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

André Furlan

Monografia de estudos especiais II

São José do Rio Preto

2023

André Furlan

Autenticação Biométrica de Locutores Drasticamente Disfônicos Aprimorada pela *Imagined Speech*

Monografia apresentada para cumprimento da disciplina de estudos especiais do curso de Doutorado em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Orientador: Prof. Dr. Rodrigo Capobianco Guido

São José do Rio Preto, SP

2023

André Furlan

Autenticação Biométrica de Locutores Drasticamente Disfônicos Aprimorada pela *Imagined Speech*

Monografia apresentada para cumprimento da disciplina de estudos especiais II do curso de Doutorado em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Orientador: Prof. Dr. Rodrigo Capobianco Guido

Comissão Examinadora

Professor Dr. Rodrigo Capobianco Guido
UNESP - Campus de São José do Rio Preto
Co-Orientador

Professora Dra. Veronica Oliveira de Carvalho
UNESP - Campus de Rio Claro

Professor Dr. Lucas Correia Ribas
UNESP - Campus de São José do Rio Preto

São José do Rio Preto, SP
2023

RESUMO

Este documento constitui a monografia produzida como resultado dos estudos especiais realizados pelo autor visando promover um levantamento bibliográfico inicial do tema de sua tese de doutorado. Foram inclusas as descrições essenciais de 14 trabalhos científicos na área de *imagined speech* e processamento de sinais de locutores disfônicos ou não, acrescidos ainda, das direções que caracterizam o tema do trabalho em questão visando a monografia de qualificação.

No capítulo 1 é feita uma breve introdução. Em seguida no Capítulo 2 são listados os objetivos e metas do estudo. Na seção 3.1 iniciando as revisões de conceitos apresenta-se a engenharia paraconsistente de características, filtros digitais *wavelets*, amostragem, quantização, caracterização dos processos de produção da voz humana e definição das redes neurais que deverão ser empregadas. Na seção 3.2 apresenta-se a revisão bibliográfica inicial. Finalmente no capítulo 4 é mostrado um cronograma com os trabalhos já realizados e uma previsão da realização dos próximos.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Com base nas pesquisas anteriores no campo do processamento de sinais de voz, como evidenciado na dissertação [Fur21], há uma demanda significativa por mecanismos de autenticação biométrica de locutores (ABLs). O entendimento é apoiado por obras científicas importantes, como [BB11] e [NP12], além do volume de artigos científicos publicados em revistas de prestígio, como o IEEE Signal Processing Magazine [HH15], o Elsevier Neurocomputing [WMWX22] e o Elsevier Computer Speech and Language [LSLR20].

Também é perceptível a demanda por uma linha diversa de pesquisa na análise de sinais de voz, que tem atraído a atenção da comunidade científica internacional, como visto em publicações recentes, como [CSBY22] e [FKO+22]. Essa linha de pesquisa se concentra nas estratégias acústico-computacionais para o pré-diagnóstico e classificação não invasiva de alterações laríngeas e outras irregularidades que afetam a fonação. Esses estudos têm se baseado em metodologias avançadas, como as redes neurais profundas [GBC16a], exemplificadas pelas referências [MS21] e [MPP21], e também na lógica paraconsistente, como descrito em [FPM+17].

É importante destacar que a combinação dessas duas linhas de pesquisa apresenta um desafio significativo: autenticar locutores afetados por distúrbios vocais de origem orgânica, funcional ou orgânico-funcional [LHA05a] [LHA05b], que alteram a impressão acústica da fala. Essa questão tem sido objeto de investigação no âmbito deste projeto de douto-

rado, financiado pela FAPESP e pelo CNPq, com alguns avanços registrados [GPP+21]. Uma abordagem mais recente considerada para aprimorar essas investigações é a análise da fala imaginada [MTGVPC19] associada às locuções degradadas.

Ao pesquisar o tema em questão nos registros do Web of Science e outras bases de dados, como ieeexplore.org, springernature.com e sciencedirect.com, observa-se uma escassez de artigos científicos. Por um lado, a maioria dos artigos é composta por propostas genéricas que envolvem metodologias para viabilizar interfaces cérebro-computador (BCIs), conforme descrito em [RG21] e [BK10], ou têm como objetivo o reconhecimento da fala imaginada em vez da autenticação de locutores, como pode ser observado nas referências de [PD22], [CFC21], [LBBPM22], [BKEM22], [LLL21] e [TMM20]. Por outro lado, os artigos mais relevantes e dedicados à biometria são encontrados nas referências [MTGVPC19], [MM18], [JCA16], [JCA17], [DPBATRT14] e [RBJL16]. No entanto, nenhum desses trabalhos oferece uma solução definitiva para o problema em questão, especialmente quando a disfonia laríngea existente é grave e de origem orgânica, funcional ou orgânico-funcional. Portanto, ainda há espaço para futuras investigações.

Dessa forma, o objetivo deste projeto de pesquisa é conceber algoritmos biométricos para autenticar indivíduos capazes de produzir apenas locuções severamente degradadas, complementando as informações da voz com os sinais cerebrais durante a fonação, ou seja, provenientes da fala imaginada. Essa proposta está na interseção das subáreas de processamento de sinais biológicos, eletrônica, sistemas inteligentes e neurociência.

Capítulo 2

Objetivos e metas

O objetivo geral deste trabalho é o de projetar e implementar algoritmos biométricos capazes de autenticar, em princípio por meio da fala, indivíduos capazes de produzir somente locuções potencialmente degradadas, adicionando informações extraídas dos sinais cerebrais durante a fonação, isto é, a *imagined speech*, ao conjunto daquelas que são acústicas e oriundas da voz prejudicada. Em um nível mais específico, os propósitos são os seguintes:

1. Realizar um amplo levantamento bibliográfico sobre o estado-da-arte na área de ABLs com disfonias laríngeas severas (DLSs), complementando-as com o estudo das estratégias do tipo BCI para decodificação de imagined speech.

2. Estudar as bases de dados públicas de *imagined speech*, tal como a referida no artigo científico [CGR17], visando possibilitar experimentos iniciais.

3. Criar, por meio do fortalecimento de parcerias com profissionais da área da saúde, uma base de dados contendo sinais de voz oriundos de indivíduos com DLS e os respectivos sinais de imagined speech, a qual deverá ser disponibilizada publicamente. Pretende-se incluir locuções pré-definidas e outras diversificadas, de centenas de indivíduos.

4. Extrair, com vistas à ABLs, características representativas dos sinais de voz e de *imagined speech* de cada locutor, experimentando e comparando as modalidades *handcrafted extraction*, considerando diferentes conceitos físicos para a análise tempo-frequência das locuções, e *feature learning*, com base em autoencoders.

5. Autenticar os locutores presentes na base de dados, representados pelos seus respectivos vetores de características (VsCs) que contemplam informações das vozes degradadas e das *imagined speechs*, utilizando estruturas profundas de aprendizagem, tais como as Redes Neurais Residuais (RNNs) e as Deep Spiking Neural Networks (DSNNs), comparando-as em termos de custo computacional para treinamento e acurácia. À medida que os dados disponíveis permitirem, considerar-se-ão duas modalidades: autenticação *text-dependent* e *text-independent*.

6. Implementar os algoritmos criados usando a linguagem de programação C/C++ para funcionamento *off-line* e, na medida do possível, em dispositivos que permitam a execução em tempo real.

7. Disseminar os resultados na literatura científica, por meio da apresentação de trabalhos em congressos e publicações em periódicos de alto impacto.

Capítulo 3

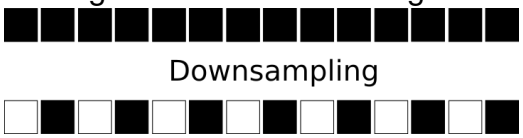
Revisão de Bibliográfica

3.1 Conceitos utilizados

3.1.1 Sinais digitais e sub-amostragem (*downsampling*)

Os sinais digitais, tanto de voz quanto aqueles vindos das medições de Eletroencefalograma (ECG), isto é, aqueles que estão amostrados e quantizados [HM11], constituem a base deste trabalho. Além do processo de digitalização, inerente ao ato de armazenar sinais em computadores, os mesmos podem sofrer, a depender da necessidade ou possibilidade, sub-amostragens ou *downsamplings* [P+96]. Isso implica em uma estratégia de redução de dimensão e, comumente, ocorre após a conversão de domínio dos sinais com base em filtros digitais do tipo *wavelet*, a serem apresentados adiante. Um exemplo consta na Figura 3.1, na qual as partes pretas contêm dados e as brancas representam os elementos removidos. Tendo em vista que este trabalho está baseado em sinais digitais de voz e ECG com base em *wavelets*, o processo de sub-amostragem é essencial.

Figura 3.1: Sub-amostragem



Fonte: Elaborado pelo autor, 2023.

3.1.2 Caracterização dos processos de produção da voz humana

A fala possui três grandes áreas de estudo: A fisiológica, também conhecida como fonética articulatória, a acústica, referida como fonética acústica, e ainda, a perceptual, que cuida da percepção da fala [KG14]. Neste trabalho, o foco será apenas na questão acústica, pois não serão analisados aspectos da fisiologia relacionada à voz, mas sim os sinais sonoros propriamente ditos.

Sinais vozeados *versus* não-vozeados

Quando da análise dos sinais de voz, consideram-se as partes vozeadas e não-vozeadas. Aquelas são produzidas com a ajuda da vibração quase periódica das pregas vocais, enquanto estas praticamente não contam com participação regrada da referida estrutura.

Frequência fundamental da voz

Também conhecida como F_0, é o componente periódico resultante da vibração das pregas vocais. Em termos de percepção, se pode interpretar F_0 como o tom da voz, isto é, a frequência de *pitch* [KG14]. Vozes agudas tem uma frequência de *pitch* alto, enquanto vozes mais graves tem baixa. A alteração da frequência (jitter) e/ou intensidade (shimmer) do *pitch* durante a fala é definida como entonação, porém, também pode indicar algum distúrbio ou doença relacionada ao trato vocal [WSA05].

A frequência fundamental da voz é o número de vezes na qual uma forma de onda característica, que reflete a excitação pulmonar moldada pelas pregas vocais, se repete por unidade de tempo. Sendo assim, as medidas de F_0 geralmente são apresentadas em Hz [Fre13].

A medição de F_0 está sujeita a contaminações surgidas das variações naturais de *pitch* típicas da voz humana [Fre13]. A importância de se medir F_0 corretamente vem do fato de que, além de carregar boa parte da informação da fala, ela é a base para construção das outras frequências que compõe os sinais de voz, que são múltiplas de F_0.

Formantes

O sinal de excitação que atravessa as pregas vocais é rico em harmônicas, isto é, frequências múltiplas da fundamental. Tais harmônicas podem ser atenuadas ou amplificadas, em função da estrutura dos tratos vocal e nasal de cada locutor. Particularmente, o primeiro formante (F_1), relaciona-se à amplificação sonora na cavidade oral posterior e à posição da língua no plano vertical; o segundo formante (F_2) à cavidade oral anterior e à posição da língua no plano horizontal; o terceiro formante (F_3) relaciona-se às cavidades à frente e atrás do ápice da língua e, finalmente, o quarto formante (F_4) relaciona-se ao formato da laringe e da faringe na mesma altura [V+14]. Formantes caracterizam fortemente os locutores, pois cada indivíduo possui um formato de trato vocal e nasal. Assim, tais frequências, que podem ser capturadas com ferramentas diversas, a exemplo da Transformada *Wavelet*, são de suma importância na área de verificação de locutores.

3.1.3 Escalas e energias dos sinais

A energia de um sinal digital $s[\cdot]$ com M amostras é definida como

$$E = \sum_{i=0}^{M-1} (s_i)^2 \qquad . \tag{3.1}$$

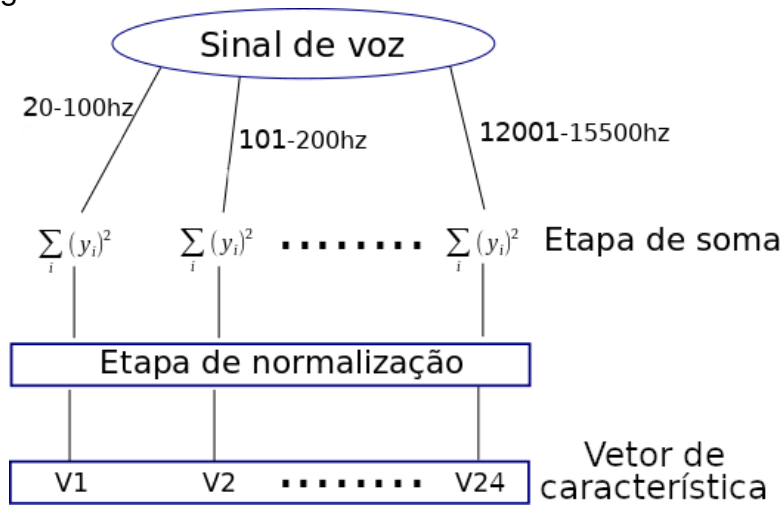
E pode ainda sofrer normalizações e ter a sua mensuração restrita a uma parte específica do sinal sob análise. Possibilidades para tais restrições podem, por exemplo, envolver a escala BARK [Zwi61] e MEL [Ber49] que serão utilizadas neste trabalho.

A escala BARK

BARK foi definida tendo em mente vários tipos de sinais acústicos. Essa escala corresponde ao conjunto de 25 bandas críticas da audição humana. Suas frequências-base de audiometria são, em Hz: **20, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500**. Nessa escala,os sinais digitais no domínio temporal atravessam filtros passa-faixas [BG02] para

os quais o início e o final da banda de passagem correspondem à frequências-base consecutivas resultando em um vetor de características com 24 coeficientes e, em seguida, as energias dos sinais filtrados são utilizadas como características descritivas de propriedades do sinal sob análise, como mostrado na Figura 3.2.

Figura 3.2: Cálculo de vetores de características com BARK



Fonte: Elaborado pelo autor, 2023.

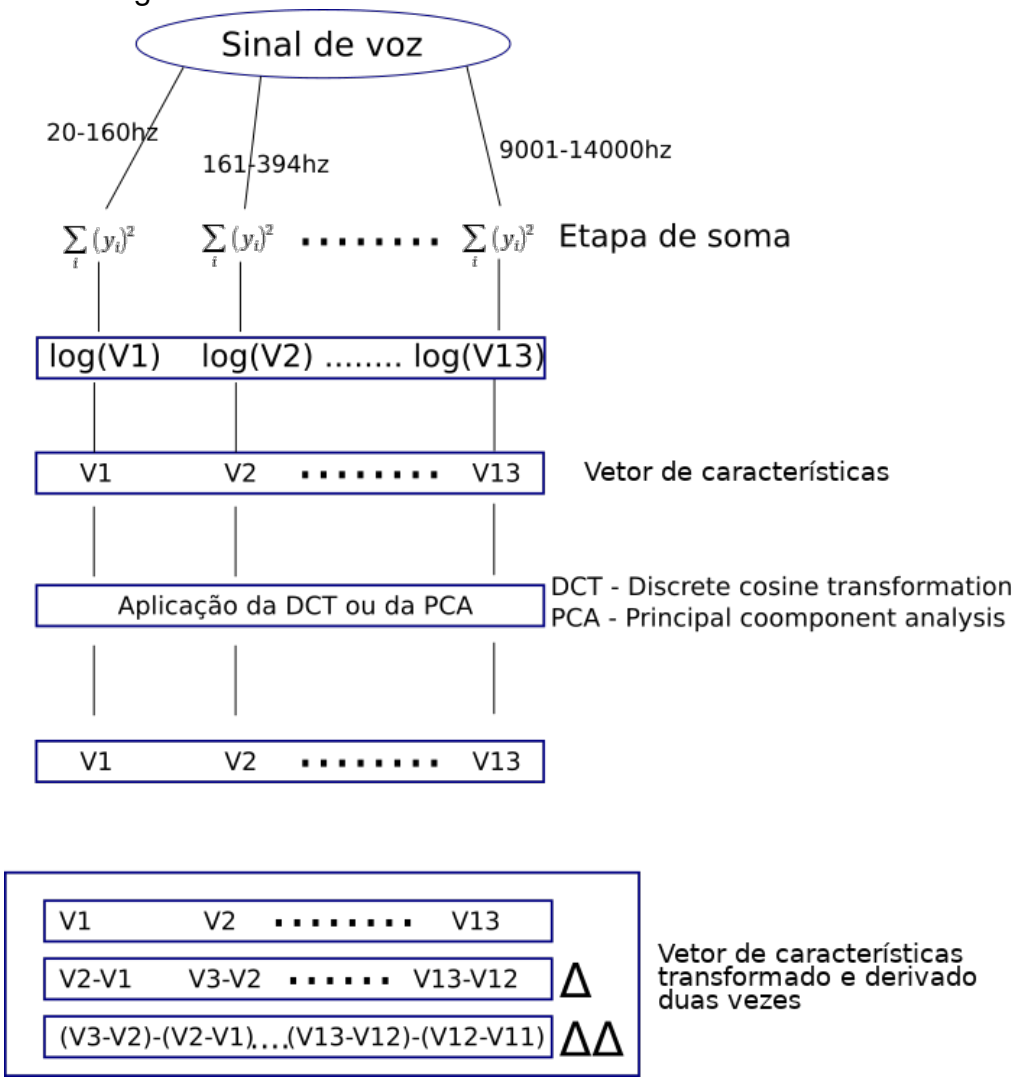
A escala MEL

Escala Mel, advinda do termo *melody*, é uma adaptação da escala Bark para sinais de voz. Dentre as várias implementações de bandas críticas a escolhida foi a implementação que contém os valores em Hz: **20, 160, 394, 670, 1000, 1420, 1900, 2450, 3120, 4000, 5100, 6600, 9000, 14000**.

A variante que será usada neste trabalho é conhecida como *Mel-frequency cepstral coefficients*(MFCC) a qual inclui, além dos intervalos definidos, uma diminuição da correlação entre os componentes gerados via aplicação da Transformada Discreta Cosseno (DCT) [SMB07] ou da Análise de Componentes Principais (PCA) [Jol02] seguida de duas derivações no vetor de características resultando em um total de 11 coeficientes. Nesse trabalho foi escolhida a DCT, no entanto, PCA poderia também ser escolhida sem prejuízos, o uso de uma ou outra depende da preferência do autor.

Novamente, desconsiderando qualquer etapa intermediária que possa ser adicionada, as energias calculadas nos intervalos definidos na escala MEL podem, por si mesmas, constituir um vetor de características, como mostrado na Figura 3.2.

Figura 3.3: Cálculo de vetores de características com MEL



Fonte: Elaborado pelo autor, 2023.

3.1.4 Filtros digitais *wavelet*

Filtros digitais *wavelet* têm sido utilizados com sucesso para suprir as deficiências de janelamento de sinal apresentadas pelas Transformadas de Fourier e de Fourier de

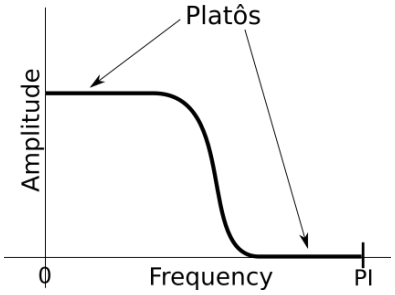
Tempo Reduzido. *Wavelets* contam com variadas funções-filtro e têm tamanho de janela variável, o que permite uma análise multirresolução [AWG09]. Particularmente, as *wavelets* proporcionam a análise do sinal de forma detalhada tanto no espectro de baixa frequência quanto no de alta contando com diferentes funções-base não periódicas diferentemente da tradicional transformada de Fourrier que utilizam somente as bases periódicas senoidal e cossenoidal.

É importante observar que, quando se trata de Transformadas *Wavelet*, seis elementos estão presentes: dois filtros de análise, dois filtros de síntese e as funções ortogonais *scaling* e *wavelet*. No tocante a sua aplicação, só a transformada direta, e não a inversa, será usada na construção dos vetores de características. Portanto, os filtros de síntese, a função *scaling* e a função *wavelet* não serão elementos abordados aqui: eles somente interessariam caso houvesse a necessidade da transformada inversa.

No contexto dos filtros digitais baseados em *wavelets*, o tamanho da janela recebe o nome de **suporte**. Janelas definem o tamanho do filtro que será aplicado ao sinal. Quando esse é pequeno (limitado), se diz que a janela tem **um suporte compacto** [P+96].

Se diz que uma *wavelet* tem boa **resposta em frequência** quando, na aplicação da mesma para filtragem, não são causadas muitas pertubações indesejadas ao sinal, no domínio da frequência. Os filtros *wavelet* de Daubechies [Dau92] se destacam nesse quesito por serem *maximamente planos* (*maximally-flat*) [But30] [Bia07] nos platôs de resposta em frequência como indicado na Figura 3.4 ao contrário do que ocorre na Figura 3.5.

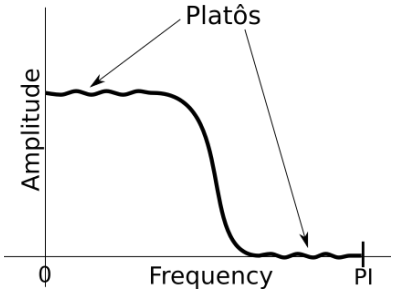
Figura 3.4: Platôs maximamente planos em um filtro digital: característica da família de Daubechies



Fonte: Elaborado pelo autor, 2023.

Além da resposta em frequência, na aplicação de um filtro digital *wavelet* também é

Figura 3.5: Platôs não maximamente planos de um filtro digital: características de outros filtros *wavelet*, distintos da família de Daubechies



Fonte: Elaborado pelo autor, 2023.

possível considerar a **resposta em fase**, que constitui um atraso ou adiantamento do sinal filtrado em relação ao sinal original, ambos no domínio temporal. Esse deslocamento pode ser **linear**, **quase linear** ou **não linear**:

- na resposta em fase **linear**, há o mesmo deslocamento de fase para todos os componentes do sinal;
- quando a resposta em fase é **quase linear** existe uma pequena diferença no deslocamento dos diferentes componentes do sinal;
- finalmente, quando a resposta é **não linear**, acontece um deslocamento significativamente heterogêneo para as diferentes frequências que compõe o sinal.

Idealmente, é desejável que todo filtro apresente boa resposta em frequência e em fase linear. Características de fase e frequência de algumas famílias de filtros *wavelet* constam na Tabela 3.1.

O algoritmo de Mallat para a Transformada *Wavelet*

Baseando-se no artigo [7079589], percebe-se que algoritmo de Mallat faz com que aplicação das *wavelets* seja uma simples multiplicação de matrizes. O sinal que deve ser transformado se torna uma matriz linear vertical. Os filtros passa-baixa e passa-alta tornam-se, nessa ordem, linhas de uma matriz quadrada que será completada segundo regras que

Tabela 3.1: Algumas das *wavelets* mais usadas e suas propriedades

Wavelet	**Resposta em frequência**	**Resposta em fase**
Haar	Pobre	Linear
Daubechies	mais próxima da ideal à medida que o suporte aumenta; *maximally-flat*	Não linear
Symmlets	mais próxima da ideal à medida que o suporte aumenta; não *maximally-flat*	Quase linear
Coiflets	mais próxima da ideal à medida que o suporte aumenta; não *maximally-flat*	Quase linear

Fonte: Elaborado pelo autor, 2023.

serão mostradas mais adiante. É importante que essa matriz quadrada tenha a mesma dimensão que o sinal a ser transformado.

Interessantemente, para que seja possível a transformação *wavelet*, basta ter disponível o vetor do filtro passa-baixas calculado a partir da *mother wavelet*, que é a função geradora desse filtro, já que o passa-alta pode ser construído a partindo-se da ortogonalidade do primeiro.

Determinar a ortogonal de um vetor significa construir um vetor, tal que, o produto escalar do vetor original com sua respectiva ortogonal seja nulo.

Considerando $h[\cdot]$ como sendo o vetor do filtro passa-baixas e $g[\cdot]$ seu correspondente ortogonal, tem-se que $h[\cdot] \cdot g[\cdot] = 0$.

Portanto, se $h[\cdot] = [a, b, c, d]$ então seu ortogonal será $g[\cdot] = [d, -c, b, -a]$ pois:

$$h[\cdot] \cdot g[\cdot] = [a, b, c, d] \cdot [d, -c, b, -a] = (a \cdot d) + (b \cdot (-c)) + (c \cdot b) + (d \cdot (-a)) = ad - ad + bc - bc = 0 \quad .$$

A título de exemplo, considera-se:

- o filtro passa baixa baseado na *wavelet* Haar: $h[\cdot] = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$
- o seu respectivo vetor ortogonal: $g[\cdot] = [\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}]$
- e também o seguinte sinal-exemplo de entrada: $s = \{1, 2, 3, 4\}$

Se o tamanho do sinal a ser tratado é quatro e se pretende-se aplicar o filtro Haar, a

seguinte matriz de coeficientes é construída:

$$\begin{pmatrix} \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0 \\ \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0 \\ 0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \\ 0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \end{pmatrix} \tag{3.2}$$

Tendo em vista que a dimensão do sinal sob análise é diferente da dimensão do filtro, basta completar cada uma das linhas da matriz de coeficientes com zeros. A matriz é montada de forma que ela seja ortogonal.

Montada a matriz de filtros, segue-se com os cálculos da transformada:

$$\begin{pmatrix} \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0 \\ \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0 \\ 0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \\ 0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \end{pmatrix} . \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} \frac{3}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ \frac{7}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix} \tag{3.3}$$

Realizada a multiplicação, é necessário montar o sinal filtrado. Isso é feito escolhendo, dentro do resultado, valores alternadamente de forma que o vetor resultante seja:

$$resultado = \left[\frac{3}{\sqrt{2}}, \frac{7}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, \frac{-1}{\sqrt{2}} \right] \quad . \tag{3.4}$$

Percebe-se que, na transformação descrita nas Equações 3.2, 3.3 e 3.4, a **aplicação dos filtros sobre o vetor de entrada ocorreu apenas uma vez**. Sendo assim, se diz que o sinal recebeu uma **transformação de nível 1**. A cada transformação, há uma separação do sinal em dois componentes: o de baixa e o de alta frequência.

Embora haja um limite, que será mencionado adiante, é possível aplicar mais de um nível de decomposição ao sinal. Para que se possa fazer isso, a Transformada *Wavelet* nível 2 deve considerar apenas a parte de baixas frequências da primeira transformada; a transformada de nível 3 deve considerar apenas a parte de baixas frequências da transformada nível 2, e assim consecutivamente.

Nos exemplos numéricos mostrados nas Tabelas 3.2, 3.3 e 3.4, usou-se um filtro normalizado cujos coeficientes são $\{\frac{1}{2}, -\frac{1}{2}\}$. Os dados destacados em **verde** correspondem ao **vetor original** que será tratado. Cada uma das linhas são os resultados das transformações nos níveis 1, 2, 3 e 4, respectivamente. As partes em **azul** correspondem à porção de **baixas frequências**, enquanto que as partes em **amarelo** correspondem às porções de **altas frequências**.

Percebe-se que na Tabela 3.2, a partir da transformação nível 2, apenas as partes de baixa frequência são modificadas. Isso implica que, no momento da implementação do algoritmo de Mallat **para níveis maiores que 1**, a abordagem será **recursiva**. Em outras palavras, a partir do nível 1 se deve aplicar Mallat apenas às porções de baixas-frequências geradas pela transformação anterior.

Tabela 3.2: Exemplo numérico da transformação *wavelet* aplicada a um vetor

Sinal	**32**	**10**	**20**	**38**	**37**	**28**	**38**	**34**	**18**	**24**	**24**	**9**	**23**	**24**	**28**	**34**
Nível 01	21	29	32,5	36	21	16,5	23,5	31	11	-9	4,5	2	-3	7,5	-0,5	-3
Nível 02	25	34,25	18,75	27,25	-4	-1,75	2,25	-3,75	11	-9	4,5	2	-3	7,5	-0,5	-3
Nível 03	29,62	23	-4,625	-4,25	-4	-1,75	2,25	-3,75	11	-9	4,5	2	-3	7,5	-0,5	-3
Nível 04	26,3125	3,3125	-4,625	-4,25	-4	-1,75	2,25	-3,75	11	-9	4,5	2	-3	7,5	-0,5	-3

Fonte: Elaborado pelo autor, 2023.

O algoritmo de Mallat e a Transformada *Wavelet-Packet*

Na Transformada *Wavelet-Packet*, os filtros aplicados são os mesmos da Transformada *Wavelet* e o procedimento recursivo de cálculo também é o mesmo, no entanto, realizada a transformação de nível 1, a transformada de nível 2 deve ser aplicada aos componentes de baixa e de alta frequência. Sendo assim a Transformada *Wavelet-Packet* obtém um nível de detalhes em todo o espectro de frequência, maior do que uma transformação regular.

Os exemplos mostrados nas Tabelas 3.3 e 3.4 permitem perceber como se dão as transformações na porção de **baixa** e de **alta** frequências, respectivamente, após a transformação *wavelet-packet* de nível 1, 2, 3 e 4.

Devido ao *downsampling* aplicado às porções de alta frequência, essas partes acabam por ficar “espelhadas” no espectro [JlCH01], ou seja, suas sequências ficam invertidas. Para resolver esse problema e preservar a ordem das sub-bandas no sinal transformado, os filtros são aplicados em ordem inversa nas porções de alta frequência. Isso altera como o algoritmo de Mallat deve ser implementado para a Transformada *Wavelet-Packet*, já que dessa vez é preciso se atentar a ordem da aplicação dos filtros passa-alta e passa-baixa.

Tabela 3.3: Exemplo numérico de *wavelet-packet* Haar aplicada ao vetor da Tabela 3.2 (porção das baixas frequências)

Sinal	**32**	**10**	**20**	**38**	**37**	**28**	**38**	**34**
Nivel 01	21	29	32,5	36	21	16,5	23,5	31
Nivel 02	25	34,25	18,75	27,25	-4	-1,75	2,25	-3,75
Nivel 03	29,62	23	-4,625	-4,25	-1,125	3	-2,875	-0,75
Nivel 04	26,3125	3,3125	-0,1875	-4,4375	0,9375	-2,0625	-1,0625	-1,8125

Fonte: Elaborado pelo autor, 2023.

Tabela 3.4: Exemplo numérico de *wavelet-packet* Haar aplicada ao vetor da Tabela 3.2 (porção das altas frequências)

Sinal	**18**	**24**	**24**	**9**	**23**	**24**	**28**	**34**
Nivel 01	11	-9	4,5	2	-3	7,5	-0,5	-3
Nivel 02	10	1,25	-5,25	1,25	1	3,25	2,25	-1,75
Nivel 03	5,625	-2	4,375	-3,25	-1,125	2	2,125	0,25
Nivel 04	1,8125	3,8125	3,8125	0,5625	0,4375	-1,5625	0,9375	1,1875

Fonte: Elaborado pelo autor, 2023.

3.1.5 Engenharia Paraconsistente de características

Nos processos de classificação, frequentemente surge a questão: “Os vetores de características criados proporcionam uma boa separação de classes?”. A Engenharia Paraconsistente de Características, recém publicada [Gui19], que usa a paraconsistência [dCBB98], [CA00] é, em meio a outras técnicas, uma ferramenta que pode ser usada para responder essa questão.

O processo inicia-se após a aquisição dos vetores de características para cada classe C_n. Se o número de classes presentes for, por exemplo, quatro então estas poderão ser representadas por C_1, C_2, C_3, C_4.

Em seguida é necessário o cálculo de duas grandezas:

- a menor similaridade intraclasse, α.
- a razão de sobreposição interclasse, β.

α indica o quanto de similaridade os dados têm entre si, dentro de uma mesma classe, enquanto β é a razão de sobreposição entre diferentes classes. Idealmente, α deve ser maximizada e β minimizada para que classificadores extremamente modestos apresentem uma acurácia interessante.

Particularmente, para calcular α e β, é necessária a normalização dos vetores de características de forma que todos os seus componentes estejam no intervalo entre 0 e 1. Em seguida, a obtenção de α se dá selecionando-se os maiores e os menores valores de cada uma das posições de todos os vetores de características para cada classe, gerando assim um vetor para os valores maiores e outro para os menores.

O **vetor de similaridade da classe**(svC_n) é obtido fazendo-se a diferença item-a-item dos maiores em relação aos menores. Finalmente, e para cada classe, é obtida a média dos valores para cada vetor de similaridade, sendo que α é o menor valor dentre essas médias. A Figura 3.6 contém uma ilustração do processo.

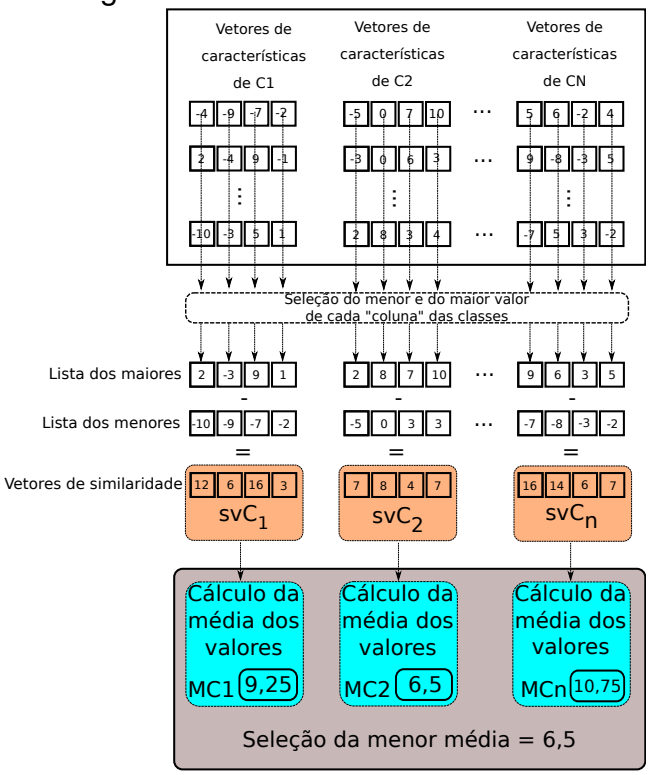
A obtenção de β, assim como ilustrado na Figura 3.7, também se dá selecionando os maiores e os menores valores de cada uma das posições de todos os vetores de características de cada classe, gerando assim um vetor para os valores maiores e outro para os menores.

Na sequência, realiza-se o cálculo de R cujo valor é a quantidade de vezes que um valor do vetor de características de uma classe se encontra entre os valores maiores e menores de outra classe.

Seja:

- N a quantidades de classes;

Figura 3.6: Cálculo do coeficiente α.



Fonte: Adaptado de [Gui19].

- X a quantidade de vetores de características por classe;
- T o tamanho do vetor de características.

Então, F, que é o número máximo de sobreposições possíveis entre classes, é dado por:

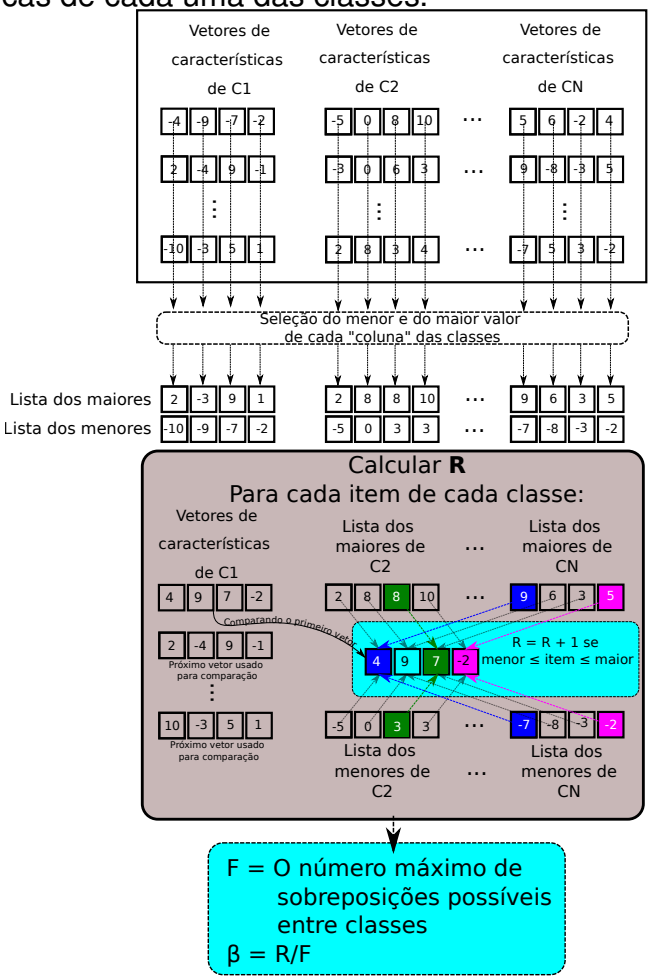
$$F = N.(N-1).X.T \quad . \tag{3.5}$$

Finalmente, β é calculado da seguinte forma:

$$\beta = \frac{R}{F} \quad . \tag{3.6}$$

Neste ponto, é importante notar que $\alpha = 1$ sugere fortemente que os vetores de características de cada classe são similares e representam suas respectivas classes precisamente. Complementarmente, $\beta = 0$ sugere os vetores de características de classes diferentes não se sobrepõe [Gui19].

Figura 3.7: Cálculo de β: Os itens destacados em azul e rosa são aqueles pertencentes a classe C1 e CN que se sobrepõe, em verde, a sobreposição é entre C1 e C2. Para cada sobreposição verificada soma-se 1 ao valor R. Essa comparação é feita para todos os vetores de características de cada uma das classes.



Fonte: Adaptado de [Gui19].

Considerando-se o plano paraconsistente [Gui19], temos:

- Verdade → fé total ($\alpha = 1$) e nenhum descrédito ($\beta = 0$)
- Ambiguidade → fé total ($\alpha = 1$) e descrédito total ($\beta = 1$)
- Falsidade → fé nula ($\alpha = 0$) e descrédito total ($\beta = 1$)
- Indefinição → fé nula ($\alpha = 0$) e nenhum descrédito ($\beta = 0$) .

No entanto, raramente α e β terão valores inteiros como os mostrados na listagem

acima: Na maioria das ocasiões, $0 \leqslant \alpha \leqslant 1$ e $0 \leqslant \beta \leqslant 1$. Por isso, se torna necessário o cálculo do **grau de certeza**, isto é, G_1, e do **grau de contradição**, isto é, G_2, conforme segue:

$$G_1 = \alpha - \beta \quad , \tag{3.7}$$

$$G_2 = \alpha + \beta - 1 \quad , \tag{3.8}$$

onde: $-1 \leqslant G_1$ e $1 \geqslant G_2$.

Os valores de G_1 e G_2, em conjunto, definem os graus entre verdade ($G_1 = 1$) e falsidade ($G_1 = -1$) e também os graus entre indefinição ($G_2 = -1$) e ambiguidade ($G_2 = 1$). Novamente, raramente tais valores inteiros serão alcançados já que G_1 e G_2 dependem de α e β.

O Plano Paraconsistente, para fins de visualização e maior rapidez na avaliação dos resultados, encontra-se ilustrado na Figura 3.8 e tem quatro arestas precisamente definidas:

- (-1,0) → falsidade;
- (1,0) → verdade;
- (0,-1) → indefinição;
- (0,1) → ambiguidade.

A propósito de ilustração na Figura 3.8, é possível ver um pequeno círculo indicando os graus dos quatro casos listados.

Para se ter ideia em que área exatamente se encontram as classes avaliadas, as distâncias (D) do ponto $P = (G_1, G_2)$ até o limites supracitados podem ser computadas. Tais cálculos podem ser feitos da seguinte forma:

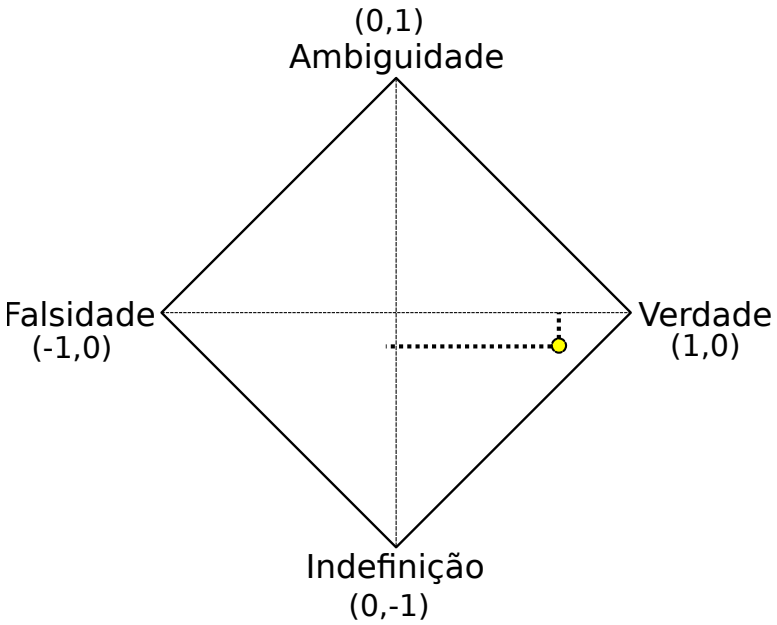
$$D_{-1,0} = \sqrt{(G_1 + 1)^2 + (G_2)^2} \quad , \tag{3.9}$$

$$D_{1,0} = \sqrt{(G_1 - 1)^2 + (G_2)^2} \quad , \tag{3.10}$$

$$D_{0,-1} = \sqrt{(G_1)^2 + (G_2 + 1)^2} \quad , \tag{3.11}$$

$$D_{0,1} = \sqrt{(G_1)^2 + (G_2 - 1)^2} \quad . \tag{3.12}$$

Figura 3.8: O plano paraconsistente: O pequeno círculo indica os graus de falsidade(-1,0), verdade(1,0), indefinição(0,-1) e ambiguidade(0,1)



Fonte: Adaptado de [Gui19].

Na prática, ou seja, para fins de classificação, geralmente considera-se a distância em relação ao ponto *"(1,0)* → *Verdade"*, que é o ponto ótimo: quanto mais próximo o ponto (G_1, G_2) estiver de $(1, 0)$, mais as os vetores de características das diferentes classes estão naturalmente separados. Isso implica, dentro da limitação de cada algoritmo, em resultados melhores sejam quais forem os classificadores usados.

3.1.6 Redes neurais

Autoencoders

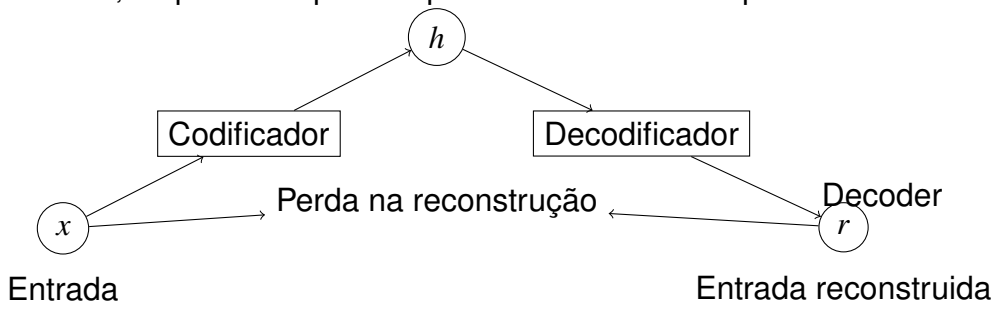
Como ilustrado na Figura 3.9 *Autoencoders* são redes neurais treinadas para reconstruir seus dados de entrada. Eles consistem em uma função codificadora, denotada como $h = f(x)$, e uma função decodificadora que produz uma reconstrução, denotada como

$r = g(h)$. A camada oculta h representa um código ou representação comprimida da entrada [GBC16b].

O principal objetivo de um *autoencoder* é aprender uma representação compactada dos dados de entrada na camada oculta e, em seguida, reconstruir os dados de entrada com a maior precisão possível usando o decodificador. No entanto, os *autoencoders* são projetados para serem incapazes de copiar perfeitamente os dados de entrada. Eles geralmente são limitados de alguma forma para apenas aproximar a entrada e priorizar certos aspectos dos dados.

Autoencoders podem ser treinados usando várias técnicas, como *gradient descent* com *minibatch* ou estocástico [GBC16b].

Figura 3.9: *Autoencoder*: x é codificado para uma dimensão menor h e, em seguida, é reconstruído em r, tal processo pode implicar ou não em uma perda na reconstrução

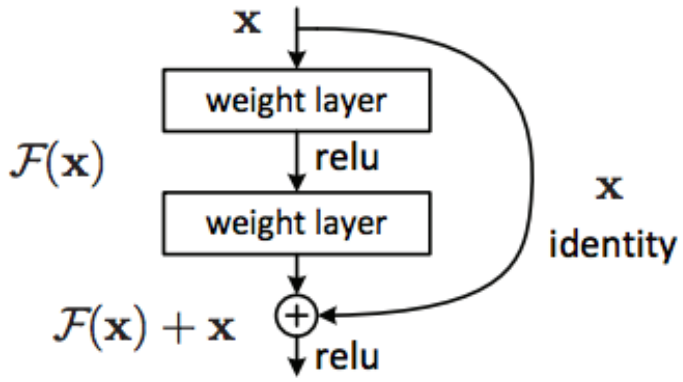


Considerando-se que a reconstrução r seja razoável, isso significa que a região h contém dados suficientes para representar a informação em sua essência, sendo assim, dentro do contexto das redes neurais, autoencoders são ótimos produtores de vetores de características.

Redes neurais residuais (ResNets)

Segundo [HZRS15] a ideia-chave por trás do *ResNets* é a inclusão de conexões de salto como ilustrado na Figura 3.10, também conhecidas como mapeamentos de identidade, que permitem que a saída de uma camada seja adicionada diretamente à entrada da

Figura 3.10: Bloco de uma Rede Neural Residual, X é uma função identidade que contorna as camada intermediárias criando "*highway connections*"



Fonte: [HZRS15]

camada subsequente. Isso contorna as camadas intermediárias e garante que redes mais profundas possam aprender. Outra vantagem no uso de conexões de salto é que essa prática diminui a ocorrência de *vanishing gradients* um problema comum em redes com muitas camadas que pode impossibilitar ou diminuir a níveis impraticáveis o aprendizado da rede.

Spike Neural Networks

Em [Kas18] se define uma *Spike Neural Network*,com um tipo de rede neural artificial inspirada na maneira como os neurônios se comunicam no cérebro por meio do uso de picos ou potenciais de ação. Nesse tipo de rede as informações são representadas pelo tempo e pela taxa de picos, e não pela força das conexões entre os neurônios.

Seu modo de operação pode ser altamente paralelo com, potencialmente, todos os neurônios operando em paralelo

Nas SNNs, o tempo dos picos e a dinâmica temporal da rede desempenham um papel crucial no processamento da informação pois cada pico carrega informações de tempo, permitindo uma codificação mais precisa de informações e comunicação assíncrona entre os neurônios.

Em se tratando do estudo atual e considerando a natureza temporal do sinais de voz e EEG, hipotéticamente, as SNNs são promissoras para classificação de tais informações.

3.2 Trabalhos correlatos

O estudo [GMJ+21] avaliou um grupo de 432 falantes da língua inglesa de variadas etnias e deficiências na fala coletando amostras e metainformações. Uma rede extratora de características composta por 8 camadas e uma rede *Long Short-Term Memory*(LSTM) classificadora de 2 foram treinadas com as falas. A entrada para o sistema ASR são as energias calculadas a partir de um banco de filtros para 80 dimensões. Os resultados superaram os transcritores humanos com ganhos médios e máximos de precisão de reconhecimento de 9% e 80%, respectivamente. A precisão dos modelos foi alta, com uma taxa de erro de palavras (WER) média de 4,6%

Em [JGX+21] se discute os desafios de reconhecer a fala disfônica e a importância das técnicas de aumento de dados no desenvolvimento de sistemas automáticos de reconhecimento de fala (ASR). Devido à complexidade das condições neuromotoras e deficiências físicas que acompanham a fala disfônica, é difícil coletar uma grande quantidade de amostras para treinar os sistemas ASR. O estudo propõe uma abordagem de aumento de dados usando redes adversariais generativas de convolução profunda (DCGAN) para modelar diferenças espectro-temporais detalhadas entre a fala disfônica e a fala normal. Experimentos realizados na base UASpeech demonstram que essa abordagem de aumento de dados supera consistentemente os métodos de aumento existentes baseados em perturbação de tempo ou velocidade, alcançando uma redução da taxa de erro de palavra (WER) de até 3,05% em comparação com o sistema de referência sem aumento de dados.

O estudo [HDTRGVP21] teve como objetivo desenvolver um sistema para detectar segmentos de palavras imaginadas em sinais EEG contínuos usando diferentes conjuntos de características e classificadores. Os pesquisadores testaram cinco conjuntos de características baseados em Transformada Discreta de Wavelet (DWT), Decomposição Empírica de Modos (EMD), características de energia, dimensões fractais e medidas de caos. Esses conjuntos de características foram usados para treinar quatro classificadores: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) e Logistic Re-

gression (LR). A avaliação de desempenho utilizou-se do *F1 score* [Tha20] e obteve um pontuação média de 0,75. Mais informações sobre o *F1 score* na Seço 5.1.

Aqui se apresenta um sistema de Reconhecimento Automático de Fala (ASR) que combina sinais de áudio e sinais de Eletroencefalograma (EEG) para aprimorar o reconhecimento de fala em sistemas de Interação Humano-Máquina (HMI) [MTG21]. O estudo explora o uso de múltiplas modalidades e aplica técnicas de Transformada Wavelet (WT) para extrair informações de fala dos sinais. Os resultados alcançam taxas de precisão de até 74,48%.

Nesta revisão [SAM+22] se examinou o uso de técnicas de inteligência artificial (IA) para decodificar a fala a partir de sinais cerebrais humanos, especificamente usando dados de eletroencefalografia (EEG). Os resultados da revisão indicaram o seguinte:

- Modalidade de Dados e Técnicas de IA: Os estudos analisaram o uso de dados de eletroencefalografia (EEG) e estímulos de palavras/frases. As técnicas de inteligência artificial (IA) utilizadas foram principalmente aprendizado de máquina e aprendizado profundo. Máquinas de vetores de suporte (SVM) e análise discriminante linear (LDA) foram comumente empregadas no aprendizado de máquina, enquanto redes neurais convolucionais (CNN) e redes neurais artificiais foram amplamente utilizadas no aprendizado profundo.

- Extração de Características e Processamento de Sinais: Devido ao ruído presente nos sinais de EEG, foram aplicadas técnicas de normalização e extração de características adequadas. A filtragem de banda passante, combinada com outras técnicas de normalização, foi frequentemente utilizada. As técnicas de extração de características incluíram padrões espaciais comuns, características estatísticas simples (como mínimo, máximo e média) e transformações discretas de wavelet.

- Conjunto de Dados e Equipamentos de Gravação: A maioria dos estudos utilizou dispositivos de EEG com 64 canais para capturar os sinais cerebrais, embora um

estudo tenha utilizado um dispositivo de 128 canais. Alguns estudos empregaram dispositivos com 32 canais ou menos.

O artigo [SSA23] propõe um sistema de verificação automática de falas (ASV) para pacientes com disartria usando recursos prosódicos (pitch, volume e probabilidade de vocalização) e aumento de dados fora do domínio. O estudo utilizou dois bancos de dados, a saber, o *Dyarthric Speech Database* (DSD) e o banco de dados *SpeechDat-Car*. Foram gerados vetores de características i-vector e x-vector usando MFCC (Mel-frequency cepstral coefficients), variáveis prosódicas e suas combinações. A combinação de MFCC, recursos de prosódia e aumento de dados produziu um EER de 11,09 para disartria leve, 13,26 na media e 11,97 para disartria grave.

O artigo [HDY+12] discute o uso de redes neurais profundas (DNNs) como criadores de modelos acústicos. Modelagem acústica é o processo de vinculação entre unidades linguísticas (como fonemas, palavras ou sentenças) e sinais de áudio. Neste artigo as DNNs são usadas para geração de vetores de características para posterior classificação usando *Hidden Markov Models* (HMM). Em relação a base dedos usada, a escolhida foi *TIMIT* que consiste em gravações de mais 630 falantes da língua inglesa. Essa combinação (DNN + HMM) atingiu uma WER de 18,5%.

Em [AFG+20] a base de dados *BREF* é composta por registros de fala francesa produzidos por 120 falantes, a mesma foi elaborada para fornecer falas contínuas para o desenvolvimento e avaliação de sistemas de Reconhecimento Automático de Fala e para modelagem de variação fonológica. Além dessa base uma própria (C2SI-LEC) contendo pacientes com falas disfuncionais também foi incluída. O artigo usou um modelo de Rede Neural Convolucional (CNN) para classificação. As características extraídas do sinal de fala foram os Coeficientes Cepstrais de Frequência Mel (MFCCs) e suas primeiras e segundas derivadas. Os dados de entrada foram normalizados subtraindo-se a média e dividindo-se pelo desvio padrão. Para lidar com a distribuição desproporcional das classes, uma técnica de subamostragem aleatória foi adotada durante a fase de treinamento. Uma taxa de

aprendizado inicial de 0,001 seguindo um cronograma de decaimento exponencial e uma estratégia de parada antecipada foi utilizada para o treinamento da CNN. O classificador acurácia de 0,68 na base *BREF* e 0,71 na base *C2SI-LEC*, os ouvintes humanos foram superados em ambas as bases.

No artigo [PD22], para obtenção dos dados foram usados 8 canais de EEG para medição dos sinais nos participantes. Múltiplas características foram extraídas simultaneamente desses sinais de EEG usando a transformadas Wavelets em cada um dos canais. Uma rede neural recorrente de memória de curto prazo (LSTM-RNN) foi usada para decodificar os sinais de EEG correspondentes a quatro comandos de áudio: para cima, para baixo, para a esquerda e para a direita. O artigo relata que o reconhecimento de padrões alcançou uma acurácia de classificação geral de 92,50%. Outras métricas como precisão, *recall* e *F1-score* também foram consideradas obtendo-se 92,74%, 92,50% e 92,62% respectivamente.

Em [CFC21] utilizou-se dados de EEG e espectroscopia de infravermelho próximo funcional (fNIRS) para coleta dos dados. Tais informações foram então separadas na categoria de tempo (média, variância, assimetria e curtose) e frequência (densidade espectral de potência e potência de banda). Os recursos extraídos foram então usados para treinar classificadores como o de análise discriminante linear (LDA), máquina de vetores de suporte (SVM) e uma rede neural convolucional (CNN). Alcançou-se uma precisão de classificação de 87,18% para fala aberta e de 53% para fala imaginada principalmente quando os estímulos foram imagens.

Neste estudo [BKEM22] registrou-se sinais de EEG correspondentes a fala imaginada de quatro vogais vindas de oito voluntários. Esse dados foram codificados em matrizes que representam a conectividade funcional entre diferentes regiões do cérebro durante a fala imaginada, de onde extraiu-se onze características afim de se detectar interações entre regiões com base no índice de localização. O índice de localização é definido como LI =

NS/NT, onde NS é o número de conexões significativas entre as regiões e NT é o número total de conexões entre as regiões. Os classificadores usados foram uma SVM e a Análise Discriminante Linear (LDA). A precisão média da classificação foi de 81,1%.

O artigo [LLL21] propõe uma estrutura baseada em aprendizado métrico profundo para decodificar a *imagined speech* usando interfaces cérebro-máquina (BCI). O método proposto foi avaliado em duas bases dados: o banco de dados *Coretto* contendo 6 classes e a base *BCI Competition* com 5 classes. Os sinais de EEG foram medidos durante a fala imaginada, e a frequência instantânea e entropia espectral foram extraídas dos sinais. A estrutura proposta usa uma rede neural siamesa que aprende a perda contrastante com base na distância determinada por uma estrutura de aprendizagem métrica profunda. O classificador obteve um precisão dentre 6 classes de 45,00 ± 3,13% e uma entre de 5 classes de 48,10 ± 3,68%.

Aqui [TMM20] se usou um conjunto de dados EEG com 15 sujeitos imaginando dizer cinco vogais (a, e, i, o, u) e seis palavras diferentes. Após a coleta, visando melhorar a etapa de classificação, os dados passaram por uma etapa de pré-processamento que incluiu os seguintes procedimentos:

- filtragem passa-faixa entre $0,5Hz$ e $100Hz$ para remover quaisquer frequências indesejadas.
- subamostragem para 100 Hz afim de reduzir a carga computacional.
- segmentação em lotes de 2 segundos, com uma sobreposição de 1 segundo entre os sinais.
- normalização para média zero e variância unitária.

Em se tratando de classificadores foi usada uma rede neural convolucional com transferência de aprendizado para classificar os sinais correspondentes às falas imaginadas atingindo uma acurácia de 23.98% (±3.08).

Segundo [AKY+18] *Hashing* sensível a locus(LSH) é frequentemente usado como um classificador para problemas relacionados a *big data*, neste trabalho é proposto uma junção de MFCC e LSH a fim de se reconhecer o locutor. Neste método o MFCC é extraído dos arquivos de sinal para posterior aplicação do LSH gerando assim uma tabela *hash*, estes valores de *hash* são então comparados identificando assim o locutor ou locutora. Nos testes realizados houve uma acurácia de 92,66%. A base de dados usada foi a *TIMIT 2018* [Con18].

3.2.1 Fechamento

Com base nos estudos mencionados nesta pesquisa, pretende-se utilizar uma camada extratora de características semelhante à abordada por [GMJ+21], bem como a aplicação de filtros wavelet packet para a decomposição dos sinais. Além disso, considera-se a possibilidade de recorrer a técnicas de aumento de dados, como aquelas empregadas por [JGX+21], devido à escassez de locutores disfônicos.

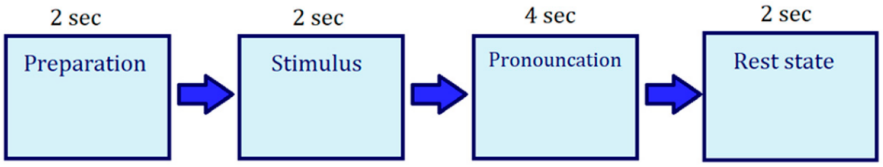
No estudo de [TMM20], é apresentado um protocolo para a obtenção de dados (Figura 3.11), o qual pode ser utilizado diretamente ou adaptado para esta pesquisa. Nesse mesmo estudo, menciona-se uma base de dados chamada UASpeech, que pode fornecer informações valiosas para a verificação de padrões.

Dentre os artigos revisados, [HDTRGVP21] parece ter a menor relevância em relação às ideias abordadas, uma vez que o método de extração de dados já está definido, ao contrário, o trabalho de [MTG21] se aproxima mais da proposta desta pesquisa, uma vez que utiliza diretamente as Transformadas Wavelet para a criação dos vetores de características.

No estudo de [SAM+22], são discutidas técnicas interessantes de normalização de dados que talvez possam ser utilizadas.

Embora os coeficientes MFCC (Mel-frequency cepstral coefficients), mencionados em [Fur21] e [SSA23], não tenham apresentado melhorias nos vetores de características, a utilização de técnicas de aumento de dados fora do domínio, mencionada em [SSA23], pode ser útil.

Figura 3.11: Protocolo para a obtenção de dados.



Fonte: [TMM20]

Embora o artigo citado [HDY+12] não se aplique diretamente ao problema desta tese, ele apresenta o conceito de Modelagem Acústica.

Já em [BKEM22] o conceito de matrizes que representam a conectividade funcional entre diferentes regiões do cérebro durante a fala imaginada parece promissor e talvez possa ajudar a produzir novas intuições acerca dos padrão cerebrais dado que, segundo o próprio estudo, diferentes regiões do cérebro são ativadas durante as falas imaginadas.

Outro estudo relevante é o de [LLL21], que propõe uma estrutura baseada em aprendizado métrico profundo para decodificar a fala imaginada. No entanto, essa abordagem difere significativamente da proposta desta pesquisa.

Em resumo, nenhum dos artigos analisados utiliza as técnicas propostas neste estudo, tais como autoencoders, redes neurais residuais, spike neural networks e vetores de características "hand-crafted". Além disso, nenhum deles tem como objetivo comparar esses métodos, o que torna esta tese potencialmente relevante para a comunidade acadêmica.

Capítulo 4

Cronograma para conclusão do doutorado

Até a presente data foram realizados os levantamentos bibliográficos e o cumprimento dos créditos necessários em disciplinas para o Doutorado.

Quanto aos trabalhos futuros os mesmo estão descritos na Tabela 4, porém primeiramente há que se separar o projeto em fases.

A metodologia descrita é dividida em cinco etapas (E1 a E5) e tem como objetivo realizar um estudo sobre a aquisição e análise de *imagined speech* por meio de eletroencefalografia (EEG) em locutores com dificuldades na fala.

Na etapa E1, foi realizado um levantamento bibliográfico sobre o estado-da-arte nas sub-áreas de aquisição e análise de *imagined speech* por meio de EEG e de reconhecimento de voz em pacientes com disfonia. Foram estudados trabalhos publicados nas referências citadas na Seção 3.2 e também cumpridos os créditos necessários das disciplinas do doutorado.

Na etapa E2, além de diálogos com profissionais das áreas de fonoaudiologia e neurologia. Também será estudada uma base de dados pública descrita em [CGR17] para compreender os padrões comumente encontrados nos sinais de interesse. Por fim, será criada uma base de dados própria contendo sinais de voz verbalizados por locutores com dificuldades na fala, bem como sinais de EEG que possam conter representações das res-

pectivas *imagined speeches*. Será definido um conjunto de frases foneticamente ricas que serão lidas em voz alta pelos locutores voluntários. Os sinais acústicos serão capturados com um gravador eletrônico, enquanto os sinais de EEG serão capturados com eletrodos posicionados sobre o couro cabeludo dos voluntários.

A etapa E3 envolverá o pré-processamento e a extração de características representativas para o objetivo proposto. Serão realizadas normalizações clássicas nos sinais, seguidas pela extração de características através de abordagens handcrafted e feature learning. A abordagem handcrafted envolverá a escolha de descritores guiados pela engenharia paraconsistente de características. Já a abordagem de feature learning investigará arquiteturas de redes neurais artificiais profundas projetadas para autocodificação. Serão extraídas características tanto dos sinais de voz quanto das *imagined speeches*.

A etapa E4 terá como foco a integração das etapas de extração de características (EC) e de autenticação. Serão empregadas estruturas de aprendizado profundo, como RNNs (Redes Neurais Recorrentes) e DSNNs (Deep Stacked Neural Networks), para o reconhecimento de padrões. Os resultados serão mensurados e comparados em termos de acurácia, levando em conta também o tempo de treinamento. Será realizada uma comparação entre as modalidades *text-dependent* e *text-independent*.

A etapa E5 consistirá na apresentação dos experimentos em eventos científicos e na publicação dos resultados em periódicos de impacto. Também serão confeccionadas a monografia de qualificação e a tese de doutorado ao longo de todo o trabalho.

Ms/Ano	E1	E2	E3	E4	E5
09/22	X				
10/22	X				
11/22	X				
12/22	X				
01/23	X	X			
02/23	X	X			
03/23	X	X			
04/23	X	X			
05/23	X	X			X
06/23	X	X			X
07/23		X	X		X
08/23		X	X		X
09/23		X	X		
10/23		X	X		
11/23		X	X		
12/23		X	X		
01/24		X	X	X	X
02/24		X	X	X	X
03/24			X	X	
04/24			X	X	
05/24			X	X	
06/24			X	X	
07/24			X	X	
08/24			X	X	

Mês/Ano	E1	E2	E3	E4	E5
09/24			X	X	
10/24			X	X	
11/24			X	X	
12/24			X	X	
01/25			X	X	
02/25			X	X	
03/25			X	X	
04/25			X	X	
05/25			X	X	
06/25			X	X	
07/25			X	X	
08/25			X	X	
09/25			X	X	
10/25			X	X	
11/25			X	X	
12/25			X	X	
01/26			X	X	
02/26			X	X	
03/26			X	X	
04/26			X	X	
05/26					X
06/26					X
07/26					X
08/26					X

Tabela 4.1: Cronograma

Capítulo 5

Apêndice

5.1 *F1 score*

O *F1 score* é uma medida de desempenho utilizada em tarefas de classificação para avaliar o equilíbrio entre precisão e *recall*. É uma métrica única que combina as duas medidas em um único valor, frequentemente usada quando lidamos com conjuntos de dados desbalanceados, nos quais a distribuição das classes é desigual.

Para entender o *F1 score*, vamos primeiro definir precisão e *recall*:

- Precisão: Mede a proporção de instâncias positivas corretamente previstas (verdadeiros positivos) em relação a todas as instâncias previstas como positivas (verdadeiros positivos mais falsos positivos).
- *Recall*: Mede a proporção de instâncias positivas corretamente previstas (verdadeiros positivos) em relação a todas as instâncias positivas reais (verdadeiros positivos mais falsos negativos). O *recall* representa a capacidade do modelo de identificar todas as instâncias positivas.

Considerando isso o *F1 score* é dados pela equação 5.1:

$$F1\ score = 2 \times \frac{precisão \times recall}{precisão + recall} \tag{5.1}$$

Essa métrica é considerada uma média harmônica, sendo assim, dá mais peso a valores mais baixos, fornecendo uma medida equilibrada que considera tanto a precisão

quanto o *recall*.

Bibliografia

[AFG+20] Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Christine Meunier, and Virginie Woisard. Towards Interpreting Deep Learning Models to Understand Loss of Speech Intelligibility in Speech Disorders — Step 1: CNN Model-Based Phone Classification. In *Proc. Interspeech 2020*, pages 2522–2526, 2020.

[AKY+18] A. Awais, S. Kun, Y. Yu, S. Hayat, A. Ahmed, and T. Tu. Speaker recognition using mel frequency cepstral coefficient and locality sensitive hashing. In *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 271–276, May 2018.

[AWG09] P. S. Addison, J. Walker, and R. C. Guido. Time–frequency analysis of biosignals. *IEEE Engineering in Medicine and Biology Magazine*, 28(5):14–29, Sep 2009.

[BB11] Homayoon Beigi and Homayoon Beigi. *Speaker recognition*. Springer, 2011.

[Ber49] Leo L. Beranek. *Acoustic Measurements*. J. Wiley, 1949.

[BG02] Marina Bosi and Richard E Goldberg. *Introduction to digital audio coding and standards*, volume 721. Springer Science & Business Media, 2002.

[Bia07] G. Bianchi. *Electronic Filter Simulation & Design*. McGraw-Hill Education, 2007.

[BK10] Katharine Brigham and BVK Vijaya Kumar. Imagined speech classification with eeg signals for silent communication: a preliminary investigation into synthetic telepathy. In *2010 4th International Conference on Bioinformatics and Biomedical Engineering*, pages 1–4. IEEE, 2010.

[BKEM22] Mohamad Amin Bakhshali, Morteza Khademi, and Abbas Ebrahimi-Moghadam. Investigating the neural correlates of imagined speech: An eeg-based connectivity analysis. *Digital Signal Processing*, 123:103435, 2022.

[But30] S. Butterworth. On the theory of filters amplifiers. 1930.

[CA00] Newton C.A. da Costa and Jair Minoro Abe. Paraconsistência em informática e inteligência artificial. *Estudos Avançados*, 14:161 – 174, 08 2000.

[CFC21] Ciaran Cooney, Raffaella Folli, and Damien Coyle. A bimodal deep learning architecture for eeg-fnirs decoding of overt and imagined speech. *IEEE Transactions on Biomedical Engineering*, 69(6):1983–1994, 2021.

[CGR17] Germán A. Pressel Coretto, Iván E. Gareis, and H. Leonardo Rufiner. Open access database of eeg signals recorded during imagined speech. 2017.

[Con18] Linguistic Data Consortium. Timit acoustic-phonetic continuous speech corpus, 2018.

[CSBY22] Mounira Chaiani, Sid Ahmed Selouani, Malika Boudraa, and Mohammed Sidi Yakoub. Voice disorder classification using speech enhancement and deep learning models. *Biocybernetics and Biomedical Engineering*, 42(2):463–480, 2022.

[Dau92] I. Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1992.

[dCBB98] N.C.A. da Costa, J.Y. Béziau, and O. Bueno. *Elementos de teoria paraconsistente de conjuntos*. Coleção CLE. Centro de Lógica, Epistemologia e História da Ciência, UNICAMP, 1998.

[DPBATRT14] Marcos Del Pozo-Banos, Jesús B Alonso, Jaime R Ticay-Rivas, and Carlos M Travieso. Electroencephalogram subject identification: A review. *Expert Systems with Applications*, 41(15):6537–6554, 2014.

[FKO+22] Shintaro Fujimura, Tsuyoshi Kojima, Yusuke Okanoue, Kazuhiko Shoji, Masato Inoue, Koichi Omori, and Ryusuke Hori. Classification of voice disorders using a one-dimensional convolutional neural network. *Journal of Voice*, 36(1):15–20, 2022.

[FPM+17] Everthon Silva Fonseca, Denis César Mosconi Pereira, Luís Fernando Castilho Maschi, Rodrigo Capobianco Guido, and Katia Cristina Silva Paulo. Linear prediction and discrete wavelet transform to identify pathology in voice signals. In *2017 Signal Processing Symposium (SPSympo)*, pages 1–4. Ieee, 2017.

[Fre13] Susana Freitas. Avaliação acústica e áudio percetiva na caracterização da voz humana. 2013.

[Fur21] André Furlan. Caracterização de voice spoofing para fins de verificação de locutores com base na transformada wavelet e na análise paraconsistente

de características. Dissertação de mestrado, Universidade Estadual Paulista - campus de São José do Rio Preto-SP, São José do Rio Preto, Brazil, 2021. Orientador: Prof Dr Rodrigo Capobianco Guido.

[GBC16a] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[GBC16b] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[GMJ+21] Jordan R. Green, Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, and Katrin Tomanek. Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases. In *INTERSPEECH 2021*, Interspeech, pages 4778–4782, 2021. Interspeech Conference, Brno, CZECH REPUBLIC, AUG 30-SEP 03, 2021.

[GPP+21] Siddhant Gupta, Ankur T Patil, Mirali Purohit, Mihir Parmar, Maitreya Patel, Hemant A Patil, and Rodrigo Capobianco Guido. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks*, 139:105–117, 2021.

[Gui19] R. C. Guido. Paraconsistent feature engineering [lecture notes]. *IEEE Signal Processing Magazine*, 36(1):154–158, Jan 2019.

[HDTRGVP21] Tonatiuh Hernandez-Del-Toro, Carlos A. Reyes-Garcia, and Luis Villasenor-Pineda. Toward asynchronous eeg-based bci: Detecting imagined words segments in continuous eeg signals. *BIOMEDICAL SIGNAL PROCESSING AND CONTROL*, 65, MAR 2021.

[HDY+12] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[HH15] John HL Hansen and Taufiq Hasan. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine*, 32(6):74–99, 2015.

[HM11] Simon Haykin and Michael Moher. *Sistemas de Comunicação-5*. Bookman Editora, 2011.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[JCA16] Isuru Jayarathne, Michael Cohen, and Senaka Amarakeerthi. Brainid: Development of an eeg-based biometric authentication system. In *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 1–6. IEEE, 2016.

[JCA17] Isuru Jayarathne, Michael Cohen, and Senaka Amarakeerthi. Survey of eeg-based biometric authentication. In *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, pages 324–329. IEEE, 2017.

[JGX+21] Zengrui Jin, Mengzhe Geng, Xurong Xie, Jianwei Yu, Shansong Liu, Xunying Liu, and Helen Meng. Adversarial Data Augmentation for Disordered Speech Recognition. In *Proc. Interspeech 2021*, pages 4803–4807, 2021.

[JlCH01] Arne Jensen and Anders la Cour-Harbo. *Ripples in Mathematics*. Springer Berlin Heidelberg, 2001.

[Jol02] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.

[Kas18] N.K. Kasabov. *Time-Space, Spiking Neural Networks and Brain-Inspired Artificial Intelligence*. Springer Series on Bio- and Neurosystems. Springer Berlin Heidelberg, 2018.

[KG14] Robinson Luis Kremer and ML d C GOMES. A eficiência do disfarce em vozes femininas: uma análise da frequência fundamental. *ReVEL*, 12:23, 2014.

[LBBPM22] Diego Lopez-Bernal, David Balderas, Pedro Ponce, and Arturo Molina. A state-of-the-art review of eeg-based imagined speech decoding. *Frontiers in Human Neuroscience*, 16, 2022.

[LHA05a] Francois Le Huche and André Allali. A voz: patologia vocal de origem funcional. In *A voz: patologia vocal de origem funcional*, pages 187–187. 2005.

[LHA05b] François Le Huche and André Allali. A voz: patologia vocal de origem orgânica. In *A voz: patologia vocal de origem orgânica*, pages 154–154. 2005.

[LLL21] Dong-Yeon Lee, Minji Lee, and Seong-Whan Lee. Decoding imagined speech based on deep metric learning for intuitive bci communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1363–1374, 2021.

[LSLR20] Kong Aik Lee, Seyed Omid Sadjadi, Haizhou Li, and Douglas A Reynolds. Two decades into speaker recognition evaluation-are we there yet? *Comput. Speech Lang.*, 61:101058, 2020.

[MM18] Luis Alfredo Moctezuma and Marta Molinas. Eeg-based subjects identification based on biometrics of imagined speech using emd. In *Brain Informatics: International Conference, BI 2018, Arlington, TX, USA, December 7–9, 2018, Proceedings 11*, pages 458–467. Springer, 2018.

[MPP21] Ioanna Miliaresi, Kyriakos Poutos, and Aggelos Pikrakis. Combining acoustic features and medical data in deep learning networks for voice pathology classification. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1190–1194. IEEE, 2021.

[MS21] Vikas Mittal and RK Sharma. Deep learning approach for voice pathology detection and classification. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 16(4):1–30, 2021.

[MTG21] P. P. Mini, Tessamma Thomas, and R. Gopikakumari. Wavelet feature selection of audio and imagined/vocalized eeg signals for ann based multimodal asr system. *BIOMEDICAL SIGNAL PROCESSING AND CONTROL*, 63, JAN 2021.

[MTGVPC19] Luis Alfredo Moctezuma, Alejandro A Torres-García, Luis Villaseñor-Pineda, and Maya Carrillo. Subjects identification using eeg-recorded imagined speech. *Expert Systems with Applications*, 118:201–208, 2019.

[NP12] Amy Neustein and Hemant A Patil. *Forensic speaker recognition*, volume 1. Springer, 2012.

[P+96] Robi Polikar et al. The wavelet tutorial, 1996.

[PD22] Dipti Pawar and Sudhir Dhage. Wavelet-based imagined speech classification using electroencephalography. *International Journal of Biomedical Engineering and Technology*, 38(3):215–224, 2022.

[RBJL16] Maria V Ruiz-Blondet, Zhanpeng Jin, and Sarah Laszlo. Cerebre: A novel method for very high accuracy event-related potential biometric identification. *IEEE Transactions on Information Forensics and Security*, 11(7):1618–1629, 2016.

[RG21] Ana-Luiza Rusnac and Ovidiu Grigore. Eeg preprocessing methods for bci imagined speech signals. In *2021 International Conference on e-Health and Bioengineering (EHB)*, pages 1–4. IEEE, 2021.

[SAM+22] Uzair Shah, Mahmood Alzubaidi, Farida Mohsen, Alaa Abd-Alrazaq, Tanvir Alam, and Mowafa Househ. The role of artificial intelligence in decoding speech from eeg signals: A scoping review. *SENSORS*, 22(18), SEP 2022.

[SMB07] D. Salomon, G. Motta, and D. Bryant. *Data Compression: The Complete Reference*. Molecular biology intelligence unit. Springer London, 2007.

[SSA23] Shinimol Salim, Syed Shahnawazuddin, and Waquar Ahmad. Automatic speaker verification system for dysarthric speakers using prosodic features and out-of-domain data augmentation. *Applied Acoustics*, 210:109412, 2023.

[Tha20] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192, 2020.

[TMM20] Markus-Oliver Tamm, Yar Muhammad, and Naveed Muhammad. Classification of vowels from imagined speech with convolutional neural networks. *Computers*, 9(2):46, 2020.

[V+14] Eugênia Hermínia Oliveira Valença et al. Análise acústica dos formantes em indivíduos com deficiência isolada do hormônio do crescimento. 2014.

[WMWX22] Xingmei Wang, Jiaxiang Meng, Bin Wen, and Fuzhao Xue. Racp: A network with attention corrected prototype for few-shot speaker recognition using indefinite distance metric. *Neurocomputing*, 490:283–294, 2022.

[WSA05] Haydée F Wertzner, Solange Schreiber, and Luciana Amaro. Análise da freqüência fundamental, jitter, shimmer e intensidade vocal em crianças com transtorno fonológico. *Revista Brasileira de Otorrinolaringologia*, 71:582–588, 2005.

[Zwi61] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248–248, 1961.