UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

RECUPERAÇÃO DE PAISAGENS SONORAS: UMA ANÁLISE DE TÉCNICAS DE FUSÃO DE VETORES DE EMBEDDINGS PRÉ-TREINADOS

ANDRÉS DAVID PERALTA DE AGUAS

RECUPERAÇÃO DE PAISAGENS SONORAS: UMA ANÁLISE DE TÉCNICAS DE FUSÃO DE VETORES DE EMBEDDINGS PRÉ-TREINADOS

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas, Campus Universitário Senador Arthur Virgílio Filho, como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador: Juan Gabriel Colonna

Manaus - AM

Setembro de 2025

Ficha Catalográfica

Elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

P426r Peralta de Aguas, Andrés David

Recuperação de paisagens sonoras: uma análise de técnicas de fusão de vetores de embeddings pré-treinados / Andrés David Peralta de Aguas. - 2025.

122 f. : il., color. ; 31 cm.

Orientador(a): Juan Gabriel Colonna.

Dissertação (mestrado) - Universidade Federal do Amazonas, Programa de Pós-Graduação em Informática, Manaus, 2025.

1. Monitoramento bioacústico. 2. Ecoacústica. 3. Modelos Deep Learning pré-treinados. 4. Base de dados vetorial. I. Colonna, Juan Gabriel. II. Universidade Federal do Amazonas. Programa de Pós-Graduação em Informática. III. Título

---

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Informática

FOLHA DE APROVAÇÃO

"RECUPERAÇÃO DE PAISAGENS SONORAS: UMA ANÁLISE DE TÉCNICAS DE FUSÃO DE VETORES DE EMBEDDINGS PRÉ-TREINADOS"

ANDRÉS DAVID PERALTA DE AGUAS

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Dr. Juan Gabriel Colonna - PPGI/UFAM - **Presidente**

Prof. Dr. Marco Antonio Pinheiro de Cristo - **Membro Interno**

Prof. Dr. André Luiz da Costa Carvalho - **Membro Externo**

Manaus, 10 de setembro de 2025.

Documento assinado eletronicamente por **Juan Gabriel Colonna**, **Professor do Magistério Superior**, em 28/09/2025, às 03:49, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015 .

Documento assinado eletronicamente por **Marco Antônio Pinheiro de Cristo**, **Professor do Magistério Superior**, em 28/09/2025, às 19:01, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015 .

Documento assinado eletronicamente por **André Luiz da Costa Carvalho**, **Professor do Magistério Superior**, em 29/09/2025, às 12:54, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015 .

Dedico este trabalho à minha mãe e ao meu pai, que sempre me impulsionaram a perseguir meus sonhos e alcançar minhas metas.

Ao meu avô Pedro, que me ensinou que devemos sempre acreditar em nós mesmos.

Agradecimentos

Este trabalho representa não apenas a conclusão de uma etapa acadêmica, mas também o resultado de uma jornada de crescimento pessoal e profissional. Muitas pessoas foram essenciais ao longo desse caminho, e expresso aqui minha mais profunda gratidão.

Agradeço, primeiramente, ao meu orientador, Profe Juan, por sua orientação, paciência e dedicação. Sua confiança, aliada ao rigor técnico e à constante disposição para compartilhar conhecimento, foram fundamentais para minha formação como pesquisador e para a realização desta dissertação.

À minha família, deixo um agradecimento especial e carregado de afeto. À minha mãe, Beatriz, por ser meu maior exemplo de força e coragem, sempre me apoiando em todos os momentos. Aos meus irmãos, Carlos e Vale, pelo carinho e incentivo constantes. À minha tia Selmira, por sue apoio incondicional. E à Mary, que sempre acreditou em mim ao longo desta trajetória.

Minha sincera gratidão a todos os professores do Programa do PPGI por compartilharem generosamente seus conhecimentos e por realizarem um trabalho de excelência que eleva a qualidade acadêmica da Universidade Federal do Amazonas. Agradeço igualmente ao IComp e a toda a equipe administrativa do PPGI, cujo suporte e infraestrutura foram essenciais para o desenvolvimento desta pesquisa. Registro meu agradecimento também à CAPES pelo apoio financeiro, sem o qual este trabalho não teria sido possível.

Por fim, deixo um agradecimento especial aos meus amigos Yune, Meng, Bruno, Euler, Giovanni e Thiago, pelo companheirismo, pelas conversas inspiradoras e pelo incentivo constante, que tornaram esta jornada mais leve e significativa. Da mesma forma, sou grato aos meus colegas e companheiros de laboratório, cuja troca de ideias, discussões e colaborações foram decisivas para o amadurecimento desta pesquisa.

Acknowledgments

This work represents not only the conclusion of an academic stage but also the result of a journey of personal and professional growth. Many people have been essential along this path, and I express here my deepest gratitude.

First and foremost, I am sincerely grateful to my advisor, Prof. Juan, for his guidance, patience, and dedication. His trust in me, combined with his technical rigor and constant willingness to share knowledge, has been fundamental to my development as a researcher and to the completion of this dissertation.

To my family, I extend a special and heartfelt acknowledgment. To my mother, Beatriz, my greatest example of strength and courage, for always supporting me unconditionally in every moment. To my siblings, Carlos and Vale, for their constant care and encouragement. To my aunt Selmira, for her unwavering support. And to Mary, who has always believed in me throughout this entire journey.

I also express my sincere gratitude to all the professors of the PPGI Program for generously sharing their knowledge and for carrying out excellent work that greatly contributes to the academic quality of the Federal University of Amazonas. I am equally thankful to IComp and the entire PPGI administrative team, whose support and infrastructure were essential for the development of this research. I also acknowledge CAPES for the financial support, without which this work would not have been possible.

Finally, I would like to express my special thanks to my friends Yune, Meng, Bruno, Euler, Giovanni, and Thiago for their companionship, inspiring conversations, and constant encouragement, which made this journey lighter and more meaningful. Likewise, I am grateful to my colleagues and lab partners, whose exchange of ideas, discussions, and collaboration were decisive for the maturation of this research.

"O tempo molda o que a pressa destrói"
(Andrés Peralta)

Resumo

A recuperação de paisagens sonoras semelhantes é essencial para o monitoramento bioacústico e ecoacústico, tarefas que continuam sendo desafiadora devido ao grande volume de dados não rotulados, ao ruído ambiental e à complexidade das cenas acústicas. Este estudo propõe um sistema eficiente que integra *embeddings* extraídos de um modelo de *deep learning* pré-treinado, combinados com uma técnica de redução de ruído e estratégias de fusão de vetores de *features* para viabilizar a recuperação baseada em similaridade em uma base de dados vetorial. Avaliamos o sistema utilizando gravações de aves, anfíbios e mamíferos em quatro metodologias experimentais, incluindo um estudo de caso focado em espécies ameaçadas. Os resultados mostram que os vetores de *embeddings* superam consistentemente as *features* tradicionais de MFCC na captura da similaridade acústica e que o algoritmo de busca aproximada (HNSW) melhora significativamente tanto a precisão da recuperação quanto a eficiência das consultas. Além disso, o sistema recupera de forma eficaz gravações da espécie criticamente ameaçada *Crax alberti*, permitindo o mapeamento de sua distribuição geográfica e destacando seu potencial para o planejamento da conservação.

Palavras-chave: Monitoramento bioacústico, Ecoacústica, Modelos Deep Learning pré-treinados, Base de dados vetorial.

Abstract

The retrieval of similar soundscapes is essential for bioacoustic and ecoacoustic monitoring, tasks that remain challenging due to the large volume of unlabeled data, environmental noise, and the complexity of acoustic scenes. This study proposes an efficient system that integrates embeddings extracted from a pre-trained deep learning model, combined with a noise reduction technique and feature vector fusion strategies to enable similarity-based retrieval in a vector database. We evaluated the system using bird, amphibian, and mammal recordings across four experimental methodologies, including a case study focused on endangered species. The results show that embedding vectors consistently outperform traditional MFCC features in capturing acoustic similarity and that the approximate search algorithm (HNSW) significantly improves both retrieval precision and query efficiency. Additionally, the system effectively retrieves recordings of the critically endangered species *Crax alberti*, allowing for the mapping of its geographic distribution and highlighting its potential for conservation planning.

Keywords: Acoustic monitoring, Bioacoustic, Deep Learning, Pre-trained models, Vector database.

Lista de Figuras

Lista de Tabelas

Sumário

1

Introdução

A degradação acelerada de ecossistemas, impulsionada por desmatamento, urbanização e poluição, tem provocado o declínio populacional de diversas espécies e exige soluções tecnológicas eficazes para a preservação da biodiversidade. Neste contexto, este trabalho investiga a interseção entre ecologia acústica, conservação ambiental e inteligência artificial, visando o desenvolvimento de um sistema eficiente para a recuperação de paisagens sonoras acusticamente semelhantes. Segundo Murray [1993], as paisagens sonoras representam o conjunto de sons de um ambiente, incluindo elementos naturais (*biophony* e *geophony*) e sons resultantes da atividade humana (*antrophony*).

A incorporação de técnicas de aprendizado de máquina tem impulsionado significativamente a bioacústica, permitindo explorar padrões complexos em dados acústicos massivos [Bianco et al., 2019]. Trabalhos recentes destacam os desafios associados à alta variabilidade e à presença de ruído nos sinais [Vargas-Masís et al., 2021, Vasconcelos et al., 2019], tornando necessária a aplicação de modelos robustos para filtrar dados relevantes [Quaderi et al., 2022]. Além desse problema, outro desafio importante está relacionado à limitação das abordagens supervisionadas [Burkov, 2019], que dependem de rótulos, enquanto técnicas não supervisionadas se mostram mais promissoras para lidar com grandes conjuntos de dados não rotulados [Xie et al., 2023]. Segundo Stowell

[2022], a recuperação acústica baseada em similaridade depende da extração de características representativas, como MFCCs, energia, entropia e centroide espectral[Cover & Hart, 1967]. No entanto, não existe uma única função $f(x)$ capaz de capturar as informações relevantes dos sinais sonoros para a tarefa de recuperação acústica.

O objetivo deste trabalho é desenvolver uma função $f(x)$ para a recuperação por similaridade de paisagens sonoras, permitindo identificar rapidamente gravações semelhantes com o intuito de facilitar o monitoramento e estudo dos ecossistemas. O desenvolvimento desta pesquisa será realizado implementando um modelo *Deep Learning* pré-treinado de forma não supervisionada para extrair vetores de características (ou *embeddings*) das gravações coletadas. O modelo que iremos avaliar é Perch, que se destaca por ter sido pré-treinado com dados bioacústicos, por gerar *embeddings* acústicos representativos e tem sido utilizado em tarefas de analise acústico [Ghani et al., 2023].

Posteriormente, os vetores extraídos pelo modelo Perch serão armazenados em um banco de dados vetorial (ex: VectorDB) junto com seus metadados, como por exemplo o local da gravação, dia e hora, tipo de gravador, etc. Por fim, serão realizadas consultas (*queries*) e mensurada a similaridade dos dados recuperados. Dessa forma, a pesquisa propõe um fluxo completo que vai da captura de uma gravação em qualquer região geográfica à sua representação vetorial, permitindo consultas eficientes por similaridade. Esperamos que os resultados obtidos revelem qual metodologia consegue os vetores mais semelhantes aos da consulta.

1.1 O problema da recuperação de paisagens sonoras acusticamente semelhantes

O monitoramento de paisagens sonoras, realizado por meio de gravadores autônomos, gera grandes volumes de dados acústicos, na sua maioria não rotulados. A criação

manual de conjuntos de dados anotados para treinar modelos supervisionados exige tempo e conhecimento de especialistas para identificar eventos acústicos de interesse [Presannakumar & Mohamed, 2023]. Este excesso de dados brutos impõe o primeiro grande desafio, o qual é como analisar e organizar grandes bibliotecas de áudio de forma eficiente e escalável.

O segundo desafio reside na própria complexidade dos sinais. Como discutido por Schafer [1969] e ilustrado na Figura 1.1 as paisagens sonoras variam drasticamente dependendo do ambiente. Gravações em contextos urbanos ou periurbanos apresentam uma sobreposição de sons muito maior do que em zonas rurais, resultando em espectrogramas com alta densidade de informação e baixa relação sinal-ruído. Além disso, áudios bioacústicos são frequentemente capturados em condições não controladas, contendo ruído ambiental e sinais que podem estar fora do espectro auditivo humano [Bradbury et al., 1998].

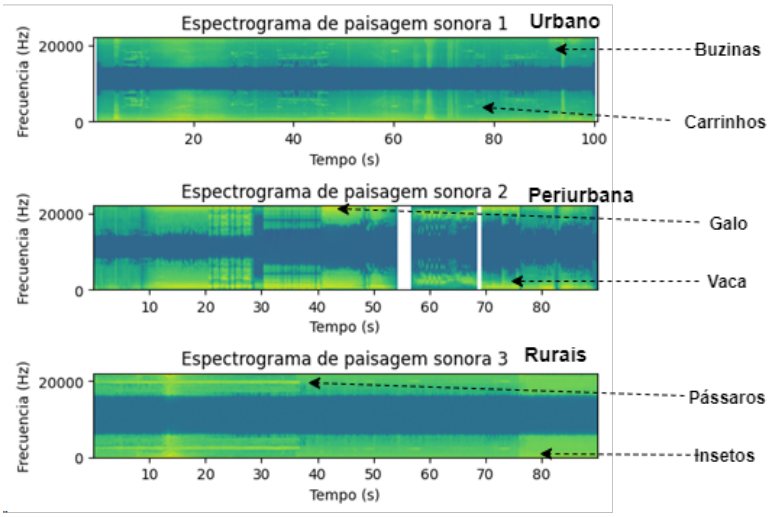


Figura 1.1. Complexidade das paisagens sonoras.

Diante dessa complexidade, o desafio central da recuperação por similaridade consiste em obter uma função de representação $f(x)$ que seja robusta o suficiente para mapear um sinal de áudio complexo a um espaço onde a proximidade entre vetores corresponda à similaridade acústica percebida. Tradicionalmente, a recuperação de

conteúdo era associada a técnicas de *hashing*, que frequentemente utilizam representações binárias. No entanto, para capturar as nuances de sinais contínuos como o som, representações baseadas em vetores de números reais (*embeddings*) são mais adequadas [Stowell, 2022]. Por este motivo, cunhamos o termo "recuperação acústica", que se baseia no cálculo de distância entre esses vetores em um espaço multidimensional.

Finalmente, um desafio prático é que as gravações de campo possuem durações variáveis. Isso impede a comparação direta e exige uma estratégia para converter uma matriz de características de tamanho variável $X \in \mathbb{R}^{D \times M}$ (onde D é a dimensão do vetor e M o número de segmentos) em uma representação vetorial única e de tamanho fixo, $\overline{x} \in \mathbb{R}^{D}$, adequada para indexação e consulta eficiente. A solução para estes desafios constitui o foco metodológico deste trabalho, que será detalhado no Capítulo 4.

1.2 Motivação ambiental

Segundo Williams & Bolitho [2003], fatores como mudanças climáticas, desmatamento, poluição, urbanização e degradação do habitat estão entre as principais causas da extinção de espécies vegetais e animais, um processo irreversível que ameaça o equilíbrio ecológico. Estima-se que a biodiversidade atual seja de aproximadamente 30 milhões de espécies, porém a perda anual alcança cerca de 30.000 espécies, indicando que estamos diante de uma possível sexta extinção em massa [Rodríguez, 2018]. De acordo com a União internacional para a Conservação da Natureza (UICN), o desmatamento representa uma ameaça significativa para muitas espécies de répteis, uma vez que elas se concentram nas florestas tropicais. Além disso, as mudanças ambientais, como as relacionadas às alterações climáticas, afetam essas espécies de forma particular [The IUCN Red List of Threatened Species, 2025].

De acordo com a União Internacional para a Conservação da Natureza (UICN), o número de espécies ameaçadas tem aumentado de forma acelerada. Na versão de

2025 da Lista Vermelha, 64.411 espécies encontram-se classificadas como ameaçadas, mais que o dobro em relação a 2007 [The IUCN Red List of Threatened Species, 2025]. Entre 2007 e 2025, os principais aumentos foram observados em peixes (+196%), corais construtores de recifes (+44%), mamíferos (+27%), répteis (+26%) e insetos (+16%). Esses dados alarmantes evidenciam a urgência por novas estratégias de monitoramento. Nesse sentido, o monitoramento acústico passivo (PAM), combinado com ferramentas de inteligência artificial, surge como uma abordagem escalável e não invasiva. Desenvolver protocolos computacionais robustos para analisar esses dados de forma automatizada é, portanto, essencial para estimar o estado das populações com maior precisão e apoiar políticas de conservação mais ágeis e informadas.

Nesse contexto, esta pesquisa busca contribuir com o desenvolvimento de uma ferramenta que auxilia na organização e análise de dados acústicos, possibilitando estudos comparativos entre diferentes ecossistemas e subsidiando ações estratégicas para proteger espécies ameaçadas e manter o equilíbrio dos ecossistemas. A pesquisa não apenas contribui para o conhecimento científico, mas também impacta diretamente na formulação e execução de políticas públicas, fortalecendo a capacidade das sociedades para enfrentar os desafios na conservação da fauna e contribuir para o desenvolvimento de estratégias sustentáveis ao nível global. A Figura 1.2 apresenta a evolução recente do número de espécies ameaçadas.

1.3 Justificativa

Os avanços tecnológicos têm impulsionado a coleta, processamento e análise de grandes volumes de dados bioacústicos, permitindo um monitoramento ambiental em escala global [Cai et al., 2007]. Nesse contexto, o uso de *Machine Learning*, especialmente redes neurais profundas, tem se destacado na identificação de espécies e na recuperação acústica de paisagens sonoras, superando as limitações da inspeção visual manual e permitindo a descoberta de padrões complexos [Farina & Li, 2022, Kurth et al., 2007,

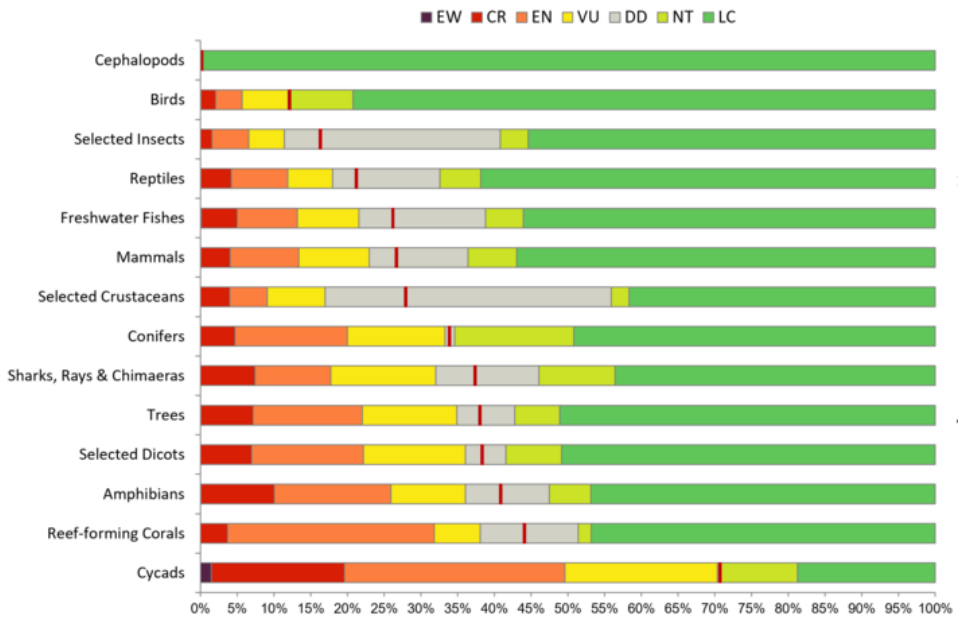


Figura 1.2. Número de espécies ameaçadas por categoria de risco. Categorias: EW – *Extinct in the Wild* (Extinto na Natureza), CR – *Critically Endangered* (Criticamente Ameaçado), EN – *Endangered* (Em Perigo), VU – *Vulnerable* (Vulnerável), NT – *Near Threatened* (Quase Ameaçado), DD – *Data Deficient* (Dados Insuficientes), LC – *Least Concern* (Pouco Preocupante). Fonte: The IUCN Red List of Threatened Species (2025).

Wang & et al., 2003]. Técnicas não supervisionadas são particularmente relevantes para lidar com os extensos conjuntos de dados gerados por monitoramento acústico passivo, dispensando a necessidade de rótulos prévios e acelerando análises em larga escala.

Contudo, a aplicação prática dessas técnicas ainda encontra barreiras significativas. Plataformas de referência como o repositório web Xeno-canto é caracterizado por armazenar gravações de sons de animais provenientes de diversas partes do mundo; o que o torna uma plataforma valiosa para profissionais na área de biodiversidade e ecologia da fauna, graças às suas funções que permitem ouvir, baixar, explorar e compartilhar gravações, por exemplo, ainda dependem de buscas baseadas exclusivamente em metadados textuais, como o nome da espécie. Essa abordagem limita fundamentalmente a descoberta de gravações acusticamente semelhantes quando a espécie é desconhecida ou não está anotada. Superar essa limitação exige o desenvolvimento de

métodos de recuperação baseada em conteúdo, que operem diretamente sobre o sinal de áudio.

É neste ponto que a presente pesquisa se justifica. Embora trabalhos recentes como os de Xie & Zhu [2023a] e Lü et al. [2024] a fusão de *features* para tarefas de classificação, uma análise comparativa e sistemática de técnicas de fusão aplicadas sobre *embeddings* para a recuperação de paisagens sonoras em bancos de dados vetoriais ainda não foi adequadamente explorada. Portanto, este trabalho propõe um método não supervisionado para preencher essa lacuna, com o objetivo de viabilizar buscas por similaridade acústica em grandes acervos de dados, como e mostrado na Figura 1.3.

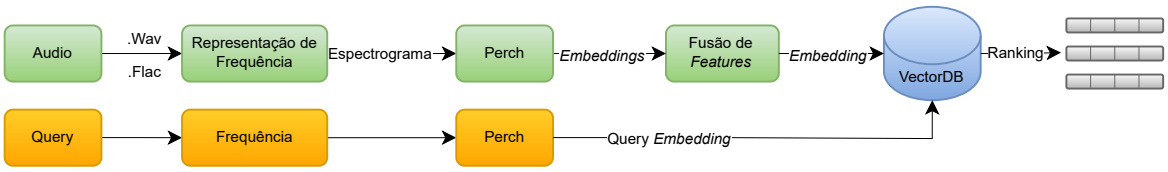


Figura 1.3. Sistema bioacústico para recuperação de paisagens sonoras acusticamente semelhantes.

1.4 Abordagem

A nossa abordagem se baseia no uso de bancos de dados vetoriais, que são sistemas especializados para a indexação e consulta eficiente de vetores de alta dimensionalidade. Para aproveitar os avanços recentes em aprendizado de máquina, primeiramente convertemos os sinais acústicos unidimensionais em representações visuais bidimensionais (espectrogramas). Essa transformação nos permite utilizar modelos de *Deep Learning*, originalmente projetados para tarefas de visão computacional, como extratores de características.

O fluxo de trabalho consiste em usar um modelo pré-treinado, o qual recebe como entrada um segmento de áudio e o transforma em um espectrograma, de onde extrai vetores de *embeddings* de dimensão fixa, que são indexados e armazenados na base de dados vetorial, possibilitando consultas rápidas e precisas por similaridade

para comparar gravações acústicas de diferentes ecossistemas. A avaliação do sistema será realizada com métricas padrão de recuperação de informação, focando tanto na precisão dos resultados quanto na eficiência computacional, quantificada através do tempo médio de resposta e seu respectivo desvio padrão.

A abordagem proposta permite recuperar eventos sonoros similares com base em uma função de similaridade no espaço de *embeddings*, independentemente das espécies presentes nas gravações, facilitando o monitoramento e estudo comparativo dos diferentes ecossistemas. Além disso, a recuperação acústica de paisagens sonoras pode contribuir para a criação de experiências imersivas e enriquecedoras para os usuários, permitindo-lhes explorar e apreciar a diversidade sonora dos diferentes ecossistemas. Os critérios utilizados para escolher a tecnologia de VectorDB encontram-se descritos na Seção 3.5.

1.5 Hipótese

Propomos que a combinação de (i) *embeddings* bioacústicos pré-treinados, agregados por fusão em nível de gravação, com (ii) indexação por busca aproximada (HNSW) produz melhor recuperação por similaridade do que representações clássicas (MFCC) sob latência viável. Além disso, postulamos que o desempenho depende criticamente da formação da *query* — incluindo seleção/fusão de segmentos e pré-processamento (*denoise*) — e que escolhas informadas nessa etapa elevam a qualidade sem custo proibitivo de tempo.

Dada a nossa hipótese, este trabalho investiga a seguinte questão central: a combinação de *embeddings* bioacústicos pré-treinados, fusão em nível de gravação e indexação por busca aproximada é superior a representações clássicas sob restrições de latência, e em que medida o desempenho depende da formação da *query*?

Para respondê-la, desdobramos em subperguntas: (PP1) qual estratégia de fusão otimiza a relação qualidade-latência; (PP2) como a formação da *query* (seleção de

segmento, pré-processamento e fusão de segmentos) altera os resultados; (PP3) qual métrica/normalização é mais adequada; (PP4) qual configuração de ANN oferece o melhor *trade-off*; e (PP5) se o sistema é operacionalmente útil no caso de espécies vulneráveis. Cada subpergunta é avaliada por métricas de Recuperação de Informação e por indicadores de engenharia (latência, *queries* atendidas por segundo e memória) ao longo dos capítulos desta dissertação.

1.6 Objetivos

Desenvolver um método não supervisionado baseado em uma arquitetura de Rede Neural Profunda pré-treinada a fim de mapear uma paisagem sonora em um vetor de *embedding* que capture melhor a similaridade entre dados ecoacústicos para poder realizar recuperação de paisagens sonoras e resolver de forma eficiente as consultas em um banco de dados vetorial.

Os objetivos específicos são descritos abaixo:

1. Avaliar sistematicamente diferentes estratégias de fusão de *features* para determinar qual otimiza a relação entre a qualidade da recuperação e a latência computacional;

2. Investigar como a formação da *query* — incluindo o pré-processamento do sinal, a seleção de segmentos e a fusão impacta o desempenho do sistema de recuperação;

3. Determinar a combinação mais eficaz entre algoritmo de busca por similaridade e métrica de distância para organizar e consultar grandes volumes de dados ecoacústicos de forma eficiente; e

4. Validar a aplicabilidade operacional do método proposto através de um caso de uso focado no monitoramento de espécies vulneráveis, utilizando dados de diferentes regiões geográficas.

1.7 Organização da proposta

Esta dissertação está organizada da seguinte forma. No Capítulo 2 são apresentados os fundamentos teóricos necessários para compreender os métodos adotados. No Capítulo 3, é apresentada uma síntese dos trabalhos relacionados, expondo as vantagens e desvantagens dos métodos existentes. A abordagem proposta é descrita no Capítulo 4; a avaliação dos resultados é mostrada nos Capítulos 5 e 6. É por fim, no Capítulo 7 apresenta as conclusões, limitações e linhas de trabalho futuro.

2

Fundamentos Teóricos

Neste capítulo, são apresentados os conceitos preliminares necessários para a compreensão e desenvolvimento do trabalho. A fundamentação teórica está dividida em doze Seções. A Seção 2.1 abrange um resumo da ecoacústicas e dos sinais sonoros, incluindo seus aspectos fundamentais, como os espectros de frequência e suas características. As Seções 2.2, 2.4 e 2.5 definem as etapas correspondentes ao pré-processamento do método, utilizadas para o aprimoramento dos dados. Nas Seções 2.6 e 2.7, é descrito o processo de extração de *embeddings*. Em seguida, na Seção 2.8 e 2.9, é explicado o mapeamento e o armazenamento das informações e mostrado como será realizado o processo de recuperação das informações e, por último, na Seção 2.12, são apresentadas as considerações finais deste capítulo.

2.1 Ecoacústica e Bioacústica

Farina & Gage [2017] definem a ecoacústica como a ciência que estuda o som nos ecossistemas, analisando os sinais ambientais principalmente naturais para compreender a composição, estrutura e dinâmica das comunidades biológicas. Essa abordagem permite avaliar a biodiversidade acústica e os impactos do ruído antropogênico na paisagem sonora, oferecendo insights sobre a saúde dos ecossistemas.

O Monitoramento Acústico Passivo tem se destacado como método não invasivo e de baixo custo, fornecendo ampla cobertura espacial e temporal por meio de técnicas modernas de processamento digital de sinais Bjorck et al. [2019]. Essa estratégia possibilita registrar, identificar e analisar padrões acústicos sem necessidade de intervenção direta.

A bioacústica, por sua vez, foca especificamente nos sons produzidos por organismos, investigando como são gerados, sua função comunicativa e seu papel na ecologia das espécies Kvsn et al. [2020]. Combinando gravações, análise espectral e técnicas computacionais, ela contribui para a compreensão da diversidade sonora e para estudar a variedade de sons biológicos, sua evolução, a interação entre indivíduos e espécies.

2.2 Sinais Sonoros e sua frequência

Os sinais sonoros são variações de pressão em um meio elástico, resultantes da vibração de uma fonte sonora, como a voz humana, instrumentos musicais ou vocalizações animais [Speaks, 2018]. Suas principais características incluem a frequência (Hz), que determina o tom percebido; a amplitude (m), associada à intensidade sonora; a duração (s), que define o tempo de emissão; e a forma de onda, que caracteriza o padrão de propagação.

No contexto da ecologia acústica, esses sinais são fundamentais para analisar interações entre espécies, estudar padrões de comunicação e monitorar ecossistemas [Frommolt et al., 2008]. Técnicas modernas de gravação permitem capturar sons ambientais de forma precisa, possibilitando a identificação de eventos acústicos de interesse. Os sinais sonoros são o principal meio de comunicação no mundo, tanto para os seres humanos como para inúmeras espécies animais. A Figura 2.1 mostra uma representação de uma onda sonora, na vertical, a intensidade ou o volume do som, enquanto na horizontal, a propagação da onda sonora no espaço.

A frequência sonora representa o número de ciclos de compressão e rarefação por

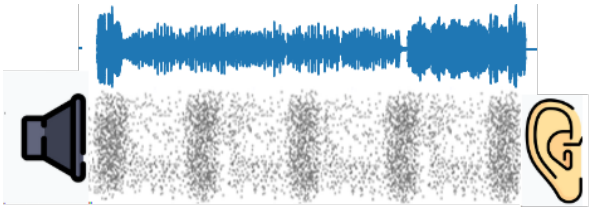


Figura 2.1. Propagação de uma onda sonora.

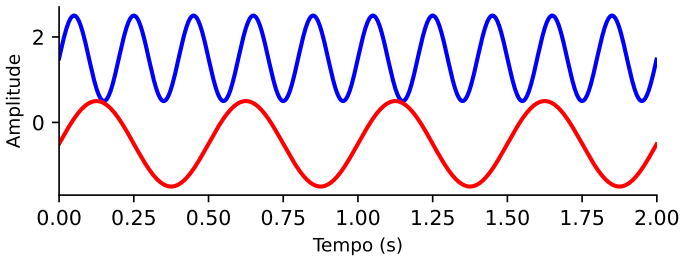


Figura 2.2. Ondas sonoras com alta cor azul e baixa frequência cor vermelho.

segundo, expressos em Hertz (Hz) [Cuadrado & Domínguez, 2019]. Frequências mais altas produzem tons agudos, enquanto frequências mais baixas estão associadas a tons graves. A Figura 2.2 ilustra visualmente essa diferença.

2.2.1 Frequência de amostragem

A frequência de amostragem f_s representa o número de amostras coletadas por segundo para converter um sinal analógico em digital [Oppenheim et al., 1999]. Por exemplo, uma taxa de $44,1\,kHz$ indica que 44.100 amostras são registradas a cada segundo. Quanto maior o f_s, maior é a fidelidade da representação digital do som. De acordo com o teorema de Nyquist, para que um sinal seja corretamente reconstruído, a frequência de amostragem deve ser no mínimo o dobro da maior frequência presente no sinal, conforme a Equação 2.1:

$$f_s = 2f_{\max} \tag{2.1}$$

onde $f_{\max}$ é o componente de frequência máxima presente no sinal. Caso essa condição não seja atendida, ocorre *aliasing*, ou seja, a sobreposição de componentes de frequência, produzindo distorções na reconstrução. A Figura 2.3 Ao amostra um sinal

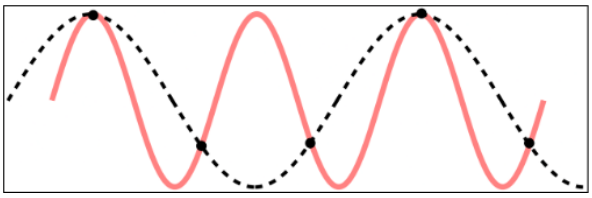


Figura 2.3. Representação do Teorema de Nyquist.

de alta frequência (linha pontilhada) com uma taxa de amostragem insuficiente (pontos pretos), os dados podem ser erroneamente interpretados como um sinal de baixa frequência (linha contínua).

2.2.2 Quantização

A quantização é o processo de converter valores contínuos de amplitude em níveis discretos predefinidos, sendo um passo essencial na digitalização de sinais [Polastre et al., 2004]. Esses níveis dependem da profundidade de bits, que define a resolução do sistema: com n bits, podem ser representados 2^n valores distintos. Quanto maior a profundidade, mais precisa é a representação do sinal. Contudo, uma profundidade de bits insuficiente resulta no chamado erro de quantização, que introduz pequenas distorções no sinal digitalizado. A Figura 2.4 ilustra esse processo, onde o sinal original (em azul) é aproximado pelo sinal quantizado (em laranja), cujos valores de amplitude são restritos a um conjunto finito de níveis.

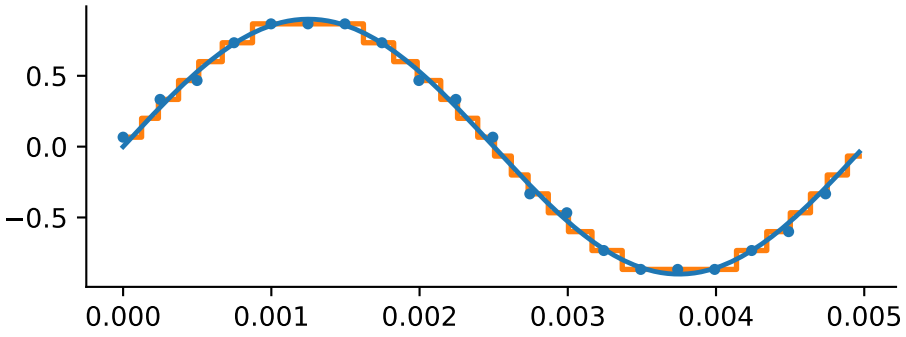


Figura 2.4. Quantificação de um sinal de áudio.

2.2.3 Qualidade de áudio vs frequência de amostragem

A qualidade de áudio refere-se à fidelidade com que um som é capturado e reproduzido, sendo influenciada principalmente pela frequência de amostragem, profundidade de bits e compressão [Lyons, 1997]. Taxas de amostragem mais altas permitem representar com maior precisão as componentes de alta frequência, aumentando a resolução do sinal. Entretanto, valores muito elevados exigem mais espaço de armazenamento e podem introduzir ruído indesejado, tornando essencial um equilíbrio entre qualidade e eficiência [Oppenheim et al., 1999].

Considerando a opinião de Oppenheim et al. [1999], a frequência de amostragem é um fator importante que determina a qualidade de um sinal de áudio porque tem um impacto direto na largura de banda do sinal. Taxas de amostragem mais altas capturam mais detalhes nas altas frequências dos sinais de áudio, ampliando o intervalo de representação entre a frequência máxima e a mínima. A Tabela 2.1 resume a relação entre taxa de amostragem, espaço ocupado e qualidade percebida. Para aplicações de bioacústica, taxas próximas a $44,1\,kHz$ são ideais, pois permitem capturar sinais até aproximadamente $22\,kHz$, cobrindo toda a faixa audível humana e a maioria das vocalizações de espécies monitoradas.

Tabela 2.1. Relação entre taxa de frequência de amostragem e bits.

Frequência de amostragem	Espaço em Mbs	Qualidade do som
48.000 Hz	Muito alto	Muita alta
44.100 Hz	Alto	Alta
22.050 Hz	Moderado	Moderada
11.025 Hz	Baixo	Baixa (impossível para capturar as espécies que vocalizam acima dos 5,5 kHz)

O ouvido humano percebe frequências entre $20\,Hz$ e $20\,kHz$; sons abaixo são classificados como *infrassom* e acima como *ultrassom* [Oppenheim et al., 1999]. Assim, amostragens abaixo de $20\,Hz$ resultam em sinais sem informação relevante, enquanto

valores muito acima de $44\,kHz$ podem incluir ruído de alta frequência, comprometendo a integridade do áudio.

2.2.4 Normalização

Antes de utilizar os modelos DL, é importante garantir que os valores de amplitude das gravações não possuam intervalos dinâmicos distintos, senão, os valores dos *embeddings* podem ser afetados. A falta de normalização da amplitude pode afetar principalmente o cálculo da distância Euclidiana entre os vetores de *embeddings*. Para evitar isso, os valores das amplitudes das gravações são normalizados para obter média zero e desvio padrão um [Theodoridis et al., 2010]. Assumindo que cada característica tem N valores, a normalização é realizada utilizando:

$$\mathrm{Z} = \frac{x_i - \mu}{\sigma}, \quad i = 1, 2, \ldots, N \tag{2.2}$$

onde Z representa o valor normalizado, x_i é o valor original da amostra temporal do áudio, μ e σ são os valores da média e do desvio padrão, respectivamente, da característica x.

2.2.5 Segmentação

De acordo com Harma [2003] a segmentação de áudio consiste em dividir um sinal contínuo em partes menores e mais fáceis de analisar, sendo um passo essencial para extrair informações relevantes. Existem diversas abordagens para segmentar um áudio, como a detecção de silêncios ou a identificação de variações de energia. No entanto, para o processamento com modelos de *Deep Learning*, a abordagem mais comum é a segmentação com janelas de comprimento fixo, pois garante que cada entrada do modelo possua a mesma dimensão [Burkov, 2019].

A escolha do tamanho do segmento é crucial, pois janelas muito curtas podem não capturar eventos acústicos completos, enquanto janelas muito longas podem conter

excesso de ruído ou múltiplos eventos, diluindo a informação relevante. Além disso, como modelos pré-treinados exigem entradas de tamanho uniforme para gerar espectrogramas consistentes, o que nos leva a manter a segmentação fixa [Farina & Li, 2022, Narasimhan et al., 2017, Vidaña-Vila et al., 2017].

2.2.6 Energia do Sinal

A energia (E) de um sinal é uma característica de análise no domínio do tempo, utilizada, por exemplo, no processo de segmentação de sinais bioacústicos [Vaca-Castaño & Rodriguez, 2010]. Esta função, que mede a força do sinal dentro de uma janela de análise, é definida como:

$$E = \sum_{n} [x(n)w(n)]^2 \tag{2.3}$$

em que $x(n)$ é o sinal e $w(n)$ é uma função de janela. Sendo L o comprimento da janela, esta característica possui uma complexidade computacional resultante de $O(L)$.

2.2.7 Valor Quadrático Médio (RMS)

O Valor Quadrático Médio (*Root Mean Square* - RMS) é uma medida estatística da magnitude de um sinal, frequentemente utilizada como um indicador da intensidade percebida dentro de uma janela [Oppenheim et al., 1999]. É calculado como a raiz quadrada da média dos quadrados dos valores das amostras:

$$x_{rms} = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} |x(n)|^2} \tag{2.4}$$

onde N é o número de amostras na janela. No contexto desta pesquisa, o valor RMS é particularmente relevante, pois é utilizado como *proxy* da relevância acústica na estratégia de fusão ponderada (*Weighted Average Pooling*), como será detalhado no Capítulo 4.

2.3 Similaridade acústica em áudios

A similaridade acústica busca quantificar o grau de semelhança entre dois sinais sonoros com base em suas características físicas ou perceptíveis, como timbre, frequência, intensidade, ritmo e forma de onda. Do ponto de vista computacional, o desafio é traduzir essas propriedades complexas em uma representação vetorial, onde a distância entre dois vetores possa refletir a semelhança acústica entre os áudios originais. Diferentemente da similaridade semântica, que se concentra no significado, a similaridade acústica fundamenta-se exclusivamente em aspectos sonoros objetivos Giordano et al. [2023] Um exemplo claro pode ser observado nos cantos de rãs: duas vocalizações podem soar semelhantes (alta similaridade acústica) mesmo pertencendo a espécies diferentes.

Do ponto de vista técnico, a similaridade acústica baseia-se em propriedades quantificáveis do sinal sonoro, como *modulações espectro-temporais*, *sonoridade*, *tom*, *timbre* ou *aspereza*. Essas propriedades permitem que a similaridade acústica seja considerada uma medida mais direta e objetiva em comparação com sua contraparte semântica. Nesse sentido, a bioacústica tradicional historicamente tem enfatizado a análise da *produção*, *propagação* e *recepção* dos sons animais a partir de suas características físicas [Odom et al., 2021].

2.4 Transformadas de domínio para os sinais acústicos

Para iniciar a etapa de pré-processamento, é necessário extrair características que representem os sinais bioacústicos. Essas características podem ser encontradas em termos de tempo ou frequência. No campo do tempo, elas podem ser calculadas diretamente, mas no campo da frequência, requer-se uma transformação prévia. As transformadas de sinais são técnicas utilizadas para converter e representar dados em diferentes for-

mas ou dimensões, para facilitar o processamento e a análise. Esta seção apresenta uma a Transformada de Fourier.

2.4.1 Transformada discreta de Fourier

A frequência é uma característica fundamental nos sinais sonoros, determinando quantas vezes a pressão oscila em um segundo. Embora isso seja válido para sinais simples, os sons bioacusticos são complexos e contêm múltiplos componentes de frequência simultaneamente. Para analisar essa composição, utiliza-se a Transformada Discreta de Fourier (DFT). Essa técnica matemática decompõe um sinal do domínio do tempo em seus componentes de frequência constituintes, revelando as amplitudes e fases de cada um [Rabiner & Gold, 1975]. Em essência, a DFT nos permite ver a composição de frequências que compõem um som em um determinado instante. Essa representação no domínio da frequência é o alicerce para a criação de espectrogramas, uma ferramenta visual essencial para a análise de áudio com redes neurais, que será abordada na próxima seção.

Para essa análise, um segmento do sinal $s(t)$ é representado por um vetor v_n, onde o índice n vai de 0 a $N-1$. Aqui, N representa o número de amostras no segmento (geralmente $N < T$) e corresponde também ao número de pontos da transformada. O espectro de frequências resultante, gerado pela DFT, representa a distribuição de amplitude do segmento em relação aos seus componentes de frequência. Formalmente, a DFT é definida como:

$$F_K = \sum_{n=0}^{N-1} v_n e^{-\frac{2\pi i k n}{N}}, \quad K = 0, 1, \ldots, N-1 \tag{2.5}$$

onde v_n é o vetor com as amostras do sinal de áudio, N é o número de pontos da transformada e F_K é o valor do componente de frequência no ponto $\omega = 2\pi k$ do

espectro. A relação entre a transformada contínua e discreta é dada por:

$$F_K = V(e^{i\omega})|\omega = \quad \frac{2\pi k}{N}, \tag{2.6}$$

onde F_K são amostras de $V(e^{iw})$ que formam uma progressão geométrica de inteiros com razão $\frac{2k}{N}$ [Oppenheim et al., 1999]. Da equação anterior, observa-se que para obter N pontos da transformada de Fourier, é necessário resolver a soma com N operações, resultando em uma ordem de complexidade $\theta = (N^2)$.

2.4.2 Transformada de tempo curto de Fourier

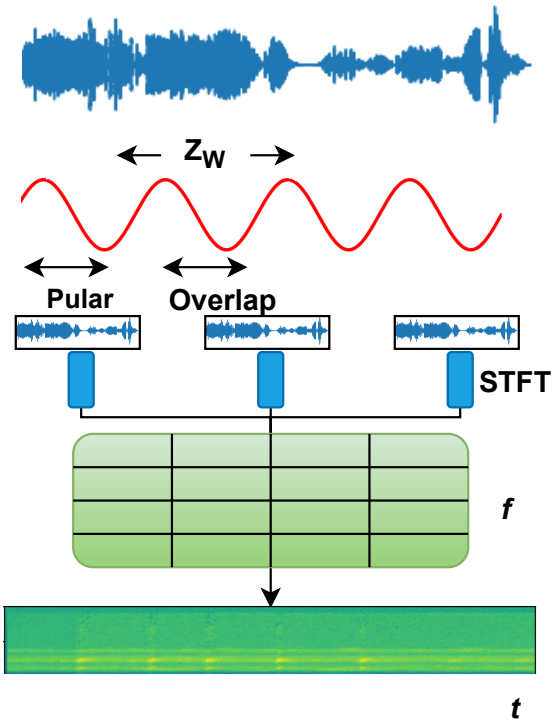
Para recuperar a informação temporal das diferentes frequências em todo o sinal, é utilizada a *short-time Fourier transform* ($STFT$) [Scholl, 2021]. A ideia básica é dividir o sinal longo em pequenos segmentos consecutivos e então realizar a transformada de Fourier nesses quadros individuais separadamente para compreender como as diversas frequências evoluem. O coeficiente k-ésimo da DFT para o t-ésimo quadro temporal de v_n é calculado como:

$$F[t,k] = \sum_{n=0}^{N-1} \omega[n]v[tH+n]e^{\frac{-i\pi kn}{N}}, \tag{2.7}$$

onde $\omega[n]$ é a função da janela que é utilizada para garantir que os quadros nos quais a DFT é calculada sejam periódicos e para levar em consideração as descontinuidades nos pontos finais que surgem devido ao número não inteiro de períodos que preenchem os quadros. Existem vários tipos de funções de janela, cuja escolha depende de diferentes aplicações. Algumas funções de janela comumente utilizadas são Retangular, Hamming, Hanning e Blackman-Harris. H é outro parâmetro que se refere ao deslocamento.

Para obter a Transformada de Fourier de Tempo Curto, podemos usar a seguinte equação aplicada ao sinal de áudio: $S = \frac{Z_w \cdot F_s}{O \cdot N} - 1$, onde Z_w é o tamanho da janela,

Figura 2.5. Aplicação da transformada de Fourier de tempo curto para obter um espectrograma de um sinal de áudio.



F_s é a frequência de amostragem do sinal, O representa a sobreposição (overlap), e N é o número de amostras. Esse cálculo é armazenado em um vetor de frequências que é então guardado em uma matriz A, e assim sucessivamente até obter todas as $STFTs$ do sinal de áudio. Cada um dos vetores de frequências representa um pixel no espectrograma resultante.

Os espectrogramas permitem visualizar as diversas contribuições de frequência de um sinal de áudio à medida que o tempo avança [Müller, 2015, Smith, 2008]. Na Figura 2.5, é capturada um sinal de áudio e suas características são extraídas no domínio do tempo e da frequência. Em seguida, o sinal é dividido em segmentos sobrepostos (overlap) usando uma janela fixa (Z_w). Para cada segmento, é calculada a $SFTF$ e suas características são obtidas e armazenadas em um vetor de frequências. Esses vetores são então guardados em uma matriz bidimensional A. Por fim, o espectrograma do sinal de áudio é gerado.

2.5 Filtro de ruído

O filtro de ruído para recuperação acústica de paisagens sonoras atua como um pré-processamento, eliminando ruídos que prejudicam a recuperação acústica, mas preservando as informações relevantes do áudio [Wang et al., 2015]. Nesta pesquisa, implementamos o algoritmo de redução de ruido *Noise Reduce* [Sainburg et al., 2021]. Este algoritmo utiliza o espectrograma do ruído presente no sinal de áudio, gerado pela *shorttime Fourier transform* ($STFT$). 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 binaria 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.

Para realizar a FFT, primeiro é necessário calcular a short-time Fourier transform ($STFT$) com N frequências que determinam a resolução no domínio da frequência do sinal de áudio. Isso é feito dividindo a taxa de amostragem fs pela resolução desejada [Smith, 2011]. Por exemplo, o modelo Perch recebe áudios com uma taxa de amostragem fs de 32 kHz a FFT é calculado com $N = 1024$, então cada banda de frequência ωf tem uma largura de $\Delta f = \frac{f_s}{N} = \frac{32000}{1024} = 31.25\,\text{Hz}$ geralmente N se escolhe como potência 2. Pelo fato da FFT ser uma transformada simétrica, a quantidade de bandas finalmente utilizadas é de 512 mas o coeficiente na frequência zero, ou seja 513 bandas totais [Yu et al., 2008].

Então, o número de quadros temporais por segundo é determinado pela equação $t = \frac{F_s}{N/4}$ considerando 50% de sobreposição entre cada quadro. Assim, o processo de passar um sinal através da $STFT$ resulta em uma matriz A com $N + 1$ linhas e t colunas por segundo:

$$A_{[f,t]} = \begin{bmatrix} a_{(1,1)} & \cdots & a_{(1,\frac{F_s}{N/4}+1)} \\ \vdots & \vdots & \ddots \\ a_{(\frac{N}{2},1)} & \cdots & a_{(\frac{N}{2},\frac{F_s}{N/4}+1)} \end{bmatrix} \tag{2.8}$$

onde cada coeficiente $a_{(f,t)} \in \mathbb{C}$ desta matriz pode ser um número complexo. Com esta matriz, calcula-se a magnitude média ao longo de todos os quadros temporais do sinal de áudio. Em seguida, procede-se ao cálculo do desvio padrão da magnitude do áudio em cada banda de frequência ao longo dos quadros temporais. Se o desvio padrão for muito alto, o dano causado ao áudio pela redução de ruído será maior do que o desejado. O próximo passo é calcular o limiar, que representa o limite do ruído, ou seja, separa o que é considerado ruído do sinal limpo. Para calcular esse limiar, soma-se o vetor da magnitude média ao produto entre o vetor de desvio padrão e um escalar determinado conforme um hyper parâmetro deste algoritmo limiar ou limite. Quanto mais próximo de 1 for o limite, maior será o número de amostras consideradas como ruído.

Depois desse cálculo, o vetor gerado é transposto e repetido nas linhas seguintes até que se alcance $\frac{n}{N/4} + 1$ linha. Em seguida, a matriz resultante é transposta, tendo a mesma forma que a matriz A. A partir da matriz que contém as magnitudes do sinal e o limiar calculado anteriormente, gera-se uma máscara binaria para o sinal. Depois aplica-se um filtro de média aritmética para suavizá-lo. Os coeficientes desse filtro podem ter diferentes pesos, contanto que a soma desses pesos seja igual a 1. Isso garante que o resultado final tenha um ganho unitário sobre o sinal filtrado, ou seja, que o sinal não sofra distorções [O'Haver, 1997].

Por outro lado, o filtro de suavização recebe como entrada as bandas de frequência B_f e bandas de tempo B_t, respectivamente, o filtro suavizará em torno da amostra em que está atuando. O valor padrão do algoritmo é $ff = 2$ e $df = 4$. Os vetores utilizados para construir o filtro de suavização em duas dimensões (frequência e tempo)

são obtidos calculando o produto externo entre dois vetores. Estes vetores são

$$\vec{f} = \left(\frac{1}{ff+1}, \frac{2}{ff+1}, \ldots, 1, \ldots, \frac{2}{ff+1}, \frac{1}{ff+1} \right), \tag{2.9}$$

$$\vec{t} = \left(\frac{1}{df+1}, \frac{2}{df+1}, \ldots, 1, \ldots, \frac{2}{df+1}, \frac{1}{df+1} \right). \tag{2.10}$$

Assim, a convolução bidimensional suaviza o sinal ao passá-lo este filtro em ambas dimensões simultaneamente, tempo e frequência [O'Haver, 1997]. O filtro $h(u, v)$ atuará sobre cada amostra do sinal, isto é, sobre os elementos da matriz A:

$$y(f,t) = h(u,v) * x(f,t) \Rightarrow y(f,t) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} h(u,v)x(f-u, t-v), \tag{2.11}$$

onde $y(f, t)$ representa a máscara após a convolução, $h(u, v)$ é a sequência que contém os coeficientes do filtro de suavização, e $x(f, t)$ é a máscara antes da suavização. Após completar o processo de redução de ruído, a função $istft$ é utilizada para realizar a *Transformada Inversa de Fourier de Tempo-Curto* ($ISTFT$), gerar um sinal temporal filtrado.

Na Figura 2.6, é realizado o processo de análise e filtragem do sinal de áudio. Primeiramente, é extraído o primeiro segmento do sinal e calculada a $SFFT$ para determinar o desvio padrão e estabelecer um limiar de referência para o áudio. Isso permite obter o perfil do áudio, conhecido como filtro de limiar (*filter threshold*). Em seguida, a $SFFT$ é aplicada a todo o sinal de áudio para gerar seu espectrograma, que representa a distribuição de energia do sinal em função do tempo e da frequência. Posteriormente, utiliza-se o filtro de limiar (*smoothing kernel*) obtido anteriormente para calcular o nível de ruído presente em todo o sinal de áudio, resultando em uma máscara binária que indica as regiões do sinal afetadas pelo ruído. Então, é aplicado um filtro de suavização tanto no domínio da frequência quanto no do tempo sobre o

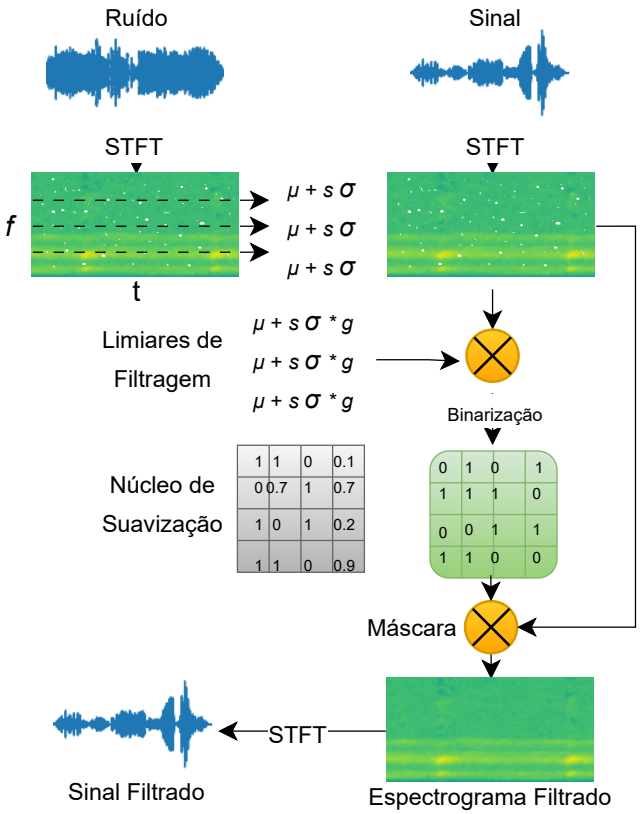


Figura 2.6. Aplicação do algoritmo de redução de ruído a um sinal de áudio. Onde μ é a média, σ o desvio padrão e g o fator de ganho.

sinal de áudio, produzindo um espectrograma filtrado que reduz o ruído e melhora a qualidade geral do sinal. Por fim, é realizada a transformada inversa da Transformada de Fourier de Tempo Curto para reconstruir o sinal de áudio filtrado, eliminando assim o ruído detectado durante o processo anterior.

2.6 Aprendizado de máquina (ML)

Burkov [2019] define o aprendizado de máquina (ML) como um ramo da ciência da computação que gera algoritmos a partir de um conjunto de exemplos de fenômenos específicos, sejam eles artificiais, naturais ou gerados por outros algoritmos. Esse processo envolve a coleta de dados e a construção algorítmica de um modelo estatístico com base nesses dados. Bianco et al. [2019] sustentam que é notável por seus recur-

sos de aprendizado de máquina para detectar e explorar automaticamente padrões nos dados. Diferentemente da abordagem tradicional no processamento de sinais e áudio, o ML depende muito dos dados. Com o conjunto de dados de treinamento correto, o aprendizado de máquina pode revelar relações complexas entre recursos e rótulos, e até mesmo entre as próprias recursos. Isso possibilita a criação de modelos que descrevem fenômenos acústicos complexos, como a voz humana, quando grandes quantidades de dados de treinamento estão disponíveis.

2.6.1 Deep learning (DL)

Lecun et al. [2015] consideram que a aprendizagem profunda permite que modelos computacionais, compostos de várias camadas de processamento, aprendam representações de dados em vários níveis de abstração. Esses métodos avançaram drasticamente o estado da arte em reconhecimento de áudio, reconhecimento de objetos visuais, reconhecimento de objetos e muitas outras áreas. A aprendizagem profunda usa algoritmos de retropropagação para descobrir estruturas complexas em grandes conjuntos de dados, nos quais a máquina ajusta os parâmetros internos usados para calcular a representação de cada camada a partir da representação da camada anterior.

Nesta pesquisa, adotamos um modelo pré-treinado baseado em DL, desenvolvido especificamente para dados bioacústicos e ecoacústicos, que permite extrair vetores de *embeddings* de alta qualidade a partir das gravações. Esses vetores são então utilizados para indexação e consultas por similaridade em nosso banco de dados vetorial, possibilitando a recuperação eficiente de arquivos acústicos com características semelhantes. A adoção desse modelo é motivada por resultados recentes que evidenciam o sucesso de abordagens de *Deep Learning* na classificação de espécies e na análise de ambientes sonoros [Barroso et al., 2023, Hagiwara et al., 2023, Mancusi et al., 2023].

2.6.2 Modelos pré-treinados

Segundo Ghani et al. [2023] os modelos de aprendizado de máquina que foram previamente treinados em grandes conjuntos de dados para realizar uma tarefa específica são chamados de modelos pré-treinados. Esses modelos são treinados em dados de treinamento e aprendem a reconhecer padrões e características relevantes para a tarefa em questão. Uma vez concluído o treinamento, os modelos pré-treinados podem ser usados como base para tarefas relacionadas, economizando tempo e recursos, já que não é necessário treinar um modelo do zero. No contexto da bioacústica, os modelos pré-treinados são usados para extrair características úteis dos dados de áudio e melhorar a eficiência e precisão dos modelos de classificação e detecção de sons, tornando-os uma ferramenta muito útil para o desenvolvimento desta pesquisa. Na Seção 2.6.3 serão apresentada a arquitetura usadas pelo modelo pré-treinado nesta pesquisa.

2.6.3 EfficientNet

A arquitetura EfficientNet B1 é uma rede neural convolucional que se baseia no conceito de escala composta para aprimorar a precisão e eficiência dos modelos de aprendizado profundo. Isso é alcançado por meio da utilização de uma combinação de blocos de expansão e blocos de profundidade. Os blocos de expansão e profundidade são utilizados para aumentar a capacidade da rede em capturar características complexas. Essa arquitetura foi desenvolvida por Tan & Le [2019]) e treinada originalmente com o conjunto de dados ImageNet, que abrange um total de 1.000 classes, produzindo vetores de *embeddings* com uma dimensão de 1280. Além disso, ela possui um total de 7.856.136,00 parâmetros treináveis e é composta por 28 camadas convolucionais.

EfficientNet B1 possui uma função dedicada à extração eficiente de *embeddings*. Inicialmente, é necessário carregar os dados, seguido da implementação da função de extração de características. Por fim, o modelo ficará responsável por obter os vetores de *embeddings* dos dados, que serão armazenados como características para uso pos-

terior. Em nossa pesquisa, a escolha de um modelo que implemente essa arquitetura é altamente benéfica, uma vez que sua eficiência computacional a torna ideal para o processamento de grandes volumes de dados, um requisito fundamental neste trabalho.

A Figura 2.7 apresenta a arquitetura da EfficientNet-B1. No lado esquerdo, são exibidos os sete blocos principais compostos por camadas MBConv com diferentes tamanhos de *kernel* e fatores de expansão, responsáveis pela extração progressiva de características até a geração dos *feature maps*. No lado direito, destaca-se a estratégia *compound scaling*, mostrando o balanceamento entre profundidade, largura e resolução para otimizar o desempenho e a eficiência do modelo.

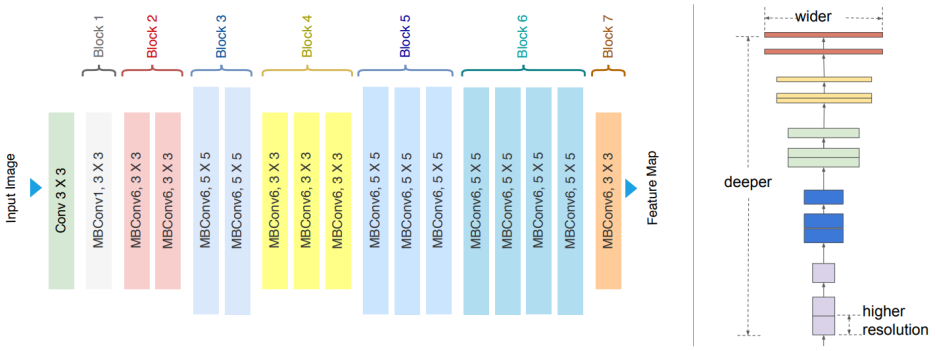


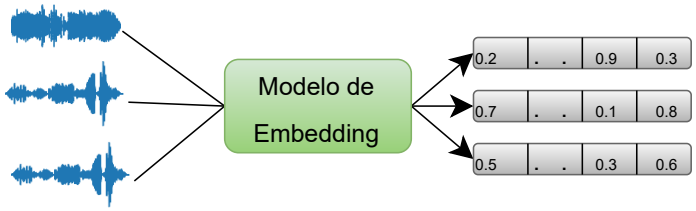
Figura 2.7. Arquitetura do EfficentNet B1. Imagem tomada de Tan, Le (2019)

2.7 Vetores de embedding

De acordo com Azar et al. [2023] os *embeddings* são vetores densos que representam dados em um espaço numérico contínuo, permitindo que algoritmos de *Machine Learning* operem de forma eficiente sobre diferentes tipos de entrada [Azar et al., 2023, Ghani et al., 2023]. Esses vetores capturam relações relevantes e possibilitam medir similaridades entre amostras por meio de distâncias no espaço vetorial [Devalraju & Rajan, 2022]. Em aplicações práticas, os *embeddings* permitem encontrar dados semelhantes aos originais por meio da comparação de suas representações numéricas, proporcionando uma ferramenta poderosa para a análise acústica e busca eficiente de informações por semelhança [Devalraju & Rajan, 2022].

No contexto desta pesquisa, utilizamos o modelo pré-treinado descrito na Seção 2.6.3 para extrair *embeddings* a partir dos espectrogramas 2D gerados a partir de segmentos de áudio. Cada vetor representa as características acústicas de uma amostra, permitindo indexação e consultas por similaridade em nosso banco de dados vetorial. As distâncias entre vetores são calculadas usando a métrica de distância euclidiana, adequada para dados com dependência tempo-frequência, conforme detalhado na Seção 2.8. Na Figura 2.8, ilustra o processo de transformação dos sinais de áudio em *embeddings* e o mapeamento dessas representações em um espaço n-dimensional, onde amostras semelhantes ficam mais próximas. Na nossa pesquisa, vamos utilizar a distância euclidiana como métrica principal para calcular a semelhança entre os *embeddings*, pois ela mede a distância entre dois pontos *embeddings* num espaço n-dimensional e a natureza dos nossos dados (áudios) possui características importantes (tempo e frequência).

Figura 2.8. Extração de vetores de *embeddings* de amostras de áudio.



2.8 Métricas para avaliar a similaridade entre consultas

Em sua pesquisa Pan et al. [2023] afirmam que as métricas para a avaliação de vetores de *embeddings* são técnicas usadas para medir a similaridade entre dois conjuntos de dados. No nosso caso, busca-se determinar quão semelhantes são os vetores de *embeddings* identificados como os mais similares à consulta realizada. Essas métricas são usadas para avaliar a capacidade dos vetores de *embeddings* em capturar a semântica e a

sintaxe dos dados. Algumas das métricas mais comuns são:

- **Distância Euclidiana:** É uma medida utilizada para avaliar a semelhança ou diferença entre pontos (*embeddings*) em um espaço euclidiano. Baseada no teorema de Pitágoras, ela é calculada como o comprimento do segmento de linha reta que une dois pontos. Nesta pesquisa, optamos pela distância euclidiana para mensurar a semelhança entre os vetores de *embeddings* devido ao seu amplo uso na literatura. A fórmula é expressa pela seguinte equação:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.12}$$

  onde x e y são dois vetores de dimensão n.

- **Similaridade de Cosseno:** É uma medida que avalia o cosseno do ângulo (θ) entre dois vetores não nulos em um espaço vetorial. Na recuperação de informação, mede a similaridade de orientação entre dois vetores, independentemente de sua magnitude. Quanto mais próximo de 1 for o valor do cosseno (ângulo pequeno), maior a similaridade. Como ilustrado na Figura 2.10, o vetor 1 (vermelho) é mais similar ao vetor 2 (verde) do que ao vetor 3 (azul), pois o ângulo entre eles é menor. A similaridade de cosseno é calculada como o produto escalar dos vetores dividido pelo produto de suas magnitudes:

$$\text{similarity}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.13}$$

  onde A e B são dois vetores de dimensão n.

Figura 2.9. Cálculo da distância euclidiana entre dois vetores de *embeddings*. Adaptado de Bishop [2006]

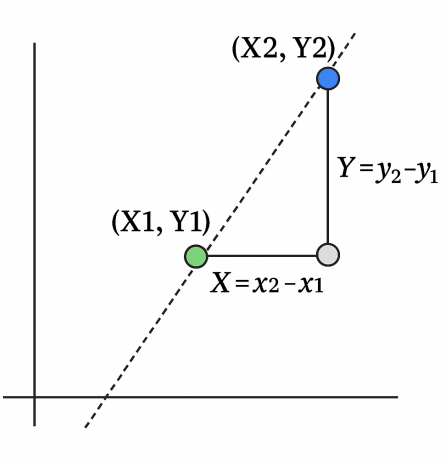
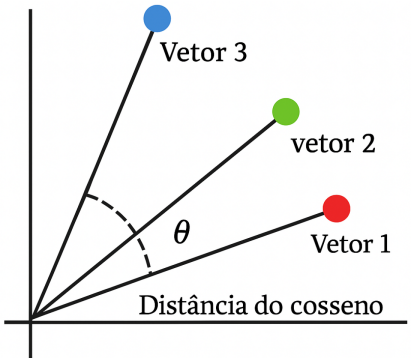


Figura 2.10. Ilustração da similaridade de cosseno, onde a semelhança é medida pelo ângulo θ entre os vetores.



2.9 Banco de dados vetoriais

As bases de dados vetoriais são sistemas projetados para armazenar, indexar e recuperar vetores de alta dimensionalidade (*embeddings*) de forma eficiente [Taipalus, 2024]. Esses vetores contêm atributos que descrevem objetos como textos, imagens ou sons. No contexto desta pesquisa, esses vetores representam características acústicas extraídas de segmentos de áudio, permitindo consultas baseadas em similaridade para identificar gravações com padrões sonoros semelhantes. Por outro lado, no contexto de consultas em bases de dados vetoriais, destaca-se a importância do processador de consultas, que desempenha um papel crucial na especificação de critérios de busca e na execução de

consultas, utilizando vários operadores, como projeção de similaridade, e otimizando o processo por meio de operadores indexados [Pan et al., 2023].

No entanto, o gerenciamento eficiente da alta dimensionalidade e dispersão dos dados vetoriais apresentam desafios, como a criação de índices quando se trata de vetores de *embeddings* (números reais) em vez de simples entradas em uma tabela. Devido à natureza densa e de alta dimensionalidade dos *embeddings*, a busca direta por varredura completa seria inviável. Para superar esse desafio, implementamos um processo de indexação que organiza os vetores em estruturas otimizadas, permitindo consultas rápidas sem necessidade de comparar todos os elementos [Pan et al., 2023]. Esse mecanismo reduz significativamente o tempo de busca e melhora a eficiência na recuperação de informações.

Outro desafio crucial que enfrentamos está relacionado com a necessidade de realizar cálculos de similaridade entre vetores de maneira eficiente para garantir resultados ótimos ao realizar consultas em nosso banco de dados vetorial. Para abordar esse problema, optamos pela implementação de algoritmos de recuperação de informações. Na Seção 2.10 fornecemos uma análise detalhada de cada um dos algoritmos, bem como as métricas que eles utilizam para obter os melhores resultados em uma consulta específica. A Seção 4 expande e justifica a escolha e implementação desses algoritmos, assim como a métrica de avaliação de similaridade utilizada.

2.9.1 Métricas para avaliar as consultas

De acordo com Pan et al. [2023], algumas das métricas mais comuns para avaliar consultas em um banco de dados vetorial são (salvo indicação em contrário, assumimos relevância binária por espécie):

- **Hit Rate:** O $H@k$ mede a proporção de resultados verdadeiramente relevantes entre os primeiros k resultados retornados pelo sistema e é definido como Krauss

et al. [2023]:

$$\text{H@k} = \frac{1}{k}\sum_{i=1}^{k} \text{r}(i) \tag{2.14}$$

onde

$$r(i) = \begin{cases} 1, & \text{se o resultado pertence à mesma espécie da consulta} \\ 0, & \text{caso contrário} \end{cases} \tag{2.15}$$

e 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 2.14 calcula a soma dos valores de $r(i)$ para os primeiros k 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 eficácia do sistema de recuperação. Um valor baixo de $H@k$ indica muitos resultados irrelevantes. Neste trabalho, adotamos $H@5$ [Krauss et al., 2023].

- **Mean Reciprocal Rank:** Mede a posição do *primeiro* relevante no Top-5, privilegiando acertos nas primeiras posições. Seja r_q a menor posição k com relevante para a consulta q ($r_q{=}\infty$ se não houver relevante em 1..5). Definimos a recíproca por consulta como $\text{RR@5}(q){=}1/r_q$ se $r_q \leq 5$ e 0 caso contrário, e reportamos

$$\text{MRR@5} \;=\; \frac{1}{N}\sum_{q} \text{RR@5}(q). \tag{2.16}$$

- **Mean Average Precision:** Resume a precisão acumulada ponderada pela ocorrência de relevantes no Top-5. Para cada consulta, $P(k){=}\frac{1}{k}\sum_{i=1}^{k} y_i$ é a precisão no corte k, e

$$\text{AP@5}(q) \;=\; \frac{1}{\min(R_q, 5)}\sum_{k=1}^{5} P(k)\, y_k, \tag{2.17}$$

onde $y_k \in \{0, 1\}$ indica relevância binária e R_q é o número total de relevantes disponíveis. Reporta-se $\text{mAP@5}{=}\frac{1}{N}\sum_q \text{AP@5}(q)$. Em cenários com um único

relevante efetivo no Top-5, AP@5 coincide com RR@5.

- **Normalized Discounted Cumulative Gain:** Avalia a qualidade do ordenamento com desconto logarítmico para posições mais baixas:

$$\mathrm{DCG@5}(q) \;=\; \sum_{k=1}^{5} \frac{y_k}{\log_2(k+1)}, \qquad \mathrm{nDCG@5}(q) \;=\; \frac{\mathrm{DCG@5}(q)}{\mathrm{IDCG@5}(q)}. \tag{2.18}$$

  A normalização usa a ordem ideal (IDCG@5); quando há um único relevante, IDCG@5=1.

- **Tempo de pesquisa (Query Time):** Mede o tempo decorrido desde o momento em que uma consulta é enviada até a obtenção dos resultados, reportado como média e desvio padrão.

- **Taxa de consultas por segundo (QPS):** Mede quantas consultas são processadas por segundo:

$$\mathbf{QPS} = \frac{\text{Número total de consultas processadas}}{\text{Tempo total em segundos}} \tag{2.19}$$

Finalmente, para medir a eficiência na velocidade das consultas, mediremos o tempo médio em milissegundos (ms) para cada consulta. Estas medições serão acompanhadas do cálculo do desvio padrão ($\pm$). Uma média de tempo pequena indica que o sistema responde rapidamente e é eficiente. Um desvio padrão pequeno indica que houve pouca variação nos tempos de resposta.

2.10 Algoritmos de recuperação de informações

De acordo com Pan et al. [2023] são projetados para localizar rapidamente os vetores mais semelhantes em grandes coleções de dados. No contexto desta pesquisa, esses algoritmos operam sobre os *embeddings* acústicos extraídos dos segmentos de áudio,

permitindo consultas eficientes no banco de dados vetorial. Processam as consultas do usuário por meio de técnicas como a ponderação de termos, a comparação de similaridades e a construção de modelos de linguagem para determinar a relevância dos dados. Existem diversos tipos de algoritmos de recuperação de informações, como o modelo de espaço vetorial, a busca booleana e a recuperação probabilística, cada um com suas próprias características e vantagens [Pan et al., 2023].

Após realizar uma revisão da literatura, constatou-se que o algoritmo canônico para a recuperação de vizinhos próximos é o k-Nearest Neighbors (k-NN) [Cover & Hart, 1967], que realiza uma busca exata ao calcular a distância entre o vetor de consulta e todos os outros vetores na base de dados. Embora simples e preciso, seu custo computacional o torna inviável para grandes volumes de dados, pois seu desempenho diminui significativamente com o crescimento do conjunto de dados, devido à necessidade de comparar todos os vetores.

Para superar essa limitação, utilizamos o algoritmo *Hierarchical Navigable Small World* (HNSW) [Malkov & Yashunin, 2018], um método baseado em grafos hierárquicos para busca aproximada dos k vizinhos mais próximos. Este algoritmo representa uma melhoria em relação ao algoritmo NSW (*Navigable-Small-World-Graph*), o qual envolve a criação de várias camadas ou níveis de grafos adjacentes conectados.O HNSW organiza os vetores em múltiplos níveis conectados e aplica uma heurística eficiente para evitar explorações desnecessárias, alcançando maior velocidade e boa precisão, mesmo em dados de alta dimensionalidade. Essa característica o torna particularmente adequado para bases compostas por *embeddings* acústicos. É importante destacar que o algoritmo HNSW se destaca por sua maior rapidez em comparação com os algoritmos kNN e IMENN.

2.10.1 Hierarchical navigable small world (HNSW)

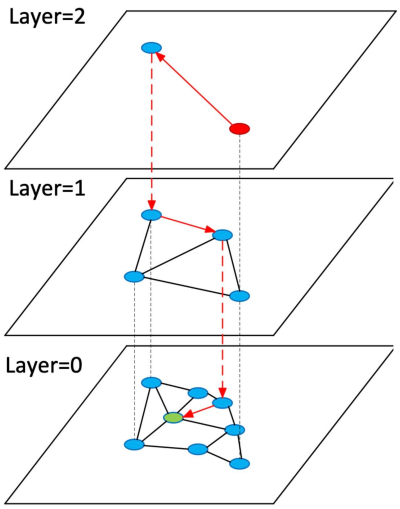
O *Hierarchical navigable small world* é um algoritmo projetado para a busca aproximada dos k vizinhos mais próximos, baseado em uma estrutura de grafos navegáveis [Malkov & Yashunin, 2018]. Sua base conceitual deriva do *k-Nearest Neighbors* (k-NN), mas o HNSW modifica a abordagem original ao introduzir uma organização hierárquica e conexões otimizadas para acelerar as consultas em espaços de alta dimensionalidade. Inicialmente, um grafo é construído representando os vetores de *embeddings*, onde cada nó é conectado aos seus vizinhos mais próximos. Esses nós são inseridos de forma incremental, criando uma estrutura em que os elementos mais semelhantes encontram-se interligados.

Uma das principais inovações do HNSW é a organização dos nós em diferentes níveis hierárquicos. Os níveis superiores contêm poucos elementos e são utilizados para iniciar a busca de forma rápida, enquanto os níveis inferiores possuem uma maior densidade de nós e são responsáveis por refinar os resultados. Durante a execução de uma consulta, o processo de busca começa em um nó localizado em um nível superior do grafo e, a partir dele, o algoritmo percorre as conexões descendo gradualmente pelos níveis. À medida que o algoritmo avança, uma lista de candidatos mais promissores é atualizada continuamente até que o nível mais baixo seja alcançado. A busca é interrompida quando não há mais candidatos relevantes ou quando um número máximo de nós é explorado.

Essa estratégia hierárquica reduz significativamente a complexidade da busca em comparação com algoritmos exatos, como o k-NN tradicional, tornando o HNSW altamente eficiente para bases de dados vetoriais extensas e de alta dimensionalidade, essa característica o torna particularmente adequado para a nossa pesquisa.

O algoritmo HNSW apresenta um desempenho eficiente e muito superior aos outros algoritmos de busca, e possui uma complexidade de $\Theta(\log n)$.

Figura 2.11. operação do algoritmo HNSW (Hierarchical Navigable Small World).



2.10.2 In-memory exactNN index (IMENN)

O *In-memory exactNN index* (IMENN) é um algoritmo heurístico para busca exata, que entra em um ciclo iterativo passando pelos passos de indexação e atualização até atingir uma condição de parada [Aoyama et al., 2020]. A base do algoritmo IMENN é o k-Nearest Neighbors (k-NN), que foi modificado para realizar uma busca exata utilizando clustres. Durante a indexação, o conjunto de vetores de *embeddings* $X = \{x_1, x_2, \ldots, x_n\}$ é dividido em C_i *clusters* (ou grupos) com vetores similares. O agrupamento é feito usando a técnica k-means. Em seguida, é criado um índice invertido para cada C_i de forma a criar uma estrutura de dados totalmente mapeada. Para realizar uma consulta, um vetor de *embedding* (X_j) é usado como *query*, e então é calculada a similaridade parcial entre este vetor e todos os centroides dos C_i *clusters*, assim o centróide com maior similaridade é escolhido. Por último, é realizada uma busca por similaridade dentro do *cluster* escolhido e são retornados os vetores mais similares com a consulta.

2.11 VectorDB

O VectorDB é um banco de dados vetorial desenvolvido pela *Jina AI*, projetado para realizar operações *CRUD* (*Create, Read, Update, Delete*) sobre objetos que contêm *embeddings*. Sua arquitetura é construída sobre o *DocArray*, que permite representar dados como *arrays* de alta eficiência e oferecendo integração com tipagem *Pydantic*, o que garante consistência nos esquemas e facilita a construção rápida de estruturas embutidas Jina AI [2023a,b].

No que se refere à indexação e à busca, o VectorDB disponibiliza dois mecanismos principais: o *InMemoryExactNNVectorDB*, que realiza busca exata e é adequado para conjuntos reduzidos devido ao seu alto custo computacional; e o *HNSWVectorDB*, baseado no algoritmo *Hierarchical Navigable Small World* Malkov & Yashunin [2018], que utiliza uma estrutura hierárquica de grafos para acelerar consultas em grandes volumes de dados, equilibrando eficiência e precisão.

Em termos de desempenho, o índice *InMemoryExactNN* proporciona exatidão máxima, porém seu custo computacional cresce de forma linear com o tamanho do conjunto de dados, o que o torna pouco adequado para cenários de grande escala. Em contraste, a abordagem *HNSW* constrói uma estrutura hierárquica que permite localizar os vizinhos mais próximos em tempos sublineares, alcançando eficiências notáveis mesmo em volumes massivos de dados [Malkov & Yashunin, 2018].

A escolha pelo VectorDB nesta pesquisa foi motivada por sua natureza *Python-native* e sua arquitetura simplificada, construída sobre o DocArray. Essa combinação facilitou a prototipagem rápida e a experimentação, permitindo alternar de forma transparente entre a busca exata (*InMemoryExactNN*) e a aproximada (*HNSW*) para fins de comparação. Além disso, sua integração com Pydantic garante a consistência dos dados, um fator importante ao lidar com os metadados associados a cada *embedding*. Enquanto outras soluções como Milvus ou Qdrant oferecem maior escalabilidade para ambientes de produção, a simplicidade e a flexibilidade do VectorDB se mostraram mais

adequadas para o escopo de uma investigação acadêmica focada na análise comparativa de algoritmos.

2.12 Considerações finais

Este Capítulo apresentou um resumo dos conceitos teóricos necessários para compreender o desenvolvimento do nosso trabalho. Essas ideias são a base das diferentes opções tomadas para estabelecer os métodos. O Capítulo 3 aborda o material teórico estudado, que possibilita compreender os trabalhos relacionados à nossa pesquisa. No próximo Capítulo são apresentados os trabalhos relacionados.

3

Trabalhos relacionados

Este capítulo oferece um resumo das soluções mais recentes para os problemas de monitoramento ambiental e recuperação de informações sobre paisagens sonoras. Vários tipos de animais, especialmente aves, são identificados nos estudos revisados. O uso do monitoramento de paisagens acústicas enfrenta alguns desafios, como perda de dados devido à aquisição e transmissão de dados e redução na taxa de classificação precisa devido a condições de sons de fontes externas. A recuperação acústica envolve uma variedade de estratégias que possibilitam a combinação de diferentes técnicas, como segmentação, filtragem e metodologias diversas, a fim de realizar eficientemente o processo de recuperação acústica. Portanto, o objetivo deste capítulo é apresentar descrições organizadas por categorias, facilitando assim a combinação e seleção de componentes para o design e planejamento de uma abordagem com propósitos específicos e/ou gerais.

3.1 Monitoramento bioacústico

O monitoramento acústico passivo (PAM) tem sido amplamente utilizado para analisar vocalizações de aves, mamíferos e outros táxons, permitindo a detecção e classificação de espécies. Recentemente, diversas pesquisas têm incorporado *Deep Learning* como

ferramenta para automatizar tarefas complexas, como segmentação, classificação e extração de padrões acústicos. Um exemplo relevante é o estudo de Bjorck et al. [2019] que investigaram a detecção acústica passiva de elefantes africanos usando DenseNet combinada com LSTM, atingindo 91% de *precision-recall*. O uso de redes recorrentes aliado a CNNs demonstrou ganhos expressivos na segmentação de vocalizações, sendo um exemplo de pipeline eficiente para espécies com sinais acústicos complexos.

A pesquisa realizada por Fanioudakis & Potamitis [2017] concentra-se na implementação de Redes Neurais Profundas que rotulam e localizam as vocalizações de aves em espectrogramas de áudio registrados no campo. Os autores aplicaram DenseNet para gerar mapas de atenção combinados ao YOLOv2 para identificar os cantos das aves, resultando em uma acurácia de 88,94% e AUC de 94,76%. A detecção acústica dos cantos das aves pode ser utilizada para o monitoramento automatizado de diversas espécies de aves e a análise de registros a longo prazo de espécies de interesse. Outro exemplo nesta linha, é o trabalho de Wolfe et al. [2023], que propuseram um classificador multilabel eficiente para cinco espécies de aves das *Great Plains*, utilizando espectrogramas Mel com normalização adaptativa (PCEN). Baseado em VGG16 pré-treinada e *transfer learning*, o sistema atingiu F1 médio de 99,6%, evidenciando o impacto de arquiteturas robustas e pré-processamentos especializados.

De forma complementar, Hagiwara et al. [2023] apresentam um referencial de 12 conjuntos de dados bioacústicos públicos e padronizados. O objetivo principal é avaliar algoritmos de aprendizado de máquina, incluindo métodos clássicos como regressão logística, SVM, árvores de decisão e XGBoost, além de arquiteturas *Deep Learning* como ResNet e VGGish. O estudo fornece métricas comparativas para classificação e detecção, servindo como base para protocolos de avaliação modernos.

Na Tabela 3.1, são apresentados os bancos de dados ecoacústicos e bioacústicos mais utilizados na literatura. A tabela apresenta os conjuntos principais de dados acústicos e bioacústicos disponíveis para a comunidade científica, destacando suas características principais, como a quantidade de amostras, a presença de metadados (nome

científico, espécie, data da gravação) e o tipo de dados que contêm.

Tabela 3.1. Principais bancos de dados ecoacústicos e bioacústicos.

Nome	Tipo	Amostras	Metadados	Referência
Xeno-Canto	Pássaros	798.336	Sim	[Gil & Donsker, 2005]
BirdCLEF	Pássaros	80.000	Sim	[Cornell Lab of Ornithology, 2014]
VGGSOUND	Vários	200.000	Sim	[Chen et al., 2020]
Macaulay	Pássaros	1.871.410	Sim	[Cornell Lab of Ornithology, 2014]
eBird	Pássaros	—	Sim	[Munson et al., 2012]
warblrb10k	Vários	8.000	Sim	[Stowell & L, 2016]
BirdCLEF+ 2025	Vários	28.564	Sim	[Klinck et al., 2025]
HumbugDB	Mosquitos	13.011	Não	[Vasconcelos et al., 2020]

3.2 Modelos pré-treinados

O uso de modelos pré-treinados tem ganhado destaque em bioacústica, principalmente para extração de *embeddings* representativos aplicados à classificação e recuperação de vocalizações. A escolha do modelo influencia diretamente a qualidade dos vetores, impactando os índices de similaridade e o desempenho em bases vetoriais. Gómez-Gómez et al. [2022] conduziram uma análise comparativa entre VGG16, ResNet50 e MobileNetV2 para a classificação de aves, utilizando 201,6 minutos de gravações com remoção de ruído de fundo. O ResNet50 obteve os melhores resultados, com *recall* macro de 85%, precisão de 85,5% e F1-score de 84%, demonstrando a viabilidade do uso de modelos pré-treinados para identificação de vocalizações.

O estudo de Ghani et al. [2023] exploram o uso de *embeddings* de características de classificadores acústicos de aves em larga escala por meio da aprendizagem por transferência por *few shot* na análise bioacústica. Os autores descobriram que os *embeddings* extraídos de modelos de vocalizações de aves superam consistentemente os *embeddings* treinados em conjuntos de dados de áudio gerais, indicando seu potencial para a classificação eficiente de novas tarefas bioacústicas com dados de treinamento limitados. Foram comparados Perch, BirdNET 2.3, AudioMAE, YAMNet e VGGish em diferentes tarefas bioacústicas. Os resultados mostram que o Perch, treinado no

Xeno-Canto com EfficientNet-B1, apresentou o melhor desempenho, atingindo AUC de 97% e *Top-1* de 86%. Este achado foi fundamental para a escolha do Perch como o extrator de características principal adotado nesta pesquisa.

O Perch utiliza janelas de 5,0 segundos e gera *embeddings* de 1280 dimensões. O BirdNET também emprega EfficientNet, com janelas de 3 segundos e *embeddings* de 1024 dimensões. O YAMNet, baseado em MobileNetV1, opera com tramas de 0,96 segundos e gera *embeddings* de 1024 dimensões a partir do AudioSet. O VGGish, por sua vez, utiliza arquitetura VGG com *embeddings* de 128 dimensões e treinamento no YouTube8M, sendo mais limitado. Já o AudioMAE aplica reconstrução espectral com janelas de 10 segundos e *embeddings* de 1024 dimensões. Essa diversidade de arquiteturas permite analisar como o tamanho da janela e a profundidade do modelo influenciam a representação vetorial para recuperação acústicab[Ghani et al., 2023].

Também é importante destacar o trabalho realizado pela equipe de pesquisa de Maclean & Triguero [2023], no qual exploraram CNNs pré-treinadas como ResNet26, ResNet50 e EfficientNet, utilizando os modelos como extratores de *embeddings* para identificar espécies de aves a partir de gravações de áudio. Embora a tarefa seja de classificação, os resultados reforçam que a escolha do modelo impacta o desempenho global, com a ResNet50 alcançando F1 micro[1] médio de 74%. Por outro lado, Tosato et al. [2023] investigaram o uso do modelos automatizados (AutoKeras) para automação da escolha de arquiteturas e hiperparâmetros. Os autores compararam o desempenho de Autokeras com modelos tradicionais, como MobileNet, ResNet50 e VGG16 alcançando bons resultados em um conjunto regional. Apesar disso, o impacto metodológico para recuperação é limitado, servindo mais como evidência de tendências no uso de *AutoML*.

Sheikh et al. [2024] apresentam o *Bird Whisperer*, um sistema que adapta o modelo *Whisper* da OpenAI, originalmente treinado para *speech recognition*, para a clas-

[1]A métrica F1 micro calcula as métricas globalmente, somando todos os verdadeiros positivos, falsos negativos e falsos positivos de todas as classes. É particularmente útil em cenários com desbalanceamento de classes.

sificação de vocalizações de aves utilizando o dataset BirdCLEF 2023, composto por 26.264 gravações de 264 espécies. O estudo demonstra que o uso do *Whisper encoder* apenas como extrator de *features* gera representações pouco discriminativas, pois muitos sinais são interpretados como ruído de fundo. Para contornar essa limitação, os autores aplicam *fine-tuning* e *data augmentation* para adaptar o modelo ao domínio bioacústico. A arquitetura integra o encoder do *Whisper*, baseado em transformers com múltiplas camadas convolucionais e *self-attention*, seguido por uma CNN de duas camadas e uma rede totalmente conectada para mapeamento das classes.

Na Tabela 3.2 estão listados os principais modelos pré-treinados em ecoacústica e bioacústica para extração de *embeddings* e suas principais características.

Tabela 3.2. Modelos pré-treinados para extração de *embeddings* acústicos e suas principais características.

Modelo	**Dataset**	**Código**	**Janela (s)**	**Entrada**	**Tamanho do embedding**
Perch	Xeno-Canto	Sim	5,0	Sons	1280
BirdNET	XC+ML+Custom	Sim	3,0	Sons	1024
AudioMAE	MAE Large	Sim	10,0	Espectrogramas	1024
YAMNet	AudioSet	Sim	0,96	Espectrogramas	1024
VGGish	YouTube8M	Sim	0,96	Espectrogramas	128

3.3 Recuperação acústica em áudios

A recuperação acústica de áudios requer *features* robustas e representação adequada dos sinais para viabilizar consultas por similaridade. Recentemente, varios estudos investigaram técnicas para melhorar o desempenho em cenários desafiadores, incluindo baixa relação sinal-ruído (SNR), número limitado de anotações e variações de domínio. Clink et al. [2023] desenvolveram um *workflow* aberto para detecção e classificação de chamadas de gibão-cinzento (*Hylobates funereus*) utilizando MFCCs e classificadores supervisionados *SVM* e *Random Forest*, alcançando F1 de 0,78. Embora voltado à classificação, o estudo evidencia a importância da seleção de *features* robustas para recuperação acústica baseada em representações consistentes.

De mesma forma, Li et al. [2023] propõem um método inovador para extração de contornos de assobios de cetáceos a partir de espectrogramas tempo-frequência, utilizando uma arquitetura baseada em *Generative Adversarial Networks* (GANs) com aprendizado *stage-wise*. O objetivo principal é superar a limitação de bases de dados com anotações reduzidas, que comprometem o desempenho de modelos de *deep learning*. Para isso, a abordagem aprende a partir dos poucos dados anotados disponíveis para sintetizar novos exemplos artificiais e controlados, atingindo um ganho médio de 1,69 no *F1-score* quando comparada a GANs tradicionais. Apesar de focar em espécies marinhas, o estudo reforça o potencial do uso de *data augmentation* para melhorar tarefas de recuperação sob dados limitados.

O estudo de Lakdari et al. [2024] analisa a discriminação individual de fêmeas de gibão-cinzento em cenários com aumento progressivo de ruído, comparando MFCCs, *embeddings* pré-treinados (BirdNET, VGGish, Wav2Vec2) e índices acústicos. Os experimentos, baseados em *playbacks* até 400 m de distância, mostram que o BirdNET atinge 98,2% de acurácia nas gravações sem distância (0 m), mas que MFCCs apresentam desempenho superior em distâncias maiores e cenários de baixo SNR. Resultados de agrupamento não supervisionado (HDBSCAN, *Affinity Propagation*) indicam que MFCCs recuperam corretamente o número de *clusters* até 150m (NMI $>$ 0,8). Esse trabalho evidencia limitações de *embeddings* genéricos e sustenta a necessidade de modelos específicos de domínio para a recuperação de vocalizações.

Por fim, Balaji & Livinza [2025] apresentam uma revisão sistemática sobre técnicas de redução de ruído em sinais bioacústicos, com análise de domínios aéreo e subaquático. O artigo propõe uma taxonomia de abordagens de tempo, frequência, tempo-frequência, espacial e métodos *ML*, *DL* e processamento clássico, destacando que ambientes subaquáticos demandam modelos mais profundos e janelas maiores, enquanto contextos terrestres preservam melhor a estrutura temporal. A revisão aponta lacunas de padronização e integração multimodal, reforçando a necessidade de desenvolver benchmarks consistentes para recuperação acústica em cenários de alto ruído.

3.4 Fusão de features na recuperação acústica em áudios

A fusão de *features* tem emergido como uma estratégia promissora para melhorar o desempenho de sistemas de recuperação acústica, permitindo combinar diferentes representações de áudio para enriquecer a discriminação entre espécies e eventos. Essa abordagem é especialmente relevante em contextos de baixa relação sinal-ruído e diversidade acústica elevada. Tolkova [2021] apresenta uma revisão abrangente sobre representações acústicas para bioacústica de conservação, analisando técnicas clássicas e modernas de extração de *features* e redução de dimensionalidade. O estudo discute desde métodos tradicionais, como MFCCs e métricas espectrais, até abordagens baseadas em aprendizado profundo, incluindo *autoencoders* variacionais, CNNs e mapeamentos não lineares como *t-SNE* e *UMAP*. A autora destaca o uso de espaços latentes para representar relações acústicas complexas, oferecendo um arcabouço conceitual essencial para integrar múltiplas representações no mesmo pipeline de recuperação.

Colonna et al. [2014] investigaram o uso de Redes de Sensores Sem Fio (WSNs) para a classificação de espécies de anuros, tratando o problema como um *ensemble* de classificadores. No sistema proposto, cada sensor atua como um classificador individual, utilizando modelos de aprendizado de máquina de baixo custo computacional som Análise Discriminante Quadrática, *Naive Bayes* e Árvores de Decisão. As decisões de cada sensor são então agregadas por meio de quatro técnicas de fusão em nível de decisão: voto majoritário, voto majoritário ponderado e as regras de combinação aritmética e geométrica. Os resultados demonstraram que a combinação de classificadores por meio da regra aritmética obteve um ganho de precisão de 11% sobre um sensor isolado, e que esse ganho aumentou para aproximadamente 20% ao aplicar a regra de rejeição para filtrar os casos de maior incerteza. O trabalho evidencia o potencial da fusão de informações, mesmo com modelos clássicos, para aumentar a robustez em

tarefas de monitoramento bioacústico.

Xie & Zhu [2023a] propõem um sistema de classificação acústica de aves baseado em um conjunto de *features profundas* extraídas de modelos de *transfer learning*. Os autores implementam uma estratégia baseada em *deep cascade features*, combinando representações intermediárias extraídas de modelos pré-treinados como *VGG16*, *ResNet50*, *EfficientNetB0*, *MobileNetV2* e *Xception* por *early fusion*, atingindo acurácia de 94,89% no dataset CLO-43DS. Este estudo evidencia o potencial da fusão precoce na melhoria da robustez em classificação bioacústica. Além disso, Lü et al. [2024] investigam a fusão dupla de *features* acústicas para reconhecimento de mamíferos marinhos, combinando MFCCs com representações *Delay-Doppler* em uma CNN profunda. O modelo atingiu 98,04% de acurácia no conjunto Watkins, superando significativamente abordagens baseadas em representações únicas. Esses resultados reforçam que a integração de *features* complementares contribui para maior generalização do sistema.

Gavali & Banu [2025] apresentam uma abordagem multimodal de fusão visual-acústica, comparando *early* e *late fusion* no dataset iBC53. A fusão precoce atingiu 95,2% de acurácia, superando tanto os modelos unimodais quanto a fusão tardia, mostrando ganhos quando diferentes modalidades são integradas. Por fim, Lu et al. [2025] introduzem o DuSAFNet, um modelo baseado em *multi-path feature fusion* e atenção espectro-temporal, alcançando 96,88% de acurácia e 96,83% de *F1-score*. O estudo demonstra que arquiteturas híbridas, que integram múltiplos níveis de abstração, oferecem resultados superiores em domínios acústicos complexos. No contexto deste trabalho, as evidências reunidas sustentam a adoção da fusão de vetores de *features* como estratégia para enriquecer a representação acústica no banco vetorial, explorando complementaridades entre diferentes domínios de informação e potencializando o desempenho do sistema proposto.

3.5 Bancos de dados vetoriais

A utilização de bancos de dados vetoriais tornou-se essencial para tarefas de recuperação baseada em similaridade, especialmente em contextos de alto volume e alta dimensionalidade, como bioacústica. Esses sistemas permitem indexar e consultar *embeddings* de maneira eficiente, possibilitando buscas aproximadas com latência reduzida e escalabilidade para milhões de vetores. Huang et al. [2022] apresentam uma revisão ampla sobre tecnologias de bases vetoriais, com foco na integração com modelos de linguagem (LLMs) e nas estratégias de armazenamento e busca. Os autores destacam algoritmos clássicos, como M-tree, Best Bin First e K-means Tree, além de discutirem desafios associados à alta dimensionalidade e à integração com frameworks de aprendizado profundo.

Por outro lado, Taipalus [2024] fornecem uma visão consolidada sobre conceitos, arquiteturas e critérios de seleção de bancos vetoriais. Os autores comparam sistemas como Pinecone, Milvus, Qdrant, Chroma e VectorDB, destacando métricas essenciais: eficiência de armazenamento, velocidade de recuperação, escalabilidade e precisão. O estudo orienta a escolha da solução mais adequada de acordo com o balanço entre custo computacional e desempenho de consulta, o que é particularmente relevante para bioacústica, onde os vetores são de alta dimensionalidade. Complementando, Jie et al. [2023] detalham técnicas modernas de indexação, execução de consultas híbridas e processamento otimizado, além de introduzir novos sistemas, como EuclidesDB, Vearch, NucliaDB e Margo. A análise reforça a importância da escolha criteriosa de índices, algoritmos e métricas, considerando trade-offs entre precisão e latência. Essa revisão sustenta a seleção do VectorDB no presente trabalho, aliado ao índice *HNSW* e à métrica de distância euclidiana, buscando eficiência na recuperação acústica.

As Tabelas 3.3 e 3.4, embasadas em pesquisas recentes, reúnem informações relevantes sobre bancos de dados vetoriais, com ênfase especial em critérios voltados para aplicações bioacústicas. A construção das tabelas fundamenta-se principalmente no

estudo de Jie et al. [2023], que fornecem elementos essenciais, como tipo de licença, natureza do banco de dados, suporte a metadados e principais casos de uso. A seção referente a variantes de consulta foi omitida para manter o foco nos aspectos mais relevantes, e o VectorDB foi incluído por ser a solução adotada neste trabalho. Além disso, parte das informações apresentadas deriva da pesquisa de Ghani et al. [2023], detalhando a categoria do tipo de código, os resultados obtidos e o formato de entrada correspondente para cada modelo. Optou-se conscientemente por não incluir o tempo de processamento de cada abordagem, concentrando a análise nos atributos mais relevantes para esta revisão, incluindo os algoritmos de busca e as métricas de similaridade empregadas pelos bancos de dados vetoriais no contexto da recuperação acústica.

Tabela 3.3. Principais bancos de dados vetoriais para bioacústica.

Nome	Licença	Tipo	Metadados	Consulta	Indexação	Casos de uso
Chroma	Apache	Lc	Sim	Aprox	HNSW	Chatbots
Milvus	OSS	Lc/Cl	Sim	Aprox	IVF, HNSW	Multimodal
Pinecone	Próprio	Cl	Sim	Aprox	IVF, PQ	Busca semântica
Qdrant	OSS	Lc/Cl	Sim	Aprox	HNSW	Texto/Imagens
Weaviate	OSS	Cl	Sim	Aprox	HNSW, PQ	RAG, NLP
VectorDB	Apache	Lc/Cl	Sim	Aprox	HNSW	Bioacústica
Vald	OSS	Lc/Cl	Sim	Aprox	HNSW, Graph	Sons/Imagens

Legenda: OSS = Open Source Software; Lc = Local; Cl = Cloud; Aprox = Aproximada.

Tabela 3.4. Bancos de dados vetoriais com seus algoritmos de busca.

Nome	Algoritmo de busca	Similaridade
Chroma	HNSW	DE, PI, SC
EuclidesDB	ANN-IVF	DE, SC
	IVF-Flat, IVF-PQ, IVF-SQ, IVF-HNSW, IMI	
Manu	HNSW, NSG, NGT	DE
	B-Tree, Sorted List	
Margo	HNSW, ANN	SC
Milvus	FLAT, IVF-FLAT, IVF-PQ, IVF-SQ8, HNSW, SCANN	DE, PI, SC
Pinecone	ANN-IVF	DE, PI, SC
Qdrant	HNSW	DE
VectorDB	HNSW	DE, SC
Vearch	Faiss	DE

Legenda: DE = Distância Euclidiana; PI = Produto Interno; SC = Similaridade do Cosseno.

3.6 Sínteses dos trabalhos relacionados

A Tabela 3.5 resume os aspectos fundamentais dos trabalhos relacionados abordados neste capítulo. A análise comparativa dos estudos evidencia uma forte tendência na utilização de arquiteturas de Aprendizagem Profunda, com destaque para as Redes Neurais Convolucionais (CNNs), que frequentemente processam espectrogramas como representação de entrada para a extração de características acústicas. Outro ponto relevante revelado pela síntese é a predominância de abordagens supervisionadas. Essa constatação reforça a importância de investigar não apenas métodos não supervisionados, mas também de focar na tarefa de recuperação (*ranking*) em vez de apenas classificação — lacunas que o método proposto nesta dissertação busca preencher.

Tabela 3.5. Síntese dos trabalhos relacionados, destacando a tarefa principal de cada um.

Autor	**Entrada**	**Abordagem**	**Tarefa Principal**	**Supervisado**
Gómez-Gómez et al. [2022]	Espectrogramas	ResNet50	Classificação	Sim
Maclean & Triguero [2023]	Sons	CNN	Classificação	Sim
Hagiwara et al. [2023]	Espectrogramas	CNN + BEANS	Classificação	Não
Ghani et al. [2023]	Espectrogramas/Sons	Modelos pré-treinados	Transfer Learning	Não
Tosato et al. [2023]	Espectrogramas	AutoKeras + CNN	Classificação	Sim
Wolfe et al. [2023]	Espectrogramas	VGG16 + Transfer Learning	Classificação	Sim
Sheikh et al. [2024]	Espectrogramas	Whisper + Fine-tuning	Classificação	Sim
Clink et al. [2023]	Sons	MFCC + ML	Classificação	Sim
Lakdari et al. [2024]	Sons	MFCC + Embeddings	Classificação	Sim
Balaji & Livinza [2025]	Sons	ML/DL	Denoising	Não
Tolkova [2021]	Espectrogramas/Sons	CNN, Autoencoders, etc.	Representação de Features	Não
Xie & Zhu [2023a]	Espectrogramas	VGG16 + Early Fusion	Fusão de Features	Sim
Lü et al. [2024]	Sons	CNN + MFCC + Delay-Doppler	Classificação	Sim
Gavali & Banu [2025]	Sons/Imagens	Inception-ResNet + Fusion	Classificação	Sim
Lu et al. [2025]	Espectrogramas	DuSAFNet + Multi-path Fusion	Classificação	Sim
Esta pesquisa	**Sons**	**Perch + Fusão de Features + VectorDB**	**Recuperação**	**Não**

3.7 Considerações finais

As pesquisas recentes sobre classificação e recuperação acústica revelam avanços significativos no uso de aprendizado profundo e modelos pré-treinados para lidar com dados bioacústicos em larga escala. Pesquisas como as de Clink et al. [2023], Lakdari et al. [2024], Li et al. [2023] demonstram que técnicas baseadas em *DL*, aliadas a *features* clássicas como MFCCs, oferecem ganhos expressivos na extração de informações relevantes mesmo em ambientes ruidosos. Da mesma forma, estudos como [Lu et al., 2025, Lü et al., 2024, Xie & Zhu, 2023a] evidenciam que a fusão de múltiplas representações espectrais e temporais potencializa os sistemas de classificação e recuperação.

Por outro lado, a literatura também revela desafios importantes a serem superados. Conforme destacado por Balaji & Livinza [2025], Gavali & Banu [2025], Tolkova [2021], a ausência de benchmarks padronizados, a heterogeneidade dos protocolos de coleta de dados e a falta de integração entre modalidades limitam a comparação direta entre métodos e a aplicabilidade em cenários complexos. Além disso, estudos como Ghani et al. [2023] mostram que, embora os modelos pré-treinados ofereçam *embeddings* representativos e resultados promissores, ainda existem limitações relacionadas à escalabilidade, ao desbalanceamento de dados e à adaptação para espécies não vistas. Nesse contexto, bancos de dados vetoriais, como o VectorDB [Huang et al., 2022, Taipalus, 2024], emergem como soluções eficientes para armazenar e recuperar *embeddings*, permitindo consultas de similaridade rápidas e escaláveis.

Os trabalhos revisados sustentam a viabilidade da fusão de vetores de *features*, combinada com bancos de dados vetoriais para recuperação eficiente, fornece uma abordagem promissora para a recuperação de paisagens sonoras acusticamente semelhantes, permitindo superar limitações de métodos tradicionais, oferecendo uma solução mais escalável e precisa para aplicações de monitoramento automatizado da biodiversidade. Os detalhes sobre a arquitetura proposta, seleção de *features* e implementação dos experimentos são apresentados no Capítulo 4.

4

Método para a recuperação acústica de paisagens sonoras

Neste capítulo, é apresentada uma descrição detalhada da metodologia proposta para a recuperação acústica de paisagens acústicas através da fusão de *features* e de um banco de dados vetorial. O percurso abrange desde a obtenção dos áudios até o procedimento de recuperação acústica por meio da busca por similaridade. A exposição geral do método em blocos, Figura 4.1, proporciona uma visão completa do funcionamento das técnicas, com explicações detalhadas das tarefas realizadas e a ordem dos passos que levam ao resultado final. A metodologia proposta utiliza vetores de *embeddings*, que representam os áudios em um espaço vetorial contínuo. Esse enfoque oferece a vantagem de reduzir a dimensionalidade dos áudios, resultando em um armazenamento mais eficiente no banco de dados vetorial. Dessa forma, facilita-se a recuperação acústica através de consultas por similaridade. É importante destacar que foi utilizado o modelos DL pré-treinado Perch para a extração dos *embeddings*. A decisão de utilizar Perch é fundamentada na revisão da literatura e nos trabalhos relacionados que mostram resultados positivos ao utilizar esse modelo. Na primeira fase, procederemos à coleta e armazenamento das gravações das paisagens sonoras, que serão submetidas a um posterior de fusão de *features* e por fim o pro-

cesso de extração de vetores de *embeddings*. Em seguida, será realizado o mapeamento e armazenamento desses vetores, culminando na fase de recuperação acústica e análise de sons. Paralelamente, será estabelecido o critério de avaliação do método a ser desenvolvido.

4.1 Base de dados

Este estudo utilizou seis conjuntos de dados bioacústicos coletados de diferentes localizações geográficas. O primeiro conjunto de dados é o `Cornell Birdcall Identification` (CBI), utilizado no desafio de identificação de aves em áudios longos, organizado pelo Centro de Bioacústica para Conservação do Laboratório de Ornitologia de Cornell. Ele contém 21.424 gravações de áudio de 264 espécies de aves em formato `.wav`, com uma taxa de amostragem de 32kHz [Klinck et al., 2020]. O segundo conjunto de dados é o `BirdCLEF 2022` (BC_{22}), que inclui 16.935 arquivos de áudio em formato `.ogg` de 179 espécies de aves, amostrados a 32kHz [Addison et al., 2022]. O terceiro conjunto de dados é o `BirdCLEF 2023` (BC_{23}), composto por 15.004 arquivos `.ogg` de 268 espécies de aves, também amostrados a 32kHz Klinck et al. [2023]. O quarto conjunto de dados, `Xeno-Canto Bird Recordings` (XBC), consiste em 14700 gravações de 264 espécies de aves, em formato `.ogg` e com taxa de amostragem de 32kHz [Nomorevotch, 2020].

O quinto conjunto de dados é o `BirdCLEF+ 2025`, uma coleção bioacústica multitaxonômica projetada para o monitoramento da biodiversidade. O conjunto é composto por um total de 28.564 gravações de áudio abrangendo 206 espécies, incluindo aves, anfíbios, mamíferos e insetos, armazenadas no formato `.ogg` com taxas de amostragem variáveis. As gravações foram coletadas a partir de três fontes principais: Xeno-Canto (XC) Vellinga, W.P. and Planqué, R. [2025], iNaturalist (iNat) iNaturalist community [2025] e o Arquivo Sonoro Colombiano (CSA) [Murillo Bedoya, D. and Buitrago-Cardona, A. and Acevedo-Charry, O. and Ochoa-Quintero, J. M., 2021].

O último conjunto de dados foi obtido a partir de gravações de campo realizadas no campus da Universidade Federal do Amazonas. Ele é composto por 60 gravações de áudio de 12 espécies de anuros, armazenadas em formato `.flac` com uma taxa de amostragem de 44.1kHz.

Essas fontes contribuem com um conjunto diversificado de eventos acústicos. Cada gravação é acompanhada por campos extensos de metadados, como *primary_label*, *type*, *collection*, *scientific_name*, *common_name*, *location* (latitude e longitude) e *author*, permitindo análises detalhadas e contextualizadas. A distribuição das gravações entre a base de dados vetorial e as consultas varia de acordo com o experimento e está explicitamente descrita na Seção 5.1.

4.2 Descrição do método

O método proposto segue um fluxo de trabalho sistemático, partindo do áudio bruto até a recuperação de gravações acusticamente semelhantes. A seguir, apresentamos a formalização do processo.

Aqui, o áudio bruto é modelado como um sinal ou série temporal unidimensional $s(t) \in \mathbb{R}^T$, onde T representa o número total de amostras e $t \in \{1, 2, \ldots, T\}$. O sinal é amostrado a uma taxa constante f_s (em Hz) e segmentado em janelas não sobrepostas com duração fixa $\Delta = 5$ segundos. Cada janela contém $N = f_s \cdot \Delta$ amostras, e o número total de janelas completas que podem ser extraídas é $M = \lfloor T/N \rfloor$.

Seja $s_i \in \mathbb{R}^N$ o i-ésimo segmento do sinal, com $i = 1, \ldots, M$. Cada segmento é processado por uma função de extração de características $f : \mathbb{R}^N \to \mathbb{R}^{1280}$, que produz um vetor de características de tamanho fixo $x_i = f(s_i)$, com $x_i \in \mathbb{R}^{1280}$. Neste trabalho, a função $f(\cdot)$ é parametrizada pelo modelo Perch Ghani et al. [2023], um extrator de *embeddings* bioacústicos baseado na arquitetura EfficientNet-B1. Dessa forma, cada gravação de áudio é representada por uma matriz de características $\mathbf{X} \in \mathbb{R}^{1280 \times M}$, construída empilhando os vetores x_i em colunas: $\mathbf{X} = [x_1\ x_2\ \ldots\ x_M]$.

Como a duração dos áudios de entrada é variável, o número de colunas M na matriz $\mathbf{X}$ também varia entre gravações. Isso impõe o desafio de obter uma representação de tamanho fixo adequada para indexação e comparação. Para resolver esse problema, aplicamos uma técnica de fusão sobre a dimensão temporal (colunas) de $\mathbf{X}$, resultando em um vetor de *embedding* global $\bar{x} \in \mathbb{R}^{1280}$. Embora estudos anteriores tenham explorado certas formas de fusão de *features* Xie & Zhu [2023b], nossa contribuição consiste em avaliar e comparar sistematicamente quatro estratégias distintas: *average pooling, weighted average pooling, sum pooling* e *max pooling*, detalhadas na Seção 4.4.

Nossa hipótese principal é que a fusão *weighted average pooling* pode melhorar o desempenho do sistema em paisagens sonoras ruidosas ou altamente variáveis. Ao atribuir maior peso aos segmentos acusticamente mais informativos, esperamos que essa técnica supere as demais abordagens de fusão que tratam todos os segmentos com igual importância. Podemos resumir esse processo em seis etapas essenciais, conforme ilustrado na Figura 4.1. Essas etapas são:

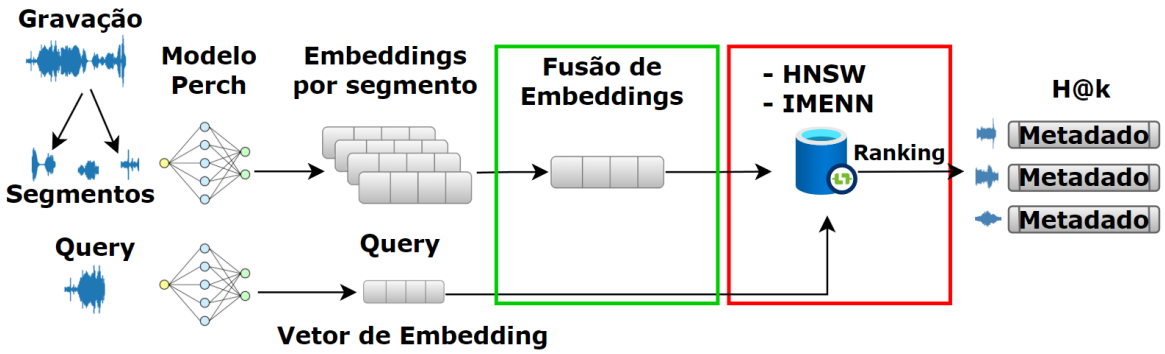


Figura 4.1. Método proposto para recuperação acústica de paisagens sonoras.

4.2.1 Pré-processamento

A etapa de pré-processamento de dados tem como objetivo preparar, limpar e transformar os dados brutos para melhorar a qualidade e adaptá-los aos modelos. O primeiro passo consiste em normalizar a frequência de amostragem de todos os áudios para 32kHz por meio de *downsampling*. Essa taxa de amostragem não é uma escolha arbi-

trária, mas sim um requisito do modelo Perch utilizado para a extração de *embeddings* [Ghani et al., 2023]. O modelo foi pré-treinado com áudios nessa frequência e, portanto, todos os dados de entrada devem ser padronizados para garantir a consistência das características extraídas. Adicionalmente, conforme o Teorema de Nyquist (discutido na Seção 2.2.1), uma taxa de 32kHz é suficiente para capturar adequadamente o espectro vocal das especies utilizadas nos conjuntos de dados desta pesquisa.

Após a normalização da amostragem, é aplicado um filtro para reduzir o ruído de fundo e aumentar a clareza dos sinais bioacústicos. Para isso, foi utilizada a biblioteca *noise reduce*, desenvolvida por Sainburg [2019]. O algoritmo opera no domínio da frequência, aplicando a transformada de Fourier para analisar o espectro do sinal e atenuar as porções identificadas como ruído, conforme detalhado na Seção 2.5.

O efeito do filtro é ilustrado na Figura 4.2, que exibe a forma de onda de um sinal antes e depois do processamento. Conforme a configuração padrão da biblioteca *noise-reduce* desenvolvida por [Sainburg, 2019], foi selecionado um clipe dos cinco primeiros segundos da gravação original (em azul) para servir como o perfil de ruído, conforme destacado pelo segmento em vermelho. Essa seleção parte da premissa de que o início de muitas gravações de campo contém principalmente o ruído ambiente. O algoritmo utiliza esta amostra de ruído para calcular a média e o desvio padrão do espectrograma do ruído para, então, estabelecer o limiar de remoção. A onda resultante (em laranja) demonstra a eficácia do filtro na eliminação de sons indesejados, o que melhora significativamente a relação sinal-ruído do áudio original.

Para ilustrar visualmente o efeito do filtro de ruído no sinal de áudio, tomamos uma gravação de nosso banco de dados e geramos um espectrograma sem aplicar o filtro de ruído, como mostrado na parte superior da Figura 4.3. Em seguida, aplicamos o filtro de ruído e obtivemos o espectrograma correspondente, representado na parte inferior da figura. Nesta última seção da figura, é possível observar claramente como os ruídos de fundo (em amarelo) são eliminados, destacando a intensidade do sinal de áudio sem os ruídos de fundo em amarelo.

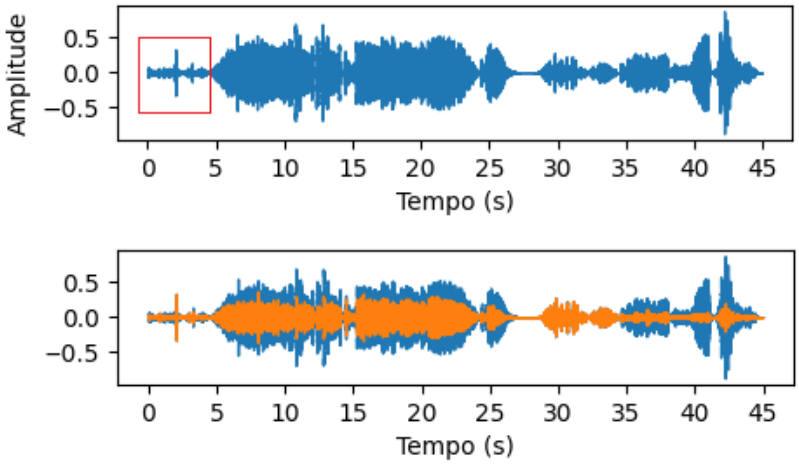


Figura 4.2. A onda de um sinal de áudio antes e depois de aplicar o filtro de redução de ruído noise reduce. O quadrado vermelho na onda azul representa o segmento de ruído utilizado para calcular o desvio padrão e o limiar para aplicar o filtro de ruído no sinal de áudio.

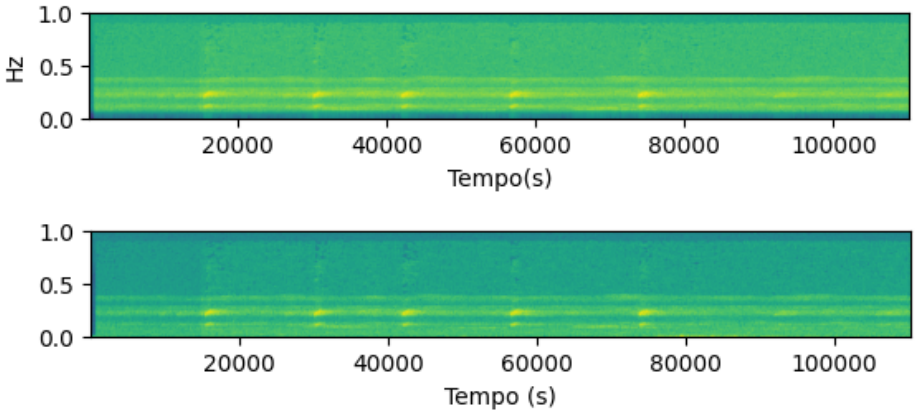


Figura 4.3. Espectrograma antes e depois da aplicação do filtro de redução de ruído a um áudio da espécie *Adenomera a.*

Para transformar os áudios em uma representação 2-D (espectrogramas), será aplicada a transformada de tempo curto de Fourier (STFT). No processamento de sinais de áudio, incluindo a análise de sinais de áudio, é frequentemente utilizada a Transformada de Fourier de Tempo Curto [Rabiner & Schafer, 2010]. A STFT é empregada para calcular o espectrograma, dividindo a sinal de áudio em segmentos de tempo mais curtos e aplicando a Transformada de Fourier a cada segmento. Isso permite analisar a distribuição de frequências ao longo do tempo e obter informações detalhadas sobre as características acústicas do sinal de áudio. Em seguida, procede-se à normalização

do sinal conforme descrito na Seção 4.2.2. Para gerar o espectrograma de um áudio, inicialmente é extraído o sinal e a taxa de amostragem, conforme detalhado na Seção 2.4.2. Posteriormente, são determinados a frequência máxima, o tamanho e a extensão da janela. Em seguida, realiza-se a segmentação do sinal de áudio, conforme exposto na Seção 4.2.3.

4.2.2 Normalização

Usamos a técnica de normalização mínima-máxima, também conhecida como *rescaling*, para ajustar os valores de amplitude do sinal de áudio a uma faixa entre -1 e 1. Essa normalização é fundamental para garantir que os valores de entrada para a rede neural estejam dentro de uma faixa adequada. A normalização dos vetores restringe os valores das características dentro de uma faixa determinada. Geralmente, os valores são normalizados para obter média zero e variância um [Theodoridis et al., 2010].

O processo de normalização é realizado dividindo cada elemento do sinal de áudio pelo seu valor máximo absoluto. Dessa forma, garantimos que a amplitude máxima do sinal esteja dentro da faixa desejada e que os valores de entrada sejam ótimos para o processamento na rede neural.

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{4.1}$$

onde $\hat{x}$ representa o valor normalizado e x é o valor original da amostra temporal de áudio. A normalização é essencial para evitar discrepâncias na amplitude dos áudios, pois essas discrepâncias poderiam afetar a precisão do modelo.

4.2.3 Segmentação dos áudios

A segmentação é o processo de dividir o sinal de áudio em segmentos ou amostras menores, com uma duração específica. Esse processo é realizado aplicando-se um algo-

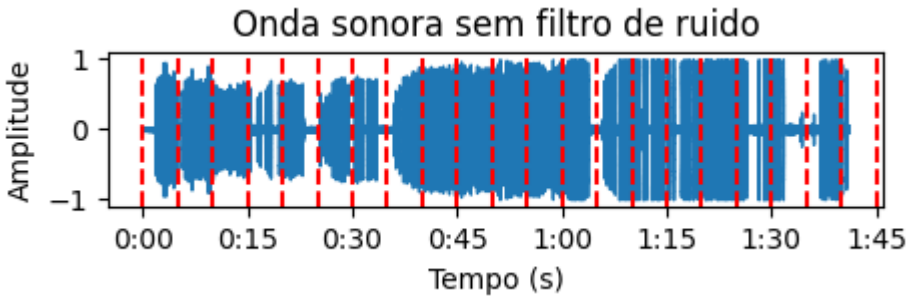


Figura 4.4. Segmentação de áudio a cada 5 segundos da espécie *Hylaedactylus*.

ritmo iterativo que, para cada segmento, multiplica a duração desejada (em segundos) pela taxa de amostragem do sinal. O número total de segmentos, M, que podem ser extraídos de uma gravação é calculado dividindo-se a duração total do áudio, D_t (em segundos), pela duração desejada para cada segmento, Δ, conforme a seguinte equação:

$$M = \left\lfloor \frac{D_t}{\Delta} \right\rfloor, \tag{4.2}$$

A segmentação dos áudios será realizada de acordo com os requisitos estabelecidos por cada um dos modelos pré-treinados que serão utilizados no desenvolvimento desta pesquisa. Na Figura 4.4, tomamos uma gravação de nosso banco de dados da espécie *Hylaedactylus* e realizamos a segmentação. A onda de cor azul representa o áudio, enquanto as linhas de cor vermelha indicam os segmentos a cada 5 segundos.

4.3 Extração dos embeddings

Nesta etapa, procederemos à extração dos vetores de *embeddings* por meio do modelo Perch mencionado na Seção 1.3. O modelo Perch é baseado na arquitetura EfficientNet B1, descrita por Tan & Le [2019], conforme explicado na Seção 2.6.3 dos fundamentos. Foi treinado com o banco de dados de cantos de aves Xeno-Canto. Perch possui como entrada áudios de 5 segundos com uma taxa de amostragem de 32 kHz. O espectrograma gerado pelo modelo é um espectrograma mel de tamanho 512 com uma janela

de 5 segundos. A dimensionalidade dos *embeddings* gerados é de 1280. Isso significa que, para este modelo, precisamos segmentar os áudios em amostras de 5 segundos.

Para extrair os *embeddings* desta arquitetura, o modelo pré-treinado é configurado com os parâmetros de entrada requeridos, como a taxa de amostragem e as dimensões do espectrograma. A abordagem consiste em utilizar o modelo como um extrator de características, acessando as representações vetoriais geradas em suas camadas intermediárias, antes das camadas finais de classificação.

O processo de extração é então aplicado a cada segmento de áudio de uma gravação. A saída da camada intermediária selecionada é capturada como o vetor de *embedding* correspondente, formando uma lista $X = \{x_1, \ldots, x_n\}$. Uma vez que todos os vetores de *embeddings* da gravação são coletados, eles são armazenados no banco de dados vetorial para a posterior etapa de recuperação acústica.

4.4 Fusão dos vetores de features

Nesta etapa os *embeddings* extraídos de cada gravação foram agregados em um único vetor representativo utilizando uma das estratégias de fusão propostas: *average pooling*, *weighted average pooling*, *sum pooling* ou *max pooling*. Para representar cada gravação com um vetor de tamanho fixo, aplicamos diversas técnicas de agregação por *pooling*, descritas a seguir.

Average Pooling: Esta técnica consiste em calcular a média aritmética sobre as colunas da matriz de *features* $X \in \mathbb{R}^{1280 \times M}$. Essa operação gera um único *embedding* global que representa o centroide de todas as características acústicas da gravação. Sua principal vantagem é a criação de uma representação suavizada, que mitiga a influência de ruído ou segmentos anômalos. A desvantagem, no entanto, é que essa abordagem pode diluir o impacto de eventos acústicos importantes, mas de curta duração.

A representação vetorial resultante para uma gravação é definida como:

$$\bar{x}_{\text{avg}} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{X}_{:,i} \in \mathbb{R}^{1280},$$

onde $\mathbf{X}_{:,i}$ representa a i-ésima coluna da matriz $\mathbf{X}$, correspondente ao vetor de *embedding* do i-ésimo segmento de áudio. O vetor $\bar{x}_{\text{avg}}$ resultante é armazenado no VectorDB e utilizado para consultas e recuperação por meio dos algoritmos IMENN e HNSW.

Weighted Average Pooling: Nesta técnica, incorporamos um peso de relevância para cada segmento com base na sua energia RMS (*Root Mean Square*), atribuindo maior influência a eventos acusticamente proeminentes, como vocalizações claras. Essa agregação ponderada enfatiza os segmentos informativos, reduzindo a contribuição de regiões de baixa energia ou ruído.

Dada a matriz de *features* $\mathbf{X} \in \mathbb{R}^{1280 \times M}$, onde cada coluna $\mathbf{X}_{:,i}$ representa o *embedding* do i-ésimo segmento, e um vetor de pesos correspondente $\mathbf{w} = [w_1, w_2, \ldots, w_M] \in \mathbb{R}^M$ com $w_i \geq 0$, o vetor resultante é calculado como:

$$\bar{x}_{\text{wavg}} = \frac{1}{\sum_{i=1}^{M} w_i} \sum_{i=1}^{M} w_i \cdot \mathbf{X}_{:,i} \in \mathbb{R}^{1280}.$$

Essa estratégia gera um *embedding* global $\bar{x}_{\text{wavg}}$ que representa de forma mais fiel os momentos informativos da gravação. Assim como no *average pooling*, os vetores resultantes são armazenados no VectorDB para avaliação com os algoritmos IMENN e HNSW.

Sum Pooling: Nesta técnica, o *embedding* global é calculado somando todos os vetores coluna da matriz de *features* $\mathbf{X} \in \mathbb{R}^{1280 \times M}$, que representam os segmentos individuais de uma gravação. O vetor resultante mantém a mesma dimensionalidade e captura a informação acústica cumulativa ao longo do tempo:

$$\bar{x}_{\text{sum}} = \sum_{i=1}^{M} \mathbf{X}_{:,i} \in \mathbb{R}^{1280}.$$

Embora esta abordagem enfatize a contribuição conjunta de todos os segmentos e mantenha a simplicidade computacional, ela apresenta uma desvantagem significativa por que o vetor resultante é altamente sensível ao número de segmentos (M), ou seja, à duração da gravação. Adicionalmente, o método é vulnerável a segmentos atípicos com valores extremos, que podem dominar o vetor agregado final.

Max Pooling: A técnica de *max pooling* constrói o *embedding* global selecionando, para cada dimensão, o valor máximo encontrado entre todas as colunas da matriz de *features* $\mathbf{X} \in \mathbb{R}^{1280 \times M}$. O objetivo é capturar o pico de ativação de cada característica, realçando assim os eventos acústicos mais proeminentes, independentemente de sua duração. A representação vetorial resultante é definida como:

$$\bar{x}_{\max} = \max_{i=1,\ldots,M} \mathbf{X}_{:,i} \in \mathbb{R}^{1280}, \tag{4.3}$$

onde a operação max é aplicada elemento a elemento entre as colunas (i.e., no tempo) para cada uma das 1280 dimensões. Essa abordagem, no entanto, funciona como uma faca de dois gumes. Ao mesmo tempo que é eficaz para destacar um evento dominante (como uma vocalização intensa), sua alta sensibilidade a valores extremos a torna particularmente vulnerável a ruídos impulsivos, como um estalo ou um clique, que podem gerar um pico de ativação anômalo e corromper a representação final do áudio.

Todas as técnicas de fusão descritas visam capturar padrões acústicos globais e locais, melhorar a discriminação entre classes e aumentar a robustez em tarefas de recuperação sob diferentes condições acústicas. Independentemente da estratégia de fusão empregada, o vetor resultante $\bar{x}$ possui uma dimensionalidade fixa de 1280, garantindo compatibilidade e eficiência computacional dentro do framework VectorDB. Cada vetor também está vinculado a metadados detalhados, conforme descrito na Seção 4.1.

Para fins de comparação com abordagens clássicas, também extraímos vetores de *features* MFCC com 128 dimensões a partir de cada segmento de 5 segundos de áudio, resultando em matrizes $\mathbf{X} \in \mathbb{R}^{128 \times M}$. As mesmas técnicas de fusão foram aplicadas

a essas matrizes MFCC, produzindo vetores agregados $\bar{x} \in \mathbb{R}^{128}$, permitindo uma comparação justa entre representações baseadas em *features* tradicionais e *embeddings* profundos.

4.5 Indexação no banco de dados vetorial

Para executar a etapa 5 do método, foi utilizada a base de dados vetorial VectorDB, fundamental para armazenar e representar os áudios de forma vetorial. Isso facilita o processamento e a análise de dados, permitindo a aplicação de algoritmos de aprendizado de máquina para extrair informações e realizar tarefas específicas no processamento de dados. No nosso caso, a implementação de uma base de dados vetoriais é especialmente útil para realizar a recuperação de vetores semelhantes, pois, ao reduzir a dimensionalidade dos dados e representá-los em vetores de *embeddings*, ocuparão menos espaço. Além disso, através de algoritmos de indexação, organizamos os vetores de *embedding* por similaridade acústica os armazenamos de maneira eficiente na base de dados vetorial. Tudo isso foi realizado com o objetivo de alcançar a recuperação da informação de paisagens acústicas da maneira mais eficaz possível.

Para a etapa de indexação, foram utilizados os algoritmos Índice de Arquivo Invertido (IVF) e Aproximação *Hierárquica Navigable Small World* (HNSW). Nas Seções 2.10.1 e 2.10.2, foi explicado o funcionamento e o pseudocódigo desses algoritmos. Em nossa pesquisa, o armazenamento é realizado usando a biblioteca VectorDB. Primeiramente, é definida uma classe generica que representa um documento com campos específicos, incluindo um campo de *embedding*. Esta classe estende a BaseDoc do vectordb e utiliza o tipo *NdArray* para representar o campo de *embedding*. Em seguida, é carregado o VectorDB usando esta classe, e uma lista de documentos (vetores de *embeddings*) é indexada. Neste caso, a lista de documentos é composta por instâncias da classe. Posteriormente, é criada a base VectorDB em memória que utiliza a classe como tipo de documento e é especificada a localização do espaço de trabalho.

Depois é criada uma lista de documentos $X = \{x_1, \ldots, x_n\}$ utilizando a classe, atribuindo valores aos campos text e *embedding* utilizando um loop iterativo para enumerar cada um dos *embeddings*. Neste caso, text é definido como o caminho do arquivo de áudio, e *embedding* é obtido da matriz de *embeddings* do conjunto de dados. Por fim, a lista de documentos X é indexada no VectorDB utilizando o método index. Após esse processo, o VectorDB mapeou e armazenou os documentos juntamente com seus *embeddings* correspondentes. Esses vetores de *embeddings* indexados serão utilizados na etapa de recuperação acústica para realizar consultas e recuperar vetores similares em termos de informação acústica.

Para garantir a reprodutibilidade e uma comparação justa entre os algoritmos de busca, foram configurados hiperparâmetros análogos para a construção dos índices. Para o algoritmo HNSW, o número máximo de conexões por nó no grafo foi definido como $M = 32$. De forma análoga, para o algoritmo IMENN, que opera sobre partições de dados, foi utilizado um valor equivalente de 32 *clusters* para a criação do índice. Os demais hiperparâmetros que controlam o esforço de busca como o fator de construção *ef_construction=200*, e o fator de busca como *ef_search=50* foram mantidos consistentes entre os algoritmos para balancear de forma equitativa o *trade-off* entre precisão e velocidade.

4.6 Recuperação acústica

Para concluir a fase final do método, foram utilizados os algoritmos de recuperação de informação por aproximação HNSW e por exatidão IVF, conforme descritos nas seções 2.10.1 e 2.10.2, respectivamente, para realizar as consultas. Esses algoritmos têm demonstrado obter bons resultados ao buscar os vizinhos mais próximos [Han et al., 2023, Stowell, 2022, Wang & et al., 2003]. O algoritmo de aproximação HNSW apresenta maior rapidez e eficiência do que o algoritmo de exatidão IVF [Gao & Long, 2023]. Na etapa experimental, os algoritmos HNSW e IVF foram avaliados para determinar

qual oferece o melhor desempenho em termos de precisão e eficácia na similitude das consultas.

Dados os vetores de *embeddings* é criado um objeto chamado consulta (*query*) que representa a consulta a ser realizada no banco de dados vetorial. O campo de texto é definido como uma *string* vazia e o campo de *embedding* recebe o *embedding* correspondente ao arquivo de áudio. Em seguida, as consultas são feitas no banco de dados vetorial utilizando o método search do VectorDB. A consulta é feita utilizando a lista de áudios (*querys*) como entrada. Por fim, as correspondências são obtidas, constituídas por objetos mais similares à consulta.

Para avaliar o método, foram tomados os tempos de recuperação das consultas no bando de dados vetorial e as taxas de H@1, H@5, na busca dos vetores mais similares à consulta (*query*), junto com os desvios padrões correspondientes ($\pm$). Nesse contexto, a similaridade refere-se a segmentos de áudio que compartilham características acústicas, seja porque pertencem à mesma espécie ou porque se originam do mesmo ambiente ou sessão de gravação. Como linha de base, implementamos os MFCCs tradicionais com 40 coeficientes por segmento. Os MFCCs foram extraídos utilizando uma FFT de 1024 pontos, com um salto (*hop length*) de 512 amostras e aplicação de janela de Hann. Os mesmos segmentos de 5 segundos utilizados para a extração dos *embeddings* foram empregados no cálculo dos MFCCs, garantindo uma comparação justa.

4.7 Considerações finais sobre o método

Cada etapa do método possui procedimentos específicos que devem ser executados da melhor forma possível para obter os melhores resultados:

1. **Pré-processamento:** A segmentação dos áudios é necessária, seguida da normalização para facilitar a extração dos *embeddings*.

2. **Vetores de embeddings:** Nesta fase, é crucial definir adequadamente qual é o

melhor modelo para extrair os vetores de *embeddings*.

3. **Mapeamento e Armazenamento:** Esta etapa desempenha um papel fundamental no método, pois realiza a indexação dos vetores de *embeddings* e os armazena na base de dados vetorial.

4. **Recuperação Acústica:** Esta última etapa do sistema determinará qual algoritmo será utilizado para realizar a recuperação acústica dos dados obtidos das paisagens sonoras.

O objetivo final destas etapas é alcançar de maneira eficiente e eficaz a recuperação acústica de paisagens sonoras implementando um banco de dados vetorial. A escolha entre velocidade e precisão nos modelos de busca apresenta um dilema central nos sistemas de recuperação de informação. O objetivo principal é manter a precisão nas consultas enquanto se alcança um tempo de resposta reduzido. Ao contrário dos bancos de dados relacionais, que se concentram na igualdade de valores e na estrutura tabular, os bancos de dados vetoriais aproveitam as propriedades dos vetores para realizar buscas baseadas em similaridade e contexto. Em ambientes com grandes conjuntos de dados, nos quais a velocidade e a precisão são críticas, surge a proposta de utilizar bancos de dados vetoriais como uma possível solução. Isso implica equilibrar a velocidade e precisão das consultas e assegurar a confiabilidade das métricas de distância em espaços de alta dimensionalidade.

Da mesma forma, a dispersão dos vetores em espaços de alta dimensionalidade presentes em bancos de dados vetoriais complica a recuperação eficiente da informação. Além disso, é crucial ter em mente que o *embedding* gerado pela rede neural foi treinado especificamente através de camadas densas para classificação, e não com o objetivo de otimizar nenhum cálculo de distância de similaridade. Isso significa que o vetor de *embedding* captura de forma eficaz as características do áudio original, mas não foi projetado para favorecer um cálculo de distância em particular. Trata-se de um vetor de *embedding* genérico que não foi ajustado para melhorar seu desempenho em

relação a nenhuma equação de distância específica, o que pode influenciar na precisão das consultas. Utilizar algoritmos de agrupamento para a indexação dos vetores de *embeddings* na base de dados vetorial surge como uma solução viável para enfrentar essa problemática.

5

Resultados

Nesta seção apresentamos os resultados dos experimentos realizados para avaliar o desempenho das diferentes estratégias propostas para a recuperação acústica de paisagens sonoras. O principal objetivo é analisar como diferentes representações acústicas, técnicas de fusão de vetores *features* e métodos de indexação vetorial influenciam a precisão e a eficiência do sistema de recuperação. Além disso, os resultados obtidos são comparados com abordagens tradicionais reportadas na literatura, buscando evidenciar as contribuições e limitações da metodologia proposta.

Todos os experimentos foram realizados em uma estação de trabalho equipada com um CPU Ryzen 5 de 2,5 GHz, 16 GB de RAM e um disco rígido de 1000 GB. Os resultados foram obtidos comparando os algoritmos *IMENN* e *HNSW* do banco de dados vetorial VectorDB. Por fim, os resultados obtidos serão discutidos detalhadamente ao longo das próximas seções, destacando os fatores que influenciam o desempenho, as vantagens e limitações de cada estratégia e a aplicabilidade prática da metodologia no contexto do monitoramento ecoacústico em larga escala.

5.1 Metodologias de avaliação

A avaliação do sistema proposto foi estruturada em três etapas experimentais sequenciais. A primeira etapa foca na definição e comparação de diferentes protocolos de recuperação. A segunda etapa utiliza o protocolo mais eficaz para realizar uma análise sistemática das técnicas de fusão de *features*. Finalmente, a terceira etapa valida a abordagem completa em um caso de uso prático para o monitoramento de espécies vulneráveis.

5.2 Etapa 1: definição do protocolo de recuperação experimental

Nesta primeira etapa, o objetivo foi estabelecer o protocolo experimental mais robusto para a tarefa de recuperação acústica. Para isso, foram comparadas três configurações distintas de consulta e particionamento de dados utilizando os cinco conjuntos de dados *CBI,* BC_{22}*,* BC_{23}*, XBC* e o de anuros, descritos na Seção 4.1.

5.2.1 Protocolo 1: Generalização entre Gravações

A primeira configuração experimental teve como objetivo avaliar a capacidade do sistema de recuperar gravações da mesma espécie capturadas em contextos diferentes, e sua configuração foi a seguinte:

- **Query:** A primeira gravação completa de cada espécie.
- **Base:** Todas as gravações restantes foram armazenadas no banco de dados vetorial.

Essa configuração nos permite avaliar a capacidade do sistema de recuperar gravações semelhantes capturadas em diferentes momentos e locais, refletindo cenários do mundo real de monitoramento da biodiversidade.

5.2.2 Protocolo 2: Consulta por Evento Proeminente

A segunda configuração experimental teve como objetivo testar se um evento acústico de alta relevância (como uma vocalização clara) pode servir como um ponto de referência eficaz para a recuperação, e sua configuração foi a seguinte:

- **Query:** O segmento com a maior energia de cada gravação.
- **Base:** Todos os segmentos restantes da mesma gravação foram armazenados no banco de dados vetorial.

Essa estratégia avalia se um evento acústico proeminente serve como um bom ponto de referência para recuperar segmentos relacionados em ambientes ruidosos, simulando uma tarefa de busca por exemplo.

5.2.3 Protocolo 3: Generalização com Particionamento 70/30

A terceira configuração experimental teve como objetivo avaliar a capacidade de generalização do sistema no cenário de teste mais rigoroso, e sua configuração foi a seguinte:

- **Query:** Um conjunto de consultas criado a partir de 30% das gravações de cada espécie.
- **Base:** Os 70% restantes das gravações de cada espécie foram armazenados no banco de dados.

Esta estratégia, que impede qualquer sobreposição de gravações entre a consulta e a base, representa o teste padrão em *Machine Learning*. Ela avalia diretamente a capacidade do sistema de generalizar, ou seja, de recuperar com precisão segmentos de áudio de uma gravação completamente inédita.

O propósito desta análise foi determinar como diferentes estratégias de seleção de consultas e organização da base de dados influenciam o desempenho da recuperação,

garantindo ao mesmo tempo a ausência de vazamento de informações entre os conjuntos de consulta e de teste.

5.3 Etapa 2: análise comparativa das técnicas de fusão de features

Para a realização dos experimentos desta etapa, foi utilizado o conjunto de dados `BirdCLEF+ 2025`. Com base nos resultados da etapa anterior, esta segunda fase experimental foi desenhada para identificar a técnica de fusão de vetores de *features* mais eficaz. Foram sistematicamente comparadas quatro estratégias (*Average Pooling*, *Weighted Average Pooling*, *Sum Pooling* e *Max Pooling*) sob três cenários de recuperação distintos, conforme ilustrado na Figura 5.2.

5.3.1 Cenário 1: Generalização com Divisão 70/30

Este cenário avaliou as técnicas de fusão sob uma condição de teste rigorosa, e sua configuração foi a seguinte:

- **Base:** 70% das gravações de cada espécie compuseram o banco de dados.
- **Query:** Os 30% restantes das gravações foram usados exclusivamente como consultas.

Ao garantir que as *queries* e os dados indexados fossem mutuamente exclusivos, esta configuração permitiu uma comparação sistemática e justa de como cada estratégia de fusão impacta a capacidade do sistema de generalizar e recuperar gravações acusticamente semelhantes da mesma espécie.

5.3.2 Cenário 2: Consulta com Evento Proeminente

Este cenário investigou o desempenho das técnicas de fusão quando a consulta representa um evento acústico de alta relevância. A configuração foi a seguinte, adaptando-se a construção da *query* conforme o objetivo da recuperação:

- **Base:** A base de dados continha vetores fusionados, onde cada vetor representava uma gravação completa.
- **Query:** A consulta foi construída de duas formas distintas:
  Para recuperação a nível de espécie: Foi criado um vetor único a partir da fusão dos segmentos de maior energia das gravações da mesma espécie para criar um único vetor-consulta.
  Para recuperação a nível de gravação: utilizou-se o vetor não fusionado do único segmento de maior energia de cada gravação.

O objetivo principal foi avaliar qual das quatro estratégias de fusão, ao ser aplicada na base de dados, era mais eficaz em corresponder a esses dois tipos de consulta proeminente, permitindo recuperar com precisão tanto gravações da mesma espécie quanto a própria gravação de origem.

5.3.3 Cenário 3: Base de Dados Compactada

O terceiro cenário experimental foi projetado para quantificar o *trade-off* entre eficiência e precisão, avaliando a capacidade do sistema de recuperar um evento acústico proeminente a partir de consultas menos salientes. A eficiência foi medida pela latência ($t \pm \sigma$) e vazão de consultas, enquanto a precisão foi avaliada pela qualidade do *ranking*. Nesta abordagem, a configuração foi a seguinte:

- **Base:** A base de dados foi compactada para conter unicamente o vetor não fusionado do segmento de maior energia de cada gravação.

- **Query:** As consultas foram criadas a partir de todos os segmentos restantes de uma gravação, aplicando a fusão sobre eles para gerar um único vetor de consulta.

Este arranjo permitiu verificar qual método de fusão na consulta é mais eficaz para agregar características de segmentos secundários e, ainda assim, identificar com sucesso o evento principal da gravação na base de dados compactada.

A volumetria de dados variou significativamente entre os cenários: o primeiro utilizou 130.245 segmentos na base de dados e 57.594 para consultas; o segundo, 187.839 na base e 26.250 para consultas; e o terceiro inverteu essa proporção, com 26.250 segmentos na base e 187.839 para consultas.

A qualidade do *ranking* de recuperação foi avaliada com base em uma relevância binária no nível de espécie, utilizando a distância Euclidiana (L2) para calcular a similaridade entre os *embeddings*. Para mensurar o desempenho, foram empregadas três métricas de ordenação padrão calculadas no Top-5: o *Mean Reciprocal Rank (MRR@5)*, para aferir a posição do primeiro resultado relevante; a *Mean Average Precision (mAP@5)*, para medir a precisão média acumulada; e o *Normalized Discounted Cumulative Gain (nDCG@5)*, para avaliar a qualidade geral da ordenação com um desconto logarítmico. Para a execução dos cálculos, o rótulo de espécie da *query* foi inferido a partir do prefixo de seu identificador, e uma distribuição de probabilidade de $(0.4, 0.3, 0.2, 0.1)$ foi utilizada para as posições de 2 a 5 ao estimar as contribuições esperadas. Adicionalmente, para quantificar a eficiência computacional de cada cenário, foram mensurados o tempo médio de consulta ($t \pm \sigma$) e a vazão do sistema, expressa em consultas por segundo (QPS).

5.4 Etapa 3: validação em caso de caso para monitoramento de espécies vulneráveis

Na Etapa 3, com o objetivo de demonstrar a aplicabilidade do sistema, foi projetado um caso de uso focado na interação funcional entre pesquisadores de bioacústica, conservação de espécies e monitoramento da vida selvagem, e o sistema proposto de recuperação. Nesse cenário, o usuário envia uma gravação de referência não rotulada, e o sistema retorna um conjunto de gravações acusticamente semelhantes, juntamente com seus respectivos metadados: espécie, localização geográfica aproximada e fonte. Essa funcionalidade, apoiada por *embeddings* pré-treinados, técnicas de fusão de vetores *features* e busca baseada em vetores,foi projetada para possibilitar uma análise mais rápida dos dados acústicos, reduzir o trabalho manual de revisão.

Utilizando a segunda configuração experimental definida na Etapa 2, juntamente com a técnica de fusão *weighted average pooling*, que apresentou os melhores resultados, o sistema foi configurado para recuperar gravações semelhantes agrupadas por espécie dentro do banco de dados vetorial. Este experimento final concentrou-se em treze espécies representativas do conjunto de dados `BirdCLEF+ 2025`: *Elaenia flavogaster*, *Penelope purpurascens*, *Megarynchus pitangua*, *Andinobates opisthomelas*, *Pyrilia pyrilia*, *Panthera onca*, *Alouatta seniculus*, *Bradypus variegatus*, *Colostethus inguinalis*, *Cerdocyon thous*, *Allobates niputidea*, *Lontra longicaudis* e *Crax alberti*. Essas espécies foram selecionadas com base no seu status de conservação e risco de declínio populacional (de acordo com as categorias da Lista Vermelha), determinado por meio do cruzamento dos metadados das gravações com a Lista Vermelha da IUCN IUCN Red List [2024].

Para cada gravação recuperada, foram analisados metadados importantes, como localização geográfica (latitude e longitude) e fonte (XC, iNat ou CSA). Além disso, a biblioteca `geopy`[1] foi utilizada em combinação com o serviço Nominatim, uma fer-

[1] `https://pypi.org/project/geopy/`

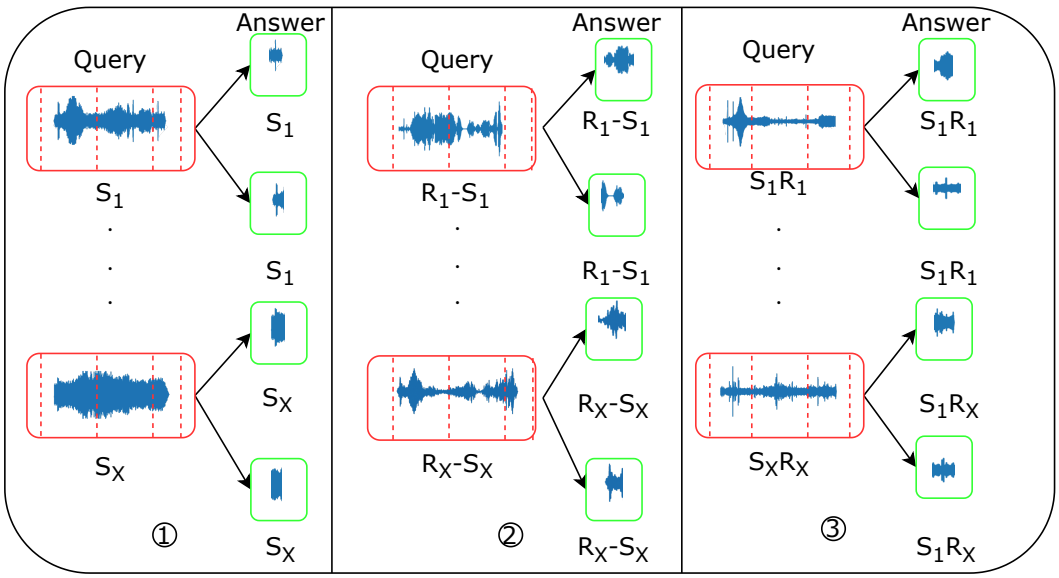


Figura 5.1. Abordagens metodológicas: (1) divisão 70%-30% no nível de gravação; (2) segmento de maior energia como consulta; (3) segmento de maior energia no banco de dados e segmentos restantes como consultas.

ramenta de geocodificação reversa que utiliza dados do OpenStreetMap, para obter a localização textual aproximada associada a cada gravação. Um mapa geográfico 2D foi gerado para visualizar a distribuição espacial das espécies e destacar possíveis padrões espaciais. O objetivo dessa abordagem é demonstrar o potencial do sistema para apoiar pesquisas ecológicas e ações de conservação, permitindo que biólogos monitorem mudanças na distribuição e identifiquem habitats críticos. A combinação do mapa e da tabela de metadados facilita uma interpretação ecológica rápida e embasada para apoiar a tomada de decisão.

5.5 Avaliando o impacto do filtro de ruido

Esta etapa tem como objetivo avaliar e determinar se a aplicação de um filtro de redução de ruído, como etapa de pré-processamento, melhora de forma quantificável a precisão da recuperação acústica, tanto para os *embeddings* quanto para os MFCCs. Nas próximas seções são apresentados os resultados e sues analise.

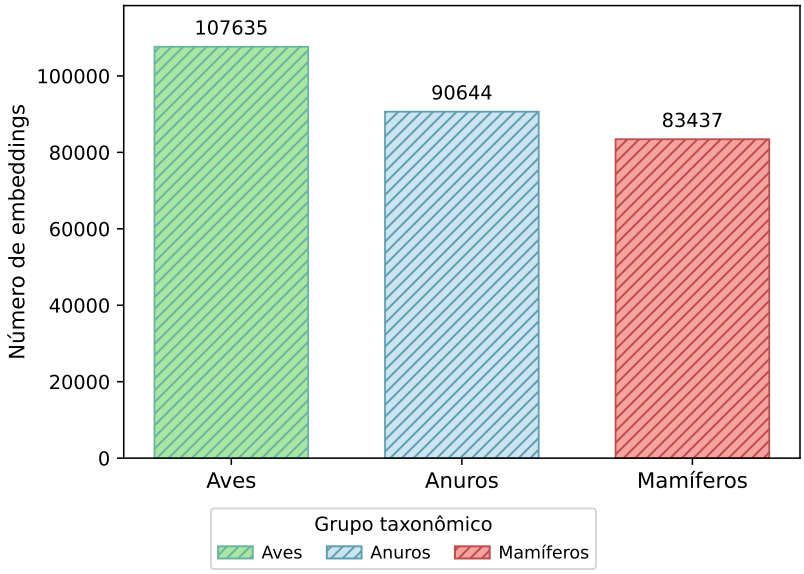


Figura 5.2. Embedings por grupo taxonômico.

5.5.1 Resultados do experimento 1

O primeiro experimento teve como objetivo avaliar o impacto do pré-processamento acústico utilizando o filtro *Noise Reduce* na tarefa de recuperação de paisagens sonoras acusticamente semelhantes. Nessa configuração, foi utilizado um cenário de consulta por exemplo, no qual uma gravação completa de uma espécie é utilizada como referência para encontrar outras gravações da mesma espécie, e foram comparadas duas representações de características distintas, utilizando *embeddings* Perch e coeficientes MFCCs tradicionais. Para a indexação e busca, foram utilizados os algoritmos *HNSW* e *IMENN*. As avaliações foram conduzidas em cinco diferentes bancos de dados e as métricas consideradas foram *Top-1* ($T1$), *Hit@1* ($H1$), *Hit@5* ($H5$), o tempo médio de consulta e o desvio padrão ($t \pm \sigma$).

Os resultados apresentados nas Tabelas 5.1 e 5.2 indicam que o filtro de ruído melhora o desempenho médio na maioria dos casos, validando uma parte crucial da hipótese. O incremento é particularmente notável nos datasets de aves BC_{22}, BC_{23}, e *XCB*, que são intrinsecamente mais ruidosos. Por exemplo, no *XCB*, o $H@5$ médio do Perch com *HNSW* aumenta de 0.46 para 0.57 após a aplicação do filtro, evidenciado

na Tabela 5.1. Embora os intervalos de confiança (IC 95%) permaneçam em 5.9% para ambos os casos como se mostra na Tabela 5.3, a ligeira redução no Coeficiente de Variação (CV) de 7.6% para 7.5% sugere uma maior consistência nos resultados após a filtragem. No conjunto BC_{23} com Perch e *HNSW*, o $H@5$ médio aumentou de 0.48 para 0.59, e o CV diminuiu de 7.1% para 6.5%, reforçando que uma maior relação sinal-ruído na entrada conduz a *embeddings* de maior qualidade e a uma recuperação mais precisa e consistente.

O segundo achado evidencia de maneira conclusiva que os *embeddings* extraídos do Perch constituem uma representação vetorial superior para esta tarefa. Comparando os melhores cenários, aplicando o filtro e *HNSW*, o Perch supera significativamente os MFCCs em todos os conjuntos de dados. No dataset BC_{23}, o Perch alcança um $H@5$ médio de 0.59, com um IC de 5.9% e CV de 6.5%; face aos 0.45 dos MFCCs com IC de 5.6% e CV de 8.4% evidenciados nas Tabelas 5.1, 5.2, 5.3. A grande diferença nas médias, aliada à notavelmente menor variabilidade relativa do Perch, demonstra a superioridade desta representação, associada à sua capacidade de capturar padrões espectro-temporais complexos que os MFCCs, de dimensionalidade reduzida, não conseguem representar adequadamente em cenários ecoacústicos diversos.

O terceiro resultado está relacionado aos algoritmos de busca. O algoritmo *HNSW* não só é significativamente mais rápido, como também oferece consistentemente maior precisão que o *IMENN*. No conjunto *XCB* aplicando o filtro de ruido e Perch, *HNSW* alcança $H@5 = 0.57$ e um CV de 7.5% enquanto *IMENN* obtém $H@5 = 0.45$, um CV: 8.7%, ver as Tabelas 5.1 e 5.3. Essa diferença substancial na média e a maior consistência, ou seja, com menor CV, confirmam a vantagem da busca aproximada via *HNSW*, cuja estrutura de grafo navegável parece capturar a topologia dos dados de forma mais eficaz que a busca exata em *clusters* do *IMENN*, especialmente em espaços de alta dimensão.

De forma geral, os resultados confirmam que o filtro de ruído produz um incremento relevante na precisão para conjuntos de dados heterogêneos. A melhoria é mais

pronunciada para o Perch do que para os MFCCs, sugerindo que as redes profundas capitalizam melhor a limpeza do sinal para gerar vetores mais robustos e representativos. Esses achados reforçam a relevância do pré-processamento na bioacústica.

Tabela 5.1. Resultados do experimento 1 com Perch: primeira gravação completa como consulta.

Sem aplicação do filtro de ruído								
	HNSW				**IMENN**			
	T1	**H**	**H5**	$t \pm \sigma$	**T1**	**H**	**H5**	$t \pm \sigma$
BC_{22}	0.31	0.34	**0.40**	2.5±2.9	0.25	0.28	0.33	2.5±4.6
BC_{23}	0.38	0.42	**0.48**	0.4±4.1	0.27	0.29	0.35	0.5±4.9
CBI	0.23	0.25	**0.27**	0.4±6.8	0.13	0.14	0.17	0.6±3.1
Anuran	0.85	0.88	**0.91**	0.2±1.7	0.81	0.85	0.87	0.3±3.6
XCB	0.36	0.38	**0.46**	3.6±3.2	0.29	0.32	0.37	5.4±4.8
Aplicando o filtro de ruído								
BC_{22}	0.34	0.37	**0.45**	2.1±2.1	0.29	0.33	0.37	2.2±3.9
BC_{23}	0.49	0.53	**0.59**	0.2±2.4	0.37	0.41	0.48	0.3±3.7
CBI	0.24	0.27	**0.29**	0.3±4.2	0.16	0.18	0.21	0.4±2.8
Anuran	0.91	0.93	**0.95**	0.1±1.4	0.86	0.89	0.91	0.2±2.3
XCB	0.41	0.46	**0.57**	3.6±2.3	0.33	0.37	0.45	3.7±3.2

Tabela 5.2. Resultados do experimento 1 com MFCCs: primeira gravação completa como consulta.

Sem aplicação do filtro de ruído								
	HNSW				**IMENN**			
	T1	**H**	**H5**	$t \pm \sigma$	**T1**	**H**	**H5**	$t \pm \sigma$
BC_{22}	0.23	0.28	**0.33**	1.9±4.9	0.15	0.17	0.21	2.4±3.6
BC_{23}	0.25	0.33	**0.35**	0.5±4.8	0.18	0.21	0.26	0.4±3.2
CBI	0.15	0.17	**0.18**	0.7±6.1	0.09	0.11	0.12	0.9±4.1
Anuran	0.46	0.55	**0.62**	0.2±2.9	0.38	0.42	0.47	0.3±3.4
XCB	0.21	0.26	**0.29**	5.5±5.6	0.13	0.15	0.18	7.2±4.5
Aplicando o filtro de ruído								
BC_{22}	0.29	0.32	**0.36**	2.7±3.7	0.18	0.22	0.25	2.5±3.6
BC_{23}	0.33	0.36	**0.45**	0.3±2.5	0.21	0.25	0.30	0.4±2.6
CBI	0.17	0.19	**0.21**	0.5±3.5	0.13	0.15	0.17	0.6±3.1
Anuran	0.53	0.58	**0.71**	0.2±2.3	0.44	0.56	0.60	0.2±2.2
XCB	0.25	0.28	**0.32**	3.6±3.5	0.16	0.19	0.22	3.8±3.2

5.5.2 Resultados do experimento 2

Neste cenário selecionou-se o trecho com maior intensidade acústica de cada gravação como *query*, com a hipótese de que essas regiões conteriam padrões mais representativos e discriminativos para recuperar outros segmentos da mesma espécie. Assim como no

Tabela 5.3. Análise estatística de métricas de precisão para o Experimento 1. A tabela apresenta os valores de Intervalo de Confiança (CI) de 95% e Coeficiente de Variação (CV) relatados como porcentagens (%).

Sem aplicar o filtro Redução de Ruído								
Dataset	**HNSW (Perch)**		**IMENN (Perch)**		**HNSW (MFCCs)**		**IMENN (MFCCs)**	
	IC	**CV**	**IC**	**CV**	**IC**	**CV**	**IC**	**CV**
BC_{22}	7.0	10.2	6.6	11.9	6.5	12.1	5.6	16.3
BC_{23}	5.9	7.1	5.5	9.3	5.5	9.2	4.9	11.7
CBI	5.2	10.7	4.3	14.9	4.5	13.8	**3.7**	17.9
Anuran	16.2	8.6	18.3	10.7	21.7	10.0	21.9	13.8
XCB	5.9	7.6	5.6	8.9	5.2	10.7	4.3	14.7
Aplicando o filtro Redução de Ruído								
BC_{22}	7.0	10.0	6.6	11.1	6.6	11.6	6.2	16.8
BC_{23}	5.9	**6.5**	5.5	8.2	5.6	8.4	5.1	12.3
CBI	5.2	10.6	4.4	14.4	4.5	13.9	3.9	17.2
Anuran	15.3	9.0	18.1	10.5	25.8	12.1	21.7	13.1
XCB	5.9	7.5	5.6	8.7	5.4	10.6	4.6	14.4

experimento anterior, foram comparados os *embeddings* Perch e MFCCs. Na busca, foram avaliados os algoritmos *HNSW* e *IMENN* sobre os mesmos cinco conjuntos de dados, BC_{22}, BC_{23}, *CBI*, *Anuran* e *XCB*, utilizando as métricas *Top-1* ($T1$), *Hit@1* ($H1$), *Hit@5* ($H5$), $t \pm \sigma$ e os dados estatísticos.

Ao comparar os resultados com os do Experimento 5.5.1, observa-se uma melhoria notável na consistência e precisão na maioria dos cenários ao usar o segmento de maior energia como consulta. Por exemplo, no dataset *XCB* com Perch, filtro e *HNSW*, evidenciado na Tabela 5.4, a precisão $H@5$ média aumenta ligeiramente de 0.57 para 0.61, mas a redução drástica no IC 95% de 5.9% para 0.8% e no CV de 7.5% para 2.1%, mostrados nas Tabelas 5.3 e 5.6, indica que os resultados são significativamente mais estáveis e confiáveis. Isto sugere que o segmento de maior energia é, efetivamente, um vetor de consulta mais potente.

O impacto do filtro de redução de ruído comparando as metades as Tabelas 5.4 e 5.5 continua a ser positivo, melhorando a consistência dos resultados na maioria dos cenários, como evidenciado pela tendência de redução do CV na Tabela 5.6. Por exemplo, utilizando Perch com *HNSW* no conjunto BC_{23}, a aplicação do filtro levou a um pequeno aumento no $H@5$ médio de 0.55 para 0.57 mas a uma redução no CV de 2.1% para 1.9%. A superioridade dos embeddings Perch sobre os MFCCs mantém-se

ou até se acentua neste cenário. No dataset BC_{23} com filtro e *HNSW*, o Perch, como se evidencia na Tabela 5.4, atinge uma precisão $H@5$ média de 0.57, um IC de 0.8% e um CV de 1.9%, enquanto os MFCCs, ver a Tabela 5.5, ficam em 0.47, com um IC de 0.8% e um CV de 3.0% mostrados na Tabela 5.6. A grande diferença nas médias e a variabilidade relativa substancialmente menor para o Perch confirmam que ele se torna ainda mais discriminativo com uma consulta forte e clara, enquanto os MFCCs permanecem mais suscetíveis ao contexto.

Na comparação dos algoritmos de busca, a vantagem do *HNSW* sobre o *IMENN* persiste tanto em velocidade quanto em precisão. No dataset *XCB* com filtro e Perch, o *HNSW* alcança $H@1 = 0.51$ e um CV de 2.1% em 2.3 ± 2.8 ms, enquanto o *IMENN* obtém $H@1 = 0.39$ e um CV de 2.9% em 2.7 ± 3.2 ms, mostrados nas Tabelas 5.4 e 5.6. A melhoria na qualidade da consulta parece facilitar a navegação eficiente do *HNSW* no seu grafo hierárquico.

De forma geral, os resultados do Experimento 5.5.2 confirmam que o uso do segmento de maior energia como consulta melhora significativamente a precisão e a consistência da recuperação em comparação com o uso da gravação completa, especialmente com os *embeddings* Perch. Esses achados destacam a importância da seleção criteriosa de regiões de consulta, sugerindo que abordagens focadas em segmentos representativos podem aumentar a eficácia de sistemas de monitoramento automatizado.

Tabela 5.4. Resultados do experimento 2 com Perch: segmento com a maior energia de cada gravação foi selecionado como consulta.

Sem aplicação do filtro de ruído								
	HNSW				**IMENN**			
	T1	**H**	**H5**	$t \pm \sigma$	**T1**	**H**	**H5**	$t \pm \sigma$
BC_{22}	0.32	0.37	**0.43**	1.5±3.3	0.27	0.30	0.35	1.7±4.7
BC_{23}	0.44	0.51	**0.55**	0.3±3.2	0.29	0.34	0.38	0.3±3.4
CBI	0.25	0.28	**0.31**	0.5±4.2	0.16	0.19	0.21	0.7±4.5
Anuran	0.91	0.92	**0.96**	0.2±3.2	0.83	0.86	0.90	0.2±4.8
XCB	0.42	0.46	**0.53**	2.6±2.3	0.34	0.37	0.42	3.9±3.9
Aplicando o filtro de ruído								
BC_{22}	0.37	0.44	**0.48**	1.3±2.9	0.29	0.33	0.37	1.5±3.3
BC_{23}	0.46	0.53	**0.57**	0.2±1.2	0.38	0.43	0.52	0.3±3.1
CBI	0.27	0.30	**0.32**	0.5±2.7	0.18	0.20	0.23	0.7±2.8
Anuran	0.93	0.95	**0.97**	0.1±1.4	0.89	0.92	0.96	0.2±2.5
XCB	0.51	0.59	**0.61**	2.3±2.8	0.39	0.45	0.54	2.7±3.2

Tabela 5.5. Resultados do experimento 2 com MFCCs: segmento com a maior energia de cada gravação foi selecionado como consulta.

Sem aplicação do filtro de ruído								
	HNSW				**IMENN**			
	T1	**H**	**H5**	$t \pm \sigma$	**T1**	**H**	**H5**	$t \pm \sigma$
BC_{22}	0.25	0.30	**0.35**	2.0±3.6	0.15	0.17	0.21	2.4±3.5
BC_{23}	0.27	0.35	**0.38**	0.4±3.4	0.20	0.24	0.29	0.4±3.8
CBI	0.14	0.16	**0.19**	0.7±3.4	0.11	0.14	0.17	0.8±4.8
Anuran	0.49	0.60	**0.65**	0.3±2.4	0.42	0.45	0.53	0.3±3.1
XCB	0.23	0.28	**0.33**	3.9±4.3	0.19	0.21	0.24	4.0±3.5
Aplicando o filtro de ruído								
BC_{22}	0.31	0.35	**0.38**	1.9±3.2	0.18	0.22	0.25	2.5±3.1
BC_{23}	0.36	0.40	**0.47**	0.3±2.3	0.24	0.29	0.34	0.4±1.9
CBI	0.19	0.20	**0.23**	0.7±2.8	0.15	0.16	0.19	0.8±2.9
Anuran	0.53	0.58	**0.71**	0.2±2.1	0.44	0.56	0.60	0.3±2.2
XCB	0.28	0.31	**0.36**	3.1±2.5	0.22	0.25	0.27	3.3±3.8

5.5.3 Resultados do experimento 3

O terceiro experimento teve como objetivo avaliar a capacidade de generalização do sistema num cenário mais rigoroso. Ao implementar uma divisão de 70% para a base de dados e 30% para consultas ao nível de gravação completa, assegura-se que não existe qualquer sobreposição temporal ou de contexto entre os dados de teste e os dados indexados. Procurou-se medir o desempenho na identificação de vocalizações da mesma espécie provenientes de gravações completamente inéditas. Foram comparados

Tabela 5.6. Análise estatística de métricas de precisão para o Experimento 2. A tabela apresenta os valores de Intervalo de Confiança (CI) de 95% e Coeficiente de Variação (CV) relatados como porcentagens (%).

Sem aplicar o filtro Redução de ruído								
Dataset	**HNSW (Perch)**		**IMENN (Perch)**		**HNSW (MFCCs)**		**IMENN (MFCCs)**	
	IC	**CV**	**IC**	**CV**	**IC**	**CV**	**IC**	**CV**
BC_{22}	0.9	2.8	0.9	3.7	0.9	4.6	0.9	6.0
BC_{23}	0.8	2.1	0.8	2.9	0.8	3.3	0.8	4.2
CBI	**0.7**	2.0	**0.7**	2.7	**0.7**	3.2	**0.7**	4.1
Anuran	7.3	3.9	8.0	4.7	8.9	5.4	9.2	6.2
XCB	0.8	2.3	0.8	3.0	0.8	3.8	0.8	4.9
Aplicando o filtro Redução de ruído								
BC_{22}	0.9	2.6	0.9	3.5	0.9	4.3	0.9	5.5
BC_{23}	0.8	**1.9**	0.8	2.7	0.8	3.0	0.8	3.9
CBI	**0.7**	**1.9**	**0.7**	2.6	**0.7**	3.0	**0.7**	3.9
Anuran	7.1	3.8	7.8	4.6	8.8	5.3	9.0	6.1
XCB	0.8	2.1	0.8	2.9	0.8	3.5	0.8	4.5

os *embeddings* Perch e MFCCs, utilizando os algoritmos *HNSW* e *IMENN* e as métricas $T1$, $H1$, $H5$, $(t \pm \sigma)$, juntamente com a análise estatística.

Os resultados nas Tabelas 5.7 e 5.8 revelam que, neste cenário de generalização estrito, os *embeddings* do Perch demonstram uma capacidade de generalização significativamente superior à dos MFCCs. Conforme esperado, os MFCCs exibem o seu desempenho mais baixo aqui. No dataset *XCB* aplicando o filtro de ruido e *HNSW*, a precisão $H@5$ média dos MFCCs cai para 0.39, com uma alta variabilidade relativa CV de 4.2%, mostrado na Tabela 5.9. Esta queda evidencia a forte dependência dos MFCCs às condições da gravação, tornando a similaridade baseada neles pouco confiável entre gravações distintas.

Em contraste, o Perch demonstra uma notável capacidade de generalização, superando as expectativas. Surpreendentemente, em comparação com o Experimento 5.5.2, o desempenho médio do Perch com filtro e *HNSW* melhorou neste cenário mais desafiador para métricas chave como $H@5$, no dataset BC_{23}= $0.57para0.77$; XCB= $0.61para0.65$, comparando Tabelas 5.4 e 5.7). Embora se observe um aumento na variabilidade, ligeiramente maiores na Tabela 5.9 vs. Tabela 5.6), o bom desempenho geral ao dos MFCCs confirma que o modelo Perch aprendeu representações abstratas das vocalizações, permitindo identificar corretamente espécies mesmo em gravações

inéditas.

Com respeito ao tempo de consulta, observa-se um ligeiro aumento geral nos tempos médios e uma maior variabilidade neste experimento, comparando $t \pm \sigma$ nas tabelas dos Experimentos 5.5.1 e 5.5.3. Este fenómeno é esperado, pois a ausência de segmentos da mesma gravação na base de dados aumenta a distância média até os vizinhos corretos, forçando os algoritmos a explorar mais o espaço de busca. Mesmo assim, o *HNSW* mantém sua vantagem sobre o *IMENN* em termos de precisão e consistência. Para XCB com filtro e Perch, *HNSW* alcança $H@5 = 0.65$, CV de 2.8% contra $H@5 = 0.59$, CV de 3.7% do IMENN, evidenciado nas Tabelas 5.7 e 5.9), justificando sua escolha apesar de uma ligeira desvantagem no tempo médio neste caso específico ($t = 2.5$ ms vs $t = 2.3$ ms).

A aplicação do filtro de ruído, por sua vez, potencializou significativamente a precisão da recuperação em praticamente todos os cenários, como já observado nos experimentos anteriores. Por exemplo, no conjunto BC_{23}, utilizando Perch com *HNSW*, a precisão $T1$ média aumentou de 0.57 para 0.74 e o $H5$ médio de 0.66 para 0.77 com a aplicação do filtro (Tabela 5.7), com uma ligeira melhoria também na consistência com um CV de 2.9% para 2.7%.

Tabela 5.7. Resultados do Experimento 3 com Perch: Divisão de 70%-30% da gravação completa.

Sem aplicação do filtro de ruído								
	HNSW				**IMENN**			
	T1	**H**	**H5**	$t \pm \sigma$	**T1**	**H**	**H5**	$t \pm \sigma$
BC_{22}	0.30	0.32	**0.37**	2.1±3.7	0.29	0.35	0.37	2.4±5.9
BC_{23}	0.57	0.62	**0.66**	0.5±4.2	0.49	0.54	0.60	0.7±3.8
CBI	0.27	0.30	**0.33**	1.0±3.7	0.17	0.19	0.22	0.9±5.2
Anuran	0.88	0.94	**0.96**	0.1±2.5	0.71	0.85	0.89	0.1±3.74
XCB	0.45	0.49	**0.57**	3.0±3.8	0.37	0.41	0.45	2.9±4.1
Aplicando o filtro de ruído								
BC_{22}	0.32	0.35	**0.51**	2.0±3.3	0.33	0.36	0.41	2.2±2.6
BC_{23}	0.74	0.75	**0.77**	0.4±2.7	0.68	0.71	0.72	0.5±3.6
CBI	0.29	0.34	**0.36**	0.9±2.8	0.17	0.19	0.22	0.8±2.6
Anuran	0.93	0.96	**0.99**	0.1±2.2	0.83	0.85	0.89	0.1±2.2
XCB	0.53	0.62	**0.65**	2.5±3.3	0.40	0.46	0.59	2.3±3.7

Tabela 5.8. Resultados do Experimento 3 com MFCCs: Divisão de 70%-30% da gravação completa.

	HNSW				IMENN			
	T1	**H**	**H5**	$t \pm \sigma$	**T1**	**H**	**H5**	$t \pm \sigma$
Sem aplicação do filtro de ruído								
BC_{22}	0.18	0.21	**0.25**	2.5±4.1	0.17	0.19	0.23	2.4±4.5
BC_{23}	0.23	0.28	0.32	0.6±3.2	0.23	0.26	**0.33**	0.7±3.8
CBI	0.16	0.18	**0.21**	1.2±3.4	0.13	0.17	0.20	1.4±3.1
Anuran	0.53	0.64	**0.68**	0.1±3.4	0.49	0.56	0.62	0.1±3.8
XCB	0.23	0.25	**0.28**	3.3±4.5	0.22	0.24	0.26	3.4±2.6
Aplicando o filtro de ruído								
BC_{22}	0.22	0.25	**0.32**	2.3±2.9	0.19	0.24	0.27	2.5±3.1
BC_{23}	0.25	0.31	**0.36**	0.5±2.5	0.27	0.30	0.35	0.6±3.1
CBI	0.20	0.22	**0.24**	1.2±2.2	0.16	0.18	0.20	1.3±2.8
Anuran	0.59	0.65	**0.78**	0.1±2.2	0.54	0.61	0.68	0.1±2.7
XCB	0.31	0.33	**0.39**	2.6±2.7	0.26	0.29	0.31	2.8±2.2

Tabela 5.9. Análise estatística de métricas de precisão para o Experimento 3. A tabela apresenta os valores de Intervalo de Confiança (CI) de 95% e Coeficiente de Variação (CV) relatados como porcentagens (%).

	Sem aplicar o filtro Redução de ruído							
Dataset	**HNSW (Perch)**		**IMENN (Perch)**		**HNSW (MFCCs)**		**IMENN (MFCCs)**	
	IC	**CV**	**IC**	**CV**	**IC**	**CV**	**IC**	**CV**
BC_{22}	1.7	3.5	1.7	4.7	1.8	5.5	1.8	6.9
BC_{23}	1.5	2.9	1.5	3.8	1.6	4.3	1.6	5.6
CBI	1.5	3.1	1.5	4.1	1.6	4.6	1.6	6.0
Anuran	12.3	6.6	13.7	8.3	15.9	9.6	16.7	11.5
XCB	1.6	3.0	1.6	4.0	1.7	4.5	1.7	5.8
Aplicando o filtro Redução de ruído								
BC_{22}	1.6	3.3	1.6	4.4	1.7	5.1	1.7	6.4
BC_{23}	**1.4**	**2.7**	**1.4**	3.6	**1.5**	4.0	**1.5**	5.3
CBI	**1.4**	2.9	**1.4**	3.9	**1.5**	4.4	**1.5**	5.7
Anuran	11.9	6.3	13.2	8.0	15.3	9.2	16.2	11.0
XCB	1.5	2.8	1.5	3.7	1.6	4.2	1.6	5.4

A análise dos resultados obtidos nos Experimentos da Etpa 1 demonstra de forma consistente que a combinação de *embeddings* Perch, o uso do filtro de redução de ruído e o algoritmo de busca aproximada *HNSW* consolida-se como a metodologia mais eficiente para a recuperação de paisagens sonoras acusticamente semelhantes, como se evidencia na Figura 5.3. O protocolo do Experimento 5.5.2, utilizando o segmento de maior energia como consulta, oferece o melhor equilíbrio entre um desempenho de alta precisão e consistência num cenário realista, enquanto o Experimento 5.5.3 (divisão 70/30) serve como a prova da capacidade de generalização do modelo Perch,

que manteve seu desempenho mesmo sem sobreposição de dados entre consulta e base.

No Experimento 5.5.1, observou-se que a aplicação do filtro melhorou consistentemente os resultados quando se utilizam gravações completas como consultas, com destaque para os conjuntos BC_{23} e *XCB*, onde a precisão média aumentou consideravelmente e a variabilidade diminuiu. O Experimento 5.5.2 mostrou que selecionar o segmento com maior energia como consulta potencializa ainda mais os ganhos de precisão e, principalmente, de consistência, uma vez que esse trecho concentra os padrões acústicos mais representativos. Finalmente, o Experimento 5.5.3 evidenciou que o sistema Perch mais *HNSW* mantém alto desempenho mesmo em cenários de generalização estrita, validando a robustez do método em contraste com a queda acentuada observada nos MFCCs.

Em síntese, os achados da Etapa 1 não apenas validam partes cruciais da hipótese (PP2 e PP4), confirmando que a configuração proposta é eficaz e escalável, mas também estabelecem a sua superioridade metodológica sobre abordagens tradicionais baseadas em MFCCs. A configuração Perch mais *HNSW*, aliada à filtragem de ruído e consultas baseadas em segmentos de alta energia, demonstrou ser a mais performática em termos de equilíbrio entre precisão, consistência e eficiência. Esses resultados consolidam uma base metodológica sólida para os experimentos da Etapa 2, que se aprofundarão nas estratégias de fusão de características.

5.6 Avaliando a fusão de features

Nesta etapa, avaliamos sistematicamente o impacto da fusão de vetores de *features* na precisão da recuperação acústica, tanto para os *embeddings* quanto para os MFCCs. Para a realização dos experimentos desta etapa, foi utilizado o conjunto de dados `BirdCLEF+ 2025`. Nas próximas seções são apresentados os resultados e sues analise.

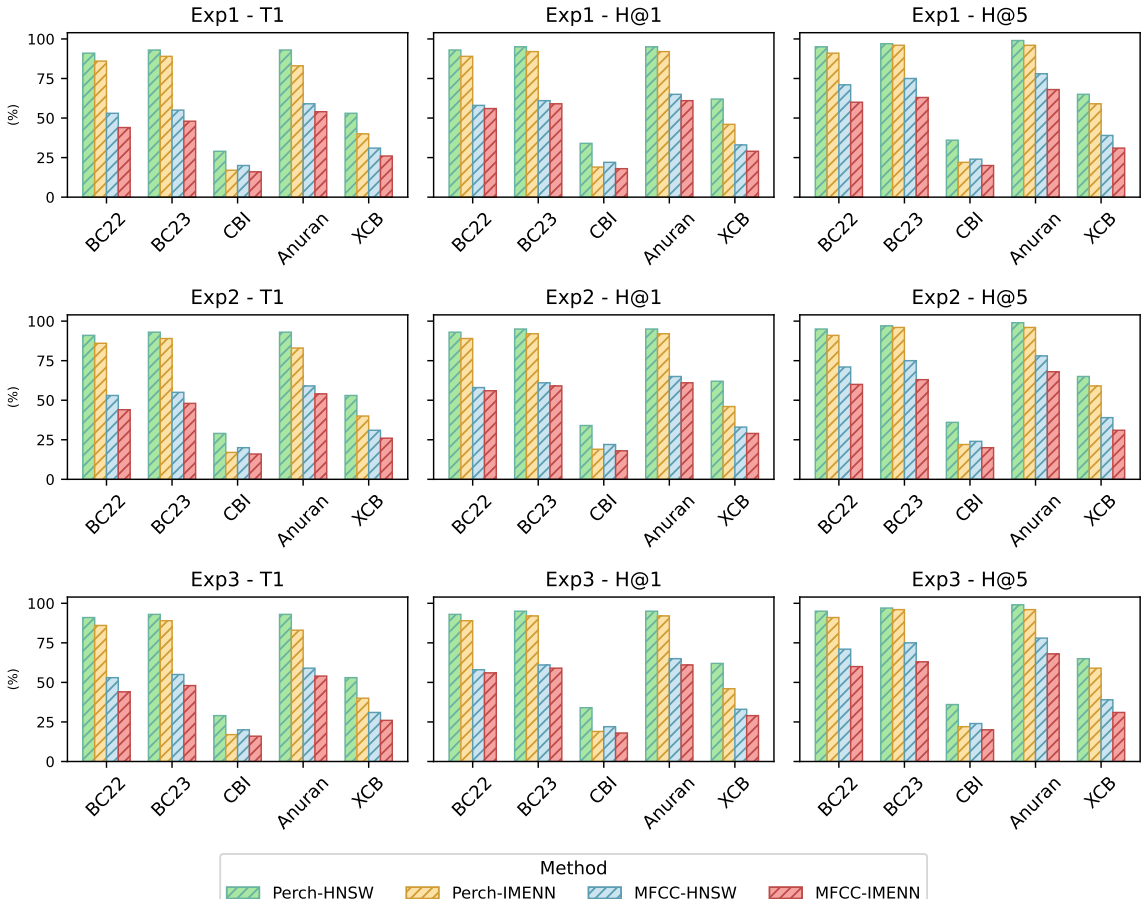


Figura 5.3. Comparação dos resultados obtidos nas três abordagens com diferentes combinações de técnicas de extração e algoritmos de busca.

5.6.1 Resultados do experimento 1

O primeiro experimento teve como objetivo avaliar o impacto das representações acústicas na tarefa de recuperação, comparando os *embeddings* extraídos do modelo Perch com os MFCCs, uma abordagem clássica amplamente adotada na literatura Lakdari et al. [2024], Winursito et al. [2018]. Para ambas as representações, foram testadas quatro técnicas de fusão de *features*: *Average Pooling*, *Weighted Average*, *Sum Pooling* e *Max Pooling*, avaliando-se o desempenho dos dois métodos de indexação vetorial *HNSW* e *IMENN* implementados no VectorDB. Os resultados quantitativos reportam as métricas $Hit@1$ e $Hit@5$ em nível de espécie, bem como o tempo médio de consulta (t) e seu desvio padrão ($\pm\sigma$). Complementarmente, são apresentadas as métricas de ordenação e as métricas estadísticas de resultados.

Os resultados detalhados nas Tabelas 5.10, 5.11 e 5.12, bem como na Figura 5.4, revelam desafios específicos impostos pela complexidade espectro-temporal das vocalizações de aves neste cenário de generalização.A análise confirma que os *embeddings* Perch possuem uma capacidade de abstração superior. Enquanto os MFCCs apresentaram uma alta variabilidade atingindo um Coeficiente de Variação $CV = 10.9\%$, indicando sensibilidade a fatores exógenos como ruído de fundo ou a distância do microfone, o Perch manteve uma consistência notável com um $CV = 6.5\%$. Esta estabilidade sugere que a arquitetura profunda captura características intrínsecas e invariantes das vocalizações, superando a rigidez das *features* clássicas, cuja aplicação prática neste cenário torna-se questionável dada a perda de informação inerente que prejudica a distinção entre cantos complexos.

A avaliação das técnicas de fusão responde à PP1, destacando o *Weighted Average Pooling* como a estratégia mais equilibrada, atingindo um $H@5_s$ de 0.42 e nas métricas de ordenação com um MRR de 0.26 e nDCG@5 de 0.29. Ao ponderar os segmentos pela energia RMS, esta técnica atua efetivamente como um mecanismo de atenção, enfatizando as sílabas proeminentes do canto e mitigando o impacto de intervalos de silêncio, o que é crucial para a sintaxe descontínua das aves. Em contraste, o fraco desempenho do *Max Pooling* com um $nDCG@5 = 0.24$ e $CV = 8.2\%$ destaca sua vulnerabilidade a ruídos impulsivos, que geram falsos picos de ativação, desviando a representação vetorial da identidade real da espécie.

A comparação entre os algoritmos de busca *HNSW* e *IMENN* informa a nossa quarta pergunta de pesquisa (PP4),consolidando o *HNSW* como uma boa escolha. Para vetores Perch fusionados, o *HNSW* superou o *IMENN* tanto em precisão alcançando um $H@5_s = 0.42$ contra 0.36 do *IMENN*, quanto vazão obtendo 58 QPS contra 23 QPS do *IMENN*. Isso indica que a estrutura de grafo hierárquico do *HNSW* navega com mais eficiência na complexa topologia do espaço de *embeddings* de 1280-d do que a quantização baseada em *clusters* do *IMENN*, oferecendo um bom *trade-off* entre latência e acurácia para monitoramento em larga escala.

Embora os MFCCs ofereçam uma vazão operacional superior, atingindo até 255 QPS com *Sum Pooling* e *HNSW*, essa vantagem torna-se irrelevante diante da sua fragilidade representacional neste cenário. A perda de informação inerente aos coeficientes cepstrais compromete a distinção de espécies com sintaxe complexa, tornando a sua aplicação prática inviável para tarefas de generalização, apesar da rapidez. Em suma, para Aves, a combinação Perch + *Weighted Average* + *HNSW* estabelece-se como a metodologia preferencial, validando a hipótese central do trabalho para este grupo.

Tabela 5.10. Resultados do Experimento 1 (divisão 70/30) para o grupo taxonômico Aves. Comparação de *embeddings* Perch e MFCCs. A tabela inclui métricas de precisão ($H@_s$), tempo médio de consulta ($t \pm \sigma$) e consultas por segundo (QPS).

Perch - Aves								
Técnica	**HNSW**				**IMENN**			
	H@1$_s$	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.28	0.38	5.76 ± 0.17	63	0.17	0.34	5.00 ± 3.37	21
Weighted Avg	0.31	0.42	6.21 ± 0.22	58	0.24	0.36	4.87 ± 3.41	23
Sum	0.27	0.37	5.12 ± 0.20	76	0.20	0.33	4.56 ± 3.62	25
Max Pool	0.25	0.31	1.33 ± 0.19	59	0.25	0.31	4.15 ± 3.65	22
MFCCs - Aves								
Avg Pool	0.17	0.34	2.78 ± 1.35	199	0.14	0.32	4.31 ± 6.86	125
Weighted Avg	0.21	0.39	3.04 ± 1.47	187	0.18	0.35	4.22 ± 6.90	121
Sum	0.19	0.36	2.56 ± 1.62	255	0.16	0.33	4.05 ± 6.79	143
Max Pool	0.13	0.29	4.55 ± 1.53	230	0.13	0.29	4.80 ± 4.22	111

Tabela 5.11. Resultados do Experimento 1 para o grupo taxonômico Aves: Métricas de ordenação de resultados utilizando *embeddings* e MFCCs com as diferentes técnicas de fusão de *features* nos algoritmos HNSW e IMENN.

Perch						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.24	0.24	0.27	0.24	0.23	0.26
Weighted Avg	0.26	0.26	0.29	0.26	0.25	0.28
Sum	0.24	0.24	0.27	0.24	0.23	0.26
Max Pool	0.21	0.21	0.24	0.21	0.20	0.23
MFCCs						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.20	0.19	0.22	0.20	0.20	0.22
Weighted Avg	0.22	0.21	0.24	0.21	0.21	0.24
Sum	0.13	0.13	0.15	0.14	0.14	0.16
Max Pool	0.18	0.17	0.20	0.18	0.18	0.20

Tabela 5.12. Análise estatística das métricas de precisão do Experimento 1 — Aves. A tabela apresenta o Intervalo de Confiança (IC) de 95% e o Coeficiente de Variação (CV), expressos em porcentagens (%).

Técnica	HNSW (Perch)		IMENN (Perch)		HNSW (MFCCs)		IMENN (MFCCs)	
	IC	CV	IC	CV	IC	CV	IC	CV
Avg Pool	2.0	7.2	2.3	8.0	3.1	11.6	3.4	12.4
Weighted Avg	1.8	6.5	2.1	7.3	2.9	10.9	3.2	11.7
Sum	1.9	6.8	2.2	7.6	3.0	11.2	3.3	12.0
Max Pool	2.4	8.2	2.6	8.8	3.5	12.8	3.7	13.6

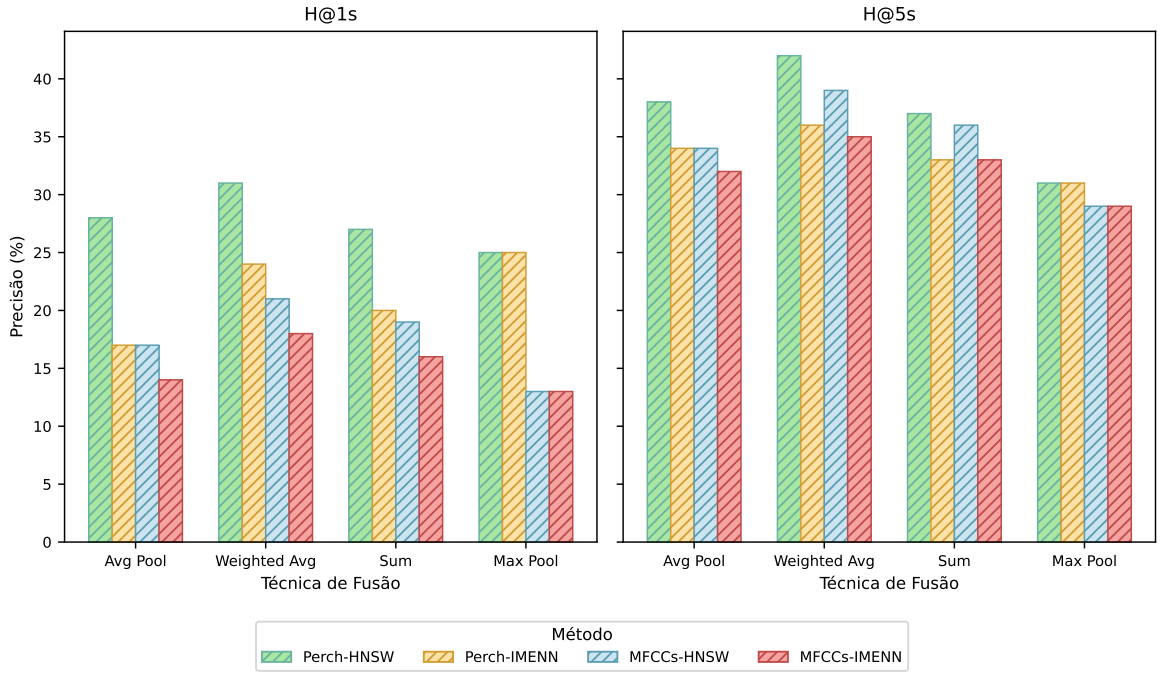


Figura 5.4. Resultados da comparação das técnicas de fusão e métodos para recuperação no nível de espécies para Aves.

Os resultados para Anuros, detalhados nas Tabelas 5.13, 5.14, 5.15 e Figura 5.5, destacam um desempenho marcadamente superior em comparação ao grupo Aves. Essa eficácia é uma implicação direta da baixa entropia do sinal de coaxar, que simplifica o problema de recuperação. A diferença entre as representações é ainda mais pronunciada. Com *HNSW* e *Weighted Average Pooling*, o Perch alcança um $H@5_s$ de 0.85 com uma consistência excelente com um CV de apenas 4.2%. Os MFCCs, em contraste, demonstram ser inadequados para este grupo, atingindo apenas $H@5_s$ de 0.57 e maior variabilidade com $CV = 6.3\%$. Este achado sustenta que, mesmo para vocalizações mais simples, a capacidade do Perch de extrair *features* discriminativas de vocalizações

supera largamente a capacidade de generalização dos coeficientes cepstrais, que podem confundir o sinal do anuro com outros ruídos ambientais constantes.

Para os Anuros, a avaliação das técnicas de fusão (PP1) revela uma nuance que se deve à natureza do sinal. Embora o *Weighted Average Pooling* proporcione a maior precisão no Top 5 com $H@5_s = 0.85$ e a melhor consistência $CV = 4.2\%$, as técnicas *Average Pooling* e *Sum Pooling* mostraram-se ligeiramente superiores nas métricas de Top 1 com um $H@1_s$ de 0.81 e 0.78 e de ordenação com $MRR = 0.78$; $nDCG@5 = 0.78$. Este comportamento sugere que, devido à uniformidade espectral dos coaxares (onde a informação é repetida em cada pulso), *Average Pooling* ou *Sum Pooling* já é suficiente para capturar a essência do sinal e obter o match correto no primeiro resultado. A ponderação do *Weighted Average Pooling*, embora refinada, oferece um ganho marginal que não compensa a complexidade de cálculo neste contexto de alta redundância. Contudo, o *Weighted Average Pooling* deve ser mantido como a escolha mais eficaz devido à sua superioridade no Top 5 e na menor dispersão dos resultados.

A comparação entre *HNSW* e *IMENN* (PP4) para Perch com *Weighted Average* reafirma a superioridade do grafo hierárquico. O *HNSW* alcança $H@5_s$ de 0.85 e $nDCG@5$ de 0.77, significativamente melhores que o *IMENN* com $H@5_s = 0.62$ e $nDCG@5 = 0.54$. Além disso, o *HNSW* oferece uma vazão maior com 78 QPS e *IMENN* 55 QPS. Essa discrepância indica que, mesmo em um espaço vetorial teoricamente mais "organizado"pela simplicidade do sinal de anuro, o *HNSW* demonstra uma melhor capacidade de navegar e encontrar os vizinhos mais próximos.

É vital ressaltar que os MFCCs alcançaram a maior vazão de consultas com 276QPS com *HNSW* e *Sum Pooling*. No entanto, a penalidade na precisão com $H@5_s = 0.53$ e na qualidade de ordenação com $nDCG@5 = 0.51$ é excessiva. A capacidade dos *embeddings* Perch de isolar as características distintivas dos coaxares justifica plenamente o uso desta representação, garantindo que o sistema seja preciso, mesmo que a vazão de consultas seja inferior. Para a recuperação de gravações de Anuros, a combinação de Perch + *Weighted Average* + *HNSW* estabelece um padrão

de eficácia e consistência elevadas.

Tabela 5.13. Resultados do Experimento 1 (divisão 70/30) para o grupo taxonômico Anuros. Comparação de *embeddings* Perch e MFCCs. A tabela inclui métricas de precisão ($H@_s$), tempo médio de consulta ($t \pm \sigma$) e consultas por segundo (QPS).

Perch - Anuros								
Técnica	**HNSW**				**IMENN**			
	H@1$_s$	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.81	0.83	2.32 ± 1.03	100	0.41	0.44	2.17 ± 3.10	89
Weighted Avg	0.76	0.85	3.51 ± 2.62	78	0.44	0.62	4.16 ± 3.27	55
Sum	0.78	0.80	3.07 ± 2.12	92	0.45	0.47	4.11 ± 4.18	48
Max Pool	0.76	0.81	3.13 ± 2.38	68	0.48	0.56	5.21 ± 3.90	36
MFCCs - Anuros								
Avg Pool	0.45	0.55	2.56 ± 3.25	220	0.37	0.42	3.87 ± 4.51	195
Weighted Avg	0.36	0.57	3.46 ± 3.86	198	0.38	0.51	6.78 ± 5.02	174
Sum	0.45	0.53	4.02 ± 4.19	276	0.32	0.44	5.78 ± 4.90	153
Max Pool	0.41	0.54	3.64 ± 4.58	250	0.36	0.52	5.20 ± 4.78	231

Tabela 5.14. Resultados do Experimento 1: Métricas de ordenação de resultados para o grupo taxonômico Anuros, utilizando *embeddings* e MFCCs com as diferentes técnicas de fusão de *features* nos algoritmos HNSW e IMENN.

Perch - Anuros						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.78	0.75	0.78	0.41	0.42	0.41
Weighted Avg	0.76	0.75	0.77	0.56	0.51	0.54
Sum	0.78	0.75	0.78	0.44	0.41	0.45
Max Pool	0.75	0.74	0.76	0.51	0.52	0.54
MFCCs - Anuros						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.50	0.49	0.51	0.38	0.38	0.39
Weighted Avg	0.43	0.42	0.44	0.45	0.43	0.47
Sum	0.50	0.49	0.51	0.36	0.35	0.37
Max Pool	0.43	0.42	0.44	0.40	0.39	0.42

Tabela 5.15. Análise estatística das métricas de precisão do Experimento 1 — Anuros. A tabela apresenta o Intervalo de Confiança (IC) de 95% e o Coeficiente de Variação (CV), expressos em porcentagens (%).

Técnica	**HNSW (Perch)**		**IMENN (Perch)**		**HNSW (MFCCs)**		**IMENN (MFCCs)**	
	IC	**CV**	**IC**	**CV**	**IC**	**CV**	**IC**	**CV**
Avg Pool	1.4	4.8	1.5	5.2	2.0	6.8	2.2	7.4
Weighted Avg	1.2	4.2	1.3	4.7	1.9	6.3	2.0	6.9
Sum	1.3	4.5	1.4	4.9	2.0	6.6	2.1	7.2
Max Pool	1.6	5.4	1.7	5.8	2.3	7.4	2.5	8.0

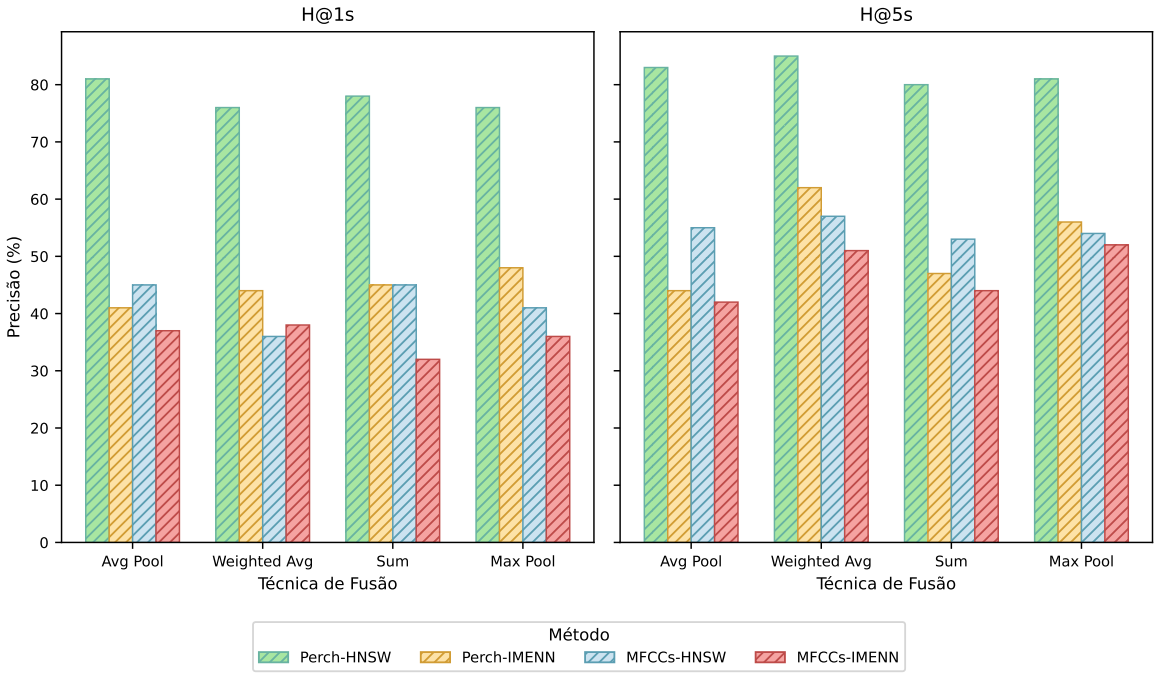


Figura 5.5. Resultados da comparação das técnicas de fusão e métodos para recuperação no nível de espécies para Anuros.

O desempenho para o grupo Mamíferos, cujos resultados estão nas Tabelas 5.16, 5.17 e 5.18 e Figura 5.6, posiciona-se em um nível intermediário entre os Anuros e as Aves. No entanto, o padrão de superioridade do Perch e do *HNSW* é mantido. A principal vantagem dos *embeddings* Perch reside na confiabilidade estatística. O Perch alcançou $H@5_s$ de 0.47 e uma boa consistência com um CV de 8.0%, enquanto os MFCCs ficaram significativamente atrás com um $H@5_s$ de 0.38 e um CV de 10.8%. Essa diferença na variabilidade é crucial para os Mamíferos, cujas vocalizações cobrem um amplo espectro de frequências tanto graves como agudas e exigem uma representação capaz de generalizar através dessa heterogeneidade sem sucumbir ao ruído de fundo que frequentemente afeta as gravações de campo. A análise estatística na Tabela 5.18 confirma que a representação Perch é fundamentalmente mais confiável para este táxon.

O *Weighted Average Pooling* emerge como a melhor estratégia geral, liderando nas métricas de ordenação com um $MRR = 0.37$ e $nDCG@5 = 0.41$ e na precisão com

um $H@5_s = 0.47$ com a menor dispersão com um $CV = 8.0\%$. A ponderação pela energia é essencial para Mamíferos, pois garante que as vocalizações de alta amplitude (como rugidos, latidos e chamados) dominem o vetor agregado, rejeitando o ruído de baixa energia. O desempenho inferior de outras técnicas, como *Average Pooling* e *Sum Pooling*, indica que a diluição do sinal *Average Pooling* ou a sensibilidade à duração *Sum Pooling* comprometem a captura da *feature* saliente (PP1).

A comparação entre *HNSW* e *IMENN* (PP4) continua a favorecer a abordagem de busca aproximada do *HNSW*. Este alcança maior precisão e consistência com um $H@5_s = 0.47$ e um $CV = 8.0\%$, superando o *IMENN* com um $H@5_s$ de apenas 0.31 e $CV = 8.6\%$. Em termos de eficiência, o *HNSW* manteve uma vazão ligeiramente superior alcançando 56 QPS e *IMENN* 49 QPS, garantindo o melhor *trade-off* entre qualidade e velocidade. Em suma, para o grupo Mamíferos no cenário de generalização, a metodologia composta por Perch + *Weighted Average Pooling* + *HNSW* demonstra ser a mais eficaz e consistente, ratificando a premissa de que a consistência taxonômica deve ser priorizada sobre a latência mínima em tarefas de monitoramento ambiental.

Tabela 5.16. Resultados do Experimento 1 (divisão 70/30) para o grupo taxonômico Mamíferos. Comparação de *embeddings* Perch e MFCCs. A tabela inclui métricas de precisão ($H@_s$), tempo médio de consulta ($t \pm \sigma$) e consultas por segundo (QPS).

Perch - Mamíferos								
Técnica	**HNSW**				**IMENN**			
	H@1$_s$	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.39	0.44	3.50 ± 2.45	75	0.29	0.36	3.79 ± 4.02	68
Weighted Avg	0.41	0.47	3.48 ± 3.14	56	0.27	0.31	4.60 ± 4.71	49
Sum	0.40	0.45	2.98 ± 3.56	64	0.23	0.31	3.71 ± 5.00	51
Max Pool	0.41	0.43	3.15 ± 4.31	81	0.22	0.27	4.14 ± 4.88	80
MFCCs - Mamíferos								
Avg Pool	0.18	0.35	3.31 ± 4.80	204	0.15	0.33	5.78 ± 5.30	145
Weighted Avg	0.26	0.38	3.23 ± 4.98	164	0.18	0.28	5.49 ± 5.12	126
Sum	0.18	0.35	3.98 ± 5.14	207	0.12	0.34	6.40 ± 6.26	158
Max Pool	0.19	0.27	3.21 ± 4.03	211	0.16	0.19	4.67 ± 4.47	186

Tabela 5.17. Resultados do Experimento 1: Métricas de ordenação de resultados para o grupo taxonômico Mamíferos, utilizando *embeddings* e MFCCs com as diferentes técnicas de fusão de *features* nos algoritmos HNSW e IMENN.

Perch - Mamíferos						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.36	0.33	0.37	0.25	0.23	0.28
Weighted Avg	0.37	0.35	0.41	0.34	0.31	0.32
Sum	0.36	0.33	0.37	0.38	0.31	0.34
Max Pool	0.31	0.29	0.34	0.21	0.22	0.23
MFCCs - Mamíferos						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.26	0.24	0.27	0.24	0.21	0.26
Weighted Avg	0.32	0.31	0.32	0.22	0.21	0.22
Sum	0.26	0.24	0.27	0.20	0.18	0.22
Max Pool	0.22	0.21	0.23	0.18	0.17	0.18

Tabela 5.18. Análise estatística das métricas de precisão do Experimento 1 — Mamíferos. A tabela apresenta o Intervalo de Confiança (IC) de 95% e o Coeficiente de Variação (CV), expressos em porcentagens (%).

Técnica	**HNSW (Perch)**		**IMENN (Perch)**		**HNSW (MFCCs)**		**IMENN (MFCCs)**	
	IC	**CV**	**IC**	**CV**	**IC**	**CV**	**IC**	**CV**
Avg Pool	2.3	8.6	2.6	9.1	3.2	11.2	3.5	12.0
Weighted Avg	2.0	8.0	2.3	8.6	3.0	10.8	3.3	11.6
Sum	2.1	8.2	2.4	8.8	3.1	11.0	3.4	11.8
Max Pool	2.5	9.1	2.8	9.7	3.4	11.7	3.7	12.6

5.6.2 Resultados do experimento 2

O objetivo do segundo experimento é avaliar qual técnica de fusão de vetores de *features* é mais eficaz para recuperar segmentos relevantes. Para isso, a abordagem foi diferenciada conforme o nível de recuperação. Para a recuperação em nível de gravação ($H@1_r$), a fusão foi aplicada para criar um único vetor representativo para cada gravação na base de dados, utilizando como consulta o vetor não fusionado do segmento de maior energia. Já para a recuperação em nível de espécie ($H@1_s, H@5_s$), a fusão foi utilizada para criar uma consulta mais representativa, os vetores dos segmentos de maior energia de diversas gravações de uma mesma espécie foram agregados e fusionados, gerando um único vetor de consulta que representa a essência acústica da espécie.

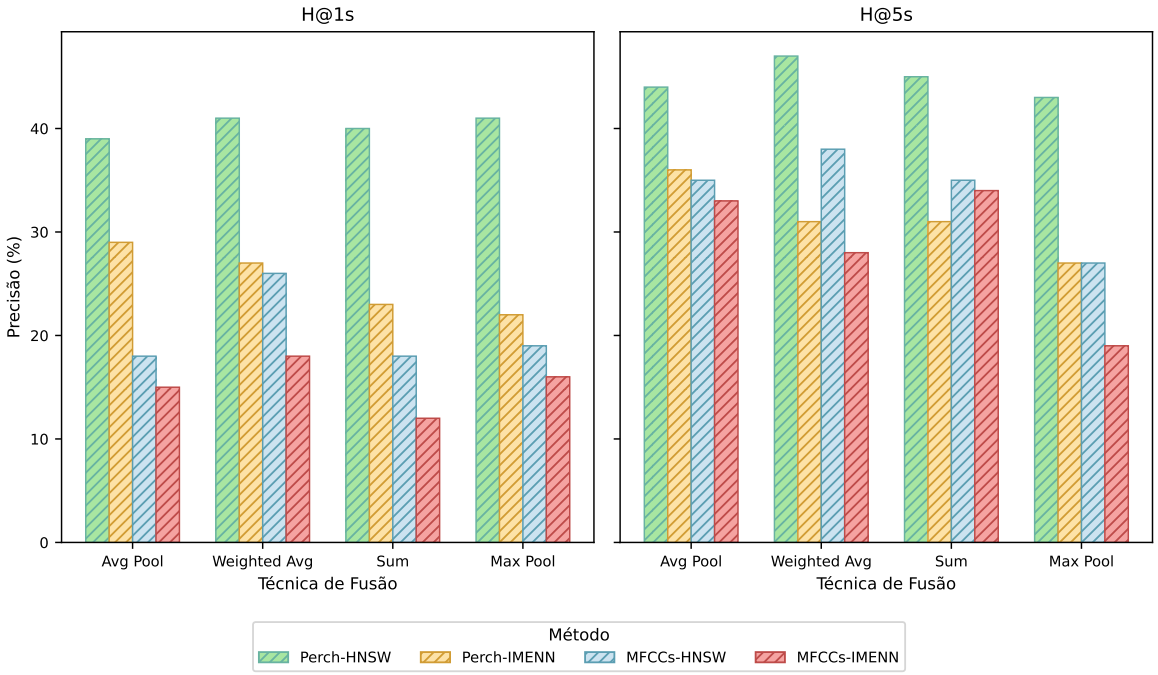


Figura 5.6. Resultados da comparação das técnicas de fusão e métodos para recuperação no nível de espécies para Mamíferos.

A hipótese subjacente é que, ao utilizar uma consulta de alta qualidade, a precisão da recuperação aumentará, melhorando as métricas de ordenação.

Neste cenário, onde a consulta é formada a partir do segmento de maior energia, representando um evento acústico proeminente e comparada com vetores da base de dados fusionados por gravação, a análise para Aves apresentados nas Tabelas 5.19, 5.20, 5.21; Figura 5.7, revela uma validação categórica da pergunta PP2. A primeira e mais crítica observação é o aumento expressivo na precisão em comparação com o Experimento 1. A precisão ao nível da espécie alcança $H@1_s$ de 0.94 e $H@5_s$ de 0.97, um salto considerável em relação aos 0.31 e 0.42 obtidos no cenário anterior. Esta melhoria drástica confirma que a principal limitação do cenário anterior não residia na representação vetorial, mas sim na incerteza introduzida pela consulta de baixa qualidade. Ao filtrar a consulta para o segmento mais claro, reduz-se o ruído do processo de *matching* e aumenta-se a consistência, com o CV a diminuir para 5.8%.

O *Weighted Average Pooling* consolida-se como a estratégia superior, alcançando o

melhor desempenho em todas as métricas de precisão com um $H@5_s = 0.97$ e ordenação $nDCG@5 = 0.86$. O sucesso reside no alinhamento estabelecido; a consulta encontra na base de dados um vetor que foi ponderado para dar máxima ênfase a esse mesmo segmento durante a fusão. Este *match* direcionado minimiza a distância no espaço latente entre o alvo e a consulta, especialmente para as vocalizações não estacionárias das aves. A queda de desempenho observada no *Max Pooling* e *Sum Pooling* reforça a fragilidade destas técnicas perante qualquer resquício de ruído remanescente.

Embora os MFCCs também apresentem uma melhoria notável alcançando um$H@5_s > 0.90$, a diferença na consistência é a chave discriminatória. O Perch com um $CV = 5.8\%$ ainda supera os MFCCs com um $CV = 11.5\%$, principalmente em termos de consistência. Isto reforça que, mesmo em condições onde a consulta e o alvo são de alta qualidade acústica, o Perch oferece uma representação mais estável e discriminativa para as aves, pois seus *features* são menos suscetíveis a variações de frequência que podem ainda confundir os MFCCs.

Na comparação dos algoritmos de busca (PP4), *HNSW* mantém a vantagem sobre *IMENN* devido à qualidade da ordenação com um $nDCG@5$ de 0.86 vs 0.76 do *IMENN*. Em termos de eficiência, ambos os algoritmos apresentaram vazões elevadas e semelhantes com 136 QPS para *HNSW* e 131 QPS para *IMENN*. Dada a superioridade em precisão e qualidade de *ranking*, o *HNSW* continua a ser a melhor escolha, validando que a sua capacidade de navegar na topologia vetorial com fidelidade é mais crítica do que uma pequena diferença de latência.

Em suma, os resultados para Aves demonstram o alto potencial da abordagem quando se utiliza uma consulta representativa e de alta qualidade. A combinação Perch + *Weighted Average* + *HNSW* atinge níveis de precisão e ordenação muito elevados, valida a aplicabilidade desta abordagem para tarefas de curadoria de dados e busca exploratória em bases de dados de grande escala, onde a recuperação precisa de exemplos é essencial.

Tabela 5.19. Resultados do Experimento 2: consulta com segmento de maior energia com comparação entre *embeddings* Perch e MFCCs para o grupo taxonômico Aves. A tabela inclui a vazão do sistema em Consultas por Segundo (QPS) para a recuperação ao nível da gravação ($H@1_r$) e da espécie ($H@1_s, H@5_s$).

Perch - Aves										
Técnica	**HNSW**					**IMENN**				
	H@1$_r$	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_r$**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.83	0.82	0.88	5.81 ± 2.50	156	0.73	0.71	0.78	5.98 ± 3.22	165
Weighted Avg	0.93	0.94	0.97	3.69 ± 2.36	136	0.81	0.81	0.88	4.78 ± 4.75	131
Sum	0.80	0.76	0.84	3.63 ± 3.18	118	0.36	0.37	0.42	3.78 ± 4.81	85
Max Pool	0.71	0.56	0.78	3.80 ± 3.10	114	0.67	0.63	0.71	4.86 ± 5.20	122
MFCCs - Aves										
Avg Pool	0.79	0.76	0.84	7.46 ± 4.61	125	0.70	0.65	0.75	5.69 ± 5.44	245
Weighted Avg	0.87	0.85	0.91	4.78 ± 4.50	129	0.81	0.76	0.84	6.98 ± 5.89	234
Sum	0.67	0.66	0.71	5.32 ± 5.13	130	0.31	0.26	0.35	6.30 ± 6.33	261
Max Pool	0.61	0.52	0.66	4.29 ± 5.61	128	0.62	0.55	0.67	6.51 ± 6.02	247

Tabela 5.20. Resultados do Experimento 2: Análise da qualidade de ordenação para consultas baseadas nos segmentos de maior energia para o grupo taxonômico Aves. São comparados os valores MRR, mAP@5 e nDCG@5 para as diferentes técnicas de fusão.

Perch - Aves						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.78	0.77	0.79	0.75	0.74	0.76
Weighted Avg	0.86	0.85	0.86	0.76	0.75	0.76
Sum	0.78	0.77	0.79	0.52	0.52	0.53
Max Pool	0.61	0.60	0.61	0.54	0.53	0.54
MFCCs - Aves						
Avg Pool	0.68	0.66	0.68	0.78	0.75	0.78
Weighted Avg	0.80	0.77	0.80	0.89	0.85	0.88
Sum	0.25	0.24	0.25	0.78	0.75	0.78
Max Pool	0.56	0.54	0.56	0.54	0.52	0.55

Tabela 5.21. Análise estatística das métricas de precisão do Experimento 2 — Aves. A tabela apresenta o Intervalo de Confiança (IC) de 95% e o Coeficiente de Variação (CV), expressos em porcentagens (%).

Técnica	**HNSW (Perch)**		**IMENN (Perch)**		**HNSW (MFCCs)**		**IMENN (MFCCs)**	
	IC	**CV**	**IC**	**CV**	**IC**	**CV**	**IC**	**CV**
Avg Pool	1.8	6.9	2.1	7.4	3.4	12.3	3.2	13.1
Weighted Avg	1.5	5.8	1.8	6.3	3.1	11.5	3.3	12.8
Sum	1.6	6.2	1.9	6.9	3.3	11.9	3.1	12.5
Max Pool	2.0	7.8	2.2	8.1	3.8	13.5	3.5	14.0

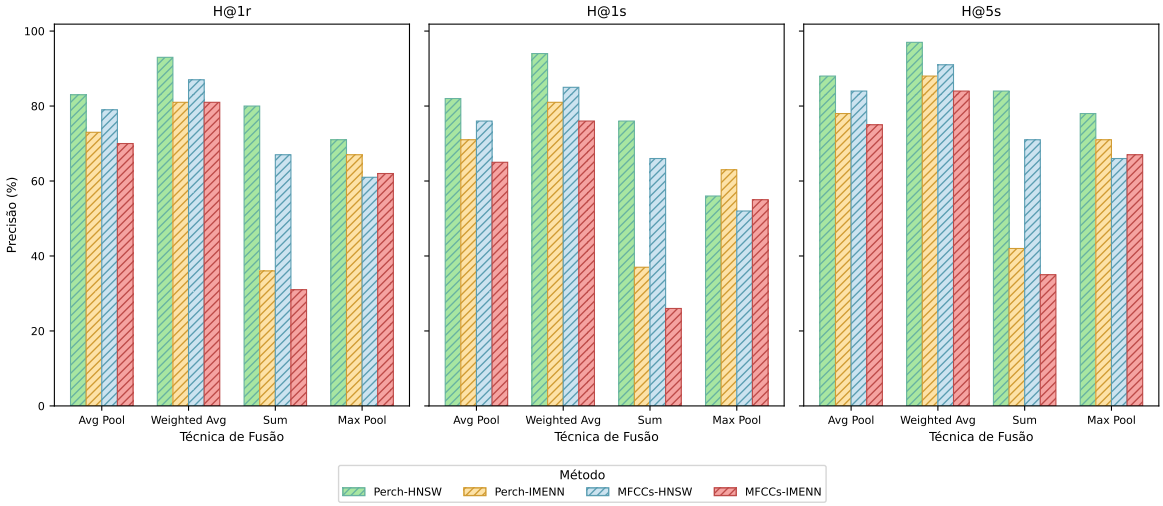


Figura 5.7. Comparativo visual do desempenho das técnicas de fusão no cenário de consulta com os segmentos de maior energia para Aves.

O cenário de consulta com evento proeminente para o grupo Anuros revela um desempenho excepcionalmente alto, detalhado nas Tabelas 5.22, 5.23, 5.24 e Figura 5.8. Essa eficácia isso ocorre porque suas vocalizações são uniformes e repetitivas, onde o segmento de maior energia é altamente representativo do som da espécie, validando fortemente a hipótese relacionada à PP2. Com a configuração Perch + *HNSW* + *Weighted Average Pooling*, o sistema alcançou níveis de precisão muitos bons $H@1_s = 0.99$ e $H@5_s = 0.99$, com métricas de ordenação igualmente elevadas $MRR = 0.98$ e $nDCG@5 = 0.96$. A notável consistência com um CV de apenas 4.0% reforça que a consulta refinada removeu a incerteza residual do Experimento 1, alinhando o vetor da *query* perfeitamente com a representação na base de dados.

Embora os MFCCs tenham apresentado uma melhoria substancial neste cenário, atingindo $H@5_s$ de 0.91, sua variabilidade $CV = 6.0\%$ permaneceu visivelmente maior que a do Perch $CV = 4.0\%$. Este achado é crucial; mesmo que a consulta seja ideal, o Perch oferece uma representação intrinsecamente mais confiável e menos suscetível a variações subtis de frequência, reforçando a superioridade do modelo Perch para isolar as *features* discriminativas dos coaxares.

No que concerne às técnicas de fusão (PP1) o *Weighted Average Pooling*

estabeleceu-se como a estratégia claramente superior, alcançando a menor variabilidade e as melhores métricas de precisão e ordenação. A degradação observada em outras fusões, como o *Sum Pooling* e o *Max Pooling*, particularmente ao usar o *IMENN*, sugere que, mesmo que o sinal seja repetitivo, a ponderação pela energia ainda é essencial para garantir que o vetor final não seja contaminado por resíduos de ruído nos segmentos secundários, otimizando o alinhamento com a consulta proeminente.

A comparação entre os algoritmos de busca revela um *trade-off* de latência acentuado e específico para este grupo (PP4). O *HNSW* demonstrou uma qualidade de recuperação superior com $H@5_s = 0.99$, $nDCG@5 = 0.96$, mas processou apenas 121 QPS. Em contrapartida, o *IMENN* foi mais rápido 162 QPS), mas sacrificou a precisão $H@5_s = 0.92$. Dada a diferença expressiva na precisão, o *HNSW* permanece a escolha preferencial. A decisão reflete um ganho em qualidade taxonômica justifica a latência ligeiramente maior, garantindo a máxima confiança na identificação da espécie. Atingir uma precisão de 0.99 valida a aplicabilidade desta metodologia para a criação de um sistema de referência ou para a automação de inventários de biodiversidade onde a identificação de anuros é crítica.

Tabela 5.22. Resultados do Experimento 2: consulta com segmento de maior energia com comparação entre *embeddings* Perch e MFCCs para o grupo taxonômico Anuros. A tabela inclui a vazão do sistema em Consultas por Segundo (QPS) para a recuperação ao nível da gravação ($H@1_r$) e da espécie ($H@1_s, H@5_s$).

Perch - Anuros										
Técnica	**HNSW**					**IMENN**				
	H@1$_r$	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_r$**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.89	0.91	0.93	3.88 ± 3.72	75	0.83	0.82	0.87	5.71 ± 5.85	158
Weighted Avg	0.95	0.99	0.99	3.76 ± 3.01	121	0.86	0.91	0.92	4.31 ± 6.21	162
Sum	0.91	0.96	0.97	4.22 ± 4.15	81	0.45	0.42	0.51	8.16 ± 7.12	159
Max Pool	0.93	0.94	0.96	4.71 ± 3.21	90	0.76	0.73	0.81	8.18 ± 5.91	150
MFCCs - Anuros										
Técnica	**HNSW**					**IMENN**				
	H@1$_r$	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_r$**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.84	0.85	0.89	5.30 ± 4.31	127	0.55	0.58	0.58	5.95 ± 6.02	247
Weighted Avg	0.88	0.92	0.91	4.53 ± 4.96	123	0.52	0.59	0.59	5.87 ± 8.15	238
Sum	0.86	0.85	0.89	4.57 ± 6.40	146	0.32	0.39	0.39	6.78 ± 8.43	258
Max Pool	0.78	0.76	0.82	5.53 ± 5.90	135	0.39	0.46	0.46	6.46 ± 6.29	245

Tabela 5.23. Resultados do Experimento 2: Análise da qualidade de ordenação para consultas baseadas nos segmentos de maior energia para o grupo taxonômico Anuros. São comparados os valores MRR, mAP@5 e nDCG@5 para as diferentes técnicas de fusão.

Perch - Anuros						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.94	0.92	0.93	0.84	0.83	0.84
Weighted Avg	0.98	0.96	0.96	0.90	0.88	0.91
Sum	0.94	0.92	0.93	0.46	0.45	0.46
Max Pool	0.94	0.92	0.94	0.76	0.76	0.77
MFCCs - Anuros						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.73	0.72	0.74	0.67	0.66	0.69
Weighted Avg	0.73	0.72	0.74	0.76	0.73	0.77
Sum	0.73	0.72	0.74	0.64	0.62	0.65
Max Pool	0.67	0.66	0.68	0.67	0.66	0.70

Tabela 5.24. Análise estatística das métricas de precisão do Experimento 2 — Anuros. A tabela apresenta o Intervalo de Confiança (IC) de 95% e o Coeficiente de Variação (CV), expressos em porcentagens (%).

Técnica	**HNSW (Perch)**		**IMENN (Perch)**		**HNSW (MFCCs)**		**IMENN (MFCCs)**	
	IC	**CV**	**IC**	**CV**	**IC**	**CV**	**IC**	**CV**
Avg Pool	1.3	4.4	1.4	4.8	1.9	6.3	2.2	7.0
Weighted Avg	1.1	4.0	1.2	4.3	1.8	6.0	2.0	6.5
Sum	1.2	4.2	1.3	4.6	1.9	6.2	2.1	6.8
Max Pool	1.5	5.1	1.6	5.5	2.3	7.2	2.5	7.9

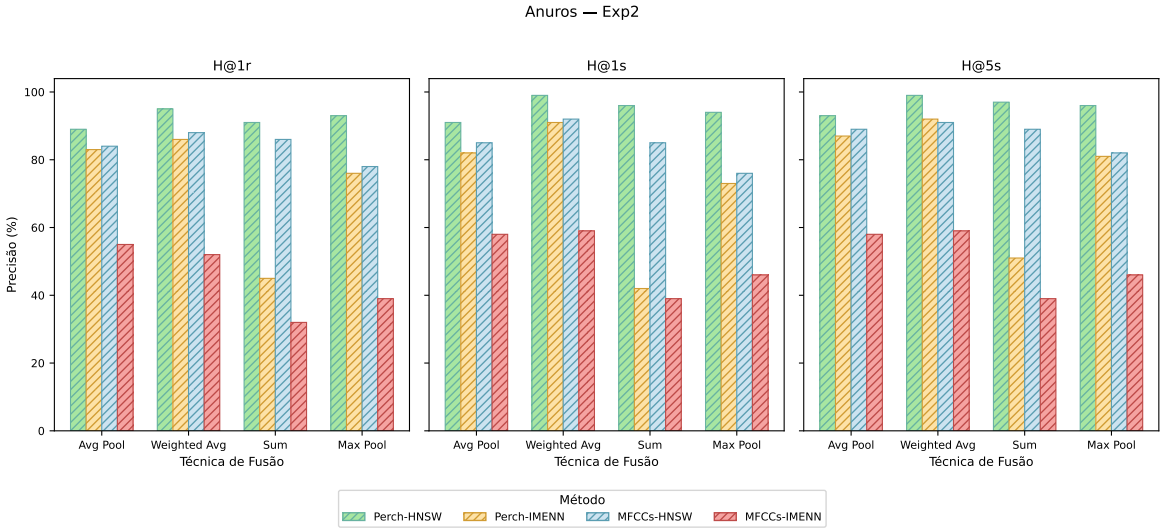


Figura 5.8. Comparativo visual do desempenho das técnicas de fusão no cenário de consulta com os segmentos de maior energia para Anuros.

A análise para o grupo Mamíferos, fundamentada nos resultados das Tabelas 5.25, 5.26 e 5.27, bem como na Figura 5.9, corrobora o impacto positivo da qualidade da consulta (PP2), embora apresente nuances estatísticas distintas dos outros grupos. Utilizando a configuração Perch + *HNSW* + *Weighted Average Pooling*, a precisão ao nível da espécie ($H@5_s$) atinge 0.93, um aumento substancial em relação aos 0.47 do cenário de generalização do Experimento 1. As métricas de ordenação também são excelentes, com MRR de 0.93 e nDCG@5 de 0.93, indicando que os resultados relevantes aparecem consistentemente no topo do *ranking*.

Curiosamente, a variabilidade relativa aumentou ligeiramente de $CV = 8.0\%$ no Experimento 1 para 8.7% aqui. Este comportamento sugere que, embora a média seja excelente, o critério de maior energia para mamíferos pode ser ocasionalmente falível. Diferente dos anuros, o ambiente de gravação de mamíferos contém frequentemente ruídos impulsivos de alta energia que podem ser selecionados erroneamente como consulta, introduzindo *outliers* que aumentam a dispersão estatística, mesmo que a precisão global seja alta.

A comparação entre representações revela que, sob condições ideais, a lacuna de desempenho diminui, mas a confiabilidade diverge. Os MFCCs atingiram um $H@5_s$ respeitável de 0.90, provando que conseguem realizar o *matching* espectral quando o sinal é limpo. Contudo, o Perch com $H@5_s = 0.93$ manteve-se superior na consistência com um $CV = 8.7\%$ contra 11.5% dos MFCCs. Isso reitera que a vantagem de Perch não é apenas atingir o pico de acurácia, mas garantir uma estabilidade operacional que as *features* clássicas não conseguem oferecer perante a heterogeneidade das vocalizações de mamíferos.

O *Weighted Average Pooling* reafirma-se como a estratégia mais eficaz, liderando em todas as métricas de recuperação e ordenação. A ponderação pela energia RMS atua como um filtro essencial para alinhar a consulta com a base de dados. Ao enfatizar os segmentos de alta amplitude, o *Weighted Average Pooling* minimiza a influência do ruído de fundo terrestre, típico deste táxon. Vale notar que o *Sum Pooling* apresentou

um CV de 9.0%, sendo, portanto, inferior ao *Weighted Average Pooling* com um CV de 8.7%, o que consolida a média ponderada como a técnica que oferece o melhor equilíbrio entre precisão e consistência estatística.

Na análise dos algoritmos de busca (PP4) com Perch e *Weighted Average*, *HNSW* oferece uma vantagem significativa em qualidade sobre o *IMENN*, apesar de uma ligeira desvantagem em vazão. *HNSW* alcança $H@5_s$ de 0.93 e nDCG@5 de 0.93 com CV de 8.7%, enquanto *IMENN* obtém $H@5_s$ de 0.85 e nDCG@5 de 0.89 com CV de 9.1%. No entanto, *IMENN* processa 141 QPS contra 132 QPS do *HNSW*. Dada a diferença notável na precisão e qualidade de ordenação, *HNSW* valida-se como a escolha recomendada, garantindo que a recuperação de espécies críticas não seja comprometida por uma economia marginal de tempo.

Para o grupo Mamíferos, a metodologia Perch + *Weighted Average Pooling* + *HNSW* atinge um patamar de recuperação de alta fidelidade. Os resultados confirmam a eficácia de utilizar consultas baseadas em eventos proeminentes para localizar gravações relevantes, estabelecendo um protocolo eficaz para a triagem de dados, mesmo considerando a ligeira sensibilidade a ruídos ambientais refletida no coeficiente de variação

Tabela 5.25. Resultados do Experimento 2: consulta com segmento de maior energia com comparação entre *embeddings* Perch e MFCCs para o grupo taxonômico Mamíferos. A tabela inclui a vazão do sistema em Consultas por Segundo (QPS) para a recuperação ao nível da gravação ($H@1_r$) e da espécie ($H@1_s, H@5_s$).

Perch - Mamíferos										
Técnica	**HNSW**					**IMENN**				
	H@1$_r$	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_r$**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.86	0.83	0.88	5.12 ± 2.44	143	0.78	0.74	0.80	6.10 ± 3.11	158
Weighted Avg	0.91	0.89	0.93	4.08 ± 2.20	132	0.82	0.79	0.85	5.02 ± 4.20	141
Sum	0.84	0.82	0.86	3.77 ± 2.86	126	0.45	0.42	0.49	4.21 ± 4.95	92
Max Pool	0.74	0.70	0.76	3.69 ± 3.02	118	0.69	0.65	0.72	5.10 ± 5.12	124
MFCCs - Mamíferos										
Avg Pool	0.80	0.76	0.83	7.01 ± 4.38	119	0.72	0.69	0.76	5.44 ± 5.11	235
Weighted Avg	0.87	0.84	0.90	5.02 ± 4.09	122	0.79	0.75	0.82	6.88 ± 5.77	226
Sum	0.65	0.63	0.69	5.20 ± 4.75	125	0.36	0.31	0.40	6.12 ± 6.01	249
Max Pool	0.59	0.54	0.63	4.10 ± 5.08	120	0.60	0.56	0.64	6.43 ± 5.92	238

Tabela 5.26. Resultados do Experimento 2: Análise da qualidade de ordenação para consultas baseadas nos segmentos de maior energia para o grupo taxonômico Mamíferos. São comparados os valores MRR, mAP@5 e nDCG@5 para as diferentes técnicas de fusão.

Perch - Mamíferos						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.91	0.90	0.91	0.88	0.87	0.88
Weighted Avg	0.93	0.92	0.93	0.89	0.88	0.89
Sum	0.91	0.90	0.91	0.81	0.80	0.81
Max Pool	0.88	0.87	0.88	0.85	0.84	0.85
MFCCs - Mamíferos						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.78	0.77	0.79	0.73	0.72	0.74
Weighted Avg	0.81	0.80	0.82	0.77	0.75	0.78
Sum	0.76	0.75	0.77	0.70	0.69	0.71
Max Pool	0.73	0.72	0.74	0.71	0.70	0.72

Tabela 5.27. Análise estatística das métricas de precisão do Experimento 2 — Mamíferos. A tabela apresenta o Intervalo de Confiança (IC) de 95% e o Coeficiente de Variação (CV), expressos em porcentagens (%).

Técnica	**HNSW (Perch)**		**IMENN (Perch)**		**HNSW (MFCCs)**		**IMENN (MFCCs)**	
	IC	**CV**	**IC**	**CV**	**IC**	**CV**	**IC**	**CV**
Avg Pool	2.5	9.3	2.7	9.8	3.5	12.1	3.8	13.5
Weighted Avg	2.2	8.7	2.5	9.1	3.2	11.5	3.6	12.8
Sum	2.3	9.0	2.6	9.4	3.3	11.8	3.7	13.1
Max Pool	2.8	9.9	3.0	10.4	3.7	12.6	4.0	13.9

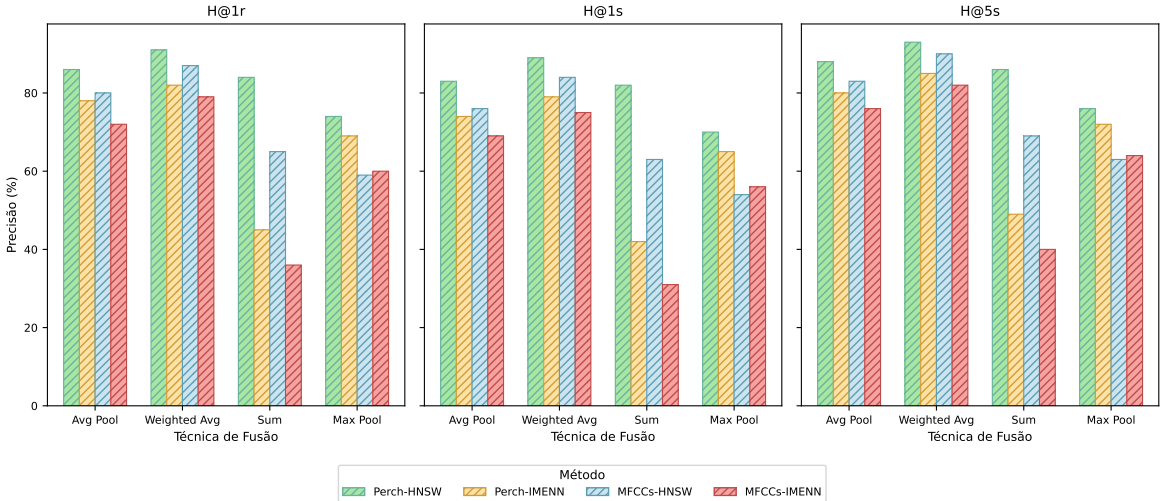


Figura 5.9. Comparativo visual do desempenho das técnicas de fusão no cenário de consulta com os segmentos de maior energia para Mamíferos.

5.6.3 Resultados do experimento 3

O objetivo deste experimento é quantificar o *trade-off* entre eficiência e precisão ao reduzir o tamanho da base de dados. Nesta abordagem, a base de dados foi compactada para conter unicamente o vetor não fusionado do segmento de maior energia de cada gravação. A recuperação foi, então, realizada utilizando os segmentos restantes como base para as consultas. Para a recuperação em nível de espécie, os vetores desses segmentos restantes de uma mesma gravação foram fusionados para criar um único vetor de consulta. Já para a recuperação em nível de gravação, cada segmento restante foi utilizado como uma consulta individual. Os resultados, apresentados nas tabelas 5.7 e 5.8, mostram que esta abordagem reduz o tempo médio de consulta de forma significativa, chegando a ganhos superiores até $3x$ no algoritmo *HNSW*.

Explorando o trade-off entre eficiência e precisão a arquitetura foi submetida a um teste de estresse, onde a base de dados foi compactada para conter apenas o segmento de maior energia, os resultados expõem a tensão fundamental entre escalabilidade e fidelidade representacional. A principal vantagem desta abordagem é a aceleração das consultas. Conforme a Tabela 5.28, a configuração Perch + HNSW + *Weighted Average Pooling* alcança uma vazão de 104 QPS, um aumento substancial comparado às 58 QPS do Experimento 1 na Tabela 5.10. No entanto, este ganho de velocidade afeto a qualidade da recuperação. A precisão $H@5_s$ caiu de 0.97 no Experimento 2 para 0.77, e a qualidade do *ranking* sofreu uma degradação crítica, com o $nDCG@5$ descendo de 0.86 para 0.71.

A precisão $H@5_s$ para Perch + HNSW + *Weighted Average* cai para 0.77 conforme na Tabela 5.28, vindo dos 0.97 obtidos no Experimento 2. A degradação mais acentuada ocorre na qualidade do *ranking*, como evidenciado pela queda do nDCG@5 de 0.86 para 0.71 comparando os resultados nas Tabelas 5.20 e 5.29. A perda da diversidade intra-gravação na base compactada dificulta a ordenação fina dos resultados relevantes. Apesar disso, a consistência dos resultados de Perch permanece alta, com um CV de

4.4% amostrado na Tabela 5.30, indicando que, embora o *ranking* piore, a capacidade de identificar a espécie correta no Top 5 é relativamente estável.

A queda no $nDCG$ indica que o sistema perdeu a capacidade de ordenar finamente os resultados. Bioacusticamente, isso ocorre porque as aves possuem repertórios vocais complexos e multifacetados. Ao reduzir a representação de uma gravação longa a um único segmento de maior energia, descarta-se a variabilidade intra-classe necessária para distinguir nuances entre espécies filogeneticamente próximas. A consulta fusionada busca uma riqueza de informação que não existe mais na base indexada.

Mesmo neste cenário com informação limitada na base, os *embeddings* Perch mantêm uma vantagem clara sobre os MFCCs. Com HNSW e *Weighted Average*, Perch com um $H@5_s = 0.77$ e um CV de 4.4% supera significativamente os MFCCs que atingiu um $H@5_s = 0.68$, que também mostram uma variabilidade muito maior com um CV de 12.8%, conforme Tabelas 5.28 e 5.30. Isso sugere que, enquanto o Perch consegue encapsular a identidade da espécie num único vetor denso, as *features* clássicas dependem de vários segmentos para estabilizar a predição.

Avaliando as técnicas de fusão aplicadas à consulta neste experimento (PP1), O *Weighted Average Pooling* manteve-se como a estratégia mais eficaz para a construção da consulta com $H@5_s = 0.77$ e um $MRR = 0.73$. A Figura 5.10 ilustra que, mesmo ao consultar uma base compactada, é vantajoso ponderar os segmentos de entrada pela energia. Isso maximiza a probabilidade de alinhar o vetor de consulta com o vetor "puro"que foi armazenado na base, mitigando parcialmente a perda de dados.

A comparação dos algoritmos de busca reafirma a hegemonia do *HNSW*. Ele superou o *IMENN* tanto na precisão, melhor *ranking* e maior vazão. O fato de o *HNSW* ser mais rápido e preciso sugere que a sua estrutura de navegação em grafo lida melhor com a distribuição espacial dos vetores compactados do que a abordagem de *clustering* do *IMENN*. Para Aves, a estratégia de base de dados compactada configura-se como uma alternativa viável apenas para triagem rápida ou aplicações de baixa latência onde a precisão não é crítica. A combinação Perch + *Weighted Average* + HNSW

oferece a melhor performance possível dentro das restrições deste cenário, mas a perda significativa de qualidade de *ranking* indica que, para monitoramento ornitológico de precisão, a preservação da diversidade temporal é insubstituível.

Tabela 5.28. Resultados do Experimento 3 (Perch): base de dados compactada com comparação entre *embeddings* Perch e MFCCs para o grupo taxonômico de Aves. A tabela inclui a vazão do sistema em Consultas por Segundo (QPS) para a recuperação ao nível da gravação ($H@1_r$) e da espécie ($H@1_s, H@5_s$).

Perch - Aves										
Técnica	**HNSW**					**IMENN**				
	H@1$_r$	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_r$**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.63	0.62	0.67	3.38 ± 2.21	98	0.56	0.59	0.63	6.61 ± 3.95	123
Weighted Avg	0.71	0.72	0.77	4.11 ± 2.14	104	0.68	0.68	0.73	4.33 ± 4.56	89
Sum	0.68	0.67	0.71	4.73 ± 2.56	86	0.60	0.58	0.65	3.91 ± 4.92	76
Max Pool	0.56	0.56	0.62	3.45 ± 2.89	33	0.50	0.48	0.55	4.78 ± 4.87	178
MFCCs - Aves										
Técnica	**HNSW**					**IMENN**				
	H@1$_r$	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_r$**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.56	0.47	0.63	4.54 ± 3.86	126	0.49	0.41	0.56	7.35 ± 5.45	241
Weighted Avg	0.65	0.53	0.68	5.98 ± 3.21	122	0.59	0.46	0.62	6.87 ± 7.10	227
Sum	0.58	0.47	0.63	0.00 ± 3.85	125	0.55	0.45	0.59	5.64 ± 6.51	243
Max Pool	0.47	0.33	0.51	4.98 ± 4.33	124	0.45	0.36	0.52	6.62 ± 5.31	247

Tabela 5.29. Resultados do Experimento 3: análise da qualidade de ordenação no cenário de base de dados compactada. A tabela apresenta os valores de MRR, mAP@5 e nDCG@5 para as diferentes técnicas de fusão no grupo taxonômico de Aves.

Perch						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.68	0.69	0.68	0.64	0.65	0.66
Weighted Avg	0.73	0.72	0.71	0.71	0.69	0.68
Sum	0.69	0.71	0.68	0.64	0.62	0.61
Max Pool	0.61	0.63	0.61	0.53	0.52	0.51
MFCCs						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.54	0.52	0.55	0.46	0.45	0.48
Weighted Avg	0.59	0.57	0.61	0.52	0.51	0.54
Sum	0.54	0.52	0.55	0.52	0.51	0.52
Max Pool	0.40	0.39	0.42	0.42	0.41	0.44

Tabela 5.30. Análise estatística das métricas de precisão do Experimento 3 — Médias. A tabela apresenta o Intervalo de Confiança (IC) de 95% e o Coeficiente de Variação (CV), relatados como porcentagens (%)..

	Sem aplicação do filtro							
Técnica	**HNSW (Perch)**		**IMENN (Perch)**		**HNSW (MFCCs)**		**IMENN (MFCCs)**	
	CI	**CV**	**CI**	**CV**	**CI**	**CV**	**CI**	**CV**
Avg Pool	0.9	4.1	1.3	5.9	2.9	14.5	2.7	15.4
Weighted Avg	1.2	4.4	1.0	4.1	2.8	12.8	3.0	15.3
Sum	0.7	3.0	1.3	5.9	2.9	14.6	2.6	13.6
Max Pool	1.2	6.0	1.3	7.1	3.4	21.6	2.9	18.1

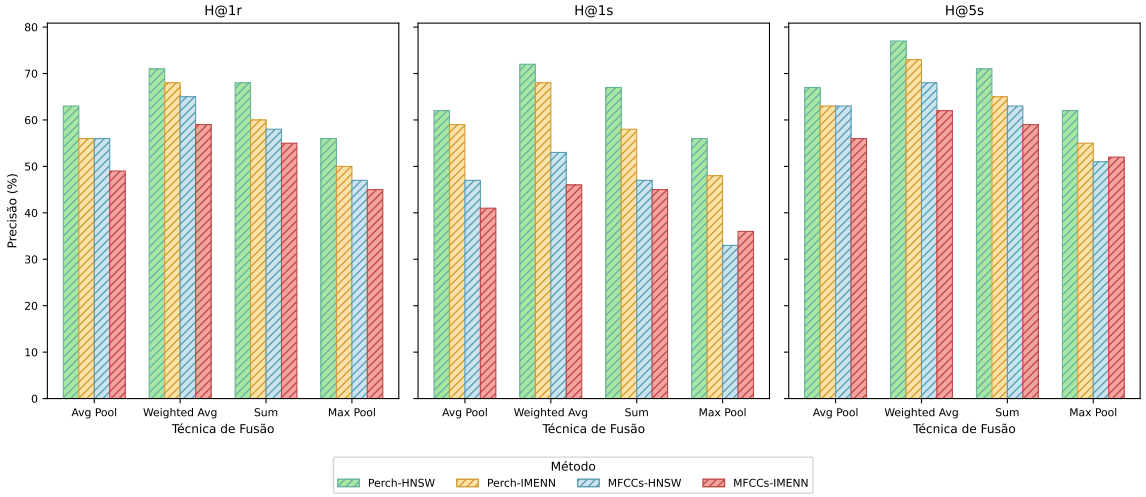


Figura 5.10. Comparativo visual do desempenho das técnicas de fusão com uma base de dados compactada para Aves.

Para o grupo Anuros, a estratégia de compactação da base de dados revela uma alta estabilidade na precisão, mas um comportamento inesperado na eficiência computacional. A análise dos resultados representados nas Tabelas 5.31 e 5.32 demonstra que a simplificação da base de dados teve um impacto marginal na qualidade da recuperação. A precisão $H@5_s$ com a configuração Perch + *HNSW* + *Weighted Average* manteve-se em 0.92, e a consistência atingiu um nível muito bom com um $CV = 3.2\%$, Tabela 5.33. Bioacusticamente, isso confirma que ao contrário das aves, onde a compactação removeu nuances vitais, para os anuros, o segmento de maior energia indexado na base contém informação suficiente para representar a espécie. A consulta fusionada consegue alinhar-se a este vetor "puro"com grande fiabilidade.

O dado mais intrigante é a queda abrupta na eficiência do *HNSW*, que registou apenas 20 QPS contra 78 QPS no Experimento 1 e 121 QPS no Experimento 2. Em contraste, o *IMENN* manteve-se rápido com 115 QPS. Este resultado sugere que, a densidade topológica dos vetores de anuros na base compactada força o grafo *HNSW* a percorrer mais arestas para encontrar os vizinhos, sacrificando velocidade por precisão. É possível que a distinção entre espécies muito próximas exija que o algoritmo percorra um número maior de nós no grafo para garantir a precisão, sacrificando a velocidade. Já o *IMENN*, por realizar buscas exatas dentro de *clusters*, não sofreu desta penalidade de navegação.

A avaliação das técnicas de fusão apresenta um *trade-off* entre pico de acurácia e estabilidade. Embora *Average* e *Sum Pooling* tenham atingido um $H@5_s$ ligeiramente superior de 0.93, o *Weighted Average Pooling* garantiu a melhor qualidade de ordenação com um $MRR = 0.91$ e $nDCG@5 = 0.89$ com a maior consistência com um $CV = 3.2\%$. A alta variabilidade do *Max Pooling* de $CV = 8.9\%$ desqualifica-o para este cenário. Ao construir a consulta, selecionar apenas os picos máximos dos segmentos restantes parece introduzir ruído ou *outliers* que desestabilizam a busca na base compactada.

A escolha do algoritmo para Anuros neste cenário impõe uma decisão binária. O *HNSW* oferece qualidade superior com um $H@5_s = 0.92$ contra 0.75 do *IMENN* e melhor *ranking* com um $nDCG = 0.89$ contra 0.72, mas é significativamente mais lento com 20 QPS contra 115 QPS. Dada a magnitude da perda de precisão com o *IMENN* com uma queda de quase 17 pontos percentuais em $H@5_s$, o *HNSW* permanece a escolha recomendada para aplicações de monitoramento, onde a identificação correta da espécie supera a necessidade de processamento em tempo real de altíssima velocidade. A estratégia de base compactada é altamente viável para Anuros em termos de qualidade de dados, validando a hipótese de que sinais estereotipados podem ser comprimidos sem grande perda de informação acústica. No entanto, a implementação deve considerar a latência do *HNSW* neste regime específico.

Tabela 5.31. Resultados do Experimento 3 (Perch): base de dados compactada com comparação entre *embeddings* Perch e MFCCs para o grupo taxonômico de Anuros. A tabela inclui a vazão do sistema em Consultas por Segundo (QPS) para a recuperação ao nível da gravação ($H@1_r$) e da espécie ($H@1_s, H@5_s$).

Perch - Anuros										
Técnica	**HNSW**					**IMENN**				
	H@1$_r$	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_r$**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.89	0.91	0.93	3.19 ± 5.54	21	0.66	0.65	0.69	5.28 ± 7.63	123
Weighted Avg	0.87	0.92	0.92	5.14 ± 4.11	20	0.69	0.71	0.75	5.87 ± 7.80	115
Sum	0.86	0.91	0.93	4.21 ± 4.90	25	0.78	0.81	0.85	5.39 ± 6.61	154
Max Pool	0.86	0.77	0.92	4.41 ± 6.20	22	0.76	0.77	0.81	5.17 ± 7.54	151
MFCCs - Anuros										
Técnica	**HNSW**					**IMENN**				
	H@1$_r$	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_r$**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.79	0.82	0.84	4.79 ± 6.85	142	0.41	0.48	0.48	6.31 ± 8.01	234
Weighted Avg	0.82	0.79	0.86	5.01 ± 6.53	134	0.68	0.71	0.75	7.16 ± 7.95	232
Sum	0.80	0.81	0.83	8.31 ± 5.89	131	0.59	0.61	0.63	6.37 ± 9.01	251
Max Pool	0.77	0.74	0.81	5.64 ± 6.06	141	0.50	0.56	0.56	6.81 ± 8.51	249

Tabela 5.32. Resultados do Experimento 3: análise da qualidade de ordenação no cenário de base de dados compactada. A tabela apresenta os valores de MRR, mAP@5 e nDCG@5 para as diferentes técnicas de fusão no grupo taxonômico de Anuros.

Perch						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.87	0.85	0.86	0.66	0.66	0.67
Weighted Avg	0.91	0.88	0.89	0.72	0.71	0.72
Sum	0.87	0.85	0.86	0.82	0.81	0.82
Max Pool	0.91	0.88	0.89	0.78	0.78	0.79
MFCCs						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.82	0.81	0.82	0.45	0.44	0.45
Weighted Avg	0.82	0.81	0.83	0.72	0.71	0.72
Sum	0.82	0.81	0.82	0.61	0.59	0.61
Max Pool	0.77	0.76	0.78	0.54	0.52	0.54

Tabela 5.33. Análise estatística das métricas de precisão do Experimento 3 — Anuros. A tabela apresenta o Intervalo de Confiança (IC) de 95% e o Coeficiente de Variação (CV), expressos em porcentagens (%).

Técnica	**HNSW (Perch)**		**IMENN (Perch)**		**HNSW (MFCCs)**		**IMENN (MFCCs)**	
	CI	**CV**	**CI**	**CV**	**CI**	**CV**	**CI**	**CV**
Avg Pool	0.7	2.2	0.7	3.1	0.6	2.1	1.4	8.8
Weighted Avg	1.0	3.2	1.1	4.3	1.3	4.3	1.3	4.9
Sum	1.3	4.0	1.3	4.3	0.6	2.1	0.7	3.3
Max Pool	2.7	8.9	0.9	3.4	1.3	4.5	1.2	6.4

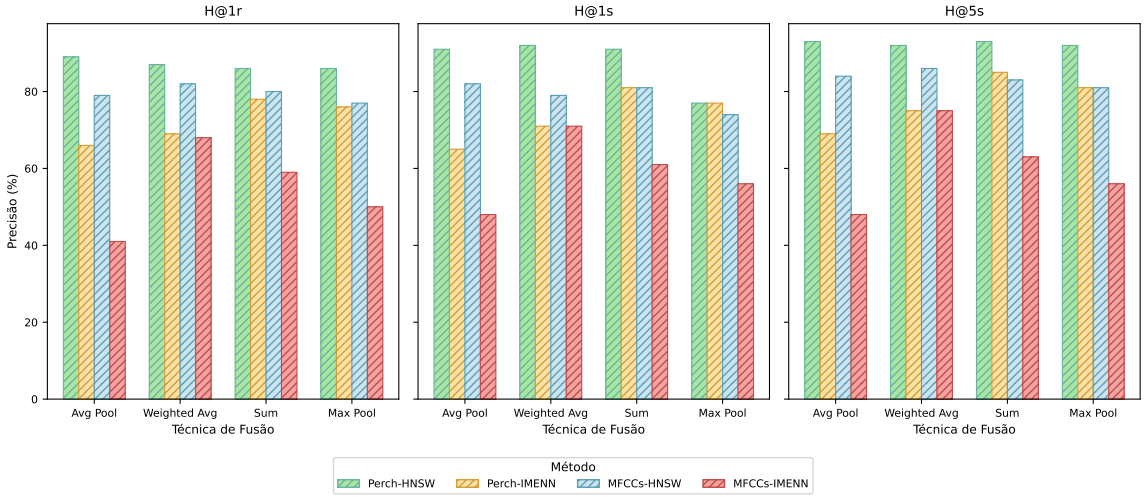


Figura 5.11. Comparativo visual do desempenho das técnicas de fusão com uma base de dados compactada para Anuros.

Encerrando a análise da Etapa 2 para os Mamíferos no Experimento 3, os resultados revelam um comportamento distinto em comparação com os outros grupos taxonômicos, especialmente no que diz respeito à consistência e eficiência. A aceleração das consultas esperada com a base reduzida é menos evidente para a melhor configuração de Perch com HNSW, mas o trade-off entre velocidade e qualidade torna-se particularmente acentuado.

A superioridade dos *embeddings* Perch sobre os MFCCs persiste neste cenário, embora a diferença na precisão média $H@5_s$ seja relativamente pequena para algumas fusões. Com HNSW e *Weighted Average Pooling*, Perch alcança $H@5_s$ de 0.76, enquanto MFCCs obtêm $H@5_s$ de 0.74 mostrados na Tabela 5.34. No entanto, a diferença crucial reside na consistência; Perch demonstra uma estabilidade excecional com um CV de apenas 2.3%, ao passo que MFCCs exibem uma variabilidade muito maior, com CV de 9.6%, como amostrado na Tabela 5.36. Esta alta consistência do Perch, mesmo com informação limitada na base, sublinha a sua capacidade de capturar características essenciais e generalizáveis das vocalizações dos mamíferos.

Avaliando as técnicas de fusão para Perch com HNSW (PP1), *Weighted Average Pooling* consolida-se novamente como a estratégia mais eficaz e confiável. Apresenta

o melhor desempenho em $H@5_s$, com valor de 0.76, lidera com clareza nas métricas de ordenação, registrando MRR de 0.73 e nDCG@5 de 0.74 na Tabela 5.35, e exibe a menor variabilidade com um CV de 2.3% evidenciados na Tabela 5.36. As fusões *Average Pooling* e *Sum Pooling* mostram uma queda acentuada na consistência, com CVs acima de 13%, tornando-as escolhas menos seguras. *Max Pooling*, embora com precisão razoável com um $H@5_s = 0.75$ mostrado na Tabela 5.34, também apresenta maior variabilidade, atingindo um CV de 4.3%. A Figura 5.12 ilustra visualmente o equilíbrio superior da fusão *Weighted Average*.

A comparação dos algoritmos de busca (PP4) para Perch com *Weighted Average* revela o trade-off mais extremo observado até agora. O HNSW oferece uma qualidade de recuperação muito superior, com $H@5_s$ de 0.76 e nDCG@5 de 0.74, e excelente consistência com um CV de 2.3%. Em contrapartida, o IMENN apresenta um desempenho significativamente inferior em qualidade $H@5_s = 0.64$ e nDCG@5 0.58, ver a Tabela 5.35 e uma consistência baixa com um CV de 11.3%, representado na Tabela 5.36. No entanto, IMENN é drasticamente mais rápido, processando 139 QPS contra apenas 20 QPS do HNSW. A escolha entre os algoritmos aqui dependeria fortemente do requisito da aplicação: HNSW para alta qualidade, IMENN para alta velocidade à custa de precisão e fiabilidade. Embora os MFCCs atinjam taxas de QPS muito elevadas até 211 com HNSW e Max Pooling, a sua precisão e consistência inferiores, especialmente comparadas ao Perch com HNSW, limitam a sua aplicabilidade.

Concluindo, para Mamíferos, a estratégia de base de dados compactada com Perch + HNSW + *Weighted Average* mantém uma boa precisão média e excelente consistência, mas com uma vazão de consultas baixa. O algoritmo IMENN oferece uma alternativa muito mais rápida, mas com uma penalidade considerável na qualidade e fiabilidade da recuperação.

Tabela 5.34. Resultados do Experimento 3 (Perch): base de dados compactada com comparação entre *embeddings* Perch e MFCCs para o grupo taxonômico de Mamíferos. A tabela inclui a vazão do sistema em Consultas por Segundo (QPS) para a recuperação ao nível da gravação ($H@1_r$) e da espécie ($H@1_s, H@5_s$).

Perch - Mamíferos										
Técnica	**HNSW**					**IMENN**				
	H@1$_r$	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_r$**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.69	0.55	0.72	4.16 ± 3.27	21	0.43	0.38	0.49	4.18 ± 5.26	146
Weighted Avg	0.73	0.73	0.76	5.86 ± 3.85	20	0.57	0.51	0.64	5.49 ± 5.01	139
Sum	0.67	0.55	0.72	4.88 ± 4.61	23	0.57	0.55	0.62	4.14 ± 6.87	177
Max Pool	0.69	0.71	0.75	4.41 ± 5.03	22	0.57	0.51	0.61	5.95 ± 6.63	167
MFCCs - Mamíferos										
Técnica	**HNSW**					**IMENN**				
	H@1$_r$	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**	**H@1$_r$**	**H@1$_s$**	**H@5$_s$**	$t \pm \sigma$	**QPS**
Avg Pool	0.66	0.69	0.69	5.37 ± 4.80	138	0.32	0.34	0.39	6.46 ± 7.40	221
Weighted Avg	0.68	0.61	0.74	6.91 ± 5.64	133	0.35	0.41	0.41	7.62 ± 8.21	245
Sum	0.63	0.67	0.69	6.63 ± 6.21	126	0.40	0.47	0.47	6.63 ± 8.96	250
Max Pool	0.57	0.32	0.61	5.37 ± 5.99	137	0.47	0.48	0.51	6.46 ± 8.44	245

Tabela 5.35. Resultados do Experimento 3: análise da qualidade de ordenação no cenário de base de dados compactada. A tabela apresenta os valores de MRR, mAP@5 e nDCG@5 para as diferentes técnicas de fusão no grupo taxonômico de Mamíferos.

Perch						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.62	0.61	0.64	0.43	0.42	0.44
Weighted Avg	0.73	0.74	0.74	0.57	0.56	0.58
Sum	0.62	0.61	0.64	0.58	0.58	0.59
Max Pool	0.64	0.61	0.65	0.52	0.51	0.54
MFCCs						
Técnica	**HNSW**			**IMENN**		
	MRR	**mAP@5**	**nDCG@5**	**MRR**	**mAP@5**	**nDCG@5**
Avg Pool	0.62	0.59	0.63	0.33	0.32	0.34
Weighted Avg	0.66	0.65	0.67	0.36	0.34	0.37
Sum	0.62	0.59	0.63	0.39	0.37	0.41
Max Pool	0.42	0.39	0.44	0.44	0.42	0.45

Tabela 5.36. Análise estatística das métricas de precisão do Experimento 3 — Mamíferos. A tabela apresenta o Intervalo de Confiança (IC) de 95% e o Coeficiente de Variação (CV), expressos em porcentagens (%).

Tecnica	**HNSW (Perch)**		**IMENN (Perch)**		**HNSW (MFCCs)**		**IMENN (MFCCs)**	
	CI	**CV**	**CI**	**CV**	**CI**	**CV**	**CI**	**CV**
Avg Pool	3.2	13.9	2.0	12.7	0.6	2.5	1.3	10.3
Weighted Avg	0.6	2.3	2.3	11.3	2.3	9.6	1.2	8.9
Sum	3.1	13.5	1.3	6.2	1.1	4.6	1.4	9.0
Max Pool	1.1	4.3	1.8	8.9	5.6	31.4	0.7	4.3

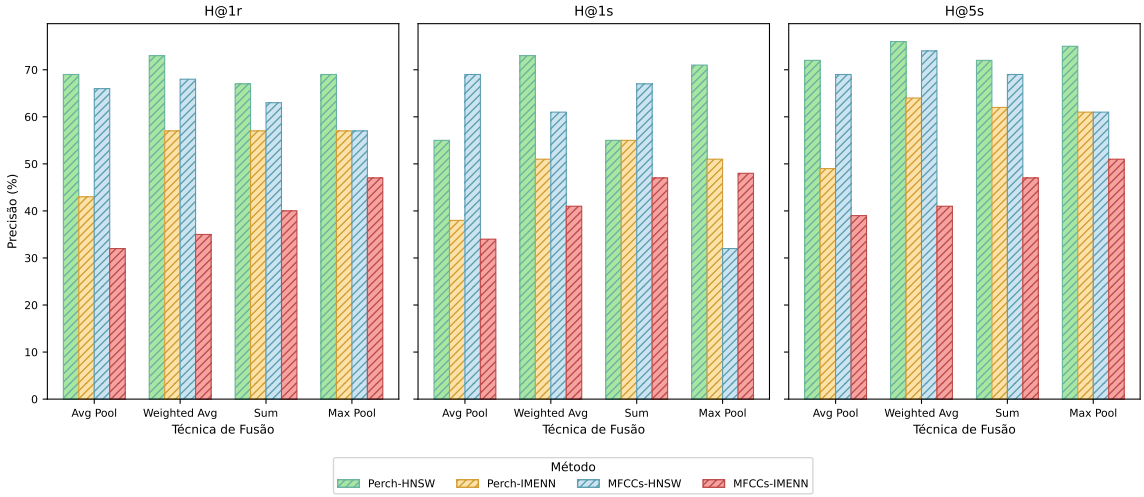


Figura 5.12. Comparativo visual do desempenho das técnicas de fusão com uma base de dados compactada para Mamíferos.

5.7 Discussão dos resultados

A integração dos resultados obtidos nos três experimentos da Etapa 2 permite explanar os mecanismos fundamentais que regem a recuperação acústica em paisagens sonoras. A análise converge para a validação da hipótese central, estabelecendo que a eficácia do sistema não depende de um único componente, mas da sinergia entre uma representação profunda eficaz, uma estratégia de agregação que preserva a saliência e uma indexação topológica eficiente.

A superioridade do *Weighted Average Pooling*, evidenciada de maneira recorrente nos resultados dos Experimentos 5.6.1, 5.6.2 e 5.6.3 não é acidental. Ao ponderar os segmentos pela sua energia RMS, esta técnica atua como um filtro de atenção não supervisionado. Em bioacústica, a energia é um *proxy* eficaz para a presença de vocalizações; portanto, o *Weighted Average Pooling* constrói um vetor representativo que maximiza a relação sinal-ruído no espaço latente. Isso explica os ganhos consistentes nas métricas de ordenação (↑MRR e ↑mAP) em comparação ao *Average Pooling*, que dilui o sinal ao incorporar segmentos de silêncio com peso igual.

Por outro lado, a análise crítica revela as falhas estruturais das outras técnicas.

O *Max Pooling* mostrou-se vulnerável a picos locais de ruído impulsivo, gerando falsos positivos que degradam o $nDCG@5$. Mais criticamente, o *Sum Pooling* provou ser incompatível com a métrica de distância Euclidiana (L_2) em cenários de duração variável. Como a norma do vetor soma escala com o número de segmentos, a distância L_2 passa a refletir a diferença de duração entre os áudios, e não a sua similaridade espectral, invalidando a recuperação.

A definição da *query* provou ser um fator tão determinante quanto o algoritmo de busca. O contraste entre o Experimento 2 e o Experimento 3 ilustra um claro *trade-off* operacional. Utilizar o segmento de maior energia como consulta gera um vetor com uma assinatura espectro-temporal "pura", livre de ruído contextual. Isso alinha perfeitamente a consulta com a representação ponderada da base, resultando no cenário ótimo de ordenação atingindo um $MRR = 0.86$ em Aves. Este protocolo valida-se como o padrão para tarefas de curadoria e busca exploratória.

A compactação da base de dados acelera drasticamente a recuperação, mas impõe uma penalidade de entropia. A perda da diversidade intra-gravação degrada o *ranking* em táxons complexos coo as Aves, embora seja tolerável para sinais repetitivas como os Anuros. Esta abordagem é, portanto, restrita a cenários de monitoramento em tempo real onde a latência é a métrica crítica.

A análise de vazão quantifica o custo da precisão. Enquanto a configuração de base completa no Experimento 2 opera a uma taxa conservadora de $\approx$ 34 QPS, a abordagem compactada Experimento 3 desbloqueia uma capacidade de processamento de até 167 QPS para recuperação em nível de gravação. Esse aumento de fator $5\times$ demonstra a viabilidade do sistema para processar fluxos contínuos de dados em grandes redes de sensores, desde que o *trade-off* na precisão de *ranking* seja aceitável para a aplicação alvo.

No que tange à indexação, o *HNSW* consolidou-se como a solução mais estável. Operando em um regime estável de latência $\approx 30\,ms$, sua estrutura de grafo hierárquico preservou a vizinhança semântica melhor que a abordagem de *clustering* do *IMENN*,

cuja busca parcial sacrificou a cobertura do espaço vetorial, deteriorando a precisão L_2.

Finalmente, os resultados decretam a perda de relevância técnica dos MFCCs para tarefas de generalização em bioacústica moderna. Embora competitivos em velocidade presentados nas Tabelas 5.10 e 5.22, a sua incapacidade de abstrair variações de ruído resultou em falhas na recuperação de gravações inéditas como se evidenciou no Experimento 3 da Etapa 1. Para ecossistemas com múltiplos táxons, as representações aprendidas pelo Perch provaram ser indispensáveis para garantir a invariância necessária.

Em suma, os experimentos validam a tríade Perch + *Weighted Average* + *HNSW* como o equilíbrio ótimo entre precisão, estabilidade e eficiência. Estes achados não apenas respondem às perguntas de pesquisa, mas fornecem a base metodológica sólida necessária para a aplicação prática do sistema no monitoramento de espécies vulneráveis, conforme será demonstrado no Capítulo 6.

5.8 Considerações finais sobre os resultados

Este capítulo teve por objetivo avaliar, de forma crítica, como diferentes estratégias de fusão de vetores de *features* afetam a recuperação acústica em bases ecoacústicas amplas. Os experimentos validaram, de forma consistente, que o *Weighted Average Pooling* representa o ponto ótimo de equilíbrio entre a eficácia à heterogeneidade intra-gravação e a precisão do *ranking*. Ao superar as limitações de diluição de sinal do *Average Pooling* e a instabilidade a *outliers* do *Max Pooling*, e ao evitar a distorção métrica inerente ao *Sum Pooling* em espaços Euclidianos, esta técnica provou ser fundamental para a consolidação de vetores representativos. A magnitude do ganho observado quando aplicada aos *embeddings* Perch sugere uma afinidade teórica, representações profundas ricas beneficiam-se desproporcionalmente de esquemas de fusão que preservam a saliência das assinaturas espectro-temporais.

A interação entre a fusão e a estratégia de consulta revelou-se um fator crítico de desempenho. A utilização do segmento de maior energia como *query* atuou como um filtro de atenção explícita, concentrando conteúdo discriminativo e reduzindo a ambiguidade do *matching*, o que elevou substancialmente as métricas MRR e nDCG. Embora a compactação da base ofereça latências compatíveis com o processamento em tempo real via *HNSW*, ela impõe um sacrifício na diversidade intra-classe que penaliza a recuperação em cenários acústicos complexos como em Aves.

Metodologicamente, os resultados apontam para um caminho evolutivo claro, a transição de pesos fixos baseados puramente em energia para pesos adaptativos, orientados por *proxies* de qualidade acústica. A adoção futura de mecanismos de atenção ou abordagens de *Multiple Instance Learning* poderia ajustar, em tempo de inferência, a contribuição relativa de trechos heterogêneos, integrando calibração de confiança para mitigar a influência de bases ruidosas.

Em síntese, a fusão de vetores de *features* não é apenas uma etapa de pré-processamento, mas um componente central da arquitetura de recuperação. Quando bem projetada, ela maximiza o valor informacional dos segmentos e preserva a eficiência exigida por sistemas de monitoramento em larga escala. A principal contribuição deste capítulo foi estabelecer, com evidências empíricas, que a tríade Perch + *Weighted Average* + *HNSW* constitui a configuração padrão para bioacústica, oferecendo a estabilidade e a precisão necessárias para a aplicação prática apresentada no capítulo seguinte.

6

Aplicação no Monitoramento de Espécies Vulneráveis

A recuperação acústica de paisagens sonoras é uma técnica que frequentemente requer uma quantidade significativa de recursos de processamento, especialmente ao empregar técnicas avançadas de análise de sinais de áudio. Para demonstrar a aplicabilidade e o valor prático do sistema, este capítulo apresenta um caso de uso focado no monitoramento de espécies vulneráveis. Empregamos a configuração de maior desempenho, validada experimentalmente nos capítulos anteriores: a recuperação baseada no segmento de maior energia como consulta, utilizando *embeddings* Perch agregados por *weighted average pooling* e indexados com *HNSW*. Conforme demonstrado, esta abordagem oferece o melhor equilíbrio entre precisão de recuperação e eficiência computacional, tornando-a ideal para a triagem rápida de grandes acervos de dados bioacústicos.

6.1 Análise de resultados do caso de uso

O estudo de caso foi delineado a partir de treze espécies selecionadas com base em critérios ecológicos e de conservação, priorizando aquelas que se encontram listadas nas

categorias *Critically Endangered* (CR), *Vulnerable* (VU), *Near Threatened* (NT) e também espécies de menor preocupação *Least Concern* (LC) na Lista Vermelha da IUCN [The IUCN Red List of Threatened Species, 2025]. O objetivo não é apenas validar a recuperação, mas também extrair *insights* ecologicamente relevantes. Diferentemente dos experimentos capítulos anteriores, aqui o foco está na análise operacional da abordagem proposta em um cenário realista, integrando dados bioacústicos heterogêneos provenientes de diferentes fontes.

Para cada espécie selecionada, foi utilizada uma única gravação de consulta, conforme definido na Etapa 2 do experimento 2. A partir dessas consultas, o sistema conseguiu identificar com precisão gravações correspondentes armazenadas na base vetorial, confirmando sua eficácia no processo de recuperação acústica. Além disso, a integração dos metadados geográficos com a biblioteca `geopy` permitiu mapear as regiões de ocorrência de cada espécie, fornecendo uma representação espacial detalhada que complementa a análise e potencializa a aplicação dos resultados para estudos ecológicos e estratégias de monitoramento.

6.1.1 Recuperação de espécies endêmicas e criticamente ameaçadas

O caso do *Crax alberti*, uma ave endêmica da Colômbia e classificada como Criticamente Ameaçada (CR), ilustra perfeitamente essa capacidade. O sistema recuperou com sucesso oito gravações exclusivas desta espécie, todas georreferenciadas na sua área de ocorrência conhecida. A Figura 6.2 apresenta o espectrograma de uma vocalização de *Crax alberti*. A análise qualitativa sugere que o sucesso da recuperação se deve à capacidade do Perch de capturar características discriminativas complexas, como a frequência fundamental grave (em torno de 200-400 Hz), a estrutura harmônica clara e a modulação temporal específica do canto, que o diferenciam de outros sons no ambiente. Esta precisão é fundamental para programas de conservação que dependem da detecção

não invasiva para estimar populações.

6.1.2 Generalização taxonômica e relevância para a conservação

Uma das validações mais significativas do sistema proposto reside na sua capacidade de generalizar a recuperação acústica através de uma ampla diversidade taxonômica. Os resultados demonstram um desempenho robusto na identificação de espécies das classes *Mammalia*, *Aves* e *Amphibia*, confirmando que os *embeddings* de Perch capturam características acústicas fundamentais que transcendem grupos biológicos específicos. Foram recuperadas com sucesso desde espécies com assinaturas vocais complexas, como a ave *Crax alberti* e o mamífero *Panthera onca*, até anfíbios com vocalizações mais estereotipadas, como *Andinobates opisthomelas* e *Allobates niputidea*. Esta capacidade de generalização entre diferentes táxons representa um avanço crucial para a realização de inventários de biodiversidade mais abrangentes e eficientes a partir de dados acústicos.

importância desta diversidade taxonômica é amplificada quando analisada sob a ótica da conservação. O sistema demonstrou não ser apenas um identificador de espécies, mas uma ferramenta eficaz para a triagem ecológica. Foram recuperadas com precisão gravações de espécies em múltiplos níveis de risco, de acordo com a Lista Vermelha da IUCN: uma espécie Criticamente Ameaçada (*Crax alberti*), uma Vulnerável (*Andinobates opisthomelas*) e quatro Quase Ameaçadas (*Lontra longicaudis*, *Pyrilia pyrilia*, *Panthera onca* e *Penelope purpurascens*). As demais espécies, classificadas como de Menor Preocupação (LC), serviram como um grupo de controle eficaz, validando a consistência do sistema mesmo em contextos de maior disponibilidade de dados.

Esta capacidade de identificar e, consequentemente, priorizar espécies de interesse para a conservação diretamente a partir de dados acústicos brutos é uma das contribuições centrais deste trabalho. O sistema não se limita a encontrar sons acusticamente semelhantes; ele organiza a informação de uma forma que é diretamente aplicável à

definição de estratégias de monitoramento, permitindo que pesquisadores e gestores foquem recursos em espécies e áreas que demandam atenção urgente.

6.1.3 Espécies de ampla distribuição

O sistema também demonstrou robustez na recuperação de espécies com vasta distribuição geográfica, como a onça-pintada (*Panthera onca*), que foram recuperados quinze registros de *Panthera onca* provenientes do Brasil, Colômbia e México; o *Crax alberti*, cujos registros foram exclusivamente recuperados na Colômbia. Este resultado é notável, pois implica que os *embeddings* do Perch são suficientemente robustos para generalizar através de diferentes biomas, como a Amazônia, Pantanal, florestas mexicanas, lidando com a variabilidade vocal intra-específica e as diferenças no ruído de fundo ambiental. Isto valida as conclusões do Experimento 3 da Etapa 1 sobre a capacidade de generalização do modelo.

6.1.4 Análise quantitativa e geográfica

A análise apresentada na Tabela 6.1 enriquecida com a coluna de Distância $L2$ Média, oferece uma visão mais aprofundada sobre a consistência acústica das espécies recuperadas. Essa métrica funciona como um *proxy* para avaliar a variabilidade intra-específica. Valores mais baixos sugerem uma assinatura acústica mais estereotipada e consistente. Por exemplo, a baixa distância média $L2$ observada para *Andinobates opisthomelas* com 0.18 corrobora a eficácia da técnica *Weighted Average Pooling* em sinais estereotipados, enquanto distâncias maiores em *Bradypus variegatus* com 0.41 sugerem uma variabilidade intra-específica que o modelo Perch consegue generalizar, mas com menor confiança.

A distribuição geográfica dos registros, visualizada na Figura 6.1, revela uma forte concentração de biodiversidade acústica em *hotspots* como a Colômbia e o Brasil. O sistema não só confirma áreas de ocorrência conhecidas, mas também permite a inte-

gração de dados de múltiplas fontes como Xeno-Canto, iNaturalist e CSA, criando um panorama mais completo que pode ser usado para identificar lacunas de monitoramento ou corredores ecológicos.

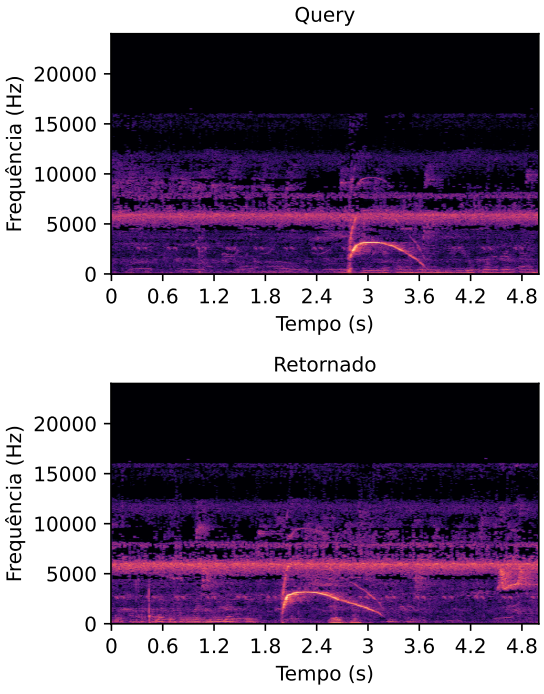
Por fim, os resultados obtidos reforçam a hipótese central deste estudo, evidenciando que a integração de *embeddings* pré-treinados, uma fusão de *features* adequada e uma base de dados vetorial eficiente possibilita a recuperação acústica de forma precisa e com alta utilidade prática para a triagem de espécies vulneráveis. Além de confirmar a eficácia do sistema no contexto aplicado, esses achados estão alinhados com trabalhos recentes na literatura, como os de Ghani et al. [2023], Stowell [2022] que demonstram que representações profundas capturam padrões espectro-temporais mais robustos e apresentam maior capacidade de generalização entre diferentes domínios, enquanto Malkov & Yashunin [2018], Pan et al. [2023], Taipalus [2024] destacam que técnicas modernas de indexação vetorial são essenciais para buscas rápidas, escaláveis e eficientes em grandes acervos bioacústicos.

Tabela 6.1. Caso de uso: espécies, gravações, distribuição geográfica e distância média $L2$.

Espécie	**# Grav.**	**Classe**	**Fonte**	**Cate.**	**Localização**	$L2$
Lontra longicaudis	2	Mammalia	XC (1) iNat (1)	NT	Brazil (2)	0.29
Allobates niputidea	2	Amphibia	CSA (2)	LC	Colombia (2)	0.22
Bradypus variegatus	2	Mammalia	iNat (2)	LC	Brazil (1) Colombia (1)	0.41
Colostethus inguinalis	2	Amphibia	iNat (2)	LC	Colombia (2)	0.37
Cerdocyon thous	2	Mammalia	iNat (2)	LC	Brazil (2)	0.38
Crax alberti	8	Aves	XC (8)	CR	Colombia (8)	0.21
Andinobates opisthomelas	10	Amphibia	iNat (10)	VU	Colombia (10)	0.18
Pyrilia pyrilia	14	Aves	XC (13) iNat (1)	NT	Colombia (12) Panama (2)	0.25
Panthera onca	15	Mammalia	XC (11) iNat (4)	NT	Brazil (7) Colombia (7) Mexico (1)	0.32
Alouatta seniculus	23	Mammalia	XC (1) iNat (22)	LC	Brazil (7) Colombia (16) Colombia (60)	0.28
Penelope purpurascens	77	Aves	XC (61) iNat (16)	NT	Costa Rica (6) Mexico (6) Panama (4) Nicaragua (1)	0.35
Elaenia flavogaster	260	Aves	XC (226) iNat (34)	LC	Argentina (8) Belize (10) Bolivia (6) Brazil (108) Colombia (60) Costa Rica (12) Ecuador (16) El Salvador (1) France (5) Honduras (4) Mexico (5) Panama (9) Paraguay (1) Peru (6) Venezuela (9)	0.31
Megarynchus pitangua	580	Aves	XC (339) iNat (169)	LC	Argentina (1) Bolivia (1) Brazil (302) Colombia (214) Costa Rica (16) Ecuador (16) El Salvador (1) Mexico (10) Panama (11) Peru (5) Venezuela (3)	0.26

Legenda: LC = Menor Preocupação; NT = Quase Ameaçada; VU = Vulnerável; CR = Criticamente Ameaçada; Grav = Gravações; Cat. = Categoria; $L2$ = Distância $L2$ Média.

Figura 6.1. Resultados da distribuição geográfica e do desempenho de recuperação das espécies selecionadas.



((a)) Espectrograma de Crax alberti.

((b)) Fotografia de Crax alberti.

Figura 6.2. Na parte (a) é apresentado o espectrograma do áudio *query* e do áudio retornado pelo sistema da espécie *Crax alberti*. Na parte (b) da figura, é mostrada uma fotografia da espécie *Crax alberti*, que está em perigo de extinção.

6.2 Implicações ecológicas e aplicações futuras

Os resultados deste caso de uso transcendem a validação técnica, oferecendo implicações diretas para a conservação. A capacidade de usar uma única gravação não rotulada para identificar rapidamente a presença de uma espécie criticamente ameaçada, como o *Crax alberti*, e mapear sua ocorrência, é uma ferramenta poderosa. Permite que gestores e pesquisadores definam áreas críticas de monitoramento e otimizem a alocação de recursos, como a instalação de gravadores autônomos.

Além disso, ao integrar dados de plataformas de ciência cidadã como o iNaturalist, o sistema pode validar e enriquecer estes conjuntos de dados, fornecendo uma camada de verificação acústica. Para espécies de ampla distribuição, como a *Panthera onca*, a análise da variabilidade acústica, refletida na distância $L2$ entre diferentes regiões pode abrir novas linhas de pesquisa sobre dialetos vocais e conectividade populacional. Em suma, o sistema proposto funciona como uma ponte eficaz entre dados acústicos brutos e conhecimento ecológico acionável, apoiando a tomada de decisões informadas para a preservação da biodiversidade.

6.3 Considerações finais sobre o monitoramento de espécies vulneráveis

O monitoramento de espécies vulneráveis enfrenta desafios crescentes devido à rápida degradação de habitats, à fragmentação ecológica e à necessidade de ferramentas capazes de lidar com grandes volumes de dados provenientes de múltiplas fontes. Nesse contexto, o sistema proposto apresenta um avanço significativo ao integrar *embeddings* acústicos pré-treinados com indexação vetorial eficiente, permitindo a recuperação rápida e precisa de gravações acusticamente semelhantes. Ao alinhar informações taxonômicas, geográficas e categorias de risco da *IUCN Red List*, o sistema oferece um suporte operacional robusto para a definição de prioridades em inventários de biodiversidade e

estratégias de conservação.

A capacidade de generalizar a recuperação entre diferentes grupos taxonômicos amplia o potencial de aplicação do sistema em diversos contextos, incluindo programas de monitoramento de longo prazo, estudos de conectividade ecológica e identificação de áreas-chave para a preservação da fauna. Além disso, ao permitir a integração com metadados georreferenciados, a abordagem proposta favorece a visualização espacial da presença de espécies, oferecendo subsídios para decisões mais corretas no planejamento de campanhas de campo e na alocação de esforços de amostragem.

A convergência entre os resultados obtidos e as evidências teóricas fortalece a aplicabilidade da abordagem proposta, posicionando o sistema como uma ferramenta estratégica para inventários de biodiversidade e para o suporte à tomada de decisão em estratégias de monitoramento e conservação. Sua adoção em cenários reais pode acelerar a detecção de espécies em risco, direcionar recursos de forma mais eficiente e apoiar ações de gestão adaptativa, contribuindo diretamente para a proteção de ecossistemas ameaçados e para o avanço das práticas de conservação baseadas em dados.

7

Conclusões

Esta pesquisa abordou o desafio de recuperar paisagens sonoras acusticamente semelhantes em grandes volumes de dados não rotulados. Para isso, foi desenvolvido um método que combina *embeddings* do modelo de *Deep Learning* Perch com técnicas de fusão de vetores *features* e um banco de dados vetorial de alta performance. O trabalho demonstra que esta abordagem não supervisionada melhora significativamente a precisão e a eficiência da recuperação, constituindo uma solução escalável para o monitoramento automatizado da biodiversidade.

7.1 Considerações finais

Nesta dissertação, abordou-se o desafio crítico da recuperação acústica em paisagens sonoras não rotuladas, propondo uma arquitetura não supervisionada que integra modelos de *Deep Learning*, estratégias de agregação de *features* e indexação vetorial aproximada. O objetivo central foi validar um sistema escalável capaz de superar as limitações das abordagens manuais e das representações clássicas, viabilizando o monitoramento automatizado da biodiversidade em grandes repositórios de dados ecoacústicos.

A investigação percorreu um fluxo metodológico rigoroso, desde o pré-processamento de sinais até a validação de algoritmos de busca. A análise integrada

das estratégias revelou que a eficácia do sistema não reside em um componente isolado, mas na sinergia entre três pilares: a robustez representacional dos *embeddings* Perch, que capturam invariâncias semânticas superiores aos MFCCs; a capacidade da fusão por *Weighted Average Pooling* em preservar a saliência de eventos acústicos em gravações heterogêneas; e a eficiência topológica do algoritmo *HNSW*, que viabilizou buscas de alta precisão com latência reduzida.

Os resultados validam a hipótese central 1.5, confirmando que a combinação de *embeddings* Perch, *fusão Weighted Average* e indexação *HNSW* supera as abordagens clássicas baseadas em MFCC, oferecendo o melhor equilíbrio entre latência e precisão. A arquitetura proposta provou ser resiliente à diversidade taxonômica abrangendo aves, anuros e mamíferos e capaz de operar com vazões compatíveis com aplicações em tempo real quando necessário. Portanto, conclui-se que a integração entre aprendizado profundo e estruturas vetoriais otimizadas constitui um paradigma viável e superior para a recuperação de informação em bioacústica.

Como validação externa da qualidade desta pesquisa, parte dos resultados e metodologias desenvolvidos foram submetidos à revisão por pares e aceitos para publicação. O trabalho foi apresentado no *XXII Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (WCAMA 2024)* , e subsequentemente no *XXXI Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia 2025).* Essas publicações reforçam a relevância científica da proposta e evidenciam seu impacto no estado da arte das soluções computacionais aplicadas à conservação da biodiversidade.

7.2 Limitações do método

Apesar dos resultados promissores obtidos na validação experimental, a abordagem proposta para a recuperação acústica de paisagens sonoras apresenta restrições intrínsecas que delimitam o seu escopo de aplicação operacional. Uma limitação central reside no viés de domínio do modelo *Perch*, utilizado para a extração dos *embeddings*, visto que

este foi treinado majoritariamente com vocalizações de aves a partir do acervo Xeno-Canto. Esse enviesamento taxonômico pode reduzir a capacidade de generalização da representação vetorial em ambientes acústicos dominados por outros grupos biológicos, como insetos ou espécies marinhos, ou em paisagens sonoras complexas onde a assinatura espectral difere significativamente dos dados de treino originais.

Adicionalmente, a eficácia do sistema é desafiada por condições de elevada densidade acústica. Embora a aplicação de filtros de redução de ruído no pré-processamento mitigue interferências estacionárias, os experimentos demonstraram que o desempenho da recuperação é sensível à polifonia intensa, caracterizada pela sobreposição simultânea de múltiplos indivíduos ou espécies. Nesses cenários, a extração de *features* pode gerar vetores de consulta ambíguos, comprometendo a precisão do *matching* e a qualidade do *ranking* na base de dados vetorial, uma vez que a mistura de fontes sonoras dilui a assinatura específica da espécie alvo.

Por fim, a arquitetura proposta impõe requisitos de infraestrutura significativos, demandando *hardware* com aceleração por GPU para executar com eficiência as etapas de inferência do modelo de *Deep Learning* e a busca vetorial em alta dimensionalidade. Essa dependência de recursos computacionais avançados pode limitar a portabilidade do método para dispositivos de borda ou plataformas de monitoramento que operam com restrições severas de consumo energético. Mesmo diante dessas limitações, as evidências empíricas confirmam que a combinação de modelos de aprendizado profundo com estruturas vetoriais otimizadas representa um avanço metodológico forte, oferecendo um caminho promissor para a escalabilidade do monitoramento bioacústico.

7.3 Trabalhos futuros

O método de recuperação desenvolvido inicia-se pela segmentação dos áudios, tornando o desempenho do sistema dependente da extração precisa de regiões de interesse. Para mitigar essa dependência e expandir as capacidades da arquitetura proposta, a evo-

lução natural da estratégia de *Weighted Average Pooling* reside na implementação de mecanismos de atenção aprendidos. Em vez de janelas fixas ou ponderação baseada puramente em energia, uma rede de atenção poderia identificar automaticamente as regiões espectro-temporais mais informativas dentro de cada gravação. Essa abordagem reduziria o impacto de rótulos fracos onde a localização exata da vocalização é desconhecida, permitindo uma utilização mais eficiente de gravações com duração variável e níveis de ruído flutuantes.

Paralelamente, para aprimorar a discriminabilidade dos vetores em cenários não supervisionados, propõe-se a adoção de técnicas de Aprendizado Contrastivo. Esta abordagem permitiria refinar o espaço latente aproximando *embeddings* de segmentos da mesma gravação e afastando segmentos de gravações distintas, aumentando a robustez intra-classe e a separabilidade inter-classe antes mesmo da etapa de indexação. Adicionalmente, futuras iterações devem abordar o problema da polifonia através de uma perspectiva multi-nível, integrando um módulo de detecção de densidade acústica seguido pela aplicação de técnicas de Separação Cega de Fontes (*Blind Source Separation*). Isso viabilizaria o isolamento de vocalizações sobrepostas antes da extração de características, melhorando a recuperação em ambientes de alta biodiversidade.

Um avanço estratégico final envolve a incorporação de modelos de áudio mais recentes, como o *Whisper*, adaptados ao domínio bioacústico através de técnicas de *fine-tuning* eficiente em parâmetros, como o LoRA. A integração subsequente com Grandes Modelos de Linguagem abriria caminho para a interpretação semântica e contextual dos eventos acústicos, elevando o sistema de um recuperador de similaridade acústica para um assistente de pesquisa ecológica capaz de inferir comportamentos. Em suma, o *pipeline* desenvolvido transcende a tarefa de recuperação, estabelecendo-se como uma métrica para avaliar a qualidade de *datasets*, a eficácia de estratégias de fusão e a detecção de eventos, com potencial para subsidiar estimativas populacionais biogeográficos de táxons como *Aves*, *Mammalia* e *Amphibia*.

Referências Bibliográficas

Addison, H.; Navine, A.; Klinck, H.; Sohier, D.; Kahl, S. & Denton, T. (2022). Birdclef 2022.

Aoyama, K.; Saito, K. & Ikeda, T. (2020). Inverted-file k-means clustering: Performance analysis. *arXiv*, pp. 1–15.

Azar, G.; Emami, M.; Fletcher, A. & Rangan, S. (2023). Estimation of embedding vectors in high dimensions. *arXiv*, 35:28708–28720.

Balaji, A. & Livinza, Z. M. (2025). A survey on bioacoustic signals denoising: Comparison of aerial and underwater signal processing techniques. *International Journal of Advanced Research*, 13(06):1547--1561.

Barroso, V.; Xavier, F. & Ferreira, C. (2023). Applications of machine learning to identify and characterize the sounds produced by fish. *ICES Journal of Marine Science*, 80(7):1854--1867.

Bianco, M.; Gerstoft, P.; Traer, J.; Ozanich, E.; Roch, M. A.; Gannot, S. & Deledalle, C. (2019). Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, 146(5):3590–3628.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Bjorck, J.; Rappazzo, B.; Chen, D.; Bernstein, R.; W., P. H. & Gomes, C. (2019). Automatic detection and compression for passive acoustic monitoring of the african

forest elephant. Em *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 476–484.

Bradbury, J. W.; Vehrencamp, S. L. et al. (1998). *Principles of animal communication*, volume 132. Sinauer Associates Sunderland, MA.

Burkov, A. (2019). *The hundred-page machine learning*, volume 22, chapter 1, p. 978. Andriy Burkov, Quebec City.

Cai, J.; Ee, D.; Pham, B.; Roe, P. & Zhang, J. (2007). Sensor network for the monitoring of ecosystem: Bird species recognition. Em *3rd international conference on intelligent sensors, sensor networks and information*, pp. 293–298.

Chen, H.; Xie, W.; V., A. & Zisserman, A. (2020). Holger klinck, sohier dane, stefan kahl, tom denton. Em *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725.

Clink, D. J.; Kier, I.; Ahmad, A. H. & Klinck, H. (2023). A workflow for the automated detection and classification of female gibbon calls from long-term acoustic recordings. *Frontiers in Ecology and Evolution*, 11:1071640.

Colonna, J. G.; Cristo, M. & Nakamura, E. F. (2014). A distribute approach for classifying anuran species based on their calls. Em *2014 22nd International Conference on Pattern Recognition*, pp. 1242–1247. IEEE.

Cornell Lab of Ornithology (2014). BirdCLEF. `https://www.macaulaylibrary.org`. Acesso em: 28 set. 2025.

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE*, 13(1):21–27.

Cuadrado, F. & Domínguez, J. (2019). *Teoría y técnica del sonido.*, volume 1, p. 220. SINTESIS.

Devalraju, D. V. & Rajan, P. (2022). Multiview embeddings for soundscape classification. *IEEE*, 30(1):1197–1206.

Fanioudakis, L. & Potamitis, I. (2017). Deep networks tag the location of bird vocalisations on audio spectrograms. *arXiv*, pp. 1–5.

Farina, A. & Gage, S. (2017). *Ecoacoustics: The ecological role of sounds*, volume 1, p. 321. John Wiley & Sons.

Farina, A. & Li, P. (2022). *Methods in ecoacoustics: the acoustic complexity indices*, volume 1, p. 127. Springer Nature.

Frommolt, K.; Bardeli, R. & Clausen, M. (2008). Computational bioacoustics for assessing biodiversity. Em *Proceedings of the International Expert meeting on IT-based detection of bioacoustical patterns, BfN-Skripten*, volume 234.

Gao, J. & Long, C. (2023). High-dimensional approximate nearest neighbor search: with reliable and efficient distance comparison operations. *Proceedings of the ACM on Management of Data*, 1(2):1–27.

Gavali, P. & Banu, J. S. (2025). A novel approach to indian bird species identification: employing visual-acoustic fusion techniques for improved classification accuracy. *Frontiers in Artificial Intelligence*, 8:1527299.

Ghani, B.; Denton, T.; Kahl, S. & Klinck, H. (2023). Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876.

Gil, F. & Donsker, D. (2005). xeno-canto :: Sharing wildlife sounds from around the world. `https://xeno-canto.org/`.

Giordano, B. L.; Esposito, M.; Valente, G. & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 26(4):664--672.

Gómez-Gómez, J.; Vidaña-Vila, E. & Sevillano, X. (2022). Western mediterranean wetlands bird species classification: evaluating small-footprint deep learning approaches on a new annotated dataset. *arXiv*, pp. 1–17.

Hagiwara, M.; Hoffman, B.; Liu, J.; Cusimano, M.; Effenberger, F. & Zacarian, K. (2023). Beans: The benchmark of animal sounds. Em *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5.

Han, Y.; Liu, C. & Wang, P. (2023). A comprehensive survey on vector database: Storage and retrieval technique, challenge. *arXiv*, pp. 1–13.

Harma, A. (2003). Automatic identification of bird species based on sinusoidal modeling of syllables. Em *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 5, pp. V–545.

Huang, P.; Xu, H.; Li, J.; Baevski, A.; Auli, M.; Galuba, W.; Metze, F. & Feichtenhofer, C. (2022). Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35(1):28708–28720.

iNaturalist community (2025). iNaturalist – Citizen science platform for biodiversity observations. Online at `https://www.inaturalist.org`.

IUCN Red List (2024). The iucn red list of threatened species.

Jie, J.; Wang, J. & Li, G. (2023). Survey of vector database management systems. *arXiv*, pp. 1–25.

Jina AI (2023a). VectorDB: A Python vector database you just need, no more, no less. `https://jina.ai/news/vectordb-a-python-vector-database-you-just-need-no-more-no-less/`.

Jina AI (2023b). Vectordb github repository. `https://github.com/jina-ai/vectordb`.

Klinck, H.; Cañas, J. S.; Demkin, M.; Dane, S.; Kahl, S. & Denton, T. (2025). Birdclef+ 2025. https://www.kaggle.com/competitions/birdclef-2025. Kaggle Competition.

Klinck, H.; Sohier, D.; Kahl, S. & Denton, T. (2023). Birdclef 2023.

Klinck, H.; Sohier, D.; Kahl, S.; Denton, T. & Addison, H. (2020). Cornell birdcall identification.

Krauss, O.; Balbino, M. & Nobre, C. (2023). Evaluation of methods of counterfactual explanation - a qualitative and quantitative analysis. Em *Anais do XI Symposium on Knowledge Discovery, Mining and Learning*, pp. 9–16.

Kurth, F.; Müller, M.; Fremerey, C.; Chang, Y. & Clausen, M. (2007). Automated synchronization of scanned sheet music with audio recordings. Em *ISMIR*, pp. 261–266.

Kvsn, R.; Montgomery, J.; Garg, S. & Charleston, M. (2020). Bioacoustics data analysis–a taxonomy, survey and open challenges. *IEEE Access*, 8:57684–57708.

Lakdari, M. W.; Ahmad, A. H.; Sethi, S.; Bohn, G. A. & Clink, D. J. (2024). Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons. *Ecological Informatics*, 80:102457.

Lecun, Y.; Bengio, Y. & Hinton, G. (2015). Deep learning. *nature*, 521(7553):436--444.

Li, P.; Roch, M. A.; Klinck, H.; Fleishman, E.; Gillespie, D.; Nosal, E.-M.; Shiu, Y. & Liu, X. (2023). Learning stage-wise gans for whistle extraction in time-frequency spectrograms. *IEEE Transactions on Multimedia*, 25:9302--9315.

Lu, Z.; Li, H.; Liu, M.; Lin, Y.; Qin, Y.; Wu, X.; Xu, N. & Pu, H. (2025). Dusafnet: A multi-path feature fusion and spectral–temporal attention-based model for bird audio classification. *Animals*, 15(15):2228.

Lyons, R. (1997). *Understanding digital signal processing*, chapter 23-54, p. 1120. Pearson Education India.

Lü, Z.; Shi, Y.; Lü, L.; Han, D.; Wang, Z. & Yu, F. (2024). Dual-feature fusion learning: An acoustic signal recognition method for marine mammals. *Remote Sensing*, 16(20):3823.

Maclean, K. & Triguero, I. (2023). Identifying bird species by their calls in soundscapes. *Applied Intelligence*, 53(19):21485–21499.

Malkov, Y. A. & Yashunin, D. (2018). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824--836.

Mancusi, M.; Zonca, N.; Rodolà, E. & Zuffi, S. (2023). Towards the evaluation of marine acoustic biodiversity through data-driven audio source separation. Em *IEEE Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–10.

Müller, M. (2015). *Fundamentals of music processing: Audio, analysis, algorithms, applications*, volume 5, p. 487. Springer.

Munson, A.; Webb, K.; Sheldon, D.; Fink, D.; Hochachka, W.; M., I.; Riedewald, M.; Sorokina, D.; Sullivan, B.; Wood, C. & Kelling, S. (2012). ebird. https://ebird.org/.

Murillo Bedoya, D. and Buitrago-Cardona, A. and Acevedo-Charry, O. and Ochoa-Quintero, J. M. (2021). Colección de Sonidos Ambientales Mauricio Álvarez-Rebolledo (IAvH-CSA). Instituto Humboldt (Colombia).

Murray, R. (1993). *Soundscape: Our Sonic Environment and the Tuning of the World*, volume 1, chapter 1, p. 320. Destiny.

Narasimhan, R.; Fern, X. & Raich, R. (2017). Simultaneous segmentation and classification of bird song using cnn. Em *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 146--150.

Nomorevotch (2020). Xeno-Canto Bird Recordings Extended (A-M). Kaggle.

Odom, K. J.; Araya-Salas, M.; Morano, J. L.; Ligon, R. A.; Leighton, G. M.; Taff, C. C.; Dalziell, A. H.; Billings, A. C.; Germain, R. R.; Pardo, M.; de Andrade, L. G.; Hedwig, D.; Keen, S. C.; Shiu, Y.; Charif, R. A.; Webster, M. S. & Rice, A. N. (2021). Comparative bioacoustics: a roadmap for quantifying and comparing animal sounds across diverse taxa. *Biological Reviews*, 96(4):1135--1159.

Oppenheim, A.; Schafer, R. & Buck, J. (1999). *Discrete-time signal processing*, volume 3, p. 1120. Pearson.

O'Haver, T. (1997). A pragmatic introduction to signal processing. *University of Maryland at College Park*.

Pan, J. J.; Wang, J. & Li, G. (2023). Survey of vector database management systems. *arXiv*, pp. 1–25.

Polastre, J.; Szewczyk, R.; Mainwaring, A. & Culler, D.and Anderson, J. (2004). Analysis of wireless sensor networks for habitat monitoring. *Springer*, 18:399–423.

Presannakumar, K. & Mohamed, A. (2023). Deep learning based source identification of environmental audio signals using optimized convolutional neural networks. *Applied Soft Computing*, 143:28708–28720.

Quaderi, S.; Labonno, S. A.; Mostafa, S. & Akhter, S. (2022). Identify the beehive sound using deep learning. *arXiv*, p. 17.

Rabiner, L. & Gold, B. (1975). Theory and application of digital signal processing. *Englewood Cliffs: Prentice-Hall*.

Rabiner, L. & Schafer, R. (2010). *Theory and applications of digital speech processing*, volume 1, p. 1056. Prentice Hall Press.

Rodríguez, A. (2018). Ciencia y divulgación sobre la sexta extinción masiva de biodiversidad,¿ es realmente el cambio climático el principal responsable? *La comunicación de la mitigación y la adaptación al Cambio Climático*.

Sainburg, T. (2019). Noisereduce. https://zenodo.org/records/3243139.

Sainburg, T.; McInnes, L. & Gentner, T. (2021). Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881--2907.

Schafer, R. M. (1969). *The new soundscape*, volume 1, p. 65. BMI Canada Limited Don Mills.

Scholl, S. (2021). Fourier, gabor, morlet or wigner: comparison of time-frequency transforms. *arXiv*, p. 10.

Sheikh, M. U.; Abid, H.; Shafique, B. S.; Hanif, A. & Haris, M. (2024). Bird whisperer: Leveraging large pre-trained acoustic model for bird call classification. Em *Proceedings of the Interspeech Conference*, pp. 5028--5032. ISCA.

Smith, J. (2008). *Mathematics of the discrete Fourier transform (DFT): with audio applications*, volume 2, chapter 1, p. 322. Julius Smith.

Smith, J. (2011). *Spectral audio signal processing*, chapter 2. Wiley Online Library.

Speaks, C. (2018). *Introduction To Sound: Acoustics for the Hearing and Speech Sciences.*, volume 4, p. 429. PLURAL PUBLISHING.

Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152.

Stowell, D. & L, W. (2016). warblrb10k$_{p}$*ublic*. *https* : *//archive.org/details/warblrb*10$_{p}$*ublic*.

Taipalus, T. (2024). Vector database management systems: Fundamental concepts, use-cases, and current challenges. *Cognitive Systems Research*, 85:101216.

Tan, M. & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. Em *International conference on machine learning*, pp. 6105–6114.

The IUCN Red List of Threatened Species (2025). Summary statistics – the iucn red list of threatened species.

Theodoridis, S.; Pikrakis, A.; Koutroumbas, K. & Cavouras, D. (2010). *Introduction to pattern recognition: a matlab approach*, volume 4, chapter 23-54, p. 217. Academic Press.

Tolkova, I. (2021). Feature representations for conservation bioacoustics: Review and discussion. Em *AI for Social Good Workshop at NeurIPS 2021*, pp. 1--12.

Tosato, G.; Shehata, A.; Janssen, J.; Kamp, K.; Jati, P. & Stowell, D. (2023). Auto deep learning for bioacoustic signals. *arXiv*, pp. 1–8.

Vaca-Castaño, G. & Rodriguez, D. (2010). Using syllabic mel cepstrum features and knearest neighbors to identify anurans and birds species. Em *2010 IEEE Workshop on Signal Processing Systems (SiPS)*, pp. 466–471. IEEE.

Vargas-Masís, R.; Segura-Sequeira, D.; Mendoza-Garro, E. & Vargas-López, D. (2021). Acoustic detection of red-capped manakin (ceratopipra mentalis) in sarapiquí, costa rica. Em *IEEE 3rd International Conference on BioInspired Processing (BIP)*, pp. 1–5.

Vasconcelos, D.; Nunes, N. & Gomes, J. (2020). An annotated dataset of bioacoustic sensing and features of mosquitoes. *Scientific Data*, 7(1):382.

Vasconcelos, D.; Nunes, N.; Ribeiro, M.; Prandi, C. & Rogers, A. (2019). Locomobis: a low-cost acoustic-based sensing system to monitor and classify mosquitoes. Em *IEEE Annual Consumer Communications e Networking Conference*, volume 1, pp. 1–6.

Vellinga, W.P. and Planqué, R. (2025). Xeno-canto – Bird sounds from around the world. GBIF Occurrence Dataset.

Vidaña-Vila, E.; Navarro, J. & Alsina-Pagès, R. (2017). Towards automatic bird detection: An annotated and segmented acoustic dataset of seven picidae species. *Data*, 2(2):18.

Wang, A. & et al. (2003). An industrial strength audio search algorithm. Em *Ismir*, pp. 7–13.

Wang, C.; Yang, H. & Meinel, C. (2015). Deep semantic mapping for cross-modal retrieval. Em *IEEE 27th International conference on tools with artificial intelligence (ICTAI)*, pp. 234--241.

Williams, S. E. & Bolitho, E.and Fox, S. (2003). Climate change in australian tropical rainforests: an impending environmental catastrophe. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1527):1887–1892.

Winursito, A.; Hidayat, R. & Bejo, A. (2018). Improvement of mfcc feature extraction accuracy using pca in indonesian speech recognition. Em *2018 International Conference on Information and Communications Technology (ICOIACT)*, pp. 379--383.

Wolfe, B.; Proctor, M. D.; Nolan, V. & Webb, S. L. (2023). An efficient acoustic classifier for high-priority avian species in the southern great plains using convolutional neural networks. *Wildlife Society Bulletin*, 47(e1492):1--17.

Xie, J. & Zhu, M. (2023a). Acoustic classification of bird species using an early fusion of deep features. *Birds*, 4(1):138--147.

Xie, J. & Zhu, M. (2023b). Acoustic classification of bird species using an early fusion of deep features. *Birds*, p. 11.

Xie, T.; Yang, Y.; Ding, Z.; Cheng, X.; Wang, X.; Gong, H. & Liu, M. (2023). Self-supervised feature enhancement: Applying internal pretext task to supervised learning. *IEEE*, 11:1708–1717.

Yu, G.; Mallat, S. & Bacry, E. (2008). Audio denoising by time-frequency block thresholding. *IEEE Transactions on Signal Processing*, 56(5):1830–1839.