Audio Engineering Society

Convention Paper

Presented at the AES 159th Convention
2025 October 23–25, Long Beach, CA, USA

Exploring the User Experience of AI-Assisted Sound Searching Systems for Creative Workflows

Haohe Liu[1], Thomas Deacon[1], Wenwu Wang[1], Matt Paradis[2], and Mark D. Plumbley[1]

[1]*Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK*
[2]*Research and Development, British Broadcasting Corporation, London, UK*

Correspondence should be addressed to Haohe Liu (haohe.liu@surrey.ac.uk)

ABSTRACT

Locating the right sound effect efficiently is an important yet challenging topic for audio production. Most current sound-searching systems rely on pre-annotated audio labels created by humans, which can be time-consuming to produce and prone to inaccuracies, limiting the efficiency of audio production. Recent works on text and audio multimodal neural networks have led to the development of contrastive language-audio pretraining (CLAP), which learns a shared embedding space for text descriptions and audio samples. Using this idea, we built a CLAP-based sound searching system (CLAP-Search) that does not rely on human annotations. To evaluate the effectiveness of CLAP-Search, we conducted comparative experiments with a widely used sound effect searching platform, the *BBC Sound Effect Library*. Our study evaluates user performance, cognitive load, and satisfaction through ecologically valid tasks based on professional sound-searching workflows. Our result shows that CLAP-Search demonstrated significantly enhanced productivity and reduced frustration while maintaining comparable cognitive demands. We also qualitatively analyzed the participants' feedback, which offered valuable perspectives on the design of future AI-assisted sound search systems.

1 Introduction

Searching for sound effects (SFX) is important in multiple sectors, including film, radio, gaming, and interactive media. High-quality sound effects can enhance storytelling and create immersive sensations in audio-visual productions [1]. For example, Kock and Louven [2] demonstrated that when participants watched films with sound effects, compared to without sound, their perceived immersion increased more than threefold. Furthermore, electroencephalogram (EEG) studies have demonstrated that well-designed sound effects can positively influence audience engagement, contributing to the formation of general mood and stimulating emotional responses [3]. With the important role of sound effects, sound effects retrieval, i.e., searching for sound effects, also becomes a key part of creative and production workflows.

Traditional sound effects libraries such as Freesound [4] and the BBC Sound Effect Library (BBC SFX) [5] have been widely used by audio professionals. BBC SFX contains over $33,000$ sound clips with text annotations, while Freesound allows users to upload sounds with tags and texts for future retrieval. Although these libraries have been widely adopted in current audio production workflows, their effectiveness heavily depends on the

quality and accuracy of manual annotations, which are time-consuming and labor-intensive. For example, an SFX of an engine sound labeled as *"Cars: Bmw 320i Convertible - BMW 320i convertible - drive with top down."* may fail to capture the primary audio event, making it difficult for users to retrieve this sound using queries like "engine". Besides, the searching systems in most traditional sound effect libraries discourage users from using complex text queries [6, 7]. These systems rely on word-matching retrieval, requiring the search query to be directly present in the metadata. As a result, users typically do not expect complex queries involving multiple sound events or specific temporal orders to work effectively. Short query-based searching behaviour is potentially not optimal for fine-grained searching, as the user cannot provide detailed text controls for sound retrieval. The issue of word matching could be improved with semantic text-to-text matching [8] to retrieve a sound with a semantically matched prompt (e.g., retrieve "barn swallow" with query "bird"). Text-to-text retrieval still requires manually annotating each audio file. To avoid this manual work, we explore how to match text queries directly with audio content.

Contrastive language-audio pre-training (CLAP) [9] has recently shown success in learning a shared embedding space between audio and text modalities with paired audio and text encoders, where the audio encoder processes sound signals and the text encoder processes natural language descriptions to create representations that enable cross-modal understanding. CLAP is usually trained on a large-scale audio-text paired dataset and learns a joint embedding space where semantically similar audio and text samples are close in distance. The aligned latent space enables CLAP to match natural language queries with the nearest audio representations to perform text-based audio retrieval [10, 11]. Since CLAP-based sound retrieval does not need human annotations, it can handle large audio collections efficiently. Users can search with natural language instead of keywords, making CLAP potentially more intuitive than traditional search methods. We test how well this approach works for professional audio production workflow.

In this work, we developed a CLAP-based sound searching system (**CLAP-Search**) that enables sound searching with natural language. We compared CLAP-Search against BBC SFX's word-matching search system [5], which we call **BBC-SFX-Search**, in a two-stage user study with audio professionals. For stage one of our study, we developed a prototype sound retrieval system and collected early feedback from participants. This stage focused on gathering user feedback about our prototype system, followed by system improvements. For stage two, we recruited additional professional audio producers and designed sound source retrieval tasks to mimic their daily workflows. Specifically, participants were tasked with finding and selecting sounds described in real radio drama scripts. This paper presents the protocol and results from our second stage of testing, as the first stage focused primarily on prototype development and initial feedback collection without comparative evaluation between systems.

To evaluate and compare the performance of CLAP-Search and BBC-SFX-Search, we collected quantitative and qualitative data in our expert user study, including retrieval task performance, general qualitative feedback, and the modified NASA task load index [12]. The results of our study indicate that participants perceived notable benefits when using the proposed CLAP-Search tool compared to the BBC-SFX-Search.

2 Study Design and Method

This section introduces the overall study design, including the background of CLAP (Section 2.1), the design of our retrieval system (Section 2.2), the protocol of the expert user study (Section 2.3), and the participant background analysis (Section 2.4).

2.1 CLAP-based Audio Retrieval

2.1.1 Model Architecture

The CLAP [9] model is a pre-trained deep neural network model consisting of an audio encoder and a text encoder. The audio encoder $f_{\text{audio}}(\cdot)$ and text encoder $f_{\text{text}}(\cdot)$ process input audio X^{a} and text X^{t}, respectively, to generate embeddings $E^{\text{a}}, E^{\text{t}} \in \mathbb{R}^D$, where D is the dimension of the embeddings, and superscripts $^{\text{a}}$ and $^{\text{t}}$ denote audio and text input. Popular options for the text encoder include BERT [13], RoBERTa [14], and BART [15]. Before performing contrastive model training, the text encoder is usually pre-trained on language modelling tasks, and the audio encoder is usually pre-trained on audio classification or other audio understanding tasks. For contrastive model training, the output of pre-trained text and audio encoders is further mapped to the shared embedding space using a

two-layer multilayer perceptron (MLP). Common architectures for the audio encoder include the pretrained audio neural networks (PANNs) [16], and the hierarchical token-semantic audio transformer (HTSAT) [17]. PANNs is a convolutional neural network (CNN) based model with downsampling and upsampling blocks, initially designed for audio classification, while HTSAT is a transformer-based model with hierarchical token processing. The calculation of audio and text embeddings can be formulated as follows:

$$E^{\mathrm{a}} = \mathrm{MLP}_{\mathrm{audio}}(f_{\mathrm{audio}}(X^{\mathrm{a}})), \quad E^{\mathrm{t}} = \mathrm{MLP}_{\mathrm{text}}(f_{\mathrm{text}}(X^{\mathrm{t}})). \tag{1}$$

In this work, we adopt the pre-trained CLAP from [9], developed based on RoBERTa and HTSAT.

2.1.2 Optimization

The audio and text encoders are trained using a contrastive loss function, which maximizes similarity between paired audio-text samples while minimizing similarity between unpaired samples, to align the audio and text embeddings. The loss ensures that paired audio-text samples are mapped closer in the shared space while non-paired samples are pushed further apart. The contrastive loss is given by

$$L = \frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{e^{E_i^{\mathrm{a}}\cdot E_i^{\mathrm{t}}/\tau}}{\sum_{j=1}^{N} e^{E_i^{\mathrm{a}}\cdot E_j^{\mathrm{t}}/\tau}} + \log\frac{e^{E_i^{\mathrm{t}}\cdot E_i^{\mathrm{a}}/\tau}}{\sum_{j=1}^{N} e^{E_i^{\mathrm{t}}\cdot E_j^{\mathrm{a}}/\tau}}\right), \tag{2}$$

where τ is a learnable temperature parameter, N is the batch size, and E_i^{a} is the embedding i-th audio in a batch. The two logarithmic terms represent audio-to-text and text-to-audio alignment, respectively.

2.1.3 Text-to-Audio Retrieval

Once pre-trained, the CLAP model can perform text-to-audio retrieval by projecting text and audio into the shared embedding space and computing their similarity. For a given text query X^{t}, the text embedding E^{t} is computed using the text encoder followed by an MLP. The system then identifies the audio embedding E^{a} in the dataset that maximizes the following cosine similarity

$$\mathrm{Similarity}(E^{\mathrm{t}}, E^{\mathrm{a}}) = \frac{E^{\mathrm{t}}\cdot E^{\mathrm{a}}}{\|E^{\mathrm{t}}\|\|E^{\mathrm{a}}\|}. \tag{3}$$

Given a text query, the retrieval result is a list of audio clips ranked by similarity score.

2.2 User Interface Design

We developed our CLAP-Search as a web application. As shown in Figure 1, the system offers three core functionalities: (1) searching with a text query, (2) uploading a sound file for search, and (3) utilizing a "search similar sound" function. These options can be used independently or in combination, allowing users to provide multiple inputs to enhance search precision. When multiple queries, such as text and audio files, are used together, the system computes the final query embedding by averaging multiple CLAP embeddings. The "uploading a sound file for search" and "search similar sound" functionalities were added in response to feedback from the prototype development stage, where participants reported challenges in describing sound requirements solely through text. To address this, we designed multimodal search capabilities that combine textual and audio-based inputs.

The audio database used in CLAP-Search is AudioSet [18], comprising 1,912,134 10-second audio clips labeled across 527 classes. During the prototype stage, feedback emphasized the importance of system responsiveness. Based on a widely used Python web demo framework Gradio [19], our initial prototype required more than 6 seconds to render the results. To enhance the system responsiveness, we redeveloped the system with javascript and Flask [20] with searching algorithm optimizations. As a result, the final system achieves an average response time of under 0.5 seconds for a dataset containing nearly two million audio files. Additional features, such as unlimited scrolling and database customization, are also implemented to enhance the user experience. However, for our experiment, users were restricted to searching within AudioSet to ensure consistency in experimental conditions.

2.3 Expert User Study Design

The experimental procedure for the user study comparing CLAP-Search and BBC-SFX-Search was conducted fully online, with participants completing tasks and providing feedback remotely. Communication was via email, which served as the primary method for sharing instructions, addressing questions, and following up after participation.

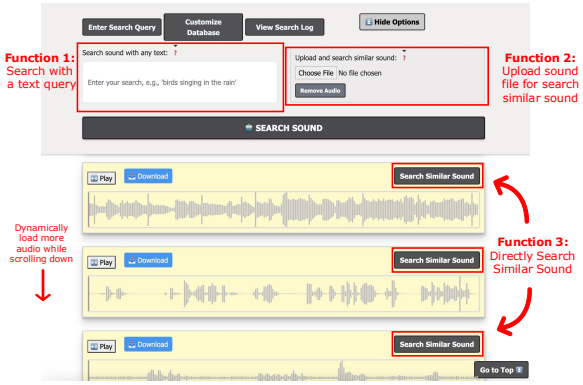

Fig. 1: The user interface of our CLAP-based sound searching system.

2.3.1 Participants and Recruitment

The study recruited participants with both professional experience in audio production and researchers in the multimedia domain. A total of 38 individuals expressed interest in participating, with 20 finally having completed the study. The study was conducted remotely with clear instructions provided via Microsoft Forms[1]. Recruitment was done through professional email networks within the BBC and the Centre for Vision, Speech, and Signal Processing (CVSSP) at the University of Surrey. Participants were given a detailed information sheet outlining the purpose of the study, procedures, and ethical considerations. Informed consent was obtained before their participation, ensuring adherence to ethical standards. To acknowledge their time and effort, all participants received reasonable compensation.

2.3.2 Task Design

The experiment was centered on a "Sound Source Retrieval" task designed to simulate real-world audio production scenarios. Participants are instructed to search for sound effects for pre-defined scripts, such as *Zoo ambience with cheering, children laughing, and people talking*, which were sourced from existing radio drama scripts to reflect real-world production settings. The 12 scripts used in the experiment that indicate target sounds are detailed in Table 1. Each participant performed the task using both the BBC-SFX-Search and CLAP-Search. For each system, participants reviewed the textual description of each script, searched for the most appropriate sound effect, and rated the difficulty of finding a relevant sound on a 0 to 10 scale, where 0 represented *Very easy to find suitable sounds* and 10 represented *Extremely hard to find suitable sounds*. After completing the 12 scripts for each system, participants' task load was assessed using a modified NASA task load index [12]. The original NASA TLX [21] requires participants to first rate six workload dimensions and then perform pairwise comparisons to determine weighting factors for calculating a weighted average score. This modified version simplifies the original NASA TLX by removing the weighting process to reduce participant burden. Instead, the overall workload is calculated as the unweighted average of individual ratings. This approach maintains validity, as research shows a high correlation between weighted and unweighted TLX scores [12]. Additionally, explanations were provided for each dimension to reduce confusion, as detailed in Table 2.

To minimize bias and ensure fairness, the order of system usage is balanced, with 10 participants beginning with the BBC-SFX-Search and the other 10 participants starting with the proposed CLAP-Search system. The same set of sound effect scripts was used for both systems to ensure consistency in evaluation.

Table 1: Scripts for the *Sound Source Retrieval* Tasks

Script ID	Description
S-1	Opens a cupboard.
S-2	Sound of an old wooden rowing boat in a still sea.
S-3	A fire is burning in a stove. A man breaks a piece of wood in two and puts it in the stove.
S-4	Distant bells.
S-5	Coin goes into 1970's phone box. We hear dialling.
S-6	Jazz piano. Footsteps walk on stage.
S-7	The wedding night, in the bed chamber. The heavy oak door slams shut.
S-8	The drip of water in an echoing stone space. The slight murmur of crowds and music far away.
S-9	Sound of taxi pulling up outside a farm.
S-10	A man starts shouting excitedly.
S-11	Walks towards the lifts and presses the button.
S-12	Background: phones, typewriters.

[1]Example: https://github.com/unkown-me/CLAP-UI-VS-BBC-UI

Table 2: Wording used for task load evaluations.

Dimension	Wording Used in the Questionnaire
Mental Demand	How easy or demanding, simple or complex was the task?
Temporal Demand	How much time pressure did you feel in performing the task? How hurried or rushed was the pace of the task?
Performance	How successful were you in accomplishing what you were asked to do?
Effort	How hard did you have to work to accomplish your level of performance?
Frustration	How insecure, discouraged, irritated, stressed, and annoyed were you during the task?

2.3.3 Post-task Surveys

After completing tasks with both systems, participants completed a post-task survey to provide both quantitative and qualitative feedback. Quantitative questions asked participants to rate their experiences on a Likert scale. For instance, participants were asked, *Would you consider using the AI-assisted Sound Searching System in your workflow? (Q1)* and *How well did you perform the task with the AI-assisted Sound Searching System compared to the BBC Sound Effect Library? (Q2)* Optional comment boxes accompanied these questions, allowing participants to elaborate on their ratings. The average completion time of the questionnaire, including the sound source retrieval task and the post-task survey, is 47 minutes 14 seconds.

The qualitative section explored detailed impressions of the AI-assisted system, with questions such as *What did you like (Q3)/dislike (Q4) most about using the AI-assisted Sound Searching System?* Participants were also asked to reflect on scenarios where the CLAP-Search performed better or worse than the BBC-SFX-Search. Additionally, demographic information was collected, including prior experience with the BBC-SFX-Search, frequency of sound library usage, educational background, age, and gender. This data enriched the analysis, providing context for participant feedback. This study received a Favourable Ethical Opinion (FEO) from the University of Surrey Ethics Committee and the Research Integrity and Governance Office (Reference Number: FEPS 22-23 016 EGA) on 27 June 2023. All procedures performed were in accordance with the Committee's guidelines, and participants provided written informed consent prior to participation.

2.4 Participant Background

A total of 20 participants were recruited, with a gender distribution of 55% male and 45% female. Age-wise, 70% of participants are within the $25-34$ age group. The participants are well-educated, with 35% holding master's degrees, 30% with doctoral degrees, and the remaining 35% with a bachelor's degree. 60% of the participants reported previous experience with the BBC-SFX-Search. 95% of the participants report prior experience with general sound effect libraries.

3 Result Analysis

This section compares the AI-assisted system (CLAP-Search) with the BBC-SFX-Search regarding the difficulty of the Sound Source Retrieval task. We analyze both qualitative feedback from participants and quantitative results from script difficulty ratings and the task load index. For clarity, participants and scripts are referred to as P-n and S-n, respectively, where n is the identifier.

3.1 Analysis Method

Due to the ordinal nature of difficulty ratings and repeated measures design, non-parametric statistics are used [22]. We used the Wilcoxon signed-rank test [23] to evaluate overall differences in task difficulty ratings between the UI systems, with alpha level $\alpha = 0.05$, which represents the significance threshold below which we reject the null hypothesis of no difference between systems. The five dimensions of the modified NASA TLX evaluation (mental demand, temporal demand, performance, effort, and frustration) were analyzed using the same test. For examining specific script-level differences between BBC-SFX-Search and CLAP-Search, Bonferroni corrections were used with Wilcoxon signed-rank tests to reduce the family-wise error rate. As we have 12 prompts in total, the Bonferroni-corrected alpha level becomes $\alpha_b = \alpha/12 \approx 0.0042$. Additionally, we report the effect size (r) for each significance test to provide further context and interpret the magnitude of the observed differences. As we have a relatively small sample size, we calculated the effect size with the rank-biserial correlation [24], which measures the proportion of the rank sum difference relative to the total rank sum.

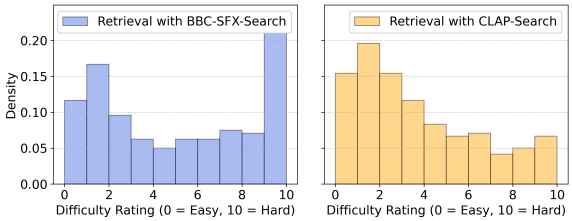

Fig. 2: Overall script difficulty distribution ($p = 1.14 \times 10^{-8}, r = 0.416$).

3.2 Task Performance

Figure 2 and Figure 3 provide insights into the overall difficulty ratings and each script, respectively. Participants consistently found it easier to locate relevant sounds with CLAP-Search, as reflected in significantly lower overall difficulty ratings ($p = 1.14 \times 10^{-8}$, $r = 0.416$). This trend is particularly evident in two specific prompts, S-6 and S-10, where CLAP-Search significantly outperformed BBC-SFX-Search ($p = 0.001, r = 0.848$ and $p = 0.0002, 0.957$, respectively). The CLAP-Search system demonstrated clear advantages over the BBC-SFX-Search in task performance.

3.2.1 Overall Script Difficulty Analysis

Figure 2 provides evidence of the comparative advantage of CLAP-Search (Proposed) over BBC-SFX-Search. The histogram and kernel density estimate (KDE) [25] plots show that difficulty ratings for CLAP-Search are skewed toward lower values, indicating that participants found it easier to locate relevant sounds overall. In contrast, difficulty ratings for the BBC-SFX-Search are more evenly distributed. To statistically validate these observations, we conducted a Wilcoxon signed-rank test, as the data is paired and non-normally distributed, confirmed by the Shapiro-Wilk test ($p < 0.05$). The test result ($p = 1.14 \times 10^{-8}, r = 0.416$) demonstrates a significant difference between the two systems and a moderate effect size.

3.2.2 Individual Scripts Difficulty Analysis

As shown in Figure 3, statistical significance after Bonferroni correction ($p < \alpha_b$) is observed for S-6 (*"Jazz piano. Footsteps walk on stage"*, $p = 0.001, r = 0.848$) and S-10 (*"A man starts shouting excitedly"*, $p = 0.0002, r = 0.957$), where the AI system demonstrated better performance compared to the BBC-SFX-Search. In contrast, no statistically significant differences were observed for the rest of the scripts ($p > \alpha_b$). To look into the statistical significance we got on S-6 and S-10, we compare the search result of these two prompt on BBC-SFX-Search and CLAP-Search. The top result when searching for *"Jazz piano"* and *"shouting excitedly"* on BBC-SFX-Search are labelled with *"Pianos: Comedy - One piano dragged along"* and *"Animals - Airedale panting excitedly."*, respectively, which have low relevance. However, CLAP-Search can give the audio required in the top result.

3.3 Preference Ratings and Qualitative Feedback

As shown in Figure 4, participants provided a range of ratings between 1 and 10 for Q1 and Q2, with average scores of 7.65 and 7.4, respectively. Most participants rate in the mid-to-high range (6–10), indicating a general preference for the usability of CLAP-Search and a likelihood of adopting it for future use. For instance, P-19 (Q1: 10, Q2: 10) gives CLAP-Search a score of 10 for both questions and states *"I can always find audio that I want to find ... The AI-assisted system is easy to use and always works for me."*

Many participants highlighted the improved relevance of CLAP-Search results compared to the BBC-SFX-Search. For example, P-10 (Q1: 9, Q2: 9) commented *"Although the AI option struggled with a couple of searches, I always felt like the results indicated an understanding of approximately what I was trying to achieve. The BBC search, time and time again, gave results that were hugely irrelevant (e.g., when searching for a man shouting excitedly [S-10], BBC gave me a full page of horse noises). Most impressive was the few occasions when the top result from the AI search was almost exactly what I had in mind!"* Similarly, P-18 (Q1: 8, Q2: 7) notes *"I think the AI system gave more relevant search results. The BBC system, for example, when searching for 'Walks towards the lifts and presses the button,' returned top results like bird and water sounds, which were completely irrelevant."*

Another commonly mentioned benefit of CLAP-Search is efficiency. Several participants described how the AI system reduced the mental overhead of searching for sounds. P-8 stated, *"Felt I could get there much faster with the AI-assisted. ... I can only imagine with an even bigger sample library it will only become*

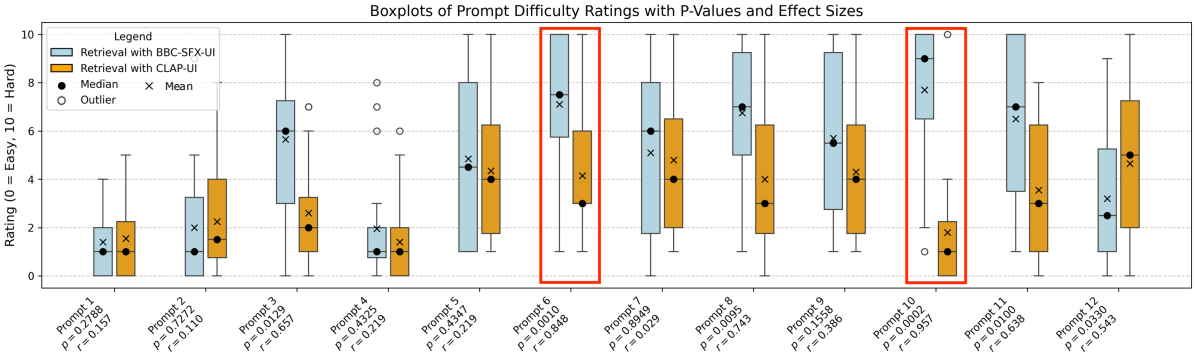

Fig. 3: Difficulty rating for each of the sound effect scripts. Red boxes mark the statistically significant results after the Bonferroni correction.

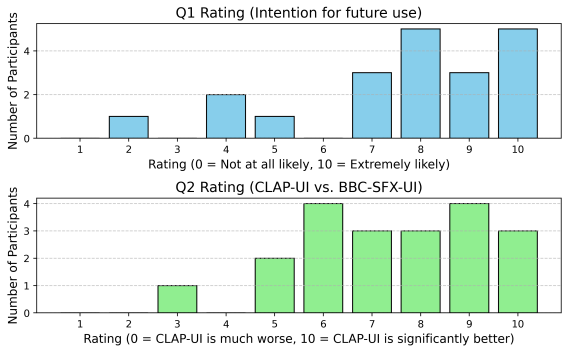

Fig. 4: Participant ratings for Q1 and Q2.

stronger." Similarly, P-9 commented, *"Yes, I can generate multiple audio clips from similar scenes using a few keywords quickly and efficiently."*

The creative flexibility offered by CLAP-Search is another area of strength. Participants appreciated how CLAP-Search allowed them to explore sound design conceptually. P-1 observed that the AI-assisted system *"allows me to be more creative in the prompts I'm asking for,"* while P-17 highlighted how the system increased creative possibilities by enabling searches for sounds that might not exist in real life, stating, *"[CLAP-Search] can help me find some sounds that you imagine, which may not be relevant or possible in real life, increasing my creative ability."*

The "Search Similar Sound" feature was also well-received, enabling users to refine their results more effectively. P-9 emphasized its utility in research requiring multiple sounds with similar environmental atmospheres: *"The search similar sound function is very useful for me, because my research requires video-level semantic relationship learning using multiple sounds."*

While the AI system demonstrated clear advantages, several limitations were noted. A recurring concern was the variability in audio fidelity. Participants involved in broadcast or radio drama production indicated that the AI-generated sounds sometimes lacked the polish of BBC-SFX-Search. For instance, P-16 commented, *"The output of the AI-assisted Sound Searching System has poorer audio quality and may not be as useful as the BBC Sound Effect Library for radio drama tasks."* However, it is essential to note that this limitation is not inherent to the CLAP-Search system itself but rather stems from the quality of the AudioSet samples used in this study. The system is compatible with higher-fidelity datasets, and its performance could be further enhanced with improved audio data.

Citing a lack of written descriptions as a major drawback, P-3 (Q1: 2, Q2: 5) rated the potential future usage of the AI system with a low score of 2, as text descriptions are the primary reference for selecting sound effects. They remarked *"often I select effects to audition based on the written description. The AI-assisted search engine didn't have this."* At the same time, P-3 gives a score of 5 for Q2 and notes that *"I use a search catalog of a local FX drive daily to find sound effects. The BBC SFX search is always too slow! I have to work extremely quickly."*. This indicates that both CLAP-Search and BBC-SFX-Search were equally unimpressive for their needs. This highlights the importance of system responsiveness and the importance of textual descriptions in sound-searching workflows.

Participants also highlighted that written descriptions in BBC-SFX-Search facilitated quicker selection, as they could filter sounds without listening to each one. P-2 remarked, *"I didn't like the fact that there weren't any descriptions (as there are in the BBC system), which*

meant you couldn't discount a sound without actually listening to it." This sentiment was echoed by P-6, who suggested, *"It would be better if the search results also come with a text description like the BBC platform. That helps with our usage even faster."* This limitation could potentially be alleviated with audio captioning systems [26, 27], which will be part of our future studies.

Furthermore, AI-assisted sound searching may not suit all scenarios. For example, P-1 (Q1: 5, Q2: 3) commented *"As an addition to the sound library, it's fine. However, for my use, I need specific sounds—e.g., bird calls or locations around the world. I can't see a scenario where we'd use AI sound for those."* Since the CLAP-Search system cannot specify the geographic location of target sounds, it may not be a helpful tool in such cases.

Another drawback was the inconsistent handling of complex scripts. While CLAP-Search excelled at simple queries, multi-element scripts often posed challenges. P-5 observed that *"When fusing more than 2 sound requirements, the AI system always ignore one requirement. Like for the jazz piano + footstep, it generates excellent jazz piano but totally ignores the footsteps."* P-14 echoed similar concerns, noting that both systems struggled with scenarios involving combinations of distinct elements. This issue may stem from limitations in the search algorithm but may also be because the specific sound effect combinations are originally absent in the dataset. However, as the CLAP-Search system is easily scalable to larger datasets, it holds greater potential for covering a broader spectrum of sound scenes in the future.

Lastly, some users observed redundancies in the results returned by CLAP-Search. For instance, P-3 remarked, *"In one example, it offered me the same sound effect multiple times. The variety of suggestions was too limited."* This highlights the need for de-duplication [28] as a potentially important step in CLAP-Search to minimize repeated results and enhance the diversity of suggestions.

These findings suggest that while CLAP-Search demonstrates considerable potential in enhancing the efficiency and creativity of sound retrieval workflows, further refinements in metadata, user interface, data curation, and complex query handling are recommended to maximize its potential.

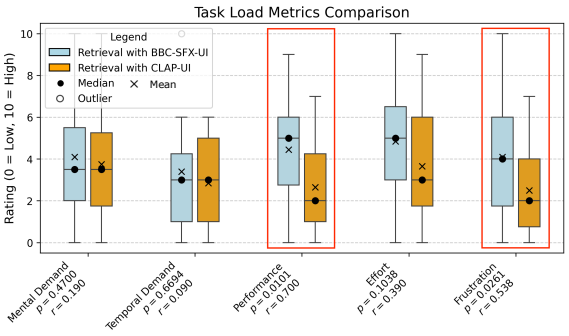

Fig. 5: Task load index evaluation result.

3.4 Workload Analysis

The workload experienced by participants during the Sound Source Retrieval task was evaluated using the dimensions defined in the modified NASA-TLX [12], encompassing five dimensions: mental demand, temporal demand, performance, effort, and frustration. The result is shown in Figure 5.

3.4.1 Mental and Temporal Demands

No significant differences were observed in mental demand ($p = 0.470$) and temporal demand ($p = 0.669$) between CLAP-Search and BBC-SFX-Search, indicating that participants perceived the cognitive complexity and time pressure of tasks to be comparable.

3.4.2 Performance and Effort

Performance ratings showed a significant improvement for CLAP-Search over BBC-SFX-Search ($p = 0.010, r = 0.700$), with participants consistently reporting that CLAP-Search provided more accurate and relevant results. This improvement aligns with qualitative insights, such as P-10's observation that *"the BBC library often gave irrelevant results, while CLAP-Search provided more useful options."* While effort did not achieve statistical significance ($p = 0.1038, r = 0.390$), a trend toward reduced effort with CLAP-Search was evident. Participants attributed this to the system's ability to retrieve results efficiently with minimal requerying.

3.4.3 Frustration

The frustration dimension showed a significant reduction with CLAP-Search compared to BBC-SFX-Search ($p = 0.026, r = 0.538$). Participants frequently cited

frustration with the BBC-SFX-Search due to irrelevant results and redundant outputs, whereas CLAP-Search's semantic relevance and natural language interface alleviated these issues.

The findings on the task load index indicate that CLAP-Search can alleviate specific aspects of workload, such as frustration and perceived performance while maintaining similar mental and temporal demands as the traditional BBC-SFX-Search. This balance suggests that CLAP-Search will be effective at integrating into existing workflows without introducing significant cognitive burdens.

4 Discussions and Limitations

This study demonstrates that CLAP-Search outperforms traditional word-matching systems, showing improved task performance, reduced frustration, and enhanced creative flexibility through natural language querying. Participants reported that results "indicated an understanding of approximately what I was trying to achieve" compared to BBC-SFX-Search's often irrelevant outputs. However, several limitations can guide future development. First, AudioSet's lower-fidelity audio affected user perceptions compared to professional-grade BBC SFX content, though CLAP-Search can work with higher-quality datasets. Second, the absence of descriptive metadata forced participants to manually audition all results, and complex multi-element queries (e.g., "jazz piano with footsteps") sometimes received only partial fulfilment. Finally, the lack of explainability sometimes can leave users uncertain about query refinement, and our limited participant pool warrants broader validation. Despite these constraints, the advantages demonstrated by CLAP-Search in semantic relevance, efficiency, and creative flexibility position it as a promising advancement with clear pathways for improvement through enhanced datasets, metadata integration, and interface refinement.

5 Conclusions

This study compared a CLAP-based sound searching system (CLAP-Search) with the word-matching-based sound searching system implemented in the BBC Sound Effects Library (BBC-SFX-Search). Our result demonstrates that CLAP-Search offers significant advantages in performance, frustration reduction, and creative flexibility, largely due to its natural language querying and better semantic relevance. However, limitations such as reliance on lower fidelity datasets, lack of metadata, and challenges in handling complex prompts highlight areas for improvement. Despite these limitations, CLAP-Search represents a promising advancement in sound retrieval technologies, with the potential to streamline workflows, reduce cognitive demands, and inspire creativity in audio production. Future work should address these limitations by enhancing metadata integration, improving query modeling, and leveraging higher-quality datasets.

6 Acknowledgments

This research was partly supported by the British Broadcasting Corporation Research and Development (BBC R&D), Engineering and Physical Sciences Research Council (EPSRC) Grant EP/T019751/1 "AI for Sound", and a PhD scholarship from the Centre for Vision, Speech and Signal Processing (CVSSP), Faculty of Engineering and Physical Science (FEPS), University of Surrey. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising. At the time of publication, Haohe Liu is affiliated with Meta, Redmond, USA, and Mark D. Plumbley is affiliated with King's College London, London, UK.

References

[1] Grimshaw, M., Tan, S.-L., and Lipscomb, S. D., "Playing with sound: The role of music and sound effects in gaming," in *The Psychology of Music in Multimedia*, Oxford University Press, 2013.

[2] Kock, M. and Louven, C., "The power of sound design in a moving picture: An empirical study with emoTouch for iPad," *Empirical Musicology Review*, 13(3-4), pp. 132–148, 2018.

[3] Kwon, Y.-S., Lee, J., and Lee, S., "The impact of background music on film audience's attentional processes: Electroencephalography alpha-rhythm and event-related potential analyses," *Frontiers in Psychology*, 13, p. 933497, 2022.

[4] Font, F., Roma, G., and Serra, X., "Freesound Technical Demo," in *ACM International Conference on Multimedia*, 2013.

[5] British Broadcasting Corporation, "BBC Rewind Sound Effects," `https://sound-effects.bbcrewind.co.uk/`, 2024.

[6] Baeza-Yates, R., Hurtado, C., Mendoza, M., and Dupret, G., "Modeling user search behavior," in *Third Latin American Web Congress*, IEEE, 2005.

[7] Weck, B. and Font, F., "The Language of Sound Search: Examining User Queries in Audio Search Engines," *arXiv preprint:2410.08324*, 2024.

[8] Jiang, J.-Y., Zhang, M., Li, C., Bendersky, M., Golbandi, N., and Najork, M., "Semantic text matching for long-form documents," in *The World Wide Web Conference*, pp. 795–806, 2019.

[9] Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S., "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[10] Elizalde, B., Zarar, S., and Raj, B., "Cross modal audio search and retrieval with joint embeddings based on text and audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4095–4099, 2019.

[11] Koepke, A. S., Oncescu, A.-M., Henriques, J. F., Akata, Z., and Albanie, S., "Audio retrieval with natural language queries: A Benchmark Study," *IEEE Transactions on Multimedia*, 25, pp. 2675–2685, 2023.

[12] Hart, S. G., "NASA-task load index (NASA-TLX); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pp. 904–908, 2006.

[13] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint:1810.04805*, 2018.

[14] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint:1907.11692*, 2019.

[15] Lewis, M., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint:1910.13461*, 2019.

[16] Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D., "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, pp. 2880–2894, 2020.

[17] Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., and Dubnov, S., "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[18] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M., "AudioSet: An ontology and human-labeled dataset for audio events," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 776–780, 2017.

[19] Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., and Zou, J., "Gradio: Hassle-free sharing and testing of ML models in the wild," *arXiv preprint:1906.02569*, 2019.

[20] Grinberg, M., *Flask Web Development*, O'Reilly Media, Inc., 2018.

[21] Hart, S. G. and Staveland, L. E., "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Advances in Psychology*, 52, pp. 139–183, 1988.

[22] Wobbrock, J. O. and Kay, M., "Nonparametric statistics in human–computer interaction," *Modern Statistical Methods for HCI*, pp. 135–170, 2016.

[23] Woolson, R. F., "Wilcoxon signed-rank test," *Encyclopedia of Biostatistics*, 8, 2005.

[24] Cureton, E. E., "Rank-biserial correlation," *Psychometrika*, 21(3), pp. 287–290, 1956.

[25] Chen, Y.-C., "A tutorial on kernel density estimation and recent advances," *Biostatistics & Epidemiology*, pp. 161–187, 2017.

[26] Liu, X., Mei, X., Huang, Q., Sun, J., Zhao, J., Liu, H., Plumbley, M. D., Kilic, V., and Wang, W., "Leveraging pre-trained BERT for audio captioning," in *European Signal Processing Conference*, pp. 1145–1149, IEEE, 2022.

[27] Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M. D., Zou, Y., and Wang, W., "WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, pp. 3339–3354, 2024.

[28] Google Inc., "Using structured data for search result deduplication," 2014.