
UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Hiago Matheus Brajato

# Reconhecimento de emoções na fala a partir da extração manual de características com validação baseada na engenharia paraconsistente

São José do Rio Preto

2022

Hiago Matheus Brajato

# Reconhecimento de emoções na fala a partir da extração manual de características com validação baseada na engenharia paraconsistente

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Campus de São José do Rio Preto.

Financiadora: CAPES
Orientador: Prof Dr Rodrigo Capobianco Guido

São José do Rio Preto, SP

2022

B814r

Brajato, Hiago Matheus

Reconhecimento de emoções na fala a partir da extração manual de características com validação baseada na engenharia paraconsistente / Hiago Matheus Brajato. -- , 2022

71 p. : tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Farmacêuticas, Araraquara,

Orientador: Rodrigo Capobianco Guido

1. Speech Emotion Recognition. 2. Reconhecimento de Emoções na Voz. 3. Engenharia Paraconsistente. 4. Seleção de características com uso de Algoritmo Genético (AG). I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Farmacêuticas, Araraquara. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Hiago Matheus Brajato

# Reconhecimento de emoções na fala a partir da extração manual de características com validação baseada na engenharia paraconsistente

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Campus de São José do Rio Preto.

Financiadora: CAPES
Orientador: Prof Dr Rodrigo Capobianco Guido

Comissão Examinadora

Prof Dr Rodrigo Capobianco Guido
UNESP - Campus de São José do Rio Preto
Orientador

Profa Dra Renata Spolon Lobato
UNESP - Campus de São José do Rio Preto

Profa Dra Luciene Cavalcanti Rodrigues
FATEC - São José do Rio Preto

24 de fevereiro de 2022

São José do Rio Preto, SP
2022

*Aos meus pais, que com suor e dedicação me criaram e continuamente incentivam e apoiam meus estudos...sem os mesmos seria impossível a confecção deste trabalho.*

# AGRADECIMENTOS

Ao meu orientador, o qual desde o início dessa jornada proveu-me gentil e competentemente os caminhos a serem seguidos no desenvolvimento do projeto.

Aos docentes e discentes conhecidos ao longo do curso, que com extrema afabilidade proporcionaram-me adquirir novos e inspiradores conhecimentos.

Aos criadores de conteúdo acadêmico digital, que com sua generosidade possibilitam o acesso à informação gratuita e qualificada.

A toda comunidade científica antepassada, sem a qual não seria possível atingir novos avanços científicos e tecnológicos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, à qual agradeço.

*"Apaixone-se pelo problema, não pela solução".*

# RESUMO

*Speech Emotion Recognition* (SER) pode ser definida como a maneira automatizada de identificar o estado emocional de um locutor a partir da sua voz. Dentre as metodologias encontradas na literatura para viabilizar o SER, as quais ainda carecem de melhor compreensão e discussão, o presente trabalho ocupa-se da abordagem *handcrafted extraction* para a composição dos vetores de características responsáveis por permitir a classificação dos sinais de voz entre sete classes emocionais distintas: raiva, tédio, desgosto, medo, felicidade, neutralidade e tristeza. Os descritores utilizados, os quais foram obtidos por meio da energia clássica, do Operador de Energia de Teager, do *zero crossing rate*, da planaridade espectral e da entropia espectral, foram submetidos à Engenharia Paraconsistente de Características, que é responsável por selecionar o melhor subgrupo de características a partir da análise de similaridades e dissimilaridades intra e interclasse, respectivamente. Finalmente, um algoritmo genético associado à uma rede neural *multilayer perceptron* foi responsável por realizar a classificação dos sinais visando a maior taxa de acurácia possível, isto é, 84.9%, considerando a base de dados pública EMO_DB com 535 sinais na modalidade *speaker-independent*. Em contraste com abordagens do tipo *feature learning*, a estratégia proposta permitiu uma melhor compreensão física do problema em questão.

Palavras-chave: *Speech Emotion Recognition* (SER), *handcrafted extraction*, Engenharia Paraconsistente de Características, redes neurais artificiais.

# *ABSTRACT*

*Speech Emotion Recognition (SER) can be defined as the automated way to identify speakers' emotional states from their voices. Considering the methodologies found in the literature, for which there is room for further research and better comprehension, this monograph considers a handcrafted feature extraction approach to create the feature vectors responsible for the classification of voice signals in one of the seven different classes: anger, boredom, disgust, fear, happiness, neutrality and sadness. The descriptors adopted, which were obtained based on regular energy, on Teager Energy Operator, on zero crossing rates, on spectral flatness and on spectral entropy, were submitted to the Paraconsistent Feature Engineering, which was responsible for selecting the best subgroup of features from the analysis of intra- and interclass similarities and dissimilarities, respectively. Lastly, a genetic algorithm associted with a multiplayer perceptron neural network was responsible for performing the classification of the described signals aiming at the highest possible accuracy rate, i.e., 84.9%, considering the well-known EMO_DB database with 535 signals in a speaker-independent approach. In contrast with feature learning strategies, the proposed approach allowed for a better comprehension of the problem being treated.*

*Keywords: Speech Emotion Recognition, handcrafted extraction, Paraconsistent Feature Engineering, artificial neural network.*

# Lista de Figuras

# Lista de Tabelas

# Sumário

# Capítulo 1

# Introdução

## 1.1 Considerações Iniciais e Objetivos

*Speech emotion recognition* (SER) pode ser descrita como uma das áreas mais emergentes dentro do campo de processamento digital de sinais. Com as primeiras investigações acerca do tema sendo iniciadas por volta dos anos 80 (BEZOOIJEN; EBRARY, 1984) (TOLKMITT; SCHERER, 1986), diversos são os âmbitos de utilização de sistemas SER, dentre os quais é possível citar a análise de satisfação de clientes, o monitoramento de estresse de pacientes dentro de sistemas de saúde, entre outros. O desenvolvimento desse tipo de sistema torna-se possível pois o sinal de voz carrega informações conectadas não somente ao seu conteúdo lexical, mas também relacionadas ao gênero, idade e estado emocional do locutor.

Diante do exposto e do desejo do autor deste trabalho em investigar os sistemas SER com base em técnicas de processamento de sinais e inteligência artificial, este trabalho tem por objetivo inicial realizar um levantamento bibliográfico sobre o tema, incluindo as dissertações acadêmicas e os conceitos relacionados. Em seguida, a intenção é a de converter os sinais de fala em vetores de características que possibilitem, por si só e independentemente de classificadores, a mínima e a máxima correlação inter e intraclasses, a fim de que os sinais de voz possam ser classificados em uma dentre sete diferentes classes de emoções presentes na base de dados utilizada. O conjunto de descritores provém dos seguintes conceitos: energia clássica (GUIDO, 2016a), Operador de Energia de Teager

(KAISER, 1990), *zero crossing rate* (ZCR) (GUIDO, 2016b), planaridade espectral (GRAY; MARKEL, 1974) e entropia espectral (SHANNON, 1948). Finalmente, a qualidade das características extraídas é aferida por meio da análise paraconsistente (GUIDO, 2019), a fim de que a melhor combinação de atributos possa ser encontrada visando o processo final de classificação das emoções, realizado por uma rede neural artificial.

Análises acerca dos resultados obtidos durante os experimentos realizados podem ser encontrados no decorrer do documento, o qual contém uma solução para o requerido problema que pode servir como base para futuras investigações.

## 1.2 Estrutura do trabalho

O restante deste trabalho está organizado da seguinte forma. No Capítulo 2 encontra-se a revisão dos conceitos envolvidos no desenvolvimento do trabalho, dentre os quais tem-se: investigação do estado-da-arte relacionado ao tema, digitalização de sinais de voz, apresentação das abordagens de extração de características baseadas no conceito de energia, Operador de Energia de Teager, *zero crossing rate*, planaridade espectral, entropia espectral, Engenharia Paraconsistente de Características e metodologias de normalização de vetores de características. Prosseguindo, o Capítulo 3 contém a descrição da base de dados utilizada, assim como a estratégia proposta para o problema abordado no documento, enquanto o Capítulo 4 contém os testes e resultados obtidos durante os experimentos. Finalmente, o Capítulo 5 contém as conclusões aferidas, as quais antecedem as referências.

# Capítulo 2

# Conceituação Teórica

## 2.1 Revisão dos Trabalhos Correlatos

No recente trabalho (AKçAY; OGUZ, 2020), o qual foi selecionado a partir de uma busca pelo termo chave *speech emotion recognition* no *Web of Science*, foi apresentada uma revisão do estado-da-arte de todas as subáreas presentes dentro de sistemas SER: bases de dados, conjunto de características, técnicas de pré-processamento, modalidades de apoio, classificadores e modelos emocionais com base em classificação discreta e em dimensões. Por meio do referido artigo, foi possível ter uma visão geral da área e delinear as direções a serem seguidas neste trabalho de mestrado. A partir deste artigo-chave, inúmeros outros trabalhos conexos foram examinados, conforme segue.

No artigo (ZVAREVASHE; OLUGBARA, 2020), realizou-se a classificação de emoções contidas no discurso a partir da seleção de características prosódicas, tais como energia, frequência fundamental e ZCR, além das espectrais, as quais foram obtidas após a transformação do sinal para o domínio de Fourier, e das cepstrais, isto é, dos *Mel Frequency Cepstral Coefficients* (MFCC). A média de acurácia obtida foi de 99.55% ao realizar os experimentos com as bases de dados *Ryerson Audio-visual Database of Emotional Speech and Song* (RAVDESS) e *Surrey Audio-visual Expressed Emotion Database* (SAVEE), sendo uma *Random Decision Forest* (RDF) escolhida como classificador.

Propõe-se, em (TAMULEVIčIUS; KARBAUSKAITė; DZEMYDA, 2019), uma abordagem de extração de características baseada em cálculos de dimensão fractal (DF) dos

sinais de voz. Visto que a DF é capaz de descrever a fragmentação e irregularidade de um sinal, oito algoritmos distintos de cálculo foram apresentados pata tal. O algoritmo de Katz, que é um dos utilizados dentro do estudo, consiste basicamente do somatório da distância Euclidiana entre amostras adjacentes, possibilitando obter valores correspondentes à presença de baixas ou altas frequências dentro do sinal. É importante citar que os cálculos de DF foram realizados após o janelamento do sinal, e por fim, medidas estatísticas foram extraídas dos cálculos. Uma taxa de acurácia média de 96,5% é reportada no artigo, ao se utilizar das bases de dados EMO-DB e *Lithuanian Spoken Language Emotions Database*.

As características extraídas de um sinal de áudio podem variar dependendo da unidade sonora, visto que diferentes fonemas respondem de diferentes formas a emoções distintas. No artigo (DEB; DANDAPAT, 2019), descreve-se a classificação de emoções tendo como etapa de pré-processamento a segmentação dos sinais entre regiões semelhantes a vogais, incluindo semivogais e ditongos, e não semelhantes a vogais, a fim de que a posterior etapa de extração de características seja realizada independentemente para cada uma dessas regiões. Dessa forma, um método de classificação baseado na alternância entre as regiões foi proposto, onde a região com maior performance é considerada para a extração de características durante o treinamento do classificador. As bases de dados EMO-DB, *Interative Emotional Dyadic Motion Capture Database* (IEMOCAP) e FAU-AIBO foram utilizadas, de forma que as respectivas taxas de acurácia de 85,1%, 64,2% e 45,2% foram atingidas pelo estudo, que procura demonstrar que a abordagem proposta se faz mais vantajosa do que o processamento do sinal completo.

O artigo (DIMITROVA-GREKOW; KLIS; IGRAS-CYBULSKA, 2019) contém a descrição de uma estratégia de reconhecimento de emoções baseada exclusivamente em características obtidas a partir da frequência fundamental (F0) dos sinais de voz. A proposta realizada pelo estudo consiste do janelamento do sinal em *frames* de 20ms com uma sobreposição de 10ms entre trechos consecutivos. Após esse processo, a *Fast Fourier Transform* (FFT), versão otimizada da *Discrete Fourier Transform* (DFT), foi aplicada a cada frame e a frequência de maior pico encontrada a fim de determinar a F0. Finalmente, medidas estatísticas são extraídas sobre os cálculos da F0, visto que os sinais possuíam diferentes

comprimentos e, dessa forma, diferentes quantidades de F0 seriam obtidas para cada um. A base de dados *Emotional Speech Corpus developed in AGH University of Science and Technology Krakow* (EMO_DB), que possui 7 diferentes emoções, foi escolhida. No entanto, os testes realizados pelo autor procuraram explorar a classificação de subgrupos de 2, 3 e 4 emoções, com uma acurácia de 89,74%, 76,14% e 62,99% para cada um dos subgrupos respectivamente.

Em (TURSUNOV; KWON; PANG, 2019), foi proposta uma abordagem de extração de características a partir do timbre dos sinais de voz aplicada à classificação de emoções discretas e na dimensão de valência. Três grupos de características foram formados pelos autores: características base, isto é, características mais populares tais como energia e suas derivadas de 1ª ordem, características de timbre e características selecionadas de timbre, isto é, o melhor subconjunto de características do grupo anterior, obtido por meio da aplicação do processo *Sequential Forward Selection* (SFS). Os experimentos foram realizados utilizando as bases de dados EMO-DB e IEMOCAP, visto que uma *Support Vector Machine* (SVM) e uma *Long Short-term Memory Neural Network* (LSTM-RNN) foram utilizadas como classificadores. A melhor taxa de acurácia obtida se deu com base em uma combinação de características-base e de timbres selecionados, implicando uma acurácia de 97,87% ao se utilizar a SVM na classificação baseada na dimensão de valência.

A proposta do artigo (GUDMALWAR; RAO; DUTTA, 2018) consiste na implementação de um SER baseado exclusivamente na utilização de características prosódicas, mais especificamente no uso de 11 medidas estatísticas sobre o tom, energia, ZCR e as três primeiras frequências formantes, com a redução da dimensão do vetor de características sendo realizada por meio do método *Principal Component Analysis* (PCA). As bases de dados EMO-DB e *LDC Emotional Speech-Transcripts* foram escolhidas para os experimentos, de forma que as respectivas taxas de acurácia foram de 75,32% e 84,5% a partir da utilização de uma *Multilayer Perceptron* (MLP) como classificador.

Em (KERKENI et al., 2018), foi realizada uma comparação entre as abordagens de extração de características MFCC e *Modulation Spectral Features* (MSF). O estudo foi realizado a partir do uso das bases de dados EMO-DB e INTER1SP *Spanish Emotional*

*Database*. Os seguintes classificadores foram usados: *Multivariate Linear Regression*, SVM e *Recurrent Neural Network*, de modo que o melhor resultado alcançado se deu com base na combinação de MFCC e MSF ao se utilizar a *Recurrent Neural Network* aplicada à base de dados INTER1SP.

O artigo (ÖZSEVEN, 2019) propõe um novo método de seleção de características como ferramenta de melhoria da acurácia do processo de classificação de emoções contidas na voz. O método proposto possui seu resultado comparado a métodos de seleção de características já estabelecidos na literatura, tais como *Principal Component Analysis* (PCA), *Sequential Foward Selection* (SFS), *Correlation Based Filter* (FCBF), entre outros. Quatro bases de dados foram utilizadas durante os experimentos, visto que os classificadores escolhidos foram os seguintes: SVM, *Multilayer Perceptron* e *k-Nearest Neighbors*. Resultados acerca da diminuição da carga de trabalho, assim como taxas de acurácia são apresentadas no decorrer do estudo, de forma que a maior taxa de acurácia obtida ao se utilizar o método de seleção proposto foi de 85.71% aplicado à base de dados EMO-DB em combinação com o classificador *Multilayer Perceptron*.

Em (JACOB, 2017) foi descrita a comparação da eficiência do reconhecimento de emoções entre os classificadores *Decision Tree* (DT) e *Logistic Regression* (LR). Levando em conta a simplicidade de tais classificadores, o artigo contém a descrição de três tipos de classificação binária com as respectivas taxas de acurácia para cada classificador: classificação de sinais emocionais vs sinais neutros (DT: 84,45% / LR: 68,06%); classificação baseada na dimensão de valência (DT: 87,76% / LR: 69,49%); classificação de emoções positivas: felicidade vs surpresa (DT: 87,31% / LR: 70%).

Uma abordagem inusitada é apresentada no artigo (KAMIńSKA; SAPIńSKI, 2017), onde a classificação das emoções é baseada na votação de um comitê de classificadores. O processo implementado consiste da extração dos descritores mais comumente utilizados em sistemas SER, tais como prosódicos, MFCC, entre outros. Após isso, dá-se início a um processo de seleção de características em que os atributos selecionados são divididos em subconjuntos de características para que finalmente cada classificador possa realizar sua classificação inicial a partir de um destes subconjuntos, visando uma classificação final

que é obtida pela classe com o maior número de predições entre os classificadores. Foram utilizadas a base de dados *Polish Acted Emotional Speech* e uma base de fala espontânea criada pelo próprio autor, com respectivas taxas médias de acuraria de 72% e 82,6%.

Com a enorme variedade de descritores aplicados ao reconhecimento de emoções encontrados na literatura, o artigo (EYBEN et al., 2016) procura estabelecer um padrão minimalista de características a serem extraídas ao invés do uso de força bruta. A escolha do padrão proposto se deu por meio do estudo do potencial das características para causar alterações fisiológicas afetivas na produção da voz, da comprovação de seu uso em estudos anteriores e sua significação teórica. Seis bases de dados foram utilizadas na avaliação das características, entre as quais está a EMO-DB. Ao se utilizar do padrão apresentado pelo artigo, taxas de acurácia de 79.71% e 66.44% foram alcançadas nas dimensões do nível de excitação e de valência, respectivamente.

Em (TRIGEORGIS et al., 2016) foi proposta a abordagem de extração de características e classificação de emoções por meio de *feature learning*. Sendo assim, os sinais de voz brutos são apresentados a uma *Convolutional Neural Network* (CNN), que é responsável por extrair a melhor representação do sinal e propagar a uma *Long Short Term Memory* (LSTM), a qual é por sua vez responsável por realização a classificação em si. A base de dados *Multimodal Corpus of Remote Collaborative and Affective Interactions* (RECOLA) foi utilizada, de forma que os modelos de classificação baseados na dimensão de valência e nível de excitação foram escolhidos para testar a eficiência do sistema.

É descrita, no artigo (VIGNOLO et al., 2016), uma proposta de otimização do banco de filtros mel utilizado na extração de características MFCC com base no uso de um algoritmo genético. A medida de acurácia do classificador é utilizada como o cenário de aptidão para a otimização do banco de filtros. As bases de dados *Simulated Stressed Speech Hindi* e *FAU-Aibo* foram utilizadas com as respectivas acurácias de 91,31% e 42,50%.

No artigo (WANG et al., 2015) foi apresentada uma abordagem de extração de características derivadas exclusivamente do espectro de Fourier, tais como valores de magnitude, harmônicos, média, máximo, mínimo e desvio padrão, entre outros, para a classificação de emoções. Foram utilizados os classificadores *Naive Bayes* e SVM aplicados às seguin-

tes bases de dados: EMO-DB, *Chinese Emotional Database* (CASIA) e *Chinese Elderly Emotional Speech Database* (EESDB). A abordagem proposta apresenta um desempenho superior quando comparada com a utilização de características tais como MFCC e prosódicas.

A análise realizada durante a leitura dos artigos revisados permite observar que, apesar das distintas abordagens adotadas na implementação de soluções para o reconhecimento de emoções no discurso, o foco da literatura relacionada ao tema encontra-se na discussão sobre qual o conjunto de características mais apropriado a estabelecer um bom nível de separabilidade entre as classes, de forma que o classificador possa realizar sua tarefa com a maior taxa de acurácia possível sem necessitar ser demasiadamente complexo. Desse modo, este trabalho procura explorar características baseadas nos conceitos de energia, operador de energia de Teager, ZCR, planaridade espectral e entropia espectral, de forma que um vetor de características final seja obtido com base na análise paraconsistente, a qual se utiliza de técnicas elementares visando uma avaliação independente do vetor de características, ou seja, de forma que o mesmo possa ser utilizado juntamente com classificadores distintos, permitindo generalização. Tal abordagem inexiste na literatura relacionada para a finalidade proposta.

## 2.2 Conceitos Essenciais

### 2.2.1 Amostragem e Quantização

A voz humana, do ponto de vista do campo de Processamento Digital de Sinais (PDS), pode ser ententida como um sinal natural de única dimensão, com valores de amplitude variando em função do tempo, formando assim o que denominamos de sinal analógico, onde ambas as grandezas citadas apresentam valores contínuos e de precisão infinita em suas dimensões, o que impossibilita sua manipulação direta por computadores digitais.

Contrariamente ao caso anterior, esta dissertação envolve a manipulação de sinais digitais, os quais podem ser definidos como sinais considerados discretos no tempo e quan-

tizados na amplitude (ANTONIOU, 2006).

A amostragem de um sinal é definida pelo Teorema da Amostragem de Nyquist: para que um sinal possa ser reconstruído sem perdas, a taxa de amostragem deve ser ao menos igual a duas vezes a maior frequencia contida no sinal. Assim, visto que a maior frequência audível ao ser humano é de 22050Hz, é comum adotar uma taxa de amostragem equivalente a 44100 amostras por segundo no caso de sinais acústicos. Para a voz humana, entretanto, a largura de banda costuma ser ainda menor e, assim, é comum que tais sinais sejam amostrados a taxas de 8000 ou 16000 amostras por segundo. Com relação à quantização, esta pode assumir intervalos variados a depender do interesse de aplicação do sinal, no entanto comumente podemos encontrar uma taxa equivalente a dois bytes por amostra no caso de sinais acústicos, o que permite $2^{16} = 65536$ valores distintos de amplitude.

### 2.2.2 Abordagens de extração de características baseadas no conceito de energia

Fisicamente, a energia de um sinal pode ser definida como a representação da sua capacidade de realizar trabalho. Se estendermos essa definição para o sinal de voz, podemos entender a energia como o trabalho que o pulmão e as cordas vocais realizam para produzir o som em função do tempo (GUIDO, 2016a). Por meio da Equação (2.1), defini-se a energia $(E)$ de um sinal discreto $s[\cdot]$ com $N$ amostras como

$$E = \sum_{i=0}^{N-1} s(i)^2 \qquad . \tag{2.1}$$

Particularmente, o artigo (GUIDO, 2016a) contém a descrição de três abordagens inovadoras, denominadas $A_1$, $A_2$ e $A_3$, para o uso do conceito de energia como base para extração de características que proporcionam análises distintas dos sinais de voz, as quais são utilizadas nesta dissertação.

### Abordagem $A_1$

O método $A_1$, que pode ser caracterizado como o mais simples entre as abordagens propostas, trata de agrupar segmentos do sinal de entrada, de forma que a energia normalizada destes possa ser calculada conforme os seguintes passos descritos abaixo e ilustrados na Figura 2.1:

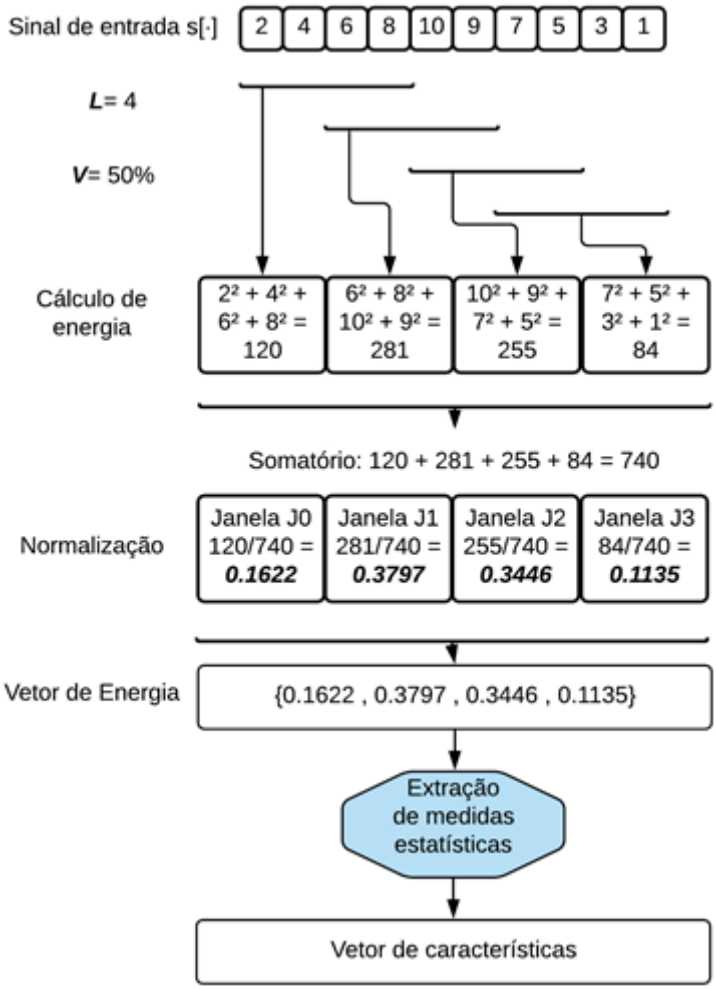
- PASSO 1: segmentar o sinal em janelas de tamanho ***L***;
- PASSO 2: definir uma taxa de sobreposição ***V*** entre as janelas, de forma que informação do sinal não seja perdida;
- PASSO 3: calcular a energia de cada janela com a equação (2.1);
- PASSO 4: normalizar a energia de cada janela;
- PASSO 5: finalmente, uma série de medidas estatísticas são extraídas do vetor de energia gerado, visto que os sinais da base de dados utilizada possuem comprimentos variáveis, proporcionando quantidades distintas de janelas.

### Abordagem $A_2$

O método $A_2$, assim como o anterior, baseia-se no janelamento do sinal de entrada, no entanto nesta abordagem podemos notar a variação do tamanho da janela ***L*** e a inexistencia de *overlaps* entre os segmentos. A motivação fundamental por trás de $A_2$ reside na inspeção do sinal em diferentes níveis de resolução, visto que o vetor de características gerado segundo este método é formado a partir da concatenação entre ***Q*** subvetores de dimensões distintas, como descrito a seguir e representado na Figura 2.2:

- PASSO 1: definir a quantidade ***Q*** de subvetores;
- PASSO 2: segmentar o sinal em janelas de acordo com a dimensão de cada subvetor ***Q***;
- PASSO 3: calcular a energia normalizada para cada janela de cada subvetor ***Q***;

Figura 2.1: Extração do vetor de características segundo o método $A_1$.




Fonte: Elaborado pelo autor, 2021.

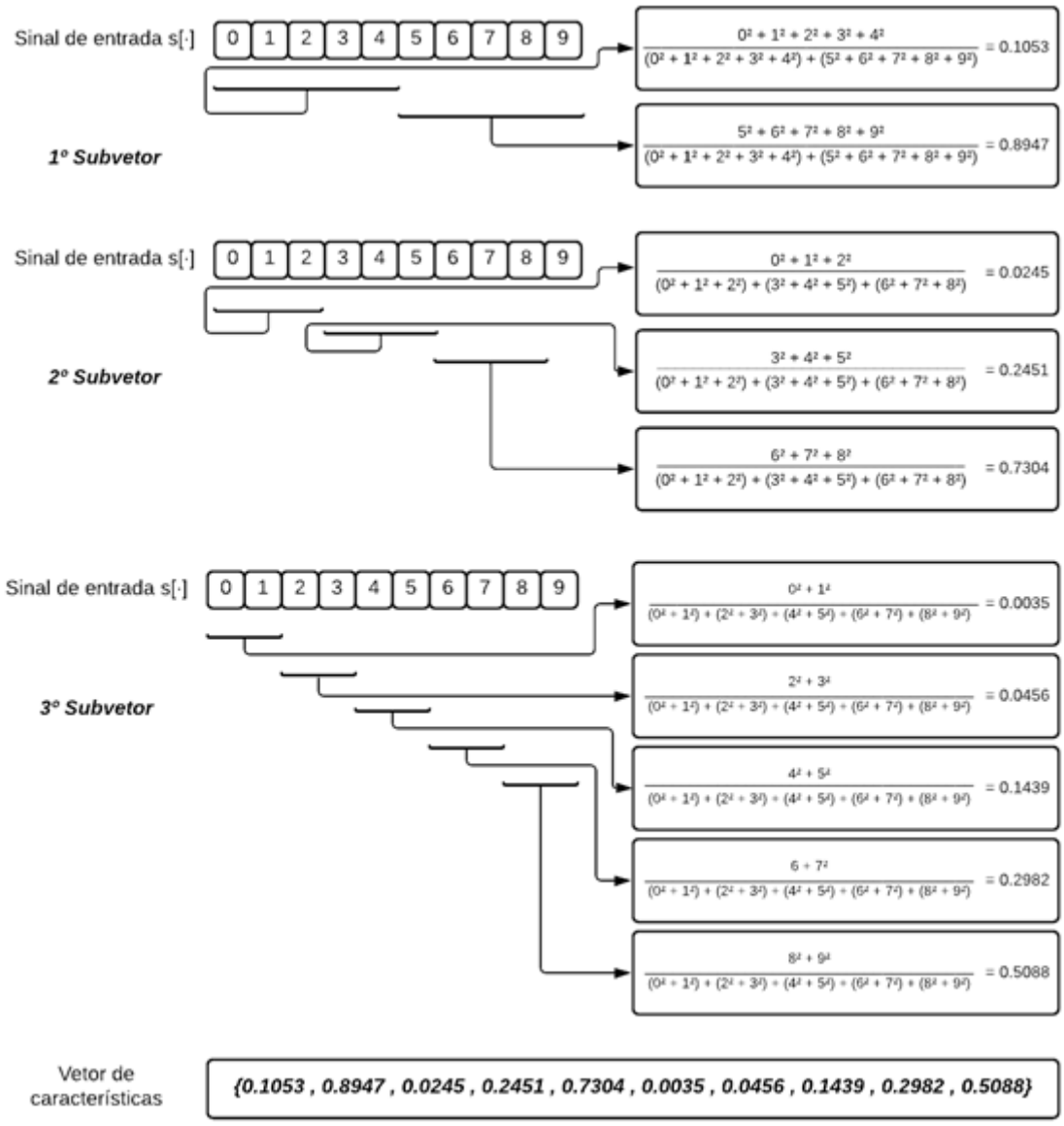
- PASSO 4: concatenar os subvetores ***Q***, o que resulta na formação do vetor de características final.

### Abordagem $A_3$

Este método consiste, diferentemente dos dois anteriores, na determinação dos comprimentos de sinal requeridos para atingir níveis pré-estabelecidos de energia. O cálculo do vetor de características segundo $A_3$ é obtido por meio das etapas a seguir descritas, como pode-se ver na Figura 2.3:

- PASSO 1: calcular a energia do sinal segundo a equação (2.1);

Figura 2.2: Extração do vetor de características segundo o método $A_2$




Fonte: Elaborado pelo autor, 2021.

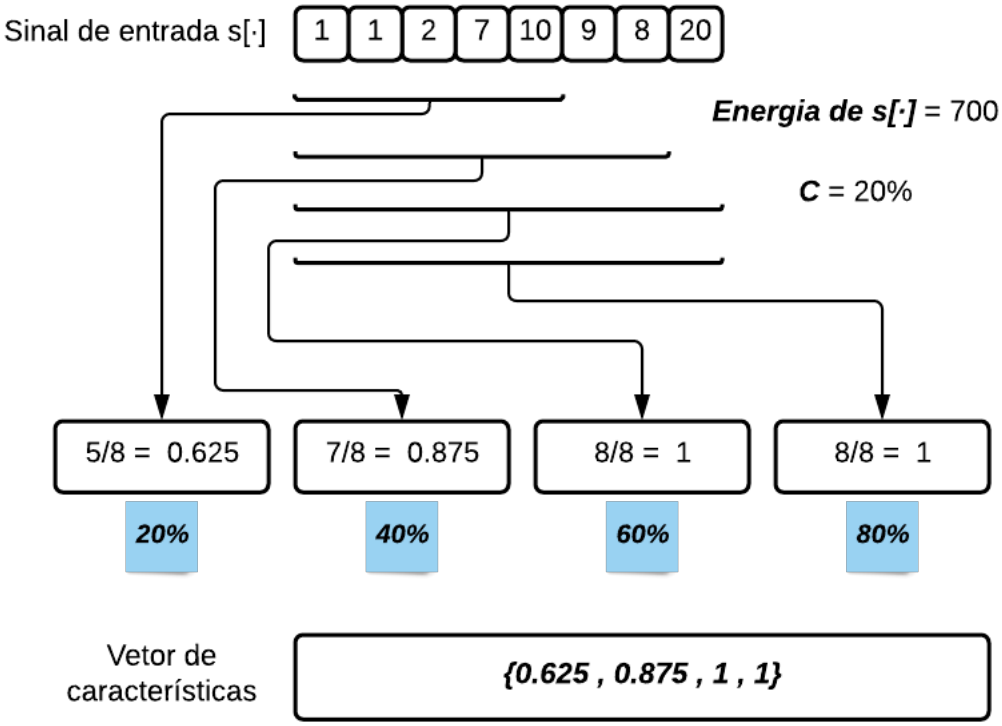
- PASSO 2: estabelecer um nível ***C*** de energia desejado;
- PASSO 3: calcular o tamanho ***T*** do vetor a ser gerado com a equação (2.2)

$$T = (\frac{100}{C}) - 1; \tag{2.2}$$

- PASSO 4: calcular os subníveis de energia a partir de ***C***;

- PASSO 5: verificar o comprimento de sinal requerido para atingir o subnível desejado;
- PASSO 6: dividir o comprimento obtido na etapa anterior pelo comprimento do sinal.

Figura 2.3: Extração do vetor de características segundo o método $A_3$




Fonte: Elaborado pelo autor, 2021.

### 2.2.3 Operador de energia de Teager

A abordagem comumente utilizada para estimar a energia de um sinal, como descrita na equação (2.1), considera apenas valores de amplitude como objeto de cálculo. Assim, ao se utilizar dessa visão convencional, pode ser notado que dois sinais de amplitude unitária, com 10Hz e 1000Hz respectivamente, possuem a mesma energia. Entretanto, a energia requerida para gerar o sinal de 1000Hz é consideravelmente maior que a mesma para o sinal de 10Hz (KAISER, 1990).

Contrariamente ao caso tradicional, o operador de energia de Teager procura estimar a energia de um sinal enfatizando a importância de analisar o mesmo do ponto de vista

de sua geração, assim como demonstrado em (KAISER, 1990). Desse modo, percebeu-se que a energia requerida para gerar um sinal é proporcional não somente ao quadrado das amplitudes, mas também ao quadrado das frequências. Dessa forma, o mesmo exemplo descrito anteriormente produziria energias distintas para os sinais de 10Hz e 1000Hz segundo esta análise.

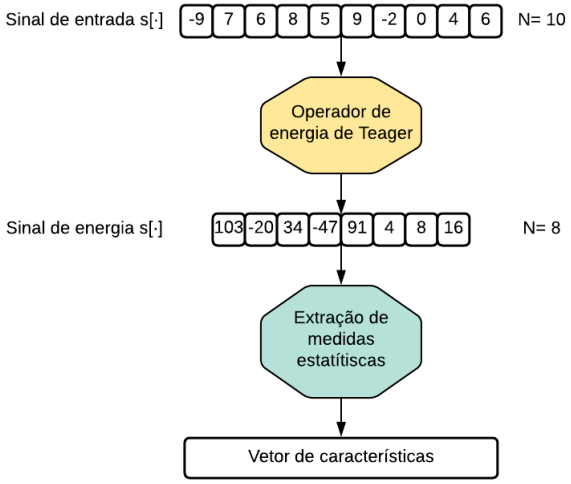
Ao se analisar um sinal discreto $s[\cdot]$, o operador de energia de Teager $(E[\cdot])$ é definido por (KAISER, 1990) como:

$$E(i) = s(i)^2 - s(i-1)s(i+1) \quad . \tag{2.3}$$

Pode-se observar a partir da equação que o operador de energia de Teager se utiliza de três amostras adjacentes de um sinal original com $N$ amostras para o cálculo de um novo sinal de energia com $N - 2$ amostras, diferentemente da abordagem comum de energia apresentada na equação (2.1), onde um valor escalar é produzido.

Observando-se esta propriedade de dependência do comprimento do sinal, o presente trabalho implementa a extração das medidas estatísticas apresentadas na seção 2.2.6 sobre o sinal de energia produzido segundo o operador de energia de Teager. Na Figura 2.4 consta um exemplo da metologia descrita.

Figura 2.4: Extração de características baseada no operador de energia de Teager

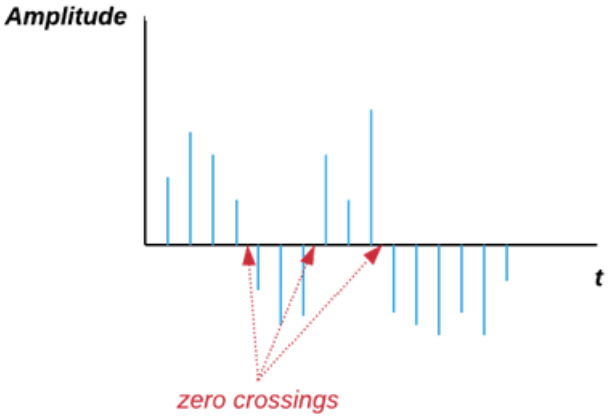



Fonte: Elaborado pelo autor, 2021.

### 2.2.4 *Zero crossing rate*

*Zero crossing rate* (ZCR) refere-se ao número de vezes que o sinal cruza o eixo horizontal em um dado intervalo de tempo, ou seja, quando amostras sucessivas possuem sinais algébricos distintos, como ilustrado na Figura 2.5. Dessa forma, apesar de ser obtida por meio da análise temporal do sinal, o ZCR pode ser entendido como uma medida simples da frequência fundamental do sinal.

Figura 2.5: *Zero crossings* em um sinal




Fonte: Elaborado pelo autor, 2021.

O ZCR de um sinal discreto $s[\cdot]$ com $N$ amostras é definido como

$$ZCR = \frac{1}{2N}\sum_{i=0}^{N-2}|sgn[s(i+1)] - sgn[s(i)]| \quad , \tag{2.4}$$

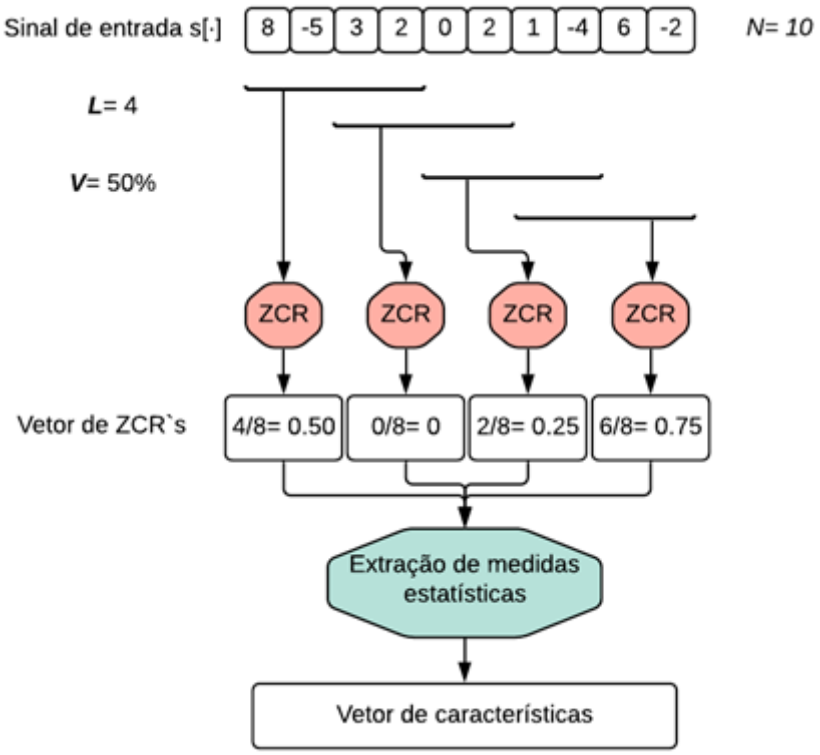
onde $sgn$ corresponde a seguinte função simbólica:

$$sgn[s(i)] = \begin{cases} 1, s(i) \geq 0 \\ -1, s(i) < 0 \end{cases} \quad . \tag{2.5}$$

Sinais de voz apresentam uma considerável variabilidade espectral e, assim, a interpretação global do ZCR sobre eles é imprecisa. Entretanto, estimativas aproximadas de propriedades espectrais podem ser obtidas usando uma representação baseada na abordagem *short-time zero crossing rate* (RABINER; SCHAFER, 1978). Uma vez que altas frequências implicam altos *zero crossing rates*, e baixas frequências implicam baixos *zero*

*crossing rates*, há uma forte correlação entre ZCRs e frequências fundamentais (RABINER; SCHAFER, 1978). Esta abordagem, a qual é utilizada no presente trabalho, é exemplificada pela Figura 2.6.

Figura 2.6: Extração de características baseada na abordagem *short-time zero crossing rate*




Fonte: Elaborado pelo autor, 2021.

### 2.2.5 Características espectrais

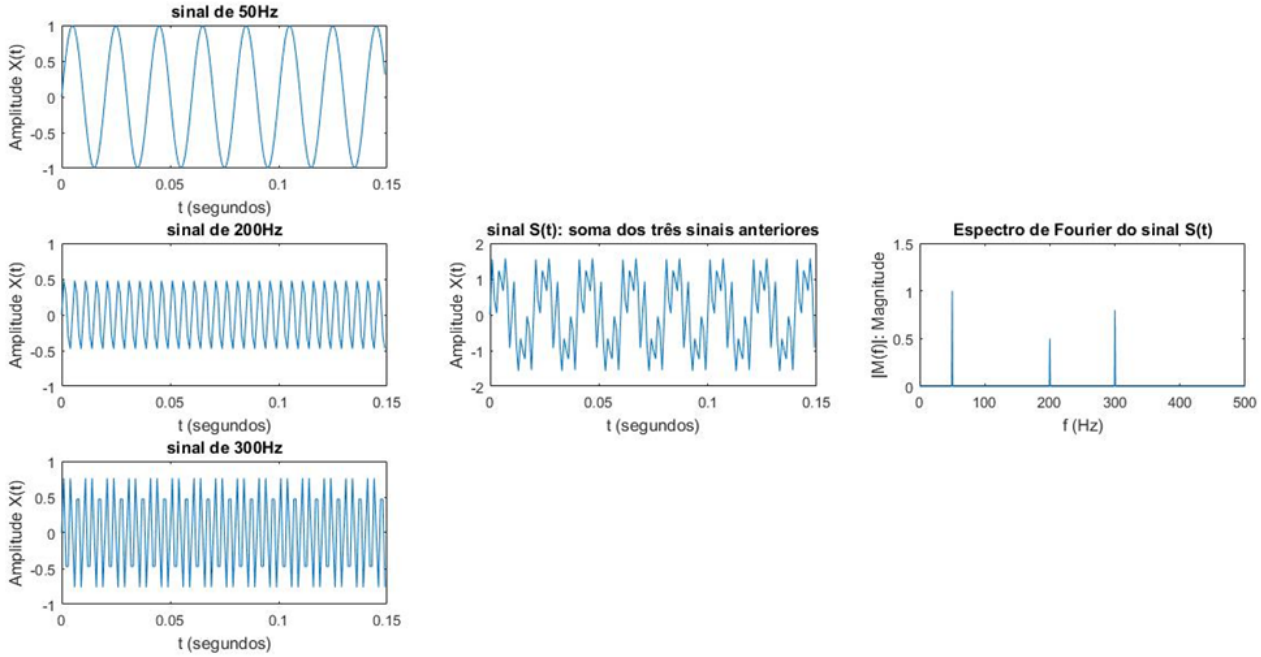
As características descritas até o momento foram extraídas a partir de uma análise temporal dos sinais, no entanto, faz-se extremamente útil que uma inspeção mais completa dos sinais, em termos de suas frequências, também seja realizada. O domínio da frequência, ou simplesmente domínio espectral, é capaz de fornecer informações detalhadas acerca dos componentes de frequência que compõem o sinal (ANTONIOU, 2006).

As ferramentas matemáticas mais comumente utilizadas no processo de transformação do sinal do domínio do tempo para o domínio da frequência são as denominadas Série

de Fourier eTransformada de Fourier, visto que a primeira é capaz de lidar com sinais contínuos periódicos, enquanto a segunda é aplicável também a sinais contínuos não-periódicos.

Segundo a Série de Fourier, qualquer sinal é composto por uma soma de senos e cossenos de diferentes frequências, assim, a partir da aplicação desta ferramenta matemática a um sinal no domínio do tempo é possível visualizar as frequências que compõem o mesmo, tornando possível a representação do sinal no domínio da frequência. A Figura 2.7 faz referência ao processo descrito até o momento.

Figura 2.7: Decomposição das frequências contidas em um sinal a partir da Série de Fourier




Fonte: Elaborado pelo autor, 2021.

A forma mais geral da série de Fourier é dada por (ANTONIOU, 2006)

$$\tilde{x}(t) = \sum_{k=-\infty}^{\infty} X_k e^{jk\omega_0 t} \quad , \tag{2.6}$$

onde $\omega_0 = \frac{2\pi}{T}$ e $e^{jk\omega_0 t}$ são a base para caracterização de uma senóide complexa, a qual pode ser decomposta a partir da Fórmula de Euler, isto é,

$$e^{jk\omega_0 t} = cos(k\omega_0 t) - jsen(k\omega_0 t) \quad , \tag{2.7}$$

sendo $j = \sqrt{-1}$.

A transformada de Fourier, diferentemente da anterior, é responsável por lidar com sinais contínuos não-periódicos. Esta extensão ocorre devido ao fato de que esta ferramenta matemática trata sinais não-periódicos como sinais periódicos de período infinito (ANTONIOU, 2006). A Transformada de Fourier é definida como (ANTONIOU, 2006)

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \qquad . \tag{2.8}$$

Enquanto as ferramentas matemáticas anteriores são responsáveis pela manipulação de sinais contínuos, a Transformada Discreta de Fourier ou *Discrete Fourier Transform* (DFT) trata da decomposição de sinais discretos, os quais são objetos de estudo do presente trabalho. Assim, a definição da DFT de um sinal discreto $s[\cdot]$ com $N$ amostras é dada por

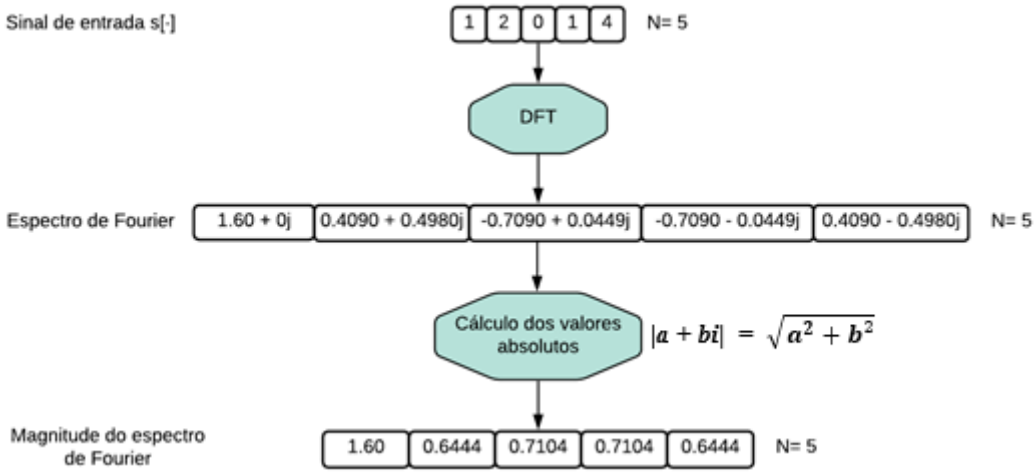
$$X(k) = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{-\frac{j2\pi kn}{N}} \qquad , \tag{2.9}$$

onde $\frac{1}{N}$ corresponde ao fator de normalização dos coeficientes complexos de frequência calculados. Finalmente, é comum que dentro do campo de processamento digital de sinais apenas valores reais sejam manipulados. Assim, costuma-se calcular os valores absolutos dos coeficientes $X(k)$, formando o que se denomina de magnitude do espectro de Fourier, de onde extrai-se as características espectrais ou de frequência. A Figura 2.8 contém um exemplo do processo descrito.

Como pode-se ver a partir da equação (2.9), o cálculo de cada coeficiente de frequência exige o processamento de todas amostras do sinal no domínio no tempo, assim sendo, pode-se aferir uma complexidade equivalente a $O(N^2)$ no cálculo da transformada discreta de Fourier. O estudo realizado no artigo (COOLEY; TUKEY, 1965) permitiu demonstrar que o cálculo realizado pela DFT acaba por processar uma quantidade considerável de redundância, assim, por meio de um método engenhoso conhecido como transformada rápida de Fourier ou *fast Fourier transform* (FFT), os autores mostraram que uma significativa quantidade de computação poderia ser eliminada sem degradar a precisão da DFT (ANTONIOU,

2006).

Figura 2.8: Obtenção da magnitude do espectro de Fourier




Fonte: Elaborado pelo autor, 2021.

Esta dissertação implementa o algoritmo FFT segundo a metodologia demonstrada em (COOLEY; TUKEY, 1965), a qual produz uma complexidade equivalente a $O(N \log N)$, como pode se observar a partir da Figura 2.9.

### Planaridade espectral

Planaridade espectral ou *spectral flatness*, introduzida no artigo (GRAY; MARKEL, 1974), pode ser definidida como a característica responsável por aferir a quantidade de ruído presente no espectro do sinal. Esta quantificação é definida como a razão entre a média geométrica e a média aritmética da magnitude do espectro do sinal, conforme demonstrado na equação (2.10):

$$Flatness = \frac{\prod_{k=0}^{K-1} x(k)^{\frac{1}{K}}}{\frac{1}{K}\sum_{k=0}^{K-1} x(k)} \quad , \tag{2.10}$$

onde $x(k)$ corresponde a cada coeficiente de frequência presente na magnitude do espectro de Fourier. Assim, espera-se que um sinal totalmente ruídoso, ou seja, com todos componentes de frequência igualmente presentes, produza um valor igual a 1, enquanto um sinal

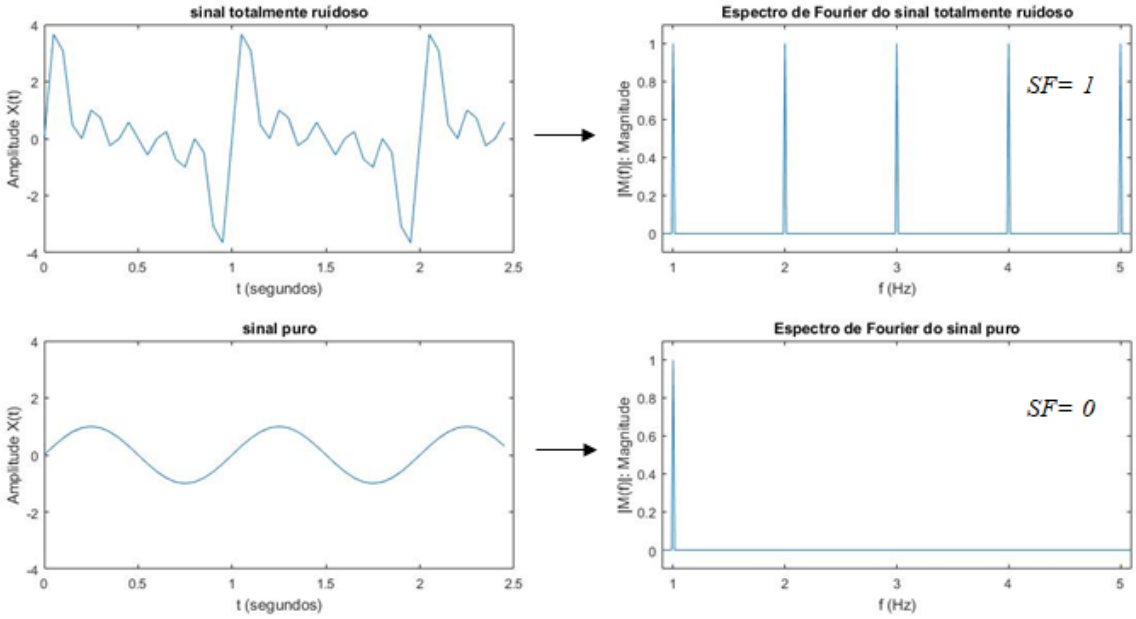
Figura 2.9: Implementação do algoritmo *fast Fourier transform* (FFT)

```
public void fft(Complex[] signal){
    int N= signal.length;

    if(N==1){
        return;
    }
    if(N%2!=0){
        throw new IllegalArgumentException("N não é potencia de 2");
    }

    Complex[] even= new Complex[(N/2)];
    Complex[] odd= new Complex[(N/2)];
    for(int i=0; i<(N/2); i++){
        even[i]= signal[(2*i)];
        odd[i]= signal[(2*i+1)];
    }
    fft(even);
    fft(odd);
    //end of recursion

    for(int k=0; k<(N/2); k++){

        Complex t = Complex.polar(1.0, -2 * Math.PI * k / N).multiply(odd[k]);
        signal[k] = even[k].add(t);
        signal[k + N/2] = even[k].sub(t);
    }
}
```

Fonte: Elaborado pelo autor, 2021.

com apenas um componente de frequência resulte em uma planaridade espectral igual a 0, como ilustrado na Figura 2.10.

Figura 2.10: Valores de *spectral flatness* para um sinal totalmente ruidoso e um sinal puro, respectivamente




Fonte: Elaborado pelo autor, 2021.

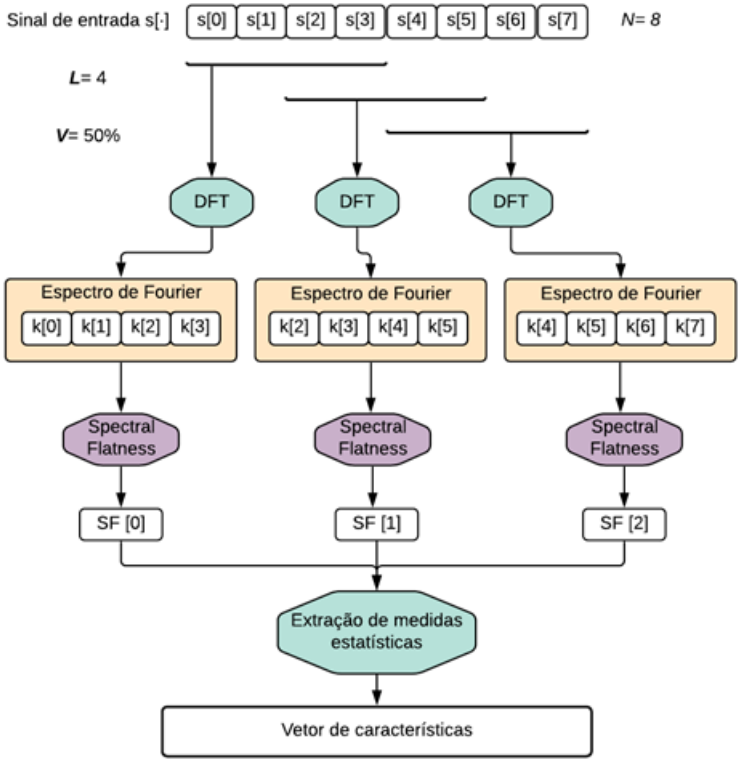
Considerando-se a variabilidade no tamanho dos sinais de voz, optou-se pela implementação do cálculo dos valores de planaridade espectral a partir de uma análise *frame-by-frame* dos sinais, conforme descrito no algoritmo abaixo e ilustrado na Figura 2.11.

- PASSO 1: segmentar o sinal em janelas de tamanho ***L***;
- PASSO 2: definir uma taxa de sobreposição ***V*** entre as janelas, de forma que informação do sinal não seja perdida;
- PASSO 3: calcular a FFT de cada janela [1] com o algoritmo ilustrado na Figura 2.9 (calcular os valores absolutos dos coeficientes de frequência);
- PASSO 4: realizar o cálculo de planaridade espectral de cada janela de coeficientes de frequência;

[1] a metodologia de calcular a FFT de janelas que percorrem o sinal é também conhecido como *Short-Time Fourier Transform* ou Transformada de Fourier de Tempo Curto

- PASSO 5: finalmente, uma série de medidas estatísticas são extraídas do vetor de planaridades espectrais gerado, visto que os sinais da base de dados utilizada possuem comprimentos variáveis, proporcionando quantidades distintas de janelas.

Figura 2.11: Extração de características a partir do cálculo de planaridade espectral




Fonte: Elaborado pelo autor, 2021.

### Entropia espectral

A entropia espectral de um sinal pode ser entendida como a medida de sua distribuição de potencia espectral. Este conceito deriva-se da entropia de informação, ou entropia de Shannon, proposta em (SHANNON, 1948).

Adaptando-se a metodologia proposta em (SHANNON, 1948), o espectro de frequências é tratado como uma distribuição de probabilidade, como pode ser visto na equação (2.11).

$$p(k) = \frac{x(k)}{\sum_{k=0}^{K-1} x(k)} \tag{2.11}$$

com

$$\sum_{k=0}^{K-1} p(k) = 1 \qquad , \tag{2.12}$$

onde $x(k)$ refere-se a cada coeficiente de frequência presente na magnitude do espectro de Fourier e $p(k)$ é probabilidade normalizada deste mesmo coeficiente.

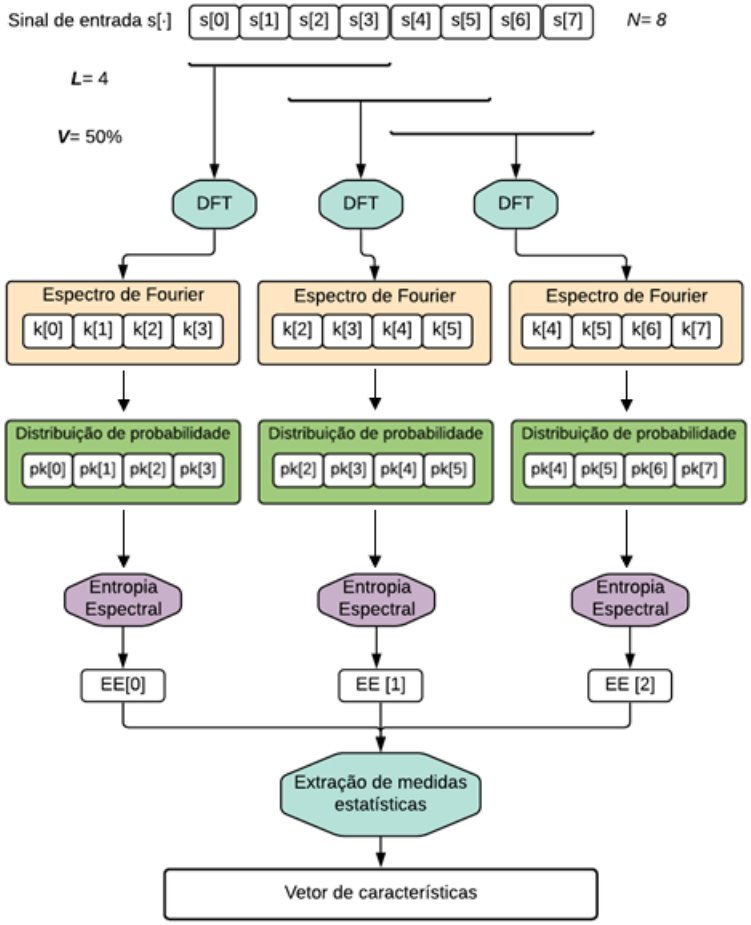
Assim, a entropia espectral $(EE)$ pode ser definida como

$$EE = -\sum_{k=0}^{K-1} p(k)\log(p(k)) \qquad . \tag{2.13}$$

Similarmente à característica descrita na seção anterior, adotou-se a abordagem de extração da entropia espectral a partir do janelamento do sinal, assim como descrito a seguir e ilustrado na Figura 2.12.

- PASSO 1: segmentar o sinal em janelas de tamanho ***L***;
- PASSO 2: definir uma taxa de sobreposição ***V*** entre as janelas, de forma que informação do sinal não seja perdida;
- PASSO 3: calcular a FFT de cada janela com o algoritmo ilustrado na Figura 2.9 (calcular os valores absolutos dos coeficientes de frequência);
- PASSO 4: calcular a distribuição de probabilidade de cada janela de espectro com a equação (2.11);
- PASSO 5: realizar o cálculo de entropia espectral de cada janela de distribuição de probabilidade, conforme a equação (2.13);
- PASSO 6: finalmente, uma série de medidas estatísticas são extraídas do vetor de entropias espectrais gerado, visto que os sinais da base de dados utilizada possuem comprimentos variáveis, proporcionando quantidades distintas de janelas.

Figura 2.12: Extração de características baseada no conceito de entropia espectral




Fonte: Elaborado pelo autor, 2021.

### 2.2.6 Medidas estatísticas

Na literatura, características são organizadas em duas classes: *low-level descriptors* (LLDs) e *functionals* (GUDMALWAR; RAO; DUTTA, 2018). Aqueles incluem características prosódicas e espectrais, enquanto estes correspondem aos seus valores estatísticos. No presente trabalho, como descrito até o momento, algumas características produzem vetores de comprimento fixo, enquanto outras acabam por gerar vetores de comprimentos variáveis, como sumarizado na Tabela 2.1.

Tabela 2.1: Variabilidade de comprimento das características extraídas no presente trabalho

| **Características** | **Vetores de comprimento fixo** |
|---|---|
| Abordagem $A_1$ de energia | NÃO |
| Abordagem $A_2$ de energia | SIM |
| Abordagem $A_3$ de energia | SIM |
| Operador de energia de Teager | NÃO |
| *Zero crossing rate* (ZCR) | NÃO |
| Planaridade espectral | NÃO |
| Entropia espectral | NÃO |

Fonte: Elaborado pelo autor, 2021.

A partir disso, faz-se comum dentro da literatura de SER, como pode-se observar em (GUDMALWAR; RAO; DUTTA, 2018) e (TAMULEVIčIUS; KARBAUSKAITė; DZEMYDA, 2019), que *functionals* sejam extraídos de LLDs objetivando-se a seguinte propriedade: geração de vetores de comprimentos fixos e que correspondam a uma generalização do sinal como um todo. As seguintes medidas estatísticas são utilizadas nesta dissertação:

- Percentis: 0 a 99%;
- Valor máximo;
- Valor mínimo;
- Desvio padrão;
- Intervalo de valores;
- Média aritmética;

- Variância.

### 2.2.7 Engenharia Paraconsistente de Características

Dentro do domínio de Reconhecimento de Padrões, uma classificação precisa dos sinais sob análise depende não exclusivamente do classificador, mas também da adequabilidade dos vetores de características extraídos ao problema em questão. A Engenharia Paraconsistente de Características (GUIDO, 2019) faz uso de técnicas elementares a fim de proporcionar uma análise quantitativa de tais vetores independentemente do classificador a ser utilizado.

A análise realizada segundo a Engenharia Paraconsistente exige a normalização prévia dos vetores de características dentro do intervalo [0, 1] para o cálculo de dois critérios independentes que fornecem a base da análise, são eles:

- $\alpha$: menor similaridade intraclasse;
- $\beta$: taxa de sobreposição interclasse.

Idealmente, espera-se que $\alpha$ e $\beta$ tenham seu valores próximos a 1 e 0 respectivamente, indicando que os vetores de características estabelecem um bom nível de separabilidade entre as classes envolvidas. Considerando que o problema abordado possua N classes, o cálculo de $\alpha$ é realizado conforme os passos descritos a seguir e ilustrados na Figura 2.13.

- PASSO 1: selecionar os maiores e menores valores de cada característica entre os vetores de uma mesma classe;
- PASSO 2: calcular ***A*** = maior valor - menor valor;
- PASSO 3: calcular ***Y*** = 1 - ***A***, de forma que ***Y*** $\approx$ 1 e ***Y*** $\approx$ 0 indicam alta e baixa similaridade intraclasse, respectivamente;
  - Exemplificando um problema com 3 classes, realizadas as etapas anteriores seriam produzidos os vetores de similaridade ***svC0***, ***svC1*** e ***svC2***.

- PASSO 4: calcular a média de cada vetor de similaridade, ou seja, ***média(svC0)***, ***média(svC1)*** e ***média(svC2)***;
- PASSO 5: obter $\alpha$ = ***min*** = [***média(svC0)***, ***média(svC1)***, ***média(svC2)***].

Terminado o cálculo da taxa de similaridade intraclasse $\alpha$, a próxima etapa da Engenharia Paraconsistente de Características foca na análise da dissimilaridade interclasse, a qual é aferida a partir do cálculo de $\beta$, como ilustrado na Figura 2.14. Conforme descrito em (GUIDO, 2019), a taxa de sobreposição interclasse $\beta$ é definida como:

$$\beta = \frac{R}{F} \quad , \tag{2.14}$$

onde ***R*** é obtido da seguinte forma:

- PASSO 1: seleciona-se os maiores e menores valores de cada característica entre todos os vetores de cada classe, formando assim dois vetores de intervalo para cada classe: um vetor de intervalo dos maiores e um vetor de intervalo dos menores;
- PASSO 2: o valor de ***R*** é obtido a partir da contagem do número de sobreposições entre cada elemento dos vetores de características de cada classe em comparação aos elementos dos vetores de intervalo das outras classes.

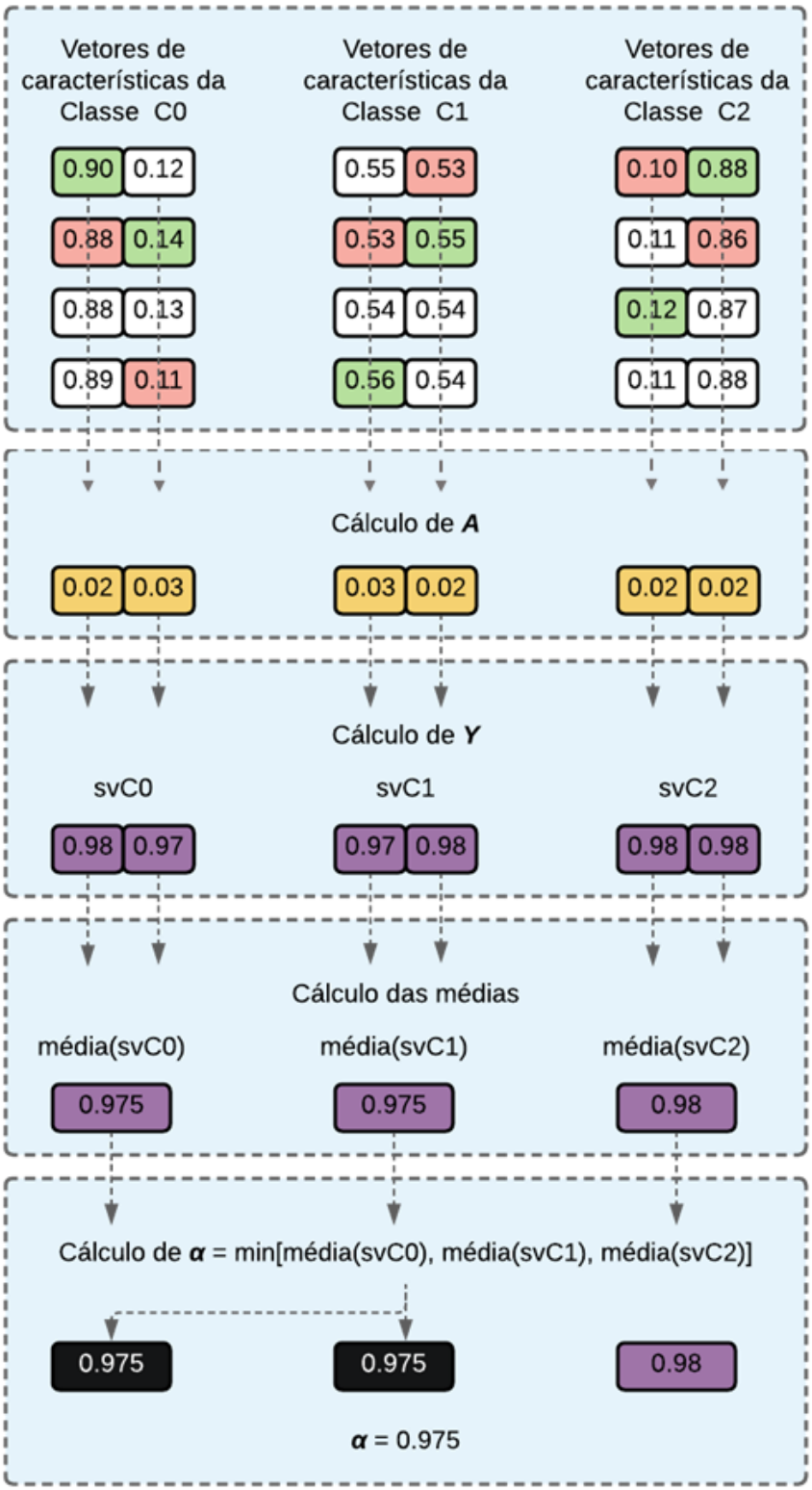
Finalmente ***F***, que é definido como o número máximo de sobreposições possíveis, é calculado conforme a equação (2.15):

$$F = N \cdot (N-1) \cdot X \cdot T \quad , \tag{2.15}$$

onde:

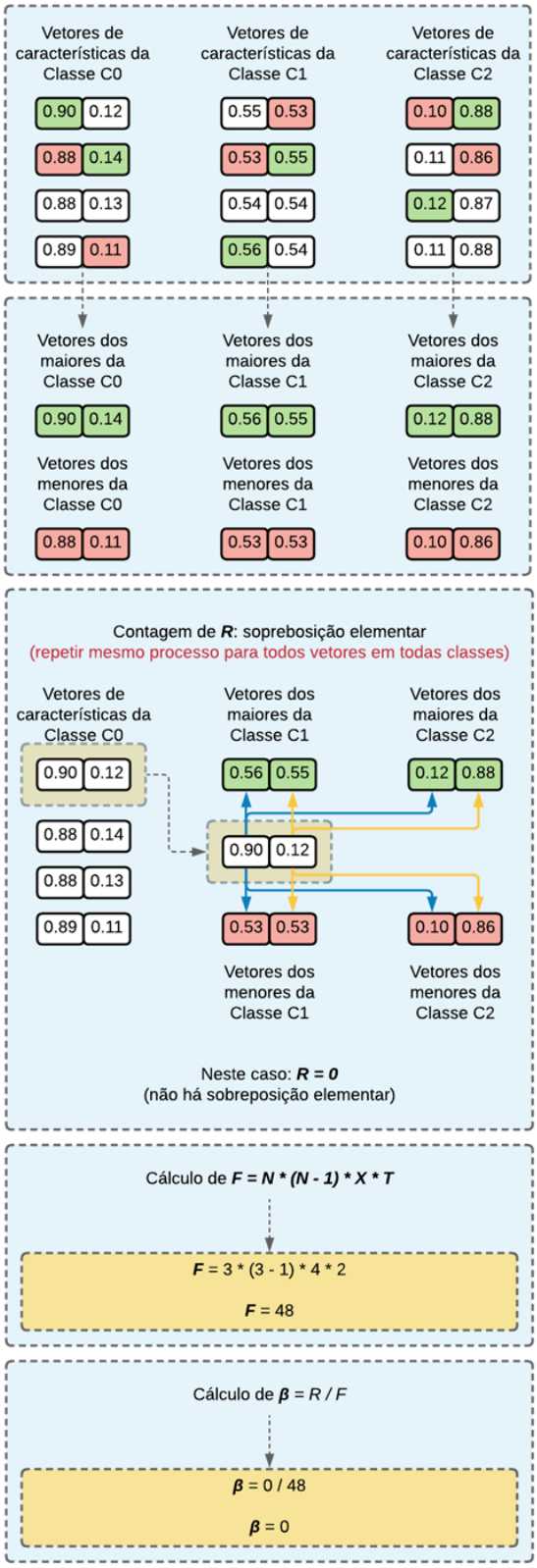
- ***N*** equivale ao número de classes;
- ***X*** representa a quantidade de amostras por classe;
- ***T*** equivale à dimensão do vetor de características.

Figura 2.13: Cálculo de $\alpha$




Fonte: Elaborado pelo autor, 2021.

Figura 2.14: Cálculo de $\beta$




Fonte: Elaborado pelo autor, 2021.

Desconsiderando o melhor caso possível produzido pela análise paraconsistente, onde $\alpha$ = 1 e $\beta$ = 0, qualquer valor dentro do intervalo [0, 1] pode ser assumido pelos critérios acima calculados, assim, o artigo (GUIDO, 2019) contém a definição do grau de certeza $G_1 = \alpha$ - $\beta$ e do grau de contradição $G_2 = \alpha + \beta$ - 1.

A partir dos graus descritos, torna-se possível plotar um ponto $P(G_1, G_2)$ dentro do plano paraconsistente, local onde é aferida a qualidade dos vetores de características, como pode ser visto na Figura 2.15.

Afim de obter maior precisão acerca da posição ocupada pelo ponto $P(G_1, G_2)$ no plano paraconsistente, a distância deste em relação aos quatro cantos ilustrados na Figura 2.15 é calculada conforme as equações (2.16), (2.17), (2.18) e (2.19).

$$D_{-1,0} = \sqrt{(G_1 + 1)^2 + (G_2)^2} \tag{2.16}$$

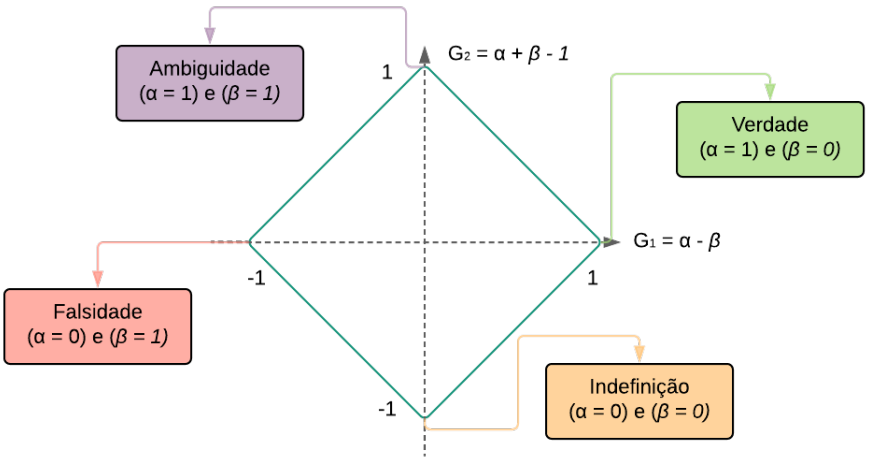
$$D_{1,0} = \sqrt{(G_1 - 1)^2 + (G_2)^2} \tag{2.17}$$

$$D_{0,-1} = \sqrt{(G_1)^2 + (G_2 + 1)^2} \tag{2.18}$$

$$D_{0,1} = \sqrt{(G_1)^2 + (G_2 - 1)^2} \tag{2.19}$$

Finalmente, a partir dos cálculos das distâncias acima descritas, espera-se que o ponto $P(G_1, G_2)$ esteja localizado mais próximo do vértice $(1, 0)$ do que em relação aos outros vértices, indicando que os vetores de características analisados são favoráveis a uma classificação precisa. Ainda, caso as menores distâncias de $P(G_1, G_2)$ sejam em relação aos vértices $(1, 0)$ e $(0, -1)$, respectivamente, a análise paraconsistente indica que as características são linearmente separáveis, ou seja, um classificador modesto pode ser usado como estratégia de solução.

Figura 2.15: Plano paraconsistente




Fonte: Elaborado pelo autor, 2021.

### 2.2.8 Normalização

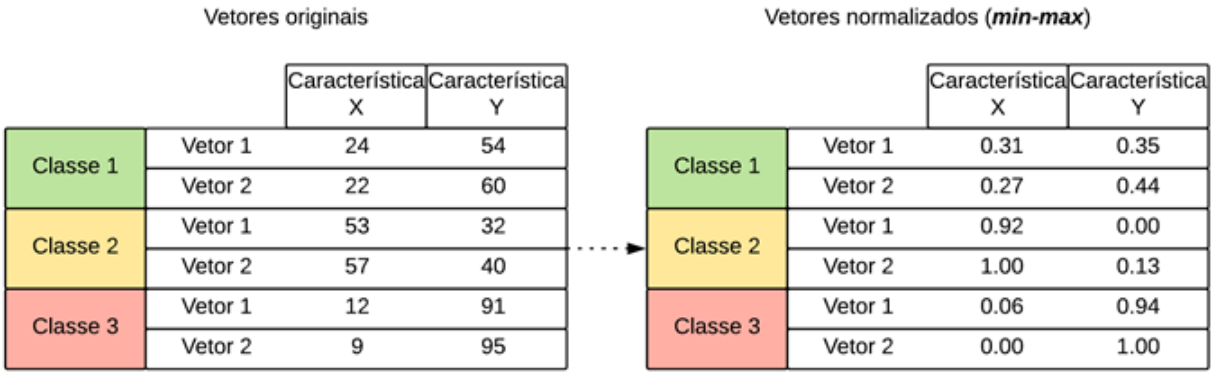
A normalização de vetores de características faz-se uma prática extremamente comum dentro da área de Reconhecimento de Padrões. Além desta etapa ser requerida pela Engenharia Paraconsistente de Características, a mesma acaba por melhorar consideravelmente a performance da maioria dos classificadores utilizados dentro da literatura.

No presente trabalho, pode-se observar que algumas abordagens de extração de características já produzem vetores com valores normalizados, portanto, este procedimento é adotado somente às características que não geram automaticamente valores dentro do intervalo [0, 1]: operador de energia de Teager e entropia espectral. A abordagem de normalização *min-max*, a qual é utilizada nesta dissertação, é definida como:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad . \tag{2.20}$$

A Figura 2.16 exemplifica numéricamente a normalização de 6 vetores de características. Faz-se importante observar que a normalização é realizada a partir do cálculo da equação (2.20) para cada coluna, ou seja, para cada característica.

Figura 2.16: Normalização de vetores de características segundo a abordagem *min-max*




Fonte: Elaborado pelo autor, 2021.

# Capítulo 3

# Abordagem Proposta

## 3.1 Base de Dados

A escolha da base de sinais de voz adotada nesta dissertação deu-se por meio do levantamento e análise do estado-da-arte dos artigos relacionados a SER, como pode-se ver na Seção 2.1. A base de dados *Berlin Emotional Speech Database* (EMO-DB) é constituída por 535 arquivos WAV inicialmente amostrados a uma taxa de 48000Hz durante as gravações com posterior *downsampling* para 16000Hz, enquanto uma resolução de 2 bytes (mono) foi admitida para a quantização dos sinais (BURKHARDT et al., 2005).

As gravações realizadas contaram com a participação de 5 atores e 5 atrizes responsáveis por expressar 10 frases em 7 distintos estados emocionais: raiva, tédio, desgosto, medo, felicidade, neutralidade e tristeza. As tabelas 3.1 e 3.2 contém informações acerca da divisão da quantidade de sinais entre as emoções e as frases pronunciadas, respectivamente. Os sinais aqui utilizados, assim como sua organização podem ser acessados gratuitamente em <http://emodb.bilderbar.info/index-1280.html>.

## 3.2 Estratégia Proposta

Este trabalho possui foco no reconhecimento de emoções na fala a partir da extração manual das características prosódicas e espectrais descritas no Capítulo anterior sucedida pela seleção dos melhores subconjuntos de características segundo à Engenha-

Tabela 3.1: Divisão da quantidade de sinais entre as emoções

| **EMO_DB** | **535 Sinais** |
|---|---|
| Raiva | 127 sinais |
| Tédio | 81 sinais |
| Desgosto | 46 sinais |
| Medo | 69 sinais |
| Felicidade | 71 sinais |
| Neutralidade | 79 sinais |
| Tristeza | 62 sinais |

Fonte: Elaborado pelo autor, 2021.

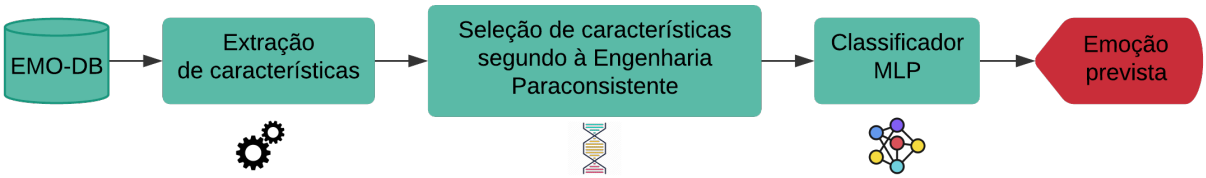
Tabela 3.2: Frases pronunciadas no EMO-DB

| **Texto original em alemão** | **Tentativa de tradução para português** |
|---|---|
| *Der Lappen liegt auf dem Eisschrank*. | O pano está na geladeira. |
| *Das will sie am Mittwoch abgeben*. | Ela entregará na quarta-feira |
| *Heute abend könnte ich es ihm sagen*. | Esta noite eu poderia contar a ele. |
| *Das schwarze Stück Papier befindet sich da oben neben dem Holzstück*. | A folha de papel preta está lá em cima, ao lado do pedaço de madeira |
| *In sieben Stunden wird es soweit sein*. | Em sete horas chegará a hora. |
| *Was sind denn das für Tüten, die da unter dem Tisch stehen?* | E aquelas malas ali embaixo da mesa? |
| *Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter*. | Eles acabaram de carrega-lo para cima e agora estão descendo novamente. |
| *An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.* | Atualmente, nos fins de semana, eu sempre vou para casa e vejo Agnes. |
| *Ich will das eben wegbringen und dann mit Karl was trinken gehen.* | Eu só quero tirar isso e depois tomar uma bebida com Karl. |
| *Die wird auf dem Platz sein*, *wo wir sie immer hinlegen*. | Estará no local onde nós sempre armazenamos. |

Fonte: Elaborado pelo autor, 2021.

ria Paraconsistente de Características. Finalmente, os vetores produzidos são separados aleatóriamente e em diversas proporções entre conjunto de treinamento e teste a fim de que uma *Multi-Layer Perceptron* (MLP) realize a classificação dos vetores. A Figura 3.1 contém a estrutura do sistema proposto nesta dissertação.

Figura 3.1: Estrutura da estratégia proposta




Fonte: Elaborado pelo autor, 2021.

## 3.3 Seleção de Características

A presente dissertação implementa a conversão de cada um dos 535 sinais descritos na Tabela 3.1 para um vetor de características correspondente conforme as metodologias de extração apresentadas na Seção 2.2. A Tabela 3.3 contém informações acerca da distribuição dos vetores de características gerados.

Tabela 3.3: Distribuição das características

| **Indices** | **Características** |
|---|---|
| 0 - 110 | Energia (A2) |
| 111 - 209 | Energia (A3) |
| 210 - 315 | Operador de energia de Teager |
| 316 - 421 | Planaridade espectral |
| 422 - 527 | Energia (A1) |
| 528 - 633 | *Zero crossing rate* |
| 634 - 739 | Entropia espectral |

Fonte: Elaborado pelo autor, 2021.

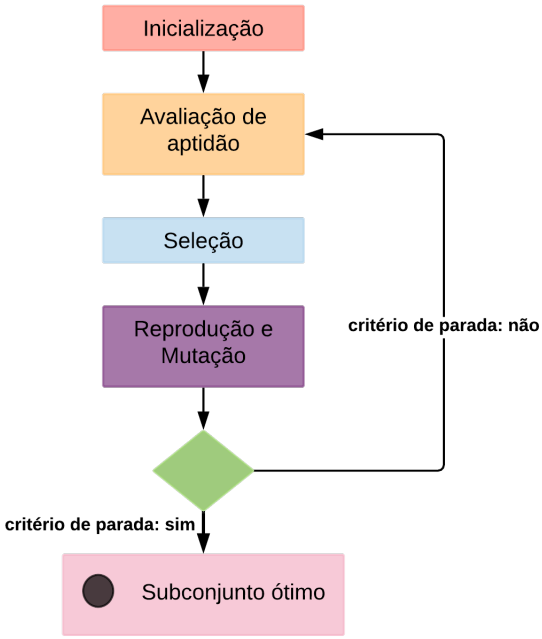
O objetivo da etapa de seleção de descritores é reduzir o número de características, identificando aquelas que apresentam o comportamento mais significativo na classificação dos sinais (GARCÍA-DOMÍNGUEZ et al., 2020). Adicionalmente, a seleção do melhor subconjunto de características desempenha importante papel na construção de classificadores mais precisos e compreenssíveis, a partir da remoção de descritores irrelevantes e redundantes.

Neste trabalho, assim como descrito anteriormente, o processo de seleção de características é realizado com base na Engenharia Paraconsistente de Características. Dessa forma, pode-se denotar matematicamente a seleção de características aqui proposta como

um problema de otimização combinatorial onde a avaliação exaustiva de todos os subconjuntos de características possíveis ($2^N$) é geralmente inviável devido à quantidade de esforço computacional necessária. Como resultado, muitos estudos de pesquisa tem se concentrado em algoritmos de busca global, como Algoritmos Genéticos (AG) para resolver o problema de seleção de recursos (AHUJA; RATNOO, 2015).

A implementação adotada nesta dissertação faz uso de um AG para o processo de seleção de características. Desse modo, apesar de não ser possível verificar qual o melhor subcojunto possível, faz-se factível a obtenção de ótimos subconjuntos de características em um tempo computacional viável. A Figura 3.2 contém o fluxograma do processo de seleção de características com o algoritmo genético.

Figura 3.2: Seleção de características com o algoritmo genético




Fonte: Elaborado pelo autor, 2021.

A etapa de inicialização do AG é responsável por gerar uma população de cromossomos (sequencia de 0‘s e 1‘s) de maneira aleatória, de forma que a dimensão de cada cromossomo é equivalente ao número total de características extraídas. Assim sendo, cada cromossomo representa um subconjunto de características, onde o valor de bit 1 representa a presença da característica, enquanto o valor de bit 0 denota sua ausência.

Após a inicialização, faz-se necessário atribuir um valor de aptidão a cada cromossomo na população. Conforme demonstrado na seção 2.2.7, a Engenharia Paraconsitente permite aferir que vetores de características mais favoráveis ao processo de classificação são aqueles que produzem o ponto P mais próximos ao vértice (1,0), assim, esta dissertação utiliza a equação (2.17) como função de avaliação da aptidão dos individuos. A partir dessa análise, o operador de seleção escolhe os indivíduos que serão recombinados para a próxima geração. O método de de seleção aqui utilizado é a Amostragem Universal Estocástica, também conhecido como *Stochastic Universal Sampling* (SUS), o qual é responsável por colocar os indivíduos em uma "roleta", com a área de cada um deles sendo proporcional à sua aptidão. Em seguida, esta roleta é girada e os indivíduos são selecionados aleatoriamente.

Finalmente, os operadores genéticos podem ser aplicados aos indíviduos selecionados, de forma que uma nova geração seja produzida. A reprodução dos cromossomos se dá a partir de um operador de *crossover* responsável por combinar os bits dos pais para a produção dos filhos, enquanto o operador de mutação é responsável por alterar o valor de alguns bits dos filhos produzidos de maneira aleatoria. A Figura 3.3 contém um exemplo das etapas presentes no AG descritas até o momento.

Inicialmente, alguns experimentos foram realizados a fim de que os parâmetros do AG fossem ajustados visando a melhor performance possível. As configurações estabelecidas encontram-se listadas na Tabela 3.4, sendo N o número de características contidas na Tabela 3.3.

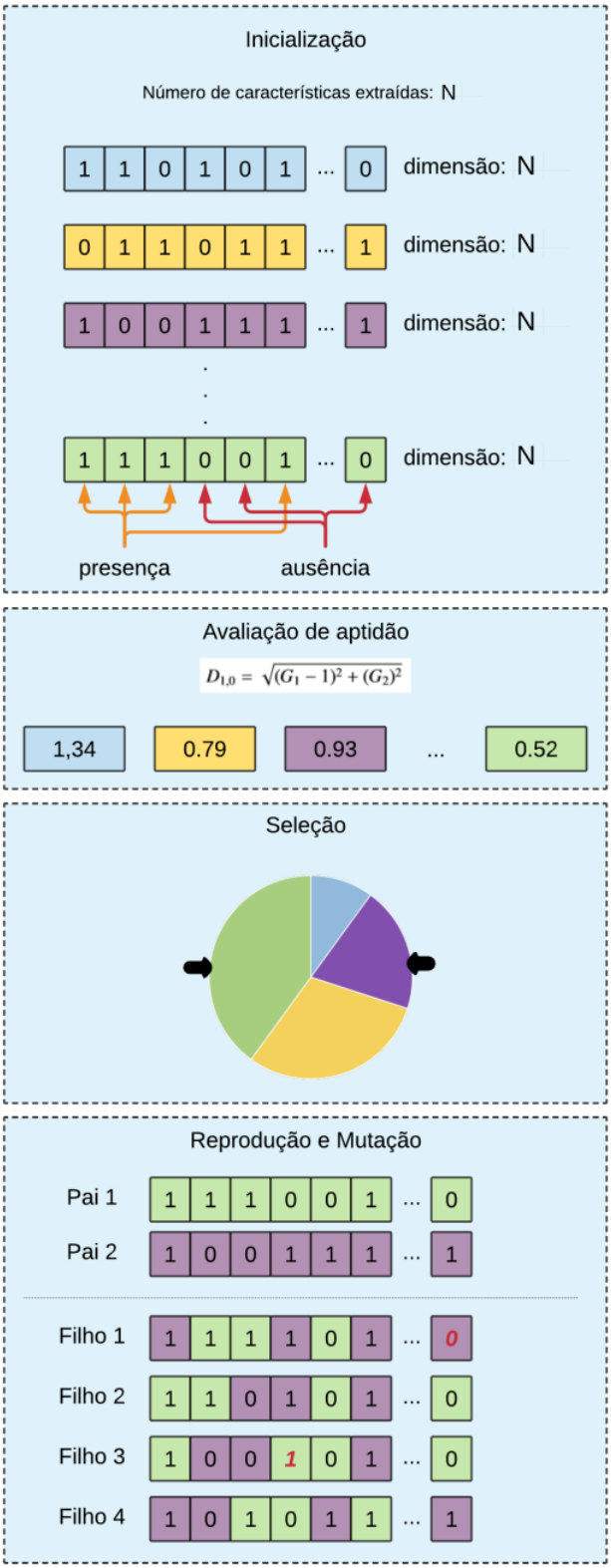
Tabela 3.4: Parâmetros do Algoritmo Genético

| Parâmetros | Quantidade |
|---|---|
| Tamanho da população | N |
| Nº de gerações | 35 |
| Nº de pais selecionados | N / 2 |
| Nº de filhos | 4 |
| Probabilidade de mutação | 1 / N |

Fonte: Elaborado pelo autor, 2021.

A Figura 3.4 contém a descrição do comportamento do AG na seleção de características. A partir da mesma, pode-se observar a convergência do AG no que diz respeito a

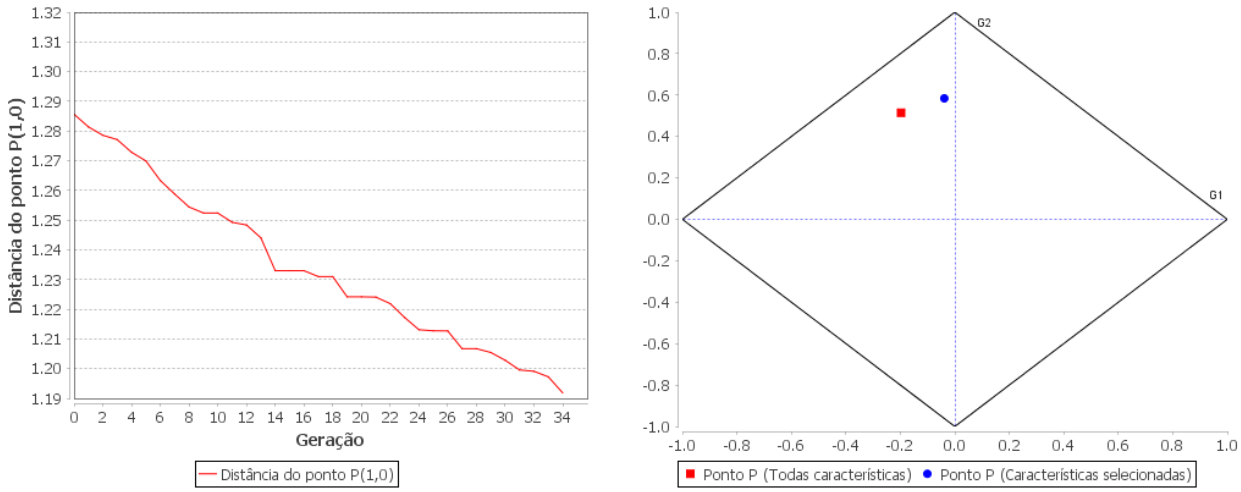
Figura 3.3: Etapas do Algoritmo Genético




Fonte: Elaborado pelo autor, 2021.

minimização da distância para $D_{1,0}$ (a), seguida pela visualização comparativa dos pontos $P(G_1, G_2)$ entre todas as características extraídas e as características selecionadas (b).

Figura 3.4: Comportamento do AG na seleção de características




Fonte: Elaborado pelo autor, 2021.

## 3.4 *Multilayer Perceptron*

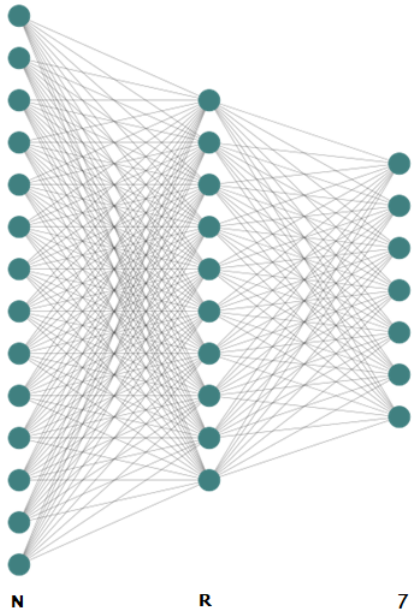
Considerando-se a não-linearidade dos vetores de características, conforme demonstrado na Figura 3.4, o presente trabalho adota o uso do mencionado classificador a partir da sua comprovada eficácia na resolução de *Speech Emotion Recognition*, como demonstrado em (IDRIS; SALAM; SUNAR, 2015) e (GUDMALWAR; RAO; DUTTA, 2018).

*Multilayer Perceptron* ou MLP é uma rede neural *feed-forward* com uma camada de entrada, múltiplas camadas ocultas e uma camada de saída (GUDMALWAR; RAO; DUTTA, 2018). A Figura 3.5 contém a ilustração da arquitetura da MLP utilizada nesta dissertação, sendo $N$ a dimensão do vetor de características apresentado ao classificador, $R$ um número arbitrário de neurônios e $7$ o número de emoções a serem classificadas.

Observada a definição da sua estrutura, o treinamento da MLP procede em duas fases (HAYKIN, 2009):

1. Fase *forward*: o sinal se propaga sempre adiante: da camada de entrada para a camada de saída através das camadas ocultas. A fim de se introduzir o aspecto

Figura 3.5: Arquitetura da MLP utilizada




Fonte: Elaborado pelo autor, 2021.

de não-linearidade à rede a função de ativação sigmóide foi utilizada nas camadas ocultas e de saída, a qual é definida por:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.1}$$

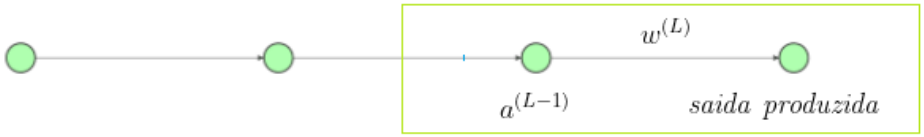
2. Fase *backward*: um sinal de erro é produzido comparando a saída produzida pela rede com uma resposta desejada. O sinal de erro resultante é propagado pela rede, camada por camada, mas desta vez a propagação é executada na direção inversa. Nesta fase, ajustes sucessivos são feitos nos pesos sinápticos da rede. O erro ou custo acima descrito é calculado utilizando a seguinte equação:

$$E_{total} = \sum \frac{1}{2}(saída\ desejada - saída\ produzida)^2 \tag{3.2}$$

Mais especificamente, a fase *backward* é realizada segundo o algoritmo de aprendizagem supervisionada denominado *backpropagation*, o qual tem como objetivo minimizar a função de custo acima descrita a partir do ajuste dos pesos sinápticos. De maneira prática,

esse algoritmo é aplicado conforme a conhecida regra da cadeia ou *chain rule*, a qual é responsável por definir o cálculo da derivada de uma função composta. Assim, tomando como base a Figura 3.6.

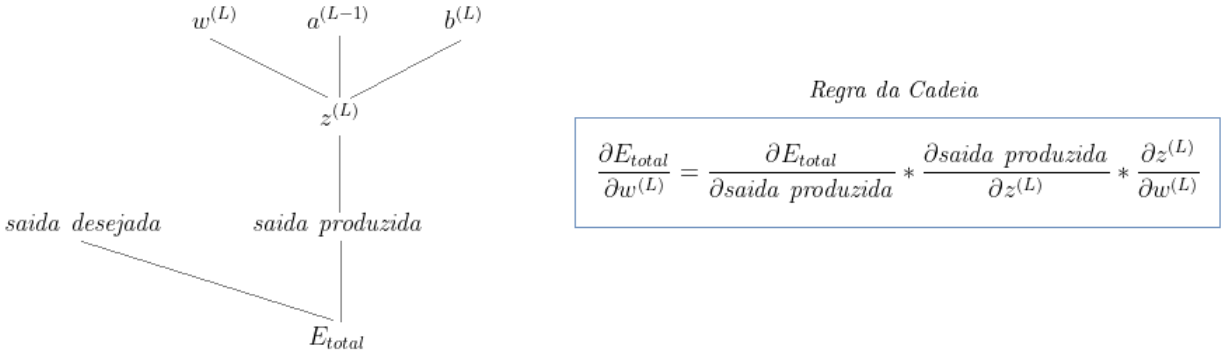
Figura 3.6: Cálculo da *saída produzida* considerando apenas as duas últimas camadas




Fonte: Elaborado pelo autor, 2021.

O objetivo do algoritmo *backpropagation* reside em entender a sensibilidade de $E_{total}$ a uma alteração do valor do peso sinaptico $w^{(L)}$, de acordo com a derivada $\frac{\partial E_{total}}{\partial w^{(L)}}$. A Figura 3.7 contém a ilustração da regra da cadeia aplicada no cálculo do *backpropagation*.

Figura 3.7: Regra da Cadeia




Fonte: Elaborado pelo autor, 2021.

Finalmente, o cálculo das derivadas parciais ilustradas na Figura 3.7 se dá conforme as equações 3.3, 3.4 e 3.5.

$$\frac{\partial E_{total}}{\partial saida\ produzida} = saida\ produzida - saida\ desejada \tag{3.3}$$

$$\frac{\partial saida\ produzida}{\partial z^{(L)}} = \sigma'(z^{(L)}) \tag{3.4}$$

$$\frac{\partial z^{(L)}}{\partial w^{(L)}} = a^{(L-1)} \tag{3.5}$$

### 3.4.1 *Early Stopping*

Um dos principais aspectos a serem considerados durante o treinamento e performance de uma rede neural reside na sua capacidade de generalização, ou seja, sua aptidão em classificar corretamente dados nunca vistos anteriormente. Tal característica está extremamente conectada à escolha do número de épocas de treinamento a serem utilizadas, a qual pode ocasionar os dois seguintes problemas caso uma escolha inapropriada seja realizada:

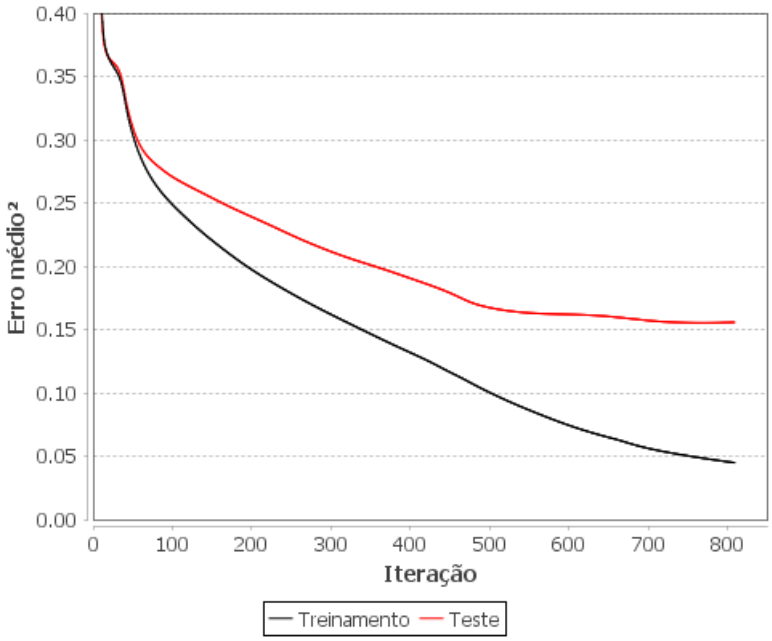
- *underfitting*: o treinamento realizado pela rede foi insuficiente, resultando em baixa performance tanto para o conjunto de treinamento quanto para o conjunto de teste;
- *overfitting*: o treinamento realizado perdeu a capacidade de generalização pois os ruídos do conjunto de treinamento foram aprendidos pela rede, resultando em alta performance no conjunto de treinamento e baixa performance no conjunto de teste.

Assim, pode-se observar no comportamento de uma rede neural como a utilizada no presente trabalho, que o erro, conforme a equação 3.2, obtido para o conjunto de treinamento tende sempre a ser minimizado, de forma que aprendizado seja extraído deste conjunto, entretanto, faz-se importante notar que este comportamento torna-se distinto quando dados do conjunto de teste são apresentados: inicialmente, o erro é minimizado, mas após um determinado número de épocas o processo inverso começa a ser observado.

Uma das ferramentas mais simples e comumente utilizadas a fim de solucionar o problema relacionado à quantidade de treinamento da rede neural é o *early stopping*. Essa abordagem consiste no seguinte processo: a cada época o erro é calculado não somente para o conjunto de treinamento, mas também para o conjunto de teste, de forma que o treinamento da rede é interrompido a partir do momento em que o erro nos dados do conjunto de teste estagna ou começa a aumentar. A Figura 3.8, a qual foi obtida a partir dos experimentos realizados neste trabalho, contém um exemplo do processo descrito neste

parágrafo. Na mesma, pode-se observar que por volta da época 800 o erro no conjunto de teste começa a crescer, resultando na finalização do treinamento da rede.

Figura 3.8: *Early Stopping*




Fonte: Elaborado pelo autor, 2021.

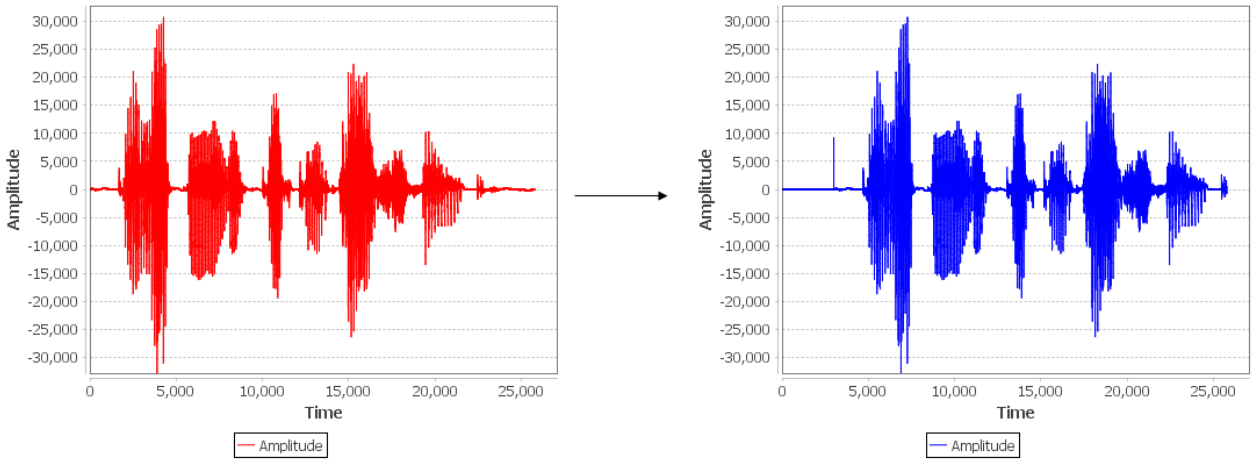
### 3.4.2 *Data Augmentation*

A capacidade de modelagem e generalização de um algoritmo para a tarefa de reconhecimento de padrões depende não somente da quantidade de treinamento a ser definida, como também da quantidade de sinais disponíveis para o treinamento. Como descrito na Seção 3.1, o EMO-DB é constituido por 535 sinais divididos em 7 emoções, uma quantidade considerada baixa se comparada com os bancos de dados utilizados no âmbito de processamento de sinais.

Assim sendo, faz-se comum dentre os estudos encontrados na área que a prática conhecida como *Data Augmentation* seja implementada a fim de que o número de vetores de características disponíveis para a modelagem da rede seja acrescido, como pode se ver em (PRASEETHA; JOBY, 2021) e (ISSA; DEMIRCI; YAZICI, 2020).

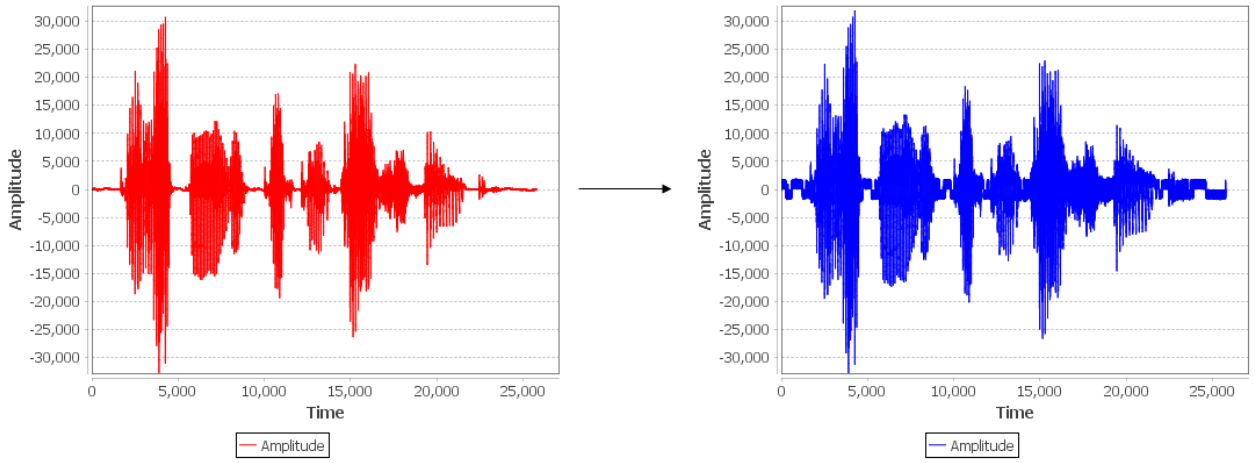
Nesta dissertação o aumento do número de amostras disponíveis no EMO-DB foi realizado com base nos estudos citados, dessa forma, as técnicas aqui utilizadas incluem as seguintes alterações no sinal de áudio adquirido: mover o início e fim dos sinais em pequenos montantes, assim como adicionar uma breve quantidade de ruído gaussiano durante seu comprimento. As Figuras 3.9 e 3.10 adquiridas a partir dos experimentos realizados pelo autor demonstram o referido processo.

Figura 3.9: *Data augmentation*: movimentação do áudio




Fonte: Elaborado pelo autor, 2021.

Figura 3.10: *Data augmentation*: acréscimo de ruído gaussiano




Fonte: Elaborado pelo autor, 2021.

## 3.5 Parâmetros de Avaliação

Neste trabalho, a tarefa de *speaker-independent* SER é performada no reconhecimento de sete emoções distintas: raiva, tédio, desgosto, medo, felicidade, neutralidade e tristeza. A fim de permitir uma análise comparativa com outros estudos três métodos de avaliação foram adotados: matriz de confusão, acurácia individual de classe e acurácia geral.

A matriz de confusão é uma ferramenta para visualização de desempenho comumente utilizada em algoritmos de aprendizagem supervisionada. Tomando como base o conjunto de teste, este método apresenta o número de classes corretas e incorretas previstas pelo modelo. Assim, a matriz de confusão é $NxN$, sendo $N$ o número de classes, com as linhas representando as instâncias em um classe real, enquanto as colunas representam as instâncias em uma classe prevista. A Tabela 3.5 contém um exemplo de uma matriz de confusão para um modelo de classificação baseado em duas classes: “Positivo” e “Negativo”.

Tabela 3.5: Exemplo de matriz de confusão

| | | **Classes Previstas** | |
|---|---|---|---|
| | | Positivo | Negativo |
| **Classes** | Positivo | $a$ | $b$ |
| **Reais** | Negativo | $c$ | $d$ |

Fonte: Elaborado pelo autor, 2021.

Para o exemplo abordado, os seguintes significados podem ser aferidos:

- $a$ é o número de previsões corretas de que uma instância é positiva;
- $b$ é o número de previsões incorretas de que uma instância é negativa;
- $c$ é o número de previsões incorretas de que uma instância é positiva;
- $d$ é o número de previsões corretas de que uma instância é negativa.

A acurácia de classificação de um modelo é definida como a porcentagem de instâncias classificadas corretamente, tomando como base todas as instâncias existentes no

problema. A equação (3.6) contém o cálculo de acurácia aqui descrito.

$$acurácia = \frac{nº\ de\ instâncias\ previstas\ corretamente}{nº\ total\ de\ instâncias} \quad . \tag{3.6}$$

Visto que a base de dados EMO-DB apresenta desbalanceamento entre a quantidade de instâncias para cada classe (ver Tabela 3.1), duas formas distintas de acurácia foram utilizadas a fim de se obter uma análise mais detalhada acerca do desempenho do modelo: acurácia individual de classe e acurácia geral.

Baseando-se nos dados contidos na matriz de confusão exemplificada pela Tabela 3.5, temos que:

$$acurácia\ individual\ \text{“}Positivo\text{”} = \frac{a}{a+b} \tag{3.7}$$

$$acurácia\ individual\ \text{“}Negativo\text{”} = \frac{d}{c+d} \tag{3.8}$$

$$acurácia\ geral = \frac{a+d}{a+b+c+d} \quad . \tag{3.9}$$

# Capítulo 4

# Testes e Resultados

## 4.1 Procedimentos de Teste e Resultados

Os testes e resultados obtidos durante os experimentos realizados para a abordagem descrita no capítulo anterior encontram-se expostas nesta seção. Realizada as etapas de extração e seleção de características, os 535 vetores de características obtidos são separados aleatóriamente para a modelagem da *Multi Layer Perceptron* segundo os seguintes montantes:

- 10% reservado para treinamento e 90% para teste;
- 20% reservado para treinamento e 80% para teste;
- 30% reservado para treinamento e 70% para teste;
- 40% reservado para treinamento e 60% para teste;
- 50% reservado para treinamento e 50% para teste;

Assim sendo, 10 experimentos foram realizados para cada caso acima listado, visando a obtenção de resultados de acurácia consistentes e estatísticos, tais como: valor mínimo e máximo, média e desvio padrão. A Tabela 4.1 contém um resumo dos experimentos realizados.

Tabela 4.1: Resultados do procedimento de classificação das quatro emoções a partir do treinamento da MLP

| Montante de treinamento | Acurácia mínima | Acurácia máxima | Média das acurácias | Desvio padrão das acurácias |
|---|---|---|---|---|
| 10% | 54.9% | 64.5% | 58.8% | 3.82 |
| 20% | 74.3% | 80.3% | 77.5% | 1.95 |
| 30% | 74.9% | 83.3% | 79.2% | 3.28 |
| 40% | 77.5% | 84.7% | 80.7% | 2.29 |
| 50% | 81.8% | 84.9% | 83.4% | 0.99 |

Fonte: Elaborado pelo autor, 2021.

Adicionalmente às informações gerais de experimentos contidos na tabela anterior, as tabelas 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.10 e 4.11 contém informações mais detalhadas acerca dos resultados obtidos a partir da apresentação das melhores e piores matrizes de confusão para cada caso.

### 4.1.1 Classificação MLP após a seleção de características utilizando montante de treinamento igual a 10%

Tabela 4.2: Melhor matriz de confusão obtida a partir da classificação realizada pela MLP com montante de treinamento igual a 10%

| *Emoção (%)* | **Raiva** | **Tédio** | **Desgosto** | **Medo** | **Felicidade** | **Neutralidade** | **Tristeza** |
|---|---|---|---|---|---|---|---|
| **Raiva** | **80.4** | 1.3 | 0.0 | 13.2 | 5.1 | 0.0 | 0.0 |
| **Tédio** | 1.7 | **75.3** | 0.0 | 0.7 | 0.8 | 12.6 | 8.9 |
| **Desgosto** | 1.0 | 14.8 | **40.4** | 20.7 | 8.9 | 5.7 | 8.5 |
| **Medo** | 3.1 | 2.9 | 7.0 | **51.4** | 7.8 | 1.9 | 25.9 |
| **Felicidade** | 11.1 | 28.2 | 8.7 | 21.0 | **21.2** | 7.8 | 2.0 |
| **Neutralidade** | 2.1 | 31.2 | 0.1 | 0.8 | 6.9 | **54.0** | 4.9 |
| **Tristeza** | 0.0 | 2.1 | 0.9 | 0.0 | 0.0 | 8.9 | **88.1** |
| **Média** | **64.5** | | | | | | |

Fonte: Elaborado pelo autor, 2021.

Tabela 4.3: Pior matriz de confusão obtida a partir da classificação realizada pela MLP com montante de treinamento igual a 10%

| *Emoção (%)* | **Raiva** | **Tédio** | **Desgosto** | **Medo** | **Felicidade** | **Neutralidade** | **Tristeza** |
|---|---|---|---|---|---|---|---|
| **Raiva** | **96.1** | 0.1 | 0.0 | 1.9 | 1.9 | 0.0 | 0.0 |
| **Tédio** | 3.1 | **56.2** | 0.0 | 0.0 | 0.8 | 36.8 | 3.1 |
| **Desgosto** | 32.2 | 0.0 | **0.0** | 24.1 | 11.7 | 25.9 | 6.1 |
| **Medo** | 55.1 | 0.0 | 0.0 | **32.2** | 6.8 | 3.7 | 2.2 |
| **Felicidade** | 57.2 | 0.0 | 0.0 | 12.3 | **5.1** | 23.7 | 1.7 |
| **Neutralidade** | 6.2 | 27.1 | 0.0 | 0.0 | 2.1 | **60.1** | 4.5 |
| **Tristeza** | 0.0 | 8.1 | 0.0 | 4.8 | 0.0 | 8.2 | **78.9** |
| **Média** | **54.9** | | | | | | |

Fonte: Elaborado pelo autor, 2021.

### 4.1.2 Classificação MLP após a seleção de características utilizando montante de treinamento igual a 20%

Tabela 4.4: Melhor matriz de confusão obtida a partir da classificação realizada pela MLP com montante de treinamento igual a 20%

| *Emoção (%)* | **Raiva** | **Tédio** | **Desgosto** | **Medo** | **Felicidade** | **Neutralidade** | **Tristeza** |
|---|---|---|---|---|---|---|---|
| **Raiva** | **89.1** | 3.8 | 0.0 | 1.1 | 6.0 | 0.0 | 0.0 |
| **Tédio** | 0.0 | **76.2** | 0.0 | 2.1 | 3.8 | 14.0 | 3.9 |
| **Desgosto** | 5.1 | 1.2 | **61.1** | 7.7 | 10.0 | 13.2 | 1.7 |
| **Medo** | 10.2 | 0.2 | 4.1 | **80.1** | 4.8 | 0.0 | 0.6 |
| **Felicidade** | 21.1 | 1.9 | 5.2 | 3.8 | **63.0** | 4.2 | 0.8 |
| **Neutralidade** | 0.0 | 9.1 | 1.1 | 0.9 | 2.1 | **83.2** | 4.3 |
| **Tristeza** | 0.1 | 2.0 | 0.0 | 1.2 | 0.0 | 1.0 | **95.7** |
| **Média** | **80.3** | | | | | | |

Fonte: Elaborado pelo autor, 2021.

Tabela 4.5: Pior matriz de confusão obtida a partir da classificação realizada pela MLP com montante de treinamento igual a 20%

| *Emoção (%)* | **Raiva** | **Tédio** | **Desgosto** | **Medo** | **Felicidade** | **Neutralidade** | **Tristeza** |
|---|---|---|---|---|---|---|---|
| **Raiva** | **96.1** | 0.0 | 0.0 | 1.8 | 1.1 | 1.0 | 0.0 |
| **Tédio** | 1.2 | **69.1** | 0.0 | 0.7 | 2.7 | 26.2 | 0.1 |
| **Desgosto** | 21.2 | 4.1 | **50.3** | 14.7 | 3.8 | 5.9 | 0.0 |
| **Medo** | 23.8 | 0.0 | 2.1 | **71.1** | 2.5 | 0.5 | 0.0 |
| **Felicidade** | 29.2 | 1.0 | 13.1 | 12.8 | **41.2** | 2.7 | 0.0 |
| **Neutralidade** | 0.1 | 7.8 | 0.0 | 1.0 | 8.1 | **82.2** | 0.8 |
| **Tristeza** | 0.0 | 5.1 | 0.8 | 2.2 | 1.1 | 7.2 | **83.6** |
| **Média** | **74.3** | | | | | | |

Fonte: Elaborado pelo autor, 2021.

### 4.1.3 Classificação MLP após a seleção de características utilizando montante de treinamento igual a 30%

Tabela 4.6: Melhor matriz de confusão obtida a partir da classificação realizada pela MLP com montante de treinamento igual a 30%

| *Emoção (%)* | **Raiva** | **Tédio** | **Desgosto** | **Medo** | **Felicidade** | **Neutralidade** | **Tristeza** |
|---|---|---|---|---|---|---|---|
| **Raiva** | **98.1** | 0.0 | 0.0 | 0.0 | 1.8 | 0.1 | 0.0 |
| **Tédio** | 0.0 | **86.2** | 0.0 | 1.8 | 0.9 | 9.1 | 2.0 |
| **Desgosto** | 10.2 | 1.8 | **68.1** | 7.9 | 3.8 | 2.9 | 5.3 |
| **Medo** | 8.1 | 1.8 | 7.9 | **75.2** | 1.7 | 1.0 | 4.3 |
| **Felicidade** | 23.2 | 0.0 | 0.8 | 3.9 | **63.2** | 8.8 | 0.1 |
| **Neutralidade** | 0.0 | 9.1 | 0.8 | 1.9 | 1.1 | **84.8** | 2.3 |
| **Tristeza** | 0.0 | 6.2 | 0.8 | 1.1 | 0.0 | 1.9 | **90.0** |
| **Média** | **83.3** | | | | | | |

Fonte: Elaborado pelo autor, 2021.

Tabela 4.7: Pior matriz de confusão obtida a partir da classificação realizada pela MLP com montante de treinamento igual a 30%

| *Emoção (%)* | **Raiva** | **Tédio** | **Desgosto** | **Medo** | **Felicidade** | **Neutralidade** | **Tristeza** |
|---|---|---|---|---|---|---|---|
| **Raiva** | **79.2** | 0.0 | 1.8 | 17.7 | 0.0 | 1.3 | 0.0 |
| **Tédio** | 0.0 | **68.0** | 0.0 | 5.2 | 0.3 | 22.8 | 3.7 |
| **Desgosto** | 2.2 | 1.2 | **56.3** | 17.8 | 4.1 | 8.8 | 9.6 |
| **Medo** | 8.2 | 0.0 | 1.7 | **83.9** | 1.1 | 2.7 | 2.4 |
| **Felicidade** | 18.0 | 1.1 | 0.8 | 14.2 | **52.1** | 13.8 | 0.0 |
| **Neutralidade** | 0.0 | 5.2 | 1.1 | 1.2 | 0.7 | **91.0** | 0.8 |
| **Tristeza** | 0.0 | 2.2 | 0.8 | 1.0 | 0.0 | 8.1 | **87.9** |
| **Média** | **74.9** | | | | | | |

Fonte: Elaborado pelo autor, 2021.

### 4.1.4 Classificação MLP após a seleção de características utilizando montante de treinamento igual a 40%

Tabela 4.8: Melhor matriz de confusão obtida a partir da classificação realizada pela MLP com montante de treinamento igual a 40%

| *Emoção (%)* | **Raiva** | **Tédio** | **Desgosto** | **Medo** | **Felicidade** | **Neutralidade** | **Tristeza** |
|---|---|---|---|---|---|---|---|
| **Raiva** | **90.0** | 0.0 | 1.3 | 2.8 | 5.9 | 0.0 | 0.0 |
| **Tédio** | 0.0 | **88.1** | 2.2 | 0.0 | 0.8 | 6.9 | 2.0 |
| **Desgosto** | 4.2 | 0.9 | **77.3** | 6.1 | 8.7 | 2.8 | 0.0 |
| **Medo** | 7.2 | 1.0 | 1.1 | **85.2** | 5.5 | 0.0 | 0.0 |
| **Felicidade** | 8.1 | 1.8 | 2.0 | 7.9 | **75.1** | 5.1 | 0.0 |
| **Neutralidade** | 0.0 | 12.2 | 2.1 | 0.8 | 0.0 | **84.9** | 0.0 |
| **Tristeza** | 0.0 | 2.1 | 1.8 | 1.0 | 0.9 | 4.8 | **89.4** |
| **Média** | **84.7** | | | | | | |

Fonte: Elaborado pelo autor, 2021.

Tabela 4.9: Pior matriz de confusão obtida a partir da classificação realizada pela MLP com montante de treinamento igual a 40%

| *Emoção (%)* | **Raiva** | **Tédio** | **Desgosto** | **Medo** | **Felicidade** | **Neutralidade** | **Tristeza** |
|---|---|---|---|---|---|---|---|
| **Raiva** | **94.7** | 0.0 | 0.0 | 1.0 | 2.9 | 1.0 | 0.4 |
| **Tédio** | 1.0 | **72.2** | 1.1 | 0.8 | 0.9 | 22.1 | 1.9 |
| **Desgosto** | 10.1 | 0.0 | **39.8** | 14.1 | 4.9 | 31.0 | 0.1 |
| **Medo** | 17.0 | 1.1 | 2.1 | **71.8** | 3.1 | 4.9 | 0.0 |
| **Felicidade** | 36.2 | 0.0 | 0.0 | 4.8 | **47.0** | 11.1 | 0.9 |
| **Neutralidade** | 1.1 | 1.8 | 0.0 | 0.0 | 1.0 | **94.9** | 1.2 |
| **Tristeza** | 0.0 | 1.0 | 0.0 | 1.2 | 0.0 | 7.9 | **89.9** |
| **Média** | **77.5** | | | | | | |

Fonte: Elaborado pelo autor, 2021.

### 4.1.5 Classificação MLP após a seleção de características utilizando montante de treinamento igual a 50%

Tabela 4.10: Melhor matriz de confusão obtida a partir da classificação realizada pela MLP com montante de treinamento igual a 50%

| *Emoção (%)* | **Raiva** | **Tédio** | **Desgosto** | **Medo** | **Felicidade** | **Neutralidade** | **Tristeza** |
|---|---|---|---|---|---|---|---|
| **Raiva** | **93.1** | 0.0 | 0.8 | 2.2 | 3.9 | 0.0 | 0.0 |
| **Tédio** | 1.1 | **81.2** | 1.9 | 0.0 | 2.8 | 12.1 | 0.9 |
| **Desgosto** | 3.2 | 0.0 | **83.1** | 8.8 | 3.9 | 0.0 | 1.0 |
| **Medo** | 4.2 | 0.0 | 6.8 | **86.1** | 2.0 | 0.9 | 0.0 |
| **Felicidade** | 9.1 | 3.0 | 4.8 | 7.1 | **72.2** | 2.8 | 1.0 |
| **Neutralidade** | 0.0 | 9.1 | 2.8 | 0.0 | 0.0 | **85.0** | 3.1 |
| **Tristeza** | 0.0 | 3.9 | 0.0 | 3.1 | 0.0 | 2.8 | **90.2** |
| **Média** | **84.9** | | | | | | |

Fonte: Elaborado pelo autor, 2021.

Tabela 4.11: Pior matriz de confusão obtida a partir da classificação realizada pela MLP com montante de treinamento igual a 50%

| *Emoção (%)* | **Raiva** | **Tédio** | **Desgosto** | **Medo** | **Felicidade** | **Neutralidade** | **Tristeza** |
|---|---|---|---|---|---|---|---|
| **Raiva** | **93.1** | 0.0 | 0.8 | 4.2 | 1.1 | 0.8 | 0.0 |
| **Tédio** | 0.0 | **79.1** | 4.2 | 0.9 | 1.8 | 7.9 | 6.1 |
| **Desgosto** | 5.1 | 2.2 | **73.8** | 7.9 | 2.2 | 3.0 | 5.8 |
| **Medo** | 9.1 | 5.9 | 1.0 | **80.2** | 3.1 | 0.7 | 0.0 |
| **Felicidade** | 13.4 | 9.1 | 9.8 | 2.2 | **61.4** | 3.0 | 1.1 |
| **Neutralidade** | 1.2 | 10.7 | 3.2 | 0.0 | 1.0 | **79.6** | 4.3 |
| **Tristeza** | 0.0 | 2.1 | 0.8 | 0.0 | 0.0 | 3.9 | **93.2** |
| **Média** | **81.8** | | | | | | |

Fonte: Elaborado pelo autor, 2021.

### 4.1.6 Síntese

A análise dos resultados obtidos a partir dos experimentos acima descritos permite observar que a melhor acurácia alcançada foi de **84.9%** quando 50% da base de dados foi utilizada para o treinamento do classificador. Nesta ocasião, a respectiva matriz de confusão reflete uma taxa de acurácia individual acima de 80% para as emoções contempladas, excetuando-se a emoção **felicidade**, a qual obteve uma acurácia individual de 72.2%.

Em suma, a partir dos experimentos sumarizados na tabela 4.1 pode-se notar que média das acurácias evolui e o desvio padrão decresce à medida que o montante de treinamento é acrescido, indicando uma maior linearidade no comportamento do algoritmo a partir da disponibilidade de uma maior quantidade de vetores para a modelagem da rede neural.

As matrizes de confusão referentes aos diferentes montantes de treinamento permitem uma análise sobre o comportamento de classificação individual de uma emoção em correlação às outras. Assim sendo, pode-se aferir que a emoção **felicidade** possui a menor acurácia individual entre as emoções observadas, enquanto **raiva** e **tristeza** são as classes com maior taxa de reconhecimento. Além disto, a partir de uma análise mais detalhada dos

testes realizados, específicamente das tabelas 4.2 e 4.3, pode-se observar que há uma forte correlação entre **felicidade** x **raiva**, assim como nas emoções **neutralidade** x **tédio**.

# Capítulo 5

# Conclusões e Trabalhos Futuros

Este trabalho procurou explorar a utilização de características prosódicas e espectrais a fim de contemplar a tarefa de reconhecimento de sete emoções a partir do sinal de voz: raiva, tédio, desgosto, medo, felicidade, neutralidade e tristeza. A partir dos resultados obtidos no Capítulo anterior, pode-se aferir a adequabilidade dos descritores extraídos ao problema em questão, visto que de forma geral, aspectos relacionados ao sinal em si, e não à características linguísticas, foram tomados como base para a formação e classificação dos vetores de características.

Pôde-se notar durante a etapa de seleção de características a relevância dos descritores extraídos a partir do domínio de Fourier, visto que grande parte dos atributos escolhidos nesta fase proveram da entropia e planaridade espectral, enfatizando-se a importância da análise do sinal de voz a partir de suas frequências. Desta forma, faz-se importante citar que a escolha da abordagem *Handcrafted Extraction* permitiu uma investigação mais precisa acerca da adequabilidade das características extraídas ao problema de classificação das sete emoções.

Ainda ao longo da seleção dos descritores, pôde-se verificar que a Engenharia Paraconsistente de Características possibilita a utilização de classificadores com uma arquitetura mais simples, além de um tempo menor de treinamento no caso específico da *Multi Layer Perceptron* (MLP) devido à quantidade reduzida de características. Complementarmente, a implementação do algoritmo genético descrito na Seção 3.3 permite que a análise paraconsistente seja executada em um tempo computacionalmente viável quando o número

de características a serem considerados é relativamente grande, como é o caso deste trabalho.

Quanto a continuidade deste trabalho, pode-se dizer que um dos aspectos a serem analisados reside no tempo de computação requerido pelas etapas de seleção de características e classificação de vetores, de forma que algum aspecto de paralelismo possa ser implementado durante os processos do algoritmo genético e da MLP. Além disso, há de se considerar práticas de extração de caraterísticas baseadas na Transformada *Wavelet* (GUIDO, 2015), como uma alternativa à construção de vetores mais adequados ao problema de classificação, ou seja, mais próximas ao vértice (1, 0) quando comparados às características extraídas no presente trabalho (Figura 3.4).

A aplicação da metodologia descrita a outras bases de dados em idiomas distintos e diferentes classificadores pode ser adotada como forma de validação dos processos de extração e seleção de características realizados, visto que os mesmos foram formulados independentemente de aspectos de linguagem ou método de classificação. No artigo (AKçAY; OGUZ, 2020) pode-se encontrar referências a respeito das bases de dados e classificadores mais comumente utilizados na tarefa de *Speech Emotion Recognition*.

# REFERÊNCIAS


AHUJA, J.; RATNOO, S. Feature selection using multi-objective genetic algorith m: A hybrid approach. *INFOCOMP Journal of Computer Science*, v. 14, p. 26–37, 06 2015.

AKçAY, B.; OGUZ, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, v. 116, 01 2020.

ANTONIOU, A. *Digital Signal Processing: Signals, Systems, and Filters*. McGraw-Hill Education, 2006. ISBN 9780071454247. Disponível em: <https://books.google.com.br/books?id=JQ4fAQAAIAAJ>.

BEZOOIJEN, R. V.; EBRARY. *Characteristics and Recognizability of Vocal Expressions of Emotion*. [S.l.]: Walter de Gruyter GmbH & Company KG, 1984. (Netherlands phonetic archives). ISBN 9783110132779.

BURKHARDT, F. et al. A database of german emotional speech. In: . [S.l.: s.n.], 2005. v. 5, p. 1517–1520.

COOLEY, J. W.; TUKEY, J. W. An Algorithm for the Machine Calculation of Complex Fourier Series. *Math. Comput.*, v. 19, p. 297–301, 1965.

DEB, S.; DANDAPAT, S. Emotion classification using segmentation of vowel-like and non-vowel-like regions. *IEEE Transactions on Affective Computing*, IEEE Computer Society, Los Alamitos, CA, USA, v. 10, n. 03, p. 360–373, jul 2019. ISSN 1949-3045.

DIMITROVA-GREKOW, T.; KLIS, A.; IGRAS-CYBULSKA, M. Speech emotion recognition based on voice fundamental frequency. v. 44, p. 277–286, 05 2019.

EYBEN, F. et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, IEEE, v. 7, n. 2, p. 190–202, 4 2016. ISSN 1949-3045. Open access.

GARCÍA-DOMÍNGUEZ, A. et al. Feature selection using genetic algorithms for the generation of a recognition and classification of children activities model using environmental sound. *Mob. Inf. Syst.*, v. 2020, p. 8617430:1–8617430:12, 2020. Disponível em: <https://doi.org/10.1155/2020/8617430>.

GRAY, A.; MARKEL, J. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 22, n. 3, p. 207–217, 1974.

GUDMALWAR, A.; RAO, C. R.; DUTTA, A. Improving the performance of the speaker emotion recognition based on low dimension prosody features vector. *International Journal of Speech Technology*, 12 2018.

GUIDO, R. C. Practical and useful tips on discrete wavelet transforms [sp tips amp; tricks]. *IEEE Signal Processing Magazine*, v. 32, n. 3, p. 162–166, 2015.

GUIDO, R. C. A tutorial on signal energy and its applications. *Neurocomput.*, Elsevier Science Publishers B. V., NLD, v. 179, n. C, p. 264–282, fev. 2016. ISSN 0925-2312. Disponível em: <https://doi.org/10.1016/j.neucom.2015.12.012>.

GUIDO, R. C. Zcr-aided neurocomputing: A study with applications. *Knowledge-Based Systems*, v. 105, p. 248–269, 2016. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705116301009>.

GUIDO, R. C. Paraconsistent feature engineering [lecture notes]. *IEEE Signal Processing Magazine*, v. 36, n. 1, p. 154–158, 2019.

HAYKIN, S. S. *Neural networks and learning machines*. Third. Upper Saddle River, NJ: Pearson Education, 2009.

IDRIS, I.; SALAM, M. S.; SUNAR, M. S. Speech emotion classification using svm and mlp on prosodic and voice quality features. *Jurnal Teknologi*, v. 78, 12 2015.

ISSA, D.; DEMIRCI, M. F.; YAZICI, A. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, v. 59, p. 101894, 05 2020.

JACOB, A. Modelling speech emotion recognition using logistic regression and decision trees. *Int. J. Speech Technol.*, Springer-Verlag, Berlin, Heidelberg, v. 20, n. 4, p. 897–905, dez. 2017. ISSN 1381-2416. Disponível em: <https://doi.org/10.1007/s10772-017-9457-6>.

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.

KAMIńSKA, D.; SAPIńSKI, T. Polish emotional speech recognition based on the committee of classifiers. *Przeglad Elektrotechniczny*, v. 2016, p. 101–106, 06 2017.

KERKENI, L. et al. Speech emotion recognition: Methods and cases study. In: . [S.l.: s.n.], 2018. p. 175–182.

PRASEETHA, V.; JOBY, P. Speech emotion recognition using data augmentation. *International Journal of Speech Technology*, 08 2021.

RABINER, L.; SCHAFER, R. *Digital Processing of Speech Signals*. [S.l.]: Englewood Cliffs: Prentice Hall, 1978.

SHANNON, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.*, v. 27, n. 3, p. 379–423, 1948. Disponível em: <http://dblp.uni-trier.de/db/journals/bstj/bstj27.html#Shannon48>.

TAMULEVIčIUS, G.; KARBAUSKAITė, R.; DZEMYDA, G. Speech emotion classification using fractal dimension-based features. *Nonlinear Analysis: Modelling and Control*, v. 24, n. 5, p. 679–695, Sep. 2019. Disponível em: <https://www.journals.vu.lt/nonlinear-analysis/article/view/14505>.

TOLKMITT, F.; SCHERER, K. *Effect of experimentally induced stress on vocal parameters*. [S.l.], 1986. 302–313 p.

TRIGEORGIS, G. et al. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: . [S.l.: s.n.], 2016. p. 5200–5204.

TURSUNOV, A.; KWON, S.; PANG, H.-S. Discriminating emotions in the valence dimension from speech using timbre features. *Applied Sciences*, v. 9, p. 2470, 06 2019.

VIGNOLO, L. et al. Feature optimisation for stress recognition in speech. *Pattern Recogn. Lett.*, Elsevier Science Inc., USA, v. 84, n. C, p. 1–7, dez. 2016. ISSN 0167-8655. Disponível em: <https://doi.org/10.1016/j.patrec.2016.07.017>.

WANG, K. et al. Speech emotion recognition using fourier parameters. *Affective Computing, IEEE Transactions on*, v. 6, p. 69–75, 01 2015.

ZVAREVASHE, K.; OLUGBARA, O. Ensemble learning of hybrid acoustic features for speech emotion recognition. *Algorithms*, v. 13, p. 70, 03 2020.

ÖZSEVEN, T. A novel feature selection method for speech emotion recognition. *Applied Acoustics*, v. 146, p. 320–326, 03 2019.

# Apêndice A

# Recursos na web

Recomenda-se, durante a leitura desta dissertação, a consulta dos códigos fontes correspondentes aos procedimentos e experimentos descritos a fim de que um melhor entendimento dos assuntos abordados possa ser obtido. Todos os códigos foram implementados utilizando a linguagem de programação JAVA a partir da IDE *NetBeans*, de forma que a seguinte organização de pacotes foi utilizada:

- **acquisition**: contém a classe responsável por efetuar a leitura dos sinais WAV;
- **analysis**: contém as classes responsáveis pelo cálculo matemático da Engenharia Paraconsistente e pela normalização dos vetores de características;
- **charts**: contém a classe responsável pela criação dos gráficos utilizados;
- **dependecies**: contém uma classe responsável pelo tratamento de números complexos utilizados na Transformada Discreta de Fourier;
- **feature extraction**: contém as classes responsáveis pela extração e composição dos vetores de características;
- **feature selection**: contém as classes responsáveis pela seleção de características utilizando o algoritmo genético;
- **view**: contém as classes responsáveis pela separação da base de dados entre treinamento e teste, assim como a implementação do classificador utilizado.

Estes recursos encontram-se disponíveis em <https://github.com/hiagomb/mestrado> para livre uso não comercial.

A fim de um melhor entedimento acerca do algoritmo de retropropagação indica-se os vídeos disponibilizados em <https://www.3blue1brown.com/>, assim como o tutorial encontrado em <https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/>.