QUALITY AUDIO PROTOTYPING: A PROTOTYPE SYSTEM FOR UNIFIED SOUND RETRIEVAL AND PROCEDURAL GENERATION.

Nelly Garcia, Aditya Bhattacharjee, Gabryel Mason-Williams, Israel Mason-Williams, Emmanouil Benetos, Joshua Reiss

Queen Mary University of London, King's College of London
London, UK
n.v.a.garcia-sihuay@qmul.ac.uk

ABSTRACT

Sound design workflows frequently oscillate between time-consuming library searches and the complexity of procedural synthesis, with practitioners typically relying on disconnected tools to address each challenge separately. This paper introduces Quality Audio Prototyping (QuAP), a working prototype that unifies content-based audio retrieval and procedural sound generation within a single interface, reducing the procedural distance between a narrative concept and its sonic realisation. The goal of QuAP is to support the design of non-speech, non-music sounds in audiovisual productions. QuAP integrates a similarity-based retrieval engine with real-time procedural audio models, complemented by a rule-based assistant that provides perceptually informed parameter guidance, offering definitions and recommendations derived from empirical optimisation rather than requiring prior synthesis knowledge. Preliminary evaluation confirms the viability of this approach: subjective assessment demonstrated statistically significant quality improvements in five of six embedded synthesis models, and an encoder ablation study established the preferred retrieval architecture on a sound effect dataset. A user evaluation with 16 practitioners confirmed the tool's workflow utility, with all participants agreeing that the parameter assistant preserved creative agency while lowering the barrier to procedural interaction.

1. INTRODUCTION

Sound design encompasses the creation and manipulation of environmental sounds, layered textures, and sound elements that compose the auditory dimension of audiovisual productions, distinct from dialogue and music [1]. To create a soundscape, practitioners rely on extensive sound libraries, Foley [2] recordings, and production stems, navigating large-scale collections under time and creative pressure.

The sound design workflow combines technical and aesthetic skills in equal measure. On the technical side, tasks such as sound retrieval, source separation, and parameter manipulation represent recurring bottlenecks where signal processing and machine learning approaches offer potential for optimisation [3]. However, existing tools typically address these bottlenecks in isolation: retrieval systems, synthesis engines, and organisational tools operate as disconnected applications, requiring practitioners to context-switch between environments and interrupting creative flow during the exploratory phase of a project.

Table 1: Comparison between traditional sound design workflows and the proposed QuAP system.

Feature	Traditional	QuAP (Proposed)
Search Method	Text / Metadata	Hybrid (Text + Embeddings)
Retrieval Speed	Manual browsing	FAISS-accelerated vector search
Sound Creation	External plugins	Integrated procedural models
Workflow State	Disconnected / linear	Unified / iterative

This paper introduces **Quality Audio Prototyping (QuAP)**, a working prototype application developed in JUCE [1] that unifies sound retrieval and procedural generation within a single interface. QuAP addresses the friction of the exploratory design phase through three integrated components:

- **Hybrid Retrieval:** Combines text-based metadata search with content-based similarity retrieval, using a MobileNet architecture deployed via ONNX Runtime to extract deep acoustic embeddings, indexed and queried using FAISS[4] for near-instantaneous results.
- **Procedural Synthesis:** Embeds six optimised synthesis models: spanning physical, modal, and subtractive approaches, enabling the generation of new sounds from scratch.
- **Intelligent Guidance:** An embedded rule-based assistant provides perceptually informed parameter recommendations for each procedural model, indicating the ranges within which synthesis parameters were perceived as most realistic by participants in a subjective evaluation. This lowers the barrier to procedural interaction and supports rapid creative exploration without removing designer agency.

By unifying similarity-based retrieval, procedural generation, and hybrid layering within a single environment, QuAP addresses the procedural distance between a narrative concept and its sonic realisation. Preliminary validation confirms the viability of this approach: the optimisation of procedural models demonstrated statistically significant quality improvements in five of the six embedded categories ($p < 0.05$); the ablation study established MobileNetV3 as the preferred encoder for sound effect retrieval, outperforming a ResNet18-IBN baseline on mean average precision (mAP: 0.449 vs. 0.412) on the FSD50K dataset; and a user evaluation with 16 practitioners confirmed the tool's workflow utility, with all participants agreeing that the parameter assistant preserved creative agency within an increasingly tool-fragmented workflow.

[1] https://juce.com

2. RELATED WORK

The evolution of sound design tools has developed along two primary axes: the ease of retrieval from large-scale audio databases, and the flexibility of interactive real-time sound synthesis. In professional practice, these two axes remain largely disconnected. Practitioners typically navigate between separate retrieval platforms, synthesis tools, generative solutions, and organisational tools to construct a single workflow. A significant limitation of existing generative and retrieval tools is their predominant orientation toward music production. The majority of large-scale audio models are trained on datasets with strong musical bias thus prioritising tonal, rhythmic, and melodic structure [5, 6]. This music-centric bias poorly generalises to the nonlinear, texture-based, and physically motivated characteristics of sound effects and environmental audio [7]. Therefore this bias represents a critical gap for sound design practitioners, whose requirements centre on parametric control, timbral specificity, and narrative fit rather than musical coherence. Commercial tools such as Splice[2], Krotos[3], ElevenLabs[4], and iZotope[5] address individual aspects of this pipeline but do not offer an integrated solution. QuAP is positioned as an attempt to unify these axes within a single environment, reducing the context-switching that interrupts the exploratory phase of sound design work.

2.1. Content-Based Audio Retrieval

Content-based audio retrieval refers to the task of retrieving audio recordings based on their acoustic characteristics of a query audio clip. The broader literature encompasses several related problem formulations. High-specific audio fingerprinting systems aim to identify near-identical matches within large databases [8, 9]. Closely related tasks include music version identification, which detects different renditions of the same musical work [10, 11], and query-by-vocal-imitation, where users retrieve sounds by producing vocal imitations of the target audio [12, 13].

Recent work has increasingly approached these problems through learned audio embeddings that map sounds into a metric space where semantically similar audio segments are located nearby. Convolutional neural networks trained for large-scale audio tagging have proven particularly effective as general-purpose audio encoders [14]. In particular, lightweight architectures such as MobileNet [15] provide a favorable balance between representational capacity and computational efficiency, making them well suited for real-time, interactive and on-device audio retrieval applications. Given its utility in related tasks [16, 17], we use MobileNet as the encoder architecture for our audio embedding system.

2.2. Procedural Audio and Intelligent Interfaces

Parallel to retrieval, procedural audio offers a dynamic alternative to static samples. By employing mathematical models to simulate physical or electronic sound-generating processes, designers can achieve a degree of parametric variability and interactive control that sample-based approaches cannot provide [18]. In this sense, procedural audio functions as a form of rule-based synthesis, in which timbral and temporal characteristics are governed by adjustable parameters rather than fixed recordings.

However, the high dimensionality of synthesis parameter spaces presents a significant barrier to adoption. Recent work in intelligent sound engineering has explored mapping complex synthesis parameters to lower-dimensional, perceptually meaningful representations, making procedural tools more accessible to practitioners without deep technical expertise [19]. QuAP main design is presented in Table 1 it builds on this direction by embedding a parameter assistant that constrains interaction to perceptually validated ranges, derived from the optimisation process described in Section 3.

In professional practice, output quality is not always the primary objective, a sound that serves the narrative and fits the production context is often sufficient, particularly under deadline pressure. Current procedural models are therefore well-suited as a foundation for layering and post-processing, rather than as a replacement for recorded material [3]. Additionally as AI-driven tools become increasingly prevalent in creative workflows, practitioners require greater transparency and control over their assets, particularly regarding the provenance of training data and the legibility of generative outputs [20].

3. METHODOLOGY

The development of QuAP required the optimisation of two independent but complementary components: the procedural audio models that underpin the synthesis engine, and the audio embedding model that drives similarity-based retrieval. For the procedural models, optimisation was guided by a feature-driven bottleneck framework [21], which identifies the acoustic features most salient for perceptual classification, with results validated through subjective evaluation. For the retrieval component, encoder architecture selection was informed by performance on content-based audio retrieval benchmarks, detailed in the ablation study presented in Section 6. Together, these methodological decisions define the technical foundation of the integrated system described in Section 4.

4. SYSTEM DESIGN

The system integrates four primary components: (1) an offline embedding and indexing pipeline, (2) a real-time query inference module, (3) an interface for interaction with procedural models, and (4) a hybrid layering stage enabling the user to combine retrieved library samples with procedurally generated audio. An overview of the system architecture is presented in Figure 1.

The system operates across two stages. In the **offline stage**, audio samples from the user's library are processed by a deep embedding model to produce fixed-dimensional feature representations, which are stored in a FAISS-indexed database. To maintain plugin responsiveness, this extraction and indexing process is executed asynchronously in a background thread, with progress notifications surfaced in the interface. In the **online stage**, a query audio clip is processed through the same embedding model, and the resulting embedding is compared against the indexed database to retrieve acoustically similar samples. More details are discussed in the next subsections.

[2] https://splice.com
[3] https://krotos.com
[4] https://elevenlabs.com
[5] https://www.izotope.com

Model	Synthesis Type	Parameters	Optimisation
Fire	Additive, modal, subtractive	Lapping, hissing, crackling, intensity	Reverb, compression and EQ in low frequencies
Explosion	Additive, modal, physically informed	Rumble, air, dust (decay, amount)	Reverb, distortion and compression
Jet	Additive, physically informed	Turbine, burn, speed	Reverb, distortion and EQ in high frequencies
Rocket	Subtractive, modal	Duration, chamber resonance	Reverb and compression
Helicopter	Physically informed	Period, frequency, distance	Reverb and distortion
Gun	Additive, modal, physically informed	Shell frequency, shell decay	Reverb and compression

Table 2: Procedural audio models embedded in QuAP, with synthesis type, exposed parameters, and applied optimisations.

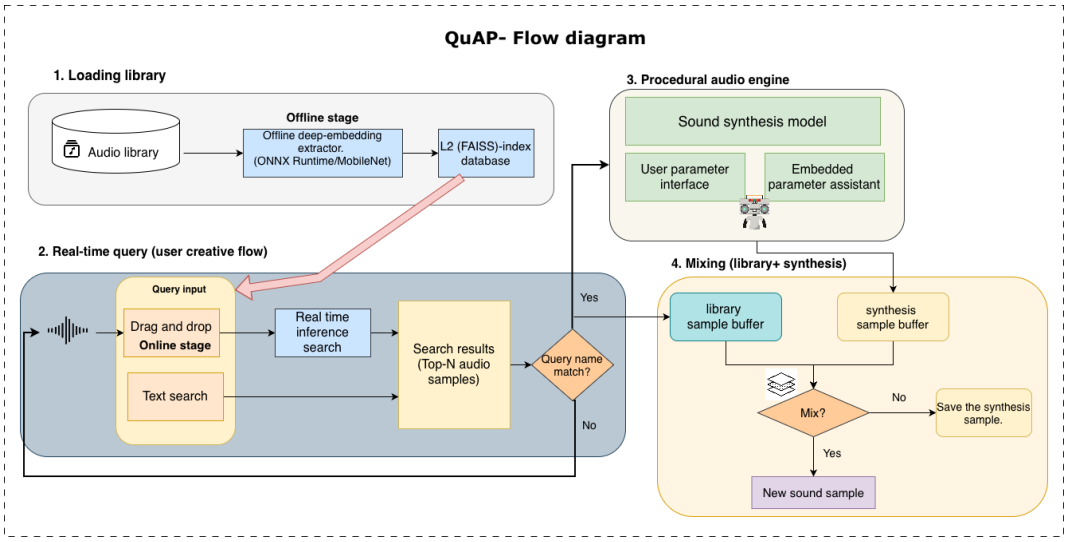

Figure 1: QuAP system architecture. The loading library stage (top left) processes the user's audio library through an offline deep-embedding extractor and FAISS indexing backend, executed asynchronously in a background thread. The real-time query stage (bottom left) accepts drag-and-drop or text input, performing similarity search against the indexed database. Where procedural models are available, the system routes to the procedural audio engine (top right) for parameter-based sound generation. The resulting synthesis output may then be layered with the retrieved library sample in the mixing stage (bottom right), allowing the designer to blend static and dynamic material within a single interface.

4.1. Procedural Audio Models

QuAP embeds six procedural audio models, available through the Nemisindo synthesis engine[6], and inspired by the design principles of Farnell [18]. The models, their synthesis types, exposed parameters, and applied optimisations are summarised in Table 2, classified following [19].

4.2. Procedural Model Optimisation

A central design challenge in embedding procedural models within a creative tool is determining which synthesis parameters are perceptually meaningful. That is, which low-level acoustic features, when modified, are most likely to influence how a user perceives the output. To address this, we applied a feature-driven bottleneck framework [21] to the six sound categories embedded in QuAP. Using low-level audio features extracted via Essentia [22] from the 6KSFX dataset [23], the framework identifies the top-K acoustic features that most reliably discriminate between recorded and synthetic audio for each category, as illustrated in Figure 2. Based on these feature importance results, targeted post-production effects were applied to each model, and the exposed parameter ranges recommended by the embedded assistant were derived from these findings. The optimisations applied to each category are listed in Table 2.

A MUSHRA subjective evaluation [24] with 20 participants was conducted to validate the optimisations. The study was conducted in accordance with institutional ethical guidelines. Participants provided informed consent prior to participation, and all data were anonymised before analysis. Statistically significant quality improvements were observed in five of the six categories ($p < 0.05$), as summarised in Table 3. The Rocket model did not reach statistical significance ($p = 0.08$), indicating that post-production effects alone (such as equalization, reverb, distortion or compression) are insufficient for this category and that synthesis-level (code based) modifications may be required. The Jet category was excluded from the final subjective evaluation, as participants consistently noted that the optimised outputs remained perceptually too synthetic relative to recorded reference material. Full optimisation details are documented at the project website[7] and in a companion study currently under review [25].

[6] `https://nemisindo.com`

[7] `https://saop-project.netlify.app`

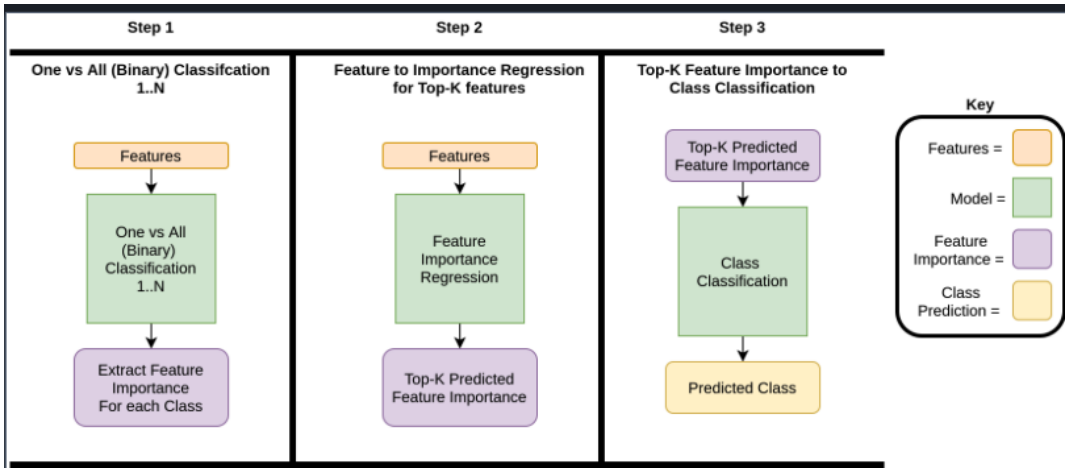

Figure 2: Feature-driven bottleneck framework used to optimise procedural audio model parameters. Steps 1–3 correspond to one-versus-all classification, top-K feature importance regression, and top-K feature classification respectively.

Table 3: MUSHRA subjective evaluation results for the six procedural models embedded in QuAP. Default refers to the unoptimised synthetic sample. Best Opt. reports the highest mean score across optimisation variants.

Model	Default	Best Opt.	F	p
Fire	28.85	40.45	15.23	< 0.001*
Explosion	56.40	52.55	11.39	< 0.001*
Helicopter	41.10	51.20	28.74	< 0.001*
Rocket	37.85	49.20	1.75	0.08
Gun	35.95	45.60	3.54	< 0.001*
Jet	Excluded — see limitations			

* $p < 0.05$

4.3. Audio Embedding Model

Audio similarity is modeled using a convolutional embedding network based on the MobileNet architecture [15]. The backbone is initialized using weights pre-trained for large-scale audio tagging on the AudioSet [26] dataset using the EfficientAT framework [27]. These models are trained to perform multi-label sound event classification and provide strong general-purpose audio representations for downstream tasks.

The network is further fine-tuned using supervised contrastive learning [28], which encourages embeddings from the same semantic class to cluster together while separating embeddings from different classes in the latent space. This is done using the FSD50K dataset [29] which provides hierarchical class labels following the AudioSet ontology.

4.4. Offline Embedding Extraction and Indexing

When a user loads a sound library into QuAP, each audio file is processed through the embedding network to generate a fixed-dimensional vector representation. These embeddings are computed offline and stored in a FAISS [4] vector database that supports fast nearest-neighbor search. This indexing step significantly reduces the computational cost of subsequent similarity queries. Because sound libraries may contain thousands of audio files, the embedding extraction and indexing procedures are executed asynchronously in a background thread. The user interface provides progress feedback during this process.

4.5. Real-Time Query and Retrieval

For similarity search, users can drag and drop an audio sample into the plugin interface. The query sample is processed through the embedding network to generate a query embedding vector. The system then performs a nearest-neighbor search against the indexed embedding database to retrieve the most acoustically similar samples. The retrieved results are ranked according to embedding similarity and displayed within the plugin interface.

To ensure compatibility with different hardware configurations, the system performs a lightweight calibration step at initialization. This step estimates the computational capacity of the host machine and adjusts inference parameters to maintain interactive performance.

4.6. Plugin Interface and Procedural Interaction

The interface dynamically adapts when procedural audio models are available for certain sound categories. In these cases, the plugin exposes additional parameter controls that allow users to manipulate the procedural sound generation model directly. The embedded assistant also recommends optimal parameter ranges whenever available, based on known perceptual knowledge [25] . This design allows QuAP to function both as a sound retrieval system and as an exploratory sound design assistant.

An example of the GUI is illustrated 3, where the procedural audio panel, the embedded assistant and the library panel can be seen.

4.7. Embedded Parameter Assistant

The embedded parameter assistant is implemented as a rule-based guidance system, distinct from large language models or generative AI. For each of the six procedural models embedded in QuAP, the assistant serves two complementary functions. First, it displays

Figure 3: QuAP graphical user interface. The left panel displays the audio library search results, supporting both text-based and drag-and-drop similarity queries. The right panel exposes the procedural audio model controls, with the embedded assistant providing perceptually informed parameter recommendations.

Table 4: Ablation study comparing encoder architectures for audio retrieval.

Encoder	mAP ↑	NDCG ↑
ResNet18-IBN	0.412	0.625
MobileNetV3 (ours)	**0.449**	**0.656**

pre-defined parameter ranges derived directly from the feature-driven bottleneck optimisation described in Section 3, specifically, the configurations under which participants in the MUSHRA subjective evaluation reported the highest perceptual similarity to recorded reference material. Second, it provides a plain-language description of what each synthesis parameter controls, allowing practitioners to understand the perceptual effect of a slider without requiring prior knowledge of the underlying synthesis model. This addresses a known barrier to procedural audio adoption [**?**], where the high dimensionality and technical naming of synthesis parameters often discourages non-expert users. The assistant does not generate or adapt recommendations dynamically; rather, it surfaces empirically validated ranges and parameter descriptions as guidance overlaid on the synthesis controls, allowing the designer to make informed adjustments while retaining full creative control over the output.

5. USER EVALUATION

Ethical approval for this study was obtained in accordance with institutional guidelines. All participants provided informed consent prior to taking part, were informed of their right to withdraw at any time, and all responses were anonymised before analysis.

QuAP was evaluated by 16 participants following up their interest in the survey: 8 sound designers, 4 audio researchers, and 4 music producers. Evaluation sessions were conducted both in-person and remotely. In-person sessions employed a cognitive walkthrough methodology [30], in which the evaluator observed how participants interacted with the interface during a guided task. The primary task required participants to locate a fire sound sample within their library using the similarity search, and subsequently generate a new variant using the procedural audio model. For remote sessions, participants were provided with the standalone plugin, a written instruction page, and a video tutorial [8].

Participants were asked to assess workflow utility, deployment context, and suggested improvements. 75% of the participants found QuAP useful and potentially helpful in their workflows. The plugin and supporting materials are available in the supplementary documents accompanying this paper.

Transcripts from all sessions were analysed using thematic analysis [31], yielding five themes and thre relevant codes are presented in Table 5. *Workflow Integration* captures how participants positioned QuAP within their existing pipeline, particularly regarding library navigation and the identification of unmet tooling needs. Participants highlighted the hybrid layering feature as a notable strength, enabling sound samples to be augmented directly with procedural model parameters within a single environment. *Interface* reflects responses to the usability and responsiveness of the plugin, with participants suggesting improvements including a visible list of available procedural models and greater prominence of the embedded assistant within the interface. *Creative Agency* encompasses the extent to which participants felt the tool supported rather than replaced their creative decision-making, particularly the role of the parameter assistant in maintaining human control over synthesis outputs. *Quality & Assessment* addresses how participants evaluated the procedural models: quality concerns were minimal, with participants instead emphasising the utility of layering synthetic outputs with recorded library material as a practical creative strategy. *Future Potential* captures expressed interest in extending the tool's capabilities, including broader model coverage and deeper integration within production workflows.

6. ABLATION STUDY

The choice of audio embedding architecture is guided by two design considerations: the model should be lightweight enough for efficient deployment, while maintaining strong performance on related content-based audio retrieval tasks. Our primary encoder is based on MobileNetV3, motivated by the findings of Greif et al. [16], who demonstrate that fine-tuning MobileNetV3 yields state-of-the-art performance in query-by-vocal-imitation retrieval.

As a baseline, we employ a ResNet18-IBN [32] encoder. This architecture represents a lightweight variant of convolutional backbones previously used in tasks such as cover song identification [32] and music sample identification [33]. To ensure a fair comparison, the baseline model is trained using the same data preprocessing and training procedure as the MobileNet-based encoder described in Section 4.3.

Both models are evaluated on a held-out test split of the FSD50K dataset. Retrieval performance is measured using mean average precision (mAP) and normalized discounted cumulative gain (NDCG). Table 4 summarizes the comparative performance of the two architectures.

7. DISCUSSION

The results across the three evaluation components — procedural, model optimisation, encoder ablation, and user evaluation — collectively support the core premise of QuAP: that unifying retrieval and synthesis within a single, interpretable environment is both technically viable and practically valuable for sound design workflows.

[8] `https://quap.netlify.app`

Table 5: Thematic framework derived from user evaluation responses.

Theme	Code	Description
Workflow Integration	Management	Library organisation and asset handling
	Similarity	Acoustic search and retrieval relevance
	Areas of opportunity	Identified gaps in current tooling
Interface	GUI	Interface usability and visual feedback
	Computational Process	System performance and responsiveness
Creative Agency	Human-in-the-loop	Designer control over generative output
	Augmentation	AI as creative support, not replacement
	Creation	Subjective aesthetic judgement
	Narrative	Fit of sound within storytelling context
Quality & Assessment	Quality	Perceived realism of synthesis models
	Assessment	Evaluation of output quality
	Limitation	Current scope and implementation gaps
Future Potential	Potential	Perceived future utility of the tool
	Possibilities	Suggested extensions and new use cases
	Waveform	Visual audio feedback and representation
	Storyboard	Integration within broader production pipeline

7.1. Procedural Model Quality and Perceptual Utility

The MUSHRA evaluation results presented in Section 4.2 reveal an important distinction between perceptual realism and workflow utility. While several models, most notably Fire (best opt: 40.45) remain within the 'slight resemblance' range of the MUSHRA scale, this does not preclude their practical value in a professional context. As confirmed by the user evaluation in section 5, practitioners do not always require acoustically perfect outputs; a sound that serves the narrative and provides a viable foundation for layering is often sufficient, particularly during the exploratory phase of a project.

The Rocket model's failure to reach statistical significance ($p = 0.08$) indicates that post-production effects alone are insufficient for certain synthesis categories, and that meaningful improvement would require modifications at the synthesis algorithmic level. This is an acknowledged limitation of the current implementation, and highlights a broader challenge in procedural audio optimisation: the ceiling of post-production enhancement is bounded by the quality of the underlying synthesis model.

The Explosion category presents a notable result: the best optimisation (52.55) performed below the default synthetic sample (56.40), suggesting that the applied post-production effects introduced perceptual artefacts rather than improvements for this category. This underscores the importance of category-specific optimisation strategies rather than uniform application of effects.

The feature-driven bottleneck framework provided a further contribution beyond model improvement: interpretability. By identifying which acoustic features most significantly differentiate real from synthetic audio, the framework generates actionable insight for both developers and practitioners. These insights directly informed the parameter assistant embedded in QuAP's interface, which constrains interaction to perceptually validated ranges , translating machine-derived feature importance into human-readable guidance that supports critical listening rather than delegating decisions to automation.

7.2. Encoder Selection and Retrieval Performance

The ablation study confirms that MobileNetV3 outperforms the ResNet18-IBN baseline on both mAP (0.449 vs 0.412) and NDCG (0.656 vs 0.625) on the FSD50K held-out test split. While the margin is modest, the choice of MobileNetV3 is additionally motivated by its computational efficiency — a critical consideration for real-time deployment within a Digital audio work station environment where latency and random access memory (RAM) constraints are non-trivial. The use of FSD50K, a sound event dataset, as the fine-tuning corpus is a deliberate departure from music-centric training data, ensuring that the embedding space is optimised for the nonlinear, texture-based characteristics of sound effects rather than tonal or rhythmic structures.

7.3. Human-in-the-Loop Design

The user evaluation finding that all participants found the embedded assistant reduced the barrier to procedural interaction, with one noting it *"gives the sensation of keeping human-in-the-loop"*, reflects a broader design principle that guided QuAP's development. The goal was not to automate sound design, but to make the technical complexity of synthesis accessible without removing the designer's creative agency. By exposing perceptually meaningful parameters and providing guidance grounded in empirical feature importance, QuAP encourages practitioners to engage with and develop their critical listening, using the tool as a creative instrument rather than a generative shortcut.

The four participants who identified limitations citing scope constraints or unaddressed bottlenecks which highlighted the boundaries of the current pilot implementation. QuAP currently supports six sound categories; expanding the procedural model library and improving the synthesis quality of categories like Rocket and Jet represents the primary direction for future development.

8. CONCLUSION

This paper introduced Quality Audio Prototyping (QuAP), a working prototype application that unifies content-based audio retrieval and procedural sound generation within a single environment. By combining a MobileNetV3-based similarity search engine with six optimised procedural audio models and an embedded parameter assistant, QuAP addresses the workflow fragmentation that characterises current sound design practice.

Three complementary evaluations validate the system. The MUSHRA subjective evaluation confirmed statistically significant quality improvements in five of the six embedded sound categories following feature-driven optimisation, demonstrating that procedural models (while not acoustically perfect) are perceptually viable as a foundation for layering and creative exploration. The ablation study established MobileNetV3 as the preferred encoder for sound effect retrieval, outperforming the ResNet18-IBN baseline on both mAP and NDCG on the FSD50K dataset. The user evaluation confirmed practitioner interest in the integrated approach, with 10 of 16 participants identifying QuAP as a viable workflow tool, and all participants agreeing that the parameter assistant lowered the barrier to procedural interaction while preserving creative agency.

The central contribution of QuAP is not the replacement of any single existing tool, but the unification of retrieval and synthesis within an interpretable, human-in-the-loop environment. To create a tool that gives designers knowledge of what each parameter means and the creative space to explore it. Future work will focus on expanding the procedural model library, improving synthesis quality for categories that did not reach statistical significance, and conducting a larger-scale user study to assess long-term workflow integration.

9. REFERENCES

[1] D. Sonnenschein, *Sound Design: The Expressive Power of Music, Voice, and Sound Effects in Cinema*. Studio City, CA: Michael Wiese Productions, 2001.

[2] V. Ament, *The Foley Grail: The Art of Performing Sound for Film, Games, and Animation*. Taylor & Francis, 2009. [Online]. Available: https://books.google.co.uk/books?id=sbcpAAAAQBAJ

[3] P. Kamath, F. Morreale, P. L. Bagaskara, Y. Wei, and S. Nanayakkara, "Sound designer-generative ai interactions: Towards designing creative support tools for professional sound designers," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024.

[4] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE transactions on big data*, vol. 7, no. 3, pp. 535–547, 2019.

[5] A. Agostinelli, T. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "Musiclm: Generating music from text," 01 2023.

[6] H. Liu *et al.*, "Audioldm: Text-to-audio generation with latent diffusion models," *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[7] M. Cámara, F. Marcos, A. R. Bargum, C. Erkut, J. Reiss, and J. L. Blanco, "Neural audio synthesis for sound effects: A scope review," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 427–445, 2026.

[8] A. Wang, "An industrial-strength audio search algorithm," in *ISMIR*, 2003.

[9] S. Chang, D. Lee, J. Park, H. Lim, K. Lee, K. Ko, and Y. Han, "Neural audio fingerprint for high-specific audio retrieval based on contrastive learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3025–3029.

[10] J. Serra, X. Serra, and R. G. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, no. 9, p. 093017, 2009.

[11] F. Yesiler, J. Serrà, and E. Gómez, "Accurate and scalable version identification using musically-motivated embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 21–25.

[12] D. S. Blancas and J. Janer, "Sound retrieval from voice imitation queries in collaborative databases," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

[13] Y. Zhang and Z. Duan, "Imisound: An unsupervised system for sound query by vocal imitation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2269–2273.

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

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[16] J. Greif, F. Schmid, P. Primus, and G. Widmer, "Improving query-by-vocal imitation with contrastive learning and audio pretraining," *arXiv preprint arXiv:2408.11638*, 2024.

[17] N. Garcia-Sihuay, D. Martilla, Y. Zhong, M. Liu, J. Groff, B. Swanson, and J. Reiss, "Querying by vocal imitation challenge," *Journal of the Audio Engineering Society*, 2026, under review.

[18] A. Farnell, *Designing Sound*. MIT Press, 2010. [Online]. Available: https://books.google.co.uk/books?id=JsuMEAAAQBAJ

[19] D. Menexopoulos, P. Pestana, and J. Reiss, "The state of the art in procedural audio," *Journal of the Audio Engineering Society*, vol. 71, pp. 825–847, 12 2023.

[20] J. Herington, R. Borasi, B. Guerrero, D. Miller, B. Koerner, Y. J. Han, Z. Borys, and R. Roberts, "Musicians' ethical concerns about ai: an interview study," *AI and SOCIETY*, vol. 41, pp. 1075–1088, 09 2025.

[21] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," 2020. [Online]. Available: https://arxiv.org/abs/2007.04612

[22] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "Essentia: an open-source library for sound and music analysis," in *Proceedings of the 21st ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2013, p. 855–858.

[23] N. Garcia and J. Reiss, "6ksfx synth dataset," no. arXiv:2501.17198, 2025.

[24] C. Mendonça and S. Delikaris-Manias, "Statistical tests with mushra data," 05 2018.

[25] N. Garcia, G. Mason-Williams, I. Mason-Williams, and J. Reiss, "Synthetic audio optimization by feature-driven importance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2026, under review.

[26] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 776–780.

[27] F. Schmid, K. Koutini, and G. Widmer, "Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation," in *ICASSP 2023-2023 IEEE international Conference on acoustics, Speech and signal processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[28] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[29] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[30] A. Blandford, D. Furniss, and S. Makri, "Qualitative HCI research: Going behind the scenes," *Synthesis Lectures on Human-Centered Informatics*, vol. 9, no. 1, pp. 1–147, 2016.

[31] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[32] X. Du, Z. Yu, B. Zhu, X. Chen, and Z. Ma, "Bytecover: Cover song identification via multi-loss training," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 551–555.

[33] A. Bhattacharjee, I. Meresman Higgs, M. Sandler, and E. Benetos, "Refining music sample identification with a self-supervised graph neural network," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*. ISMIR, 2025.