Victor Gabriell Ribeiro da Silva

ANÁLISE DO SINAL DE FALA PARA RECONHECIMENTO DE EMOÇÕES UTILIZANDO REPRESENTAÇÃO SEMÂNTICA

São Cristóvão - SE

Dezembro de 2022

Victor Gabriell Ribeiro da Silva

ANÁLISE DO SINAL DE FALA PARA RECONHECIMENTO DE EMOÇÕES UTILIZANDO REPRESENTAÇÃO SEMÂNTICA

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica - PROEE, da Universidade Federal de Sergipe, como parte dos requisitos necessários para a obtenção do título de Mestre em Engenharia Elétrica

Universidade Federal de Sergipe – UFS

Programa de Pós-Graduação em Engenharia Elétrica - PROEE

Orientador: Prof. Dr. Jugurta Rosa Montalvão Filho

São Cristóvão - SE

Dezembro de 2022

TERMO DE APROVAÇÃO

" Análise do sinal de fala para reconhecimento de emoções utilizando representação semântica"

Discente:

Victor Gabriell Ribeiro da Silva

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Sergipe, como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica.

Aprovada pela banca examinadora composta por:

gov.br Documento assinado digitalmente
JANIO COUTINHO CANUTO
Data: 23/12/2022 21:42:16-0300
Verifique em https://verificador.iti.br

Prof. Dr. Jânio Coutinho Canuto (PROEE/UFS)
Presidente

gov.br Documento assinado digitalmente
ALEXANDRE LUIS MAGALHAES LEVADA
Data: 21/12/2022 18:48:53-0300
Verifique em https://verificador.iti.br

Prof. Dr. Alexandre Luis Magalhães Levada (UFSCAR)
Examinador Externo

gov.br Documento assinado digitalmente
LEONARDO NOGUEIRA MATOS
Data: 21/12/2022 17:28:49-0300
Verifique em https://verificador.iti.br

Prof. Dr. Leonardo Nogueira Matos (DCOMP/UFS)
Examinador Externo

gov.br Documento assinado digitalmente
VICTOR GABRIELL RIBEIRO DA SILVA
Data: 26/12/2022 10:57:24-0300
Verifique em https://verificador.iti.br

Victor Gabriell Ribeiro da Silva
Candidato

Cidade Universitária "*Prof. José Aloísio de Campos*", 21 de dezembro de 2022.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Silva, Victor Gabriel Ribeiro da
S586a Análise do sinal de fala para reconhecimento de emoções utilizando representação semântica / Victor Gabriel Ribeiro da Silva ; orientador Jugurta Rosa Montalvão Filho. - São Cristóvão, 2022.
93 f.: il.

Dissertação (mestrado em Engenharia Elétrica) – Universidade Federal de Sergipe, 2022.

1. Engenharia elétrica. 2. Emoções. 3. Fala. 4. Voz. 5. Semântica. I. Montalvão Filho, Jugurta Rosa orient. II. Título.

CDU 621.3: 81’37

Agradecimentos

Sou eternamente grato à toda minha família, que sempre me incentivou e apoio durante toda minha trajetória acadêmica, especialmente, minha mãe Josefina.

Ao meu orientador Jugurta, pelos ensinamentos e confiança. Pelas suas ótimas explicações simples para conceitos tão complexos. Por todo apoio, que mesmo durante a pandemia, foram essenciais para finalizar esse trabalho. Agradeço também pela atenção e dedicação durante a correção deste trabalho, que me fizeram aprender ainda mais.

Aos colegas que ganhei no mestrado: Israel e Vitor Magno, pela colaboração, conversas e incentivos. Também pelas reuniões aos sábados de manhã, que me ajudaram a entender vários assuntos importantes para esse trabalho.

Aos membros da banca examinadora pelo interesse, disponibilidade, além dos seus comentários pertinentes. A todos os professores que tive a oportunidade de conhecer durante o mestrado, em especial, Jânio e Eduardo, que contribuíram de forma significativa na formação do meu conhecimento sobre otimização, processamento de imagens e grafos.

À Mayane, por toda paciência e gentileza em me explicar os processos do mestrado, em todos os momentos que precisei.

A todos os professores incríveis que tive a sorte de encontrar na minha jornada e que contribuíram de várias formas para que eu chegasse até aqui, principalmente, Alan, Dami, Felipe e Lívia.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo suporte financeiro durante o mestrado.

Resumo da Dissertação apresentada ao PROEE/UFS como parte dos requisitos necessários para a obtenção do grau de Mestre (Me.)

ANÁLISE DO SINAL DE FALA PARA RECONHECIMENTO DE EMOÇÕES UTILIZANDO REPRESENTAÇÃO SEMÂNTICA

Victor Gabriell Ribeiro da Silva

Dezembro/2022

Orientador: Prof. Dr. Jugurta Rosa Montalvão Filho

O interesse em analisar automaticamente as emoções humanas tem crescido nos últimos anos, principalmente devido as suas possíveis aplicações na sociedade. Ao longo dos últimos 20 anos, diversos trabalhos foram propostos a fim de determinar características nos sinais de voz capazes de representar as emoções. A maior parte dos trabalhos utiliza atributos associados à prosódia e ao espectro do sinal de voz. Neste trabalho, são apresentadas metodologias de análise do sinal de voz, explorando principalmente as técnicas de pré-processamento, extração de características e classificação. Além disso, é utilizada uma metodologia baseada na representação semântica, a fim de obter um mapeamento não-linear do espaço de características que seja útil para acrescentar informações complementares, consequentemente, aumentando a discriminação entre as classes. Os resultados obtidos apontam que, quando ajustada adequadamente, a representação semântica é capaz de aumentar significativamente a acurácia da classificação de emoções expressadas em sinais de voz.

Palavras-chaves: Reconhecimento de emoções; Voz; Representação semântica.

Abstract of Dissertation presented to PROEE/UFS as a partial fulfillment of the requirements for the degree of Master

Speech signal analysis for emotion recognition using semantic representation

Victor Gabriell Ribeiro da Silva

Dezembro/2022

Advisor: Prof. Dr. Jugurta Rosa Montalvão Filho

The interest in automatically analyzing human emotions has grown in recent years, mainly due to its possible applications in society. Over the last 20 years, several works were proposed in order to extract features in speech signals capable of representing emotions. Most works use features associated to prosody and the spectrum of the speech signal. In this work, speech analysis methodologies are presented, exploring mainly pre-processing, feature extraction and classification techniques. Furthermore, a methodology based on semantic representation is used in order to obtain a non-linear mapping of the feature space. Then, it is possible to use the semantic representation to add complementary information and, consequently, increase the discrimination between the classes. The results obtained show that semantic representation is able to significantly increase the accuracy in the task of speech emotion recognition.

Keywords: Emotion recognitio; Speech; Semantic representation.

Sumário

Lista de ilustrações . . . . . **i**
Lista de tabelas . . . . . **ii**
Lista de símbolos . . . . . **iii**

1 INTRODUÇÃO **1**
1.1 Problemática **2**
1.2 Contribuições **3**
1.3 Objetivos **3**
1.4 Organização do trabalho **4**

2 AS EMOÇÕES **5**
2.1 Aspectos teóricos sobre as emoções **5**
2.1.1 Perspectiva darwinista 5
2.1.2 Perspectiva jamesiana 6
2.1.3 Perspectiva cognitiva 7
2.2 Relações entre emoção e voz **7**

3 ANÁLISE DO SINAL DE VOZ **10**
3.1 Introdução sobre o sinal de voz **10**
3.2 Pré-processamento do sinal de voz **11**
3.2.1 Janelamento 11
3.2.2 Detecção de atividade vocal 12
3.2.2.1 Energia 12
3.2.2.2 Taxa de cruzamento por zero 12
3.2.2.3 Autocorrelação 13
3.2.2.4 Outras técnicas 13
3.3 Extração de características acústicas **14**
3.3.1 Prosódia 14
3.3.1.1 Frequência fundamental 14
3.3.1.2 *Jitter* e *shimmer* 15
3.3.1.3 Energia 15

3.3.2 Características espectrais . . . 16
3.3.2.1 Coeficientes mel-cepstrais . . . 16
3.3.2.2 Coeficientes LFPC . . . 17

4 REPRESENTAÇÃO SEMÂNTICA . . . 18
4.1 Quantização vetorial . . . 19
4.2 Contexto semântico . . . 19
4.2.1 Contexto baseado em vizinhança . . . 20
4.2.2 Contexto baseado no TF-IDF . . . 21
4.3 Otimização . . . 23

5 CLASSIFICAÇÃO . . . 25
5.1 Modelo de mistura de Gaussianas . . . 25
5.2 Redes neurais artificiais . . . 26
5.3 k-vizinhos mais próximos . . . 29

6 REVISÃO BIBLIOGRÁFICA . . . 31

7 METODOLOGIA . . . 36
7.1 Conjunto de dados . . . 36
7.2 Pré-processamento do sinal de voz . . . 37
7.3 Extração de características . . . 39
7.3.1 Frequência fundamental . . . 39
7.3.2 *Jitter* e *shimmer* . . . 41
7.3.3 Energia . . . 44
7.3.4 Taxa de cruzamento por zero . . . 45
7.3.5 Coeficientes mel-cepstrais . . . 45
7.3.6 Coeficientes LFPC . . . 46
7.3.7 Normalização . . . 48
7.4 Representação semântica . . . 48
7.4.1 Contexto baseado em vizinhança . . . 49
7.4.2 Contexto baseado em TF-IDF . . . 49
7.5 Métodos de classificação . . . 50
7.5.1 Modelo de mistura gaussianas . . . 50
7.5.2 Rede neural artificial . . . 52
7.5.3 k-vizinhos mais próximos . . . 53

8 RESULTADOS E DISCUSSÕES . . . 54
8.1 Resultados para o contexto baseado em vizinhança . . . 55
8.1.1 Primeiro experimento . . . 55
8.1.2 Avaliação dos parâmetros do contexto de vizinhança . . . 56

8.1.2.1 Número de dimensões 56
8.1.2.2 Janela de contexto 57
8.1.3 Avaliação do k-NN 58
8.2 Resultados para o contexto baseado no TF-IDF 60
8.2.1 Primeiro experimento 60
8.2.2 Avaliação dos parâmetros do contexto TF-IDF 61
8.2.2.1 Número de dimensões 61
8.2.2.2 Número de símbolos com maiores TF-IDF por classe 61
8.2.3 Avaliação do k-NN 62
8.3 Comentários finais 63

9 CONCLUSÃO 67

REFERÊNCIAS 69

Lista de ilustrações

Figura 1.1 – Número de publicações relacionadas a reconhecimento de emoções através de sinais de voz indexadas no *ScienceDirect* entre os anos de 2000 e 2021. 1
Figura 2.1 – Ordem de eventos para acontecer uma emoção, de acordo com a perspectiva jamesiana. 6
Figura 2.2 – Oito emoções apresentadas no modelo bidimensional. 8
Figura 3.1 – Sinal de voz ilustrado como contínuo no tempo. 11
Figura 3.2 – Sinal de voz ilustrado como amostrado. 11
Figura 3.3 – Sinal de voz ilustrado como amostrado, entre as amostras 1 e 100. . . . 11
Figura 3.4 – Ilustração de segmento com voz. 12
Figura 3.5 – Ilustração de segmento com ruído. 12
Figura 3.6 – Ciclos de um segmento de voz. O período de um ciclo é indicado por T. 15
Figura 3.7 – Ciclos de um segmento de voz, com destaque para a amplitude pico a pico definida como A. 16
Figura 3.8 – Filtros passa-bandas para determinar os coeficientes mel-cepstrais. . . . 17
Figura 4.1 – Contexto de vizinhança simétrico e assimétrico. 20
Figura 4.2 – Matriz de coocorrências para o conjunto de palavras da frase *João ganhou uma bicicleta azul, mas João não queria uma bicicleta azul*, escolhendo a primeira palavra (antes e depois) como vizinha. 21
Figura 5.1 – Diagrama de uma rede neural do tipo mais popular, com apenas uma camada oculta. A camada de entrada, oculta e saída são representadas pelos nós, e os pesos são representados pelos arcos entre cada círculo. As setas indicam o sentido do fluxo da informação na rede durante o processo de ida (conhecido como *feedfoward*). 27
Figura 5.2 – Ilustração do algoritmo k-NN com duas classes com amostras sobrepostas. O círculo azul representa uma amostra com classe desconhecida. Já o losango laranja e o círculo cinza representam amostras de classes diferentes. 30
Figura 7.1 – (a) Janela do sinal de voz no domínio do tempo; (b) Domínio da frequência. 38

Figura 7.2 – Espectrograma do sinal 03a01Fa do banco de dados EmoDB. 39
Figura 7.3 – Espectrograma do sinal 03a01Fa do banco de dados EmoDB após o VAD. 39
Figura 7.4 – Função de autocorrelação para uma determinada janela de um sinal de voz. O dm_{pico}, sinalizado com um retângulo vermelho, indica o valor de dm para o primeiro pico da função de autocorrelação. Note que até o atraso $Fs/560 = 16000/560 \approx 28$, a autocorrelação é 0, devido a limitação de valores para $f0$, como descrito no passo 2. 40
Figura 7.5 – Ilustração da estimação da $f0$. No item (a), é mostrada a $f0$ estimado através da técnica da autocorrelação. No item (b), é apresentada a $f0$ após o processo de redução de grandes variações. Já no item (c), é mostrada a $f0$ obtida após o filtro de média móvel. 42
Figura 7.6 – Comparativo entre a $f0$ (em azul) estimada e o espectrograma entre 1 e 560 Hz para o sinal de voz 03a01Nc do banco de dados EmoDB. . . . 42
Figura 7.7 – Ilustração dos máximos locais, $Ml(i)$ e suas respectivas amostras (ou localizações), $n_{Ml(i)}$. Também é mostrado o mínimo local de cada ciclo, $ml(i)$. Idealmente, $Ml(i)$ deve representar o pico e $ml(i)$ deve representar o vale do ciclo i. 43
Figura 7.8 – Filtros passa-bandas usados no passo 2 para a determinação dos MFCCs. 46
Figura 8.1 – Esquema para a classificação independente do orador. Para cada grupo, é ajustado os parâmetros do classificador, em seguida, o classificador é testado com o grupo de teste para obter a acurácia. Por fim, a acurácia média é determinada. 55
Figura 8.2 – Comparação entre acurácias para a representação semântica ao variar a dimensão da representação semântica, e a média da acurácia para a representação natural. São usados 12 coeficientes MFCC e a rede neural artificial como classificador. O número de vizinhos do contexto semântico é 100 para todos os experimentos. 57
Figura 8.3 – Comparação entre acurácias para a representação semântica ao variar o tamanho da janela de contexto, e a média da acurácia para a representação natural. São usados 12 coeficientes MFCC e a rede neural artificial como classificador. A dimensão é 2 para todos os experimentos. 58
Figura 8.4 – Comparação entre acurácias para a representação semântica e natural ao variar o número de vizinhos, k, do o classificador k-NN. São usados 12 coeficientes MFCC como características. O número de vizinhos para o contexto baseado em vizinhança é 100, e a dimensão da representação semântica é 2 para todos os experimentos. 59

Figura 8.5 – Comparação entre acurácias para a representação semântica ao variar a dimensão da representação semântica, e a média da acurácia para a representação natural. São usados 12 coeficientes MFCC como características e o GMM como classificador. O número de vizinhos é 100 para todos os experimentos. 62
Figura 8.6 – Comparação entre acurácias para a representação semântica ao variar o número de símbolos com maiores TF-IDF por classe, e a média da acurácia para a representação natural. São usados 12 coeficientes MFCC como características e o GMM como classificador. A dimensão é 2 para todos os experimentos. 62
Figura 8.7 – Comparação entre acurácias para a representação semântica e natural ao variar o número de vizinhos, k, do o classificador k-NN. São usados 12 coeficientes MFCC. O número de símbolos com maiores TF-IDF por classe é 500, e a dimensão da representação semântica é 2 para todos os experimentos. 63
Figura 8.8 – Disposição dos sinais na representação semântica com contexto baseado na vizinhança, usando 12 MFCCs e 200 vizinhos. Cada círculo representa a média de todos vetores de um áudio. 65
Figura 8.9 – Disposição dos sinais na representação semântica com contexto baseado no TF-IDF, usando 12 MFCCs e 700 símbolos por classe. Cada círculo representa a média de todos vetores de um áudio. 66

Lista de tabelas

Tabela 1 – Descrição de trabalhos relevantes relacionados à área de classificação de emoções através de sinais de voz, em ordem cronológica. 34
Tabela 2 – Quantidade de arquivos de áudio por emoção. 37
Tabela 3 – Codificação *one-hot* para as emoções. 52
Tabela 4 – Oradores utilizados para cada grupo. 55
Tabela 5 – Acurácia para três classificadores diferentes (GMM, RNA e k-NN), para cada vetor características, usando o contexto semântico baseado em vizinhança. 56
Tabela 6 – Acurácia para três classificadores diferentes (GMM, RNA e k-NN), para cada vetor características, usando o contexto semântico baseado no TF-IDF. 60

Lista de símbolos

$s(n)$	Sinal de voz amostrado, definido em $\mathbb{R}$
N	Número de amostras ou observações de um determinado vetor.
E	Energia, definido em $\mathbb{R}$
tcz	Taxa de cruzamento por zero, definido em $\mathbb{R}$
R_{ss}	Autocorrelação do sinal s, definido em $\mathbb{R}$
J	Função de custo, definida em $\mathbb{R}$
V	Tamanho de um vocabulário com estados discretos, definido em $\mathbb{N}$
w	Estado discreto, definido em $\mathbb{N}$
$TFIDF$	*Term frequency - inverse document frequency*, em português, frequência do termo - inverso da frequência nos documentos, definido em $\mathbb{R}$
$dim_{\mathbf{x}}$	Dimensão de um determinado vetor $\mathbf{x}$, definido em $\mathbb{N}$
MFCC	Vetor de coeficientes mel-cepstrais, definidos em $\mathbb{R}^{dim_{\mathbf{MFCC}}}$
LFPC	Vetor de coeficientes LFPC, definido em $\mathbb{R}^{dim_{\mathbf{LFPC}}}$
$\mathbf{x}$	Vetor de características, definido em $\mathbb{R}^{dim_{\mathbf{x}}}$
$\mathbf{g}$	Vetor obtido na representação semântica, definido em $\mathbb{R}^{dim_{\mathbf{g}}}$
$\boldsymbol{\mu}$	Centroide, definido em $\mathbb{R}^{dim_{\boldsymbol{\mu}}}$
$\mathbf{S}$	Matriz de similaridade, definida em $\mathbb{R}$
$\mathbf{D}$	Matriz de dissimilaridade, definida em $\mathbb{R}$
$\mathbf{B}$	Matriz de relacionamento entre classes e símbolos, definido em $\mathbb{R}$
$\mathbf{Q}$	Matriz de distância euclidianas, definido em $\mathbb{R}$

Capítulo

1 Introdução

Nos últimos anos, tem crescido o interesse em se investigar automaticamente as emoções humanas e como elas se manifestam devido às suas potenciais aplicações em dispositivos inteligentes, especialmente em dispositivos que requerem uma alta interação entre o homem e a máquina, além de aplicações em robótica [1, 2] e *call centers* [3]. Estudos na área de reconhecimento automático de emoções não são recentes, os primeiros trabalhos surgiram ainda na década de 1990, incluindo classificação de emoções através de expressões faciais ou voz. Na Figura 1.1, é apresentado o número de publicações relacionadas ao reconhecimento automático de emoções através da voz, indexadas no *ScienceDirect* entre os anos de 2000 e 2021, é possível observar um crescente número de trabalhos nessa área. Embora o número de trabalhos tenha crescido muito ao longo dos últimos anos, a área continua bastante desafiadora, pois ainda é necessário estabelecer relações bem definidas entre as emoções e as suas manifestações na voz, para que seja possível extrair as informações mais relevantes para a classificação automática [4].

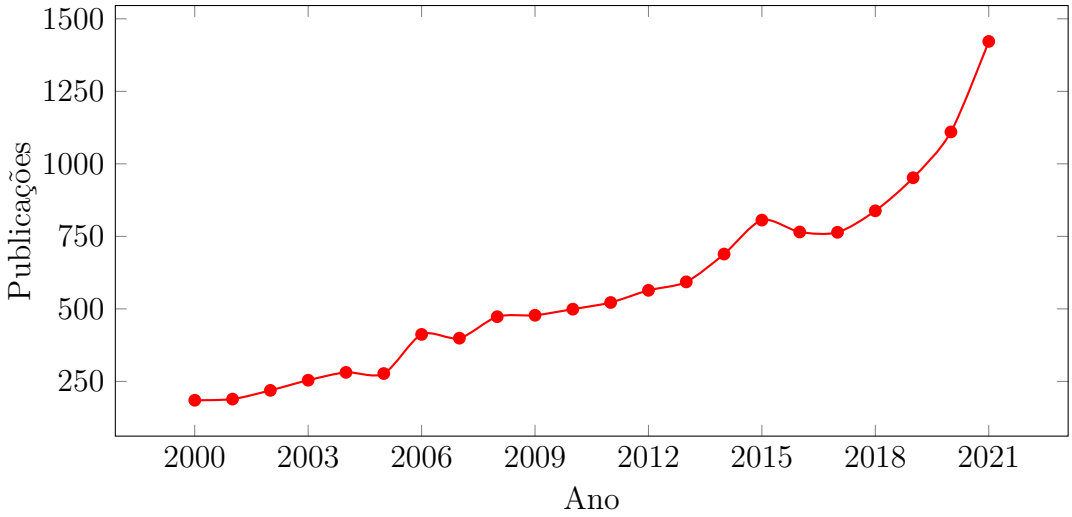


Figura 1.1 – Número de publicações relacionadas a reconhecimento de emoções através de sinais de voz indexadas no *ScienceDirect* entre os anos de 2000 e 2021.

A maior parte dos estudos relacionados as emoções humanas estão concentrados

na identificação, interpretação ou síntese das emoções. Essa área de pesquisa é conhecida como Computação Afetiva. Um dos temas bastante explorados é sobre a relação entre as emoções e suas manifestações na voz, face ou gestos corporais. Os autores da área presumem a existência de uma correspondência entre um estado emocional e determinadas expressões corporais [5]. Para a classificação de emoções através da voz, as propriedades vocais são profundamente analisadas, grande parte dos trabalhos relacionados baseia-se nos princípios da paralinguística, essa é uma área da linguística que estuda os aspectos não verbais associados a comunicação humana, estando interessada em *como* a informação foi transmitida [6].

A informação paralinguística transmitida através da voz está intimamente relacionada ao estado emocional [7], além de estar associada a outras características, como idade, gênero e algumas condições do trato vocal. Na voz, as características paralinguísticas são comumente associadas à prosódia, esse termo refere-se a uma classe de propriedades vocais relacionadas ao ritmo e entonação, essas propriedades são importantes na comunicação humana, pois são capazes de expressar algumas informações associadas ao contexto linguístico, no português, por exemplo, na seguinte frase interrogativa: *a apresentação é hoje?*, mudamos a entonação na palavra *hoje* para deixarmos claro para o ouvinte que se trata de uma pergunta. O mesmo tipo de variação acontece nas manifestações de emoções, por exemplo, quando estamos felizes, tendemos a falar de forma mais excitada do que quando estamos tristes. Dessa forma, é possível observar que a prosódia faz parte da comunicação humana e tem papel fundamental no auxilio da interpretação da informação [8].

Os aspectos da prosódia estão frequentemente associados ao ritmo e entonação, na linguística, esses aspectos são considerados características suprassegmentais, pois seus padrões são independentes das características segmentais (consoantes ou vogais) [9]. As propriedades suprassegmentais se relacionam com a percepção auditiva de altura, *pitch*, ênfase e duração de fonemas ou sílabas, sendo esses, representados por propriedades acústicas do sinal de fala, como frequência fundamental, amplitude e duração de intervalos na fala (intervalos com voz e sem voz) [10]. Essas características são comumente usadas para descrever a prosódia em sinais de voz. Além das propriedades relacionadas à prosódia, características relacionadas ao espectro também são usadas, essas características são capazes de descrever o comportamento do sinal de fala no domínio da frequência, técnicas como coeficientes mel-cepstrais ou coeficientes LFPC são amplamente usadas.

1.1 Problemática

Diversos autores apontam que um dos principais desafios na classificação automática de emoções através da voz é determinar propriedades relevantes no sinal de fala. Sabendo disso, formas alternativas de extrair informações têm sido estudadas. Neste trabalho, é

estudada uma técnica que realiza o mapeamento não-linear das características da voz para um outro espaço, essa metodologia funciona através de relações de similaridade (ou dissimilaridade) entre as observações do conjunto de dados. Essas relações são determinadas de acordo com um contexto semântico, previamente definido [11]. A partir dessas relações, é possível determinar um novo espaço vetorial para o conjunto de dados [12]. Esse novo espaço, na área de Processamento de Linguagem Natural, é conhecido como representação semântica, e representa os dados conforme o contexto semântico estabelecido [13].

Nas implementações mais comuns, os principais descritores acústicos refletem as propriedades de um curto período de tempo do sinal de voz, conhecido como janela. Devido a isso, não levam em consideração informações associadas ao contexto temporal, dessa forma, as relações de vizinhanças não são capturadas. Essas relações podem conter informações relevantes a respeito da emoção expressada no sinal de voz, como consequência natural, a acurácia no processo de classificação pode ser melhorada. As relações temporais podem ser levadas em consideração com a utilização da representação semântica.

1.2 Contribuições

Neste trabalho, é desenvolvida uma metodologia baseada na representação semântica adaptada para sinais de voz. A principal contribuição do trabalho é a elaboração de um tipo de contexto semântico baseado no TF-IDF, que é originalmente modificado para lidar com o problema de classificação de emoções através da voz.

Além disso, o trabalho apresenta uma análise sobre o efeito da representação semântica (usando contexto baseado em vizinhança; e outro contexto baseado no TF-IDF modificado) na classificação de emoções através de sinais voz. Até onde foi pesquisado, não há evidências sobre trabalhos que tenham realizado esse tipo de análise.

1.3 Objetivos

De forma geral, a principal finalidade do presente trabalho é analisar o sinal de voz de diferentes formas, a fim de identificar emoções através da voz, explorando conceitos fundamentais de pré-processamento, extração de características e classificação dos sinais de fala. Além disso, é avaliado o desempenho da representação semântica na tarefa de classificação de emoções usando sinais de voz, descrevendo aspectos relacionados ao contexto semântico e à otimização dos vetores dessa representação.

Os objetivos específicos desse trabalho são apresentados a seguir:

1. Implementar metodologias clássicas para classificação de emoções através de sinais

de voz, incluindo as etapas de pré-processamento, extração de características e classificação;

2. Investigar a utilização da representação semântica como um mapeamento não-linear das características acústicas, observando o efeito desse método na tarefa de classificação;

3. Analisar e discutir os resultados obtidos com as metodologias clássicas e a representação semântica, descrevendo o possível potencial da metodologia proposta.

1.4 Organização do trabalho

O presente trabalho está organizado em diferentes capítulos, inicialmente, no **Capítulo 2**, são discutidos aspectos teóricos relacionados as emoções humanas, abordando diferentes perspectivas, além das relações entre emoções e voz. Já no **Capítulo 3**, são descritas as principais metodologias para o processamento do sinal de voz, incluindo pré-processamento e extração de características. No **Capítulo 4**, são apresentadas as técnicas de representação semântica para o sinal de voz, discutindo quantização, contexto semântico e otimização dos vetores. Já no **Capítulo 5**, é apresentada uma breve descrição sobre alguns dos principais classificadores utilizados em sistemas de reconhecimento de emoções. No **Capítulo 6** é apresentada uma revisão bibliográfica a respeito de reconhecimento automático de emoções através da voz. Já no **Capítulo 7**, é mostrada a metodologia a ser utilizada nesse trabalho. Por fim, nos **Capítulos 8 e 9**, são apresentados respectivamente os resultados e discussões e a conclusão do trabalho.

Capítulo

2 As emoções

A emoção é considerada uma experiência mental de curta duração que não pode ser observado diretamente por outra pessoa, sendo particular à pessoa que a experimenta. Além disso, as emoções são parcialmente dependentes de eventos concretos, objetos e situações. Segundo o trabalho de Scherer [14], uma hipótese da motivação para termos emoções é para que seja possível saber como lidar em diferentes situações, com diferentes estímulos, tornando nosso comportamento mais adaptado e flexível. Neste capítulo, são apresentados aspectos teóricos sobre as emoções, além das relações entre emoção e voz.

2.1 Aspectos teóricos sobre as emoções

Diversos autores, em diferentes áreas do conhecimento (incluindo psicologia, biologia, neurociência, psiquiatria e filosofia) estudam as emoções, entretanto, ainda não existe um consenso entre os pesquisadores sobre determinados aspectos teóricos sobre as emoções, em especial, sobre o que causa as emoções, ou seja, o que acontece entre o estímulo (entrada) e a emoção (saída) [15, 16]. Dessa forma, diversas teorias foram desenvolvidas a fim de explicar a emoção. Historicamente, diferentes perspectivas foram apresentadas, três destas ganharam bastante notoriedade científica: (1) darwiniana, (2) jamesiana, e (3) cognitiva. Cada uma dessas define e explica a natureza das emoções de um ponto de vista diferente [17].

2.1.1 Perspectiva darwinista

A perspectiva darwinista tem origens nos trabalhos de Charles Darwin [18], essa perspectiva assume que as emoções, de certa forma, contribuíram para a nossa sobrevivência, possuindo um papel importante para o processo evolutivo. Sendo assim, segundo a teoria, é esperado notar as mesmas emoções em todos os seres humanos. Essa perspectiva tem sido bastante explorada, diversos pesquisadores contemporâneos tentam associar as emoções

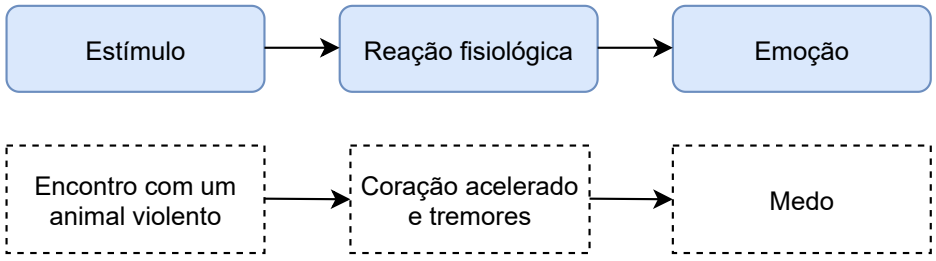
com aspectos evolutivos, principalmente, buscando encontrar um conjunto de emoções que são observáveis em todos os humanos.

Diferentes trabalhos seguiram com o objetivo de demonstrar que esse conjunto de emoções existe e que suas manifestações são universais, ou seja, todos os seres humanos expressam da mesma forma [19, 20]. Para esses autores, esse conjunto é composto pelas seguintes emoções: alegria, tristeza, medo, desgosto, raiva e surpresa. Segundo os autores, isso pôde ser comprovado observando as expressões faciais dos indivíduos estudados. Essas são consideradas por eles como emoções *fundamentais*, *primárias* ou *básicas*, além disso, de acordo com os autores, são a partir dessas que outras emoções são formadas [21].

Embora muitos autores concordem que existem emoções fundamentais, e que essas podem ser observadas de forma universal, muitas críticas têm sido feitas a esses trabalhos, especialmente pela falta de evidência da universalidade nas manifestações das emoções consideradas como básicas. De acordo com alguns autores, aspectos sociais e culturais podem influenciar bastante como os seres humanos expressam as emoções [22]. Os autores que defendem a influência dos aspectos culturais afirmam que as emoções devem ser experimentadas e expressadas de acordo com as regras sociais e culturais [21].

2.1.2 Perspectiva jamesiana

A perspectiva jamesiana foi proposta por William James [23], nessa teoria, a emoção é uma experiência consciente da resposta fisiológica a um determinado estímulo. Dessa forma, um estímulo provoca reações fisiológicas, e então, a emoção é a resposta à essas reações. Nessa perspectiva, não seria possível haver emoção sem que as reações fisiológicas acontecessem primeiro. É importante destacar a relação dessa teoria com a perspectiva darwiniana: para James, as emoções ocorrem pois nosso corpo responde automaticamente e adaptadamente ao ambiente, sendo assim, possuindo um importante significado evolutivo [21]. Na Figura 2.1, é apresentada a ordem dos eventos para a emoção ocorrer, seguido de um exemplo.



(Adaptado de Houwer e Hermans [15])

Figura 2.1 – Ordem de eventos para acontecer uma emoção, de acordo com a perspectiva jamesiana.

O trabalho de James foi revolucionário para a época, pois definiu uma nova ordem

de eventos para que a emoção aconteça, até então, acreditava-se que a emoção vinha antes das reações fisiológicas (*nós fugimos porque sentimos medo*). Entretanto, segundo a perspectiva jamesiana, as respostas fisiológicas vêm antes da experiência emocional (*nós sentimos medo porque fugimos*). Essa teoria foi bastante criticada, especialmente por reduzir as emoções a somente experiências a reações fisiológicas [15]. Diversos autores propõem que existe um fator cognitivo associado as emoções, que atribui algum sentido as reações fisiológica [24].

2.1.3 Perspectiva cognitiva

A perspectiva cognitiva parte do princípio que pensamentos e emoções são inseparáveis, desse ponto de vista, as emoções são dependentes de um processo de *avaliação* dos estímulos recebidos. Esse processo permite que os estímulos sejam interpretados de alguma forma, sendo assim, para cada emoção existe um determinado padrão de avaliação. De acordo com os pesquisadores, esse processo de avaliação é não deliberativo, ou seja, não há uma decisão consciente sobre qual será a avaliação encaminhada para cada estímulo, sendo esse processo classificado como: direto, imediato, não intelectual e automático [19, 25].

Uma teoria foi proposta por Schachter [26], para este, inicialmente o estímulo produz uma a excitação fisiológica, essa excitação é então interpretada e associada à alguma causa, e é esse processo de interpretação da excitação que constitui a emoção. Dessa forma, existe um processo cognitivo relacionado à excitação fisiológica, por exemplo, um confronto com um animal violento e um encontro com uma pessoa amada podem produzir reações fisiológicas similares no início, é só depois da associação dessa reação com o perigo (no caso do animal selvagem) ou com a felicidade (no outro caso) que as emoções de medo ou de alegria são formadas [15].

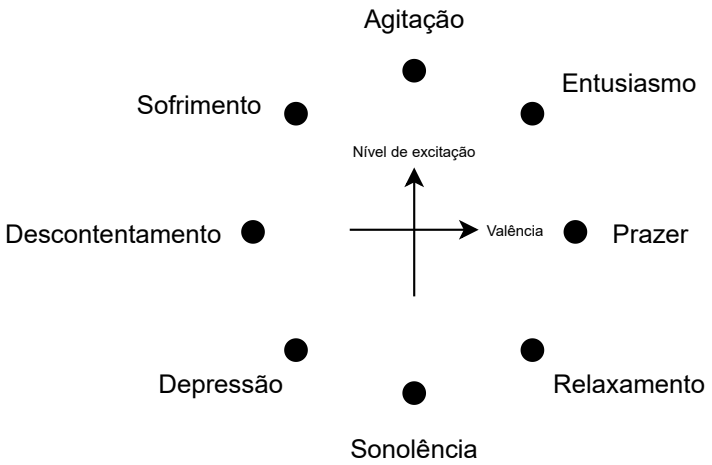
2.2 Relações entre emoção e voz

Para que seja possível identificar automaticamente as emoções a partir de sinais de voz, é preciso estabelecer uma conexão entre as emoções e a voz humana. A voz é uma das formas pela qual é possível expressar emoções, além disso, a voz também transmite diversas informações a respeito do orador, como o tamanho do trato vocal, sexo, idade e origem geográfica, tudo isso em paralelo ao conteúdo linguístico. Diferentes pesquisas mostram que as emoções fundamentais (alegria, tristeza, medo, desgosto, raiva e surpresa [20]) podem ser demonstradas por meio de expressões faciais e pela voz [27, 28].

Diversos autores têm estudado propriedades acústicas no sinal de voz, a fim de determinar padrões que estejam associados a estados emocionais. Ainda não existe um único conjunto de características acústicas comumente aceito e utilizado, entretanto, algumas

propriedades são amplamente utilizadas e exploradas, entre elas, estão as informações associadas à prosódia, espectro de frequências, intensidade e energia, essas se mostraram bastante importantes, pois são capazes de relacionar a voz com determinadas emoções [29, 27, 10].

Alguns trabalhos apresentam um modelo dimensional para as emoções, um dos modelos mais utilizados é baseado em duas dimensões: *excitação* e *avaliação* [30]. A excitação descreve o nível de excitação do estado emocional, por exemplo, as emoções de medo, tensão e raiva podem provocar alta excitação, enquanto as emoções associadas ao cansaço e sonolência provocam baixa excitação. Já a avaliação (ou valência), descreve se a emoção é positiva (por exemplo, felicidade e contentamento) ou negativa (por exemplo, tristeza, estresse e medo). Esse modelo é apresentado na Figura 2.2.



(Adaptado de Russell [30])

Figura 2.2 – Oito emoções apresentadas no modelo bidimensional.

Existe ainda um modelo tridimensional. Neste, além da excitação e avaliação, é incluída também a *dominância*. Essa terceira dimensão é adicionada para distinguir, por exemplo, entre medo e raiva, visto que, ambas as emoções possuem alta excitação e valência negativa. Segundo os autores desse modelo, a dominância é uma forma de caracterizar como lidamos com a situação, podendo ser de forma dominante (como durante a raiva), ou de forma submissa (como durante o medo). Esse modelo ficou conhecido como PAD (Prazer, Excitação e Dominância) [31, 32].

Estes modelos ajudam a diferenciar as emoções e relacioná-las a determinadas características acústicas. Vários trabalhos mostram que há uma forte correlação entre excitação e intensidade do sinal de voz, um exemplo disso é que em emoções com alta excitação (como a raiva), é observada uma alta intensidade no sinal de voz. O contrário acontece em emoções de baixa excitação (como a tristeza), essas apresentam normalmente baixa intensidade [10]. Diversas outras correlações têm sido investigadas a fim de tornar possível o reconhecimento automático de emoções por meio de sinais de voz, em todos

estes trabalhos as propriedades acústicas da voz são bastante exploradas, as principais propriedades são apresentadas no próximo capítulo.

Capítulo

3 Análise do sinal de voz

Neste capítulo, serão apresentadas as principais etapas para a análise do sinal de voz, cobrindo os seguintes tópicos: (1) uma breve introdução sobre os sinais de voz, (2) pré-processamento, e (3) extração de características. Inicialmente, são apresentados conceitos fundamentais sobre a produção vocal e o sinal digital de voz. Também são discutidos os métodos de pré-processamento do sinal de voz, descrevendo técnicas de janelamento e detecção de atividade vocal. Em seguida, são apresentadas as características acústicas amplamente utilizadas por pesquisadores da área.

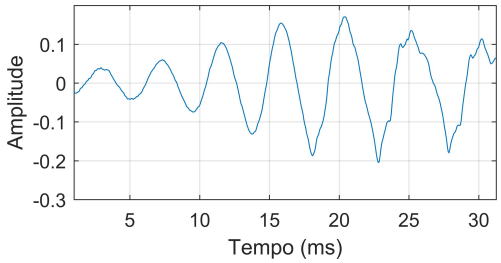
3.1 Introdução sobre o sinal de voz

Um sinal de voz é produzido através da atividade conjunta de diferentes elementos da anatomia humana, como pulmões, traqueia, laringe, faringe, cavidade oral e cavidade nasal. Inicialmente, o ar vindo dos pulmões passam pela traqueia, em seguida, passa pela laringe e entra na cavidade oral. Enquanto passa pela laringe, o ar é modulado pelas cordas vocais, formando ondas acústicas periódicas (mais precisamente, quase periódicas), que são posteriormente modificadas quando passam pela boca e nariz. Para modular o ar, as cordas vocais vibram em determinada frequência, comumente, essa frequência é definida como a frequência fundamental da voz [33].

Para que esse sinal de voz seja explorado computacionalmente, é preciso transformá-lo em um tipo de sinal que o computador consiga analisar. Isso é feito através da conversão do sinal de voz analógico para um sinal digital. O sinal analógico pode ser interpretado como o som da voz que sai da boca, que varia continuamente no tempo [34].

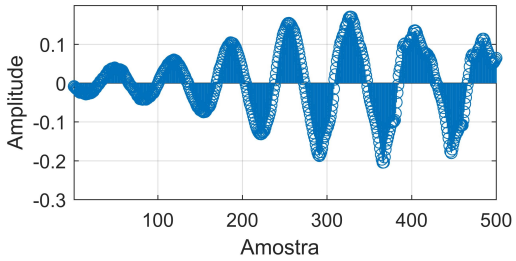
Já o sinal digital é uma estimação do sinal analógico, que pode ser processado computacionalmente. O sinal digital representa um sinal analógico através de uma sequência de números quantizados, essa sequência é formada por amostras do sinal analógico que são, normalmente, capturadas regularmente com uma determinada periodicidade. Na Figura 3.1, é apresentada uma ilustração de um sinal de voz contínuo no tempo, já na Figura 3.2,

é mostrada uma ilustração do sinal amostrado. Na Figura 3.3, é apresentado a mesma ilustração do sinal de voz amostrado, agora entre as amostras 1 e 100, é possível observar com mais clareza as amostras.



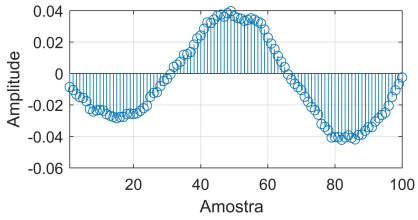
(Fonte: Banco de vozes emotivas Emo-DB [35])

Figura 3.1 – Sinal de voz ilustrado como contínuo no tempo.



(Fonte: Banco de vozes emotivas Emo-DB [35])

Figura 3.2 – Sinal de voz ilustrado como amostrado.



(Fonte: Banco de vozes emotivas)

Figura 3.3 – Sinal de voz ilustrado como amostrado, entre as amostras 1 e 100.

Neste trabalho, o sinal de de voz, denotado como $s(n)$, é sempre digital, com N amostras, onde $0 \leq n \leq N-1$.

3.2 Pré-processamento do sinal de voz

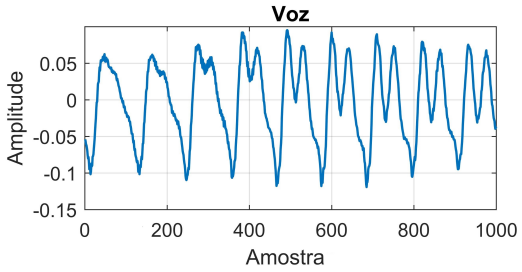
3.2.1 Janelamento

Devido à lenta variação em sinais de voz, é bastante comum se processar o sinal de voz em pequenos segmentos de áudio (janelas de áudio), dessa forma, é possível assumir que as propriedades da forma de onda do sinal se mantêm constantes nestes blocos [34]. Esse processo é útil para a extração de informações em cada segmento. Normalmente, são utilizadas janelas com duração entre 20 e 40 milissegundos, além disso, pode haver sobreposição entre as janelas [29].

3.2.2 Detecção de atividade vocal

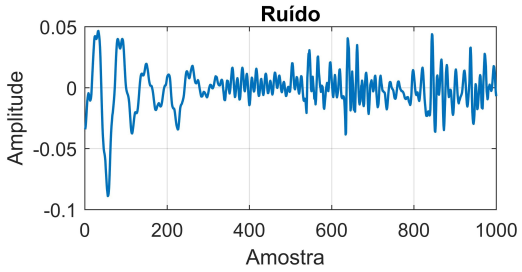
A partir de uma determinada janela de áudio, é necessário decidir se o segmento possui voz ou não. Isso é extremamente importante para a identificação das emoções, uma vez que, intervalos do sinal sem voz (como o silêncio e o ruído) podem estar presentes nas locuções de diferentes emoções, caso essas partes não sejam removidas, são causados sérios prejuízos à classificação. A ideia central das metodologias de detecção de atividade vocal é selecionar características capazes de distinguir entre voz e o ruído [36].

A detecção de atividade vocal pode ser feita de diferentes formas, é muito comum que sejam utilizadas técnicas baseadas em: (1) energia, (2) taxa de cruzamento por zero e (3) autocorrelação [34, 37, 29]. Estes três métodos são simples de implementar, além disso, são eficientes para diferenciar entre segmentos com voz e sem voz. Nas Figuras 3.4 e 3.5, são apresentadas, respectivamente, uma ilustração de segmento de sinal com voz e uma ilustração de segmento sem voz, que chamaremos de ruído. É possível perceber diversas características em cada sinal, principalmente sobre a sua periodicidade e amplitude. A seguir, é discutido como estas características podem ser determinadas.



(Fonte: Banco de vozes emotivas Emo-DB [35])

Figura 3.4 – Ilustração de segmento com voz.



(Fonte: Banco de vozes emotivas Emo-DB [35])

Figura 3.5 – Ilustração de segmento com ruído.

3.2.2.1 Energia

A energia, como definida aqui, corresponde à soma de todas as amostras elevadas ao quadrado. Em muitos casos, segmentos de áudio com voz apresentam alta amplitude (observe a Figura 3.4) quando comparadas a partes sem áudio (Figura 3.5), isso faz com que a energia de segmentos com voz seja maior do que em segmentos sem voz.

3.2.2.2 Taxa de cruzamento por zero

A taxa de cruzamento por zero (também conhecido como ZCR, do inglês *zero-crossing rate*) determina a taxa que o sinal de voz $s(n)$ muda de sinal algébrico durante a janela. É comum que segmentos de áudio com ruído variem de forma mais rápida do que o

sinal de voz, isso faz com que o sinal cruze o zero mais vezes (observe a Figura 3.5). Já em segmento de áudio com voz, é comum que a taxa de cruzamento por zero seja menor, como mostrado na Figura 3.4. Dessa forma, a partir da determinação da taxa de cruzamento por zero, é possível estimar se existe voz ou silêncio no segmento de áudio [38].

3.2.2.3 Autocorrelação

A autocorrelação determina o quão semelhante o sinal é dele mesmo, comparando um sinal com diferentes intervalos de tempo do mesmo sinal. O resultado é uma medida de similaridade entre um sinal $s(n)$ em função do atraso, dessa forma, é possível medir o nível de semelhança entre o sinal $s(n)$ e o mesmo sinal deslocado dm amostras $s(n + dm)$, como é mostrado na na Equação 3.1 [39, 40].

$$R_{ss}(dm) = \sum_{n=0}^{N-1-dm} s(n)s(n + dm) \tag{3.1}$$

Em que N é o comprimento do sinal $s(n)$, e $dm = 0, 1, 2, ..., max_{dm}$, em que max_{dm} é o atraso máximo associado à correlação. É importante observar que quando $dm = 0$, a correlação é máxima. Além disso, caso $s(n)$ seja aproximadamente periódico, ao se aumentar o valor de dm até a metade do período da onda, a correlação tende a diminuir. Já quando dm cresce novamente até o período da onda, a correlação tende a aumentar. Logo, o primeiro pico na autocorrelação R_{ss} pode indicar o período da onda [40].

É a partir da função de autocorrelação $R_{ss}(dm)$ que é possível se estimar a frequência fundamental no segmento de voz $s(n)$, caso ele seja periódico, ou o nível de semi-periodicidade desse segmento [41]. Em segmentos de voz (vide Figura 3.4) existem padrões de onda que se repetem de forma quase periódica, diferente de segmentos sem voz (Figura 3.5), em que não há padrões que se repetem. Por isso a função de autocorrelação é útil para distinguir entre segmentos de voz e sem voz.

3.2.2.4 Outras técnicas

A energia, a taxa de cruzamento por zero e a autocorrelação são técnicas clássicas para a detecção de atividade vocal, entretanto, diversas outras informações podem ser extraídas, como a relação sinal-ruído ou coeficientes de predição linear. Além disso, é comum que sejam utilizadas metodologias de reconhecimento de padrões, essas funcionam extraindo diversas características do sinal, a partir dessas informações, é elaborado um modelo capaz de distinguir entre segmentos com voz e sem voz [36, 42].

3.3 Extração de características acústicas

Uma importante parte do processo de de reconhecimento automático de emoções é a extração de características adequadas do sinal de voz. Na literatura, as características associadas à prosódia e ao espectro são as mais utilizadas [43]. É muito comum que essas características sejam extraídas de cada segmento de áudio (como discutido na seção sobre janelamento 3.2.1).

3.3.1 Prosódia

A prosódia pode ser definida como a análise de propriedades não verbais da comunicação humana pela voz, como o ritmo e a entonação. Essas propriedades podem ser utilizadas para expressar emoções, atitudes e interesse, por exemplo. Além disso, a prosódia é muito importante para a organização linguística, inserindo pausas em determinados momentos da fala, e ainda para dar sentido às orações [44, 29].

Os seguintes indicadores acústicos são comumente associadas à prosódia: frequência fundamental, energia, além de propriedades relacionadas à duração dos momentos com voz e de momentos de silêncio [43]. Neste trabalho, também são consideradas as variações na frequência fundamental (*jitter*) e variações nas amplitudes de cada ciclo (*shimmer*) presentes em um segmento de voz. A seguir, são descritos alguns desses indicadores.

3.3.1.1 Frequência fundamental

A frequência fundamental ($f0$) é uma propriedade da onda sonora, é o inverso do período fundamental (duração de um ciclo completo da onda, como mostrado na Figura 3.6). A $f0$ está associada ao termo *pitch*, que pode ser entendido como a altura (frequência) percebida em um sinal de áudio, sendo essa uma característica subjetiva associada a percepção humana do som. A $f0$ pode ser estimada no domínio do tempo ou da frequência, diversas técnicas podem ser utilizadas, um método bastante conhecido é o da autocorrelação (como mostrado na Equação 3.1) [44].

A partir da função de autocorrelação $R_{ss}(m)$, é possível se calcular a $f0$ [44]. A $f0$ é um importante parâmetro pra indicar a entonação presente no segmento de voz. No contexto emocional, emoções com baixa excitação (como a tristeza) tendem a ter $f0$ mais baixos do que emoções com alta excitação (como raiva ou alegria) [10].

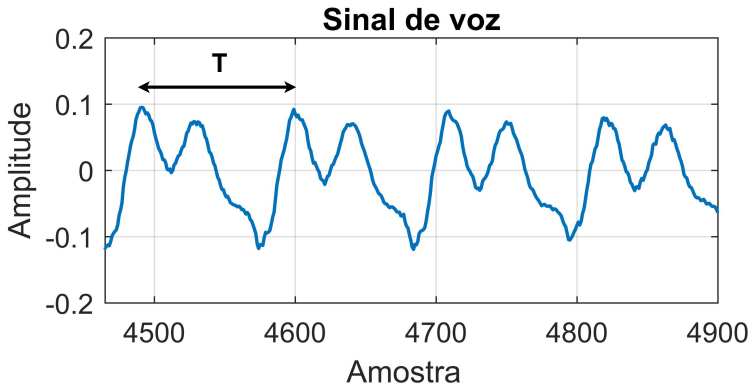


Figura 3.6 – Ciclos de um segmento de voz. O período de um ciclo é indicado por T.

3.3.1.2 *Jitter* e *shimmer*

O *jitter* e o *shimmer* são comumente utilizados como indicadores de qualidade vocal, sendo bastante utilizados na identificação de vozes patológicas [45, 46] e para verificação e reconhecimento de orador [47, 48]. Partindo do princípio que as emoções devem alterar aspectos associados à qualidade vocal [49], diversos trabalhos também utilizaram estes parâmetros para distinguir entre diferentes emoções em sinais de voz [29, 49, 50, 51].

É conhecido que o sinal de voz não é inteiramente periódico, existem alguns tipos de perturbações na forma de onda desse sinal. Variações na frequência fundamental são chamadas de *jitter*, já variações na amplitude da onda são conhecidas como *shimmer* [52]. O *jitter* mede o nível de variação entre sucessivos períodos T (como mostrado na Figura 3.6), dessa forma, é possível observar o quanto o período está variando ao longo do tempo.

O *jitter* pode ser calculado somando todas as diferenças entre períodos consecutivos, dividindo pelo número de ciclos, e em seguida, dividindo pela média dos períodos dos ciclos presentes na janela observada [48, 52]. Já o *shimmer*, mede o nível de variação entre as amplitudes pico a pico de ciclos consecutivos, como mostrado na Figura 3.7. De forma semelhante ao *jitter*, o *shimmer* pode ser calculado somando todas as diferenças entre as amplitudes pico a pico, dividindo pelo número de ciclos, e dividindo pela média das amplitudes dos ciclos presentes na janela observada.

3.3.1.3 Energia

A energia é uma das básicas e mais importantes medidas de sinais de voz. A utilização da energia como característica acústica para reconhecimento automático de emoções parte do principio que emoções com baixa excitação (como a tristeza e desgosto) tendem a ter energia baixa, ao contrário de emoções de alta excitação (como raiva e surpresa) [10]. Isso pode ser facilmente observado no cotidiano, quando estamos com raiva, é comum falar com maior intensidade, como forma de demonstrar nossa insatisfação com

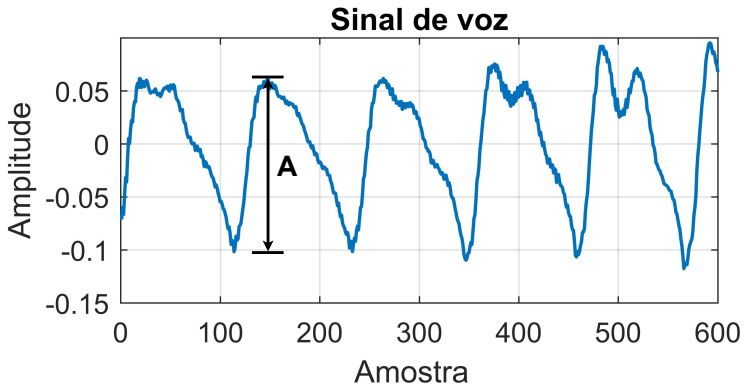


Figura 3.7 – Ciclos de um segmento de voz, com destaque para a amplitude pico a pico definida como A.

algo, já quando estamos tristes, a intensidade da voz é normalmente mais baixa. Diversos trabalhos já observaram isso [53, 54, 6]. Dessa forma, é possível, por exemplo, distinguir entre emoções de alta excitação e de baixa excitação utilizando a energia.

3.3.2 Características espectrais

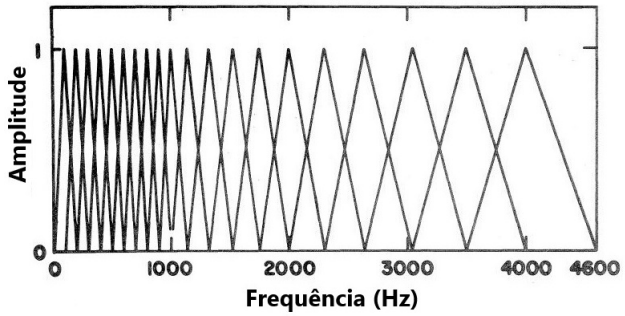
As características espectrais são baseadas no espectro de frequências do sinal de voz, esse espectro é obtido transformando o sinal do domínio do tempo para o domínio da frequência, comumente através da transformada rápida de Fourier (também conhecida como FFT, do inglês *fast Fourier transform*) [43]. Diversos extratores de características partem do espectro do sinal, sendo um dos mais populares em processamento de fala os coeficientes mel-cepstrais.

3.3.2.1 Coeficientes mel-cepstrais

Os coeficientes mel-cepstrais (também conhecidos como MFCCs, do inglês *mel-frequency cepstrum coefficients*) foram formulados por Davis e Mermelstein em 1980 [55], estes são uma das características mais conhecidas e utilizadas em processamento de voz. A ideia dos MFCCs é basicamente realizar uma análise espectral através de um banco de filtros passa-faixas. Esses filtros possuem uma interessante inspiração biológica baseada nas bandas críticas do sistema auditivo, descritas detalhadamente no trabalho de Fletcher [56].

No trabalho de Davis e Mermelstein [55], 20 filtros passa-bandas triangulares são definidos de modo que se assemelhem às bandas críticas, como mostrado na Figura 3.8. A escala mel é adotada para a criação dos filtros passa-faixas, assim, é possível que os filtros sejam distribuídos de forma não linear na frequência, semelhante às bandas críticas.

Ao longo do tempo, diferentes implementações para obter os MFCCs foram desenvolvidas, na implementação mais popular, os filtros possuem largura constante até a frequência de 1 kHz, a partir desse valor, a largura cresce de forma logarítmica [34], como pode ser observado na Figura 3.8.



(Fonte: Davis e Mermelstein [55].)

Figura 3.8 – Filtros passa-bandas para determinar os coeficientes mel-cepstrais.

Detalhes sobre a implementação original podem ser obtidos no trabalho de Davis e Mermelstein [55]. Neste trabalho, é utilizada uma implementação semelhante à proposta por Slaney [57], que utiliza 40 filtros com áreas iguais. Mais detalhes sobre a implementação usada no presente trabalho são apresentados na Seção 7.3.5, no Capítulo 7.

3.3.2.2 Coeficientes LFPC

Os coeficientes LFPC (*log frequency power coefficients*, coeficientes do logaritmo da potência, em uma possível tradução para o português) é uma técnica simples utilizada para analisar o espectro do sinal de voz. Nessa técnica, o espectro do sinal é delimitado por sub bandas, então, é calculada a energia do espectro dentro de cada sub banda. Na implementação proposta por Nwe et al. [58], as sub bandas são distribuídas de maneira logarítmica, a fim de simular a forma que os humanos percebem as frequências do som, de forma análoga ao proposto nos MFCCs.

Mais detalhes sobre os coeficientes LFPC podem ser obtidos no trabalho de Nwe et al. [58]. Neste trabalho, é usada a mesma implementação apresentada no trabalho de Nwe et al., como apresentado na Seção 7.3.6, no Capítulo 7.

Capítulo

4 Representação semântica

A representação semântica tem sido usada amplamente na área de Processamento de Linguagem Natural (NLP), sendo aplicada para determinar uma representação vetorial para palavras. De acordo com Rettig et al. [59], essa representação é capaz de descrever a semântica das palavras através de vetores numéricos, popularmente conhecidos como *word vectors* ou *word embbedings*. As principais metodologias baseiam-se na *hipótese de semântica estatística*: os padrões estatísticos das palavras utilizadas na comunicação humana podem ser usados para identificar o que as pessoas querem dizer [11].

Na área de NLP, foi observado que padrões estatísticos associados à frequência de palavras em contextos pré-definidos arbitrariamente podem ser muito relevantes para entender a semântica das palavras. Por exemplo, duas palavras que acontecem frequentemente em uma mesma frase devem estar relacionadas [60, 61]. Tais relações, nos problemas de processamento de linguagem natural, se convertem em relações semânticas entre as palavras, resultando em espaços vetoriais que as capturam, permitindo assim dar um tratamento numérico aos dados [62]. Neste sentido, ideias similares foram desenvolvidas para sinais de outras naturezas, como imagens [63].

Para a representação semântica, é necessário estabelecer relações de similaridade ou dissimilaridade entre todos os símbolos (e.g. palavras, letras, notas musicais, entre outros) $w_n (n = 1, 2, ..., V)$ presentes no conjunto de dados, sendo que V é o tamanho do vocabulário. Dessa forma, é determinada uma matriz $\mathbf{S}$, com dimensões VxV, que indica as relações de similaridade entre os símbolos. As relações estabelecidas em $\mathbf{S}$ podem ser determinadas de diferentes formas, variando de acordo com tipo de contexto semântico de interesse [11]. Em NLP, é comum a utilização de contexto baseado em janela de n palavras, de forma que, quanto mais a palavra w_i acontece no contexto da palavra w_j, maior será o grau de similaridade entre w_i e w_j [64, 12].

Utilizando métodos de otimização, é possível determinar vetores numéricos $\mathbf{g}_n (n = 1, 2, ..., V)$, definidos no $\mathbb{R}^{dim_{\mathbf{g}}}$ que mantenham as mesmas relações de similaridade ou dissimilaridade apresentadas em $\mathbf{S}$, de forma que a distância entre $\mathbf{g}_i$ e $\mathbf{g}_j$ seja aproximadamente igual à distância entre w_i e w_j estabelecida em $\mathbf{S}$ [65, 12]. Sendo assim, caso um

determinado símbolo (palavra, ou qualquer estado discreto) w_i seja, em algum contexto, similar a outro símbolo w_j, ambos estarão próximos na representação semântica gerada.

É provável que a principal utilidade da representação semântica para a classificação de dados esteja associada ao poder de definir relações entre as observações $\mathbf{x}_i$ através do contexto semântico utilizado. A hipótese levantada é que, dependendo do contexto escolhido em projeto (e arbitrariamente), é possível deformar o espaço de características, inclusive para que as classes sejam melhor discriminadas, que é um dos objetivos deste trabalho. Nesta seção, serão apresentadas metodologias para a utilização da representação semântica em sinais de fala, descrevendo a quantização vetorial, e em seguida dois tipos de contextos semântico que podem ser utilizados.

4.1 Quantização vetorial

A representação semântica é aplicada em conjuntos finitos, como sequências de símbolos, palavras, caracteres, entre outros. Entretanto, o sinal de voz é contínuo, dessa forma, é preciso quantizar o sinal para que as técnicas de representação semântica sejam aplicadas. Para tornar o sinal discreto, é possível aplicar metodologias de quantização vetorial, essas são responsáveis por mapear cada vetor de características a um determinado símbolo.

Uma das formas de realizar a quantização vetorial é através do método conhecido como *k-means* [66]. Nessa técnica, é obtido um conjunto de k centroides que se adaptem aos dados. Após isso, cada vetor de características é associado ao centroide mais próximos. Para cada vetor de características $\mathbf{x}$, é calculada a distancia euclidiana para o centroide $\boldsymbol{\mu}_n (n = 1, 2, ..., k)$, o centroide que tiver menor distância é associado ao vetor, como mostrado na Equação (4.1). Dessa forma, cada vetor $\mathbf{x}$ é associado a um símbolo $w_n (n = 1, 2, ..., k)$ mais próximo.

$$w_n = \underset{n}{\operatorname{argmin}} \left(\sqrt{(\mathbf{x} - \boldsymbol{\mu}_n)^\top (\mathbf{x} - \boldsymbol{\mu}_n)} \right) \tag{4.1}$$

4.2 Contexto semântico

Após a quantização vetorial, uma etapa de definição do contexto semântico é necessária para determinar as relações entre os símbolos, essas relações podem ser definidas de diversas formas, e resultam em uma matriz $\mathbf{S}_{V \mathrm{x} V}$ de similaridade entre os V símbolos no vocabulário. Nesta seção, duas técnicas para a definição do contexto semântico são apresentadas: (1) contexto baseado em vizinhança, e (2) contexto baseado no TF-IDF.

Antes de apresentar os aspectos teóricos sobre os tipos de contextos semântico utilizados nesse trabalho, é importante destacar os seguintes pontos:

1. A representação natural corresponde à representação original do vetor de características $\mathbf{x}$, definido em $\mathbb{R}^{dim_{\mathbf{x}}}$.
2. A representação natural quantizada é obtida através da técnica de agrupamento *k-means*, com k níveis de quantização. Sendo assim, um vetor de características $\mathbf{x}_i$ é associado a um determinado símbolo w_n, como mostrado na Equação 4.1.
3. É usado $w_n(n = 1, 2, ..., V)$ para representar o vocabulário, ou seja, todos os símbolos possíveis. Cada elemento de w é único.

4.2.1 Contexto baseado em vizinhança

O contexto baseado em vizinhança é bastante utilizado em NLP para associar vetores a palavras, mantendo as relações semânticas entre as palavras [67, 68, 12]. Esse contexto é baseado na ideia da *hipótese distribucional* [69]. Nessa hipótese, a linguagem é descrita através da ocorrência de palavras no contexto de outras, de forma que, palavras que ocorrem no mesmo contexto tendem a ter maiores relações semânticas ou sintáticas [69, 13].

O contexto baseado em vizinhança funciona da seguinte forma: dada uma palavra w_n, o seu contexto pode ser definido como os Nv símbolos vizinhos, esses vizinhos podem ser igualmente divididos entre antecedentes e subsequentes, formando uma janela simétrica (vide Figura 4.1). A janela também pode ser assimétrica, com mais símbolos em uma determinada direção do que em outra, como mostrado na Figura 4.1, em que w_n é o símbolo central, e os símbolos vizinhos são considerados como contexto.

Figura 4.1 – Contexto de vizinhança simétrico e assimétrico.

Para esse contexto semântico, a matriz de similaridade $\mathbf{S}$ torna-se uma matriz de coocorrência entre os símbolos. Sendo assim, $\mathbf{S}_{ij}$ indica quantas vezes o símbolo w_i acontece dentro do contexto de w_j, ou seja, quantas vezes w_i aparece na vizinhança de w_j. Por exemplo, dada a seguinte frase: *João ganhou uma bicicleta azul, mas João não queria uma bicicleta azul*, e considerando o contexto como a primeira palavra que ocorre antes e depois, é obtido a matriz de coocorrências apresentada na Figura 4.2.

	João	ganhou	uma	bicicleta	azul	mas	não	queria
João	0	1	0	0	0	1	1	0
ganhou	1	0	1	0	0	0	0	0
uma	0	1	0	2	0	0	0	1
bicicleta	0	0	2	0	2	0	0	0
azul	0	0	0	2	0	1	0	0
mas	1	0	0	0	1	0	0	0
não	1	0	0	0	0	0	0	1
queria	0	0	1	0	0	0	1	0

João ganhou uma bicicleta azul, mas João não queria uma bicicleta azul

Figura 4.2 – Matriz de coocorrências para o conjunto de palavras da frase *João ganhou uma bicicleta azul, mas João não queria uma bicicleta azul*, escolhendo a primeira palavra (antes e depois) como vizinha.

Esse forma de construir a matriz **S** foi utilizada no trabalho de Pennington et al. [12] para determinar representação vetorial para palavras, obtendo um método bastante utilizado e conhecido como *GloVe (Global Vectors)*. De acordo com os autores, o tamanho e a forma (simétrica ou assimétrica) da janela influenciam na representação semântica das palavras. Janelas simétricas e pequenas são melhores para capturar a informação sintática, já janelas grandes e simétricas são mais adequadas para descrever informações semânticas.

No âmbito de NLP, a informação semântica está relacionada ao significado das palavras, já a informação sintática diz respeito as regras de construção das frases. Não é possível associar diretamente esses conceitos à analise de sinais de voz, entretanto, assim como um texto, o sinal de voz também é sequencial e existem estruturas temporais complexas incorporadas ao mesmo. São essas estruturas que o contexto baseado em vizinhança tenta capturar.

4.2.2 Contexto baseado no TF-IDF

O TF-IDF (*term frequency - inverse document frequency*, em português, frequência do termo - inverso da frequência nos documentos) é uma técnica utilizada para determinar uma medida de ocorrência para símbolos. Nessa técnica, a frequência de um símbolo, w_n, em um determinado documento, $do(i), i = 1, 2, ..., \mathcal{L}$, é reduzida por um fator que leva em consideração a ocorrência de w_n em outros documentos, como mostrado na Equação 4.2, em que $f_{w_n,do(i)}$ representa a frequência de w_n em $do(i)$, $\mathcal{L}$ é o número total de documentos, e nd_{w_n} é o número de documentos em que w_n ocorre [6, 70].

$$TFIDF_{w_n,do(i)} = f_{w_n,do(i)} \log\Big(\frac{\mathcal{L}}{nd_{w_n}}\Big) \tag{4.2}$$

Na Equação 4.2, $f_{w_n,do(i)}$ corresponde ao TF (frequência do termo), e $\log(\frac{\mathcal{L}}{nd_{w_n}})$ corresponde ao IDF (frequência inversa do documento). É possível observar que $TFIDF_{w_n,do(i)}$ é inversamente proporcional a nd_{w_n}. Dessa forma, em quanto mais documento w_n ocorrer simultaneamente, menor é $TFIDF_{w_n,do(i)}$.

A Equação 4.2 apresenta a ideia geral do TF-IDF. Para a utilização desse princípio para a classificação de emoções através da voz, alguns detalhes podem ser incorporados. Inicialmente, os documentos se tornam as classes, sendo assim, é definido $TFIDF_{w_n,c(i)}$, em que $c(i)$ representa a i-ésima classe, então, $TFIDF_{w_n,c(i)}$ corresponde a uma medida de ocorrência do símbolo w_n na classe $c(i)$.

Além disso, é possível utilizar uma forma diferente para determinar o IDF, neste trabalho, o termo $\log\left(\frac{\mathcal{L}}{nd_{w_n}}\right)$ é substituído por $\frac{1}{\log(f_{w_n,*}+\gamma)}$, em que $f_{w_n,*}$ representa o número de ocorrências de w_n em todas as classes, ou seja, caso w_n ocorra 100 vezes ao total (considerando todas as classes), $f_{w_n,*} = 100$. Já o termo γ é usada para evitar problemas de divergência com o logaritmo quando o $f_{w_n,*}$ é zero, nesse trabalho, é usado $\gamma = 2$.

A mudança proposta nesse trabalho não altera a ideia geral do TF-IDF, entretanto, com a modificação, quanto mais vezes w_n ocorrer fora da classe $c(i)$, menor é $TFIDF_{w_n,c(i)}$. As alterações sugeridas são apresentadas na Equação 4.3.

$$TFIDF_{w_n,c(i)} = f_{w_n,c(i)} \frac{1}{\log(f_{w_n,*} + \gamma)} \tag{4.3}$$

A principal motivação para essa alteração está relacionada ao efeito da ocorrência de um símbolo w_n fora da classe $c(i)$. Note que no TF-IDF tradicional (Equação 4.2) não importa quantas vezes o símbolo w_n ocorre fora do documento $do(i)$, a equação apenas contabiliza em quantos documentos w_n ocorre. Por outro lado, no modificação proposta (Equação 4.3), a ocorrência de w_n em todas as classes ($f_{w,*}$) é levada em consideração, assim, é possível ter uma medida mais fina sobre relevância de w_n dentro da classe $c(i)$.

Sendo assim, para cada classe emotiva $c(i)(i = 1, 2, ..., C)$, em que C é o número total de classes, são determinados os Ns símbolos com maiores TF-IDF (calculados a partir da Equação 4.3). Assim, é possível definir uma matriz $\mathbf{B} \in \mathbb{N}^{C \times Ns}$ (Equação 4.4), em que cada linha de $\mathbf{B}$ indica classe, e as colunas representam os símbolos com maiores TF-IDF. Portanto, a linha c da matriz $\mathbf{B}$ apresenta Ns símbolos com maiores TF-IDF para a classe c. Dessa forma, é possível considerar somente os símbolos mais representativos de cada classe.

$$\mathbf{B} = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,Ns} \\ \vdots & \vdots & \ddots & \\ w_{C,1} & w_{C,2} & \dots & w_{C,Ns} \end{bmatrix} \tag{4.4}$$

A partir de $\mathbf{B}$, é construída a matriz simétrica de similaridade $\mathbf{S}$ entre cada símbolo $w \in \mathbf{B}$. A matriz $\mathbf{S}$ é inicializada com valores zero, em que S_{ij} significa o quão semelhante o símbolo w_i é do símbolo w_j.

Supondo dois símbolos pertencentes a matriz $\mathbf{B}$ e que ocorrem na mesma classe: $w_{c,1}$ e $w_{c,2}$, é determinado que $S_{w_{c,1},w_{c,2}} = \frac{1}{n_{w_{c,1}} n_{w_{c,2}}}$, em que $n_{w_{c,1}}$ representa o número de classes em que o símbolo $w_{c,1}$ acontece. Esta medida de semelhança tem os seguintes comportamentos extremos, quando $w_{c,1}$ e $w_{c,2}$ acontecem na mesma classe:

1. No caso dos símbolos $w_{c,1}$ e $w_{c,2}$ serem exclusivos de apenas uma classe, cada, então $n_{w_{c,1}} = n_{w_{c,2}} = 1$, dessa forma, a semelhança $S_{w_{c,1},w_{c,2}}$ é máxima, igual a 1;
2. Por outro lado, se os símbolos são compartilhados por todas as classes, então $n_{w_{c,1}} = n_{w_{c,2}} = C$, e a semelhança entre esses símbolos será igual a $\frac{1}{C^2}$.

Além disso, caso $w_{c,1}$ e $w_{c,2}$ não aconteçam na mesma classe, não existe semelhança entre os símbolos, dessa forma, $S_{w_{c,1},w_{c,2}} = 0$.

Esse tipo de contexto é capaz de atribuir propriedades similares para observações distantes entre si (na representação natural), porém pertencentes a uma mesma classe, e com alta relevância dentro da classe (devido ao TF-IDF). Até onde foi pesquisado, não há evidências da construção da matriz similaridade $\mathbf{S}$ usando a metodologia descrita, o que torna essa estratégia de construção uma contribuição original desse trabalho.

4.3 Otimização

Para ambas as formas de contexto definidas anteriormente, é necessária a etapa de otimização para obter a representação semântica. Inicialmente, a matriz de similaridade $\mathbf{S}$ é convertida em uma matriz de dissimilaridade $\mathbf{D}$ (com as mesmas dimensões de $\mathbf{S}$). A conversão é feita como mostrado em 4.5. É utilizado o logaritmo para reduzir a variação de $\mathbf{S}$, alguns elementos de $\mathbf{S}$ podem ser grandes, enquanto outros, muito pequenos.

$$\mathbf{D} = -log(\mathbf{S} + 1) \tag{4.5}$$

A matriz $\mathbf{S}$ é somada a um para evitar problemas de divergência com o logaritmo, quando o argumento for zero. Além disso, $\mathbf{D}$ é normalizada como, $\mathbf{D} = \frac{(\mathbf{D} - \min(\mathbf{D}))}{\max(\mathbf{D}) - \min(\mathbf{D})}$, para que seu valor mínimo seja 0 e o máximo seja 1.

Após obter a matriz de dissimilaridade $\mathbf{D}$, são determinados os vetores $\mathbf{g}_n (n = 1, 2, ..., V)$ () que representam os símbolos $w_n (n = 1, 2, ..., V)$. Os vetores $\mathbf{g}_n \in \mathbb{R}^{dim_{\mathbf{g}}}$, e

sua dimensão é arbitrária. Esses vetores são obtidos através de técnicas de otimização que reduzam a função de custo apresentada na Equação 4.6.

$$J(\mathbf{Q}, \mathbf{D}) = \sum_{ij}^{V} (Q_{ij} - D_{ij})^2 \tag{4.6}$$

Sendo que, Q_{ij} representa as distâncias euclidianas entre os vetores $\mathbf{g}_i$ e $\mathbf{g}_j$, como mostrado na Equação (5.9).

$$Q_{ij} = \sqrt{(\mathbf{g}_i - \mathbf{g}_j)^\top (\mathbf{g}_i - \mathbf{g}_j)} \tag{4.7}$$

Os vetores $\mathbf{g}_i$ são inicializados de forma aleatória, com valores entre 0 e 1 com distribuição uniforme. A redução da diferença entre $\mathbf{Q}$ e $\mathbf{D}$ é um problema de otimização, em que $\mathbf{Q}$ é alterado através de adaptações iterativas em $\mathbf{g}_i$. O método do gradiente descendente pode ser usado para realizar as adaptações, como mostrado em (4.8).

$$\mathbf{g}_i \leftarrow \mathbf{g}_i - \alpha \nabla_{\mathbf{g}_i} J(\mathbf{Q}, \mathbf{D}) \tag{4.8}$$

Em que α é o passo de adaptação (ou taxa de aprendizado), e $\nabla_{\mathbf{g}_i} J(\mathbf{Q}, \mathbf{D})$ é o gradiente da função de custo com relação a $\mathbf{g}_i$, podendo ser calculado como apresentado em (4.9).

$$\nabla_{\mathbf{g}_i} J(\mathbf{Q}, \mathbf{D}) = 2 \sum_{j \neq i}^{V} (Q_{ij} - D_{ij}) \frac{(\mathbf{g}_i - \mathbf{g}_j)}{\sqrt{Q_{ij}}} \tag{4.9}$$

Capítulo

5 Classificação

Após a etapa de extração de características, os atributos são utilizados para estimar a emoção em um dado sinal de voz. Existem diversos métodos de classificação, desde técnicas tradicionais, até metodologias que utilizam aprendizado profundo [29]. Assim como em vários outros desafios de reconhecimento de padrões, ainda não existe um único classificador comumente aceito para sistemas de identificação de emoções, a escolha depende de diversos fatores, como o tipo de dado extraído e quantidade de informação disponível.

No contexto de classificação de emoções a partir de sinais de voz, a entrada $\mathbf{x}_i$, corresponde a um vetor de características extraído do sinal do voz, para cada vetor $\mathbf{x}_i$, o método de classificação irá estimar uma determinada emoção $c_i(i = 1, 2, ...C)$, em que C é o número total de emoções (ou de classes). A metodologia de classificação pode ser supervisionada, ou não supervisionada. Na primeira, no processo de treinamento, cada entrada precisa estar rotulada com sua verdadeira emoção. Já no aprendizado não supervisionado, a entrada não precisa estar rotulada, o algoritmo sozinho irá agrupar os dados de acordo com suas características. Nesta seção, serão apresentados alguns dos principais classificadores utilizados na área, incluindo o Modelo de Mistura de Gaussianas, Redes Neurais Artificiais e k-NN (k-Vizinhos mais próximos).

5.1 Modelo de mistura de Gaussianas

O modelo de mistura de Gaussianas, conhecido também como GMM (do inglês, *Gaussian Mixture Models*), pode ser utilizado para solucionar o problema de estimar funções de densidade de probabilidade a partir de um conjunto de dados $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$, com N observações, e dimensão D. O GMM assume que uma determinada população de dados corresponde a instâncias aleatórias independentes de uma mistura de distribuições Gaussianas, essa mistura é obtida através da combinação linear entre duas ou mais distribuições Gaussianas. Essa mistura torna possível modelar dados com diferentes aglomerados, algo que com uma simples Gaussiana não seria apropriado. A mistura de

Gaussianas pode ser definida como mostrado na Equação 5.1 [71].

$$p(\mathbf{x}|\lambda) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5.1}$$

Em que, $\mathbf{x}$ é um vetor de características com dimensão $dim_{\mathbf{x}}$, e λ corresponde ao modelo de mistura de Gaussiana de uma determinada emoção. Além disso, π_k representa o peso para a Gaussiana k, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{dim_{\mathbf{x}} \times dim_{\mathbf{x}}}$ corresponde a matriz de covariância da Gaussiana k, já $\boldsymbol{\mu}_k \in \mathbb{R}^{dim_{\mathbf{x}}}$, indica o vetor de média da Gaussiana k. O termo $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ está associado à verossimilhança da observação $\mathbf{x}$ dada a Gaussiana k, como mostrado na Equação 5.2 a seguir.

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{dim_{\mathbf{x}}/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^{\top}(\boldsymbol{\Sigma}_k)^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)) \tag{5.2}$$

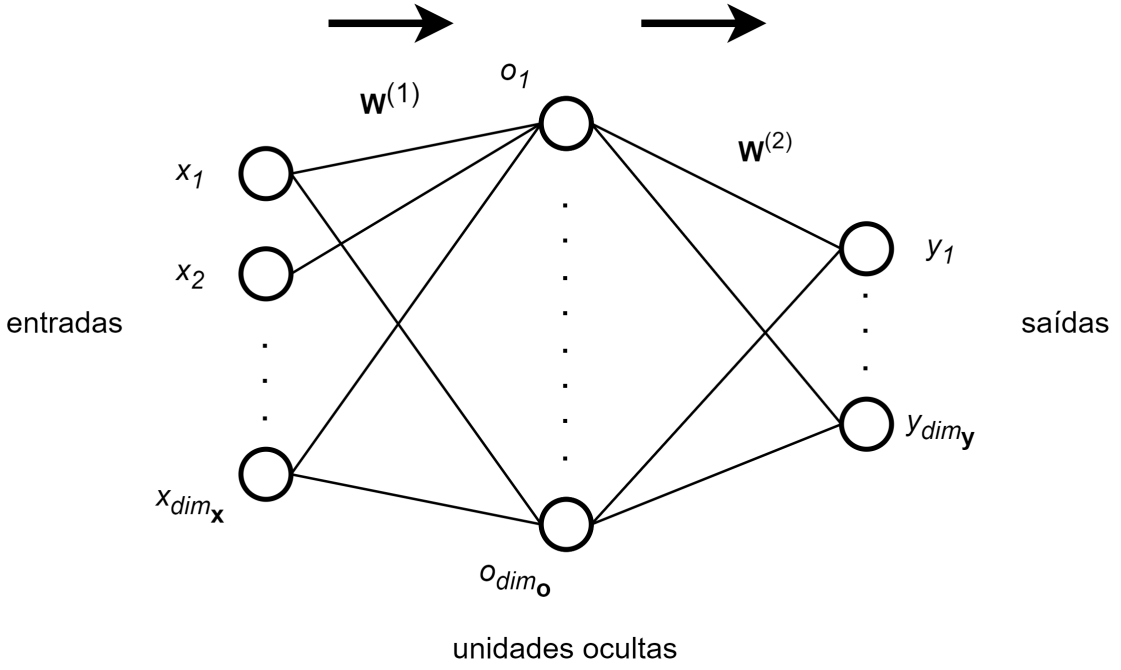
É importante observar que o termo $(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}(\boldsymbol{\Sigma}_k)^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)$ corresponde a distância de Mahalanobis, e que quando $\boldsymbol{\Sigma}_k$ é uma matriz identidade, o termo torna-se distância Euclidiana. Além disso, $\sum_{k=1}^{K} \pi_k = 1$ e $\pi_k \in [0, 1]$. O modelo de mistura de Gaussianas possui três parâmetros: π_k, $\boldsymbol{\Sigma}_k$ e $\boldsymbol{\mu}_k$, que são convenientemente ajustados de acordo com o método conhecido como *Expectation-Maximization*, apresentado em [71] e descrito na Seção 7.5.1.

5.2 Redes neurais artificiais

As redes neurais artificiais são bastante utilizadas em diversos problemas de classificação. Em sua versão popular, é composta por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída, nessa sequência. Cada camada é conectada com a próxima, usando pesos que são inicializados aleatoriamente, no final do processo de treinamento, esses pesos são otimizados a fim de reduzir uma função de custo derivável que reflete o erro de classificação. Na Figura 5.1, é apresentado o diagrama de uma rede neural artificial.

As unidades da camada oculta $(o_1, o_2, ..., o_{dim_{\mathbf{o}}})$ são formadas por uma combinação entre as entradas $(x_1, x_2, ..., x_{dim_{\mathbf{x}}})$ e os pesos $\mathbf{W}^{(1)} \in \mathbb{R}^{dim_{\mathbf{o}} \times dim_{\mathbf{x}}}$, em que $dim_{\mathbf{o}}$ e $dim_{\mathbf{x}}$ são os número de elementos na camada oculta e na camada de entrada, respectivamente. Na Equação 5.3, é mostrado como $\mathbf{o}$ pode ser calculado.

$$o_j = h\Big(\sum_{i=1}^{dim_{\mathbf{x}}} W_{ji}^{(1)} x_i + W_{j0}^{(1)}\Big) = h\Big(\sum_{i=0}^{dim_{\mathbf{x}}} W_{ji}^{(1)} x_i\Big) = h(\mathbf{W}_j^{\intercal}\mathbf{x}) \tag{5.3}$$



(Fonte: Adaptado de [71].)

Figura 5.1 – Diagrama de uma rede neural do tipo mais popular, com apenas uma camada oculta. A camada de entrada, oculta e saída são representadas pelos nós, e os pesos são representados pelos arcos entre cada círculo. As setas indicam o sentido do fluxo da informação na rede durante o processo de ida (conhecido como *feedfoward*).

Na Equação 5.3, o termo $W_{j0}^{(1)}$ representa o viés de cada neurônio da camada oculta, o_j. O viés é capaz de determinar se o neurônio será ou não ativado. Note que pode ser adicionado uma entrada $x_0 = 1$, que é multiplicada por $W_{j0}^{(1)}$, permitindo que o cálculo do neurônio o_j seja feito de forma matricial [72]. Além disso, $h(\cdot)$ indica uma função não linear diferenciável. Caso essa função seja linear, é possível determinar uma rede neural sem camada oculta que seja equivalente.

A função $h(\cdot)$ precisa ser diferenciável para que os parâmetros da rede possam ser otimizados por meio do cálculo de gradientes. A função também deve apresentar continuidade e suavidade, para que possa ser definida em todo intervalo de seu argumento. Outra propriedade desejável é que $h(\cdot)$ sature, ou seja, tenha máximos e mínimos, para que seja possível limitar sua saída. É muito comum a utilização de uma função sigmoide, como a apresentada na Equação 5.4, que é limitada entre 0 e 1, além disso, possui todas as propriedades desejáveis para uma função de ativação [71].

$$\sigma(\phi) = \frac{1}{1 + \exp(-\phi)} \tag{5.4}$$

Além da sigmoide, recentemente, também tornou-se comum a utilização da função *ReLU*, que não é limitada nem derivável em todo seu domínio, porém compensa isso com

seu baixo custo computacional. A função $ReLU$ definida na Equação 5.5.

$$ReLU(\phi) = \max(\phi, 0) \tag{5.5}$$

De forma semelhante ao cálculo das unidades da camada oculta, as unidades da saída $(y_1, y_2, ..., y_{dim_{\mathbf{y}}})$ também são formadas a partir de uma combinação entre as unidades da camada oculta e os pesos $\mathbf{W}^{(2)} \in \mathbb{R}^{dim_{\mathbf{y}} \times dim_{\mathbf{o}}}$, em que $dim_{\mathbf{y}}$ é o número de elementos na camada de saída da rede. Na Equação 5.6 é mostrado como $\mathbf{y}$ pode ser determinado. Além disso, agora o viés $W_{k0}^{(2)}$ é associado ao o_0 com valor 1.

$$y_k = h(\sum_{j=1}^{dim_{\mathbf{o}}} W_{kj}^{(2)} o_j + W_{k0}^{(2)}) = h(\sum_{j=0}^{dim_{\mathbf{o}}} W_{kj}^{(2)} o_j) = h(\mathbf{W}_k^{\intercal}\mathbf{o}) \tag{5.6}$$

A saída é transformada de acordo com uma função de ativação especifica, essa função é determinada de acordo com a natureza dos dados e do problema. Como recomendado em [71], para problemas de classificação multiclasse, a função *softmax* (apresentada na Equação 5.7) pode ser utilizada, essa é uma generalização da função sigmoide (Equação 5.4) para problemas multiclasse, que satisfaz $0 \leq y_k \leq 1$ e $\sum_k y_k = 1$, o que permite que cada saída da rede neural seja tratada como uma probabilidade estimada para uma determinada classe.

$$h(\phi_k) = \frac{\exp(\phi_k)}{\sum_j \exp(\phi_j)} \tag{5.7}$$

O processo de aprendizagem funciona a partir de uma sequência de *exemplos* apresentados à rede, cada exemplo é composto por uma entrada e um rótulo associado a essa entrada, o rótulo pode ser entendido como a classe correta para aquela entrada. Para cada exemplo, a rede faz o caminho de ida (*feedfoward*) que resulta em uma saída. Essa saída é comparada com o rótulo, que deve ser codificado numericamente, e então é determinado o erro.

No caso de uma rede com só uma camada oculta, são os parâmetros $\mathbf{W}^{(1)}$ e $\mathbf{W}^{(2)}$ que são aprendidos pela rede neural artificial. Quando devidamente aprendidos, a rede é capaz de estimar a saída correta a partir de uma nova entrada. O processo de aprendizagem é feito atualizando esses parâmetros, através do método de otimização baseado na retropropagação do erro, conhecido como *backpropagation* [71, 72]. No *backpropagation*, os parâmetros $\mathbf{W}$ são atualizados de acordo com o gradiente descendente da função de custo utilizada [71, 72].

Para problemas de otimização multiclasse, em que cada entrada $\mathbf{x}_n$ está associada a somente uma das C classes, é apropriado utilizar a função de custo da entropia cruzada categórica [71], apresentada na Equação 5.8, em que t_{kn} é a k-ésima saída verdadeira para a entrada $\mathbf{x}_n$, além disso, $t_{kn} \in [0, 1]$ e $\sum_{k=1}^{C} t_{kn} = 1$. O termo $y_k(\mathbf{x}_n, \mathbf{W})$ corresponde à

k-ésima saída estimada pela rede neural para a entrada $\mathbf{x}_n$, essa saída também depende dos pesos $\mathbf{W}$.

$$J(\mathbf{W}) = -\sum_{n=1}^{N}\sum_{k=1}^{dim_{\mathbf{y}}} t_{kn}\ln(y_k(\mathbf{x}_n, \mathbf{W})) \tag{5.8}$$

5.3 k-vizinhos mais próximos

O k-vizinhos mais próximos, também conhecido como *k-nearest neighbors* (k-NN) é uma abordagem simples que classifica uma nova observação de acordo com as k observações de treinamento mais próximas, de forma que a classe da observação com menor distância é associada a nova observação. A distância pode ser calculada de diferentes formas, porém é muito comum a utilização da distância Euclidiana [73]. Sendo $\mathbf{x}_i$ o vetor de características da nova observação, a distância Euclidiana para um determinado vetor $\mathbf{x}_j$ de treinamento pode ser calculada como mostrado na Equação (5.9).

$$\mathbf{Q}_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top(\mathbf{x}_i - \mathbf{x}_j)} \tag{5.9}$$

O número de vizinhos, k, é um parâmetros muito importante para este classificador. Quando $k = 1$, o algoritmo não é capaz de analisar bem o contexto local, pois somente a amostra mais próxima é levada em consideração. Esse fenômeno pode ser observado na Figura 5.2, em que existem duas classes com amostras sobrepostas. Quando $k = 1$, o algoritmo erra pois a amostra mais próxima não é a correta. Entretanto, quando $k = 7$, o algoritmo acerta, pois 6 amostras da classe correta são selecionadas, nesse caso, o algoritmo observa mais amostras ao redor da amostra de teste, o que melhora a percepção sobre o contexto local, isso pode aumentar a qualidade do modelo de classificação.

Na prática, é comum que o valor de k seja escolhido arbitrariamente, dentro de um determinado intervalo de valores. Ao final, é selecionado o k que tenha obtido o melhor desempenho [37].

Embora a implementação do k-NN seja simples (quando comparado aos outros classificadores citados nesse trabalho), essa metodologia possui algumas desvantagens, devido à necessidade de se computar a distância entre uma amostra de teste e cada amostra de treinamento, cujo custo computacional pode ser muito alto, dependendo do número de amostras de treinamento. Além disso, todas as amostras de treinamento devem ficar armazenadas, o que pode ser um problema, caso haja pouca disponibilidade de memória [73].

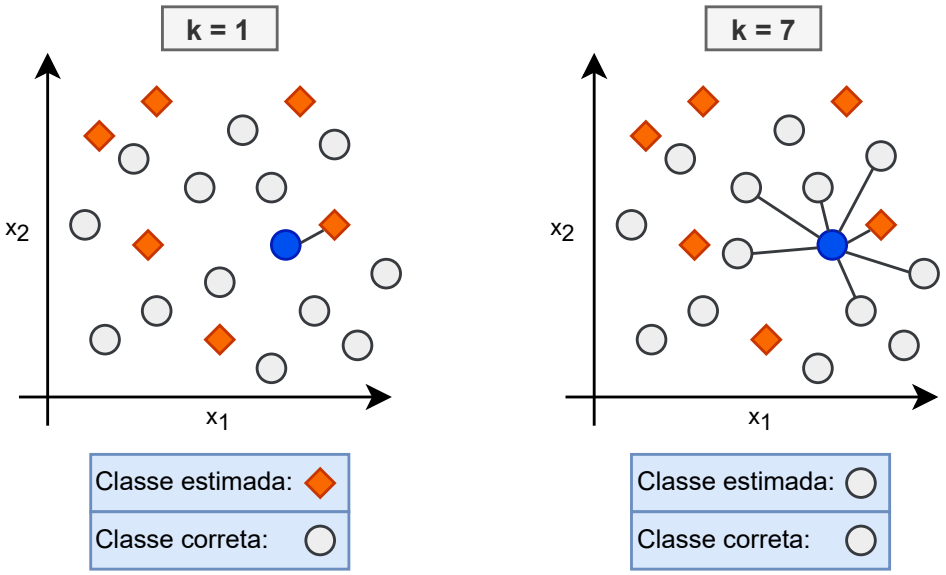


Figura 5.2 – Ilustração do algoritmo k-NN com duas classes com amostras sobrepostas. O círculo azul representa uma amostra com classe desconhecida. Já o losango laranja e o círculo cinza representam amostras de classes diferentes.

Capítulo

6 Revisão bibliográfica

Diversas abordagens para o reconhecimento automático de emoções através de sinais de voz já foram propostas. Neste capítulo, será apresentada uma breve revisão da literatura, mostrando alguns aspectos a respeito dos trabalhos relacionados ao tema.

Na maior parte dos trabalhos, a extração de características é um processo fundamental. Em vários estudos, características associadas à prosódia e ao espectro são bastante exploradas, além de apresentaram os melhores resultados na identificação de emoções [29, 43]. Um dos primeiros trabalhos relacionados ao tema foi desenvolvido por Deallaert et al. em 1996 [74]. Nesse estudo, parâmetros associados à prosódia foram extraídos, como a média, desvio padrão, máximo e mínimo do *pitch*, além da estimação da velocidade de fala, que foi feita a partir da duração dos segmentos de voz. Segundo os autores, os resultados obtidos na classificação de quatro emoções (alegria, tristeza, raiva e medo) foram próximos da performance humana.

No trabalho apresentado por Busso et al. em 2004 [75], além das medidas estatísticas do *pitch*, medidas relacionadas à energia e a relação entre duração de segmentos com voz e sem voz, também foram exploradas. De forma semelhante, no estudo feito por Luengo et al. [76], essas características também são extraídas, entretanto, são incluídas características a respeito do *shimmer* e *jitter*, estes parâmetros refletem duas propriedades importantes da onda do sinal de voz: variação da amplitude e da frequência dos padrões de onda. De forma geral, a utilização de parâmetros associados a prosódia são bastante utilizados até hoje, normalmente, determinando diversas medidas estatísticas sobre esses parâmetros, como feito em diversos trabalhos [77, 78, 79, 53, 37].

Além da utilização da prosódia, muitos trabalhos extraem atributos relacionados ao espectro de frequências do sinal de voz. Diversos trabalhos apontam que as características do trato vocal são bem representadas nesse domínio, e que isso pode ajudar a identificar as emoções através da voz [43, 80, 81]. As características espectrais mais comuns são os coeficientes mel-cepstrais (MFCCs), LFPCs (*log frequency power coefficients*) e coeficientes da predição linear (LPCCs) [82, 29, 83, 84].

No trabalho proposto por Milton [85], 24 MFCCs são utilizados a fim de identificar sete emoções diferentes através da voz, além de medidas estatísticas sobre esses coeficientes, como média, desvio padrão, curtose e assimetria. De forma similar, no trabalho de Lalitha [86], os MFCCs também são usados, além de informações associadas ao cepstrum do sinal. Vários estudos apontam que as características espectrais e de prosódia, quando usadas em conjunto, podem melhorar a acurácia de sistemas de identificação de emoções [37, 29, 78], sendo essa a abordagem mais comum em trabalhos mais recentes [87, 88, 89].

De forma geral, as características extraídas do sinal de voz são cruciais para a identificação correta da emoção. Além disso, um outro ponto importante é sobre os métodos de classificação. É nessa etapa que a emoção é de fato identificada. Nos sistemas de classificação tradicionais, alguns métodos são bastante utilizados: modelo de mistura de gaussianas (GMM) [90, 91], e modelos ocultos de Markov (HMM, do inglês *hidden Markov models*) [92, 93, 58]. Outras metodologias também são usadas, embora menos frequentes: redes neurais artificiais, k-vizinho mais próximos (k-NN), e máquina de vetores de suporte (SVM do inglês *support vector machines*) [43, 94, 29].

Ao se analisar os trabalhos relacionados, é possível notar que não há consenso sobre qual metodologia de classificação é a mais adequada, além disso, os trabalhos utilizam conjuntos de características e bancos de dados diferentes, dificultando uma comparação justa entre os classificadores. De acordo com Lanjewar et al. [95], para o banco de dados do EMO-DB (*Berlin emotional database*, em português: banco de dados de vozes emotivas de Berlim), o GMM é mais robusto e apresenta acurácias melhores, quando comparado ao k-NN. Ainda segundo o autor, a velocidade na classificação do k-NN é superior à do GMM, indicando que o k-NN pode ser uma opção quando há restrições associadas ao tempo. Segundo os estudos apresentado por Ayadi et al. [96] e Sailunaz et al. [43], o HMM e GMM são os métodos mais usados para classificação de emoções, além de serem amplamente usados também em outras aplicações em sinais voz.

No trabalho de Iriya [37], GMM, HMM, k-NN e SVM foram comparados com relação à acurácia na classificação de sete emoções (com separação por sexo), para o banco de dados EMO-DB. Os resultados apontaram que o GMM e HMM obtiveram resultados semelhantes, sendo que o HMM apresenta resultados ligeiramente superiores, com um custo computacional maior. Já o k-NN e SVM apesentaram resultados inferiores, quando comparados ao GMM e HMM. Ainda segundo o autor, a classificação entre homens e mulheres tem aspectos diferentes, tais que as mesmas características não podem ser usadas com resultados equivalentes para ambos os sexos.

Todos as metodologias de classificação citadas nos parágrafos anteriores são usadas em sistemas tradicionais de classificação de emoções [97]. Em trabalhos mais recentes, é notória a utilização de técnicas de aprendizado profundo para a classificação de emoções. Classificadores como redes neurais recorrentes (RNNs, do inglês *recurrent neural networks*)

e LSTM (*long short-term memory*, em português: memória de longo prazo) têm sido bastante utilizados para a classificação de emoções através da voz [29]. Ambas foram desenvolvidas para lidar com informações sequenciais (sinal de fala, texto, vídeo ou séries temporais). Segundo os autores de trabalhos recentes, a possibilidade de analisar dependências temporais complexas tornam esses modelos mais robustos e capazes de obter resultados melhores [98, 99, 97, 100].

No trabalho de Eyben et al. [101], características associadas ao espectro e à prosódia são usadas para alimentar uma LSTM-RNN, usada para classificação de emoções em tempo real. Segundo os autores, uma rede neural tradicional não é adequada para classificar séries temporais, pois são classificadores estáticos, classificando apenas as janelas de áudio, e não levam em consideração informações sobre janelas vizinhas.

Além das técnicas descritas anteriormente, redes neurais convolucionais (CNNs, do inglês *convolutional neural networks*) também têm sido utilizadas, devido à sua capacidade de capturar relações espaciais e temporais dos sinais de entrada [29]. Em muitos trabalhos, são utilizadas propriedades relacionadas ao espectro do sinal de fala na entrada da CNN, como MFCCs e mel-espectrogramas.

No trabalho realizado por Mao et al. [102], a CNN é utilizada para aprender características relevantes associadas às emoções, através do espectrograma. Nesse trabalho, três estratégias de variação de oradores foram comparadas: único orador (em que o orador utilizado no treinamento do modelo é o mesmo do usado no teste), sistema dependente do orador (em que os oradores utilizados no treinamento do modelo também são usados durante o teste), e sistema independente do orador (em que os oradores utilizados no treinamento do modelo são diferentes dos utilizados durante o teste). Para um único orador foi obtido acurácia de 93,7%, para o sistema dependente do orador foi obtido acurácia de 88,3%, já para o sistema independente do orador foi obtido acurácia de 85,2%, todos os resultados são para o banco de dados EMO-DB, com sete emoções. Segundo os autores, a utilização da CNN permite aprender características que são invariantes a determinados fatores.

Já no trabalho proposto por Issa et al. [103], são extraídos dos sinais de fala os coeficientes mel-cepstrais, *chromagram*, mel-espectrograma, representação de Tonnetz e contraste espectral, totalizando 193 características. Segundo os autores, a combinação destas características é capaz de descrever o *pitch*, timbre e harmonia do sinal de fala. Essas características são usadas como entrada para uma CNN unidimensional, obtendo 86,1% de acurácia para o banco de dados EMO-DB, com sete emoções.

De forma semelhante, no trabalho de García-Ordás et al. [104], é proposta a utilização de uma rede neural convolucional completa (também conhecida como *fully convolutional network*) para a classificação de emoções. Com essa metodologia, é possível utilizar entradas com diferentes tamanhos, os autores utilizaram o MFCC e mel-espectrograma, obtendo

Tabela 1 – Descrição de trabalhos relevantes relacionados à área de classificação de emoções através de sinais de voz, em ordem cronológica.

Referência	**Conjunto de dados**	**Características**	**Classificador**	**Resultados (acurácia)**
Schüller et al. (2003) [93]	Próprio	Características associadas à prosódia e energia	HMM e GMM	Para GMM, 86,8% Para HMM, 77,8%
Nwe et al. (2003) [58]	Próprio	12 LFPC 12 MFCC 16 LPCC	HMM	Para LFPC, 77,1% Para MFCC, 59,0% Para LPCC, 56,1%
Ayadi et al. (2007) [105]	EMO-DB	12 MFCC 12 delta MFCC Energia	k-NN RNA HMM GMVAR	Para k-NN, 67,3% Para RNA, 55,0% Para HMM, 71,0% Para GMVAR, 76,0% Todos usando 6 emoções
Seehapoch et al. (2013) [106]	EMO-DB	F_0 Energia ZCR LPC MFCC	SVM	Para F0, 53,47% Para Energia 62,26% Para ZCR, 51,10% Para LPC, 58,45% Para MFCC, 78,04% Características combinadas, 89,80%
Iriya (2014) [37]	EMO-DB	Características associadas à prosódia e energia	GMM HMM k-NN SVM	Para homens: Para GMM, 52,79% Para HMM, 59,23% Para k-NN, 50,64% Para SVM, 57,08% Para mulheres: Para GMM, 59,10% Para HMM, 61,26% Para k-NN, 48,01% Para SVM, 53,31%
Mao et al. (2015)	EMO-DB SAVEE DES MES	Espectrograma	CNN	Para o EMO-DB, 85,2% Para SAVEE, 73,6% Para o DES, 79,9% Para o MES, 78,3%
Issa et al. (2020) [103]	EMO-DB RAVDESS IEMOCAP	MFCC Chromagram Mel-espectrograma Representação de Tonnetz Contraste espectral	CNN	Para EMO-DB, 86,1% Para RAVDESS, 71,61% Para IEMOCAP, 64,3%
Er (2020) [107]	EMO-DB RAVDESS IEMOCAP	Características associadas ao domínio da frequência	CNN e SVM	Para EMO-DB, 90,21% Para RAVDESS, 79,41% Para IEMOCAP, 85,37%
Sahoo et al. (2021) [108]	EMO-DB RAVDESS SAVEE	Mel-espectrograma	TLEFuzzyNet	Para EMO-DB, 99,38% Para RAVDESS, 99,66% Para SAVEE, 98,57%

resultado superior com o MFCC.

Na Tabela 1, é apresentada uma breve descrição sobre trabalhos relevantes relacionados à área de classificação de emoções através de sinais de fala. É mostrado o conjunto de dados, as características e as metodologias de classificação utilizadas, além dos resultados obtidos em cada experimento.

Os trabalhos descritos anteriormente apresentam diferentes soluções para o problema de classificação de emoções através de sinais de fala. A maioria dos trabalhos recentes busca lidar com a informação sequencial (ou temporal) através de metodologias baseadas em aprendizado profundo, incluindo RNNs e LSTM. Além da utilização do potencial das CNNs para aprender características relevantas a respeito do sinal de fala.

Em trabalhos relacionados à área de Processamento de Linguagem Natural (NLP), têm sido utilizadas representações vetoriais para palavras, de forma que, para cada palavra,

é associado um determinado vetor n-dimensional. Em algumas metodologias, os vetores levam em consideração a semântica das palavras. Dessa forma, o espaço dimensional obtido gera uma representação semântica. Nessa representação, é comum a utilização de um contexto temporal como base para sua construção. A representação criada é capaz de levar em consideração informações associadas a sequência da informação, e essas informações podem ser úteis para classificação de emoções através de sinais de voz.

No trabalho proposto por Bengio et al. [67], uma rede neural artificial é utilizada para obter representações vetoriais para palavras, utilizando o contexto baseado em palavras anteriores. Após o treinamento da rede, os parâmetros otimizados são capazes de representar as palavras em um espaço contínuo, de forma que, palavras com significado semelhantes estão próximas no espaço vetorial obtido pela rede.

De forma similar, no trabalho de Mikolov et al. [68], uma rede neural também é usada para obter as representações, entretanto, os autores propõem uma metodologia mais simples e eficiente para o treinamento, além de introduzir dois modelos para o aprendizado de representações vetoriais para palavras.

Uma abordagem diferente é utilizada no trabalho apresentado por Pennington et al. [12]. Nesse trabalho os autores propõem a utilização de uma matriz de coocorrências entre as palavras, essa matriz reflete as relações de ocorrência (dentro de um mesmo contexto local) entre todas as palavras do *corpus* utilizado. Com essa metodologia, é possível determinar vetores n-dimensionais para cada palavra, reduzindo uma função de custo através de técnicas de otimização.

Em 2018, o trabalho desenvolvido por Chung et al. [109] propôs uma versão do *Word2Vec* para sinais de voz, o *Speech2Vec*. No trabalho, é utilizado um *autoencoder* baseado em Rede Neural Recorrente (RNN), que estima um vetor com dimensão fixa para um determinado segmento de voz. Segundo os autores, o vetor é capaz capturar a informação semântica a respeito das palavras ditas durante o segmento de voz. Em 2021, o trabalho proposto por Tzirakis et al. [110] utiliza o *Speech2Vec* em conjunto com o *Word2Vec* para capturar informação semântica e paralinguística a respeito da voz com o intuito de classificar emoções, os resultados obtidos apontam que há uma melhora no reconhecimento de emoções ao aplicar o método proposto.

Atualmente, há poucos registros na literatura sobre trabalhos que exploram metodologias usadas em NLP para lidar com características extraídas do sinal de voz, especialmente em problemas de classificação de emoções. Neste trabalho, é avaliado o desempenho da representação semântica na tarefa de classificação de emoções usando sinais de voz, analisando aspectos relacionados ao contexto semântico, características vocais e metodologias de classificação.

Capítulo

7 Metodologia

Neste capítulo, é descrita a metodologia utilizada no presente trabalho, sendo dividida em cinco partes:

1. Conjunto de dados;
2. Pré-processamento do sinal de voz;
3. Extração de características;
4. Representação semântica; e
5. Métodos de classificação.

7.1 Conjunto de dados

Os dados utilizados para todos os experimentos deste trabalho são provenientes do Banco de Dados de Vozes Emotivas de Berlim (Emo-DB) [35], produzido pela Universidade Técnica de Berlim. Essa base é formada por locuções produzidas por dez atores, cinco do sexo masculino e cinco do sexo feminino. Além disso, os áudios possuem frequência de amostragem de 16 kHz.

Esse conjunto de dados já foi amplamente utilizado em trabalhos relacionados, o que deve facilitar a comparação dos resultados. Foram gravadas 535 frases em sete diferentes emoções: raiva, tédio, nojo, medo, alegria, tristeza e neutralidade. As frases são faladas em alemão e seus conteúdos linguísticos não têm nenhuma relação com a emoção expressada pelo ator. Além disso, 20 pessoas avaliaram o conjunto de dados em duas abordagens: reconhecimento e naturalidade das emoções expressadas, obtendo 80% de acurácia para o reconhecimento, e julgadas como natural por 60% dos entrevistados.

A quantidade de arquivos de áudio para cada emoção é apresentada na Tabela 2 a seguir. Observe que a emoção Raiva possui a maior quantidade de arquivos, sendo mais

Tabela 2 – Quantidade de arquivos de áudio por emoção.

Emoção	**Quantidade**	**Quantidade utilizada**
Raiva	127	90
Tédio	81	81
Nojo	46	46
Medo	69	69
Alegria	71	71
Tristeza	62	62
Neutralidade	79	79
Total	**535**	**498**

que o dobro da quantidade de arquivos para a classe Nojo. Devido a esse desbalanceamento na quantidade de arquivos de áudio por classe, neste trabalho, para a emoção Raiva, são utilizados apenas 90 arquivos de áudio. Dessa forma, não é possível comparar diretamente os resultados deste trabalho com resultados obtidos em outros trabalhos que utilizam o mesmo conjunto de dados.

7.2 Pré-processamento do sinal de voz

Esta etapa é responsável por preparar os sinais de voz para que sejam analisados pelas técnicas seguintes. Cada áudio é dividido em janelas de 40 ms, com sobreposição de 10 ms. Neste trabalho, o sinal janelado é apresentado como $\hat{s}(n)$, em que $0 \leq n < Wn$, onde Wn significa o número de amostras da janela.

Para técnicas em que a análise espectral é necessária (VAD, MFCC e LFPC), é utilizada a função de Hamming para o janelamento do sinal de voz, essa função melhora a representação espectral [111]. A função de Hamming é apresentada na Equação 7.1. Além disso, também é aplicado um filtro de pré-ênfase, esse filtro tem finalidade de balancear o espectro de frequências, melhorando a representação espectral do sinal de voz [112, 113]. O filtro de pré-ênfase usado é mostrado na Equação 7.2.

$$w_{hamm}(n) = \begin{cases} 0,54 - 0,46\cos(2\pi n/(W_n - 1)), & \text{se } 0 \leq n < Wn \\ 0, & \text{outro} \end{cases} \tag{7.1}$$

$$s(n) \leftarrow s(n) - 0.97s(n-1) \tag{7.2}$$

É também realizada a detecção de atividade vocal (VAD), funcionando da seguinte forma: para cada janela l, é calculado o espectro do sinal, através da transformada rápida de Fourier (também conhecida FFT (*fast Fourier transform*) [111]). Em seguida, é calculada a média da magnitude do espectro entre as frequências de 80 e 560 Hz, $\mu_E(l)$. Caso $\mu_E(l)$

seja superior à média da magnitude do espectro entre as frequências de 80 e 560 Hz para todas as janelas de $s(n)$, ϵ, é considerado que a janela l possui voz, como mostrado a seguir

$$VAD(l) = \begin{cases} 1, & \text{se } \mu_E(l) \geq \epsilon \\ 0, & \text{se } \mu_E(l) < \epsilon \end{cases} \tag{7.3}$$

Dessa forma, a média do espectro entre 80 e 560 Hz para todas as janelas de uma locução específica serve como limiar de decisão.

A principal motivação para a utilização dessa técnica está relacionada a frequência fundamental da voz ($f0$). Durante a comunicação habitual dos humanos, a $f0$ está entre 80 e 560 Hz [114, 115]. A frequência fundamental pode ser uma manifestação da existência de voz em uma determinada janela, pois para fonemas vocálicos, a vibração das cordas vocais produz uma onda sonora que possui estrutura semi-periódica [36]. Através disso, é possível identificar se uma determinada janela l possui atividade vocal.

O domínio da frequência de sinais periódicos é composto por impulsos na frequência fundamental ($f0$) e em seus múltiplos (harmônicos) [34], essa característica pode ser observada no item (b) da Figura 7.1, que ilustra picos proeminentes na frequência fundamental e em seus harmônicos. Dessa forma, para janelas com $f0$ entre 80 e 560 Hz, a média da magnitude do espectro entre essas frequências é mais alta, quando comparada a janelas em que o sinal não possui voz, apenas ruído.

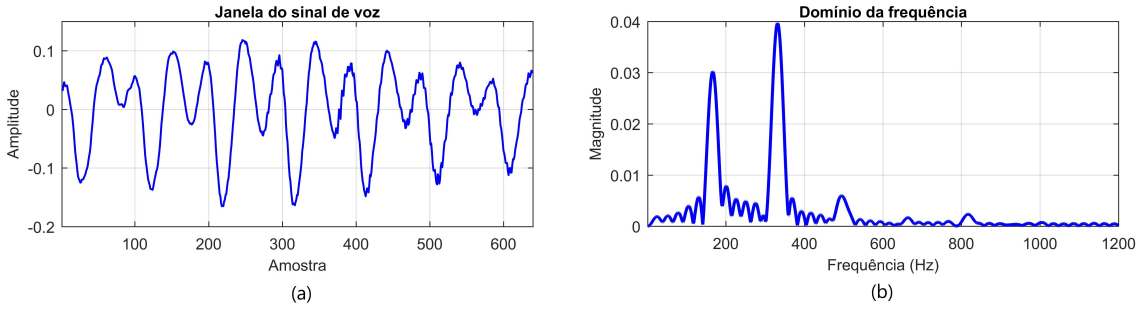


Figura 7.1 – (a) Janela do sinal de voz no domínio do tempo; (b) Domínio da frequência.

Na Figura 7.2, é apresentado o espectrograma do sinal do arquivo 03a01Fa do banco de dados EmoDB. Já na Figura 7.3, é mostrado o espectrograma do mesmo sinal após a aplicação da técnica de detecção de atividade vocal proposta. É possível observar que janelas com baixa magnitude entre as frequências de 80 e 560 Hz são removidas. Ainda sobre a Figura 7.3, é possível observar que as janelas consideradas com voz são concatenadas, provocando uma alteração na dinâmica da fala.

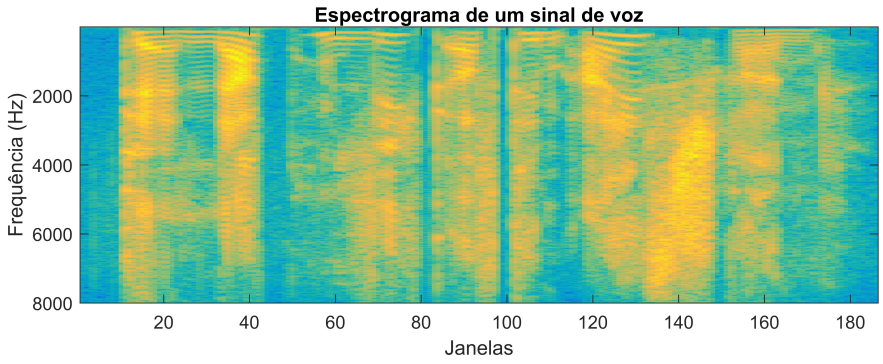


Figura 7.2 – Espectrograma do sinal 03a01Fa do banco de dados EmoDB.

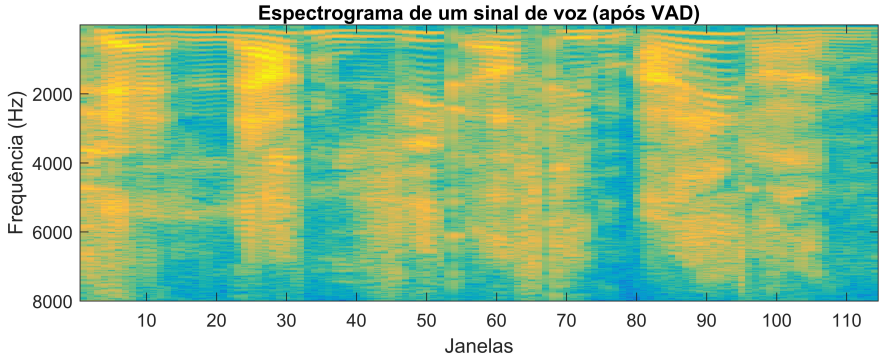


Figura 7.3 – Espectrograma do sinal 03a01Fa do banco de dados EmoDB após o VAD.

7.3 Extração de características

Após a classificação de janelas com atividade vocal, são extraídas características dos sinais de voz. São utilizadas características associadas à prosódia e ao espectro. As características relacionadas a prosódia são: (1) frequência fundamental ($f0$), (2) *jitter*, (3) *shimmer*, (4) energia e (5) taxa de cruzamento por zero. Já as características associadas ao espectro são: (1) coeficientes mel-cepstrais e (2) coeficientes LFPC.

As características são extraídas para todas as janelas (classificadas com voz pelo VAD) do sinal de voz $s(n)$. A seguir, é detalhado como cada característica é determinada.

7.3.1 Frequência fundamental

A metodologia utilizada para calcular frequência fundamental, $f0$, é baseada na autocorrelação (apresentada em 3.2.2.3). Os seguintes passos são executados para obter $f0(l)$, $l = 1, 2..., Ln$, em que $f0(l)$ corresponde à $f0$ da $l-$ésima janela analisada, e Ln representa o número de janelas do sinal de voz $s(n)$.

1. Para cada janela de áudio l, calcula-se a média das amostra, $\mu(l)$, em seguida,

subtrai-se $\mu(l)$ de todas as N amostras da janela l, como mostrado na Equação 7.4.

$$\hat{s}(n) \leftarrow \hat{s}(n) - \mu(l), \quad n = 1, 2, ..., N \tag{7.4}$$

2. Em seguida, é determinada a função de autocorrelação para cada janela l do sinal de voz, como mostrado na Equação 7.5 a seguir, em que dm inicia no inteiro mais próximo a $\frac{Fs}{560}$, e é incrementado de forma unitária até o inteiro mais próximo de $\frac{Fs}{80}$ (Fs corresponde a frequência de amostragem do sinal de voz, para os dados utilizados nesse trabalho $Fs = 16$ kHz). Esses valores são usados para limitar a $f0(l)$ entre 80 e 560 Hz.

$$R_{ss}(dm) = \sum_{n=0}^{Wn-1} \hat{s}(n)\hat{s}(n + dm) \tag{7.5}$$

3. Determina $f0(l)$ como mostrado na Equação 7.6, onde dm_{pico} corresponde ao valor de dm para o primeiro pico da função de autocorrelação $R_{ss}(dm)$, como ilustrado na Figura 7.4.

$$f0(l) = Fs/dm_{pico} \tag{7.6}$$

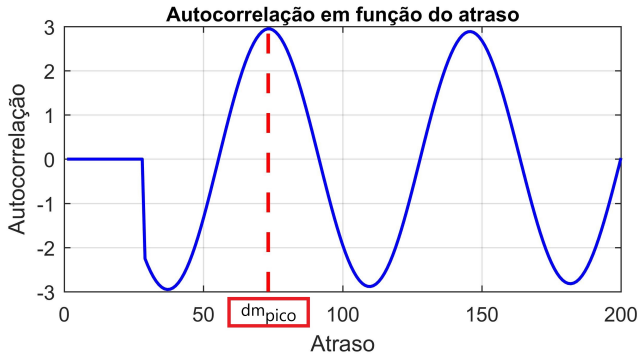


Figura 7.4 – Função de autocorrelação para uma determinada janela de um sinal de voz. O dm_{pico}, sinalizado com um retângulo vermelho, indica o valor de dm para o primeiro pico da função de autocorrelação. Note que até o atraso $Fs/560 = 16000/560 \approx 28$, a autocorrelação é 0, devido a limitação de valores para $f0$, como descrito no passo 2.

4. Em seguida, é aplicado um algoritmo para reduzir grandes variações do vetor $f0$ obtido, essas variações são provocadas por erros na estimação da frequência fundamental. Inicialmente, calcula-se a mediana de $f0$, Me.

5. Caso a diferença absoluta entre $f0(i)$ e Me seja maior que 50 Hz, $f0(i)$ é substituído pela média entre 10 amostras vizinhas, 5 anteriores e 5 posteriores. Caso contrário, $f0(i)$ é mantido, como mostrado na Equação 7.7.

$$f0(i) = \begin{cases} \frac{1}{10}\sum_{j=-5, j\neq 0}^{5} f0(i+j), & \text{se } |f0(i) - Me| > 50 \\ f0(i), & \text{se } |f0(i) - Me| \leq 50 \end{cases} \tag{7.7}$$

6. Por fim, é aplicado um filtro de média móvel a fim de suavizar o vetor $f0$ obtido. O filtro calcula a média de 5 amostras anteriores a $f0(i)$, como mostrado na Equação 7.8.

$$f0(i) = \frac{1}{5}\sum_{j=1}^{5} f0(i-j), \quad i = 6, 7, ..., N \tag{7.8}$$

A Figura 7.5 apresenta o processo de estimação da $f0$ com o arquivo de áudio 03a01Nc do banco de dados EmoDB. O processo é realizado conforme os passos definidos acima. O item (a) é obtido através do passo 1, 2 e 3, resultando em um vetor de $f0$. O item (b) é obtido através do passo 4 e 5, que resulta em um vetor com menos variações, visto que, frequências fundamentais discrepantes são substituídas por uma média das $f0$ vizinhas (anteriores e posteriores). Por fim, o item (c) é conseguido por meio do passo 6, que suaviza o vetor obtido através de um filtro de média móvel.

A Figura 7.6 apresenta uma comparação entre a $f0$ estimada e o espectrograma (entre 1 e 560 Hz), é possível observar que a $f0$ estimada segue o primeiro máximo local do espectrograma. Segundo autores, esse primeiro máximo local representa a frequência fundamental do sinal de voz [6]. Dessa forma, é possível observar que a metodologia aplicada é capaz de determinar aproximadamente a frequência fundamental para cada janela de um sinal de voz.

7.3.2 *Jitter* e *shimmer*

A seguir, são apresentados os passos realizados para obter $jitter(l)$ e $shimmer(l)$, $l = 1, 2..., Ln$, em que $jitter(l)$ e $shimmer(l)$ correspondem ao *jitter* e *shimmer* da $l-$ésima janela analisada, respectivamente.

1. Inicialmente, o sinal de voz $s(n)$ passa por um filtro FIR passa-faixas, mantendo frequências entre 80 e 560 Hz.

2. Para cada janela de áudio l, calcula-se a média das amostra, $\mu(l)$, em seguida, subtrai-se $\mu(l)$ de todas as N amostras da janela l, como mostrado na Equação 7.4.

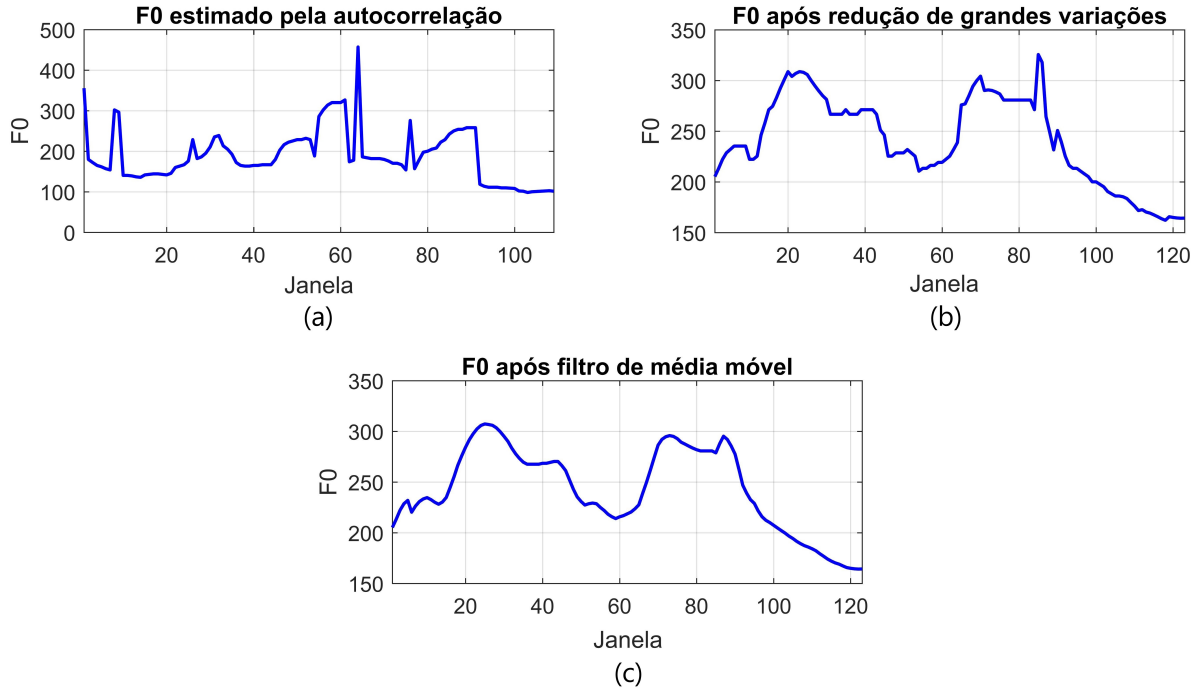


Figura 7.5 – Ilustração da estimação da $f0$. No item (a), é mostrada a $f0$ estimado através da técnica da autocorrelação. No item (b), é apresentada a $f0$ após o processo de redução de grandes variações. Já no item (c), é mostrada a $f0$ obtida após o filtro de média móvel.

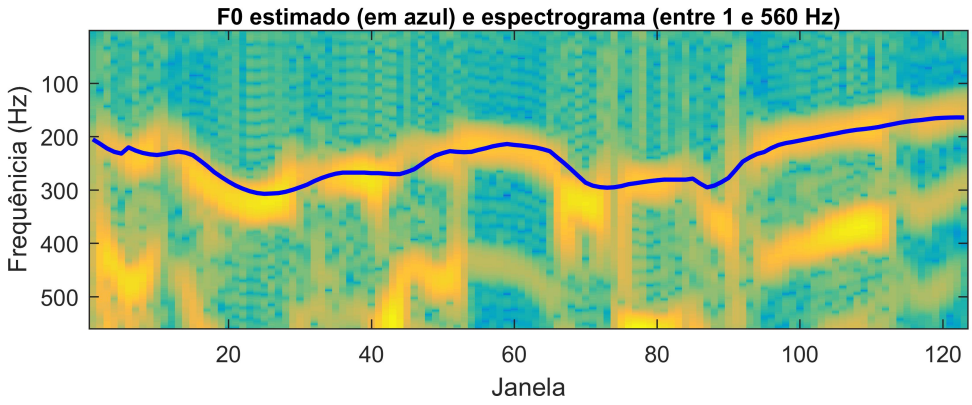


Figura 7.6 – Comparativo entre a $f0$ (em azul) estimada e o espectrograma entre 1 e 560 Hz para o sinal de voz 03a01Nc do banco de dados EmoDB.

3. Para cada janela de áudio l, são determinados máximos locais, $Ml(i)$, no sinal janelado, $\hat{s}(n)$, que obedecem aos seguintes critérios:

   a) Amplitude mínima de $0,5\max(\hat{s}(n))$;

   b) A diferença entre as posições de cada máximo local $Ml(i)$ deve ser de no mínimo $0.85Fs/f0(l)$ amostras.

Os critérios estabelecidos acima têm como propósito filtrar máximos locais mais prováveis de serem as cristas de cada ciclo de onda, como mostrado na Figura 7.7.

O critério b) adiciona uma restrição que não permite a existência de máximos locais muito próximos. Para isso, é determinada a quantidade de amostras que corresponde ao período dos ciclos para a janela l (fazendo $Fs/f0(l)$), então 85% dessa quantidade é usada como distância mínima entre os picos.

Caso não existam máximos locais que atendam a esses critérios, é atribuído valor nulo para $jitter(l)$ e $shimmer(l)$, o valor nulo é posteriormente corrigido no passo 5.

Também são calculados mínimos locais $ml(i)$ (vide Figura 7.7), da mesma forma que são definidos os máximos locais, entretanto usando $-s(n)$.

Sobre a Figura 7.7, é possível observar que cada máximo local possui uma localização correspondente chamada de $n_{Ml(i)}$, que indica em qual amostra $Ml(i)$ ocorre.

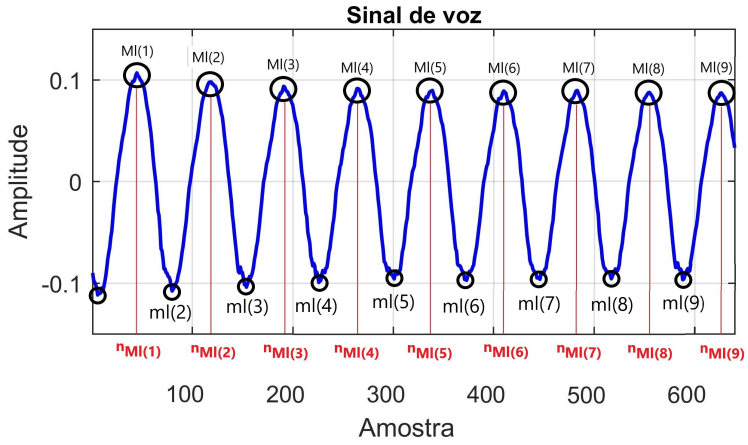


Figura 7.7 – Ilustração dos máximos locais, $Ml(i)$ e suas respectivas amostras (ou localizações), $n_{Ml(i)}$. Também é mostrado o mínimo local de cada ciclo, $ml(i)$. Idealmente, $Ml(i)$ deve representar o pico e $ml(i)$ deve representar o vale do ciclo i.

4. Após definir os máximos locais, $Ml(i)$, $i = 1, 2, ..., Ci$, em que Ci corresponde ao número de ciclos no sinal janelado $\hat{s}(n)$, é estimado o número de amostras que corresponde ao período de cada ciclo i, Na_i, calculando a diferença entre as localizações $n_{Ml(i)}$ e $n_{Ml(i-1)}$, como mostrado na Equação 7.9.

$$Na_i = n_{Ml(i)} - n_{Ml(i-1)} \tag{7.9}$$

Já para estimar a amplitude pico-a-pico, A_i, para cada ciclo i, é determinada a diferença entre o máximo local do ciclo i, $Ml(i)$ e o mínimo local do ciclo i, $ml(i)$, como mostrado na Equação 7.10.

$$A_i = Ml(i) - ml(i) \tag{7.10}$$

5. Para cada janela de áudio l, é calculado o $jitter(l)$ e o $shimmer(l)$, como apresentado nas Equações 7.11 e 7.12, respectivamente.

$$jitter(l) = \frac{\frac{1}{Ci}\sum_{i=1}^{Ci}|Na_i - Na_{i-1}|}{\frac{1}{Ci}\sum_{i=1}^{Ci}Na_i} \tag{7.11}$$

$$shimmer(l) = \frac{\frac{1}{Ci}\sum_{i=1}^{Ci}|A_i - A_{i-1}|}{\frac{1}{Ci}\sum_{i=1}^{Ci}A_i} \tag{7.12}$$

6. Por fim, uma metodologia de correção e suavização dos valores obtidos para $jitter(l)$ e $shimmer(l)$ é realizada. Caso não existam máximos locais que satisfaçam aos critérios apresentados no passo 3, o valor nulo é atribuído para $jitter(l)$ e $shimmer(l)$, para corrigir isso, a média dos valores definidos para 10 janelas vizinhas (5 anteriores e 5 posteriores) são utilizados, como mostrado nas Equações 7.13 e 7.14, para a correção do $jitter(l)$ e $shimmer(l)$, respectivamente.

$$jitter(i) = \begin{cases} \frac{1}{10}\sum_{j=-5,j\neq 0}^{5} jitter(i+j), & \text{se } jitter(i) = nulo \\ jitter(i), & \text{se } jitter(i) \neq \text{nulo} \end{cases} \tag{7.13}$$

$$shimmer(i) = \begin{cases} \frac{1}{10}\sum_{j=-5,j\neq 0}^{5} shimmer(i+j), & \text{se } shimmer(i) = nulo \\ shimmer(i), & \text{se } shimmer(i) \neq \text{nulo} \end{cases} \tag{7.14}$$

7.3.3 Energia

Para determinar a energia $En(l)$, $l = 1, 2, ..., Ln$, são executados os passos descritos a seguir.

1. Inicialmente, para cada janela de áudio l, calcula-se a média das amostra, $\mu(l)$, em seguida, subtrai-se $\mu(l)$ de todas as N amostras da janela l, como mostrado na Equação 7.4.

2. Em seguida, é calculada a energia, $En(l)$, somando todas as amostras ao quadrado presentes no sinal de voz janelado $\hat{s}(n)$, em seguida, é determinado o logaritmo do resultado, como mostrado na Equação 7.15.

$$En(l) = log_{10}(\sum_{n=0}^{N-1}\hat{s}(n)^2) \tag{7.15}$$

7.3.4 Taxa de cruzamento por zero

Para determinar a taxa de cruzamento por zero $tcz(l), l = 1, 2, ..Ln$, são executados os seguintes passos:

1. Para cada janela de áudio l, calcula-se a média das amostra, $\mu(l)$, em seguida, subtrai-se $\mu(l)$ de todas as N amostras da janela l, como mostrado na Equação 7.4.

2. Em seguida, a taxa de cruzamento por zero pode ser calculada como mostrado na Equação 7.16.

$$tcz(l) = \sum_{n=1}^{N-1} 0.5|sinal\{\hat{s}(n)\} - sinal\{\hat{s}(n-1)\}| \tag{7.16}$$

   Pode ser observado que $0.5|sinal\{\hat{s}(n)\} - sinal\{\hat{s}(n-1)\}|$ só é 1 quando $\hat{s}(n)$ e $\hat{s}(n-1)$ possuem sinais diferentes, sendo assim, $tcz(l)$ representa a soma de todas as vezes em que o sinal $\hat{s}(n)$ cruzou o zero, dentro da janela l. Além disso, o operador $sinal\{\hat{s}\}$ indica o sinal algébrico de $\hat{s}(n)$, como mostrado na Equação 7.17.

$$sinal\{\hat{s}\} = \begin{cases} 1, & \text{se } \hat{s}(n) \geq 0 \\ -1, & \hat{s}(n) < 0 \end{cases} \tag{7.17}$$

7.3.5 Coeficientes mel-cepstrais

A implementação dos coeficientes mel-cepstrais é baseada no trabalho de Slaney [57], entretanto, são usados 37 filtros passa-bandas, essa modificação é feita para que o último filtro tenha limite superior na última componente de frequência do espectro (8 kHz), como mostrado na Figura 7.8. Além disso, $MFCC_k(l)$ corresponde a um vetor com 12 coeficientes mel-cepstrais para a janela l. A seguir, são descritos os passos realizados para obter o $MFCC_k(l)$.

1. Inicialmente, para cada janela l do sinal de voz é determinada a transformada rápida de Fourier.

2. Em seguida, para cada janela analisada, é calculado o produto escalar entre o espectro dessa janela (ao quadrado), $X(m)^2$ e o banco de filtros passa-faixas, $F(m, j)$, como mostrado na Equação 7.18, em que, $j = 1, 2, ..., J$, sendo J o número de filtros, e $m = 1, 2, ..., M$, onde M significa o número de componentes do espectro X. São

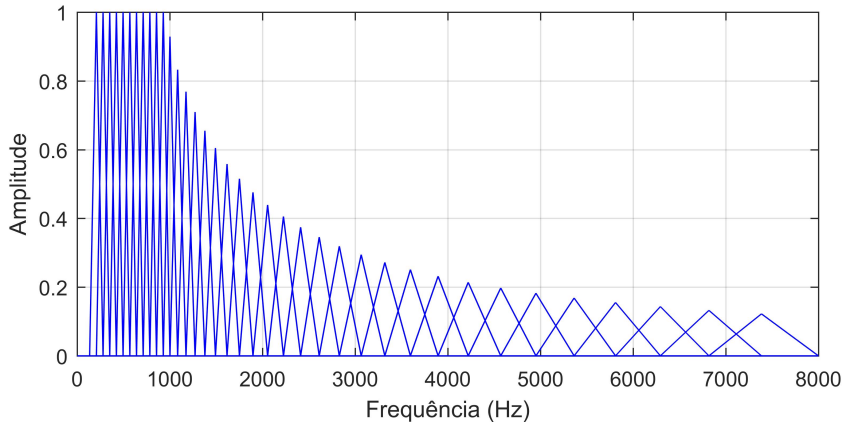


Figura 7.8 – Filtros passa-bandas usados no passo 2 para a determinação dos MFCCs.

utilizados 37 filtros triangulares com áreas aproximadamente iguais, como mostrado na Figura 7.8.

$$El(j) = log_{10}(\sum_{m=1}^{M} X(m)^2 F(m,j)) \tag{7.18}$$

3. Em seguida, o vetor $MFCC_k(l)$ pode ser obtido a partir da Equação 7.19. Lembramos que o primeiro coeficiente (quando $k = 1$) representa a informação referente a energia do sinal de voz [116], e não é usado nesse trabalho.

$$MFCC_k(l) = \sum_{j=1}^{J} El(j) \cos[k(j - 0,5)\frac{\pi}{J}] \quad k = 2, 3, ...13 \tag{7.19}$$

7.3.6 Coeficientes LFPC

A implementação usada para obter os coeficientes LFPC é baseada no trabalho de Nwe et al. [58]. Os passos a seguir demonstram como obter $LFPC_j(l)$. Neste trabalho, $LFPC_j(l)$ corresponde a um vetor com 12 coeficientes LFPC para a janela l.

1. Inicialmente, para cada janela l do sinal de voz é determinada a transformada rápida de Fourier.

2. Após isso, são definidos 12 filtros passa-faixas. Inicialmente é determinado a largura $b(1), b(2), ..., b(12)$ para os 12 filtros. A Equação 7.20 mostra como calcular a largura $b(j)$ através de $b(j-1)$, é necessário determinar o valor de $b(1)$ para calcular as larguras de banda posteriores, nesse trabalho é usado $b(1) = 54$ Hz.

Ainda sobre a Equação 7.20, o α é usado para que a largura e banda $b(j)$ seja maior que a largura de banda $b(j-1)$, isso faz com que os filtros sejam dispostos de forma não linear ao longo do espectro. Nesse trabalho é utilizado $\alpha = 1{,}4$.

$$b(j) = \alpha b(j-1) \tag{7.20}$$

3. Em seguida, é definido a frequência central $fc(1), fc(2), ..., fc(12)$ para os 12 filtros. A Equação 7.21 mostra como determinar $fc(j)$. É necessário determinar $fc(1)$ para calcular as frequências centrais posteriores, é usado $fc(1) = 127$ Hz.

$$fc(j) = fc(1) + \sum_{i=1}^{j-1} b(i) + \frac{b(j) - b(1)}{2} \tag{7.21}$$

4. Os filtros filtros passa-faixas W podem ser determinados como apresentado na Equação 7.22, sendo que, $j = 1$, 2,..., 12, $m = 1, 2, ..., M$, e M significa o número de componentes do espectro X para a janela l. Além disso, $l_{inf}(j)$ e $l_{sup}(j)$ correspondem, respectivamente, ao limite inferior e superior do j-ésimo filtro passa-faixas.

$$W(m, j) = \begin{cases} 1, & l_{inf}(j) \leq m < l_{sup}(j) \\ 0, & \text{outro} \end{cases} \tag{7.22}$$

Em que,

$$l_{inf}(j) = fc(j) - \frac{b(j)}{2} \tag{7.23}$$

$$l_{sup}(j) = fc(j) + \frac{b(j)}{2} \tag{7.24}$$

Cada filtro é usado para obter a energia logarítmica de uma determinada faixa de espectro, como mostrado na Equação 7.25.

$$E(j) = 10\log_{10}(\sum_{m=1}^{M} X(m)^2 W(m, j)) \tag{7.25}$$

5. Por fim, para cada filtro j, a energia logarítmica é dividida pelo número de componentes espectrais presentes na faixa do espectro que $E(j)$ foi obtido, como mostrado na Equação 7.26.

$$LFPC_j(l) = \frac{E(j)}{b(j)} \quad j = 1, 2, ...12 \tag{7.26}$$

7.3.7 Normalização

Todas as características extraídas são normalizadas, para que todas tenham a mesma importância nas metodologias de classificação utilizadas a seguir. Na literatura, diferentes técnicas para a normalização são utilizadas, entretanto, os trabalhos apresentados por Bock et al [117] e Sefara [118], mostram que a utilização do escore padrão (também conhecido como *z-score*) pode melhorar a classificação das emoções de forma significativa.

A normalização de acordo com o escore padrão é feita como mostrado na Equação 7.27, em que:

- $x(i,k)$ é a característica k da observação i;
- $\mathbf{x}(*,k)$ são todas as observações da característica k;
- μ_k é a média das observações da característica k;
- $\sigma(\cdot)$ representa uma função que extrai o desvio padrão amostral dos elementos de um vetor dado como entrada.

Dessa forma, todas as características terão média 0 e desvio padrão 1.

$$x(i,k) = \frac{x(i,k) - \mu_k}{\sigma(\mathbf{x}(*,k))} \tag{7.27}$$

7.4 Representação semântica

Nesta etapa, é aplicada a metodologia da representação semântica, explicada no Capítulo 4. Para obter a representação semântica, são utilizados dois tipos de contexto semântico: contexto baseado em vizinhança e o contexto baseado no TF-IDF. Inicialmente, é utilizado o método de agrupamento *k-means* para determinar 2048 centroides que se adaptam aos dados. Dessa forma, o vetor de características para uma determinada janela l, $\mathbf{x}(l)$, é associado a um símbolo $w(l) \in \{1, 2, ..., 2048\}$.

O valor de 2048 centroides indica o número de níveis de quantização. Valores menores também podem ser usados, pois facilitam a otimização da matriz de dissimilaridade $\mathbf{D}$, entretanto, a resolução dos dados quantizados é menor [119]. Nesse trabalho, é usado 2048 por conseguir manter uma boa relação entre a resolução dos dados e custo de otimização da matriz de dissimilaridade.

A seguir, é detalhado a implementação para cada tipo de contexto.

7.4.1 Contexto baseado em vizinhança

1. Inicialmente, para cada janela, o vetor de características $\mathbf{x}(l)$ é quantizado como mostrado na Equação 4.1, apresentada na Seção 4.1.

2. Em seguida, para o símbolo correspondente a janela l, $w(l)$, é determinado Nv símbolos vizinhos (anteriores e posteriores) a $w(l)$. Assim é possível determinar $\mathbf{S}_{w(l),w(i)}$, para cada vizinho $w(i)$. Lembrando que, $\mathbf{S}_{w(l),w(i)}$ corresponde ao número de vezes que o símbolo $w(i)$ ocorre na vizinhança de $w(l)$.

3. Ao final, utilizando a matriz de similaridade $\mathbf{S}$, são determinados os vetores da representação semântica, $\mathbf{g}_n(n =, 1, 2, ...V)$. Essa etapa é feita utilizando a metodologia de otimização apresentada na Seção 4.3.

7.4.2 Contexto baseado em TF-IDF

1. Para cada janela, o vetor de características $\mathbf{x}(l)$ é quantizado como mostrado na Equação 4.1.

2. Em seguida, para o símbolo correspondente a janela l, $w(l)$, é calculado $TFIDF_{w(l),c(i)}$, em que $c(i)$ corresponde a i-ésima classe emotiva, como mostrado na Equação 4.3. Nessa etapa, o objetivo é determinar a relevância do símbolo $w(l)$ para cada classe $c(i)(i = 1, 2, ...C)$.

3. Em seguida, para cada classe classe, são determinados Ns símbolos com os maiores TF-IDF, formando a matriz $\mathbf{B}$ apresentada na Equação 4.4, na Seção 4.2.2.

4. Feito isso, é determinada a matriz de similaridade $\mathbf{S}$ entre cada símbolo $w \in \mathbf{B}$, como mostrado na Seção 4.2.2.

5. Por fim, utilizando a matriz de similaridade $\mathbf{S}$, são determinados os vetores da representação semântica, $\mathbf{g}_n(n =, 1, 2, ...V)$. Essa etapa é feita utilizando a metodologia de otimização apresentada na Seção 4.3.

É importante destacar que a dimensão de $\mathbf{g}$ é definida de forma arbitrária, para ambos os contextos. Normalmente, quando há intenção em visualizar os vetores obtidos, é utilizado dimensão 2, também é possível visualizar nas dimensões 1 ou 3. Quando não há essa intenção, dimensões maiores podem ser usadas. Neste trabalho diferentes dimensões são testadas.

Além disso, também é relevante ressaltar que para obter os vetores $\mathbf{g}$, são usados somente dados de treino. Durante a classificação, os dados de teste precisam ser mapeados

para a representação semântica. Para isso, considerando um sinal de voz de teste $s(n)$, são executados os seguintes passos:

1. Para cada janela l, quantizar o vetor de características $\mathbf{x}(l)$, como mostrado na Equação 4.1. Assim, o vetor de características para uma determinada janela l, $\mathbf{x}(l)$, é associado a um símbolo $w(l) \in \{1, 2, ..., 2048\}$.

2. Cada símbolo $w_1, w_2, ...w_V$ $(V = 2048)$ está associado a um vetor da representação semântica $\mathbf{g}_n(n = 1, 2, ..., V)$, logo, é possível relacionar $w(l)$ a um determinado vetor $\mathbf{g}_n$.

7.5 Métodos de classificação

Nesta seção, são apresentados os métodos utilizados para a classificação das propriedades extraídas na etapa de extração de características e representação semântica. Para fins comparativos, serão utilizados três métodos diferentes: (1) modelo de mistura de gaussianas (GMM), (2) rede neural artificial (RNA), e (3) k-vizinhos mais próximos (k-NN). Os detalhes a respeito da metodologia de cada técnica são descritos a seguir. Essas técnicas foram escolhidas pois são amplamente utilizadas em trabalhos relacionados [29, 37, 4].

7.5.1 Modelo de mistura gaussianas

Para o GMM, é utilizado o método conhecido como *Maximization-Expectation* para ajustar os parâmetros das gaussianas. Para a inicialização dos centros da gaussianas, são usadas amostras aleatórias do conjunto de dados. É utilizado um modelo GMM para cada classe emotiva, dessa forma, 7 modelos GMM são usados. São utilizadas 32 gaussianas para cada modelo GMM.

O *Expectation-Maximization* (EM) é composto por duas etapas: *Expectation* (E), em que os parâmetros atuais das gaussianas são usados para determinar as probabilidades *a posteriori*, e *Maximization* (M), em que as probabilidades *a posteriori* são utilizadas para atualizar as médias, matrizes de covariâncias e os pesos para cada gaussiana. A seguir, são descritos os passos utilizados para implementar o EM.

1. Inicialmente, são inicializados os centros $\boldsymbol{\mu}_k$ de cada gaussiana k, utilizando uma amostra aleatória do conjunto de dados.

2. Em seguida, são determinadas as matrizes de covariância $\mathbf{\Sigma}_k$ de cada gaussiana k, como mostrado na Equação 7.28, em que $\boldsymbol{\mu}_\mathbf{x}$ representa a média de todos os vetores de características, definido em $\mathbb{R}^{dim_\mathbf{x}}$.

$$\mathbf{\Sigma}_k = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu}_\mathbf{x})(\mathbf{x}_i - \boldsymbol{\mu}_\mathbf{x})^\top \tag{7.28}$$

   É importante lembrar que é utilizado somente a diagonal de $\mathbf{\Sigma}_k$. De acordo com o trabalho de Reynolds et al. [120], a diagonal é capaz de alcançar os mesmos resultados de quando usada a matriz $\mathbf{\Sigma}_k$ completa, além disso, a utilização das diagonais diminui o custo computacional durante o treinamento.

3. Após isso, são calculadas as probabilidades *a posteriori*, $P_k(i)$, para a gaussiana k e o vetor de características $\mathbf{x}_i$, como mostrado na Equação 7.29

$$P_k(i) = \frac{\pi_k \mathcal{N}(\mathbf{x_i}|\mu_k, \mathbf{\Sigma}_k)}{\sum_{j=1}\pi_j \mathcal{N}(\mathbf{x_i}|\mu_j, \mathbf{\Sigma}_j)} \tag{7.29}$$

4. Em seguida, são atualizadas as médias ($\boldsymbol{\mu}_k$), covariâncias ($\mathbf{\Sigma}_k$) e os pesos (π_k) para cada gaussiana, como mostrado nas expressões seguintes.

$$\pi_k^{novo} = \frac{U_k}{N} \tag{7.30}$$

$$\boldsymbol{\mu}_k^{novo} = \frac{1}{U_k}\sum_{i=1}^{N} P_k(i)\mathbf{x}_i \tag{7.31}$$

$$\mathbf{\Sigma}_k^{novo} = \frac{1}{U_k}\sum_{i=1}^{N} P_k(i)(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\intercal \tag{7.32}$$

$$U_k = \sum_{i=1}^{N} P_k(i) \tag{7.33}$$

5. Por fim, é determinado a verossimilhança $L(\mathbf{x}, \lambda)$, como mostrado na Equação a seguir, em que N corresponde ao número de vetores.

$$L(\mathbf{x}, \lambda) = \sum_{i=1}^{N} ln\Big\{\sum_{k=1}^{K}\pi_k \mathcal{N}(\mathbf{x_i}|\mu_\mathbf{k}, \mathbf{\Sigma}_k)\Big\} \tag{7.34}$$

   Para o treinamento, os passos 3, 4 e 5 são repetidos 20 vezes. É ajustado um GMM para cada classe emotiva, dessa forma, cada modelo, com k gaussianas, modela um conjunto de dados correspondente a uma emoção.

   Para a classificação de um vetor de características $\mathbf{x}_i$, é determinada a verossimilhança $L(\mathbf{x}_i, \lambda)$, para cada modelo GMM, o modelo que obtiver a maior verossimilhança é apontado como resultado.

O GMM foi implementado no software MATLAB [121], de acordo como descrito nessa seção.

7.5.2 Rede neural artificial

Para a RNA, é utilizada uma camada oculta com 100 neurônios e função de ativação *ReLu*. Na camada de saída da rede é utilizada a função de ativação *softmax* (Equação 5.7), adequada para problemas de de classificação multiclasse [71].

A saída é rotulada na forma de codificação *one-hot*, muito utilizada para lidar com dados categóricos. Nessa codificação, um grupo de *bits* pode ter somente um elemento com valor 1, os outros devem ser 0. Dessa forma, para a classificação de 7 emoções, a codificação pode ser feita como mostrado na Tabela 3. Essa codificação funciona bem em conjunto com a função de ativação *softmax* na saída [72].

Tabela 3 – Codificação *one-hot* para as emoções.

Emoção	**Código**
Felicidade	1000000
Medo	0100000
Neutralidade	0010000
Tristeza	0001000
Desgosto	0000100
Cansaço	0000010
Raiva	0000001

Para a otimização da rede neural, sera utilizado o método de *backpropagation*, essa é uma técnica muito utilizada, pois é capaz de otimizar de forma eficiente os parâmetros da rede neural. Será utilizada a função de custo da entropia cruzada categórica. Segundo diversos autores, essa função de custo pode ser mais apropriada para problemas de classificação multiclasse, sendo capaz de melhorar a performance e reduzir o tempo de otimização dos parâmetros [122, 71]. A função é apresentada na seção 5.2, na Equação 5.8.

Através do *backpropagation*, os parâmetros $\mathbf{W}$ são atualizados utilizando o gradiente descendente da função de custo $J(\mathbf{W})$ com relação a cada parâmetro W_{ij}, como mostrado na Equação 7.35, onde α corresponde ao passo de adaptação.

$$W_{ij} \leftarrow W_{ij} - \alpha \nabla_{W_{ij}} J(\mathbf{W}) \tag{7.35}$$

A implementação para a rede neural artificial pode ser resumida nos seguintes passos:

1. Rotular em codificação *one-hot* as observações de treinamento e guardar seus vetores de características.

2. Otimizar a rede neural com os dados treino, usando as configurações para a rede neural descritas anteriormente.

Neste trabalho, é usada a implementação do rede neural artificial da biblioteca *scikit-learn* [123] para a linguagem de programação Python.

7.5.3 k-vizinhos mais próximos

Para o k-NN, será utilizado a distância euclidiana como medida de distância. O critério de desempate utilizado é baseado no aumento de vizinhos, ou seja, caso haja empate entre duas ou mais classes, é escolhida outra amostra mais próxima para desempatar. Em resumo, a implementação pode ser executada com os seguintes passos:

1. Rotular as observações de treinamento e guardar seus vetores de características.
2. Calcular a distância euclidiana da observação nova ($\mathbf{x}_i$) para todas as observações de treinamento.
3. Determinar as k observações com menores distâncias.
4. Escolher a classe que mais acontece entre as k observações mais próximas.
5. Em caso de empate, escolher outra observação mais próxima para desempatar.

Neste trabalho, é usada a implementação do k-NN da biblioteca *scikit-learn* [123] para a linguagem de programação Python.

Para todos os métodos de classificação, a classificação de um determinado sinal de voz com N vetores de características é feita classificando cada vetor, a classe com mais vetores é apontada como resultado.

Capítulo

8 Resultados e discussões

Nesta capítulo, são apresentados os resultados e discussões dos experimentos realizados. Os experimentos estão divididos de acordo com o contexto utilizado para a representação semântica, e são apresentados na seguinte ordem: (1) contexto baseado em vizinhança, e (2) contexto baseado no TF-IDF. Com o intuito de comparar os resultados obtidos na representação semântica, também são apresentados os resultados obtidos com a representação natural (representação original dos dados).

A avaliação das representações (natural e semântica) é feita de acordo com a acurácia de classificação. A classificação é realizada para cada arquivo de áudio, um arquivo de áudio possui somente uma classe emotiva correta (raiva, tédio, nojo, medo, alegria, tristeza ou neutralidade).

Ao total, neste trabalho, existem 498 arquivos de áudio, que são divididos em dois grupos: (1) arquivos de treino, e (2) arquivos de teste. Os arquivos de treino são usados para ajustar a representação semântica, e também para otimizar os parâmetros dos classificadores. Os arquivos de teste são usados apenas para avaliar a acurácia para cada classificador. Além disso, os arquivos de treino e teste correspondem a 75% e 25% do total de arquivos de áudio, respectivamente.

É utilizada a estratégia de classificação independente do orador, ou seja, os oradores presentes nos arquivos de treino são diferentes dos oradores dos arquivos de teste. O conjunto de dados usado possui 10 oradores, são utilizados 4 grupos, em cada grupo, são selecionados 7 oradores para treino, e 3 oradores para teste, como descrito na Tabela 4 mostrada a seguir (cada orador é representado por um número). Dessa forma, é possível testar todos os oradores da base. Na Figura 8.1, é apresentado o esquema de classificação proposto nesse trabalho.

A acurácia, ac, é determinada dividindo o número de áudios corretamente classificados, $n_{acertos}$, pelo número total de áudios classificados, n_{total}, como mostrado na Equação

Tabela 4 – Oradores utilizados para cada grupo.

Grupo	Oradores para treino	Oradores para teste
1	3, 9, 10, 11, 13, 15 e 16	8, 12 e 14
2	8, 10, 11, 12, 13, 14 e 16	3, 9 e 15
3	3, 8, 9, 12, 14, 15 e 16	10, 11 e 13
4	3, 8, 10, 11, 13, 14 e 15	9, 12 e 16

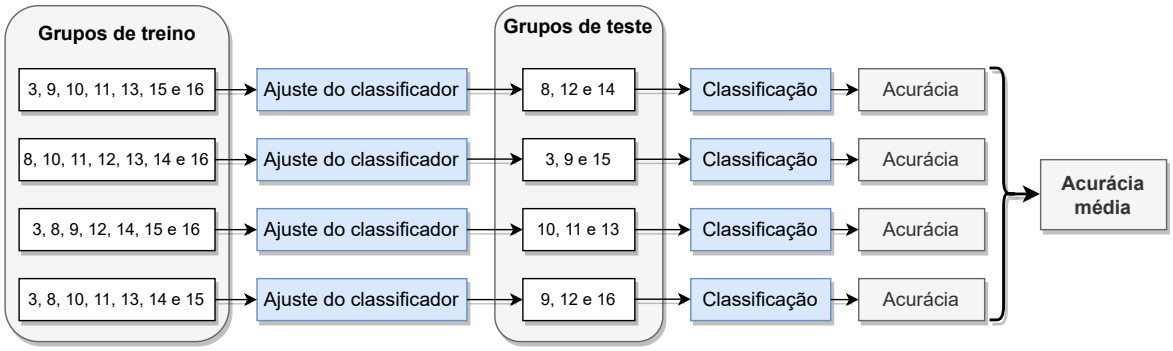


Figura 8.1 – Esquema para a classificação independente do orador. Para cada grupo, é ajustado os parâmetros do classificador, em seguida, o classificador é testado com o grupo de teste para obter a acurácia. Por fim, a acurácia média é determinada.

8.1.

$$ac = \frac{n_{acertos}}{n_{total}} \tag{8.1}$$

8.1 Resultados para o contexto baseado em vizinhança

8.1.1 Primeiro experimento

No primeiro experimento, é usada a representação semântica com 2 dimensões e 100 vizinhos (50 anteriores e 50 posteriores). Os resultados são divididos de acordo com o classificador utilizado. O objetivo desse experimento é avaliar de forma geral o desempenho da representação natural e semântica, para cada classificador, e para cada vetor de características, observando se a representação semântica é capaz de trazer alguma informação relevante a respeito das emoções veiculadas nos sinais de voz.

Na Tabela 5, são apresentadas as acurácias utilizando os seguintes classificadores: Modelo de mistura de gaussianas (GMM), Rede neural artificial (RNA) e k-Vizinhos mais próximos (k-NN). É importante destacar que devido a inicialização aleatória do GMM e da RNA, os resultados podem variar de um experimento para outro, por isso é calculada a média de 3 experimentos para esses dois classificadores. Além disso, lembramos que a

prosódia corresponde a 5 características: $f0$, *jitter*, *shimmer*, energia e taxa de cruzamento por zero. Para o k-NN, é usado $k = 1$, já para os outros classificadores, são utilizadas as mesmas configurações apresentadas no Capítulo 7.

É possível observar que, no geral, a acurácia para a representação natural é superior à média da acurácia para a representação semântica. Baseado nisso, uma hipótese pode ser considerada: os parâmetros utilizados na representação semântica podem não ter sido capazes de capturar, de forma útil, a informação temporal incorporada aos sinais de voz. Estes parâmetros são: janela de contexto (número de vizinhos) e dimensão da representação semântica. Variar esses dois parâmetros é essencial para descobrir o efeito de cada um deles, isso é feito no próximo experimento.

Tabela 5 – Acurácia para três classificadores diferentes (GMM, RNA e k-NN), para cada vetor características, usando o contexto semântico baseado em vizinhança.

Acurácia para GMM

Modalidade	Prosódia	MFCC	LFPC	Média
Representação natural	42,27	50,51	52,59	**48,46**
Representação semântica	21,84	38,06	42,06	**33,99**

Acurácia para RNA

Modalidade	Prosódia	MFCC	LFPC	Média
Representação natural	37,32	39,96	42,42	**39,90**
Representação semântica	28,97	46,42	32,58	**35,99**

Acurácia para k-NN

Modalidade	Prosódia	MFCC	LFPC	Média
Representação natural	44,08	47,55	45,39	**45,67**
Representação semântica	12,53	19,27	12,53	**14,78**

8.1.2 Avaliação dos parâmetros do contexto de vizinhança

Como observado no primeiro experimento, é necessário avaliar o efeito dos parâmetros do contexto baseado em vizinhança: (1) número de dimensões, e (2) janela de contexto. Para isso, foi repetido o experimento utilizando a RNA como classificador e o MFCC como vetor de características, esse foi escolhido por resultar no melhor desempenho para a representação semântica (como mostrado na Tabela 5).

8.1.2.1 Número de dimensões

Na Figura 8.2, é apresentado um gráfico que compara a acurácia entre a representação semântica (variando de acordo com o número de dimensões), e a média da acurácia para a representação natural, usando o vetor de características MFCC e RNA

como classificador. Observando o gráfico, é possível notar que aumentar a dimensão da representação semântica não resulta, necessariamente, em um aumento na acurácia. Uma possível explicação para isso pode está relacionada às dificuldades em se obter um modelo de classificação confiável quando é usado um conjunto de dados com alta dimensão (maldição da dimensionalidade). Segundo alguns autores [71, 72], aumentar a dimensão de um conjunto de dados requer um aumento também no número de amostras de treinamento para que o modelo de classificação seja confiável, isso também ajuda a explicar o motivo da acurácia não ter aumentado quando a dimensão da representação semântica aumenta.

Além disso, também é possível observar que quando a dimensão da representação semântica é igual a 1, a acurácia obtém seu menor valor. Uma hipótese para isso é que, na dimensão 1, é mais difícil otimizar os vetores da representação semântica, o que resulta em uma representação má ajustada, que não bem captura as relações desejadas.

Outro ponto importante é que a acurácia atinge seu maior valor quando o número de dimensões é igual a 2. Uma possível explicação para isso é que a dimensão intrínseca da representação semântica seja igual a 2, ou seja, a informação que a representação semântica carrega pode ser representada em apenas 2 dimensões.

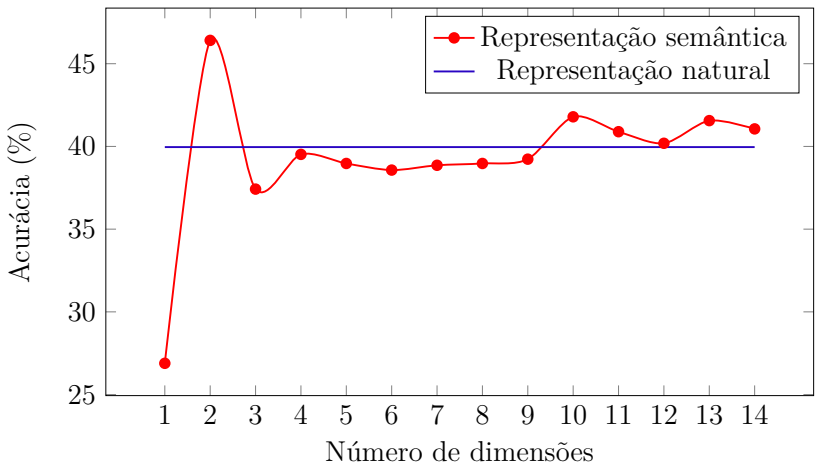


Figura 8.2 – Comparação entre acurácias para a representação semântica ao variar a dimensão da representação semântica, e a média da acurácia para a representação natural. São usados 12 coeficientes MFCC e a rede neural artificial como classificador. O número de vizinhos do contexto semântico é 100 para todos os experimentos.

8.1.2.2 Janela de contexto

Na Figura 8.3, é mostrada a acurácia para a representação semântica ao variar o tamanho da janela do contexto baseado em vizinhança. De acordo com o experimento, é possível observar que aumentar o tamanho da janela não aumenta monotonicamente, a acurácia.

Ao observar a Figura 8.3, percebemos que a acurácia atinge o valor máximo quando é usada uma janela com 200 vizinhos (100 anteriores e 100 posteriores). Esses 200 vizinhos correspondem a aproximadamente uma janela de 2 segundos do sinal de voz. A duração média de cada sinal de voz da base EmoDB é 2,8 segundos. Levando em consideração que algumas janelas são removidas pelo VAD (que diminui a duração média de cada áudio), podemos perceber que a duração da janela de vizinhança (2 segundos) e duração média de cada áudio (∼2,8 segundos) estão próximas. Em sinais de voz com duração menor ou igual a 2 segundos, todos os símbolos são considerados como vizinhos, quando são utilizados 200 vizinhos no contexto baseado em vizinhança.

Ao aumentar o número de vizinhos no contexto de vizinhança, é esperado que a representação semântica seja capaz de capturar dinâmicas temporais (ou padrões sequenciais) maiores. Nesse experimento, observamos que usar um número de vizinhos que corresponda a uma janela cuja duração seja próxima a duração média de cada sinal de voz parece ser melhor.

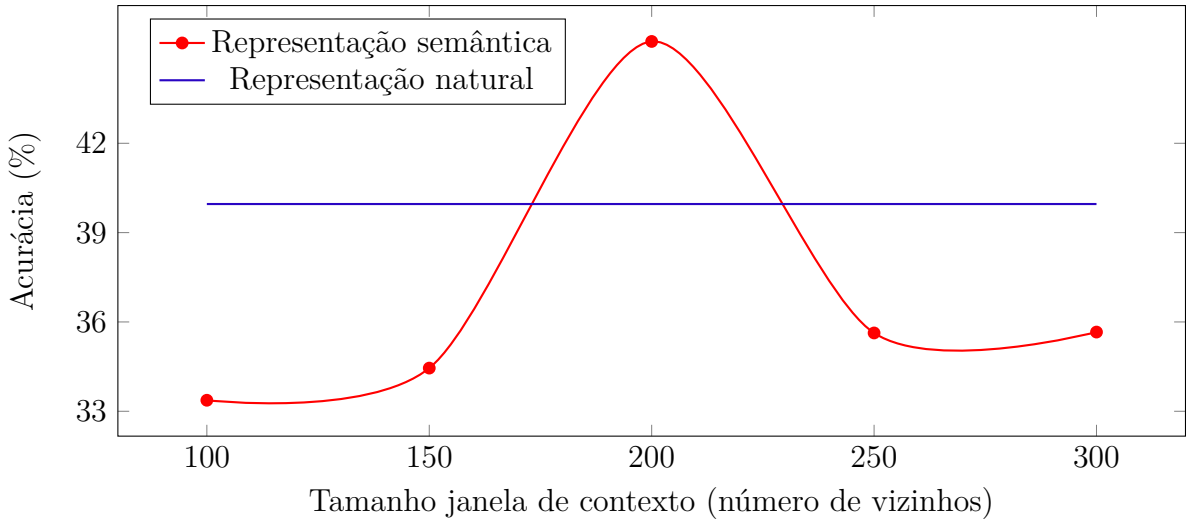


Figura 8.3 – Comparação entre acurácias para a representação semântica ao variar o tamanho da janela de contexto, e a média da acurácia para a representação natural. São usados 12 coeficientes MFCC e a rede neural artificial como classificador. A dimensão é 2 para todos os experimentos.

8.1.3 Avaliação do k-NN

Ao observar a Tabela 5, é possível perceber que a média da acurácia para o classificador k-NN (com $k = 1$) é muito menor que a acurácia da mesma representação em outros classificadores. Uma possível explicação para isso é que o valor de k, que corresponde ao número de vizinhos do k-NN, não está adequado para a representação semântica. Por

isso, nesse experimento, o número de vizinhos do k-NN é variado para observar o efeito deste parâmetro na classificação.

Na Figura 8.4, é apresentada uma comparação entre as acurácias para cada representação ao variar o número de vizinhos do classificador k-NN. É possível notar que a acurácia para a representação natural varia muito pouco, entre 45% e 49%, para todos os valores de k avaliados. Já para a representação semântica, a acurácia cresce de forma acentuada de acordo com o número de vizinhos, chegando a atingir 59,86% quando $k = 33$.

Percebemos que o valor de k pode ter um grande impacto na classificação. Geralmente, k é variado e o valor que resultar na melhor acurácia é selecionado. Ao aumentar o valor de k, estamos utilizando estimadores de densidade mais regularizados (i.e. funções mais suaves). Uma explicação para que a acurácia tenha atingido valores baixos para valores pequenos de k está associada a distribuição das amostras de cada classe. Caso existam classes com amostras sobrepostas, usar um k muito pequeno deve levar a erros, pois o algoritmo não é capaz de analisar bem os vizinhos da amostra de teste, como mostrado na Figura 5.2 da seção 5.3.

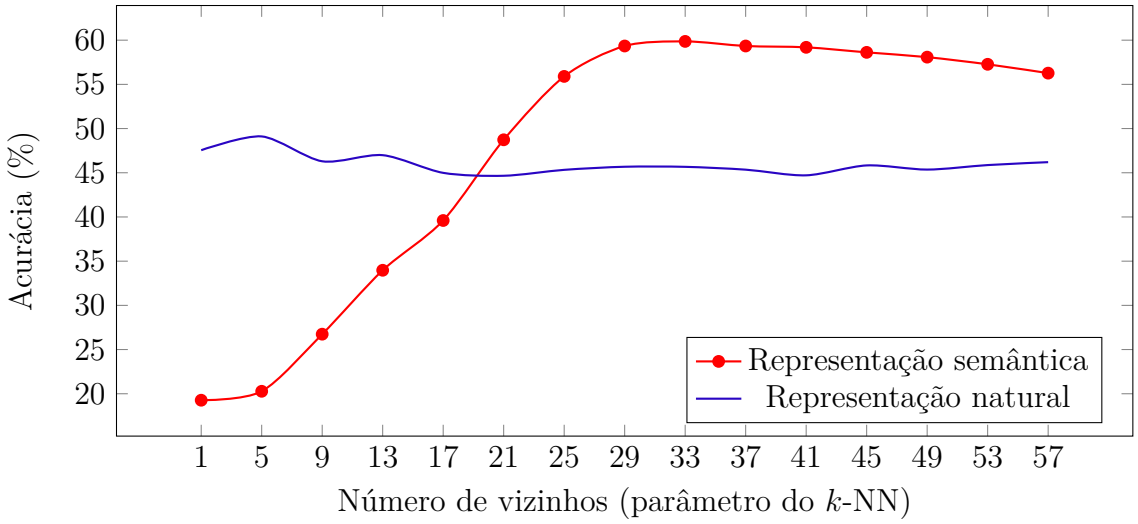


Figura 8.4 – Comparação entre acurácias para a representação semântica e natural ao variar o número de vizinhos, k, do o classificador k-NN. São usados 12 coeficientes MFCC como características. O número de vizinhos para o contexto baseado em vizinhança é 100, e a dimensão da representação semântica é 2 para todos os experimentos.

8.2 Resultados para o contexto baseado no TF-IDF

8.2.1 Primeiro experimento

Neste experimento, é usada a representação semântica com 2 dimensões e 500 símbolos com maiores TF-IDF por classe (Ns). Os resultados são divididos de acordo com o classificador utilizado. O objetivo desse experimento é avaliar de forma geral o desempenho da representação natural e semântica, para cada classificador, e para cada vetor de características.

Na Tabela 6, são apresentadas as acurácias utilizando os seguintes classificadores: GMM, RNA e k-NN. Para o k-NN, é usado $k = 1$, para os outros classificadores, são utilizadas as mesmas configurações apresentadas no Capítulo 7.

É possível observar que, no geral, a acurácia para a representação natural é superior a média da acurácia para a representação semântica, o mesmo padrão observado para o contexto baseado em vizinhança. Isso nos leva a considerar a hipótese de que os parâmetros associados ao contexto TF-IDF (dimensão e número de símbolos com maiores TF-IDF por classe) não foram capazes de capturar a informação semântica desejada. Ao variar esses parâmetros, podemos observar o efeito de cada um para a classificação de emoções, isso é realizado no próximo experimento.

Tabela 6 – Acurácia para três classificadores diferentes (GMM, RNA e k-NN), para cada vetor características, usando o contexto semântico baseado no TF-IDF.

Acurácia para vetor de características usando GMM

Modalidade	Prosódia	MFCC	LFPC	Média
Representação natural	43,87	50,32	53,13	**49,10**
Representação semântica	24,90	58,39	44,97	**42,75**

Acurácia para vetor de características usando RNA

Modalidade	Prosódia	MFCC	LFPC	Média
Representação natural	35,57	40,00	39,65	**38,41**
Representação semântica	31,62	35,69	35,05	**34,12**

Acurácia para vetor de características usando k-NN

Modalidade	Prosódia	MFCC	LFPC	Média
Representação natural	44,08	47,40	45,39	**45,63**
Representação semântica	12,53	23,92	12,53	**16,33**

8.2.2 Avaliação dos parâmetros do contexto TF-IDF

Como discutido anteriormente, é preciso avaliar o efeito da variação dos parâmetros do contexto TF-IDF: (1) número de dimensões, e (2) número de símbolos com maiores TF-IDF (Ns). Para isso, foi repetido o experimento utilizando o GMM como classificador e o MFCC como vetor de características, esse foi escolhido por resultar no melhor desempenho para a representação semântica (vide Tabela 6).

8.2.2.1 Número de dimensões

Na Figura 8.5, é apresentado um gráfico que compara a acurácia entre a representação semântica (variando de acordo com o número de dimensões), e a média da acurácia para a representação natural usando 12 coeficientes MFCC e GMM como classificador. Podemos notar os seguintes pontos: (1) a acurácia não melhora ao aumentar a dimensão, (2) a acurácia é maior quando a dimensão é igual a 1.

De forma semelhante ao observado no contexto de vizinhança, a acurácia não cresce ao aumentar o número de dimensões da representação semântica, e os motivos para isso devem ser os mesmos: problemas relacionados a maldição da dimensionalidade.

Além disso, a acurácia é maior quando a dimensão é igual a 1 (71,99%). Isso indica que a informação que a representação semântica carrega pode ser melhor representada em apenas uma dimensão. Este resultado é surpreendente, pois indica que a representação semântica baseada no TF-IDF é capaz de capturar elementos relevantes a respeito de cada classe emotiva e consegue representar essas informações em uma dimensão muito menor que a dimensão original dos dados.

8.2.2.2 Número de símbolos com maiores TF-IDF por classe

Na Figura 8.6, é mostrada uma comparação entre a representação semântica (ao variar Ns), e a média da representação natural. Podemos observar que a acurácia cresce de forma acentuada ao aumentar Ns, chegando em 73,81% quando $Ns = 700$.

Como já explicado, o parâmetro Ns controla o número de símbolos mais relevantes para cada classe (através do TF-IDF). Ao aumentar Ns, é esperado que a representação semântica seja capaz de carregar mais informações importantes sobre cada classe, isso deve ser favorável no processo de classificação, visto que, essas informações são essenciais na discriminação entre as classes emotivas.

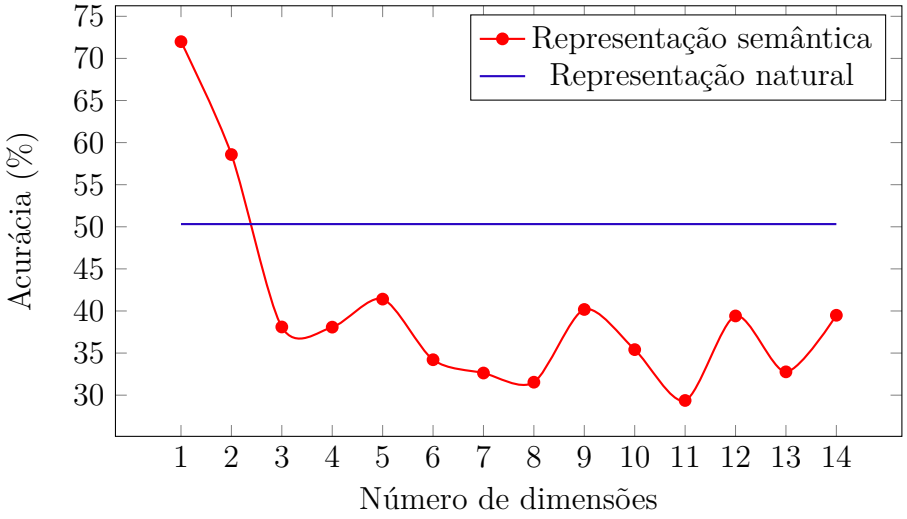


Figura 8.5 – Comparação entre acurácias para a representação semântica ao variar a dimensão da representação semântica, e a média da acurácia para a representação natural. São usados 12 coeficientes MFCC como características e o GMM como classificador. O número de vizinhos é 100 para todos os experimentos.

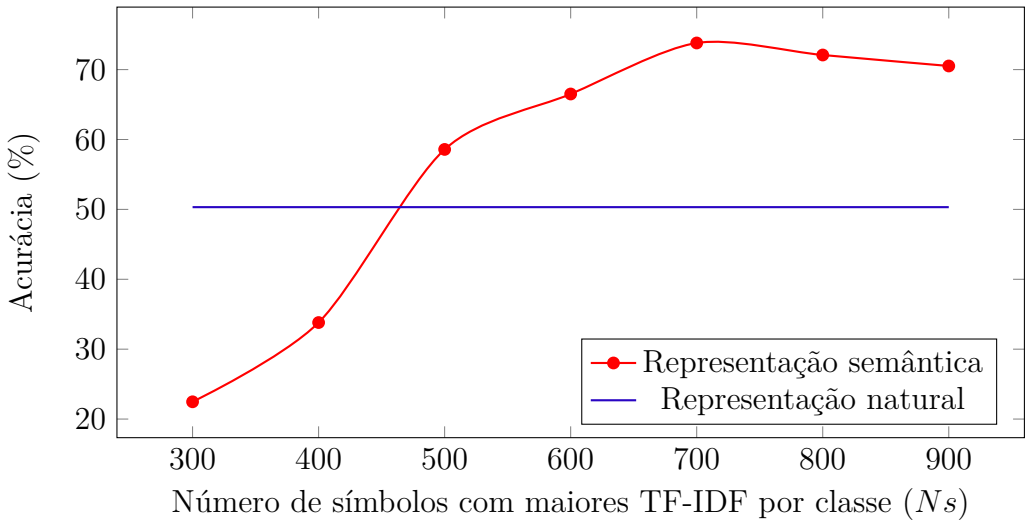


Figura 8.6 – Comparação entre acurácias para a representação semântica ao variar o número de símbolos com maiores TF-IDF por classe, e a média da acurácia para a representação natural. São usados 12 coeficientes MFCC como características e o GMM como classificador. A dimensão é 2 para todos os experimentos.

8.2.3 Avaliação do k-NN

De forma semelhante ao ocorrido no contexto baseado em vizinhança, o classificador k-NN também apresenta acurácia menor para a representação semântica, quando comparado aos outros classificadores. Como explicado anteriormente, é possível que o valor de k não está adequado para a representação semântica.

Na Figura 8.7, é apresentada uma comparação entre a acurácia para as duas representações ao variar o número de vizinhos do k-NN. É possível observar que a acurácia, para a representação natural, não varia muito, ficando entre 45% e 49%. Já para a representação semântica, a acurácia cresce bastante ao aumentar o número de vizinhos, chegando até 47.59% quando $k = 29$. O motivo da acurácia crescer ao aumentar k deve ser o mesmo apresentado para o contexto de vizinhança, ou seja, usar um k muito pequeno pode resultar em erros de classificação, porque as distribuições de probabilidades condicionadas às classes não estão bem representadas na vizinhança de cada amostra de teste.

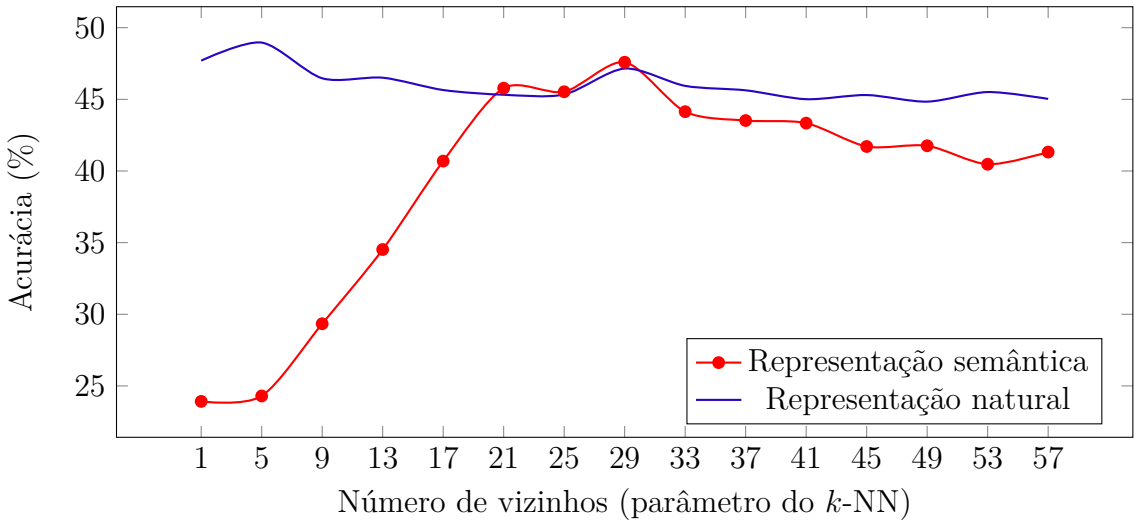


Figura 8.7 – Comparação entre acurácias para a representação semântica e natural ao variar o número de vizinhos, k, do o classificador k-NN. São usados 12 coeficientes MFCC. O número de símbolos com maiores TF-IDF por classe é 500, e a dimensão da representação semântica é 2 para todos os experimentos.

8.3 Comentários finais

Ao analisar os experimentos, podemos perceber que as características de prosódia apresentam as menores acurácias para todos os experimentos, em contrapartida, é notável o destaque das características espectrais, sobretudo dos coeficientes MFCC. Os resultados ajudam a comprovar que as características espectrais são capazes de representar, melhor que as demais características aqui estudadas, as informações importantes sobre as emoções expressadas nos sinais de voz, isso também pode ser observado em diversos outros trabalhos relacionados a classificação de emoções através da voz, que fizeram comparações entre as características espectrais e as características relacionadas à prosódia [106, 82, 124].

Também é possível observar que o modelo de mistura de gaussianas apresenta resultados superiores em quase todos experimentos. Uma possível explicação para isso pode estar relacionada à capacidade da mistura de gaussianas em modelar com restrições

as distribuições de probabilidade, onde essas restrições parecem se adequar aos tipos de dados observados no problema em questão. É válido destacar que o GMM tem sido usado extensivamente em diferentes problemas envolvendo a classificação de características vocais (em especial, identificação ou reconhecimento de orador e classificação de emoções) desde os anos 1990, tornando-se um dos métodos de classificação mais comuns para esses problemas [120, 4, 29].

Na Figuras 8.8 e 8.9, é apresentado como os sinais de áudio estão dispostos na representação semântica, com o contexto baseado em vizinhança (Figura 8.8), e o contexto baseado em TF-IDF (Figura 8.9). São usadas as configurações que resultaram nas melhores acurácias para ambos os contextos. É possível observar que existe uma certa sobreposição entre algumas classes, nos dois contextos.

Ao observar a Figura 8.8, percebemos que, mesmo que contexto baseado em vizinhança não leve em consideração a classe, há um padrão de agrupamento entre as emoções. Já na Figura 8.9, o contexto baseado no TF-IDF considera as classes e isso é refletido em sua representação, é possível observar que as classes raiva, alegria e aversão estão bem agrupadas.

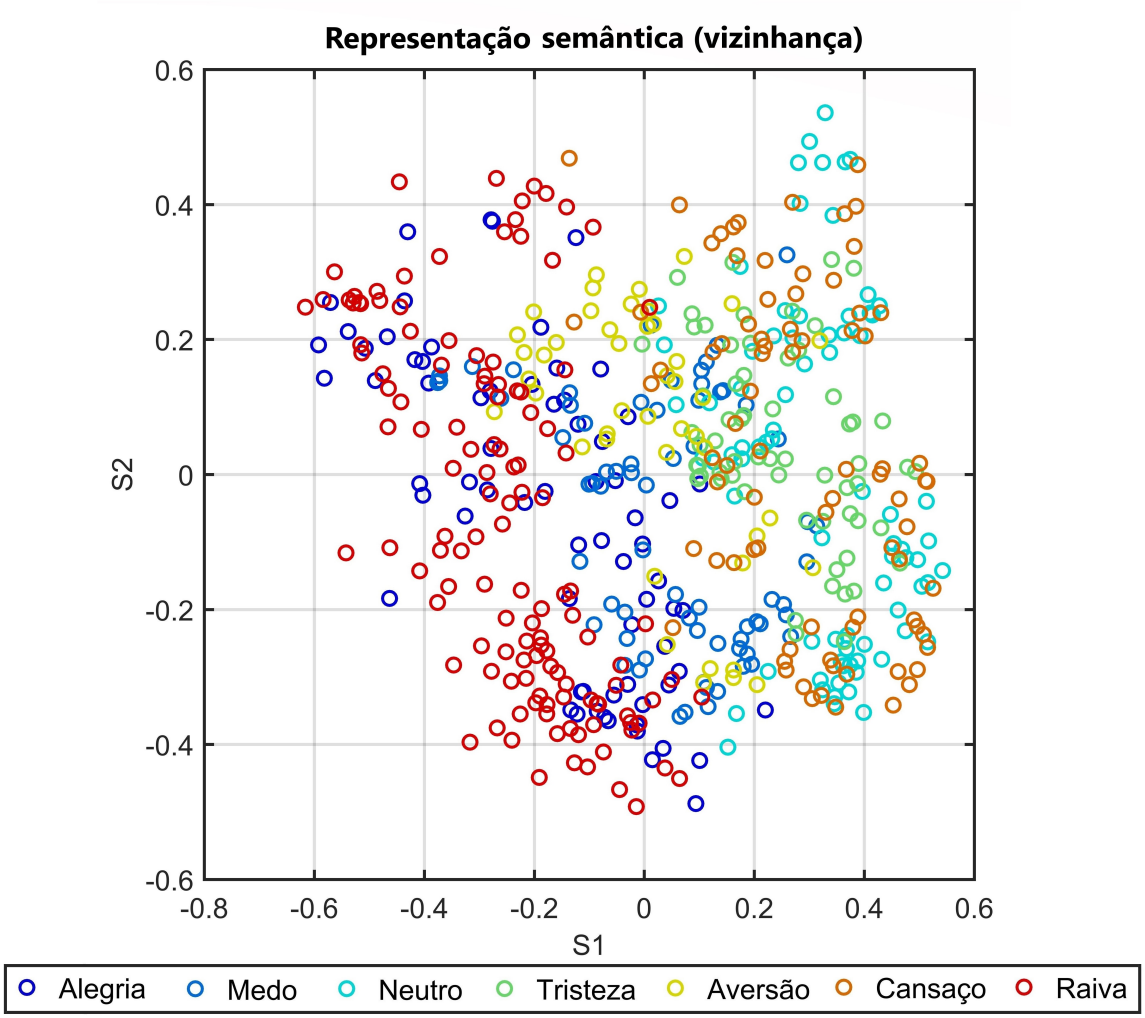


Figura 8.8 – Disposição dos sinais na representação semântica com contexto baseado na vizinhança, usando 12 MFCCs e 200 vizinhos. Cada círculo representa a média de todos vetores de um áudio.

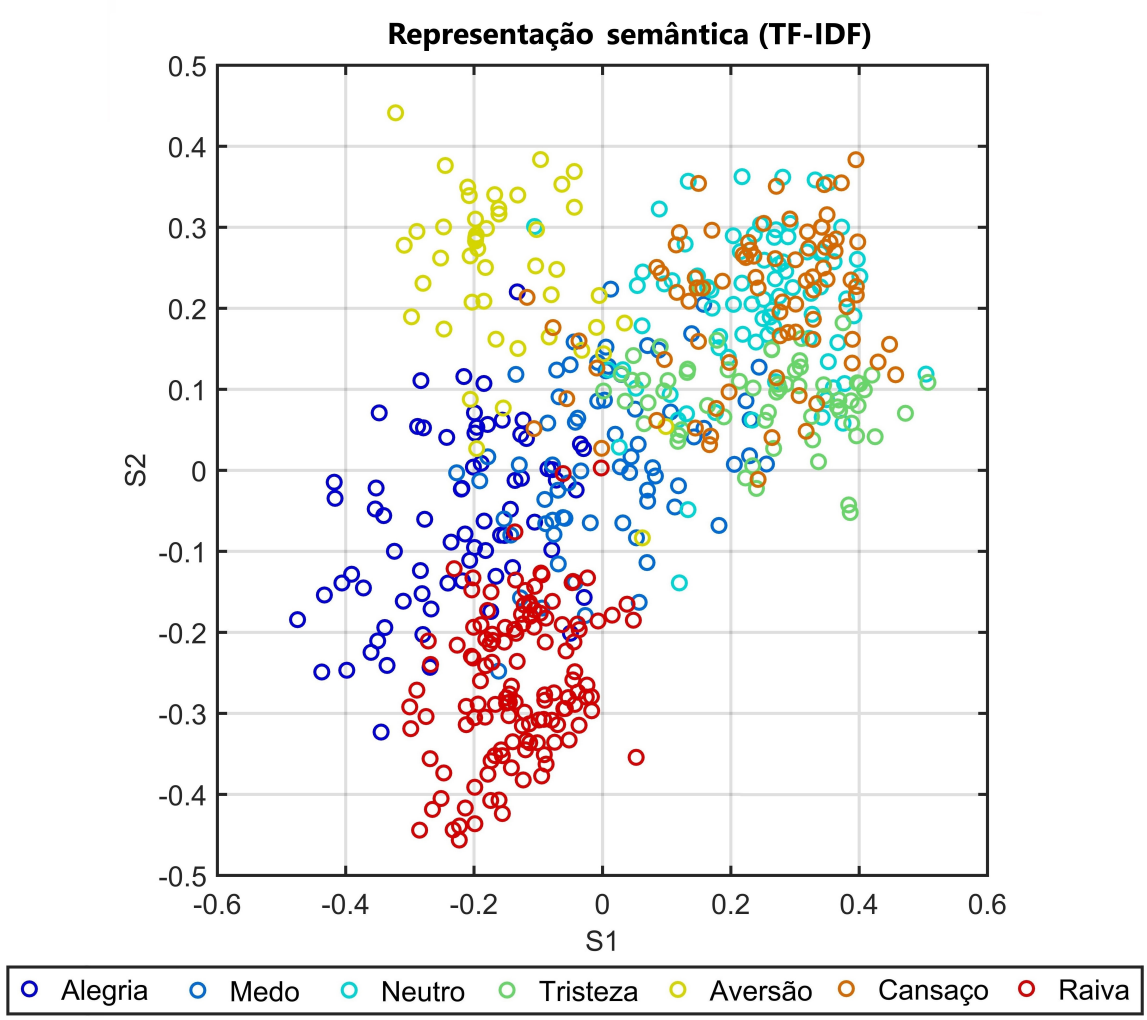


Figura 8.9 – Disposição dos sinais na representação semântica com contexto baseado no TF-IDF, usando 12 MFCCs e 700 símbolos por classe. Cada círculo representa a média de todos vetores de um áudio.

Capítulo

9 Conclusão

Neste trabalho, sinais de voz foram analisados de diferentes formas, utilizando metodologias clássicas para a extração de características, com o objetivo de classificar as emoções expressadas nos sinais de voz. Foram exploradas técnicas amplamente utilizadas de pré-processamento, extração de características e classificação. O trabalho também apresentou duas metodologias baseadas na representação semântica: (1) contexto baseado em vizinhança, (2) e contexto baseado em TF-IDF. Foram investigados e discutidos os efeitos das duas metodologias no processo de classificação de emoções através de sinais de fala.

De acordo com os resultados obtidos no presente trabalho, podemos observar que, em casos específicos, a utilização da representação semântica é capaz de melhorar bastante a discriminação entre as emoções, apresentando acurácias maiores no processo de classificação, quando comparada a representação natural. Isso indica que, quando ajustada corretamente, a representação semântica é capaz de carregar informações relevantes a respeito do sinal de voz. Além disso, os parâmetros utilizados para determinar a representação semântica (número de dimensões, tamanho da janela ou número de símbolos por classe) têm um impacto significativo na classificação de emoções.

Quando comparadas as duas metodologias de representação semântica utilizadas, concluímos que o contexto baseado no TF-IDF, proposto originalmente neste trabalho, destaca-se na tarefa de classificação de emoções, provavelmente pelo seu potencial em identificar elementos relevantes para cada emoção.

Com relação às características utilizadas, notamos que as características associadas ao espectro apresentam melhor desempenho no processo de classificação, quando comparadas às características relacionadas à prosódia, esse fenômeno também pode ser observado em diversos outros trabalhos relacionados. Já com relação aos classificadores, observamos que o modelo de mistura de gaussianas apresenta resultados superiores quando comparado à rede neural artificial e o k-vizinhos mais próximos.

Ainda com relação aos classificadores, também podemos concluir que o número de vizinhos utilizados no método k-NN é muito significativo para a classificação utilizando a representação semântica. No geral, usando a representação semântica, ao aumentar o número de vizinhos, é observado um aumento significativo da acurácia.

Como sugestão para trabalho futuros, podem ser utilizadas metodologias para combinar a representação natural e semântica. É esperado que ao combinar as representações, de forma cuidadosa, acurácias maiores possam ser obtidas. Além disso, observamos que a quantização vetorial é fundamental para a obtenção da representação semântica, por isso, torna-se interessante investigar o efeito do nível de quantização, bem como as implicações da utilização de outras técnicas de quantização vetorial.

Também observamos que ainda é necessário elaborar metodologias para ajustar automaticamente os parâmetros da representação semântica para que a acurácia da classificação aumente. Além disso, também é interessante que trabalhos futuros avaliem os efeitos da representação semântica com outros métodos de classificação.

Os algoritmos utilizados neste trabalho estão publicados em um repositório público no GitHub (`https://github.com/victoribeir0/emorec_repsem`). Os algoritmos foram escritos usando as linguagens de programação Matlab e Python.

Referências

[1] M. Leo, M. Del Coco, P. Carcagni, C. Distante, M. Bernava, G. Pioggia, and G. Palestra, "Automatic emotion recognition in robot-children interaction for asd treatment," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 145–153, 2015. Citado na página 1.

[2] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots," *International Journal of Social Robotics*, vol. 11, no. 4, pp. 555–573, 2019. Citado na página 1.

[3] M. Bojanić, V. Delić, and A. Karpov, "Call redistribution for a call center based on speech emotion recognition," *Applied Sciences*, vol. 10, no. 13, p. 4653, 2020. Citado na página 1.

[4] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021. Citado nas páginas 1, 50 e 64.

[5] A. Konar and A. Chakraborty, *Emotion recognition: A pattern analysis approach*. John Wiley & Sons, 2015. Citado na página 2.

[6] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013. Citado nas páginas 2, 16, 21 e 41.

[7] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Language and cognitive processes*, vol. 25, no. 7-9, pp. 905–945, 2010. Citado na página 2.

[8] E. Noth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "Verbmobil: The use of prosody in the linguistic components of a speech understanding system," *IEEE Transactions on Speech and Audio processing*, vol. 8, no. 5, pp. 519–532, 2000. Citado na página 2.

[9] J. Cole, "Prosody in context: a review," *Language, Cognition and Neuroscience*, vol. 30, no. 1-2, pp. 1–31, 2015. Citado na página 2.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001. Citado nas páginas 2, 8, 14 e 15.

[11] F. Günther, L. Rinaldi, and M. Marelli, "Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions," *Perspectives on Psychological Science*, vol. 14, no. 6, pp. 1006–1033, 2019. Citado nas páginas 3 e 18.

[12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014. Citado nas páginas 3, 18, 20, 21 e 35.

[13] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of artificial intelligence research*, vol. 37, pp. 141–188, 2010. Citado nas páginas 3 e 20.

[14] K. R. Scherer, "Emotion as a multicomponent process: a model and some cross-cultural data.," *Review of personality & social psychology*, 1984. Citado na página 5.

[15] J. De Houwer and D. Hermans, "Cognition and emotion: Reviews of current research and theories," 2010. Citado nas páginas 5, 6 e 7.

[16] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1-2, pp. 5–32, 2003. Citado na página 5.

[17] K. Mulligan and K. R. Scherer, "Toward a working definition of emotion," *Emotion Review*, vol. 4, no. 4, pp. 345–357, 2012. Citado na página 5.

[18] C. Darwin, *The expression of the emotions in man and animals*. University of Chicago press, 2015. Citado na página 5.

[19] R. R. Cornelius, *The science of emotion: Research and tradition in the psychology of emotions.* Prentice-Hall, Inc, 1996. Citado nas páginas 6 e 7.

[20] P. Ekman, "Facial expressions of emotion: New findings, new questions," 1992. Citado nas páginas 6 e 7.

[21] R. R. Cornelius, "Theoretical approaches to emotion," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000. Citado na página 6.

[22] J. A. Russell, J.-A. Bachorowski, J.-M. Fernández-Dols, *et al.*, "Facial and vocal expressions of emotion," *Annual review of psychology*, vol. 54, no. 1, pp. 329–349, 2003. Citado na página 6.

[23] W. James, F. Burkhardt, F. Bowers, and I. K. Skrupskelis, *The principles of psychology*, vol. 1. Macmillan London, 1890. Citado na página 6.

[24] M. Power and T. Dalgleish, "Cognition and emotion: From order to disorder," 2015. Citado na página 7.

[25] M. B. Arnold, "Emotion and personality.," 1960. Citado na página 7.

[26] S. Schachter, "The interaction of cognitive and physiological determinants of emotional state," in *Advances in experimental social psychology*, vol. 1, pp. 49–80, Elsevier, 1964. Citado na página 7.

[27] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *Quarterly Journal of Experimental Psychology*, vol. 63, no. 11, pp. 2251–2272, 2010. Citado nas páginas 7 e 8.

[28] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003. Citado na página 7.

[29] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020. Citado nas páginas 8, 11, 12, 14, 15, 25, 31, 32, 33, 50 e 64.

[30] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980. Citado na página 8.

[31] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, 2007. Citado na página 8.

[32] A. Mehrabian, "Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies," 1980. Citado na página 8.

[33] B. Kovačević, M. M. Milosavljevic, M. Veinovic, and M. Marković, *Robust digital processing of speech signals*. Springer, 2017. Citado na página 10.

[34] L. R. Rabiner and R. W. Schafer, *Introduction to digital speech processing*, vol. 1. Now Publishers Inc, 2007. Citado nas páginas 10, 11, 12, 17 e 38.

[35] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005. Citado nas páginas 11, 12 e 36.

[36] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–15, 2015. Citado nas páginas 12, 13 e 38.

[37] R. Iriya, *Análise de sinais de voz para reconhecimento de emoções.* PhD thesis, Universidade de São Paulo, 2014. Citado nas páginas 12, 29, 31, 32, 34 e 50.

[38] R. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in *Advanced techniques in computing sciences and software engineering*, pp. 279–282, Springer, 2010. Citado na página 13.

[39] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. Citado na página 13.

[40] D. Gerhard *et al.*, *Pitch extraction and fundamental frequency: History and current techniques*. Department of Computer Science, University of Regina Regina, Canada, 2003. Citado na página 13.

[41] P. Boersma *et al.*, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, pp. 97–110, Citeseer, 1993. Citado na página 13.

[42] M.-W. Mak and H.-B. Yu, "A study of voice activity detection techniques for nist speaker recognition evaluations," *Computer Speech & Language*, vol. 28, no. 1, pp. 295–313, 2014. Citado na página 13.

[43] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion detection from text and speech: a survey," *Social Network Analysis and Mining*, vol. 8, no. 1, pp. 1–26, 2018. Citado nas páginas 14, 16, 31 e 32.

[44] T. Polzehl, *PERSONALITY IN SPEECH.* Springer, 2016. Citado na página 14.

[45] M. Brockmann-Bauser, *Improving jitter and shimmer measurements in normal voices.* PhD thesis, Newcastle University, 2012. Citado na página 15.

[46] J. Kreiman and B. R. Gerratt, "Jitter, shimmer, and noise in pathological voice quality perception," in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003. Citado na página 15.

[47] M. Farrús and J. Hernando, "Using jitter and shimmer in speaker verification," *IET Signal Processing*, vol. 3, no. 4, pp. 247–257, 2009. Citado na página 15.

[48] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Eighth annual conference of the international speech communication association*, 2007. Citado na página 15.

[49] J.-A. Bachorowski and M. J. Owren, "Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context," *Psychological science*, vol. 6, no. 4, pp. 219–224, 1995. Citado na página 15.

[50] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman, "Stress and emotion classification using jitter and shimmer features," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–1081, IEEE, 2007. Citado na página 15.

[51] J.-A. Bachorowski, "Vocal expression and perception of emotion," *Current directions in psychological science*, vol. 8, no. 2, pp. 53–57, 1999. Citado na página 15.

[52] I. R. Titze, Y. Horii, and R. C. Scherer, "Some technical considerations in voice perturbation measurements," *Journal of Speech, Language, and Hearing Research*, vol. 30, no. 2, pp. 252–260, 1987. Citado na página 15.

[53] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Error weighted semi-coupled hidden markov model for audio-visual emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 142–156, 2011. Citado nas páginas 16 e 31.

[54] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home*, vol. 6, no. 2, pp. 101–108, 2012. Citado na página 16.

[55] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980. Citado nas páginas 16 e 17.

[56] H. Fletcher, "Auditory patterns," *Reviews of modern physics*, vol. 12, no. 1, p. 47, 1940. Citado na página 16.

[57] M. Slaney, "Auditory toolbox," Citado nas páginas 17 e 45.

[58] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003. Citado nas páginas 17, 32, 34 e 46.

[59] L. Rettig, J. Audiffren, and P. Cudré-Mauroux, "Fusing vector space models for domain-specific applications," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1110–1117, IEEE, 2019. Citado na página 18.

[60] Y. Li and T. Yang, "Word embedding for understanding natural language: a survey," in *Guide to big data applications*, pp. 83–104, Springer, 2018. Citado na página 18.

[61] Y. Goldberg, "Neural network methods for natural language processing," *Synthesis lectures on human language technologies*, vol. 10, no. 1, pp. 1–309, 2017. Citado na página 18.

[62] O. Shahmirzadi, A. Lugowski, and K. Younge, "Text similarity in vector space models: a comparative study," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 659–666, IEEE, 2019. Citado na página 18.

[63] B. Neupane, T. Horanont, and J. Aryal, "Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis," *Remote Sensing*, vol. 13, no. 4, p. 808, 2021. Citado na página 18.

[64] C. V. Gysel, M. De Rijke, and E. Kanoulas, "Neural vector spaces for unsupervised information retrieval," *ACM Transactions on Information Systems (TOIS)*, vol. 36, no. 4, pp. 1–25, 2018. Citado na página 18.

[65] J. Montalvao, L. Miranda, and B. Dorizzi, "Straightforward working principles behind modern data visualization approaches," *IEEE Access*, vol. 9, pp. 4242–4252, 2020. Citado na página 18.

[66] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982. Citado na página 19.

[67] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *The journal of machine learning research*, vol. 3, pp. 1137–1155, 2003. Citado nas páginas 20 e 35.

[68] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013. Citado nas páginas 20 e 35.

[69] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954. Citado na página 20.

[70] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of documentation*, 2004. Citado na página 21.

[71] C. M. Bishop, "Pattern recognition," *Machine learning*, vol. 128, no. 9, 2006. Citado nas páginas 26, 27, 28, 52 e 57.

[72] P. E. Hart, D. G. Stork, and R. O. Duda, *Pattern classification*. Wiley Hoboken, 2000. Citado nas páginas 27, 28, 52 e 57.

[73] M. Kuhn, K. Johnson, *et al.*, *Applied predictive modeling*, vol. 26. Springer, 2013. Citado na página 29.

[74] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3, pp. 1970–1973, IEEE, 1996. Citado na página 31.

[75] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*, pp. 205–211, 2004. Citado na página 31.

[76] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, "Automatic emotion recognition using prosodic parameters," in *Ninth European conference on speech communication and technology*, 2005. Citado na página 31.

[77] S. G. Koolagudi, A. Barthwal, S. Devliyal, and K. S. Rao, "Real life emotion classification using spectral features and gaussian mixture models," *Procedia engineering*, vol. 38, pp. 3892–3899, 2012. Citado na página 31.

[78] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in *2009 international conference on information engineering and computer science*, pp. 1–4, IEEE, 2009. Citado nas páginas 31 e 32.

[79] A. Zhu and Q. Luo, "Study on speech emotion recognition system in e-learning," in *International Conference on Human-Computer Interaction*, pp. 544–552, Springer, 2007. Citado na página 31.

[80] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019. Citado na página 31.

[81] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011. Citado na página 31.

[82] S. Kuchibhotla, H. D. Vankayalapati, R. Vaddi, and K. R. Anne, "A comparative analysis of classifiers in emotion recognition through acoustic features," *International Journal of Speech Technology*, vol. 17, no. 4, pp. 401–408, 2014. Citado nas páginas 31 e 63.

[83] S. Renjith and K. Manju, "Speech based emotion recognition in tamil and telugu using lpcc and hurst parameters—a comparitive study using knn and ann classifiers," in *2017 International conference on circuit, power and computing technologies (ICCPCT)*, pp. 1–6, IEEE, 2017. Citado na página 31.

[84] J. Ancilin and A. Milton, "Improved speech emotion recognition with mel frequency magnitude coefficient," *Applied Acoustics*, vol. 179, p. 108046, 2021. Citado na página 31.

[85] A. Milton, S. S. Roy, and S. T. Selvi, "Svm scheme for speech emotion recognition using mfcc feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013. Citado na página 32.

[86] S. Lalitha, D. Geyasruti, R. Narayanan, and M. Shravani, "Emotion detection using mfcc and cepstrum features," *Procedia Computer Science*, vol. 70, pp. 29–35, 2015. Citado na página 32.

[87] L. Abdel-Hamid, "Egyptian arabic speech emotion recognition using prosodic, spectral and wavelet features," *Speech Communication*, vol. 122, pp. 19–30, 2020. Citado na página 32.

[88] F. Daneshfar, S. J. Kabudian, and A. Neekabadi, "Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and gaussian elliptical basis function network classifier," *Applied Acoustics*, vol. 166, p. 107360, 2020. Citado na página 32.

[89] K. Zvarevashe and O. Olugbara, "Ensemble learning of hybrid acoustic features for speech emotion recognition," *Algorithms*, vol. 13, no. 3, p. 70, 2020. Citado na página 32.

[90] R. Subhashree and G. Rathna, "Speech emotion recognition: Performance analysis based on fused algorithms and gmm modelling," *Indian Journal of Science and Technology*, vol. 9, no. 11, pp. 1–8, 2016. Citado na página 32.

[91] P. Patel, A. Chaudhari, R. Kale, and M. Pund, "Emotion recognition from speech with gaussian mixture models & via boosted gmm," *International Journal of Research In Science & Engineering*, vol. 3, 2017. Citado na página 32.

[92] S. Mao, D. Tao, G. Zhang, P. Ching, and T. Lee, "Revisiting hidden markov models for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6715–6719, IEEE, 2019. Citado na página 32.

[93] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 2, pp. II–1, Ieee, 2003. Citado nas páginas 32 e 34.

[94] S. Ramakrishnan, "Recognition of emotion from speech: A review," *Speech Enhancement, Modeling and recognition–algorithms and Applications*, vol. 7, pp. 121–137, 2012. Citado na página 32.

[95] R. B. Lanjewar, S. Mathurkar, and N. Patel, "Implementation and comparison of speech emotion recognition system using gaussian mixture model (gmm) and k-nearest neighbor (k-nn) techniques," *Procedia computer science*, vol. 49, pp. 50–57, 2015. Citado na página 32.

[96] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011. Citado na página 32.

[97] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019. Citado nas páginas 32 e 33.

[98] D. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Systems with Applications*, vol. 173, p. 114683, 2021. Citado na página 33.

[99] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017. Citado na página 33.

[100] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5089–5093, IEEE, 2018. Citado na página 33.

[101] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Online emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, 2010. Citado na página 33.

[102] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014. Citado na página 33.

[103] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020. Citado nas páginas 33 e 34.

[104] M. T. García-Ordás, H. Alaiz-Moretón, J. A. Benítez-Andrades, I. García-Rodríguez, O. García-Olalla, and C. Benavides, "Sentiment analysis in non-fixed length audios using a fully convolutional neural network," *Biomedical Signal Processing and Control*, vol. 69, p. 102946, 2021. Citado na página 33.

[105] M. M. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using gaussian mixture vector autoregressive models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–957, IEEE, 2007. Citado na página 34.

[106] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," in *2013 5th international conference on Knowledge and smart technology (KST)*, pp. 86–91, IEEE, 2013. Citado nas páginas 34 e 63.

[107] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221640–221653, 2020. Citado na página 34.

[108] K. K. Sahoo, I. Dutta, M. F. Ijaz, M. Woźniak, and P. K. Singh, "Tlefuzzynet: Fuzzy rank-based ensemble of transfer learning models for emotion recognition from human speeches," *IEEE Access*, vol. 9, pp. 166518–166530, 2021. Citado na página 34.

[109] Y.-A. Chung and J. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," *arXiv preprint arXiv:1803.08976*, 2018. Citado na página 35.

[110] P. Tzirakis, A. Nguyen, S. Zafeiriou, and B. W. Schuller, "Speech emotion recognition using semantic information," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6279–6283, IEEE, 2021. Citado na página 35.

[111] J. H. McClellan, R. W. Schafer, and M. A. Yoder, *Dsp first*. Pearson Education, 2017. Citado na página 37.

[112] K. Kumar, R. Aggarwal, and A. Jain, "A hindi speech recognition system for connected words using htk," *International Journal of Computational Systems Engineering*, vol. 1, no. 1, pp. 25–32, 2012. Citado na página 37.

[113] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient mfcc extraction method in speech recognition," in *2006 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 4–pp, IEEE, 2006. Citado na página 37.

[114] M. Behlau, *Voz: O livro do especialista: volume II*. Revinter, 2005. Citado na página 38.

[115] H. Hollien, "On vocal registers," *Journal of Phonetics*, vol. 2, no. 2, pp. 125–143, 1974. Citado na página 38.

[116] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, "Features for content-based audio retrieval," in *Advances in computers*, vol. 78, pp. 71–150, Elsevier, 2010. Citado na página 46.

[117] R. Böck, O. Egorow, I. Siegert, and A. Wendemuth, "Comparative study on normalisation in emotion recognition from speech," in *International Conference on Intelligent Human Computer Interaction*, pp. 189–201, Springer, 2017. Citado na página 48.

[118] T. J. Sefara, "The effects of normalisation methods on speech emotion recognition," in *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pp. 1–8, IEEE, 2019. Citado na página 48.

[119] L. Wang *et al.*, "Toward a discriminative codebook: Codeword selection across multi-resolution.," in *CVPR*, pp. 1–8, 2007. Citado na página 48.

[120] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000. Citado nas páginas 51 e 64.

[121] MATLAB, *version 9.0.0.341360 (R2016a)*. Natick, Massachusetts: The MathWorks Inc., 2016. Citado na página 52.

[122] G. P. Zhang, "Neural networks for classification: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451–462, 2000. Citado na página 52.

[123] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. Citado na página 53.

[124] Y. Li, L. Chao, Y. Liu, W. Bao, and J. Tao, "From simulated speech to natural speech, what are the robust features for emotion recognition?," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 368–373, IEEE, 2015. Citado na página 63.