applied sciences

MDPI

*Technical Note*

# HearIt: Auditory-Cue-Based Audio Playback Control to Facilitate Information Browsing in Lecture Audio

**Jeongmin Hong [1], Hyesoo Jeon [2], Hana Lee [1], Daehyun Kim [2] and Minsam Ko [2,*]**

[1] Department of Applied Artificial Intelligence, Hanyang University, Seoul 133-791, Korea; jeongminhong@hanyang.ac.kr (J.H.); hanalee@hanyang.ac.kr (H.L.)
[2] Department of Human-Computer Interaction, Hanyang University, Seoul 133-791, Korea; shellingphod@hanyang.ac.kr (H.J.); eogus2116@hanyang.ac.kr (D.K.)
* Correspondence: minsam@hanyang.ac.kr

**Abstract:** Students often utilize audio media during online or offline courses. However, lecture audio data are mostly unstructured and extensive, so they are more challenging in information browsing (i.e., chaining, linking, extraction, and evaluation of relevant information). Conventional time-level skip control is limited in auditory information browsing because it is hard to identify the current position and context. This paper presents HearIt, which provides semantic-level skip control with auditory cues for auditory information browsing. With HearIt, users can efficiently change the playback position in the paragraph-level. Furthermore, two auditory cues (positional cue and topical cue) help grasp the current playback and its context without additional visual support. We conducted a pilot study with the prototype of HearIt, and the results show its feasibility and design implications for future research.

**Keywords:** auditory information browsing; lecture audio; auditory feedback

check for updates

**Citation:** Hong, J.; Jeon, H.; Lee, H.; Kim, D.; Ko, M. HearIt: Auditory-Cue-Based Audio Playback Control to Facilitate Information Browsing in Lecture Audio. *Appl. Sci.* **2021**, *11*, 3803. https://doi.org/10.3390/app11093803

Academic Editor: Carlos Alario-Hoyos

Received: 31 March 2021
Accepted: 20 April 2021
Published: 22 April 2021

## 1. Introduction

Nowadays, audio contents contain various helpful information. For example, in many classrooms, students record lectures and use them for their active learning. Especially, such audio recording and re-playing are essential for visually-impaired people. Many people with visual impairment benefit from auditory guidance for their daily activities such as studying [1], filling out a form [2], taking a pictures [3,4]. There have been many studies to support information access of visually-impaired people. Some studies have proposed tools that enable users to process information by converting written texts to spoken text. For example, screen readers provide an audio-based interface to navigate a screen [5–7]. FingerReader [8] reads aloud printed-texts to help blind users aware of the information.

As audio content has become prevalent and its accessibility has been improved, there have been studies to process auditory information effectively. For example, some studies have proposed audio skimming based on the structure of original texts [1], text summarization [9], and audio processing [10]. However, auditory information browsing has been relatively under-studied. Information browsing is a kind of information exploration and represents the behaviors that combine (chain or link) relevant information, extract, and evaluate it. Information browsing is also related to *active reading* [11] that frequently involves seeking, highlighting, comparison, and non-sequential navigation. Conventionally, an audio seekbar helps users explore information in audio by providing playback time and a progress bar. However, it basically supports time-level playback control (e.g., skip 10 s), so the user needs to know the specific playback time positions where the relevant information is placed. Especially, this can be more challenging to browse lecture audio because it mostly has a long length and contains several parts explaining similar but different ones.

This work proposes HearIt, a tool for auditory information browsing based on semantic-level playback control with auditory cues. HearIt is designed to support key components in the behavioral model of information browsing: (1) linking and chaining and (2) extraction and evaluation. First, to facilitate users' linking and chaining data, it provides the *paragraph-level skip control* that repositions the current audio playback based on semantic chunks. Such semantic-based audio playback control can help users find relevant contexts for linking and chaining, effectively more than existing *time-level skip controls* (e.g., skip forward in 10 s). Second, to help users quickly grasp the context of the current playback and evaluate it without any visual supports, HearIt provides two *auditory cues*: *positional cue* and *topical cue*. The positional cue is the paragraph number to offer the perception of the current position. The topical cue is a set of spoken keywords representing a paragraph and helps users determine whether to hear it more.

We conducted a user study to evaluate the effectiveness of HearIt by the within-subject experiment. Twelve participants were asked to perform an auditory information browsing task that explores a lecture audio and points out the specific positions relevant to given contexts. The participants used three browsing methods to perform the task: (1) HearIt, (2) Partial-HearIt, which provides the same function as HearIt, except for topical cue, and (3) Baseline, without the paragraph-level skip and auditory cues. We statistically compared the efficiency (task completion time and accuracy) and the usability of the three methods. We also conducted a survey that contains open-ended questions about the experiences with the browsing methods.

The results show that the proposed method is significantly efficient and effective in auditory information browsing without visual supports. The participants with HearIt mostly found the correct answer in a shorter time compared to those with Baseline. Most of the participants said that the paragraph-level skip is very efficient to reach the targeted position by ignoring irrelevant parts of the audio. Furthermore, the two auditory cues play a crucial role in enabling users to quickly grasp the current context. Our study addresses the needs of studying auditory information browsing and provides design implications for further research.

## 2. Related Work

### *2.1. Audio-Based Information Behavior*

Audio data are closely related to information behaviors in online learning. Many online courses provide lecture audio, and students often record offline lectures by themselves. There have been many studies about the use of lecture audio data. O'Callaghan et al. [12] addressed the advantages and necessity of recorded lecture audio. Lecture audio can effectively supplement the conventional offline learning process. For example, with lecture audio, the students can overcome constraints due to space and time, and lecturers can improve their teaching methods by monitoring the lecture and analyzing their vocabulary selection [13]. In addition, audio media is often used for note-taking. Nakayama et al. [14] examined the effectiveness of audio-based note-taking during the online course. Similarly, Audio Notebook [15] aimed at capturing knowledge directly from conversations as a physical device to facilitate note-taking by audio structuring techniques coupled with the note-taking behavior.

Earlier studies revealed that audio media could improve general asynchronous communication [16]. For example, Voicelist was designed to overcome widespread communication challenges such as cost, textual literacy, and data connections through an interactive voice response [17]. Especially, there have been studies that auditory information can help academic interactions during learning. Auditory information can support interaction and enhance learning opportunities [18]. Merry et al. [19] investigated the modality of academic feedback in e-learning environments and found that the students prefer auditory feedback more than written texts. That was because auditory feedback is mostly more detailed and personalized, so it is easier to understand. In addition, auditory feedback can support more versatile communication, so the instructors can reduce their cognitive and physical loads when recording the feedback [20].

Finally, audio-only media can create spaces for information sharing. Ackerman et al. [21] addressed that audio-only media have much potential of creating a useful social space. From an audio usage perspective in a collaborative space, Metatla et al. [22] found empirical evidence that audio can support non-visual collaboration and collaborators' interactions by helping them extract information about others' actions and current positions. Wang et al. [23] use audio as the interaction medium for a wiki. It is useful for users to navigate the linear structure of the wiki using the audio version.

### 2.2. *Accessibility to Auditory Information*

Many studies have improved information accessibility via audio media. For example, Text-to-Speech (TTS) [1] increases the accessibility to textual content in smart devices. Screen readers convert digital text on the screen into spoken text, enabling users to navigate the web interface and access the text in it, such as documents, menus, icons, and web pages without sight [1,7,24]. BlindReader [25] is designed to help a visually-impaired reader understand the materials effectively by using haptic feedback that provides a sense of touch.

In addition, there has been paid much attention to the accessibility of printed textual materials like books or newspapers. Many studies have proposed user interfaces for delivering printed information to visually-impaired people. For example, finger-worn designs for the blinds have been proposed to control reading in comfortable ways [8,26,27]. FingerReader [8], a wearable device with a small finger-worn form, helps blind users read the printed texts by scanning a single line and reading out the words as synthesized speech, along with the finger.

Furthermore, there have been studies to interpret other types of visual materials (e.g., graph, map, and picture). OrCam [28] is a voice-activated device that attaches to virtually any glasses. It helps blind users live a more independent life by processing a book text, smartphone screen, and recognize faces. Access Lens [29] harnesses computer vision technologies to enable users to use accessible gestures on paper documents and other physical objects, such as product packages. Given the printed image material, Access Lens locates the text and reads specific content where the fingertips touch.

### 2.3. *Auditory Information Processing*

People often use *skimming*, which quickly identifies the gist or general idea of a large volume of contents, and it has been known to be helpful in learning [30–32]. However, unlike the visual contents, skimming on the audio contents is more challenging because information is processed linearly and sequentially. Combining and extracting relevant information is very limited [33].

There have been computational tools for efficient information processing. For example, some studies proposed audio skip controls based on the structures of the audio content. Tyflos [34] utilizes the pyramid structure of the document for guiding users from overview to the details. In addition, Digital Accessible Information SYstem (DAISY) [1] uses the structure of the documents like paragraphs, headings, or sections. With DAISY, users can control the current playback position by the level of the document components, as long as the input audio can be formatted. Similarly, Job Access With Speech by Freedom Scientific (JAWS) is a screen reader that reads out the first sentence of each paragraph sequentially to understand the entire text quickly [35,36]. However, according to the study results [35,36], each paragraph's first sentence is not sufficient for understanding the context.

Alternatively, it has been studied to summarize the audio contents by highlighting essential parts. Summate [9] is a FireFox-based tool that summarizes web pages and presents the summary in an alert box for blind individuals. AcceSS [37] removes the clutter and retains the important sections to give the user a preview of the page. Some studies proposed a skimming method based on audio processing; for example, Imai et al. [10] proposed a method of extracting essential parts of audio content based on voice pitch.

However, from the perspective of information browsing, the prior studies have some limitations. For information browsing (e.g., studying over an audio lecture), it is important to find the relevant details and link them to a specific context by exploring throughout audio content. As a result that information browsing contains a dynamic process, the existing methods that mainly focus on quickly delivering entire content could be limited. In addition, using the text structure is limited because all the audio contents cannot be easily in a specific format. Furthermore, the summarization of the entire text may not be effective because it cannot preserve the mood and style of the original text [36] and did not itself offer playback controls to respond to the users' dynamic information needs. This work focuses on auditory information browsing and explores the supportive design for it.

## 3. Design Space for Auditory Information Browsing

Waterworth and Chignell [38] explained information exploration behavior by searching and browsing. Searching is a behavior that starts with target identification followed by query formulation, search, extraction, and evaluation. On the other hand, browsing is a behavior that begins with context, followed by chaining or linking, and then by extraction and evaluation. The existence of the explicit query (specificity of information needs) distinguishes searching and browsing.

This work aims to support information browsing behavior on lecture audio data. Figure 1 presents the design space for an auditory information browser. We considered four design subspaces according to two dimensions: (1) the level of the playback skip control and (2) the type of cue about the current playback.

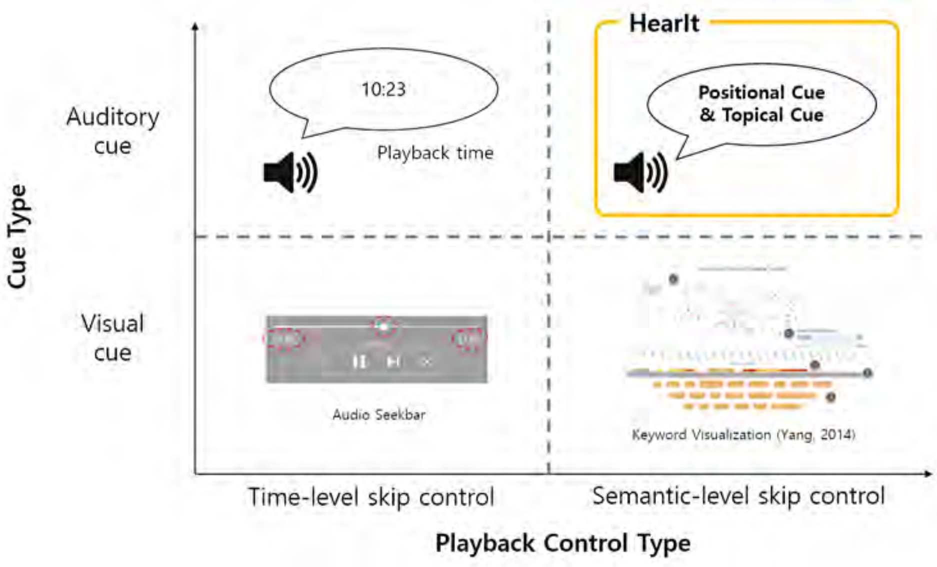


**Figure 1.** Design space for auditory information browsing.

First, the design for the auditory information browser can be specified by the level of the playback skip control. The user should control the audio playback position to find and link relevant information. The time-level skip control (e.g., skip 10 s) has been widely used in conventional audio seekbar, but it does not consider the current playback contexts. Some studies have proposed semantic-level skip control. For example, Yang [39] presented the segment-level keyword search function by the timeline representing the linear-structure of the audio content and visualizing relevant keywords.

Second, providing appropriate cues is important in the design to increase awareness of the current playback. Audio content has diverse advantages, but the audio is often limited to identifying the current playback's position and surrounding contexts because the user depends on hearing only and process the content sequentially. Therefore, the auditory information browser should provide appropriate feedback about the audio playback.

We classified the awareness cue into two modality types: (1) visual cue and (2) auditory cue. There are many graphical user interfaces for audio playback. The most common one is an audio seekbar. It can help users monitor the current playback and easily move to the targeted position using its slider bar. In addition, there have been graphical interfaces for visualizing keywords in audio [39]. Most of the studies in this design subspace utilize texts corresponding to the audio content.

On the other hand, the auditory cue can be helpful in particular situations. The auditory cue can be delivered easily, even on a simple device without a screen, and also it allows users to have multitasking. Especially, it would be helpful for the visually-impaired people's information behavior. However, providing auditory cues for information browsing is more challenging. Information browsing behavior requires more dynamic processes, and the limitations become more serious when audio content is not well structured or does not have available meta-data.

In this design space, HearIt aims to provide semantic-level audio skip control with auditory cues. The following section describes the details about the implementation of HearIt.

## 4. Prototype of HearIt

Figure 2 represents the prototype of HearIt. HearIt has a pen shape to reduce the interface's complexity for audio controls by utilizing intuitive gestures instead of more buttons. HearIt has a play/pause button and a scroll wheel for the playback control. In addition, its pen point is used to recognize gestures of drawing lines. HearIt is designed to support key components of information browsing behavior [38]: (1) linking and chaining and (2) extraction and evaluation. First, HearIt provides paragraph-level skip control to facilitate efficient chaining and linking by enabling the user to skip forward or backward in semantic chunks. In addition, it provides auditory cues that enable the user to extract and evaluate the information based on a quick understanding of its positional and topical context.

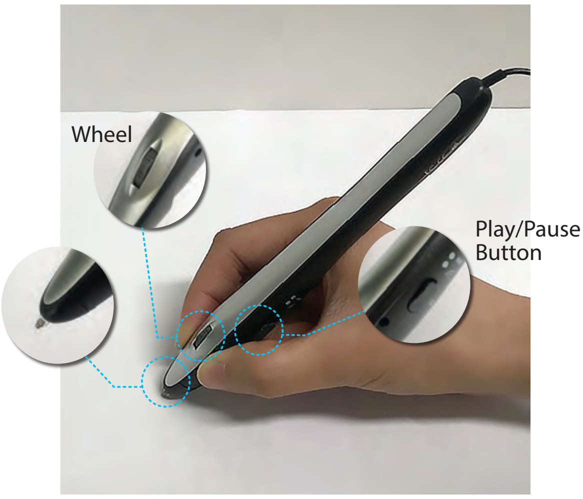


**Figure 2.** HearIt apparatus.

### *4.1. Paragraph-Level Skip Control*

The audio seekbar is a standard interface for playback control. With the seekbar, the user can monitor the current position and quickly change it to a specific position by laying a bar there. However, such GUI-based audio seekbar is not suitable for visually-impaired people because controlling the slider bar is limited. Therefore, the user with visual impairment mostly depends on the time-level skip control (e.g., skip forward in five seconds). The time-level skip control takes a long time to move the playback position corresponding to the context where the user wants to go, and this can cause difficulty in chaining and linking process.

Figure 3 shows the time-level skip control and paragraph-level skip control in HearIt. HearIt provides the paragraph-level skip control that allows the user to move to the beginning of the other paragraph, so the user can quickly reach out to the targeted playback position by skipping irrelevant audio parts. Specifically, a HearIt user can begin to play the audio by the play button. Next, scrolling the wheel changes the playback position to the beginning of the previous or next paragraph. It is also possible to skip multiple paragraphs by holding the scroll wheel in a direction for two or more seconds. HearIt also supports the time-level skipping for sophisticated controls like a typical seekbar. The gesture of drawing a line controls the time-level skip. The direction of the line determines whether to skip forward or backward, and its length determines the length of the skip interval.

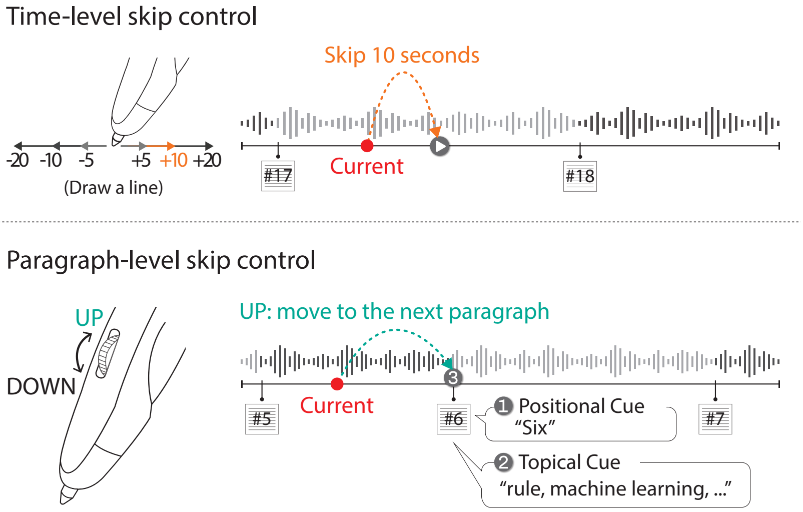


**Figure 3.** Time-level skip vs. paragraph-level skip.

### *4.2. Auditory Cues*

Without a visual audio seekbar, the awareness of the current playback position is limited. It is also hard to know how far the targeted position that the user wants to check is from the current position. Moreover, unlike visual processing, nearby information cannot be simultaneously processed, and this is not effective for information browsing, especially in extracting key ideas and evaluating their relevance.

In HearIt, two auditory cues, (1) positional cue and (2) topical cue, give hints about the context of the current position. The positional cue is a paragraph number to inform where the current playback is, and the topical cue is a set of keywords extracted from the current paragraph. This is to help the user overview the current paragraph, evaluate its relevance, and determine whether to hear more. The keywords for each paragraph are extracted by [40]. First, all the nouns were extracted by morphological analysis of the *konlpy* module. Next, each term is weighted by TF-IDF (term frequency and Inverse document frequency), and the top seven terms, which have the highest weights, are selected for each paragraph. The formula for the weight $w_{t,p}$ of the term $t$ in the paragraph $p$ is as follows:

$$w_{t,p} = TF_{t,p} \times \log \frac{N}{DF_t} \tag{1}$$

where $TF_{t,p}$ is the frequency of $t$ in $p$ and $DF_t$ is the number of paragraphs containing $t$, and $N$ is the total number of paragraphs. Note that the term is excluded if it only occurs once.

When the paragraph-level skip control moves the playback position, HearIt reads out the two auditory cues before the main audio. The positional cue is spoken aloud first, and a set of keywords is followed as the topical cue. The user can hear the topical cue's keywords one by one, starting with the highest weighted term. It is also possible to stop the cue playback and jump to the main audio by drawing a line from left to right.

## 5. Methodology

We conducted a pilot study to evaluate the feasibility of HearIt. In this study, we compared three variants: (1) HearIt, (2) Partial-HearIt, which provides the same function as HearIt, except for topical cue, and (3) Baseline, without the paragraph-level skip and auditory cues. We designed the repeated measures experiment that each participant experiences the three methods. The participant was asked to explore lecture audio and find the specific positions related to the given contexts (searching without explicit queries). We prepared three audio files of a graduate lecture on AI (about 20 min for each). As a result that we simply recorded actual slide-based lectures in a University, it did not have explicit structure at first. To structuralize the audio, we first partitioned the audio by using the page numbers of the slide that the audio covers. Next, several partitioned audio intervals have a longer length than four minutes, so we further divided them based on the context. Finally, the partitioned audio interval was regarded as the paragraph, and there were 31 paragraphs from the three audio lectures. Under the lecturer's supervision, three contexts per lecture audio to be browsed and playback positions to each context (ground truth) were selected. An example of the contexts is "Differences between rule-based systems and machine learning".

### *5.1. Procedure*

First, we had an orientation to explain the procedure and the methods. The participants had their own time to get used to controlling audio playback by the three methods. After the orientation, the participants performed three sessions, and in each session, they were asked to use one of the methods for auditory information browsing. At the beginning of the session, the participant listened to one of the audio without any interruption. Next, the participant browsed the audio by a given method and was asked to find specific playback positions corresponding to each context within five minutes. We blocked the participants' sight by an eyepatch to force them to perform auditory information browsing without any visual support. A session was finished if the participants pointed out all the positions or the time was up. After three sessions had been repeated with a different method and different lecture audio, the participant was asked to respond to the survey.

### *5.2. Participants*

We recruited 12 participants via the school bulletin board as shown in Table 1. The participants were 2 males and 10 females, and their age was 22.9 on average. They were undergraduate students and majored in IT-related, but never took the course we used for this study. We conducted the randomization to control the difficulty of each audio and the order of browsing method.

**Table 1.** Study participants.

| Participant ID | Major | Position | Gender | Age |
|---|---|---|---|---|
| P1 | Culture Tech | Undergraduate | F | 23 |
| P2 | Design Tech | Undergraduate | F | 22 |
| P3 | Media Tech | Undergraduate | F | 23 |
| P4 | Media Tech | Undergraduate | M | 23 |
| P5 | Media Tech | Undergraduate | F | 23 |
| P6 | Media Tech | Undergraduate | F | 22 |
| P7 | Media Tech | Undergraduate | F | 24 |
| P8 | Media Tech | Undergraduate | F | 23 |
| P9 | Culture Tech | Undergraduate | F | 23 |
| P10 | Media Tech | Undergraduate | F | 22 |
| P11 | Media Tech | Undergraduate | F | 23 |
| P12 | Software | Undergraduate | M | 24 |

We note that the number of participants in our pilot study still satisfied the minimum requirement for the statistical test. G*Power (http://www.gpower.hhu.de/, accessed on 10 September 2020), a power analysis application, revealed that a sample size of 12 is enough to provide statistical significance with 80% power ($1 - \beta$) for large effect size ($f^2 > 0.4$) and with an alpha level set at 0.05. In addition, the earlier HCI studies [41,42] used 12 participants to perform a within-subject (repeated measure) experiment to evaluate the task performance with the apparatus.

### 5.3. Data Analysis

We analyzed experimental data in quantitative and qualitative ways. First, we quantitatively compared the following.

- Task completion time: The seconds to point out all playback positions corresponding to three given contexts.
- Accuracy: The percentage of the correct playback positions against the ground truth. We regard the position marked by the participant when it is within five sentences from the ground truth.
- Usability: *System Usability Scale* (SUS) score on 10 5-point Likert scale items [43].

We conducted one-way repeated measures ANOVA (RM ANOVA) to compare the three methods on the task completion time, accuracy, and usability. By Levene's test, we found all the variables do not violate the assumption on the homogeneity of variance. When the ANOVA result is significant, we conducted a series of paired *t*-test (two-tailed) with Bonferroni correction as a posthoc test. All the hypothesis testing was performed with a significance level of 5%.

On the other hand, the survey responses to the open-ended questions were qualitatively analyzed. In the survey, the participants were asked to describe limitations in auditory information browsing without visual supports and compare their browsing experience with the browsing methods.

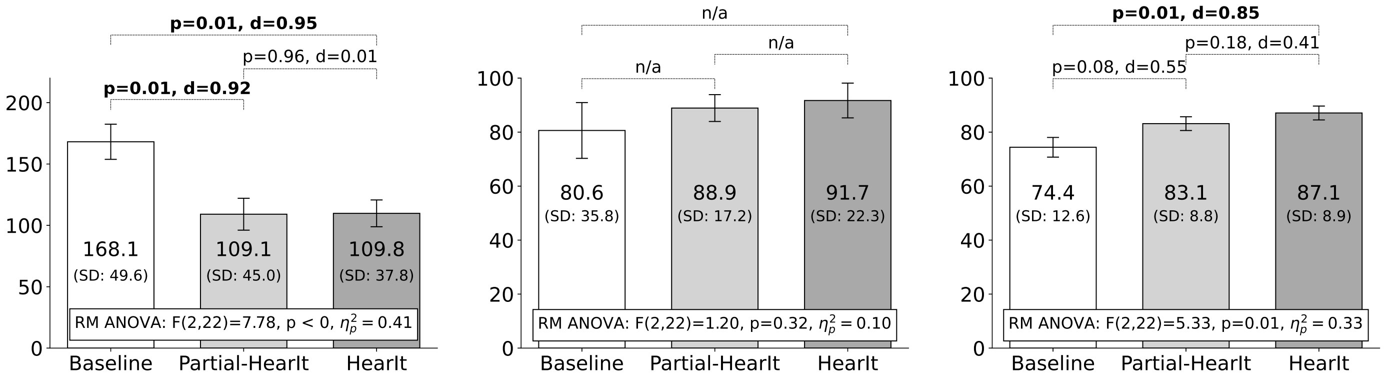
## 6. Results

### 6.1. Auditory Information Browsing without Visual Supports

In the survey, all the participants mentioned that they felt considerable limitations in auditory information browsing without visual support. Many participants reported that tracking the current playback is difficult. For example, P7 said, *"It was hard to figure out how much the playback position needs to be moved to where I wanted to go."* Another participant (P10) mentioned, *"(When skipping multiple intervals) It was difficult to identify how far the position was moved from the start"*. In addition, it tended to be challenging to grasp the current context quickly. For example, P8 commented, *"It was confusing because I had to deal with both what I was hearing and what I wanted to go"*. P11 said, *"I had to be highly concentrated because it requires to process the parts of the lectures sequentially"*. P4 also said, *"It was hard to memorize audio contexts because I could not map each context to the slider bar's position (in an audio seekbar)"*.

### 6.2. Efficiency: Task Completion Time and Accuracy

RM ANOVA revealed a significant and considerable effect of the browsing method on task completion time. As shown in Figure 4, the task completion time with Baseline is significantly longer than Partial-HearIt and HearIt. This indicates that paragraph-level skip control increases the speed in browsing. However, there was no significant difference between Partial-HearIt and HearIt, possibly meaning that the type of auditory cue we considered does not affect the time. HearIt achieved the highest accuracy on average (91.7%), followed by Partial-HearIt (88.9%) and Baseline (80.6%), even though there was no significant difference.



(**a**) Task completion time (second) (**b**) Accuracy (%) (**c**) Usability (SUS score)

**Figure 4.** One-way repeated measures ANOVA and posthoc test results.

### 6.3. Usability: SUS Score and Survey Responses

Table 2 presents SUS scores for each method. RM ANOVA on the SUS score showed a significant difference among the browsing method. HearIt achieved the highest SUS on average (87.1), followed by Partial-HearIt (83.1) and Baseline (74.4). The posthoc test revealed a significant difference between HearIt and Baseline, while the other pairwise comparison tests were not significant.

**Table 2.** System Usability Scale results: mean (SD).

| Category | Item | Baseline | Partial-HearIt | HearIt |
|---|---|---|---|---|
| Satisfaction | I think that I would like to use this system frequently. | 3.25 (1.06) | 3.92 (0.51) | 4.75 (0.45) |
| | I felt very confident using the system. | 3.92 (0.90) | 4.25 (0.75) | 4.75 (0.45) |
| Ease of use | * I found the system unnecessarily complex. | 1.75 (1.06) | 1.50 (0.52) | 1.83 (1.19) |
| | I thought the system was easy to use. | 3.90 (1.00) | 4.50 (0.67) | 4.50 (0.52) |
| | I found the various functions in this system were well integrated. | 3.08 (0.79) | 3.92 (0.51) | 4.58 (0.51) |
| | * I thought there was too much inconsistency in this system. | 1.58 (0.79) | 1.33 (0.49) | 1.50 (0.52) |
| | * I found the system very cumbersome to use. | 2.08 (1.16) | 1.58 (0.51) | 1.50 (0.52) |
| Ease of Learning | * I think that I would need the support of a technical person to be able to use this system. | 1.75 (0.75) | 1.58 (0.67) | 2.00 (1.21) |
| | I would imagine that most people would learn to use this system very quickly. | 4.67 (0.65) | 4.50 (0.52) | 4.33 (0.78) |
| | * I needed to learn a lot of things before I could get going with this system. | 1.42 (0.67) | 1.83 (0.83) | 1.75 (0.75) |

*: Reversed items.

Our qualitative data supplement the results. First, all the participants preferred the paragraph-level skip control over the time-level skip control. There were many responses that the paragraph-level skip control is convenient to explore the information quickly. For example, P1 said, *"It helped me skip the unnecessary information when I remember the content to some extent"*. P11 also commented, *"I could organize the content of information easily in my head when I use paragraph-level skip control"*. One participant (P7) mentioned that the time-level skip is still needed by saying, *"For the more sophisticate playback control, the time-level skip control is still useful"*.

Second, all the participants liked to use the auditory cues in their auditory information browsing. The positional cue is helpful when the user approximately remembers the targeted position related to the given context. P4 said *"I could skip the unnecessary paragraph quickly based on the positional cue"*. Similarly, P8 commented, *"Because I listened to the entire audio at first, the paragraph number is enough for me to know what topics were covered here"*.

However, most of the participants (91.6%) preferred using both additional cues. The topical cue helped the participant quickly evaluate the relevance of the current playback position. P12 said, *"With the topical cue, I could be more confident in guessing the topics of* *the following part"*. P5 commented, *"Topical cue helped me explore what I forgot"*. Finally, the participants mostly heard the top three keywords in the topical cues and decided whether to skip or play its main audio.

## 7. Discussion

The study results reveal the limitations in auditory information browsing without visual support in terms of monitoring the playback position and understanding the current context. HearIt helps the users overcome these limitations by reducing the task completion time but maintaining accuracy. Furthermore, the users felt more comfortable with controlling the audio playback as what they wanted to. The users also became confident in their guesses about the following audio content to decide whether to keep playing.

In this study, we focused on supporting information browsing, while information exploration should be a flexible combination of browsing and querying [38]. Further studies for supporting and balancing both browsing and querying would be helpful. For example, using the audio acceleration and bookmarking functions in [1] can be considered with the paragraph-level skip and the auditory cues. In addition, the current HearIt uses auditory feedback, but it is possible to utilize other modalities, such as leveraging haptic or tactile interaction. For example, [44] proposed wearable devices utilizing vibration that helps visually-impaired people communicate with information. We think that there is much potential for utilizing multimodal interaction in information exploration.

In addition, the current topical cue can be improved in several ways. First of all, it is possible to examine the ideal number and structure of keywords for topical cue to save the time to play them. The keywords' extraction can also be improved, for example, utilizing text features with audio features (e.g., voice pitch [10]). Lastly, the current term weighting is based on the original text's static structure (e.g., paragraph), but it would be possible to use semantic structures by advanced text analysis techniques [45,46].

Finally, we note that the results of this study should be carefully interpreted. The number of participants in the pilot study may not be sufficient to prove the general effectiveness. In addition, we conducted the lab-controlled experiment, so observing the in-situ behavior with the methods would be necessary further. Nevertheless, we believe that our study results show the feasibility of the proposed method and directions for further improvement.

## 8. Conclusions

This paper presents HearIt which supports auditory information browsing by providing the paragraph-level skip control and two auditory cues. Our pilot study with 12 participants showed the potential of the proposed browsing methods and possible research directions. In recent years, auditory information access has been considerably improved, thanks to the advances in smart handheld devices and AI technologies. However, there is still much room for improvement for fostering specific information behavior. We believe that it will be helpful to pay attention to understanding detailed types of information behavior and designing specialized supports for each.

**Author Contributions:** Design and implementation, J.H. and H.J.; analysis, D.K. and H.L.; investigation, D.K. and H.L.; resources, J.H.; writing—original draft preparation, J.H. and H.J.; project administration, supervision, methodology, writing—review, and editing, M.K.; visualization, H.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2018R1C1B5044211) and a grant (21CTAP- C152247-03) from Technology Advancement Research Program funded by Ministry of Land, Infrastructure, and Transport of Korean government. We greatly appreciate the support.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Argyropoulos, V.; Paveli, A.; Nikolaraizi, M. The role of DAISY digital talking books in the education of individuals with blindness: A pilot study. *Educ. Inf. Technol.* **2019**, *24*, 693–709. [CrossRef]
2. Feiz, S.; Billah, S.M.; Ashok, V.; Shilkrot, R.; Ramakrishnan, I. Towards Enabling Blind People to Independently Write on Printed Forms. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Scotland, UK, 4–9 May 2019; pp. 1–12.
3. White, S.; Ji, H.; Bigham, J.P. EasySnap: Real-time audio feedback for blind photography. In Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, New York, NY, USA, 3–6 October 2010; pp. 409–410.
4. Ahmetovic, D.; Sato, D.; Oh, U.; Ishihara, T.; Kitani, K.; Asakawa, C. ReCog: Supporting Blind People in Recognizing Personal Objects. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–12.
5. Buzzi, M.C.; Buzzi, M.; Leporini, B.; Akhter, F. Is Facebook really "open" to all? In Proceedings of the IEEE International Symposium on Technology and Society, Wollongong, Australia, 7–9 June 2010; pp. 327–336.
6. Wu, S.; Adamic, L.A. Visually impaired users on an online social network. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Toronto, ON, Canada, 26 April–1 May 2014; pp. 3133–3142.
7. Bragg, D.; Bennett, C.; Reinecke, K.; Ladner, R. A large inclusive study of human listening rates. In Proceedings of the CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–12.
8. Shilkrot, R.; Huber, J.; Meng Ee, W.; Maes, P.; Nanayakkara, S.C. FingerReader: A wearable device to explore printed text on the go. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Korea, 18–23 April 2015; pp. 2363–2372.
9. Harper, S.; Patel, N. Gist summaries for visually impaired surfers. In Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility, Baltimore, MD, USA, 9–12 October 2005; pp. 90–97.
10. Imai, A.; Tazawa, N.; Takagi, T.; Tanaka, T.; Ifukube, T. A new touchscreen application to retrieve speech information efficiently. *IEEE Trans. Consum. Electron.* **2013**, *59*, 200–206. [CrossRef]
11. Tashman, C.S.; Edwards, W.K. Active reading and its discontents: The situations, problems and ideas of readers. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, 7–12 May 2011; pp. 2927–2936.
12. O'Callaghan, F.V.; Neumann, D.L.; Jones, L.; Creed, P.A. The use of lecture recordings in higher education: A review of institutional, student, and lecturer issues. *Educ. Inf. Technol.* **2017**, *22*, 399–415. [CrossRef]
13. Glass, J.; Hazen, T.J.; Hetherington, L.; Wang, C. Analysis and processing of lecture audio data: Preliminary investigations. In Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004, Boston, MA, USA, 6 May 2004; pp. 9–12.
14. Nakayama, M.; Mutsuura, K.; Yamamoto, H. Effectiveness of audio information for note-taking and learning activities during a fully online course. In Proceedings of the 20th International Conference Information Visualisation (IV), Lisbon, Portugal, 19–22 July 2016; pp. 196–202.
15. Stifelman, L.; Arons, B.; Schmandt, C. The audio notebook: Paper and pen interaction with structured speech. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, WA, USA, 31 March–5 April 2001; pp. 182–189.
16. Schmandt, C.; Arons, B. A conversational telephone messaging system. *IEEE Trans. Consum. Electron.* **1984**, *30*, 21–24. [CrossRef]
17. Cervantes, R.; Sambasivan, N. Voicelist: User-driven telephone-based audio content. In Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services, Amsterdam, The Netherlands, 2–5 September 2008; pp. 499–500.
18. Tomlinson, B.J.; Walker, B.N.; Moore, E.B. Auditory Display in Interactive Science Simulations: Description and Sonification Support Interaction and Enhance Opportunities for Learning. In Proceedings of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–12.
19. Merry, S.; Orsmond, P. Students' attitudes to and usage of academic feedback provided via audio files. *Biosci. Educ.* **2008**, *11*, 1–11. [CrossRef]
20. Heimbïrger, A.; Isomöttönen, V.; Nieminen, P.; Keto, H. How do academics experience use of recorded audio feedback in higher education? A thematic analysis. In Proceedings of the IEEE Frontiers in Education Conference (FIE), San Jose, CA, USA, 3–6 October 2018; pp. 1–5.
21. Ackerman, M.S.; Starr, B.; Hindus, D.; Mainwaring, S.D. Hanging on the 'wire: A field study of an audio-only media space. *ACM Trans. Comput. Hum. Interact. (TOCHI)* **1997**, *4*, 39–66. [CrossRef]
22. Metatla, O.; Bryan-Kinns, N.; Stockman, T. "I Hear You" Understanding Awareness Information Exchange in an Audio-only Workspace. In Proceedings of the CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–13.
23. Wang, L.; Roe, P.; Pham, B.; Tjondronegoro, D. An audio wiki supporting mobile collaboration. In Proceedings of the ACM Symposium on Applied Computing, Ceara, Brazil, 16–20 March 2008; pp. 1889–1896.
24. Voykinska, V.; Azenkot, S.; Wu, S.; Leshed, G. How blind people interact with visual content on social networking services. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, San Francisco, CA, USA, 27 February–2 March 2016; pp. 1584–1595.

25. Sabab, S.A.; Ashmafee, M.H. Blind reader: An intelligent assistant for blind. In Proceedings of the 19th International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 18–20 December 2016; pp. 229–234.
26. Stearns, L.; Du, R.; Oh, U.; Jou, C.; Findlater, L.; Ross, D.A.; Froehlich, J.E. Evaluating haptic and auditory directional guidance to assist blind people in reading printed text using finger-mounted cameras. *ACM Trans. Access. Comput. (TACCESS)* **2016**, *9*, 1–38. [CrossRef]
27. Nanayakkara, S.; Shilkrot, R.; Yeo, K.P.; Maes, P. EyeRing: A finger-worn input device for seamless interactions with our surroundings. In Proceedings of the 4th Augmented Human International Conference, Stuttgart, Germany, 7–8 March 2013; pp. 13–20.
28. Waisbourd, M.; Ahmed, O.M.; Newman, J.; Sahu, M.; Robinson, D.; Siam, L.; Reamer, C.B.; Zhan, T.; Goldstein, M.; Kurtz, S.; et al. The Effect of an Innovative Vision Simulator (OrCam) on Quality of Life in Patients with Glaucoma. *J. Vis. Impair. Blind.* **2019**, *113*, 332–340. [CrossRef]
29. Kane, S.K.; Frey, B.; Wobbrock, J.O. Access lens: A gesture-based screen reader for real-world documents. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 347–350.
30. Yusuf, Q.; Yusuf, Y.Q.; Yusuf, B.; Nadya, A. Skimming and scanning techniques to assist EFL students in understanding English reading texts. *IRJE Indones. Res. J. Educ.* **2017**, *1*, 43–57.
31. Dhillon, B.P.S.; Herman, H.; Syafryadin, S. The Effect of Skimming Method to Improve Students' Ability in Reading Comprehension on Narrative Text. *Linguist. J. Linguist. Lang. Teach.* **2020**, *6*, 77–88.
32. Fauzi, I.; Raya, F. The effectiveness of skimming and scanning strategies in improving comprehension and reading speed rates to students of English study programme. *Regist. J.* **2018**, *11*, 101–120. [CrossRef]
33. Gilmour, A.F.; Fuchs, D.; Wehby, J.H. Are students with disabilities accessing the curriculum? A meta-analysis of the reading achievement gap between students with and without disabilities. *Except. Child.* **2019**, *85*, 329–346. [CrossRef]
34. Keefer, R.; Dakapoulos, D.; Esposito, A.; Bourbakis, N. An interaction based approach to document segmentation for the visually impaired. In *International Conference on Universal Access in Human-Computer Interaction, Proceedings of the 5th International Conference, UAHCI 2009, Held as Part of HCI International 2009, San Diego, CA, USA, 19–24 July 2009*; Springer: Berlin/Heidelberg, Germany, 2009, pp. 540–549.
35. Ahmed, F.; Soviak, A.; Borodin, Y.; Ramakrishnan, I. Non-visual skimming on touch-screen devices. In Proceedings of the International Conference on Intelligent User Interfaces, Santa Monica, CA, USA, 19–22 March 2013; pp. 435–444.
36. Ahmed, F.; Borodin, Y.; Puzis, Y.; Ramakrishnan, I. Why read if you can skim: Towards enabling faster screen reading. In Proceedings of the International Cross-Disciplinary Conference on Web Accessibility, Lyon, France, 16–17 April 2012; pp. 1–10.
37. Parmanto, B.; Ferrydiansyah, R.; Saptono, A.; Song, L.; Sugiantara, I.W.; Hackett, S. AcceSS: Accessibility through simplification & summarization. In Proceedings of the International Cross-Disciplinary Workshop on Web Accessibility (W4A), Chiba, Japan, 10–14 May 2005; pp. 18–25.
38. Waterworth, J.A.; Chignell, M.H. A model for information exploration. *Hypermedia* **1991**, *3*, 35–58. [CrossRef]
39. Yang, H.; Meinel, C. Content based lecture video retrieval using speech and video text information. *IEEE Trans. Learn. Technol.* **2014**, *7*, 142–154. [CrossRef]
40. Yao, L.; Pengzhou, Z.; Chi, Z. Research on news keyword extraction technology based on TF-IDF and TextRank. In Proceedings of the IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 17–19 June 2019; pp. 452–455.
41. Yin, G.; Otis, M.J.D.; Fortin, P.E.; Cooperstock, J.R. Valuating Multimodal Feedback for Assembly Tasks in a Virtual Environment. *ACM Hum.-Comput. Interact.* **2019**, *3*, 1–11. [CrossRef]
42. Olthuis, R.; Kamp, J.v.d.; Lemmink, K.; Caljouw, S. Touchscreen Pointing and Swaiping: The Effect of Background Cues and Target Visibility. *Motor Control* **2020**, *24*, 422–434. [CrossRef] [PubMed]
43. Bangor, A.; Kortum, P.T.; Miller, J.T. An empirical evaluation of the system usability scale. *Intl. J. Hum. Comput. Interact.* **2008**, *24*, 574–594. [CrossRef]
44. Cosgun, A.; Sisbot, E.A.; Christensen, H.I. Evaluation of rotational and directional vibration patterns on a tactile belt for guiding visually impaired people. In Proceedings of the IEEE Haptics Symposium (HAPTICS), Houston, TX, USA, 23–26 February 2014; pp. 367–370.
45. Lai, C.; Farrús, M.; Moore, J.D. Integrating Lexical and Prosodic Features for Automatic Paragraph Segmentation. *Speech Commun.* **2020**, *121*, 44–57. [CrossRef]
46. Yang, X.; Yumer, E.; Asente, P.; Kraley, M.; Kifer, D.; Giles, C.L. Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4342–4351. [CrossRef]