Enhancing the Podcast Browsing Experience through Topic Segmentation and Visualization with Generative AI

Jimin Park
parkjimin520@ewhain.net
Ewha Womans University
Seoul, South Korea

Chaerin Lee
crlee@hcil.snu.ac.kr
Seoul National University
Seoul, South Korea

Eunbin Cho
echo45@ewhain.net
Ewha Womans University
Seoul, South Korea

Uran Oh*
uran.oh@ewha.ac.kr
Ewha Womans University
Seoul, South Korea

ABSTRACT

Podcasts present challenges in information retrieval due to their non-visual nature and extended length. To understand these challenges, we conducted interviews with 12 podcast users and identified difficulties in grasping the overall podcast content with metadata alone, highlighting the necessity of navigating to specific segments. Based on this finding, we propose a browsing method that utilizes Large Language Models (LLMs) and image generation models to segment podcast contents, integrating visual cues for supporting efficient navigation. To investigate how this new method differs from conventional approaches and to evaluate its effectiveness, we conducted another user study with 12 participants. The results revealed that keyword search is ineffective when dealing with unfamiliar or inaccurate keywords. Additionally, it requires thorough examination of the script to comprehend the overall content of each episode. On the other hand, segmenting the contents and labeling the topic for each segment facilitated was found to be helpful for understanding of the overall content, enabling easy navigation to desired topics. Furthermore, we found that providing an image enabled participants to easily distinguish one segment from another, which was preferred by participants. This multimodal browsing approach is expected to establish a foundational framework for the effective browsing and comprehension of audio content, extending its applicability beyond podcasts to various forms of audio files.

CCS CONCEPTS

• **Information systems** → Speech / audio search; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

Podcast, Audio Content Browsing, Content Visualization, Generative AI

*The corresponding author.

IMX '24, June 12–14, 2024, Stockholm, Sweden

ACM ISBN 979-8-4007-0503-8/24/06
https://doi.org/10.1145/3639701.3656324

ACM Reference Format:
Jimin Park, Chaerin Lee, Eunbin Cho, and Uran Oh. 2024. Enhancing the Podcast Browsing Experience through Topic Segmentation and Visualization with Generative AI. In *ACM International Conference on Interactive Media Experiences (IMX '24), June 12–14, 2024, Stockholm, Sweden.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3639701.3656324

1 INTRODUCTION

Due to the inherent nature of audio data, which lacks visual cues, the exploration of specific segments within spoken content poses a significant challenge [9, 18]. These challenges are exacerbated for users aiming to seamlessly browse content in long-form audio formats, such as podcasts [5, 11].

To support users for finding information in audio content through browsing, previous studies have primarily focused on accessing information in audio files through keyword searches by converting speech into text [10, 33]. However, even when they know the target keywords, users often refrain from direct keyword searches and prefer a step-by-step browsing method [17, 26] which is found to enhance result understanding and reduces cognitive load of users as it keeps users' awareness of their location while browsing [27]. Furthermore, while podcast creators commonly offer titles or descriptions (i.e., metadata) for episodes, often these details do not accurately reflect the actual content of the podcast [19]. Therefore, there arises a need for a novel approach allowing listeners to understand the overall content before delving into the episode, considering potential disparities between listeners' expectations and the actual content.

With the continuous rise in the popularity of podcasts as a medium for sharing information and entertainment [22, 28], there is an escalating demand for the development of effective podcast-centric methods that enable users to access and browse content seamlessly. In this context, understanding the demands of podcast users is paramount [6]. To address this, we executed an formative interview study targeting podcast users to investigate browsing behaviors of podcast users and identify existing challenges. Our interview results show that while we expected that podcasts are usually listened to without active browsing, we found that podcast users have the desire to locate and revisit specific segments whenever they lost focus as they often multitask while listening to podcasts.

Based on interview results, we propose a segment-based podcast player to support participants with efficient content browsing and navigation. This approach segments the entire content into multiple sections based on the topic of the content and provide a title and representative images for each section using generative AI. The inclusion of images in addition to providing topics for segments is inspired by the idea of using several scene images in video navigation tasks [34], which we expected to ease the content navigation of audio content with this additional visual cue.

To investigate the advantages and limitations of this approach, we conducted a user study with 12 participants with three conditions: (1) Keyword Search, which is a baseline condition, (2) Topic Segmentation, which provides a topic sentence for each segment and (3) Image Segmentation, which provides a representative image along with the topic sentence for each segment. The research questions are as follow:

- *RQ1: What challenges do podcast users encounter during information browsing, if any?*
- *RQ2: What are the benefits and limitations of segmenting podcast episodes based on topics compared to existing keyword based navigation approach?*
- *RQ3: To what extent the integration of visual cues enhance the browsing experience within podcasts?*

To find the answer for RQ1, we conducted a formative interview study with 12 podcast users and found that there is a need for efficient content searching within podcast episodes due to heavy multitasking while listening to podcast. Then we conducted a main user study to answer RQ2 and RQ3 with 12 participants. As a result, participants preferred content segmentation over keyword search, suggesting that segmentation-based approach could potentially improve the effectiveness of podcast navigation since the search keyword can appear multiple times within the same episodes which need to be scanned one by one. Moreover, participants prefer having representative images of each content in addition to its textual topic sentences since images are more intuitive than text, enables users to quickly understand information and distinguish between sections. Our contributions can be summarized as follows:

(1) The identification of the current discomforts in browsing podcast for end-users.
(2) A proposal of topic-based content segmentation and navigation using generative AI.
(3) A design guideline based on an empirical evaluation of the proposed approach of browsing podcast cotents.

2 RELATED WORK

2.1 Challenges in Podcast Information Access

Research on podcasts has predominantly focused on recommendation algorithms [3, 15, 16, 31] or podcast production [2, 21], with insufficient discussion regarding the navigational difficulties stemming from the non-visual nature of audio content, specifically podcasts. In this regard, there is a need for more research to enhance information access in podcasts and facilitate users in finding desired segments within episodes. Audio files, such as podcasts, can be accessed for information by converting them into text using Speech-to-Text (STT). However, Jones et al. [11] indicated that podcasts, with various audio formats and often include transcripts with substantial noise [7, 25], posing difficulties in information retrieval distinct from conventional searches for written text. To underscore these challenges, a word error rate of up to 18% was identified in a large-scale dataset of 100k Spotify podcasts [8]. This highlights the limitations of providing scripts in podcast information retrieval tasks. Additionally, due to the typically long length of podcasts, transcripts can become extensive documents. Podcast listeners must rely on metadata to determine whether investing time in a particular content is worthwhile. This challenge arises because podcasts are not uniformly short, and they often do not provide a clear overview of the content in the beginning. Various studies have explored summarization methods using audio transcripts to support podcast navigation [4, 20, 25, 29]. However, relying solely on audio transcripts not only leads to information loss but also has limitations in understanding the context of content, thus restricting optimization of information access. Summaries of podcast transcripts help listeners decide whether to engage with a particular episode. Nevertheless, an inherent challenge lies in the potential discrepancy between these summaries and the actual content of the podcast [24]. Consequently, Jones et al. advocate for a holistic approach utilizing multimodal strategies in podcast processing. In this context, we aim to enhance users' podcast browsing experience by applying a new approach.

2.2 Long-form Audio Content Browsing

Li et al. [13] emphasized podcasts as examples of long-form audio dialogues, highlighting issues with traditional browsing methods. Exploring lengthy audio content poses challenges in terms of skimming and comprehension. While 2x fast-forwarding can save time, it may compromise understanding and impose cognitive load. Reading scripts can be time-consuming and less engaging. Therefore, Li et al. emphasized the need for a method to swiftly access areas of interest in long-form voice conversations, proposing a structured audio dialogue browsing system using automatic summarization and voice modeling. This system provides recursive summaries when memory limitations of language models prevent summarizing extensive audio documents at once. In this way, the summary acts as an overview, providing clues about the sought-after information, playing a crucial role in deciding whether to stay in the current location or move to another one for further exploration. Research supporting audio file browsing has been conducted in various ways beyond keyword searches and summary utilization [1, 23]. Zhi et al. [35] found that navigating audio by displaying keywords extracted from the script alongside the audio file is intuitive and useful. Moreover, research has been conducted to enhance the user experience of audio content, particularly podcasts. Xia et al. [30] improved the user experience of travel podcasts by proposing Crosscast, which automatically adds visual information. This method uses natural language processing and text mining techniques to identify relevant and visually important entities from podcast text, selecting appropriate images and maps (locations) from online sources. The experiment showed that 85.7% of users prefer podcasts with visual content rather than audio-only. Also, Laban et al. [12] proposed NewsPod, automatically generating a conversational form of news podcasts through advancements in natural language processing

and text-to-speech technology. This divides each news event into segments, and each segment is composed of questions and answers so that listeners can engage dynamically with the content. However, research on effectively accessing information that podcast users want and understanding the overall content is still lacking. We aim to investigate the difficulties faced by podcast users, apply various script-based browsing methods to user studies, and derive podcast browsing solutions by combining the benefits and limitations of each.

3 FORMATIVE STUDY: A SEMI-STRUCTURED INTERVIEW

To identify the challenges and the needs while browsing and searching for specific segments within spoken audio content, we first conducted a semi-structured interview study with frequent podcast users.

3.1 Participants

We recruited participants through the university community and conducted interviews with 12 participants (2 males) with an average age of 29.4 (*SD* = 6.4). Eight participants had been listening to podcasts for over five years, while the remaining four participants (P3, P6, P7, and P11) had been listening to podcasts for over two years. All but three (P4, P6, P12) responded that they listen to podcasts *'while taking public transportation'*, and four answered *'while walking'* (P2, P5, P7, P9). Other responses include *'before bedtime'* (P1, P6), *'doing household chores'* (P4, P5), *'during mealtime'* (P3, P12), and working from home (P8).

3.2 Procedure

The interview sessions were conducted via video conferencing, recorded with participants' consent, and transcribed. Starting with basic questions such as podcast usage behaviors, we presented tasks involving (1) browsing an episode from multiple episodes, and (2) searching for specific information or segments within an episode. To observe participants' behaviours, we also asked participants to find memorable segments from episodes that they recently listened to and locate specific segments within a new episode.

3.3 Results

After conducting an investigation, encompassing episodes of varying lengths (over 40 minutes, under 20 minutes) and diverse listening situations (unheard episodes, previously listened episodes), we observed that users encountered more challenges in browsing longer and unfamiliar episodes. We identified three main categories of participants' browsing needs: (1) direct navigation to a specific segment, (2) being able to skip particular segments, and (3) detailed overview beyond metadata to grasp the content before starting listening.

3.3.1 Users wish to go to specific segment. Participants expressed the need to navigate directly to specific segments. Some participants reported navigating to find sections corresponding to the episode title or podcast creator-provided episode descriptions (P2, P8, P9, P11). Additionally, P3 sought segments featuring guest appearances, while P8 navigated to re-listen to parts related to the topic. Furthermore, participants mentioned difficulty accessing topics when there were unknown or forgotten keywords during searches. Consequently, they tended to search with broader terms when words were difficult to recall (P4, P10). Focused on the quest to find specific segments, we conducted mobile screen recording observations. The functionalities across various podcast platforms displayed a consistent pattern. Participants, in navigating episodes, employed features such as Jump back/next controls, Seek bar, and 2x Fast-forward. For instance, P7 shared an experience from their daily commute, where these tools were employed in seeking out a particular segment of interest. However, despite diligent efforts, the challenge persisted, leading to the discontinuation of podcast listening upon reaching the office.

> *"During my commute, time was tight, so I was using the 15-second rewind button to find that content (real estate investment-related talk). I was looking for it, but the information didn't come up. Then I got to the office, I figured I'd find it later and just turned off the podcast." (P7)*

3.3.2 Users wish to skip specific segment. Participants articulated a desire not solely to find particular segments but also to skip parts such as introductory parts, advertisements, or sections perceived as less relevant (*N*=9, all but P1, P3, P11). They aimed to skip intros, commonly featuring repetitive content such as self-introductions, program explanations, and theme songs. Participants also expressed a wish to skip through advertisements that appear during the episode. Additionally, there were demands to bypass irrelevant personal chatter, personal stories unrelated to the topic, and segments perceived as uninteresting or unnecessary.

> *"For example, there's an intro, there's chit-chat, and it's a bit time-consuming to listen to those, so I think I start listening from when important content comes out." (P4)*

3.3.3 Metadata alone is insufficient to grasp the entire content. Podcast creators provide titles and episode descriptions when uploading episodes. However, participants mentioned the subjective nature of episode descriptions, and many episodes lacked detailed information about the content. Due to the inadequacy of metadata, nine participants found it challenging to grasp the content of episodes (all but P1, P2, P10). Participants typically used metadata to get an overview of the content before listening. With longer episodes, they expressed the need to grasp the content beforehand to avoid disappointment and reluctance to listen to the entire episode. P11 shared discomfort when encountering episodes with only basic information like date and episode number, emphasizing the inconvenience of having to listen to understand the content and decide whether to listen or not.

> *"Apart from the title, I wanted to know more about the content, but it felt like there was a limit. So, I ended up listening to see what this episode was about." (P11)*

3.3.4 Other Comments. Participants were asked about their reasons for using their current podcast platforms. Three participants mentioned that the intuitive user interface of podcast platforms influenced their decision (P2, P8, P12). This insight prompted an investigation into preferred streaming platforms for podcasts. The

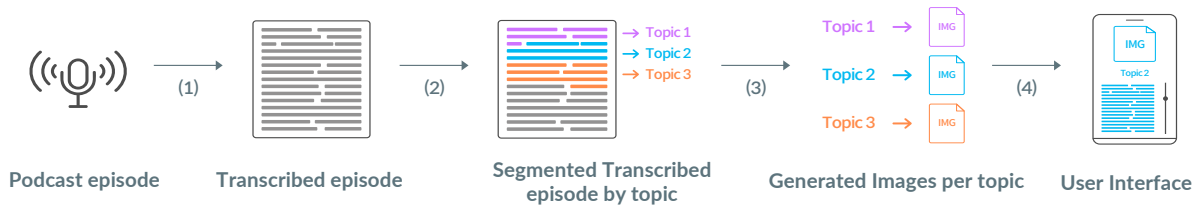

Figure 1: The 4-Step Segmentation Process: (1) STT, (2) Topic Segmentation, (3) Text-to-Image Generation, and (4) Audio Synchronization.

study explored the top ten podcast platforms, including Spotify, Apple Podcasts, and Google Podcasts, based on their popularity in 2023 [32]. As a result, these platforms provided only navigation features such as playback speed adjustment and n-second jump back/next controls. Exploring podcasts with the highest revenue as of 2019 (e.g., The Joe Rogan Experience, My Favorite Murder), none of them provided transcripts or timestamps. However, the availability of transcripts may vary for each program.

4 SEGMENT-BASED PODCAST PLAYER

Based on the interview results that there is a need to explore specific segments and understand the content of podcast episodes, we designed and implemented a prototype that utilizes scripts to segment podcasts and applies images to provide visual cues.

4.1 Segmentation Process

As shown in Figure 1, we divided a podcast episode into topical segments through a 4-step process, applying topics and images to each segment.

- *Step 1: Speech-to-text.* The audio files of podcast episodes were transcribed using CLOVA Note[1]'s Speech-to-Text (STT), with the study focusing on Korean podcasts, necessitating the use of a Korean STT program for accuracy. The resulting transcript includes timestamps and speakers for each sentence.
- *Step 2: Topic Segmentation.* Subsequently, employing the Anthropic's[2] Large Language Model, we divided the script into 6 to 8 segments based on the topics discussed within the episode.
- *Step 3: Image Generation per Topic.* Using the representative topics obtained from the previous step (e.g., for the 'Enjoying London' episode, topics like 'London Market', 'London Gallery', 'London Park' were generated), we utilized the Stable Diffusion's illustrative style image generation model to create images.
- *Step 4: Content Syncing.* Finally, leveraging the timestamps of each segment generated as an output of the Large Language Model, we mapped the topics and images to the podcast audio file and script. This application allows for the incorporation of visual information cues, showcasing the changing topics and images throughout the progression of the episode in the podcast content.

[1]https://clovanote.naver.com/
[2]https://www.anthropic.com/

4.2 User Interface

As shown in Figure 2, the user interface consists of six parts. First, (1) *Keyword Search* enables users to locate specific keywords within podcast scripts. When you search for a keyword, the matching word light up in red, and you can browse to the previous or next keyword using the Next Keyword Button. (2) *Thumbnail View* showcases topics alongside relevant images. This changes based on the current playback time, aligning with the image or topic of that segment. Buttons on either side of the thumbnail facilitate movement to the previous/next segment. (3) *Play Control Buttons* offer play, pause, reset options, along with 15-second skip-forward and skip-backward controls. (4) *Scripts* section entails the podcast episode transcribed through Speech-to-Text, accompanied by timestamps. Clicking on a timestamp enables users to navigate to the corresponding minute and second. (5) *Seek Bar*, arranged vertically, prevents users from covering the script with their hands while navigating. Turning on the Parallel Toggle Button above the Seek Bar synchronizes the Seek Bar's scroll for audio playback control with the script in the Text View. You can activate or deactivate this feature using the toggle button. Lastly, (6) *Touch Panel* at the bottom provides a holistic view of all segments, allowing users to navigate to a specific topic by selecting the corresponding segment. The user interface was implemented using the JAVA language in Android Studio. The implementation was tested on the SM-G981N model, and a user study was conducted using this model.

5 MAIN STUDY: USABILITY TESTING

To understand how to best support audio content browsing and searching, we designed a single-session within-subject study with three conditions: *Keyword Search*, *Topic Segmentation*, and *Image Segmentation* Each participant completed two tasks under all three conditions. The presentation order of these conditions was fully counterbalanced.

5.1 Participants

We recruited the participation of 12 individuals (3 male) with an average age of 27.6 years (SD = 8.3) through the university community. Recruitment targeted those who engage with podcasts for a minimum of 30 minutes per week via mobile platforms. Four participants reported listening to podcasts once a week (P2, P7-P9), and the rest responded twice a week. Additionally, two participants were involved in both the Section 3 interview and the user study, with their involvement deemed non-interfering due to the divergent nature of the data and objectives (P5, P10). The study duration

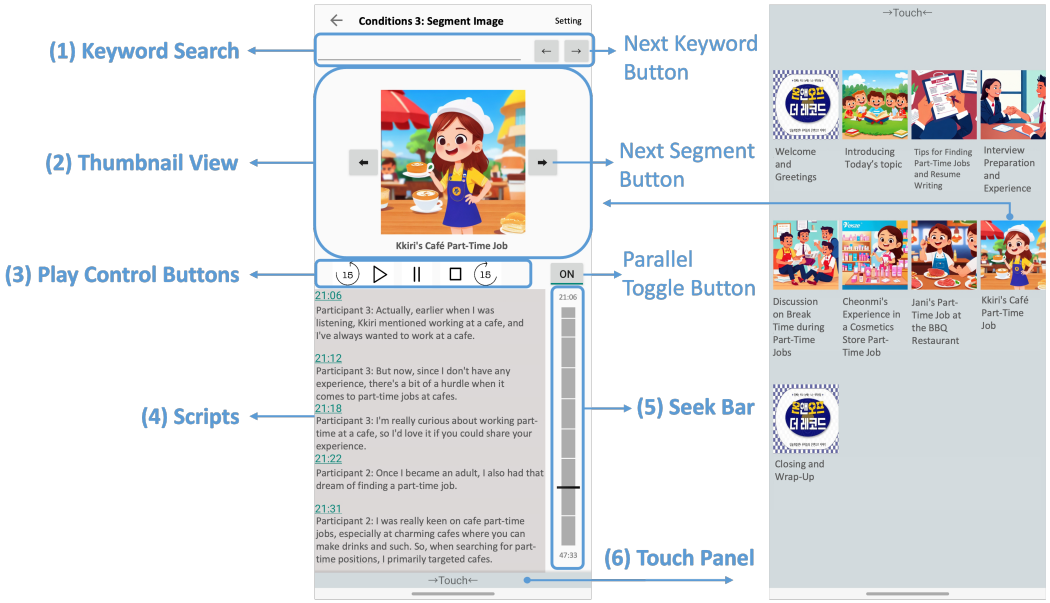

Figure 2: User Interface with six parts. The seek bar shows segmented content by topic, and the thumbnail view shows the title and the representative image of the segmented content that a user is currently listening to.

spanned 50 minutes, with participants receiving a compensation of approximately 10 dollars.

5.2 Conditions

The three conditions we compared in this study were:

- *Keyword Search.* It provides a podcast script with a keyword search function, which is a baseline condition. The content is not segmented by topic and thus it does not provide a title or a representative image in the thumbnail view nor the touch panel.
- *Topic Segmentation.* It segments the script by topics using a Large Language Model and provides topics representing each segment. The topic of each segment is shown in the thumbnail view and the touch panel.
- *Image Segmentation.* It provides images generated using an image generation model in addition to the topic of each segment in the thumbnail view and the touch panel.

In the *Keyword Search* and the *Topic Segmentation*, we utilized original thumbnails from the podcast program to convey the atmosphere and identity of the podcast channel. Therefore, as for the *Topic Segmentation*, the same original thumbnail image was used for all segments throughout a podcast.

Additionally, in the *Image Segmentation*, since segments indicating the beginning and end of each episode were generated, original thumbnails were used at the start and end, while images generated from topics were used in other parts. Moreover, although the *Topic Segmentation* and the *Image Segmentation* include a keyword search function, participants were prohibited from using it during the study.

Recognizing the potential diversity and ambiguity in interpreting images, we opted to present them alongside text. Through pilot testing, we explored the effectiveness of providing either summaries or keywords with images. Summaries, being relative lengthy and challenging to grasp at a glance, disrupted content flow; keywords offered limited information, hindering comprehension. Consequently, we decided to draw inspiration from YouTube podcasts providing chapters and chose to present topics that represent each segment. These topics play a crucial role in emphasizing the core ideas or themes of each segment, facilitating a swift understanding of its content. For these reasons, we concluded that providing topics is the most fitting approach.

5.3 Procedure

Initial interactions involved obtaining fundamental insights through interviews, probing participants regarding their podcast consumption frequency and encountered challenges in exploration. Subsequent to this, we conducted concise practice tutorials for each condition. Tasks 1 and 2 were constrained by a 3-minute time limit, informed by pilot test findings that skimming a 40-50 minute podcast script takes about 3 minutes. To examine how podcast segmentation improves comprehension of content compared to broader skimming, we opted for a 3-minute temporal constraint.

In Task 1, participants were assigned the task of exploring a podcast episode for a duration of 3 minutes under each condition. Their mandate was to articulate the content and overarching themes encapsulated in a singular episode. Participants were provided with illustrative responses derived from the tutorial episode of the pilot test to aid their understanding of the task.

Task 2, on the other hand, entailed participants pinpointing the timestamp when a specified topic, presented by the researcher,

commenced. This task also adhered to a 3-minute temporal boundary. Participants were directed to initiate and respond in tandem with the researcher's cues during each task, with the temporal duration spanning from initiation to response measured. Based on interview results, we divided the task into scenarios with explicit and ambiguous keywords to discern the efficacy of keyword search and segmentation when finding specific sections. In situations where keywords were unclear, we paraphrased words (e.g., changing 'dislike' to 'unfavorable'). Additionally, to prevent a direct correspondence between the topics generated for each segment and the segments participants were asked to find (action items), we also paraphrased them. The discretion of whether to play the recording during exploration was left to the volition of the participants.

The episode data used in Task 1 and Task 2 was meticulously curated to be distinct, and following the completion of all tasks, evaluations for each condition were conducted. Participants were prompted to assess, on a Likert scale ranging from 1 to 7 (Figure 3), the extent to which browsing under the given condition facilitated task resolution, the perceived difficulty of browsing, the alleviation of discomfort compared to their typical browsing behaviors, and their overall satisfaction with the browsing. We implemented a Balanced Latin Square design for the randomization of condition order. Upon the conclusion of all conditions, an interview regarding user experience ensued. Participants were tasked with ranking their preferences for the three conditions and browsing potential use cases where the browsing methods of their favored condition might prove advantageous. Subsequently, insights were gathered on the perceived appropriateness of segmentation and image quality, culminating in the conclusion of the study.

5.4 Apparatus

Podcast users consistently expressed greater difficulty in navigating longer episodes and those they had not encountered before. Consequently, these user opinions were considered during the study design phase. Episodes introducing books, movies, or providing informational content such as history, vocabulary, or news—material likely to be influenced by participants' prior knowledge—were excluded from consideration. Instead, episodes conducted in a conversational format were chosen. Furthermore, with reference to the statistical average podcast length ranging from 30 minutes to one hour [8], episodes falling within the 40 to 50-minute duration were selected. To minimize the likelihood of participants having encountered these programs, we ensured that they were not included in the rankings of popular podcast platforms in Korea[3].

5.4.1 Model used for segmentation. When employing the Large Language Model (LLM) to segment podcasts, we encountered concerns about potential variations in the number of segments across episodes. To mitigate this concern, an empirical approach was undertaken by referencing popular podcasts known to provide chapter information in their descriptions. An analysis was conducted on the three most recent videos from a cohort of 10 YouTube podcasts, each boasting over 1 million subscribers. The resulting average was 6.008, indicating an average chapter length of approximately 6 minutes. Consequently, we set the segmentation criteria for a 40-minute podcast to around 6.66 chapters and for a 50-minute podcast to approximately 8.33 chapters, resulting in segmentation criteria set at 6 to 8 chapters. This criterion was established as a reference, and various segmentation options can be considered based on the actual content and flow of the podcast. For the segmentation process, we chose the Anthropic model, considering both the length of the podcast script and the model's performance as crucial factors. Due to input token limitations, the GPT-4-32K model from OpenAI could handle episodes of about 30 minutes or less. The Llama 2 model was not used due to issues with segmenting that did not align with the episode's content. Therefore, we opted to use the Anthropic model. Input prompt settings:

- Prompt message: Your task is to divide the podcast into 6 to 8 segments by topic.
- User message: Divide the podcast into segments by topic. [Podcast Scripts]

The Temperature parameter was set to 0, considering the interview results, to encompass a broad range of topics when dividing the segments. For the Max Tokens parameter, using the default value of 1000 caused an issue where the segmentation did not extend to the later part of the podcast (Outro and Closing). Therefore, it was set to a maximum of 4000. Additionally, although the model specified the minutes and seconds for each segment when generating the encompassing topic, this information did not align perfectly with the actual script. Consequently, the outputted timestamps were used to adjust the minutes and seconds for each segment.

5.4.2 Model used to generate the image. An issue arose with the DALL.E 2 image generation model when incorporating segment-specific topics through the prompt message, resulting in the generation of images accompanied by meaningless text. Consequently, we opted to forego this image model and instead employed the Stable Diffusion model, chosen for its reduced instances of nonsensical text generation, its flexibility in considering various models fine-tuned in diverse styles and the openness and the ability to run locally unlike DALL.E 2.

In an effort to maintain a semblance of coherence with existing podcast thumbnails, we sought a balance between realistic appearance and stylization. Consequently, the decision was made to utilize the KIDS-ILLUSTRATION-LSH model [4], fine-tuned with illustrative images, for podcast thumbnails. We leveraged segment-specific topics generated using the Anthropic model for the prompt message. We set the num inference steps parameter to 100 and the guidance scale to 10, carefully considered both image quality and processing speed. Additionally, acknowledging the model's relative proficiency in generating content in English compared to other languages, topics in Korean were pre-translated into English before being integrated into the prompt messages.

6 FINDINGS

All participants (N=12) evaluated the three conditions as more helpful than their previously used exploration methods, given that none of them had previously utilized podcasts with provided scripts.

[3]https://www.podbbang.com/

[4]https://huggingface.co/sbarcelona11/KIDS-ILLUSTRATION-LSH

6.1 Task 1: Grasp the Overall Content

6.1.1 Descriptive evaluation. When evaluating the responses submitted by participants in Task 1, a rubric assessment sheet was devised to evaluate them based on the extent to which they described the covered topics. This takes into account how many of the topics within a single episode were articulated. This was used as an indication of how well participants grasped various topics within podcast episodes. The rating criteria was slightly different depending on the number of segments for each podcast episode which varied from 6 to 8. As for the one with 6 segments, for example, the rating criteria were as follows: 5-6 segments were classified as *High*, 3-4 segments as *Middle*, and 1-2 segments as *Low*. As for 8 segments, 7-8 segments were classified as *High*, 4-6 segments as *Middle*, and 1-3 segments as *Low*. *High* indicates a comprehensive understanding of the entire content, while *Middle* indicates an understanding of the main points. *Low* indicates only being able to infer the episode's title.

Except for P3, P4, P10, and P11, participants were rated as *High* for the *Topic Segmentation*. In the *Image Segmentation*, all participants except for P3, P5, P6, and P11, were rated as *High*, with the majority of participants receiving a *High* rating in both conditions. In contrast, for the *Keyword Search*, six participants were rated as *Low* (P1, P2, P4, P7, P8, P12). To be specific, for the *Keyword Search*, we had three for both *High* (P5, P9, P11) and *Middle* (P3, P6, P10), and six for *Low*. For the *Topic Segmentation*, eight of them were rated as *High*, one as *Middle* (P4), and three as *Low* (P3, P10, P11). In the *Image Segmentation*, eight were rated as *High*, three as *Middle* (P3, P5, P6), and one as *Low* (P11).

6.1.2 Effectiveness. As shown in Figure 3: *Question 1*, all participants evaluated how helpful each condition was in understanding the overall content of the episode using a 7-point scale (7 is the best). The results showed that the *Image Segmentation* was rated as the most helpful condition with an average of 6.5 (SD = 0.52), and the *Topic Segmentation* followed closely with an average of 6.41 (SD = 0.51). The *Keyword Search*, with an average of 3.58 (SD = 1.50), was revealed as the least helpful browsing condition. The Friedman test results indicated a significant difference among the three conditions ($\chi^2_{(2,12)} = 16.62$, $p < .01$). Post-hoc pairwise comparisons revealed no significant difference between the *Topic Segmentation* and the *Image Segmentation* (p=.77), and both conditions were significantly higher than the *Keyword Search* ($p < .01$). The *Keyword Search* was identified as the least helpful condition (P1, P4, P6, P10-P12), given the inconvenience of having to skim through the entire script to understand the content. For instance, one participant stated,

> *"Although there is a keyword search, there were parts where I had to look closely at the script to understand the content, so it took more time, I guess." (P6)*

On the contrary, participants expressed a favorable perception of the *Topic Segmentation*, highlighting its utility in comprehending the episode's essence solely based on the presented topics (P4, P7-P10, P12). For example, one commented that,

> *"Since each segment clearly shows the topic, there's a clear sense of how the story unfolds. Even without looking at the details, I can predict how it will develop, so it was very helpful." (P7)*

Likewise, the *Image Segmentation* played a contributory role in facilitating an encompassing understanding of the content. Participants articulated that the incorporation of images offered an intuitive means to grasp the themes embedded within each segment (P1, P3, P6, P8-P10). For instance, P9 said,

> *"Basically, the topics are there, and appropriate images are also there, so even without reading the text, I could easily grasp what this content is about. It was very helpful for understanding the main points." (P9)*

6.2 Task 2: Go to Specific Segments

For Task 2, the information participants were prompted to find was structured to encompass a broad range. This decision was informed by responses from interviews where participants indicated a preference for seeking sections related to broad topics or titles when searching for specific segments (P2, P3, P8-P11).

6.2.1 Task completion time. In Task 2, which involved participants locating specific segments within podcast episodes, the time taken by participants to complete the task was measured. In the sub-task Task 2-1, which involves participants finding segments under accurate keyword indications, the condition with the least average completion time was the *Image Segmentation*, taking an average of 10.66 seconds (SD = 8.58). Following this, the *Topic Segmentation* took an average of 15.5 seconds (SD = 13.33), while the condition with the longest average execution time was the *Keyword Search*, with 24.66 seconds (SD = 24.70). A normality test confirmed that the data did not follow a normal distribution, thus a Kruskal-Wallis test was conducted. The results indicated that the p-value was 0.12, signifying no statistically significant difference among the conditions. Continuing with Task 2-2, in which participants were tasked with locating specific segments in a scenario involving paraphrased keywords, the condition with the least mean completion time was the *Topic Segmentation*, registering an average of 24.75 seconds (SD = 34.36). Following closely, the *Image Segmentation* recorded an average of 25.25 seconds (SD = 32.08), revealing minimal differences. Conversely, the condition with the most prolonged completion time was the *Keyword Search*, averaging 113.41 seconds (SD = 56.08). Given the nonparametric distribution of Task 2-2 results, a Kruskal-Wallis test was conducted, indicating a significant difference with a p-value of 0.0002. Post-hoc pairwise comparisons using the Mann-Whitney U test revealed no significant difference between the *Topic Segmentation* and the *Image Segmentation*. Concerning this aspect, participants articulated that the efficacy of the *keyword search* hinges on the specificity of the provided keywords for pinpointing particular segments. They highlighted that if the keywords are overly broad, the utility diminishes, primarily due to the challenge posed by retrieving multiple keywords. This, in turn, complicates the prompt identification of specific segments (N=10, all but P2, P12).

6.2.2 Effectiveness. In addition, a 7-point scale evaluation was conducted on how helpful each condition was in finding specific segments of the episode (Figure 3: *Question 2*). The condition that received the highest score was the *Topic Segmentation*, with an average of 6.5 (SD = 0.52), indicating it as the most beneficial condition. The *Image Segmentation* had an average of 6.25 (SD = 0.86),

while the *Keyword Search* was rated the lowest with an average of 4 (SD = 1.20), marking it as the least helpful condition in Task 2 as well. The Friedman test revealed a significant difference among the conditions ($\chi^2_{(2,12)}$ = 10.29, $p < .01$). Post-hoc pairwise comparisons indicated no significant difference between the *Topic Segmentation* and the *Image Segmentation* (p=.77). However, there were statistically significant differences between the *Keyword Search* and the other two conditions, notably with the *Topic Segmentation* showing a significantly higher rating compared to the *Keyword Search* ($p < .01$). The *Image Segmentation* also demonstrated a significant difference with $p < .05$. Ten participants except for P1 and P12 emphasized the challenges of the *Keyword Search*, highlighting that important keywords often appear multiple times in the script, making it difficult to find specific segments without accurate knowledge of the keywords. In this context, one expressed the following about the *Keyword Search*:

> *"Important keywords often repeat, and when that happens, I have to check each keyword one by one, making the search not so easy." (P9)*

However, two participants mentioned that in the pursuit of explicit and precise terminologies or information, the *Keyword Search* seems to be helpful (P1, P3).

Regarding the *Topic Segmentation*, participants mentioned that they could quickly access specific segments they were looking for through segments divided by topics (P4, P7, P8, P11, P12). Additionally, P12 stated that the segmentation's length facilitated an understanding of the predominant themes addressed in the episode.

> *"In the keyword search condition, there was no segment division on the progress bar, but it was more convenient when there was. I can easily gauge the duration dedicated to each topic and discern their overall priorities in the episode by observing the length of each segment." (P12)*

The *Image Segmentation* proved beneficial in intuitively distinguishing segments through the utilization of images (P1, P3, P6, P8-P10). Leveraging visual information for clear differentiation among various segments, participants mentioned their adeptness at precisely locating desired sections. For instance,

> *"When browsing, the images displayed prominently for each segment seemed to make differentiation easier. When there's only text, a significant amount of cognitive energy is expended on typographic recognition. However, with images, the differences are more intuitively apparent, which I found preferable." (P10)*

6.3 User Experience Evaluation

6.3.1 Preference. Participants were asked to rank their preferences for the different conditions. The most preferred condition was the *Image Segmentation*, with 6 participants selecting it as their top choice (P3, P5-P7, P9, P10). Following closely, the *Topic Segmentation* was ranked first by 5 participants (P2, P4, P8, P11, P12), while only one participant chose the *Keyword Search* as their top preference (P1). On the other hand, the least preferred condition was the *Keyword Search* for all participants except P1 (N=11) who chose the *Topic Segmentation* as their least preferred, and no participants indicated the *Image Segmentation* as their least preferred. All participants conveyed their intent to incorporate their preferred condition into their regular podcast consumption habits.

For those favoring the *Topic Segmentation*, participants favored its efficacy in providing a swift grasp of the episode's overarching content through segmented topics, consequently streamlining the exploration process. They also expressed that it was helpful in understanding the flow of the episode, allowing them to anticipate upcoming discussions. For those favoring the *Image Segmentation*, the integration of segmented topics with relevant images was underscored as a noteworthy aspect, enhancing intuitive comprehension of the content. Specifically, two participants emphasized that the inclusion of visual components renders a more perspicuous apprehension of the overall atmosphere (e.g., positive or negative) in comparison to the exclusive consumption of textual information (P5, P7).

However, a subset of participants expressed reservations about the concrete utility of images, asserting that the *Image Segmentation* may not provide substantial assistance (P4, P8, P12). According to their viewpoint, the *Topic Segmentation* was deemed more comprehensible.

6.3.2 Difficulty, Satisfaction, Discomfort Alleviation. In addition, as depicted in Figure 3, participants were surveyed on Browsing Difficulty (*Question 3*), Browsing Satisfaction (*Question 4*), and the extent to which discomfort was alleviated compared to their usual browsing methods (*Question 5*). The browsing difficulty for the *Keyword Search* condition received relatively low ratings, averaging 3.58 (SD = 1.08), with satisfaction averaging 3.92 (SD = 0.79). In contrast, the *Topic Segmentation* received ratings of 5.91 (SD = 0.79) for browsing difficulty and 6 (SD = 0.73) for satisfaction. Similarly, the *Image Segmentation* received an average rating of 6.33 (SD = 0.49) for browsing difficulty and matched the *Topic Segmentation* with an average satisfaction rating of 6 (SD = 0.85). The scores for the *Topic Segmentation* and the *Image Segmentation* suggest a prevailing tendency that participants found browsing relatively easy and satisfying. The Friedman test results indicated significant differences in both difficulty ($\chi^2_{(2,12)}$ = 17.37, $p < .01$) and satisfaction ($\chi^2_{(2,12)}$ = 12.5, $p < .01$). According to post-hoc pairwise comparisons, there were no significant differences between the *Topic Segmentation* and the *Image Segmentation* (difficulty: p=.24, satisfaction: p=1). However, both conditions significantly outperformed the *Keyword Search* in terms of difficulty and satisfaction (p <.01 for both).

In contrast, regarding discomfort alleviation, participants shared that discomfort was somewhat alleviated in all conditions compared to their usual podcast platforms, where scripts were not provided. Specific mean results for each condition were the *Keyword Search* 4.66 (SD = 0.88), the *Topic Segmentation* 6.16 (SD = 0.57), and the *Image Segmentation* 6.25 (SD = 0.62). Similar to difficulty and satisfaction, there were significant differences in discomfort alleviation among the conditions ($\chi^2_{(2,12)}$ = 12.5, $p < .01$). Post-hoc pairwise comparisons revealed no significant difference between the *Topic Segmentation* and the *Image Segmentation* (p= 1). However, both conditions significantly differed from the *Keyword Search* (p <.01 for both).

Following the study phase, participants were queried if they had previously listened to the podcast used in the study, and none

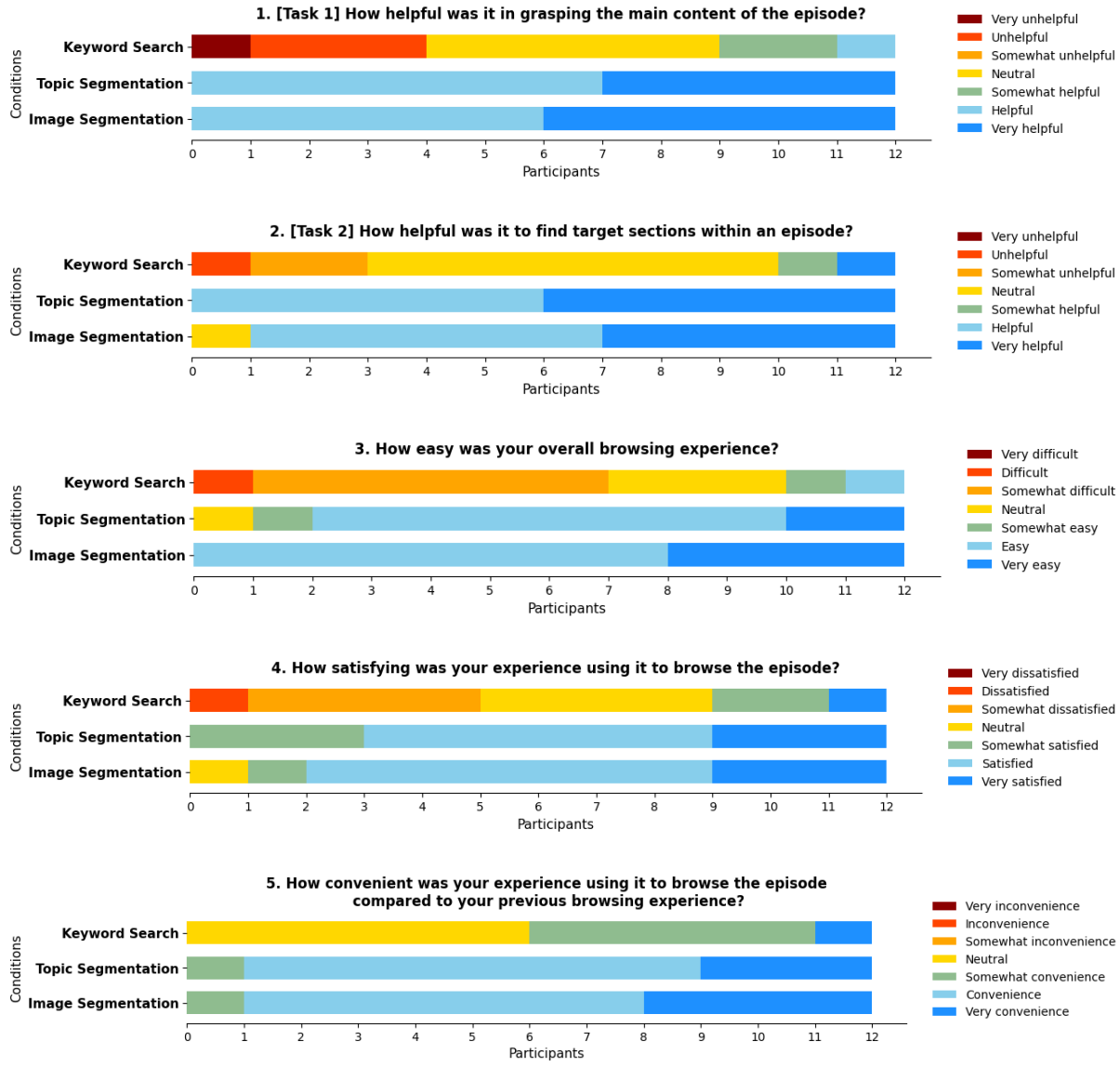

Figure 3: 7-point scale results of subjective ratings

reported prior exposure to it. Furthermore, participants were solicited for insights into the perceived variations in difficulty related to the researcher-requested questions (action items) across different conditions. The findings revealed a unanimous consensus among participants, indicating that they did not discern any disparities in difficulty when responding to the posed inquiries.

6.3.3 Segmentation and Image Quality Evaluation. Segmentation and Image Quality were evaluated using a 7-point scale. The results concerning the clarity of segmented sections in the episode yielded an average score of 6.33 (SD = 0.77). Subsequently, the evaluation of whether the number of segmented sections was appropriate resulted in an average score of 5.75 (SD = 0.75). Participants generally expressed positive evaluations with an average score of 6.5 (SD = 0.52) regarding whether each segmented section effectively encapsulated the content's topic. Finally, the Image Quality assessment, determining whether the representative image for each section aligned well with the content, yielded an average score of 5.5 (SD = 1.08). Participants generally perceived that the segmented sections were clearly divided according to the topics and succinctly encapsulated the content related to the respective themes. Participants attributed this positive assessment to the ease with which they could navigate to the starting point of each topic by tapping on the corresponding segment (P4, P8-P12). Regarding the opinion on the number of sections, participants noted that the segments were appropriately set to be visible within a single page, facilitating

convenient topic identification (P7-P9). However, one participant suggested that detailed distinctions in segmentation did not contribute to a better understanding of the content, proposing that a broader scope for the segments might be beneficial (P12). Some participants expressed that they could not tell if the content is well-segmented or if the number of segment is appropriate as they are unfamiliar with the episodes (P3, P6). Opinions on the alignment of images with content varied among participants. Some participants expressed the view that more specific and detailed images would better convey the content than overly generic visuals (P1, P5). Conversely, other participants asserted that the images appropriately represented the content and aided in understanding the topics (P2, P3, P4, P10, P11).

6.3.4 Use Case. Participants were queried about the anticipated utility of their top-ranked condition in specific scenarios. Regarding the *Topic Segmentation*, one participant articulated that it could facilitate quick decision-making on whether to listen to a podcast or not when there is insufficient information in the title (P12). Furthermore, the division of topics could be advantageous in situations where users specifically desire to listen to certain segments only (P2, P4, P8) or for language learners listening to podcasts to identify and revisit missed sections (P11).

> *"In situations where episode titles or thumbnails don't provide enough details like there's only podcast titles or 'episode 1' whatsoever, having these segmented topics could streamline the decision-making process of whether it's worth investing time in or not." (P12)*

Regarding the *Image Segmentation*, two participants highlighted its potential usefulness in situations where there is limited time to read and quick information retrieval is essential (P3, P7). One participant mentioned that images could be helpful in situations with many speakers (P7), while another participant expressed that it might be beneficial when quickly browsing to understand specific parts by visually identifying them. Moreover, two other participants posited that incorporating images could prove advantageous for individuals with visual impairments (P5, P10). They further suggested that such a feature could be particularly beneficial for young children. One participant said,

> *"For instance, for people with poor eyesight. My eyesight has deteriorated, and text appears a bit hazy. However, with images, in a podcast talking about fruits, as they go from apples to pears and then bananas, you could just click on 'Hey, I wanna hear about bananas' and jump straight to that part. So, it would be helpful in such situations." (P5)*

Lastly, a participant who ranked the *Keyword Search* as their top choice mentioned that the keyword search feature would be useful when sharing a podcast with friends after listening (P1).

7 DISCUSSION

Based on the results of the user study conducted to examine the advantages and limitations under three conditions, we propose an optimal podcast browsing design for users.

7.1 Keyword search is effective for obtaining detailed information

The evaluation results of Task 1 and Task 2, along with the preference outcomes, indicate a disfavor towards the *Keyword Search* in comparison to conditions involving Segmentation, such as the *Topic Segmentation* and the *Image Segmentation.* Although it was anticipated that the *Keyword Search* would excel in swiftly navigating to desired segments, the repetitiveness of core keywords led to a lack of favorability. Participants articulated the challenges encountered in locating desired segments when the search keywords were excessively broad, imprecise, or lacking in specificity. Especially in Task 2-2, the *Keyword Search* took significantly more time to locate specific segments compared to other segmentation conditions, emphasizing these limitations. Furthermore, in the event of multiple instances of identical keywords being queried, a comprehensive grasp of the intended target necessitates an understanding of the contextual information within the transcript, accentuating the necessity of scrutinizing the script. This procedural verification introduces an added layer of intricacy to the navigation process. Moreover, the *Keyword Search* proves challenging in precisely determining the location of the information being explored, making it difficult to pinpoint the starting and ending points of the desired segments. The absence of clear information about where the story begins and ends, or its categorization, accentuates the difficulties in browsing. Additionally, depending solely on the *Keyword Search* to grasp the overall content of an episode presents challenges. While it is feasible to understand the episode's content by skimming the script, the difficulty arises, especially in the context of longer podcasts, where it becomes intricate to capture the content at a glance. This challenge is linked to the observation that there is uncertainty about which keywords to use for the search. The searching process becomes more complex when attempting to locate a specific section without a clear understanding of the keywords, particularly when lacking an overall comprehension of the content. Nonetheless, the *Keyword Search* proves helpful in discerning specific details that defy categorization under overarching themes. Therefore, it is judicious to employ the *Keyword Search* in conjunction with the podcast script. This methodology facilitates access to precise information extending beyond segmented content, cultivating a more profound comprehension of the podcast's substance. This integrative approach proves particularly advantageous in instances where precision is paramount, endeavoring to navigate to information retained from a previous episode.

7.2 Topic segmentation is effective for understanding the overall content

In the case of lengthy podcasts, breaking down the transcript into segments is necessary. Users prefer getting an overview of the content before deciding to listen, highlighting the importance of transcript segment separation. Providing the *Topic Segmentation* allows users to grasp the main points of the episode without reading the entire transcript. This makes it easy to decide whether to listen to the episode without actually diving into it, making the listening process more convenient. This setup also allows users to predict the overall flow of episode content, making it easy to understand what the content is about before listening. Especially, this structure

allows users to predict the overall flow of the episode, making it easy to understand the content before listening to the podcast. It not only simplifies the identification of desired information but also facilitates smoother browsing. Furthermore, facilitating the verification of segment-specific lengths affords the opportunity to preconceive the predominant topics addressed. This feature contributes substantively to an enriched listening experience. The *Topic Segmentation* offers several benefits that simplify and enhance the podcast browsing process. Notably, its standout feature is the easy access it provides to content aligned with specific topics. This makes it a breeze to find the starting point of particular categories, handy for skipping less engaging content like intros and ads. Additionally, assessing the topic flow helps in deciding which parts of the podcast are worth listening to. Furthermore, segmenting narrows down the range of sections users have to navigate, allowing for efficient exploration. This advantage is particularly helpful in overcoming the challenges posed by the interactive nature often found in podcast content. Therefore, the *Topic Segmentation* constitutes a pivotal element in furnishing enhanced user experiences, affording users a positive encounter in browsing and comprehending podcasts more effectively. Particularly pronounced with lengthier episodes, the emphasis shifts from merely furnishing scripts to adopting an approach that categorizes content into segmented topics. However, the *Topic Segmentation* may exhibit limitations when delving into granular details. In cases where a singular segment encompasses multiple topics, there exists a propensity for other subjects to be overlooked. Consequently, for individuals seeking more precise and detailed information, recourse to consulting the transcript or employing the *Keyword search* may prove to be more efficacious. It is imperative to recognize that the advantages of segmenting may vary contingent upon the content and intrinsic nature of the program.

7.3 Image segmentation is more intuitive than text

Despite concerns about the potential interference of image integration in browsing, the *Image Segmentation* exhibited similar levels of satisfaction, preference, and task completion time as the *Topic Segmentation*. Additionally, there was no significant difference in terms of difficulty and alleviation of discomfort between the two methods. Participants who preferred images emphasized the intuitive nature of the *Image Segmentation*, allowing for quick comprehension. Images facilitate easy differentiation between segments compared to text. Moreover, visually representing topics through images aids in quickly identifying desired subjects. This aligns with participants' opinions that the inclusion of images enhances readability. The combination of images and text expedites information understanding and significantly contributes to grasping the overall content. Interestingly, the application of visual cues such as images, which proves advantageous for individuals grappling with content comprehension solely through text. This demographic includes seniors, illiterate individuals, or those afflicted by visual impairments. This design can enhance the accessibility of podcasts.

Moreover, the *Image Segmentation* facilitates a nuanced comprehension of the atmospheric nuances embedded in the content of an episode. In contrast to the sole perusal of textual elements, images possess the capacity to articulate the mood of the content with heightened sensory and vivid expressions. Additionally, in the context of progressively protracted podcast scripts featuring an augmented number of segments, the utility of images as discerning tools is further accentuated. Amid scenarios characterized by extensive scripts and an abundance of segments, wherein information retrieval may be complicated due to the expanded segment count, the employment of images emerges as a valuable asset, facilitating an instinctive demarcation between individual segments. Furthermore, images significantly contribute to enhancing the overall enjoyment of podcasts. While podcasts are primarily consumed for information, participants articulated that the introduction of images confers heightened visual allure, thereby augmenting the pleasurable aspects of topic discovery. The inclusion of images not only furnishes participants with a gratifying exploration experience but is also posited to simplify navigation process compared to the *Topic Segmentation*.

Images afford diverse possibilities in presentation. For instance, when depicting iconic movie scenes, actor images, or visuals from specific locations, the versatility of image utilization becomes evident. This showcases the adaptability of images in various forms and styles, enhancing podcast content. Consequently, the efficacy of *Image Segmentation* depends on individual preferences or the nature of the podcast program. Additionally, images may inadvertently disrupt focus on the main topic, potentially complicating the swift navigation through desired information.

7.4 Limitations and Future work

While the study aimed to evaluate Segmentation and Image Quality among participants, the appropriateness of segmenting content can vary based on the nature and characteristics of the program. Users may seek different information in podcasts (e.g., interesting segments, previously listened sections), and some might browse without specific goals. Users can also employ various functions, engaging in complex navigation processes across multiple steps. The single-session user study, while prompting participants to find specific segments, may not precisely align with everyday usage patterns. Moreover, while we tried to set the style of images identical across segments within an episode by using a fine-tuned model with illustrated images, it would be ideal to guarantee the coherence of the images over the duration of an episode as in prior work [14]. Additionally, since participants were already familiar with the functionality of keyword search, participants may felt it easier to use this, which makes direct comparison not possible. Moreover, while statistical results were presented, a sample size of 12 may not be sufficient. We intend to address this limitation in future work by conducting the study with a larger participant pool, aiming to achieve more statistically significant results. In addition, exploring whether the findings can be generalized to individuals who do not listen to podcasts would be of interest.

8 CONCLUSION

This paper proposes an optimal podcast browsing design by comparing various script-based browsing methods and investigating their respective advantages and limitations, based on the results of

interviews with podcast users. Our evaluation shows that segmenting podcast scripts assists in comprehending the overall content of episodes and navigating to target points. Contrary to expectations, keyword search proved less effective in reaching specific targets, which is stemmed both from the prospect of keywords being dispersed across various segments of the script and the requirement for a meticulous understanding of the script. Furthermore, the integration of images heightened the intuitiveness of each segment and conveyed a mood conducive to effective browsing. This multimodal browsing approach not only holds the promise of enhancing the podcast user experience but also exhibits versatility in its application to diverse audio files. Moreover, we can offer various browsing methods through different combinations. As a consequence, users are anticipated to proficiently comprehend content and expeditiously browse essential segments across a spectrum of situations.

ACKNOWLEDGMENTS

This research was supported by the Basic Research Lab Program through the National Research Foundation of Korea(NRF-2021R1A4A1032582) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2022-00155966, Artificial Intelligence Convergence Innovation Human Resources Development (Ewha Womans University)).

REFERENCES

[1] Fahmi Abdulhamid and Stuart Marshall. 2013. Treemaps to visualise and navigate speech audio. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*. 555–564.

[2] Pedro Almeida, Pedro Beça, Telmo Silva, Marcelo Afonso, Iulia Covalenco, and Carolina Duarte Nicolau. 2022. A Podcast Creation Platform to Support News Corporations: Results from UX Evaluation. In *ACM International Conference on Interactive Media Experiences*. 343–348.

[3] Geetha Sai Aluri, Paul Greyson, and Joaquin Delgado. 2023. Optimizing Podcast Discovery: Unveiling Amazon Music's Retrieval and Ranking Framework. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1036–1038.

[4] Barry Arons. 1997. SpeechSkimmer: a system for interactively skimming recorded speech. *ACM Transactions on Computer-Human Interaction (TOCHI)* 4, 1 (1997), 3–38.

[5] Jana Besser, Martha Larson, and Katja Hofmann. 2010. Podcast search: User goals and retrieval technologies. *Online information review* (2010).

[6] Sylvia Chan-Olmsted and Rang Wang. 2022. Understanding podcast users: Consumption motives and behaviors. *New media & society* 24, 3 (2022), 684–704.

[7] Amelia Chelsey. 2021. Is There a Transcript? Mapping Access in the Multimodal Designs of Popular Podcasts. In *The 39th ACM International Conference on Design of Communication*. 46–53.

[8] Ann Clifton, Aasish Pappu, Sravana Reddy, Yongze Yu, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. The spotify podcast dataset. *arXiv preprint arXiv:2004.04270* (2020).

[9] Tatsuya Ishibashi, Yuri Nakao, and Yusuke Sugano. 2020. Investigating audio data visualization for interactive sound recognition. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 67–77.

[10] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth JF Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu TREC. 2020. Podcasts Track Overview. In *The 29th Text Retrieval Conference (TREC) notebook. NIST, Gaithersburg, MD, USA*.

[11] Rosie Jones, Hamed Zamani, Markus Schedl, Ching-Wei Chen, Sravana Reddy, Ann Clifton, Jussi Karlgren, Helia Hashemi, Aasish Pappu, Zahra Nazari, et al. 2021. Current challenges and future directions in podcast information access. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1554–1565.

[12] Philippe Laban, Elicia Ye, Srujay Korlakunta, John Canny, and Marti Hearst. 2022. Newspod: Automatic and interactive news podcasts. In *27th International Conference on Intelligent User Interfaces*. 691–706.

[13] Daniel Li, Thomas Chen, Alec Zadikian, Albert Tung, and Lydia B Chilton. 2023. Improving Automatic Summarization for Browsing Longform Spoken Dialog. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[14] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6329–6338.

[15] Yu Liang, Aditya Ponnada, Paul Lamere, and Nediyana Daskalova. 2023. Enabling Goal-Focused Exploration of Podcasts in Interactive Recommender Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 142–155.

[16] Zahra Nazari, Praveen Chandar, Ghazal Fazelnia, Catherine M Edwards, Benjamin Carterette, and Mounia Lalmas. 2022. Choice of implicit signal matters: Accounting for user aspirations in podcast recommendations. In *Proceedings of the ACM Web Conference 2022*. 2433–2441.

[17] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[18] Abhishek Ranjan, Ravin Balakrishnan, and Mark Chignell. 2006. Searching in audio: the utility of transcripts, dichotic presentation, and time-compression. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 721–730.

[19] Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. 2021. Detecting extraneous content in podcasts. *arXiv preprint arXiv:2103.02585* (2021).

[20] Rezvaneh Rezapour, Sravana Reddy, Rosie Jones, and Ian Soboroff. 2022. What Makes a Good Podcast Summary?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2039–2046.

[21] Jemily Rime, Jon Francombe, and Tom Collins. 2022. How do you pod? A study revealing the archetypal podcast production workflow. In *ACM International Conference on Interactive Media Experiences*. 11–18.

[22] Jemily Rime, Chris Pike, and Tom Collins. 2022. What is a podcast? Considering innovations in podcasting through the six-tensions framework. *Convergence* 28, 5 (2022), 1260–1282.

[23] Deb K Roy and Chris Schmandt. 1996. NewsComm: a hand-held interface for interactive access to structured audio. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 173–180.

[24] Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Fei Liu. 2022. Towards abstractive grounded summarization of podcast transcripts. *arXiv preprint arXiv:2203.11425* (2022).

[25] Damiano Spina, Johanne R Trippas, Lawrence Cavedon, and Mark Sanderson. 2017. Extracting audio summaries to support effective spoken document search. *Journal of the Association for Information Science and Technology* (2017), 2101–2115.

[26] Lucille Alice Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.

[27] Jaime Teevan, Christine Alvarado, Mark S Ackerman, and David R Karger. 2004. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 415–422.

[28] Stephanie J Tobin and Rosanna E Guadagno. 2022. Why people listen: Motivations and outcomes of podcast listening. *Plos one* 17, 4 (2022), e0265806.

[29] Lorenzo Vaiani, Moreno La Quatra, Luca Cagliero, and Paolo Garza. 2022. Leveraging multimodal content for podcast summarization. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. 863–870.

[30] Haijun Xia, Jennifer Jacobs, and Maneesh Agrawala. 2020. Crosscast: adding visuals to audio travel podcasts. In *Proceedings of the 33rd annual ACM symposium on user interface software and technology*. 735–746.

[31] Longqi Yang, Michael Sobolev, Christina Tsangouri, and Deborah Estrin. 2018. Understanding user interactions with podcast recommendations delivered via voice. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 190–194.

[32] YouGov. 2023. Preferred podcast platforms by age U.S. 2023. https://www.statista.com/statistics/1385092/preferred-podcast-platforms-by-age/.

[33] Peng Yu, Kaijiang Chen, Lie Lu, and Frank Seide. 2005. Searching the audio notebook: keyword search in recorded conversation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. 947–954.

[34] Baoquan Zhao, Shujin Lin, Xin Qi, Zhiquan Zhang, Xiaonan Luo, and Ruomei Wang. 2017. Automatic generation of visual-textual web video thumbnail. In *SIGGRAPH Asia 2017 Posters*. 1–2.

[35] Qiyu Zhi, Suwen Lin, Shuai He, Ronald Metoyer, and Nitesh V Chawla. 2018. VisPod: Content-Based Audio Visual Navigation. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. 1–2.