



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Campus de Ilha Solteira

PATRICIA GABRIEL VIEIRA

**IDENTIFICAÇÃO DE PESSOAS UTILIZANDO ATRIBUTOS DE
LÍDERES WAVELET EXTRAÍDOS DOS SINAIS DE VOZ EM
MODELOS DE APRENDIZADO DE MÁQUINA**

Ilha Solteira
2021





UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Campus de Ilha Solteira

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

PATRICIA GABRIEL VIEIRA

**IDENTIFICAÇÃO DE PESSOAS UTILIZANDO ATRIBUTOS DE
LÍDERES WAVELET EXTRAÍDOS DOS SINAIS DE VOZ EM
MODELOS DE APRENDIZADO DE MÁQUINA**

Tese de Doutorado apresentada à
Faculdade de Engenharia do Campus
de Ilha Solteira - UNESP, para
obtenção do Grau de Doutora em
Engenharia Elétrica.

Área de conhecimento: Automação.

Prof. Dr. Jozue Vieira Filho
Orientador

Ilha Solteira
2021

FICHA CATALOGRÁFICA

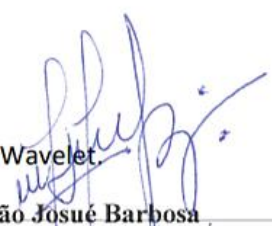
Desenvolvido pelo Serviço Técnico de Biblioteca e Documentação

V658i Vieira, Patricia Gabriel.
Identificação de pessoas utilizando atributos de líderes wavelet extraídos dos sinais de voz em modelos de aprendizado de máquina / Patricia Gabriel Vieira. -- Ilha Solteira: [s.n.], 2021
92 f. : il.

Tese (Doutorado em Engenharia Elétrica) - Universidade Estadual Paulista. Faculdade de Engenharia de Ilha Solteira. Área de conhecimento: Automação, 2021

Orientador: Jozue Vieira Filho
Inclui bibliografia

1. Reconhecimento de Locutor. 2. Wavelet.



João Josué Barbosa

Serviço Técnico de Biblioteca e Documentação
Diretor Técnico
CRB 8-5642


CERTIFICADO DE APROVAÇÃO


TÍTULO DA TESE: Identificação de Pessoas Utilizando Atributos de Líderes Wavelet Extraídos dos Sinais de Voz em Modelos de Aprendizado de Máquina


AUTORA: PATRÍCIA GABRIEL VIEIRA

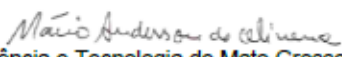
ORIENTADOR: JOZUE VIEIRA FILHO

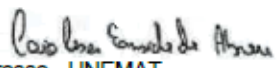
Aprovada como parte das exigências para obtenção do Título de Doutora em ENGENHARIA ELÉTRICA, área: Automação pela Comissão Examinadora:

Prof. Dr. JOZUE VIEIRA FILHO (Participação Virtual) 
Coordenadoria Executiva / Câmpus Experimental de São João da Boa Vista - UNESP

Prof. Dr. ANNA DIVA PLASENCIA LOTUFO (Participação Virtual) 
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Ilha Solteira - UNESP

Prof. Dr. MARCO APARECIDO QUEIROZ DUARTE (Participação Virtual) 
Departamento de Matemática / Universidade Estadual de Mato Grosso do Sul - UEMS

Prof. Dr. MÁRIO ANDERSON DE OLIVEIRA (Participação Virtual) 
Departamento de Eletroeletrônica / Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso - IFMT

Prof. Dr. CAIO CESAR ENSIDE DE ABREU (Participação Virtual) 
Departamento de Computação / Universidade do Estado de Mato Grosso - UNEMAT

Ilha Solteira, 09 de setembro de 2021

*“Para os olhos da mente, um fractal é
uma maneira de entrever o infinito.”*

James Gleick

AGRADECIMENTOS

Agradeço a Deus pelas bênçãos e sabedoria recebidas.

À nossa senhora de Aparecida, minha intercessora durante esta jornada.

Aos meus pais José e Maria pelo apoio e conselhos fornecidos para vencer os obstáculos da vida.

À minha irmã pelas alegrias proporcionadas com a sua companhia.

Ao meu orientador Jozue Vieira Filho pela oportunidade, paciência e dedicação neste trabalho.

Aos professores do PPGEE - FEIS - UNESP pelas valiosas contribuições na minha formação acadêmica. Meus agradecimentos aos professores que compuseram a banca de qualificação e de defesa desta tese pelas sugestões e colaborações recebidas, em especial ao professor Marco Aparecido Queiroz Duarte pelo acompanhamento desde o mestrado.

Aos meus amigos Natalia, Hugo e Marco pela companhia e forças recebidas desde a graduação.

Ao Fabrício e aos demais colegas do grupo de *wavelets* e do Laboratório de Processamento de Sinais e Instrumentação (LabPSI - FEIS - UNESP) pelas trocas de conhecimentos e experiências.

Às pessoas que direta ou indiretamente me ajudaram a superar os momentos difíceis durante este período.

Por fim, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, pelo suporte financeiro concedido.

RESUMO

Este trabalho apresenta uma metodologia para identificação de locutores baseada na inserção de um novo atributo de áudio, denominado Média Máxima dos Líderes Wavelet (*Maximum Mean Wavelet Leaders – MMWL*), extraídos e concatenados com os Coeficientes Mel-Cepstrais (*Mel-Frequency Cepstrum Coefficients – MFCC*) em modelos de aprendizado de máquina. A extração de características dos sinais de voz é fundamental para o reconhecimento de locutor, tanto para a identificação, como para a verificação. Independentemente da aplicação, é essencial ter um sistema que seja capaz de reunir, distinguir e classificar características extraídas dos sinais de voz com alta taxa de acurácia. Neste sentido, o principal objetivo deste trabalho é propor uma metodologia usando atributos confiáveis de sinais de voz para a identificação. A base do trabalho é a extração dos atributos da MMWL aliada a um processo de aprendizado de máquina. Os resultados indicam que a inserção da MMWL destaca características multifractais dos sinais de voz, aumenta a precisão dos modelos baseados nos MFCC e melhora o percentual de confiança na identificação de locutores. Para validar o método proposto, um estudo detalhado é realizado envolvendo atributos clássicos de sinais de voz para comparação com os resultados obtidos usando MMWL Espectral + MFCC.

Palavras-chave: reconhecimento de locutor; análise multifractal; wavelet.

ABSTRACT

This work presents a methodology for speaker identification based on the insertion of a new audio attribute, called Maximum Mean Wavelet Leaders (MMWL), extracted and concatenated with Mel-Frequency Cepstrum Coefficients (MFCC) in machine learning models. Feature extraction from speech signals is crucial for speaker recognition, both for identification and verification. Regardless of the application, a speaker identification system must be able to gather, distinguish and classify features extracted from speech signals with a high accuracy rate. Therefore, the main objective of this work is to propose a methodology using reliable attributes of speech signals for identification. The basis of the work is the extraction of MMWL attributes associated to a machine learning process. The results indicate that the insertion of MMWL highlights multifractal features of speech signals, increases the accuracy of MFCC-based models, and improves the percentage of confidence in speaker identification. To validate the proposed method, a detailed study is conducted involving classical attributes of speech signals for comparison with the results obtained using Spectral MMWL + MFCC.

Keywords: speaker recognition; multifractal analysis; wavelet.

LISTA DE FIGURAS

Figura 1 - Estrutura do aparelho fonador.....	20
Figura 2 - Conversão Analógico-Digital.	23
Figura 3 - Árvore de decomposição da DWT, em três níveis, e o sinal após a DWT....	34
Figura 4 - Árvore de decomposição da WPT, em três níveis, e o sinal após a WPT.	36
Figura 5 - Líderes Wavelet.	40
Figura 6 - Exemplo do cálculo dos Líderes Wavelet.....	41
Figura 7 - Esquema da WTMM, WLMF e MWL/MMWL para análise de sinais.....	43
Figura 8 - Tarefas de identificação e verificação de locutor.	45
Figura 9 - Estrutura geral de um sistema de reconhecimento de voz.	48
Figura 10 - Matriz de Confusão.....	54
Figura 11 - Análise do: (a) sinal de voz “Zero” usando (b) Líderes Wavelet.	57
Figura 12 - Análise dos sinais de voz usando: (a) Espectro Multifractal e (b) Líderes Wavelet Máximo.	58
Figura 13 - Espectro multifractal dos sinais de voz com seleção de banda.....	59
Figura 14 - Matriz de confusão para dois locutores do banco <i>Speech Commands</i>	61
Figura 15 - Identificação de locutor proposta.....	64
Figura 16 - Processamento do sinal: (a) Processo de sobreposição de janelas e (b) Processo completo de extração dos atributos.	65
Figura 17 - Matriz de confusão usando a MMWL Espectral + MFCC (banco AN4)....	70
Figura 18 - Matriz de confusão usando Pitch + MFCC (banco AN4).....	70
Figura 19 - Matriz de confusão obtida por arquivo (banco AN4).	71
Figura 20 - Histogramas para diferentes arquivos de áudio usando: (a) MMWL espectral e (b) Entropia espectral.....	74
Figura 21 - Características extraídas do: (a) sinal de voz, usando (b) MMWL Espectral e (c) Entropia Espectral.	75
Figura 22 - MMWL Espectral posicionada no sinal de voz.	76
Figura 23 - Espectrogramas e Histogramas da MMWL Espectral, respectivamente para: (a), (b) sinal de voz original e (c), (d) sinal de voz com remoção do silêncio.	77
Figura 24 - Detecção da parte vozeada do: (a) sinal de voz “Two”, usando (b) MMWL Espectral com Limiar e (c) Contorno de Pitch com Energia e ZCR.	78

Figura 25 - Características para emoções de felicidade e tristeza de locutores femininos (F) e masculinos (M) obtidas com MMWL Espectral (a, b, c, d) e Pitch (e, f, g, h).....	79
Figura 26 - Decaimento Espectral e MMWL Espectral, respectivamente, para: (a), (c) Violão e (b), (d) Bateria.....	81
Figura 27 - Comparação Espectral do: (a) áudio limpo, aplicando (b) MMWL Espectral e (c) Nivelamento Espectral; (d) áudio corrompido com ruído branco, aplicando (e) MMWL Espectral e (f) Nivelamento Espectral.....	82

LISTA DE TABELAS

Tabela 1 - Nível de decisão para segregação de componentes.....	29
Tabela 2 - Descrição dos arquivos de áudio.	58
Tabela 3 - Estatísticas aplicadas à MWL para sinais de voz de dois locutores.	60
Tabela 4 - Percentual de confiança para a palavra “Zero”.	62
Tabela 5 - Acurácia dos classificadores (banco <i>Speech Commands</i>).	62
Tabela 6 - Comparação entre modelos de classificadores (banco <i>Speech Commands</i>). ...	63
Tabela 7 - Comparação entre os tipos de modelo KNN (banco <i>Speech Commands</i>). ...	64
Tabela 8 - Distribuição dos arquivos para o banco AN4.....	67
Tabela 9 - Acurácia dos atributos para o banco AN4.....	68
Tabela 10 - Distribuição do conjunto de teste completo (banco TIMIT).....	72
Tabela 11 - Acurácia de validação para conjunto de teste completo (banco TIMIT). ...	72

LISTA DE ABREVIATURAS E SIGLAS

CWT	<i>Continuous Wavelet Transform</i> - Transformada Wavelet Contínua
DAV	Detector de Atividade de Voz
DCT	<i>Discrete Cosine Transform</i> - Transformada Cosseno Discreta
DFT	<i>Discrete Fourier Transform</i> – Transformada de Fourier Discreta
DWT	<i>Discrete Wavelet Transform</i> - Transformada Wavelet Discreta
FIR	<i>Finite Impulse Response</i> - Filtro de Resposta Finita ao Impulso
FT	<i>Fourier Transform</i> – Transformada de Fourier
FFT	<i>Fast Fourier Transform</i> – Transformada Rápida de Fourier
IDFT	<i>Inverse Discrete Fourier Transform</i> - Transformada de Fourier Discreta Inversa
KNN	<i>K Nearest Neighbors</i> - K Vizinhos Mais Próximos
LPC	<i>Linear Prediction Coefficients</i> - Coeficientes de Predição Linear
MFCC	<i>Mel-Frequency Cepstrum Coefficients</i> - Coeficientes Mel-Cepstrais.
MMWL	<i>Maximum Mean Wavelet Leaders</i> – Média Máxima dos Líderes Wavelet.
MWL	<i>Maximum Wavelet Leaders</i> - Líderes Wavelet Máximos
RAL	Reconhecimento Automático de Locutor
WPT	<i>Wavelet Packet Transform</i> - Transformada Wavelet Packet
WL	<i>Wavelet Leaders</i> - Líderes Wavelet
WLMF	<i>Wavelet Leaders Multifractal Formalism</i> - Líderes Wavelet - Formalismo Multifractal
WT	<i>Wavelet Transform</i> - Transformada Wavelet
WTMM	<i>Wavelet Transform Modulus Maxima</i> - Transformada Wavelet Módulos Máximos

SUMÁRIO

1.	INTRODUÇÃO	15
1.1	CONTRIBUIÇÃO DO TRABALHO E OBJETIVOS.....	18
1.2	ORGANIZAÇÃO DO TRABALHO	18
2.	ANÁLISE DE SINAIS DE VOZ	19
2.1	SINAL DE VOZ	19
2.2	PRÉ-PROCESSAMENTO	23
2.2.1	Amostragem	24
2.2.2	Pré-ênfase	25
2.2.3	Segmentação	25
2.2.4	Janelamento.....	26
2.2.5	Transformada de Fourier discreta	26
2.2.6	Transformada do cosseno discreta	27
2.3	DETECTOR DE ATIVIDADE DE VOZ.....	28
2.4	FREQUÊNCIA FUNDAMENTAL - PITCH.....	29
2.5	COEFICIENTES MEL-CEPSTRAIS	30
3.	ANÁLISES WAVELET E MULTIFRACTAL	32
3.1	ANÁLISE WAVELET	32
3.1.1	Transformada wavelet contínua (CWT)	33
3.1.2	Transformada wavelet discreta (DWT).....	33
3.1.3	Transformada wavelet packet (WPT)	34
3.2	ANÁLISE MULTIFRACTAL	37
3.2.1	Espectro multifractal	37
3.2.2	Líderes Wavelet	39
3.2.3	Média Máxima dos Líderes Wavelet	41
4.	RECONHECIMENTO DE LOCUTOR	45

4.1	VERIFICAÇÃO DE LOCUTOR	46
4.2	IDENTIFICAÇÃO DE LOCUTOR	48
4.3	APRENDIZADO DE MÁQUINA	49
4.3.1	Seleção do modelo de aprendizagem	50
4.3.2	Classificador KNN.....	52
4.3.3	Avaliação e validação do modelo	53
5.	METODOLOGIA PROPOSTA E RESULTADOS	56
5.1	MMWL NA IDENTIFICAÇÃO DE LOCUTORES	56
5.2	MODELO DE IDENTIFICAÇÃO DE LOCUTOR.....	61
5.3	METODOLOGIA PROPOSTA	64
5.4	RESULTADOS E DISCUSSÕES	66
5.5	EXPERIMENTOS ADICIONAIS.....	73
5.5.1	Entropia Espectral <i>versus</i> MMWL Espectral.....	74
5.5.2	Pitch <i>versus</i> MMWL Espectral	77
5.5.3	Decaimento Espectral <i>versus</i> MMWL Espectral	80
5.5.4	Nivelamento Espectral <i>versus</i> MMWL Espectral.....	81
6.	CONCLUSÕES.....	83
	REFERÊNCIAS	84

1. INTRODUÇÃO

Diversas tecnologias têm sido implementadas utilizando sinais de voz. Dentre elas, o reconhecimento de locutor (orador) tem obtido grande destaque nos últimos anos devido à praticidade na identificação de pessoas e o crescente desenvolvimento de dispositivos controlados por comandos de voz.

Segundo Aizat *et al.* (2020), o processamento de sinais de voz tornou-se fundamental em aplicações para as quais são utilizados sistemas de identificação e verificação de locutor, tais como a fonética forense para identificação de suspeitos de crime, a robótica para tornar possível a interação homem-máquina, a segurança através da biometria etc.

O reconhecimento de locutor consiste em distinguir e identificar pessoas por meio das características presentes em suas vozes, com aplicações que envolvem, normalmente, dois objetivos básicos: verificar e identificar. A verificação de locutor (*Speaker Verification - SV*) é um método utilizado para aceitar ou rejeitar a reivindicação de identidade de um locutor individual, enquanto que a identificação do locutor (*Speaker Identification - SI*) determina qual locutor registrado fornece um determinado enunciado entre um conjunto de locutores conhecidos (RABINER; JUANG, 1993; NSTC, 2006). Em termos gerais, no reconhecimento de locutor, o objetivo é extrair atributos dos sinais de voz para formar e comparar diferentes padrões.

Sistemas de reconhecimento de locutor são complexos do ponto de vista de implementação e podem ser afetados por vários aspectos inerentes ao seu desenvolvimento, como tamanho de bancos de dados, dificuldade de encontrar padrões nos parâmetros extraídos do sinal de voz, baixa qualidade de locuções e ruídos presentes em canais de comunicação ou captação da fala. Além disso, os sistemas devem ser capazes de lidar com a possibilidade de que, com o tempo, novos dados sofram alterações como, por exemplo, uma rouquidão da voz ocasionada por algum distúrbio nas pregas vocais devido a uma doença patológica.

Os principais fatores que aumentam a probabilidade de erro e acarretam degradação no desempenho dos sistemas de reconhecimento de locutor estão relacionados à baixa variabilidade das características na distinção dos locutores, uma vez que a identificação da pessoa de forma exclusiva pelo sistema é dificultada pela falta de padrões únicos encontrados nos sinais. Dessa forma, para lidar com este tipo de situação e analisar os

atributos mais representativos dos sinais de voz, foi inserida neste trabalho uma ferramenta oriunda da inteligência artificial (*Artificial Intelligence - AI*) conhecida como aprendizado de máquina ou aprendizado automático.

O aprendizado de máquina (*machine learning*) consiste em programar computadores para otimizar o critério de performance da extração de padrões usando dados de exemplos e informações prévias (LARRAÑAGA *et al.*, 2006; FABRIS, MAGALHÃES; FREITAS, 2017).

Na última década, o aprendizado de máquina foi aplicado com sucesso no reconhecimento de padrões em sinais de voz, facilitando a tomada de decisões e previsões. Em Prakash e Nithya (2014), bons resultados foram obtidos analisando os atributos do áudio em algoritmos de aprendizado semi-supervisionado. Em Bhavsar e Ganatra (2012), um estudo comparativo entre modelos de aprendizado de máquina e seus algoritmos de classificação existentes na literatura é feito, analisando a velocidade, a acurácia e outras características. Também é feita uma revisão dos atributos utilizados para o reconhecimento de locutor em Kinnunen e Li (2010).

Um modelo de identificação de locutor deve ser capaz de encontrar um conjunto único de padrões e características das impressões vocais, diferenciando-as do resto de todo o conjunto de locutores. Uma abordagem de alto nível para isso consiste em recolher amostras da voz de uma pessoa, extrair as características adequadas do áudio para o classificador, construir o modelo treinando o classificador e, por fim, realizar a classificação para o reconhecimento (SHARMA, 2019).

A acurácia dos modelos de classificação e o desempenho dos sistemas identificação estão extremamente relacionados com as características extraídas do áudio e dos classificadores (DHANALAKSHMI; PALANIVEL; RAMALINGAM, 2011; ZUBAIR; YAN; WANG, 2013). Por isso, a extração de características (recursos/parâmetros ou atributos) é um dos aspectos mais importantes da identificação de locutor, com influência significativa no desempenho do processo.

Um estudo realizado por Tirumala *et al.* (2017) indicou que abordagens baseadas na extração dos Coeficientes Mel-Cepstrais (*Mel-Frequency Cepstrum Coefficients – MFCC*), variações dos MFCC, ou a fusão dos MFCC com outras características são amplamente utilizadas nos sistemas de reconhecimento de locutor. Em Li *et al.* (2001), os MFCC também demonstraram proporcionar uma melhor precisão na classificação em relação a outras características temporais extraídas do sinal de voz.

Entretanto, mesmo realizando boas previsões, estes modelos não são 100% assertivos. Assim, neste trabalho, com o objetivo de aumentar o percentual de acurácia dos modelos de aprendizado de máquina para a identificação de locutor, propõe-se a extração de um novo atributo, chamado Média Máxima dos Líderes Wavelet (*Maximum Mean Wavelet Leaders* - MMWL), baseada nos Líderes Wavelet, obtidos por meio da análise multifractal.

Atualmente, a análise multifractal tem sido muito utilizada para análise de dados empíricos. A análise multifractal empírica consiste, essencialmente, em estimar os expoentes de escala a partir de um determinado conjunto de dados. Estes expoentes de escala são comumente envolvidos em várias tarefas de análise de dados, como a detecção, a identificação ou a classificação (WENDT; ABRY, 2007).

Algumas pesquisas utilizando técnicas multifractais vêm sendo aplicadas na área de processamento de sinais de voz. Por exemplo, os autores Sant'Ana, Coelho e Alcain (2006) propuseram um sistema de reconhecimento automático de voz independente de texto, utilizando características estatísticas obtidas por meio de um estimador multidimensional wavelet. Em Zhang, Guo e Zhang (2009) é proposto um método de extração de características não-lineares com base no método WTMM (*Wavelet Transform Modulus Maxima*), a fim de facilitar a extração de características do espectro multifractal de sinais de voz.

Na pesquisa de Zhou e Zhang (2018), um método novo foi proposto com base no WTMM, que facilitou a extração de características do espectro multifractal dos sinais de voz. Os resultados dessa experiência indicaram que a composição das características de espectro multifractal combinada com MFCC e LPC (*Linear Prediction Coding*) foi capaz de diminuir a taxa de erro no reconhecimento de voz, melhorar desempenho computacional e, além disso, eliminar perturbações de outras características redundantes.

As características multifractais dos sinais de voz motivaram o estudo da inserção de atributos baseados nos Líderes Wavelet, oriundos do formalismo multifractal, na identificação de pessoas usando aprendizado de máquina. Um estudo comparativo com outros atributos foi realizado para destacar a importância da inserção da MMWL no reconhecimento de locutor, bem como suas contribuições no desempenho dos modelos de classificação.

1.1 CONTRIBUIÇÃO DO TRABALHO E OBJETIVOS

Os objetivos deste trabalho consistem em apresentar um novo atributo confiável do sinal de voz, denominado MMWL, e propor uma metodologia usando esse atributo para a identificação de pessoas. Na abordagem usando modelos supervisionados de aprendizado de máquina para identificação de locutor, verifica-se que usando apenas os atributos de MFCC é possível obter uma performance razoavelmente boa na predição.

Entretanto, ao inserir e concatenar a MMWL com o MFCC, a performance do modelo de identificação para os conjuntos de dados avaliados melhora em termos de acurácia e supera a performance quando outros atributos são usados. Desta forma, uma metodologia baseada no novo atributo proposto concatenado com MFCC se mostra vantajosa e fortemente indicada para sistemas de identificação de locutor.

Além disso, o atributo proposto permite extrair informações significativas que não são possíveis de serem obtidas usando os atributos de áudio existentes na literatura. Por isso, várias sugestões usando os atributos da MMWL são feitas ao longo do texto em aplicações voltadas para o processamento de sinais de voz.

1.2 ORGANIZAÇÃO DO TRABALHO

Para facilitar a compreensão da proposta, o trabalho está organizado em seis capítulos. No primeiro capítulo, já apresentado, é feita uma introdução geral para entender o contexto do trabalho. No segundo capítulo é apresentada uma breve explanação dos aspectos conceituais da voz, a fim de compreender o mecanismo de produção da fala, bem como os fundamentos do processamento digital de sinais de voz. No terceiro são apresentados os conceitos básicos que servirão de subsídios para compreender as análises wavelet e multifractal presentes na metodologia proposta neste trabalho. No quarto capítulo são apresentadas as aplicações envolvendo reconhecimento de locutor e um resumo das técnicas de aprendizado de máquina usadas na identificação de locutor. Já no quinto capítulo são apresentadas a construção e a definição da metodologia proposta, bem como os resultados usando a mesma. Também no quinto capítulo são apresentados alguns resultados experimentais usando o novo atributo de áudio proposto. Por fim, no sexto são feitas as considerações finais e apresentadas sugestões para trabalhos futuros.

2. ANÁLISE DE SINAIS DE VOZ

Neste capítulo, primeiramente, serão apresentados os aspectos conceituais da fala, desde a definição do sinal de voz até os mecanismos de produção da fala. Na sequência, para melhor compreender a análise de sinais de voz, serão revisados os conceitos básicos de processamento de sinais.

2.1 SINAL DE VOZ

A fala é o processo de comunicação mais utilizado pelo homem, gerada a partir de um sistema vocal próprio de cada ser humano, permitindo-lhe a troca de ideias, expressão de opiniões ou revelação do seu pensamento. Com o advento do computador e os avanços na área de robótica, a interação do ser humano com uma máquina através da voz passou a ser estudada e implementada em diferentes contextos (RABINER; 1978), exigindo sistemas e algoritmos complexos de reconhecimento automático de fala (MACIEL, 2007).

As características básicas dos sinais de voz apresentam diversas variações. Por exemplo, um mesmo locutor pode apresentar variações na pronúncia de sua voz provocadas por diversas fontes de uma locução, como o contexto da frase, o seu estado emocional, a velocidade de locução, a dicção ou o grau de clareza da pronúncia das palavras. Também podem ocorrer variações relacionadas a diferenças fisiológicas entre locutores, como idade, sexo, sotaque, entre outras (BORGES, 2001).

Quando uma pessoa fala, são emitidos sons e suas interpretações gráficas são as letras. Dessa forma, a voz é um som produzido pelo aparelho fonador. O som possui parâmetros perceptuais, tais como:

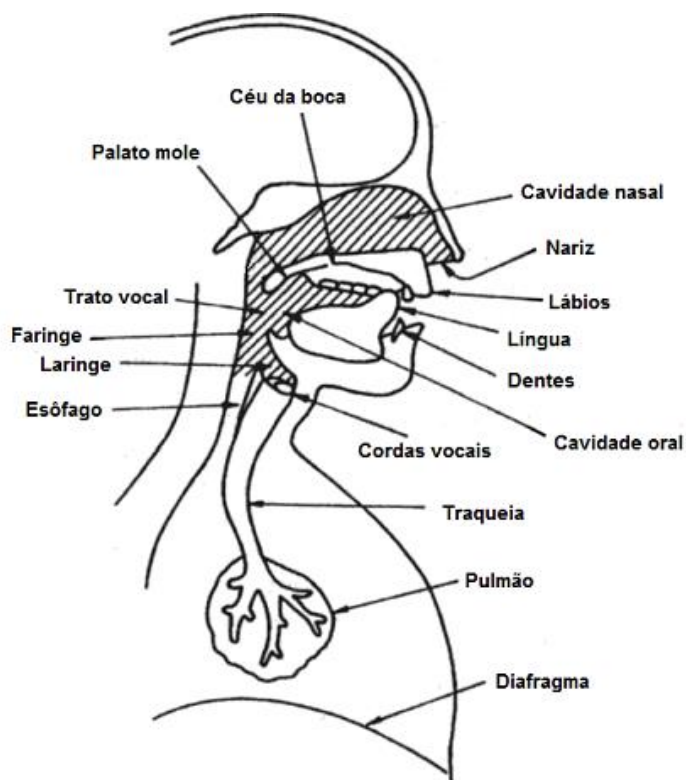
- Intensidade (fraca ou forte) - percepção da amplitude e da energia;
- Altura (aguda ou grave) - percepção da frequência fundamental;
- Fase - normalmente imperceptível;
- Timbre (origem do som) - percepção da complexidade.

Os sinais de voz são compostos por uma sequência de sons que servem como uma representação simbólica da mensagem produzida pelo locutor para o ouvinte (RABINER, 1978).

A voz é uma característica que somente os humanos possuem e baseia-se na produção de sons articulados, originando assim uma linguagem que é a fonte da comunicação. Assim, a voz não só transmite informação léxica, mas também expressa emoções, como dor e alegria, através de sua entonação.

Para a produção da fala, vários músculos e órgãos que constituem o aparelho fonador (pulmões, laringe e boca, por exemplo) são acionados. A posição, forma e tamanho desses elementos definem as propriedades físicas da voz e variam de indivíduo para indivíduo (FURUI, 2001). Um diagrama esquemático da estrutura do aparelho fonador humano é mostrado na Figura 1.

Figura 1 - Estrutura do aparelho fonador.



Fonte: Furui (2001).

A produção da voz se inicia com o fluxo de ar que flui desde os pulmões, impulsionado pelo diafragma, atravessando a laringe, onde se encontram dois pequenos tendões ou membranas chamadas cordas vocais. O fluxo aéreo respiratório, ao passar pelos ciclos de abertura e fechamento das pregas vocais, constituirá uma vibração que irá ressonar pelo trato vocal (FUKUYAMA, 2001).

Os sons da fala, dependendo da presença ou ausência de vibração das cordas vocais, podem se dividir em sonoros e surdos, com características distintas (ALCAIM; OLIVEIRA, 2012). Em um processo de comunicação, a ausência desses componentes indica trechos de ruído de fundo.

A componente sonora é composta por sinais de características periódicas, onde a sonoridade (som sonoro) ocorre quando a corrente de ar que vem dos pulmões encontra as cordas vocais fechadas, fazendo-as vibrar. Por exemplo, na palavra “Bato”, percebe-se esta sonoridade devido ao fonema /B/. O fonema é a unidade sonora mais simples da língua e divide-se em vogais, semivogais e consoantes.

Já a componente surda é formada por sinais que se assemelham a um ruído colorido, cuja energia concentra-se em frequências mais elevadas do que a componente sonora. O som surdo (não sonoro) ocorre quando a corrente de ar que vem dos pulmões encontra as cordas vocais relaxadas (abertas), não ocorrendo vibração. Por exemplo, na palavra “Prato” percebe-se este som surdo devido ao fonema /P/.

Uma importante distinção na locução pode ser feita com relação a produção da fala a partir de um texto, ou seja, a leitura. Uma locução emitida a partir de um texto pré-fixado é dita dependente de texto e, em caso contrário, independente de texto. Em um processo de reconhecimento de fala, sistemas dependentes de texto tendem a apresentar performances superiores aos independentes de texto. Tal característica deve-se ao fato de que os sistemas dependentes de texto se apoiam inicialmente no reconhecimento da locução para, então, a partir da divergência entre a locução de teste e um modelo selecionado, identificar o locutor (CAMPBELL, 1997).

O aparelho fonador de cada locutor apresenta uma estrutura peculiar que dá origem a um sinal de voz com características exclusivas. Da componente útil deste sinal, ou seja, dos sinais sonoros e surdos, podem ser extraídos os coeficientes mel-cepstrais que são utilizados na modelagem do correspondente aparelho fonador. Os sistemas de identificação usualmente empregam coeficientes extraídos de bancos de filtros, mas é possível usar, também, a predição linear (TOKUDA, 1994).

A componente útil do sinal de voz é isolada do ruído de fundo empregando-se um detector de atividade de voz (DAV). Os DAVs são sensíveis à relação sinal-ruído, sendo necessário realizar adaptações em sua estrutura de acordo com a qualidade do sinal de voz utilizado. Além do uso dos DAVs, também é importante realizar a redução de ruído dos sinais de voz (DUARTE, 2005).

A redução de ruído em sinais de voz tem como objetivo melhorar a qualidade do sinal. Existem vários métodos para a redução de ruído, sendo alguns usando a transformada de Fourier, como em Vieira Filho (1996), e outros usando a Transformada Wavelet como, por exemplo, em Duarte (2005), que utilizou a WT em uma metodologia baseada em limiar, obtendo assim, uma ferramenta que considera como ruído os coeficientes do sinal cujos valores absolutos são menores que um determinado valor. Na literatura também existem os que utilizaram a WT sem o uso do Limiar, por exemplo, Soares *et al.* (2011).

A eficiência dos métodos utilizados na redução de ruído é constatada verificando os níveis de redução de ruído e de distorção, e para isso, podem ser utilizadas duas medidas de qualidade objetivas, a SNR (*Signal to Noise Ratio*) e a PESQ (*Perceptual Evaluation of Speech Quality*) (ITU-T, 1996).

No processamento de sinais no domínio do tempo é possível realizar diversas operações no sinal de voz, tais como, armazenar, recuperar, cortar, copiar, colar, realçar, atenuar e mixar segmentos de arquivos de áudio. Já o processamento de sinais no domínio da frequência pode ser feito para diversos fins, por exemplo, quando se deseja realizar a filtragem digital, recuperar gravações, fazer ajustes de duração e de altura das amostras de som e até para aplicações, tais como, identificação, síntese ou reconhecimento de voz.

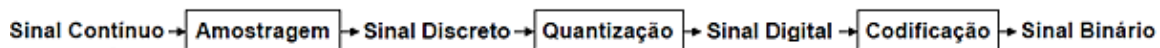
Em suma, uma locução é captada através de um processo de aquisição, filtragem, amostragem, quantização. O pré-processamento de sinal de voz é feito para que seja possível extrair informações úteis para o sistema que se está trabalhando.

De acordo com Oppenheim (1999), o sinal pode ser visto como uma variável física aplicada geralmente a algo que transmite informação, sendo representado matematicamente como função de uma ou mais variáveis independentes. Por exemplo, um sinal de voz é representado matematicamente como uma função do tempo (OPPENHEIM, 1999).

Os tipos de sinais podem ser definidos de acordo com o tipo das variáveis:

- Sinal analógico: todas as variáveis são contínuas.
- Sinal amostrado: discretização das variáveis independentes (amostragem).
- Sinal quantizado: discretização da variável dependente (quantização).
- Sinal digital: todas as variáveis são discretas (amostragem e quantização).

Um resumo do processo de conversão de um sinal analógico em um sinal digital é ilustrado na Figura 2.

Figura 2 - Conversão Analógico-Digital.

Fonte: Elaborado pela própria autora.

A amostragem é feita quando o sinal contínuo é convertido em sinal discreto no domínio do tempo, como mostrado na Figura 2. A frequência de amostragem de um sinal é um fator determinante no seu processo de digitalização. Para reconstruir o sinal original a partir do sinal amostrado, sem perda de informação, é necessário atender o teorema de Nyquist. O teorema de Nyquist é a base teórica do processamento digital de sinais contínuos, pois garante que um sinal contínuo pode ser amostrado, processado digitalmente e reconstruído de forma correta sem que ocorra distorção do tipo *aliasing*. Para tal, a frequência de amostragem deve ser igual ou maior do que duas vezes a frequência máxima do sinal original, assegurando que o sinal discretizado será constituído de repetições não sobrepostas do sinal original (LATHI, 2006).

Considerando também o processo descrito na Figura 2, após a amostragem vem a quantização, onde a amplitude analógica é convertida em um conjunto de amplitudes numéricas. Assim, o quantizador converte um sinal de tempo discreto em um sinal digital com representação por um número finito de bits. Por fim, é feita a codificação, que consiste em associar uma sequência de bits a cada um dos níveis em que se tem a quantização do sinal.

Os codificadores de fala permitem uma ampla gama de aplicações, incluindo telefonia de banda larga, dispositivos de comunicação, protocolo de Internet, segurança, criptografia, armazenamento de fala, entre outros. Às vezes esses codificadores utilizam aspectos dos processos de fala e de percepção da fala, e, portanto, podem não ser úteis para sinais de áudio mais gerais, como a música (RABINER, 1978).

2.2 PRÉ-PROCESSAMENTO

As principais informações contidas em sinal podem ser selecionadas por meio de uma filtragem adequada do sinal amostrado e depende da aplicação. De acordo com Oppenheim (1999), um filtro digital é um sistema discreto projetado para selecionar o conteúdo espectral de um sinal, limitando-o a uma determinada banda de frequências, isto é, a função de transferência do filtro forma uma janela espectral através da qual

somente é permitida a passagem da parte desejada do espectro do sinal. Com base na resposta da função de transferência, os filtros são classificados de acordo com as frequências que são atenuadas. Seguem alguns tipos:

- **Passa-baixa:** permite passar as frequências menores que a frequência de corte, ou seja, as baixas frequências;

- **Passa-alta:** permite passar as frequências maiores que a frequência de corte, ou seja, as altas frequências do sinal;

- **Passa-faixa:** permite passar as frequências de uma determinada faixa, ou seja, certa banda do sinal.

Existem duas classes principais de filtros digitais, que são os filtros de resposta ao impulso infinita (*Infinite Impulse Response - IIR*) e os filtros de resposta ao impulso finita (*Finite Impulse Response - FIR*).

A filtragem faz parte do pré-processamento, que serve para limitar as informações representativas do sinal, eliminando ruídos e outras interferências. Sendo assim, o sinal de voz é primeiramente pré-processado e então são utilizados algoritmos para extração dos parâmetros do sinal vocal. O conjunto de parâmetros de uma amostra de voz compõe um padrão que pode ser classificado. O pré-processamento visa extrair informações úteis do sinal e pode incluir etapas como: pré-ênfase, segmentação, janelamento e mudança de domínio (tempo – frequência, por exemplo).

2.2.1 Amostragem

Tipicamente, o conteúdo mais explorado de um sinal de voz varia de algumas dezenas de Hertz a 4 kHz (3,4 kHz é, por exemplo, o limite dos canais telefônicos analógicos), de modo que, considerando o Teorema de Nyquist, uma amostragem típica deve ser realizada com taxa de amostragem em torno de 10 kHz (DINIZ, 1997). Atenuações de componentes indesejáveis do sinal também podem ser feitas por meio de um filtro passa-baixas, também importante para evitar o fenômeno de *aliasing* (Teorema de Nyquist) (CARDOSO, 2009).

2.2.2 Pré-ênfase

De acordo com Cardoso (2009), a pré-ênfase consiste em filtrar um sinal para enfatizar informações que estão presentes em frequências mais elevadas da componente útil do sinal de voz, pois se sabe que essas frequências apresentam menor energia em relação às frequências mais baixas. O filtro de pré-ênfase no domínio da transformada Z possui a função de transferência dada pela Equação (1), onde ω é a frequência de corte escolhida no intervalo de 0,95 a 0,98 (CARDOSO, 2009).

$$H(z) = 1 - az^{-1}. \quad (1)$$

O filtro digital FIR passa-alta de primeira ordem da equação (1) é aplicado a fim de compensar os efeitos dos pulsos glotais e ressaltar as frequências das formantes (MORENO, 1996).

As formantes são as frequências de ressonância do trato vocal. O trato vocal é o nome genérico dado ao conjunto de cavidades e estruturas que participam da produção sonora, tendo como limite inferior a região glótica e como limite superior os lábios (RABINER, 1978). Por exemplo, para cada vogal, o trato vocal assume uma configuração relativamente estável, determinando frequências de ressonância específicas (CARDOSO, 2009).

2.2.3 Segmentação

Um sistema de segmentação de voz tem o objetivo de determinar as fronteiras que separam os elementos essenciais da fala, como palavras, sílabas ou fonemas de uma determinada locução. Eles podem ser usados para codificação de voz, como é o caso dos codificadores fonéticos em sistemas de reconhecimento automático e síntese de fala, entre outros.

De acordo com Campbell (1997), os parâmetros da voz para curtos intervalos de tempo da ordem de 10 ms a 30 ms (milissegundos) podem ser considerados invariantes no tempo. Sendo assim, o sinal de voz pode ser dividido em partes de tamanho fixo num período de tempo escolhido dentro de um determinado intervalo, permitindo caracterizar a cada instante o trato vocal como um filtro digital (CARDOSO, 2009).

Dessa forma, a segmentação consiste em particionar o sinal de voz em segmentos, selecionados por janelas ou quadros (frames) de duração perfeitamente definida (RABINER, 1978). Normalmente, o número de amostras é da forma 2^n , $n \in \mathbb{Z}$, o que facilita o cálculo da Transformada de Fourier em etapas posteriores do processamento do sinal (CARDOSO, 2009).

2.2.4 Janelamento

Após o processo de segmentação do sinal de voz é necessário reduzir o efeito das variações bruscas de amplitude presentes no início e término de cada quadro, atenuando o valor das amostras que se localizam nas extremidades do quadro. Isto pode ser feito multiplicando o quadro por uma janela suavizada e sobrepondo janelas adjacentes ao longo do processamento (CARDOSO, 2009). Assim, para um quadro com $x(n)$ amostras e uma função janela $w(n)$, o sinal suavizado $y(n)$ é dado por:

$$y(n) = x(n)w(n). \quad (2)$$

Há diversas funções de janelamento bem definidas na literatura. Uma das janelas mais usadas é a de Hamming (OPPENHEIM, 1999). A janela de Hamming é dada pela Equação (3), na qual M corresponde ao comprimento da janela.

$$w(n) = 0,54 - 0,46 \cos\left(\frac{2n\pi}{M-1}\right), \quad (3)$$

Considerando que o sinal de voz não é estacionário, o espectro de potência do sinal é calculado a cada janela usando a transformada de Fourier discreta (OPPENHEIM, 1999).

2.2.5 Transformada de Fourier discreta

O cálculo da transformada de Fourier discreta (*Discrete Fourier Transform - DFT*) de um quadro de comprimento M é definido pela Equação (4). Considerando um

sinal complexo $x(n)$, são necessárias M^2 operações de multiplicações complexas e $M.(M-1)$ somas complexas para o seu cálculo direto.

$$X(k) = \sum_{n=0}^{M-1} x(n) e^{-j \frac{2\pi k}{M} n}, \text{ para } k = 0, 1, \dots, M - 1. \quad (4)$$

Para reduzir o elevado número de operações da DFT é utilizado um conjunto de algoritmos proposto por Cooley-Tukey em 1965, conhecido como transformada rápida de Fourier (*Fast Fourier Transform - FFT*).

2.2.6 Transformada do cosseno discreta

A transformada do cosseno discreta (*Discrete Cosine Transform - DCT*), de forma análoga à DFT corresponde a uma transformada ortogonal, diferenciando-se desta última pelo fato de só ser aplicável a sequências reais. Com a DCT busca-se formar uma sequência periódica e simétrica a partir de uma sequência finita de tal forma que esta última possa ser corretamente recuperada (OPPENHEIM, 1999). Como é possível realizar esta transformada de formas diversas, há várias formulações para a DCT, no entanto, para processamento de sinais e imagens, é frequentemente empregada a DCT do tipo 2, que concentra mais energia nos seus primeiros coeficientes, permitindo uma maior compactação da informação presente no sinal. Outras características da DCT podem ser encontradas em Smith (1999). De acordo com Rao e Yip (1990), a DCT tipo 2 ou DCT II é definida pela Equação (5).

$$X^{c2}(k) = \sum_{n=0}^{M-1} x(n) \cos\left(\frac{\pi}{M} \left(n + \frac{1}{2}\right) k\right), \text{ para } k = 0, 1, \dots, M - 1. \quad (5)$$

A DFT está relacionada com a DCT conforme a Equação (6), permitindo que todo o conhecimento adquirido no cálculo rápido da DFT seja aproveitado para obter a DCT.

$$X^{c2}(k) = 2 \operatorname{Re} \left\{ X(k) e^{-j \frac{\pi k}{2M}} \right\}, \text{ para } k = 0, 1, \dots, M - 1. \quad (6)$$

2.3 DETECTOR DE ATIVIDADE DE VOZ

Não é possível extrair informações úteis para caracterizar o aparelho fonador dos trechos do sinal de voz em que está presente apenas o ruído de fundo. Assim, esses trechos de sinal devem ser removidos antes que o sinal seja enviado ao sistema de identificação. Os Detectores de Atividade de Voz (DAVs) permitem identificar e eliminar esses trechos a partir de técnicas como verificação do número de cruzamentos por zero, detecção da diferença do nível de energia entre a componente surda e a sonora, dentre outras. Independentemente da técnica empregada, devem-se separar os quadros que trazem informação útil daqueles que são compostos essencialmente por ruído de fundo, sendo aproveitados apenas os quadros com a componente útil.

Os detectores abordados neste trabalho são baseados no número de cruzamentos por zero e na energia. A técnica de cruzamento por zero consiste em verificar a frequência de alternância entre sinais positivos e negativos ao longo de uma janela de amostras. Como o ruído de fundo e a componente surda do sinal de voz apresentam um elevado grau de aleatoriedade, suas taxas de cruzamento por zero tendem a ser mais elevadas do que para os trechos sonoros. O uso exclusivo da técnica de cruzamentos por zero não permite por si só um isolamento adequado da componente útil, devendo-se ainda considerar outras informações, como a energia do sinal (RABINER, 1975).

A energia de um sinal pode ser calculada usando a Equação (7) em um intervalo de medições centrado em uma amostra.

$$E(n) = \sum_{i=-n_0}^{n_0} |s(n+i)|^2. \quad (7)$$

Segundo Rabiner (1975), ao se empregar o conjunto das técnicas de cruzamento por zero e energia é necessário fixar limiares diferenciados de energia para cada tipo de locução, o que exige adaptações no DAV para garantir uma decisão correta e compatível com sinais de diferentes qualidades.

Na Tabela 1 são apresentadas as características de cruzamento por zero e de energia de forma comparativa em diferentes componentes de um sinal de voz, evidenciando a necessidade de monitoramento de ambas as características apontadas.

Tabela 1 - Nível de decisão para segregação de componentes.

Decisão	Nível de cruzamento por zero	Nível de energia
Componente sonora	Baixo	Alto
Componente surda	Alto	Médio
Ruído de fundo	Alto	Baixo

Fonte: Elaboração da própria autora.

Além da energia do sinal, existem outras formas de extrair as características do sinal, como a autocorrelação, que pode ser um bom parâmetro para a discriminação entre sons sonoros e surdos.

2.4 FREQUÊNCIA FUNDAMENTAL - PITCH

A frequência fundamental, também denominada Pitch, é a frequência de vibração das cordas vocais, e está relacionada à percepção humana de sinais acústicos (HARTMANN, 1996). O tempo de duração de um pulso glotal é conhecido como período de Pitch. O período de Pitch é o inverso da frequência fundamental. Para um ouvinte, um Pitch maior ou menor significa, em termos de percepção auditiva, que o som é mais agudo ou mais grave, respectivamente. Valores de frequência baixos, entre 85 Hz e 180 Hz, são frequentes no sexo masculino (adulto em fala normal). Os valores femininos podem variar de 165 Hz a 255 Hz. Crianças e, principalmente bebês, podem atingir valores acima de 300 Hz.

A informação do período do Pitch é usada em várias aplicações, tais como:

- Identificação e verificação de locutores;
- Análise e síntese de fala;
- Diagnósticos de doenças da voz (laringite, nódulos da prega vocal e disfonias).

Um exemplo simples de método de extração da Pitch é o método da autocorrelação. Para um sinal de voz, define-se o estimador de autocorrelação de curto prazo pela Equação (8) (DELLER; HANSEN; PROAKIS, 2000).

$$r_s(\mu; m) = \frac{1}{N} \sum_{n=m-N+|\mu|+1}^m x(n)x(n - |\mu|). \quad (8)$$

Na Equação (8), a autocorrelação mede o grau de similaridade entre o mesmo sinal sem defasagem e defasado de um total de μ amostras, dentro de uma janela terminando na amostra m , para um intervalo de tempo limitado. Para trechos sonoros de sinais de voz, pode-se mostrar que a autocorrelação assumirá valores máximos quando houver máxima similaridade com respeito a periodicidade, construindo-se assim um detector e rastreador de Pitch eficiente.

2.5 COEFICIENTES MEL-CEPSTRAIS

Os coeficientes mel-cepstrais (*Mel-Frequency Cepstral Coefficients* - MFCC), introduzidos por Davis e Mermelstein (1980), são os coeficientes cepstrais obtidos com base na escala mel ao invés da escala linear de frequências.

Na estimação dos MFCC, os sinais de voz devem ser pré-processados para obter um conjunto de dados com informações úteis e sem a presença de ruído. Dessa forma, o sinal é submetido a um filtro de pré-ênfase para em seguida ser segmentado, onde cada segmento é então ponderado espectralmente por uma janela de Hamming e submetido ao DAV para que sejam descartados os quadros que não contêm informações úteis (CARDOSO, 2009). Por fim, é feita a extração dos coeficientes mel-cepstrais utilizando abordagens como a de banco de filtros em escala mel ou de predição linear.

A escala mel surgiu com o intuito de mapear a percepção de Pitch em uma escala não linear e foi concebida a partir de experimentos, onde se mapearam os incrementos subjetivos constantes de Pitch em suas correspondentes frequências (PICONE, 1993; CARDOSO, 2009).

Neste trabalho, a estimação dos coeficientes mel-cepstrais é feita utilizando bancos de filtros. Assim, após a fase de pré-processamento de sinais de voz, é calculado o módulo da transformada de Fourier dos quadros com informação útil e os espectros destes segmentos são submetidos a um banco de filtros em escala mel, para enfatizar as componentes presentes nas frequências centrais, atenuando as demais. Em seguida, é tomado o logaritmo do espectro resultante e calculada a sua DCT (CARDOSO, 2009). A saída dos filtros de escala mel é denotada por $X(m)$ e os MFCC, denotados por C_n , onde n é o índice do coeficiente cepstral, são obtidos da seguinte forma:

$$C_n = \sum_{m=1}^M [\log X(m)] \cos\left(\frac{\pi n}{m} \left(m + \frac{1}{2}\right)\right) \quad m = 1, \dots, M. \quad (9)$$

Ao fim da estimação dos MFCC, é possível aplicar técnicas para a remoção de distorções presentes no sinal de voz, como por exemplo, a técnica de Subtração da Média Cepstral (SMC) de Kermorvant (1999).

3. ANÁLISES WAVELET E MULTIFRACTAL

A transformada de Fourier é amplamente utilizada em processamento digital de sinais. Entretanto, ao se determinar todas as componentes de frequência de um sinal, não é possível saber o instante em que elas ocorrem, ou seja, não há informação de tempo.

A análise de sinais com wavelets permite a extração de dados coerentes tanto no domínio da frequência quanto no domínio do tempo (ou espaço, para imagens). A análise com wavelets pode ser vista como uma “decomposição atômica”, onde se buscam os componentes básicos dos sinais, “os átomos”. Uma vez descritos os átomos do sinal, a combinação e a produção de novas moléculas se tornam mais fáceis. Por exemplo, numa partitura musical temos um arranjo de átomos (as notas) que possuem duração e frequência determinadas (FARIA, 1997). As wavelets também podem ter caráter fractal com padrões que se repetem em escalas diferentes.

A seguir são apresentados alguns conceitos e definições das análises wavelet e multifractal.

3.1 ANÁLISE WAVELET

Para que uma função $\psi(t)$ seja considerada uma wavelet, as seguintes condições básicas são necessárias:

- a) $\psi(t) \in L^2(\mathfrak{R})$, ou seja, a função pertence ao espaço das funções de quadrado integrável ou, ainda, o espaço das funções de energia finita, no sentido que:

$$\int_{-\infty}^{+\infty} |\psi(t)|^2 dt < \infty. \quad (10)$$

- b) Transformada de Fourier $\hat{\psi}(\omega)$ satisfaz a condição de admissibilidade (DAUBECHIES, 1992):

$$C_\psi = \int_{-\infty}^{+\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty. \quad (11)$$

Da condição de admissibilidade tem-se:

$$\lim_{\omega \rightarrow 0} \hat{\psi}(\omega) = 0. \quad (12)$$

Assim, se $\hat{\psi}(\omega)$ é contínua então, $\hat{\psi}(0) = 0$, ou seja,

$$\int_{-\infty}^{+\infty} \psi(t) dt = 0. \quad (13)$$

3.1.1 Transformada wavelet contínua (CWT)

A Transformada Wavelet Contínua (CWT), $W_f(a, b)$, de um sinal $f(t)$ é definida como:

$$W_f(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t) \psi^* \left(\frac{t-b}{a} \right) dt, \quad (14)$$

onde ψ^* é o conjugado complexo da função wavelet mãe ψ , a é o parâmetro de dilatação ou escala e b , o parâmetro de translação (deslocamento), sendo ambos os parâmetros reais, e a positivo.

3.1.2 Transformada wavelet discreta (DWT)

A implementação da Transformada Wavelet Discreta (DWT) reduz o custo computacional de executar a CWT, pois na DWT os parâmetros de dilatação e translação são discretizados, usando a escala diádica da seguinte forma:

$$a = 2^j, b = 2^j k, \quad (15)$$

sendo j e k inteiros positivos.

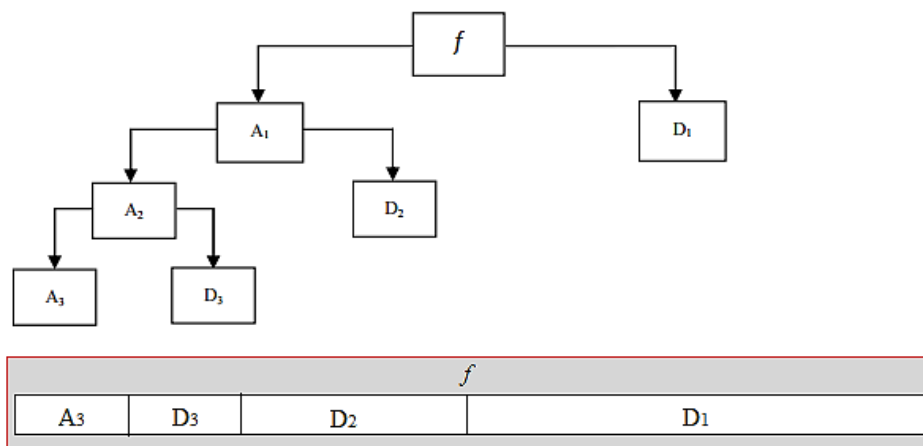
De acordo com Mallat (1989), a DWT pode ser vista como uma operação de filtragem através de um filtro wavelet em uma banda particular de frequência, que varia ao longo do tempo. A faixa de frequência do filtro depende do nível de decomposição, sendo possível realizar o exame local do sinal pelo deslocamento no domínio do tempo. Assim, o sinal pode ser decomposto em sub-bandas em uma estrutura de árvore, que

pode ser vista como uma estrutura de banco de filtros. Uma representação da $f(t)$, em um processo inverso, pode ser feita usando os coeficientes wavelet de detalhes e aproximações em vários níveis, como segue,

$$f_{DWT}(t) = \sum_{i=1}^j D_i(t) + A_j(t). \quad (16)$$

Na Equação (16), D representa os detalhes wavelet e A é a aproximação wavelet no j -ésimo nível de decomposição. Dessa forma, por exemplo, se $j = 2$, $f_{DWT}(t) = D_1(t) + D_2(t) + A_2(t)$. Uma representação gráfica da decomposição de um sinal f em três níveis pela DWT é mostrada na Figura 3.

Figura 3 - Árvore de decomposição da DWT, em três níveis, e o sinal após a DWT.



Fonte: Elaboração da própria autora.

3.1.3 Transformada wavelet packet (WPT)

Uma wavelet packet é representada como uma função $\psi_{j,k}^i(t)$, onde i é o parâmetro de modulação, j é o parâmetro de dilatação, k é o parâmetro de translação e n é o nível de decomposição na árvore gerada pela decomposição wavelet packet, dada por:

$$\psi_{j,k}^i(t) = 2^{-j/2} \psi^i(2^{-j}t - k), \text{ para } i = 1, 2, \dots, j^n. \quad (17)$$

A wavelet ψ^i é obtida pelas relações recursivas das Equações (18) e (19):

$$\psi^{2i}(t) = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} h(k) \psi^i\left(\frac{t}{2} - k\right). \quad (18)$$

$$\psi^{2i+1}(t) = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} g(k) \psi^i\left(\frac{t}{2} - k\right). \quad (19)$$

A função ψ^i é conhecida como a wavelet mãe e os filtros discretos $h(k)$ e $g(k)$ são filtros em quadratura espelhados, associados, respectivamente, à função escala e à função wavelet mãe (DAUBECHIES, 1992).

Os coeficientes da wavelet packet $c_{j,k}^i$, correspondentes para o sinal $f(t)$, podem ser obtidos pela Equação (20), satisfazendo a condição de ortogonalidade (WICKERHAUSER, 1994).

$$c_{j,k}^i = \int_{-\infty}^{\infty} f(t) \psi_{j,k}^i(t) dt. \quad (20)$$

A componente wavelet packet do sinal em um nó particular pode ser obtida pela Equação (21).

$$f_j^i(t) = \sum_{k=-\infty}^{\infty} c_{j,k}^i \psi_{j,k}^i(t). \quad (21)$$

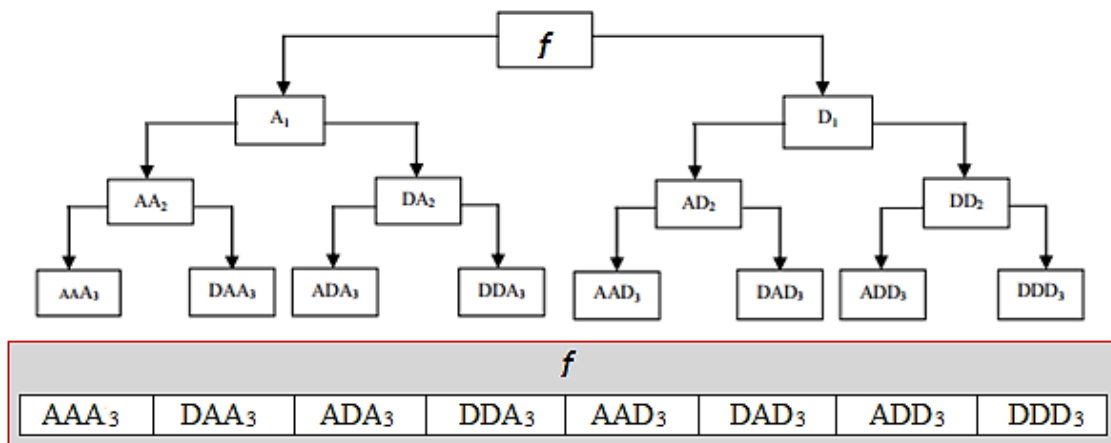
Após realizar a decomposição wavelet packet até o j -ésimo nível, o sinal $f(t)$ pode ser representado como uma soma de todos os componentes wavelet no nível j , como mostrado na Equação (22).

$$f_{WPT}(t) = \sum_{i=1}^{2^j} f_j^i(t). \quad (22)$$

Uma representação gráfica da árvore de decomposição de um sinal pela WPT, em três níveis, é mostrada na Figura 4.

Assim, a WPT pode ser vista como um banco de filtros (MALLAT, 1989), servindo para isolar os diferentes componentes de frequência de um sinal e funcionando como um arranjo de filtros passa-faixa que decompõe o sinal em diversas componentes.

Figura 4 - Árvore de decomposição da WPT, em três níveis, e o sinal após a WPT.



Fonte: Elaboração da própria autora.

Como a WT possui várias possíveis funções de decomposição, a aplicação da WPT nos sinais de voz pode ser otimizada de acordo com a seleção da função wavelet mais adequada para o processamento.

Para explorar os coeficientes da WPT na identificação de locutor, neste trabalho foram calculadas a energia, a entropia e a sensibilidade da WPT em sinais de voz.

A energia de um nó da wavelet packet representa a energia armazenada em uma faixa de frequência específica e é usada principalmente para extrair os componentes de frequência dominantes do sinal. A energia do nó da wavelet packet tem maior potencial para uso na classificação do sinal em comparação com o uso apenas de seus coeficientes (YEN; LIN, 2000).

A entropia E é uma função de custo aditivo tal que $E(0) = 0$. A entropia indica a quantidade de informação armazenada no sinal, ou seja, quanto maior a entropia maior é a informação armazenada no sinal e vice-versa. Existem várias definições de entropia na literatura especializada (WICKERHAUSER, 1994).

A entropia de energia da wavelet packet de um sinal em um determinado nó n da árvore wavelet packet é um caso especial da entropia P – norma definida na Equação (23), onde $P = 2$ e $c_{j,k}^i$ são os coeficientes WPT.

$$E_n = \sum_k |c_{j,k}^i|^P, \text{ para } P \geq 1. \quad (23)$$

A entropia de Shannon é definida pela Equação (24).

$$E_n = -\sum_k (c_{j,k}^i)^2 \log[(c_{j,k}^i)^2]. \quad (24)$$

A estimação da sensibilidade dos coeficientes da WPT, apresentada em Vieira (2016), consiste em comparar os coeficientes da WPT no último nível, denotado por C , obtidos entre dois sinais e, baseada nesta métrica, a estimação realizada neste trabalho foi adaptada para sinais de voz, onde os coeficientes C_B de um desses sinais servirão como base para verificar se os coeficientes C_I do sinal que se deseja identificar possuem semelhanças por meio da Equação (25),

$$\lambda = \frac{||C_I| - |C_B||}{C_B}. \quad (25)$$

Assim, λ pode representar a variação dos coeficientes da WPT devido a alguma diferença entre os sinais de voz do locutor de referência e o confrontado.

3.2 ANÁLISE MULTIFRACTAL

Esta seção apresenta os principais conceitos dos Líderes Wavelet baseados no Formalismo Multifractal necessários para construir e compreender a Média Máxima dos Líderes Wavelet aqui proposta. Para mais detalhes sobre a teoria multifractal, o leitor pode consultar, por exemplo, Jaffard (2004), Wendt e Abry (2007), Wendt *et al.* (2009) e Leonarduzzi *et al.* (2014)

A análise multifractal, que consiste principalmente na medição de expoentes de escala, é uma técnica padrão disponível na maioria das caixas de ferramentas de análise de dados empíricos. Essa análise é baseada na estimativa dos Líderes Wavelet, que é uma elaboração sobre os coeficientes wavelet (WENDT; ABRY, 2007).

3.2.1 Espectro multifractal

Com o espectro multifractal é possível fazer uma caracterização do sinal X , descrevendo as variações ao longo do tempo da regularidade do trajeto da sua amostra. Tal regularidade local é medida por meio dos expoentes Hölder $h(t)$. O expoente Hölder

quantifica a força do comportamento singular de X em torno de t_0 . Para uma introdução detalhada da análise multifractal, o leitor pode consultar Jaffard (2004).

Em vez de fazer uso de uma função do tempo $h(t)$, normalmente descreve-se a variabilidade do intervalo dos expoentes Hölder encontrado em X através do espectro multifractal (ou singularidade) $D(h)$, que é definida como a dimensão de Hausdorff dos conjuntos de pontos t_i nos quais $h(t_i) = h$ (WENDT; ABRY, 2007). Para mais detalhes sobre a dimensão de Hausdorff e a teoria fractal, o leitor pode consultar Mandelbrot e Frame (2003).

Assim, por meio da análise multifractal é possível inferir $D(h)$ a partir de uma duração finita e única de dados. Este processo é comumente referido como formalismo multifractal (JAFFARD, 2004).

Por muito tempo, os coeficientes wavelet foram considerados como as principais formas de análise multifractal empírica (RIBEIRO; RIEDI; BARANIUK, 2001). No entanto, em Jaffard (2004) é mostrado que a wavelet baseada no formalismo multifractal sofre duas grandes desvantagens: não permite alcançar todo o espectro multifractal do processo em análise e não é válida para todos os tipos de processos multifractais. Notadamente, processos contendo oscilações de singularidades são incorretamente analisados (WENDT; ABRY, 2007).

Um formalismo multifractal mais relevante pode ser obtido se os coeficientes wavelet forem substituídos pelos Líderes Wavelet (*Wavelet Leaders Multifractal Formalism* – WL, WLMF, ou também MF-WL). Por construção, os WL são monotonicamente crescentes com escala 2^j , cuja propriedade tem sido a chave para a concepção de um formalismo multifractal (JAFFARD, 2004). De fato, de acordo com Wendt e Abry (2007), sob leves condições de regularidade da trajetória da amostra, os WL reproduzem exatamente o expoente Hölder de $X(t)$ em t_0 , ou seja, $h(t_0)$ é o supremo de todos os valores h , com o limite de escalas finas, fazendo $2^j \rightarrow 0$.

Nas últimas décadas os Líderes Wavelet têm se mostrado uma ótima ferramenta para observar singularidades dos sinais e imagens, como observado em Gadhoumi *et al.* (2018) e Wendt *et al.* (2009), sendo possível com a análise multifractal ressaltar as informações contidas nas partes não estacionárias do sinal avaliado.

3.2.2 Líderes Wavelet

Considerando algumas definições e conceitos sobre a DWT descritos neste trabalho, será feita nesta seção a definição dos Líderes Wavelet.

Seja $X(t)$ o processo em análise e n a sua duração de observação, com $t \in [0, n)$. Supõe-se que a wavelet mãe $\psi_0(t)$ possui suporte compacto no tempo, então é feita a indexação $\lambda_{j,k} = \{k 2^j, (k+1) 2^j\}$ e a união $3\lambda_{j,k} = \lambda_{j,k-1} \cup \lambda_{j,k} \cup \lambda_{j,k+1}$.

Logo, de acordo com Wendt e Abry (2007), os Líderes Wavelet (*Wavelet Leaders* - *WL*) são definidos como:

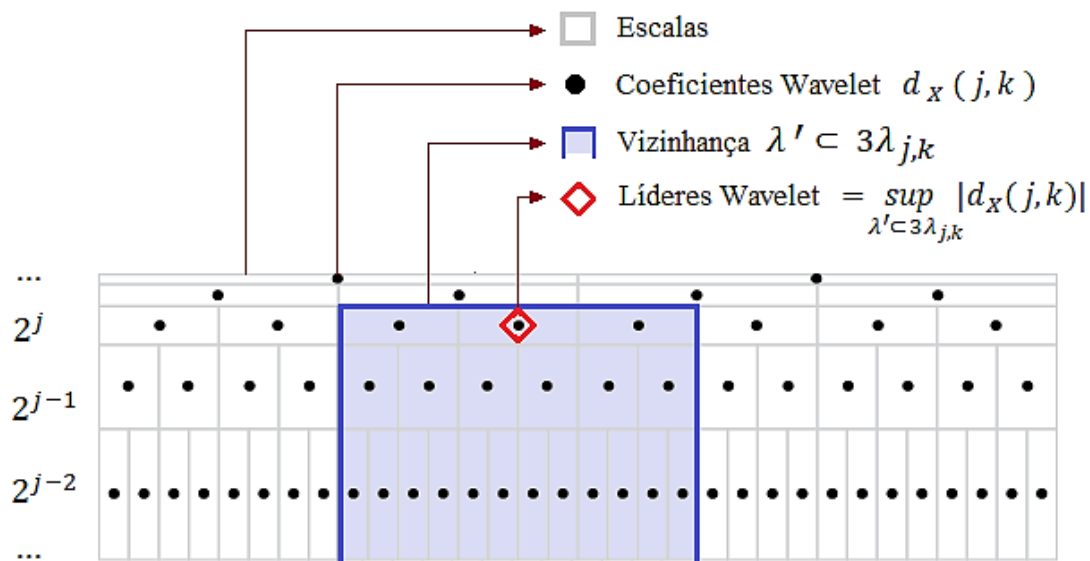
$$L_X(j, k) = \sup_{\lambda' \subset 3\lambda_{j,k}} |d_X(j, k)|, \quad (26)$$

onde o supremo é tomado dos coeficientes wavelet discretos $d_X(j, k)$ na vizinhança temporal $3\lambda_{j,k}$ sobre a escala e em todas as escalas mais finas $2^{j'} < 2^j$.

Os Líderes Wavelet L_X são calculados a partir dos coeficientes wavelet discretos $d_X(j, k)$ tomando o supremo na vizinhança de tempo $3\lambda_{j,k}$, sobre todas as escalas mais finas $2^{j'} < 2^j$. Cada escala mais fina tem o dobro do número de coeficientes da próxima escala mais grossa. Cada intervalo diádico na escala 2^j pode ser escrito como uma união de dois intervalos em uma escala mais fina.

De maneira simplificada, os WL quantificam a regularidade local e o Espectro de Singularidade indica o tamanho do conjunto de expoentes líderes nos dados (SERRANO; FIGLIOLA, 2009). A identificação dos Líderes Wavelet está ilustrada na Figura 5.

Figura 5 – Líderes Wavelet.



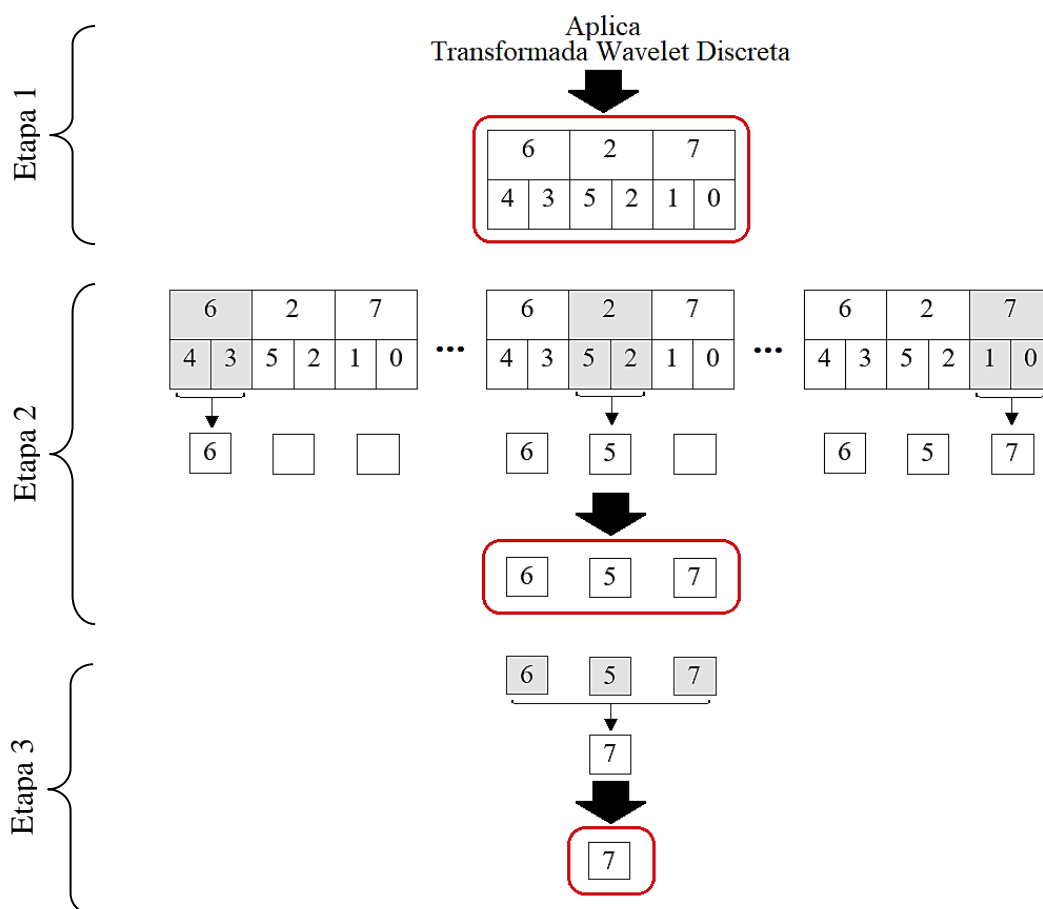
Fonte: Adaptado de Wendt e Abry (2007).

Em termos de algoritmo, seguindo a definição descrita em Wendt e Abry (2007), o cálculo dos Líderes Wavelet pode ser feito seguindo os seguintes passos:

1. Calcule os coeficientes wavelet $d_X(j, k)$, usando a DWT, e salve o valor absoluto de cada coeficiente para cada escala;
2. Comece na escala que é um nível mais grosso que a maior escala obtida;
3. Compare o primeiro valor com todos os seus maiores intervalos diádicos e obtenha o valor máximo;
4. Vá para o próximo valor e compare seu valor com todos os seus valores de escalas mais finas;
5. Continue comparando os valores com seus valores alinhados, obtendo o máximo;
6. A partir dos valores máximos obtidos para essa escala, examine os três primeiros valores e obtenha o máximo desses vizinhos. Esse valor máximo é um líder para essa escala;
7. Continue comparando os valores máximos para obter os outros líderes dessa escala;
8. Vá para a próxima escala mais grossa e repita o processo.

Na Figura 6 é apresentado um exemplo de como é feito o cálculo dos Líderes Wavelet, com o intuito de ilustrar o algoritmo de forma mais resumida.

Figura 6 – Exemplo do cálculo dos Líderes Wavelet.



Fonte: Elaboração da própria autora.

3.2.3 Média Máxima dos Líderes Wavelet

A análise dos sinais usando diretamente os Líderes Wavelet não é prática, pois exige uma metodologia específica para selecionar a escala dos WL mais apropriada. Assim, neste trabalho propõe-se, inicialmente, os Líderes Wavelet Máximos (*Maximum Wavelet Leaders - MWL*), evitando, assim, o uso de métodos mais complexos.

Os MWL_l são definidos como o valor absoluto (módulo) do máximo logaritmo dos Líderes Wavelet 1-D no nível $l = i + 1$, com i representando os Líderes Wavelet no nível l , ou escala $(2)^{i+1}$. Vale salientar que os Líderes Wavelet não foram definidos para o nível $l = 1$. Assim, se x é um sinal de áudio e $WL_{\{1,l\}}(x)$ são os Líderes Wavelet L_x para a dimensão 1, logo os MWL_l são definidos como:

$$MWL_l(x) = |\max(\log(WL_{\{1,l\}}(x)))|. \quad (27)$$

Nos estudos envolvendo estatísticas descritivas deste trabalho, foi observado que a média dos MWL poderia oferecer vantagens como um atributo de áudio em aplicações de reconhecimento de locutor.

Assim, o atributo de áudio proposto neste trabalho e denominado de Média Máxima dos Líderes Wavelet (*Maximum Mean Wavelet Leaders - MMWL*), é obtido por meio da média dos Líderes Wavelet Máximos do valor espectral s_k para o *bin* k , denotado por $\mu = \frac{1}{l} \sum_{i=1}^l MWL_l(s_k)$ onde l é o nível dos Líderes Wavelet 1-D, cuja definição é dada por:

$$MMWL = \sum_{k=b_1}^{b_2} s_k \mu. \quad (28)$$

Na Equação (28), b_1 e b_2 são as bordas da banda do sinal, em *bins*, sobre as quais é calculada a MMWL Espectral.

Diferentes funções wavelet podem ser utilizadas para obter a MMWL, mas, neste trabalho foi utilizada a wavelet *biorthogonal 1.5*, cuja função já vem pré-definida como padrão para o cálculo dos Líderes de Wavelet 1-D no *software* MATLAB.

Além disso, todos os atributos (características) de áudio utilizados neste trabalho, incluindo a MMWL Espectral, foram obtidos de segmentos da voz de acordo com um processo de janelamento/segmentação usando a janela de Hamming, em quadros com 496 amostras e sobreposição de 94%. Esta percentagem alta de sobreposição assegura uma análise espectral fina. Além disso, foi adotado o espectro de potência para o cálculo das características espectrais, embora o espectro de magnitude também pudesse ser usado.

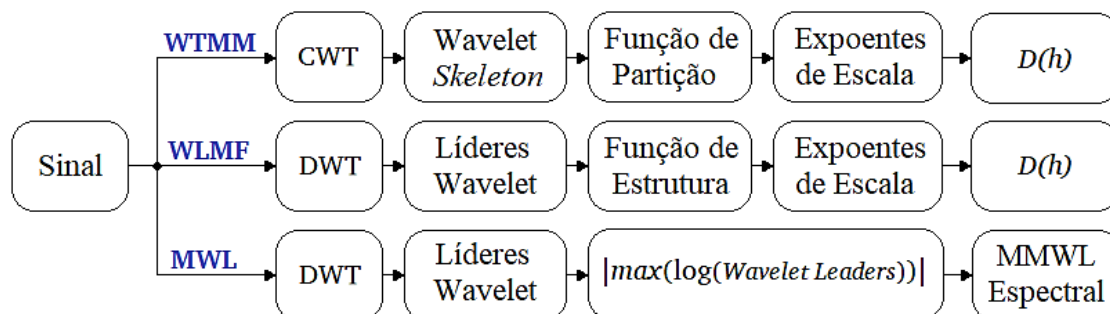
Os algoritmos de pré-processamento de sinais de áudio foram obtidos a partir do pacote da *Audio Toolbox* do MATLAB, cujas ferramentas tornaram possível o encapsulamento de múltiplos extratores de características de áudio. Dessa forma, a implementação da MMWL Espectral desenvolvida neste trabalho foi incorporada a esta extração de atributos de áudio otimizada do MATLAB. Alguns desses extratores/atributos, também chamados de descritores espectrais, podem ser encontrados em Peeters (2004).

Vale lembrar que a extração da MMWL Espectral depende do comprimento da janela usada para segmentar o sinal de áudio. Em particular, ao usar o MATLAB, é importante observar que os algoritmos usados para calcular as estimativas dos Líderes

Wavelet 1-D são oriundos da teoria do Formalismo Multifractal. Dessa forma, dependendo do número de amostras (comprimento da janela) do sinal, o nível mínimo de regressão dos WLMF pode não ser atingido.

Também vale a pena recapitular as ferramentas wavelet e multifractal discutidas neste trabalho e mencionar as diferenças entre os Líderes Wavelet Máximo (*Maximum Wavelet Leaders – MWL*), a Média Máxima dos Líderes Wavelet (*Maximum Mean Wavelet Leaders - MMWL*), os Líderes Wavelet - Formalismo Multifractal (*Wavelet Leaders Multifractal Formalism – WLMF*) e a abordagem apresentada em Arneodo *et al.* (2002) chamada Transformada Wavelet Módulos Máximos (*Wavelet Transform Modulus Maxima - WTMM*). Enquanto que os WLMF e a WTMM estão relacionados com o expoente Hölder h . Os MWL baseado nos WLMF não têm qualquer relação com h demonstrada até o presente momento. Além disso, o atributo da MMWL proposto neste trabalho, também denominada por MMWL Espectral, é obtido a partir da média quadro-por-quadro dos MWL. Um diagrama de blocos resumindo os algoritmos de WTMM, WLMF e MWL/MMWL está ilustrado na Figura 7.

Figura 7 – Esquema da WTMM, WLMF e MWL/MMWL para análise de sinais.



Fonte: Elaboração da própria autora.

Em resumo, a WTMM retorna uma estimativa do expoente Hölder usando a Transformada Wavelet Contínua (*Continuous Wavelet Transform - CWT*). Já os WLMF retornam os Líderes Wavelet, que também estimam os expoentes Hölder usando a Transformada Wavelet Discreta (*Discrete Wavelet Transform - DWT*). Os MWL propostos neste trabalho retornam os valores máximos absolutos dos logaritmos dos Líderes Wavelet obtidos de forma análoga aos WLMF e, conseqüentemente, a partir dos MWL é obtida a MMWL Espectral.

De acordo com Puchalski (2019), as implementações usando a WTMM oferecem desvantagens no formalismo multifractal para alguns tipos de sinais, uma vez que os

coeficientes wavelet podem dificultar a garantia de estabilidade numérica por estarem concentrados em valores próximos de zero, o que não é o caso dos métodos baseados em WLMF.

Além disso, foi verificado neste trabalho que o tempo de processamento para calcular a WTMM de um sinal de áudio é três vezes maior do que para obter os WLMF. Assim, como os MWL são baseados nos WLMF, a MMWL Espectral se torna uma ferramenta de baixo custo computacional. Vale ressaltar que a implementação da MMWL, na prática, é simples, pois não depende de uma estrutura contínua, como a WTMM.

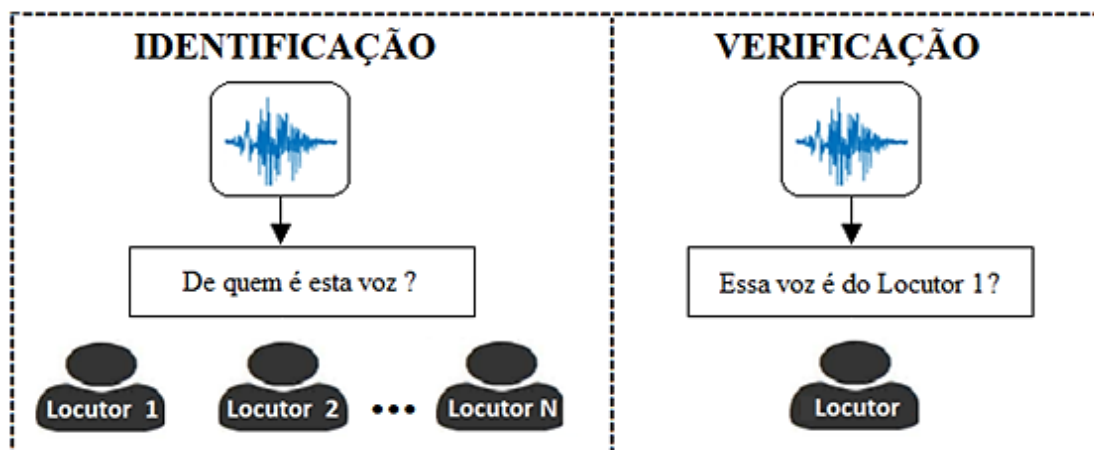
4. RECONHECIMENTO DE LOCUTOR

A área de reconhecimento de locutor possui uma grande variedade de aplicações e tem se expandido muito nos últimos anos. Nos sistemas de reconhecimento de locutor são usadas as técnicas de verificação ou de identificação de locutor e para compreendê-las é necessário definir alguns termos:

- **Locutor:** indivíduo que gera uma determinada locução.
- **Verificação de locutor:** processo em que se verifica se um determinado locutor é ou não o mesmo indivíduo indicado.
- **Identificação de locutor:** processo em que se identifica um determinado locutor a partir de um conjunto limitado de locutores.

A diferença entre os sistemas de identificação e verificação de locutor pode ser exemplificada pelas tarefas realizadas em cada sistema na Figura 8. Assim, em termos de aplicações práticas, um sistema de verificação poderia ser empregado para confirmar a identidade de uma pessoa que tenta acessar a sua conta, utilizando a própria voz como chave de acesso. Já para o caso da identificação, pode-se imaginar uma situação em que a partir de um banco de vozes de criminosos e de uma locução emitida por um criminoso desconhecido, deseja-se descobrir sua identidade (CARDOSO, 2009).

Figura 8 – Tarefas de identificação e verificação de locutor.



Fonte: Elaboração da própria autora.

O número ideal de componentes dos vetores utilizados na composição dos modelos de cada locutor não pode ser determinado analiticamente. Contudo, dependendo do classificador, quanto maior for a disponibilidade de locuções para a

realização dos treinamentos, maior deve ser o número de componentes a ser empregado para que se obtenham melhores índices na correta identificação (REYNOLDS, 1995).

Assim, para não ter o desempenho do sistema de identificação reduzido, a combinação de modelos com elevado número de componentes e pouco material de treinamento deve ser evitada, uma vez que, os modelos assim gerados acabam por apresentar uma correspondência grosseira com o respectivo locutor.

Outros aspectos observados e que devem ser levados em conta nos modelos de aprendizado de máquina são:

- **Tamanho das locuções:** locuções mais extensas garantem um maior refinamento na modelagem do aparelho fonador, uma vez que são disponibilizadas mais características do sinal de voz ao sistema de treinamento;
- **Condições de treinamento:** o treinamento dos modelos deve ser replicado para a situação de teste do sistema, pois se isto não acontecer, ocorrerá uma significativa degradação dos índices na correta identificação;
- **Canal de comunicação:** Deve-se empregar o mesmo canal de comunicação tanto na fase de treinamento quanto na de identificação, pois modelos, quando treinados a partir de locuções captadas com o emprego de determinado microfone, podem se mostrar incompatíveis com sinais amostrados utilizando-se outros equipamentos com características distintas na fase de teste;
- **Relação sinal-ruído da locução:** As características do sinal de voz, quando extraídas de um sinal contaminado por ruído, não refletem de forma fidedigna o locutor que as gerou e, portanto, acabam por comprometer a performance do sistema de identificação. Uma alternativa quando se trabalha com sinais de baixa relação sinal ruído é o emprego de técnicas de melhoria do sinal (CARDOSO, 2009).

4.1 VERIFICAÇÃO DE LOCUTOR

A verificação é composta basicamente pelas etapas de treinamento e reconhecimento. A etapa de treinamento normalmente é realizada antes de se tentar reconhecer qualquer amostra de voz. A etapa de verificação abrange a aquisição dos sinais de voz dos locutores que serão cadastrados no sistema, extração de informações

úteis dessas amostras, geração dos padrões de voz que serão utilizados como referências na etapa de reconhecimento e identificação dos limiares de similaridade associados aos padrões gerados (PETRY; SOARES; BARONE, 2003).

Uma das aplicações destinadas à verificação de locutor é o reconhecimento automático de locutor, onde uma pessoa pode interagir com máquinas através da fala. Os sistemas de reconhecimento automático de locutor (RAL) têm como objetivo a determinação automática do indivíduo emissor de uma locução, materializada num sinal de voz, através da comparação entre características extraídas da locução atual e locuções anteriores. Esses sistemas seguem basicamente etapas de verificação e de autenticação.

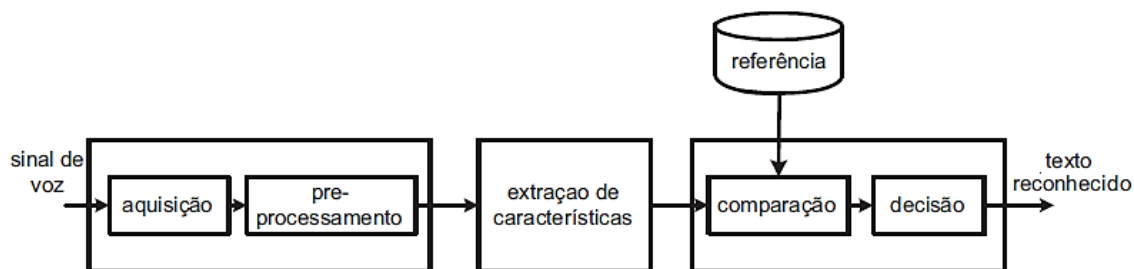
A etapa de autenticação ou reconhecimento automático abrange a aquisição do sinal de voz que será avaliado, a extração de informações úteis e a comparação dessas informações com os padrões gerados nas etapas anteriores. Nesse momento, o limiar de similaridade indicará se a identidade foi ou não aceita (PETRY; SOARES; BARONE, 2003).

Em aplicações de RAL, o sinal de voz não é apenas amostrado e classificado utilizando técnicas de reconhecimento de padrões. A identidade de um locutor está intrinsecamente associada às características fisiológicas e comportamentais da voz. As principais informações e características da voz devem ser extraídas do sinal amostrado. Para tal, a voz é inicialmente pré-processada e então são utilizados algoritmos para extração dos parâmetros do sinal vocal. O conjunto de parâmetros de uma amostra de voz compõe um padrão que pode ser classificado (PETRY; ZANUZ; BARONE, 1999).

As características mais utilizadas para RAL são os MFCC (DAVIS; MERMELSTEIN, 1980; PEETERS, 2004), principalmente se forem de dados controlados. Os dados controlados são obtidos a partir de sinais com boa relação sinal-ruído, captados com microfones de boa qualidade ou em canais cujas distorções são conhecidas ou invariantes no tempo (REYNOLDS, 1995).

Nos sistemas de RAL também pode ser inserida a tecnologia de reconhecimento automático da voz. A estrutura geral de um sistema de reconhecimento de voz é apresentada na Figura 9. É importante destacar que o reconhecimento automático de locutor não tem a mesma finalidade do reconhecimento automático de voz, pois um é destinado para reconhecer a pessoa e o outro para reconhecer a locução (texto dito).

Figura 9 - Estrutura geral de um sistema de reconhecimento de voz.



Fonte: Adaptado de Rabiner e Schafer (2010).

4.2 IDENTIFICAÇÃO DE LOCUTOR

A identificação de pessoas baseada em características individuais que são extraídas diretamente dos usuários é denominada biometria. A biometria por voz vem sendo empregada em diversas aplicações, tais como a autenticação de locutor, identificação criminal, controle de acesso, entre outras (ROSE, 2002; CORDELLA, 2003). A vantagem ao se utilizar informações biométricas extraídas de uma locução deve-se a facilidade com que esta é capturada do ambiente e processada, não exigindo equipamentos de elevada complexidade ou custo.

O sistema de identificação pode ser dividido em dois grandes blocos: o primeiro realiza a extração, onde cada característica representa um elemento do vetor no espaço de características, obtendo assim, as informações parametrizadas de cada locutor; já o segundo realiza as tarefas de identificação de padrões nas atividades de teste e treinamento dos modelos de cada locutor. A eficiência do sistema baseia-se na capacidade de reconhecer e classificar os sinais de áudio de acordo com as características extraídas (DHANALAKSHMI; PALANIVEL; RAMALINGAM, 2011).

Desta forma, no processo de identificação de locutor, as características do aparelho fonador são extraídas e convenientemente categorizadas. Assim, na fase de treinamento desses sistemas, o algoritmo empregado busca a cada iteração obter modelos cada vez mais correlacionados com os vetores de características fornecidos inicialmente. Por fim, na fase de teste e identificação é calculada a probabilidade de que determinado conjunto de vetores possa ter sido gerado por cada um dos modelos

presentes no sistema e, dessa forma, o locutor cujo modelo for o mais provável é identificado.

Existem diversas características dos sinais de voz, tais como, características temporais, espectrais e prosódicos. Entretanto, dependendo do sistema, nem todas são essenciais (BABAE *et al.*, 2017). O número de dimensões do espaço de características é igual ao número de características extraídas. Se a quantidade de atributos selecionados é muito alta, um problema de dimensionalidade ocorre (JAIN; DUIN; JIANCHANG, 2000).

Os vetores característicos do sinal de voz são necessários para gerar os modelos de cada um dos locutores, bem como servir de suporte para que o sistema possa identificar o respectivo locutor a partir de uma locução qualquer que lhe seja disponibilizada. O algoritmo empregado na fase de treinamento garante que a cada iteração sejam obtidos modelos cada vez mais correlacionados com os vetores de coeficientes dados inicialmente.

Antes da extração é de suma importância realizar o pré-processamento de sinais de áudio, com intuito de obter uma representação do sinal robusta e adequada, eliminando problemas como, por exemplo, ruídos ou redundâncias nas amostras. A etapa de pré-processamento envolve a redução de ruído, equalização, filtragem passa-baixa, e segmentação do sinal de áudio original em eventos de voz e silêncio para serem utilizados na extração de características (BABAE *et al.*, 2017).

A identificação de locutor é crucial em diferentes áreas e aplicações, como biometria, segurança, forense, entre outros. Ao comparar essas aplicações, é possível notar que a extração de características está presente em todas elas e isso permite supor que a extração de atributos de áudio poderia ser útil em diversos sistemas que utilizam sinais de voz. Entretanto, este trabalho estudou a extração dos atributos propostos somente para o sistema de identificação de locutor.

4.3 APRENDIZADO DE MÁQUINA

Uma das dificuldades encontradas na identificação de locutores está na forma de exploração das informações presentes nos dados, pois dependendo da aplicação, a quantidade de dados pode aumentar de forma expressiva. Para analisar estes dados, uma das ferramentas utilizadas atualmente é o aprendizado de máquina (*machine learning*).

No aprendizado de máquina, os dados são estruturados para formar conjuntos de dados de treinamento e de teste. Estes conjuntos contêm instâncias, geralmente representadas por um conjunto de tamanho fixo e com variáveis numéricas ou nominais (características associadas a cada instância), chamadas de atributos (FABRIS; MAGALHÃES; FREITAS, 2017).

O aprendizado de máquina pode ser dividido em duas técnicas básicas:

- **Aprendizado supervisionado:** O treinamento do modelo é feito com dados conhecidos de entrada e saída para que após essa etapa ele possa prever resultados futuros. Resumidamente, essa técnica utiliza dados rotulados vinculados a cada instância para a construção de modelos capazes de realizar previsões em dados não rotulados;
- **Aprendizado não supervisionado:** O algoritmo realiza uma busca por padrões ocultos ou estruturas intrínsecas nos dados de entrada. Assim, nessa técnica não há utilização de dados rotulados, sendo necessárias etapas adicionais de rotulagem ao final do processo, de modo a identificar potenciais elementos no conjunto de dados.

Além dessas duas técnicas, há também a técnica conhecida como método semi-supervisionado. O método semi-supervisionado é um método intermediário, no qual o algoritmo realiza o treinamento com um conjunto de dados rotulados (conjunto menor) para gerar um modelo capaz de rotular os dados de um conjunto não rotulado (conjunto maior). Em seguida, os novos dados rotulados são adicionados iterativamente ao conjunto de treinamento, melhorando assim o modelo a cada ciclo realizado (LIBBRECHT; NOBLE, 2015).

4.3.1 Seleção do modelo de aprendizagem

Para escolher o tipo de técnica mais apropriada, é necessário compreender o funcionamento de cada técnica de aprendizado, exemplificadas a seguir.

De acordo com Kubat (2015), o aprendizado supervisionado se concentra na indução de classificadores, enquanto que o aprendizado não supervisionado está interessado em descobrir propriedades úteis dos dados disponíveis.

No aprendizado supervisionado, duas técnicas que se destacam são a de regressão, para predição de dados contínuos, e o da classificação, no caso de dados discretos. Nas técnicas de regressão é possível prever respostas contínuas, como por exemplo, mudanças de temperatura em um determinado processo; já nas técnicas de classificação são utilizados modelos que classificam os dados de entrada em categorias para prever respostas discretas como, por exemplo, para decidir se um determinado e-mail é genuíno ou *spam*, em um processo de classificação de e-mails. Já o aprendizado não supervisionado tenta encontrar padrões ocultos ou estruturas intrínsecas nos dados, obtendo dados de entrada sem respostas rotuladas. Neste tipo de aprendizado, o agrupamento (*clusters*) é a técnica mais popular, sendo utilizada em aplicações como pesquisa de mercado, análise de sequência genética, entre outros (KUBAT, 2015; THE MATHWORKS INC, 2016).

Vale ressaltar que existem diversas abordagens com algoritmos de aprendizado de máquina e que a escolha do melhor método não é direta, pois depende de vários fatores, como o tamanho, os tipos de dados, dos *insights* obtidos dos dados, da forma como esses *insights* serão usados, etc.

Entretanto, em Babae et al. (2017) são apresentados alguns resultados com relação à classificação de sinais de áudio utilizando aprendizado de máquina, onde os melhores resultados são produzidos por algoritmos de aprendizado supervisionado, tendo alcançado uma média na acurácia de 90,13%, enquanto que nos métodos semi-supervisionados e sem supervisão a acurácia foi de 82,99% e 81,07%, respectivamente.

Pelo fato de a metodologia supervisionada ser considerada um método de alta precisão na classificação e detecção para sinais de áudio, ela é utilizada neste trabalho para identificação de locutor e nele são estudados o desempenho dos algoritmos mais abordados na literatura.

A fim de obter um bom desempenho dos modelos na predição com aprendizado de máquina supervisionado, os seguintes procedimentos devem ser adotados:

- Realizar a seleção dos atributos, que são as características vinculadas a cada instância, antes deles serem utilizados nos modelos de aprendizado de máquina, para obter uma boa predição;
- Fazer ajustes e filtrar as instâncias a serem utilizadas, com intuito de remover os dados redundantes e os dados discrepantes (*outliers*);
- Normalizar os dados para manter a proporcionalidade;

- Inserir as instâncias de forma aleatória para prevenir a inserção de viés;
- Realizar o balanceamento dos dados;
- Separar as instâncias, ou seja, colocar de forma separada o conjunto de treinamento, o conjunto de teste e o conjunto de validação, para garantir a eficácia do modelo;
- Definir a quantidade de instâncias a serem utilizadas, pois um número pequeno pode acarretar baixa performance do modelo e um número elevado pode melhorar a performance, mas aumenta significativamente o custo computacional.

Destes procedimentos, o de maior importância é a seleção prévia dos atributos. Em Libbrecht e Noble (2015) é demonstrado que realizar tais procedimentos, como seleção a priori dos atributos, possibilita uma melhora significativa na performance do classificador, em termos de acurácia.

Além disso, a seleção de atributos é importante para prevenir o *overfitting* e fornecer modelos rápidos e de menor custo efetivo computacional (GUYON; ELISSEEFF, 2003). O termo *overfitting*, muito utilizado na área estatística, descreve uma situação na qual um modelo gerado se ajusta somente na base de dados que o gerou, obtendo baixa performance quando aplicado em uma base de dados desconhecida (ALMEIDA, 2018).

4.3.2 Classificador KNN

Vários algoritmos utilizados no aprendizado de máquina são baseados em classificadores, tais como, análise de discriminante (*Discriminant Analysis*), Máquina de Vetores de Suporte (*Support Vector Machine - SVM*), árvores de decisão (*Decision Trees*), regressão logística, K Vizinhos Mais Próximos (*K Nearest Neighbors - KNN*), redes Bayesianas (*Naive Bayes*), classificação de conjuntos (*Ensemble Classification*), entre outros (KANDOI; ACENCIO; LEMKE, 2015).

Neste trabalho, a escolha do classificador levou em consideração o modelo que obteve maior percentual de acurácia analisando os WL extraídos de um pequeno conjunto de dados de dois locutores masculinos com o aplicativo *Classification Learner* do software MATLAB® R2019b – versão estudante, sendo escolhido o algoritmo de classificação K Vizinhos Mais Próximos.

Os algoritmos de classificação KNN são considerados os mais simples entre aqueles utilizados no aprendizado de máquina. Ao contrário da maioria das técnicas, o KNN é uma técnica “preguiçosa”, o que significa que ele faz as avaliações somente quando necessário, mas de forma rápida e com boa predição. Embora esta abordagem pareça simples, ela é eficaz em uma série de algoritmos de aprendizado de máquina (SHARMA, 2019).

O KNN é utilizado como um método de predição que decide o valor previsto de X_{t+1} ao encontrar os k -vizinhos mais próximos dos dados de entrada P_{t+1} e usando as saídas observadas. A distância Euclidiana é normalmente usada para avaliar a semelhança (HUANG *et al.*, 2009). Assumindo que v_i são os valores de saída dos k -vizinhos encontrados, o valor previsto X_{t+1} pode ser determinado através do cálculo da média ponderada dos vizinhos pelo seguinte modo (LIN; LI; SADEK, 2013):

$$X_{t+1} = \frac{1}{k} \sum_{i=1}^k v_i. \quad (29)$$

Desta forma, o KNN basicamente armazena todos os dados de treinamento, compara novos pontos com estes dados e retorna a classe mais frequente dos K - pontos mais próximos. O KNN é também uma abordagem robusta que é capaz de segmentar e classificar fluxos de áudio em fala, música, som ambiente e silêncio (LIE; HONG-JIANG; HAO, 2002).

Segundo Babae *et al.* (2017), dentre os trabalhos envolvendo algoritmos de aprendizado supervisionados, aqueles que utilizavam KNN obtiveram alto desempenho com relação a outras técnicas, como em Khunarsal, Lursinsap e Raicharoen (2013), onde foi obtida uma acurácia de 90,57%.

4.3.3 Avaliação e validação do modelo

Para ilustrar de forma simples a performance dos algoritmos de aprendizado de máquina, bem como mostrar o desempenho da metodologia proposta, os resultados obtidos neste trabalho foram apresentados usando a matriz de confusão (*confusion matrix* ou *confusion matrix chart*). Basicamente, essa matriz mostra a relação entre a classe verdadeira (*true class*) conhecida como categoria real (*actual class*) e a classe prevista (*predicted class*), fornecendo os resultados do modelo para cada classe ao

descrever o cumulativo de sucessos (êxitos) e fracassos (erros) da predição sobre o conjunto de dados, como ilustrado na Figura 10.

Figura 10 – Matriz de Confusão.

Positivo Verdadeiro <i>(True Positive)</i> TP	Negativo Falso <i>(False Negative)</i> FN	(Positivo) P	Classe Verdadeira (Real Classe)
Positivo Falso <i>(False Positive)</i> FP	Negativo Verdadeiro <i>(True Negative)</i> TN	(Negativo) N	
(Positivo) P	(Negativo) N	Classe Prevista	

Fonte: Elaboração da própria autora.

Na Figura 10, são esquematizados os seguintes valores:

- **Positivo Verdadeiro, também chamado Verdadeiro Positivo (*True Positive* — TP):** ocorre quando, no conjunto real, a classe que se está buscando foi prevista corretamente;
- **Positivo Falso, também chamado Falso Positivo (*False Positive* — FP):** ocorre quando, no conjunto real, a classe que se está buscando foi prevista incorretamente;
- **Negativo Verdadeiro, também chamado Verdadeiro Negativo (*True Negative* — TN):** ocorre quando, no conjunto real, uma classe que não se está buscando foi prevista corretamente;
- **Negativo Falso, também chamado Falso Negativo (*False Negative* — FN):** ocorre quando, no conjunto real, uma classe que não se está buscando foi prevista incorretamente.

As medidas de desempenho são calculadas de acordo com a capacidade do modelo de prever corretamente os rótulos do conjunto de teste (HAN; KAMBER, 2012; WITTEN; FRANK; HALL, 2011). Neste trabalho, a avaliação do desempenho preditivo dos modelos nos conjuntos de teste foi feita calculando-se a acurácia (*accuracy*), mas existem outras métricas, como sensibilidade (*recall*), *F-score*, ROC (*Receiver Operating Characteristic*), especificidade, precisão, taxa de falsa descoberta, coeficiente de correlação de Matthews, entre outras.

A acurácia foi escolhida como parâmetro de avaliação de desempenho porque é uma das métricas mais utilizadas na literatura voltada para identificação de locutor, o que facilita comparações. Entretanto, vale ressaltar que o uso de outras métricas em alguns conjuntos de dados pode levar a uma conclusão mais precisa dos resultados. Para obter a acurácia dos modelos é usada a seguinte expressão:

$$Acurácia = \frac{TP+TN}{TP+TN+FP+FN}. \quad (30)$$

Vale lembrar que a medida de acurácia é calculada para o conjunto de treinamento e para o conjunto de teste. No entanto, este último é mais relevante porque informa a capacidade do classificador de prever novos dados desconhecidos.

Na validação da metodologia proposta é utilizado o algoritmo de validação cruzada (*cross-validation*). Essa técnica separa aleatoriamente o conjunto de dados em 10 subconjuntos contendo uma proporção aproximada de igualdade das classes. Dessa forma, cada subconjunto é isolado por vez, para que o algoritmo possa ser aplicado em cada um dos 9 subconjuntos restantes, gerando assim, um submodelo que será aplicado no subconjunto previamente isolado. Por fim, após esse processo ser executado 10 vezes, uma estimativa de erro do algoritmo de aprendizagem é obtida.

5. METODOLOGIA PROPOSTA E RESULTADOS

Neste capítulo é apresentada a metodologia proposta para a identificação de locutor e seus resultados, bem como os fundamentos que a motivaram. Assim, nas primeiras seções são apresentados os estudos prévios e os seus progressos até chegar na formulação e obtenção de uma metodologia eficiente e robusta. Após essa definição, são apresentados os resultados obtidos aplicando a metodologia proposta neste trabalho em dois bancos de voz da literatura. Por fim, são apresentados os resultados de experimentos adicionais realizados com o atributo de áudio construído neste trabalho.

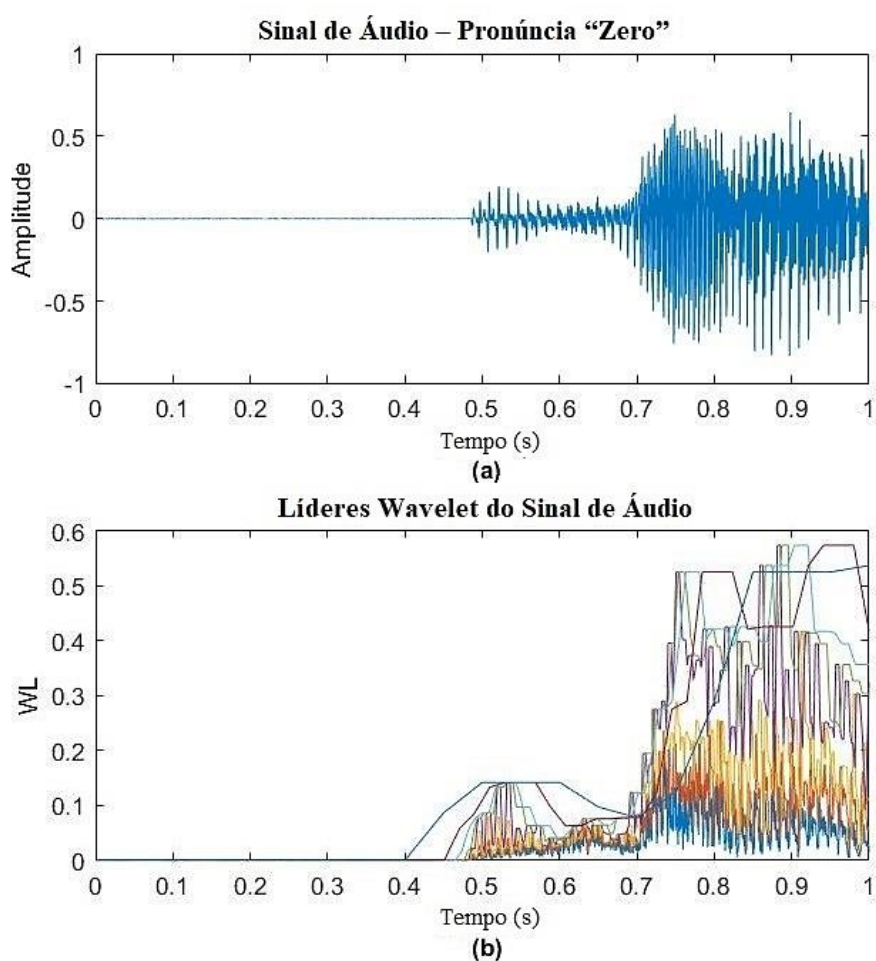
5.1 MMWL NA IDENTIFICAÇÃO DE LOCUTORES

Esta seção apresenta os estudos iniciais feitos usando Líderes Wavelet para obter o atributo proposto chamado neste trabalho de MMWL. Assim, a Figura 11 ilustra o sinal de áudio e os Líderes Wavelet por níveis/escalas (indicados pelos espectros de várias cores da Figura 11) para um sinal de voz de um locutor masculino pertencente ao banco de dados *Speech Commands* de Warden (2018), cuja pronúncia(sentença) registrada foi: “Zero”.

A base de dados de Warden (2018) contém arquivos de áudio em formato *wav*, na qual cada arquivo contém a pronúncia de uma única palavra em inglês, amostrada a uma taxa de 16 kHz e com duração de aproximadamente 1 segundo. Os arquivos de áudio foram coletados em locais não controlados por pessoas do mundo todo.

Na Figura 11 pode-se verificar que a análise do sinal usando diretamente os WL não é prática, pois exige um procedimento específico para selecionar a escala dos WL mais apropriada. Dessa forma, para evitar métodos complexos e manter um padrão único de análise, neste trabalho a proposta inicial foi extrair os Líderes Wavelet Máximos.

Figura 11 – Análise do: (a) sinal de voz “Zero” usando (b) Líderes Wavelet.



Fonte: Elaboração da própria autora.

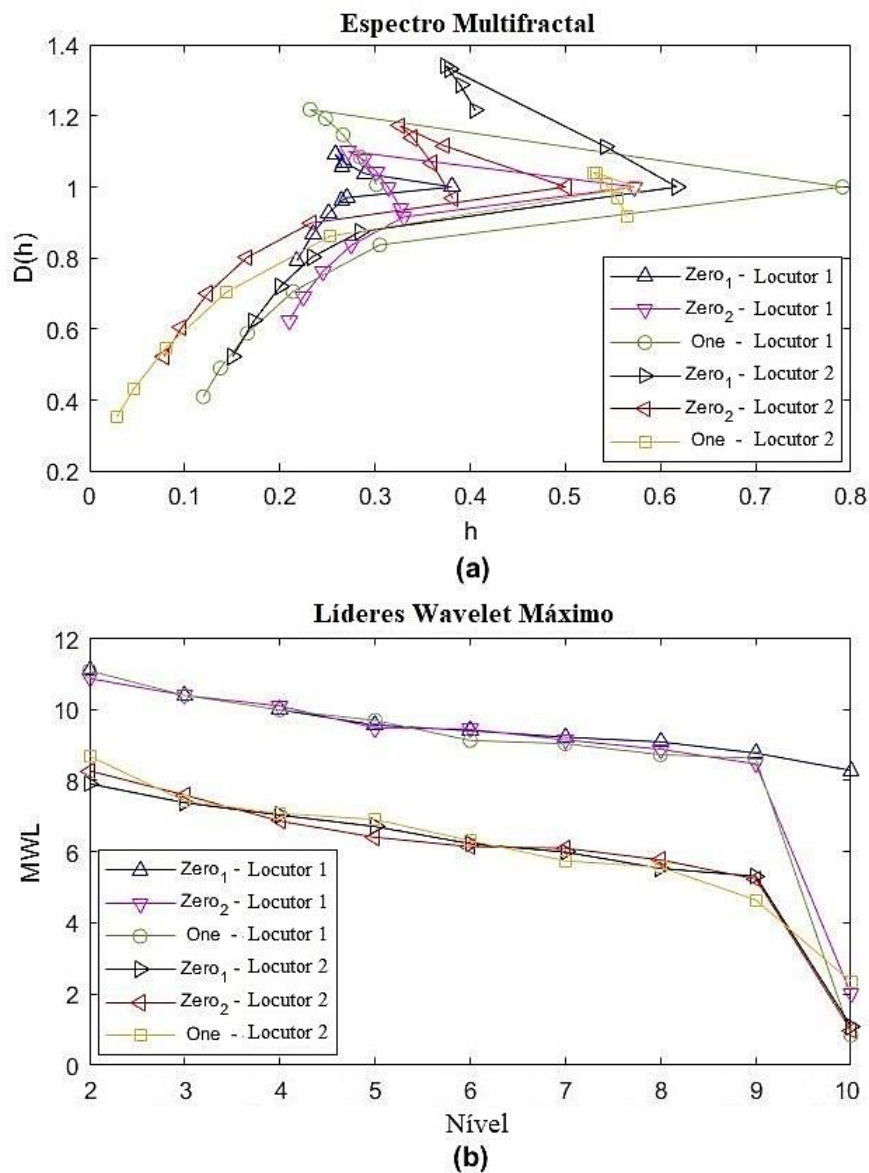
A partir do MWL é possível obter resultados interessantes para sinais de voz e ainda melhores do que os fornecidos pelo Espectro de Singularidades. Por exemplo, na Figura 12, tanto o MWL como o Espectro Multifractal são ilustrados para seis locuções de dois locutores masculinos, no formato de arquivo de áudio *wav*, pertencentes ao banco de dados *Speech Commands* descritos na Tabela 2. Os áudios das pronúncias de palavras que se repetem na Tabela 2 foram gravados em momentos diferentes, contendo diferenças na tonalidade e na entonação em que foram pronunciados. Cada arquivo de áudio possui as seguintes especificações:

- Número de canais de áudio codificados: 1;
- Taxa de amostragem, em hertz: 16000;
- Número total de amostras: 16000;
- Duração do arquivo, em segundos: 1;
- Número de bits por amostra codificada: 16.

Tabela 2 – Descrição dos arquivos de áudio.

Locutor	Palavra pronunciada	Nome do arquivo
Pessoa 1 (p1)	“Zero”	Zero ₁ – Locutor 1
	“Zero”	Zero ₂ – Locutor 1
	“One”	One – Locutor 1
Pessoa 2 (p2)	“Zero”	Zero ₁ – Locutor 2
	“Zero”	Zero ₂ – Locutor 2
	“One”	One – Locutor 2

Fonte: Elaboração da própria autora.

Figura 12 – Análise dos sinais de voz usando: (a) Espectro Multifractal e (b) Líderes Wavelet Máximo.

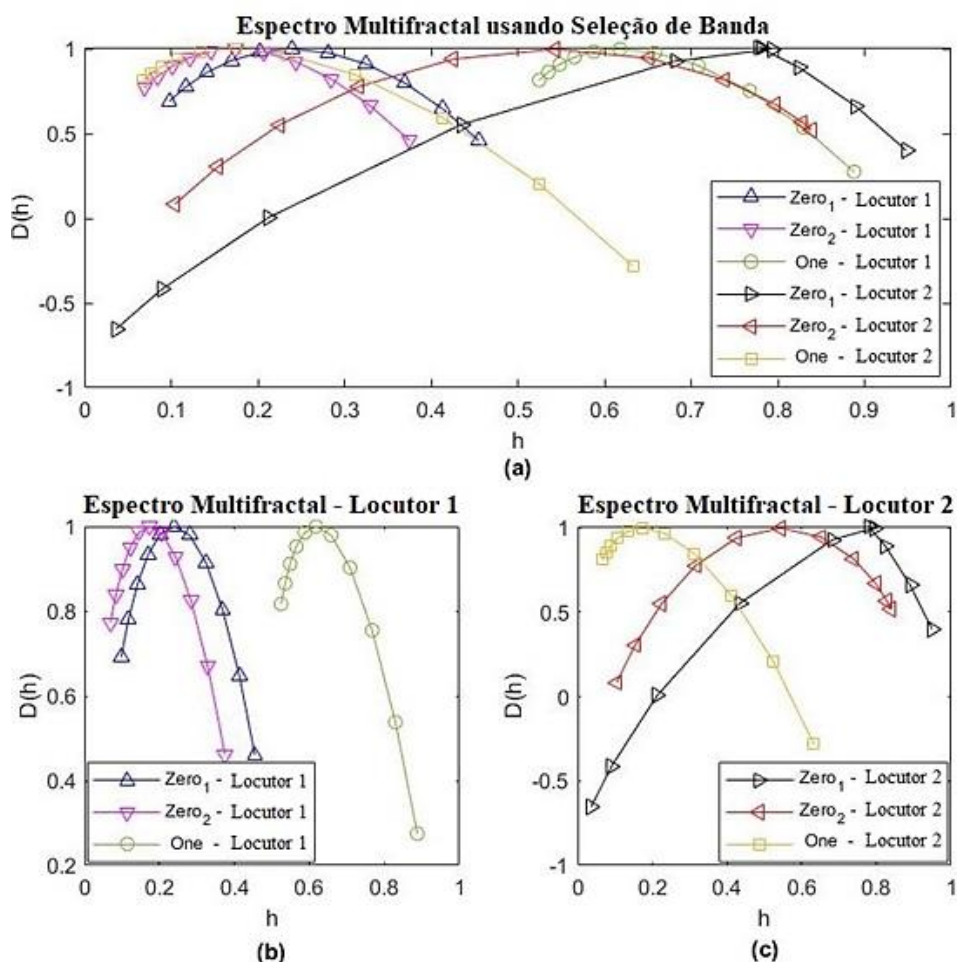
Fonte: Elaboração da própria autora.

Na Figura 12, os espectros de MWL do locutor 1 e do locutor 2 apresentam alguns padrões e semelhanças. Já no espectro multifractal $D(h)$ não é fácil visualizar semelhanças.

No início deste trabalho, o espectro multifractal foi estudado para encontrar semelhanças entre os sinais. Para isso, foi explorada a regularidade local nos áudios dos locutores, mas usando apenas algumas bandas do sinal. A sensibilidade da WPT foi usada como método de seleção de bandas neste trabalho. Essa métrica é uma adaptação da entropia da WPT para sinais de voz proposta por Vieira Filho e Duarte (2016).

A banda mais sensível dos sinais utilizando os valores do espectro multifractal foi obtida considerando o $D(h)$ do sinal denominado *Zero₁ - Locutor 1* como base e comparando-a com os $D(h)$ obtidos de outros sinais descritos anteriormente na Tabela 2. Os resultados estão na Figura 13(a). Também na Figura 13 são ilustrados os resultados contendo apenas o espectro multifractal do locutor 1, Figura 13(b), e locutor 2, Figura 13(c), respectivamente.

Figura 13 – Espectro multifractal dos sinais de voz com seleção de banda.



Fonte: Elaboração da própria autora.

Com a seleção da banda, Figura 13, é possível identificar os sinais de voz mais próximos da locução base. Estes resultados podem ser promissores se destinados à verificação do locutor dependente de texto. No entanto, o processo de seleção da banda pode elevar o custo computacional do sistema, de modo que a extração do MWL em sinais de voz acaba sendo a mais apropriada.

Embora o MWL não indique diretamente linearidade ou multifractalidade do sinal, o mesmo pode ser usado para estimar características relevantes dos sinais de fala dependentes do discurso e do locutor.

Portanto, como observado nestes resultados preliminares, o MWL é mais adequado para detectar semelhanças e características relevantes entre os sinais de voz sem utilizar métodos de seleção de banda, cujas implementações poderiam aumentar a complexidade e o custo computacional do sistema, especialmente no aspecto do reconhecimento de locutores.

A fim de melhor distinguir o discurso e os locutores utilizando características obtidas do MWL, foram calculadas e analisadas as seguintes estatísticas descritivas do MWL: soma, média, mediana e desvio padrão (*Standard Deviation - STD*). Os resultados são apresentados na Tabela 3.

Tabela 3 – Estatísticas aplicadas à MWL para sinais de voz de dois locutores.

Sinal de Áudio	Soma	Média	Mediana	STD
Zero ₁ – Locutor 1	85,81	9,53	9,40	0,85
Zero ₂ – Locutor 1	78,80	8,75	9,44	2,63
One – Locutor 1	77,54	8,61	9,12	3,02
Zero ₁ – Locutor 2	53,17	5,90	6,22	1,99
Zero ₂ – Locutor 2	53,36	5,92	6,14	2,07
One – Locutor 2	54,71	6,07	6,31	1,82

Fonte: Elaboração da própria autora.

Na Tabela 3 é possível notar que a mediana apresenta valores muito próximos para o mesmo locutor, independentemente da palavra pronunciada, o que é muito importante nas aplicações de identificação de locutor. Por outro lado, a média fornece resultados interessantes quando considerada a possibilidade de se explorar outras aplicações além do reconhecimento de locutor, como o reconhecimento de voz (fala). Em outras palavras, o parâmetro estatístico de média dos MWL mostra que a Média

Máxima dos Líderes Wavelet tem potencial para ser explorada tanto para o reconhecimento de locutor quanto para o reconhecimento de voz.

Assim, estes resultados promissores motivaram a extração da MMWL como atributo de áudio em modelos de aprendizado de máquina na identificação de pessoas.

5.2 MODELO DE IDENTIFICAÇÃO DE LOCUTOR

Focando em obter uma identificação automática, buscou-se uma metodologia robusta baseada no aprendizado de máquina. Para tal, foi realizado um estudo aplicando o aprendizado de máquina na identificação de locutores, utilizando os atributos da MMWL Espectral. Assim, o modelo com maior percentual de acurácia foi escolhido.

A MMWL Espectral foi extraída dos seis sinais de voz descritos na Tabela 2 para dois locutores e testada inicialmente em um algoritmo de classificação supervisionado escolhido aleatoriamente. O modelo obtido utilizou o classificador KNN para os seis arquivos de áudio, sendo usados dois desses arquivos de cada locutor para treino e um para teste. Os resultados podem ser interpretados com a matriz de confusão apresentada na Figura 14.

Figura 14 – Matriz de confusão para dois locutores do banco *Speech Commands*.

Acurácia de Validação

74	2	97.4%	2.6%	Classe Verdadeira p1
	82	100.0%		
100.0%	97.6%			
	2.4%			
p1 p2				
Classe Prevista				

Fonte: Elaboração da própria autora.

Na Figura 14 é possível observar que nos quadros com amostras de voz dos locutores, o modelo:

- Previu a pessoa 1 (p1): 74 vezes corretamente e 2 vezes incorretamente;
- Previu a pessoa 2 (p2): 82 vezes corretamente e 0 vezes incorretamente.

É importante ressaltar que a quantidade total de vezes que o modelo realiza a predição em cada locutor depende da quantidade de arquivos disponíveis para treino e teste e também da quantidade de quadros obtidos da extração janelada das amostras de voz dos arquivos.

Ao detalhar o desempenho do modelo é possível observar a eficiência na identificação por meio do percentual de confiança. Na Tabela 4 é ilustrado o percentual de confiança na identificação para a locução “Zero” dos locutores (p1 e p2).

Tabela 4 – Percentual de confiança para a palavra “Zero”.

Palavra Pronunciada	Locutor Verdadeiro/Real	Locutor Previsto	Percentual de Confiança
“Zero”	p1	p1	87,32
“Zero”	p2	p2	100,00

Fonte: Elaboração da própria autora.

A acurácia de validação foi de 98,73%, o que indicou ser uma boa característica para ser inserida na identificação de locutor. A partir disso, foi iniciado um estudo para encontrar os melhores modelos de classificação dentre os disponíveis no pacote de ferramentas do MATLAB para analisar as amostras dos sinais de voz usando aprendizado de máquina. Os resultados dos modelos de classificação usados nos testes para dois locutores do banco *Speech Commands* são apresentados na Tabela 5.

Tabela 5 – Acurácia dos classificadores (banco *Speech Commands*).

Classificador	Acurácia
<i>Decision Trees</i>	95,6%
<i>Discriminant Analysis</i>	94,3%
<i>Ensemble Classification</i>	100,0%
<i>KNN</i>	100,0%
<i>Naive Bayes</i>	95,6%
<i>Support Vector Machines (SVM)</i>	98,1%

Fonte: Elaboração da própria autora.

Na Tabela 5 é possível observar que os classificadores KNN e os de Classificação de Conjuntos (*Ensemble Classification*) apresentaram acurácia de 100%. Esses dois modelos possuem as seguintes definições:

- **Modelo KNN, número 5.14 do MATLAB:** Tipo *fine*, com somente 1 vizinho e utilizando a métrica de distância Euclidiana com pesos iguais de distância;
- **Modelo Classificação de Conjuntos, número 6.23 do MATLAB:** Método de subespaço KNN com dimensão 7 e utilizando 30 aprendizes do tipo KNN.

Para escolher qual dos dois classificadores seria o mais apropriado, foram analisados e comparados a velocidade de predição e o tempo de treinamento dos modelos para dois locutores do banco *Speech Commands*, Tabela 6.

Tabela 6 – Comparação entre modelos de classificadores (banco *Speech Commands*).

Modelo	Velocidade de predição (em observações/segundo)	Tempo de treinamento (em segundos)	Acurácia
KNN	~6600	0,34034	100,0%
Classificação de Conjuntos	~410	2,43590	100,0%

Fonte: Elaboração da própria autora.

Na Tabela 6 é observado que o modelo do classificador KNN treina e realiza a predição em menor tempo em comparação com o de Classificação de Conjuntos. Além disso, a Classificação de Conjuntos utiliza como melhor subespaço para treinamento justamente o KNN, concluindo assim, que o classificador KNN é o mais indicado.

Com o classificador KNN escolhido, restou verificar se o modelo KNN do tipo *fine* utilizado no treinamento era o mais indicado para os dados. Novamente, levando em consideração a acurácia, o tempo de treinamento e a velocidade de predição, foram analisados os tipos de modelos que usam o classificador KNN disponíveis no MATLAB. Os resultados dessas análises para dois locutores do banco *Speech Commands* são ilustrados na Tabela 7.

Na Tabela 7, ao observar os resultados, é possível notar que o modelo *Fine KNN* se destaca, indicando, portanto, que a escolha do mesmo é uma das mais apropriadas para a identificação de locutor.

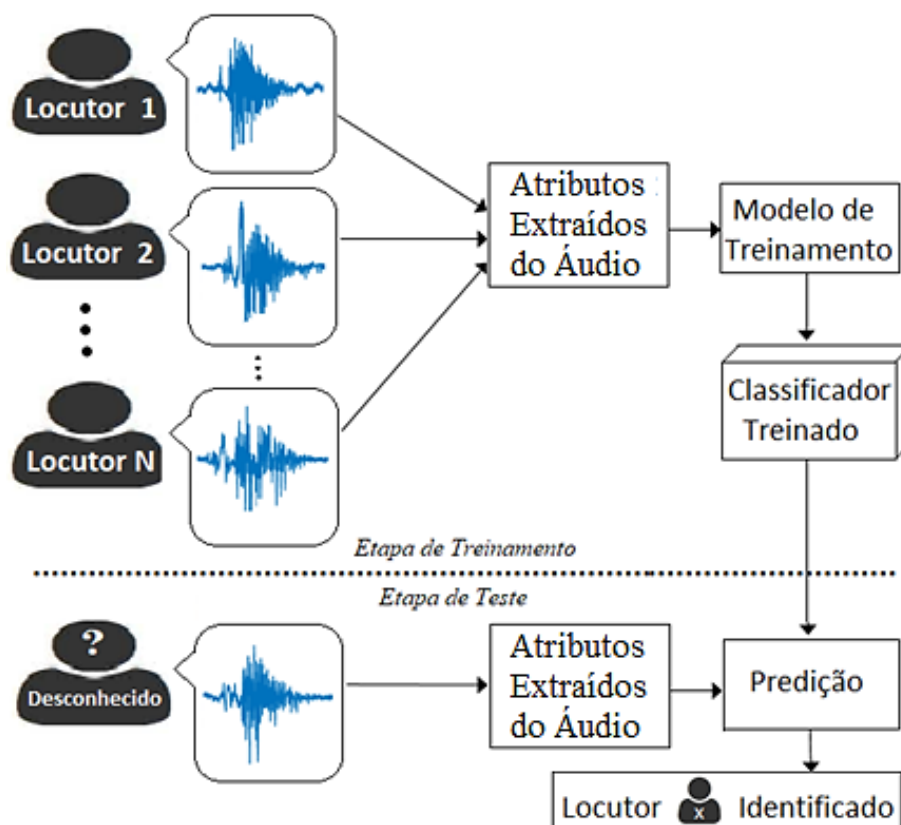
Tabela 7 – Comparação entre os tipos de modelo KNN (banco *Speech Commands*).

Tipo do Modelo (predefinido)	Acurácia	Tempo de treinamento (em segundos)	Velocidade de predição (em observações/segundo)
<i>Fine KNN</i>	100,0%	0,28189	~8400
<i>Medium KNN</i>	95,6%	0,31330	~7900
<i>Coarse KNN</i>	53,2%	0,20293	~8200
<i>Cosine KNN</i>	94,9%	0,20935	~8500
<i>Cubic KNN</i>	96,2%	0,26713	~7100
<i>Weighted KNN</i>	97,5%	0,19856	~7200

Fonte: Elaboração da própria autora.

5.3 METODOLOGIA PROPOSTA

A identificação de locutor proposta neste trabalho utiliza o aprendizado de máquina para identificar pessoas com base nos atributos extraídos do sinal de voz. Essa metodologia é ilustrada na Figura 15.

Figura 15 – Identificação de locutor proposta.

Fonte: Elaboração da própria autora.

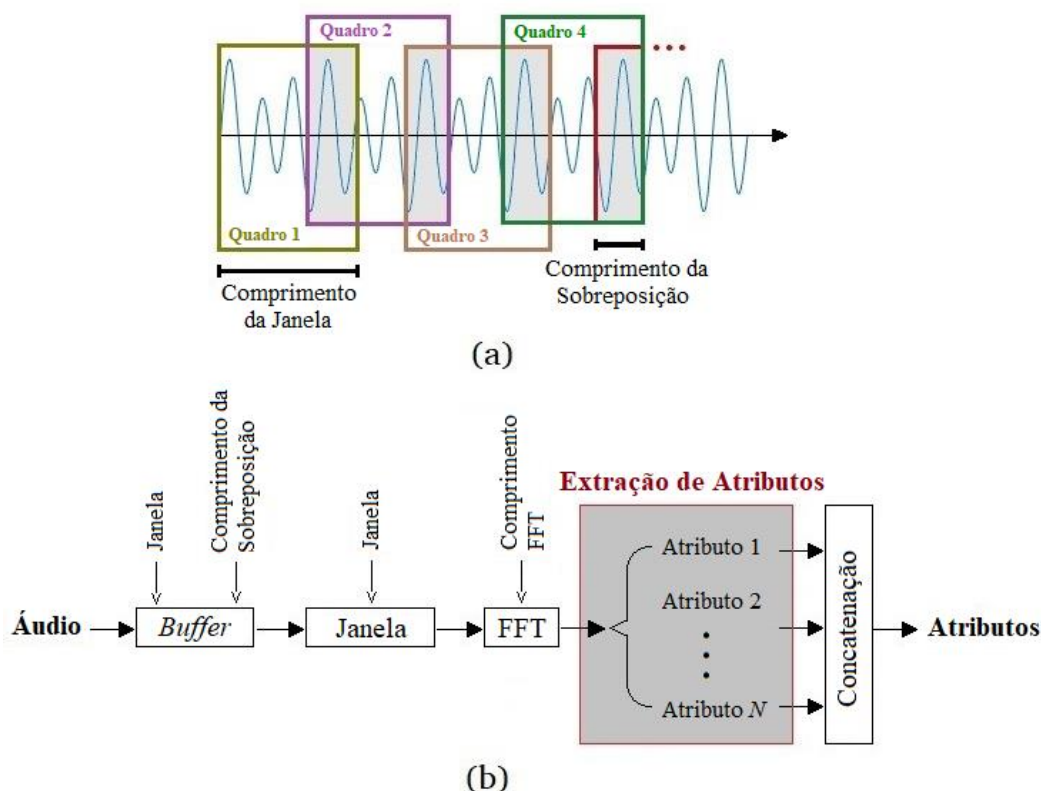
Na Figura 15, observa-se que a metodologia proposta para identificação de locutor envolve duas etapas básicas: treinamento e teste.

Na etapa de treinamento, são extraídos os atributos dos sinais de voz gravados de N locutores, sendo N finito. Esses atributos extraídos dos sinais são usados para treinar o classificador KNN.

Após o treinamento, vem a etapa de teste, onde novos sinais de voz de locutores desconhecidos que precisam ser classificados passarão pelo mesmo processo de extração de atributos. Por fim, com o classificador treinado, o modelo indicará qual dos N locutores é o mais próximo, realizando assim, a identificação do locutor.

Neste trabalho, todos os atributos de áudio, incluindo a MMWL e os MFCC, foram extraídos de um sinal de voz segmentado de acordo com um processo de sobreposição de janelas (*windowing-overlap*), conforme mostrado na Figura 16.

Figura 16 – Processamento do sinal: (a) Processo de sobreposição de janelas e (b) Processo completo de extração dos atributos.



Fonte: Elaboração da própria autora.

Os termos usados na Figura 16 são definidos como:

- Comprimento da sobreposição (*overlap length*) - Número de amostras sobrepostas entre janelas adjacentes;

- Janela (*window*) - Janela aplicada no domínio do tempo;
- Comprimento da janela (*window length*) - comprimento do quadro (*frame*);
- Comprimento FFT (*FFT length*) - Número de *bins* usados na FFT.

A etapa de extração de atributos é representada pela região destacada pelo bloco na cor cinza da Figura 16(b). Esses atributos podem ser apenas a MMWL Espectral ou um conjunto de atributos. O conjunto de atributos, também chamado combinação (*match features*), fusão ou concatenação de atributos, ocorre ao usar dois ou mais tipos de atributos extraídos para serem concatenados, como por exemplo, MFCC + MMWL Espectral.

Para uma extração mais adequada dos atributos, foi realizado um pré-processamento de sinais de voz que incluiu: eliminação do ruído, seleção da parte vozeada do sinal e eliminação da parte não vozeada.

5.4 RESULTADOS E DISCUSSÕES

Após a escolha do modelo de classificação, a metodologia proposta para a identificação de locutor foi aplicada usando treinamento supervisionado nas seguintes especificações:

- Bancos de dados: TIMIT e AN4;
- Função wavelet mãe usada na MMWL: Bior 1.5 (padrão do MATLAB);
- Número de vizinhos do classificador KNN do tipo *fine*: 5;
- Métrica de distância utilizada na classificação: Euclidiana ponderada ao quadrado e inversa;
- Método de validação do classificador: validação cruzada com predição por perda de classificação;
- Divisão do armazenamento dos dados: 80% selecionados aleatoriamente para compor o conjunto de dados para treinamento e 20% restantes alocados como o conjunto para teste.

Vale lembrar que, para identificação utilizando essa abordagem de aprendizado de máquina, inicialmente deve ser feita a identificação das classes, ou seja, para cada conjunto de dados, as instâncias (sinais de voz) devem ser rotuladas com os seus

respectivos nomes de locutor. No caso dos conjuntos das bases TIMIT e AN4, todos os arquivos já são distribuídos rotulados. Também é importante fazer a normalização dos valores após a extração dos atributos.

Por fim, com estas definições feitas, as características foram extraídas em quadros de amostras de 32 milissegundos com uma sobreposição de 94% dos sinais de voz gravados dos N locutores presentes no banco de dados e utilizadas para treinar o classificador. Após isso, os novos sinais de voz de locutores desconhecidos (oriundos dos 20% que foram separados para teste) passam pela mesma extração dos atributos da etapa de treinamento e por fim o classificador treinado prevê qual dos N locutores é o mais parecido com a do locutor desconhecido.

Para encontrar as características de áudio mais apropriadas no modelo de identificação foram extraídos e comparados 16 atributos de áudio usando em cada um o mesmo procedimento de extração descrito na metodologia proposta. Os atributos, em inglês, foram: *Spectral Rolloff Point*, *Pitch*, *Spectral Spread*, *Spectral Crest*, *Spectral Flatness*, *Spectral Kurtosis*, *Spectral Skewness*, *Spectral Centroid*, *Spectral Flux*, *Spectral Decrease*, *Spectral Entropy*, *Spectral Slope*, *Harmonic Ratio*, *LPC*, *MFCC* e *Spectral MMWL*.

O banco de dados usado no primeiro estudo foi o *Census*, também conhecido como banco de dados AN4 do *CMU Robust Speech Recognition Group* de Acero (1992). Esse conjunto de dados, descrito na Tabela 8, contém gravações de cinco pessoas masculinas e cinco femininas no formato *flac*, na língua inglesa americana e pronunciando frases diferentes em todos os discursos.

Tabela 8 – Distribuição dos arquivos para o banco AN4.

Locutores (rótulos)	Gênero	Quantidade de arquivos (treino)	Quantidade de arquivos (teste)	Total de arquivos (banco)
fejs	Feminino	10	3	13
fmjd	Feminino	10	3	13
fsrb	Feminino	10	3	13
ftmj	Feminino	10	3	13
fwxs	Feminino	10	2	12
mcen	Masculino	10	3	13
mrcb	Masculino	10	3	13
msjm	Masculino	10	3	13
msjr	Masculino	10	3	13
msmn	Masculino	7	2	9

Fonte: Elaboração da própria autora.

A Tabela 8 mostra a quantidade de arquivos em cada conjunto de dados de acordo com a porcentagem de divisão definida nas especificações da metodologia proposta. Os rótulos identificando os locutores na Tabela 8 serão as classes da matriz de confusão.

Dadas as definições da metodologia e as descrições dos sinais, os resultados são obtidos. A Tabela 9 ilustra a dimensão do vetor de atributos, tempo de processamento e o percentual de acurácia dos modelos treinados para cada atributo ou conjunto de atributos concatenados, cujos resultados foram obtidos testando um de cada vez na metodologia da Figura 15 para o banco de dados AN4.

Tabela 9 – Acurácia dos atributos para o banco AN4.

Atributos	Dimensão do vetor	Tempo de processamento (em segundos)	Acurácia
<i>Spectral Rolloff Point</i>	1	~009,115	09,46%
<i>Pitch</i>	1	~013,317	11,59%
<i>Spectral Spread</i>	1	~008,711	12,54%
<i>Spectral Crest</i>	1	~008,797	12,69%
<i>Spectral Flatness</i>	1	~008,746	13,01%
<i>Spectral Kurtosis</i>	1	~008,711	13,50%
<i>Spectral Skewness</i>	1	~009,713	13,70%
<i>Spectral Centroid</i>	1	~007,690	14,40%
<i>Spectral Flux</i>	1	~008,653	14,54%
<i>Spectral Decrease</i>	1	~008,186	14,61%
<i>Spectral Entropy</i>	1	~008,751	14,90%
<i>Spectral Slope</i>	1	~008,544	15,27%
<i>Spectral MMWL</i>	1	~010,836	15,59%
<i>Harmonic Ratio</i>	1	~013,189	23,08%
<i>LPC</i>	249	~308,755	98,14%
Todos os 16 Atributos	276	~566,449	99,53%
<i>MFCC + LPC + Pitch</i>	263	~333,980	99,60%
<i>MFCC + Pitch</i>	14	~028,974	99,61%
<i>MFCC + LPC</i>	262	~317,361	99,65%
<i>MFCC</i>	13	~021,572	99,72%
<i>MFCC + Spectral MMWL</i>	14	~021,995	99,81%

Fonte: Elaboração da própria autora.

Na Tabela 9, a combinação MMWL Espectral + MFCC obteve o maior percentual de acurácia, aumentando a assertividade na previsão e demonstrando ser uma excelente combinação de características destinadas para a identificação de locutor. Estes

resultados confirmam os estudos realizados por Tirumala *et al.* (2017), cujas abordagens baseadas na extração de fusões do MFCC com outras características proporcionaram uma melhor acurácia de classificação.

Também na Tabela 9 é possível observar que a inserção de uma quantidade excessiva de atributos não contribui significativamente na predição e pode aumentar consideravelmente o tempo de processamento na execução dos algoritmos de extração.

Desta forma, estes resultados confirmam o fato de que a eficiência dos sistemas de reconhecimento de locutores depende de características representativas dos sinais e que, a partir de uma combinação de características, também é possível fornecer atributos eficientes, gerando sistemas precisos e robustos. Vale lembrar que a combinação de características deve corresponder ao modelo, já que o classificador buscará padrões em características concatenadas e normalizadas.

Além disso, vale ressaltar que embora a MMWL Espectral (*Spectral MMWL*) não possa competir diretamente com o MFCC, a inserção de características que diminuam o erro dos modelos na identificação é sempre necessária.

Na Tabela 9, é observado também que a acurácia do modelo usando MMWL Espectral + MFCC é superior à do modelo usando Pitch + MFCC, que são os atributos mais utilizados na literatura especializada para identificação de locutores atualmente, por exemplo: Ezzaidi e Rouat (2004), Hanifa, Isa e Mohamad (2020), Shao, Milner e Cox (2003) e The MathWorks, Inc. (2019).

Os resultados e contribuições dos atributos de áudio propostos neste trabalho podem ser analisados em detalhes comparando as matrizes de confusão das características concatenadas clássicas da literatura, Pitch + MFCC, com a combinação de atributos mais vantajosa obtida neste trabalho, MMWL Espectral + MFCC, ilustradas nas Figuras 17 e 18, respectivamente. É importante ressaltar que os resultados apresentados nas Figuras 17 e 18 consideram uma previsão por quadros de cada arquivo, ou seja, cada quadro é considerado como um sinal a ser classificado pelo modelo.

No conjunto de teste, a moda das previsões, ou seja, as previsões mais frequentes obtidas para cada arquivo, levaram aos mesmos resultados na matriz de confusão tanto para os atributos de Pitch + MFCC quanto para os de MMWL Espectral + MFCC, conforme apresentado na Figura 19.

**Figura 17 – Matriz de confusão usando a MMWL Espectral + MFCC (banco AN4).
Acurácia de Validação - MMWL Espectral + MFCC (Por Quadro)**

4429	2		2	5					1	99.8%	0.2%	fejs
2	7764		5	2			4			99.8%	0.2%	fmjd
1	6	7018	2	1			1			99.8%	0.2%	fsrb
5	6	2	6731	11	3		1			99.6%	0.4%	ftmj
3	14		2	7098	7	1	7		1	99.5%	0.5%	fwxs
	1		1		4724	3		1		99.9%	0.1%	mcen
		1		2	1	4781		3	1	99.8%	0.2%	mrcb
		1		5	5		4861		5	99.7%	0.3%	msjm
				1	4	1	1	2692		99.8%	0.2%	msjr
	1			4		1	6		5203	99.8%	0.2%	msmn

99.8%	99.6%	99.9%	99.8%	99.6%	99.6%	99.8%	99.6%	99.9%	99.8%
0.2%	0.4%	0.1%	0.2%	0.4%	0.4%	0.2%	0.4%	0.1%	0.2%
fejs	fmjd	fsrb	ftmj	fwxs	mcen	mrcb	msjm	msjr	msmn

Classe Prevista

Fonte: Elaboração da própria autora.

**Figura 18 – Matriz de confusão usando Pitch + MFCC (banco AN4).
Acurácia de Validação - Pitch + MFCC (Por Quadro)**

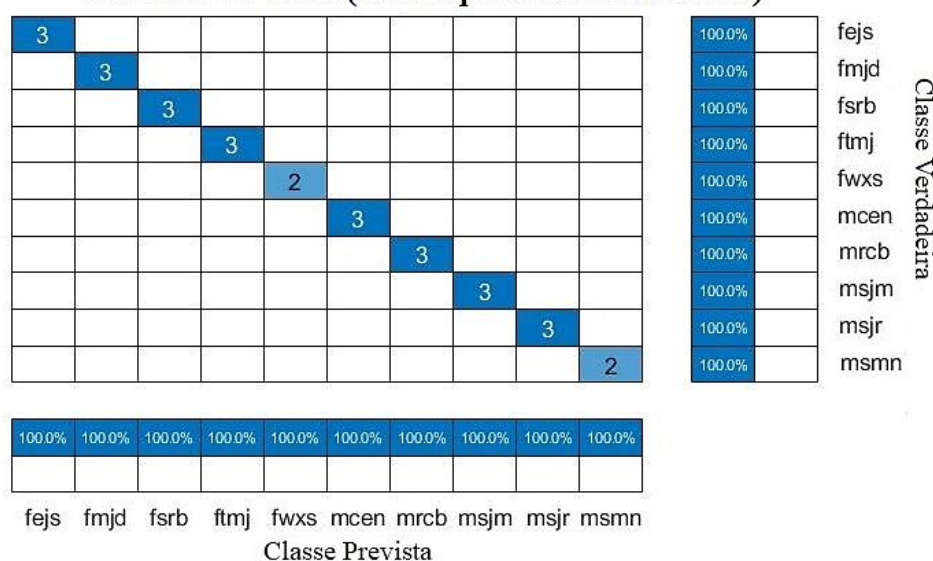
4418	2	1	10	5			1		2	99.5%	0.5%	fejs
2	7762	1	7	3			1		1	99.8%	0.2%	fmjd
4	9	7012	1	2				1		99.8%	0.2%	fsrb
7	14	4	6724	7		2	1			99.5%	0.5%	ftmj
5	15		2	7105	1		2		3	99.6%	0.4%	fwxs
					4718	5	4	1	2	99.7%	0.3%	mcen
1		2		3	5	4771	3	4		99.6%	0.4%	mrcb
1	2	3	1	4	7	1	4858			99.6%	0.4%	msjm
		4			1	6	1	2686		99.6%	0.4%	msjr
	1	2		3	1	1	4		5203	99.8%	0.2%	msmn

99.5%	99.4%	99.8%	99.7%	99.6%	99.7%	99.7%	99.7%	99.8%	99.8%
0.5%	0.6%	0.2%	0.3%	0.4%	0.3%	0.3%	0.3%	0.2%	0.2%
fejs	fmjd	fsrb	ftmj	fwxs	mcen	mrcb	msjm	msjr	msmn

Classe Prevista

Fonte: Elaboração da própria autora.

Figura 19 – Matriz de confusão obtida por arquivo (banco AN4).
Acurácia do Teste (Por Arquivo Usando a Moda)



Fonte: Elaboração da própria autora.

Na Figura 19, os locutores previstos correspondem aos locutores esperados para todos os arquivos do teste. Embora o resultado final seja o mesmo para ambos os atributos de áudio, é essencial usar o modelo treinado com o maior percentual de acurácia de validação para garantir que o sistema preditivo seja o mais eficiente possível.

O desempenho preditivo dos atributos da MMWL Espectral + MFCC e dos atributos Pitch + MFCC também foram confirmados em um segundo estudo feito para as amostras de voz do banco de dados *DARPA TIMIT Acoustic - Phonetic Continuous Speech Corpus* de Garofolo *et al.* (1993). Esta base de dados contém 6300 arquivos de áudio gravados por 630 pessoas em dialetos de 8 grandes regiões dos Estados Unidos, com 10 frases pronunciadas em inglês americano por cada locutor. Os arquivos estão disponíveis em formato *wav*, com duração de aproximadamente 4 segundos, amostrados a 16 kHz com uma boa relação sinal-ruído ($SNR > 20$ dB). A base de dados TIMIT tem sido amplamente utilizada em sistemas de reconhecimento automático de voz.

Os testes e os treinamentos dos modelos foram feitos com o conjunto completo de testes do banco TIMIT. Este conjunto de dados contém 168 locutores e 1344 expressões, em 624 textos diferentes, representando cerca de 27% do material total de voz. Na Tabela 10 são discriminados os números de locutores e gênero para os dialetos de 8 regiões.

Tabela 10 – Distribuição do conjunto de teste completo (banco TIMIT).

Dialeto da Região (DR)	Masculino	Feminino	Total
1	7	4	11
2	18	8	26
3	23	3	26
4	16	16	32
5	17	11	28
6	8	3	11
7	15	8	23
8	8	3	11

Fonte: Elaboração da própria autora.

Os dialetos das 8 regiões da Tabela 10 são:

- DR 1: Nova Inglaterra;
- DR 2: Norte;
- DR 3: Norte de Midland;
- DR 4: Sul de Midland;
- DR 5: Sul;
- DR 6: Cidade de Nova Iorque;
- DR 7: Ocidente;
- DR 8: *Army Brat* (que se move de uma região para outra).

Os percentuais de acurácia dos modelos utilizando MFCC + MMWL Espectral e do MFCC + Pitch nos quadros dos sinais de voz dessas oito regiões são mostrados na Tabela 11.

Tabela 11 – Acurácia de validação para conjunto de teste completo (banco TIMIT).

Dialeto da Região (DR)	MMWL Espectral + MFCC	Pitch + MFCC
1	99,89%	99,79%
2	99,78%	99,52%
3	99,75%	99,47%
4	99,75%	99,59%
5	99,82%	99,65%
6	99,86%	99,69%
7	99,85%	99,68%
8	99,87%	99,70%

Fonte: Elaboração da própria autora.

Na Tabela 11, é possível observar que a metodologia utilizando a concatenação do atributo de áudio proposto com o MFCC apresentou uma acurácia de validação superior aos atributos de áudio da literatura. Também foi observado que no modelo treinado, e destinado apenas a um dialeto entre as oito regiões, a acurácia média foi superior ao conjunto com todas as regiões.

Além disso, os testes realizados com modelos treinados apenas para grupos masculinos ou femininos separadamente obtiveram uma acurácia média de 99,95% para o grupo masculino e 99,98% para o grupo feminino, cujo percentual é maior do que para o modelo treinado com ambos os grupos. Esses resultados indicam que há uma pequena influência no desempenho ao fazer previsões com modelos gerados para grupos específicos.

Quanto à quantidade de arquivos de áudio usados para extração, foi observado que houve uma diminuição no percentual de validação inferior a 1% ao reduzir a quantidade de áudios disponíveis de cada locutor ou reduzir o tamanho do conjunto de dados para treinamento. Supondo que o objetivo do sistema seja usar pouco espaço de armazenamento e obter baixo custo computacional para treinar os modelos, nesse caso, o uso de poucos sinais na geração dos modelos pode ser uma alternativa que trará resultados sem perdas significativas na eficiência do sistema.

Neste trabalho também foi feito um estudo para determinar a função wavelet mais apropriada para extrair a MMWL Espectral em sinais de voz. Os resultados obtidos neste teste apresentaram uma percentagem variando em média 0,02% entre as famílias wavelet de Daubechies (db), Symlet (sym), Coiflet (coif), e Biorthogonal (bior), indicando assim, que o desempenho dos modelos não sofre influência significativa com a escolha da função wavelet mãe.

5.5 EXPERIMENTOS ADICIONAIS

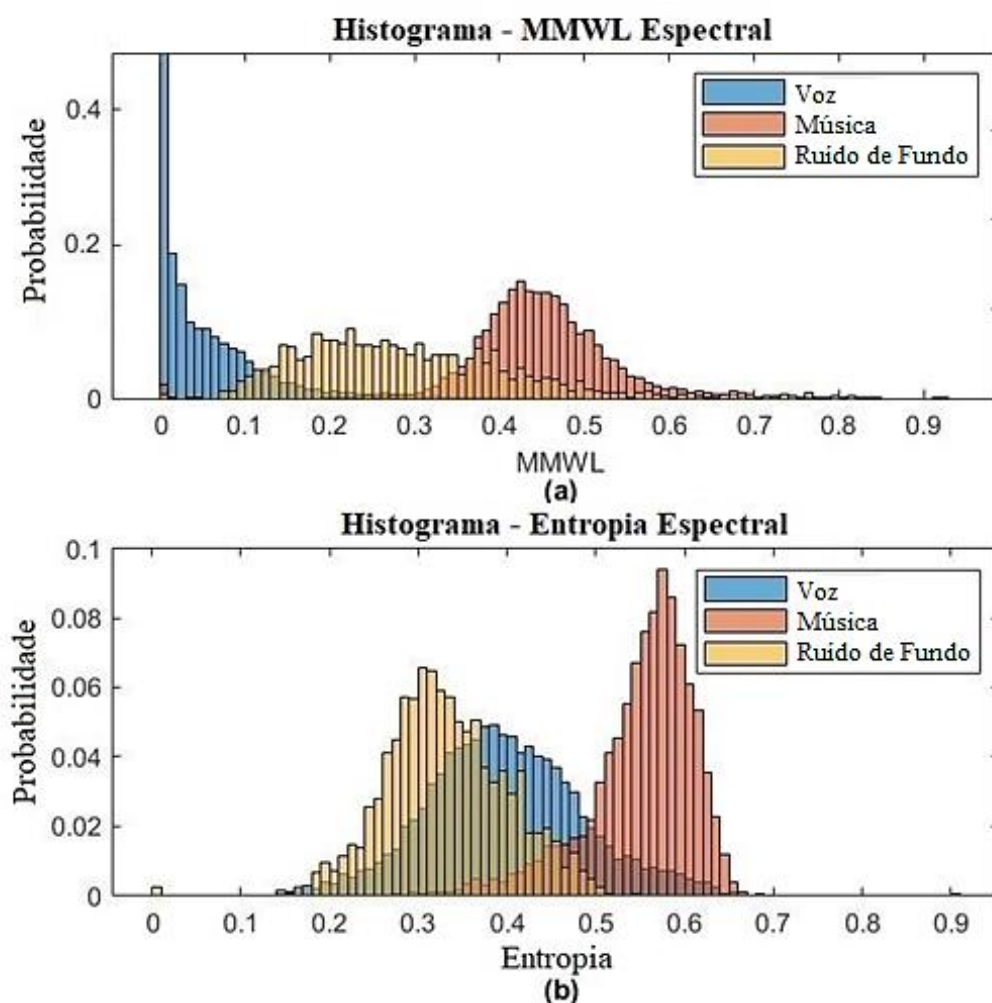
Os experimentos ilustrados nesta seção visam apresentar o comportamento da MMWL Espectral em comparação com os atributos clássicos de áudio que são frequentemente utilizados em diversas aplicações. Dessa forma, os experimentos adicionais deste trabalho permitem apresentar os resultados iniciais usando o novo atributo de áudio e indicar as possibilidades de adotar a MMWL em pesquisas futuras para obter novas contribuições na área de processamento de sinais de áudio.

5.5.1 Entropia Espectral versus MMWL Espectral

A Entropia Espectral (*Spectral Entropy*) de Misra *et al.* (2004) mede o pico do espectro do sinal e tem sido utilizada com sucesso para detectar segmentos vozeados e não vozeados (*voiced/unvoiced decision*) em aplicações de reconhecimento automático de voz. Essa característica também tem apresentado bons resultados quanto à discriminação entre voz e música (PIKRAKIS; GIANNAKOPOULOS; THEODORIDIS, 2008).

Para testar e comparar o potencial em distinguir sinais de voz e música da MMWL Espectral com a Entropia Espectral, são apresentados na Figura 20 os histogramas da MMWL e da Entropia computados para os arquivos de voz, de música e de ruído de fundo.

Figura 20 – Histogramas para diferentes arquivos de áudio usando: (a) MMWL espectral e (b) Entropia espectral.

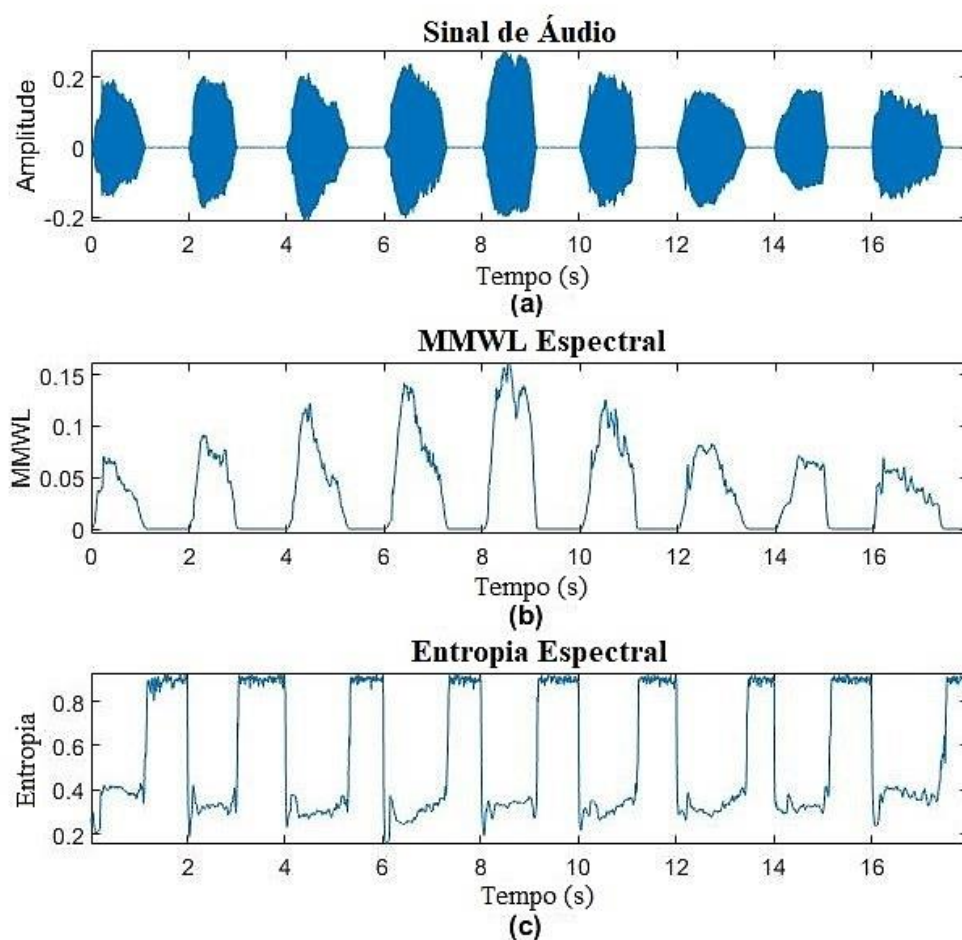


Fonte: Elaboração da própria autora.

Na Figura 20, é observado que a MMWL Espectral separa com mais precisão as três classes envolvidas. Além disso, o histograma da MMWL apresenta valores de probabilidade muito maiores do que a da Entropia Espectral, especialmente para o sinal de voz.

Apenas para ilustração, na Figura 21 é apresentado um sinal de voz e as características MMWL Espectral e Entropia Espectral. O sinal de voz usado é de um locutor do gênero feminino cantando em tom Lá Maior (*A Major*), amostrado à 16 kHz, contendo na frase um solfejo em escala musical de Lá Maior (*A Major Scale*) para diferentes notas na tonalidade de Lá Maior (*A's tonality*).

Figura 21 – Características extraídas do: (a) sinal de voz, usando (b) MMWL Espectral e (c) Entropia Espectral.



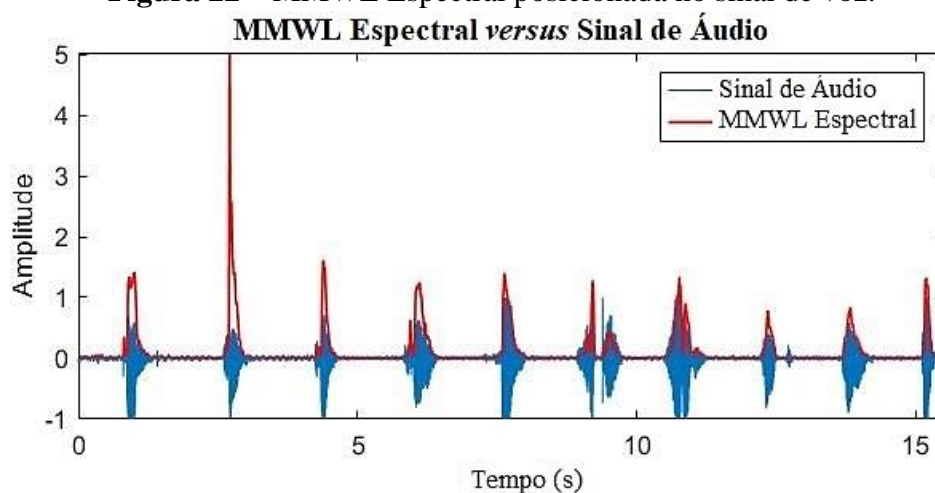
Fonte: Elaboração da própria autora.

Comparando as características de áudio da Figura 21 é possível notar que estes descritores de áudio são úteis para determinar regiões do sinal vozeados e não vozeados.

Também é possível observar que a MMWL Espectral contorna as regiões da fala vozeada com mais detalhes do que a Entropia Espectral.

Além disso, ao estabelecer limiares ajustados para as necessidades específicas de cada aplicação, a MMWL Espectral pode ser promissora na criação de algoritmos para detectar os limites/regiões de fala ou na detecção da presença/atividade de voz nos sinais de áudio. Para mostrar esta possibilidade, na Figura 22 é ilustrado o sinal de voz posicionado em cima da MMWL Espectral aplicada no mesmo. O sinal de voz usado é o sinal chamado *counting*, de um locutor masculino, amostrado a 8 kHz, contendo uma frase gravada da pessoa contando de um até dez na língua inglesa, da seguinte forma: “One two three four five six seven eight nine ten.”

Figura 22 – MMWL Espectral posicionada no sinal de voz.



Fonte: Elaboração da própria autora.

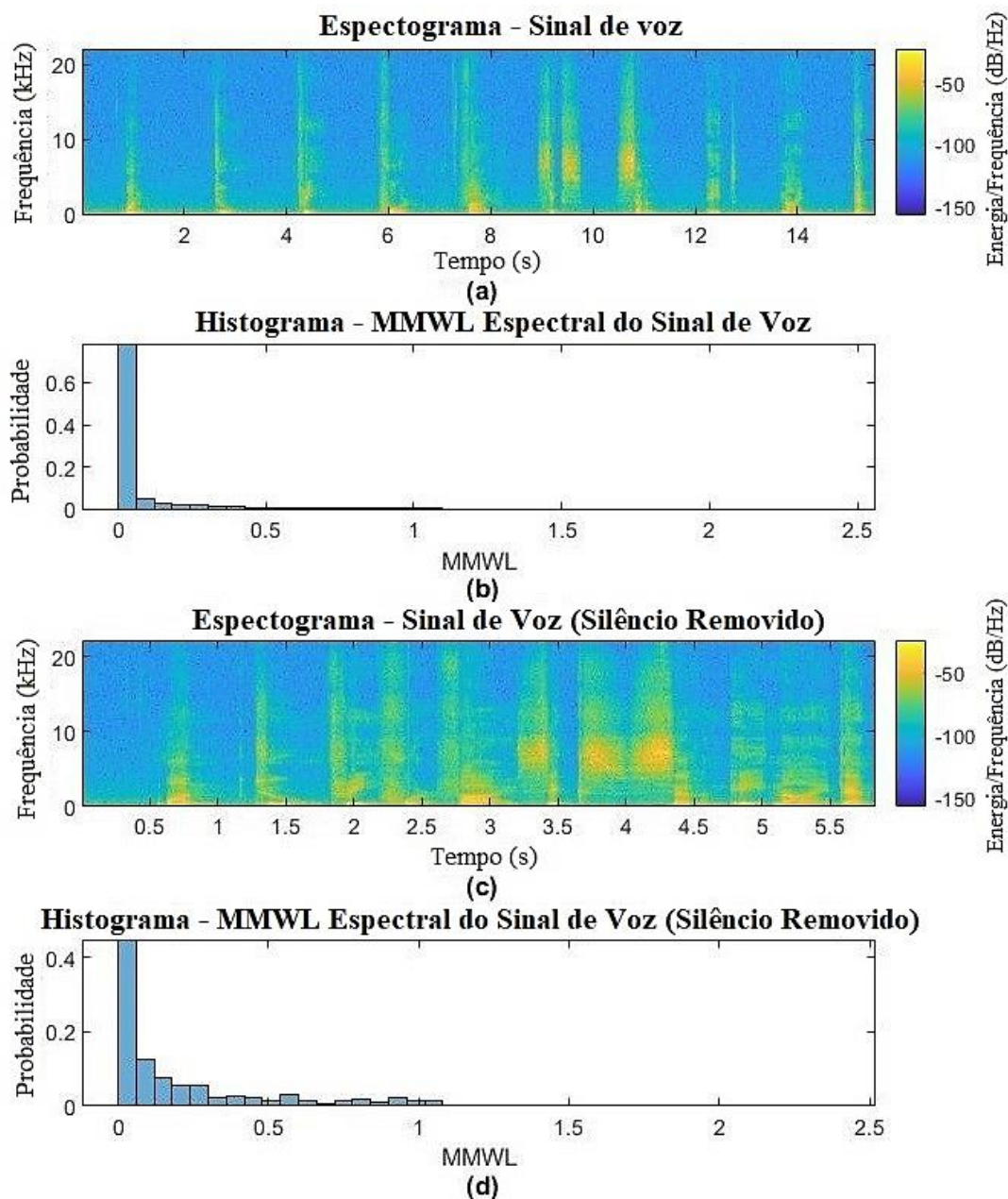
Vale lembrar que o procedimento de detecção de limites de fala é amplamente utilizado para aumentar a eficiência computacional de algoritmos como os de aprendizagem profunda (*deep learning*), principalmente se aplicado a grandes conjuntos de dados em tempo real.

Um exemplo usando limites de fala é a remoção das regiões silenciosas do sinal de voz. A remoção de silêncio permite obter estatísticas das características mais representativas da voz ao invés do canal do sinal.

Para checar isso, na Figura 23 são apresentados os espectrogramas e os histogramas da MMWL Espectral para o sinal *counting* original e sua versão sem silêncio, respectivamente. O algoritmo usado para remover as regiões de silêncio foi baseado na energia e dispersão espectral de Giannakopoulos (2009). Na Figura 23 é

observado que após isolar as regiões com voz no sinal de áudio, as distribuições da MMWL bem como as regiões vozeadas aparecem enfatizadas.

Figura 23 – Espectrogramas e Histogramas da MMWL Espectral, respectivamente para: (a), (b) sinal de voz original e (c), (d) sinal de voz com remoção do silêncio.



Fonte: Elaboração da própria autora.

5.5.2 Pitch versus MMWL Espectral

O Pitch é uma das características clássicas de sinais de voz, utilizada principalmente em aplicações de reconhecimento automático de locutor (ATAL, 1972).

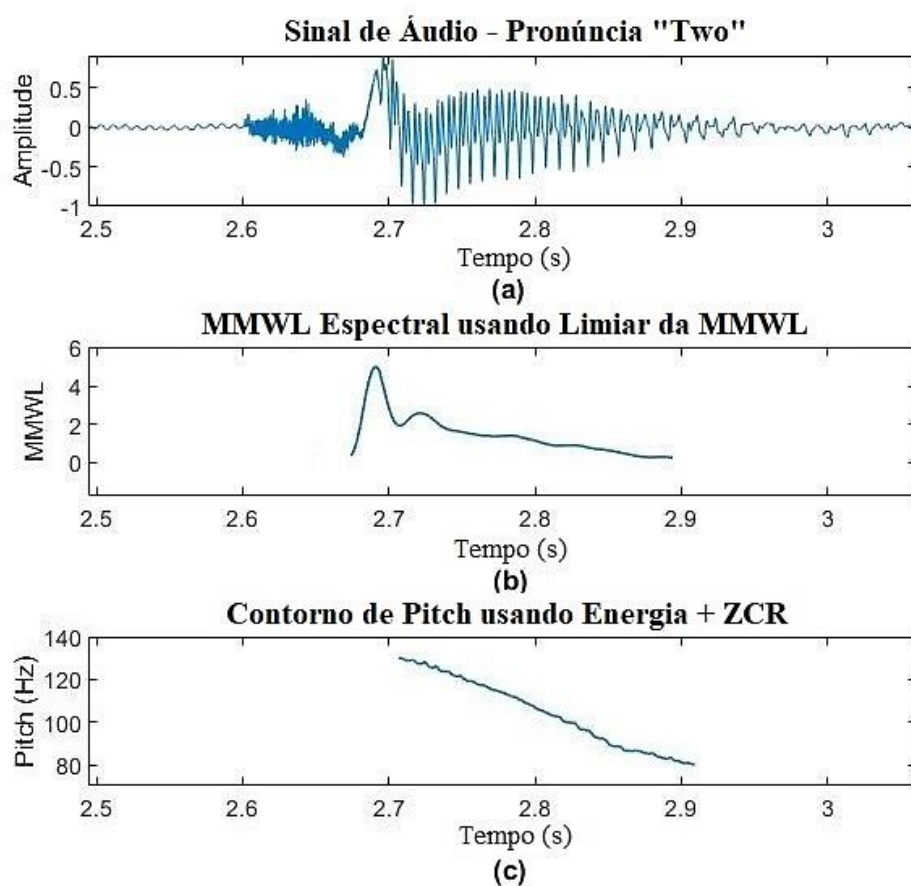
O método utilizado neste trabalho para estimar o Pitch foi baseado na Função de Autocorrelação (*Autocorrelation Function - ACF*) (HUI; DAI; WEI, 2006; RAKESH, DUTTA; SHAMA, 2011).

O Pitch e a MMWL Espectral foram comparados e usados para detectar os segmentos vozeados/não-vozeados do sinal de voz *counting* correspondente à pronúncia “Two”, Figura 24.

Para seleccionar os quadros vozeados através do Pitch, foi utilizado o método proposto por Giannakopoulos (2009) combinado com a Taxa de Cruzamentos por Zero (*Zero-Crossing Rate - ZCR*) de Bachu *et al.* (2010) para distinguir entre silêncio e fala.

Por outro lado, para determinar os quadros que contêm segmentos vozeados usando a MMWL Espectral, foi definido apenas um limiar de 0,25. A escolha do limiar foi simples e empírica, mas o resultado mostra o potencial da MMWL em separar segmentos vozeados ou não. Um estudo da forma mais apropriada para determinar o limiar deve ser realizado de acordo com a aplicação.

Figura 24 – Detecção da parte vozeada do: (a) sinal de voz “Two”, usando (b) MMWL Espectral com Limiar e (c) Contorno de Pitch com Energia e ZCR.



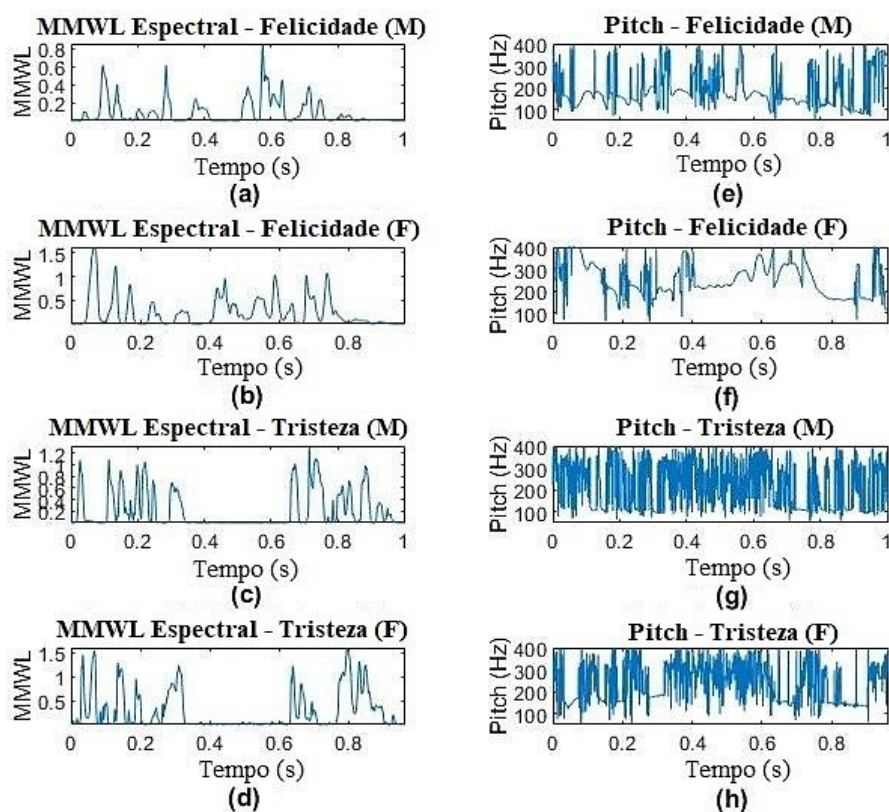
Fonte: Elaboração da própria autora.

Na Figura 24 é observado que as técnicas alcançaram resultados satisfatórios para o caso dos sons sonoros(vozeados). Também é possível observar que a MMWL Espectral oferece vantagens para distinguir entre partes vozeadas e não vozeadas com o uso de limiares diretamente na extração do atributo de áudio, evitando assim o uso de métodos adicionais para selecionar os quadros com voz. Portanto, a MMWL também tem potencial para ser usada como um recurso para identificar segmentos vozeados/não vozeados.

Em Gupta, Fahad e Deepak (2020), o Pitch também foi utilizado para Reconhecimento de Emoção por Voz (*Speech Emotion Recognition - SER*). Atualmente, o maior desafio em SER é a extração de características (MUSTAQEEM; SAJJAD; KWON, 2020).

A fim de identificar a emoção, o Pitch e a MMWL Espectral foram investigados e comparados para sinais com emoções de felicidade e tristeza de um locutor masculino (M) e de um feminino (F) do *Database of Emotional Speech* de Burkhardt *et al.* (2005), Figura 25.

Figura 25 – Características para emoções de felicidade e tristeza de locutores femininos (F) e masculinos (M) obtidas com MMWL Espectral (a, b, c, d) e Pitch (e, f, g, h).



Fonte: Elaboração da própria autora.

As frases para estas duas emoções estão na língua alemã, onde a sentença gravada foi “Die wird auf dem Platz sein, wo wir sie immer hinlegen”, cuja tradução é “Estará no lugar onde sempre o colocamos”.

Na Figura 25, podem ser observadas semelhanças entre os espectros da MMWL para a mesma emoção melhor do que o Pitch. Por exemplo, na emoção de tristeza, é possível notar que a região central apresenta valores menores da MMWL tanto para os espectros masculinos como para os femininos.

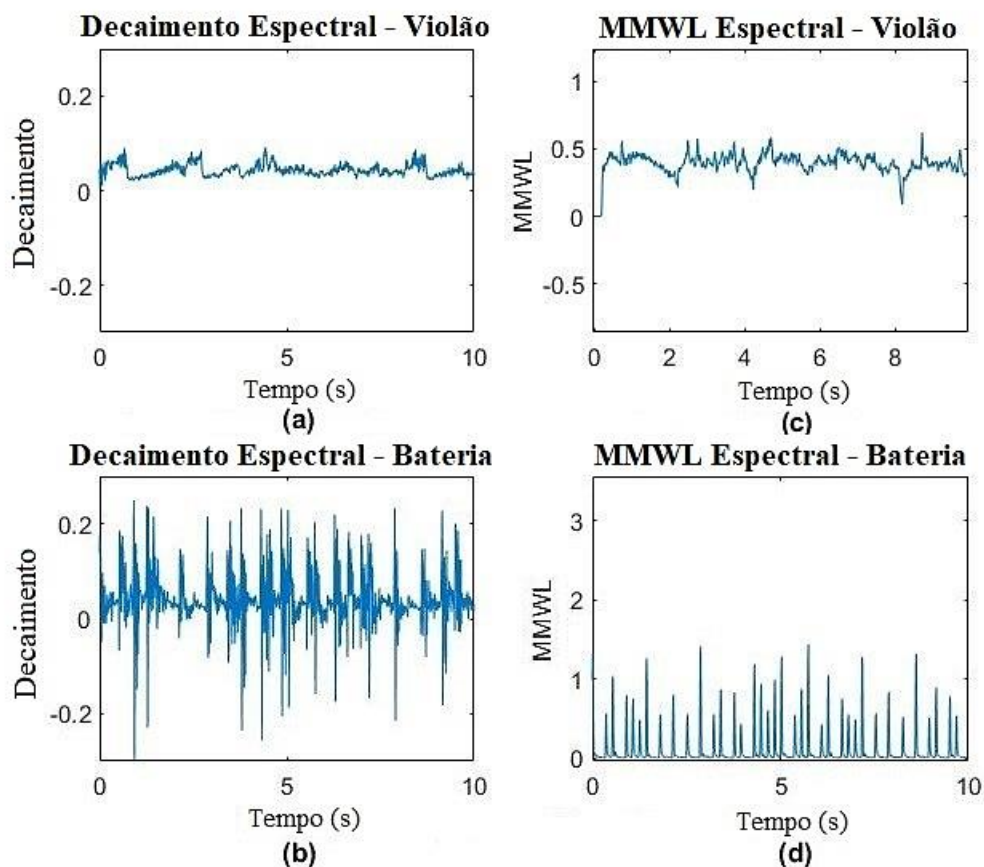
Estudos para obtenção de um parâmetro baseado na MMWL, tais como, um limiar da MMWL ou seleção de banda do sinal, são sugeridos para os sistemas de reconhecimento de emoção.

5.5.3 Decaimento Espectral *versus* MMWL Espectral

O Decaimento Espectral (*Spectral Decrease*) de Peeters (2004) representa a quantidade de diminuição do espectro, permitindo enfatizar as inclinações das frequências mais baixas, e tem sido utilizado para reconhecimento de instrumentos (*instrument recognition*) (ESSID; RICHARD; DAVID, 2006).

Na Figura 26, a MMWL Espectral é comparada com o Decaimento Espectral para os instrumentos de violão elétrico e de bateria. É observado nos espectros na Figura 26 que ambas as características de áudio estão de acordo com o objetivo de distinguir os dois instrumentos. Um estudo usando a MMWL Espectral também é sugerido em outras aplicações, tais como, o reconhecimento acústico de cena (*acoustic scene recognition*) e a classificação do gênero musical (*music genre classification*).

Figura 26 – Decaimento Espectral e MMWL Espectral, respectivamente, para: (a), (c) Violão e (b), (d) Bateria.



Fonte: Elaboração da própria autora.

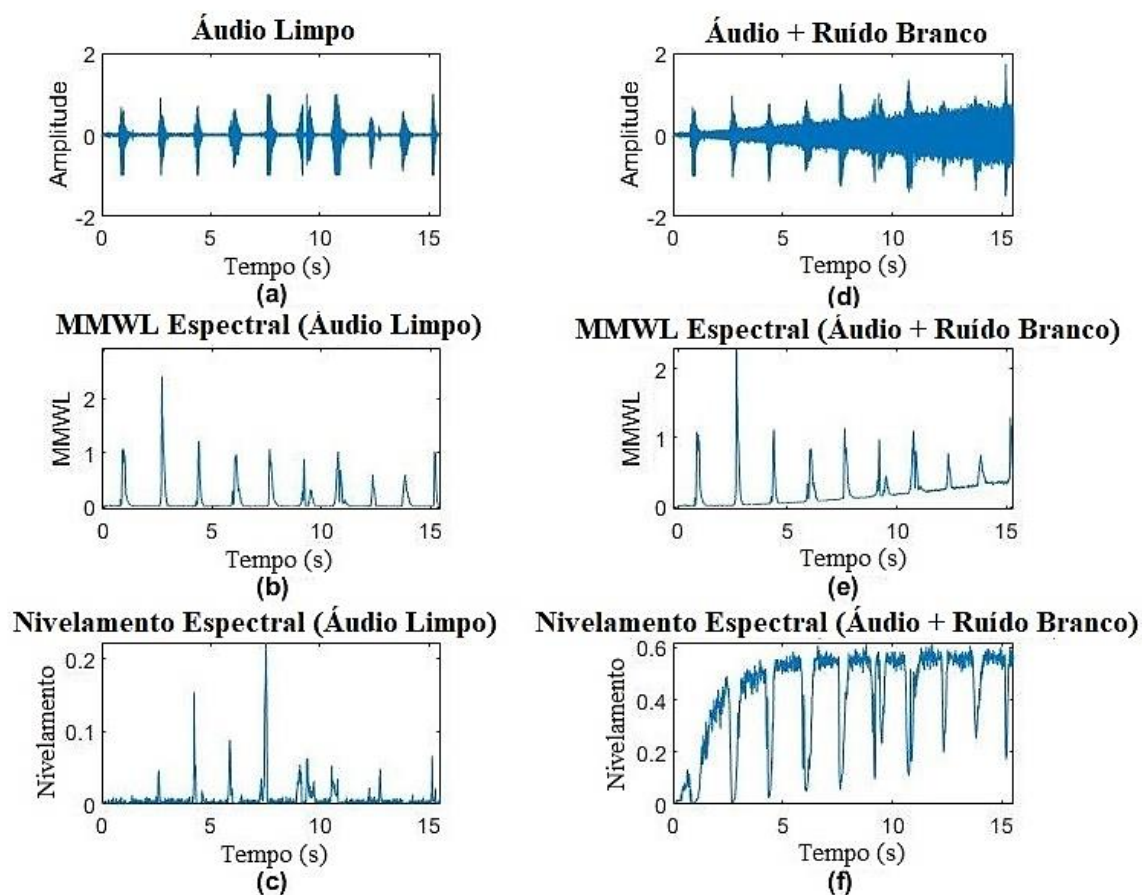
5.5.4 Nivelamento Espectral versus MMWL Espectral

O Nivelamento Espectral (*Spectral Flatness*) de Johnston (1988) mede a razão entre a média geométrica e a média aritmética do espectro. O Nivelamento Espectral pode ser usado para indicar o ruído e a tonalidade de um sinal de voz e até de outros tipos de áudios. Resultados satisfatórios usando o Nivelamento Espectral são encontrados na detecção de voz cantada (*singing voice detection*) em Lehner, Widmer e Sonnleitner (2014) e no reconhecimento de cenas por áudio (*audio scene recognition*) em Petetin, Laroche e Mayoue (2015).

Ao analisar a possibilidade de indicar ruído no sinal foi realizada uma comparação entre o Nivelamento Espectral e a MMWL Espectral, para o sinal *counting* limpo e corrompido por ruído branco, Figura 27. Conforme mostrado na Figura 27, à medida que o ruído aumenta no sinal de voz, o Nivelamento Espectral e a MMWL

Espectral também aumentam. Entretanto, é importante observar pelos espectros dos atributos de áudio que a MMWL Espectral indicou o ruído sem perder as características originais extraídas do sinal limpo.

Figura 27 – Comparação Espectral do: (a) áudio limpo, aplicando (b) MMWL Espectral e (c) Nivelamento Espectral; (d) áudio corrompido com ruído branco, aplicando (e) MMWL Espectral e (f) Nivelamento Espectral.



Fonte: Elaboração da própria autora.

Portanto, os resultados preliminares apresentados nesta seção indicam que a MMWL tem potencial para ser explorada em uma variedade de etapas e aplicações do processamento de sinais de voz, bem como para outros tipos de áudio.

6. CONCLUSÕES

A identificação de locutor possui uma variedade de aplicações. Muitos sistemas e técnicas utilizadas nessas aplicações ainda apresentam limitações que precisam ser resolvidas a fim de melhorar o seu uso. Vale ressaltar que quanto mais esses sistemas forem práticos, robustos e eficazes, melhor eles serão para a identificação. Por isso, é necessário estar em busca constante de sistemas mais eficientes.

Sendo assim, neste trabalho buscou-se aumentar a acurácia do modelo de identificação de locutor por meio da extração de um novo atributo de sinal de áudio, chamado MMWL Espectral, e do atributo clássico de MFCC, conseguindo dessa forma, um sistema preditivo eficiente. A concatenação MMWL Espectral + MFCC usada como atributo em modelos de aprendizado de máquina obteve a maior acurácia dentre os atributos da literatura estudados para os bancos de dados TIMIT e AN4, o que indica que a inserção dos WL contribuiu de forma significativa para uma melhoria na taxa de acerto do sistema.

Os resultados obtidos mostram que a MMWL Espectral pode ser considerada uma boa característica a ser explorada e inserida em sistemas de identificação. O atributo proposto também tem potencial para diversas aplicações envolvendo processamento de sinais. Outras pesquisas também podem ser feitas com a implementação da autenticação de locutor via dispositivos em tempo real.

Para trabalhos futuros, pode-se aplicar o sistema proposto em sinais de áudio coletados de pessoas com e sem o uso de máscaras faciais, a fim de estudar o desempenho do modelo em arquivos de áudio coletados em sons abafados.

REFERÊNCIAS

- ACERO, A. **Acoustical and environmental robustness in automatic speech recognition**. Kluwer Academic Publishers, USA, 1992. Disponível em: <http://www.speech.cs.cmu.edu/databases/an4/>. Acesso em: 14 dez. 2018.
- AIZAT, K.; MOHAMED, O.; ORKEN, M.; AINUR, A.; ZHUMAZHANOV, B. Identification and authentication of user's voice using DNN features and i-vector. **Cogent Engineering**, Abingdon, v. 7, n. 1, p. 1751557, 2020.
- ALCAIM, A.; OLIVEIRA, C. **Fundamentos do processamento de sinais de voz e imagem**. Rio de Janeiro: Interciência, 2012.
- ALMEIDA, R. O. **Predição de rotas metabólicas de enzimas utilizando aprendizado de máquina**. 2018. Tese (Doutorado) - Instituto de Biociências, Universidade Estadual Paulista, Botucatu, 2018. Disponível em: <https://repositorio.unesp.br/handle/11449/157299>. Acesso em: 04 jan.2021.
- ARNEODO, A.; AUDIT, B.; DECOSTER, N.; MUZY, J. F.; VAILLANT, C. Wavelet based multifractal formalism: applications to DNA sequences, satellite images of the cloud structure, and stock market data. *In*: BUNDE, A.; KROPP, J.; SCHELLNHUBER, H. J. (ed.). **The science of disasters: climate disruptions, heart attacks, and market crashes**. Berlin: Heidelberg: Springer Berlin Heidelberg, 2002. p. 26–102.
- ATAL, B. S. Automatic speaker recognition based on pitch contours. **The Journal of the Acoustical Society of America**, Melville, v. 52, n. 6B, p. 1687–1697, 1972.
- BABAEE, E.; ANUAR, N. B.; WAHAB, A. W. A.; SHAMSHIRBAND, S.; CHRONOPOULOS, A. T. An overview of audio event detection methods from feature extraction to classification. **Applied Artificial Intelligence**, New York, v. 31, n. 9/10, 661-714, 2017.
- BACHU, R. G.; KOPPARTHI, S.; ADAPA, B.; BARKANA, B. D. Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy. *In*: ELLEITHY, K. (ed.). **Advanced techniques in computing sciences and software engineering**. Dordrecht: Springer, 2010.
- BHAVSAR, H.; GANATRA, A. A comparative study of training algorithms for supervised machine learning. **International Journal of Soft Computing and Engineering (IJSCE)**, Madhya Pradesh, v. 2, n. 4, p. 2231–2307, 2012.
- BORGES, L. A. **Sistema de adaptação de locutor utilizando auto-vozes**. 2001. Dissertação (Mestrado) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2001. Disponível em: <https://www.teses.usp.br/teses/disponiveis/3/3142/tde-05052003-104044/pt-br.php>. Acesso em: 04 jan. 2021.

BURKHARDT, F.; PAESCHKE, A.; ROLFES, M.; SENDLMEIER, W.; WEISS, B. **A database of German emotional speech**. 2005. Disponível em: https://www.researchgate.net/publication/221491017_A_database_of_German_emotional_speech. Acesso em: 04 jan. 2021.

CAMPBELL, J. P. Speaker recognition: a tutorial. **Proceedings of the IEEE**, Piscataway, v. 85, n. 9, p. 1437–1462, 1997.

CARDOSO, D. P. **Identificação de locutor empregando modelos de mistura gaussianas**. 2009. Dissertação (Mestrado) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2009. Disponível em: <https://www.teses.usp.br/teses/disponiveis/3/3142/tde-13072009-155208/publico/monografia.pdf>. Acesso em: 04 jan. 2021.

CORDELLA, L. P.; FOGGIA, P.; SANSONE, C.; VENTO, M. A real-time text-independent speaker identification system. *In: PROCEDIMENTOS DA CONFERÊNCIA INTERNACIONAL SOBRE ANÁLISE E PROCESSAMENTO DE IMAGENS*, 12, 2003, Massachusetts. **Proceedings of the [...]**. Massachusetts: IEEE, 2003. p. 632.

DAUBECHIES, I. **Ten lectures on wavelets**. Disponível em: <https://jqichina.files.wordpress.com/2012/02/ten-lectures-of-waveletsefbc88e5b08fe6b3a2e58d81e8aeb2efbc891.pdf>. Acesso em: 04 jan. 2021.

DAVIS, S. B.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. **IEEE Transactions on Acoustics, Speech and Signal Processing**, Piscataway, v. 28, n.4, p. 357-366, 1980.

DELLER, J. R.; HANSEN, J. H. L.; PROAKIS, J. G. **Discrete-time processing of speech signals**. New York: Wiley-Interscience, 2000.

DHANALAKSHMI, P.; PALANIVEL, S.; RAMALINGAM, V. Pattern classification models for classifying and indexing audio signals. **Engineering Applications of Artificial Intelligence**, Oxford, v. 24, n. 2, p. 350–357, 2011.

DINIZ, S. S. **Uso de técnicas neurais para o reconhecimento de comandos à voz**. 1997. Dissertação (Mestrado) - Instituto Militar de Engenharia, Rio de Janeiro, 1997.

DUARTE, M. A. Q. **Redução de ruído em sinais de voz no domínio wavelet**. 2005. Tese (Doutorado) – Faculdade de Engenharia, Universidade Estadual Paulista, Ilha Solteira, 2005. Disponível em: <https://repositorio.unesp.br/handle/11449/100369>. Acesso em: 04 jan. 2021.

ESSID, S.; RICHARD, G.; DAVID, B. Instrument recognition in polyphonic music based on automatic taxonomies. **IEEE Transactions on Audio, Speech and Language Processing**, Piscataway, v. 14, n. 1, p. 68–80, 2006.

EZZAIDI, H.; ROUAT, J. Pitch and MFCC dependent GMM models for speaker identification systems. *In: CANADIAN CONFERENCE ON ELECTRICAL AND COMPUTER ENGINEERING, CANADIAN CONFERENCE, 2004, Niagara Falls. Proceedings of the [...].* Niagara Falls: IEEE, 2004. V. 1, p. 43-46.

FABRIS, F.; MAGALHÃES, J. P.; FREITAS, A. A. A review of supervised machine learning applied to ageing research. **Biogerontology**, Dordrecht, v. 18, p. 171-188, 2017.

FARIA, R. R. A. **Aplicação de Wavelets na análise de gestos musicais em timbres de instrumentos acústicos tradicionais**. 1997. Dissertação (Mestrado) - Escola Politécnica, Universidade de São Paulo, São Paulo, 1997. Disponível em: <https://www.teses.usp.br/teses/disponiveis/3/3142/tde-18072013-104904/pt-br.php>. Acesso em: 04 jan. 2021.

FUKUYAMA, E. E. Análise acústica da voz captada na faringe próximo à fonte glótica através de microfone acoplado ao fibrolaringoscópio. **Revista Brasileira de Otorrinolaringologia**, São Paulo, v. 67, n. 6, p. 776–786, 2001.

FURUI, S. **Digital speech processing, synthesis and recognition**. 2. ed. New York: Springer, 2001. Signal processing and communications series.

GADHOUMI, K.; DO, D.; BADILINI, F.; PELTER, M. M.; HU, X. Wavelet leader multifractal analysis of heart rate variability in atrial fibrillation. **Journal of Electrocardiology**, Philadelphia, v. 51, n. 6S, p. S83–S87, 2018.

GAROFOLO, J. S.; LAMEL, L. E; FISHER, W. M.; FISCUS, J. G.; PALLETT, D. S.; DAHLGREN, N. L. **The DARPA TIMIT acoustic-phonetic continuous speech corpus**. Philadelphia: Linguistic Data Consortium, 1993.

GIANNAKOPOULOS, T. **A method for silence removal and segmentation of speech signals, implemented in Matlab**. Athens: University of Athens, Athens, 2009.

GUPTA, S.; FAHAD, M. S.; DEEPAK, A. Pitch-synchronous single frequency filtering spectrogram for speech emotion recognition. **Multimedia Tools and Applications**, New York, v. 79, n. 31, p. 23347–23365, 2020.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research**, Cambridge, v. 3, p. 1157-1182, 2003.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques**. 3. ed. New York: Elsevier, 2012.

HANIFA, R. M.; ISA, K.; MOHAMAD, S. Speaker ethnic identification for continuous speech in malay language using pitch and MFCC. **Indonesian Journal of Electrical Engineering and Computer Science**, Padang, v. 19, n. 1, p. 207, 2020.

HARTMANN, W. M. Pitch, periodicity, and auditory organization. **Journal of the Acoustical Society of America**, Melville, v. 100, n. 6, p. 3491–3502, 1996.

HUANG, C. J.; YANG, Y. J.; YANG, D. X.; CHEN, Y. J. Frog classification using machine learning techniques. **Expert Systems with Applications**, Oxford, v. 36, n. 2, Part 2, p. 3737–3743, 2009.

HUI, L.; DAI, B. Q.; WEI, L. A pitch detection algorithm based on AMDF and ACF. *In: ICASSP, IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING*, 2006, Toulouse. **Proceedings of the [...]**.Toulouse: IEEE, 2006.

ITU-T. Percentual Evaluation of Speech Quality (PESQ). **International Telecommunication Union (ITU)**. Geneve, Switzerland, 1996. 862 p.

JAFFARD, S. Wavelet techniques in multifractal analysis. **Proceedings of Symposia in Pure Mathematics**, Providence, v. 72, n. 2, p. 91–152, 2004,

JAIN, A. K.; DUIN, R. P. W.; JIANCHANG, M. Statistical pattern recognition: A review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Piscataway, v. 22, n. 1, p. 4–37, 2000.

JOHNSTON, J. D. Transform coding of audio signals using perceptual noise criteria. **IEEE Journal on Selected Areas in Communications**, Piscataway, v. 6, n. 2, p. 314–323, 1988.

KANDOI, G.; ACENCIO, M. L.; LEMKE, N. Prediction of druggable proteins using machine learning and systems biology: a mini-review. **Frontiers in Physiology**, Lausanne, v. 6, p. 366, 2015.

KERMORVANT, C. **A comparison of noise reduction techniques for robust speech recognition.** IDIAP-RR 10, 1999. Disponível em: citeseer.ist.psu.edu/article/kermorvant99comparison.html. Acesso em: 20 abr. 2019.

KHUNARSAL, P.; LURSINSAP, C.; RAICHAROEN, T. Very short time environmental sound classification based on spectrogram pattern matching. **Information Sciences**, Philadelphia, v. 243, p. 57–74, 2013.

KINNUNEN, T.; LI, H. An overview of text-independent speaker recognition: from features to supervectors. **Speech Communication**, Amsterdam, v. 52, n. 1, p. 12–40, 2010.

KUBAT, M. **An introduction to machine learning.** New York: Springer international publishing, 2015. 273 p.

LARRAÑAGA, P.; CALVO, B.; SANTANA, R.; BIELZA, C.; GALDIANO, J.; INZA, I.; LOZANO, J.A.; ARMAÑANZAS, R.; SANTAFÉ, G.; PÉREZ, A.; ROBLES, V. Machine learning in bioinformatics. **Briefings in Bioinformatics**, Oxford, v. 7, p. 86–112, 2006.

LATHI, B. P. **Sinais e sistemas lineares.** 2. ed. New York: Bookman, 2006. 680 p.

LEHNER, B.; WIDMER, G.; SONNLEITNER, R. On the reduction of false positives in singing voice detection. *In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP) 2014, Florence. Proceedings of the [...].* Florence: IEEE, 2014.

LEONARDUZZI, R.; WENDT, H.; JAFFARD, S.; ROUX, S.; TORRES, M. E.; ABRY, P. Extending multifractal analysis to negative regularity: P-exponents and P-leaders. *In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2014, Florence. Proceedings of the [...].* Florence: IEEE, 2014.

LI, D.; SETHI, I. K.; DIMITROVA, N.; MCGEE, T. Classification of general audio data for content-based retrieval. **Pattern Recognition Letters**, Amsterdam, v. 22, n. 5, p. 533–544, 2001.

LIBBRECHT, M. W.; NOBLE, W. S. Machine learning applications in genetics and genomics. **Nature Reviews Genetics**, London, v. 16, p. 321-332, 2015.

LIE, L.; HONG-JIANG, Z.; HAO, J. Content analysis for audio classification and segmentation. **IEEE Transactions on Speech and Audio Processing**, Piscataway, v. 10, n. 7, p. 504–16, 2002.

LIN, L.; LI, Y.; SADEK, A. A k nearest neighbor based local linear wavelet neural network model for online short-term traffic volume prediction. **Procedia - Social and Behavioral Sciences**, Amsterdam, v. 96, n. 20, p. 66–77, 2013.

MACIEL, A. M. A. **Investigação de um ambiente de processamento de voz utilizando VoiceXML**. 2007. 83 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Pernambuco, Recife, 2007.

MALLAT, S. G. A theory for multiresolution signal decomposition: the wavelet representation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Piscataway, v. 11, n. 7, p. 674–693, 1989.

MANDELBROT, B. B.; FRAME, M. Fractals. *In: MEYERS, R. A. (ed.). Encyclopedia of physical science and technology*. 3. ed. New York: Academic Press, 2003. p. 185–207.

MISRA, H.; IKBAL, S.; BOURLARD, H.; HERMANSKY, H. Spectral entropy based feature for robust ASR. *In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING 2004, Montreal. Proceedings of the [...].* Montreal: IEEE, 2004.

MORENO, P. **Speech recognition in noisy environments**. 1996. PhD (Thesis) - Carnegie Mellon University, 1996.

MUSTAQEEM; SAJJAD, M.; KWON, S. Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. **IEEE Access**, Piscataway, v. 8, p. 79861–79875, 2020.

NATIONAL SCIENCE AND TECHNOLOGY COUNCIL (NSTC). **Speaker recognition**. 2006. P. 15-17, 2006.

OPPENHEIM, A. V. **Discrete-time signal processing**. 2. ed. New Jersey: Prentice Hall, 1999. 589 p.

PEETERS, G. **A large set of audio features for sound description (similarity and classification) in the CUIDADO project**. 2004. 25 p. Disponível em: http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf. Acesso em: 04 jan. 2021.

PETETIN, Y.; LAROCHE, C.; MAYOUE, A. Deep neural networks for audio scene recognition. *In: EUROPEAN SIGNAL PROCESSING CONFERENCE (EUSIPCO) 2015, Nice. Proceedings of the [...]*. Nice: IEEE, 2015.

PETRY, A.; SOARES, S. S.; BARONE, D. A. C. **Sistema para autenticação de usuários por voz em redes de computadores**. Porto Alegre: Ed. da UFRGS, 2003.

PETRY, A.; ZANUZ, A.; BARONE, D. A. C. **Utilização de técnicas de processamento digital de sinais para a identificação automática de pessoas pela voz**. Porto Alegre: Ed. da UFRGS, 1999.

PICONE, J. W. Signal modeling techniques in speech recognition. **Proceedings of the IEEE**, Piscataway, v. 81, n. 9, 1993.

PIKRAKIS, A.; GIANNAKOPOULOS, T.; THEODORIDIS, S. A Speech/music discriminator of radio recordings based on dynamic programming and bayesian networks. **IEEE Transactions on Multimedia**, Piscataway, v. 10, n. 5, p. 846–857, 2008.

PRAKASH, V. J.; NITHYA, D. L. A survey on semi-supervised learning techniques. **International Journal of Computer Trends and Technology (IJCTT)**, Trichy, v. 8, n. 1, p. 25–29, 2014.

PUCHALSKI, A. Application of the wavelet multifractal analysis of vibration signal for rotating machinery diagnosis. **Vibrations in Physical Systems**, v. 30, n. 2, p. 1-8, 2019.

RABINER, L. **Fundamentals of speech recognition**. New Jersey: Prentice Hall PTR, 1993.

RABINER, L. An algorithm for determining the endpoints of isolated utterances. **The Bell System Technical Journal**, Piscataway, v. 54, n. 2, p. 297–315, 1975.

RABINER, L. **Digital processing of speech signals**. New Jersey: Prentice Hall, 1978. 364 p.

RABINER, L.; SCHAFER, R. **Theory and applications of digital speech processing**. New Jersey: Prentice Hall, 2010.

RAKESH, K.; DUTTA, S.; SHAMA, K. Gender recognition using speech processing techniques in LABVIEW. **International Journal of Advances in Engineering & Technology**, Chennai, v. 1, n. 2, p. 51, 2011.

RAO, K. R.; YIP, P. **Discrete cosine transform: algorithms, advantages, applications**. San Diego: Academic Press, 1990.

REYNOLDS, D. A. Robust text-independent speaker identification using Gaussian mixture speaker models. **IEEE Transactions on Speech and Audio Processing**, Piscataway, v. 3, n. 1, p. 1094–1105, 1995.

RIBEIRO, V. J.; RIEDI, R. H.; BARANIUK, R. G. Wavelets and multifractals for network traffic modeling and inference. *In: PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING - ICASSP, 2001, Salt Lake City. Proceedings of the [...]*. Salt Lake City: IEEE, 2001.

ROSE, P. **Forensic speaker identification**. Nova Iorque: Taylor and Francis, 2002.

SANT'ANA, R.; COELHO, R.; ALCAIM, A. Text-Independent Speaker Recognition Based on the Hurst Parameter and the Multidimensional Fractional Brownian Motion Model. **IEEE Transactions on Audio, Speech and Language Processing**, Piscataway, v. 14, n. 3, p. 931-940, 2006.

SERRANO, E.; FIGLIOLA, A. Wavelet Leaders: A new method to estimate the multifractal singularity spectra. **Physica A: Statistical Mechanics and Its Applications**, Amsterdam, v. 388, n. 14, p. 2793–2805, 2009.

SHAO, X.; MILNER, B.; COX, S. **Integrated pitch and MFCC extraction for speech reconstruction and speech recognition applications**. *In: EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY - EUROSPEECH - INTERSPEECH*, p. 1725–1728, 2003. Disponível em: https://www.isca-speech.org/archive/pdfs/eurospeech_2003/shao03_eurospeech.pdf. Acesso em: 04 jan. 2021.

SHARMA, A. M. **Speaker recognition using machine learning techniques**. 2019. Thesis. (Master's Projects) - San José State University, San José, 2019.

SMITH, S. W. **The scientist and engineer's guide to digital signal processing**. San Diego: California Technical Publishing, 1999.

SOARES, W. C.; VILLARREAL, F.; DUARTE, M. A. Q.; VIEIRA FILHO, J. Wavelets in a problem of signal processing. **Novi Sad Journal of Mathematics**, Novi Sad, v. 41, n. 1, p. 11-12, 2011.

THE MATHWORKS INC. **Interactive machine learning with Matlab**. MathWorks, 2016. Disponível em: <https://explore.mathworks.com/interactive-machine-learning-with-matlab/>. Acesso em: 20 abr. 2016.

THE MATHWORKS INC. **Speaker identification using pitch and mfcc - matlab and simulink.** Disponível em: <https://www.mathworks.com/help/audio/examples/speaker-identificationusing-pitch-and-mfcc.html>. Acesso em: 16 abr. 2019.

TIRUMALA, S. S.; SHAHAMIRI, S. R.; GARHWAL, A. S.; WANG, R. Speaker identification features extraction methods: A systematic review. **Expert Systems with Applications**, Oxford, v. 90, p. 250-271, 2017.

TOKUDA, K. **Recursive calculation of Mel-cepstrum from LP coefficients.** p. 1–7, 1994. p. 1-7. Disponível em: https://www.sp.nitech.ac.jp/~tokuda/tips/mgceptr_sa2.pdf. Acesso em: 04 jan. 2021.

VIEIRA, P. G. **Seleção da faixa de frequência usando wavelets para detecção de danos em sistemas SHM baseados no princípio da EMI.** 2016. Dissertação (Mestrado) – Faculdade de Engenharia, Universidade Estadual Paulista, Ilha Solteira, 2016. Disponível em: <https://repositorio.unesp.br/handle/11449/148608?locale-attribute=es>. Acesso em: 04 jan. 2021.

VIEIRA, P. G.; VIEIRA FILHO, J.; DUARTE, M. A. Q. Selection of the optimal frequency band using the wavelet packet transform in shm systems. *In: NATIONAL MEETING ON COMPUTATIONAL MODELING, 19, MEETING ON MATERIALS SCIENCE AND TECHNOLOGY, 7, 2016. Proceedings of the [...].* [s.l: s.n.], 2016.

VIEIRA FILHO, J. **Redução de ruído em sinais de voz nos sistemas rádio móveis veiculares.** 1996. Tese (Doutorado) – Faculdade de Engenharia Elétrica e Computação, Universidade Estadual de Campinas, Campinas, 1996.

WARDEN, P. Speech commands: a dataset for limited-vocabulary speech recognition. **ArXiv**, [s. l.], 2018.

WENDT, H.; ABRY, P. Multifractality tests using Bootstrapped Wavelet Leaders. **IEEE Transactions on Signal Processing**, Piscataway. v. 55, n. 10, p. 4811–4820, 2007.

WENDT, H.; ROUX, S. G.; JAFFARD, S.; ABRY, P. Wavelet leaders and bootstrap for multifractal analysis of images. **Signal Processing**, Amsterdam, v. 89, n. 6S, p.1100-1114, 2009.

WICKERHAUSER, M. V. **Adapted wavelet analysis from theory to software.** New York: AK Peters, 1994. p.237-272.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining: practical machine learning tools and techniques.** [S.l.]: Morgan Kaufmann, 2011.

YEN, G. G.; LIN, K. C. Wavelet packet feature extraction for vibration monitoring. **IEEE Transactions on Industrial Electronics**, Piscataway, v. 47, p. 650-667, 2000.

ZHANG, S.; GUO, Y.; ZHANG, Q. Robust voice activity detection feature design based on spectral kurtosis. *In: FIRST INTERNATIONAL WORKSHOP ON EDUCATION TECHNOLOGY AND COMPUTER SCIENCE 2009*, Wuhan, Hubei.

Proceedings of the [...]. Wuhan: IEEE, 2009.

ZHOU, Y.; ZHANG, L. Speaker recognition system based on multifractal spectrum feature and characters selection policy. **Journal of Frontiers of Computer Science and Technology**, Beijing, v. 8, n. 11, p. 1752-1761, 2018.

ZUBAIR, S.; YAN, F.; WANG, W. Dictionary learning based sparse coefficients for audio classification with max and average pooling. **Digital Signal Processing**, Waltham, v. 23, n. 3, p. 960–70, 2013.