Journal of Information Systems Engineering and Management
2024, 9(3), 28013
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Caring for Special Participants in the Digital Media Era: A Study on Enhancing the Blind User Experience on Short Video Platforms Through Auditory Cues

Xin Wang [1,2], Anping Cheng [3], Kiechan Namkung [4], Younghwan Pan [5*]

[1] Ph.D candidate, Department of Smart Experience Design, Graduate School of Techno Design, Kookmin University, Seoul, Republic of Korea
[2] Lecturer, Shandong Vocational College of Special Education, Jinan, China
[3] Doctor, Department of Smart Experience Design, Graduate School of Techno Design, Kookmin University, Seoul, Republic of Korea
[4] Professor, Department of AI Design, Graduate School of Techno Design, Kookmin University, Seoul, Republic of Korea
[5] Professor, Department of Smart Experience Design, Graduate School of Techno Design, Kookmin University, Seoul, Republic of Korea
* **Corresponding Author:** peterpan@kookmin.ac.kr

Citation: Wang, X., Cheng, A., Namkung, K., & Pan, Y. (2024). Caring for Special Participants in the Digital Media Era: A Study on Enhancing the Blind User Experience on Short Video Platforms Through Auditory Cues. *Journal of Information Systems Engineering and Management*, *9*(3), 28013. https://doi.org/10.55267/iadt.07.14774

ARTICLE INFO

Received: 29 Apr 2024
Accepted: 16 May 2024

ABSTRACT

Screen readers for the visually impaired and blind and short video platforms have conflicting functionalities. In particular, blind users encounter information access barriers when searching for video content, which reduces their user experience. We embed auditory cues at the beginning of a short video corresponding to its content to help blind users identify the video type. The experimental design and evaluation results reveal the significant impact of these auditory cues. By embedding auditory cues, we can significantly enhance the user's usability, recognition efficiency, and emotional experience, surpassing traditional short videos' experience. Speech had the shortest response time and highest accuracy, while auditory icons provided a better emotional experience. In addition, some participants expressed concerns about the potential social privacy issues associated with Speech. This study provides auditory cue-matching solutions for a wide range of short videos. It offers a beacon of hope for enhancing the experience of short video platforms for the blind user. By doing so, we contribute to the well-being of people with disabilities and provide highly versatile user experience design recommendations for a broader range of digital media platforms.

Keywords: Blind User Experience, Digital Media, Auditory Cues, Short Video Platforms.

INTRODUCTION

Short videos, popularized by platforms like TikTok, Instagram and YouTube Shorts, are often captured, edited, or enhanced using smartphones for social sharing (Kaye, Chen, & Zeng, 2021). This mode of dissemination has gained immense popularity, even among blind individuals (Liu, Carrington, Chen, & Pavel, 2021). Blind individuals rely primarily on assistive technologies, especially screen readers, to discern and comprehend short videos. Screen readers convey label information of short videos through a text-to-speech (TTS) mechanism (Kuber, Hastings, & Tretter, 2012). However, when short videos lack label information and audio descriptions, the screen reader's screen reading efficiency degrades to varying degrees, significantly reducing the usability of short video software (Encelle et al., 2021). Subsequently, this diminishes the overall user experience (Liu et al., 2021).

Another auxiliary technology is auditory cues (Mynatt, 1994). The technology conveys information or

instructions to users through sound signals. Existing research has involved this technology in blind intelligent navigation (Bilal Salih et al., 2022), Blind Education and Training (Dulyan & Edmonds, 2010), Blind Entertainment Interaction, and so on (Voykinska, Azenkot, Wu, & Leshed, 2016). It is noteworthy that auditory cues have significantly enhanced the quality of life and experiences for individuals who are blind or visually impaired, such as in remote navigation assistance systems (Chaudary, Pohjolainen, Aziz, Arhippainen, & Pulli, 2023). Regarding applying auditory cues in multimodal interaction technology, most studies have focused on developing basic life necessities for individuals with disabilities (Hussain, L. Chen, Mirza, G. Chen, & Hassan, 2015). However, short videos are increasingly important in social media and information dissemination (Wu et al., 2021). The selection of effective hearing assistance technology for blind people and promoting the leapfrog innovation of science and technology are critical issues in the current field (Mankoff, Fait, & Tran, 2005).

Moreover, studies have demonstrated the positive physiological and emotional feedback of blind people when they receive auditory cues from a medical point of view (Klinge, Röder, & Büchel, 2010); these discussions suggest that the field has tremendous research potential. There is currently insufficient research on improving the user experience of blind individuals on short video platforms. This study explores how auditory cues enhance the usability, response time, accuracy, and emotional response of blind users in identifying different types of short videos. The experimental results are then analyzed. We propose diverse design suggestions to help enhance the overall user experience for blind users. These design recommendations provide diverse options for embedding auditory cues in short video platforms from a Multi-modal technology and interaction design perspective but also provide valuable insights and innovative solutions for enhancing the user experience and engagement of blind users in the digital media space.

This study employed a mixed research method divided into three phases. Firstly, we conducted semi-structured interviews to understand the pain points of blind users of Douyin (Chinese version of TikTok). Secondly, we selected three auditory cues and five representative short videos through literature analysis. Expert sound designers then precisely reproduced the auditory cue prototypes. Lastly, using a quantitative research approach, 80 blind students were randomly assigned to identify short video types using a laptop, with their response time and results recorded. The study confirmed the effectiveness of auditory cues in enhancing the user experience for blind individuals in identifying short videos. Random sampling interviews also revealed preferences. The paper introduces the research background in the "Introduction." Then, it provides a detailed description of the research methodology in the "Methods" section. The "Results" section presents the research findings through ANOVA analysis. In the "Discussion and Conclusion" section, the authors discuss the recommendations and limitations of the research design, summarize the research findings, and propose future research directions.

The main contributions of this paper are (1) Innovatively applied auditory cues to a short video platform and validated that it can effectively enhance the user experience for blind people. (2) A detailed design process and interaction experiment results of auditory cues will help practitioners and researchers understand their application in various fields (business, education, leisure, and rehabilitation). (3) Utilizing a rigorous mixed-method approach, develop comprehensive and rational design recommendations tailored for blind users on both short video and large-scale digital media platforms, ensuring universal design principles to enhance user experience for individuals with visual impairments and promote inclusivity for all users.

LITERATURE REVIEW

Classification and Accessibility of Short Videos for Blind People

Short videos are typically limited to 2 minutes or less, and possess social attributes (Kaye et al., 2021). According to the survey conducted by Guo and Gurrin (2012), previous studies have mainly focused on the theory of multimodal video classification. These studies extract features from video text, audio, and visual characteristics. Subsequently, mainstream algorithms such as Support Vector Machines (SVM), k-nearest Neighbors (K-NN), Naive Bayes classifier (NB), and Decision Trees (DT), among others, are employed to define the classification results of video streams. Scholars have researched ways to make video content more accessible for Visually impaired and blind. Gregory Frazier first proposed Audio Description (AD) in the 1970s in the United States. It is a literary form, similar to haiku poetry, that uses brief and vivid language to convey visual images that are otherwise inaccessible. This expression helps blind or visually impaired individuals better understand visual experiences (Snyder, 2005). Wang et al. (2021) worked on creating a system to automatically generate audio descriptions for video content. Liu et al. (2021) created a video search interface to make it easier for all users, including those with visual impairments, to find videos. Encelle, Ollagnier-Beldame, Pouchot, and Prié (2011)

suggested using speech synthesis and Earcons in the ACAVA project to make videos more accessible for blind individuals. Palmer, Schloss, Xu, and Prado-León (2013) proposed Cross-modal correspondences can be further enhanced through intermediary factors such as culture (e.g., pitch in musical scores), language (e.g., the term "high" used to describe both sound and height), or emotional connotations (certain sounds and colors can match in emotional valence). In conclusion, how to break through the limitations of traditional short video attributes and enable blind people to access content through better multimodal interactions needs to be further explored as an adaptive solution that suits the characteristics of short videos.

Auditory Cues and Their Application Effects

Auditory cues are a simple concept, referring to using sound signals to convey information or instructions to users (Mynatt, 1994), enhancing user experience and operational efficiency (Garzonis, Bevan, & O'Neill, 2009). Artificial auditory cues, like alarm clock beeps or email notifications, are commonly used to provide information (Stephan, Smith, Martin, Parker, & McAnally, 2006). Auditory cues are increasingly popular among the blind community for accessibility.

Research indicates that auditory cues exhibit various cognitive and physical features. Regarding cognitive features, specific mappings, and metaphors can assist users in better identifying particular information. Different alert sounds may influence users' perception and response to information (Walker & Kramer, 2005). Ease of learning, intuitiveness, memorability, and user preferences can impact user experience and satisfaction (Garzonis et al., 2009); Pitch, volume, duration, frequency, and rhythm are important physical features (Nees & Liebman, 2023). Pitch conveys emotion and information, volume attracts attention, duration creates momentary or sustained cues, frequency generates sensory effects, and rhythm enhances alertness or provides a sense of pace. Therefore, Brown, Newsome, and Glinert (1989) proposed that diverse auditory cues can replace traditional visual cues. In social media, previous research on auditory cues has mainly focused on webpages (Donker, Klante, & Gorny, 2002), Electronic Games (Roth, Petrucci, Pun, & Assimacopoulos, 1999) and Applications on mobile devices (Csapó, Wersényi, Nagy, & Stockman, 2015). Additionally, researchers have employed multimodal technologies that combine auditory and tactile cues to guide blind individuals in performing computer-related tasks (Shimomura, Hvannberg, & Hafsteinsson, 2010). However, from the perspective of assisting blind individuals in browsing short videos, auditory cues are more compact, flexible, and convenient than tactile feedback (Csapó et al., 2015). With the high prevalence of smartphones among the blind population (Abraham, Boadi-Kusi, Morny, & Agyekum, 2022), however, the built-in screen reader alert method on smartphones is singular, lacks fun, and has low privacy, making it a common expectation among the current blind community to enhance these aspects of the experience (A. Khan & Khusro, 2021). Due to the diversity of auditory cue types, previous studies have demonstrated their usability in various scenarios.

The widely recognized auditory cues comprise Speech, Auditory Icons, and Earcons (Garzonis et al., 2009; Dingler, Lindsay, & Walker, 2008). Speech conveys linguistic information through sound signals (Adebiyi et al., 2017). Garzonis et al. (2009) suggest that in developing navigation assistance technology for blind people, providing action prompts through speech has demonstrated excellent usability and user satisfaction. Research results indicate that simple verbal communication is the most explicit way to describe objects. However, it is essential to note that this approach may raise privacy concerns when used in public settings. Hussain et al. (2015) showed that repeating verbal commands increases cognitive load. Auditory icons use sounds that mimic real-world objects or actions to convey information or indicate operations (Gaver, 1987). Auditory icons mimic sounds of real-world objects. They capture the fundamental characteristics of events in a complex yet intuitive manner (Cabral & Remijn, 2019). Cabral and Remijn (2019) argue that when applying Auditory icons, one should adhere to sound design characteristics, including recognizability, conceptual mapping, physical parameters, and user preferences. Edworthy, Parker, and Martin (2022) in their research on the design of alarms in clinical settings, fully utilized the unique physical characteristics of Auditory icons, supporting their application in creating accurate and usable multi-contextual alerts. However, Šabić, Chen, and MacDonald (2021) in a study on in-car warning sounds, validated that the performance of auditory icons is least stable in different environments. Earcons are a method of conveying auditory information using abstract, musical-like tones. They consist of short, rhythmic sequences of pitches with variable volume, timbre, and pitch range (Brewster, Wright, & Edwards, 1993). Designers can create earcons that apply to various objects, operations, or interactions, offering high flexibility. For instance (Leplâtre & Brewster, 2000) using Earcons to indicate hierarchical menus on the phone has enhanced navigation task performance within the menu. However, Earcons lack intuitive associations (Sanderson, Wee, Seah, & Lacherez, 2006) and have high learning costs. In summary, there's a research gap in adapting auditory cues to short videos.

Usability, Recognition Efficiency, and Emotional Experience

Usability refers to the ease with which a product or system can be used, learned, and memorized by specific

user groups in particular contexts, involving effectiveness, efficiency, and satisfaction (Donker et al., 2002; Jordan, 2020). In the study, usability testing demonstrated that the auditory web browser performed significantly in effectiveness, efficiency, and satisfaction. Therefore, usability is an essential metric for enhancing the user experience of blind individuals on social media platforms (Chaudary et al., 2023). The investigation confirmed the usability of remote navigation assistance for visually impaired individuals. The study of Theodorou et al. (2022) proposed usability testing as a critical dimension. It introduced research metrics for a smartphone training app that simulates outdoor navigation for blind people. Through a combination of user experience evaluations and emotional analysis, the study concluded that it enhances the overall user experience. Brooke (1986) introduced the System Usability Scale (SUS) 1986 which is widely used for evaluating system usability (Lewis, 2018). Our study will evaluate the usability of short video audio cues. Verifying whether this new technology can enhance the user experience is essential (Finstad, 2010).

In studies related to auditory cues, response time and accuracy are often considered core data metrics for assessing the efficiency of human interaction with technology (referred to as "recognition efficiency" to distinguish it from the efficiency dimension in usability). Users use these metrics to measure performance in task execution (Townsend & Altieri, 2022). Garzonis, Bevan, and O'Neill (2008) studied the intuitiveness of semantics in audio stimuli through testing based on response time and accuracy. This research provides crucial reference data for the design of mobile service notifications. Mieda, Kokubu, and Saito (2019) verified that blind soccer players can quickly identify the direction of sounds through auditory cues, demonstrating unique fundamental cognitive skills as tested for response time and accuracy. M. A. Khan, Paul, Rashid, Hossain, and Ahad (2020) evaluated the performance of an AI-based visual assistance system in controlled environments with real scenarios encountered by blind individuals using response time and accuracy. This study demonstrated that the system provides greater accessibility, comfort, and navigational convenience for visually impaired individuals. In the current literature, we lack research evaluating the efficiency of recognizing auditory cues embedded in short videos, using response time and accuracy as metrics.

Emotional experience is a part of user experience, reflecting an individual's emotional response (Saariluoma & Jokinen, 2014). According to the research findings by Rokem and Ahissar (2009), Blind individuals possess higher auditory and memory capabilities than sighted individuals. Meanwhile, the study by Klinge et al. (2010) focused on the amygdala in the brains of blind individuals, providing further evidence that blind individuals exhibit superior activation in emotion, memory, and cognition compared to sighted individuals. In the absence of visual or emotional experiences, the amygdala, particularly when exposed to auditory emotional signals, develops responses, serving sensory patterns and becoming the most reliable source of emotional information, thus providing theoretical value for research on emotional experiences in blind individuals. Mehrabian and Russell (1974) first proposed the PAD three-dimensional emotion model in 1974, which posits that emotions have three dimensions: pleasure, arousal, and dominance. These dimensions can effectively explain human emotions. Subsequently, the Self-Assessment Manikin (SAM) provided a more intuitive representation of Mehrabian and Russell's (1974) three PAD dimensions. Researchers designed it as an alternative method to sometimes cumbersome verbal self-report measurements (Lang, 2019). Redondo, Fraga, Padrón, and Piñeiro (2008) utilized the Self-Assessment Manikin (SAM) to assess 111 sounds, measuring emotional ratings on the Pleasure, Arousal, and Dominance dimensions. Soares et al. (2013) employed the Self-Assessment Manikin (SAM) to collect emotional dimension data for auditory stimuli, providing standardized emotional sounds for the Portuguese language system and investigating emotional dimension ratings across different cultures. Iturregui-Gallardo and Méndez-Ulrich (2020) designed a version of the Self-Assessment Manikin (SAM) more suitable for visually impaired individuals, creating the Tactile Self-Assessment Manikin (T-SAM). This development scientifically evolved the tool for measuring emotional experiences from an accessibility and inclusive design perspective. Unlike the applications mentioned earlier, there is a notable research gap in measuring emotional experience scores using SAM for short videos with embedded auditory cues, highlighting an unexplored and noteworthy research direction. Despite extensive research on auditory cues and emotional experiences, the effects of embedding auditory cues in short videos need further understanding. Therefore, we address the following research questions:

RQ1: Does embedding auditory cues in short videos enhance users' effectiveness, efficiency, and satisfaction in identifying types?

RQ2: Does embedding auditory cues in short videos enhance users' recognition efficiency (response time, accuracy)?

RQ3: Does embedding auditory cues in short videos improve users' emotional experiences?

RQ4: Which type of auditory cue do they prefer the most?

METHODOLOGY

This study employed a mixed research method, conducted in three phases. Interviews and surveys obtained qualitative data, while experiments collected quantitative data. A chi-square test was used during data analysis to confirm the absence of proportional bias between gender and experimental group assignments. We used stats to understand response time, accuracy, and survey data. To conduct difference testing, we used a multifactorial analysis of variance to compare response times and accuracy, followed by a simple effects analysis. Survey results underwent one-way analysis of variance, including post-hoc tests. We assessed the reliability and validity of the survey results through reliability testing and validity analysis. We also used a Think-Aloud Protocol (TAP) (van Someren, Barnard, & Sandberg, 1994), asking participants to verbalize their thoughts, feelings, and opinions during the data collection. This method provided a more precise understanding of participants' thoughts and experiences throughout the experiment.

First Phase of the Study: Interviews

We recruited ten blind participants (**Table 1**) who had over three years of experience using Douyin (the Chinese version of TikTok) and demonstrated normal cognitive and hearing abilities. They were familiar with all the features of the application.

Table 1. Statistical Information on Interview Participants

NO.	Gender	Age	Vocation	Visual characteristics	Usage time（years）
1	Male	18	Student	No vision	3
2	Male	26	Worker	No vision	4
3	Female	45	Self-employment	No vision	3.5
4	Male	34	physical therapist	No vision	5
5	Female	30	Self-employed"	No vision	6
6	Female	18	Student	No vision	3
7	Male	17	Student	No vision	5
8	Female	25	Performer	No vision	4
9	Female	22	streamer	No vision	3.5
10	Male	16	Student	No vision	4.5

We designed questions to understand participants' experiences, evaluations, and challenges when using the short videos platform in a traditional way to browse short videos (**Figure 1**). With participants' consent, we recorded the interview process using mobile phones and sought detailed answers by posing additional questions. The interviews continued until theoretical saturation was achieved (Rowlands, Waddell, & McKenna, 2016). Data analysis revealed the participants' primary pain points: 1. Screen readers were influenced by label information when identifying the types of short video content, resulting in reduced usability and recognition efficiency; 2. Poor user experience satisfaction.

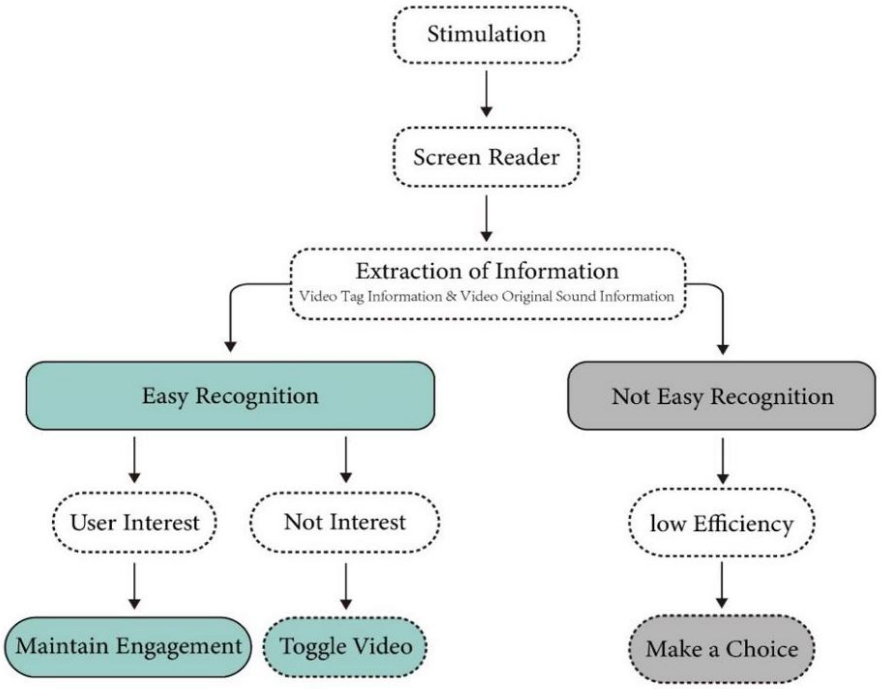


Figure 1. The Process of Recognizing Short Video Types Using Traditional Methods

Second Research Phase: Preparation of Experimental Materials

Selection of Experimental Materials: Short Videos

In a previous study, Dinh, Dorai, and Venkatesh (2002) employed a sound feature-based approach, extracting audio elements such as Energy, Variance, and zero-crossing rate. They utilized SVM and k-nearest neighbor classifiers, achieving classifier recognition accuracy rates of 96.1% and 88.8%. Several studies have used five types of short videos: "news," "commercials," "sports," "cartoons," and "music" and have obtained scientifically rigorous results. The study concluded that accurate video categorization could be achieved solely based on audio information, and video duration had minimal impact on classification effectiveness. The above study coincides with the auditory discrimination characteristics of blind users and the temporal attributes of short videos, thus further validating the rationality of choosing "news," "advertisement," "sports," "cartoon," and "music" as the experimental materials.

Production of Experimental Materials: Auditory Cue

As a foundational research endeavor, we selected well-known auditory cue types as independent variables, including Speech, Auditory icons, and Earcons (Dingler et al., 2008). We collected these three types of auditory cues from the British Broadcasting Corporation (BBC). Subsequently, sound design experts faithfully replicated them according to design prototypes outlined in authoritative literature (Blattner, Sumikawa, & Greenberg, 1989), ensuring that all production parameters adhered to standards. This meticulous process guaranteed the scientific integrity of the auditory materials.

The final presentation comprised an auditory cue library crafted based on the production of five video genres (**Table 2** and **Figure 2**).

Table 2. The 5 Types of Short Videos and the Corresponding Sound Descriptions

No.	Short video type	Speech (Chinese)	Auditory icon	Earcons
1	News	新闻	BBC News ident (Garzonis et al., 2009)	Piano-monophonic going up (Garzonis et al., 2009)
2	Advertisement	广告	Sound of a TV switching on to white noise (Garzonis et al., 2008)	Piano-polyphonic going down (Garzonis et al., 2009)

No.	Short video type	Speech (Chinese)	Auditory icon	Earcons
3	Sports	体育	Stadium crowd (Garzonis et al., 2009)	Piano-monophonic going down (Garzonis et al., 2009)
4	Cartoon	卡通	Wind chimes (Garzonis et al., 2009)	Violin-varying pitch chords and single notes (Garzonis et al., 2009)
5	Music	音乐	Audience applauding (e.g. in a theatre) (Garzonis et al., 2009)	Piano-polyphonic going down (Garzonis et al., 2009)

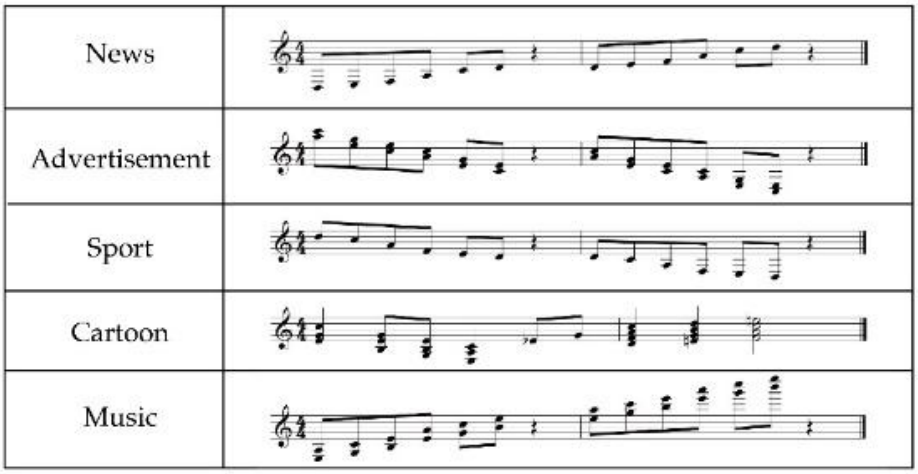


Figure 2. Examples of Earcons

Design of the Novel Approach

Following the design principles of auditory cues (Blattner et al., 1989), we set the duration of all three auditory cues at 2624 ms. Additionally, we kept the volume, quality, and frequency consistent with the design parameters proposed by Garzonis et al. (2008). The net duration of each of the five short videos was 90 seconds, ensuring the absence of text descriptions or other label information within the content. We embedded auditory cues at the beginning of the short videos. After rationally pairing the three types of auditory cues with the five video genres (**Figure 3**), we generated an experimental material library, as shown in **Table 3**. The primary characteristic of the auditory cues embedded in short videos is their reliance on auditory content for matching, serving as a novel auditory clue specifically designed to assist users in identifying short video genres.

Table 3. Twenty Short Videos Used in the Testing Tasks

Short Video clips for Auditory recognition tasks			
Group A	**Group B**	**Group C**	**Group D**
S1+V1	A1+V1	E1+V1	V1
S2+V2	A2+V2	E2+V2	V2
S3+V3	A3+V3	E3+V3	V3
S4+V4	A4+V4	E4+V3	V4
S5+V4	A5+V5	E5+V3	V5

S = Speech, A =Auditory icon, E = Earcons, 1 = News, 2 = Announcement, 3 = Sports,4 = Cartoon, 5 = Music, V = Short video.

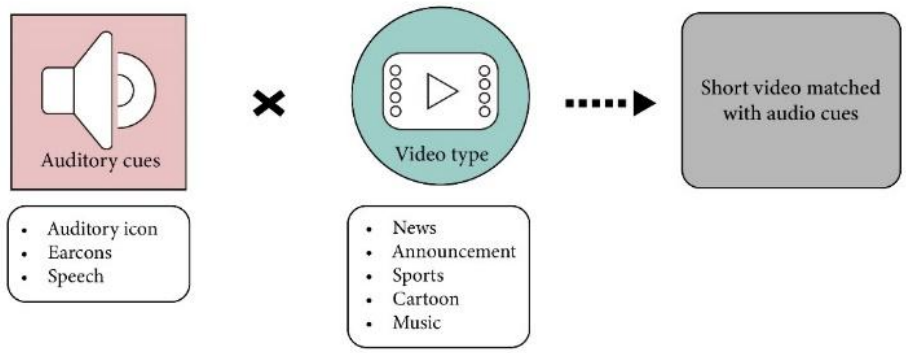


Figure 3. Design Approach Schematic

Third Research Phase: Experimental Design and Methodology

Experimental Methodology

The experiment used a between-group design, randomly assigning 20 participants to each of the four groups. The first three groups were experimental groups (A/B/C), receiving three different short video auditory cues: speech, auditory icons, and earcons, respectively. The fourth group was the control group (D), which received no auditory cues. We recorded participants' response times and accuracy rates as objective data. Each group provided subjective data regarding effectiveness, efficiency, and satisfaction through usability assessment questionnaires. Additionally, we utilized emotional experience questionnaires to collect data from all four groups. The final section employs a random sampling interview approach to understand user preferences. The remaining portions of this section delineate the experimental methodology, participants, testing environment, tasks to be executed, roles of the test personnel, and the dependent variables under study. It is noteworthy that we employed the Think-Aloud protocol (van Someren et al., 1994). It combines questionnaire surveys, brief interviews, and user observations. We document any verbal comments and expressions indicating user emotions during the testing process for later use in interviews. Throughout the usability testing, we record user video performances to gather additional information about their behavior. We further explain the overview of the research design below, as shown in **Figure 4**.

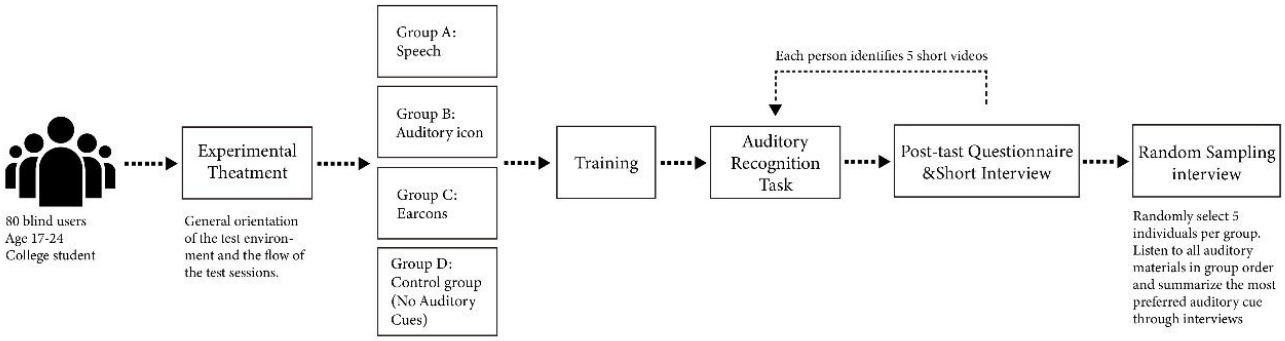


Figure 4. Experimental Flowchart

Testing Phase Overview: Provides participants a general introduction to the testing environment and procedures.

Post-test Questionnaires and Brief Interviews: Questions encompass user satisfaction with the effectiveness

and efficiency of task execution and emotional experiences. Test personnel conduct brief interviews immediately after each task based on observational outcomes.

Random Sampling Interviews: After completing the post-test questionnaires, we randomly select 20 participants from the four groups. They first listen to auditory cues from other groups' short videos and then participate in interviews, elucidating their user preferences.

Participants

We recruited 80 participants (47 males, 33 females) with an average age of 18.8 years (SD ± 1.8), ranging from 17 to 24 years old, all of whom were students from Shandong Special Education Vocational College (**Table 4**). They sequentially completed tasks I, II, and III. All participants had to be utterly blind with normal hearing (certified by disability cards) and pass the Cognitive Test for the Blind (Nelson, Dial, & Joyce, 2002). They need at least three years of Douyin experience (verified through open-source data monitoring within the software). Participants were informed about the nature of the study and data collection methods. We informed participants that their participation was voluntary and without compensation, with the right to withdraw at any time without consequences. They signed written consent forms and waivers based on informed consent principles. Indeed, the Ethics Committee at Kookmin University of Korea approved this study.

Table 4. Profile of Participants

Experimental Group	Age Range (Number)	Gender
Speech	17-18(14)	Male(8)/Female(6)
	19-20(4)	Male(2)/Female(2)
	21-22(2)	Male(1)/Female(1)
	23-24(1)	Male(1)/Female(0)
Auditory	17-18(14)	Male(9)/Female(5)
	19-20(2)	Male(1)/Female(1)
	21-22(1)	Male(0)/Female(1)
	23-24(1)	Male(1)/Female(0)
Earcons	17-18(12)	Male(7)/Female(5)
	19-20(3)	Male(3)/Female(0)
	21-22(4)	Male(2)/Female(2)
	23-24(2)	Male(1)/Female(1)
Control Group	**Age Range (Number)**	**Gender**
No auditory cues	17-18(10)	Male(6)/Female(4)
	19-20(4)	Male(2)/Female(2)
	21-22(2)	Male(1)/Female(1)
	23-24 (4)	Male(2)/Female(2)

Experimental Environment

We conducted the experiment in a quiet classroom where participants sat in front of a ThinkPad laptop with their hands on an external keyboard (Logitech G610) on the desk. The Windows Media Player software played the required short videos on the display screen. Throughout the experiment, participants wore TONEMAC H3 headphones, and we uniformly adjusted all sound levels to a comfortable 75 decibels (dB) (a).

Task Protocol

All participants underwent pre-test group training, including repeated listening to auditory cues corresponding to short video types and proficient operation of the external keyboard to control video playback.

In Task I, after signing the consent form, participants followed instructions to sit in front of the screen, wear headphones, and place their fingers on the Spacebar key of the keyboard to control video playback and pause. Researchers explained the auditory recognition task again, where participants randomly identified five short videos (**Figure 5**). They quickly determined the video type based on auditory cues, verbally stating their answers immediately after pressing the pause key. We employed randomized playback with a 5-second interval between tasks. At the start of each task, researchers recorded participants' response times and identification results.

(a) Screen Display Elements

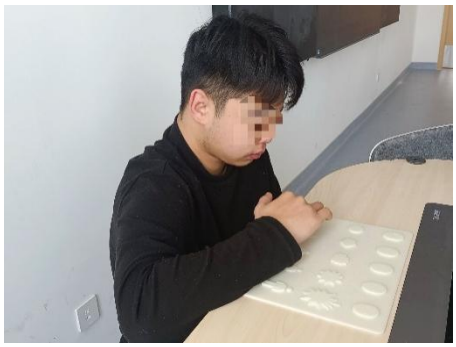
(b) Task site

Figure 5.

In Task II, each participant who had just completed the auditory recognition task immediately filled out the System Usability Scale (SUS) questionnaire (Jordan, Thomas, McClelland, & Weerdmeester, 1996). The SUS provided a five-point scale evaluation for effectiveness, efficiency, and satisfaction, ranging from 1=Strongly Disagree to 5=Strongly Agree. Precisely, we followed the scoring methodology of Jordan et al. (1996) with some appropriate modifications. For instance, we altered the seventh question from 'Most People' to 'Most Blind Individuals' to reflect the subjects and objectives of our study more accurately. Subsequently, participants utilized the T-SAM Emotional Scale to obtain emotional experience scores.

Participants used a tactile 9-point T-SAM questionnaire featuring raised markings in graphics and Braille to rate their emotional experiences. After each short video playback, participants framed the pleasure assessment as "How pleasant, happy, and satisfied do you currently feel?" with responses ranging from 1=Strongly Disagree to 9=Strongly Agree. Participants assessed emotional arousal by responding to the question, "How do you currently feel stimulated, excited, or frenzied?" on a scale from 1=Strongly Disagree to 9=Strongly Agree. Participants assessed dominance by responding to the question, "How do you currently feel, in control of what might happen next, making your own choices, feeling important?" with responses ranging from 1=Strongly Disagree to 9=Strongly Agree (**Figure 6**).

(a) Tactile Version of the Self-assessment Manikin (T-SAM)



(b) Usage site

Figure 6.

In Task III of the random sampling interviews, five participants were randomly selected from each of the four groups, resulting in 20 individuals. Participants from other groups experienced experimental materials and, after each set of materials, they provided evaluations. Their feedback included sentiments such as "feels easier to identify than the materials from their group" and "feels more cheerful than the previous group". Finally, each participant expressed their favorite short video auditory cue and provided reasons for their preference.

Quality Characteristics and Metrics

This experiment, by the definition of recognition efficiency outlined by Townsend and Altieri (2022), measured response time and accuracy as objective data. Subjective data, on the other hand, encompassed usability, efficiency, satisfaction, pleasure, arousal, dominance, and other dimensions of user emotional experience based on usability metrics described by Lewis and Sauro (2017) and emotional experience dimensions

proposed by Mehrabian and Russell (1974). Multiple indicators were employed, aligning with the criteria for usability (Donker et al., 2002), recognition efficiency (Townsend & Altieri, 2022), and emotional experience (Saariluoma et al., 2014), all of which contribute to influencing the user experience (**Figure 7**).

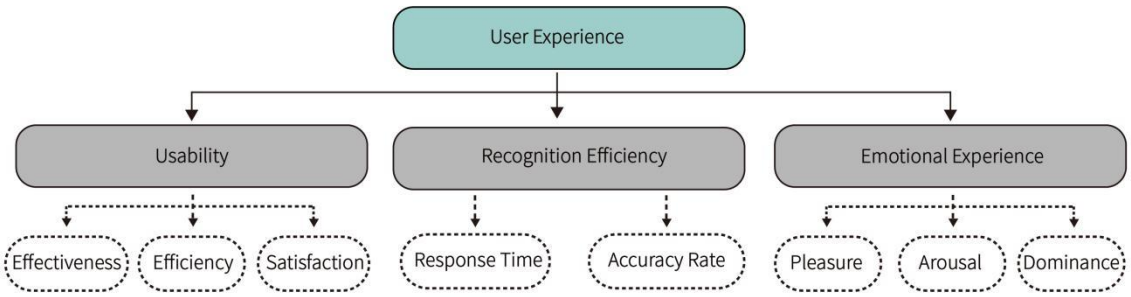


Figure 7. Dependent Variables and Indicators

Data Processing and Analysis Methods

We employed experimental and questionnaire survey methods to obtain relevant data. In the empirical analysis phase, we used the chi-square distribution test to examine whether the sample distribution adhered to the principle of randomness. When the chi-square difference was insignificant, it indicated no proportional deviation between gender and experimental group treatment. We applied descriptive statistics to analyze the response time, accuracy, and questionnaire survey results, including mean, standard deviation, and other parameters. The purpose was to preliminarily understand the primary distribution forms of data results, such as centrality and dispersion. In the difference analysis phase, we utilized factorial analysis of variance (ANOVA) to compare response time and accuracy and examine within-group and between-group effects. The above method aims to test the main effects of grouping, the impact of different auditory recognition materials, and interaction effects. We conducted post hoc analyses on significant interaction effects. We used a one-way ANOVA to compare the mean scores of different experimental treatments in each dimension for the questionnaire survey results, followed by post-hoc tests. Additionally, we performed reliability tests and validity analyses to assess the reliability and effectiveness of the questionnaire survey results (Laugwitz, Held, & Schrepp, 2008).

RESULTS

Participant Characteristics

Among all 80 sample participants in this study, there were 47 males and 33 females (**Table 5**). Divide into 4 groups using random grouping, with 20 people in each group. The first three groups were the experimental group, receiving prompt sounds for Speech, Auditory icon, and Earcons respectively. The fourth group was the control group, not receive any prompt sounds. The cross chi-square test results of gender and different grouping results were insignificant ($p=0.452>0.05$), indicating no gender proportion difference among the groups.

Table 5. Participant Characteristics

Gender	ALL	Speech	Auditory icon	Earcons	Control	p-Value
Men/Women	47/33	14/6	13/7	10/10	10/10	0.452
Men (%)	58.8	70.0	65.0	50.0	50.0	

Response Time and Accuracy of Participants Under Different Short Video Types

Using a mixed experimental design, we compared the differences in response time and accuracy among participants under different cue tone conditions and the differences in response time and accuracy among participants using different listening materials. **Table 6** presents the descriptive statistics of overall response time and accuracy for each group, and the appendix provides the descriptive statistics of different listening materials.

Table 6. Response Time and Accuracy of Participant

Speech			Auditory icon			Earcons			Control		
User	RT	ACC	User	RT	ACC	User	RT	ACC	User	RT	ACC
1	1.04	1.00	21	6.49	0.60	41	4.11	0.40	61	7.18	0.40
2	3.30	1.00	22	2.77	0.40	42	8.15	0.60	62	3.73	0.40
3	1.30	0.80	23	3.57	0.60	43	7.06	0.60	63	15.59	0.40
4	2.54	1.00	24	6.97	1.00	44	4.63	0.80	64	5.11	0.60
5	1.32	1.00	25	2.98	1.00	45	5.42	0.80	65	9.91	0.20
6	2.47	0.80	26	3.37	0.60	46	9.81	0.40	66	14.00	0.40
7	3.86	1.00	27	6.76	0.60	47	4.05	0.80	67	7.37	0.60
8	1.25	0.80	28	6.73	0.80	48	5.30	0.60	68	3.26	0.40
9	0.91	0.80	29	4.75	0.80	49	9.26	0.80	69	3.30	0.20
10	1.06	0.80	30	4.09	0.60	50	4.58	0.60	70	4.78	0.60
11	3.13	0.80	31	2.87	1.00	51	3.20	1.00	71	8.27	0.60
12	1.33	1.00	32	3.28	1.00	52	3.20	0.40	72	14.37	0.40
13	3.69	0.80	33	3.12	0.80	53	8.87	0.20	73	3.03	0.00
14	1.22	1.00	34	2.47	0.20	54	8.74	0.40	74	11.59	0.20
15	3.05	1.00	35	5.11	0.40	55	3.33	0.20	75	14.69	0.20
16	1.58	0.80	36	2.83	0.60	56	4.41	0.40	76	9.60	0.40
17	1.45	1.00	37	2.31	0.80	57	5.40	0.40	77	8.02	0.00
18	2.92	1.00	38	2.53	0.80	58	4.14	0.60	78	4.38	1.00
19	1.88	1.00	39	3.57	0.60	59	6.56	0.60	79	3.83	0.00
20	3.68	0.80	40	2.39	1.00	60	5.92	0.40	80	2.50	0.20
Average	2.15	0.91	Average	3.95	0.71	Average	5.81	0.55	Average	7.73	0.36
SD	1.03	0.10	SD	1.61	0.23	SD	2.15	0.21	SD	4.38	0.25

Using different Auditory Recognition Tasks as intra-group variables and different groups as inter-group variables, a two-factor analysis of variance method was used to compare the main and interactive effects of the two variables on participants' response time. The calculation results (**Table 7**) show that the main effect of different Auditory Recognition Tasks on Response Time is insignificant, $F=0.289$, $p=0.593>0.05$. The main effect of different groups on Response Time is significant, $F=16.805$, $p=0.0001<0.001$. The post hoc test results showed that the Speech group had the shortest Response Time (**Table 8**) ($M\pm SD=2.15\pm1.03$), followed by the Auditory icon group ($M\pm SD=3.95\pm1.61$). Among the experimental group, the Earcons group had the most extended Response Time ($M\pm SD=5.81\pm2.15$), while the control group had significantly higher Response Time than the three experimental groups ($M\pm SD=7.73\pm4.38$).

Table 7. Two-way Analysis of Response Time

	SS	df	MS	F	p-Value	η2
Auditory Recognition Task	0.505	1	0.505	0.289	0.593	0.004
Group	1728.621	3	576.207	16.805	0.0001	0.399
Auditory Recognition Task × Group	210.136	3	70.045	40.098	0.0001	0.613

In addition, the interaction between Auditory Recognition Tasks and Group significantly impacts Response Time, with $F=40.098$ and $p=0.0001<0.001$. In addition, the interaction between Auditory Recognition Tasks and Group significantly impacts Response Time, with $F=40.098$ and $p=0.0001<0.001$. The above results indicate differences in reaction times between different groups of participants on various auditory discrimination tasks. The simple effects test results showed no significant difference among the Auditory Recognition Tasks in the Speech group, and the mean was at a lower level. There was no significant difference in the Auditory Recognition Tasks among the control group, and the mean values were all high. In the Auditory icon group, the response time of News, Announcement, and Sports Auditory Recognition Tasks was significantly lower than that of Cartoon and Music. However, the Music group had the shortest response time in the Earcons group, while the other four Auditory Recognition Tasks had longer response times.

Table 8. Simple Effect Analysis of Response Time

	Speech	Auditory icon	Earcons	Control	Total
1.News	2.20±1.26	3.00±1.21	6.58±2.62	7.64±4.52	4.85±3.56
2.Announcement	2.04±1.23	2.93±1.32	6.63±3.14	7.52±4.89	4.78±3.80
3.Sports	2.15±1.36	3.14±1.54	6.43±2.77	7.54±4.10	4.81±3.46
4.Cartoon	2.19±1.10	5.36±2.38	6.37±3.02	7.88±4.56	5.45±3.64
5.Music	2.18±1.18	5.31±2.43	3.03±0.87	8.06±5.05	4.64±3.65
Total	2.15±1.03	3.95±1.61	5.81±2.15	7.73±4.38	
Post Hoc	1=2=3=4=5	1=2=3<4=5	1=2=3=4>5	1=2=3=4=5	

Using different Auditory Recognition Tasks as intra-group variables and different groups as inter-group variables, a two-factor analysis of variance method was used to compare the main and interactive effects of the two variables on participants' Accuracy (**Table 9**). The calculation results show that the main effect of different Auditory Recognition Tasks on Accuracy is insignificant, $F=0.280$, $p=0.598>0.05$. The main effect of different groups on Accuracy is significant, $F=24.797$, $p=0.0001<0.001$. The post hoc test results showed that the Speech group had the highest Accuracy (**Table 10**) (M±SD=0.91±0.10), followed by the Auditory icon group (M±SD=0.71±0.23). Among the experimental group, the Earcons group had the lowest Accuracy (M±SD=0.55±0.21), while the control group had significantly lower Accuracy than the three experimental groups (M±SD=0.36±0.25).

Table 9. Two-way Analysis of Accuracy

	SS	df	MS	F	p-Value	η2
Auditory Recognition Tasks	0.045	1	0.045	0.280	0.598	0.004
Group	16.768	3	5.589	24.797	0.0001	0.495
Auditory Recognition Task × Group	3.745	3	1.248	7.770	0.0001	0.235

In addition, the interaction between Auditory Recognition Tasks and Group significantly impacts Accuracy, with $F=7.770$ and $p=0.0001<0.001$. This result suggests that participants in different groups differed in Accuracy on various auditory discrimination tasks. The simple effects test results showed no significant difference among the Auditory Recognition Tasks in the Speech group, and the mean was at a high level. There was no significant difference in the Auditory Recognition Tasks among the control group, and the mean values were all low. In the Auditory icon group, the Accuracy of News, Announcement, and Sports Auditory Recognition Tasks was significantly higher than that of Carton and Music. However, within the Earcons group, the Music group demonstrated the highest Accuracy, whereas the remaining four Auditory Recognition Tasks exhibited lower Accuracy.

Table 10. Simple Effect Analysis of Accuracy

	Speech	Auditory icon	Earcons	Control	Total
1.News	0.90±0.31	0.85±0.37	0.45±0.51	0.35±0.49	0.64±0.48
2.Announcement	0.90±0.31	0.85±0.37	0.45±0.51	0.35±0.49	0.64±0.48
3.Sports	0.90±0.31	0.90±0.31	0.50±0.51	0.40±0.50	0.67±0.47
4.Cartoon	0.90±0.31	0.50±0.51	0.40±0.50	0.35±0.49	0.54±0.50
5.Music	0.95±0.22	0.45±0.51	0.85±0.37	0.35±0.49	0.65±0.48
Total	0.91±0.10	0.71±0.23	0.55±0.21	0.36±0.25	
Post Hoc	1=2=3=4=5	1=2=3>4=5	1=2=3=4<5	1=2=3=4=5	

SAM and SUS Score Under Different Short Video Types

Reliability and Validity Analysis

The reliability test results (**Table 11**) of the System Usability Questionnaire and Self-Assessment Manikin show that the overall reliability and reliability values of each dimension of the two questionnaires are above 0.8, indicating good stability and high reliability of the questionnaire survey results.

Table 11. Reliability Test

	Factor	N	Cronbach's α	Total α
SAM	Pleasure	3	0.924	0.856
	Arousal	3	0.932	
	Dominance	3	0.913	
SUS	Validity	3	0.856	0.913
	Efficiency	4	0.929	
	Satisfaction	3	0.897	

In the validity test results (**Table 12** and **Table 13**), the KMO values of both questionnaires were above 0.7, and Bartlett's Test of Sphericity reached a significant level, making the questionnaire very suitable for factor analysis. The classification results of the rotated load value matrix are entirely consistent with the theoretical classification results of the questionnaire. Therefore, the questionnaire has a good level of validity.

Table 12. Validity Test of SAM

	Component		
	1	2	3
SAM1	0.160	0.898	0.119
SAM2	0.265	0.884	0.067
SAM3	0.254	0.916	0.130
SAM4	0.901	0.216	0.095
SAM5	0.899	0.254	0.083
SAM6	0.924	0.197	0.057
SAM7	0.131	0.014	0.915
SAM8	0.132	0.155	0.909
SAM9	-0.037	0.131	0.921
Eigenvalue	2.669	2.617	2.568
Variance	29.656%	29.077%	28.529%
KMO		0.762	
Bartlett's Test		<0.001	

Table 13. Validity Test of SUS

	Component		
	1	2	3
SUS4	0.386	0.154	0.772
SUS5	0.108	0.306	0.834
SUS10	0.187	0.069	0.891
SUS2	0.844	0.267	0.196
SUS3	0.774	0.356	0.223
SUS7	0.859	0.195	0.195
SUS8	0.881	0.267	0.212
SUS1	0.224	0.878	0.141
SUS6	0.315	0.808	0.205
SUS9	0.303	0.855	0.180
Eigenvalue	3.263	2.584	2.350
Variance	32.631	25.837	23.504
KMO		0.858	
Bartlett's Test		<0.001	

Descriptive Statistics and One-way ANOVA

Using different prompt sound groups as the difference test indicators, one-way ANOVA was used to compare the differences in scores of different groups of participants in various dimensions. The descriptive statistics and difference test results are shown in **Table 14**. According to the data in the table, there are significant differences

in scores among different groups for each dimension, with p<0.001. The post hoc test results showed that the Auditory icon group scored the highest in the Pleasure and Arousal dimensions, followed by the Earcons group. The Speech score was the lowest in the experimental group, while the control group scored significantly lower than the experimental group.

In the test results of Dominance, Validity, and Efficiency dimensions, the mean values of each group from high to low are Speech>Auditory icon>Earcons>Control, respectively. In the Satisfaction dimension, the mean values of each group from high to low are Auditory icon>Speech> Earcons>Control.

Table 14. Descriptive Statistics and One-way ANOVA

	Speech	Auditory icon	Earcons	Control	Total	F	p-Value	Post Hoc
Pleasure	5.62±1.71	7.70±0.81	6.72±1.37	4.70±1.25	6.18±1.73	19.450	<0.001	2>3>1>0
Arousal	4.12±1.42	6.08±1.13	5.03±1.40	2.88±1.16	4.53±1.73	22.497	<0.001	2>3>1>0
Dominance	7.15±1.18	6.02±1.30	5.23±1.07	4.19±0.85	5.65±1.54	25.338	<0.001	1>2>3>0
Validity	33.63±1.28	28.50±6.76	23.75±6.95	18.88±6.04	26.19±7.89	24.233	<0.001	1>2>3>0
Efficiency	46.63±2.19	40.75±8.32	33.38±9.74	24.00±8.68	36.19±11.45	31.265	<0.001	1>2>3>0
Satisfaction	29.50±4.84	34.13±1.86	24.63±7.83	16.38±5.99	26.16±8.58	37.111	<0.001	2>1>3>0

DISCUSSION

While numerous studies have addressed accessibility design research for blind individuals related to auditory cue technologies to the best of our knowledge (Bilal et al., 2022; Dulyan & Edmonds, 2010), this study represents the first demonstration that auditory cues can assist blind individuals in recognizing themes in short video content.

Evaluation of the Effectiveness and Usability of Auditory Cues in Short Video Recognition

The results demonstrate a significant performance advantage in the experimental group over the control group in short video recognition, showcasing faster response times and higher accuracy. As shown by Edworthy et al. (2022), assertions regarding the benefits of auditory cues confirm the consistent and reliable performance of auditory cues in identifying themes in short videos. It suggests their potential application across various digital media platforms with similar contexts. This application could significantly enhance the visually impaired community's recognition efficiency, including online streaming, music playback, and audio literature. Regarding enhancing accuracy and reducing response time, speech outperformed auditory icons, followed by earcons, while the control group exhibited the least satisfactory performance. Furthermore, considering subjective scores from the System Usability Scale (SUS) analysis, the experimental group's overall performance surpassed that of the control group. Users rated speech the highest for efficiency and effectiveness but surprisingly lower for satisfaction than auditory icons. Previously, studies reported have shown that blind users increasingly value social entertainment and privacy (Ahmed, Hoyle, Connelly, Crandall, & Kapadia., 2015). They express excitement about the potential of new mobile technologies to enhance privacy and independence. Continuous efforts in designing and introducing differentiated, diverse, and personalized auditory cues are crucial to meeting user demands. Compared to other auditory cues, speech exhibited significantly superior performance in recognition efficiency; the findings are directly in line with the previous findings (Adebiyi et al., 2017), the observation that speech represents the most semantically rich acoustic medium. We argue that auditory cues' semantic solid and intuitive nature plays a crucial role in assisting blind individuals in rapidly and accurately understanding information, constituting a significant factor influencing recognition efficiency. We speculate that prioritizing using voice as a cue in situations where users need to recognize short video content helps them quickly access critical information, especially when managing breaking news such as natural disasters, wars, and pandemics. These enhancements improve the timeliness of information dissemination on short video platforms and increase their practicality in emergencies.

Empirical Study on the Effectiveness of Auditory Cues in Different Short Video Types

The most intriguing discovery lies in our meticulous comparison of the recognition efficiency results across three auditory cues in the experimental group and their alignment with five categories of matched short videos. Firstly, when embedded with auditory icons, short videos in the news, advertising, and sports genres exhibited recognition efficiency second only to Speech. This aligns with the findings of (Maes, Giacofci, & Leman, 2015), who asserted that the advantage of auditory icons lies in their representation of sound sources from real-world environments. The strong metaphorical nature of auditory cues in these three video categories, which depict genuine life scenarios, effectively stimulates users' memory and associative capabilities. Consequently, in future design considerations for matching, short videos depicting real-life situations should be prioritized using auditory icons.

Furthermore, recognition efficiency is notably improved when embedding auditory icons in music-oriented short videos. A popular explanation is that auditory icons' unique rhythm and melodic features establish a high congruence with music videos, facilitating rapid empathetic connection for users (Brewster et al., 1993). While our results align with certain aspects of Adebiyi et al. (2017), discrepancies arise in terms of the High cost of learning time leading to low evaluations. We emphasize that variations in the study subjects may contribute to this inconsistency. In short, if auditory icons are tailored explicitly for music-oriented content, these drawbacks will likely be mitigated to the fullest extent.

Therefore, we can skillfully utilize perceptual features such as semanticity, intuition, metaphorical elements, and consistency to improve the matching effect between auditory cues and short videos while at the same time significantly improving the recognition efficiency of visually impaired people when categorizing short video types.

Significant Association Between Emotion Experience and Short Video Auditory Cues in Blind Users

Based on subjective ratings, we analyzed the impact and usability of auditory cues on emotional experiences across different types of short video content. Surprisingly, the evaluation results exhibited a significant disparity from the trends observed in recognition efficiency, a novel observation not explored in previous research and a key innovation in this study. Notably, across multiple dimensions such as pleasantness, arousal, dominance, effectiveness, efficiency, and satisfaction, the control group scored significantly lower than the experimental group. Users perceived the highest levels of pleasure and arousal when using auditory icons, followed by Earcons. This result broadly aligns with the performance of auditory icons in recognition efficiency. It supports the findings of (Cabral & Remijn, 2019), who asserted that the more closely the content of audio-visual materials is associated with real-world events, the more prominent the role of auditory icons becomes. Accordingly, we posit that auditory icons not only provide blind individuals with familiar auditory information but also stimulate their proactive engagement. To maximize its advantages, researchers should thoroughly research the design and application, considering the habits and actual needs of blind users.

Similarly, the scores for Earcons showcase their uniqueness. Blattner et al. (1989) Highlighted that Earcons can create a creative and artistic space, and Scherer (2004) proposed that highly arousing auditory stimuli can induce relaxation. We believe that Earcons, with their distinctive melodic rhythm and brief and abstract nature, strongly correlated with themes in music and cartoon short videos, effectively enhance the attention and arousal of blind users, reinforcing perceptual intuitiveness and interaction. The above points imply that Earcons can be used as effective auditory cue material for designers and developers and are particularly suitable for matching with music and short cartoon videos.

Interestingly, Speech received the highest ratings in preference, effectiveness, and efficiency, followed by auditory icons, Earcons, and the control group. The findings align with previous findings by Adebiyi et al. (2017), who explicitly stated that Speech enhances the audience's sense of control. This is because Speech directly conveys the core information of the video, avoiding complex decoding. For blind and visually impaired individuals accustomed to the feedback audio "Spearcon" from screen readers, this further strengthens their trust in Speech. Therefore, this concise and clear prompting method enhances information processing efficiency and boosts the autonomy and confidence of blind users. Consequently, Speech may serve as a universal alternative, providing auditory cues for short videos that currently lack matching solutions.

In terms of satisfaction, our random sampling interviews corroborated the results. The highest to lowest scores were auditory icons, Speech, Earcons, and the control group. Auditory icons were highly favored and are likely to be the preferred choice for short video auditory cues. Surprisingly, despite the overall sound performance of Speech, several participants pointed out that it appeared too mechanical and monotonous, expressing privacy concerns. A similar pattern of results was obtained by Garzonis et al. (2008), who discovered that mechanized voice prompts may not be suitable for all applications. In addition to the three auditory cues used in the study, a

future direction could involve designing a composite auditory cue with speech elements as the primary component and incorporating current popular elements. This result emphasizes the importance of integrating diverse auditory cues into short video platforms iteratively to enhance user experience and user retention.

The Matching Scheme and Future Considerations

Table 15. Auditory Cue Matching Scheme for Different Types of Short Videos

Short Video Type	The best auditory cues	The second best auditory cue	Matching Suggestions
News	Speech	Auditory icon	Speech directly conveys the core information, although monotonous, but more in line with the needs of news; Auditory icons have good emotional experience, but the recognition efficiency is slightly slow.
Advertisement	Auditory icon	Speech	Auditory icons sound authentic and are recognised efficiently; speech conveys core information and is monotonous.
Sports	Auditory icon	Speech	Auditory icons come from real sports scenarios and are recognised efficiently. Speech can be an alternative.
Music	Earcons	Auditory icon	Earcons have a consistency with the rhythm and melody of music-based videos for a better emotional experience. Auditory icons may be slightly less effective in this regard.
Cartoon	Earcons	Auditory icon	Earcons allow for high recognition efficiency and emotional experience. Auditory icons will be slightly less.

This study has several limitations to consider when interpreting and applying the results to provide practical guidelines for designing auditory cues for short videos. Firstly, the participants in this study were blind college students from Shandong Special Education Vocational College. Research indicates that young people in the college age group are the most active users of the internet (Zhou, Fong, & Tan, 2014), making the results of this study persuasive for the auditory cue experiences of blind users of short videos. Secondly, cultural differences may impact the acceptance of auditory cues, making the results more reflective of specific cultural user preferences. Lastly, the library of auditory cue materials for short videos did not encompass a broader range of auditory cue types, requiring further development and supplementation. In future research, we plan to expand the sample recruitment, including blind users from different nationalities, age groups, educational backgrounds, and cultural contexts. This expansion aims to diversify the types of auditory cue materials for short videos, conduct in-depth analyses of factors influencing user preferences, and design varied auditory cue-matching solutions for a broader range of short video types. This will become a sustainable, iterative update of the application, which can combine tactile feedback to unlock more excellent research value, supporting the synchronized development of short video platforms in the digital media era (see **Table 15** for details).

CONCLUSION

The issue of information loss resulting from functional conflicts between screen readers and short video software significantly impacts the user experience of blind individuals. We employed a mixed research method to conduct a detailed experimental design and evaluate the innovative application of auditory cues, validating their potential to enhance the user experience of blind individuals on short video platforms across usability, recognition efficiency, and emotional experience. A total of 80 blind college students participated in this study, undertaking a series of experiments and surveys. The results indicate that the incorporation of auditory cue technology significantly improves the user experience of blind individuals, constituting the core contribution of this study. The research also identified optimal design solutions for speech, auditory icons, and Earcons when matching different types of short videos and application scenarios. As digital media continues to evolve, the application will also be able to leverage its sustainable and iteratively updated nature to bring value to more areas. These contributions have expanded the application scope of auditory cue technology and demonstrated its potential to enhance the experience of blind individuals in digital media environments.

AUTHOR CONTRIBUTIONS

Conceptualization, K.N.; methodology, X.W.; software, A.C.; validation, Y.P.; formal analysis, X.W.; investigation, K.N.; resources, X.W.; data curation, A.C.; writing—original draft preparation, X.W.; writing—review and editing, K.N.; visualization, X.W.; supervision, Y.P.; project administration, X.W. All authors have read and agreed to the published version of the manuscript. All authors contributed equally to this work.

FUNDING

This research received no external funding.

INSTITUTIONAL REVIEW BOARD STATEMENT

The research was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Kookmin University Institutional Review Board: KMUIRB (protocol code KUM-202308-HR-373 and the date of approval was 12 November 2023).

INFORMED CONSENT STATEMENT

Informed consent was obtained from all subjects involved in the study.

DATA AVAILABILITY STATEMENT

Not applicable.

ACKNOWLEDGMENTS

The authors would like to sincerely thank all the participants in this study for their time and willingness to share their experiences and feelings. Thanks are also due to all the researchers for their hard work.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

Abraham, C. H., Boadi-Kusi, B., Morny, E. K. A., & Agyekum, P. (2022). Smartphone usage among people living with severe visual impairment and blindness. *Assistive Technology*, *34*(5), 611-618.

Adebiyi, A., Sorrentino, P., Bohlool, S., Zhang, C., Arditti, M., Goodrich, G., & Weiland, J. D. (2017). Assessment of feedback modalities for wearable visual aids in blind mobility. *PloS One*, *12*(2), e0170531.

Ahmed, T., Hoyle, R., Connelly, K., Crandall, D., & Kapadia, A. (2015, April). Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM conference on human factors in computing systems* (pp. 3523-3532). New York, NY: Association for Computing Machinery.

Bilal Salih, H. E., Takeda, K., Kobayashi, H., Kakizawa, T., Kawamoto, M., & Zempo, K. (2022). Use of auditory cues and other strategies as sources of spatial information for people with visual impairment when navigating unfamiliar environments. *International Journal of Environmental Research and Public Health*, *19*(6), 3151.

Blattner, M. M., Sumikawa, D. A., & Greenberg, R. M. (1989). Earcons and icons: Their structure and common design principles. *Human–Computer Interaction*, *4*(1), 11-44.

Brewster, S. A., Wright, P. C., & Edwards, A. D. (1993, May). An evaluation of earcons for use in auditory human-computer interfaces. In *Proceedings of the INTERACT'93 and CHI'93 conference on human factors in computing systems* (pp. 222-227). New York, NY: Association for Computing Machinery.

Brooke, J. (1986). System usability scale (SUS): A quick-and-dirty method of system evaluation user information. *Reading, UK: Digital Equipment Co Ltd*, *43*, 1-7.

Brown, M. L., Newsome, S. L., & Glinert, E. P. (1989). An experiment into the use of auditory cues to reduce visual workload. *ACM SIGCHI Bulletin*, *20* 339-346.

Cabral, J. P., & Remijn, G. B. (2019). Auditory icons: Design and physical characteristics. *Applied Ergonomics*, *78*, 224-239.

Chaudary, B., Pohjolainen, S., Aziz, S., Arhippainen, L., & Pulli, P. (2023). Teleguidance-based remote navigation assistance for visually impaired and blind people—Usability and user experience. *Virtual Reality*, *27*(1), 141-158.

Csapó, Á., Wersényi, G., Nagy, H., & Stockman, T. (2015). A survey of assistive technologies and applications for blind users on mobile platforms: A review and foundation for research. *Journal on Multimodal User Interfaces*, *9*, 275-286.

Dingler, T., Lindsay, J., & Walker, B. N. (2008, June). Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. In *Proceedings of the 14th International Conference on Auditory Display, Paris, France* (pp. 1-6). Retrieved from http://hdl.handle.net/1853/49940

Dinh, P. Q., Dorai, C., & Venkatesh, S. (2002). Video genre categorization using audio wavelet coefficients. In *ACCV 2002: The 5th Asian conference on computer vision* (pp. 1-6). Retrieved from https://staff.itee.uq.edu.au/lovell/aprs/accv2002/accv2002_proceedings/Dinh69.pdf

Donker, H., Klante, P., & Gorny, P. (2002, October). The design of auditory user interfaces for blind users. In *Proceedings of the second Nordic conference on human-computer interaction* (pp. 149-156). New York, NY: Association for Computing Machinery.

Dulyan, A., & Edmonds, E. (2010, November). AUXie: Initial evaluation of a blind-accessible virtual museum tour. In *Proceedings of the 22nd conference of the computer-human interaction special interest group of Australia on computer-human interaction* (pp. 272-275). New York, NY: Association for Computing Machinery.

Edworthy, J. R., Parker, C. J., & Martin, E. V. (2022). Discriminating between simultaneous audible alarms is easier with auditory icons. *Applied Ergonomics*, *99*, 103609.

Encelle, B., Ollagnier-Beldame, M., Pouchot, S., & Prié, Y. (2011, October). Annotation-based video enrichment for blind people: A pilot study on the use of earcons and speech synthesis. In *The proceedings of the 13th international ACM SIGACCESS conference on computers and accessibility* (pp. 123-130). New York, NY: Association for Computing Machinery.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323-327.

Garzonis, S., Bevan, C., & O'Neill, E. (2008, December). Mobile Service Audio Notifications: Intuitive semantics and noises. In *Proceedings of the 20th Australasian conference on computer-human interaction: Designing for habitus and habitat* (pp. 156-163). New York, NY: Association for Computing Machinery.

Garzonis, S., Jones, S., Jay, T., & O'Neill, E. (2009, April). Auditory icon and earcon mobile service notifications: Intuitiveness, learnability, memorability and preference. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1513-1522). New York, NY: Association for Computing Machinery.

Gaver, W. W. (1987). Auditory icons: Using sound in computer interfaces. *ACM SIGCHI Bulletin*, *19*(1), 74.

Guo, J., & Gurrin, C. (2012, November). Short user-generated videos classification using accompanied audio categories. In *Proceedings of the 2012 ACM international workshop on Audio and multimedia methods for large-scale video analysis* (pp. 15-20). New York, NY: Association for Computing Machinery.

Hussain, I., Chen, L., Mirza, H. T., Chen, G., & Hassan, S. U. (2015). Right mix of speech and non-speech: Hybrid auditory feedback in mobility assistance of the visually impaired. *Universal access in the Information Society*, *14*, 527-536.

Iturregui-Gallardo, G., & Méndez-Ulrich, J. L. (2020). Towards the creation of a tactile version of the Self-Assessment Manikin (T-SAM) for the emotional assessment of visually impaired people. *International Journal of Disability, Development and Education*, *67*(6), 657-674.

Jordan, P. W. (2020). *An introduction to usability*. Boca Raton, FL: Crc Press.

Jordan, P. W., Thomas, B., McClelland, I. L., & Weerdmeester, B. (Eds.). (1996). *Usability evaluation in industry*. Boca Raton, FL: CRC Press.

Kaye, D. B. V., Chen, X., & Zeng, J. (2021). The co-evolution of two Chinese mobile short video apps: Parallel platformization of Douyin and TikTok. *Mobile Media & Communication*, *9*(2), 229-253.

Khan, A., & Khusro, S. (2021). An insight into smartphone-based assistive solutions for visually impaired and blind people: Issues, challenges and opportunities. *Universal Access in the Information Society*, *20*(2), 265-298.

Khan, M. A., Paul, P., Rashid, M., Hossain, M., & Ahad, M. A. R. (2020). An AI-based visual aid with integrated reading assistant for the completely blind. *IEEE Transactions on Human-Machine Systems*, *50*(6), 507-517.

Klinge, C., Röder, B., & Büchel, C. (2010). Increased amygdala activation to emotional auditory stimuli in the blind. *Brain*, *133*(6), 1729-1736.

Kuber, R., Hastings, A., & Tretter, M. (2012). Determining the accessibility of mobile screen readers for blind users. In *Proceedings of IASTED conference on human-computer interaction*. https://userpages.umbc.edu/~rkuber/pubs/IASTED2012b.pdf

Lang, P. J. (2019). The cognitive psychophysiology of emotion: Fear and anxiety. In *Anxiety and the anxiety disorders* (pp. 131-170). Abingdon, UK: Routledge.

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *HCI and usability for education and work: 4th symposium of the workgroup human-computer interaction and usability engineering of the Austrian computer society* (pp. 63-76). Berlin, Germany: Springer.

Leplâtre, G., & Brewster, S. A. (2000). Designing non-speech sounds to support navigation in mobile phone menus. In *6th International Conference on Auditory Display (ICAD)* (pp. 190-199). Retrieved from https://eprints.gla.ac.uk/3210/1/icad20001.pdf

Lewis, J. R. (2018). The system usability scale: Past, present, and future. *International Journal of Human-Computer Interaction*, *34*(7), 577-590.

Lewis, J. R., & Sauro, J. (2017). Can I leave this one out? The effect of dropping an item from the SUS. *Journal of Usability Studies*, *13*(1), 28-46.

Liu, X., Carrington, P., Chen, X. A., & Pavel, A. (2021, May). What makes videos accessible to blind and visually impaired people?. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-14). New York, NY: Association for Computing Machinery.

Maes, P. J., Giacofci, M., & Leman, M. (2015). Auditory and motor contributions to the timing of melodies under cognitive load. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(5), 1336.

Mankoff, J., Fait, H., & Tran, T. (2005, April). Is your web page accessible? A comparative study of methods for assessing web page accessibility for the blind. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 41-50). New York, NY: Association for Computing Machinery.

Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. Cambridge, MA: MIT Press.

Mieda, T., Kokubu, M., & Saito, M. (2019). Rapid identification of sound direction in blind footballers. *Experimental brain research*, *237*, 3221-3231.

Mynatt, E. D. (1994, April). Designing with auditory icons: How well do we identify auditory cues?. In *Conference companion on human factors in computing systems* (pp. 269-270). New York, NY: Association for Computing Machinery.

Nees, M. A., & Liebman, E. (2023). Auditory icons, earcons, spearcons, and speech: A systematic review and meta-analysis of brief audio alerts in human-machine interfaces. *Auditory Perception & Cognition*, *6*(3-4), 300-329.

Nelson, P. A., Dial, J. G., & Joyce, A. (2002). Validation of the cognitive test for the blind as an assessment of intellectual functioning. *Rehabilitation Psychology*, *47*(2), 184.

Palmer, S. E., Schloss, K. B., Xu, Z., & Prado-León, L. R. (2013). Music–color associations are mediated by emotion. *Proceedings of the National Academy of Sciences*, *110*(22), 8836-8841.

Redondo, J., Fraga, I., Padrón, I., & Piñeiro, A. (2008). Affective ratings of sound stimuli. *Behavior Research Methods*, *40*, 784-790.

Rokem, A., & Ahissar, M. (2009). Interactions of cognitive and auditory abilities in congenitally blind individuals. *Neuropsychologia*, *47*(3), 843-848.

Roth, P., Petrucci, L., Pun, T., & Assimacopoulos, A. (1999, May). Auditory browser for blind and visually impaired users. In *CHI'99 extended abstracts on Human factors in computing systems* (pp. 218-219). New York, NY: Association for Computing Machinery.

Rowlands, T., Waddell, N., & McKenna, B. (2016). Are we there yet? A technique to determine theoretical saturation. *Journal of Computer Information Systems*, *56*(1), 40-47.

Saariluoma, P., & Jokinen, J. P. (2014). Emotional dimensions of user experience: A user psychological analysis. *International Journal of Human-Computer Interaction*, *30*(4), 303-320.

Šabić, E., Chen, J., & MacDonald, J. A. (2021). Toward a better understanding of in-vehicle auditory warnings and background noise. *Human factors*, *63*(2), 312-335.

Sanderson, P., Wee, A., Seah, E., & Lacherez, P. (2006). Auditory alarms, medical standards, and urgency. In *Proceedings of the 12th International conference on auditory display*. London, UK: University of London.

Scherer, K. R. (2004). Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them?. *Journal of New Music Research*, *33*(3), 239-251.

Shimomura, Y., Hvannberg, E. T., & Hafsteinsson, H. (2010). Accessibility of audio and tactile interfaces for young blind people performing everyday tasks. *Universal Access in the Information Society*, *9*, 297-310.

Snyder, J. (2005, September). Audio description: The visual made verbal. In *International congress series* (Vol. 1282, pp. 935-939). Amsterdam, Netherlands: Elsevier.

Soares, A. P., Pinheiro, A. P., Costa, A., Frade, C. S., Comesaña, M., & Pureza, R. (2013). Affective auditory stimuli: Adaptation of the international affective digitized sounds (IADS-2) for European Portuguese. *Behavior research methods*, *45*, 1168-1181.

Stephan, K. L., Smith, S. E., Martin, R. L., Parker, S. P., & McAnally, K. I. (2006). Learning and retention of associations between auditory icons and denotative referents: Implications for the design of auditory warnings. *Human factors*, *48*(2), 288-299.

Theodorou, P., Tsiligkos, K., Meliones, A., & Filios, C. (2022). A training smartphone application for the simulation of outdoor blind pedestrian navigation: Usability, UX evaluation, sentiment analysis. *Sensors*, *23*(1), 367.

Townsend, J. T., & Altieri, N. (2012). An accuracy–response time capacity assessment function that measures performance against standard parallel predictions. *Psychological review*, *119*(3), 500.

van Someren, M. W., Barnard, Y. F., & Sandberg, J. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London, UK: Academic Press.

Voykinska, V., Azenkot, S., Wu, S., & Leshed, G. (2016, February). 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* (pp. 1584-1595). New York, NY: Association for Computing Machinery.

Walker, B. N., & Kramer, G. (2005). Mappings and metaphors in auditory displays: An experimental assessment. *ACM Transactions on Applied Perception (TAP)*, *2*(4), 407-412.

Wang, Y., Liang, W., Huang, H., Zhang, Y., Li, D., & Yu, L. F. (2021, May). Toward automatic audio description generation for accessible videos. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-12). New York, NY: Association for Computing Machinery.

Wu, Y., Wang, X., Hong, S., Hong, M., Pei, M., & Su, Y. (2021). The relationship between social short-form videos and youth's well-being: It depends on usage types and content categories. *Psychology of Popular Media*, *10*(4), 467.

Zhou, R., Fong, P. S., & Tan, P. (2014). Internet use and its impact on engagement in leisure activities in China. *PloS one*, *9*(2), e89598.