UNIVERSIDADE ESTADUAL PAULISTA

“JÚLIO DE MESQUITA FILHO”

CAMPUS DE SÃO JOSÉ DO RIO PRETO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

HIAGO MATHEUS BRAJATO

MONOGRAFIA DE ESTUDOS ESPECIAIS I

São José do Rio Preto – São Paulo

2020

RECONHECIMENTO DE EMOÇÕES NA FALA A PARTIR DA EXTRAÇÃO MANUAL DE CARACTERÍSTICAS COM VALIDAÇÃO BASEADA NA ENGENHARIA PARACONSISTENTE

Monografia apresentada para cumprimento da disciplina de estudos especiais do curso de Mestrado 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.

Orientador: Prof. Dr. Rodrigo Capobianco Guido.

Banca Examinadora

Professor Dr. Rodrigo Capobianco Guido (UNESP) - Campus de São José do Rio Preto Coorientador

Professora Dra. Renata Spolon Lobato (UNESP) - Campus de São José do Rio Preto

Professora Dra. Roberta Spolon (UNESP) - Campus de Bauru

São José do Rio Preto – São Paulo

2020

RESUMO

A presente monografia contém a descrição dos conceitos essenciais estudados para viabilizar, em futuro próximo, a confecção da dissertação de mestrado do autor, a qual possui foco no reconhecimento de emoções na fala, isto é, *Speech Emotion Recognition* (SER). Este documento traz uma breve introdução sobre o tema da dissertação, assim como a análise da relevância dos sistemas SER no cenário atual com base em um levantamento bibliográfico de 15 artigos científicos do estado-da-arte. Esta monografia encontra-se organizada na seguinte forma: O Capítulo 1 foi dedicado à introdução e ao histórico do problema a ser tratado, o Capítulo 2 se encarrega da revisão dos conceitos envolvidos no desenvolvimento do futuro trabalho, desde o processo de digitalização de sinais de voz, passando pelas possíveis abordagens de extração artesanal de características e pela análise paraconsistente de vetores de características, até a discussão dos artigos correlatos. Finalmente, no Capítulo 3, encontra-se o cronograma de etapas realizadas e das fases futuras.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Um sistema para reconhecimento de emoções no discurso (SER) pode ser definido como uma maneira automatizada de identificar o estado emocional de uma pessoa a partir da sua voz. O desenvolvimento de sistemas SER se faz possível pois o sinal de voz carrega informações conectadas não somente ao seu conteúdo lexical, mas também relacionadas à idade, gênero e estado emocional do falante. Tal afirmação pode ser encontrada em [16] onde mudanças acústicas produzidas na voz são analisadas em quatro emoções distintas: tristeza, raiva, felicidade e neutralidade. O estudo demonstra que, por exemplo, a fala associada à raiva e felicidade é caracterizada por uma maior duração, menor silêncio entre as palavras e um tom mais alto em comparação com falas associadas à tristeza e neutralidade.

As primeiras investigações acerca dos sistemas SER foram conduzidas por volta de meados da década de 80, como pode ser visto em [21] e [19]. Desde então é grande a variedade de aplicações desenvolvidas no contexto de reconhecimento de emoções a partir da voz, desde o uso para análise de satisfação de clientes no *e-commerce*, sistemas de saúde, até o monitoramento de estresse ou dor de pacientes, a fim de que doenças possam ser tratadas precocemente e/ou tratamentos sejam melhorados.

A voz, fisicamente, pode ser entendida como uma onda mecânica, a qual precisa de um meio para se propagar. Tratando a voz como uma manifestação na natureza, podemos visualizá-la como um sinal de uma dimensão, onde temos uma função ***f(t)*** descrita como valores de amplitude em função do tempo ***t***. Tendo isso em vista, faz-se interessante que analisemos tal sinal a partir de sistemas de processamento digital de sinais (DSP), desde que esse sinal seja previamente digitalizado.

O objetivo do estudo em questão, como dito anteriormente, consiste do reconhecimento de emoções presentes em um sinal de fala digitalizado a partir de um vetor de características extraído manualmente. A revisão apresentada a seguir engloba os principais conceitos envolvidos dentro deste estudo, entre os quais temos a digitalização de sinais de voz, a estrutura dos sistemas SER, a extração artesanal de características , a análise de vetores de características com base na engenharia paraconsistente, uma discussão contemplando o estado-da-arte na área e, ainda, a contextualização do andamento do projeto de mestrado.

2 LEVANTAMENTO BIBLIOGRÁFICO

2.1 Revisão dos conceitos utilizados

2.1.1 Digitalização do Sinal de Voz e Arquivos WAVE

Como mencionado na seção anterior, a voz pode ser vista como um sinal de uma única dimensão como uma função com valores de amplitude no eixo $\boldsymbol{Y}$ como imagem do domínio $\boldsymbol{X}$, que corresponde ao tempo. Visto que esse sinal é continuo nos dois eixos, faz-se necessário que façamos a digitalização do mesmo a fim de que possamos processá-lo por um computador. O Teorema de *Nyquist* afirma que a taxa de amostragem necessária para que o sinal seja reconstruído sem perdas deve ser pelo menos igual à duas vezes a maior frequência contida no sinal. Sendo assim, é comum que adotemos uma taxa de amostragem igual a 44100Hz, visto que a maior frequência audível ao ser humano corresponde a 22050Hz. A resolução da quantização comumente adotada é de 16 bits por canal, o que equivale a 65536 valores de aproximação.

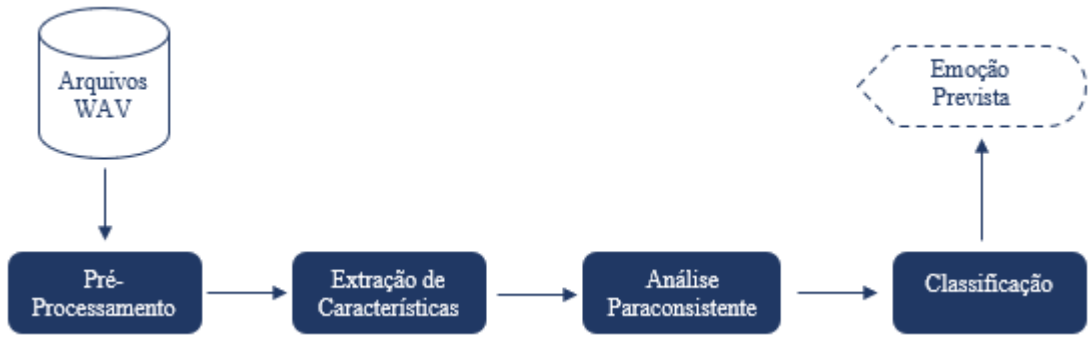
Neste estudo serão utilizados arquivos no formato WAV, o qual comumente usa as taxas mencionadas. Resumidamente, sem entrar em especificações mais aprofundadas, um arquivo WAV é composto por um cabeçalho com informações gerais sobre o arquivo e uma lista com os valores de amplitude para cada amostra do eixo temporal.

Informações acerca do banco de dados utilizado e suas características serão descritas na seção 2.2.1, a qual trata da contextualização das ferramentas e tecnologias utilizadas no andamento do projeto.

2.1.2 Estrutura de um Sistema SER

Apesar de existirem diversas variações quanto à sua estrutura, um sistema tradicional para o reconhecimento de emoções a partir da fala é composto de três partes principais: pré-processamento, extração de características e classificação. No presente estudo uma etapa é acrescentada antes do processo de classificação, chamada análise paraconsistente dos vetores de características. Na Figura 2.1.2 mostra-se o fluxo do sistema SER composto pelas partes acima citadas.

Figura 2.1.2. Estrutura do sistema SER



Fonte: Elaborado pelo Autor

A etapa de pré-processamento é responsável por ajustar detalhes, preparando os dados brutos para os processamentos vindouros. Nela, podemos notar a aplicação de filtros para a remoção de ruídos, segmentação do sinal entre partes vozeadas e não-vozeadas, entre outros. A fase de extração de características pode ser compreendida como parte crucial do sistema, visto que o seu objetivo é extrair descritores que sejam capazes de caracterizar os sinais, ou as classes em si, de forma que a etapa posterior de classificação possa apresentar resultados condizentes sem necessitar ser demasiadamente complexa. Visando possibilitar um maior grau de interpretabilidade, a abordagem adotada no presente estudo para a extração de características consiste no procedimento artesanal, isto é, *handcraft extraction*, ao invés de *feature learning*.

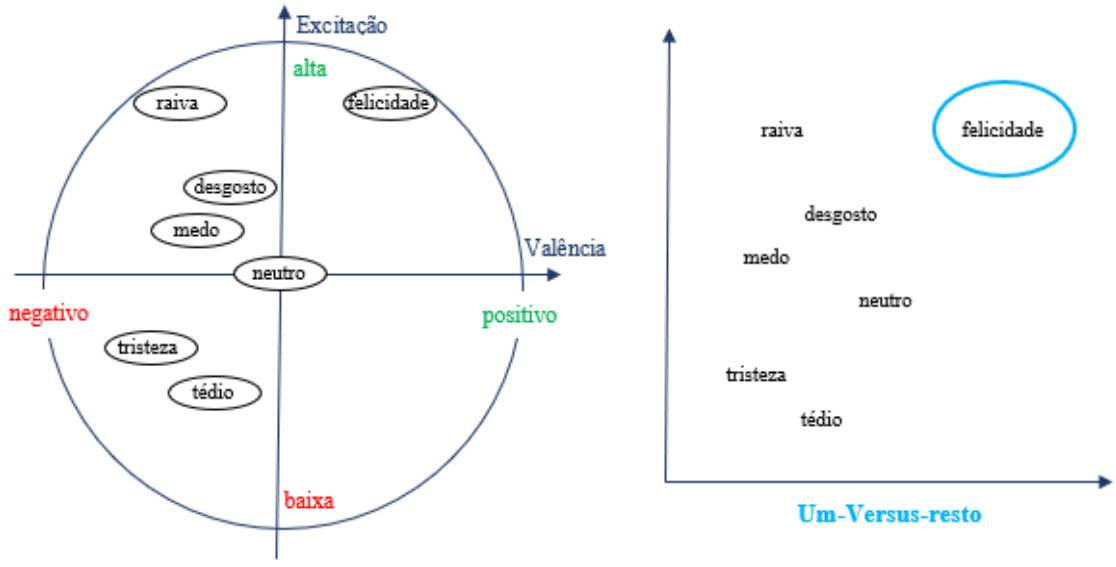
Prosseguindo, o ciclo de análise paraconsistente é responsável por inspecionar os vetores de características, independentemente do classificador a ser utilizado, visando mensurar o grau de adequabilidade das mesmas e, possivelmente, selecionar sub-conjuntos de características mais adequadas ao objetivo em questão. Tal análise é feita a partir de métricas de similaridade intraclasse e dissimilaridade interclasse, como será no visto na seção 2.1.4 destinada a esse assunto, a qual trará detalhadamente a explicação dos cálculos e métricas presentes dentro da engenharia paraconsistente.

A última etapa do sistema SER trata da classificação dos sinais, por meio dos vetores de características gerados e analisados anteriormente, os quais são fornecidos ao algoritmo classificador. Duas abordagens podem ser utilizadas nessa fase: *pattern-matching* (PM) e *knowledge-based* (KB). Naquela, uma ou mais amostras são escolhidas como modelos para cada classe de interesse a fim de que vetores de entrada sejam classificados a partir da comparação com tais modelos. Diferentemente, nesta abordagem, faz-se uso de um processo de aprendizagem no qual os vetores isolados para treinamento são utilizados para que as

métricas ou funções de aproximação do classificador sejam ajustadas a fim de que elementos de teste nunca "vistos" sejam classificados corretamente. Exemplos de classificadores baseados em PM são *hard thresholds* (HT) e *distance measures* (DM), enquanto Redes Neurais Artificiais (RNAs) e Máquinas de Vetor de Suporte (SVMs) são caracterizadas como KB.

Outra característica importante dentro de sistemas SER está relacionada ao objetivo de classificação dos sinais de voz, visto que três diferentes tipos de abordagem são comumente encontradas na literatura relacionada ao tema: classificação binária com base na dimensão de valência, classificação binária com base no nível de excitação e classificação discreta de emoções. Resumidamente, a dimensão de valência pode ser vista como um plano onde emoções consideradas positivas são posicionadas de um lado, enquanto emoções negativas são dispostas no outro. A lógica da classificação com base no nível de excitação é a mesma da anterior, com a diferença de que o plano divide emoções onde o nível de excitação é alto em comparação com outras que podem ser consideras mais apáticas. Por fim, a classificação de emoções discretas, a qual é adotada neste projeto, procura realizar o modelo de classificação um-versus-resto, ou seja, procura classificar cada emoção em sua devida categoria, isto é, felicidade, tristeza, neutralidade, e assim por diante. A Figura 2.1.2.1 contém, à esquerda, emoções nas dimensões de valência e excitação e, à direita, o modelo de classificação um-versus-resto.

Figura 2.1.2.1. Esquerda: emoções nas dimensões de valência e nível de excitação. Direita: emoções discretas.



Fonte: Elaborado pelo Autor

2.1.3 Exemplo de abordagem *handcrafted extraction* baseada no conceito de Energia

Um dos fatores mais críticos durante o desenvolvimento de sistemas SER está relacionado à escolha das características mais apropriadas e aptas a discriminar as emoções contidas no sinal de fala. A fim de resolver este problema, diversos artigos científicos procuraram estabelecer padrões de extração minimalistas, como pode se ver em [6] , enquanto outras publicações procuram se basear no conceito de timbre [2], considerando o formato da onda como um fator essencial para discriminar as emoções. De fato, inúmeras são as publicações que procuram trazer soluções quanto ao conjunto de descritores a serem extraídos do sinal, no entanto, todos afirmam que apesar das considerações feitas, a escolha da extração manual desses atributos é uma tarefa considerada empírica, e em geral, as extrações apresentadas nesses estudos provêm de uma combinação entre características estatísticas, considerando o sinal globalmente, e locais, isto é, baseadas no conceito de janelamento.

Em [14] encontra-se uma abordagem interessante e inovadora a respeito do uso do conceito de energia como base para extração de vetores de características com uma complexidade relativamente baixa, no entanto com um grande potencial de estabelecer padrões capazes de serem distinguidos de maneira eficiente pelo classificador, como pode se ver neste mesmo artigo, onde três diferentes problemas reais são tratados a partir de diferentes abordagens de extração que se utilizam da energia do sinal.

Conforme afirmado em [14], a energia de um sinal, fisicamente, representa sua capacidade de realizar trabalho. Quando estendemos 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 [14]. Por meio da Equação (1), defini-se a energia (E) de um sinal discreto s[.] com N amostras.

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

Após essa contextualização, apresentam-se as três abordagens propostas em [14] para extração de características baseada no conceito de energia, as quais devem ser usadas no desenvolvimento da dissertação de mestrado. A seguir, cada uma delas será detalhada.

2.1.3.1 Abordagem A1

Esta abordagem pode ser considerada como a mais simples entre as propostas no artigo em questão. A ideia é que o vetor extraído seja calculado a partir dos seguintes passos:

- segmentar o sinal em janelas de tamanho $\boldsymbol{L}$;
- definir uma taxa de sobreposição $\boldsymbol{V}$ entre as janelas, de forma que informação do sinal não seja perdida;
- calcular a energia de cada janela com a equação descrita anteriormente;
- realizar a normalização da energia de cada janela, dividindo pelo somatório da energia de todas as janelas.

A Figura 2.1.3.1 contém um exemplo real de extração do vetor de características a partir da abordagem A1, com base no exemplo numérico disponibilizado pelo artigo.

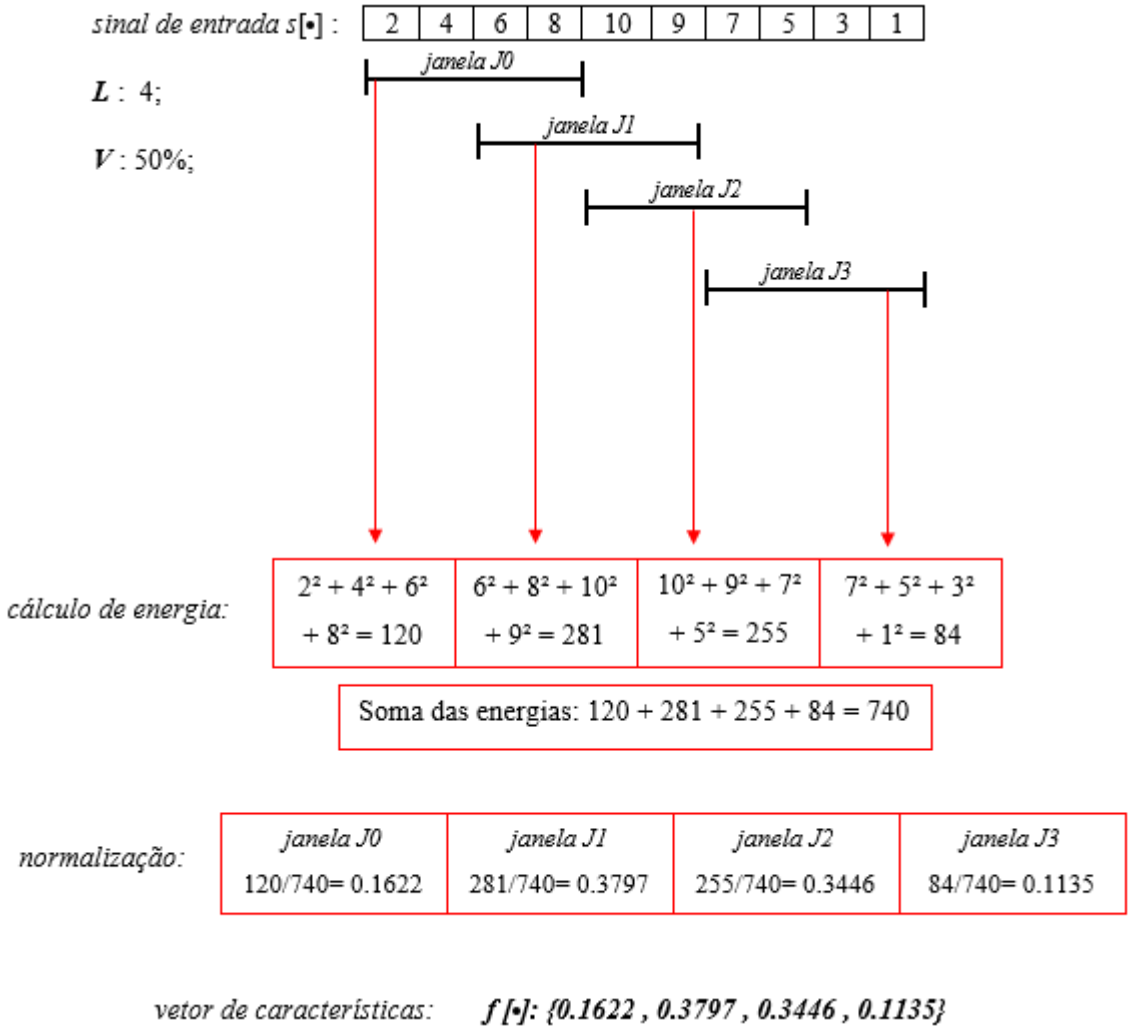
2.1.3.2 Abordagem A2

A ideia por trás dessa abordagem, assim como na anterior, consiste do janelamento do sinal, no entanto em A2 o tamanho da janela é variável e não existe taxa de sobreposição entre as amostras. A motivação para essa abordagem é a análise do sinal em diferentes níveis de resolução a partir da criação de sub-vetores de energia com base em diferentes tamanhos de janelas. As seguintes etapas devem ser cumpridas para o cálculo do vetor de energia:

- definir a quantidade de sub-vetores $\boldsymbol{Q}$;
- segmentar o sinal em janelas de acordo com a dimensão de cada sub-vetor $\boldsymbol{Q}$;
- calcular a energia normalizada para cada janela de cada sub-vetor $\boldsymbol{Q}$;
- concatenar os sub-vetores $\boldsymbol{Q}$ gerados a fim de obter um vetor final.

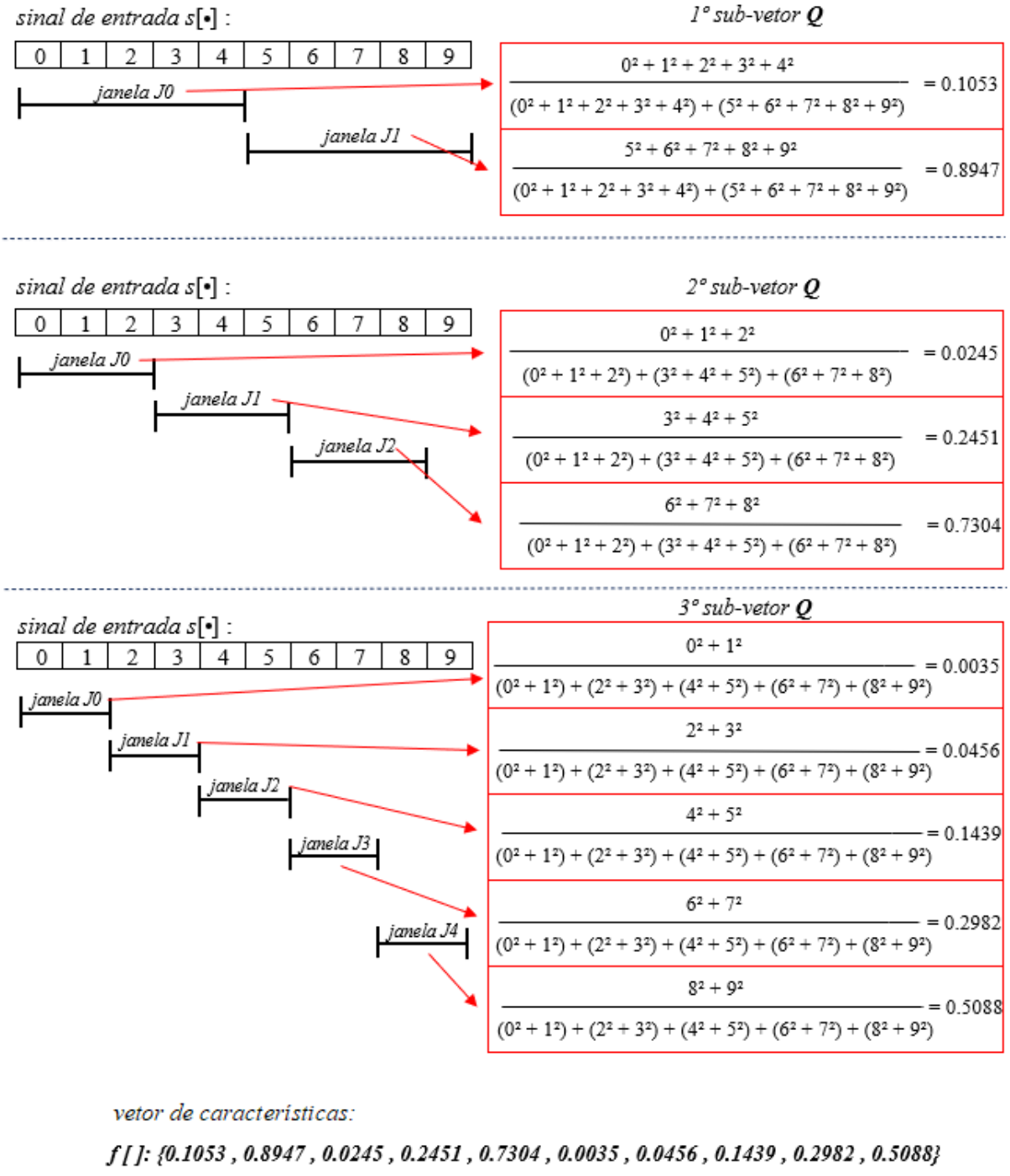
A Figura 2.1.3.2 contém um exemplo real de extração do vetor de características a partir da abordagem A2, baseado no exemplo numérico disponibilizado pelo artigo com $\boldsymbol{Q} = 3$.

Figura 2.1.3.1. Exemplo de extração do vetor de características a partir da abordagem A1



Fonte: Elaborado pelo Autor

Figura 2.1.3.2. Exemplo de extração de características baseada na abordagem A2



Fonte: Elaborado pelo Autor

2.1.3.3 Abordagem A3

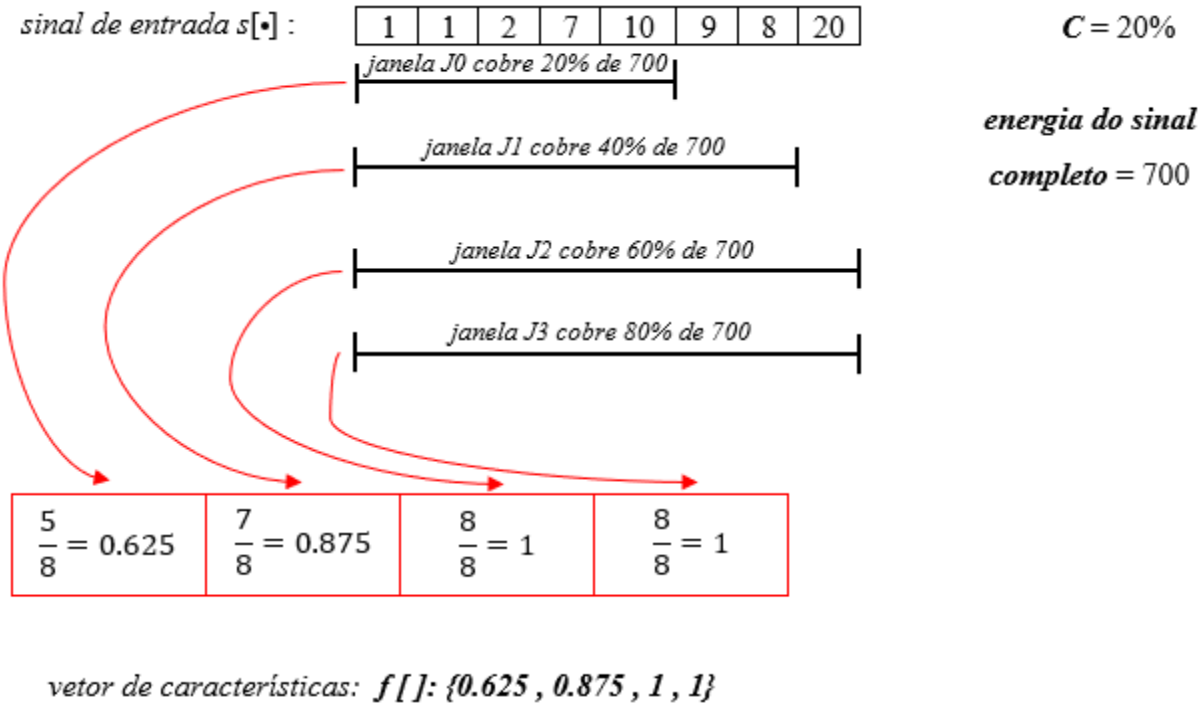
A última abordagem proposta no artigo possui um propósito diferente das duas anteriores. O objetivo de A3 é que possamos calcular o comprimento de sinal necessário para

cumprir níveis ***C*** pré-estabelecidos de energia. Para que o vetor de características possa ser gerado deve-se:

- definir um nível ***C*** de energia desejado;
- calcular o tamanho ***T*** do vetor a ser gerado com a equação $\boldsymbol{T} = (\frac{100}{\boldsymbol{C}}) - 1$;
- realizar um cálculo comum de energia do sinal completo com a equação de energia;
- calcular os subníveis de energia a partir de ***C*** (energia correspondente à 20%, 40%, 60% e 80%);
- verificar o comprimento de sinal necessário para atingir o subnível requerido;
- dividir o comprimento obtido no passo anterior pelo comprimento total do sinal.

A Figura 2.1.3.3 contém um exemplo de extração do vetor de características com base em A3, a partir do exemplo numérico disponibilizado pelo artigo com ***C*** = 20%.

Figura 2.1.3.3. Exemplo de extração de características baseada na abordagem A3 com C=20%



Fonte: Elaborado pelo Autor

2.1.4 Engenharia Paraconsistente de Características

Como mencionado nos capítulos anteriores, o presente projeto se baseia na extração de características realizada de forma artesanal. Visto que essa tarefa, apesar dos diversos estudos e padrões minimalistas encontrados na literatura, constitui uma escolha empírica dos atributos, faz-se necessário que façamos uma avaliação do vetor de características extraído, a fim de que possamos verificar a qualidade dos descritores, assim como sua aptidão para estabelecer algum nível de separabilidade entre as classes.

Uma vez que, atualmente, grande parte dos sistemas de processamento digital que possuem seu foco no reconhecimento de padrões se utilizam de algoritmos baseados em Inteligência Artificial para estabelecer funções de separabilidade entre as classes, torna-se interessante que usemos uma lógica alternativa à lógica clássica a fim de que situações conflitantes e indeterminadas possam ser tratadas.

Em [15], a lógica paraconsistente, a qual é capaz de lidar com situações como as descritas acima, é utilizada como uma ferramenta de avaliação de vetores de características extraídos manualmente, a fim de que a utilidade dos atributos extraídos seja provada para o problema a ser tratado, independentemente do classificador a ser utilizado.

A engenharia paraconsistente de características possui apenas como requisito que os vetores de características estejam previamente normalizados, com valores dentro do intervalo {0,1}. Após isso, inicia-se o cálculo das duas métricas estatísticas independentes que fornecem a base da análise. São elas:

- $\boldsymbol{\alpha}$: valor de similaridade intraclasse;
- $\boldsymbol{\beta}$: taxa de dissimilaridade interclasse.

Essas duas métricas se encontram dentro do intervalo {0,1}, e, idealmente, espera-se que o valor de $\boldsymbol{\alpha}$ se aproxime de 1, enquanto o valor de $\boldsymbol{\beta}$ se aproxime de 0. O cálculo de $\boldsymbol{\alpha}$ segue os seguintes passos para cada uma das N classes existentes no problema:

a. seleciona-se o maior e menor valor de cada posição, ou seja, cada característica, entre os vetores de características de uma mesma classe;
b. realiza-se o cálculo de $\boldsymbol{A = (maior - menor)}$;
c. calcula-se $\boldsymbol{Y = (1 - A)}$, de forma que $\boldsymbol{Y \approx 1}$ indica alta similaridade, enquanto $\boldsymbol{Y \approx 0}$ demonstra baixa similaridade;

- Tomando como exemplo que tenhamos 3 classes no problema, após o cálculo dos passos anteriores para cada classe teremos os vetores de similaridade ***svC0, svC1*** e ***svC2***;

d. realiza-se o cálculo da média de cada vetor, de forma que obteremos ***média(svC0), média(svC1)*** e ***média(svC2)***;

e. defini-se ***α= min[média(svC0), média(svC1), média(svC2)]***.

A Figura 2.1.4 contém um exemplo do processo descrito acima a partir da demonstração de um exemplo numérico.

Prosseguindo, define-se $\boldsymbol{\beta = \frac{R}{F}}$, onde o cálculo de ***R*** segue as seguintes etapas:

a. calcula-se dois vetores de intervalo para cada classe: um vetor com os elementos de menor valor e o outro com elementos de maior valor;

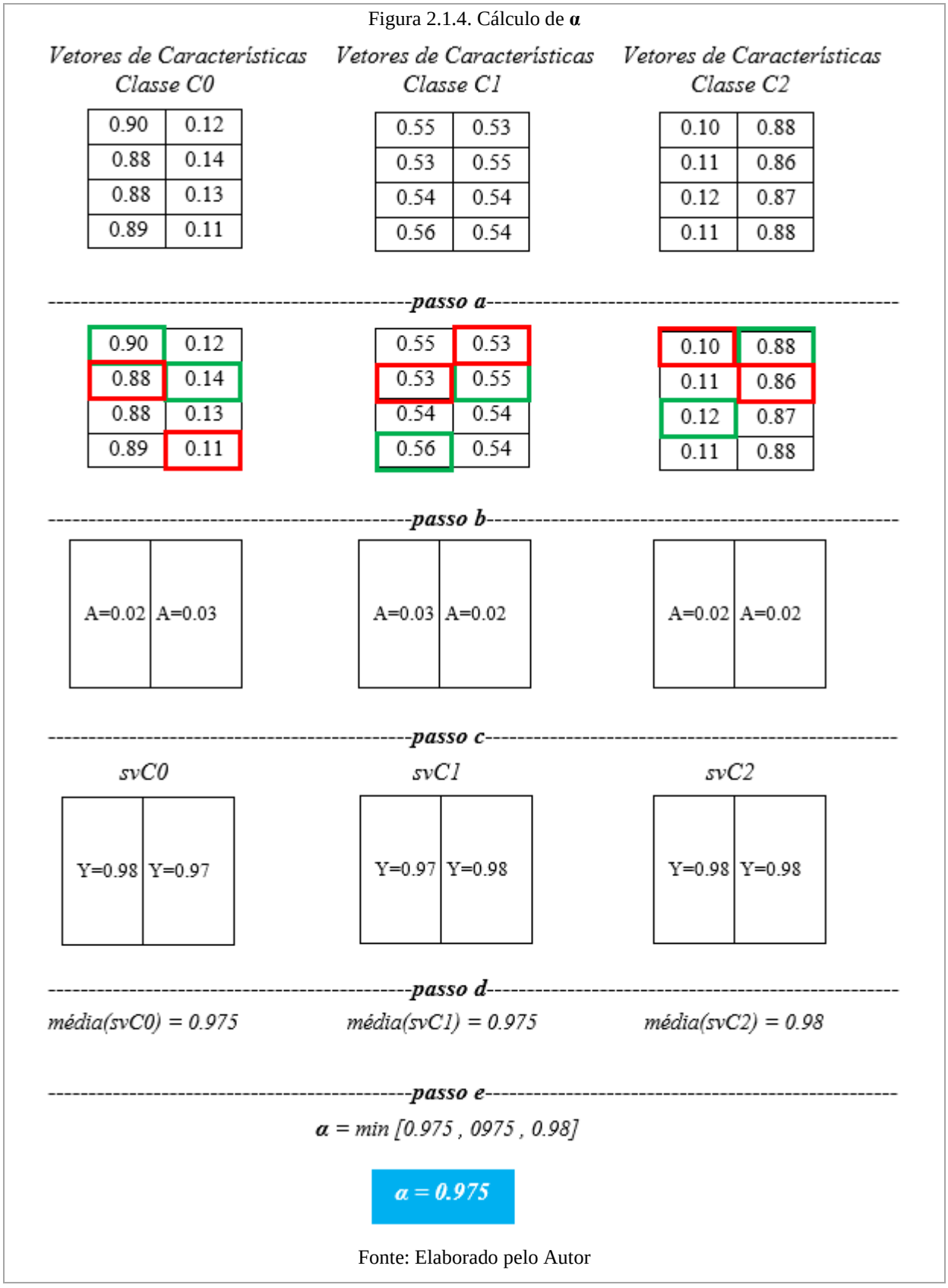
b. o valor de ***R*** é obtido através da contagem do número de sobreposições de intervalo para cada característica em comparação com a mesma característica nas outras classes.

O cálculo de ***F*** se dá a partir da seguinte equação: ***F = N * (N - 1) * X * T***. Temos que ***N*** equivale ao número de classes, ***X*** representa a quantidade de amostras em cada classe e ***T*** a dimensão do vetor de características. A Figura 2.1.4.1 contém um exemplo numérico do processo de obtenção de ***β***.

Visto que os parâmetros ***α*** e ***β*** podem assumir qualquer valor dentro do intervalo {0,1}, a engenharia paraconsistente implementa o cálculo do ***grau de certeza (G1)*** e do ***grau de contradição (G2)***, onde:

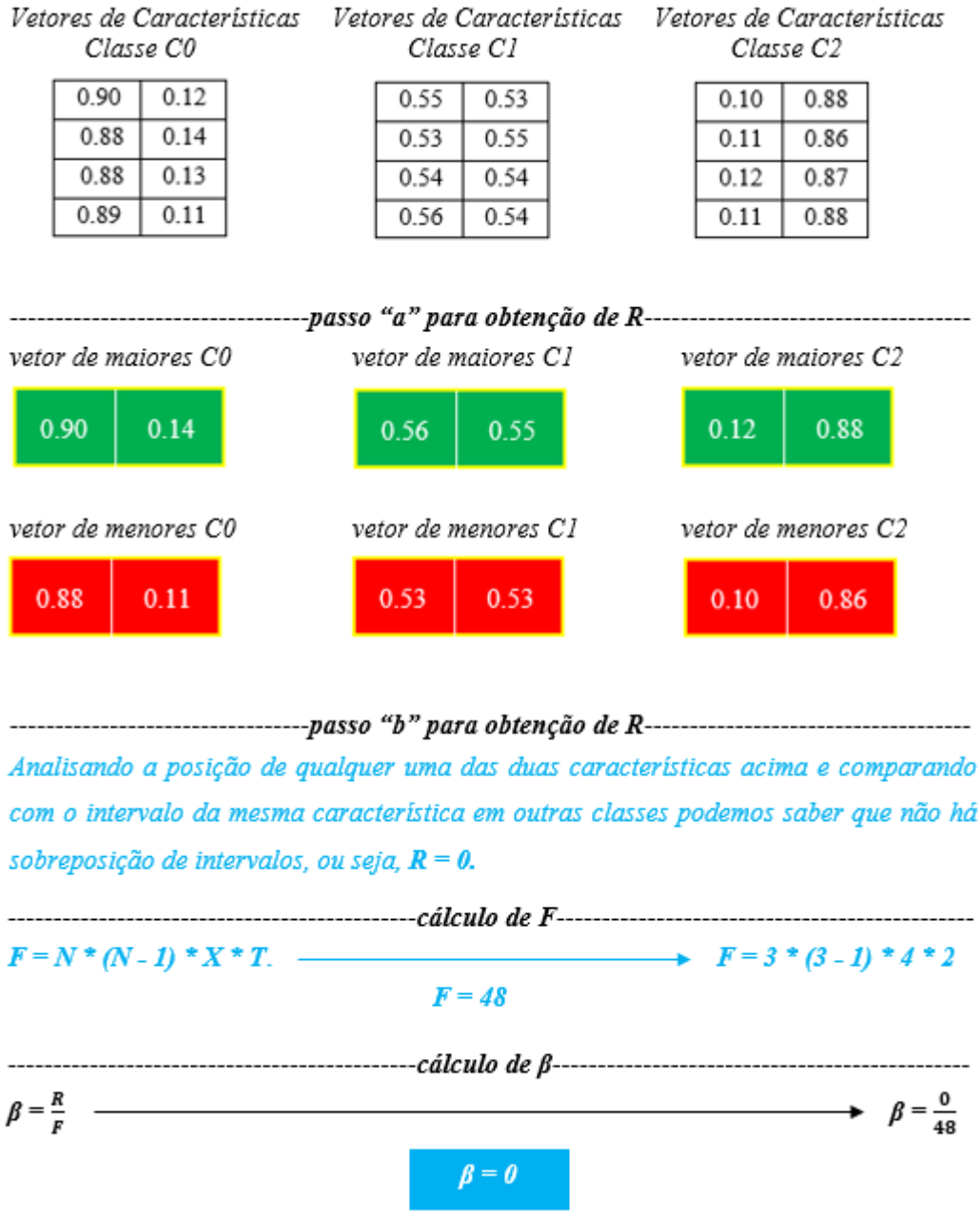
- ***G1 = α – β;***
- ***G2 = α + β – 1.***

Figura 2.1.4. Cálculo de **α**



Fonte: Elaborado pelo Autor

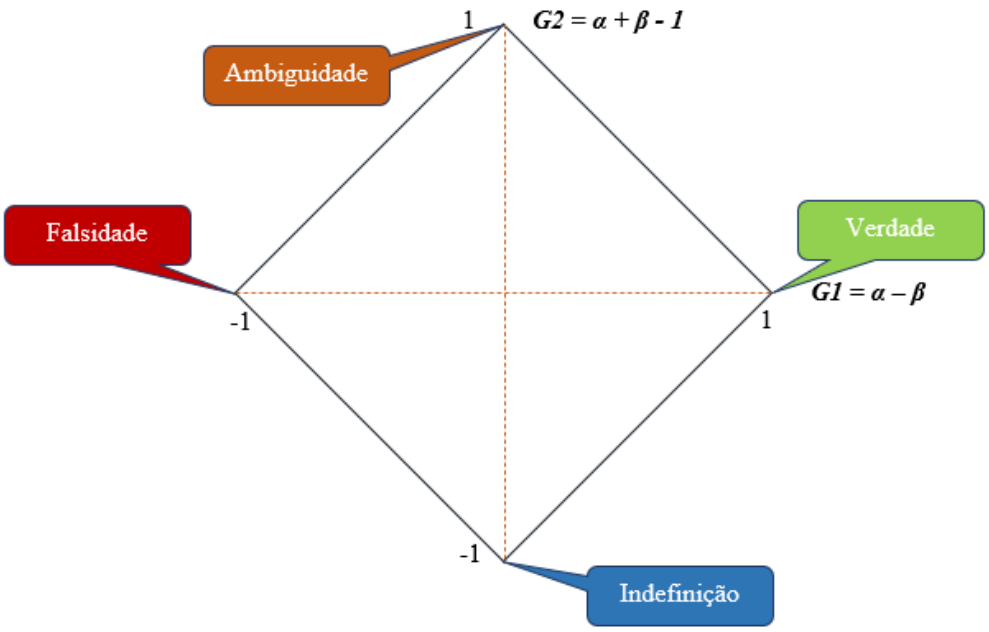
Figura 2.1.4.1. Cálculo de **β**



Fonte: Elaborado pelo Autor

Realizados os cálculos anteriores, estamos aptos a plotar um ponto ***P(G1, G2)*** no plano paraconsistente, onde podemos analisar a qualidade dos nossos vetores de características a partir da posição de ***P*** dentro do plano, conforme ilustrado na Figura 2.1.4.2.

Figura 2.1.4.2. Plano paraconsistente



Fonte: Elaborado pelo Autor

A fim de que possamos ter uma avaliação numérica da qualidade dos vetores de características, o artigo [15] implementa a abordagem do cálculo de distância do ponto $\boldsymbol{P}$ em relação aos quatro cantos indicados na Figura acima. Como pode ser visto, vetores de características mais favoráveis a estabelecer uma boa separabilidade entre as classes se localizam-se mais próximos ao canto (1,0) do que em relação aos outros cantos. A lista abaixo traz as equações para cálculo de distância do ponto $\boldsymbol{P}$ em relação a cada canto.

- $\boldsymbol{D}_{(1,0)} = \sqrt{(\boldsymbol{G}1 - 1)^2 + (\boldsymbol{G}2)^2}$
- $\boldsymbol{D}_{(0,-1)} = \sqrt{(\boldsymbol{G}1)^2 + (\boldsymbol{G}2 + 1)^2}$
- $\boldsymbol{D}_{(-1,0)} = \sqrt{(\boldsymbol{G}1 + 1)^2 + (\boldsymbol{G}2)^2}$
- $\boldsymbol{D}_{(0,1)} = \sqrt{(\boldsymbol{G}1)^2 + (\boldsymbol{G}2 - 1)^2}$

Finalmente, o plano paraconsistente fornece uma informação valiosa em relação ao uso de possíveis classificadores: desde que a menor distância do ponto $\boldsymbol{P}$ seja para $\boldsymbol{D}_{(1,0)}$ e que $\boldsymbol{P}$ esteja localizado no quadrante inferior direito, as características são linearmente separáveis, ou seja, um classificador mais simples pode ser utilizado a fim de solucionar o problema.

2.2 Estado-da-arte: revisão de artigos científicos relacionados ao projeto

O artigo [9] (2020) procura realizar a classificação de emoções contidas no discurso a partir da combinação (*feature selection*) entre características prosódicas (energia, frequência fundamental e *zero crossing rate* - ZCR), espectrais (obtidas após a transformação do sinal para o domínio de Fourier) e cepstrais (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.

Em [13] (2020) é apresentada uma revisão do estado de arte de todas as subáreas presentes dentro de sistemas SER: bases de dados, conjunto de características, técnicas de pré-processamento, isto é, *framing*, normalização, entre outros, modalidades de apoio, ou seja, uso de outras categorias de sinal como apoio a sistemas SER, classificadores e modelos emocionais, tais como classificação discreta e baseada em dimensões. Por meio do referido artigo, foi possível ter uma visão geral da área e delinear as direções a serem seguidas na dissertação de mestrado.

Em [7] (2019) é proposta uma abordagem de extração de características baseada em cálculos da dimensão fractal (DF) de sinais de voz. Visto que a dimensão fractal é capaz de descrever a fragmentação e irregularidade de um sinal, oito algoritmos distintos de cálculo da DF foram apresentados. Exemplificando a DF, o algoritmo de Katz, que é um dos utilizados dentro do estudo, consiste basicamente do somatório da distância euclidiana entre duas amostras adjacentes, possibilitando obter valores correspondentes à presença de baixas ou altas frequências dentro do sinal. Quanto maior a soma das distâncias, maior a presença de altas frequências. É importante citar que os cálculos de DF foram realizados após o janelamento do sinal, e por fim, medidas estatísticas foram extraídas dos cálculos de DF. O artigo apresenta taxas interessantes de reconhecimento, com uma acurácia média de **96,5%** 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 diferentes emoções. O artigo [17] (2019) procura realizar a classificação das emoções tendo como etapa de pré-

processamento a segmentação do sinal entre regiões semelhantes a vogais, incluindo semivogais e ditongos, e não semelhantes a vogais, de forma que a etapa de extração de características é realizada independentemente para cada uma dessas regiões. Dessa forma, um método de classificação baseado na alternância entre as regiões é 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 [18] (2019) procura realizar o reconhecimento de emoções se baseando exclusivamente em características obtidas a partir da frequência fundamental (F0) do sinal. A proposta realizada pelo estudo consiste do janelamento do sinal em *frames* de 20ms com uma sobreposição de 10ms entre os *frames*. Após esse processo a *Fast Fourier Transform* (FFT), versão otimizada da DFT, foi aplicada a cada frame e a frequência de maior pico encontrada a fim de determinar a F0. Finalmente, medidas estatísticas foram extraídas sobre os cálculos da F0, visto que os sinais possuíam diferentes comprimentos e dessa forma diferentes quantidades de F0 eram obtidas para cada áudio. A base de dados *Emotional Speech Corpus developed in AGH University of Science and Technology Krakow*, que possui 7 diferentes emoções, foi escolhida, no entanto os testes realizados pelo autor procuram explorar a classificação de subgrupos de 2, 3 e 4 emoções, a fim de provar a persistência de F0 na separabilidade de emoções distintas, com uma acurácia de **89,74%**, **76,14%** e **62,99%** para cada um dos subgrupos respectivamente.

Em [2] (2018) é proposta uma abordagem de extração de características a partir do timbre dos sinais de voz, como ferramenta de melhoria da taxa de acurácia na 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, ou seja, extraídas do sinal completo e através de janelamentos, e características selecionadas de timbre, isto é, o melhor subconjunto de características do grupo anterior, obtido através da aplicação de *Sequential Forward Selection*. Os experimentos foram feitos 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

através da combinação de características base e de timbres selecionadas, implicando a acurácia de **97,87%** ao se utilizar a SVM na classificação baseada na dimensão de valência.

No artigo [3] (2018) consta a 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 através do método *Principal Component Analysis* (PCA). As bases de dados EMO-DB e *Linguistic Data Consortium* (LDC) 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 *Multi Layer Perceptron* (MLP) como classificador.

Em [12] (2018) foi realizada a comparação de duas abordagens de extração de características (*Mel Frequency Cesptral Coefficients*- MFCC e *Modulation Spectral Features*- MSF), de forma que uma solução obtida a partir da combinação dessas abordagens é proposta no final do artigo. O estudo é realizado a partir do uso das bases de dados EMO-DB e INTER1SP *Spanish Emotional Database*. Os seguintes classificadores foram usados: *Multivariate Linear Regression* (MLR), SVM e *Recurrent Neural Network* (RNN), de modo que o melhor resultado alcançado pelo artigo se deu através da combinação de MFCC e MSF ao se utilizar a RNN para a base de dados INTER1SP.

O artigo [20] (2018) propõe um novo método de seleção de características como forma de melhorar a acurácia 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), *Fast-Correlation Based Filter* (FCBF), entre outros. Quatro bases de dados foram utilizadas durante os experimentos, visto que SVM, MLP *e k-Nearest Neighbors* (k-NN) foram os classificadores escolhidos. Resultado acerca da diminuição da carga de trabalho, assim como taxas de acurácia são apresentados no artigo, de forma que a maior taxa de acurácia utilizando o método de seleção proposto foi de **85.71%** quando a base EMO-DB foi utilizada em combinação com o classificador MLP.

Em [1] (2017) foi realizada a comparação da capacidade de classificação de emoções entre dois classificadores relativamente simples: *Decision Tree (DT)*, baseado em aprendizado de máquina, e *Logistic Regression (LR)*, baseado em cálculos estatísticos. Levando em conta a

simplicidade dos classificadores, o artigo realiza três tipos de classificação binária, com as respectivas taxas de acurácia:

i. classificação de sinais emocionais vs sinais neutros (DT: 84,45% / LR: 68,06%);

ii. classificação baseada na dimensão de valência (DT: 87,76% / LR: 69,49%);

iii. classificação de emoções positivas: felicidade vs surpresa (DT: 87,31% / LR: 70%).

Uma abordagem não muito comum é apresentada no artigo [5] (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 um processo de seleção de características é efetuado, de forma que os atributos selecionados são divididos em subconjuntos de características. Finalmente, cada classificador realiza a sua classificação inicial a partir de um desses subconjuntos, de forma que a classificação final é obtida pela classe com o maior número de predições (votos) entre os classificadores. Foi utilizado no processo uma base de dados de fala espontânea criada pelo próprio autor e uma base gravada com atores denominada *Polish Acted Emotional Speech*, com uma taxa média de acuraria de 82,6% e 72% para as respectivas bases de dados.

Com a enorme variedade de descritores extraídos na literatura sobre SER, o artigo [6] (2016) procura estabelecer um padrão minimalista de características a serem extraídas ao invés do uso de força bruta. A escolha desse padrão se deu através do estudo do potencial dos atributos para causar alterações fisiológicas afetivas na produção da voz, da comprovação de seu uso em estudos anteriores e sua significação teórica. Entre as características recomendadas podemos notar: atributos baseados em energia/amplitude, em frequência e em parâmetros espectrais. Seis bases de dados foram utilizadas na avaliação das características, entre as quais está a EMO-DB. No padrão proposto pelo artigo, taxas de acurácia de **79.71%** e **66.44%** foram alcançadas nas dimensões do nível de excitação e de valência respectivamente.

Em [8] (2016) é proposta a abordagem de extração de características e classificação de emoções através de *Feature Learning*, sendo assim, o sinal de áudio bruto é apresentado 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 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 classificações baseadas na dimensão de valência e nível de excitação foram escolhidas para testar a eficiência do sistema.

Em [11] (2016) é proposta a otimização do banco de filtros mel utilizado na extração de características MFCC através do uso de um algoritmo evolutivo ou 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 usadas com respectivas acurácias de **91,31%** e **42,50%**.

No artigo [10] (2015) é proposta a abordagem de extração de descritores de Fourier, tais como valores de magnitude, harmônicos, média, máximo, mínimo e desvio padrão das amplitudes dos parâmetros de Fourier, entre outros, como única fonte de características para a classificação de emoções. Foram utilizados os classificadores *Naive Bayes* (NB) e uma SVM, enquanto as bases de dados escolhidas foram as seguintes: EMO-DB, *Chinese Emotional Database* (CASIA) e *Chinese Elderly Emotional Speech Database* (EESDB). A abordagem proposta foi comparada com a utilização de MFCC e características prosódicas, tais como ZCR, energia e F0. Nos experimentos realizados, os descritores de Fourier conseguem taxas de acurácia melhores que as duas outras classes de características, principalmente para a base EMO-DB.

2.2.1 Contextualização do andamento do projeto

A escolha do banco de dados a ser utilizado, parte fundamental no desenvolvimento do projeto, foi realizada com base no estado de arte da literatura relacionada ao tema. A base de dados pública *Berlin Emotional Speech Database* (EMO-DB) consiste de 535 arquivos *WAV* com apenas um canal (mono) à uma resolução de 16 bits, visto que as gravações foram feitas com uma taxa de amostragem de 48Khz com a posterior realização de um *downsampling* para 16Khz, a fim de que processamento desnecessário seja evitado.

Esta base de dados foi gravada com a participação de 10 atores, sendo 5 homens e 5 mulheres, que são responsáveis por falar 10 frases em 7 estados emocionais diferentes, sendo que do total de 535 áudios tem-se:

- raiva, com 127 sinais;
- felicidade, com 71 sinais;
- neutralidade, com 79 sinais;
- tristeza, com 62 sinais;

- medo, com 69 sinais;
- tédio, com 81 sinais;
- desgosto, com 46 sinais.

A Tabela 2.2.1 contém as 10 frases disponibilizadas pela base de dados. Mais informações a respeito do ***EMO-DB*** podem ser encontradas em [4] e em *http://emodb.bilderbar.info/index-1280.html.*

A biblioteca de aquisição e leitura de arquivos *wave* criada e disponibilizada pelo orientador deste trabalho foi melhorada e implementada em uma abordagem orientada a objetos. As abordagens de extração de características baseadas no conceito de energia, assim como a análise dos vetores com base na engenharia paraconsistente também foram implementas e encontram-se disponibilizadas em https://github.com/hiagomb/special_studies_1. Encontra-se em andamento a implementação de um processo teste de extração, avaliação e construção de um classificador.

Tabela 2.2.1. Frases da base de dados EMO-DB

texto original (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

3 CRONOGRAMA DE DESENVOLVIMENTO

A Figura 3.1 contém o cronograma que deve ser seguido até o término do curso de mestrado do autor.

Figura 3.1. Cronograma

Fonte: Elaborado pelo Autor

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Agnes Jacob. Modelling speech emotion recognition using logistic regression and decision trees. International Journal of Speech Technology: 897-905, 2017.

[2] Anvarjon Tursunov, Soonil Kwon and Hee-Suk Pang. Discriminating Emotions in the Valence Dimension from Speech Using Timbre Features. Applied Sciences, Jun 2019.

[3] Ashishkumar Gudmalwar, Chevula Rama Rao, Anirban Dutta. Improving the performance of the speaker emotion recognition based on low dimension prosody features vector. International Journal of Speech Technology, 2018.

[4] Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W. F. & Weiss, B. A database of German emotional speech, in 'INTERSPEECH', ISCA: 1517-1520, 2005.

[5] Dorota Kamińska, Tomasz Sapiński. Polish Emotional Speech Recognition Based on the Committee of Classifiers. Przeglad Elektrotechniczny: 101-106, 2017.

[6] Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.;Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing.IEEE Trans. Affect. Comput: 190-202, 2016.

[7] Gintautas Tamulevičius, Rasa Karbauskaitė, Gintautas Dzemyda. Speech emotion classification using fractal dimension-based features. Nonlinear Analysis: Modelling and Control: 679-695, 2019.

[8] G. Trigeorgis et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 5200-5204, 2016.

[9] Kudakwashe Zvarevashe, Oludayo Olugbara. Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition, Algorithms: 1-24, 2020.

[10] Kunxia Wang, Ning An, Bing Nan Ali, Yanyong Zhang. Speech Emotion Recognition Using Fourier Parameters. Affective Computing, IEEE Transactions on: 69-75, 2015.

[11] Leandro Daniel Vignolo, S.R.Mahadeva Prasanna, Samarendra Dandapat, Hugo Rufiner, Diego Milone. Feature optimisation for stress recognition in speech. Pattern Recognition Letters: 1-7, 2016.

[12] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub. Speech Emotion Recognition: Methods and Cases Study. International Conference on Agents and Artificial Intelligence: 175-182, 2018.

[13] Mehmet Berkehan Akçay, Kaya Oguz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication: 56-76, 2020.

[14] Rodrigo Capobianco Guido. A tutorial on signal energy and its applications. Neurocomputing 179: 264-282, Dez 2015.

[15] Rodrigo Capobianco Guido. Paraconsistent feature engineering [Lecture notes]. IEEE Signal Processing Magazine, 36 (1): 154–158, Jan 2019.

[16] Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth S. Narayanan. An acoustic study of emotions expressed in speech. In Proceedings of Inter Speech: 2193–2196, Out 2004.

[17] Suman Deb, Samarendra Dandapat. Emotion Classification Using Segmentation of Vowel-Like and Non-Vowel-Like Regions. in *IEEE Transactions on Affective Computing*: 360-373, 2019.

[18] Teodora Dimitrova-Grekow, Aneta Klis, Magdalena Igras-Cybulska. Speech Emotion Recognition Based on Voice Fundamental Frequency: 277-286, 2019.

[19] Tolkmitt, F.J., Scherer, K.R. Effect of experimentally induced stress on vocal parameters. J. Exp. Psychol. [Hum.Percept.] 12 (3): 302–313, 1986.

[20] Turgut Ozseven. A novel feature selection method for speech emotion recognition. Applied Acoustics: 320-326, 2018.

[21] Van Bezooijen, R. The Characteristics and Recognizability of Vocal Expression of Emotions. Foris, Drodrecht, The Netherlands, 1984.