of the music played at certain instances. Some participants were dissatisfied with segments that were high in arousal, even though the valence matched, as it broke the immersion. They described how this music distracted them from reading and drew too much attention to the track itself. This is likely an important factor to consider for future work on soundtracking books.

**Movie-based music cues stronger than emotion-based cues.** At the end of the interview, participants were asked to describe their favourite and least favoured pieces of music from the two chapters, if any. The best segments were split between music retrieved via movie cues and those retrieved via emotion matching, with the former being more prominent. On the other hand, the least favoured segments, often mentioned only after the author's insistence, were typically emotion-based matches that were described in terms of their neutrality. This finding suggests that our approach can effectively retrieve narrative-specific music and that such a system is perhaps well-suited for book soundtracking, as opposed to pure emotion-based methods.

## 5. CONCLUSION

We presented a novel system to automatically weave a book-length soundtrack, using the music present in the relevant movie adaptation. Perceptual validation of the constructed soundtrack for the first book and movie pair of the *Harry Potter* series provided very positive feedback that validated several proposed design decisions and techniques for text, movie, and music processing. Participants in our perceptual study were particularly receptive to music that was fetched via the movie and uniformly stated that it improved the immersiveness of the reading experience, and even transported them to the fictional world.

**Future work.** While our method has been successful in generating a soundtrack for a book with a movie adaptation, it needs to be modified to make it work on books without adaptations. Future work can potentially use our approach to investigate common trends across books to determine new cross-narrative rules. Our approach can also be extended for human-in-the-loop collaborative soundtrack construction applications, though such a use case is beyond the scope of this work.

## 6. ACKNOWLEDGMENTS

We thank Siddarth Baasri for his valuable input in analysing motifs present in the *Harry Potter* soundtrack and Saravanan Senthil for illustrating Fig. 1.

## 7. REFERENCES

[1] M. G. Boltz, "Musical soundtracks as a schematic influence on the cognitive processing of filmed events," *Music Perception*, vol. 18, pp. 427–454, 2001. 1

[2] M. C. Pennington and R. P. Waxler, *Why reading books still matters: The Power of Literature in Digital Times*. Routledge, 2017. 1

[3] M. Chełkowska-Zacharewicz and M. Paliga, "Music emotions and associations in film music listening: The example of leitmotifs from The Lord of the Rings movies," *Annals of Psychology (Polish)*, 2019. 1

[4] J. Salas, "Generating music from literature using topic extraction and sentiment analysis," *IEEE Potentials*, vol. 37, pp. 15–18, 2018. 2

[5] H. Davis and S. M. Mohammad, "Generating music from literature," in *Workshop on Computational Linguistics for Literature @ European Association of Computational Linguistics (CLfL@EACL)*, 2014. 2, 4

[6] S. Harmon, "Narrative-inspired generation of ambient music," in *International Conference on Computational Creativity (ICCC)*, 2017. 2

[7] M. Thorogood and P. Pasquier, "Computationally created soundscapes with audio metaphor," in *International Conference on Computational Creativity (ICCC)*, 2013. 2

[8] A.-M. Oncescu, A. S. Koepke, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries," in *Interspeech*, 2021. 2

[9] M. Won, S. Oramas, O. Nieto, F. Gouyon, and X. Serra, "Multimodal Metric Learning for Tag-Based Music Retrieval," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. 2

[10] R. Cai, C. Zhang, C. Wang, L. Zhang, and W.-Y. Ma, "MusicSense: Contextual Music Recommendation using Emotional Allocation Modeling," in *ACM International Conference on Multimedia (ACM MM)*, 2007. 2

[11] M. Won, J. Salamon, N. J. Bryan, G. J. Mysore, and X. Serra, "Emotion embedding spaces for matching music to stories," in *International Society for Music Information Retrieval (ISMIR)*, 2021. 2, 4

[12] S. Reddy and J. Mascia, "Lifetrak: Music in Tune with your Life," in *International Workshop on Human-Centered Media (HCM)*, 2006. 2

[13] L. Baltrunas, M. Kaminskas, B. Ludwig, O. Moling, F. Ricci, A. Aydin, K.-H. Lüke, and R. Schwaiger, "InCarMusic: Context-Aware Music Recommendations in a Car," in *International Conference on Electronic Commerce and Web Technologies (EC-Web)*, 2011. 2

[14] A. Stupar and S. Michel, "Picasso: Automated soundtrack suggestion for multi-modal data," in *International Conference on Information and Knowledge Management (CIKM)*, 2011. 2

[15] A. Zehe, L. Konle, L. K. Dümpelmann, E. Gius, A. Hotho, F. Jannidis, L. Kaufmann, M. Krug, F. Puppe, N. Reiter, A. Schreiber, and N. Wiedmer, "Detecting Scenes in Fiction: A new Segmentation Task," in *European Association of Computational Linguistics (EACL)*, 2021. 2, 9

[16] S. Sarfraz, N. Murray, V. Sharma, A. Diba, L. Van Gool, and R. Stiefelhagen, "Temporally-weighted hierarchical clustering for unsupervised action segmentation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 9

[17] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and Permuted pre-training for Language Understanding," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 9

[18] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 2

[19] M. A. Hearst, "Text tiling: Segmenting text into multi-paragraph subtopic passages," *Comput. Linguistics*, vol. 23, pp. 33–64, 1997. 2, 9

[20] M. Riedl and C. Biemann, "Text Segmentation with Topic Models ," *Journal for Language Technology and Computational Linguistics (JLCL)*, vol. 27, no. 47-69, pp. 13–24, 2012. 2, 9, 10

[21] E. Gómez, "Tonal description of polyphonic audio for music content processing," *INFORMS J. Comput.*, vol. 18, pp. 294–304, 2006. 3, 4

[22] P. N. Juslin and J. A. Sloboda, *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press, 2011. 3, 4

[23] O. Lartillot, P. Toiviainen, and T. Eerola, "A matlab toolbox for music information retrieval," in *Data Analysis, Machine Learning and Applications*, 2008. 3

[24] J. Foote and M. L. Cooper, "Media segmentation using self-similarity decomposition," in *IS&T/SPIE Electronic Imaging*, 2003. 3, 4

[25] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, "StoryGraphs: Visualizing Character Interactions as a Timeline," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 3, 9

[26] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, "Book2Movie: Aligning Video Scenes with Book Chapters," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3, 5, 10

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021. 3, 4

[28] M. Brysbaert, A. B. Warriner, and V. Kuperman, "Concreteness ratings for 40 thousand generally known english word lemmas," *Behavior Research Methods*, vol. 46, pp. 904–911, 2014. 3

[29] A. Wang, "An industrial strength audio search algorithm," in *International Society for Music Information Retrieval (ISMIR)*, 2003. 4

[30] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," in *Association of Computational Linguistics (ACL)*, 2020. 4

[31] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *International Society for Music Information Retrieval (ISMIR)*, 2009. 4

# Supplementary Material

We present some additional details on segmentation (Sec. A) and alignment (Sec. B). We also provide additional details of the reading app (Sec. C) and round up the appendix with the questionnaire used in the perceptual study (Sec. D).

## A. SEGMENTING THE BOOK, MOVIE, AND MUSIC

### A.1 Book Segmentation

We segment the book using the hierarchical clustering approach described in Sarfraz *et al*. [16], using the official implementation [2]. We extract our text features using the `sentence-transformers` package, which provides several pretrained models for sentence embeddings. Our chosen model, MPNet [17], ranked first on the leaderboard as of writing, on sentence embedding and semantic search tasks. Specifically, we use the model named `all-mpnet-base-v2`. For our final segments, we use the segments from the third partition, as it avoided the over-segmentation present at lower partitions while ensuring that segments were homogeneous largely. A histogram of word counts per chapter segment can be seen in Fig. 6.

At one point, we also considered annotating the chapters with text segments to evaluate the efficacy of our approach but quickly came to realise that it would be misleading. As described in Zehe *et al*. [15], segment boundaries in fiction can be attributed to many, often overlapping reasons. For instance, a shift in time during a flashback may overlap with a change in location. Further, depending on the differentiating factor used, segments could overlap; a section divided by an emotion may belong to one when considering time alone. As such, some of our preliminary evaluations of this approach, using a few annotated chapters, undersold the efficacy of our approach and reported poor F1/accuracy scores. In reality, we find that our text segments are highly plausible and reflective of topical changes in the text, which was also validated by the participants of the perceptual study. We leave a thorough investigation of fiction text segmentation to future work, as it falls outside the scope of this paper.

We also tried TextTiling [19], using the NLTK implementation [3], but found that it uniformly partitioned the text typically, with every chapter containing 3-5 segments regardless of content (see Fig. 7). In addition, we tried TopicTiling [20] which uses Latent Dirichlet Allocation topic models to segment text, using its official implementation [4]. This produces a series of depth score, indicating the likelihood of having a segment boundary at that instant. As seen in Fig. 8, TopicTiling seemed to over-segment the text, and produced several short segments. Both these methods have also been shown to be perform poorly on a German variant of the fiction segmentation task [15].

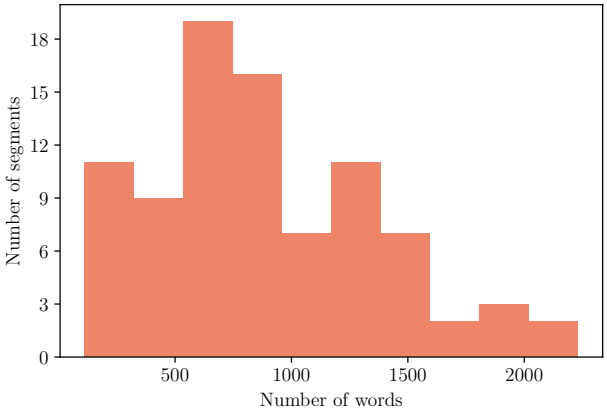


**Figure 6**. Histogram showing the word count distribution for chapter segments.

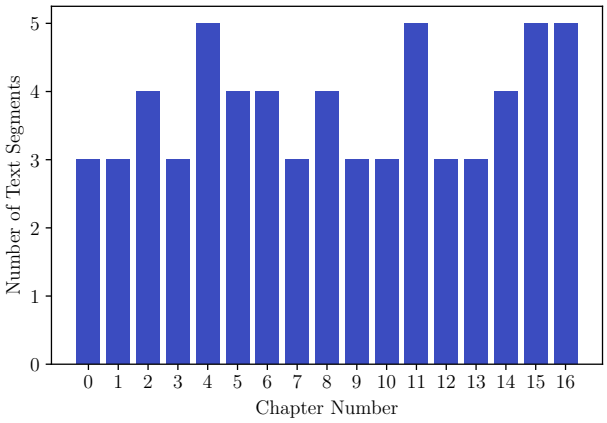


**Figure 7**. Number of chunks per chapter of *Harry Potter* on using TextTiling [19].

### A.2 Movie Segmentation

We segment the movie using the shot threading approach described in Tapaswi *et al*. [25], with its official `MATLAB` implementation [5]. As this approach requires pre-computed video shots, we use an open-source package [6] to obtain shots. For *Harry Potter*, we obtain 2,525 shots. We also tried *PySceneDetect* for shot/scene detection but found that it resulted in over-segmentation, which is undesirable when trying to obtain high-confidence matches between the text and movie.

### A.3 Music Segmentation

Our choice of novelty based segmentation is a reflection of how tracks undergo emotional and tonal shifts. After multiple rounds of testing, we settled on a kernel width of

[2] https://github.com/ssarfraz/FINCH-Clustering/tree/master/TW-FINCH

[3] https://www.nltk.org/_modules/nltk/tokenize/texttiling.html

[4] https://github.com/riedlma/topictiling

[5] https://github.com/makarandtapaswi/Video_ShotThread_SceneDetect/

[6] https://github.com/makarandtapaswi/shotdetection/

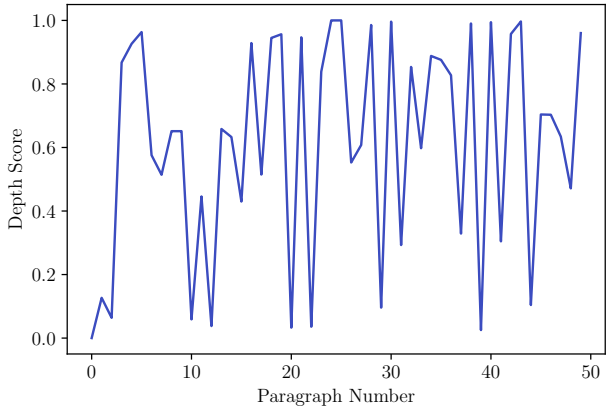


**Figure 8**. Depth score of various segment boundaries, as computed by TopicTiling [20], for a single chapter of *Harry Potter*. Higher values suggest the presence of a boundary.

64 for this, as it produced perceptually cohesive segments that were sufficiently long.

## B. ALIGNING THE BOOK, MOVIE, AND MUSIC

### B.1 Chapter-Scene Coarse Alignment

In order to align book chapters with movie scenes, we largely reproduce the approach described in Tapaswi *et al*. [26]. Their approach parses each chapter and video scene to collect a collection of dialogues and characters, identified by POS tagging on the text/script/subtitles. A similarity score is then computed between the two on two counts - the similarity of character histograms and length of the longest common subsequence (LCS) between movie and book dialogues. We refer the reader to the original paper for more details on its implementation.

Our implementation did however, differ in a few aspects. We align the subtitles with the movie transcript (obtained online [7]) using Dynamic Time Warping with LCS as the distance function, to obtain speaker identities for all dialogues, as the subtitles did not provide this. The original paper used facetracks instead.

To obtain character names, we process the entire book using BookNLP [8], which detects quotes and performs speaker attribution. We only count those characters who have dialogue, unlike in POS Tagging, as they are more likely to be seen in the movie. We use the character names obtained here for our histogram. We automatically match characters names in the movie and book based on LCS, to obtain a common space for our character histogram.

Next, we compute an inverse character frequency to prioritise rare characters who can provide strong alignment cues and scale our character histogram accordingly.

We set $\alpha = 1$ (Eq. 6 in Tapaswi *et al*. [26]) and use equal weights for the character and dialogue similarity scores. The resulting alignment is visualised in Fig. 9.

[7] https://warnerbros.fandom.com/wiki/Harry_Potter_and_the_Philosopher%27s_Stone/Transcript

[8] https://github.com/booknlp/booknlp

### B.2 Movie Audio-Soundtrack Alignment

We use Shazam API [9] in *Python* to identify music from the movie audio. We also tried an open-source implementation of audio fingerprinting and music identification, DejaVu [10], by registering the soundtrack album in a custom database. However, this approach failed to effectively recognise music in the presence of noise, though different parameters may yield better results.

## C. READING APPLICATION

In order to facilitate a smooth reading experience, we built a custom reading app to ensure that readers with *varying* reading speeds could enjoy our soundtrack. We built our reading app using React [11] and used simple state manipulation to change the track played based on scroll position. We dynamically cross-fade songs based on scroll position using the Web Audio API.

A restricted version of this application is provided in the supplementary material. Please refer to the `README` for installation / setup instructions.

## D. STUDY QUESTIONNAIRE

As part of the perceptual study, we ask our volunteers several questions in a semi-structured oral interview:

1. What is your general feedback with what you've read/heard?
2. Compare your reading experience with music to one without.
3. How was the music placed with respect to the text / How aligned is the text and music?
4. Would you say that the the music is repetitive? Describe your thoughts with respect to the diversity of music played as well as how often a single song plays.
5. Different music is played at different points in the book, did this have any effect on your reading experience?
6. How did the music transitions affect your reading experience?
7. How did the music transitions align with changes in text?
8. Did the soundtrack reflect the emotion present in the text? How well did it do?
9. Were there any instances that you particularly liked? Were there any instances that you didn't like?

[9] https://github.com/Numenorean/ShazamAPI

[10] https://github.com/worldveil/dejavu/

[11] https://reactjs.org/

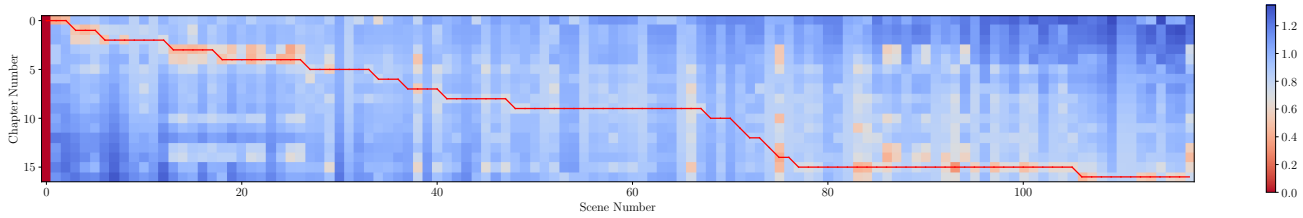


**Figure 9**. Visualisation of the chapter-scene (17 chapters, 120 scenes) similarity scores computed for *Harry Potter*. The red line on the matrix shows the shortest path on the graph, which aligns the book and movie.