Recuperação semântica de paisagens sonoras usando banco de dados vetoriais

Andrés D. Peralta[1], **Eulanda Miranda dos Santos**[1],
Jie Xie[2], **Juan G. Colonna**[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Escola de Inteligência Artificial - Universidade Normal de Nanji

andres, emsantos, juancolonna {@icomp.ufam.edu.br}

xiej8734@gmail.com

Abstract. *The semantic recovery of soundscapes emerges as a crucial component for monitoring ecosystems. However, due to the continuous nature of monitoring over time, we face considerable challenges due to the vast volume of collected audio records. In addition to the large data volume, we also encounter the inconvenience of the lack of labels in the recordings. Currently, there are several proposals based on supervised machine learning to recognize and classify animal species based on their vocalizations. However, there is a shortage of studies implementing semantic recovery of soundscapes through the application of pre-trained Deep Learning models and vector databases (ex: VectorDB). In this study, we develop a vector database for querying and retrieving similar acoustic landscapes with anuran calls.*

Resumo. *A recuperação semântica de paisagens sonoras emerge como um componente crucial para monitorar ecossistemas. No entanto, devido à natureza contínua do monitoramento ao longo do tempo, enfrentamos desafios consideráveis devido ao vasto volume de registros de áudio coletados. Além do grande volume de dados, também nos deparamos com a falta de rótulos nas gravações. Atualmente, existem várias propostas baseadas em aprendizado de máquina supervisionado para reconhecer e classificar espécies animais com base em suas vocalizações. No entanto, há uma escassez de estudos que implementam a recuperação semântica de paisagens sonoras por meio da aplicação de modelos de Deep Learning pré-treinados e bancos de vetoriais (por exemplo, VectorDB). Neste estudo, desenvolvemos um banco de vetoriais para consultar e recuperar paisagens acústicas semelhantes com vocalizações de anuros.*

1. Introdução

De acordo com Schafer [1993] as paisagens sonoras compreendem o conjunto de sons de um ambiente específico, composta pela agregação de todos os sons de um lugar, incluindo os sons naturais *(biophony e geophony)* e os produzidos pela atividade humana *(antrophony)*. Também pode incluir uma grande variedade de sons, como o canto dos pássaros, o fluxo dos rios, o vento nas árvores, o tráfego da cidade, o murmurar das conversas, o ruído industrial e muitos outros [Devalraju and Rajan, 2022]. Além disso, Pijanowski et al. [2011] destacam a possibilidade de analisar e estudar as paisagens sonoras a

partir de diversas perspectivas, como ecologia acústica, bioacústica, geografia acústica e composição musical.

Bianco et al. [2019] destacam o surgimento de métodos de coleta de dados bioacústicos impulsionados por avanços em tecnologia e informática. A aplicação de ferramentas computacionais e métodos de *Machine Learning* no campo da bioacústica tem recebido crescente atenção. No entanto, a complexidade das paisagens sonoras apresenta desafios em termos de análise de dados, para os quais o aprendizado de máquina se mostra como um potencial solução, conforme sustentado pelo [Quaderi et al., 2022]. O objetivo deste trabalho surge, portanto, da necessidade de desenvolver uma sistema capaz de resolver buscas ou consultas por similaridade, de forma permitir a recuperação de paisagens sonoras, com o intuito de facilitar o monitoramento e estudo dos ecossistemas.

Nesta pesquisa implementamos um modelo *Deep Learning* pré-treinado para extrair vetores de *embeddings* das gravações de áudio ambiental, as quais foram armazenadas em um banco da dados vetorial, o *VectorDB*. Esta tecnologia de banco de dados vetoriais permite armazenar de forma compacta dados ecoacústicos e também realizar busca por similaridade sonora de forma eficiente. O modelo pré-treinado que usamos para extrair os vetores foi Perch [Ghani et al., 2023]. Este modelo se destaca pelo uso da arquitetura de rede neural EfficientNet [Tan and Le, 2019].

Sendo assim, podemos concluir que a finalidade primordial desta pesquisa consiste em alcançar a recuperação semântica por semelhança de paisagens sonoras de qualquer região geográfica, através da captura de uma gravação, extração dos vetores de *embeddings* e posterior realização de consultas à base de dados vetorial com esses vetores.

2. Fundamentos Teóricos

Uma das tarefas mais desafiadoras ao lidar com a recuperação semântica de paisagens sonoras é a alta dimensionalidade inerente a este tipo de dados. A coleta de som ambiental gera volumes massivos de dados não rotulados, principalmente quando as gravações são de alta qualidade. Esse tipo de dados são essenciais para o treinamento e avaliação de algoritmos de *Deep Learning*. A rotulagem desses conjuntos de dados requer tempo e a participação de especialistas na identificação de eventos acústicos. [Barrington et al., 2007, Koepke et al., 2023]. A recuperação semântica em áudios permite descrever os sinais sonoros e seu significado utilizando técnicas de aprendizado de máquina para compreender o conteúdo do áudio, facilitando a análise do som e recuperação de informação similar [Slaney, 2002].

Existem várias técnicas para realizar o processo de recuperação de paisagens sonoras [Presannakumar and Mohamed, 2023]. Segundo Xu [2020], a aplicação de Hash de cadeia tipo *MD5* ou *SHA256* para recuperar áudios similares não é viável devido às características únicas do som e sua variabilidade. Essas sutilezas impedem o uso de *Hashing* desse tipo para recuperação por similaridade, sendo adequados apenas para recuperação exata. Por outro lado, Jin et al. [2023] afirmam que o uso de *Hashing* Semântico em *Machine Learning* para mapear eficientemente dados de alta dimensionalidade para espaços de menor dimensão é viável, desde que a qualidade do som seja alta e não haja ruído de fundo, a fim de evitar viés na informação. Neste contexto, será implementada a técnica de recuperação de arquivos de áudio baseada em vetores de *embeddings*, indexando esses vectores com os algoritmos HNSW e IMENN para melhorar a eficácia e precisão das

consultas na base de dados vetorial [Jina-Ai, 2023].

As gravações ecoacústicas, geralmente feitas com microfones omnidirecionais, captam sons do ambiente, incluindo ruídos indesejáveis que dificultam o processamento e a análise dos eventos acústicos. O filtro de ruído para recuperação semântica de paisagens sonoras atua como um pré-processamento, eliminando ruídos que prejudicam a recuperação semântica, mas preservando as informações relevantes do áudio [Wang et al., 2015]. Nesta pesquisa, implementamos o algoritmo de redução de ruído *Noise Reduce* [Sainburg et al., 2020]. Este algoritmo utiliza o espectrograma do ruído presente no sinal de áudio, gerado pela *Transformada de Fourier de Tempo Curto* (STFT). A partir deste espectrograma, calcula-se a média e o desvio padrão de cada banda de frequência, permitindo distinguir entre eventos acústicos e ruído de fundo. Após isso, uma máscara booliana diferencia as regiões do espectrograma com sinal e com ruídos. Essa máscara é suavizada e aplicada ao espectrograma completo para reduzir ou eliminar as regiões de ruído. Por fim, a STFT inversa é aplicada para reconstruir o sinal filtrado.

2.1. Algoritmos HNSW e IMENN

O algoritmo *Hierarchical Navigable Small* (HNSW) foi projetado para a busca aproximada dos *k*-vizinhos mais próximos baseado dentro de um grafo [Malkov and Yashunin, 2020]. Inicialmente um vetor de *embedding* é selecionado como ponto inicial e, em seguida, o grafo é construído conectando cada nó aos seus vizinhos mais próximos. Depois, os nós são organizados em níveis hierárquicos, onde os níveis superiores contêm somente alguns nós e são usados para iniciar a busca aproximada, enquanto os níveis inferiores têm mais nós e são usados para refinar os resultados. Finalmente, para fazer uma consulta, começa-se em um nó de nível superior e desce-se pelo grafo em direção aos níveis inferiores. A busca é interrompida quando se alcança um nível suficientemente baixo ou quando um número máximo de nós é explorado.

In Memory ExactNN Index (IMENN) é um algoritmo exato conhecido como k-NN [Jina-Ai, 2023]. Durante a indexação, o conjunto de vetores de *embeddings* $X = \{x_1, x_2, \ldots, x_n\}$ é armazenado como um tabela plana em memória. Para realizar uma consulta, um novo vetor de *embedding* (x_j) é usado como *query*, e então é calculada a similaridade entre este vetor e todos os vetores armazenados. Os vetores que apresentarem maior similaridade com a *query* são retornados como os resultados da consulta.

2.2. Métricas de avaliação

O desempenho de um banco de dados vetorial pode ser avaliado por meio de um conjunto de métricas que evidenciam a qualidade das consultas que este retorna. As métricas comumente utilizadas na literatura de recuperação semântica são o Precisão-*k* ($P@k$), o *Top-k* e o tempo medido em milissegundos. A $P@k$ mede a proporção de resultados relevantes entre os k primeiros resultados retornados pelo sistema e é definida como [Yadav et al., 2014]:

$$\mathrm{P@k} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{r}(i) \tag{1}$$

sendo

$$r(i) = \begin{cases} 1, & \text{se o resultado pertence ao mesmo áudio} \\ 0, & \text{caso contrário,} \end{cases} \tag{2}$$

onde k representa o número de resultados retornados pelo sistema, e $r(i)$ é uma função indicadora que retorna 1 se o i-ésimo resultado for relevante e 0 caso contrário. A equação 1 calcula a soma dos valores de $r(i)$ para os k primeiros resultados e divide por k, produzindo a fração de resultados relevantes. Esta fração é obtida para cada consulta e finalmente a média é calculada para indicar a eficacia do sistema proposto para recuperação semântica ecoacústica. Um valor baixo de $P@k$ indica que o sistema possui um alto número de resultados não relevantes. Neste trabalho adotamos $P@5$.

A métrica Top-k é definida como:

$$f(q,k) = \begin{cases} 1, & \text{se pelo menos um resultado é do mesmo áudio} \\ 0, & \text{caso contrário} \end{cases} \tag{3}$$

onde $1 \leq i \leq k$. A função $f(q,k)$ recebe uma consulta q e os k elementos recuperados k. Esta função retorna 1 se existir pelo menos um elemento recuperado $r(i)$ que possua o mesmo rótulo que a consulta q. Caso contrário, retorna 0. Essa função é aplicada a cada uma das consultas e, após isso, calculamos a média dividindo pelo número total de consultas [Petersen et al., 2022]. Para avaliar os resultados de nosso sistema foram utilizadas a Top-1 e a Top-5.

Finalmente, para mensurar a eficiência na velocidade das consultas medimos o tempo médio em milissegundos (ms) para cada consulta. As medições foram acompanhadas do cálculo do desvio padrão. Uma média de tempo pequena indica que o sistema responde rápido e é eficiente. Um desvio padrão pequeno indica que houve pouca variação nos tempos das respostas.

3. Trabalhos relacionados

O impacto dos métodos baseados em *Machine Learning* no campo da bioacústica e a atenção recente que têm recebido nos motivaram a realizar esta pesquisa sobre recuperação semântica de paisagens sonoras [Bjorck et al., 2019, Fanioudakis and Potamitis, 2017, Hagiwara et al., 2023].

Jati and Emmanouilidou [2020] conduziram uma pesquisa para avaliar a eficácia da técnica de *Deep Hashing* na recuperação eficiente de eventos de áudio. Os autores avaliaram os modelos pré-treinados: VGGish e TLWeak [Hershey et al., 2017, Kumar et al., 2018]. Além disso, eles propuseram um sistema parcialmente supervisionado que obtém embeddings em um espaço de baixa dimensionalidade enquanto otimiza os códigos *Hashing*. Este sistema foi treinado e avaliado com os datasets DCASE e ESC-50 [Fonseca et al., 2018, Piczak, 2015], e avaliado com a métrica Top-1 do vizinho mais próximo. Os autores utilizaram como ponto de partida o modelo DNQ Deep (*Product Quantization Module*) utilizado na recuperação semântica de imagens [Meihan et al., 2020]. Os resultados finais indicam que o sistema proposto é promissor e que a aplicação de algoritmos de busca não exaustiva resulta em uma recuperação mais precisa e eficiente.

Lin et al. [2023] propuseram um método para identificação e classificação de eventos sonoros baseado em vetores de atributos de som (SAVs). Este método recebe um som como entrada, gera um espectrograma de tamanho fixo e utiliza o VGGish como codificador para extrair um novo conjunto de características. Em seguida, um módulo aprende as características globais discriminativas, enquanto outro módulo é responsável

por aprender as características espectro-temporais locais dos atributos de cada classe. Os autores treinaram o VGGish sem transferência de aprendizagem usando o conjunto de dados RWCP-SSD, que é composto por sons ambientais reais [Nakamura et al., 2000]. Os resultados dos experimentos demonstram que o modelo é viável para a identificação de eventos sonoros, embora seja necessário mais pesquisa para melhorar a precisão na classificação.

No estudo realizado por Ghani et al. [2023] foram empregados modelos pré-treinados em tarefas de classificação de áudios para realizar a recuperação semântica por meio de vetores de *embeddings*. Os autores exploraram a viabilidade de utilizar os *embeddings* para identificar classes bioacústicas distintas daquelas para as quais os modelos foram originalmente treinados. Eles avaliaram os modelos em diversos conjuntos de dados, que incluíam diferentes cantos de aves, morcegos, mamíferos marinhos e anfíbios. Os resultados indicaram que os vetores de *embeddings* extraídos dos modelos pré-treinados em vocalizações de aves são melhores do que aqueles treinados em conjuntos de áudios gerais.

Os trabalhos anteriores em recuperação semântica de áudio apresentam certas limitações que precisam ser abordadas. Por exemplo, alguns métodos são computacionalmente lentos e custosos, o que dificulta sua aplicação em grandes conjuntos de dados ou em tempo real. Além disso, a representação de sons pode ser limitada em termos de capturar toda a informação semântica necessária para caracterizar a similaridade entre diferentes gravações. Outro desafio reside na falta de robustez desses métodos diante do ruído ambiental, o que pode gerar resultados imprecisos.

4. Matérias e métodos

4.1. Base de dados

Nesta pesquisa utilizamos um conjunto de dados bioacústicos próprio, obtido por meio de gravações realizadas no campus universitário da Universidade (UFAM). Este conjunto de dados contem 60 gravações de áudio de 12 espécies de anuros no formato .FLAC com uma taxa de amostragem de 44,100 kHz. Estas gravações foram filtradas para eliminar ruídos indesejados aplicando o filtro descrito na Seção 2 e, posteriormente, foram segmentadas em janelas de 5 segundos. A segmentação produziu 890 segmentos, dos quais 623 segmentos (70%) foram armazenados no banco de dados vetorial para responder às consultas e 267 segmentos (30%) foram reservados para realizar consultas *(querys)*.

Ambos conjuntos de segmentos provêm das mesmas gravações, mas é importante destacar que os 267 segmentos usados para as consultas não estão armazenados no banco de dados nem possuem sobreposição temporal com aqueles segmentos que já estão no banco. Este procedimento foi aplicado a todas as gravações com o objetivo de encontrar segmentos semelhantes durante as consultas.

4.2. Método proposto

Nosso método começa recebendo várias gravações de áudio e as segmentando em janelas temporais. Essas janelas são então inseridas no modelo *Deep Learning* Perch para extrair vetores de 1280 dimensões. Originalmente, o Perch foi treinado com cantos de aves do banco de dados Xeno-Canto [Tan and Le, 2019]. As gravações utilizadas para o treinamento possuem frequência de amostragem de 32 kHz. O Perch utiliza a arquitetura de

rede Neural EfficientNet B1 ilustrada na Figura 1, otimizada para equilibrar o desempenho e a demanda por recursos computacionais, o que é crucial para a rápida extração de embeddings para novos dados de consulta.

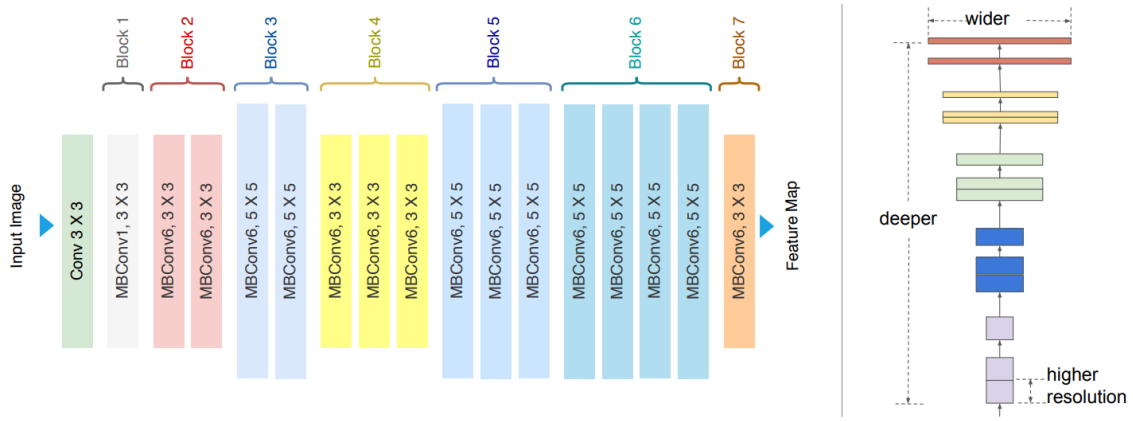

Figura 1. Rede neural Efficient Net adotada pelo modelo Perch. Figura adaptada de Ahmed TSabab [2021]. Lado esquerdo ilustra a sequencia de camadas, enquanto o lado direto mostra a composição dos blocos MBConv.

O banco de dados vetorial escolhido para armazenar os embeddings foi o VectorDB [Jina-Ai, 2023], que pode ser implementado localmente ou na nuvem. Além disso, o VectorDB é licenciado sob a licença Apache 2.0, permite a incorporação de metadados e suporta consultas aproximadas ou precisas (IMENN ou HNSW).

Para realizar uma nova consulta a partir de um áudio desconhecido, primeiro ele é mapeado para um vetor de *embedding* usando Perch e, em seguida, este vetor é consultado no VectorDB, que retorna as gravações semanticamente mais similares junto com seus metadados. O que consideramos como semelhantes são os segmentos de áudio dentro da mesma gravação. Ou seja, se um segmento de uma gravação pode ser semelhante a outros segmentos dessa mesma gravação. Podemos resumir esse processo em quatro etapas essenciais, conforme ilustrado na Figura 2. Estas etapas são:

- **Pré-processamento de dados:** redução de ruídos ambientais é realizada utilizando um filtro no domínio da frequência para melhorar a qualidade das gravações, e é feita uma subamostragem para normalizar a frequência para 32 kHz;
- **Segmentação de áudio:** a segmentação dos áudios foi realizada em janelas consecutivas de 5s segundos;
- **Extração de embeddings:** nesta etapa, os segmentos são inseridos no modelo *Deep Learning* pré-treinado Perch para extrair os vetores de *embedding*; e
- **Banco de dados vetorial e recuperação semântica:** primeiramente, o banco de dados vetorial é construído utilizando o VectorDB com exemplos de áudios e seus metadados, posteriormente, os algoritmos IMENN e HNSW são utilizados para resolver novas consultas.

5. Experimentos e resultados

Todos os experimentos foram conduzidos em uma estação de trabalho equipada com uma CPU Ryzen 5 de 2,5 GHz, 16 GB de memória RAM e um disco rígido de 1000 GB. Os

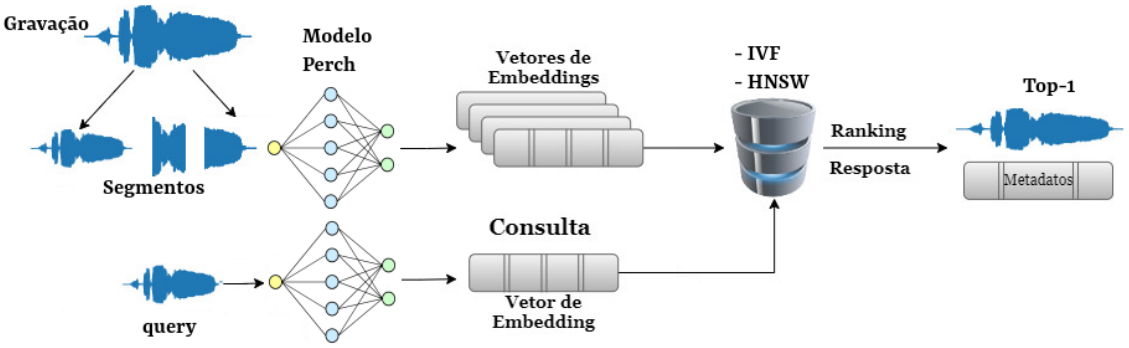

Figura 2. Ilustração do método proposto para recuperação semântica de paisagens sonoras.

resultados foram obtidos comparando os algoritmos IMENN e HNSW do banco de dados vetorial VectorDB com e sem filtro de ruídos aplicado às gravações. Foram realizadas 267 consultas no banco de dados vetorial. O desempenho foi avaliado utilizando as métricas Top-1, Top-5, Precisão-k (P@5), e o tempo médio (T) em milissegundos necessário para resolver cada consulta junto com o desvio padrão.

Tabela 1. Resultados das 267 consultas.

Algoritmo	Filtro de ruídos	Top-1	Top-5	P@5	T (ms)
IMENN	Não	56%	63%	0,58	11,02±5,63
	Sim	77%	84%	0,79	7,73±4,77
HNSW	Não	80%	85%	0,83	8,11±4,03
	Sim	94%	98%	0,96	7,08±2,58

Os resultados apresentados na Tabela 1 indicam que o algoritmo IMENN alcança uma precisão de 77% na busca do melhor resultado Top-1. O desempenho experimenta um aumento ao buscar o Top-5, com uma precisão de 84%, uma precisão-k P@5 de 0,58, um tempo de execução de 7,73 milissegundos e um desvio padrão de 4,77. Observa-se uma melhoria nos resultados ao aplicar o filtro de ruído. Além disso, evidencia-se que os tempos de execução são menores para o Top-1 em comparação com o Top-5, já que o algoritmo termina quando encontra a primeira resposta à consulta, enquanto o Top-5 ainda precisa buscar outras quatro possíveis respostas. Por outro lado, o algoritmo HNSW com filtro de ruído exibe uma precisão consideravelmente superior, atingindo 94% para o Top-1 e 98% para o Top-5. Além disso, ele se destaca por seu menor tempo de execução em contraste com o algoritmo IMENN, registrando um valor de 7,08 milissegundos, uma precisão-k de 0,96 e um desvio padrão de 2,58.

Os resultados obtidos dos experimentos respaldam a viabilidade de utilizar a técnica de vetores de *embeddings* para a recuperação semântica de paisagens sonoras, evidenciando o seu potencial aplicável no âmbito da bioacústica, graças à sua alta precisão e eficiência computacional. Além disso, observa-se que a combinação da aplicação de redução de ruído aos áudios e a implementação de um algoritmo de busca por aproximação (HNSW) melhora significativamente a recuperação semântica dessas paisagens sonoras. Também foi confirmado que o modelo pré-treinado Perch extrai de forma

eficaz *embeddings* das gravações de áudio. Presume-se que os resultados obtidos com nosso sistema poderiam gerar novas hipóteses para desenvolver métodos adicionais destinados a melhorar a recuperação semântica de paisagens sonoras, o que, por sua vez, poderia contribuir para o aumento da preservação e monitoramento da fauna.

6. Conclusões

A recuperação semântica de áudios demanda uma quantidade significativa de recursos de processamento e memória, o que se torna uma limitação, especialmente em ambientes onde tais recursos são escassos. Especificamente, a recuperação semântica de dados ecoacústicos ou bioacústicos não rotulados é um campo pouco explorado na literatura existente, tornando este trabalho pioneiro ao propor uma solução para organizar dados de monitoramento ambiental nesse contexto.

Neste estudo, introduzimos um sistema inovador de recuperação semântica de paisagens sonoras que se baseia em modelos de *Deep Learning* pré-treinados e em um banco de dados vetoriais de código aberto. Essa abordagem busca resolver o desafio de organizar grandes volumes de dados coletados de forma autônoma, mas que não foram rotulados, permitindo ao mesmo tempo consultas por similaridade, facilitando a comparação de amostras de diferentes locais de coleta ou épocas do ano.

Concluímos que nossa proposta é promissora para estudos ecológicos espaço-temporais. Identificamos, contudo, que ainda há margem para melhorias nos resultados, como a comparação com outros algoritmos de *Deep Learning* pré-treinados ou a utilização de diferentes bases de dados vetoriais. No futuro, planejamos avaliar mais detalhadamente os requisitos de tempo e memória de diferentes versões de bancos de dados vetoriais disponíveis.

7. Agradecimentos

Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Código Financeiro 001. Este trabalho foi parcialmente apoiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas - FAPEAM - por meio do projeto POSGRAD 2024. Esta pesquisa, realizada no âmbito do Projeto Samsung-UFAM de Ensino e Pesquisa (SUPER), de acordo com o Artigo 39 do Decreto n°10.521/2020, foi financiada pela Samsung Eletrônica da Amazônia Ltda, nos termos da Lei Federal n°8.387/1991, através do convênio 001/2020 firmado com a UFAM e FAEPI, Brasil. Esta pesquisa foi realizada no âmbito do PROGRAMA DA FAPEAM EDITAL N. 013/2022 - PRODUTIVIDADE-CT&I. Projeto: Diferentes Abordagens Computacionais Para Monitoramento Ecoacústico Autónomo da Região Amazônica. This work is also supported by National Natural Science Foundation of China (Grant No: 32371556)

Referências

N. H. N. Ahmed TSabab. Classification and understanding of cloud structures via satellite images with efficientunet. *SN Computer Science*, 2021. doi: 10.1007/s42979-021-00981-2.

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, volume 2, pages II–725–II–728, 2007. doi: 10.1109/ICASSP.2007.366338.

M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle. Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, 146:3590–3628, 2019. doi: 10.1121/1.5133944.

J. Bjorck, B. H. Rappazzo, D. Chen, R. Bernstein, P. H. Wrege, and C. P. Gomes. Automatic Detection and Compression for Passive Acoustic Monitoring of the African Forest Elephant. pages 476–484, 2019. doi: 10.1609/aaai.v33i01.3301476.

D. V. Devalraju and P. Rajan. Multiview embeddings for soundscape classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1197–1206, 2022. doi: 10.1109/TASLP.2022.3153272.

L. Fanioudakis and I. Potamitis. Deep Networks tag the location of bird vocalisations on audio spectrograms. *arXiv.org*, 2017. doi: 10.48550/arXiv.1711.04347.

E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, and X. Serra. General-purpose tagging of freesound audio with audioset labels: Task description,dataset, and baseline. *arXiv*, 2018.

B. Ghani, T. Denton, S. Kahl, and H. Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. 2023. doi: 10.1038/s41598-023-49989-z.

M. Hagiwara, B. Hoffman, J.-Y. Liu, M. Cusimano, F. Effenberger, and K. Zacarian. Beans: The benchmark of animal sounds. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096686.

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. pages 131–135. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP), 2017. doi: 10.1109/ICASSP.2017.7952132.

A. Jati and D. Emmanouilidou. Supervised deep hashing for efficient audio event retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4497–4501, 2020. doi: 10.1109/ICASSP40776.2020.9053766.

L. Jin, Z. Li, and J. Tang. Deep Semantic Multimodal Hashing Network for Scalable Image-Text and Video-Text Retrievals. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. doi: 10.1109/TNNLS.2020.2997020.

Jina-Ai. Jina-ai/vectordb: A Python vector database you just need - no more, no less., 2023. URL `https://github.com/jina-ai/vectordb`.

A. S. Koepke, A.-M. Oncescu, J. F. Henriques, Z. Akata, and S. Albanie. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685, 2023. doi: 10.1109/TMM.2022.3149712.

A. Kumar, M. Khadkevich, and C. Fügen. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 326–330, 2018. doi: 10.1109/ICASSP.2018.8462200.

Y. Lin, X. Chen, R. Takashima, and T. Takiguchi. zero-shot sound event classification using a sound attribute vector with global and local feature learning. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 5, 2023. doi: 10.1109/ICASSP49357.2023.10096367.

Y. A. Malkov and D. A. Yashunin. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 824–836, 2020. doi: 10.1109/TPAMI.2018.2889473.

L. Meihan, D. Yongxing, B. Yan, and D. Ling-Yu. Deep product quantization module for efficient image retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4382–4386, 2020. doi: 10.1109/ICASSP40776.2020.9054175.

S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada. In *LREC*, pages 965–968, 2000.

F. Petersen, H. Kuehne, C. Borgelt, and O. Deussen. Differentiable top-k classification learning. In *39 th International Conference on Machine Learning*, 2022.

K. J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, page 1015–1018. Association for Computing Machinery, 2015. doi: 10.1145/2733373.2806390.

B. C. Pijanowski, L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, S. H. Gage, and N. Pieretti. Soundscape Ecology: The Science of Sound in the Landscape. *BioScience*, 61:203–216, 2011. doi: 10.1525/bio.2011.61.3.6.

K. Presannakumar and A. Mohamed. Deep learning based source identification of environmental audio signals using optimized convolutional neural networks. *Applied Soft Computing*, 2023. doi: 10.1016/j.asoc.2023.110423.

S. J. S. Quaderi, S. A. Labonno, S. Mostafa, and S. Akhter. Identify the beehive sound using deep learning. *arXiv.org*, 2022. doi: 10.48550/arXiv.2209.01374.

T. Sainburg, M. Thielk, and T. Q. Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10), 2020.

R. M. Schafer. *The Soundscape*. Amazon, Rochester, Vt. : United States, Oct. 1993. ISBN 978-0-89281-455-8.

M. Slaney. Semantic-audio retrieval. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–4108–IV–4111, 2002. doi: 10.1109/ICASSP.2002.5745561.

M. Tan and Q. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6105–6114, 09-15 Jun 2019.

C. Wang, H. Yang, and C. Meinel. Deep semantic mapping for cross-modal retrieval. In *IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 234–241, 2015. doi: 10.1109/ICTAI.2015.45.

H. Xu. Cross-Modal Sound-Image Retrieval Based on Deep Collaborative Hashing. In *5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, pages 188–197, 2020. doi: 10.1109/ISCTT51595.2020.00041.

P. Yadav, P. Sujatha, P. Dhavachelvan, and K. Prasad. Weight based precision oriented metrics for multilingual information retrieval system. In *IEEE International Conference on Advanced Communications, Control and Computing Technologies*, pages 1114–1119, 2014. doi: 10.1109/ICACCCT.2014.7019271.