Evaluation of Deep Audio Representations for Semantic Sound Similarity

Recep Oguz Araz, Dmitry Bogdanov, Pablo Alonso-Jiménez, Frederic Font
Music Technology Group, Universitat Pompeu Fabra
Barcelona, Spain
{recepoguz.araz, dmitry.bogdanov, pablo.alonso, frederic.font}@upf.edu

***Abstract*—Navigating large audio collections presents a significant challenge due to the intricate nature of sound properties and the varied needs of users. To enhance user experience, audio-sharing platforms offer a sound similarity function, which leverages vector-based representations of audio clips to facilitate the retrieval of sounds. This study evaluates the retrieval performances of one manually engineered audio representation and thirteen deep audio embeddings in the semantic sound similarity task. By employing a diverse range of models, our research investigates the effects of utilizing different input modalities and training objectives. In the process, we explore various design choices for integrating embeddings into sound similarity systems. Our evaluation is based on objective ranked performance metrics that incorporate sound classes and sound families, complemented by preliminary subjective assessments. We observe that the multimodal models using audio and language modalities outperform audio-only models by a significant margin, which in turn outperform audio and image models. Notably, the state-of-the-art models on the sound event classification task are not the top-performing models on the semantic sound similarity task. In addition, our findings in embedding processing methods and similarity search functions provide insights broadly applicable to information retrieval systems across different modalities.**

***Index Terms*—sound similarity, semantic similarity, audio information retrieval, content-based retrieval, query-by-example, QbE, deep embeddings, representation learning, multimodal representation learning**

I. Introduction

Sounds can create an immersive sensation by themselves or complement other senses. As a result, a large number of professionals and hobbyists work with audio clips, using numerous types of sounds in their productions. For accessing audio clips, online platforms such as Freesound[1] serve as crucial resources. However, their vast volume of audio presents significant challenges to retrieving sounds [1]–[3].

To address these challenges, Audio Information Retrieval (AIR) systems have been developed to aid users in accessing audio clips within extensive collections. An effective AIR system should store each item efficiently in terms of digital size and allow accurate and fast retrieval, which becomes challenging for large collections. As the amount of information on the internet keeps growing, improving the search results of AIR systems becomes increasingly important.

This work is supported by "IA y Música: Cátedra en Inteligencia Artificial y Música" (TSI-100929-2023-1) funded by the Secretaría de Estado de Digitalización e Inteligencia Artificial and the European Union-Next Generation EU, under the program Cátedras ENIA.

[1]https://freesound.org/

One way to retrieve sounds in an AIR system is by providing an example sound to which similar sounds are needed. This method, termed Query-by-Example (QbE), retrieves similar items to a query item [4]. For QbE retrieval, an AIR system stores vector representations for each item and uses a vector space operation such as the inner product to rank the similarities between pairs of items [5].

Considerable efforts have been directed towards the creation of representations driven by acoustic and semantic motivations [4]–[8]. Nonetheless, disparities between acoustically similar sounds and their semantic equivalence have been recognized, indicating that sounds similar in acoustic properties do not necessarily share semantic similarity [9].

Recently, deep representation learning models for audio understanding have demonstrated the ability to create semantic representations of sounds through sound event classification tasks [10]–[12]. Subsequently, sound class labels have been used in contrastive learning to semantically enrich the representations [13]. Building on this progress, researchers have proposed deep representation learning models for obtaining representations tailored to sound similarity systems [14]–[16].

Later approaches have started leveraging natural language instead of relying on class labels, further enriching the audio representations semantically. Notably, contrastive language-audio pre-training (CLAP) approaches have enabled new capabilities such as cross-modal audio retrieval and audio captioning [17]–[19]. This progress has been further augmented by the introduction of large language models, enabling new capabilities such as audio reasoning [20].

However, recent studies have pointed out the shortcomings of the CLAP approach in leveraging natural language beyond keywords and capturing the temporal sequence of sound events [21]. Despite such limitations, improvements in deep representation learning continue to show promise. In this work, we evaluate thirteen state-of-the-art models with different input modalities and training objectives. In the process, we explore multiple design choices for building sound similarity systems that utilize representations from these models, commonly referred to as deep embeddings.

The rest of the paper is organized as follows: Section II details the various sound similarity systems that we build, Section III describes our evaluation methodology throughout these experiments, Section IV reports the results of the experiments, and Section V provides final remarks to conclude the work.

II. Sound Similarity Systems

In this section, we describe the methodology used in building sound similarity systems that utilize the audio representations of Freesound or the embeddings taken from numerous representation learning models. Initially, we outline the process of replicating the manually engineered audio representation of Freesound. Then, we report the chosen models and compare their relevant properties. We follow by exploring various design choices to find the optimal setting for each embedding. This includes examining the methods for extracting the embeddings, applying post-processing to the embeddings, and implementing the similarity search. The code to run the experiments is available on our project repository.[2]

A. Freesound's audio representation

Freesound extracts and processes a mixture of low- and high-level features for audio representation [1]. Specifically, it uses a subset of the features available in the FreesoundExtractor[3] algorithm of the Essentia audio analysis library [22]. This set includes perceptually motivated features such as Mel Frequency Cepstral Coefficients (MFCC), dissonance, and pitch, alongside others like spectral centroid and spectral rolloff, which are not based on perception.

In the original implementation, features are extracted at the frame level and aggregated to the clip level by computing statistical measures. Each clip-level feature is independently scaled to the range $[0, 1]$ by finding its minimum and maximum value across all audio clips. The scaled features are then concatenated into a vector of 846 dimensions. Subsequently, this vector undergoes dimensionality reduction to 100 dimensions using Principal Component Analysis (PCA). We believe that dimensionality reduction is applied to reduce the embedding size, a practical consideration in real-world systems.

B. Deep audio representations

There have been many developments in the audio understanding field. Previously, the focus was on training models that use only the audio modality as an input. Recently, other modalities, such as image or language, have been introduced to complement the audio modality. There is a notable variation in the training objectives employed by these models, ranging from supervised classification to supervised contrastive learning and extending to self-supervised learning techniques. Although there are models designed for sound similarity systems, to the best of our knowledge, they are not publicly available [14]–[16]. Therefore, we chose thirteen of the available models trained with different objectives. Below, we provide a review of the models used in our experiments. However, this list is by no means complete.

1) Audio-only Models: VGGish [10] trains a convolutional neural network (CNN) for audio classification on the YouTube100M [23] dataset using the video titles. YAMNet [24] is a CNN trained to classify the audio clips in the AudioSet dataset [25], which contains about 1.8M videos of 10 seconds long duration. FSD-SINet [12] implements signal processing-inspired pooling methods to improve a CNN for audio classification. They evaluate the model's performance on the FSD50K dataset [26]. BEATs [27] is a transformer architecture pre-trained with the discrete label prediction self-supervised task on AudioSet. After pre-training, it is fine-tuned on AudioSet.

2) Multi-modal Models: CLAP2022 [17] jointly trains an audio encoder and a text encoder using contrastive learning, leveraging both natural language and audio supervision on 128K audio-text pairs. They use a pre-trained PANNs [28] as the audio encoder. Similarly, LAION-CLAP [18] employs this approach on 2.5 million audio-text pairs but with different encoders, where the audio encoder is a pre-trained HTS-AT [29]. The authors of CLAP2022 later use the same audio encoder as LAION-CLAP and train a new model, CLAP2023, on 4.6M pairs, along with additional improvements [19]. Pengi [20] also trains the same audio encoder, but specifically to prompt a pre-trained GPT-2 base model with the question-answering task, utilizing 3.4 million audio-text pairs.

Wav2CLIP [30] uses knowledge distillation to leverage the CLIP [31] model originally trained to match images with text descriptions. As the audio encoder, they pre-train a ResNet18 [32] on 200K videos of the VGGSound [33] dataset. AudioCLIP [34] also uses a pre-trained CLIP model but with a multi-stage training process. They first pre-train an ESResNeXt [35] audio encoder on AudioSet and then train the unfrozen CLIP model jointly with the audio encoder on AudioSet. Conversely, ImageBind [36] aligns the latent spaces of 5 distinct modalities, including audio and language, separately to the latent space of the image modality using contrastive learning. Its audio encoder employs the AST [11] transformer-based architecture, and the audio data is sourced from AudioSet.

OpenL3 [37] is a publicly available implementation of the L3 [38] model which uses a CNN as an audio encoder. It is trained on various subsets of AudioSet to identify the temporal correspondence between audio and image segments in videos, known as audio-visual correspondence (AVC) learning. CAV-MAE [39] is also trained on the AudioSet with a combination of the masked auto-encoder framework and contrastive learning. It utilizes the AST for encoding audio.

C. Integrating deep audio representations

We obtain audio embeddings by extracting the outputs of the models' intermediate or final layer based on their training objectives. Depending on the model, these raw embeddings may be on the frame level and require aggregating in time to yield clip-level embeddings. Models such as YAMNet and FSD-SINet provide frame-level embeddings, necessitating aggregation over time. Conversely, models such as LAION-CLAP or CAV-MAE produce clip-level embeddings for variable-length audio clips, bypassing the need for aggregation.

The prevalent approach for frame aggregation is using averaging [10], [37]. However, this approach may smear

[2]https://github.com/raraz15/semantic-sound-similarity

[3]https://essentia.upf.edu/reference/std_FreesoundExtractor.html

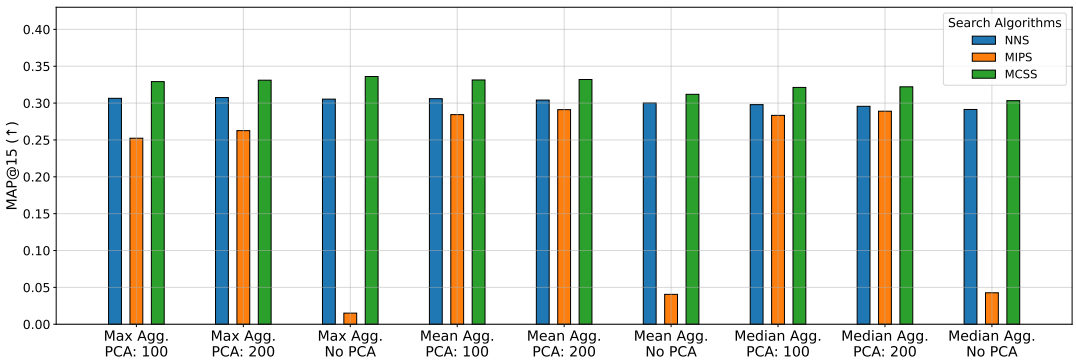

Fig. 1. Semantic sound similarity performances of systems that use FSD-SINet VGG42-tlpf for audio representation. The x-axis contains embedding processing parameters, while the y-axis corresponds to the average of the class-wise MAP@15 values. The color of a block indicates the similarity search algorithm that was used with the corresponding embedding processing system.

important characteristics in longer audio clips due to the smoothing effect of averaging. In contrast, the median of a distribution is robust to the number of its elements, and the maximum is less susceptible to the weight of the distribution. Therefore, we experiment with mean, median, and maximum frame aggregation methods.

Each model produces embeddings with varying dimensions, as detailed in Table II. To ensure that all models are evaluated under similar conditions and to be consistent with the original Freesound implementation, we explored the impact of dimensionality reduction using PCA after creating the clip-level embeddings. In our experiments, we considered three approaches: keeping 100 principal components, keeping 200 principal components, and not applying PCA.

The vectors resulting from this process served as the final representations of audio clips. For models that produce frame-level embeddings, nine different setups were tested, reflecting combinations of frame aggregation and dimensionality reduction. For models producing clip-level embeddings, only three setups were tested since no aggregation was required.

It is common to provide different variations of a model. The same architecture trained with different datasets, or the modified versions of an architecture are commonly provided in the literature. For example, the FSD-SINet model has four different versions. For such models, we independently looked for the best design choices for each variation. After the variations were tuned to their best performance, the variation that outperformed the rest was chosen to represent that model.

D. Similarity Search

Besides optimizing the embedding processing procedure for each model, we also searched for the best similarity search function. We experimented with Maximum Inner Product Search (MIPS), Maximum Cosine Similarity Search (MCSS), and Nearest Neighbor Search (NNS) with the Euclidean distance. The choice to evaluate both MIPS and MCSS was driven by the domain-specific nature of similarity scores and the potential relevance of vector magnitudes in their application. Given the initial uncertainty regarding the importance of vector magnitudes, a comparative analysis of both functions was conducted. Moreover, it can be shown that when the vectors in a subset of a vector space have equal norms, MCSS, MIPS, and NNS become equivalent. Nevertheless, we continue to seek the best similarity search function, as it determines whether normalized vectors should be stored to reduce computational demands during queries.

III. Evaluation Methodology

A sound similarity system is intended to help humans; therefore, its performance should be evaluated by users. However, an extensive subjective evaluation requires the participation of many users to capture the vast use cases of the system. Moreover, the time it takes for a single participant to perform the test is quite high. Therefore, in this work, we use objective evaluations to reduce the search space before performing an extensive subjective evaluation with multiple participants. Concurrent with the objective evaluations, we perform preliminary subjective evaluations on ourselves to validate if the objective metrics correlate well with our experiences. The remainder of this section details the evaluation dataset and both evaluation methods.

A. Evaluation dataset

We use the FSD50K evaluation set to compare the sound similarity system performances [26]. Its 10,231 audio clips were manually annotated using 200 labels from the AudioSet ontology in a multi-label fashion. The AudioSet ontology represents a large set of sound events using 632 unique classes from seven sound families: “Animal”, “Natural sounds”, “Sounds of things”, “Music”, “Source-ambiguous sounds”, “Human sounds”, “Channel, environment, and background” [25]. The ‘Channel, environment, and background” node and its descendants are excluded in the FSD50K dataset.

The sound classes in the AudioSet ontology are organized into a hierarchical graph, where 474 unique classes are represented as leaf nodes. Within this structure, some nodes are children of multiple intermediary nodes. For such nodes, the FSD50K dataset assigns a single parent node. For instance, “Squeak”, can be a descendant of “Sounds of things”

or "Source-ambiguous sounds", but is assigned to "Source-ambiguous sounds". A file detailing these decisions is provided by the authors.[4]

Investigating the label hierarchy revealed that 22 nodes of the FSD50K dataset have multiple parents in the AudioSet ontology. For an accurate evaluation, we searched for the right parent class and sound family for these labels. We listened to at least five audio clips per problematic label and reported some of our findings below.

- "Clapping" can have "Hands" or "Human group actions" as a parent, and FSD50K suggests the former. We observed that audio clips labeled with "Hands", e.g., with Freesound IDs 340356, 119634, and 410868, may include crowd recordings. A single pair of hands clapping and a crowd clapping are both acoustically and semantically different.
- "Hiss" can be the child of "Onomatopoeia", "Cat", "Snake" or "Steam", while FSD50K suggests "Onomatopoeia". Clips 1942, 89447, and 265581 have all different hiss sounds. Although they are similar acoustically, none of these sounds are semantically similar due to being produced by different animals.
- "Growling" can be the child of "Dog", "Cat", "Roaring cats (lions and tigers)", and "Canidae and dogs and wolves"; but there is no suggested parent. Clips 389617, 256646, and 276577 contain a wide range of acoustically similar but semantically dissimilar sounds.

Based on our observations, we decided not to include these 22 labels within any specific sound family. Instead, our analyses treat them as individual labels, and their hierarchical relations are not considered to ensure accuracy. For a more detailed analysis, please refer to [40] and the project repository. Given these findings, we advocate for caution when using FSD50K labels to measure nuanced sound semantics.

B. Objective evaluation

Our objective evaluations of a model's semantic sound similarity performance are based on the labels from the FSD50K dataset. Although these labels have certain limitations, as outlined in Section III-A, they are employed in our experiments to approximate coarse sound semantics.

We evaluate a system's response to a single query using the Average Precision at the N^{th} position (AP@N) and the Rank of the 1^{st} Relevant Item (R1) metrics. We use N=15 and N=150 in our experiments, which correspond to the number of displayed sounds on the first and tenth page of Freesound, respectively. Both metrics require a function that evaluates the relevance of two audio clips, which we define based on the relationship between their labels. We consider a retrieved sound to be relevant to a query sound if the retrieved sound is labeled with the sound class we are interested in the query.

For broader evaluations beyond individual queries, we utilize sound class and family hierarchies to compute summarized metrics. We query the audio clips for each sound class and assess the results using the AP@15, AP@150, and R1 metrics. The averages of these metrics across the corresponding classes yield the Mean Average Precision@15 (MAP@15), Mean Average Precision@150 (MAP@150), and the Mean Rank of the 1st Relevant Item (MR1), respectively. These class-averaged metrics are critical for assessing a QbE system's performance.

Given the presence of 200 sound classes in the evaluation dataset, directly comparing the metrics of individual classes across multiple models can be challenging. Therefore, we consider two distinct summary approaches. The first averages the class-wise scores of a metric across all classes, producing a single composite value representing the metric over the entire dataset. The second averages the class-wise scores within sound families, summarizing performance for specific families. As discussed in Section III-A, the family-averaged summaries exclude the problematic sound classes.

TABLE I
BEST PERFORMING VARIATIONS OF MULTIPLE VARIATION MODELS

Model	Variation
FSD-SINet	vgg42-tlpf-1
BEATs	iter3_plus_AS2M
OpenL3	env-mel256-emb512
Imagebind	huge
LAION-CLAP	630k-fusion-best
AudioCLIP	full-Training
CAV-MAE	as_46.6

C. Preliminary subjective evaluation

We develop a web interface that implements the sound similarity system on the FSD50K evaluation set. It takes a sound class input from the user and randomly selects an audio clip from this class. The sound is then queried against up to four different user-selected QbE systems, and the retrieved sounds are displayed side-by-side. The similarity score of each retrieved sound and the AP@15 score of the results for each system's retrieved sounds are also displayed. We use this interface throughout our experiments to listen to the retrieved sounds and subjectively evaluate the semantic similarity capabilities of the models. In particular, after narrowing the search to three models, we use the interface extensively to compare their retrieval performance with Freesound.

IV. RESULTS & DISCUSSION

As described in Section II-C and Section II-D, we optimize the embedding processing parameters for creating audio representations from each raw embedding and identify the best similarity search function for the resulting representations. Fig. 1 displays the evaluation results for the FSD-SINet's VGG42-tlpf variation, showing the average of the class-wise MAP@15 scores across 27 unique QbE systems that utilize the same raw embeddings. While the observations from this figure are detailed below, similar trends were noted in other models and variations, as well as in the MAP@150 and MR1 metrics.

[4]https://github.com/xavierfav/Freesound-data-set/blob/master/ontology/ontology_crowd.json

TABLE II

SEMANTIC SOUND SIMILARITY PERFORMANCES USING AVERAGES OF CLASS-WISE METRICS. MODALITIES ARE DENOTED AS, A: AUDIO, I: IMAGE, L: LANGUAGE, O: OTHERS. ONLY AUDIO ENCODER PARAMETERS ARE REPORTED. 'DATA SIZE' DENOTES TRAINING DATA VOLUME; FOR IMAGEBIND, ONLY AUDIO DATA IS PROVIDED. ORIGINAL EMBEDDING DIMENSIONS ARE GIVEN BUT FOR RETRIEVAL, PROCESSED EMBEDDINGS WERE USED.

Modality	Model	Pre-training Objective	# Param.	Data Size	# Dim.	MAP@15 ↑	MAP@150 ↑	MR1 ↓
A	Freesound	-	-	-	846	0.09	0.03	43
A	VGGish [10]	Classification	62M	80M	128	0.20	0.11	27
	YAMNet [41]	Classification	4M	2M	1024	0.27	0.15	25
	FSD-SINet [12]	Classification	5M	51K	512	0.33	0.19	18
	BEATs [27]	Classification	90M	1.8M	768	0.37	0.20	13
A & I	OpenL3 [37]	Audiovisual correspondence	5M	296K	512	0.15	0.06	31
	CAV-MAE [39]	Audiovisual correspondence	85M	1.8M	768	0.33	0.18	15
A & L	CLAP2022 [17]	Contrastive alignment	81M	128K	1024	0.38	0.25	20
	LAION-CLAP [18]	Contrastive alignment	31M	2.5M	512	**0.53**	**0.37**	**8**
	CLAP2023 [19]	Contrastive alignment	31M	4.6M	1024	0.50	0.34	10
	Pengi [20]	Question-answering	31M	3.4M	1024	0.49	0.35	12
A & I & L	AudioCLIP [34]	Contrastive alignment	30M	1.8M	1024	0.06	0.02	56
	Wav2CLIP [30]	Contrastive alignment	12M	200K	512	0.12	0.04	36
A & I & L & O	ImageBind [36]	Contrastive alignment	85M	1.8M	1024	0.29	0.17	22

Corresponding figures for each model and each variation are available in our project repository.

- MCSS scored the highest while MIPS scored the lowest for each embedding processing system, indicating that the vector norms negatively affected the similarity score.
- The aggregation type did not influence the metric considerably. We posit that this is because the corresponding models were trained with mean aggregation, shaping the resulting distributions to concentrate around the mean.
- Reducing the dimensions to 100 did not significantly decrease the score across any combination of aggregation type and search algorithm. In fact, for mean and median aggregation, it increased the score.
- When no dimensionality reduction was applied, MIPS scored particularly lower. This observation requires further exploration.

As explained in Section II-C, several models have multiple variations. For each variation, we identify the optimal QbE system and report the top-performing variations in Table I. Finding the optimal embedding processing parameters showed that applying mean aggregation and reducing the dimensions to 100 principal components typically yielded each model's best or near-best performance. Therefore, we fixed these parameters for all models, except for aggregation, which is required only for a subset of the models. For the similarity search, we chose MCSS since it consistently performed better across models and embedding processing systems.

In Table II, performances of all models are compared after fixing the model variations, embedding processing parameters, and similarity search function. The presented scores are averages of the class-wise metrics. This table reports the semantic sound similarity performance of the deep representation learning models categorized by input modality and includes the baseline method, Freesound, for comparison.

Among the audio-only models, there is significant variability between all three metrics. Notably, YAMNet and BEATs, which were both trained on AudioSet, exhibit different outcomes due to their distinct training methods. YAMNet used supervised classification, while BEATs benefited from self-supervised pre-training and supervised fine-tuning, highlighting the potential advantages of this approach. In addition, the minimal performance difference between FSD-SINet and BEATs requires further investigation. One possible explanation is that FSD-SINet was developed and evaluated on the FSD50K dataset.

OpenL3 and CAV-MAE, which utilize both audio and image modalities were trained with the AVC task. Specifically, OpenL3 variations were trained on videos featuring either musical performances or environmental sounds, which may lack sufficient semantic relations with numerous sound classes. In contrast, CAV-MAE, which leveraged self-supervised pre-training with a large parameter count and was fine-tuned on AudioSet, delivers competitive performance.

The performance of AudioCLIP and Wav2CLIP, which incorporated audio, image, and language modalities together, are notably low. Both models utilized the CLIP model, pre-trained on language-image pairs, for audio representation learning. Their performance suggests that models derived from a pre-trained CLIP may not possess the necessary training objectives for learning effective semantic representations of sounds. Since our experiments with joint audio, language, and image modalities are limited to the models that utilize CLIP, we cannot draw conclusions about the effectiveness of this tri-modal input.

It can be observed that the best-performing models use only audio and language modalities, together. Notably, the LAION-CLAP model performs the best, while CLAP2023, trained with more data and a different text encoder, slightly underperforms. Both models were trained with supervised contrastive learning, which may indicate the effectiveness of this learning paradigm. In contrast, Pengi, a question-answering model that trained the same audio encoder, also underperforms. These

differences could be attributed to the labeling issues within the FSD50K dataset discussed in Section III-A, or potentially to the insufficient size of the FSD50K evaluation set to draw definitive conclusions. Nonetheless, the comparative success observed among these models may be attributed to utilizing supervision from both natural language and audio, or it could be a consequence of utilizing larger volumes of training data. Additionally, the comparative success of the CLAP2022 model, considering its small training data volume is notable.

Now we focus on input modality combinations of only audio, language, and image, and exclude the tri-modal models derived from a pre-trained CLIP. We find that the modality combination performances rank as follows: audio-language, audio, and audio-image. Nonetheless, it is important to recognize that this comparison involves various factors including different training paradigms; isolating the effects of each factor would require testing additional models. Moreover, our observations are based on objective metrics and our preliminary subjective evaluations. Consequently, we advocate for thorough subjective evaluations.

The ImageBind model, distinguished by its training that aligned the representations of five distinct modalities separately to the representations of images, demonstrates an impressive ability to capture sound semantics. This proficiency is notably achieved despite the model not being exposed to audio and language pairs together during training. Its capability to outperform various models highlights the potential of indirect, modality-agnostic learning approaches in capturing complex inter-modal relationships.

It is notable that models such as CAV-MAE and BEATs, which excel in the sound event recognition task tend to underperform in the semantic sound similarity task, compared to models such as LAION-CLAP or CLAP2023. This disparity between classification and retrieval performance of embeddings aligns with findings reported in the literature, highlighting the different demands of the two tasks [18], [20].

As described in Section III-B, we also compute metrics for sound families. Fig. 2 shows the model performance for sound families using the family averaged MAP@15 scores. It can be seen that, for each model, the performance peaks in the "Music" family. However, the sound classes related to music in the FSD50K dataset are confined to the presence of music or the identification of a range of musical instruments, which may not capture the nuanced semantics necessary for certain applications. Additionally, all models perform lowest on the "Source-ambiguous sounds", which demand a distinct type of semantics compared to other categories.

We additionally conducted a preliminary subjective evaluation using the web interface described in Section III-C. This involved a four-way comparison among the baseline method and the most promising models from each modality category identified in Table II. Specifically, we compared Freesound's performance with LAION-CLAP, BEATs, and CAV-MAE. In these tests, LAION-CLAP consistently produced results that were semantically closer to the query sounds. BEATs and CAV-MAE's rankings were consistent with their objective

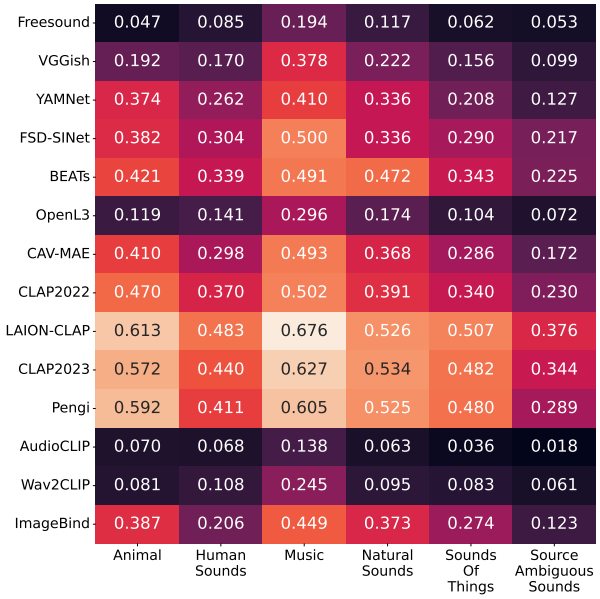

Fig. 2. Heatmap illustrating the semantic sound similarity performance across different sound families, using the family averaged MAP@15 scores.

analysis results, with all models surpassing the Freesound baseline in terms of subjective sound similarity. Additionally, comparisons involving LAION-CLAP, CLAP2023, and Pengi showed no significant performance differences. It is crucial to acknowledge the limitations of this preliminary subjective evaluation and to interpret its outcomes with caution.

V. Conclusion

In this study, we evaluated the retrieval performance of one manually engineered audio representation and thirteen deep audio embeddings on the semantic sound similarity task through a comprehensive, objective evaluation. These evaluations were conducted across multiple levels, considering sound hierarchies from sound classes to sound families, and were complemented by preliminary subjective assessments. In the process, we explored numerous design choices regarding embedding processing methods and similarity search functions to determine the optimal setting for each embedding.

We found that the input modalities are crucial for learning semantic representations of sounds. Specifically, representations obtained from multi-modal models that integrate audio and language had significantly higher retrieval performance compared to audio-only models, which, in turn, outperformed audio and image models. Interestingly, the models that excel in the sound event classification task underperformed in the semantic sound similarity task, highlighting the unique challenges of each task. Lastly, we observed that substantially reducing the embedding dimensionality with PCA can enhance retrieval performance slightly.

We plan to leverage the extensive Freesound data to develop an open-source model for semantic sound similarity and conduct a comprehensive user-based evaluation within the Freesound environment to deepen our understanding.

REFERENCES

[1] F. Font, "Design and evaluation of a visualization interface for querying large unstructured sound databases," PhD Thesis, Universitat Pompeu Fabra, Aug. 2010.

[2] X. Favory, "Improving Sound Retrieval in Large Collaborative Collections," PhD Thesis, Universitat Pompeu Fabra, Mar. 2021.

[3] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, Oct. 2008.

[4] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet, "Audio Information Retrieval using Semantic Similarity," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. IEEE, Apr. 2007.

[5] B. Mechtley, G. Wichern, H. Thornburg, and A. Spanias, "Combining semantic, social, and acoustic similarity for retrieval of environmental sounds," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.

[6] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE MultiMedia*, 1996.

[7] X. Yu, J. Zhang, J. Liu, W. Wan, and W. Yang, "An audio retrieval method based on chromagram and distance metrics," in *2010 International Conference on Audio, Language and Image Processing*. IEEE, Nov. 2010.

[8] Q. Wu, X. Zhang, P. Lv, and J. Wu, "Perceptual similarity between audio clips and feature selection for its measurement," in *2012 8th International Symposium on Chinese Spoken Language Processing*. IEEE, Dec. 2012.

[9] M. Slaney, "Mixtures of probability experts for audio retrieval and indexing," in *Proceedings. IEEE International Conference on Multimedia and Expo*. Lausanne, Switzerland: IEEE, 2002.

[10] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[11] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021.

[12] E. Fonseca, A. Ferraro, and X. Serra, "Improving Sound Event Classification by Increasing Shift Invariance in Convolutional Neural Networks," Jul. 2021, arXiv:2107.00623 [cs, eess].

[13] X. Favory, K. Drossos, T. Virtanen, and X. Serra, "COALA: Co-Aligned Autoencoders for Learning Semantically Enriched Audio Representations," in *ICML 2020 Workshop on Self-supervision in Audio and Speech*, 2020.

[14] P. Manocha, R. Badlani, A. Kumar, A. Shah, B. Elizalde, and B. Raj, "Content-Based Representations of Audio Using Siamese Neural Networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2018.

[15] M. Sert and A. M. Basbug, "Combining Acoustic and Semantic Similarity for Acoustic Scene Retrieval," in *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2019.

[16] J. Fan, E. Nichols, D. Tompkins, A. E. Mendez Mendez, B. Elizalde, and P. Pasquier, "Multi-Label Sound Event Retrieval Using A Deep Learning-Based Siamese Structure With A Pairwise Presence Matrix," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[17] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP Learning Audio Concepts from Natural Language Supervision," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[18] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[19] B. Elizalde, S. Deshmukh, and H. Wang, "Natural Language Supervision For General-Purpose Audio Representations," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[20] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An Audio Language Model for Audio Tasks," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[21] H.-H. Wu, O. Nieto, J. P. Bello, and J. Salamon, "Audio-Text Models Do Not Yet Leverage Natural Language," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[22] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "ESSENTIA: an Audio Analysis Library for Music Information Retrieval," in *Proceedings - 14th International Society for Music Information Retrieval Conference*, Nov. 2013.

[23] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A Large-Scale Video Classification Benchmark," Sep. 2016, arXiv:1609.08675 [cs].

[24] "Sound classification with YAMNet | TensorFlow Hub." [Online]. Available: https://www.tensorflow.org/hub/tutorials/yamnet

[25] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017.

[26] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 30, Dec. 2021, publisher: IEEE Press.

[27] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023.

[28] 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," Aug. 2020, arXiv:1912.10211 [cs, eess].

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

[30] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2CLIP: Learning Robust Audio Representations from Clip," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and others, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[33] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A Large-Scale Audio-Visual Dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[34] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending Clip to Image, Text and Audio," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[35] A. Guzhov, F. Raue, J. Hees, and A. R. Dengel, "ESResNe(X)t-fbsp: Learning Robust Time-Frequency Transformation of Audio," *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021.

[36] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind One Embedding Space to Bind Them All," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[37] A. L. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2019.

[38] R. Arandjelović and A. Zisserman, "Look, Listen and Learn," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[39] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. R. Glass, "Contrastive Audio-Visual Masked Autoencoder," in *The Eleventh International Conference on Learning Representations*, 2023.

[40] R. O. Araz, "Semantic Sound Similarity with Deep Embeddings for Freesound," Master's thesis, Universitat Pompeu Fabra, 2023.

[41] "models/research/audioset/yamnet at master · tensorflow/models." [Online]. Available: https://github.com/tensorflow/models