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

Professor Dr. Aleardo Manacero Júnior
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 ????? trabalhos científicos na área de *imagined speech* e processamento de sinais de locutores disfônicos, 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. Iniciando as revisões de conceitos apresenta-se a engenharia paraconsistente de características, filtros digitais *wavelets*, amostragem, quantização e caracterização dos processos de produção da voz humana no capítulo 2 seção 1, a seguir, na seção 2 deste mesmo capítulo apresenta-se a revisão bibliográfica inicial. Finalmente no capítulo 3 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 [GBC16], 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 doutorado, 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. Os detalhes estão descritos neste documento, que está organizado da seguinte forma: os objetivos gerais e específicos são apresentados na Seção 2, a metodologia e o cronograma são descritos na Seção 3, os resultados esperados são discutidos na Seção 4, e a relevância e o impacto deste projeto são abordados na Seção 5.

Capítulo 2

Revisão de Bibliográfica

2.1 Conceitos utilizados

2.1.1 Engenharia paraconsistente de características

Dentro do processo de classificação frequentemente surge a questão:
Os vetores de características criados proporcionam uma boa separação de classes?

O método de cálculo do plano paraconsistente é 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) onde n é o índice de cada uma delas. 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 será 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, β mostra a razão de sobreposição entre diferentes classes. Idealmente α deve ser maximizada e β minimizada para um desempenho ótimo dos classificadores.

Inicialmente é necessária a normalização dos vetores de características de forma que a soma de todos os seus valores seja um.

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 de cada classe gerando assim um vetor para os valores maiores e outro para os menores.

O **vetor de similaridade**(svC_n) é obtido fazendo-se a diferença item a item dos maiores em relação aos menores.

Finalmente e para cada classe é tirada a média dos valores de cada vetor de similaridade, α é o menor valor dentre essas médias.

A figura **??** ilustra este processo.

A obtenção de β, assim como ilustrado na figura **??**, também se dá selecionando-se 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 se segue com o cálculo de R cujo valor é a quantidade de vezes que um valor do vetor de características de uma classe se encontra no intervalo de valores maiores e menores de outra classe.

É necessário o cálculo de F que é o número máximo de sobreposições possíveis entre classes e é dado por:

$$F = N.(N-1).X.T \tag{2.1}$$

onde:

- N é a quantidades de classes.
- X é quantidade de vetores de características por classe.
- T é o tamanho do vetor de características.

Finalmente, β é calculado:

$$\beta = \frac{R}{F} \tag{2.2}$$

Nesse 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 [**?**].

- 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 descrédito total ($\beta = 0$)

No entanto, raramente α e β terão tais valores, na maioria do tempo $0 \leqslant \alpha \leqslant 1$ e $0 \leqslant \beta \leqslant 1$, por isso, se torna necessário o cálculo do **grau de certeza**(G_1) e do **grau de contradição**(G_2).

$$G_1 = \alpha - \beta \tag{2.3}$$

$$G_2 = \alpha + \beta - 1 \qquad , \tag{2.4}$$

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 e falsidade, ou seja, $G_1 = -1$ e $G_1 = 1$ respectivamente e também os graus entre indefinição e ambiguidade, ou seja, $G_2 = -1$ e $G_2 = 1$ respectivamente.

O plano paraconsistente para fins de visualização e maior rapidez na avaliação dos resultados como ilustrado na figura **??** tem quatro cantos definidos:

- (-1,0) → Falsidade.
- (1,0) → Verdade.
- (0,-1) → Indefinição.
- (0,1) → Ambiguidade.

É importante perceber que na figura **??** existe um pequeno círculo, este indica onde se encontram as classes nos graus explicados da listagem anterior.

Para se ter ideia em que área exatamente se encontram as classes avaliadas se deve calcular as distâncias(D) do ponto $P = (G_1, G_2)$ dos limites supracitados. Tal cálculo pode ser feito da seguinte forma:

$$D_{-1,0} = \sqrt{(G_1 + 1)^2 + (G_2)^2} \quad , \tag{2.5}$$

$$D_{1,0} = \sqrt{(G_1 - 1)^2 + (G_2)^2} \quad , \tag{2.6}$$

$$D_{0,-1} = \sqrt{(G_1)^2 + (G_2 + 1)^2} \quad , \tag{2.7}$$

$$D_{0,1} = \sqrt{(G_1)^2 + (G_2 - 1)^2} \quad , \tag{2.8}$$

2.1.2 Filtros digitais wavelets

Filtros digitais *wavelet* vem para suprir as deficiências de janelamento de sinal apresentadas pelas transformadas de Fourier e pelas transformadas curtas de Fourier. *Wavelets* contam com variadas funções filtro e tem tamanho de janela variável o que permite uma análise multirresolução [**?**].

As *wavelets* proporcionam a análise do sinal de forma detalhada tanto no espectro de baixa frequência quanto no de alta frequência.

É importante observar que, quando se trata de transformada *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 então os filtros de síntese, a função *scaling* e a função *wavelet* não serão elementos abordados aqui, pois, esses só interessariam caso houvesse a necessidade da transformada inversa. A abordagem usada será baseada nos filtros de análise digitais que proporcionarão a decomposição do sinal com o uso de filtros passa baixas e passa altas estritamente no domínio discreto.

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 se diz que a janela tem **um suporte compacto** [**?**].

Se diz que uma *wavelet* tem boa **resposta em frequência** quando, na aplicação da mesma na filtragem das frequências não são causadas muitas pertubações indesejadas ao sinal, as wavelets de Daubechies se destacam neste quesito por serem *maximamente*

planas (Maximally-flat) nos platôs de resposta em frequência como indicado na figura 2.1.

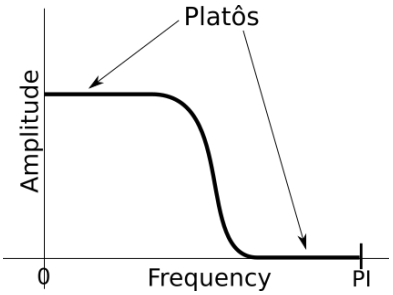

Figura 2.1: Platôs maximamente planos em um filtro digital

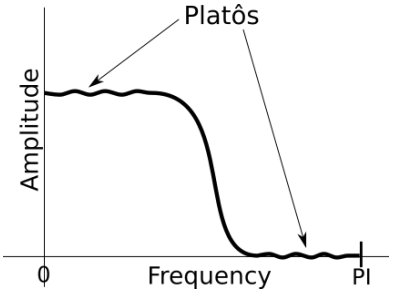

Figura 2.2: Platôs não maximamente planos de um filtro digital

Além da resposta em frequência a aplicação de um filtro digital *wavelets* também pode gerar o que se chama de **resposta em fase**, 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 formantes do sinal.

Idealmente é desejável que todo filtro apresente boa resposta em frequência e resposta em fase linear.

Wavelet	Resposta em frequência	Resposta em fase
Haar	Pobre	Linear
Daubechies	Quanto maior o suporte, melhor. *Maximally-flat*	Não linear
Symmlets	Quanto maior o suporte, melhor. Não *Maximally-flat*	Quase linear
Coiflets	Quanto maior o suporte, melhor. Não *Maximally-flat*	Quase linear

Tabela 2.1: Algumas wavelets mais populares e suas propriedades

O algoritmo de Malat

O algoritmo de Malat torna aplicação das *wavelets* no sinal em uma simples multiplicação de matrizes, o sinal que deve ser transformado se torna uma matriz linear vertical já os filtros passa-baixa e passa-alta tornam-se, nessa ordem, linhas de uma matriz quadrada que será completada segundo regras que serão mostradas mais adiante.

É importante que essa matriz quadrada tenha de aresta a mesma quantidade de itens que o nosso sinal, ou seja, se o sinal tem quatro elementos então a matriz de filtros deve ser uma de 4x4.

Algo interessante a se notar é que, para que seja possível a transformada wavelet, basta ter disponível o filtro passa-baixa construído a partir da *mother wavelet* já que o filtro passa-alta pode ser construído a partir da ortogonalidade do primeiro.

A título de exemplo considere:

O filtro passa baixa baseado na wavelet Haar: $h[\cdot] = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$,

E seu respectivo valor ortogonal: $g[\cdot] = [\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}]$,

Considere também o seguinte sinal: $sinal = [1, 2, 3, 4]$.

Se o tamanho do sinal a ser tratado é quatro, ou seja, o sinal tem quatro pulsos, e se pretende-se aplicar o filtro Haar, a seguinte matriz é construída:

$$\begin{pmatrix} \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0 \\ \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0 \\ 0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \\ 0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \end{pmatrix} \tag{2.9}$$

No entanto, filtros Haar tem apenas dois valores e, necessariamente, a linha da

matriz deve ter quatro itens. Para resolver este problema basta completar cada uma das linhas com zeros. A matriz é montada de forma que a mesma seja ortogonal.

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

$$\begin{pmatrix} \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0 \\ \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0 \\ 0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \\ 0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \end{pmatrix} . \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} \frac{3}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \\ \frac{7}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix} \tag{2.10}$$

realizada a multiplicação é necessário agora montar o sinal filtrado, isso é feito escolhendo, dentro do resultado, valores alternadamente de forma que o vetor resultante seja:

$$resultado = \left[\frac{3}{\sqrt{2}}, \frac{7}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, \frac{-1}{\sqrt{2}} \right] \quad . \tag{2.11}$$

2.1.3 Amostragem, quantização e o formato do arquivo Wave

Serão usados arquivos no formato *wave* usando *pulse-code modulation* (PCM), neste esquema os dados são armazenados sem perdas. O arquivo, segundo [**?**], se estrutura como o ilustrado na figura **??**.

A taxa de amostragem de 44100hz permite, segundo o teorema de Nyquist, que seja realizada a quantização de frequências de até 22050hz a uma resolução de 16bits.

A estrutura de interesse se localiza na última parte do arquivo, mais especificamente no bloco "data", aqui os dados são organizados como um grande vetor de números, cada um deles, indicando a intensidade do sinal naquele ponto.

2.1.4 Caracterização dos processos de produção da voz humana

A fala possui três grandes áreas de estudo: fisiológica (ou fonética articulatória), acústica (ou fonética acústica) e perceptual (ou comumente chamada percepção da fala) [**?**].

Neste trabalho o foco será apenas na acústica, já que não serão analisados aspectos da fisiologia relacionada a voz e sim o sinal sonoro propriamente dito.

Vozeada versus não-vozeada

Quando da análise de voz se pode levar em consideração as partes vozeadas e/ou não-vozeadas do sinal. As partes vozeadas são aquelas produzidas com ajuda das pregas vocais, as partes não-vozeadas não tem participação desta 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 (pitch) [**?**].

Vozes agudas tem um pitch alto, vozes mais graves tem um pitch baixo, a alteração do pitch durante a fala é definido como entonação.

A frequência fundamental da voz é a velocidade na qual uma forma de onda se repete por unidade de tempo, ou seja, o número de ciclos vibratórios produzidos pelas pregas vocais, num segundo, sendo assim, as medidas de f_0 geralmente são apresentadas em Hz [**?**].

A medição de f_0 está sujeita a contaminações surgidas das variações naturais de *pitch* tipicas da voz humana [**?**].

A importância de se medir f_0 corretamente vem do fato de que, além de carregar boa parte da informação da fala, f_0 é a base para construção das outras frequências, pois essas são múltiplas de f_0.

Formantes

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 cavida-

des à frente e atrás do ápice da língua; o quarto formante (f_4), ao formato da laringe e da faringe na mesma altura [**?**].

2.2 Trabalhos correlatos

O estudo [GMJ+21] avaliou um grupo de 432 falantes da língua inglesa de variadas etnias e deficiências na fala forneceu amostras e metainformações. Aqui foram avaliadas a precisão do reconhecimento automático de fala (ASR) em vocabulário aberto. Os modelos tiveram um desempenho melhor do que os transcritores humanos especialistas e dois modelos ASR independentes. 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%, melhor do que a dos modelos independentes (WER média de 31%). As melhorias mais significativas foram observadas nos falantes com disfonias mais graves.

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 Regression (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 seção 4.1.

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 caracterís-

ticas 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. Recomendou-se o uso de conjuntos de dados maiores para melhorar o desempenho dos modelos, mas a disponibilidade de conjuntos de dados EEG publicamente acessíveis para decodificação da fala é limitada devido a preocupações com privacidade e segurança.

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%.

———-

No artigo [**?**] foi apresentado um esquema de diferenciação entre a fala comum e aquela vinda de um dispositivo reprodutor. O foco da análise se dá na distorção causada pelo alto-falante segundo a energia e outras várias características do espectro do sinal. Uma base com 771 sinais de fala foi criada para cada um dos quatro dispositivos de gravação usados totalizando 3084 trechos de áudio. Uma *support vector machine* (SVM) foi usada como classificador. De acordo com os experimentos a *taxa de verdadeiros positivos* é de 98,75% e a *taxa de verdadeiros negativos* é de 98,75%.

Em [**?**] é mostrado um método para diferenciar a voz de um locutor verdadeiro da voz gerada por sistemas usando sintetizadores baseados no *modelo oculto de Markov* (HMM). SAS[**?**] foi a escolha de base de dados. Este método usa coeficientes de características logarítmicos extraídos de wavelets que são apresentados a um classificador SVM. Os resultados obtidos tiverem, em média, mais de 99% de acurácia.

Usando uma decomposição por espalhamento baseada em wavelets e convertendo o resultado em coeficientes cepstrais (SCCs) o artigo [**?**] cria um vetor de características que é avaliado por modelos de mistura Gaussiana (GMM). SAS e ASVspoof 2015 [**?**] foram as bases de dados escolhidas para testes. Em relação aos resultados foram usadas a *taxa de falsos verdadeiros* (FAR) que representa a taxa de ocorrências falsas classificadas

como verdadeiras e a *taxa de falsos falsos* (FRR) que é a taxa de ocorrências verdadeiras classificadas como falsas. Aos pontos em que FAR é igual a FRR chamou-se de pontos de taxa de erros iguais e a *taxa de erros iguais* (ERR) é o valor de $\frac{FAR}{FRR}$. Considerando isso nos experimentos foi obtida uma ERR geral de 0,18.

Já em [**?**] os autores usam o *"Zero time windowing"* ou janelamento de tempo zero (ZTW), conceito esse que deve ser melhor entendido durante a confecção da dissertação, para, em conjunto com a análise cepstral do espectro gerado, fazer a análise do sinal. Os experimentos foram feitos usando-se a base ASVspoof 2017[**?**] com um classificador GMM, a taxa geral de ERR dos experimentos foi de 0,1475.

Em [**?**] é citado que existe uma diferença entre as propriedades espectrais da voz original e da voz gravada. No escrito são usados coeficientes cepstrais sobre os quais são aplicados uma média e uma normaliza de variância para diminuir o impacto dos ruídos na classificação. Uma GMM foi usada como classificador. A base de dados usada é a ASVspoof 2017. Quanto aos resultados se obteve uma EER geral menor que 0,1.

A proposta de [**?**] é usar sinais residuais de predição linear, para, juntamente com coeficientes cepstrais criar características que serão apresentas a um classificador GMM. Novamente, a base de dados usada foi a ASVspoof 2015 e os resultados em ERR geral foram de 5,249.

Para detecção de voice spoofing [**?**] importa do campo de processamento de imagens o conceito de textura, para o processamento de voz esse conceito é chamado de "texturas de voz". Padrões binários locais (LBP) e seu respectivo histograma são usados para a construção do vetor de características que será avaliado por uma SVM. A base de dados usada para testes foi a ASVspoof 2015. A taxa máxima de acurácia conseguida foi de 0,7167.

Uma abordagem que combina análise de sinal de fala usando a *transformada de constante Q* (CQT) com o processamento cepstral é mostrada em [**?**], essa técnica resulta no que se chama *coeficientes cepstrais de constante Q*(CQCCs), segundo o artigo, a vantagem destes coeficientes é a resolução espectro temporal variável. As base de dados usadas foram a RedDots [**?**], ASVspoof 2015 e AVSpoof. Foram usados três classificado-

res:

- DA-IICT: Uma fusão de dois classificadores GMM, sendo que um deles usa *coeficientes cepstrais de frequência MEL*(MFCC) e o outro usa características CFCC-IF [**?**].
- STC [**?**].
- SJTU [**?**].

Na seção de experimentos são feitos testes para cada uma das bases com os seguintes resultados: ASVspoof 2015 → EER geral de 0.026; AVspoof → EER geral de 0; RedDots → EER geral de 0,185.

O artigo [**?**] propõe uma aproximação usando reverberação e as partes não vozeadas da fala, três GMMs foram definidos para a classificação, esses classificadores votam se uma ocorrência é ou não verdadeira, ganhado sempre a classificação que obtiver mais votos. A base de dado utilizada foi a ASVSpoof 2017. O sistema de avaliação de desempenho escolhido, novamente, foi o ERR e esta alcançou um valor de 2,99.

A principal ideia de [**?**] é capturar a amplitude instantânea vinda de flutuações instantâneas de energia. Segundo o artigo as modulações de amplitude são mais suscetíveis ao ruído inserido no sinal original por uma fonte reprodutora. O estudo usa a base de dados ASVSpoof 2017 e GMM como classificador. Os resultados apresentados chegaram a uma EER de 0.0019.

No trabalho [**?**] foram usadas as diferenças entre bandas de frequências específicas para diferenciar um sinal legítimo de um usado em ataques de falsificação. Neste trabalho é proposta a *predição linear em domínio de frequência*(FDLP) juntamente com GMMs para classificação dos dados presentes na base ASVspoof 2017. Os resultados apresentados chegaram a uma EER de 0.0803.

Em [**?**] se propõe duas novas características que visam interpretar as componentes estáticas e dinâmicas do sinal, essas características complementam as características de tempo restrito no espectro. São elas a *"modulation spectral centroid frequency"* e a *long*

term spectral average. O sistema usa como classificador um GMM juntamente com a base dados ASVSpoof 2017. Os resultados chegaram a um valor de EER de 0,0654.

Considerando o envelopamento das amplitudes e das frequências instantâneas em cada banda estreita filtrada, [**?**] discute como diferenciar um sinal legítimo de um falso. A base de dados usada foi a ASVSpoof 2015. Um GMM foi usado como classificador e, em relação ao desempenho, o método chegou a ter um EER de 0,045.

Neste trabalho [**?**] é proposto o uso do *gammatone frequency cepstral coefficients*(MGFCC). O gammatone é o produto de uma distribuição gamma com um sinal senoide e é usado na construção de filtros auditivos que, neste caso, são usados para extrair características do sinal de voz. A base de dados usada foi a ASVspoof 2015. O classificador usado foi um GMM e o EER chegou a 0,02556.

Segundo [**?**] *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 [**?**].

2.2.1 Contextualização

No trabalho proposto para dissertação de mestrado do autor a intenção é encontrar um conjunto de características que demonstrem ser as mais disjuntas possíveis para fins de separação entre as classes, com o objetivo de melhorar a acurácia de classificadores para detecção de ataques de *voice spoofing*. Caraterísticas essas com base na transformada *wavelet*, devido a sua boa resolução em relação às dimensões de tempo e frequência. Essas características serão avaliadas usando a análise paraconsistente de acordo com o trabalho [**?**] recentemente publicado.

Capítulo 3

Cronograma para conclusão do curso de mestrado

Até a presente data foram realizados os primeiros levantamentos para construção da base de dados com as vozes que serão objeto de pesquisa. Essa base conta com um total de 21 gravações originais e outras 21 de *playback* gravadas de pessoas com variados gêneros e idades pronunciando os dígitos de zero a nove em Inglês. A ideia é que essa base possa crescer e abarcar boa parte dos tipos de vozes existentes na região geográfica próxima ao campus da UNESP de São José do Rio Preto. Também foram estudados e implementados filtros passa baixas e passa altas, além de testadas metodologias de criação de vetores de características usando os intervalos espectrais pré-definidos pelas bandas MEL e BARK.

Foi melhorada a biblioteca criada e fornecida pelo orientador deste trabalho, a qual facilita a manipulação dos arquivos de áudio no formato *wave*. Do mesmo modo, código-fonte para análise paraconsistente de características foi também desenvolvido e complementado com a devida documentação, a qual foi norteada por instruções recebidas do orientador.

Quanto aos trabalhos futuros a tabela 3.1 mostra o cronograma previsto.

09/03 – 31/03	Início dos trabalhos: Coleta de dados para constituição da base de dados e estudo da base bibliográfica. Escrita da dissertação.
01/04 – 15/04	Coleta de dados para constituição da base de dados e estudo da base bibliográfica. Início dos experimentos. Escrita da dissertação.
16/04 – 16/04	Reunião de validação com o orientador.
17/04 – 29/04	Experimentos e escrita da dissertação.
30/04 – 30/04	Reunião de validação com o orientador.
01/05 – 20/05	Experimentos e escrita da dissertação.
21/05 – 28/05	Reunião de validação com o orientador.
29/05 – 10/06	Experimentos e escrita da dissertação.
11/06 – 11/06	Reunião de validação com o orientador.
12/06 – 17/06	Experimentos e escrita da dissertação.
18/06 – 18/06	Reunião de validação com o orientador.
19/06 – 24/06	Experimentos e escrita da dissertação.
25/06 – 25/06	Reunião de validação com o orientador.
26/06 – 01/07	Experimentos e escrita da dissertação.
02/07 – 02/07	Reunião de validação com o orientador.
03/07	Entrega da dissertação.

Tabela 3.1: Cronograma

Capítulo 4

Apêndice

4.1 *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 4.1:

$$F1\ score = 2 \times \frac{precisão \times recall}{precisão + recall} \tag{4.1}$$

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.

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

[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.

[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.

[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.

[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.

[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.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[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.

[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.

[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.

[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.

[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.

[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.

[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.