BACHAREL EM CIÊNCIA DA COMPUTAÇÃO

ESTUDO DE EXTRATORES DE CARACTERÍSTICAS PARA CLASSIFICAÇÃO DE SONS AMBIENTAIS

MATHEUS JOSEPH MARQUES ARAÚJO

Rio Verde, GO

2026

INSTITUTO FEDERAL GOIANO - CAMPUS RIO VERDE
BACHAREL EM CIÊNCIA DA COMPUTAÇÃO

ESTUDO DE EXTRATORES DE CARACTERÍSTICAS PARA CLASSIFICAÇÃO DE SONS AMBIENTAIS

MATHEUS JOSEPH MARQUES ARAÚJO

Trabalho de Conclusão de Curso apresentado ao Instituto Federal Goiano - Campus Rio Verde, como requisito parcial para a obtenção do Grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Douglas Cedrim Oliveira

Rio Verde, GO

Junho, 2026

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema Integrado de Bibliotecas do IF Goiano - SIBi

Araújo, Matheus
A663e Estudo de extratores de características para classificação de sons ambientais: Study of feature extractors for environmental sound classification / Matheus Araújo. Rio Verde 2026.

32f. il.

Orientador: Prof. Dr. Douglas Cedrim Oliveira.
Tcc (Bacharel) - Instituto Federal Goiano, curso de 0219201 - Bacharelado em Ciência da Computação - Integral - Rio Verde (Campus Rio Verde).
1. Sons Ambientais. 2. Classificação de Sons. 3. MLP. I. Título.

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

TERMO DE AUTORIZAÇÃO PARA DISPONIBILIZAR PRODUÇÃO TÉCNICA NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Repositório Institucional do IF Goiano - RIIF Goiano Sistema Integrado de Bibliotecas

- Profissional de Educação do IF Goiano -

Com base no disposto na Lei Federal nº 9.610/98, e manual sobre a Produção Técnica, publicado pela DAV/CAPES/MEC*, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano, a disponibilizar gratuitamente o documento no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada eletronicamente abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

Identificação da Produção Técnica – DAV/CAPES

[ ] Editoria [ ] Material Didático

[ ] Curso de Formação Profissional [ ] Projetos de Extensão à Comunidade

[ ] Relatório Técnico Conclusivo [ ] Atividade Técnica/Tecnológica

[ ] Disseminação do Conhecimento Técnico/Tecnológico [ ] Produto Bibliográfico

[X] Outras Produções Técnicas - Tipo: TCC (Graduação)

Nome Completo do Autor/a: Matheus Joseph Marques Araújo

Matrícula: 2017102201910471

Título do Trabalho: ESTUDO DE EXTRATORES DE CARACTERÍSTICAS PARA CLASSIFICAÇÃO DE SONS AMBIENTAIS

Restrições de Acesso ao Documento

Documento confidencial: [X] Não [ ] Sim

Justifique: _____

Informe a data que poderá ser disponibilizado no RIIF Goiano: 10 / 07 / 2026

O documento está sujeito a registro de patente? [ ] Sim [X] Não

O documento pode vir a ser publicado como livro e/ou artigo? [ ] Sim [X] Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a docente e/ou autor/a declara que:

1 - o documento é seu trabalho original, detém os direitos autorais da produção técnica e não infringe os direitos de qualquer outra pessoa ou entidade;

2 - obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;

3 - cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Rio Verde, 2 de julho de 2026.

(Assinado Eletronicamente)

Matheus Joseph Marques Araújo (Autor)

(Assinado Eletronicamente)

Douglas Cedrim Oliveira (Orientador)

1058004

(Assinatura do Docente, Autor e/ou Detentor dos Direitos Autorais)

Documento assinado eletronicamente por:

- **Douglas Cedrim Oliveira**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 02/07/2026 22:44:10.
- **Matheus Joseph Marques Araújo**, **2017102201910471 - Discente**, em 02/07/2026 22:45:26.

Este documento foi emitido pelo SUAP em 02/07/2026. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 837385
Código de Autenticação: c56708b5dd

INSTITUTO FEDERAL GOIANO
Campus Rio Verde
Rodovia Sul Goiana, Km 01, Zona Rural, 01, Zona Rural, RIO VERDE / GO, CEP 75901-970
(64) 3624-1000

Regulamento de Trabalho de Conclusão de Curso (TCC) – IF Goiano - Campus Rio Verde

ANEXO V - ATA DE DEFESA DE TRABALHO DE CURSO

Aos vinte e dois dias do mês de junho de dois mil e vinte e seis às catorze horas, reuniu-se a Banca Examinadora composta por: Prof. Dr. Douglas Cedrim Oliveira (orientador), Prof. Dr. Márcio Antonio Ferreira Belo Filho (membro interno) e Prof. Dr. André da Cunha Ribeiro (membro interno), para examinar o Trabalho de Conclusão de Curso (TCC) intitulado “Estudo de extratores de características para classificação de sons ambientais” de Matheus Joseph Marques Araújo, estudante do curso de bacharelado em Ciência da Computação do IF Goiano – Campus Rio Verde, sob Matrícula nº 2017102201910471. A palavra foi concedida ao estudante para a apresentação oral do TC, em seguida houve arguição do candidato pelos membros da Banca Examinadora. Após tal etapa, a Banca Examinadora decidiu pela Aprovação do estudante. Ao final da sessão pública de defesa foi lavrada a presente ata, que segue assinada pelos membros da Banca Examinadora.

Rio Verde, 22 de junho de 2026.

(Assinado eletronicamente)

Douglas Cedrim Oliveira

Orientador(a)

(Assinado eletronicamente)

Márcio Antonio Ferreira Belo Filho

Membro da Banca Examinadora

(Assinado eletronicamente)

André da Cunha Ribeiro

Membro da Banca Examinadora

Observação:

Documento assinado eletronicamente por:

- **Douglas Cedrim Oliveira**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 22/06/2026 15:37:08.
- **Marcio Antonio Ferreira Belo Filho**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 22/06/2026 15:38:15.
- **Andre da Cunha Ribeiro**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 22/06/2026 15:40:40.

Este documento foi emitido pelo SUAP em 22/06/2026. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 833658
Código de Autenticação: 85efd278dc

Dedico esse trabalho a minha família.

AGRADECIMENTOS

Primeiramente, agradecer minha família, meu pai Francisco de Assis, minha mãe Helen e minha irmã Esther Mary, pelo apoio que recebi do inicio ao fim, sem eles eu não teria chegado até aqui.

Agradecer meus professores que também me apoiaram durante o curso, mas agradecer em especial ao professor e orientador Douglas Cedrim, pelo imenso apoio dado principalmente nessa fase difícil do TCC, me ajudou muito, disponibilizando seu tempo e também a paciência.

E também agradecer ao meus amigos que conheci durante o curso, que fizeram parte dessa trajetória, que me ajudaram desde o inicio. Um agradecimento em especial ao meus amigos Mateus Farias e ao Athos, que foram muito importante nessa fase final do curso.

RESUMO

ARAÚJO, Matheus. **Estudo de extratores de características para classificação de sons ambientais.** Junho, 2026. 32 f. Monografia – (Curso de Bacharel em Ciência da Computação), Instituto Federal Goiano - Campus Rio Verde. Rio Verde, GO.

O universo sonoro ao nosso redor vai muito além da linguagem falada ou de músicas estruturadas; ele é composto por uma infinidade de sons ambientais, desde o latido de um cachorro até o ruído de máquinas. Ensinar os computadores a ouvir e interpretar esses eventos — uma capacidade conhecida como percepção acústica — é fundamental para o avanço de tecnologias modernas, como sistemas de segurança inteligentes e veículos autônomos. No entanto, o áudio ambiental é complexo e não estruturado, o que torna a extração de suas características um desafio crítico. Este trabalho apresenta um estudo comparativo de extratores de características para a classificação de sons ambientais, com o objetivo de analisar como essas técnicas mapeiam e separam os áudios por classe no espaço computacional. Utilizando bases de dados públicas e consolidadas no estado da arte, a separação e a distribuição das classes foram avaliadas visualmente, e os atributos extraídos alimentaram um modelo de rede neural do tipo *Multilayer Perceptron* (MLP). Os resultados demonstram a eficácia da abordagem proposta, alcançando uma acurácia de 77,50% em um dos datasets testados, evidenciando o potencial das representações espectro-temporais para a evolução do reconhecimento acústico automatizado.

Palavras-chave: Sons Ambientais, Classificação de Sons, MLP.

ABSTRACT

ARAÚJO, Matheus. **Study of feature extractors for environmental sound classification**. Junho, 2026. 32 f. Trabalho de Conclusão de Curso – Bacharel em Ciência da Computação, Instituto Federal Goiano - Campus Rio Verde. Rio Verde, GO.

The soundscape surrounding us goes far beyond spoken language or structured music; it is composed of a myriad of environmental sounds, ranging from a dog barking to machinery noise. Teaching computers to listen to and interpret these events—a capability known as acoustic perception—is fundamental for the advancement of modern technologies, such as intelligent security systems and autonomous vehicles. However, environmental audio is complex and unstructured, making feature extraction a critical challenge. This work presents a comparative study of feature extractors for environmental sound classification, aiming to analyze how these techniques map and separate audio events into distinct classes within the computational space. Using public datasets consolidated in the state-of-the-art, the class separation and distribution were visually evaluated, and the extracted attributes were used to train a Multilayer Perceptron (MLP) neural network. The results demonstrate the effectiveness of the proposed approach, achieving an accuracy of 77.50% on one of the tested datasets, highlighting the potential of spectro-temporal representations for the evolution of automated acoustic recognition.

Keywords: Ambient Sounds, Sound Classification, MLP.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Hoje, os computadores conseguem reconhecer objetos em imagens em tempo real e compreender a fala humana com uma precisão impressionante. Porém, se pararmos para ouvir o ambiente ao nosso redor, perceberemos que o universo sonoro vai muito além da linguagem falada ou de músicas estruturadas. O latido de um cachorro, o barulho da chuva, uma buzina no trânsito ou o som de passos — todos esses eventos compõem o que chamamos de sons ambientais.

A Classificação de Sons Ambientais é a área da computação que tenta ensinar os computadores a ouvir, interpretar e categorizar esses ruídos do cotidiano. Essa capacidade de "percepção acústica" é a base de uma série de tecnologias modernas. Ela é fundamental para sistemas de segurança inteligentes (capazes de detectar estilhaços de vidro ou pedidos de socorro), para veículos autônomos (que precisam reagir à aproximação de uma sirene antes mesmo de visualizá-la) e para dispositivos de assistência à saúde e automação residencial.

Apesar do enorme potencial prático, a análise computacional de sons ambientais historicamente enfrentou barreiras complexas. Diferente de áreas consagradas como a visão computacional, que há muito tempo se consolidou com bases de dados públicas massivas e padronizadas, o processamento de áudio ambiental sofria com a fragmentação. Durante anos, os pesquisadores dependiam de amostras pequenas, privadas ou excessivamente específicas, o que dificultava a comparação justa entre diferentes algoritmos e limitava a evolução da área.

A mudança nesse campo ocorreu com o surgimento de iniciativas de padronização metodológica e a criação de *"benchmarks"* de código aberto. A disponibilização de datasets públicos, balanceados e categorizados permitiu que a comunidade científica passasse a tratar o áudio não apenas como um sinal bruto unidimensional, mas sim por meio de representações espectro-temporais que alimentam redes neurais profundas, elevando o patamar da precisão automatizada.

Sendo assim, este trabalho tem por objetivo investigar diferentes extratores de características de áudio para identificar a eficácia de cada característica em diferenciar espécies dessa ordem de outros sons ambientais. Ao avaliar essas técnicas, busca-se facilitar e aprimorar o monitoramento ambiental, contribuindo para a conservação e proteção dos ecossistemas. Mais especificamente, pretende-se investigar se a utilização dos Coeficientes Cepstrais de Frequência Mel, juntamente com a Taxa de Cruzamento por Zero podem classificar sons ambientais corretamente. Também pretende-se investigar como a utilização de redes neurais para o treinamento, juntamente com a utilização de visualização de dados, pode auxiliar nesse processo.

O restante do texto está estruturado de forma que no Capítulo 2 será abordado a

fundamentação teórica, onde será discutido temas que abordam assuntos relevantes para a compreensão do trabalho. No Capítulo 3 será discutido os trabalhos relacionados, que usam técnicas e métodos relevantes que ajudaram no desenvolvimento desse trabalho. No Capítulo 4, materiais e métodos será mostrado os extratores usados, base de dados, técnicas de projeção que foram aplicados no trabalho. No capítulo 5, será discutido os resultados obtidos pelas técnicas utilizadas, também comparando resultados com outro trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Neste tópico será discutido temas para apresentar e explicar conceitos que ajudaram a melhor compreensão para o trabalho apresentado.

2.1 Som

O som pode ser descrito como uma onda mecânica que se propaga através de um meio, causando vibração em moléculas do ar por exemplo, essa onda se espalham por diferentes direções, assim podendo chegar os ouvidos humanos, essas vibrações chegam ao tímpano. O som possui uma variação em sua frequência, o sistema auditivo humano consegue perceber o som que varia de uma frequência de 20 Hz até 20 kHz ou 20000 Hz (HALLIDAY; RESNICK; WALKER, 2014).

A frequência é, basicamente, a característica do som que o nosso cérebro interpreta como a "altura" ou o "tom" daquele áudio. É ela que nos diz se um som é grave ou agudo. Para exemplificar, podemos pensar no som como uma onda que viaja pelo ar, subindo e descendo o tempo todo. Frequência é a velocidade dessa oscilação: Ela mede quantas vezes essa onda completa um ciclo inteiro (uma subida e uma descida) no período de 1 segundo. A unidade de medida é o Hertz (Hz): Se uma onda vibra 100 vezes em um segundo, dizemos que ela tem uma frequência de 100 Hz (DONOSO, 2014).

Baseado no Donoso (2014), a amplitude é a característica física do som que o nosso cérebro interpreta como a intensidade ou, no linguagem do dia a dia, o volume do áudio. É ela que nos diz se um som é forte (barulhento) ou fraco (silencioso).

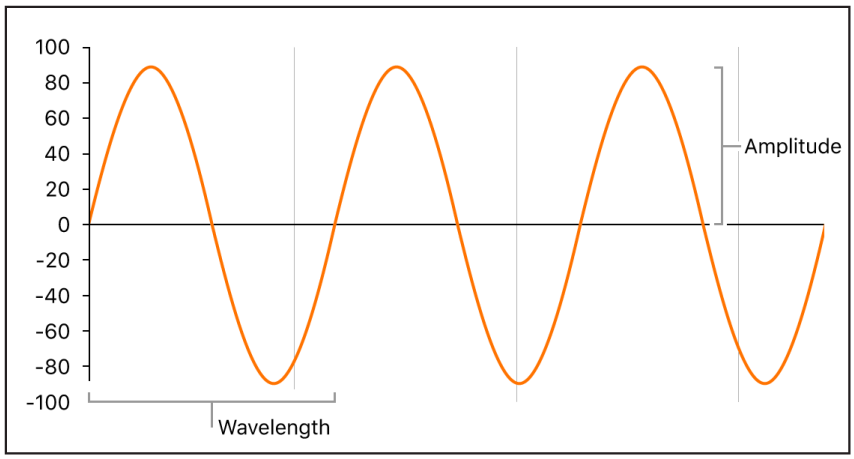
Para entender o conceito, imagine que o som viaja pelo ar empurrando as moléculas, criando ondas de pressão.

Amplitude pode ser descrita como tamanho ou a altura dessa onda: Ela mede o grau de perturbação que a onda causa no meio. Em termos práticos, é a distância entre o ponto de repouso (o silêncio absoluto) e o pico máximo da onda (seja para cima ou para baixo).

Quanto maior a força aplicada para gerar o som, maior será a amplitude: Se você tocar uma corda de violão de leve, ela vibra pouco, desloca pouco ar e a amplitude é baixa. Se você puxar a corda com força, ela vai oscilar muito mais, criando uma onda com grande amplitude.

Na Figura 1 podemos ver a representação de um *waveform*, mostrando um áudio com seu comprimento de onda.

Figura 1 – Ilustração de um som através do seu formato de onda com seu comprimento e amplitude (*waveform*).



Fonte: Extraída de Bäckström et al. (2022).

2.2 Sons ambientais

Os sons ambientais, também conhecidos como sons de fundo ou sons de ambiente, são um componente essencial na criação de ambientes acústicos. Eles se referem aos sons naturais ou artificiais que estão presentes em um determinado espaço e que podem afetar a percepção e a experiência das pessoas que estão naquele ambiente. A análise e o entendimento dos sons ambientais envolvem várias disciplinas, incluindo acústica, psicologia, ecologia e design de som (CHU; NARAYANAN; KUO, 2009).

Sons Naturais, são todos os sons gerados de forma orgânica pela própria natureza, sem qualquer intervenção humana. Eles funcionam como um termômetro da saúde de um ecossistema e são subdivididos em dois grupos, Geofonia: Os sons de elementos não-vivos, como o barulho do vento nas folhas, o impacto da chuva no solo, o trovão ou o murmúrio constante de um riacho. Biofonia: Os sons produzidos por seres vivos, como o canto dos pássaros ao amanhecer, o coaxar dos sapos e o zumbido dos insetos. Na maioria das vezes, o cérebro humano interpreta os sons naturais como sinais de segurança e relaxamento, sendo amplamente utilizados para reduzir o estresse.

Sons Antropogênicos, também conhecidos na ciência como antropofonia, englobam todo e qualquer som criado pelas atividades humanas e pelo desenvolvimento tecnológico. Exemplos: O fluxo contínuo do tráfego de veículos, o barulho de britadeiras em construções, o zumbido de máquinas industriais, sistemas de ventilação e até a música ambiente de um shopping. Diferente dos sons naturais, o impacto dos sons antropogênicos depende profundamente do contexto e da intensidade. Ponto Negativo: Quando são muito intensos ou repetitivos (como o ruído de uma obra ou trânsito pesado), transformam-se em poluição sonora, gerando estresse, ansiedade e fadiga mental. Ponto Positivo: Quando são contextualizados (como uma música suave em um café ou o burburinho de uma praça

movimentada), podem trazer uma sensação de pertencimento, conforto e dinamismo social. (PIJANOWSKI et al., 2011).

2.3 Som digital / Áudio

O som digital, ou simplesmente áudio, refere-se à representação computacional de uma onda sonora captada por um receptor, como um dispositivo equipado com microfone. Na prática, a gravação realizada por um equipamento eletrônico envolve a conversão de ondas sonoras — que são vibrações físicas no ar — em sinais elétricos. Para que esses sinais possam ser processados, armazenados e reproduzidos por computadores, eles precisam passar por um processo de digitalização (OPPENHEIM; SCHAFER; BUCK, 1999).

Esse processo começa no microfone, que transforma a pressão da onda sonora em um sinal elétrico analógico (contínuo no tempo). Em seguida, esse sinal é enviado para um Conversor Analógico-Digital (ADC, do inglês *Analog-Digital Converter*. A conversão ocorre por meio de duas etapas fundamentais:

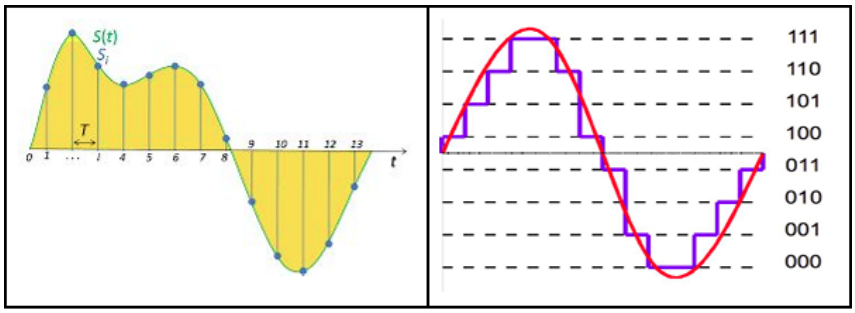
Amostragem: onde a intensidade do som é medida em intervalos regulares de tempo (ritmo definido pela taxa de amostragem).

Quantização: onde o valor da amplitude do som em cada um desses intervalos é convertido em um número binário discreto.

Como o sinal analógico original é contínuo e o mundo digital trabalha com valores finitos (discretos), esse processo de conversão introduz pequenas aproximações, gerando uma sutil perda de fidelidade conhecida como erro de quantização ou alteração na amostra de áudio original.

Na Figura 2 temos a representação do sinal de áudio digital, através do *waveform*.

Figura 2 – Ilustração de um (*waveform*) de um áudio digital.



Fonte: Extraída de Instituto de Engenharia (2020).

A qualidade do áudio digital é influenciada por diversos fatores, como a taxa de amostragem e a profundidade de bits. A taxa de amostragem, medida em hertz (Hz), determina quantas vezes por segundo o som é amostrado. Uma taxa de amostragem mais alta geralmente resulta em uma reprodução de som de melhor qualidade, pois captura mais detalhes do som original. Já a profundidade de bits, medida em bits, define a precisão

com que cada amostra de som é representada. Uma maior profundidade de bits permite uma maior faixa dinâmica e menos distorção no áudio.

Apesar de ser um termo não muito comum, o áudio também possui um frame, ou quadro. Um frame de áudio, no contexto digital, pode ser entendido como um "pacote" mínimo de informação sonora em um instante específico. Imagine o som como uma sequência de fotografias; cada uma dessas "fotos" individuais seria o equivalente a uma amostra (sample), que captura a intensidade do som naquele momento. O frame, então, é o conjunto que agrupa todas as amostras de todos os canais de áudio (como esquerdo e direito em um som estéreo) que ocorrem exatamente ao mesmo tempo. Portanto, em um áudio estéreo, um frame conterá duas amostras: uma para o canal esquerdo e outra para o direito.

A utilidade dos frames está na forma como o áudio digital é processado e reproduzido. Computadores e dispositivos de áudio leem uma sequência contínua desses frames para reconstruir a onda sonora original que ouvimos. A quantidade de frames lidos por segundo está diretamente relacionada à taxa de amostragem (medida em Hertz), que define a fidelidade do áudio. Além das amostras, um frame de áudio também é caracterizado pela profundidade de bits *(bit depth)*, que determina a precisão com que a intensidade do som é representada. Em resumo, os frames são os blocos fundamentais que, quando organizados em sequência, formam a totalidade de um arquivo de som digital.

Com base nos conceitos de áudio digital, o janelamento (windowing) é uma técnica matemática essencial utilizada para analisar ou processar pequenos segmentos de um sinal de áudio, ou seja, os próprios frames. Quando "recortamos" um frame de um fluxo de áudio contínuo para analisá-lo, criamos bordas artificiais no início e no fim do segmento. Se fôssemos analisar as frequências desse trecho diretamente (geralmente com uma operação chamada Transformada de Fourier), essas bordas abruptas introduziriam frequências falsas que não existem no sinal original, um fenômeno chamado de vazamento espectral (spectral leakage), que contamina a análise (OPPENHEIM; SCHAFER; BUCK, 1999).

Além disso, o áudio digital permite uma série de manipulações que não são possíveis com o áudio analógico. Por exemplo, o som pode ser editado, remixado e melhorado usando softwares de edição de áudio.

A compressão de áudio também é comum no áudio digital, onde o tamanho do arquivo é reduzido sem perda significativa de qualidade, permitindo um armazenamento e transmissão mais eficientes. O formato em que o arquivo de áudio está armazenado influencia diretamente na qualidade e na quantidade de informação presente no trecho de áudio, pois cada formato possui um tipo de compressão específico.

Para entender como isso funciona na prática, os formatos de áudio são divididos em três grandes grupos:

Sem Compressão: O exemplo mais famoso é o WAV (desenvolvido pela Microsoft e IBM). Ele salva o áudio exatamente como ele foi capturado pelo conversor analógico-digital,

bit por bit. Por não perder nenhuma informação, oferece a máxima qualidade possível, mas gera arquivos gigantescos. É o formato ideal para edição profissional e para criar datasets de pesquisa (como o ESC-50).

Com Compressão Sem Perdas *(Lossless)*: O formato mais conhecido é o FLAC. Ele funciona como um "arquivo ZIP" para áudio: consegue reduzir o tamanho do arquivo quase pela metade, mas, quando o computador vai tocar o som, ele descompacta o arquivo e recupera 100% da qualidade original, sem nenhuma perda de dados.

Com Compressão Com Perdas *(Lossy)*: É aqui que entram o MP3 e o OGG (muito usado no Spotify). Esses formatos usam algoritmos de psicoacústica para reduzir drasticamente o tamanho do arquivo (chegando a 10% do tamanho de um WAV). Para fazer esse milagre, o compressor joga fora os sons que o ouvido humano dificilmente consegue escutar — como frequências agudas demais ou sons muito baixos que acontecem logo após um estrondo muito alto.

Para projetos de classificação de sons por inteligência artificial, a escolha do formato é vital. Arquivos com compressão *lossy* (MP3/OGG) podem descartar frequências sutis que seriam valiosas para a rede neural identificar um ruído de fundo ou um detalhe acústico importante (BOSI; GOLDBERG, 2002).

2.4 Extratores de Características

Em termos simples, os extratores de características são algoritmos matemáticos projetados para funcionar como "filtros inteligentes". A função deles é ler um arquivo de dados brutos (como uma imagem pixel por pixel ou um sinal de áudio segundo por segundo), identificar quais padrões ali dentro realmente importam para resolver um problema e descartar todo o resto, que é considerado redundância ou ruído.

No processamento de áudio digital, a extração de características é uma etapa obrigatória. Um sinal de áudio bruto (no domínio do tempo) é uma sequência muito grande de oscilações de amplitude. Tentar alimentar uma inteligência artificial diretamente com esses milhares de pontos numéricos por segundo exige um poder computacional muito grande e pode confundir o algoritmo, porque muita informação ali não ajuda a diferenciar um som do outro. O extrator de características resolve isso: ele condensa o sinal, reduzindo sua dimensionalidade, e gera um vetor de características *(feature vector)* compacto que descreve a "identidade" daquele som.

2.4.1 *Zero Crossing Rates* (ZCR)

Zero Crossing Rates (ZCR) é o método para analisar a quantidade de vezes em que o sinal de um áudio passa pelo valor zero, ou seja, em um áudio temos o tempo e seu comprimento de onda, analisando o a onda do áudio, é possível ver quantas vezes o sinal passa pelo valor zero, indo de valores positivos para valores negativos, ZCR tem sido

estudado para reconhecimento de notas musicais e em alguns casos para reconhecimento de fala.

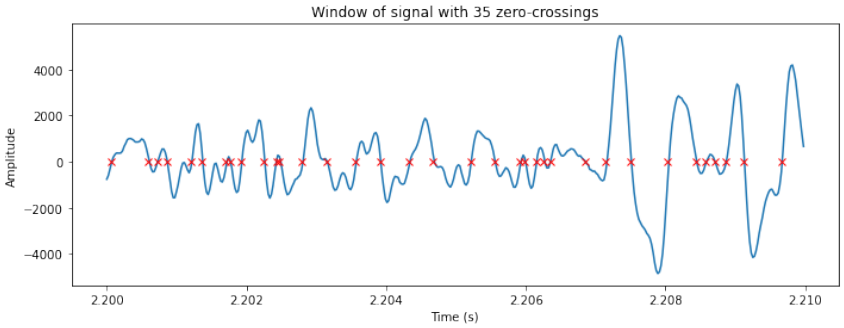
O ZCR é um dos extratores de características que será estudado.O ZCR possui aplicações em algumas áreas como por exemplo: **Classificação de sons:** O ZCR é uma métrica útil para diferenciar sons com características distintas. Sons com frequências altas, como ruídos ou sons percussivos, geralmente apresentam valores elevados de ZCR, enquanto sons mais graves ou tons sustentados tendem a apresentar valores mais baixos. **Reconhecimento de fala e música:** No reconhecimento de fala, o ZCR pode ajudar a distinguir entre vogais e consoantes, pois as consoantes geralmente têm taxas de cruzamento zero mais altas. Na música, o ZCR é usado para identificar padrões de notas ou instrumentação, tornando-se uma ferramenta em aplicações como síntese de som e análise de conteúdo musical. **Detecção de ruídos:** Por meio do ZCR, é possível identificar ruídos ou interferências em sinais de áudio, já que sons não harmônicos frequentemente possuem cruzamentos zero mais frequentes. **Segmentação de áudio:** Essa técnica pode ser utilizada para dividir sinais de áudio em segmentos, ajudando a identificar pontos de transição, como pausas na fala ou mudanças em trilhas musicais.

$$\text{zcr} = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{\mathbb{R}<0}\left(s_t s_{t-1}\right) \tag{1}$$

O Zero Crossing Rate pode ser utilizado de forma simples, somando o total de vezes em que o sinal do áudio passa pelo valor zero, depois fazendo a normalização pela amostragem . Analisando a composição da equação, é possível observar que o termo $\frac{1}{T-1}$ faz normalização a pelo número total de amostras, enquanto o termo $\sum_{t=1}^{T-1}$ realiza o somatório do total das amostras por onde existiu o cruzamento pelo zero. A função 1 avalia a alternância de sinal entre amostras consecutivas, assumindo o valor 1 sempre que ocorre um cruzamento pelo zero (mudança de polaridade positiva/negativa) e 0 quando o sinal permanece na mesma faixa.

Na Figura 3 podemos ver um exemplo de aplicação do ZCR.

Figura 3 – Ilustração de um Áudio no ZRC (*waveform*). Em vermelho todas as 35 ocorrências de cruzamento do valor zero, ou seja, zcr=1.



Fonte: Extraída de Bäckström et al. (2022).

2.4.2 Mel-frequency cepstral coefficients (MFCCs)

Os coeficientes cepstrais de frequência Mel (MFCCs) são parâmetros que compõem coletivamente uma representação de um sinal de áudio conhecida como cepstrum de frequência Mel. Eles são derivados de uma transformação não linear do espectro de um clipe de áudio, resultando em um "espectro de um espectro". A principal diferença entre o cepstrum de frequência Mel e o cepstrum tradicional é que, no primeiro, as bandas de frequência são igualmente espaçadas na escala Mel, que é uma escala perceptual que aproxima a resposta do sistema auditivo humano mais de perto do que as bandas de frequência linearmente espaçadas usadas no espectro tradicional.

Essa transformação de frequência permite uma melhor representação do som, sendo particularmente útil em aplicações como a compressão de áudio, onde pode reduzir a largura de banda de transmissão e os requisitos de armazenamento de sinais de áudio. Os MFCCs são amplamente utilizados como características em sistemas de reconhecimento de fala, como aqueles que podem reconhecer automaticamente números falados em um telefone.

Além disso, os MFCCs estão encontrando cada vez mais aplicações em sistemas de recuperação de informações musicais, como a classificação de gêneros musicais, medidas de similaridade de áudio, entre outros.

Para vizualização do MFCC usaremos os espectro de áudio, como o exemplo ilustrado na Figura 4:

Figura 4 – Ilustração de um som através do seu formato de onda (*waveform*).

Fonte: Extraída de Blocks (2026).

2.5 Técnicas de Projeção Multidimensional

As técnicas de projeção multidimensional são essenciais para a análise e visualização de dados complexos, frequentemente encontrados em diversas áreas como bioinformática, economia, e aprendizado de máquina. Esses métodos visam reduzir a dimensionalidade dos dados, preservando o máximo de informações significativas possível. Entre as principais

técnicas utilizadas, destacam-se a Análise de Componentes Principais (PCA), a Análise de Componentes Independentes (ICA), e o t-Distributed Stochastic Neighbor Embedding (t-SNE).

2.5.1 Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (PCA) é uma técnica estatística amplamente utilizada para a redução de dimensionalidade de conjuntos de dados complexos. O objetivo principal do PCA é transformar um grande conjunto de variáveis possivelmente correlacionadas em um conjunto menor de variáveis não correlacionadas, chamadas de componentes principais, enquanto retém a maior parte da variabilidade presente nos dados originais.

A seguir temos os passos que o PCA realiza até a visualização dos dados.

Padronização dos Dados: Antes de aplicar o PCA, é essencial padronizar os dados, especialmente se eles possuem diferentes unidades de medida. A padronização implica subtrair a média e dividir pelo desvio padrão para cada variável, garantindo que todas as variáveis tenham a mesma escala.

Cálculo da Matriz de Covariância: A matriz de covariância é calculada para entender como as variáveis do conjunto de dados se relacionam umas com as outras. A covariância indica a direção e a magnitude da relação linear entre duas variáveis.

Determinação dos Autovalores e Autovetores: A partir da matriz de covariância, são calculados os autovalores e autovetores. Os autovalores indicam a quantidade de variância explicada por cada componente principal, enquanto os autovetores determinam a direção de cada componente principal no espaço original dos dados.

Formação das Componentes Principais: As componentes principais são formadas ordenando-se os autovalores de forma decrescente. Cada componente principal é uma combinação linear das variáveis originais, ponderada pelos autovetores correspondentes. As primeiras componentes principais retêm a maior quantidade de variação dos dados.

Transformação dos Dados: Finalmente, os dados originais são transformados para o novo espaço de componentes principais, reduzindo a dimensionalidade ao selecionar apenas as primeiras componentes principais que explicam a maior parte da variância dos dados.

A Análise de Componentes Principais é uma ferramenta poderosa e versátil para a simplificação e interpretação de dados complexos. Sua capacidade de reduzir a dimensionalidade mantendo a maior parte da variância torna-a uma escolha popular em diversas áreas, desde a bioinformática até a análise de mercados financeiros.

2.5.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

O t-Distributed Stochastic Neighbor Embedding (t-SNE) é uma técnica de redução de dimensionalidade especialmente projetada para a visualização de dados em alta

dimensionalidade. Desenvolvida por Laurens van der Maaten e Geoffrey Hinton em 2008, essa técnica é amplamente usada em aplicações de aprendizado de máquina, bioinformática e análise de dados, devido à sua capacidade de preservar a estrutura local dos dados ao mapeá-los para um espaço de baixa dimensionalidade.

Agora os passos do t-SNE para entender como ele funciona até a representação da visualização dos dados:

Distribuição Probabilística dos Pontos no Espaço Original: O t-SNE começa calculando uma distribuição probabilística para a proximidade dos pontos no espaço original de alta dimensionalidade. Isso é feito para cada ponto, baseando-se na similaridade entre pontos vizinhos próximos, geralmente usando uma distribuição gaussiana.

Distribuição Probabilística no Espaço de Baixa Dimensionalidade: Em seguida, o t-SNE define uma distribuição similar para a proximidade dos pontos no espaço de baixa dimensionalidade, mas utiliza uma distribuição t (*t-distribution*) ao invés de uma gaussiana. Essa escolha ajuda a capturar a estrutura local dos dados de maneira mais eficaz, especialmente em altas dimensões.

Minimização da Divergência KL: O t-SNE então minimiza a divergência de Kullback-Leibler (KL) entre as duas distribuições probabilísticas. A minimização é feita por meio de métodos de otimização iterativa, ajustando as posições dos pontos no espaço de baixa dimensionalidade para preservar a estrutura local e as similaridades dos dados originais.

Vantagens na utilização do t-SNE. Visualização de Estruturas Locais: O t-SNE é excepcional na preservação das relações locais entre os pontos, tornando-o ideal para visualizar clusters e subestruturas em dados complexos. Redução Não-Linear da Dimensionalidade: Diferente de métodos lineares como PCA, o t-SNE é capaz de capturar relações não-lineares entre os dados, oferecendo uma representação mais fiel de sua estrutura intrínseca. Flexibilidade: O t-SNE pode ser usado em uma ampla variedade de tipos de dados e tem se mostrado eficaz em situações como a visualização de dados de expressão gênica, análise de imagens, e representação de palavras *embeddings* de palavras em processamento de linguagem natural.

Desvantagens e Limitações:

Complexidade Computacional: O t-SNE é computacionalmente mais intensivo do que métodos lineares, o que pode ser um problema para conjuntos de dados muito grandes. Sensibilidade a Hiperparâmetros: A qualidade da projeção gerada pelo t-SNE pode ser sensível à escolha de hiperparâmetros, como a perplexidade e a taxa de aprendizado, exigindo experimentação cuidadosa.

Aplicações Comuns: Análise de Dados de Alta Dimensionalidade: O t-SNE é frequentemente usado para explorar e visualizar dados de alta dimensionalidade, como dados de expressão gênica em bioinformática ou características de imagens em aprendizado de máquina. Identificação de Clusters: Em muitos casos, o t-SNE ajuda a identificar e visu-

alizar clusters ou grupos naturais dentro de um conjunto de dados, sendo particularmente útil em análises exploratórias.

O t-Distributed Stochastic Neighbor Embedding (t-SNE) é uma ferramenta poderosa para a visualização e análise de dados de alta dimensionalidade, oferecendo uma visão detalhada das estruturas locais e das similaridades intrínsecas dos dados. Embora tenha algumas limitações, suas vantagens na preservação de estruturas não-lineares e na visualização clara de clusters tornam-no uma técnica indispensável em muitas áreas de pesquisa e aplicação prática.

As técnicas de projeção multidimensional desempenham um papel crucial na compreensão e interpretação de dados complexos. A escolha da técnica mais apropriada depende da natureza dos dados e do objetivo específico da análise. Com o avanço contínuo das tecnologias de computação, espera-se que novas e mais sofisticadas técnicas de projeção surjam, proporcionando ainda mais precisão e *insight* na análise de dados multidimensionais.

3 TRABALHOS RELACIONADOS

O trabalho de Chu, Narayanan e Kuo (2009) aborda a importância da extração de características de sons ambientais, por ser uma área relevante. Nele é abordado diferentes técnicas, como MFCC. O artigo propõe o uso do algoritmo *Matching Pursuit* (MP) por ser eficaz no domínio tempo-frequência, enquanto o MFCC é usado como complemento por ser mais preciso em reconhecimento de sons ambientais.

Os autores utilizam *features* amplamente utilizadas, como o MFCCs e o ZCR. O MFCC, popular por modelar aspectos da percepção auditiva humana e descrever a forma espectral, são eficazes para sons estruturados como fala e música. No entanto, Chu, Narayanan e Kuo (2009) apontam que sua performance pode ser limitada na presença de ruído ou para sinais com espectro plano, característicos de muitos sons ambientais. O ZCR, por sua vez, é uma medida temporal que quantifica a frequência com que o sinal cruza o eixo zero, sendo útil para distinguir certas classes de sons, mas limitada para outras com características temporais semelhantes.

Em busca de superar as limitações das *features* de espectro ou temporais, Chu, Narayanan e Kuo (2009) propõem o uso do algoritmo MP como uma ferramenta para extrair *features* eficazes no domínio tempo-frequência. A abordagem MP decompõe o sinal de áudio utilizando um dicionário de funções elementares, chamadas átomos (no estudo, foram utilizados átomos de Gabor), selecionando iterativamente aqueles que melhor representam a estrutura do sinal. As *features* propostas são derivadas dos parâmetros (como escala e frequência) dos átomos mais significativos selecionados pelo MP. Segundo os autores, essa técnica resulta em um conjunto de *features* flexível, interpretável e capaz de capturar características tempo-frequência que o MFCC isoladamente não consegue, oferecendo uma representação mais completa para sons ambientais não estruturados.

Os resultados experimentais apresentados no artigo, conduzidos em 14 classes distintas de sons ambientais, corroboram a proposta. A análise comparativa demonstrou que, embora as *features* MP isoladamente já apresentassem um desempenho geral superior ao dos MFCCs em diversas classes, a combinação das *features* MP com os MFCCs (MP+MFCC) produziu a maior acurácia de 83,9%. Este desempenho da abordagem combinada foi significativamente superior ao uso isolado das *features* e também ao uso de um conjunto maior de *features* convencionais, indicando a forte complementaridade entre a representação espectral do MFCC e a representação tempo-frequência do MP. Adicionalmente, o desempenho do sistema automático com *features* combinadas mostrou-se comparável ao de ouvintes humanos na mesma tarefa de classificação.

De forma resumida:

- MFCC obteve desempenho variável, sendo bons para algumas classes mas extremamente ruins para outras (taxa de reconhecimento de 0% para 4 classes);

- Features MP tiveram um desempenho geral melhor e mais consistente.
- A combinação MP junto MFCC obteve a maior precisão média (83.9% com GMM), mostrando que as features se complementam.
- Usar todas as features juntas resultou em desempenho pior do que apenas MP+MFCC, reforçando que mais features não é sempre melhor.

Já o trabalho de Karol (2015) realizou a criação de uma base de dados (*dataset*), para classificação de sons ambientais, o *dataset* possui 2.000 áudios curtos, esses áudios são divididos em 50 classes, sons comuns, como ondas do mar, cachorro, aves e etc. O *dataset* foi construído através do projeto Freesound. O artigo também avalia a precisão humana na classificação desses sons do *dataset* e compara com o desempenho do algoritmo na classificação utilizando extratores de características como MFCC e ZCR. Foram utilizadas três técnicas de classificação supervisionada para treinamento: kNN, Random Forest e SVM.

Criação de Dataset de Referência (ESC-50 e ESC-10)

O artigo de Karol (2015) introduz o dataset `ESC-50`, composto por 2.000 gravações ambientais rotuladas, balanceadas entre 50 classes (40 clipes por classe). As classes são agrupadas em 5 categorias principais: sons de animais, paisagens sonoras naturais e sons de água, sons humanos (não fala), sons internos/domésticos e ruídos externos/urbanos. Nessa base o artigo obteve as seguintes acurácias: kNN = 32,2%, random forest = 44,3% e SVM = 39,6%, evidenciando a dificuldade de obter bons resultados com uma grande quantidade de classes.

Também é apresentado o `ESC-10`, um subconjunto do `ESC-50` com 10 classes, concebido como um *benchmark* simplificado, representando sons transientes/percussivos, eventos sonoros com forte conteúdo harmônico e ruídos/paisagens sonoras estruturadas. Nessa base o artigo obteve as seguintes acurácias: kNN = 66,7%, random forest = 72,7% e SVM = 67,5%, melhorando os resultados para uma quantidade menor de classes.

A criação desses *datasets* públicos é uma contribuição significativa, pois a maioria dos estudos anteriores utilizava *datasets* específicos, pequenos ou proprietários, dificultando a comparabilidade e reprodutibilidade das pesquisas na área.

Análise da taxa de acerto da audição Humana como base: O estudo estabelece uma estimativa da capacidade humana no reconhecimento dos sons do *dataset*, servindo como um ponto de referência para sistemas de classificação automática. A acurácia média humana alcançada foi de 95,7% para o *dataset* `ESC-10` e 81,3% para o `ESC-50`.

Observou-se que a dificuldade de classificação variava entre os tipos de eventos sonoros, com categorias como sons humanos e de animais sendo mais fáceis, e paisagens sonoras e ruídos mecânicos sendo mais difíceis para os participantes humanos.

4 MATERIAIS E MÉTODOS

Neste capítulo são indicados: a base de dados utilizada; os processamentos feitos na base; e todas as implementações utilizadas para a análise dos dados.

4.1 Características de áudio

Para a realização deste trabalho, são extraídas algumas características dos áudios, que objetivam descrever o áudio de uma forma mais compacta e que possibilite distinguir bem entre as diferentes classes disponíveis, para que assim seja possível fazer a classificação.

Para extração de características foram utilizados: Zero-crossing rate (ZCR), para calcular o número médio de vezes que o sinal muda em um curto prazo; Mel-Frequency Cepstral Coe (MFCC). Serão utilizados também técnicas de visualização de dados multidimensionais, para investigar o comportamento dos diferentes extratores, através da visualização da dissimilaridade dos vetores de características extraídos. Dessa forma, espera-se contribuir para a identificação de quais descritores são mais adequados para espécies dessa ordem. Esses são alguns extratores de características encontrado durante pesquisas, cada método de extração de características tem funcionalidades diferentes.

Dos métodos acima, devemos estudar quais são melhores para a aplicação, o MFCC por exemplo se destaca em reconhecimento de fala. Para obter o maior número de informação possível, precisamos trabalhar com todo som possível, ou seja, em um áudio de 5 segundos, que está classificado como sons de onda do mar, pode conter também som de alguma ave.

4.2 Base de dados

O conjunto de dados *Environmental Sound Classification - ESC* de Karol (2015) é uma coleção de gravações ambientais curtas, cada áudio do conjunto de dados possui a duração de 5 segundos. Cada áudio possui taxa de amostragem de 44100 Hz, canal único, utilizando codec Ogg Vorbis e taxa de bits (*bitrate*) de 192 kbps. Todos os clipes foram extraídos e rotulados manualmente por Karol (2015) a partir de gravações de campo públicas disponíveis através do projeto Freesound.org.

Para o trabalho foram usados as bases de dados do ESC-2, ESC-10 e ESC-12, cada base de dados possuindo áudios de duração de 5 segundo e no formato WAV.

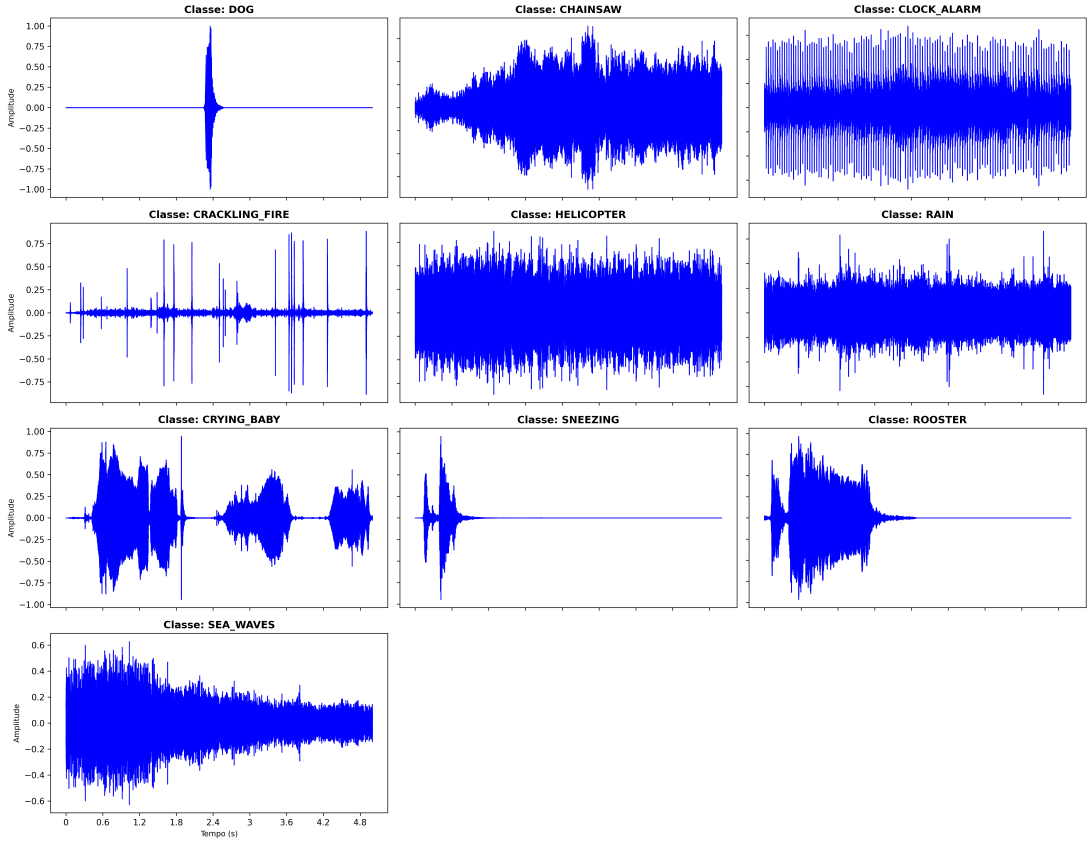
Começando pelo ESC-2, é uma base de dados que possui 80 áudios, dividido em duas classes, *Sea Waves* e *Dog*, que são sons de ondas do mar e sons de latidos de cachorros respectivamente.

4.2.1 ESC-10

O `ESC-10` é uma base de dados com 400 gravações de sons ambientais (10 classes, 40 clipes por classe) sendo um subconjunto do `ESC-50`, essas classes são divididas em 3 grupos. O primeiro grupo temos sons que são sons percussivos, sons que acontecem repentinamente ou sons com padrões significativos, as classes presente são, espirro (*sneezing*), latido de cachorro (*dog barking*) e tic tac do relógio (*clock ticking*). O segundo grupo são evento de sons que possuem um forte conteúdo harmônico, bebê chorando *(*crying baby) e canto do galo (*crowing rooster*). O terceiro e último grupo dessa base de dados é composta por sons da natureza e barulhos, chuva (*rain*), ondas do mar (*sea waves*), estalo de fogo (*fire crackling*), helicóptero (*helicopter*) e serra elétrica (*chainsaw*).

Na Figura 5 temos a representação de *waveform* de um áudio de cada classe presente na base de dados ESC-10 usada no trabalho.

Figura 5 – Imagem que ilustra o waveform de cada classe de áudio presente na base de dados ESC-10.



Fonte: Autoria própria.

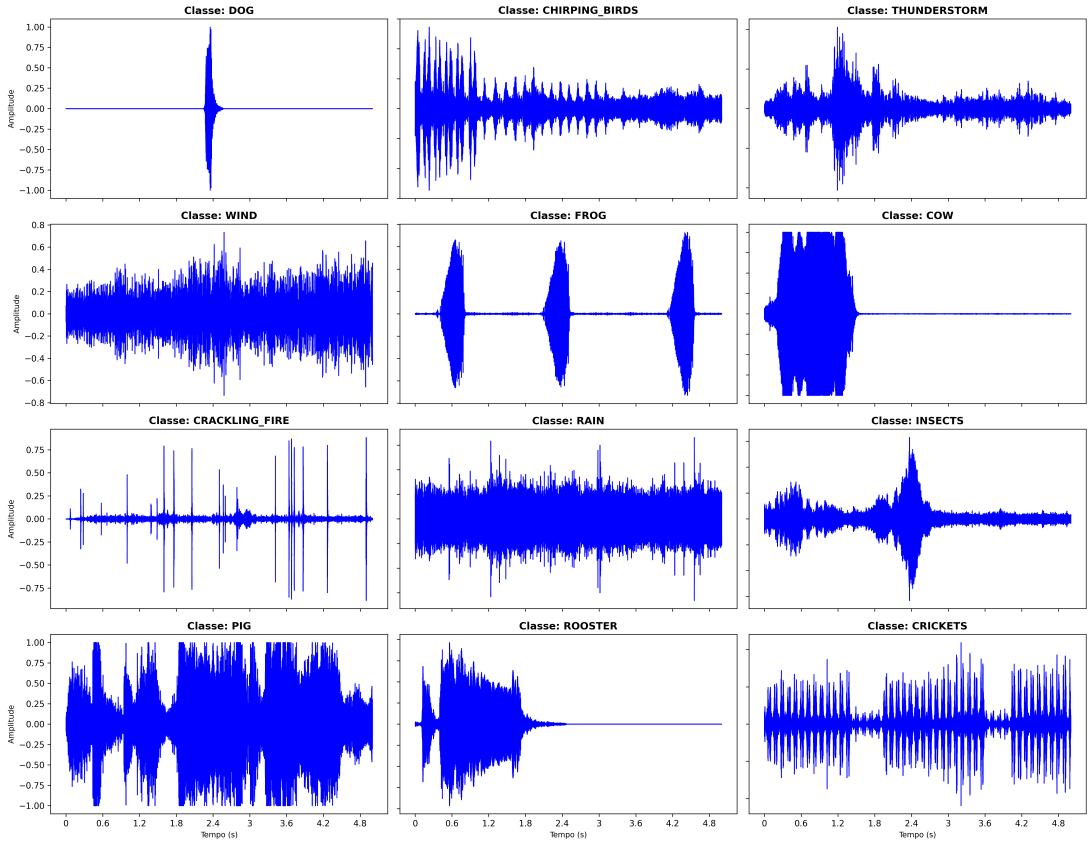
4.2.2 ESC-12

O `ESC-12` é a base de dados com um acréscimo de algumas classes para doze, mas essa base de dados não usa todas as classes da base de dados anterior, o número de arquivos de áudio também são maiores, sendo no total 480 áudio para essa base de dados.

As classes do `ESC-12` são: *Crackling Fire* (Algo como sons de estalo de fogo, quando está queimando algo), *Crickets* (Grilo), *Insects* (Insetos), *Frog* (Sapo), *Cow* (Vaca), *Rain* (Chuva), *Chirping Birds* (Canto dos Pássaros), *Rooster* (Galo), *Pig* (Porco), *Wind* (Vento), *Thunderstorm* (Tempestade), *Dog* (Cachorro).

Na Figura 6 temos a representação de *waveform* de um áudio de cada classe presente na base de dados ESC-12 usada no trabalho.

Figura 6 – Imagem que ilustra o waveform de cada classe de áudio presente na base de dados ESC-12.



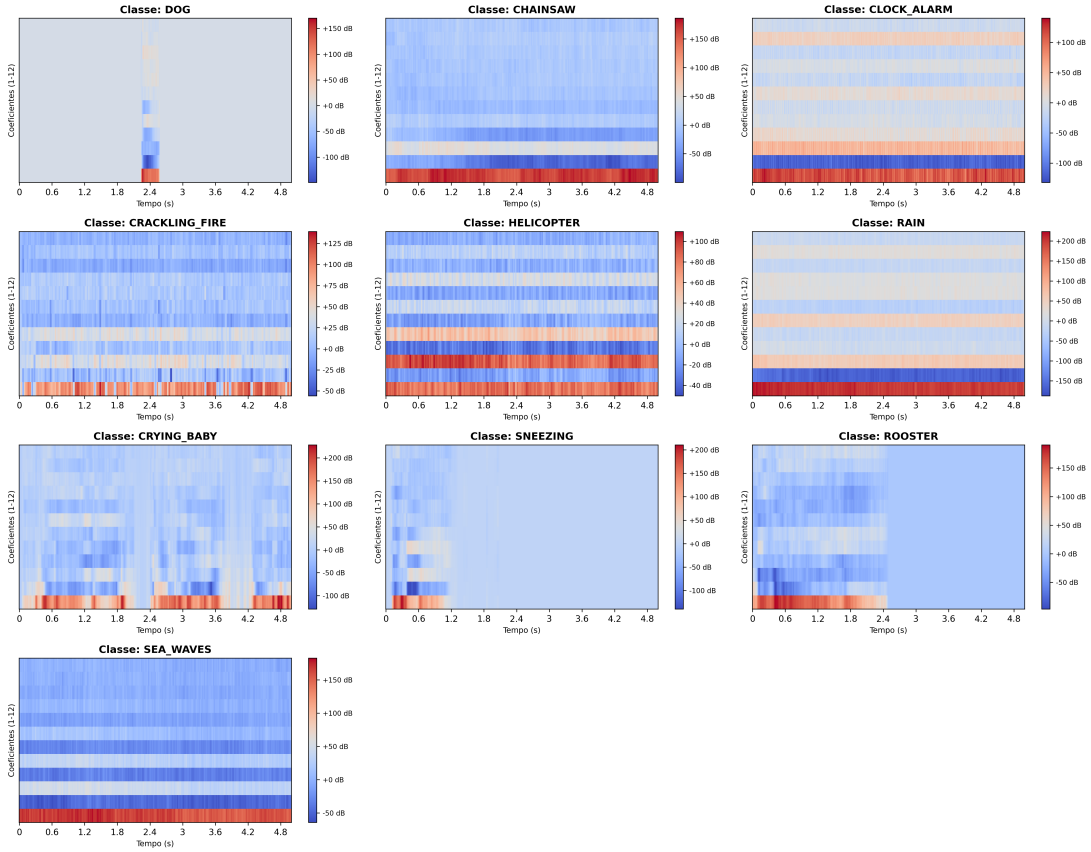
Fonte: Autoria própria.

4.3 Vetor de características e projeção multidimensional

Para realizar a visualização dos dados e o treinamento será utilizado um vetor de características composto pelo MFCC e o ZCR. A Implementação do MFCC no trabalho será definida o mesmo número de coeficientes do trabalho de Karol (2015), o MFCC será calculado com 13 coeficientes, e também para seguir a mesma metodologia do trabalho citado, será descartado o primeiro coeficiente, o autor descarta o coeficiente zero, pois, segundo ele garante que o classificador avalie as características reais do áudio, já que o coeficiente 0 sofre variações no volume que podem causar pertubação na análise, então na prática após a realização do cálculo o MFCC passa a ter 12 vetores com 216 colunas.

Na Figura 7 temos a representação de cada áudio que foi apresentado na Figura 5, porém a representação do MFCC.

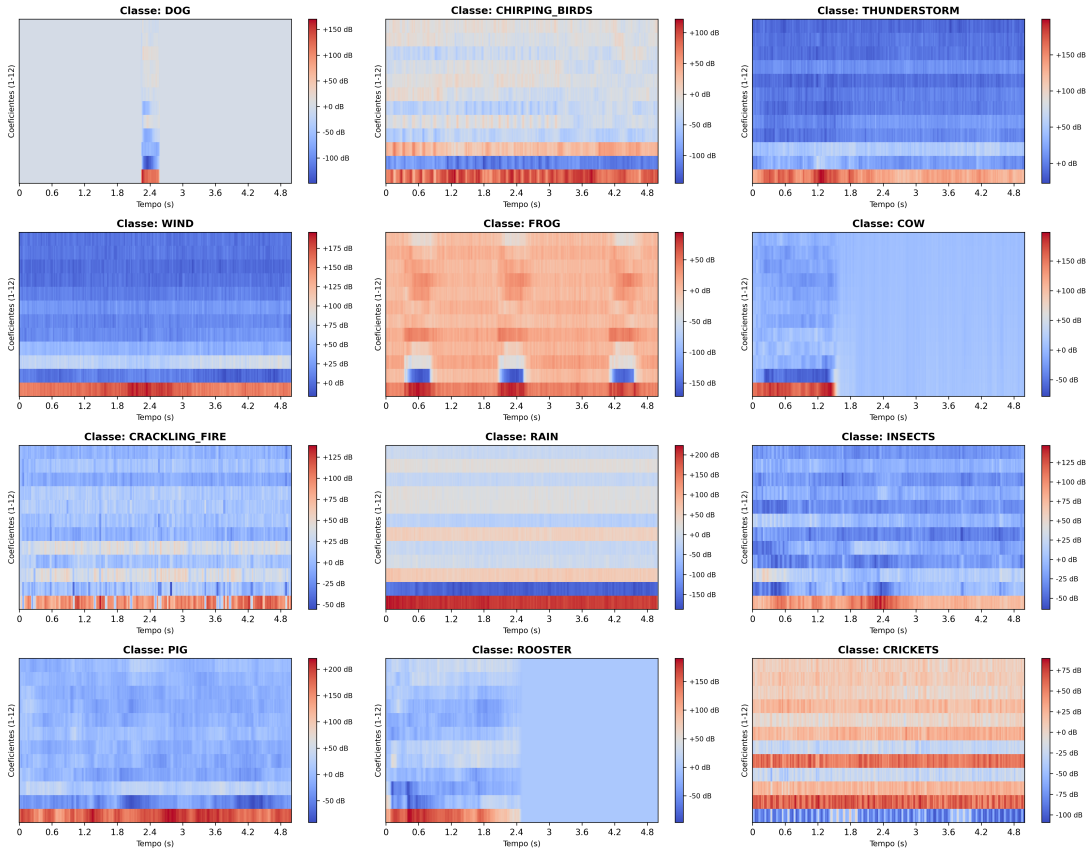
Figura 7 – Imagem que ilustra o MFCC de cada classe de áudio presente na base de dados ESC-10.



Fonte: Autoria própria.

Na Figura 8 temos a representação de cada classe presente na base de dados ESC-12.

Figura 8 – Imagem que ilustra o MFCC de cada classe de áudio presente na base de dados ESC-12.

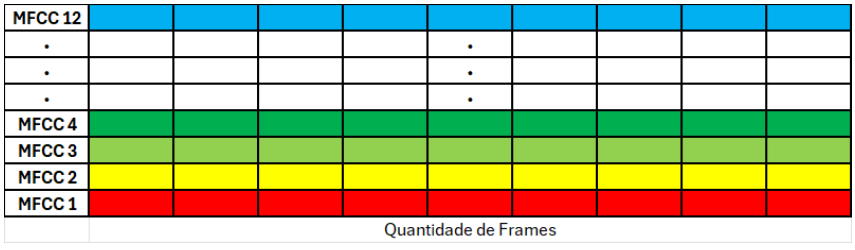


Fonte: Autoria própria.

O MFCC passará por 2 cálculos, um vai calcular a média do MFCC, todas as features serão calculadas e permanecerão com o mesmo tamanho, o outro calculo será para calcular seu desvio padrão, que também mantém o tamanho do MFCC. Após concluído as duas etapas, teremos 2 MFCC, MFCC média e MFCC desvio padrão.

Na Figura 9 é possivel observar como é feito a montagem do vetor de características, esse ilustração mostra como fica o MFCC bruto.

Figura 9 – Ilustração de como é feita a montagem dos vetores de características: Cada linha representa um coeficiente do MFCC e cada coluna um frame do áudio.

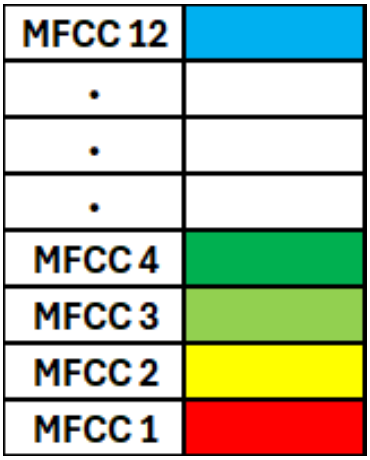


Fonte: Autoria própria.

Na Figura 10 podemos ver a ilustração da transformação do vetor do MFCC após

a realização dos cálculos de média e desvio padrão.

Figura 10 – Ilustração dos vetores de características de um áudio, onde cada coeficiente (linha) é resumida a um único valor (por exemplo: média, desvio padrão) ao longo de todos os seus frames.



Fonte: Autoria própria.

O ZCR também será calculado com média e desvio padrão, porém, como ele não tem os coeficiente a estrutura dele é um pouco diferente, o ZCR terá um vetor de cada, um vetor de média e um vetor de desvio padrão.

Para a visualização foi utilizada a técnica de projeção multidimensional t-SNE (MAATEN; HINTON, 2008). Foi utilizada a implementação disponível na biblioteca scikit-learn versão 1.4.2. A técnica de projeção receberá os vetores de características multidimensionais que foram construídos em forma de uma matriz, os 12 vetores do MFCC média com 12 vetores do MFCC desvio padrão, somados com um vetor do ZCR média e um vetor do ZCR desvio padrão, totalizando uma matriz com 26 colunas e 400 linhas, onde cada linha representa um áudio diferente.

Na Figura 11 é ilustrado como ficam os vetores de características de cada áudio do *dataset*.

Figura 11 – Representação dos vetor de características construídos, contendo 26 dimensões: 2 dimensões para ZCR, 12 dimensões para média dos MFCCs e 12 dimensões para desvio padrão dos MFCCs.

	MFCC_MÉDIA_1	...	MFCC_MÉDIA_12	...	MFCC_STD_1	...	MFCC_STD_12	ZCR_MÉDIA	ZCR_STD
Áudio_1									
Áudio_2									
Áudio_3									
.									
.									
.									
Áudio_n-1									

Fonte: Autoria própria.

4.3.1 Silhueta

O Coeficiente de Silhueta (*Silhouette Score*) é uma métrica quantitativa usada para avaliar a qualidade de um agrupamento (*clustering*), medindo o quão bem cada amostra de áudio foi pontuada em sua respectiva classe. O cálculo gera um índice que varia de -1 a 1 para cada ponto, baseado na relação entre a coesão (a distância média do ponto em relação aos elementos do seu próprio grupo) e a separação (a distância média em relação ao grupo vizinho mais próximo). Na prática, um resultado próximo a 1 indica que os extratores de características agruparam os sons de forma ideal e bem isolada, valores próximos a 0 revelam sobreposição de classes na fronteira, e valores negativos apontam que o som foi classificado no grupo incorreto.

4.3.2 Normalização

As *features* dos vetores de características tem valores em intervalos de valores diferentes (MFCC médio x desvio padrão; zcr médio x desvio padrão), então para lidar isso é feito um processo de normalização dos valores das *features* antes da projeção. São feitas duas normalizações, ambas com implementação na Scikit-learn: `Standard scaler` e `Robust scaler`.

`Standard scaler` normaliza os dados utilizando o z-score, de forma que após a normalização cada *feature* será ajustada para uma distribuição centrada na média dos valores e com desvio padrão controlado, mais especificamente, uma distribuição normal de média 0 e desvio padrão 1: $N(0,1)$. Sua equação é mostrada na Equação 2:

$$v_{normalizado} = \frac{v - \mu}{\sigma}, \tag{2}$$

onde μ é o valor média da *feature* e σ seu desvio padrão.

`Robust scaler` parte de uma ideia semelhante mas utilizando a mediana e o intervalo interquartil (IQR), com o objetivo de preservar possíveis anomalias (*outliers*). Sua equação é mostrada na Equação 3:

$$v_{normalizado} = \frac{v - m}{IQR}, \tag{3}$$

onde m é o valor mediana da *feature* e IQR seu valor de intervalo interquartil entre os quartis 3 e 1 (Q3 e Q1), região que concentra 50% dos dados.

4.3.3 Treinamento supervisionado

Além da projeção convencional dos vetores de características, neste trabalho analisamos como ficam os *clusters* dos áudios após uma etapa de treinamento supervisionado. Para isso, foi utilizada a metodologia proposta por Rauber et al. (2016), que constrói um vetor multidimensional a partir dos valores das ativações dos neurônios em uma

determinada camada. Isso possibilita visualizar quão separados estão os *clusters* em cada época do treinamento.

Para o treinamento foi utilizada uma rede neural perceptron multicamadas *Multilayer Perceptron* (MLP), com as seguintes características: A MLP possui 3 camadas, a primeira camada possui 256 neurônios, sua segunda camada possui 56 neurônios, a teceira e última camada possui 32 neurônios. As 3 camadas presente na MLP estão usando a ativação Unidade Linear Retificada ou *Rectified Linear Unit* (ReLU), a camada de saída usa a ativação *softmax*, usada para problemas que possuem múltiplas classes, como é o caso do trabalho. Essa configuração utilizada no trabalho foi a que obteve os melhores resultados, também foram testadas outras configurações de neurônios como por exemplo, camada 1 com 128 neurônios, camada 2 com 64 neurônios e camada 3 com 16 neurônios, esse foi um exemplo de testes feito na MLP, mas foram realizados outros testes com outras configurações, porém com resultados inferiores.

Para realizar a visualização das imagens geradas, foi utilizada a penúltima camada da MLP.

4.4 Linguagem de programação e bibliotecas

A linguagem de programação usada foi o Python, a versão usada no trabalho foi a 3.12. A seguir será mostrado quais foram as bibliotecas que foram utilizadas para a construção do trabalho apresentado.

Librosa é uma biblioteca Python para análise de música e áudio. Ele fornece todo o processamento de áudio para a analise, a Librosa permite extrair características como coeficiente da escala Mel(MFCC), além do uso do MFCC com a Librosa, o trabalho também usará a extração de características com ZCR. A Versão da biblioteca Librosa utilizada para o trabalho foi a 0.10.1.

O NumPy *Numerical Python* é uma biblioteca em Python que permite manipular arrays e matrizes multidimensionais. As vantagens em usar o NumPy se deve pelo fato dele ser mais eficiente com a memória e sua velocidade de processamento, que é mais rápido que a solução nativa do Python. A versão do NumPy usada para a o trabalho foi a 1.26.4.

O Matplotlib é uma biblioteca do Python para realizar a visualização de dados, para a aplicação do matplotlib para o trabalho é fornecido uma coordenada, assim ele cria um gráfico em 2D para que seja feito a visualização dos dados para comparação. A versão do matplotlib utilizada do trabalho foi a 3.8.4.

O Seaborn é uma biblioteca Python para a visualização de dados, similar e baseada no matplotlib, porém ela possui algumas diferenças, sua estética é um pouco melhor, com uma paleta de cores diferentes, e a vantagem maior é o foco em estatística comparado com o matplotlib, para o trabalho a versão usada foi a 0.13.2.

O Scikit-learn (também conhecida como sklearn) é uma biblioteca Python comumente usada para tarefas relacionada ao aprendizado de máquina, no trabalho porém,

foi usada algumas ferramentas como o sklearn silhouette, ele consiste em pegar os dados gerados para a visualização e calcular uma média de agrupamento das classes, assim dando uma pontuação para poder classificar o resultado obtido. Outra ferramenta que está presente no sklearn é o sklearn manifold, nele tem a ferramenta t-SNE que foi usada no trabalho para fazer a visualização dos dados e o cálculo de silhueta. A versão usada para o trabalho foi a 1.4.2

A Plotly é uma das bibliotecas mais poderosas do Python para a criação de gráficos interativos e de qualidade profissional. Ao contrário de bibliotecas mais tradicionais como a Matplotlib ou Seaborn — que geram gráficos estáticos, a Plotly foca na experiência do usuário e na exploração de dados através da interatividade. Ela serve para criar gráficos onde o usuário pode passar o mouse por cima para ver os valores exatos (tooltips), dar zoom em áreas específicas, isolar dados clicando na legenda e arrastar o gráfico para mudar a perspectiva. O plotly permite construir desde gráficos simples (barras, linhas, dispersão) até visualizações altamente complexas, como mapas geográficos interativos, gráficos 3D, diagramas de redes e gráficos financeiros. A versão do plotly usada no trabalho é a 6.7.0.

5 Resultados

5.1 Visualização dos resultados

A visualização dos resultados é feita de duas formas: uma quantitativa, através das métricas de acurácia e da silhueta dos *clusters*; e outra qualitativa, através das projeções multidimensionais com a t-SNE.

Para validar os experimentos, inicialmente foi feito o cálculo da acurácia do ESC-10 utilizando a MLP proposta nesse trabalho, com o objetivo de verificar ela se aproxima dos valores obtidos por Karol (2015). O resultado desse experimento pode ser visto na Tabela 1, na linha *ESC10 - Matheus*, onde a acurácia obtida foi de 77,5%, sendo até um pouco superior ao melhor resultado obtido por Karol (2015), que foi através do kNN, indicado na primeira linha dessa Tabela.

Como esses resultados aproximam a MLP proposta do treinamento efetuado por Karol (2015), foi efetuado um novo experimento para entender como seria o comportamento dessa mesma MLP nas classes de áudio selecionadas para esse trabalho que constituíram a ESC12. A acurácia obtida pode ser vista na última linha da Tabela 1, indicando que essas classes são um pouco mais difíceis de classificar do que a ESC10 de Karol (2015).

Tabela 1 – Resultados da métrica de acurácia nas bases ESC-10 e ESC-12. Em negrito os melhores resultados, por base de dados.

	kNN	Random Forest	SVM	MLP
ESC10 (KAROL, 2015)	72,70%	44,3%	67,5%	-
ESC10 - Matheus	-	-	-	**77,50%**
ESC12 - Matheus	-	-	-	**68,75%**

Fonte: Autoria própria.

Na Tabela 2 são indicados os valores de silhuetas obtidos após as projeções, indicando o quão bem agrupados e distintos ficaram os *clusters*, tanto para o ESC10 quanto para o ESC12.

Tabela 2 – Resultados do cálculo da silhueta nas bases de dados ESC-10 e ESC-12 após as projeções com a TSNE, antes e após o treinamento com MLP. Em negrito os melhores resultados.

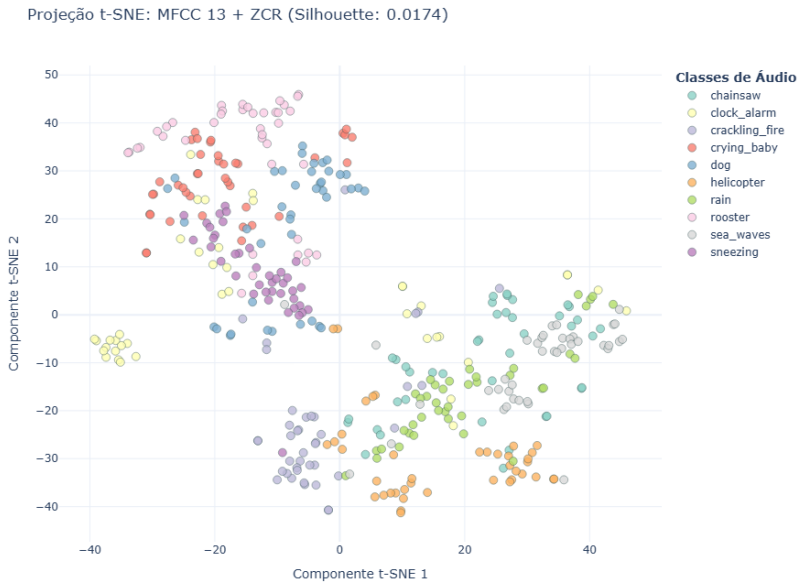
	Scaler	Vetor de características	Ativações da MLP (treino)
ESC10	Standard	**0,0174**	0,3770
ESC10	Robust	-0,0063	**0,4511**
ESC12	Standard	**0,0063**	0,3015
ESC12	Robust	0,0038	**0,3485**

Fonte: Autoria própria.

5.1.1 Projeções do ESC-10

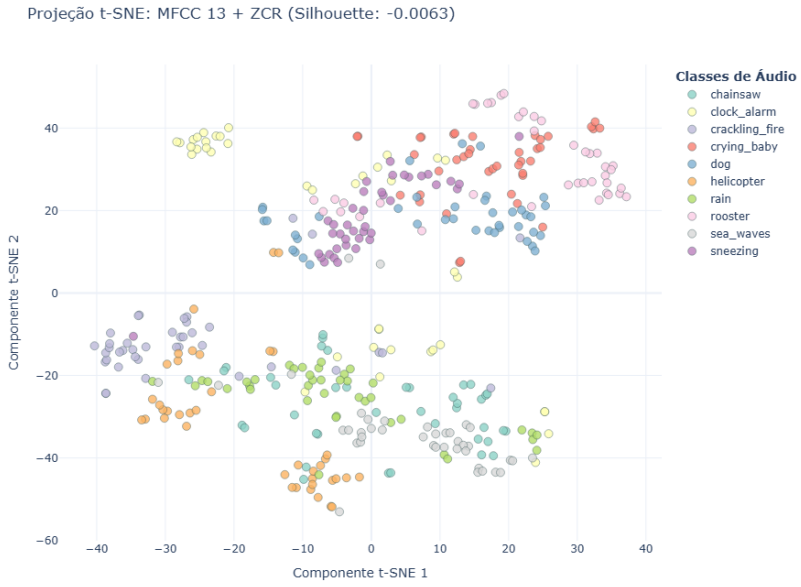
O primeiro resultado obtido foi com a base de dados ESC-10.Como é possível observar nas Figuras 12 e 13, usando o *Standard Scaler* e *Robust Scaler* as classes ficaram muito misturadas, somente o cálculo do MFCC e ZCR não foi suficiente para realizar a distinção das classes, o que pode ser visto nos seus valores de silhuetas próximos de zero.

Figura 12 – Projeção base de dados ESC-10 com *Standard Scaler* utilizando t-SNE.



Fonte: Autoria própria.

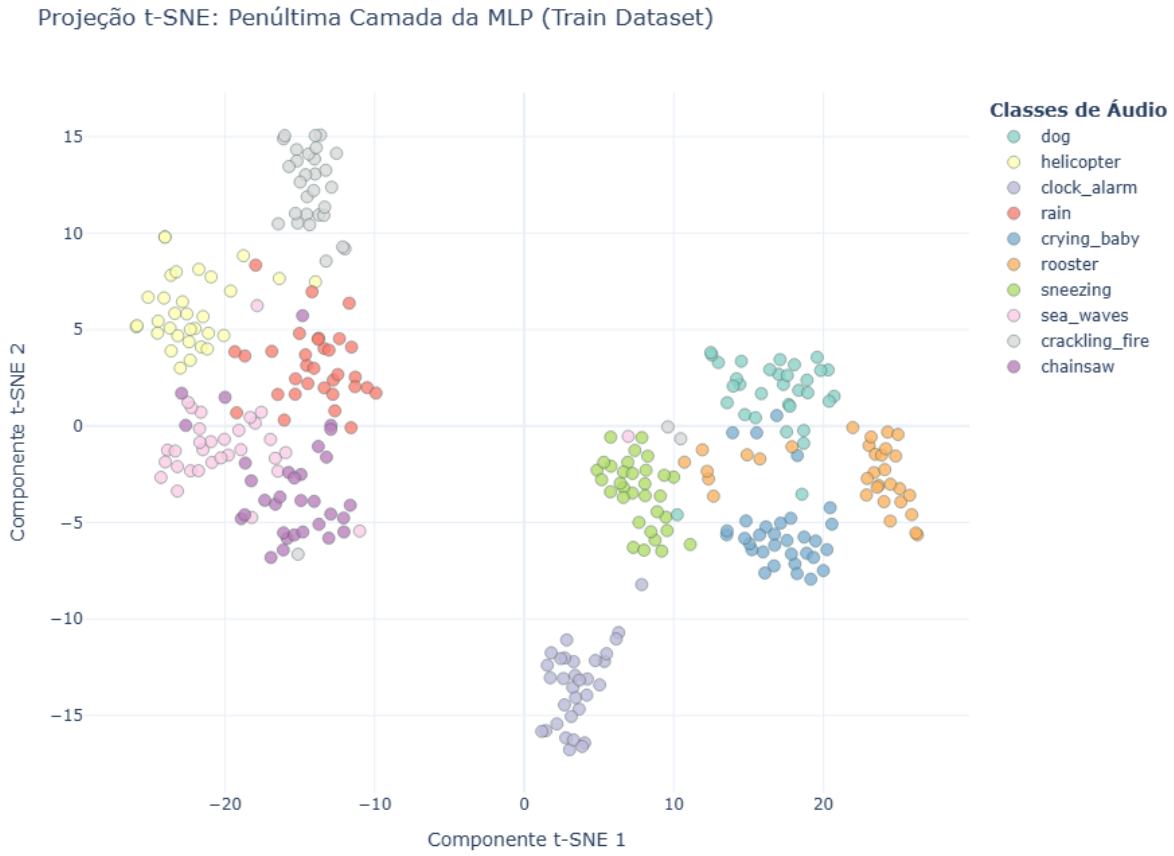
Figura 13 – Projeção base de dados ESC-10 com *Robust Scaler* utilizando t-SNE.



Fonte: Autoria própria.

Na Figura 14 gerada após o treinamento com a MLP, é possível observar que as classes se separam melhor, facilitando a visualização do agrupamento das classes. Analisando um caso mais específico, podemos ver por exemplo, que no agrupamento de sons da classe *rain* aparece um áudio da classe *chainsaw* no meio deles, ouvindo o áudio é possível observar que eles possuem uma semelhança, por exemplo, o momento em que tem um ruído no áudio é parecido, o que pode fazer com que o espectrograma seja parecido também.

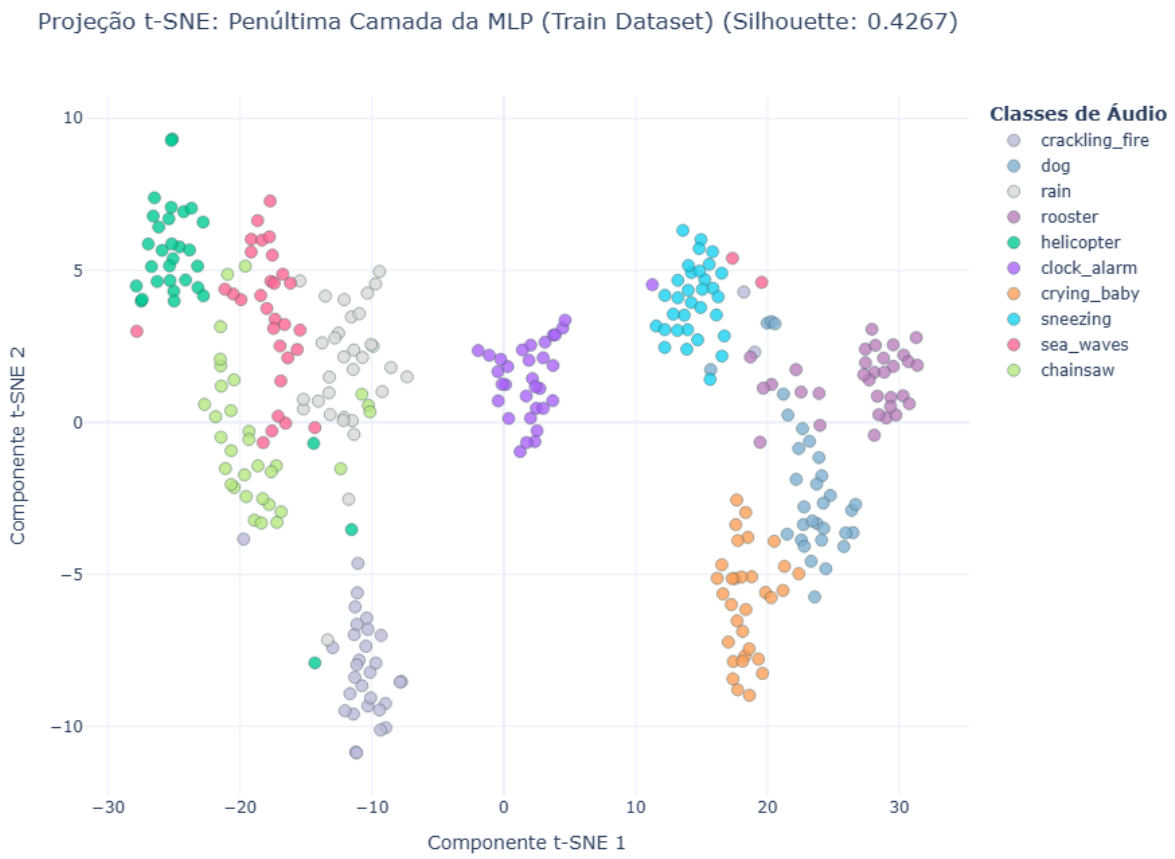
Figura 14 – Resultado da projeção do treinamento ESC-10 após o vetor multidimensional passar pelo MLP, usando o *Standard Scaler*.



Fonte: Autoria própria.

Na Figura 15 vemos uma pequena melhora na separação das classes, onde elas estão um pouco mais agrupada, mas ainda sendo possivel notar que em alguns casos ainda tem um áudio que se mistura em uma classe diferente, como foi o áudio da classe *rain* e *sneezing* se misturando na classe *crackling fire*, essa mistura acontecendo por causa do momento em que acontece o evento no áudio, juntamente com ruído presente neles, fazendo com que fique mais difícil a separação das classes.

Figura 15 – Resultado da projeção do teste ESC-10 após o vetor multidimensional passar pelo MLP, usando o *Robust Scaler*.



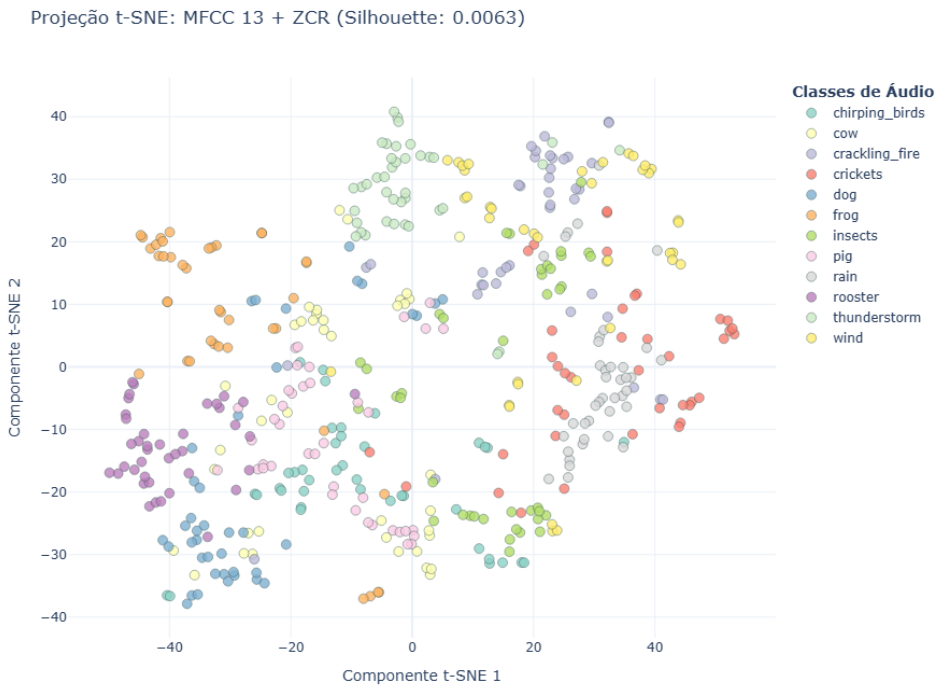
Fonte: Autoria própria.

É possível comparar as Figuras 14 e 15, onde podemos notar que o *Robust Scaler* se saiu melhor quando comparado com o *Standard Scaler*, apesar dos dados iniciais somente com o vetor de características, serem um pouco superior em favor do *Standard Scaler*.

5.1.2 Projeções do ESC-12

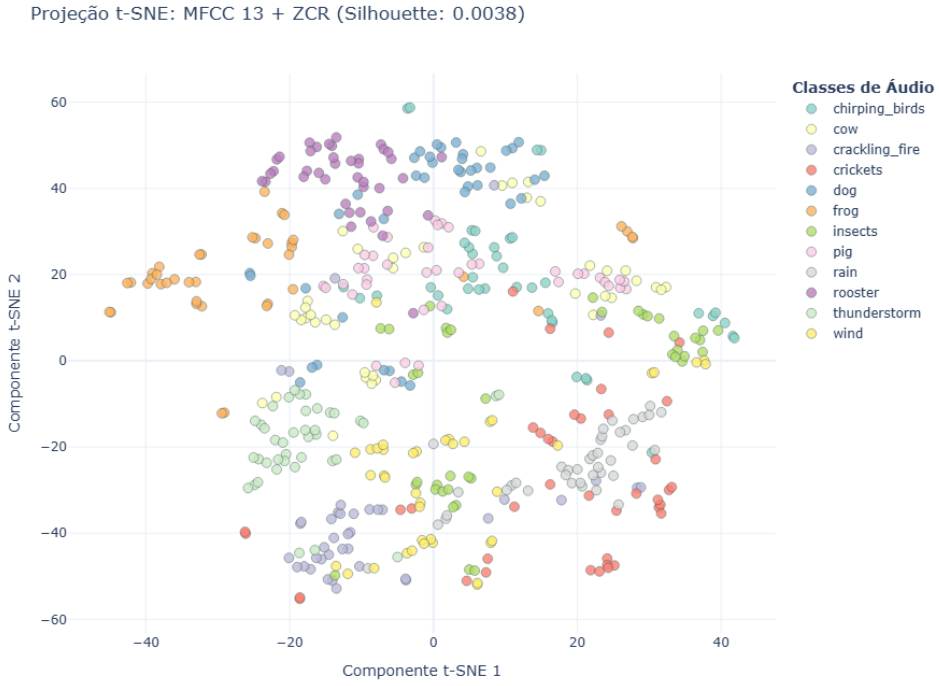
Também foi realizado a exbição do resultado da base de dados ESC-12, assim como o ESC-10, foi utilizado o MFCC junto com o ZCR. Como podemos observar nas Figuras 16 e 17, assim como na base de dados ESC-10, temos uma imagem que todas as classes estão misturadas, sendo que das 12 classes presente na base de dados, podemos destacar a classe *thunderstorm* e a classe *wind*, que são as classes mais agrupadas se comparado com as outras.

Figura 16 – Resultado do plot da base de dados ESC-12 com o *Standard Scaler*.



Fonte: Autoria própria.

Figura 17 – Resultado do plot da base de dados ESC-12 com o *Robust Scaler*.

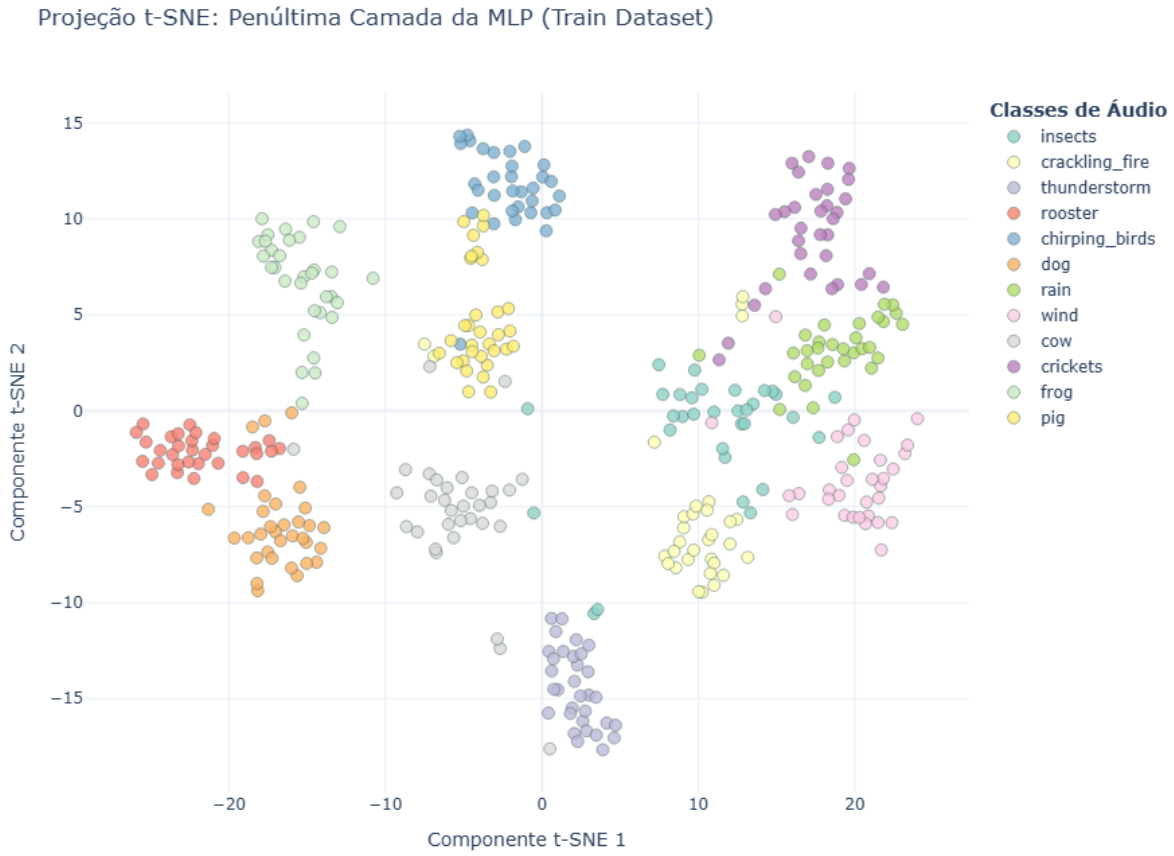


Fonte: Autoria própria.

Após o resultado obtido com o cálculo do MFCC junto com o ZCR e a projeção multidimensional, os mesmos dados foram colocados em uma MLP para que fosse realizado o treinamento com a base de dados, para realizar uma melhor separação das classes na projeção.

Na Figura 18 temos o resultado da base de dados ESC-12 com *Standard Scaler* após passar pela MLP, com a nova figura gerada é possível observar a melhora no agrupamento das classes, porém acontecendo alguns casos de um áudio misturar em outra classe, como o áudio da classe *chirping birds* que ficou junto com áudios da classe *pig*.

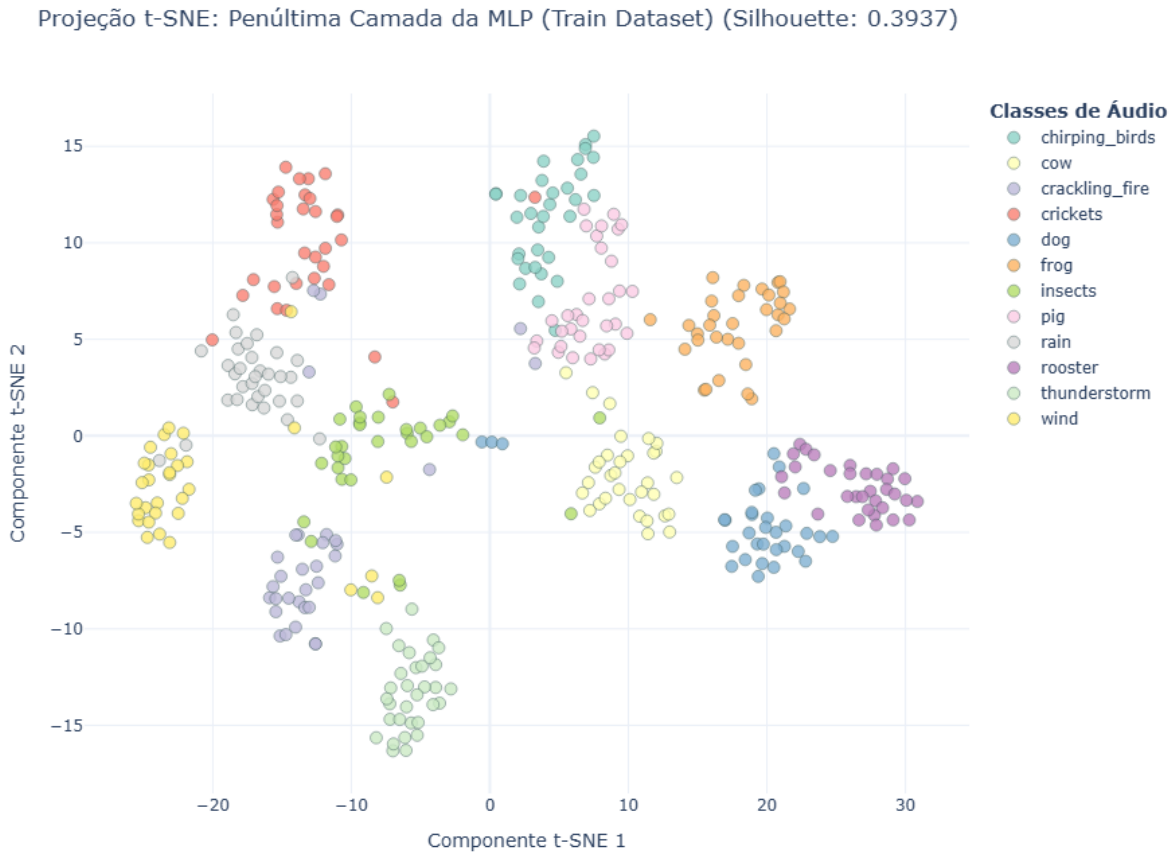
Figura 18 – Resultado da projeção do ESC-12 com *Standard Scaler* após pegar o vetor multidimensional e passar pela MLP.



Fonte: Autoria própria.

E na Figura 19 da base de dados ESC-12 com *Robust Scaler*, que as classes estão mais agrupadas se comparado com o *Standard Scaler*.

Figura 19 – Resultado da projeção do ESC-12 com *Robust Scaler* o teste da MLP.



Fonte: Autoria própria.

Assim como o resultado do ESC-10, no ESC-12 tivemos uma conclusão parecida com relação ao resultado final do *Standard Scaler* e *Robust Scaler*, se observado nas Figuras 18 e 19, e também pelo valor de sua silhueta, mais próximo de um. Assim,é possível observar que teve um resultado mais favorável ao *Robust Scaler*, mesmo que nas Figuras 16 e 17 tenha mostrado que o *Standard Scaler* tenha saído superior ao *Robust Scaler*.

6 CONCLUSÃO

O objetivo do trabalho foi realizar a visualização da separação de classes das base de dados do ESC-10 e ESC-12, utilizando as técnicas MFCC+ZCR. O trabalho utilizou o MFCC e o ZCR para extrair as informações mais importante de cada áudio para que fosse feito a classificação do mesmo.

Como resultado, obtivemos que se usado somente MFCC+ZCR, temos uma imagem onde todos os áudios estão misturado entre eles, não sendo possível separar os grupos de áudio visualmente. Usando a *silhouette score*, facilitou a compreensão através de números, onde se o valor da*silhouette score* fosse mais próximo de 1 é melhor, a visualização do ESC-10 usando somente o cálculo MFCC+ZCR teve 0,0174 de pontuação, e o ESC-12 sua pontuação foi de 0,0063, as duas pontuações baixas das duas bases dados refletem na imagem gerada no t-SNE, que nenhuma classe ficou agrupada.

Resultado diferente quando pegamos os mesmos dados de MFCC+ZCR e passamos por uma MLP, onde os resultados obtiveram uma melhora significativa, sendo refletida na pontuação da *silhouette score*, do ESC-10 saiu de 0,0174 para 0,3770 (Resultado considerado do Standard Scaler), e do ESC-12 que era de 0,0063 para 0,3015 (Standard Scaler). A melhoria obtida no *silhouette score* também foi vista nos resultados gerados pelo visualizador t-SNE, nessas novas imagens, que de fato as classes estão melhores agrupado entre elas, facilitando a visualização do agrupamento dos áudios de mesma classe, a acurácia dos dados de ESC-10 e ESC-12 foram de 77,50% e 68,75% respectivamente.

Além disso, cabe notar que com o trabalho foi possível visualizar qualitativamente os resultados obtidos pelo artigo de Karol (2015), o que pode abrir espaço para melhorias nas suas métricas de classificação.

Conclusão, se comparado ao cálculo base do vetor de características o modelo MLP aplicado no trabalho obteve uma melhoria expressiva nos resultados base gerados, com o `Robust Scaler` apresentando o melhor resultado.

6.1 Trabalhos Futuros

Para trabalhos futuros, realizar a investigação de outras técnicas e descritores de áudios visando melhorar os resultados obtidos nesse trabalho, também realizando a comparação dos descritores de áudio com técnicas que usam aprendizado de máquinas. Buscando outros trabalhos e autores para que seja possível estudar outra metodologia e aprofundar pesquisas, como o trabalho de Salamon e Bello (2017).

Referências

BLOCKS, D. *Representação gráfica de onda sonora digitalizada (Waveform)*. 2026. Ilustração conceitual de domínio público. Acessado em: 15 jun. 2026. Citado na página 9.

BOSI, M.; GOLDBERG, R. E. *Introduction to Digital Audio Coding and Standards*. Boston, MA: Kluwer Academic Publishers, 2002. Citado na página 7.

BäCKSTRöM, T. et al. *Introduction to Speech Processing*. 2. ed. Zenodo, 2022. Disponível em: <https://speechprocessingbook.aalto.fi/Representations/Zero-crossing_rate.html>. Citado 2 vezes nas páginas 4 e 8.

CHU, S.; NARAYANAN, S.; KUO, C.-C. J. Environmental sound recognition with time–frequency audio features. *Audio, Speech, and Language Processing, IEEE Transactions on*, v. 17, p. 1142 – 1158, 09 2009. Citado 2 vezes nas páginas 4 e 13.

DONOSO, J. P. Som e acústica. *Instituto de Física de São Carlos, Universidade de São Paulo (IFSC/USP)*, v. 1, p. 1 – 34, 11 2014. Acessado em: 24 maio 2026. Disponível em: <https://www.ifsc.usp.br/~donoso/fisica_arquitetura/12_som_acustica_1.pdf>. Citado na página 3.

HALLIDAY, D.; RESNICK, R.; WALKER, J. *Fundamentos de Física: Gravitação, Ondas e TermodInâmica*. 10. ed. Rio de Janeiro: LTC, 2014. v. 2. Tradução de Ronaldo Sérgio de Biasi. Citado na página 3.

Instituto de Engenharia. *Como os computadores reconhecem a voz humana?* 2020. Acessado em: 15 jun. 2026. Disponível em: <https://www.institutodeengenharia.org.br/site/2012/12/10/como-os-computadores-reconhecem-a-voz-humana/>. Citado na página 5.

KAROL, P. *ESC: Dataset for Environmental Sound Classification*. 2015. <https://github.com/karolpiczak/paper-2015-esc-dataset/blob/master/Notebook/ESC-Dataset-for-Environmental-Sound-Classification.ipynb>. Citado 5 vezes nas páginas 14, 15, 17, 24 e 31.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, v. 9, p. 2579–2605, 2008. Citado na página 20.

OPPENHEIM, A. V.; SCHAFER, R. W.; BUCK, J. R. *Discrete-Time Signal Processing*. 2. ed. Upper Saddle River, NJ: Prentice Hall, 1999. Citado 2 vezes nas páginas 5 e 6.

PIJANOWSKI, B. C. et al. Soundscape ecology: The science of sound in the landscape. *BioScience*, v. 61, n. 3, p. 203–216, 2011. Citado na página 5.

RAUBER, P. E. et al. Visualizing the hidden activity of artificial neural networks. *IEEE transactions on visualization and computer graphics*, IEEE, v. 23, n. 1, p. 101–110, 2016. Citado na página 21.

SALAMON, J.; BELLO, J. P. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, v. 24, n. 3, p. 279–283, 2017. Citado na página 31.