PAGES: Enhancing Literary Experience with Ambient Music

Cayden Blake, Grant Lewis, Chase Westhoff and Dan Ventura
Computer Science Department
Brigham Young University
cayden.karl.blake@gmail.com, grant@lewisfamily.org, chasemwesthoff@gmail.com, ventura@cs.byu.edu

Abstract

We present a proof-of-concept system aimed at enhancing the audio book listening experience through the generation of ambient music tailored to tone and content. The system utilizes the Whisper model to extract text from audio data, the LLaMA large language model to analyze that text and extract musical qualities and the MusicGen model to generate music reflecting those qualities. The system captures the feel of the text while creating seamless transitions across tonal shifts in the text, enhancing the integration of music with the narrative tone. The result is an ambient music generation system that dynamically adapts to the content for which it is generating.

Introduction

The integration of ambient music with literary experiences, such as reading books or listening to audiobooks, can significantly enhance the experience of readers and listeners. Theories such as Juslin and Västfjäll's BRECVEMA model (2008) explain how music evokes emotions through brain stem reflexes, evaluative conditioning, and musical expectancy. Effectively evoking the correct emotions at the appropriate times enhances immersion, subtly enriching the narrative without drawing attention away from the text; conversely, inappropriate ambient sounds can disrupt immersion and distract the audience.

Creating ambient music that aligns with the varied tones of long-format books or audiobooks poses substantial creative and financial challenges. Composing hours of music to match extensive narratives would not only be tedious and creatively draining for musicians but also financially impractical for audiobook production companies.[1] To address these challenges, we introduce the PAGES (Progressive Ambient Generative Environmental Soundtracks) system, a proof-of-concept designed to automate the creation of ambient music that dynamically adapts to the tone of literary content.[2]

PAGES utilizes state-of-the-art advancements in artificial intelligence, combining the Whisper API for audio-to-text conversion (Gat et al. 2023), the LLaMA model for textual analysis (Touvron et al. 2023), and the MusicGen model for music generation (Copet et al. 2023). The integration of these technologies into a single system allows for the automatic generation of ambient music that complements the emotional and thematic elements of any audiobook, regardless of length, thereby enhancing the auditory reading experience affordably and efficiently. Through this approach, PAGES offers a new take on multi-sensory literary consumption and opens up new possibilities for creative AI applications in art and entertainment.

PAGES uniquely combines technologies for generating music and processing text to serve the specific needs of audiobook listeners. There are currently no systems of which we are aware that solve this problem. Unlike general-purpose music generation systems, PAGES is tailored for literary works, adapting not only to the emotional cues but also the thematic and narrative structures of the text, setting it apart from existing technologies. In particular, the system is characterized by the following capabilities:

- **Adaptive Music Generation:** automatic generation of music that adapts to the changing emotional and thematic elements of the input content.
- **Contextual Harmony:** generated music complements the literary context, enhancing the listener's emotional and intellectual engagement without overshadowing the primary content.
- **Seamless Transitions:** fluid musical transitions between different sections of text to prevent jarring changes that could disrupt the listener's experience.

Methods

The PAGES system utilizes a combination of AI technologies for extracting text from audio, performing textual analysis and performing musical generation conditioned on a text prompt. Figure 1 illustrates the complete process.

Audio Input to Text Converting spoken language from audio into text is done using Distil Whisper (Gandhi, von Platen, and Rush 2023), in particular, the model *distil-whisper/distil-large-v2* publicly available on HuggingFace. Whisper processes the audio data by transcribing and segmenting it into chunks—in the case of PAGES, between 15

[1] An informal google search suggests a reasonable estimate for music composition is US \$100/minute or more, depending on quality and complexity of the music and experience of the composer. Given that an average audio book is around 10 hours long, costs for augmenting a single book would likely be at least US $\$60,000$.

[2] `https://github.com/chasewes/PAGES`

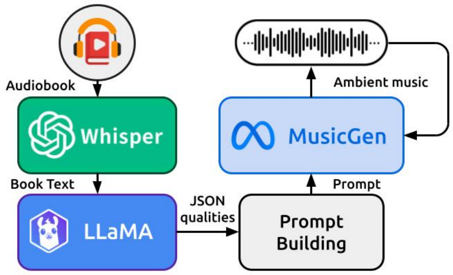

Figure 1: The PAGES Model Pipeline illustrating the flow from audio input through text analysis to music generation.

and 20 seconds in length, allowing the system to capture smaller shifts in emotion and tone across the larger text. This model worked well out-of-the box and was easily integrated.

Text Analysis After the audio has been converted to text it is analyzed to extract mood, tone, thematic elements, and emotional cues using LLaMA2-7b (Touvron et al. 2023). The process involves formatting a query to the LLaMA model that encapsulates the text segment and requests specific musical qualities that would best complement the text, structured as a JSON object. To structure the output of the LLaMA model, primarily to ensure it suggests musical qualities in an accessible and consistent format, the LMFormatEnforcer tool(Gat 2023) was employed. This tool allows the restriction of the model's output to fit a predefined JSON structure by filtering permissible tokens during the generation process. The resulting JSON is used to construct a prompt for guiding the MusicGen model in synthesizing music that reflects the specified attributes, thus prodding the output to be thematically and emotionally synchronized with the narrative content (see Figure 2 for an example).

Music Synthesis The transformer-based MusicGen model is tasked with synthesizing music that is aligned with the prompt generated from the analyzed text. To ensure a seamless auditory experience, MusicGen is also prompted with the last two seconds of the previous music segment. This enables a smooth transition between consecutive music segments, helping maintain immersion and thus augmenting narrative flow and emotional consistency. By overlapping segments in this manner, MusicGen effectively creates a continuous stream of background music.

While the tone and tempo of the music may vary to reflect changes in the narrative's mood and pacing, certain musical attributes such as instrumentation and key are kept consistent throughout a substantial portion of the composition. This consistency prevents the music from becoming jarring or overly conspicuous, which could potentially disrupt the listener's engagement with the text. For instance, if a particular chapter of a book is set in a somber and contemplative environment, the music generated will likely maintain a minor key with string instruments throughout the chapter to complement the setting and tone, even if the intensity or tempo changes from section to section.

While the music should generally flow continuously, there are strategic points within the text where a "musical reset" is appropriate. These points typically occur at major structural breaks in the text, such as new chapters or distinct shifts in the narrative's setting or tone. At these junctures, the MusicGen prompt does not include the previous 2 seconds of music, to facilitate a clean break and better alignment with the new narrative elements and allow the music to adapt to significant changes in the story, enhancing the overall impact and maintaining listener interest.

```
A) Text Input:
"The stormy sea was a raging monster,
unforgiving and brutally cold."

B) Formatted Input to LLaMA:
TEXT: "The stormy sea was a raging
monster, unforgiving and brutally cold."
TASK: In JSON Format, A piece of music
generated as background ambience for the
above text would have these qualities:

C) LLaMA's Generated Continuation:
{
"tone":"dark",
"intensity":"high",
"setting":"nautical",
"tempo":"slow",
"musical_instrument":"string instruments",
"is_major_key":false
}

D) Prompt for Music Generation:
"Ambient Background music with a dark
tone and high intensity, using string
instruments to create a nautical setting.
The piece moves at a slow pace, in a minor
key, evoking an immersive atmosphere."
```

Figure 2: Example evolution of the prompt text as it passes through different stages of the pipeline. The text input (A) is the output of the Whisper system; it is formatted as a query prompt (B) and passed into LLaMa which outputs a JSON object (C) containing musical qualities related to the text; those qualities are inserted into a template prompt (D), which is ultimately used to prompt the MusicGen system.

Integration and Output The PAGES system outputs generated ambient music and input audio as separate tracks, allowing a user to adjust the background music's relative volume, ensuring the music enhances rather than distracts from the narrative experience. The system provides simple controls designed to adjust the music's volume in relation to the audiobook's audio and can be set to automatically remember preferences for future listening sessions.

Results

Contextualized Language-Audio-Pretraining embeddings (CLAP) (Elizalde et al. 2023) are used for evaluating sentiment, textual similarity, and continuity, facilitating a nuanced understanding of emotional resonance, coherence, and alignment of the generated ambient tracks.

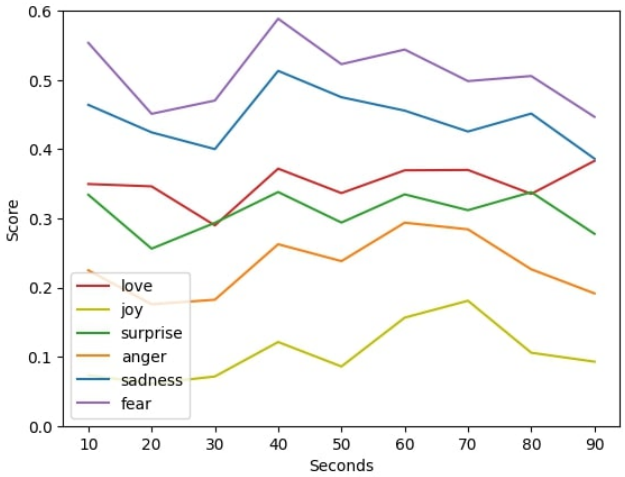

Figure 3: Cosine similarity between primary emotions and music generated by PAGES over time for an excerpt from *The Lord of the Rings*. The word "song" was appended to the text for each emotion before embedding.

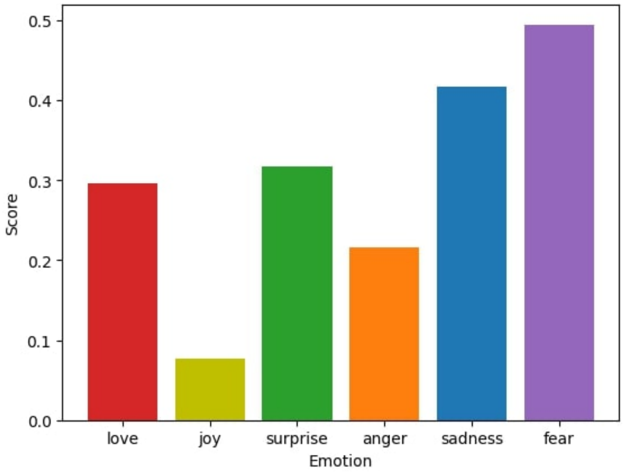

Figure 4: Normalized cosine similarity between primary emotions and the complete generated audio by PAGES for an excerpt from *The Lord of the Rings*.

Sentiment and Tone Analysis To gauge the ability of PAGES to align generated music with the tone and emotional content of input text, the generated music is embedded into a joint text-audio latent space using the CLAP audio embeddings, and measure cosine similarity between the music and a set of primary emotions: love, joy, surprise, anger, sadness, and fear (Parrott 2001). Plotting these similarity scores for multiple sections of the generated audio provides insight into the extent to which the music captures the intended emotional nuances of the narrative, and how that tone evolves over the course of the song.

The generated music was divided into 10-second increments $\iota_1, \ldots, \iota_k$ and embedded each into a latent space using the CLAP audio encoder ϕ: $\phi(\iota_1), \ldots, \phi(\iota_k)$. The name of each of the emotions (with the word "song" appended) was then individually embedded into the same latent space using the CLAP text encoder ψ: $\psi(\text{“love song”}), \ldots, \psi(\text{“fear song”})$, and the cosine similarity δ between each emotion embedding $\psi(e)$ and each audio increment embedding $\phi(\iota_j)$ was computed:

$$\delta(\psi(e), \phi(\iota_j)) = \frac{\psi(e)\phi(\iota_j)}{||\psi(e)||||\phi(\iota_j)||}$$

Figure 3 shows an example of this when generating music to accompany an excerpt from the *Lord of the Rings: The Fellowship of the Ring*. In this particularly intense scene, Gandalf faces the Balrog one would expect fear to be one of the dominant emotions, potentially with some surprise and sadness toward the end.

For comparison, Figure 4 shows the result of embedding the entire example piece as a single point and comparing it to the different emotion embeddings. Note that the cosine similarity scores in this figure are normalized to lie between 0 and 1. As expected, given the source material's content, the most dominant emotions in the generated music are generally sadness, fear, and love, demonstrating that the PAGES model is able to successfully extract some tonal features from the input text.

Textual Similarity To gauge the ability of PAGES to align generated music with the input text itself, the analysis above is repeated using 5 text excerpts of 240 words from fairly popular literary works, including *Cat in the Hat*, *The Fellowship of the Ring*, *Romeo and Juliet*, *Star Wars: Revenge of the Sith*, and *The Way of Kings*. For each excerpt, a 10 second background track was generated, and subsequently embedded (as a single point) into the CLAP latent space. Then, the corresponding text excerpts used as prompts to the PAGES model were embedded as well (again, as a single point), and the cosine similarity scores between each text-audio pair were computed. The results of that comparison can be seen in the similarity matrix in Figure 5.

As can be seen in the figure, the PAGES model does not perform as well on this metric. Only the songs generated from the *Revenge of the Sith* and *Romeo and Juliet* excerpts could be correctly identified. However, *The Fellowship of the Ring* song has a high similarity to its corresponding text, and the similarity between *The Fellowship of the Ring* and *The Way of Kings* is not unreasonable. Further, it is important to recognize the limitations of using this metric to evaluate the model: comparing directly with the text prompt is quite fine-grained and many texts can look similar at this level of analysis; embedding both text and music as single points exacerbates this further by encoding the fine-grained information in a single high-level abstraction; and the CLAP embedding space was not trained for this particular task but rather for the adjacent task of labelling audio samples, and therefore expectations of how the embeddding model generalizes to the more specific case of book text and ambient music should be tempered. A more specialized embedding (if it existed) would likely give a more nuanced and informative viewpoint.

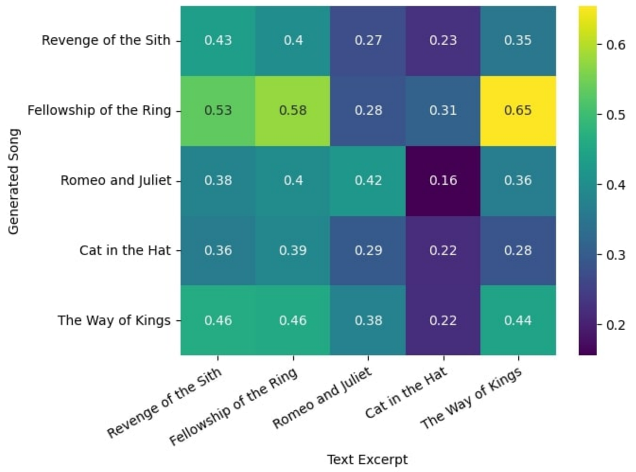

Figure 5: Cosine similarity scores between songs from the PAGES model and the text used to generate those songs.

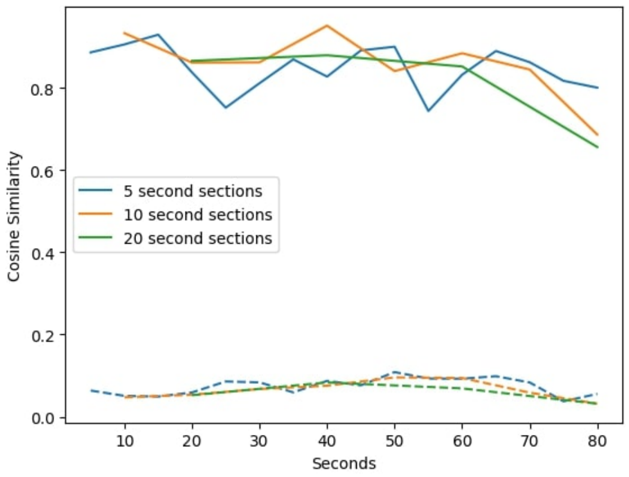

Figure 6: Cosine similarity scores between adjacent sections of a song generated by the PAGES model. Different colors indicate different increment sizes, and the dashed lines are average similarity scores of the music increments compared to 10 random audio samples from the ESC-50 dataset.

Audio Consistency To gauge the ability of PAGES to generate audio exhibiting long-term coherence, the cosine similarity between adjacent musical increments is computed: $\delta(\phi(\iota_j), \phi(\iota_{j+1})), 0 \leq j < k$, expecting to see a high level of similarity between each adjacent section. The similarity scores over time are shown in Figure 6. As expected, there is a high level of similarity between the adjacent sections of audio, indicating that the music generated by the PAGES model remains consistent over time. As a baseline for this metric, 10 random audio clips from the ESC-50 dataset (Piczak 2015) were sampled and compared to each section of the generated music. The average similarity score over time between these audio clips and the music is shown by the dashed lines in Figure 6.

Discussion

The PAGES system represents a significant advancement in the intersection of artificial intelligence and the creative arts. By combining existing models, the system provides a novel solution that adapts ambient music to align with the mood and themes of literary content. The goal of this integration is the enhancement of emotional and thematic immersion, and the creation of a more engaging experience for users. While initial demonstrations and analyses are encouraging, the success of the PAGES project depends on its effectiveness in achieving the following criteria:

- **Technical Robustness:** The system should consistently produce high-quality music that is free from technical defects such as clipping, unnatural transitions, and other audio artifacts. Early results are encouraging.
- **Operational Efficiency:** The generation process should be efficient, with minimal latency and the ability to operate in real-time. Generation is currently 3x real-time.
- **Emotional and Thematic Alignment:** The music generated should accurately reflect and enhance the emotional cues and thematic elements of the text, as validated through user feedback and sentiment analysis metrics. Early results are encouraging.
- **User Satisfaction and Engagement:** The system should improve user engagement, as measured by user retention rates, session times, and qualitative user feedback. This has yet to be evaluated.

The flexibility and effectiveness of PAGES suggest its applicability beyond audiobooks, including enhancement of narrative experiences in movies and video games and personalized auditory environments that adapt in real time to user interactions or narrative shifts.

Looking forward, there are several avenues for enhancing the PAGES system: real-time processing—improving the system's capability to operate in real-time will allow for live adjustments and interactions, making the system more responsive and dynamic; interactivity and personalization—developing interactive features that allow users to customize the music style and thematic elements could significantly increase user engagement and satisfaction; musical themes and motifs—incorporating recurring musical themes or motifs, akin to what a human composer might do, could provide a more cohesive and immersive experience, particularly over longer texts such as novels; model fine-tuning—fine-tuning MusicGen on a broader range of background music could enhance the quality and diversity of music generation; advanced prompting techniques—experimenting with more complex prompting strategies to ensure the generated music consistently matches the narrative context; comprehensive evaluation—conducting extensive evaluations on larger sections of music and across varied genres and books to robustly assess the system's performance and areas for improvement; exploration of alternative music models—testing other music generation models could uncover potential improvements in music quality and system adaptability.

References

[Copet et al. 2023] Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, 47704–47720.

[Elizalde et al. 2023] Elizalde, B.; Deshmukh, S.; Ismail, M. A.; and Wang, H. 2023. CLAP learning audio concepts from natural language supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.

[Gandhi, von Platen, and Rush 2023] Gandhi, S.; von Platen, P.; and Rush, A. M. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv* abs/2311.00430.

[Gat et al. 2023] Gat, N.; Elonen, J.; Fuchs, B.; and Erdem, A. 2023. Lm format enforcer. Available online at `https://github.com/noamgat/lm-format-enforcer`.

[Gat 2023] Gat, N. 2023. LM format enforcer: A tool to enforce output formats for large language models. `https://github.com/noamgat/lm-format-enforcer`.

[Juslin and Västfjäll 2008] Juslin, P. N., and Västfjäll, D. 2008. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences* 31(5):559–621.

[Parrott 2001] Parrott, W. G. 2001. *Emotions in Social Psychology: Essential Readings*. Psychology Press.

[Piczak 2015] Piczak, K. J. 2015. ESC: Dataset for Environmental Sound Classification. `https://www.karolpiczak.com/papers/Piczak2015-ESC-Dataset.pdf`.

[Touvron et al. 2023] Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. Llama: Open and efficient foundation language models. *arXiv:2302.13971*.