

Universidade Estadual Paulista  
Instituto de Biociências, Letras e Ciências Exatas  
Departamento de Ciência da Computação e Estatística

Mateus dos Santos Pereira

Identificação De Emoções Em Sinais De Voz Com  
Base No Operador De Energia De Teager Aprimorado

São José do Rio Preto - SP

2021

Mateus dos Santos Pereira

Identificação De Emoções Em Sinais De Voz Com  
Base No Operador De Energia De Teager  
Aprimorado

Monografia apresentada ao Programa de  
graduação em Ciência da Computação da  
UNESP para obtenção do título de Bacharel.

Orientador: Prof. Dr. Rodrigo Capobi-  
anco Guido

São José do Rio Preto - SP

2021

P436i	<p>Pereira, Mateus dos Santos</p> <p>Identificação de emoções em sinais de voz com base no operador de energia de teager aprimorado / Mateus dos Santos Pereira. -- São José do Rio Preto, 2021</p> <p>47 p. : il., tabs.</p> <p>Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto</p> <p>Orientador: Rodrigo Capobianco Guido</p> <p>1. Processamento de sinais. 2. Reconhecimento de emoções. 3. Acústica. 4. Operador de energia de Teager aprimorado. 5. Aprendizado de máquina. I. Título.</p>
-------	---

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

# Agradecimentos

Agradeço a Deus e a todos aqueles que me ajudaram a percorrer esse trecho da minha vida. Agradeço à minha família por me dar o apoio necessário e estar presente. Aos amigos que fiz e compartilharam momentos de alegria comigo. Agradeço também àqueles que já não fazem parte do meu convívio mas que deixaram um impacto significativo, impacto esse que contribuiu para experiências únicas nessa caminhada e que me ajudou a me tornar uma pessoa melhor e mais forte. Por fim, agradeço ao Prof. Dr. Rodrigo Capobianco Guido pela oportunidade e apoio na elaboração desse projeto.

*“No fim tudo dá certo, e se não deu certo é porque ainda não chegou ao fim.”*

**Fernando Sabino**

# Resumo

PEREIRA, M. *Identificação De Emoções Em Sinais De Voz Com Base No Operador De Energia De Teager Aprimorado*. 2021. TCC UNESP 2021.

Atualmente, as pessoas estão cada vez mais conectadas, seja com seu computador pessoal, seja com seu celular, ou seja com qualquer outro tipo de tecnologia presente no seu dia a dia. A interface humano-computador, apesar de sua evolução, ainda enfrenta desafios e obstáculos em busca de uma experiência mais intuitiva e ubíqua. O estudo e desenvolvimento de aplicações com foco em reconhecimento de emoções em sinais de fala consegue diminuir essa distância entre nós humanos e as máquinas, tornando aquilo algo mais natural. O reconhecimento e classificação de emoções em sinais de voz é possibilitado através da extração de características do sinal de fala e sua respectiva classificação emocional com base nessas características. Neste trabalho é abordado como o aprendizado de máquina possibilita o reconhecimento de emoções na fala, qual é o seu estado atual, e discute trabalhos futuros para o seu aperfeiçoamento. O Operador de Energia de Teager Aprimorado é analisado no contexto de classificação de emoções na fala utilizando uma base de dados com expressões emocionais simuladas e observando qual o seu impacto na extração de características cepstrais por meio dos coeficientes Mel-Cepstrais. O algoritmo de Máquina de Vetores de Suporte foi escolhido para a implementação de um classificador fazendo uso de aprendizado de máquina com base nos dados observados de outros estudos apresentados neste trabalho.

Palavras-chave: Processamento de sinais. Reconhecimento de Emoções. Acústica. Operador de Energia de Teager Aprimorado. Aprendizado de máquina.

# Abstract

PEREIRA, M. *Emotion Identification in Voice Signals based on Improved Teager's Energy Operator*. 2021. TCC UNESP 2021.

Nowadays, people are increasingly connected, either with their personal computer, either with their cell phone, or with any other type of technology present in their daily lives. The human-computer interface, despite its evolution, still faces challenges and obstacles in search of a more intuitive and ubiquitous experience. The study and development of applications focused on recognizing emotions in speech signals manages to reduce this distance between us humans and machines, making it feel more natural. The recognition and classification of emotions in voice signals is made possible by extracting characteristics of the speech signal and their respective emotional classification based on these characteristics. This work discusses how machine learning enables the recognition of emotions in speech, what their current state is, and discusses future work for its improvement. The Enhanced Teager Energy Operator is analyzed in the context of emotion classification on speech by making use of a database with simulated emotional expressions and observing its impact on cepstral feature extraction through Mel-Frequency Cepstral Coefficients. The Support Vector Machine algorithm was chosen for the implementation of a machine learning classifier based on the data observed from other studies shown in this work.

Keywords: Signal processing. Emotion Recognition. Acoustics. Enhanced Teager Energy Operator. Machine Learning.

# Lista de Figuras

Figura 2.1 - Representação de uma curva senoidal. . . . .	17
Figura 2.2 - Representação da amostragem de um sinal. . . . .	18
Figura 2.3 - Processo de comunicação. . . . .	20
Figura 2.4 - Comparação entre $T(x_B[\cdot])$ em marrom e $\mathcal{T}(x_B[\cdot])$ em azul . . . . .	23
Figura 2.5 - Visualização da separação de duas classes por meio de um hiperplano. . . . .	24
Figura 3.1 - Arquitetura Geral do Sistema . . . . .	31
Figura 4.1 - Representação do sinal de voz após a aplicação do Operador de Energia de Teager (em azul) e do Operador de Energia de Teager Aprimorado (em vermelho) . . . . .	37
Figura 4.2 - Matriz de confusão dos resultados obtidos sem o uso do Operador de Energia de Teager . . . . .	39
Figura 4.3 - Matriz de confusão dos resultados obtidos com o uso do Operador de Energia de Teager . . . . .	39
Figura 4.4 - Matriz de confusão dos resultados obtidos com o uso do Operador de Energia de Teager Aprimorado . . . . .	40



# Lista de Tabelas

Tabela 2.1 - Estrutura de uma matriz de confusão . . . . .	25
Tabela 2.2 - Exemplo de uma matriz de confusão com valores fictícios. . . . .	26
Tabela 4.1 - Métricas . . . . .	40

# Lista de Abreviaturas

<b>AM</b>	<i>Amplitude Modulation</i>
<b>EMD</b>	<i>Empirical mode decomposition</i>
<b>FM</b>	<i>Frequency Modulation</i>
<b>GMM</b>	<i>Gaussian Mixture Model Classifier</i>
<b>HSA</b>	<i>Hilbert Spectral analysis</i>
<b>LPCC</b>	<i>Linear prediction cepstral coefficients</i>
<b>MFC</b>	<i>Mel-frequency cepstrum</i>
<b>MFCC</b>	<i>Mel-frequency cepstral coefficients</i>
<b>PCM</b>	<i>Pulse Code Modulation</i>
<b>RIFF</b>	<i>Resource Interchange File Format</i>
<b>RNN</b>	<i>Recurrent neural network</i>
<b>SER</b>	<i>Speech Emotion Recognition</i>
<b>SVM</b>	<i>Support Vector Machine</i>
<b>TEMFCC</b>	<i>Teager-energy based Mel-frequency cepstral coefficients</i>
<b>TEO</b>	<i>Teager energy operator</i>
<b>TKEO</b>	<i>Teager-Kaiser energy operator</i>
<b>WAVE</b>	<i>Waveform</i>

# Sumário

<b>1</b>	<b>Introdução</b>	<b>12</b>
1.1	Considerações iniciais . . . . .	12
1.2	Objetivos . . . . .	13
1.3	Organização do trabalho . . . . .	14
<b>2</b>	<b>Conceituação Teórica</b>	<b>15</b>
2.1	Considerações iniciais . . . . .	15
2.2	Sinais Digitais e o Formato WAVE de Arquivos de Áudio . . . . .	16
2.3	A Estrutura Biofísica de Produção da Voz . . . . .	19
2.4	Caracterização das Emoções . . . . .	21
2.5	O Operador de Energia de Teager Aprimorado . . . . .	21
2.6	Máquinas de Vetores de Suporte . . . . .	23
2.7	Métricas Utilizadas . . . . .	25
2.7.1	Matriz de Confusão . . . . .	25
2.8	Trabalhos Correlatos em Reconhecimento de Emoções pela Voz . . . . .	27
2.9	Considerações Finais . . . . .	29
<b>3</b>	<b>A Abordagem Proposta</b>	<b>30</b>
3.1	Considerações iniciais . . . . .	30
3.2	Arquitetura Geral do Sistema . . . . .	30
3.2.1	Base de Dados Utilizada . . . . .	32
3.3	Módulos . . . . .	32
3.3.1	Módulo de Extração de Características . . . . .	32
3.3.2	Módulo de mesclagem . . . . .	33

3.3.3	Módulo classificador . . . . .	34
3.3.4	Resultados . . . . .	34
3.4	Considerações finais . . . . .	34
<b>4</b>	<b>Testes e Resultados</b>	<b>36</b>
4.1	Considerações Iniciais . . . . .	36
4.2	Testes Realizados . . . . .	36
4.2.1	Seleção de dados . . . . .	37
4.2.2	Conjunto de testes e treinamento . . . . .	38
4.2.3	Resultados Obtidos . . . . .	38
4.3	Discussões . . . . .	40
<b>5</b>	<b>Conclusões e Trabalhos Futuros</b>	<b>42</b>
5.1	Conclusões Gerais . . . . .	42
5.2	Problemas Encontrados . . . . .	43
5.3	Trabalhos Futuros . . . . .	43
	<b>Referências</b>	<b>43</b>

# Capítulo 1

## Introdução

### 1.1 Considerações iniciais

O reconhecimento de emoções na fala é a tarefa de reconhecer os aspectos emocionais da fala independente do conteúdo semântico da mesma. Tal tarefa é concebida como algo natural para nós seres humanos como parte do processo de comunicação, porém a capacidade de realizá-la por meio de dispositivos programáveis ainda é um assunto recorrente em pesquisas da área (LECH et al., 2020).

O reconhecimento de emoções na fala é um tópico que tem uma vasta gama de aplicações em nosso cotidiano. A interação humano-computador, apesar do avanço nos últimos anos, ainda está longe de ser algo totalmente natural. Atualmente encontramos diversas áreas com aplicabilidade, como por exemplo sistemas de ensino à distância que avaliam as emoções de seus estudantes para que os professores possam agir de acordo com as necessidades de engajamento. Outro exemplo seria um sistema de computador de bordo em carros que assegurasse a segurança do passageiro de acordo com seu estado emocional. Uma aplicação na área da saúde seria o uso por parte de terapeutas, para que eles possam avaliar como o paciente está se sentindo. Apesar desses e outros exemplos, o reconhecimento de emoções na fala ainda é um desafio, por conta da complexidade e variáveis no campo das emoções, que mudam de locutor para locutor, cultura para cultura.

O foco do presente trabalho é classificar a emoção presente em amostras de falas em calmo, feliz, temeroso ou aborrecido por meio de algoritmos de aprendizado de máquina a fim de analisar o comportamento do Operador de Energia de Teager Aprimorado neste cenário. O método discutido neste trabalho utiliza características cepstrais extraídas a partir da fala sob condições emocionais simuladas a fim de realizar a sua classificação emocional.

## 1.2 Objetivos

Motivado pelas considerações anteriores, este trabalho busca estudar o atual estado do campo de reconhecimento de emoções, analisar os desafios atuais e avaliar os testes realizados utilizando o Operador de Energia de Teager em um contexto de aprendizado de máquina para determinar o seu comportamento na área de classificação de emoções na fala.

Diversos trabalhos que envolvem a área de reconhecimento de emoções na fala foram realizados e estudados para a elaboração deste trabalho. Em (QADRI et al., 2021) o autor elabora um sistema de reconhecimento de emoções na fala com base nos coeficientes cepstrais de frequência mel (MFCCs) e operador de energia de teager para obter características a serem utilizadas no treinamento de uma rede neural para a classificação das emoções. O objetivo deste trabalho é classificar sinais de fala com expressões emocionais simuladas em suas respectivas emoções por meio de algoritmos de aprendizado de máquina, com base em características cepstrais extraídas da fala junto da aplicação do Operador de Energia de Teager Aprimorado. A análise e classificação das falas em uma das emoções definidas é realizada com base nas características extraídas e falas rotuladas de acordo com as expressões simuladas do conjunto de treinamento, junto de métricas utilizadas nos trabalhos relacionados abordados.

### **1.3 Organização do trabalho**

O texto vindouro do presente trabalho está organizado da seguinte forma:

- No Capítulo 2 apresenta-se uma série de trabalhos publicados envolvendo a área de reconhecimento de emoções, mostrando como são inúmeras as possibilidades de se realizar essa tarefa. Expõe-se, também, os principais conceitos e teorias que estão relacionados com o trabalho que foi desenvolvido.
- No Capítulo 3 apresenta-se, com detalhes, todo o desenvolvimento do trabalho proposto e de que forma os conceitos discutidos no capítulo anterior foram utilizados.
- No Capítulo 4 relatam-se todos os resultados obtidos no trabalho, a partir dos testes de reconhecimento que foram realizados.
- No Capítulo 5 apresentam-se as conclusões sobre o trabalho, bem como propostas para pesquisas futuras.

# Capítulo 2

## Conceituação Teórica

### 2.1 Considerações iniciais

A fundamentação teórica necessária para a compreensão deste trabalho é apresentada nas seguintes seções, junto de trabalhos relacionados ao seu tema. Na seção 2.2, o detalhamento de como o som se comporta é caracterizado e é apresentado como os sinais de som são representados em um meio digital, junto da estrutura do formato WAVE amplamente utilizado. Na sequência, a seção 2.3 trata a respeito de como a fala é produzida, abordando estruturas biológicas responsáveis pela produção da voz. A seção 2.4 apresenta como as emoções são caracterizadas, e como elas se relacionam com diferentes características presentes na fala. Já na seção 2.5, é detalhado o operador não-linear denominado de Operador de Energia de Teager-Kaiser, que permite o cálculo da energia necessária para a produção de um sinal, fundamental para extrair características de um sinal de fala. As seções 2.6 e 2.7.1 abordam a fundamentação teórica sobre o aprendizado de máquina que será utilizado para a classificação da fala em diferentes emoções de acordo com as características extraídas. Por fim, a seção 2.8 traz os trabalhos correlatos em reconhecimento de emoções pela voz, detalhando o objetivo e resultados de diferentes estudos relacionados ao tema.



## 2.2 Sinais Digitais e o Formato WAVE de Arquivos de Áudio

Em (BHATNAGAR, 2002), o som é descrito como perturbações no ar, ou outro meio, o que é característico de uma onda mecânica que necessita de um meio para a sua propagação. Tal detalhe é contrastante com as ondas eletromagnéticas, que não necessitam de um meio de propagação para a transmissão de energia. O som é caracterizado como uma onda longitudinal, que se propaga por meio de compressões e rarefações no ar.

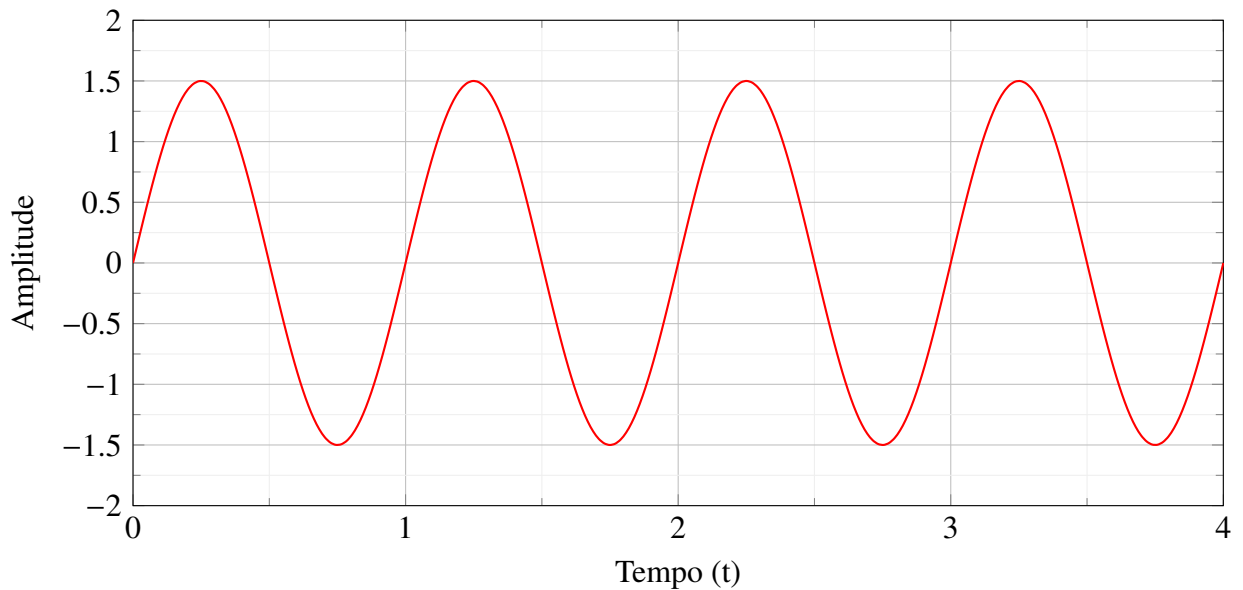
A representação matemática das ondas sonoras se dá por meio das ondas senoidais, que descrevem uma oscilação periódica suave. Um sinal periódico é caracterizado por um sinal que se repete a cada  $T$  segundos, em que  $T$  é denominado como o período deste sinal. A sua periodicidade assume que esta repetição é verdadeira no domínio do tempo, se estendendo tanto para o passado quanto para o futuro (LYNN, 1989).

Temos a seguinte curva em função do tempo, definida pela função 2.1 a seguir (HAZEWIN-KEL, 2002):

$$y(t) = A \sin(2\pi ft + \varphi) = A \sin(\omega t + \varphi) \quad (2.1)$$

Em que  $A$  representa a amplitude da função, ou seja, o maior pico a partir do ponto zero;  $f$  representa a frequência da onda, ou seja, o número de oscilações por unidade de tempo;  $\omega = 2\pi f$  representa a frequência angular em radianos por segundo;  $\varphi$  representa, em radianos, a fase, que especifica o ciclo da oscilação quando  $t = 0$ .

Figura 2.1 – Representação de uma curva senoidal.



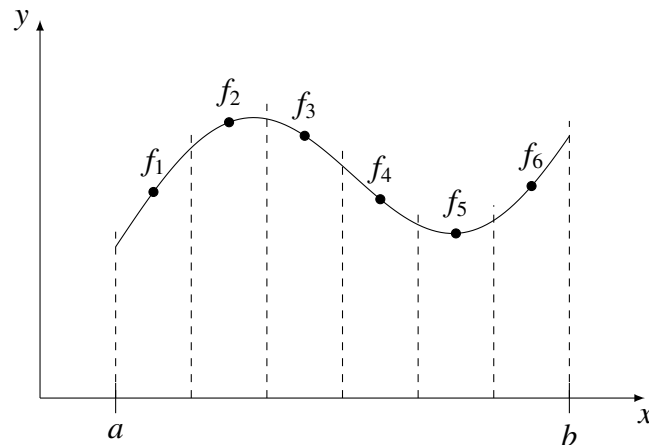
Fonte: Elaborado pelo autor.

Na figura 2.1 é possível observar a periodicidade da curva senoidal, uma de suas principais características fundamentais. Para fins de simplificação, foi utilizado um período de 1 segundo e amplitude de 1,5 decibéis (dB).

Os sinais digitais são representações discretas de sinais da natureza. Um exemplo disso é o sinal de som, que é amplamente armazenado em uma estrutura de arquivo denominada como WAVE, sigla para Waveform Audio File Format (IBM / MICROSOFT, 1991). Os sinais de som no meio digital podem ser representados por meio da amostragem da forma de onda, em um processo denominado por PCM (*Pulse Code Modulation*), que armazena os dados do áudio sem compressão (BHATNAGAR, 2002).

Pelo processo de amostragem é possível transformar um sinal analógico em um sinal digital por meio de sua discretização. A partir desse processo, uma função de tempo contínuo é convertida em uma função de tempo discreto por meio de uma sequência de valores discretos que representam o sinal original. A taxa de amostragem desse sinal representa o intervalo de tempo  $t$  em que é capturado os valores de sua amplitude, ou seja, a cada  $t$  segundos é capturada a amplitude atual do sinal analógico, de forma que seja possível obter uma versão digital do mesmo.

Figura 2.2 – Representação da amostragem de um sinal.



Fonte: Elaborado pelo autor.

Ao transformar um sinal de natureza analógica em um sinal digital por meio de sua amostragem, ocorre uma perda de informações, pois ao passar do domínio contínuo, que trata de intervalos infinitos, para o domínio discreto, que trata de intervalos finitos, é impossível replicar todas as informações presentes no sinal analógico. Para tratar esse problema, foi elaborado o teorema da amostragem de Nyquist–Shannon, que mantém a fidelidade de um sinal ao passar pelo processo de amostragem.

O Teorema da amostragem de Nyquist–Shannon é definido da seguinte forma:

**Teorema 1** (Teorema da amostragem de Nyquist–Shannon). *Seja um sinal, limitado em banda, e seu intervalo de tempo dividido em partes iguais, de forma que se obtenham intervalos tais que, cada subdivisão compreenda um intervalo com período  $T$  segundos, onde  $T$  é menor do que  $f_m/2$ , e se uma amostra instantânea é tomada arbitrariamente de cada subintervalo, então o conhecimento da amplitude instantânea de cada amostra somado ao conhecimento dos instantes em que é tomada a amostra de cada subintervalo contém toda a informação do sinal original (SHANNON, 1949).*

O teorema 1 nos diz que, ao realizar a amostragem de um sinal de um meio analógico para um meio discreto, a taxa de amostragem necessária para que o sinal possa ser discretizado e transformado de volta em sinal analógico sem perda de fidelidade é de duas vezes a maior frequência, que é representada por  $f_m$ . Para a representação de um sinal com uma frequência

máxima de 440Hz, por exemplo, é necessário uma taxa mínima de  $2f_m = 2 \times 440 = 880$  amostras por segundo.

Quando a amostragem do sinal não satisfaz as condições do teorema, como a falta de uma limitação de banda, dois sinais diferentes podem ser indistinguíveis quando amostrados, o que se dá o nome de *aliasing*.

Um sinal de som armazenado digitalmente conta com, além de sua taxa de amostragem, a sua profundidade de bits, em que cada bit representa a amplitude do sinal naquele momento. Em uma profundidade de 8-bits, os valores positivos de amplitude são representados por valores superiores a 128, enquanto que os valores negativos são representados por valores inferiores a 128. O valor 128, nessa situação, é tratado efetivamente como o "zero" da amplitude na representação digital do sinal (BHATNAGAR, 2002).

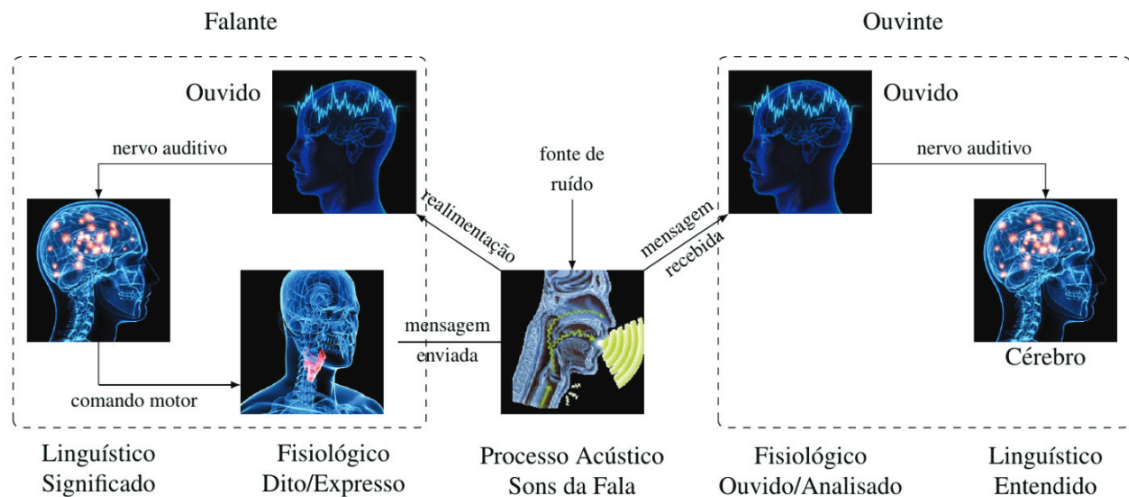
Os arquivos no formato WAVE, além de armazenarem os dados do sinal em questão, armazenam também metadados sobre o mesmo, como taxa de amostragem, canais e tipo dos dados de áudio. O formato é uma variação do formato RIFF (*Resource Interchange File Format*), um formato de contêiner de arquivos genérico para armazenamento de dados multimídia (IBM / MICROSOFT, 1991).

### **2.3 A Estrutura Biofísica de Produção da Voz**

O princípio da fala parte do processo de comunicação verbal. Esse processo inicia-se com a formação de palavras, frases e seus respectivos significados em nosso cérebro. Em seguida, o processo fisiológico de comunicação inicia a reprodução da voz por meio de sinais acústicos. Tal processo se dá por meio de flutuações da pressão de ar, gerados pelas pregas vocais, moduladas pelo trato vocal e irradiadas pela boca (SILVA; SOUZA; MAY, 2017). O sinal gerado por um interlocutor será captado pelo ouvido de um indivíduo receptor por meio de seu processo fisiológico, que converterá o sinal acústico em sinais elétricos no cérebro para o entendimento da fala e assim prosseguir com a comunicação. A figura 2.3 ilustra o processo descrito e suas

respectivas etapas.

Figura 2.3 – Processo de comunicação.



Fonte: (SILVA; SOUZA; MAY, 2017).

A fala é resultado de movimentos voluntários dos aparelhos respiratório e digestivo de nosso corpo. A fonte desses movimentos reside nas musculaturas torácicas e abdominais, que controlam a passagem de ar pelos pulmões e traqueia. Assim que o ar é expelido dos pulmões, ele percorre a traqueia até chegar na faringe. No topo da traqueia, temos a laringe, onde se encontram as pregas vocais, compostas por ligamento e músculo, responsáveis pela produção da voz. A produção de sinais sonoros por esse meio é denominada fonação. O movimento das pregas vocais se assemelha à vibração da palheta de um instrumento de sopro (FLANAGAN, 1972).

O estudo a respeito da fala e seus sons é dividido em duas categorias, uma delas estando associada à área da fonética, e a outra à área da fonologia. A área de fonética estuda as propriedades articulatórias, acústicas e perceptivas do processo de produção e recepção da voz. Já a área de fonologia estuda as diferenças fônicas intencionais que geram significados que dão origem a morfemas, palavras e frases (CALLOU, 1990).

## 2.4 Caracterização das Emoções

As emoções humanas desempenham um importante papel na expressão vocal. No entanto, as emoções da fala nem sempre são fáceis de serem distinguidas, até mesmo para nós humanos, devido a ambiguidades presentes na expressão emocional (XIAO et al., 2005). Uma tarefa importante na classificação de emoções na fala é obter uma clara expressão e distinção de emoções nas características a serem analisadas.

Existem 3 grandes categorias de amostras emocionais de fala, sendo elas expressão vocal natural, expressão emocional induzida e expressão emocional simulada (SCHERER, 2003). As expressões naturais são emoções naturais e autênticas, gravadas de acontecimentos do dia a dia ou algum programa de TV ou rádio. Já as expressões emocionais induzidas são obtidas por meio de substâncias psicoativas ou jogos com eventos intencionalmente planejados com o intuito de obter uma resposta emocional. Por fim as expressões emocionais simuladas são aquelas em que atores tentam imitar uma determinada emoção, dessa forma o processo pode ser mais controlado a fim de obter as amostras desejadas para determinada necessidade. As amostras emocionais de fala utilizadas neste trabalho se enquadram na terceira categoria, as expressões simuladas.

## 2.5 O Operador de Energia de Teager Aprimorado

Diversos problemas da área de processamento de sinais podem ser modelados por meio de uma abordagem com modelos lineares, com o uso de filtros sobre o sinal de entrada, criando um novo sinal de saída em função desse sinal anterior. Com o uso do operador não-linear chamado de *Operador de Energia de Teager* (também conhecido como Operador de Energia de Teager-Kaiser), é possibilitado a modelagem de problemas não-lineares em processamento de sinais (KVEDALEN, 2003).

De acordo com Teager, o modelo de fala em meados do ano de 1983 era impreciso. Teager argumenta que os modelos elaborados com base em filtros lineares não eram suficientes para descrever todos os processos envolvidos na produção da fala, devido à presença de processos não-lineares. Esses processos foram descritos mais detalhadamente em (TEAGER; TEAGER, 1990), pontuando sobre a energia que origina determinado som junto de seu gráfico. No entanto, nenhum algoritmo é detalhado para permitir o cálculo dessa energia, o que é elaborado posteriormente por Kaiser.

Kaiser apresenta seu algoritmo em 1990, definindo a versão discreta do Operador de Energia de Teager da seguinte maneira (KAISER, 1990):

$$\Psi[x[n]] = x^2[n] - x[n-1]x[n+1] \quad (2.2)$$

Para se estimar a energia de um sinal por meio do Operador de Energia de Teager, apenas algumas amostras são utilizadas, o que acarreta em imprecisões quando há ruído no sinal de entrada. Em (BOVIK; MARAGOS; QUATIERI, 1993), é proposto um algoritmo que faz uso de um conjunto de filtros passa-faixa, ou banco de filtros, para a supressão de ruído nos sinais.

Em (GUIDO, 2019) o Operador de Energia de Teager Aprimorado é apresentado. Neste trabalho, é demonstrado como a entidade mecânica e o sinal digital possuem uma massa ( $\eta$ ) que expressa a resistência em mudar a amplitude do sinal no decorrer do tempo.

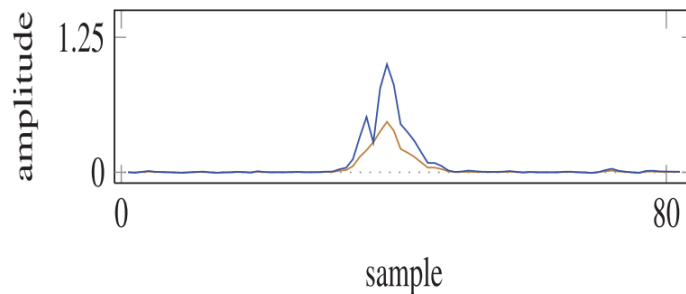
Temos inicialmente o Operador de Energia de Teager  $T$  que modela a energia de um sinal discreto  $x[\cdot]$  como o produto do quadrado de sua amplitude pela frequência instantânea. (GUIDO, 2019) mostra como  $T$  é modificado de forma a obter  $A^2 \cdot \Omega^2$  exatamente em função de  $x[\cdot]$ . Essa modificação de  $T$  dá origem ao Operador de Energia de Teager Aprimorado, expresso da seguinte forma:

$$\mathcal{T} = \frac{T}{\frac{1}{2}\eta} = \frac{x_n^2 - x_{n-1} \cdot x_{n+1}}{\text{sinc}^2(\Omega)} = A^2 \cdot \Omega^2 \quad (2.3)$$

O artigo detalha como o operador de energia de teager original é aprimorado, de forma a ser representado de forma análoga a um sistema mecânico. Tal aprimoramento permite análises mais precisas e detalhadas, útil em campos como reconhecimento de fala, identificação de

locutor, detecção de emoção na fala e análise de sinais biomédicos.

Figura 2.4 – Comparação entre  $T(x_B[\cdot])$  em marrom e  $\mathcal{T}(x_B[\cdot])$  em azul



Fonte: (GUIDO, 2019)

## 2.6 Máquinas de Vetores de Suporte

O Aprendizado de Máquina é um sistema que realiza um determinado "aprendizado" sobre um conjunto de dados para que ele possa prever, ou no caso, classificar esses dados de acordo com as informações adquiridas.

O seu uso nos dias atuais é amplamente difundido, sendo encontrado nos mais diversos sistemas presente no nosso cotidiano. Desde o *smartphone* até sistemas médicos, sua ubiquidade é inquestionável. É possível realizar diagnósticos médicos (ZWETSCH et al., 2006) por meio de aplicações de técnicas e algoritmos de Aprendizado de Máquina que seriam quase impossíveis sem o seu uso.

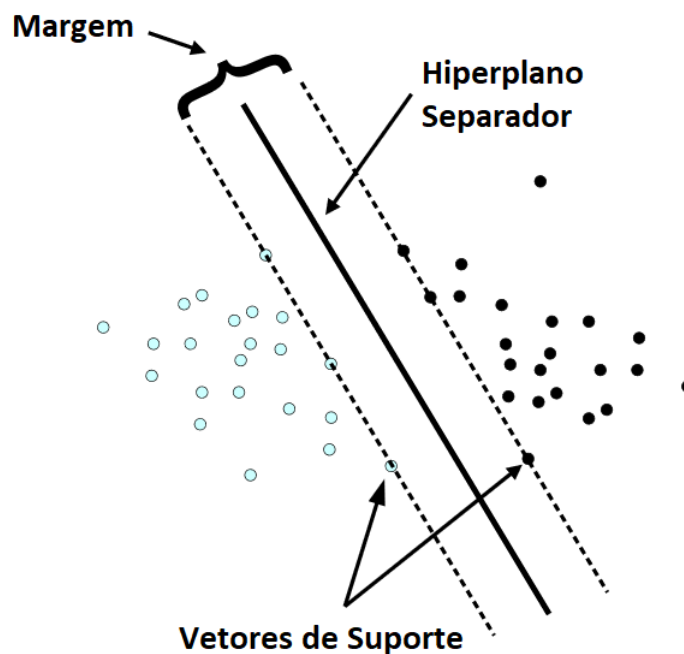
Um dos princípios fundamentais do aprendizado de máquina é a indução, que permite inferir regras gerais a partir de um conjunto de dados específico para um dado modelo, sendo ela a tarefa primária do aprendizado de máquina (MOONEY, 2000). Tal aprendizado é categorizado em supervisionado e não-supervisionado (LORENA; CARVALHO, 2007). O aprendizado supervisionado tem como característica o aprendizado de um mapeamento de uma entrada a uma saída por meio de pares entrada-saída (HAYKIN; NETWORK, 2004). Ou seja, o algoritmo é



"ensinado" a partir de uma base de dados pré existente e, com a introdução de novos dados, é capaz de produzir saídas corretamente com base nesse aprendizado. Já no caso do aprendizado não-supervisionado não existem exemplos prontos para o aprendizado, o algoritmo aprende a representar os dados de entrada a partir de medidas de qualidade (LORENA; CARVALHO, 2007).

As Máquinas de Vetores de Suporte (SVM) representam uma técnica poderosa para a classificação, regressão e detecção de anomalias em aprendizado de máquina, com modelagem de diversos espaços dimensionais gerados a partir de diferentes métricas analisadas. SVMs foram desenvolvidas por Cortes e Vapnik (1995) para classificação binária de classes. A abordagem geral pode ser dita como: Encontrar um hiperplano ótimo ao se maximizar a margem dos pontos mais próximos. Os pontos que se encontram nas margens são chamados de vetores de suporte, e o centro da margem é o hiperplano ótimo que separa as duas classes (MEYER, 2001). Uma visualização dessa descrição é representada na figura 2.5

Figura 2.5 – Visualização da separação de duas classes por meio de um hiperplano.



Fonte: (MEYER, 2001) (Adaptado).

As Máquinas de Vetores de Suporte são embasadas na teoria de aprendizado estatístico,

que estabelece princípios para que seja possível obter classificadores com uma generalização que permite prever classes de novos dados corretamente, no mesmo domínio utilizado para o aprendizado. Com esses princípios, o SVM se mostra uma técnica com boa capacidade de generalização (LORENA; CARVALHO, 2007).

## 2.7 Métricas Utilizadas

Nesta seção são detalhadas as métricas utilizadas nos testes e análise dos resultados obtidos posteriormente neste trabalho. Tais métricas permitem uma melhor visualização e comprovação dos dados analisados.

### 2.7.1 Matriz de Confusão

As matrizes de confusão, também conhecidas como matrizes de erro, são matrizes ou tabelas que permitem visualizar a performance de um algoritmo de aprendizado supervisionado (STEHMAN, 1997), que no nosso caso será o algoritmo de Máquina de Vetores de Suporte. Essa matriz detalha os resultados esperados e os resultados obtidos a partir do aprendizado de um determinado conjunto de dados.

Tabela 2.1 – Estrutura de uma matriz de confusão

Emoção	Verdadeiramente Feliz	Verdadeiramente Triste
Classificação: Feliz	Verdadeiros Felizes	Falsos Felizes
Classificação: Triste	Falsos Tristes	Verdadeiros Tristes

A tabela 4.2 mostra um exemplo com apenas duas emoções a fim de simplificar o entendimento. Em casos com três ou mais emoções, a lógica permanece a mesma, porém com uma matriz  $n \times n$ , sendo  $n$  a quantidade de emoções a serem classificadas.

Cada item da tabela representa o seguinte:

**Verdadeiros Felizes:** Se referem a casos em que emoções felizes são classificadas como verdadeiramente felizes.

**Falsos Felizes:** Casos em que emoções tristes são classificadas como felizes, mas são verdadeiramente tristes.

**Verdadeiros Tristes:** Se referem a casos em que emoções tristes são classificadas como verdadeiramente tristes.

**Falsos Tristes:** Casos em que emoções felizes são classificadas como tristes, porém, são verdadeiramente felizes.

Tabela 2.2 – Exemplo de uma matriz de confusão com valores fictícios.

Emoção	Verdadeiramente Feliz	Verdadeiramente Triste
Classificação: Feliz	53	19
Classificação: Triste	20	82

Com a matriz de confusão em mãos, é possível atribuir uma pontuação à classificação realizada pelo algoritmo de aprendizado de máquina. Dessa forma, temos valores a serem utilizados em métricas para avaliação da classificação.

**Acurácia:** Define a proximidade entre o valor esperado e o valor obtido experimentalmente. A acurácia mede a frequência em que o classificador está correto. Na equação 2.4, temos que a acurácia é calculada pela soma dos verdadeiros positivos com verdadeiros negativos, sobre o total de classificações realizadas.

$$P = \frac{VP + VN}{Total} \quad (2.4)$$

**Precisão:** Mede, daqueles classificados como corretos, quantos de fato estavam corretos. Na equação 2.5, temos que a precisão é calculada pelos verdadeiros positivos, sobre a soma dos verdadeiros positivos com falsos positivos.

$$P = \frac{VP}{VP + FP} \quad (2.5)$$

**Recall:** Também denominado como revocação, mede a frequência com que o classificador

encontra exemplos de uma classe específica. Na equação 2.6, é calculada pelos verdadeiros positivos, sobre a soma dos verdadeiros positivos com falsos negativos.

$$Recall = \frac{VP}{VP + FN} \quad (2.6)$$

**F1 Score:** Combina a precisão e *recall* de maneira que a pontuação é definida como duas vezes a média harmônica entre eles. Na equação 2.7, é representado o seu cálculo.

$$F1 = 2 * \frac{P * R}{P + R} \quad (2.7)$$

## 2.8 Trabalhos Correlatos em Reconhecimento de Emoções pela Voz

Alguns dos diversos estudos correlatos a respeito do reconhecimento de emoções em sinais de fala são apresentados nessa seção, com foco em análise e classificação de atributos extraídos a partir da voz.

Em (KERKENI et al., 2019), os autores apresentam dois algoritmos de aprendizado de máquina para a classificação de sete emoções distintas: alegria, raiva, tristeza, surpresa, medo, desgosto e neutralidade. O estudo faz uso de atributos baseados em demodulação AM-FM e análises não-lineares combinando o Operador de Energia de Teager-Kaiser com Decomposição do modo empírico (EMD). Os paradigmas de aprendizado de máquina utilizados são RNN, sigla para Redes Neurais Recorrentes, e SVM, sigla para Máquinas de Vetores de Suporte. Foram utilizados duas bases de dados, uma em Espanhol, e outra de Berlin. Para a extração de atributos, são de importância os atributos cepstrais e de modulação AM-FM para a classificação posterior do sinal. Baseando-se nos resultados e análises dos experimentos do estudo em questão, o autor conclui que para que o classificador RNN tenha bons resultados de reconhecimento de emoções, é necessário uma maior extração de atributos e um tempo mais longo de treina-

mento do sistema classificador. Já o classificador SVM demonstra um maior potencial para seu uso prático devido a essas características.

O trabalho apresentado por Jain et al. (2020) aborda uma classificação de emoções na fala utilizando máquinas de vetores de suporte (SVM) como o seu algoritmo de classificação. Neste estudo, foram utilizados os estados emocionais de tristeza, raiva, medo e felicidade para a classificação das amostras de falas de duas bases de dados distintas. Os atributos importantes extraídos de cada amostra foram energia, tom, MFCC, LPCC e taxa de fala. As estratégias de classificação utilizadas foram Um-Contra-Todos e classificador dependente de gênero. O autor, por fim, conclui que, a partir dos resultados obtidos em seus experimentos, a extração de atributos MFCC junto do classificador dependente de gênero alcançou a melhor taxa de reconhecimento de emoções.

Sandhya et al. (2020) propõem em seu estudo métodos para a identificação de locutores sob a influência de emoções na fala. O autor argumenta que sistemas de identificação de locutores funcionam bem em condições neutras, mas se deterioram em condições emocionais, devido ao impacto que as emoções causam ao sinal de voz. Neste estudo são extraídas diversas características cepstrais da fala como coeficientes cepstrais de frequência mel (MFCCs) e coeficientes cepstrais de frequência mel baseados em energia de Teager (TEMFCCs) a fim de analisar quais apresentam melhor resultado entre diversos classificadores. Em seus resultados, o autor obtém uma melhor acurácia de 100% em condições neutras e 87,0967% em condições emocionais.

Em seu trabalho, Iriya (2014) aborda a classificação de emoções nos sinais de fala com diferentes atributos como frequência fundamental, energia de curto prazo, formantes e coeficientes cepstrais. São utilizados diversos algoritmos classificadores para a comparação entre eles, tais como K-vizinhos mais próximos (KNN), Máquinas de Vetores de Suporte (SVM), Modelos de Misturas Gaussianas (GMM) e Modelos Ocultos de Markov, tendo o GMM como o principal devido ao seu desempenho e custo computacional. Iriya faz o uso de um sistema de classificação de estágio único e um sistema de classificação sequencial em três estágios baseado na teoria de emoções da área de psicologia. No melhor caso, o autor conseguiu 100% de acurácia na identificação da emoção Tristeza com ambos os sexos masculino e feminino.

Pereira, Pádua e Silva (2015) apresentam um reconhecimento de emoções multimodal, fo-

cado em participantes de telejornais. O sistema analisa, junto da fala, as expressões faciais dos apresentadores, para auxiliar na classificação das emoções. Para a extração de características do sinal de voz, os autores utilizaram a ferramenta *openSMILE*, analisando a intensidade sonora e frequência fundamental dos sinais.

O trabalho realizado por Jahangir et al. (2021) trata a respeito de aprendizagem profunda no contexto do reconhecimento de emoções na fala. Nele são analisadas diversas características extraídas do sinal de fala e como eles se comportam em diferentes abordagens de aprendizagem profunda. O autor menciona que o Operador de Energia de Teager é indicado para detectar estresse na fala com base em suas observações apresentadas no estudo. Ele também afirma que características cepstrais como MFCCs são as mais capazes para reconhecimento de fala.

## **2.9 Considerações Finais**

Com as informações apresentadas neste capítulo, é possível ter o conhecimento necessário a respeito do tema deste trabalho para a compreensão dos capítulos seguintes, que tratarão a abordagem proposta, seus experimentos e respectivos resultados. Os trabalhos correlatos mostram que a utilização de características relacionadas à energia do sinal são amplamente abordadas em diversos sistemas de classificação, incluindo o sistema de Máquinas de Vetores de Suporte (SVM).

# Capítulo 3

## A Abordagem Proposta

### 3.1 Considerações iniciais

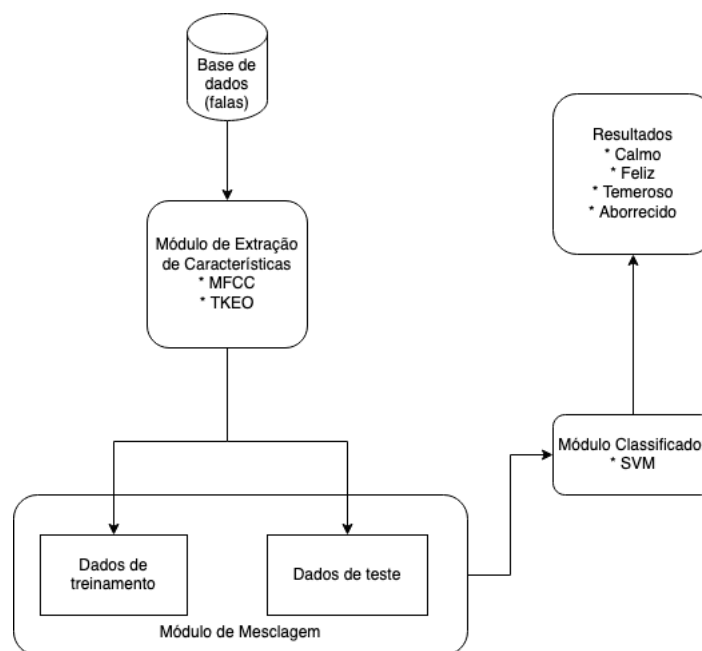
Neste capítulo, a metodologia utilizada na elaboração deste projeto é apresentada. Na seção 3.2, a arquitetura do sistema elaborado é discutida e detalhada, proporcionando uma visão geral sobre o que será trabalhado. Na sequência, são abordados a base de dados utilizada para a classificação, os módulos de extração de características e classificador, junto das métricas escolhidas para a análise posterior dos resultados no capítulo seguinte. Então, por fim, temos as considerações finais do que foi apresentado neste capítulo.

### 3.2 Arquitetura Geral do Sistema

A arquitetura do sistema tem como principais componentes três módulos fundamentais para o reconhecimento de emoções. Conforme demonstrado na figura 3.1, o projeto é composto pelo módulo de extração de características, módulo de mesclagem e módulo classificador. O primeiro módulo é responsável por aplicar o Operador de Energia de Kaiser e sua versão aprimorada e extrair as características dos sinais de fala da base de dados que serão fundamentais

para a classificação do sistema. Já o módulo de mesclagem trata de separar as características em duas bases diferentes, uma que será utilizada para treinar o módulo classificador e outra utilizada para testar a eficácia do algoritmo de aprendizado de máquina. Na sequência temos o módulo classificador, que tem como entrada os dados de teste e os dados de treinamento, responsável por classificar determinada fala em uma das quatro emoções escolhidas para o projeto (calmo, feliz, temeroso, aborrecido). Por fim, obtemos os resultados da classificação e analisamos as métricas do sistema.

Figura 3.1 – Arquitetura Geral do Sistema



Fonte: Elaborado pelo autor.

Para o desenvolvimento do sistema proposto, optou-se pela utilização da linguagem de programação *Python* e suas respectivas bibliotecas. A escolha desta linguagem de programação para projetos de aprendizado de máquina é observada em diversas aplicações, devido à facilidade de uso e bibliotecas que auxiliam na visualização, processamento de dados e cálculos estatísticos essenciais para a implementação de um sistema de aprendizado de máquina (MÜLLER; GUIDO, 2016).

A biblioteca de principal importância para a elaboração deste projeto é a versão 0.8.1 da biblioteca *librosa* (MCFEE et al., 2021), voltada especialmente para processamento de sinais



de áudio. Com ela, é possível extrair as características dos arquivos de áudio necessárias para a análise dos dados, tais como características espectrais (MCFEE et al., 2015).

Para a implementação dos algoritmos de aprendizado de máquina, foi utilizada a biblioteca *scikit-learn*, proporcionando algoritmos para a implementação do módulo classificador e métricas de eficácia do sistema.

### **3.2.1 Base de Dados Utilizada**

Como primeira etapa para a elaboração do projeto, foi escolhida uma base de dados com amostras de sinais de voz para o aprendizado e classificação baseada em emoções. A base em questão é a *Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)* 2018. Ela é composta por 7356 arquivos, totalizando 24,8 *gigabytes*. Nela temos gravações de 24 atores, sendo metade masculinos e metade femininos, no formato *.wav* com taxa de amostragem de 48kHz e profundidade de 16 bits. As falas estão disponíveis nas emoções de Alegria, Tristeza, Raiva, Medo, Surpresa, Desgosto e Neutro. Cada expressão é produzida em duas intensidades, normal e alta, com uma neutra adicional (LIVINGSTONE; RUSSO, 2018). Para este projeto, o foco é nas emoções neutro, alegria, medo e tristeza.

## **3.3 Módulos**

### **3.3.1 Módulo de Extração de Características**

Este módulo é responsável por extrair características a partir do sinal de voz e representá-los em um vetor de características. A primeira etapa é rotular a emoção a partir do nome do

arquivo, conforme descrito na base de dados utilizada. Na sequência, o Operador de Energia de Teager e o Operador de Energia de Teager Aprimorado são aplicados nos sinais de voz. Em seguida, esses novos sinais passam por métodos de extração de características que são executados em cada arquivo de som. O adotado por este trabalho foi o MFCC. Tais informações carregam informações distintas a respeito do sinal de voz, e auxiliam na classificação posterior. Os métodos utilizados são descritos na sequência.

**Coeficientes Mel-Cepstrais (MFCC):** Os coeficientes Mel-Cepstrais proporcionam características importantes e poderosas na área de reconhecimento de fala. A função do MFCC é imitar o comportamento da audição humana com a aplicação da análise cepstral. (KISHORE; SATISH, 2013). De acordo com (ATTNEAVE; OLSON, 1971), a escala de Mel foi desenvolvida para descrever uma relação entre a frequência real e o que era interpretado pela audição. Para este projeto, foram utilizados 40 coeficientes mel-cepstrais para a classificação de emoções.

**Operador de Energia de Teager (TEO):** O Operador de Energia de Teager modela a energia de um sinal como o produto do quadrado da sua amplitude ( $A$ ) pela frequência instantânea ( $\Omega$ ) (GUIDO, 2019). Tal energia nos dá uma representação mais detalhada da fala quando a dicção apresenta determinado estresse e trata o seu comportamento nos domínios da frequência e do tempo (AOUANI; AYED, 2020). Características baseadas no Operador de Energia de Teager refletem melhor as características da passagem de ar na fala sob condições emocionais (ZHOU; HANSEN; KAISER, 1998).

### 3.3.2 Módulo de mesclagem

O módulo de mesclagem separa os vetores de características em grupos distintos, um com os dados que serão utilizados para treinar o classificador, e outro com dados que serão utilizados para testar o classificador. A definição de quais dados estarão presentes em cada grupo se dá de forma aleatória, sendo 25% dos dados reservados para testes e os 75% restantes utilizados para

treinamento.

### **3.3.3 Módulo classificador**

Este módulo é responsável por analisar e classificar os sinais de áudio conforme a emoção reconhecida, separados pelo módulo anterior. Para a sua implementação foi utilizado o algoritmo de aprendizado de máquina conhecido como Máquinas de Vetores de Suporte, detalhado no capítulo 2, seção 2.6. O treinamento do módulo classificador se dá pelos dados de treino separados pelo módulo de mesclagem. Os sinais de voz podem ser classificados como calmo, feliz, temeroso ou aborrecido por este módulo.

### **3.3.4 Resultados**

Os resultados obtidos com a classificação de emoções a partir do módulo anterior são comparados com os resultados esperados e as métricas são calculadas com base nesses valores. São analisadas a acurácia, precisão, *recall* e *F1 score*. É apresentada também a matriz de confusão obtida a partir das métricas calculadas, referente às 4 emoções abordadas.

## **3.4 Considerações finais**

Neste capítulo foi apresentada a abordagem proposta para a realização do projeto, bem como o funcionamento de cada módulo do sistema. A arquitetura do sistema foi inspirada nos trabalhos de (DENDUKURI; HUSSAIN, 2019), (KERKENI et al., 2019) e (AOUANI; AYED,

2020). Nos capítulos seguintes, essa abordagem será colocada em prática e os testes necessários para a análise do projeto serão detalhados, junto dos resultados obtidos.

# Capítulo 4

## Testes e Resultados

### 4.1 Considerações Iniciais

Neste capítulo encontram-se os dados referentes à implementação do sistema descrito neste trabalho e os resultados de classificação dos algoritmos de aprendizado de máquina com a base de dados escolhida. Para fins análise e comparação, são apresentadas métricas apresentadas no capítulo 2 referentes aos resultados obtidos com o sistema desenvolvido.

### 4.2 Testes Realizados

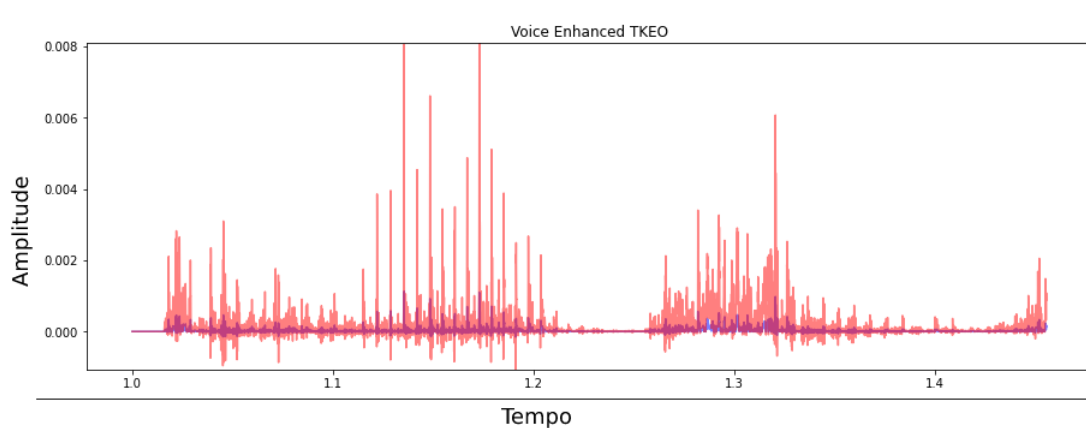
Esta seção aborda as etapas realizadas para obtenção dos resultados, desde a seleção de dados até os testes, bem como os resultados obtidos e as métricas analisadas.

### 4.2.1 Seleção de dados

A partir da base de dados *RAVDESS* (LIVINGSTONE; RUSSO, 2018), foram obtidos 7356 arquivos audiovisuais, gravados por 24 atores. Entre eles, arquivos de vídeo e áudio, apenas vídeo, e apenas áudio. Para este trabalho foram utilizados apenas os arquivos falados, os arquivos de canto não foram utilizados.

Os arquivos de áudio de fala foram gravados de forma que cada ator gravou duas frases distintas em 6 tonalidades diferentes, expressando as emoções neutro, calmo, feliz, aborrecido, raivoso, temeroso, surpreso, e enojado. Tal descrição enquadra a base de dados na categoria de expressões emocionais simuladas, pois não foram espontâneas e nem induzidas por algum agente externo. As expressões emocionais simuladas são mais utilizadas em trabalhos como este por serem mais controlados e previsíveis. Para os testes realizados posteriormente com o classificador, foram selecionadas apenas os áudios correspondentes às emoções calmo, feliz, temeroso e aborrecido da base de dados.

Figura 4.1 – Representação do sinal de voz após a aplicação do Operador de Energia de Teager (em azul) e do Operador de Energia de Teager Aprimorado (em vermelho)



Fonte: Elaborado pelo autor.

Para fins de análise, foram selecionadas três versões deste grupo de dados. Uma delas consiste no sinal de voz original, a segunda consiste nos sinais de voz após a aplicação do operador de energia de Teager, e a terceira consiste nos sinais de voz após a aplicação do operador de

energia de Teager aprimorado. Cada versão é analisada para fins de comparação entre elas e análise do impacto do operador de energia de Teager na classificação das emoções.

#### **4.2.2 Conjunto de testes e treinamento**

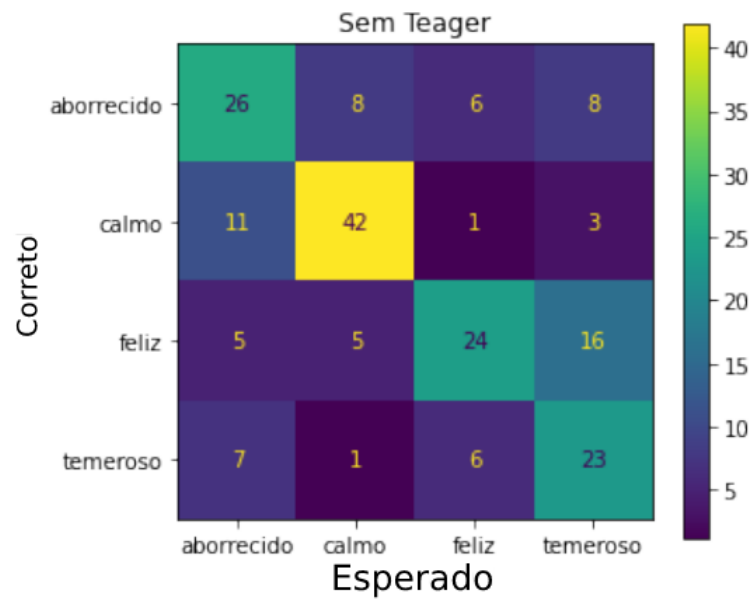
Após a seleção dos dados de áudio e a aplicação dos operadores de energia abordados anteriormente, é feita a separação entre conjunto de teste e conjunto de treinamento. Um conjunto representando 25% da seleção de áudios é destinada ao conjunto de teste, enquanto que os 75% restantes são destinados ao conjunto de treinamento.

São extraídos os coeficientes Mel-Cepstrais de cada arquivo, tanto do conjunto de testes quanto do conjunto de treinamento, obtendo-se 40 coeficientes que possibilitam o classificador SVM trabalhar nesta base de dados. As características extraídas são então separadas no conjunto de treinamento que é utilizada para supervisionar o aprendizado do classificador de emoções. A partir disso o algoritmo de máquina de vetores de suporte tem os dados necessários para que seja possível rotular os sinais de voz do conjunto de teste.

#### **4.2.3 Resultados Obtidos**

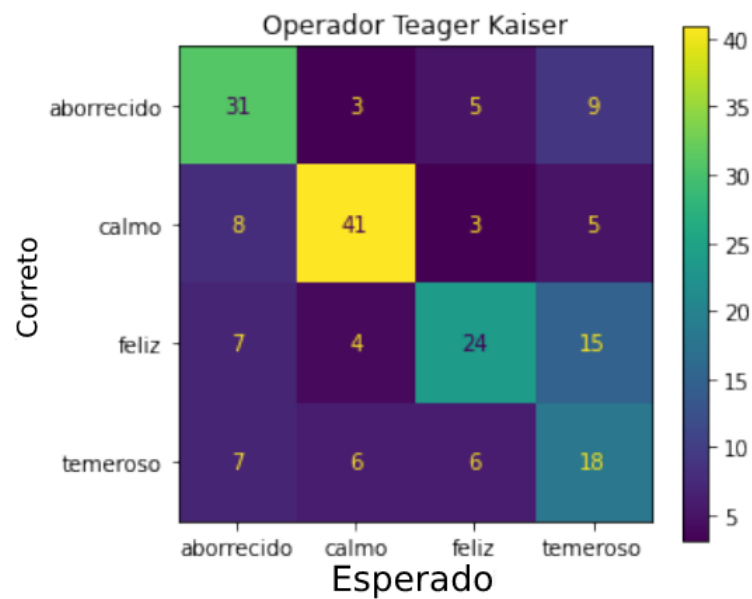
Após o preparo dos dados, a separação em conjunto de treinamento e conjunto de teste, e a extração de características por meio do MFCC, o algoritmo *SVM* (Máquina de Vetores de Suporte) é executado. A execução ocorre na base de dados original, a base após a aplicação do operador de energia de Teager, e a base após a aplicação do operador de energia de Teager aprimorado. As figuras a seguir mostram as matrizes de confusão dos resultados obtidos com a classificação dos dados, bem como tabelas com cada métrica discutida no capítulo 2: acurácia, precisão, *recall* e *F1 score*.

Figura 4.2 – Matriz de confusão dos resultados obtidos sem o uso do Operador de Energia de Teager



Fonte: Elaborado pelo autor.

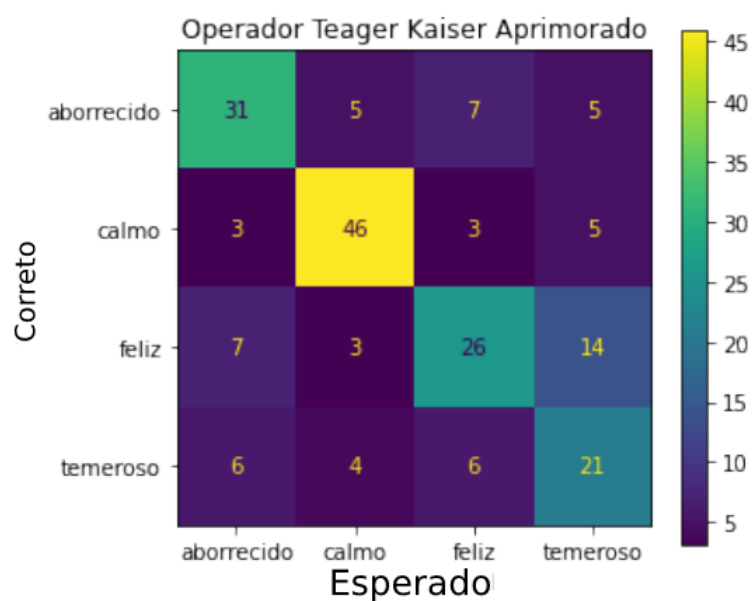
Figura 4.3 – Matriz de confusão dos resultados obtidos com o uso do Operador de Energia de Teager



Fonte: Elaborado pelo autor.



Figura 4.4 – Matriz de confusão dos resultados obtidos com o uso do Operador de Energia de Teager Aprimorado



Fonte: Elaborado pelo autor.

Após a realização dos testes, foram obtidas as seguintes métricas:

Tabela 4.1 – Métricas

Métrica	Sem Teager	Teager	Teager Aprimorado
Acurácia	59,90%	59,38%	64,58%
Precisão	61,29%	60,99%	65,15%
Recall	59,90%	59,38%	64,58%
F1 Score	60,03%	59,74%	64,66%
Média	60,28%	59,87%	64,74%

### 4.3 Discussões

Com os testes e resultados abordados neste capítulo, é possível observar que as matrizes de confusão e métricas apresentadas mostram que o Operador de Energia de Teager aprimorado fornece uma maior porcentagem nas métricas de acurácia, precisão, *recall* e *f1 score*. Apesar do

Operador de Energia de Teager original fornecer um sinal mais propício para reconhecimento de emoções baseados em energia, como por exemplo falas sob estresse, os dados dos testes nos mostram que, com a extração baseada em MFCC junto do uso do classificador SVM, não há diferença significativa entre a extração sem o Operador de Energia de Teager e a extração com ele. Contudo, ao olharmos os dados do Operador de Energia de Teager Aprimorado, temos melhores resultados quando comparados com os outros dois cenários.

Ao olharmos também para as matrizes de confusão, podemos observar que a emoção calmo foi a mais bem classificada entre as outras emoções avaliadas. Além disso, é possível notar que a emoção temeroso é confundida pela emoção feliz pelo algoritmo classificador SVM. Tal resultado pode ser causado pela forma como as emoções foram simuladas pela base de dados utilizada, como também pela forma como o classificador se comporta com as características extraídas e fornecidas durante a etapa de treinamento do sistema.

## Capítulo 5

# Conclusões e Trabalhos Futuros

### 5.1 Conclusões Gerais

Com a expansão de sistemas de inteligência artificial, nossa sociedade está cada vez mais imersa nesse tipo de tecnologia em seu dia a dia (MAKRIDAKIS, 2017). O reconhecimento de emoções é um pilar com aplicações existentes e também conta com grande potencial para o futuro. Tendo isso em mente, neste trabalho foi implementado um sistema de aprendizado de máquina que permite identificar e classificar emoções em sinais de fala. Tal implementação faz uso de análise de características da fala, bem como uma nova abordagem do Operador de Energia de Teager, como descrito em (GUIDO, 2019). Com esse operador de energia, suas aplicações vão desde sistemas de processamento de fala, até análise de sinais biomédicos. Na área de reconhecimento de emoções, é possível ter uma análise mais precisa e detalhada, principalmente quando a dicção apresenta estresse, o que auxilia na extração de características na implementação do sistema de aprendizado de máquina.

De uma maneira geral, os resultados obtidos mostram que ao utilizar o Operador de Energia de Teager Aprimorado, é possível obter uma capacidade melhor de extração de características, o que condiz com o que é apresentado em (GUIDO, 2019). Este trabalho nos mostra que, dado uma base de dados adequada para o treinamento da classificação de emoções, o Operador de Energia de Teager Aprimorado nos dá um conjunto de dados mais propício para extração

de características e, conseqüentemente, um sistema com melhor capacidade de classificar as emoções de sinais de fala.

## **5.2 Problemas Encontrados**

Um dos problemas encontrados durante a realização deste trabalho é que algumas emoções são mais bem reconhecidas que outras, conforme exemplificado na matrizes de confusão no Capítulo 4, o que torna a metodologia proposta mais favorável a alguns tipos de emoções, como aborrecido e calmo, enquanto outras emoções acabam sendo confundidas por outras, como é o caso das emoções feliz e temeroso.

## **5.3 Trabalhos Futuros**

A partir do estudo realizado neste trabalho, é possível expandir o tema com algumas propostas. No trabalho (QADRI et al., 2021), é explorado a detecção de emoções em bases de dados multi-linguais. Nele, o autor utiliza uma fusão do Operador de Energia de Teager com o MFCC para obter melhores resultados, junto de aprendizagem profunda com redes neurais. Em (JABLOUN, 2017), o operador de energia de Teager é modificado pelo autor para aplicações na biomedicina, sendo interessante analisar como ele se difere do operador de energia de teager aprimorado apresentado neste trabalho. Já no trabalho (KERKENI et al., 2019), é utilizado o TKEO também no contexto de reconhecimento de emoções, porém com o método de decomposição do modo empírico.

Com esses trabalhos junto com o conteúdo abordado neste, é possível explorar novas formas de aprimorar a extração de características de sinais de voz a fim de obter um sistema de reconhecimento de emoções mais robusto e fiel.

# Referências

AOUANI, H.; AYED, Y. B. Speech emotion recognition with deep learning. *Procedia Computer Science*, v. 176, p. 251–260, 2020. ISSN 1877-0509. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1877050920318512>>.

ATTNEAVE, F.; OLSON, R. K. Pitch as a medium: A new approach to psychophysical scaling. *The American Journal of Psychology*, University of Illinois Press, v. 84, n. 2, p. 147–166, 2021/08/02/ 1971. ISSN 00029556. Full publication date: Jun., 1971. Disponível em: <<https://doi.org/10.2307/1421351>>.

BHATNAGAR, G. *Introduction to multimedia systems*. [S.l.]: Academic Press, 2002.

BOVIK, A. C.; MARAGOS, P.; QUATIERI, T. F. Am-fm energy detection and separation in noise using multiband energy operators. *IEEE Transactions on Signal Processing*, v. 41, n. 12, p. 3245–3265, Dec 1993. ISSN 1941-0476.

CALLOU, D. *Iniciação à fonética e à fonologia*. [S.l.]: Editora Schwarcz-Companhia das Letras, 1990.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

DENDUKURI, L. S.; HUSSAIN, S. J. Statistical feature set calculation using teager energy operator on emotional speech signals. In: *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*. [S.l.: s.n.], 2019. p. 192–195.

FLANAGAN, J. L. *Speech analysis: synthesis and perception*. [S.l.]: Springer, 1972.

GUIDO, R. C. Enhancing teager energy operator based on a novel and appealing concept: Signal mass. *Journal of the Franklin Institute*, Elsevier, v. 356, n. 4, p. 2346–2352, 2019.

HAYKIN, S.; NETWORK, N. A comprehensive foundation. *Neural networks*, v. 2, n. 2004, p. 41, 2004.

HAZEWINKEL, M. *Encyclopaedia of mathematics*. Berlin; New York: Springer-Verlag, 2002. OCLC: 123132388. ISBN 9781402006098. Disponível em: <<http://eom.springer.de/default.htm>>.

- IBM / MICROSOFT. *Multimedia Programming Interface and Data Specifications 1.0*. USA, 1991. Disponível em: <<http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/WAVE/Docs/riffmci.pdf>>. Acesso em: 12 ago. 2020.
- IRIYA, R. *Análise de sinais de voz para reconhecimento de emoções*. Tese (Doutorado) — Universidade de São Paulo, 2014.
- JABLOUN, M. A new generalization of the discrete teager-kaiser energy operator - application to biomedical signals. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2017. p. 4153–4157.
- JAHANGIR, R. et al. Deep learning approaches for speech emotion recognition: state of the art and research challenges. *Multimedia Tools and Applications*, Springer, p. 1–66, 2021.
- JAIN, M. et al. *Speech Emotion Recognition using Support Vector Machine*. 2020.
- KAISER, J. F. On a simple algorithm to calculate the 'energy' of a signal. In: *International Conference on Acoustics, Speech, and Signal Processing*. [S.l.: s.n.], 1990. p. 381–384 vol.1. ISSN 1520-6149.
- KERKENI, L. et al. Automatic speech emotion recognition using an optimal combination of features based on emd-tkeo. *Speech Communication*, Elsevier, v. 114, p. 22–35, 2019.
- KISHORE, K. K.; SATISH, P. K. Emotion recognition in speech using mfcc and wavelet features. In: *2013 3rd IEEE International Advance Computing Conference (IACC)*. [S.l.: s.n.], 2013. p. 842–847.
- KVEDALEN, E. Signal processing using the teager energy operator and other nonlinear operators. *Master, University of Oslo Department of Informatics*, Citeseer, v. 21, 2003.
- LECH, M. et al. Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Frontiers in Computer Science*, Frontiers, v. 2, p. 14, 2020.
- LIVINGSTONE, S. R.; RUSSO, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, Public Library of Science San Francisco, CA USA, v. 13, n. 5, p. e0196391, 2018.
- LORENA, A. C.; CARVALHO, A. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. ISSN 21752745. Disponível em: <[https://www.seer.ufrgs.br/rita/article/view/rita\\_v14\\_n2\\_p43-67](https://www.seer.ufrgs.br/rita/article/view/rita_v14_n2_p43-67)>.
- LYNN, P. A. *An Introduction to the Analysis and Processing of Signals*. 3. ed. [S.l.]: Macmillan Education UK, 1989. (New Electronics). ISBN 978-0-333-48887-4, 978-1-349-19719-4, 196-197-199-2.
- MAKRIDAKIS, S. The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures*, v. 90, p. 46–60, 2017. ISSN 0016-3287. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0016328717300046>>.

- MCFFEE, B. et al. *librosa/librosa: 0.8.1rc2*. Zenodo, 2021. Disponível em: <<https://doi.org/10.5281/zenodo.4792298>>.
- MCFFEE, B. et al. *librosa: Audio and music signal analysis in python*. In: CITESEER. *Proceedings of the 14th python in science conference*. [S.l.], 2015. v. 8, p. 18–25.
- MEYER, D. Support vector machines. *Porting R to Darwin/X11 and Mac OS X*, Citeseer, v. 1, p. 23, 2001.
- MOONEY, R. J. Integrating abduction and induction in machine learning. *Abduction and Induction*, Springer, p. 181–191, 2000.
- MÜLLER, A. C.; GUIDO, S. *Introduction to machine learning with Python: a guide for data scientists*. [S.l.]: "O'Reilly Media, Inc.", 2016.
- PEREIRA, M. H. R.; PÁDUA, F. L. C.; SILVA, G. D. Abordagem multimodal para reconhecimento automático de emoções aplicada ao estudo de níveis de tensão em telejornais. *Brazilian Journalism Research*, v. 11, n. 2, p. 160–183, 2015.
- QADRI, S. A. A. et al. Speech emotion recognition using deep neural networks on multilingual databases. In: JIZAT, J. A. M. et al. (Ed.). *Advances in Robotics, Automation and Data Analytics*. Cham: Springer International Publishing, 2021. p. 21–30. ISBN 978-3-030-70917-4.
- SANDHYA, P. et al. Spectral features for emotional speaker recognition. In: *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAEECC)*. [S.l.: s.n.], 2020. p. 1–6.
- SCHERER, K. R. Vocal communication of emotion: A review of research paradigms. *Speech communication*, Elsevier, v. 40, n. 1-2, p. 227–256, 2003.
- SHANNON, C. E. Communication in the presence of noise. *Proceedings of the IRE, IEEE*, v. 37, n. 1, p. 10–21, 1949.
- SILVA, A.; SOUZA, F.; MAY, V. Identificação de padrões de vogais em registros acústicos: análise por componentes cepstrais e redes neurais. 04 2017.
- STEHMAN, S. V. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, Elsevier, v. 62, n. 1, p. 77–89, 1997.
- TEAGER, H. M.; TEAGER, S. M. A phenomenological model for vowel production in the vocal tract. *Speech Science: Recent Advances*, College-Hill, p. 73–109, 1983.
- TEAGER, H. M.; TEAGER, S. M. Evidence for nonlinear sound production mechanisms in the vocal tract. In: *Speech production and speech modelling*. [S.l.]: Springer, 1990. p. 241–261.
- XIAO, Z. et al. Features extraction and selection for emotional speech classification. In: *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005*. [S.l.: s.n.], 2005. p. 411–416.

ZHOU, G.; HANSEN, J.; KAISER, J. Classification of speech under stress based on features derived from the nonlinear teager energy operator. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*. [S.l.: s.n.], 1998. v. 1, p. 549–552 vol.1.

ZWETSCH, I. C. et al. Processamento digital de sinais no diagnóstico diferencial de doenças laríngeas benignas. *Scientia Medica*, Pontificia Universidad Católica de Río Grande del Sur, v. 16, n. 3, p. 109–114, 2006.