Universidade de Brasília - UnB
Faculdade UnB Gama - FGA
Engenharia Eletrônica

Desafios na Implementação de Modelos de *Speech Transformer* e *Conformer* para Reconhecimento de Fala Silenciosa com Eletromiografia

Autor: Mariana Martins de Souza
Orientador: Prof. Dr. Gerardo Antonio Idrobo Pizo

Brasília, DF
2023

Mariana Martins de Souza

Desafios na Implementação de Modelos de *Speech Transformer* e *Conformer* para Reconhecimento de Fala Silenciosa com Eletromiografia

Monografia submetida ao curso de graduação em (Engenharia Eletrônica) da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em (Engenharia Eletrônica).

Universidade de Brasília - UnB

Faculdade UnB Gama - FGA

Orientador: Prof. Dr. Gerardo Antonio Idrobo Pizo

Brasília, DF

2023

Mariana Martins de Souza

Desafios na Implementação de Modelos de *Speech Transformer* e *Conformer* para Reconhecimento de Fala Silenciosa com Eletromiografia/ Mariana Martins de Souza. – Brasília, DF, 2023-

89 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Gerardo Antonio Idrobo Pizo

Trabalho de Conclusão de Curso – Universidade de Brasília - UnB
Faculdade UnB Gama - FGA , 2023.

1. Amplificador. 2. Fala silenciosa. I. Prof. Dr. Gerardo Antonio Idrobo Pizo. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Desafios na Implementação de Modelos de *Speech Transformer* e *Conformer* para Reconhecimento de Fala Silenciosa com Eletromiografia

CDU 02:141:005.6

Mariana Martins de Souza

Desafios na Implementação de Modelos de *Speech Transformer* e *Conformer* para Reconhecimento de Fala Silenciosa com Eletromiografia

Monografia submetida ao curso de graduação em (Engenharia Eletrônica) da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em (Engenharia Eletrônica).

Trabalho aprovado. Brasília, DF, 11 de setembro de 2023:

Prof. Dr. Gerardo Antonio Idrobo Pizo
Orientador

Prof. Dr. Daniel Mauricio Muñoz Arboleda
Convidado 1

Prof. Dr. Marcus Vinícius Chaffim Costa
Convidado 2

Brasília, DF
2023

Agradecimentos

Agradeço primeiramente a Deus por me fornecer a força e a perseverança necessárias para alcançar meus objetivos.

Quero expressar minha profunda gratidão à minha família, meus amigos e todos aqueles próximos a mim, pelo constante encorajamento, confiança e amor incondicional.

Quero estender meus agradecimentos aos professores do curso de Engenharia Eletrônica da Faculdade do Gama, com destaque para meu orientador, Gerardo Idrobo, por me proporcionarem oportunidades, compartilharem conhecimento e demonstrarem disponibilidade e motivação ao longo desta jornada.

Por fim, expresso minha gratidão a todos que, de uma forma ou de outra, contribuíram direta ou indiretamente para a realização deste trabalho, demonstrando apoio e me fazendo continuar, apesar das adversidades.

Resumo

O principal meio de comunicação entre os seres humanos é a fala, possibilitando a transmissão de ideias, emoções e informações. No entanto, há situações em que a comunicação por meio da fala não é viável devido à perda da capacidade de falar ao longo da vida, ambientes desfavoráveis ou a necessidade de privacidade. Nesse contexto, surge a necessidade de aplicar técnicas de reconhecimento de fala silenciosa, que permitem identificar o que está sendo dito com base na movimentação dos órgãos articulatórios, músculos faciais e do pescoço, utilizando métodos como leitura labial, ultrassom e outros sensores.

Para abordar esse tema, realizou-se um estudo abrangente sobre o funcionamento da fala e as principais abordagens de reconhecimento de fala silenciosa desenvolvidas até o momento, bem como os métodos de processamento de sinais e reconhecimento de fala mais comuns. Com base nas referências consultadas, foi delineado um estudo de caso que servirá como base para experimentos futuros. Detalhes técnicos foram apresentados sobre a técnica de eletromiografia, que envolve a captura de dados por meio de eletrodos, assim como os principais métodos de amplificação de sinal e pré-processamento no contexto da eletromiografia. Também foram explorados métodos como STFT, MFCC, LSTM bidirecional, DTW, CCA, RNN, Transformer, Vocoder e CTC, com um foco específico no trabalho de Gaddy e Klein, que foi selecionado como referência para o estudo experimental.

Na fase prática do trabalho, foram analisados dois modelos de redes neurais naturais: o Speech Transformer e o Conformer. Devido a limitações de hardware, não foi possível implementar e avaliar adequadamente o Speech Transformer. No entanto, o modelo Conformer foi implementado, embora tenha apresentado desafios, incluindo uma alta taxa de erro por palavra. Diversos testes foram conduzidos com diferentes otimizadores e taxas de aprendizado, mas não resultaram em melhorias substanciais.

Acredita-se que as dimensões menores da arquitetura Conformer e o tamanho reduzido da base de dados possam ter contribuído para os resultados menos satisfatórios. Além disso, a escolha de utilizar o sinal EMG parametrizado manualmente em vez do EMG bruto pode ter impactado negativamente no processo de aprendizado.

Para pesquisas futuras, é recomendado testar as hipóteses mencionadas e investigar a eficácia do código do modelo Conformer. Também seria benéfico conduzir experimentos com outras bases de dados e explorar configurações de modelos mais robustas, com o objetivo de aprimorar o desempenho do reconhecimento de fala silenciosa.

Palavras-chaves: fala silenciosa. interfaces. parametrização. reconhecimento. eletromiografia. redes neurais. speech transformer. conformer.

Abstract

The main means of communication among human beings is speech, enabling the transmission of ideas, emotions, and information. However, there are situations where speech communication is not feasible due to the loss of the ability to speak over one's lifetime, unfavorable environments, or the need for privacy. In this context, there is a need to apply silent speech recognition techniques, which allow for the identification of what is being said based on the movement of articulatory organs, facial muscles, and the neck, using methods such as lip reading, ultrasound, and other sensors.

To address this topic, a comprehensive study was conducted on the functioning of speech and the primary approaches to silent speech recognition developed to date, as well as the most common signal processing and speech recognition methods. Based on the consulted references, a case study was outlined, which will serve as a basis for future experiments. Technical details were provided on the electromyography technique, involving data capture through electrodes, as well as the main signal amplification and preprocessing methods in the context of electromyography. Methods such as STFT, MFCC, bidirectional LSTM, DTW, CCA, RNN, Transformer, Vocoder, and CTC were also explored, with a specific focus on the work of Gaddy and Klein, which was selected as a reference for the experimental study.

In the practical phase of the work, two models of natural neural networks were analyzed: the Speech Transformer and the Conformer model. Due to hardware limitations, it was not possible to implement and adequately evaluate the Speech Transformer. However, the Conformer model was implemented, although it presented challenges, including a high word error rate. Several tests were conducted with different optimizers and learning rates, but substantial improvements were not achieved.

It is believed that the smaller dimensions of the Conformer architecture and the reduced size of the database may have contributed to the less satisfactory results. Additionally, the choice to use manually parameterized EMG signals instead of raw EMG may have negatively impacted the learning process.

For future research, it is recommended to test the mentioned hypotheses and investigate the effectiveness of the Conformer model's code. It would also be beneficial to conduct experiments with other databases and explore more robust model configurations to enhance the performance of silent speech recognition.

Key-words: silent speech. interfaces. parameterization. recognition. electromyography. neural network. speech transformer. conformer.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

A_V	Sinal do áudio vocalizado
E_S	Sinal do EMG silencioso
E_V	Sinal do EMG vocalizado
Adam	Estimativa Adaptativa de Momento, do inglês *Adaptive Moment Estimation*
Ag	Prata
AgCl	Cloreto de prata
ANN	Rede Neural Artificial, do inglês *Artificial Neural Network*
ASE	Codificação Avançada de Fala, do inglês *Advanced Speech Encoding*
BASE	Motor de Busca Acadêmico de Bielefeld, do inglês *Bielefeld Academic Search Engine*
BN	Normalização de Lote, do inglês *Batch Normalization*
CCA	Análise de Correlação Canônica, do inglês *Canonical Correlation Analysis*
CNN	Rede Neural Convolucional, do inglês *Convolutional Neural Network*
CTC	Classificação Temporal Conexionista, do inglês *Connectionist Temporal Classification*
DARPA	Agência de Projetos de Pesquisa Avançada de Defesa, do inglês *Defense Advanced Research Projects Agency*
DTW	Alinhamento Temporal Dinâmico, do inglês *Dynamic Time Warping*
ECoG	Eletrocorticografia
EEG	Eletroencefalografia
EGG	Eletroglotografia
EMA	Articulografia eletromagnética
EMG	Eletromiografia

FFT	Transformada de Fourier Rápida, do inglês *Fast Fourier Transform*
GAN	Perda Generativa Adversarial, do inglês *Generative Adversarial Network*
GMM	Modelo de Mistura Gaussiana, do inglês *Gaussian Mixture Model*
GPU	Unidade de Processamento Gráfico, do inglês *Graphics Processing Unit*
HMM	Modelos Ocultos de Markov, do inglês *Hidden-Markov Models*
LDA	Análise Discriminante Linear, do inglês *Linear Discriminant Analysis*
LSTM	Memória Longa de Curto Prazo, do inglês *Long Short-Term Memory*
MFCC	Coeficientes Cepstrais de Frequência Mel, do inglês *Mel-Frequency Cepstral Coefficients*
NAM	Murmúrio Não Audível
NIRS	Espectroscopia de Infravermelho Próximo, do inglês *Near-Infrared Spectroscopy*
PAUM	Potencial de Ação da Unidade Motora
PICNN	Rede Neural Convolucional de Inicialização Paralela, do inglês *Parallel Initiation Convolutional Neural Network*
PMA	Articulografia de Ímã Permanente
PRISMA	Itens de Relato Preferencial para Revisões Sistemáticas e Meta-Análises, do inglês *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*
ReLU	Unidade Linear Retificada, do inglês *Rectified Linear Unit*
RF	Floresta Aleatória, do inglês *Random Forest*
RMS	Valor Médio Quadrático, do inglês *Root-Mean-Square*
RMSProp	Propagação da Raiz quadrada da Média, do inglês *Root-Mean-Square Propagation*
RNN	Rede Neural Recorrente, do inglês *Recurrent Neural Network*
SAD	Detecção de Atividade de Fala, do inglês *Speech Activity Detection*
sEMG	Eletromiografia de Superfície
SGD	Gradiente Descente Estocástisco, do inglês *Stochastic Gradient Descent*

SNAP	Potencial de Ação do Nervo Sensorial
SSI	Interface de Fala Silenciosa, do inglês *Silent Speech Interface*
STFT	Transformada de Fourier de Curto Prazo, do inglês *Short Time Fourier Transform*
SVM	Máquina de Vetores de Suporte, do inglês *Support Vector Machine*
T	Prompt de Texto
TCC	Trabalho de Conclusão de Curso
UM	Unidade Motora
VAD	Detecção de Atividade de Voz, do inglês *Voice Activity Detection*
VCN	Velocidade de Condução Nervosa
WER	Taxa de Erro por Palavra, do inglês *Word Error Rate*
WT	Transformada Wavelet

Sumário

1 Introdução

1.1 Contextualização

A denominada fala silenciosa consiste na comunicação similar à fala, sem a exigência de sons produzidos audíveis. Dessa forma, com o uso de sensores é possível captar esses enunciados que foram articulados sem a emissão de som audível. Com os dados captados da fala silenciosa, é possível gerar uma voz sintética ou um texto digital para transmitir ou reproduzir a mensagem desejada, tornando o mecanismo acessível aos dispositivos atuais e seus respectivos assistentes digitais, aproveitando a alta qualidade existente de sistemas de fala para texto baseados em áudio (WADKINS, 2019).

O projeto em questão tem uma ampla gama de aplicações potenciais. Por exemplo, poderia ser usado para criar um dispositivo análogo a um fone de ouvido *Bluetooth* que permite que as pessoas conversem pelo celular sem fazer barulho e atrapalhar os que estão ao seu redor. Tal dispositivo também poderia ser útil em situações onde o ambiente está com muito barulho para se manter uma conversa audível ou onde manter o silêncio é importante. Mas, principalmente, a tecnologia pode ser usada por pessoas que não são mais capazes de produzir fala ou que devem evitar fazê-lo.

Em termos de patologias e condições clínicas, tem-se que a comunicação através da fala silenciosa poderia ser útil, por exemplo, para pacientes com afasia, que possuem dificuldade de se comunicar, com apraxia de fala, que possuem dificuldades motoras na produção de fala, com disartria, fraqueza nos músculos da fala, ou que passaram por laringectomia, que pode acarretar em perda de fala ou rouquidão (GONZALEZ-LOPEZ et al., 2020).

Até o presente momento, diferentes métodos de monitoramento de fala silenciosa foram desenvolvidos, obtendo informações de partes diferentes do corpo e com interfaces variadas. Uma interface de fala silenciosa é um sistema que permite a comunicação de fala quando um sinal acústico audível não está disponível. Ao adquirir dados do sensor de elementos do processo de produção da fala humana, seja dos articuladores, suas vias neurais ou do cérebro em si, ela produz uma representação digital da fala que pode ser sintetizada diretamente, interpretada como dados ou encaminhada para uma rede de comunicações (DENBY et al., 2010).

Várias modalidades de biosinais foram estudadas no contexto do desenvolvimento desses sistemas de comunicação de fala silenciosa, incluindo ultrassom, imagens ópticas, eletroencefalografia (EEG) e eletromiografia de superfície (sEMG). Diversos trabalhos foram capazes de obter o reconhecimento da fala silenciosa com resultados promissores,

como o de Wadkins (2019), Wang et al. (2021), Gaddy e Klein (2020), Tran et al. (2010), entre outros. Contudo, utilizando vocabulários restritos. Se trata então de uma área de estudo em desenvolvimento, com limitações a serem superadas e técnicas a serem aprimoradas, mesmo com o emprego do aprendizado de máquina para alavancar os avanços (MELTZNER et al., 2011).

1.2 Motivação

Desde os primórdios, a comunicação foi essencial para a convivência e o desenvolvimento dos seres humanos. Com o auxílio da tecnologia, os meios de comunicação se tornaram ainda mais amplos e eficientes, assim como as tecnologias de maneira geral, trazendo soluções e derrubando barreiras.

Nesse sentido, a principal motivação do presente trabalho foi verificar a possibilidade e a efetividade do monitoramento da fala silenciosa, tido como alternativa de comunicação para pessoas que possuem limitações físicas e para situações onde a manifestação da voz não é bem recebida ou possibilitada.

1.3 Justificativa

A não existência de dispositivos comerciais que sejam capazes de transmitir a fala humana sem o uso da vocalização ou da digitação faz com que essa seja uma temática promissora. Ademais, a maior parte das pesquisas realizados na área é de outros países, existindo portanto uma demanda nacional de estudos na temática, à qual esse trabalho busca contribuir.

1.4 Objetivo

1.4.1 Objetivo Geral

O foco principal do trabalho é a implementação dos métodos *Speech Transformer* e *Conformer*, duas redes neurais profundas, consideradas o estado da arte no contexto, para reconhecimento de sinais emitidos durante a fala silenciosa, de forma com que os sinais elétricos captados sejam transformados em palavras e frases.

1.4.2 Objetivos Específicos

- Compreender a execução da fala silenciosa e os parâmetros que podem ser utilizados para seu monitoramento;

- Identificar os principais métodos de monitoramento da fala silenciosa;
- Estabelecer um estudo de caso a ser utilizado como base para o experimento;
- Implementar e analisar o comportamento do *Speech Transformer* no contexto do estudo de caso;
- Implementar e analisar o comportamento do *Conformer* no contexto do estudo de caso.

1.5 Divisão do Trabalho

Esse trabalho está organizado em sete capítulos, estando inclusa a introdução. O capítulo 2 consiste na revisão bibliográfica, com a apresentação do modelo de revisão sistemática, fluxograma PRISMA e demais artefatos utilizados. O 3 é o desenvolvimento do referencial teórico, levantado a partir da revisão bibliográfica. O 4 é a consolidação dos entendimentos e proposta experimental de monitoramento da fala silenciosa com base em um estudo de caso. O 5 trata da metologia a ser empregada para execução do experimento, cujos resultados e descrições constam no capítulo 6. Por fim, a última parte do trabalho trata-se da conclusão, englobando também sugestões para trabalhos futuros.

2 Revisão Bibliográfica

O presente capítulo visa apresentar os resultados de uma revisão sistemática, além das análises qualitativas e quantitativas que permearam a seleção final de artigos.

O monitoramento da fala silenciosa pode ser realizado com o uso de diferentes técnicas eletrônicas, captando sinais diferentes, com aparatos variados. Mesmo que tenham o mesmo objetivo final, a reprodução da fala humana, ou objetivos correlatos, os resultados obtidos, bem como as metodologias empregadas são plurais. Entender os processos envolvidos e a diversidade de conclusões obtidas, é, portanto, fundamental para a escolha do método a ser empregado experimentalmente.

Diante desse contexto, foi elaborada uma revisão sistemática de literatura para melhor análise dos estudos de casos realizados e consolidação dos fatores que são entendidos como cruciais para o monitoramento da fala silenciosa, a serem utilizados como referencial teórico posteriormente.

2.1 Método de Revisão Sistemática

A etapa de revisão sistemática de literatura foi dividida em três fases: planejamento, condução e resultados. Durante o planejamento, foram realizadas a definição dos objetivos da revisão sistemática, a criação das questões de estudo, termos de busca e fontes de pesquisa, além dos critérios de inclusão e exclusão, conforme pode ser observado na Seção 2.2.

Na fase de condução, foi executada a busca nas bases de dados selecionadas, sendo essas *BASE – Bielefeld Academic Search Engine* e *IEEE Xplore*, utilizando-se os termos de pesquisa definidos na fase anterior e os filtros com alguns critérios de exclusão. Os resultados de busca obtidos então passam aos estudos primários, sendo realizada a leitura dos resumos das publicações para aplicação dos demais critérios de inclusão e exclusão. Também foi realizado um estudo bibliométrico para compreensão do cenário da área de pesquisa, considerando a quantidade de publicações encontradas por autores e artigos mais citados, sendo possível identificar alguns dos principais autores e publicações da área. Os métodos de condução da revisão sistemática estão detalhados na Seção 2.3.

A etapa de resultados, por sua vez, contemplou a leitura integral das publicações selecionadas para o estudo das técnicas de captação de dados, processamento dos sinais coletados e aplicação das soluções, sendo realizada a extração e a síntese desses dados. Também foram analisadas as amostras utilizadas pelos autores, bem como a taxa de êxito das técnicas empregadas e sua complexidade. Os resultados encontrados estão descritos

na Seção 2.4.

2.2 Planejamento

2.2.1 Estratégia de busca

Durante o planejamento, foi definido como objetivo de pesquisa o estudo de monitoramento de fala silenciosa. Foi estabelecida então a questão de pesquisa "quais as principais técnicas utilizadas para o reconhecimento da fala silenciosa?". Por fim, as perguntas secundárias foram definidas para auxiliar a solução da questão principal, sendo elas: (1) "como é realizada a aquisição de dados?"; (2) "quais são os métodos de processamento de sinais empregados?"; e (3) "quais foram as condições de vocabulário utilizadas?".

Para as fontes de pesquisa, foram selecionadas a IEEE Xplore[1] e a BASE - *Bielefeld Academic Search Engine*[2]. A primeira foi selecionada por ser referência entre os cursos de engenharia elétrica, eletrônica e de áreas da computação, contendo algumas da publicações mais relevantes e citadas. Por sua vez, a BASE foi escolhida por ser uma plataforma interdisciplinar com mais de 300 milhões de documentos e dezenas de milhares de provedores de conteúdo. A definição dos termos para a busca contemplou as seguintes palavras: (1) fala silenciosa; (2) perda de voz; (3) interface; (4) sensor; (5) processamento; (6) parametrização; (7) reconhecimento.

2.2.2 Critérios de inclusão

Para a identificação dos métodos empregados no monitoramento da fala silenciosa, foram definidos os seguintes critérios de inclusão com base nas questões secundárias de pesquisa:

- O artigo apresenta técnicas de aquisição de dados de fala silenciosa;
- O artigo apresenta métodos de processamento de sinais de fala silenciosa;
- O artigo apresenta condições de vocabulário para delimitação experimental do monitoramento de fala silenciosa.

2.2.3 Critérios de exclusão

Os critérios de exclusão foram selecionados considerando a possibilidade de remoção de artigos que não respondem ao menos uma das questões secundárias e que não possuíssem relação com a área de estudo, que estivessem disponíveis em idiomas além

[1] Pode ser acessado em: https://ieeexplore.ieee.org/
[2] Pode ser acessado em: https://www.base-search.net/

dos escolhidos ou que não pudessem ser acessados de forma integral, impossibilitando a leitura. Portanto, os critérios de exclusão definidos foram:

- O artigo não se enquadra nos critérios de aceitação;
- O artigo é de outra área de estudo;
- O artigo está duplicado;
- Não é possível a leitura integral do artigo;
- A publicação está em um idioma diferente de inglês e português;
- A publicação é uma versão mais antiga de outro estudo.

2.3 Condução da Pesquisa

Com base na questão de pesquisa definida e nas questões secundárias, foi realizada a busca nas fontes de pesquisa IEEE Xplore e BASE - *Bielefeld Academic Search Engine*, sendo encontrados 1027 artigos relacionados. Com a aplicação de filtros baseados nos critérios de exclusão e inclusão, o número de artigos selecionados reduziu para 261, sendo a maior parte das exclusões decorrente de estudos de em outros idiomas além dos selecionados e pela impossibilidade da leitura integral das publicações, justificativas categorizadas em "assinalados como não elegíveis pelas ferramentas automatizadas" no fluxograma PRISMA da Fig. 1.

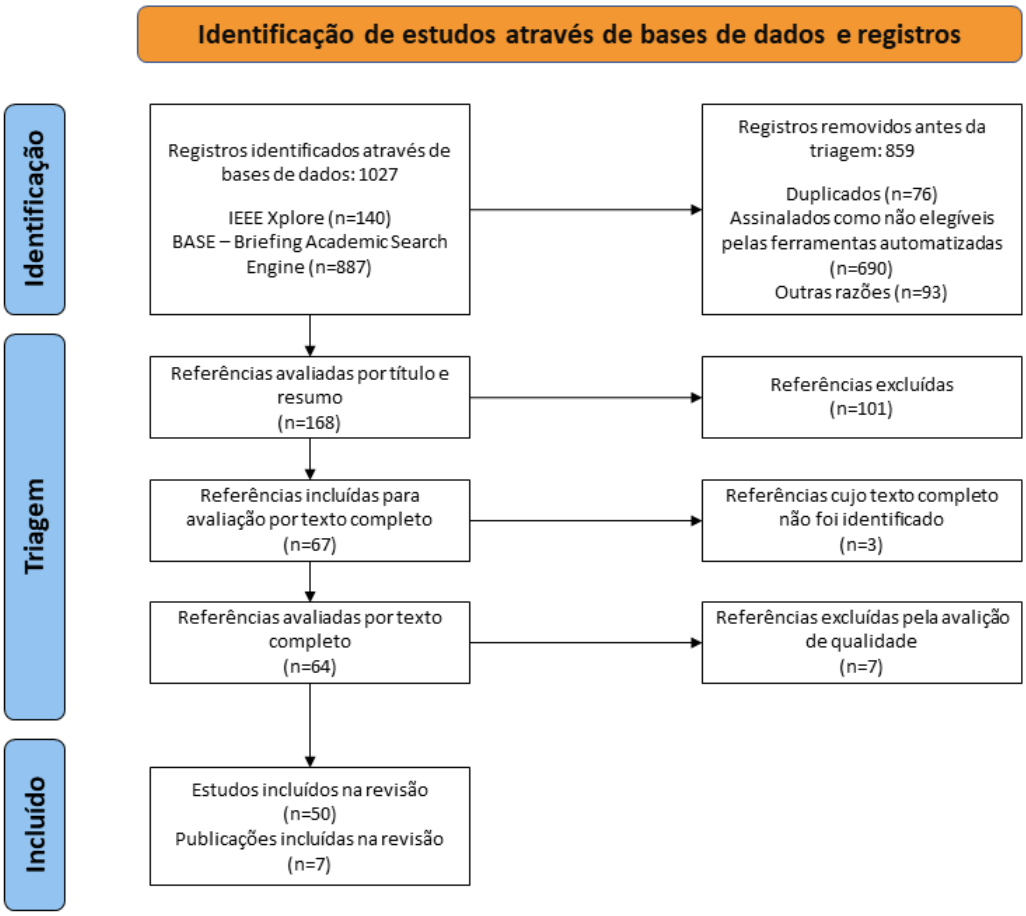

Figura 1 – Fluxograma PRISMA adaptado
Fonte: Autoria própria

Antes da triagem, ainda foi realizada a exclusão de referências com base no estudo bibliométrico realizado, sendo selecionados apenas os artigos de maior relevância dentre os buscados na plataforma da IEEE Xplore, conforme descrito na Seção 2.3.1. Foram incluídas somente as publicações com maior número de citações, de autores tidos como referência da área ou de autores com diversas publicações da temática, sendo assim, as referências removidas estão identificadas como "outras razões" no fluxograma PRISMA. Dessa forma, ao todo, considerando as duas fontes de pesquisa, foram selecionadas 168 referências para a etapa de triagem.

Durante a triagem, foi realizada a leitura do título e do resumo e cada publicação, sendo aplicados novamente os critérios de exclusão e inclusão. Nessa etapa, a maior parte das exclusões se deu por se tratarem de estudos de outras áreas, não sendo relevantes, portanto, para a revisão em questão. A partir da leitura completa dos textos, ainda foi possível excluir algumas publicações com base na avaliação de qualidade mostrada na Seção 2.3.2. Dessa forma, foram incluídas, ao todo, 57 referências na revisão, sendo 50 artigos e 7 teses de mestrado ou doutorado.

2.3.1 Estudos bibliométricos

O estudo bibliométrico foi realizado na base de dados IEEE Xplore, sendo utilizadas as ferramentas da própria plataforma. Não foi possível aplicar a técnica na base de dados BASE devido à ausência de informações, como o número de citações das publicações identificadas na busca. Os termos utilizados para o estudo bibliométrico contemplaram as palavras: fala silenciosa, reconhecimento, processamento, aquisição de sinal, interface. Sendo essas identificadas em títulos, resumos e palavras-chaves.

A listagem obtida pela busca continha 140 artigos, com um total de 1.190 citações. No entanto, desses artigos, 40 não eram citados e os 9 artigos que possuíam mais de 30 citações somaram 46,21% das citações totais, conforme pode ser observado na Figura 2. Dos artigos restantes, 22 possuíam 10 ou mais citações, sendo esses responsáveis por 32,69% das citações totais. Os artigos citados menos de 10 vezes representaram 49,29% das publicações encontradas e representavam 21,1% das citações totais, não sendo considerada uma quantidade significativa.

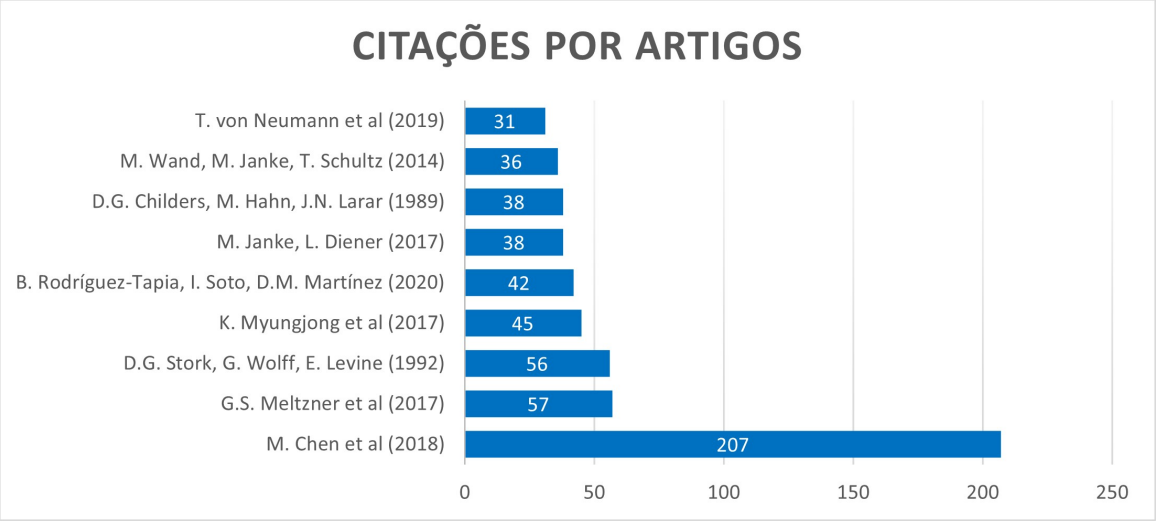

Figura 2 – Gráfico de análise de quantidade de citações por artigos
Fonte: Autoria própria

Fazendo a análise tendo os autores como foco, 14 deles publicaram mais de 2 artigos dentre os selecionados na busca, desses foram destacados os que possuíam 4 ou mais publicação para criação do gráfico da Figura 3. Conjuntamente, esses autores representam apenas 17,14% das citações. Realizando então a análise dos autores com mais citações, independente do número de publicações, foi possível obter o gráfico da Figura 4, ressaltando-se que os 11 autores destacados representam 40,17% das citações.

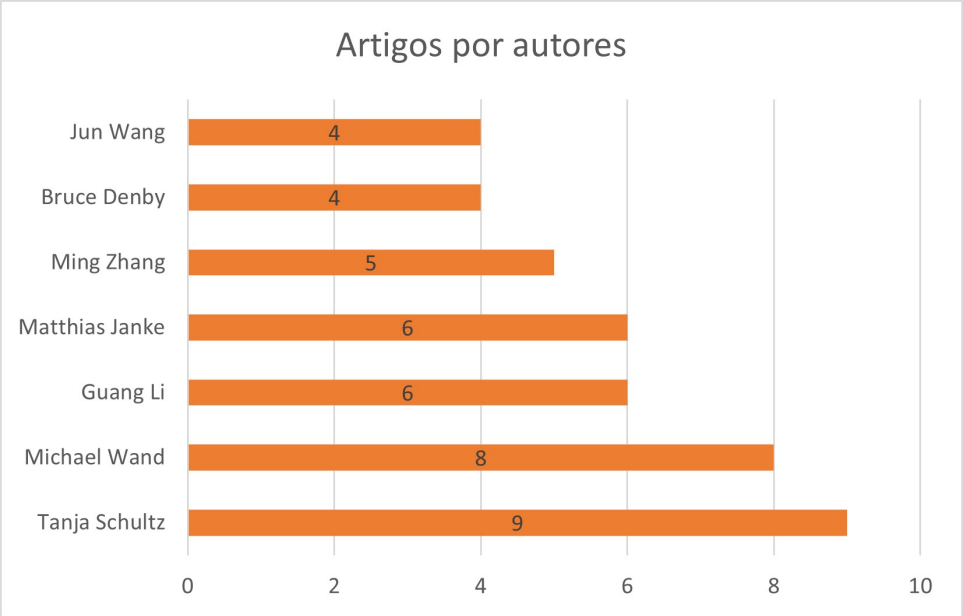

Figura 3 – Gráfico de análise de quantidade de artigos por autores
Fonte: Autoria própria

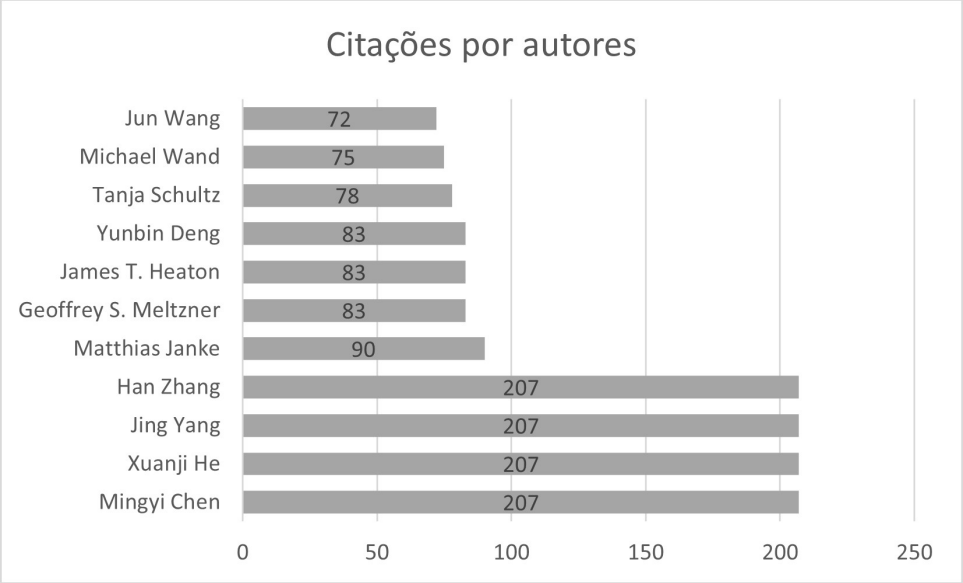

Figura 4 – Gráfico de análise de quantidade de citações por autores
Fonte: Autoria própria

2.3.2 Avaliação de Qualidade

Para definir a importância dos artigos selecionados para o estudo, foram levantadas questões para a métrica de qualidade, sendo essas:

- Os objetivos do artigo e questões de pesquisa são evidentes e pertinentes?
- As condições de realização da pesquisa estão devidamente descritas?

- Os dados coletados estão relacionados com a questão de pesquisa?
- Os resultados estão declarados de maneira clara?
- As limitações do estudo foram explicitadas?
- As descobertas e conclusões foram claramente relatadas?

Com base nos critérios de qualidade, foram excluídos mais 9 artigos, sendo a principal justificativa a descrição insuficiente das condições e contexto de realização da pequisa e as limitações encontradas.

2.4 Resultados da Revisão

Com o estudo das 57 referências selecionadas, foi possível identificar quais eram as principais técnicas abordadas pelos autores, respondendo assim a questão principal e as secundárias de maneira quantitativa.

Primeiramente, foram levantados, a partir da leitura dos artigos, quais seriam os principais meios utilizados para a captação dos dados. Considerando que não houve limitações a respeito da origem dos dados, como movimentos dos músculos da face ou pescoço, murmúrios não audíveis ou estímulos cerebrais, foram identificadas diversas interfaces, conforme mostrado na Figura 5.

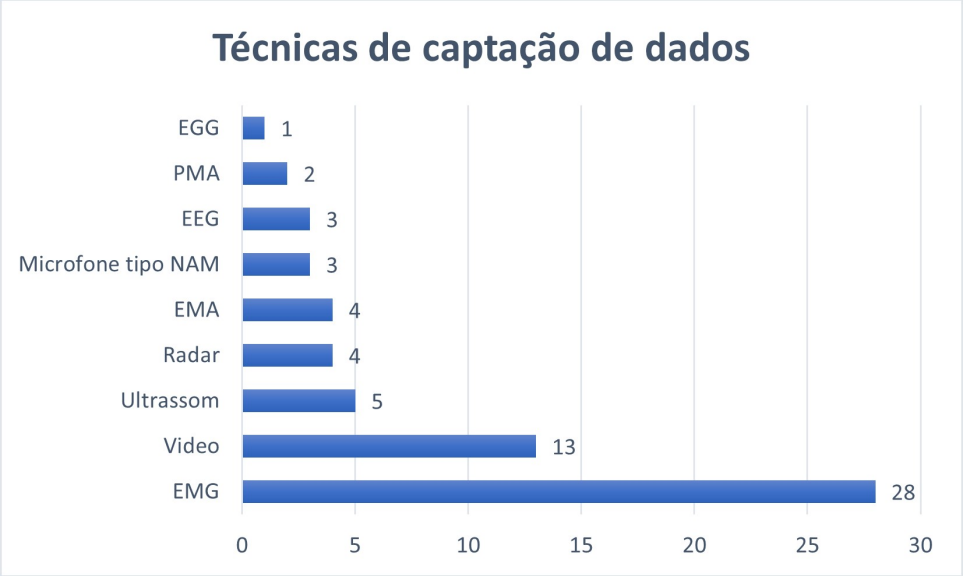

Figura 5 – Gráfico das técnicas de captação de dados identificadas
Fonte: Autoria própria

A técnica mais utilizada foi a eletromiografia (EMG), seguida por análise de vídeos. Além dessas, também foram identificados o uso de ultrassonografias, radares de

diferentes tipos, articulografia eletromagnética (EMA, do inglês *electromagnetic articulography*), captação de murmúrio não audível (NAM, do inglês *non-audible murmur*) com microfone, eletroencefalografia (EEG), articulografia de ímã permanente (PMA, do inglês *permanent magnet articulography*) e eletroglotografia (EGG). Algumas técnicas foram utilizadas isoladamente, enquanto alguns estudos utilizaram uma combinação, como o caso de ultrassom e vídeo, que foram usados de maneira complementar para maior robustez do experimento.

Para o reconhecimento de fala, são utilizadas primeiramente dois métodos de detecção de atividades nos sinais, sendo uma baseada em fala (SAD, do inglês *speech activity detection* e outra baseada em voz (VAD, do inglês *voice activity detection*), cujas diferenças serão abordadas na Seção 3.3.1. O processamento dos sinais, por sua vez, é realizado através de diversas técnicas, dentre as quais se destacam coeficientes cepstrais de frequência mel (MFCC - *Mel-Frequency Cepstral Coefficients*), análise discriminante linear (LDA - *linear discriminant analysis*), modelo oculto de Markov (HMM - *Hidden-Markov-Models*) e modelos de mistura gaussiana (GMM - *gaussian mixture model*), como pode ser observado na Figura 6.

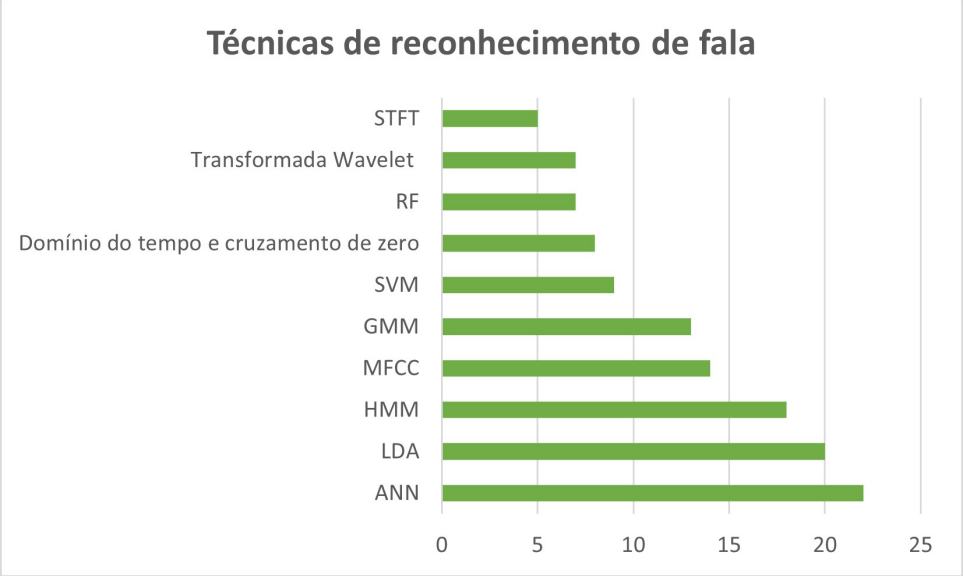

Figura 6 – Gráfico das técnicas de processamento de sinais e reconhecimento de fala identificadas

Fonte: Autoria própria

Foram identificadas ainda utilizações de redes neurais artificiais (ANN), como a recorrente (RNN), e memória longa de curto prazo(LSTM, do inglês *long short-term memory*), a convolucional (CNN), assim como algumas de suas variações, implementadas pelos autores dos estudos, como é o caso da rede neural convolucional de iniciação paralela (PICNN) (WU et al., 2022).

Em termos de amostras e condições de vocabulário, houve uma diversificação ainda

maior. Alguns estudos tiveram como propósito a leitura de poucas palavras por uma pessoa ou um grupo pequeno, enquanto outros se basearam em centenas de frases ditas por grupos de dezenas de pessoas. No entanto, é possível observar que a maior parte se manteve limitado a um conjunto de palavras ou frases pré-definidos, não abordando, portanto, a fala espontânea.

2.5 Considerações parciais

A revisão sistemática resultou no levantamento de dados de 57 publicações que obedeceram aos critérios de inclusão, exclusão e qualidade, a partir dos quais foi possível identificar as principais técnicas utilizadas, conforme ilustrado de maneira gráfica anteriormente. No entanto, para o aprofundamento do referencial teórico, foram utilizados apenas algumas publicações, que possuíam mais relevância para o estudo, como Wadkins (2019), Denby et al. (2010), Wang et al. (2021),Ocarino et al. (2005), assim como Gaddy (2022). Além das citadas, as publicações de Dong, Xu e Xu (2018) e Gulati et al. (2020) também foram amplamente consultadas, principalmente no que se refere a arquitetura das redes neurais *Speech Transformer* e *Conformer*, respectivamente.

3 Referencial Teórico

3.1 Funcionamento da fala humana

A produção da voz é o resultado da interação de componentes de três sistemas do corpo humano, o respiratório (com a atuação do pulmão, laringe, cavidade nasal, entre outros), o digestório (pelo uso da faringe, cavidade oral, mandíbula, dentes e lábios) e o nervoso (abrangendo encéfalo, medula espinhal e os nervos que conduzem o impulso elétrico para os músculos utilizados na articulação).

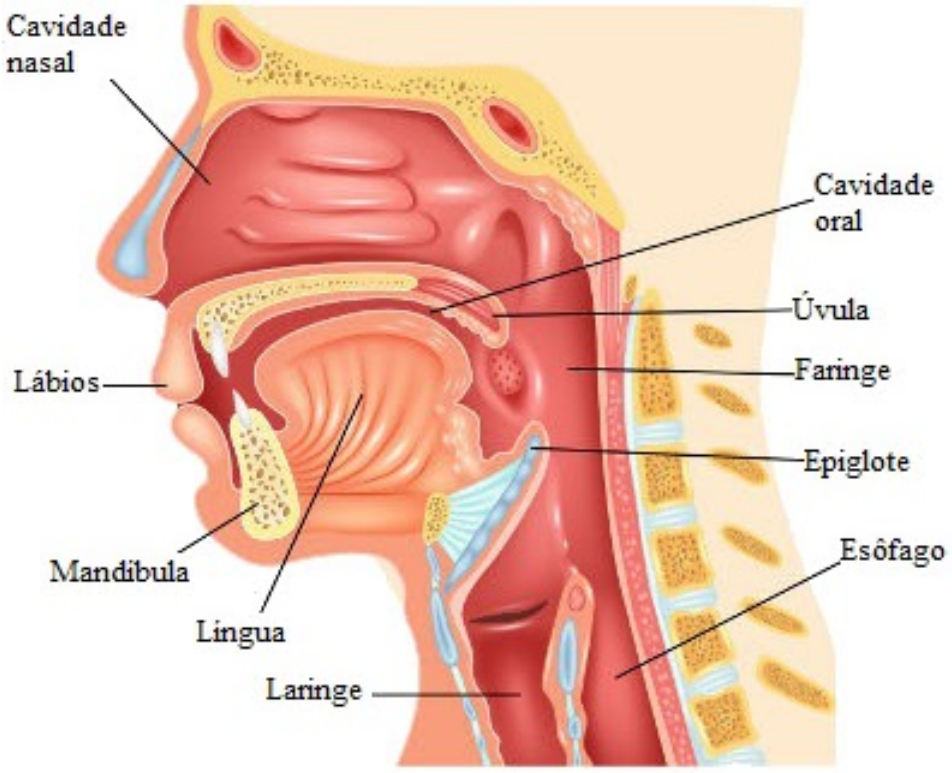

Figura 7 – Anatomia da fala
Fonte: (MUNDO EDUCAÇÃO, 2016)

Dentre as estruturas citadas, pode-se dizer que a laringe é o principal centro de vocalização. É uma estrutura formada por cartilagem (osso da laringe), membranas mucosas e ligamentos, e é influenciada pelos músculos que se originam e se inserem na própria laringe (músculos intrínsecos), bem como pelos músculos que conectam a laringe a outras estruturas na cabeça, pescoço e tórax (músculos extrínsecos). Em geral, os músculos intrínsecos da laringe são responsáveis pela abertura, fechamento, alongamento, encurtamento e tensão longitudinal das cordas vocais.

Assim, a musculatura intrínseca desempenha um papel crucial no mecanismo de

vocalização, afetando a vibração glótica e parâmetros relacionados à qualidade do som, frequência fundamental e intensidade. Por outro lado, os músculos extrínsecos estão envolvidos no movimento vertical da laringe no pescoço e de forma mais indireta afetam o posicionamento da laringe, a geometria da glote, a configuração da cavidade na glote, e os parâmetros vocais (FALE - UFMG, 2021).

3.2 Interfaces de fala silenciosa

Por envolver diversos componentes corporais, conforme mostrado na Seção 3.1, a obtenção de dados da fala silenciosa pode ser realizada através de múltiplas técnicas, com sensores e tecnologias diferentes. A respeito do termo "fala silenciosa", ressalta-se ainda que pode se referir tanto à fala subvocal ou subvocalização, quanto à ativação neurológica deliberada. No primeiro caso, é possível identificar algum movimento visível ou fala audível, sendo uma derivação da fala natural. Por sua vez, no caso de ativação neurológica, não há movimento visível ou som, sendo fruto apenas dos estímulos enviados aos articuladores internos de fala, fenômeno comum durante a leitura muito atenta (WADKINS, 2019).

Os estudos sobre subvocalização começaram na década de 70, com Eriksen, Pollack e Montague (1970) e Klapp, Anderson e Berrian (1973) definindo o conceito de "fala implícita" na literatura. Foi sugerido que a subvocalização pode desempenhar um papel na compreensão de palavras desconhecidas, pois muitas vezes os voluntários de seus estudos pausavam e subvocalizavam novas palavras antes de lê-las em voz alta. Dessa forma, a subvocalização foi considerada uma precursora da língua falada, ou outra representação da mesma.

A primeira interface de fala silenciosa (referida como SSI, do inglês *Silent Speech Interface*), no entanto, é datada da década de 80, no Japão. Primeiramente proposta como aprimoramento do reconhecimento de fala em ambientes ruidosos, a leitura labial visual automática foi o enfoque das primeiras patentes, com equipamentos tidos como capazes de interpretar comandos simples falados. Embora fossem sistemas limitados, serviram como precursores para o avanço do reconhecimento da fala silenciosa. Em 1985, foi utilizada a técnica de eletromiografia, com 3 sensores faciais empregados para o reconhecimento de 5 vogais japonesas, sendo obtida 71% de precisão na reprodução dessas em tempo real (DENBY et al., 2010). No mesmo período, Morse publicou nos Estados Unidos um estudo similar, obtendo 97% de acurácia na reprodução de duas palavras em inglês (WANG et al., 2021). Alguns anos depois, um sistema baseado em imagens, no qual movimentos do lábio e da língua foram extraídos do vídeo do rosto do locutor, retornou 91% de reconhecimento em um experimento semelhante aos anteriores.

Embora a possibilidade de robustez dos dispositivos de fala silenciosa ao ruído ambiente já fosse discutida em alguns dos primeiros artigos, a ideia de também recuperar

sinais de excitação glótica da fala expressa em ambientes ruidosos foi um desenvolvimento um pouco posterior. Um grande motivador foi o programa ASE (*Advanced Speech Encoding*, traduzido livremente como "Codificação Avançada de Fala") da DARPA do início dos anos 2000, que financiou pesquisas sobre síntese de fala de baixa taxa de bits “com inteligibilidade aceitável, qualidade e reconhecibilidade do alto-falante aural em ambientes acusticamente agressivos”, estimulando assim os desenvolvimentos no processamento da fala usando uma variedade de atividade glotal mecânica e sensores eletromagnéticos.

Não foi até o advento dos telefones celulares, no entanto, que as SSIs em sua forma atual começaram a ser discutidas. No Japão, em 2002, um comunicado de imprensa da NTT DoCoMo anunciou um protótipo silencioso celular usando EMG e captura óptica do movimento labial. Segundo a empresa, o estímulo para desenvolver um telefone desse tipo estava em eliminar o ruído excessivo de locais públicos, além de ser uma tecnologia para ajudar pessoas que perderam permanentemente a voz. Os primeiros trabalhos de pesquisa da SSI mencionando explicitamente a privacidade do celular como uma meta também começaram a aparecem nessa época (DENBY et al., 2010).

Em 2003, o Centro de Pesquisa Ames da NASA sobre sistemas de reconhecimento de fala subvocal, foi responsável pelo treinamento de um modelo de rede neural para o reconhecimento de 6 a 10 palavras subvocais, incluindo dígitos de zero a nove. O sistema deles usava fala sub-audível (ou *sotto voce*) e dependia de EMG do movimento muscular visível. Usando 100 amostras de treinamento de cada palavra, alcançaram uma precisão de 92% em amostras de teste gravadas. Sua conclusão ao experimentar vocabulários maiores foi que uma tarefa de reconhecimento de 20 palavras parecia viável, mas o reconhecimento de um vocabulário maior não era. Assim, sugeriram que o sistema fosse usado com um pequeno vocabulário especializado para a tarefa em questão. Mencionaram ainda que o sistema sofre por ser específico para um único usuário e sensível a ruídos, localizações de eletrodos e alterações fisiológicas no usuário (WADKINS, 2019).

A possibilidade de ir mais longe hoje do que em alguns dos projetos SSI anteriores se deve em grande parte pelos avanços em instrumentação feitos pela comunidade de pesquisa de produção de fala. Muitas das tecnologias de detecção propostas para uso em SSIs foram desenvolvidas ao longo de muitos anos para extrair informações detalhadas e em tempo real sobre o processo de produção da fala humana. Existe, portanto, hoje uma riqueza de recursos disponíveis para a aplicação de ultrassom, raio-X, imagem por ressonância magnética funcional, EMA, EMG, EEG, entre outras técnicas (DENBY et al., 2010).

Além disso, tem havido uma exploração significativa no uso de sistemas invasivos com implantes cerebrais para recuperar os sinais gerados durante a fala subvocal. O uso de matrizes de microeletrodos intracorticais mostra resultados promissores em testes em estágio inicial com o objetivo de fornecer uma prótese de fala. Da mesma forma, o método

de sintetizar fala estável a partir de sinais de eletrocorticografia (ECoG) de alta densidade também tem sido explorado.

Também houve um grade número de pesquisas de sistemas não invasivos, com a esperança de utilização além do uso clínico. Diversos trabalhos usam um sistema não invasivo para transcrever sinais EMG da fala subvocal para o texto, alcançando uma taxa de erro por palavra bastante baixa em um vocabulário amplo, mas com movimentos de boca claramente visíveis. Tais trabalham foram ainda continuados, avaliando também sistemas de reconhecimento de fala subvocal contínuo, embora os testes ainda impliquem em movimentos faciais significativos e uma alta taxa de erro por palavra (WADKINS, 2019).

Pode-se observar, portanto, que pesquisadores em fonética e fonoaudiologia, juntamente com pesquisadores médicos e profissionais responsáveis por muito do que se sabe sobre as deficiências de fala, e especialistas em engenharia biomédica lançaram muitas das bases necessárias para o desenvolvimento de aplicações SSI bem-sucedidas. No entanto, apesar de terem se desenvolvido bastante ao longo dos anos, as técnicas de SSI ainda estão em processo de evolução, para resultados cada vez mais precisos e cômodos para o uso diário.

3.2.1 Tecnologias de interface de fala silenciosa

Para a obtenção de dados da fala silenciosa sem a presença de sinais acústicos audíveis, vários estudos já foram realizados com sensores e tecnologias sendo empregados. A lista a seguir cita algumas das técnicas exploradas em tais estudos.

- Eletromiografia de superfície (sEMG): opera registrando os sinais elétricos gerados pela contração dos músculos articulatórios durante a produção da fala, usando eletrodos colocados sobre superfície da pele, nos músculos da face e do pescoço (VOJTECH et al., 2021);

Figura 8 – Eletromiografia de superfície
Fonte: (WAND; JANKE; SCHULTZ, 2014)

- Interfaces cérebro-computador baseadas em eletroencefalografia (EEG), espectroscopia de infravermelho próximo (NIRS) ou implantes no córtex motor da fala: a atividade cerebral é registrada para tentar obter informações sobre o processo de produção da fala, podendo ocorrer de maneira invasiva ou não invasiva, embora o segundo modo seja recomendado por envolver menos riscos (VORONTSOVA et al., 2021);

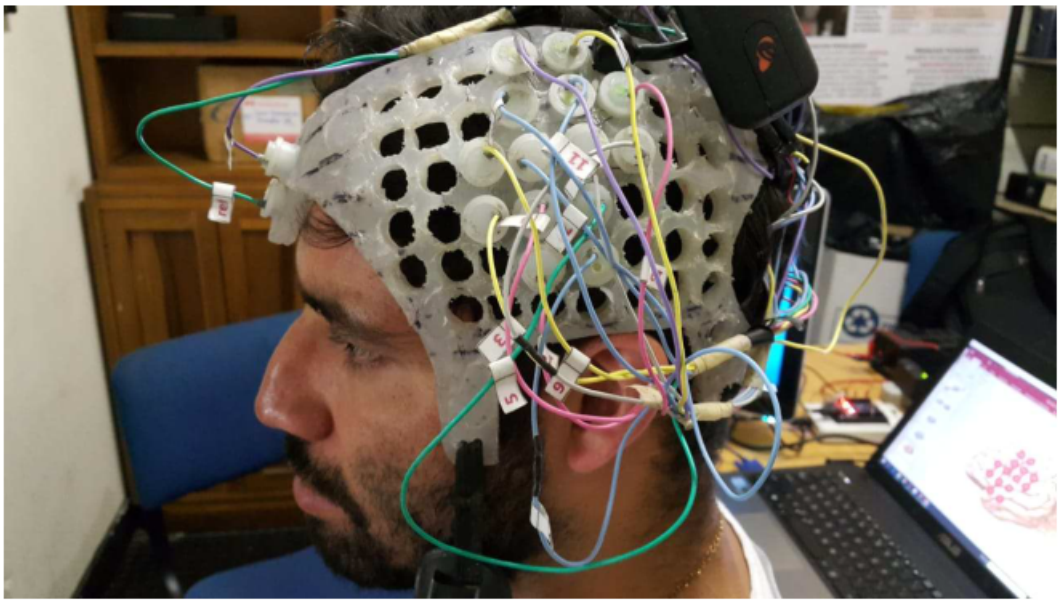

Figura 9 – Eletroencefalografia
Fonte: (DELGADO, 2020)

- Leituras labiais baseadas em câmera de vídeo: uma câmera de vídeo captura o movimento da boca e as palavras faladas são inferidas usando técnicas de processamento de imagem;

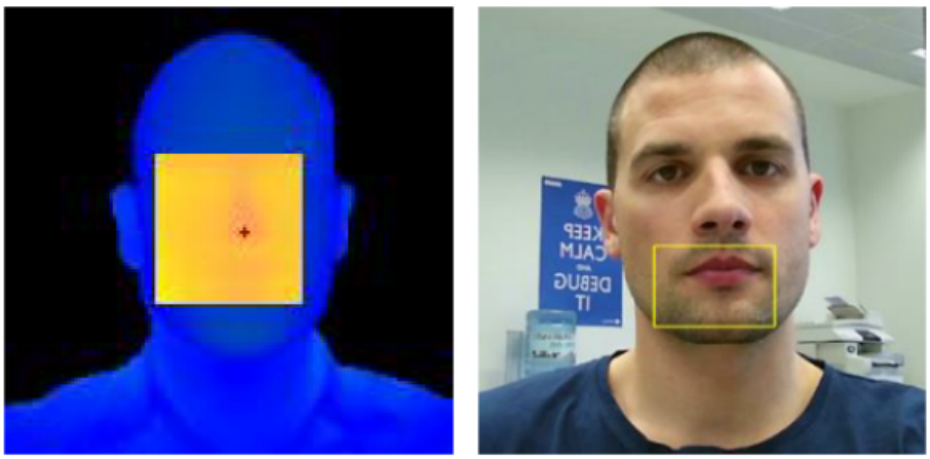

Figura 10 – Técnica de leitura labial
Fonte: (ABREU, 2014)

- Articulografia eletromagnética (EMA) ou com ímãs permanentes (PMA): sensores posicionados nos articuladores, geralmente lábios e língua, são utilizados para monitorar os movimentos com dispositivos de campo magnético (MEIRELES, 2017);

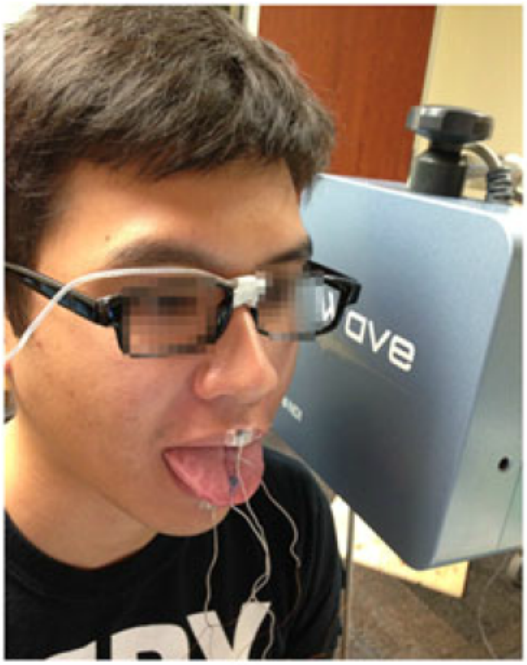

Figura 11 – Articulografia eletromagnética
Fonte: (KIM et al., 2017)

- Ultrassonografia da língua: ultrassons capturados em tempo real são utilizados para analisar o movimento da língua durante a fala com o posicionamento da sonda abaixo do queixo; a técnica costuma ser utilizada em conjunto com outras, como a leitura labial com câmera de vídeo;

Figura 12 – Técnica de ultrassonografia da língua em conjunto com a leitura labial por câmera
Fonte: (WANG; ROUSSEL; DENBY, 2021)

- Microfone de murmúrio não audível: ondas acústicas de baixa amplitude e quase inaudíveis conduzidas pelo corpo são medidas com um tipo de microfone estetoscópico posicionado na pele, atrás da orelha (NAKAJIMA et al., 2003);

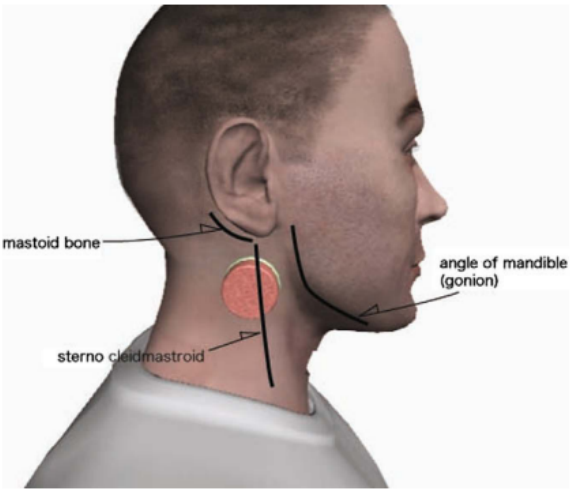

Figura 13 – Posicionamento do microfone NAM
Fonte: (TRAN et al., 2010)

- Detecção de atividade glótica com base em eletroglotografia (EGG) ou vibrometria: a atividade elétrica ou a vibração na área da laringe é medida para inferir a atividade da glote (DENBY et al., 2010);

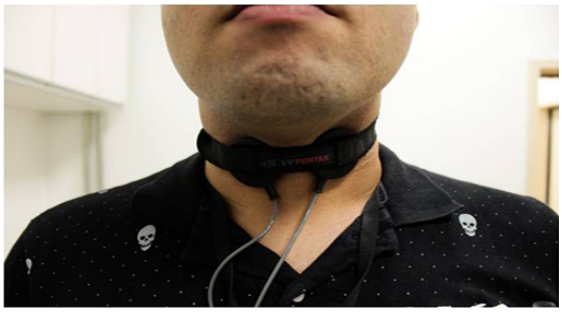

Figura 14 – Eletroglotografia
Fonte: (MATTA et al., 2021)

- Detecção baseada em radar: antenas são colocadas em frente dos articuladores visíveis ou em contato com a pele, para que os movimentos sejam inferidos a partir das ondas eletromagnéticas refletidas (WAGNER et al., 2022).

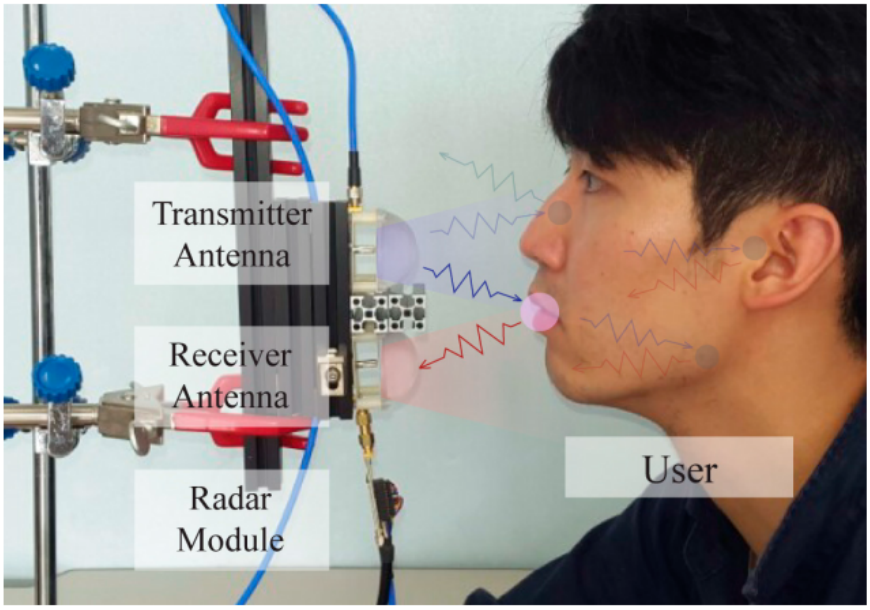

Figura 15 – Sistema baseado em radar
Fonte: (SHIN; SEO, 2016)

Todos os métodos citados, no entanto, possuem limitações, sendo necessário avaliar qual o melhor uso para as condições e resultados almejados. Métodos invasivos, por exemplo, exigem eletrodos duráveis desenvolvidos para o uso prolongado e sem prejuízos à saúde. De maneira similar, EMA e PMA demandam cuidados extras para evitar problemas de saúde por terem sensores posicionados na boca. Além disso, o posicionamento desses sensores pode acarretar também em alterações na fala, dificultando o reconhecimento dessa. Embora não seja uma solução totalmente eficaz, um dos métodos de amenizar esse efeito, é a utilização de sensores sem fio, que também tem impacto em questões estéticas, tornado as soluções mais apropriadas para o uso cotidiano.

A respeito do contexto de uso, é importante ressaltar que algumas patologias ou condições físicas tornam algumas técnicas inadequadas, como é o caso do microfone NAM para usuários que passaram por remoção da laringe, uma vez que o fluxo de ar proveniente dos pulmões pode não produzir a ressonância necessária no trato vocal, exigindo o uso de fontes sonoras alternativas. Da mesma maneira, o uso do EGG também seria impossibilitado em usuários com tais condições.

Em todas as tecnologias apresentadas, os sensores utilizados devem ser cuidadosamente posicionados para obter a melhor resposta. Em um sistema sensível à orientação da superfície da língua em uma imagem de ultrassom, por exemplo, qualquer movimento da sonda consiste em uma mudança do quadro de referência da imagem, que deve ser levado em consideração. EMA, EMG, EGG e EEG também são sensíveis a variações no posicionamento do sensor, e os pesquisadores que usaram microfones com tecnologia NAM relataram a necessidade de encontrar “pontos ideais” para colocar seus dispositivos

e obter melhores resultados.

Outra preocupação comum é a independência do orador. Apesar dos avanços no reconhecimento de fala baseado em áudios na questão de independência do locutor, sistemas que dependem de anatomia do orador ou de codificação sináptica exata inerente ao movimento dos músculos articulatórios ainda necessitam de mais desenvolvimentos (DENBY et al., 2010). Os rostos humanos têm características complexas, como geometria de superfície, suavidade, comportamentos dinâmicos e deformações. Muitos dos eletrodos empregados atualmente, portanto, não obedecem às texturas da pele, formando lacunas instáveis e reduzindo correspondentemente a relação sinal-ruído. Por mais que seja possível contornar o problema com o uso de áreas de contato maiores e materiais de forte adesão para enrolar os eletrodos, tais soluções restringem severamente os movimentos dos músculos e podem causar desconforto, provocando até alterações na fala (WANG et al., 2021).

Em relação às alterações na fala, é importante destacar que as pessoas tendem a articular suas palavras de maneira diferente quando não conseguem ouvir claramente sua própria fala, especialmente em ambientes ruidosos, um fenômeno conhecido como o efeito Lombard. Isso pode representar um desafio para os SSIs, a menos que possam fornecer um áudio de alta qualidade e sincronizado por meio de fones de ouvido. Além disso, o uso de sistemas de *feedback* auditivo pode trazer benefícios adicionais em certas situações. No contexto do EEG, estudos demonstraram que essa técnica pode induzir mudanças nas ondas cerebrais dos usuários para tornar a compreensão das palavras mais fácil (BIRBAUMER et al., 2000).

Além do efeito Lombard, contudo, existe a questão adicional de saber se os falantes se articulam de maneira diferente quando falam silenciosamente, seja em silêncio ou em ambientes ruidosos, e a maioria das indicações é que a articulação silenciosa e vocalizada não são de fato idênticas. Em qualquer caso, a melhor prática seria treinar sistemas SSI em fala silenciosa, em vez de fala audível, uma vez que este é o contexto em que eles acabarão por operar. No entanto, fazer isso é experimentalmente muito mais difícil, uma vez que a ausência de um fluxo de áudio impede o uso de ferramentas de reconhecimento automático de fala padrões para rotular e segmentar os dados captados pelos sensores, além de dificultar o desenvolvimento de um sintetizador de fala de saída para a SSI.

Visando o uso diário, existem ainda alguns pontos de melhoria a serem trabalhados, como a portabilidade das tecnologias de SSI, um bom tratamento de sinais para lidar com a sobreposição de ruídos e sinais captados, como o caso de atividades cerebrais correlacionadas ao sinal de interesse no EEG, e o vocabulário restrito. O reconhecimento de fala silenciosa ou até mesmo o reconhecimento automático de fala é uma tarefa difícil, se tratando de fala contínua e ainda mais complexa se for realizada de maneira interativa em tempo real e em sistemas portáteis. Por isso, as primeiras aplicações SSI úteis

devem se concentraram no objetivo mais facilmente realizável de reconhecimento de fala de vocabulário limitado. Um comum desafio para todas as tecnologias potenciais é então a criação de dicionários que sejam de tamanho limitado, mas ricos o suficiente para serem genuinamente úteis para as tarefas e cenários de SSIs para os quais são adaptadas, por exemplo, telefonia, recursos de fala pós-laringectomia, reconhecimento de comando verbal e afins (DENBY et al., 2010).

3.3 Técnicas de reconhecimento de fala

Uma vez que os sinais de interesse são captados, por qualquer uma das tecnologias de SSI, esses precisam ser tratados para a eliminação de ruído, amplificados, transformados e/ou comparados com o vocabulário para a identificação do que foi falado. Dessa maneira, diversas técnicas podem ser empregadas em cada uma dessas etapas, sendo algumas das principais destacada na Seção 2.4 e detalhadas a seguir.

3.3.1 Detecção de atividade

A maior parte dos sistemas utiliza detecção de atividade de voz acústica, o denominado VAD, no entanto, alguns estudos focaram na detecção de atividade de fala (SAD), principalmente os que utilizam EMG. O processo de fala acontece antes da vocalização em si, com os estímulos musculares e contrações, contudo, definir o início e o fim da atividade relacionada à fala com SAD é mais difícil, considerando que as contrações musculares precedem a produção da fala por intervalos de tempo variáveis. Se tratando de sistemas que não envolvem palavras ou sentenças audíveis, o desafio é ainda maior, porque sem uma sugestão acústica, é difícil diferenciar os sinais associados com a atividade de fala daqueles associados com movimentos não verbais, exigindo a elaboração de algorítimos mais precisos e técnicas aprimoradas (MELTZNER et al., 2011).

3.3.2 Parametrização

Após a detecção da atividade de fala, é crucial proceder com a extração de informações dos sinais captados a fim de realizar a classificação. Nesse processo, são utilizadas técnicas de parametrização, amplamente empregadas no campo de aprendizado de máquina, para identificar os atributos relevantes. Com base na revisão bibliográfica, foi possível identificar algumas dessas técnicas amplamente utilizadas no âmbito do reconhecimento de fala, e a seguir, estando essas descritas sucintamente a seguir.

- Transformada Wavelet (WT - *Wavelet Transform*) : *wavelets* são ferramentas matemáticas para decompor funções hierarquicamente. Em geral, permitem a representação por funções, seja de uma imagem, uma curva, superfície ou forma de onda.

Ao contrário da transformada de Fourier, possibilita obter informação do domínio do tempo, além de informação do domínio da frequência, sendo uma boa técnica para reconstrução de imagens, por não propiciar compressão dos dados durante o mapeamento para outro domínio (STOLLNITZ; DEROSE; SALESIN, 1995);

- Transformada de Fourier de curto prazo (STFT - *Short Time Fourier Transform*): assim como a WT, a STFT difere da transformada de Fourier por fornecer as informações de frequência localizadas no tempo para situações nas quais os componentes de frequência de um sinal variam ao longo do tempo, enquanto a transformada padrão de Fourier fornece as informações de frequência calculadas em todo o intervalo de tempo do sinal. A técnica é usada principalmente no processamento de sinais de áudio (DINIZ; SILVA; NETTO, 2010);

- Parâmetros de domínio do tempo e cruzamento de zero: as análises no domínio do tempo consistem na utilização de um conjunto de funções matemáticas com relação ao tempo. Por sua vez, parâmetros de cruzamento de zero indicam mudanças de sinais de funções matemáticas, sendo assim, a contagem de cruzamentos é usada no processamento de fala para estimar a frequências fundamental da fala, enquanto no processamento de imagens são usados para identificar pontos de borda potenciais (HSUE; SOLIMAN, 1990);

- Coeficientes Mel-cepstrais: a técnica de MFCC aplica um conjunto de filtros digitais não espaçados linearmente no domínio da frequência ao espectro real do sinal, antecedendo o processo de utilização da função logarítmica. Podem ser definidos então como coeficientes derivados da representação cepstral, sendo utilizados para representar a resposta do sistema auditivo humano, devido à percepção naturalmente não linear dos sinais sonoros (RIBEIRO et al., 2014);

- Modelo de mistura gaussiana (GMM): o modelo consiste em uma função paramétrica de densidade de probabilidade, representada como a soma ponderada de densidades de componentes gaussianos. A técnica também é capaz de formar aproximações para densidades arbitrárias. Seu uso em sistemas de reconhecimento de fala, principalmente os que envolvem dados acústicos, pode ser justificado por sua capacidade de representar uma grande classe de distribuições amostrais (REYNOLDS et al., 2009).

3.3.3 Classificação

Após a parametrização dos sinais, é realizada a classificação desses, uma vez que já foram extraídas as informações que permitem uma comparação de padrões com a base dados e identificação das palavras que foram expressas naquele sinal. Abaixo estão

descritas resumidamente as principais técnicas de classificação identificadas na revisão bibliográfica.

- Modelo oculto de Markov (HMM): é um modelo probabilístico do aprendizado de máquina, destinado a detectar padrões em dados sequenciais. Possui esse nome por ser um modelo estatísticos para capturar informações ocultas de símbolos sequenciais observáveis. Sua aplicação de maior sucesso tem sido no processamento de linguagem natural, destacando-se por ter sido usado pela primeira vez para o reconhecimento de fala, sendo então uma técnica muito usada para esse fim desde a década de 80 (FRANZESE; IULIANO, 2019);

- Redes neurais artificiais (ANN): consistem de técnicas computacionais compostas por diversas unidades de processamento que estão conectadas, as camadas ocultas ou intermediárias, e que são capazes de obter conhecimento por experiência. Atualmente é muito usada para visão computacional, reconhecimento de voz, processamento de linguagem natural, entre outros. Sua versatilidade está relacionada com a existência de diferentes tipos, com unidades de processamento variadas, capazes de tornar a arquitetura dessas redes neurais especializadas para um uso específico (CASTRO; CASTRO, 2001);

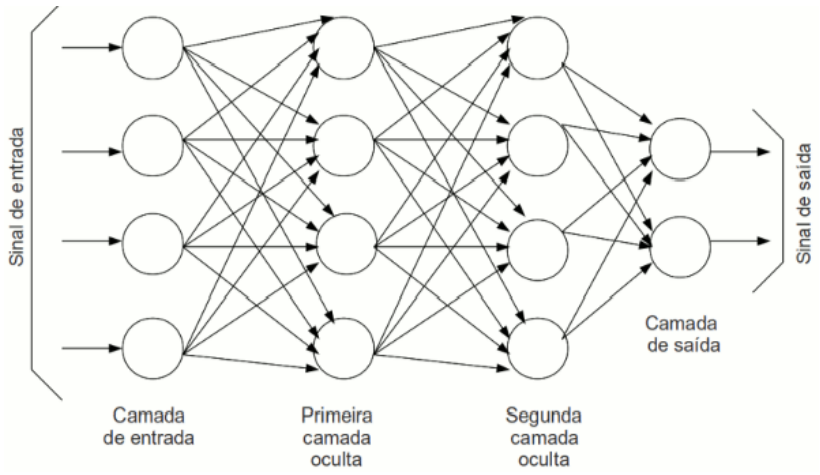

Figura 16 – Exemplo de rede neural artificial
Fonte: (MONOLITO NIMBUS, 2017)

- Máquina de vetores suporte (SVM - *Support Vector Machine*): a técnica pode ser considerada outra categoria de rede neural, sem a realimentação das unidades de processamento. Se trata essencialmente de uma máquina linear, por fazer a separação dos padrões lineares para composição do hiperplano de superfície de decisão, contudo, pode também ser usada para padrões não-lineares, com o tratamento dos dados de entrada. Também é uma técnica amplamente utilizada em visão computacional, mas também é aplicável em classificação textual (GONÇALVES, 2010);

- Análise discriminante linear (LDA): mais uma técnica de estatística e aprendizado de máquina para reconhecimento de padrões, a LDA tem como foco a redução de dimensões, podendo ser utilizada em dados com milhares de dimensões e dezenas de grupos. É empregada em conjuntos de dados lineares, o que acaba por ser seu maior limitador. Ainda assim, a técnica é muito utilizada para reconhecimento de fala e imagens, como rostos (BOELTER, 2021);

- Floresta aleatória (RF - *Random Forest*): a técnica de RF é baseada em múltiplas árvores de decisões, servindo tanto como um sistema para classificar, quanto para aproximações lineares. As árvores de decisão que compõe a floresta classificam de maneira independente, com base em aspectos diferentes de um banco de dados e esse conjunto de respostas é utilizado em uma função matemática para determinar a classificação final. Dessa forma, o método pode ser usado inclusive em dados não lineares e de maneira mais assertiva que as árvores de decisões isoladamente (BOELTER, 2021).

3.4 Considerações parciais

Essa seção teve como objetivo introduzir conceitos essenciais para o entendimento da fala silenciosa, com o funcionamento da fala, bem como explicitar algumas as técnicas levantadas na revisão bibliográficas. No que diz respeito às interfaces de fala, a simulação deste trabalho terá como foco a eletromiografia. Para processamento dos sinais, será utilizado MFCC, STFT, além de parâmetros do domínio do tempo e cruzamento de zero para a parametrização e uma rede neural artificial para o reconhecimento de fala. Também serão utilizadas outras técnicas que não foram apresentadas na presente seção por não se destacarem na amostra de publicações selecionadas na revisão bibliográfica. Dessa forma, o Capítulo 4 terá como foco o detalhamento dos recursos a serem utilizados no experimento, tanto os já abordados superficialmente nessa seção, quanto os demais.

4 Estudo de caso

Ao identificar as principais técnicas empregadas no monitoramento da fala silenciosa, foi viável selecionar aquelas mais pertinentes para uma análise aprofundada. Este aprofundamento não se limitou apenas à compreensão teórica, mas também abrangeu o planejamento experimental, visando-se a escolha do estudo de caso a ser utilizado neste trabalho.

Em relação às interfaces de fala silenciosa apresentadas, a eletromiografia de superfície foi escolhida por se tratar de um método efetivo e de fácil implementação, amplamente utilizado na literatura e com alta acurácia. Em seu trabalho, Gaddy e Klein (2020) implementaram o reconhecimento da fala silenciosa, em cenários de vocabulário aberto e fechado, com o uso tanto de fala vocalizada, quanto de fala silenciosa. Além de descreverem a metodologia empregada com detalhes, ainda disponibilizaram todos os dados obtidos experimentalmente, bem como os códigos utilizados e materiais complementares, como amostras dos resultados previstos em seu estudo. Dessa forma, seu trabalho será usado como base, para fins de realização, comparação e possíveis melhorias nos pontos destacados pelos autores. Nas seções a seguir, temos, portanto, um detalhamento das técnicas utilizadas pelos autores.

4.0.1 Eletrodos

Em seu trabalho,Gaddy e Klein (2020) utilizaram oito canais, com eletrodos de prata banhados a ouro e pasta condutiva Ten20. Foi usada uma configuração de eletrodo monopolar, com um eletrodo de referência compartilhado atrás de uma orelha. Um eletrodo conectado ao pino de polarização da placa *OpenBCI Cyton Biosensin* também foi colocado atrás da outra orelha para cancelar ativamente interferência de modo comum. A localização dos eletrodos está descrita na Tabela 1 e pode ser observada na Figura 17.

Tabela 1 – Localização dos eletrodos

	Localização
1	bochecha esquerda, logo acima da boca
2	canto esquerdo do queixo
3	abaixo do queixo, para trás 3 cm
4	garganta, 3 cm à esquerda do pomo de Adão
5	maxilar médio direito
6	bochecha direita, logo abaixo da boca
7	bochecha direita, a 2 cm do nariz
8	atrás da bochecha direita, 4 cm na frente da orelha
ref	abaixo da orelha esquerda
bias	abaixo da orelha direita

Fonte: (GADDY; KLEIN, 2020), traduzido pela autora

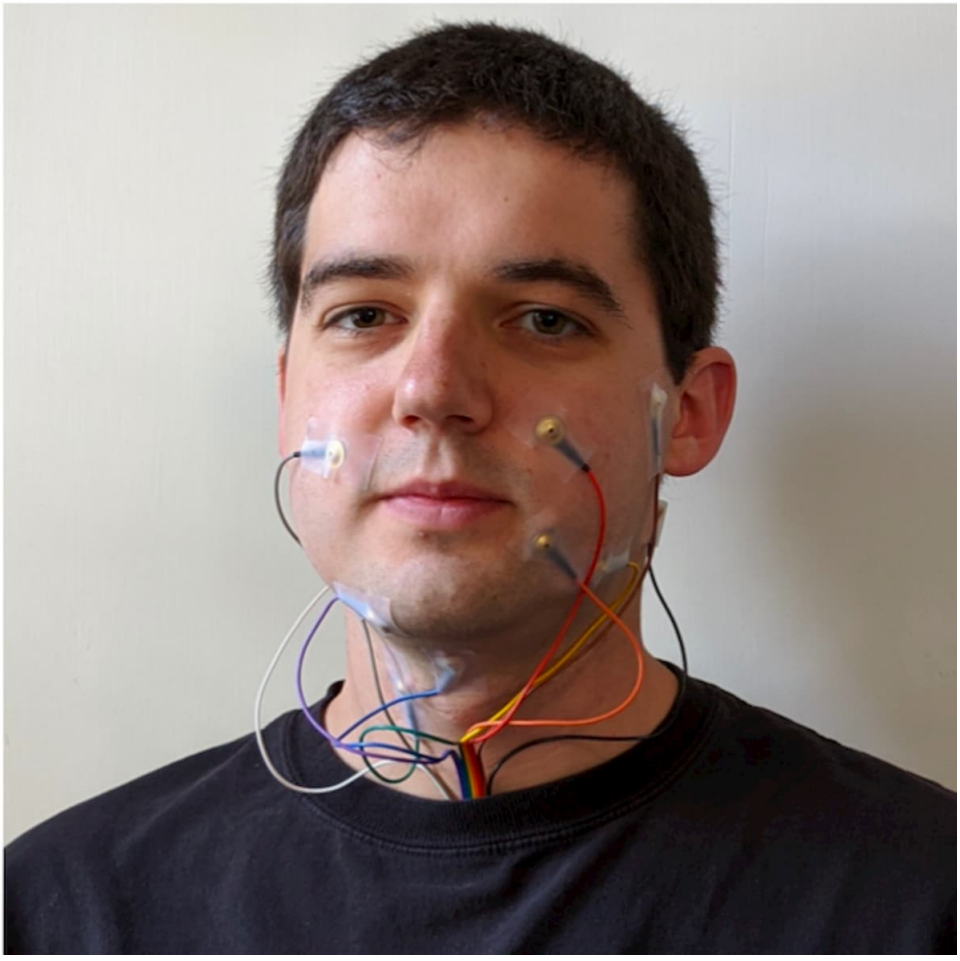

Figura 17 – Posição dos eletrodos no rosto e pescoço
Fonte: (GADDY; KLEIN, 2020)

A utilização do ouro na composição do eletrodo pode garantir maior condutividade, em comparação com os eletrodos Ag/AgCl comumente utilizados. Porém, além de possuir um custo mais elevado, o ouro é um elemento mais moldável, podendo ter alterações de

forma durante o uso. Acredita-se portanto que esse foi o motivo para o uso de eletrodos de prata banhados a ouro, em vez de eletrodos de ouro puramente ditos.

4.1 Coleta de dados

A coleta de dados do estudo de Gaddy e Klein foi realizada através de sinais alinhados no tempo de EMG e de áudio, captado de um único alto-falante, tanto durante a fala silenciosa, quanto durante a vocalizada. A porção primária do conjunto de dados consistiu em paralelos entre os dados silenciosos e os vocalizados, onde os mesmos enunciados foram gravados usando ambos os modos de fala. Esses exemplos podem ser vistos como tuplas (E_S, E_V, A_V, T) de EMG silencioso, EMG vocalizado, áudio vocalizado e o *prompt* do texto, onde E_V e A_V são alinhados no tempo. Ambos os modos de fala de um enunciados foram coletados em uma única sessão para certificar que a colocação dos eletrodos fosse consistente entre eles. Para alguns enunciados, foi registrado apenas o modo de fala vocalizada, essas instâncias são referidas como dados não paralelos e representadas com a tupla (E_V, A_V, T). Os exemplos foram segmentados no nível do enunciado. O texto lido foi incluído com cada instância no conjunto de dados e usado como referência ao avaliar a inteligibilidade posteriormente.

Para comparação, foram registrados dados de dois domínios: um vocabulário fechado e uma condição de vocabulário aberto, que são descritos nas seções a seguir.

4.1.1 Condição de vocabulário fechado

Como outros sinais relacionados à fala, o sinal capturado no EMG de um determinado fonema pode parecer diferente dependendo de seu contexto. Por esta razão, os experimentos iniciais usaram um conjunto de vocabulário mais focado antes de expandir para um vocabulário maior na etapa seguinte.

Para criar uma condição de dados de vocabulário fechado, foi gerado um conjunto de expressões de data e hora para leitura. Essas expressões são provenientes de um pequeno conjunto de modelos como "`<dia da semana> <mês> <ano>`" que são preenchidos com valores selecionados aleatoriamente (mais de 50.000 enunciados únicos são possível a partir deste esquema). A Tabela 2 resume as propriedades dos dados coletados nesta condição no estudo. Um conjunto de validação de 30 enunciados e um conjunto de teste de 100 enunciados foram selecionados aleatoriamente, deixando 370 enunciados para treinamento.

4.1.2 Condição de vocabulário aberto

A maioria das frases selecionadas para o vocabulário aberto foi coletada a partir de livros. Ao contrário dos dados de vocabulário fechado que são coletados em uma única

Tabela 2 – Resumo de dados de vocabulário fechado

Condição de vocabulário fechado
Fala paralela silenciosa/vocalizada
(E_S, E_V, A_V)
26 minutos em silêncio / 30 minutos vocalizado
Sessão única
500 enunciados
Média de 4 palavras por enunciado
67 palavras no vocabulário

Fonte: (GADDY; KLEIN, 2020), traduzido pela autora

sessão, os dados de vocabulário aberto foram divididos em várias sessões em que os eletrodos foram recolocados antes de cada sessão, o que pode ter implicado em pequenas alterações de posição entre as diferentes sessões. Além disso, há sessões com enunciados paralelos silenciosos e vocalizados, também foram coletados dados de sessões não paralelas com apenas enunciados vocalizados. Um resumo do conjunto de dados é mostrado na Tabela 3. Foram selecionados uma validação e um teste, definidos aleatoriamente a partir dos dados de EMG silencioso paralelo, com 100 e 200 enunciados, respectivamente. Durante o teste, foram usados apenas os sinais E_S do EMG silencioso, sendo descartadas as gravações vocalizadas do enunciados de teste, por não serem significativas.

A variedade de palavras abrangidas na condição de vocabulário aberto propicia que nem todas as palavras usadas durante a etapa de teste e validação tenham sido aprendidas durante o treinamento, aumentando a dificuldade de previsão do modelo.

Sistemas de reconhecimento de fala podem ser classificados em 4 tipos, de acordo com a robustez do sistema e o nível de compreensão da fala natural: palavras isoladas, palavras conectadas, fala contínua e fala espontânea, estando esses listados do mais limitado ao mais inclusivo, em se tratando da compreensão da maioria. Os níveis de palavras isolada e palavras conectadas diferem entre si pelo fato do segundo reconhecer várias palavras em sucessão com pausas suficientes entre elas. Tal tipo, no entanto, não configura uma fala natural. Por sua vez, a fala contínua não necessita dessas pausas, sendo mais aplicável no contexto cotidiano. Contudo, a fala contínua, ao contrário da fala espontânea, abrange apenas um vocabulário específico, com as sentenças selecionadas pelos desenvolvedores e passadas para os sistemas de reconhecimento de fala (SAKSAMUDRE; SHRISHRIMAL; DESHMUKH, 2015). Dessa forma, o trabalho feito por Gaddy e Klein pode ser enquadrado no tipo de fala contínua, equivalente à maior parte dos estudos feitos na área.

Dada a variação nos sinais eletromiográficos durante a fala silenciosa e a fala audível, influenciada, por exemplo, pelo efeito Lombard, é benéfico que o estudo tenha coletado

Tabela 3 – Resumo de dados de vocabulário aberto

Condição de vocabulário aberto
Fala paralela silenciosa/vocalizada
(E_S, E_V, A_V)
3,6 horas em silêncio / 3,9 horas vocalizado
Sessão média com 30 min. de cada modo
1588 enunciados
Fala vocalizada não paralela
(E_V, A_V)
11,2 horas
Sessão média com 67 min. de duração
5477 enunciados
Total
18,6 horas
Média de 16 palavras por enunciado
9828 palavras no vocabulário

Fonte: (GADDY; KLEIN, 2020), traduzido pela autora

dados de ambas as modalidades de fala, juntamente com o áudio correspondente. Além de permitir a comparação entre esses dois tipos de fala, a inclusão dos sinais vocalizados também pode ser vantajosa para aprimorar o reconhecimento da fala silenciosa.

4.2 Amplificação do sinal

A energia produzida em um músculo, conhecida como atividade elétrica muscular, tem um valor muito pequeno, medido em microvolts. Portanto, para visualização, o sinal EMG, uma vez detectado pelos eletrodos, deve ser amplificado. Durante as primeiras décadas de uso da EMG, todo o sinal capturado pelos eletrodos, incluindo atividade muscular (EMG) e atividade eletromagnética externa (ruído), foi amplificado. Por causa disso, a quantidade de interferência externa era muito grande, ou seja, a proporção de ruído no sinal era considerável. Este fato exigia que a coleta ocorresse em local especial ("gaiola de cobre") com mínima interferência de sinais eletromagnéticos externos. Em 1950, a engenharia biomédica introduziu esquemas de amplificação diferencial e rejeição de modo comum, fazendo com que o uso de EMGs não mais se limitasse a "gaiolas de cobre".

Para se realizar a amplificação diferencial, três eletrodos são utilizados: dois eletrodos para detecção do sinal e um de referência. Os eletrodos de detecção são posicionados no músculo de interesse, seguindo a orientação das fibras, e o de referência (ou eletrodo terra), em contato com qualquer proeminência óssea do corpo. Dessa forma, é possível comparar a bioenergia que atinge o eletrodo sensor (potencial de ação da unidade motora

+ sinal eletromagnético externo) com a energia que chega ao eletrodo de referência (sinal eletromagnético externo). Apenas a energia específica do eletrodo sensor (PAUM) é amplificada e registrada (amplificação diferencial).

O processo de amplificação diferencial é possível porque quando os eletrodos são colocados paralelamente às fibras musculares, os potenciais de ação resultantes são alcançados pelas unidades motoras em momentos diferentes. Desta forma, a energia detectada por cada eletrodo é diferente. Os sinais exclusivos de cada eletrodo são amplificados e a energia compartilhada pelos eletrodos de detecção e referência - o modo comum - é removida do processo (OCARINO et al., 2005).

Além da utilização de uma técnica de cancelamento da interferência do modo comum com a utilização do eletrodo, Gaddy e Klein não citam o uso de nenhuma outra técnica. Contudo, a placa OpenBCI Cyton utilizada (Figura **??**) possui o conversor ADS1299 da *Texas Instruments*, com amplificador de ganho programável com as seguintes opções de ganho: 1, 2, 4, 6, 8, 12 e 24, entre outras funcionalidades (DATASHEET..., 2017).

4.3 Pré-processamento do sinal

Antes de serem analisados, os dados eletromiográficos passam por uma série de processamentos. Pesquisadores desenvolveram métodos de processamento com o objetivo de minimizar interferências de outras fontes que não o campo eletromagnético externo, e de permitir a quantificação do sinal da EMG.

A primeira etapa de processamento é a filtragem do sinal. A maioria dos instrumentos de EMG de superfície possui filtro de 60 Hz (muito usada para o funcionamento de lâmpadas e outros equipamentos elétricos), denominado "*notch filter*". Ele pode ser encontrado no circuito eletrônico do instrumento (filtro analógico) ou no software por ele utilizado (filtro digital). O "*notch filter*" é um filtro de rejeição de uma banda de frequência específica (de 59-61 Hz). A finalidade deste filtro é remover qualquer interferência no ambiente que exceda a capacidade de rejeição do modo comum.

Outro filtro amplamente utilizado é chamado de banda passante, que permite a seleção de amplitudes de frequência específicas para análise. Por exemplo, um filtro de banda passante típico permite que toda a energia acima de 20 Hz passe e interrompe acima de 300 Hz. Este filtro geralmente é escolhido porque aproximadamente 80% da energia muscular é encontrada nesta faixa de frequência. O corte mais baixo elimina sinais de frequência muito baixa que são frequentemente associados a cabos em movimento ou outros parâmetros relacionados a movimentos mais lentos. O corte superior, por sua vez, elimina sinais com frequências maiores que a frequência de atividade muscular (OCARINO et al., 2005).

No estudo utilizado como referência experimental, é possível notar a utilização de um filtro de Butterworth passa-alta, com frequência de corte 2 Hz, para remoção da tensão de *offset* (deslocamento) e desvio nos sinais coletados. Os ruídos elétricos foram removidos com o filtro de *notch* (rejeita faixa) a 60 Hz e seus harmônicos. Adicionalmente, filtros *Forward-backward* foram usados para evitar atraso de fase. Como o áudio foi gravado a partir de um microfone de laptop embutido a 16kHz, o ruído de fundo foi reduzido usando um algoritmo de disparo espectral e o volume foi normalizado entre as sessões com base nos níveis RMS (*root-mean-square* - valor médio quadrático) (GADDY; KLEIN, 2020).

4.4 Reconhecimento da fala

Em seu trabalho, Gaddy e Klein utilizaram dois métodos para o reconhecimento de fala, sendo o primeiro o modelo baseado em transdução neural recorrente de recursos EMG para recursos de áudio alinhados no tempo, utilizando dessa forma os 3 sinais captados experimentalmente (áudio e EMG vocalizado, bem como o EMG Silencioso). Além desse, realizaram o reconhecimento de fala através das características EMG captados diretamente. Em ambos os métodos, no entanto, os dados de fala vocalizada foram utilizados (tanto das gravações vocalizadas paralelas, quanto das não paralelas) para garantir maior robustez ao treinamento do modelo, garantindo melhores resultados nas etapas de teste e validação.

Também é importante ressaltar que o trabalho desenvolvido pelos autores consistiu em diversas etapas, sendo apresentadas versões preliminares em artigos, o primeiro de 2020 e o segundo de 2021, e uma versão completa na tese de doutorado de David Gaddy, divulgada em 2022, trazendo maiores informações sobre a versão apresentada no segundo artigo. As divergências entre os modelos apresentados também serão abordadas a seguir.

No método de transdução, utilizaram um conjunto de enunciados obtidos tanto no modo de fala silenciosa, quanto na vocalizada, associando características de fala da instância vocalizada (A_V) com os sinais de EMG (E_S e E_V) através de alinhamentos entre as duas gravações, sendo utilizada a mesma taxa de quadros. O alinhamento foi inicialmente encontrado usando a distorção dinâmica no tempo entre os sinais EMG e, em seguida, refinado usando análise de correlação canônica (CCA - *canonical correlation analysis*) e áudio previsto de um modelo parcialmente treinado.

Na versão de 2020, as características foram extraídos através de artifícios do domínio no tempo e cruzamento de zero. Também foram utilizadas STFT e MFCC para parametrização do sinal. Em termos de reconhecimento de padrões para o alinhamento no tempo entre os sinais EMG e o áudio, foi usado o algoritmo de DTW (alinhamento temporal dinâmico, do inglês *dynamic time warping*). Foi utilizada ainda a CCA para a otimização do DTW e a arquitetura do projeto consistiu em um LSTM bidirecional. Ao

fim da síntese da fala, foi gerado ainda um áudio de saída com o uso do decodificador *WaveNet* (GADDY; KLEIN, 2020).

Já na segunda versão do modelo, a arquitetura da rede neural se baseou em um conjunto de blocos convolucionais residuais seguidos por uma camada *Transformer*. As camadas convolucionais foram utilizadas para extrair as características do EMG através do aprendizado, substituindo o método de extração manual anterior. Como não foi comprovada nenhuma contribuição aos resultados com o uso da técnica de CCA em conjunto com a de DTW, foi realizado apenas o alinhamento dinâmico no tempo. Também acrescentaram um sistema preditivo auxiliar de perda de fonema (GADDY; KLEIN, 2021).

Por sua vez, o modelo apresentado na tese de doutorado mantém muitas das características do apresentado em 2021, sendo a principal modificação o uso do *HiFi-GAN vocoder* no lugar do *WaveNet*. Nesse trabalho também é apresentado o modelo de reconhecimento de fala por EMG, sem utilizar os recursos de áudio. Para tal, o modelo de arquitetura, com o uso de *Transformer* foi aproveitado, sendo implementado um *softmax* sobre o vocabulário de caracteres no lugar da projeção linear para características de fala. Para o reconhecimento de fala foi utilizado um modelo de predição de caracteres treinado com uma perda de classificação temporal conexionista (CTC, do inglês, *connectionist temporal classification*) (GADDY, 2022).

Esses métodos serão melhor abordados nas subseções seguintes.

4.4.1 STFT

A transformada de Fourier de curto prazo, como dito na Seção 3.3.2, é uma ferramenta utilizada principalmente no processamento de áudio, isso por definir uma classe de distribuições tempo-frequência que especificam amplitude complexa pelo tempo e frequência para qualquer sinal. As principais aplicações da STFT geralmente incluem aproximar a análise de tempo-frequência realizada pelo ouvido para fins de exibição espectral e medir os parâmetros do modelo em um espectro de tempo curto. Uma generalização da STFT foi proposta por Dennis Gabor e pode ser descrita como (DINIZ; SILVA; NETTO, 2010):

$$X_F(\Omega_0, b) = \int_{-\infty}^{\infty} x(t)g(t-b)e^{-j\Omega_0 t}\mathrm{d}t \tag{4.1}$$

A STFT é equivalente à transformada de Fourier da função janelada $x(t)g(t-b)$. A janela $g(t)$ é em geral "concentrada" em torno de $t = 0$, e sua finalidade é isolar os valores da função $x(t)$ em torno de $t = b$ antes do cálculo da transformada de Fourier. A STFT possui duas variáveis independentes: a frequência e a posição b da janela de dados. Para cada valor de b, a transformada fornece o conteúdo espectral $X_F(\Omega_0, b)$ de $x(t)$ em torno de $t = b$ (DINIZ; SILVA; NETTO, 2010).

No trabalho de Gaddy e Klein, a STFT foi usada em conjunto com a extração das características com artifícios do domínio no tempo e cruzamento de zero. Após ser realizada a extração no domínio do tempo, obtiveram 16 amostras para cada janela de 27 ms, para essas, foi calculado um intervalo de 16 pontos STFT, o que forneceu 9 características adicionais (GADDY; KLEIN, 2020).

4.4.2 MFCC

Os coeficientes cepstrais de frequência mel são representações paramétricas de sinais acústicos, resultantes de uma transformada de cosseno do logaritmo real do espectro de energia a curto prazo, expresso em uma escala de frequência mel. Antes do uso da função logarítmica, é aplicado um banco de filtros digitais não espaçados linearmente no domínio do tempo ao espectro real do sinal, o que torna os coeficientes mel-cepstrais diferentes dos coeficientes cepstrais propriamente ditos. Para calcular o MFCC, é necessário seguir as quatro etapas (RIBEIRO et al., 2014):

- Usar o módulo da transformada de Fourier ($|FFT(x(n))|^2$) para calcular o espectro de magnitude do sinal, x(n);
- Aplicar o banco de filtros trianguladores em escala mel. O espaçamento entre esses filtros digitais, no entanto, deve respeitar a escala Mel, podendo ser usada a função de mapeamento da frequência acústica (f) em Hz para a escala de frequência em mels mostrada na Equação 4.2. Onde F_{mel} é a frequência percebida, em mels e F_{linear} é a frequência linear, em Hz;

$$F_{mel} = 2595 \log_{10}(1 + \frac{F_{linear}}{700}) \tag{4.2}$$

- Obter o cepstro calculando o logaritmo da energia de saída de cada filtro;
- Calcular os coeficientes MFCC de acordo com a Equação 4.3

$$c_{mel}(n) = \sum_{k=1}^{N_f} \log(S_f(k)) \cos[n(k - \frac{1}{2})] \frac{\pi}{N_f} \qquad n = 0, 1, \ldots, N_f \tag{4.3}$$

onde N_f é a quantidade de filtros utilizados, $c_{mel}(n)$ é o n-ésimo coeficiente mel-cepstral e $S_f(k)$ é a saída do banco de filtros, obtida com a Equação 4.4

$$S_f(k) = \sum_{j=1}^{NFFT} W_k(j) X(j) \qquad k = 1, \ldots, N_f \tag{4.4}$$

onde $W_k(j)$ representa as janelas de ponderação triangulares associadas às escalas Mel e $X(j)$ é o espectro de magnitude da FFT de N pontos (NFFT).

4.4.3 LSTM bidirecional

A memória longa de curto prazo é uma variação da RNN, a rede neural recorrente, dessa forma, primeiramente serão elucidadas as características da RNN para posteriormente serem tratados os atributos da LSTM e da LSTM bidirecional.

As RNN são redes neurais artificiais utilizadas para a identificação de padrões em dados sequenciais, como textos e áudios, por exemplo. Elas possuem dimensão temporal, utilizando memórias ou estados para processar as sequências de informações de entrada em relação ao tempo. Dessa forma, são muito usadas quando o contexto é importante, ou seja, quando decisões de iteração ou amostras passadas podem influenciar as atuais. Para isso, as informações da sequência de entrada são difundidas para a RNN individualmente a cada passo de tempo. Os dados são acumulados e cada célula além de se conectar com outra, ainda é realimentada, possibilitando assim a classificação e a previsão das informações sequenciais.

A LSTM, por sua vez, memoriza as informações em intervalos arbitrários, sendo indicada para casos em que é necessária a classificação e a predição em sequencias temporais com espaços de tempo desconhecidos. A composição da LSTM conta com uma cadeia de quatro segmentos de células, além de três camadas de redes neurais, denominadas portões e são ativadas com uma função sigmoide de valores entre 0 e 1. A Figura 18 representa uma única célula LSTM, podendo ser observados os três portões, denominados *forget*, *input* e *output* (NIELSEN, 2015).

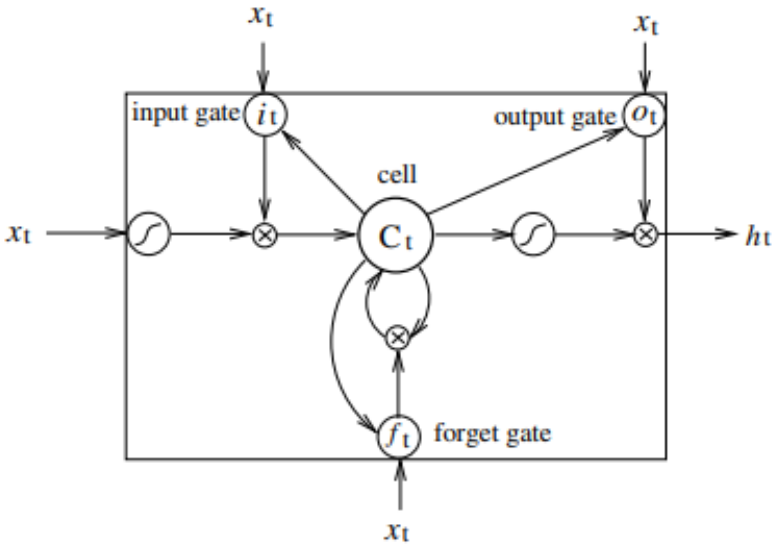

Figura 18 – Célula de uma LSTM
Fonte: (HUANG; XU; YU, 2015)

O *forget gate* é responsável por eliminar os dados que forem considerados inúteis. As duas entradas, x_t (entrada no instante específico) e h_{t-1} (saída da célula em um instante anterior), que alimentam o portão são multiplicadas por matrizes de peso e somadas ao

bias. O resultado passa pela função de ativação cuja saída é binária, como mencionado. Caso a saída seja 0, a informação é descartada, caso seja 1, a informação é mantida por ser julgada relevante.

O *input gate* adiciona as informações úteis ao estado da célula. A função sigmoide é utilizada para regular as informações, filtrando os valores a serem lembrados, de maneira análoga ao *forget gate*, com o uso das entradas x_t e h_{t-1}. Posteriormente, é criado um vetor através da função *tanh*, que fornece saídas entre -1 e 1, abrangendo todos os valores possíveis de x_t e h_{t-1}. A saída da função sigmoide é multiplicada pelos valores do vetor, para se obter quais são informações consideradas úteis do estado da célula.

Por sua vez, o *output gate* extrai as informações úteis filtradas do estado da célula atual e determina qual será a saída. De forma análoga ao que acontece no portão de *input*, a função sigmoide multiplica os valores do vetor gerado pela função *tanh*, filtrando as informações que serão apresentadas como a saída da célula atual (h_t) e entrada para próxima, o h_{t-1} da célula seguinte.

Matematicamente, o funcionamento da LSTM pode ser representado pelas Equações 4.5 até 4.9 a seguir (NIELSEN, 2015):

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{4.5}$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{4.6}$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{4.7}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{4.8}$$

$$h_t = o_t \tanh(c_t) \tag{4.9}$$

Onde f, i e o são as funções dos portões *forget*, *input* e *output*, respectivamente, c é o estado da célula, σ é a função sigmoide e h é a saída da célula. W representa a matriz que armazena os pesos dos portões, b_f, b_c, b_i e b_o são os bias.

A LSTM bidirecional, no entanto, é capaz de processar os dados em duas direções, conforme mostrado na Figura 19, ou seja, as informações de entradas são processadas em duas camadas, uma avançando e outra retornando, de forma com que o treinamento da rede seja aprimorado (NIELSEN, 2015).

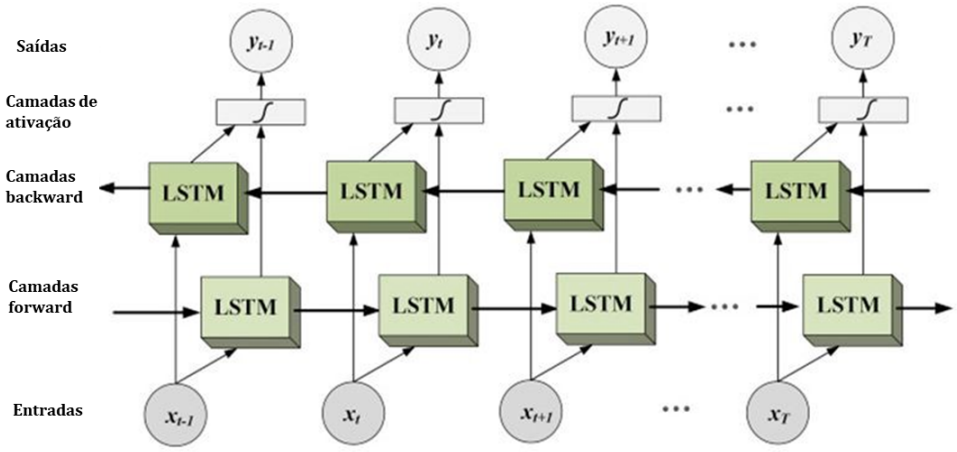

Figura 19 – LSTM bidirecional
Fonte: (SAMPAIO, 2022)

Gaddy e Klein (2020) utilizaram um modelo com 3 camadas de LSTM bidirecionais, cada uma contendo 1024 unidades ocultas, seguidas por uma projeção linear para a dimensão das características de fala, conforme ajuste realizado o conjunto de validação. Também aplicaram um *dropout* de 0,5 entre todas as camadas de LSTM, bem como antes da primeira camada e após a última camada de LSTM.

4.4.4 DTW

O algoritmo DTW (*Dynamic Time Warping*) é baseado em programação, com abordagem de resolução de problemas e é usado para medir a similaridade entre duas sequências, que podem variar no tempo ou no espaço. A similaridade é medida pelo cálculo da distância entre duas séries temporais. Como qualquer algoritmo baseado em programação dinâmica, a computação necessária para tal, é polinomial por natureza. O DTW foi originalmente desenvolvido para reconhecimento de fala, mas foi aplicado em vários contextos como biometria, verificação de impressões digitais, aplicativos de escrita, reconhecimento de formas, mineração de dados e outros (YADAV; ALAM, 2018).

No estudo de Gaddy e Klein, o DTW foi usado para encontrar o alinhamento monotônico de custo mínimo entre duas sequências, s_1 e s_2. Dessa forma, o algoritmo construiu uma tabela $d[i, j]$ do custo mínimo de alinhamento entre os primeiros itens i de s_1 e os primeiros itens j de s_2. Embora fosse um alinhamento monotônico, uma posição i podia se repetir várias vezes com valores crescentes de j, portanto, pegaram o primeiro par de qualquer sequência para fazer um mapeamento $a_{s1s2}[i] \rightarrow j$ de cada posição i em s_1 para uma posição j em s_2. Para a transferência de áudio, usaram o DTW com $s_1 = E'_S$ e

$s_2 = E'_V$ (os sinais EMG silencioso e vocalizado já parametrizados manualmente), usando distâncias euclidianas para o custo do alinhamento, conforme a Equação 4.10.

$$\delta_{EMG}[i, j] = \|E'_S[i] - E'_V[j]\| \tag{4.10}$$

Dessa forma, o DTW resultou em um alinhamento $a_{SV}[i] \rightarrow j$, que determina a posição j em E'_V para cada i em E'_S. Com o mesmo método, também foi realizado o alinhamento de uma sequência de áudios distorcidos $\widetilde{A}'_S$ com E'_S usando $\widetilde{A}'_S = A'_V[a_{SV}[i]]$. Assim, puderam utilizar $\widetilde{A}'_S$ para cálculo de perda durante a transdução. De maneira análoga, após o início do treinamento do modelo de transdução de E'_S para as características de áudio previstas $\widehat{A}'_S$, fizeram o alinhamento entre $\widehat{A}'_S$ e as características de áudio vocalizado A_V com o custo da Equação 4.11.

$$\delta_{audio}[i, j] = \|\widehat{A}'_S[i] - A'_V[j]\| \tag{4.11}$$

Tais alinhamentos foram tratados como fixos e os erros foram retropropagados a partir das características pareadas pelo alinhamento. Como a perda de treinamento também possui distância l_2 no espaço de características de áudio, a métrica δ é usada tanto para o alinhamento em si, quanto para a perda. À medida que o treinamento progride e as previsões melhoram, os alinhamentos também se aprimoram, fornecendo um sinal de aprendizado mais eficaz para o modelo. O treinamento com exemplos vocalizados ajuda a iniciar o processo, uma vez que esses exemplos já possuem saídas alinhadas, assim é realizado o treinamento nos dois modos de fala simultaneamente (GADDY; KLEIN, 2020).

4.4.5 CCA

A análise de correlação canônica (CCA) é um método de correlação linear entre duas variáveis multidimensionais. A CCA faz uso de duas visões do mesmo objeto semântico para extrair a representação da semântica. A análise de correlação depende do sistema de coordenadas no qual as variáveis são descritas, portanto, mesmo que haja uma relação linear muito forte entre dois conjuntos de variáveis, dependendo do sistema de coordenadas utilizado, esta relação pode não ser visível como uma correlação. A análise de correlação canônica busca um par de transformações lineares, uma para cada um dos conjuntos de variáveis, tal que quando o conjunto de variáveis é transformado, as coordenadas correspondentes são maximamente correlacionados (HARDOON; SZEDMAK; SHAWE-TAYLOR, 2004).

Estando dois vetores (v_1 e v_2) emparelhados, a CCA encontra projeções lineares P_1 e P_2 que maximizam a correlação entre as dimensões correspondentes de $P_1 v_1$ e $P_2 v_2$. Para obter esses pareamentos iniciais exigidos pela CCA, Gaddy e Klein usaram alinhamentos

encontrados com o DTW (E'_S e E'_V) para obter as projeções P_S e P_V, usando 15 dimensões do espaço. Então, com as projeções do CCA, definiram um novo custo para o DTW, sendo esse representado na Equação 4.12.

$$\delta_{CCA}[i,j] = \|P_S E'_S[i] - P_V E'_V[j]\| \tag{4.12}$$

4.4.6 Blocos convolucionais

O modelo de blocos convolucionais é baseado na arquitetura de redes neurais convolucionais (CNN, do inglês *convolutional neural network*), que possuem esse nome justamente por serem compostas por diversas camadas que realizam a operação de convolução através de filtros convolutivos. Ao percorrer todos os pontos da entrada da rede neural, os filtros (ou *kernels*) são multiplicados com cada elemento correspondente da entrada para que a soma dos resultados represente a função de saída, que corresponde às características de entrada processadas, conforme mostrado na Figura 20. Os valores de cada filtro podem ser ajustados para obtenção de características diferentes da entrada, permitindo que essa técnica seja muito utilizada em processamento de sinais e imagens (GOODFELLOW; BENGIO; COURVILLE, 2016).

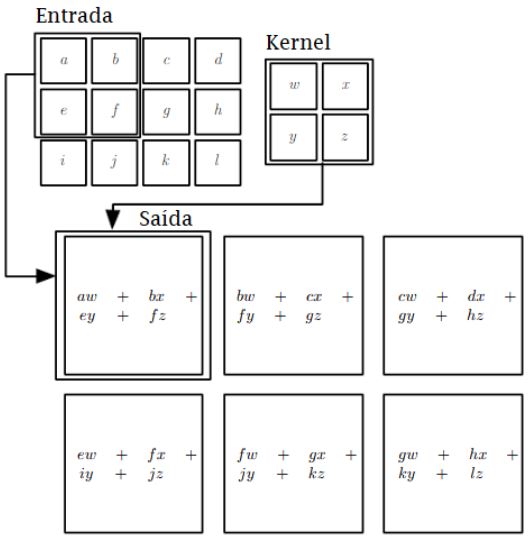

Figura 20 – Operação de convolução
Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016), traduzido pela autora

Gaddy e Klein (2021) usaram um conjunto de camadas de blocos convolucional para aprender e extrair as características de EMG. Essas camadas foram treinadas junto com o modelo de transdução, possibilitando que o modelo consiga aprender suas próprias características. A arquitetura foi formada por uma pilha de 3 blocos de convolução residuais, com condições unidimensionais. Em cada bloco, existem 2 caminhos de computação, sendo o primeiro formado por 2 camadas de convolução com largura de 3 e ativação por

unidade linear retificada (ReLU, do inglês *Rectified Linear Unit*). Já no segundo caminho, existe apenas 1 convolução de largura 1, implicando em uma transformação linear sem agregação de sequência. Cada convolução é seguida de uma operação de normalização de lote (BN, do inglês *batch normalization*). Os dois caminhos são somados ao fim, sendo utilizada então mais uma ativação ReLU, como pode ser visto na Figura 21.

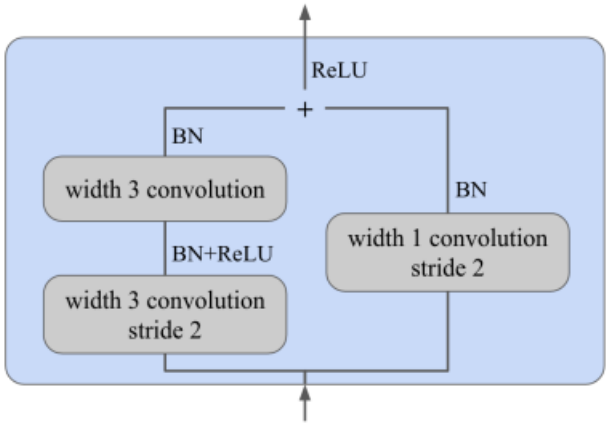

Figura 21 – Bloco convolucional
Fonte: (GADDY; KLEIN, 2021)

Para complementar o modelo, ainda implementaram no algoritmo de treinamento uma mudança aleatória dos sinais EMG em até 8 amostras, de forma com que as camadas convolucionais tivessem visões ligeiramente diferentes das entradas.

4.4.7 *Transformer*

A arquitetura *Transformer* foi empregada como alternativa à LSTM bidirecional por também ser capaz de acessar informações distantes, mas através de um mecanismo de atenção, tido como mais eficiente. O modelo proposto por Vaswani et al. (2017) está ilustrado na Figura 22, sendo composto por uma estrutura do tipo codificador-decodificador. Por não se tratar de um modelo de recorrência, as informações sobre a posição relativa ou absoluta dos itens na sequência de entrada devem ser fornecidas pelo *positional encoding*, um vetor de dimensão d, definido com o uso de seno e cosseno.

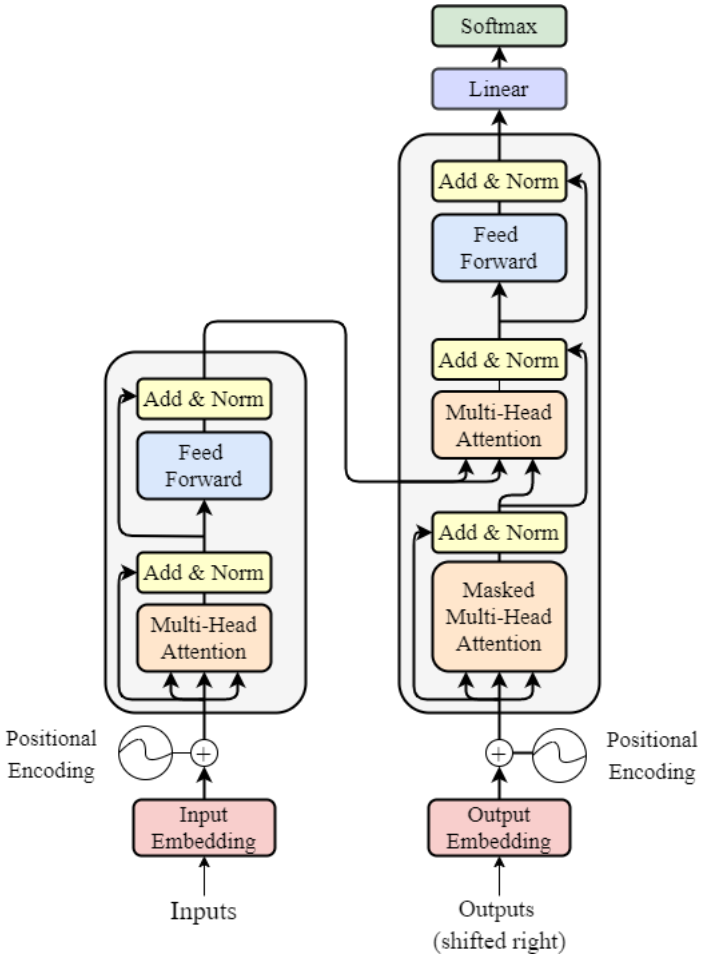

Figura 22 – Arquitetura Transformer
Fonte: (VASWANI et al., 2017)

O codificador (representado pelo lado esquerdo da figura) é formado por uma pilha de N camadas iguais, compostas por 2 subcamadas, cada. A primeira sub-camada se trata do mecanismo de auto-atenção de múltiplas cabeças (*multi-head*), que agrega informações ao logo da sequência, e a segunda, da rede *feed-forward* simples, que processa as informações em cada posição. Na saída de cada subcamada, é empregada a conexão residual, bem como a normalização da camada, para conferir maior estabilidade e otimização ao treinamento da rede neural. Dessa forma, a saída das subcamadas (*Add & Norm* na figura) pode ser caracterizada como $LayerNorm(x + Sublayer(x))$, onde $Sublayer(x)$ é a função implementada pela subcamada.

O decodificador (representado pelo lado direito da figura), por sua vez, é composto pelas mesmas subcamadas do codificador, juntamente com uma subcamada adicional, que executa o mecanismo de atenção *multi-head* na saída do codificador. A pilha do decodificador também é formada por N camadas idênticas, mas difere do codificador por não considerar posições subsequentes na subcamada de auto-atenção, além de deslocar os *embeddings* de saída por uma posição, garantindo que as previsões para a posição em questão dependa apenas das saídas conhecidas em posições anteriores (VASWANI et al.,

2017)

O mecanismo de auto-atenção em ambos os casos estima a importância que cada elemento possui em relação aos demais elementos da sequência posicional, por exemplo, a relevância de uma palavra em determinado ponto da frase, após o uso de outras palavras específicas. Para tal, o vetor de entrada do mecanismo é subdividido em outros 3 vetores com dimensão d: o vetor de pesquisa q, o de chaves k e o de valores v. Após isso, os vetores resultantes de diferentes entradas são organizados em três matrizes distintas, denominadas Q, K e V. Por fim, a função de atenção é calculada seguindo as etapas (LIMA et al., 2023):

- As pontuações entre os vetores de entrada são calculadas com a Equação 4.13. Essas pontuações determinam o nível de atenção em relação aos outros itens ao codificar o item na posição atual.
$$S = Q \cdot K^T \tag{4.13}$$
- As pontuações são normalizadas com a Equação 4.14 para melhorar a estabilidade do gradiente e aprimorar o treinamento.
$$S_n = \frac{S}{\sqrt{d_k}} \tag{4.14}$$
- As probabilidades são calculadas com a Equação 4.15, de forma com que a soma de todas as probabilidades seja 1, permitindo a interpretação como a distribuição de probabilidades.
$$P = softmax(S_n) \tag{4.15}$$
- A matriz de atenção é gerada com a Equação 4.16, garantindo que os vetores com maiores probabilidades recebam uma relevância adicional nas camadas subsequentes.
$$Z = V \cdot P \tag{4.16}$$

De maneira simplificada, o mecanismo de auto-atenção pode ser definido pela Equação 4.17.

$$Atenção(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \tag{4.17}$$

Para complementar o modelo, no entanto, foi utilizado o modelo *multi-head*, em que vários mecanismos de atenção são utilizados em vez de um único, possibilitando que a arquitetura lide simultaneamente com informações de diferentes subespaços de representação em posições distintas, melhorando o desempenho da camada de atenção. Dessa forma, cabeças diferentes utilizam diferentes matrizes de consulta, chave e valor, permitindo que essas matrizes projetem os vetores de entrada em subespaços de representação distintos.

Gaddy e Klein (2021) utilizaram 6 camadas *Transformer* com 8 cabeças, uma dimensão de modelo de 768, uma dimensão de consulta (d) de 96 e uma dimensão de *feed-forward* de 3072. Foi aplicado um *dropout* de 0,2 após a não-linearidade do *feed-forward* e nos valores de atenção. A saída da última camada *Transformer* passou por uma projeção linear final para reduzir a dimensão para 80, obtendo as previsões das características de áudio ou EMG geradas pelo modelo, a depender do tipo de método empregado para o reconnhecimedo de fala.

Para capturar a natureza invariante ao tempo, utilizaram *embeddings* de posição relativa, em vez das mais comuns, de posição absoluta. Nessa abordagem, adicionaram um vetor aprendido p aos vetores de chave durante o cálculo dos pesos de atenção, o qual depende da distância relativa entre as posições de consulta e chave. Portanto, a atenção foi calculada conforme a Equação 4.18.

$$Atenção(Q, K, V) = softmax(\frac{Q \cdot (K + p)^T}{\sqrt{d_k}}) \cdot V \tag{4.18}$$

Onde p é uma busca de *embedding* com índice $i - j$, até uma distância máxima k em cada direção. No modelo, definiram $k = 100$, dando a cada camada aproximadamente 1 segundo de visão em cada direção, e todos os pesos de atenção com distância maior que k foram definidos como zero.

4.4.8 *Vocoder*

Ao final da técnica de transdução, foi utilizado um algoritmo para processar e sintetizar os sinais de áudio, denominado *vocoder*. Sendo uma espécie de codificador de voz, o *vocoder* analisa e modifica as características do sinal de voz para criar efeitos sonoros, alterar a qualidade da voz, harmonizar e até transmitir informações de voz de maneira mais eficiente, sendo muito usados portanto para produção de áudio, música, telecomunicações e processamento de voz.

Na sua forma mais básica, o *vocoder* utiliza duas fontes de áudio: um sinal de voz, geralmente humano, e um sinal de portadora, geralmente um som musical ou ruído. O sinal de voz é subdividido em diversas bandas de frequência e essas bandas são então usadas para modular o sinal de portadora. O resultado é uma síntese sonora onde a qualidade e as características do sinal de voz são transferidas para o sinal da portadora (ANUMANCHIPALLI; CHARTIER; CHANG, 2019).

Para sintetizar áudio a partir de características de fala no primeiro modelo criado, Gaddy e Klein (2020) usaram um decodificador WaveNet, que gera a amostra de áudio por amostra condicionada às características de fala do MFCC de A'. O WaveNet é capaz de gerar fala com sonoridade bastante natural, no entanto, se trata de um modelo regressivo, gerando uma amostra de cada vez. Sendo assim, na versão final do modelo apresentada

em 2022, Gaddy optou pelo uso do *HiFi-GAN vocoder*, um modelo neural que é treinado para prever formas de onda de áudio a partir de amostras de vozes. O modelo gerador do *HiFi-GAN* é um modelo de rede neural convolucional que gera todas as amostras de áudio em paralelo.

O *HiFi-GAN* se fundamenta principalmente em um modelo de rede generativa adversarial (GAN, do inglês *Generative Adversarial Network*), em que um conjunto de discriminadores é treinado para distinguir entre amostras reais e sintéticas, enquanto o gerador gera amostras sintéticas com o objetivo de enganar os discriminadores. No contexto específico do HiFi-GAN, o sistema emprega oito discriminadores distintos, cada um operando em diferentes períodos e escalas, para produzir visualizações variadas do áudio gerado, todas as quais se assemelham ao áudio real. O algoritmo também incorpora a perda de correspondência de características dos discriminadores e a perda de reconstrução de espectrograma para aprimorar e estabilizar o processo de treinamento. Dessa forma, foi possível alcançar resultados de alta qualidade e uma saída sonora natural que corresponde ao alto-falante utilizado no experimento (GADDY, 2022).

4.4.9 CTC

A CTC (Classificação Temporal Conexionista, do inglês *Connectionist Temporal Classification*) é uma função de custo genérica que permite treinar sequências em que o alinhamento entre a entrada e a saída é desconhecido. Essa abordagem permite a sincronização temporal da sequência de saída em relação à sequência de entrada. A CTC converte uma sequência de rótulos com informações temporais em uma sequência mais curta de rótulos, removendo informações de sincronismo e alinhamento. Dessa forma, possibilita o treinamento de modelos de sequência sem a necessidade de alinhamento prévio entre os dados de entrada e saída, sendo, portanto, ideal para casos onde é necessário realizar o treinamento em dados não alinhados.

Para o treinamento do modelo, é usado um conjunto de exemplos X, onde cada componente é um par de sequências $(x : y)$. A sequência de entrada $x = (x_1, x_2, \ldots, x_T)$ possui tamanho T e pertence ao espaço de entrada X, já a sequência de saída $y = (y_1, y_2, \ldots, y_U)$ possui tamanho U e pertence ao espaço de saída Y, onde $U \leq T$. O modelo gera uma distribuição de probabilidades sobre o espaço de todos os possíveis rótulos pertencentes ao alfabeto de rótulos Y'. Essas probabilidades estimam uma distribuição sobre os caminhos π pertencentes a Y', sedo cada caminho uma saída gerada pela rede.

A sequência de saída (z) é obtida pelo mapeamento β de caminhos π para o conjunto de rótulos possíveis Y', sendo representada como $z = \beta(\pi)$. Para o mapeamento dos caminhos, é empregada a remoção de rótulos repetidos consecutivos, assim como os "vazios". Também é utilizado o algoritmo *forward-backward* para calcular o gradiente da função de perda. Por fim, é realizada a decodificação, sendo possível rotular uma sequência

de entrada x escolhendo a rotulação mais provável para cada dado (SANTANA, 2017).

Em seu trabalho, Gaddy (2022) otimizou o treinamento do CTC com a normalização do texto, convertendo-o para letras minúsculas apenas e removendo pontuações, de forma com que o vocabulário possuísse as 26 letras minúsculas do alfabeto inglês, 10 dígitos e um caractere de espaço. Durante a inferência, a busca em feixe (*beam search*) foi usada para procurar a sequência de saída com a maior probabilidade quando somada sobre os caminhos possíveis. Um modelo de linguagem também foi utilizado durante a inferência (um modelo de linguagem de 5-gramas com suavização Kneser-Ney modificada), multiplicando as probabilidades do modelo de linguagem com as probabilidades de caracteres de saída do modelo.

4.5 Considerações parciais

Nesse capítulo, foi apresentado o estudo de caso a ser desenvolvido experimentalmente, baseado no trabalho realizado por Gaddy e Klein. A Seção A descreve o funcionamento da EMG, técnica utilizada para obtenção dos dados de fala silenciosa, trazendo também informações sobre eletrodos e o posicionamento desses, não apenas em termos gerais, mas no caso específico do trabalho utilizado como referência. A Seção 4.1 é voltada para a metodologia de coleta de dados, descrevendo as limitações de vocabulário impostas e os sinais coletados, tanto na fala silenciosa, quanto na vocalizada. Na Seção 4.2 é abordada a necessidade de amplificação do sinal no uso da EMG, assim como a Seção 4.3 descreve os principais métodos para o processamento do sinal captado para eliminação de ruídos. Novamente, é feito um paralelo entre o que é recomendado na literatura e o que foi realizado por Gaddy e Klein.

Por fim, na Seção 4.4 são descritas as técnicas utilizadas no trabalho de referência para o reconhecimento da fala, tanto em termos de parametrização dos dados, quanto de classificação. Algumas dessas técnicas citadas foram abordadas anteriormente na Seção 3.3 por estarem entre as mais utilizadas nos trabalhos analisados no estudo bibliométrico.

5 Metodologia

A primeira parte desse trabalho foi dedicada ao levantamento das principais técnicas utilizadas o estudo de monitoramento de fala silenciosa, sendo escolhido um trabalho de referência para a execução do experimento proposto no Trabalho de Conclusão de Curso (TCC) 2. Neste capítulo serão apresentadas as etapas seguidas para a realização deste trabalho.

5.1 Revisão bibliográfica sobre métodos de parametrização e classificação para o reconhecimento da fala

Nesta fase, foram examinados os procedimentos e instrumentos empregados na parametrização e classificação da fala silenciosa, especialmente no contexto da coleta de dados por meio da Eletromiografia (EMG). A seção 2 do estudo trata mais pormenorizadamente dos métodos empregados na compilação da revisão bibliográfica. É importante destacar que essa revisão enfoca predominantemente a análise do estado atual das técnicas, considerando suas aplicações em outros estudos e cenários. Isso se faz necessário para melhor compreender a eficácia dessas técnicas e identificar eventuais desafios e questões a serem abordadas.

5.2 Conjunto de dados

Considerando a proposta experimental do trabalho, não foi realizada nenhuma obtenção de dados adicional, todo o conjunto de dados foi proveniente do repositório no GitHub fornecido por David Gaddy. Lá estão disponibilizadas as gravações de áudio das sessões paralelas e não paralelas em formato *FLAC*, os sinais EMG em formato *NPY* e as informações dos *prompts* de texto correspondentes em formato *JSON*.

5.3 Avaliação do código original

Nessa etapa, o código original proposto o trabalho de referência foi executado, sendo seguidas todas as instruções fornecidas, tanto de configuração do ambiente, quanto de instalação de pacotes adicionais. A avaliação é necessária para maior compreensão dos *scripts* em Python, além de permitir identificar se existem problemas no algoritmo, ferramentas ou até mesmo em versionamento dos pacotes.

5.4 Correção de problemas na execução

Com os resultados obtidos na avaliação anterior, foram levantadas maneiras de corrigir os problemas encontrados. Para essa etapa, foi preciso levar em consideração soluções que não tivessem impacto significativo nos algoritmos originais, para evitar maiores alterações nos resultados.

5.5 Implementação e avaliação dos métodos de parametrização e classificação propostos

Para a avaliação dos modelos levantados em 5.1, o conjunto de dados de sinais EMG foi empregado para o reconhecimento da fala. Por se tratar de uma verificação, apenas parte da amostra foi utilizada, sendo validado o comportamento apresentado pelos métodos individualmente e sua relevância para o contexto.

5.6 Treinamento e avaliação do modelo

Com a implementação completa do modelo proposto, este pôde ser treinado, sendo utilizada uma amostra maior do conjunto de dados. A avaliação, por sua vez, abrangeu os mesmos parâmetros utilizados no experimento original, sendo a taxa de erro por palavra o principal deles, para efeitos de comparação com as informações fornecidas.

6 Experimentos e resultados

Este capítulo apresenta a execução do experimento, bem como os resultados obtidos. No que diz respeito às ferramentas utilizadas, o trabalho foi desenvolvido utilizando a linguagem de programação *Python*, assim como no experimento de referência. Todo o código-fonte encontra-se disponível em um repositório no *GitHub* em formato de Jupyter Notebook[1].

6.1 Seleção de métodos de reconhecimento de fala silenciosa

Durante esta etapa, foram identificados os métodos que seriam capazes de executar a tarefa de reconhecimento de fala silenciosa de maneira análoga à realizada originalmente no trabalho de referência. Considerando que são feitas duas abordagens, uma com o modelo de transdução e outra com o reconhecimento através dos sinais EMG diretamente, foi selecionado apenas o segundo método, não sendo utilizada a técnica de transdução. A escolha se baseia no resultado obtido no experimento original, onde o modelo de transdução demonstrou um resultado pior, com taxa de erro por palavra (WER, do inglês *word error rate*) de 36,2%, em comparação com a taxa de 28,8% do reconhecimento do EMG direto, sendo o erro calculado através da Equação 6.1:

$$WER = \frac{S + I + E}{N} \tag{6.1}$$

onde S representa o número de substituições, ou seja, palavras erradas ou substituídas, I são as inserções, palavras que não estão presentes na referência, E são as exclusões, palavras que existem na transcrição de referência, mas estão ausentes na predição. Por fim, N representa o número de palavras na transcrição de referência.

Sendo assim, a Figura 23 ilustra a proposta experimental do presente trabalho, estando as técnicas selecionadas destacadas em azul. É importante ressaltar, no entanto, que existem diversos métodos desenvolvidos além dos dispostos no diagrama, estando listado apenas os que foram citados anteriormente no referencial teórico e os escolhidos para implementação no estudo de caso.

[1] Pode ser acessado em: https://github.com/martinssmariana/TCC2.git

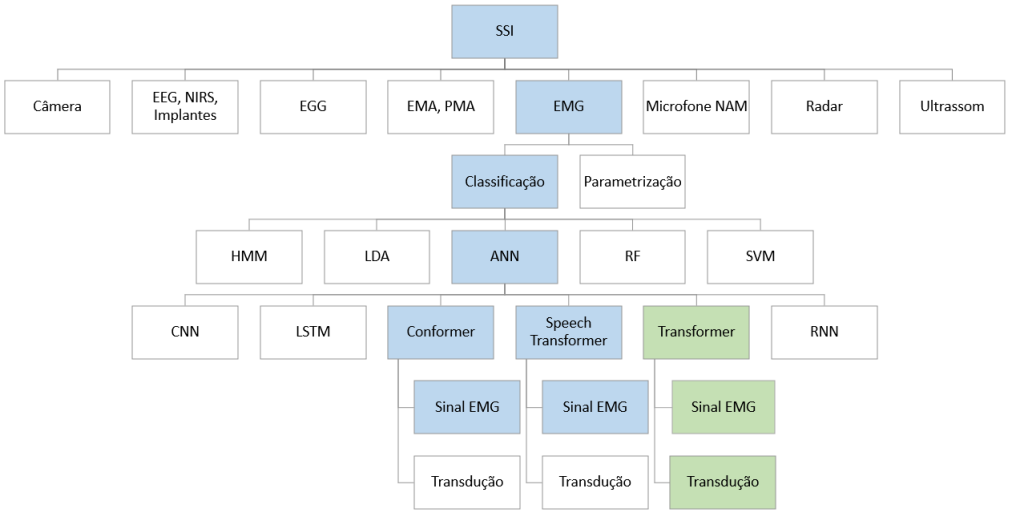

Figura 23 – Diagrama da proposta experimental
Fonte: Autoria própria

6.1.1 *Speech-Transformer*

Dong, Xu e Xu (2018) elaboraram um modelo de reconhecimento automático de fala que transforma sequências de recursos de fala para a sequência de caracteres correspondente. Para a criação da arquitetura, se basearam no *Transformer* e propuseram um mecanismo de atenção 2D, inspirado na LSTM, mas substituindo a recorrência tempo-frequência por dependências temporais e espectrais capturadas pela atenção. A Figura 24 ilustra a arquitetura do modelo, sendo o lado esquerdo referente ao codificador e o direito ao decodificador, possibilitando fazer uma comparativa com a arquitetura *Transformer* original.

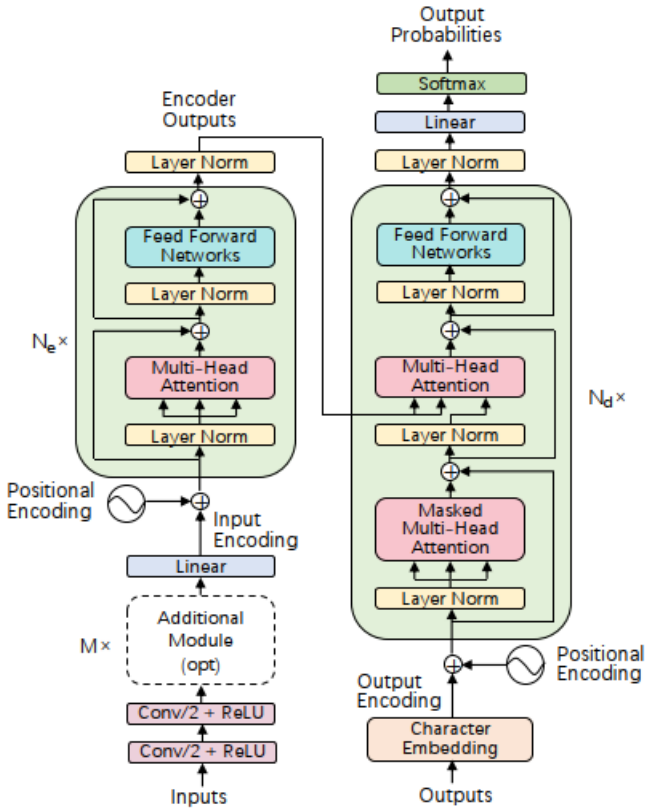

Figura 24 – Arquitetura do *Speech-Transformer*
Fonte: (DONG; XU; XU, 2018)

As redes convolucionais utilizadas no codificador do modelo têm como objetivo lidar com as discrepâncias de comprimento entre a sequência de características de fala e a de caracteres. Para tal, as entradas são representadas como espectrogramas bidimensionais com eixos de tempo e frequência, dessa forma, as redes convolucionais exploram a estrutura local dos espectrogramas, avançando ao longo do tempo. Além de produzir uma representação oculta com comprimento aproximado ao da sequência de caracteres, as duas camadas de CNNs 3x3 com passo 2 também evitam problemas de falta de memória na GPU.

Para o módulo adicional mencionado na Figura 24, os autores fizeram testes com 3 tipos diferentes de estrutura, dentre as quais o módulo de atenção 2D, representado na Figura 25 resultou em melhores resultados. Primeiramente, o módulo executa 3 redes convolucionais nos espectrogramas com n canais para extrair as representações de consultas, chaves e valores, onde os canais de saída de cada rede convolucional são c.

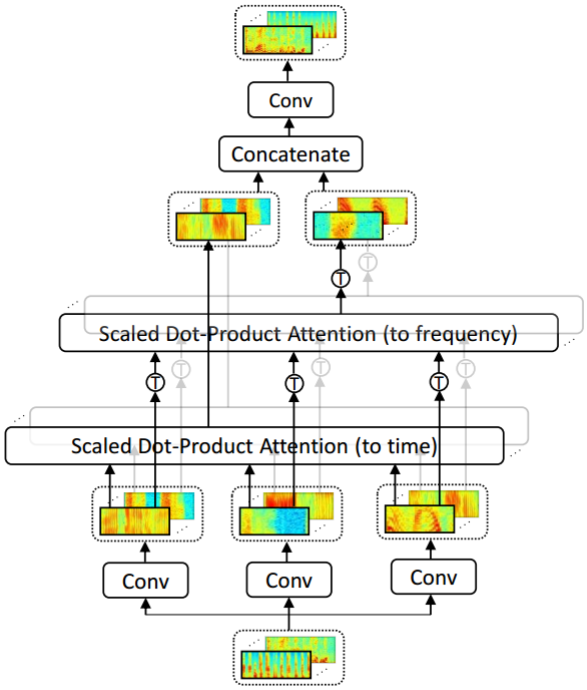

Figura 25 – Módulo de atenção 2D
Fonte: (DONG; XU; XU, 2018)

Em seguida, são introduzidos 2 tipos de atenção para capturar dependências temporais e espectrais, sendo que cada atenção lida com as consultas, chaves e valores do canal correspondente no eixo do tempo diretamente e de maneira transposta para o eixo da frequência. Por fim, as saídas da atenção 2D são concatenadas e alimentadas para outra rede convolucional, que é utilizada para as saídas finais dos n canais do bloco, de acordo com as Equações 6.2 a 6.4:

$$\begin{aligned} \textit{Atenção2D} = W^O * \text{Concat}(\textit{canal}_1^t, \ldots, \textit{canal}_c^t, \\ \textit{canal}_1^f, \ldots, \textit{canal}_c^f) \end{aligned} \tag{6.2}$$

$$\textit{canal}_i^t = \text{Atenção}((W_i^Q * I), (W_i^K * I), (W_i^V * I)) \tag{6.3}$$

$$\textit{canal}_i^f = \text{Atenção}((W_i^Q * I)^T, (W_i^K * I)^T, (W_i^V * I)^T)^T \tag{6.4}$$

onde I representa as entradas dos n canais, $*$ representa a operação convolucional, W_i^Q, W_i^K e W_i^V representam os filtros aplicados em I para obter as consultas, chaves e valores do canal i, respectivamente, por fim, W^O representa os filtros aplicados nos 2 canais c concatenados para fornecer as saídas finais dos n canais.

Após a saída do módulo, é aplicada uma transformação linear às saídas do mapa de características, resultando em vetores de dimensão d_{modelo}, intitulados codificação de entrada e acrescidos a uma codificação posicional, também de dimensão d_{modelo}, que permite que a arquitetura considere as posições relativas. No *Speech-Transformer*, é utilizado um esquema de codificação posicional que combina funções seno e cosseno para capturar as dependências temporais. Essa codificação é somada à codificação de entrada e alimentada em blocos codificadores.

O decodificador também utiliza uma codificação posicional, juntamente com atenção mascarada e atenção *multi-head*, para gerar as saídas finais. A normalização de camada e a conexão residual são aplicadas em cada sub-bloco para melhorar o treinamento. Por fim, as saídas do decodificador são transformadas em probabilidades através de uma projeção linear e uma função *softmax*.

6.1.2 *Conformer*

O modelo *Conformer* se trata de uma combinação orgânica de convoluções com auto-atenção, focado em contexto de reconhecimento automático de fala. No caso específico, a auto-atenção aprende as interações globais, enquanto as convoluções capturam eficientemente as correlações locais baseadas em deslocamento relativo. A arquitetura do *Conformer* está representada na Figura 26, estando a auto-atenção e convolução intercaladas entre dois módulos de alimentação direta (GULATI et al., 2020).

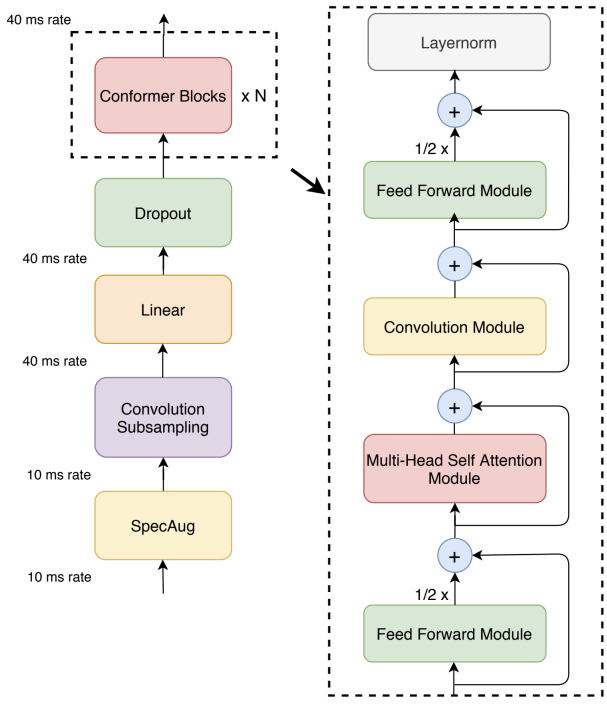

Figura 26 – Arquitetura do modelo do codificador *Conformer*
Fonte: (GULATI et al., 2020)

Para seu funcionamento, primeiramente a entrada é processada com uma camada de subamostragem de convolução e somente depois segue para os N blocos *Conformer*. Cada bloco desses é composto por 4 módulos empilhados: *feed-forward*, auto-atenção, convolução e um segundo módulo de alimentação direta, sendo que os módulos *feed-forward* seguem a proposta original do *Transformer*, com normalização de camada dentro da unidade residual e na entrada antes da primeira camada linear, mas com meio passo e ativação *Swish*.

O módulo de auto-atenção *multi-head* foi desenvolvida a partir do esquema de codificação posicional senoidal relativa, que permite que o módulo crie melhores generalizações em diferentes comprimentos de entrada, tornando o codificador mais robusto em se tratando de variação do comprimento das falas. Também foram utilizadas unidades residuais pré-normalizadas com *dropout*, para auxiliar o treinamento e a regularização de modelos mais profundos.

O módulo de convolução é baseado em uma convolução ponto a ponto e uma unidade linear com controle, além de uma única camada de convolução unidimensional por profundidade. Ao final, também é realizada uma normalização em lote, com foco no aprimoramento do treinamento de modelos profundos.

Dessa forma, a entrada x_i do bloco *Conformer* é tratada nos módulos de acordo com as Equações 6.5 a 6.7, sendo representada então como a Equação 6.8 na saída y_i do bloco:

$$\widetilde{x}_i = x_i + \frac{1}{2}\mathrm{FFN}(x_i) \tag{6.5}$$

$$x'_i = \widetilde{x} + \mathrm{MHSA}(\widetilde{x}_i) \tag{6.6}$$

$$x''_i = x'_i + \mathrm{Conv}(x'_i) \tag{6.7}$$

$$y_i = \mathrm{Layernorm}(x''_i + \frac{1}{2}\mathrm{FFN}(x''_i)), \tag{6.8}$$

onde FFN se refere ao módulo de *Feed Forward*, MHSA se refere ao módulo de *Multi-Head Self-Attention* e Conv se refere ao módulo de Convolução, conforme descrito anteriormente.

6.2 Teste do código original

Devido a limitações computacionais de *hardware*, para o teste do modelo original, foi necessário o uso de plataformas online que possibilitassem o uso de uma GPU (unidade

de processamento gráfico, do inglês *Graphics Processing Unit*) capaz de suportar a memória requisitada pela rede neural. Dessa forma, foram utilizadas as plataformas Kaggle e Google Colab para o funcionamento do experimento.

Nos testes, foi observado um consumo muito alto de memória RAM, ultrapassando os 13 GB das máquinas virtuais utilizadas, o que exigiu o uso do *Garbage Collector*, ou coletor de lixo, ferramenta que gerencia alocação e desalocação de memória na execução de programas, evitando problemas de vazamento. Também foi observado que o uso de 100% da base de dados fornecida pelos autores provocava o desligamento da máquina virtual pelo uso exacerbado da GPU. Diante desse contexto, a base de dados foi reduzida para 80% dos exemplos fornecidos, sendo utilizada uma função *subset* já fornecida no código original.

Com as mudanças citadas, além de outras pequenas modificações realizadas para adequação do código ao ambiente virtual, o código funcionou, podendo então servir de parâmetro para os novos modelos testados. No entanto, ressalta-se que o ambiente virtual, tanto do Kaggle quanto do Colab, é baseado na versão 3.10 do Python e algumas das bibliotecas utilizadas estão limitadas a suas versões mais atuais, sendo assim, havia a possibilidade de incompatibilidade entre versões dos pacotes usados. Tal problema se deu principalmente com a biblioteca NumPy, conforme mostrado na Figura 27, um *log* do sistema na execução do código. A consequência real dessa incompatibilidade não foi identificada diretamente nos resultados, no entanto.

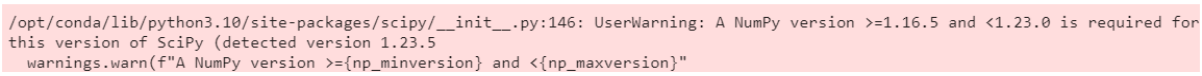

Figura 27 – Incompatibilidade de versão da biblioteca NumPy
Fonte: Autoria própria

6.3 Implementação e avaliação do *Speech-Transformer*

O código do modelo do *Speech-Transformer* implementado foi baseado no elaborado por Kim, Bae e Won (2021). Para o experimento, os *scripts* originais de Gaddy (2022) foram utilizados, para leitura do sinal EMG e preparação dos dados. No código principal do reconhecimento, a maior parte da estrutura foi mantida, sendo substituído o modelo de rede neural e apenas os pontos necessários para a execução do programa. Na simulação no Kaggle, no entanto, ocorreu o erro de memória da GPU excedida, para os dois modelos fornecidos pela plataforma, P100 e T4, que forneciam o valor máximo de 16 GB. Para contornar o problema, o tamanho do lote (*batch size*) foi reduzido para o treinamento, mas a solução não teve impacto significativo na resolução do problema. A dimensão da rede neural em si e o número de cabeças também foi reduzido, sendo correspondente à referência de tamanho pequeno para a rede, com dimensão de 144 e 4

cabeças. As alterações também não foram capazes de diminuir o consumo da memória da GPU para o limite aceitável, mesmo combinadas.

No ambiente do Google Colab, foi usada uma GPU V100, que dispõe de aproximadamente 40 GB de memória. As alterações realizadas anteriormente no código para menor consumo da memória da GPU foram mantidas, contudo o erro de memória da GPU excedida também ocorreu na execução do código na plataforma, conforme mostrado na Figura 28.

```
---------------------------------------------------------------------------
OutOfMemoryError                          Traceback (most recent call last)
<ipython-input-26-e28588e2bfda> in <cell line: 202>()
    204         evaluate_saved()
    205     else:
--> 206         main()

                                14 frames
/content/attention.py in forward(self, query, key, value, mask)
     35
     36     def forward(self, query: Tensor, key: Tensor, value: Tensor, mask: Optional[Tensor] = None) ->
Tuple[Tensor, Tensor]:
---> 37         score = torch.bmm(query, key.transpose(1, 2)) / self.sqrt_dim
     38
     39         if mask is not None:

OutOfMemoryError: CUDA out of memory. Tried to allocate 26.00 MiB (GPU 0; 39.56 GiB total capacity; 37.41 GiB
already allocated; 26.56 MiB free; 37.92 GiB reserved in total by PyTorch) If reserved memory is >> allocated
memory try setting max_split_size_mb to avoid fragmentation.  See documentation for Memory Management and
PYTORCH_CUDA_ALLOC_CONF
```

Figura 28 – Erro de falta de memória da GPU
Fonte: Autoria própria

Assim como foi realizado no código original, a base de dados utilizada foi reduzida, sendo selecionados apenas 10% dos dados de EMG, porém nem mesmo a redução da quantidade de dados foi capaz de solucionar o problema de falta de memória da GPU.

As demais alternativas para a solução do erro, como diminuição do sinal de entrada, foram descartadas pela possibilidade de interferência nos resultados. Dessa forma, pelas limitações de hardware, o modelo de *Speech-Transformer* não pôde ser implementado e avaliado de fato.

6.4 Implementação do *Conformer*

O modelo do *Conformer* teve como base o código criado por Soohwan Kim[2] e foi implementado com dimensão de 144 e 4 cabeças, equivalente ao modelo pequeno, uma vez que o modelo com dimensões maiores também acarretou no erro de memória excedente da GPU.

Ao implementar o modelo para o reconhecimento da fala silenciosa, o código foi composto de maneira similar à que ocorreu no *Speech Transformer*, sendo utilizados os

[2] Pode ser acessado em: https://github.com/sooftware/conformer.git

scripts originais para preparação dos dados e avaliação da rede neural. Por se tratar de um teste, inicialmente, foi utilizada apenas 10% da base de dados, para maior agilidade no treinamento do *Conformer*. Nesse ponto, foram identificados alguns erros de código, principalmente de dimensionamento dos argumentos, que necessitaram de uma adequação para pleno funcionamento do modelo.

Avaliando o código original, foi identificado que o valor do EMG bruto foi usado como entrada do *Transformer*, o que se justifica pela parametrização por aprendizado citada pelos autores do estudo original, realizada pela própria rede neural. No entanto, ao utilizar o EMG bruto, com 8 canais, como entrada do *Conformer*, foram encontrados novos erros de incompatibilidade de dimensões, geradas principalmente pela existência dos 8 canais. Uma vez resolvidas as incompatibilidades, não foi alertado mais nenhum erro de código, porém foi obtido um erro acesso ilegal à memória da GPU, conforme Figura 29.

```
---------------------------------------------------------------------------
RuntimeError                              Traceback (most recent call last)
<ipython-input-20-1f8ce38388e7> in <cell line: 281>()
    283         evaluate_saved()
    284     else:
--> 285         main()

                                7 frames
/usr/local/lib/python3.10/dist-packages/torch/nn/modules/module.py in convert(t)
   1141                 return t.to(device, dtype if t.is_floating_point() or t.is_complex() else None,
   1142                             non_blocking, memory_format=convert_to_format)
-> 1143             return t.to(device, dtype if t.is_floating_point() or t.is_complex() else None, non_blocking)
   1144
   1145         return self._apply(convert)

RuntimeError: CUDA error: an illegal memory access was encountered
CUDA kernel errors might be asynchronously reported at some other API call, so the stacktrace below might be incorrect.
For debugging consider passing CUDA_LAUNCH_BLOCKING=1.
Compile with `TORCH_USE_CUDA_DSA` to enable device-side assertions.
```

Figura 29 – Erro de acesso ilegal à memória
Fonte: Autoria própria

Pela falta de especificidade no motivo do erro, foi definido que seriam utilizados os valores de EMG parametrizados manualmente através de domínio no tempo e cruzamento de zero, em vez do EMG bruto. Pelo fato das dimensões dos dois tipos de tensores serem diferentes, foram necessárias novas correções de erros de dimensionamento.

6.5 Treinamento e avaliação do *Conformer*

Após a validação de algumas épocas rodadas sem problemas de código, ainda que com a taxa de erro de 100 (seguindo o mesmo cálculo da Equação 6.1), a base de dados foi aumentada para sua totalidade. Novamente, o consumo da memória da GPU ultrapassou a disponibilizada, sendo preciso diminuir a base de dados para 80%, assim como foi realizado na validação do código original. Também foi necessário o uso da ferramenta *Garbage Collector* para liberação da memória RAM.

Devido à limitação de tempo de execução da máquina virtual existente no Kaggle, não sendo possível usar o mesmo ambiente virtual por mais de 12 horas seguidas, o

treinamento foi realizado por partes, sendo utilizada a técnica de *checkpoint* para armazenamento dos parâmetros do aprendizado. O aprendizado foi feito por 3 etapas, cada uma com 80 épocas, totalizando então 240 épocas, um número até maior que o original, de 200 épocas. Ao longo das épocas finalizadas, a menor taxa de erro por palavra registrada foi 99,45. Por se tratar de um aprendizado com mais épocas, era esperada uma taxa de erro menor, dessa forma, foram realizadas algumas alterações para identificar possíveis pontos de melhoria do código.

O treinamento foi então realizado com diferentes otimizadores, sendo que cada otimizador foi usado em dois treinamentos, o primeiro com a taxa de aprendizado original e o segundo com a taxa de 0,001. O número de épocas definido foi 50, que no modelo original já foi o suficiente para observar uma evolução na taxa de erro, que ficou em 78,73%. Os resultados de WER obtidos em cada uma das simulações podem ser observados na Tabela 4.

Tabela 4 – WER com diferentes otimizadores após 50 épocas

	Taxa de aprendizado	
Otimizador	0,001	0,0003
AdamW	99,57	99,74
Adam	99,73	99,87
SGD	98,82	98,93
RMSProp	99,76	99,95
Adadelta	99,95	99,97

Fonte: Autoria própria

Os otimizadores testados são amplamente utilizados em arquiteturas de redes neurais. O SGD (gradiente descente estocástisco, do inglês *Stochastic Gradient Descent*) é considerado uma das alternativas mais robustas e confiáveis, principalmente se tratando de arquiteturas mais simples. O otimizador Adam (estimativa adaptativa de momento, do inglês *Adaptive Moment Estimation*) é um dos mais populares para a tarefa de reconhecimento de fala, por combinar as vantagens do SGD com adaptações de taxa de aprendizado, que introduz uma correção de peso de decaimento (*weight decay*), L2, diretamente no algoritmo de otimização. O RMSProp (propagação da raiz quadrada da média, do inglês *Root Mean Square Propagation*), por sua vez, é usado principalmente quando há preocupações com a convergência devido a diferentes escalas de gradiente. Por fim, o Adadelta é uma variação do Adagrad (gradiente descendente adaptativo, do inglês *Adaptive Gradient Descent*), utilizado para evitar a redução da taxa de aprendizado.

Como pode ser observado, não houveram mudanças significativas entre os otimizadores testados e as taxas de aprendizado. Analisando a evolução das épocas, é possível notar que não houve melhora contínua nos parâmetros, do mesmo modo que foi possível identificar a predominância de algumas letras específicas a cada época, de forma com que

as predições com uma taxa de erro menor foram decorrentes de palavras monossilábicas com grande incidência nas frases em inglês, como *"i", "to", "in", "a"* e *"of"*. Na Figura 30 é possível observar a predição do modelo para o melhor resultado de WER, obtido com o uso do otimizador SGD e taxa de aprendizado 0,001.

```
predictions:  ['edition ', 'opinion in ', 'codification of ', 'i i', 'utopian ', 'upon ', 'utopiani', 'i ', 'i ', 'i ', 'i ', 'p
edantic ', 'i ', 'coeducation ', 'i i', 'i ', 'oucanasta', 'i ', 'i ', 'dedication to ', 'apocalyptica', 'cacafue', 'pointing to
a ', 'audaciti', 'i incautiously i', 'i o', 'i i', 'i ', 'popinot i', 'i i', 'opinion ', 'i ', 'opinion ', 'i i', 'coincident ',
'i ononti', 'coupeau ', 'i ', 'aficionado', 'i ', 'oceanic ', 'i economise ', 'codificatio', 'poincilit', 'i pointing to a ', 'd
ocilit', 'i ', 'deificatio', 'oedipus ', 'i ', 'ofici', 'pointing to a p', 'education ', 'opinionati', 'codification of ', 'i ',
'dedication to a ', 'i ', 'i ', 'pointing to a ', 'i o', 'education in ', 'oedipus i ', 'pointing to a p', 'doing ', 'outdistanc
i', 'opinion i', 'opinion in ', 'pointing to a i', 'i ', 'concepcion ', 'opinionati', 'onontio initiation', 'i ', 'i ', 'adoptio
n of a ', 'i ', 'anointi', 'codification ', 'i ', 'cocoeni', 'conciliation ', 'i i', 'opening it i ', 'coeducatio', 'i ', 'oedip
us i ', 'odonai', 'caecilia', 'pointing to a ', 'odeon ', 'i ', 'i ', 'pointing to a i', 'i ', 'dedication to ', 'upon ', 'oedip
us i ', 'i ', 'opening a ', 'cocoanut ', 'i ', 'opinion in ', 'i ', 'i ', 'oedipus i ', 'dedication to ', 'i ', 'i ', 'pietati',
'oucanasta i', 'cocoanut ', 'i ', 'codification of ', 'utopian o', 'i cincinnati', 'codification of ', 'coincident ', 'people ',
'i ', 'deputations ', 'i ', 'i ', 'cocoanut ', 'opinion in ', 'i ', 'oecolampadius ', 'codification of i', 'i ', 'oedipus i ',
'i ', 'depending on ', 'i ', 'i ', 'coeducation i', 'i i', 'cecilia ', 'i opinion ', 'i opinion ', 'opinionati', 'economic ', 'p
opocatapetl ', 'i ', 'i ', 'oedipus i i', 'i ', 'podingto', 'i ', 'i ', 'i ', 'edition ', 'opinion ', 'opinion in ', 'opinion in
', 'i i', 'i ', 'decontamination ', 'i ', 'opinion i', 'opinionati', 'oedipus i i', 'opinion ', 'i ', 'codification of ', 'oedip
us i ', 'opinion ', 'opinionati', 'adeimantus ', 'i ', 'i ', 'i ', 'i initiatio', 'pointing to a ', 'opinionati', 'i ', 'adonai
', 'coeducatio', 'petitioni', 'toinon ', 'i ', 'coeducation ', 'i opinionati', 'ouistiti', 'i ', 'pointing to a ', 'piepenbrink
', 'i ', 'dondindac i ', 'i ', 'i ', 'coeducation ', 'cecilia ', 'pointing to a i', 'coeducation ', 'pointing to a ', 'pointing
to a ', 'i papini', 'i ', 'aeolian ']
finished epoch 39 - training loss: 0.8014 validation WER: 98.82
```

Figura 30 – Predições do *Conformer* com otimizador SGD e taxa de aprendizado 0,001
Fonte: Autoria própria

Devido à alta taxa de erro observada durante o treinamento com diferentes conjuntos de hiperparâmetros, bem como considerando que o modelo *Conformer* geralmente é empregado com entradas no formato de espectrogramas, foi realizado um experimento no qual os dados EMG passaram por um processo de conversão. Antes de serem fornecidos como entrada para o modelo, os sinais EMG foram submetidos a uma transformação para calcular o espectrograma correspondente. Esse cálculo do espectrograma foi realizado utilizando a biblioteca PyTorch, com a seguinte configuração: uma taxa de amostragem de 516Hz, um tamanho de janela de 31 milissegundos, uma sobreposição de 37,5%, e uma transformada de Fourier de 256 pontos (FFT), parâmetros similares ao utilizados no modelo de transdução.

Pela demanda de memória RAM do modelo, foi selecionado o ambiente virtual do Google Colab, configurado com a GPU V100 e RAM alta. Com a base de dados reduzida, o treinamento foi realizado por 109 épocas, sendo interrompido na plataforma sem nenhuma mensagem de erro que pudesse indicar o motivo do desligamento da máquina virtual. Ao longo das épocas, a menor taxa de erro por palavra obtida foi 99.95%. Alterando o método de cálculo do espectrograma, para que fosse realizado previamente para toda a base de dados e armazenado para consulta posterior, o problema de desligamento da máquina virtual não se repetiu, no entanto, não foi identificada nenhuma melhora no aprendizado.

Comparativamente, o modelo com espectrograma se mostrou mais lento que o normal, que faz uso direto do sinal EMG tratado. Outra diferença entre os dois modelos de reconhecimento de fala com *Conformer* é que o com espectrograma fez predições com letras mais variadas, conforme pode ser observado nas Figuras 31 e 32.

```
predictions:  ['i iriti', 'vivifi', 'iridi', 'rimini', 'vivifi', 'rivervi', 'iridi', 'rivervi', 'civi
lizi', 'i imihi', 'iridi', 'civilizi', 'vivifi', 'i crimini', 'imihi', 'imihi', 'rivi', 'rivervi', 'i
ridi', 'iridi', 'imihi', 'ivori', 'iridi', 'i imihi', 'viridi', 'rivervi', 'vivifi', 'civilizi', 'riv
ervi', 'rivervi', 'i civilizi', 'rivi', 'orivi', 'iridi', 'imihi', 'crimini', 'iroquoi', 'irvi', 'ivo
ri', 'i imihi', 'iridi', 'civilizi', 'ivori', 'iriti', 'imihi', 'rimini', 'imihi', 'irvi', 'vivifi',
'iridi', 'civilizi', 'i oxidizi', 'rimini', 'ivi', 'iridi', 'viridi', 'viciniti', 'iriti', 'iridi',
'i iridi', 'rimini', 'i inciviliti', 'mimicri', 'imihi', 'civilizi', 'iriti', 'iridi', 'rivi', 'riv
i', 'rimini', 'rimini', 'orivi', 'rivi', 'imihi', 'rimini', 'iridi', 'iridi', 'civilizi', 'rivi', 'ri
mini', 'rimini', 'rimini', 'iridi', 'rimini', 'iridi', 'rimini', 'iridi', 'civilizi', 'vivifi', 'imih
i', 'iridi', 'rimini', 'imihi', 'vivifi', 'rivi', 'iridi', 'civilizi', 'i iridi', 'vivifi', 'iridi',
'rimini', 'imihi', 'rimini', 'rimini', 'i iriti', 'primitivi', 'i iridi', 'iridi', 'i iridi', 'imih
i', 'viridi', 'iridi', 'ivori', 'civilizi', 'iridi', 'rimini', 'crimini', 'iridi', 'mimicri', 'rimin
i', 'i iridi', 'iridi', 'rivi', 'imihi', 'iridi', 'iridi', 'irvi', 'minimizi', 'rimini', 'minimizi',
'rivi', 'rimini', 'rimini', 'civilizi', 'civilizi', 'crimini', 'ivori', 'diviniti', 'mimicri', 'irid
i', 'imihi', 'civilizi', 'iridi', 'orivi', 'viciniti', 'mimicri', 'mimicri', 'mimicri', 'iridi', 'rim
ini', 'rimini', 'civilizi', 'rimini', 'imihi', 'rivervi', 'rimini', 'i vivifi', 'viridi', 'mimicri',
'iridi', 'mimicri', 'iridi', 'iridi', 'croixili', 'i minimizi', 'iridi', 'iriti', 'imihi', 'iridi',
'rivi', 'viciniti', 'i iridi', 'rimini', 'iridi', 'imihi', 'ivori', 'i imihi', 'oiri', 'iridi', 'rive
rvi', 'imihi', 'vivifi', 'iriti', 'irvingi', 'iriti', 'iridi', 'vivifi', 'rimini', 'iridi', 'iridi',
'viciniti', 'imihi', 'iridi', 'iridi', 'iriti', 'irvingi', 'civilizi', 'rimini', 'rimini']
finished epoch 50 - training loss: 0.7919 validation WER: 99.76
```

Figura 31 – Predições do *Conformer* com EMG
Fonte: Autoria própria

```
predictions:  ['vyv', 'sassacus', 'phth', 'jej', 'coxcom', 'exege', 'grippa', 'symptomat', 'tsisk', 'tuttu', 'cxxi
x', 'juju', 'mukluks', 'didn', 'titiv', 'amahag', 'ysay', 'agafia', 'ninepins', 'rarer', 'kek', 'vev', 'bookk','nku
nk', 'the ', 'kokovtsov', 'herefordsh', 'tattleta', 'highg', 'ullullo', 'eveleg', 'kukuleyo', 'titm', 'shapham','mu
mu', 'zarz', 'abanaza', 'geoghega', 'vivipa', 'daddl', 'mumf','papae', 'myrm', 'tentat', 'zizz', 'gigg', 'hexh', 'v
egg', 'mtb', 'whych', 'hsh', 'ethelbertha', 'lyly', 'q', 'bannin', 'lalala', 'hutuktu', 'scaccia', 'pipp', 'waw','t
otopoto', 'loualaba', 'non', 'excesses', 'popop', 'cawa', 'rorr', 'zbysz', 'giojo', 'kek', 'clel', 'halfoff', 'cxx
x', 'diodoro', 'hh', 'greegree', 'usuf', 'wowow', 'juj', 'pupp', 'biq', 'vedettes', 'cuju', 'nucky', 'nund', 'bub',
'cyg', 'hamadcha', 'panpanga', 'xxxv', 'twt', 'jij', 'xax', 'sluss', 'xy', 'mdcccxc', 'dymond', 'klik', 'hoh', 'spe
s', 'xcix', 'snowstor', 'nonn', 'umgungun', 'yny', 'snaphan', 'pottowott', 'nandana', 'nonsins', 'kidgi','soss', 'p
urpu', 'lillyw', 'testifiet', 'ssshh', 'flails', 'cocksc', 'whych', 'diddimus', 'dd', 'onon', 'quaq', 'zzzz', 'stut
t', 'kakik', 'landlad', 'typifyi', 'aq', 'altsta', 'dagda', 'torturo', 'hammam', 'kyk', 'titbit', 'nonentiti', 'mi
m', 'popap', 'disd', 'quok', 'bobsb', 'gallipoli', 'fyf', 'nunqu', 'precipic', 'oooo', 'takata', 'excheque', 'ogall
alla', 'diddu', 'huju', 'affidav', 'vyvya', 'swow', 'zoz', 'heugh', 'ngenyan', 'wowow', 'unvacc','lowdow', 'koloko',
'satiat', 'mamam', 'aggag', 'cece', 'intuiti', 'shamashna', 'amalgama', 'rococo', 'vev', 'xxx', 'iff', 'xo', 'cecc',
'jarj', 'joz', 'wiriw', 'summum', 'lalal', 'yaroslav', 'roycrof', 'bassist', 'cutbu', 'wowow', 'minimizing', 'dord
r', 'odysseys', 'moem', 'bieb', 'sts', 'hinz', 'menehwehn', 'qr', 'syby', 'papadop', 'errore', 'bayaya', 'bibbie',
'zzzz', 'qri']
finished epoch 111 - training loss: 0.0621 validation WER: 99.97
```

Figura 32 – Predições do *Conformer* com espectrograma
Fonte: Autoria própria

Apesar das alterações de parâmetros, o modelo não apresentou melhorias no treinamento, mesmo após horas de treinamento e mais de 100 épocas de simulação.

Por fim, o comparativo entre os resultados com as diferentes técnicas abordadas encontra-se na Tabela 5. Sendo que os resultados ressaltados representam a menor taxa de erro por palavra obtida no treinamento com 50 épocas para cada uma das técnicas.

Tabela 5 – WER com diferentes técnicas após 50 épocas

Técnica	WER
Transformer	78,73%
Conformer	99,74%
Conformer com espectrograma	99,95%
Conformer com novos parâmetros	98,82%

Fonte: Autoria própria

A técnica denominada como *Transformer*, se trata do código baseado no experimento original, apenas com as alterações de adaptação aos ambientes virtuais de execução. A denominada *Conformer* utiliza os mesmos parâmetros do experimento original, ou seja,

otimizador AdamW e taxa de aprendizado de 0,0003. No *Conformer* com espectrograma os parâmetros foram mantidos, mas foi efetuado o cálculo do espectrograma na entrada da rede neural. Por sua vez, o *Conformer* com novos parâmetros está representando o modelo com otimizador SGD e taxa de aprendizado 0,001, obtido no teste dos diferentes otimizadores.

7 Conclusões

Neste trabalho, foi realizada uma revisão bibliográfica das técnicas usadas para o reconhecimento da fala silenciosa, sendo selecionado o trabalho de Gaddy (2022) para o estudo de caso por abordar algumas das técnicas amplamente utilizadas no contexto de SSI, além da disponibilidade das informações do experimento, não apenas no âmbito de resultados, como também de metologia, base de dados e algoritmos.

Um dos principais objetivos do trabalho foi o uso de técnicas de aprendizado de máquina diferentes da empregada no experimento tido como base, a fim de analisar o comportamento de outras redes neurais naturais, para isso, foram selecionados o *Speech Transformer* e o *Conformer*, cuja arquitetura tem similaridades com a do *Transformer* usado por Gaddy (2022).

Em termos práticos, os algoritmos originais criados por David Gaddy foram testados primeiramente, sendo comprovada a eficácia do modelo baseado na arquitetura *Transformar*, ainda que com uma redução da base de dados utilizada. Justamente por se tratarem de arquiteturas baseadas em camadas de atenção e com presença de blocos convolucionais, o resultado esperado para o *Speech Transformer* e para o *Conformer* era um comportamento similar ao do experimento original, com redução da taxa de erro por palavra após poucas dezenas de épocas de treinamento.

No caso do *Speech Transformer*, as limitações de *hardware* tiveram grande impacto no experimento por não possibilitarem a execução do modelo ao ser implementado nas plataformas virtuais, mesmo com parâmetros bem inferiores aos utilizados no estudo original. Dessa forma, não foi possível obter resultados conclusivos a partir dessa abordagem pela alta demanda de memória de GPU do modelo, superior à das máquinas virtuais utilizadas no experimento.

Para o *Conformer*, embora o modelo tenha sido executado sem impedimentos, a WER ficou muito acima do esperado, mesmo para os casos de treinamento com mais épocas que o experimento original. Sua implementação, no entanto, propiciou informações o suficiente para melhor aprendizado a respeito da técnica e algumas inferências sobre a arquitetura empregada. Algumas hipóteses foram levantadas para justificar a divergência entre a taxa de erro esperada e a obtida, as quais serão tratadas melhor a seguir.

Em termos comparativos com o experimento executado por Gaddy e Klein, o modelo da rede neural empregada neste trabalho está com dimensões menores. Enquanto o *Transformer* tem 768 de dimensão e 8 cabeças, sendo considerado de tamanho grande, o *Conformer* implementado tem 144 e 4, respectivamente, o que o configura como pequeno. A configuração menor implica em menor capacidade de aprendizado da rede, sendo ne-

cessário um maior número de épocas para a obtenção dos resultados, que, ainda assim, podem não alcançar aos obtidos em uma rede de capacidade maior.

Outro fator de impacto no resultado é o tamanho da base de dados. Utilizar uma porcentagem menor da base de dados pode implicar em uma maior dificuldade de aprendizado por parte da rede neural, tornando-o menos robusto. No entanto, ao ser analisado o comportamento do *Transformer* do experimento original com 80% do *dataset*, foi verificado que o modelo continuou convergindo bem após algumas dezenas de épocas, de forma com que apenas a redução da base de dados não deveria implicar em uma taxa de erro tão grande por parte do *Conformer*.

A escolha do uso do sinal EMG parametrizado manualmente em vez do EMG bruto também pode ter acarretado em divergências no aprendizado da rede neural, contudo, não foi possível testar o impacto do uso do sinal tratado na arquitetura original para efeitos de comparação. Ao substituir os argumentos do *Transformer*, fazendo as mudanças de dimensões necessárias, foi obtido o erro de acesso ilegal à memória, de maneira similar a que ocorreu com o *Conformer* no início de sua implementação (Figura 28.

Por fim, o modelo do *Conformer* utilizado como base da arquitetura do experimento pode não ter sido implementado corretamente, sendo necessária a verificação com outras bases de dados, preferencialmente amplamente utilizadas para contextos de reconhecimento de fala, como a *Librispeech* para avaliação do seu comportamento. A partir da análise com uma base de dados comumente usada, é possível comprovar sua eficácia e realizar os ajustes necessários para o uso no contexto do reconhecimento de fala silenciosa.

7.1 Trabalhos futuros

Embora os resultados obtidos não estejam próximos dos obtidos no experimento original de Gaddy e Klein, o presente trabalho proporcionou aprendizado em diversas áreas de conhecimento, principalmente no que se refere a redes neurais. A contribuição do trabalho se deu pelo levantamento das principais técnicas usadas para o reconhecimento de fala silenciosa, sendo propostos dois métodos para aprendizado dos sinais de fala obtidos por eletromiografia de superfície, baseados nas arquiteturas *Speech Transformer* e *Conformer*.

Para trabalhos futuros, as hipóteses levantadas anteriormente podem ser testadas, assim como as redes neurais citadas. No caso do *Speech Transformer*, melhores configurações de *hardware* podem ser empregadas para a validação do modelo, bem como outras técnicas de otimização de uso da memória da GPU. No caso do *Conformer*, podem ser realizados novos testes com outros parâmetros e maiores dimensões de arquitetura, além de um número maior de épocas, para análise da convergência do modelo. A respeito do *Conformer*, também pode ser estudado o comportamento do método com outras bases de

dados para que adaptações sejam feitas na sua implementação.

Em ambas as situações, é importante ressaltar a flexibilidade dos modelos para serem ajustados visando um melhor desempenho no contexto da fala silenciosa, seguindo o exemplo dos autores do experimento original, que incorporaram recursos como *embeddings* de posição relativa, entre outros métodos. Além disso, é válido mencionar que existem diversas abordagens adicionais, conforme ilustrado na Figura 23, permitindo a exploração de diferentes estratégias e a investigação das variadas possibilidades promissoras oferecidas por esse campo de pesquisa.

Referências

ABREU, H. P. M. *Visual speech recognition for European Portuguese*. Tese (Doutorado) — Universidade do Minho (Portugal), 2014. Citado na página 32.

ANUMANCHIPALLI, G. K.; CHARTIER, J.; CHANG, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature*, Nature Publishing Group, v. 568, n. 7753, p. 493–498, 2019. Citado na página 58.

BIRBAUMER, N. et al. The thought translation device (ttd) for completely paralyzed patients. *IEEE Transactions on rehabilitation Engineering*, IEEE, v. 8, n. 2, p. 190–193, 2000. Citado na página 36.

BOELTER, J. d. S. Classificação de sinais eletromiográficos utilizando redes neurais artificiais, análise discriminante linear e floresta aleatória. Universidade do Vale do Rio dos Sinos, 2021. Citado na página 40.

CASTRO, F. D.; CASTRO, M. D. Redes neurais artificiais. *Porto Alegre, RS: Pontifícia Universidade Católica do Rio Grande do Sul*, 2001. Citado na página 39.

DATASHEET ADS1299-x Low-Noise, 4-, 6-, 8-Channel, 24-Bit, Analog-to-Digital Converter for EEG and Biopotential Measurementsaspberry Pi 4 Model B. [S.l.]: Texas Instrument, 2017. <https://www.ti.com/lit/ds/symlink/ads1299.pdf>. Acesso em: 07 de janeiro 2023. Citado na página 46.

DELGADO, S. I. V. Development of algorithms to improve the technical efficiency of capturing, processing, and identification of eeg signals in the word imagery task. Bogotá-Ingeniería-Doctorado en Ingeniería-Ingeniería Eléctrica, 2020. Citado na página 32.

DENBY, B. et al. Silent speech interfaces. *Speech Communication*, Elsevier, v. 52, n. 4, p. 270–287, 2010. Citado 7 vezes nas páginas 15, 26, 28, 29, 34, 36 e 37.

DINIZ, P. S.; SILVA, E. A. D.; NETTO, S. L. *Digital signal processing: system analysis and design*. [S.l.]: Cambridge University Press, 2010. Citado 2 vezes nas páginas 38 e 48.

DONG, L.; XU, S.; XU, B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: IEEE. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. [S.l.], 2018. p. 5884–5888. Citado 4 vezes nas páginas 26, 64, 65 e 66.

ERIKSEN, C. W.; POLLACK, M. D.; MONTAGUE, W. E. Implicit speech: Mechanism in perceptual encoding? *Journal of Experimental Psychology*, American Psychological Association, v. 84, n. 3, p. 502, 1970. Citado na página 28.

FALE - UFMG. *Anatomia e fisiologia da voz*. 2021. Disponível em: <http://www.letras.ufmg.br/lbass/>. Acesso em: 13 de setembro de 2022. Citado na página 28.

FRANZESE, M.; IULIANO, A. Hidden markov models. Elsevier, 2019. Citado na página 39.

GADDY, D.; KLEIN, D. Digital voicing of silent speech. *arXiv preprint arXiv:2010.02960*, 2020. Citado 11 vezes nas páginas 16, 41, 42, 44, 45, 47, 48, 49, 52, 53 e 58.

GADDY, D.; KLEIN, D. An improved model for voicing silent speech. *arXiv preprint arXiv:2106.01933*, 2021. Citado 4 vezes nas páginas 48, 54, 55 e 58.

GADDY, D. M. *Voicing Silent Speech*. [S.l.]: University of California, Berkeley, 2022. Citado 6 vezes nas páginas 26, 48, 59, 60, 69 e 76.

GONÇALVES, A. R. Máquina de vetores suporte. *Acesso em*, v. 21, 2010. Citado na página 39.

GONZALEZ-LOPEZ, J. A. et al. Silent speech interfaces for speech restoration: A review. *IEEE access*, IEEE, v. 8, p. 177995–178021, 2020. Citado na página 15.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016. Citado na página 54.

GRAINGER, J. Book review: Muscles alive: Their functions revealed by electromyography. *The Canadian Journal of Occupational Therapy*, SAGE PUBLICATIONS, INC., v. 54, n. 3, 1987. Citado na página 85.

GULATI, A. et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020. Citado 2 vezes nas páginas 26 e 67.

HAKONEN, M.; PIITULAINEN, H.; VISALA, A. Current state of digital signal processing in myoelectric interfaces and related applications. *Biomedical Signal Processing and Control*, Elsevier, v. 18, p. 334–359, 2015. Citado na página 89.

HARDOON, D. R.; SZEDMAK, S.; SHAWE-TAYLOR, J. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, MIT Press, v. 16, n. 12, p. 2639–2664, 2004. Citado na página 53.

HSUE, S.-Z.; SOLIMAN, S. S. Automatic modulation classification using zero crossing. In: IET. *IEE Proceedings F (Radar and Signal Processing)*. [S.l.], 1990. v. 137, n. 6, p. 459–464. Citado na página 38.

HUANG, Z.; XU, W.; YU, K. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015. Citado na página 50.

KIM, M. et al. Speaker-independent silent speech recognition from flesh-point articulatory movements using an lstm neural network. *IEEE/ACM transactions on audio, speech, and language processing*, IEEE, v. 25, n. 12, p. 2323–2336, 2017. Citado na página 33.

KIM, S.; BAE, S.; WON, C. Kospeech: Open-source toolkit for end-to-end korean speech recognition. *SIMPAC*, ELSEVIER, p. Volume 7, 100054, February 2021. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2665963821000026>. Citado na página 69.

KLAPP, S. T.; ANDERSON, W. G.; BERRIAN, R. W. Implicit speech in reading: Reconsidered. *Journal of Experimental psychology*, American Psychological Association, v. 100, n. 2, p. 368, 1973. Citado na página 28.

LIMA, D. L. S. et al. Classificação de imagens de exames de endoscopia por cápsula utilizando transformers. Universidade Federal do Maranhão, 2023. Citado na página 57.

MATTA, R. S. d. et al. Multidimensional voice assessment: the immediate effects of lax vox® in singers with voice complaints. *Revista CEFAC*, SciELO Brasil, v. 23, 2021. Citado na página 34.

MEIRELES, A. R. O uso do magnetômetro (ema) na análise de dados articulatórios da prosódia da fala. *Prosódia da fala: pesquisa e ensino. São Paulo: Blucher*, 2017. Citado na página 32.

MELTZNER, G. S. et al. Signal acquisition and processing techniques for semg based silent speech recognition. In: IEEE. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. [S.l.], 2011. p. 4848–4851. Citado 2 vezes nas páginas 16 e 37.

MILLS, K. R. The basics of electromyography. *Journal of Neurology, Neurosurgery & Psychiatry*, BMJ Publishing Group Ltd, v. 76, n. suppl 2, p. ii32–ii35, 2005. Citado na página 86.

MONOLITO NIMBUS. *Redes neurais artificiais*. 2017. Disponível em: <https://www.monolitonimbus.com.br/redes-neurais-artificiais/>. Acesso em: 27 de novembro de 2022. Citado na página 39.

MUNDO EDUCAÇÃO. *Epiglote*. 2016. Disponível em: <https://mundoeducacao.uol.com.br/biologia/epiglote.htm>. Acesso em: 08 de setembro de 2022. Citado na página 27.

NAKAJIMA, Y. et al. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In: IEEE. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*. [S.l.], 2003. v. 5, p. V–708. Citado na página 34.

NIELSEN, M. A. *Neural networks and deep learning*. [S.l.]: Determination press San Francisco, CA, USA, 2015. v. 25. Citado 2 vezes nas páginas 50 e 51.

OCARINO, J. de M. et al. Eletromiografia: interpretação e aplicações nas ciências da reabilitação. 2005. Citado 2 vezes nas páginas 26 e 46.

REAZ, M. B. I.; HUSSAIN, M. S.; MOHD-YASIN, F. Techniques of emg signal analysis: detection, processing, classification and applications. *Biological procedures online*, Springer, v. 8, n. 1, p. 11–35, 2006. Citado na página 85.

REYNOLDS, D. A. et al. Gaussian mixture models. *Encyclopedia of biometrics*, Berlin, Springer, v. 741, n. 659-663, 2009. Citado na página 38.

RIBEIRO, G. et al. Análise mel-cepstral na discriminação de patologias laríngeas. In: *XXIV Congresso Brasileiro de Engenharia Biomédica*. [S.l.: s.n.], 2014. Citado 2 vezes nas páginas 38 e 49.

SAKSAMUDRE, S. K.; SHRISHRIMAL, P.; DESHMUKH, R. A review on different approaches for speech recognition system. *International Journal of Computer Applications*, Citeseer, v. 115, n. 22, 2015. Citado na página 44.

SAMPAIO, D. B. Predição da evapotranspiração de referência usando rede lstm bidirecional e var+ lstm. Serra, 2022. Citado na página 52.

SANTANA, L. M. Q. d. Aplicação de redes neurais recorrentes no reconhecimento automático da fala em ambientes com ruídos. Pós-Graduação em Ciência da Computação, 2017. Citado na página 60.

SHIN, Y. H.; SEO, J. Towards contactless silent speech recognition based on detection of active and visible articulators using ir-uwb radar. *Sensors*, MDPI, v. 16, n. 11, p. 1812, 2016. Citado na página 35.

STOLLNITZ, E. J.; DEROSE, A.; SALESIN, D. H. Wavelets for computer graphics: a primer. 1. *IEEE computer graphics and applications*, IEEE, v. 15, n. 3, p. 76–84, 1995. Citado na página 38.

TRAN, V.-A. et al. Improvement to a nam-captured whisper-to-speech system. *Speech communication*, Elsevier, v. 52, n. 4, p. 314–326, 2010. Citado 2 vezes nas páginas 16 e 34.

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Citado 3 vezes nas páginas 55, 56 e 57.

VOJTECH, J. M. et al. Surface electromyography–based recognition, synthesis, and perception of prosodic subvocal speech. *Journal of Speech, Language, and Hearing Research*, ASHA, v. 64, n. 6S, p. 2134–2153, 2021. Citado na página 30.

VORONTSOVA, D. et al. Silent eeg-speech recognition using convolutional and recurrent neural network with 85% accuracy of 9 words classification. *Sensors*, MDPI, v. 21, n. 20, p. 6744, 2021. Citado na página 31.

WADKINS, E. J. *A continuous silent speech recognition system for AlterEgo, a silent speech interface*. Tese (Doutorado) — Massachusetts Institute of Technology, 2019. Citado 6 vezes nas páginas 15, 16, 26, 28, 29 e 30.

WAGNER, C. et al. Silent speech command word recognition using stepped frequency continuous wave radar. *Scientific Reports*, Nature Publishing Group UK London, v. 12, n. 1, p. 4192, 2022. Citado na página 34.

WAND, M.; JANKE, M.; SCHULTZ, T. Tackling speaking mode varieties in emg-based speech recognition. *IEEE transactions on biomedical engineering*, IEEE, v. 61, n. 10, p. 2515–2526, 2014. Citado na página 31.

WANG, H.; ROUSSEL, P.; DENBY, B. Representation learning of tongue dynamics for a silent speech interface. *IEICE TRANSACTIONS on Information and Systems*, The Institute of Electronics, Information and Communication Engineers, v. 104, n. 12, p. 2209–2217, 2021. Citado na página 33.

WANG, Y. et al. All-weather, natural silent speech recognition via machine-learning-assisted tattoo-like electronics. *npj Flexible Electronics*, Nature Publishing Group UK London, v. 5, n. 1, p. 20, 2021. Citado 5 vezes nas páginas 16, 26, 28, 36 e 87.

WU, J. et al. A novel silent speech recognition approach based on parallel inception convolutional neural network and mel frequency spectral coefficient. *Frontiers in Neurorobotics*, Frontiers Research Foundation, 2022. Citado na página 25.

YADAV, M.; ALAM, M. A. Dynamic time warping (dtw) algorithm in speech: a review. *International Journal of Research in Electronics and Computer Engineering*, v. 6, n. 1, p. 524–528, 2018. Citado na página 52.

YOUNG, A. J.; HARGROVE, L. J.; KUIKEN, T. A. The effects of electrode size and orientation on the sensitivity of myoelectric pattern recognition systems to electrode shift. *IEEE Transactions on Biomedical Engineering*, IEEE, v. 58, n. 9, p. 2537–2544, 2011. Citado na página 89.

Apêndices

APÊNDICE A – Eletromiografia

O tecido muscular conduz potenciais elétricos de maneira análoga aos nervos, resultando na geração de sinais elétricos denominados "potencial de ação muscular". A eletromiografia de superfície é um método utilizado para registrar e analisar as informações contidas nesses potenciais de ação muscular (REAZ; HUSSAIN; MOHD-YASIN, 2006).

Um músculo é composto de feixes de células especializadas capazes de contração e relaxamento. A função primária dessas células especializadas é gerar forças, movimentos e a capacidade de se comunicar, como fala ou escrita ou outros modos de expressão. O tecido muscular tem extensibilidade e elasticidade. Tem a capacidade de receber e responder a estímulos e pode ser encurtado ou contraído. O tecido muscular tem quatro funções principais: produzir movimento, mover substância dentro do corpo, fornecer estabilização e gerar calor. Três tipos de tecido muscular podem ser identificados com base em propriedades e mecanismos de controle: esquelético, liso e cardíaco, dentre os quais, o esquelético é tido como alvo da EMG.

A contração do músculo esquelético é iniciada por impulsos nos neurônios para o músculo e geralmente está sob controle voluntário. As fibras musculares esqueléticas são bem supridas de neurônios para sua contração. Esse tipo específico de neurônio é chamado de “neurônio motor” e se aproxima do tecido muscular, mas não está realmente conectado a ele. Um neurônio motor geralmente fornece estimulação para muitas fibras musculares.

O corpo humano como um todo é eletricamente neutro; tem o mesmo número de cargas positivas e negativas. Mas no estado de repouso, a membrana da célula nervosa é polarizada devido a diferenças nas concentrações e composição através da membrana plasmática. Existe uma diferença de potencial entre os fluidos intracelulares e extracelulares da célula. Em resposta a um estímulo do neurônio, uma fibra muscular despolariza à medida que o sinal se propaga ao longo de sua superfície e a fibra se contrai. Essa despolarização, acompanhada por um movimento de íons, gera um campo elétrico próximo a cada fibra muscular (GRAINGER, 1987).

Uma eletromiografia é um sinal registrado a partir desse biopotencial gerado por músculos, mais especificamente das unidades motoras (UM). É comum se referir à unidade motora como o complexo que inclui o neurônio motor que desce da medula espinhal e o músculo que este motor neurônio inerva. A eletromiografia é a técnica utilizada para capturar os estímulos elétricos geradas a partir das unidades motoras. Ressalta-se porém que os músculos são inervados por uma infinidade de UMs, variando de algumas dezenas, como seria o caso de um pequeno músculo localizado na mão, a mais de mil para os maiores.

A quantidade de força que é gerada pelos músculos está diretamente relacionada a dois fatores distintos: o número de UMs que estão ativas (o chamado recrutamento de UM) e o *status* de UM, ou seja, o número de vezes por segundo que elas "disparam"(a chamada taxa de disparo de UM, isto é, a frequência com que as UMs são ativadas pelo impulso nervoso ou, equivalentemente, o número de despolarizações por segundo de uma UM).

Com relação ao método de gravação, é preciso distinguir entre gravações intramusculares, quando fios finos ou agulhas são colocados no músculo, e registros de superfície, quando os eletrodos são colocados na superfície do corpo, próximo ao músculo. Neste último caso, a técnica é por vezes referida como eletromiografia de superfície (ou sEMG), para distingui-la da EMG invasiva.

Na eletromiografia invasiva, o registro visa estudar a fisiologia da UM e identificar possíveis alterações associadas a patologias, por exemplo, como a UM varia seu comportamento quando há perda de suprimento nervoso para o músculo. Já nas gravações de superfície, o foco está no comportamento muscular como um todo e quando usadas como ferramenta de diagnóstico, fornecem informações sobre anomalias do comportamento muscular, por exemplo, a capacidade de um músculo para produzir força, quando necessário e com a quantidade necessária.

Ao registrar a atividade elétrica de uma UM, este biopotencial é chamado de Potencial de Ação da Unidade Motora (PAUM). Algumas informações podem ser extraídas diretamente da forma do PAUM, tendo como base o número de picos, voltas e cruzamentos de zero e a presença de possíveis potenciais de satélite geralmente usados para caracterizar a UM.

Outro parâmetro clinicamente relevante na análise de PAUM é a taxa de disparo. Este valor não é constante ao longo do tempo, e uma série de parâmetros podem ser extraídos da taxa de disparo instantânea ao longo do tempo, que é o inverso do tempo decorrido entre duas despolarizações sucessivas.

Um parâmetro interessante e clinicamente significativo é obtido ao analisar múltiplas UMs: dado um par de UMs, é interessante coletar informações sobre o comportamento compartilhado entre elas, observando a correlação cruzada entre as taxas de disparo do PAUM. Isto é chamado de unidade comum, e seu valor depende de uma série de fatores, incluindo o efeito de envelhecimento por comando proveniente do sistema nervoso central (MILLS, 2005).

EMG são frequentemente usadas clinicamente para determinar a velocidade de condução nervosa (VCN), que é uma medida a taxa na qual um nervo pode transportar informações de um local de estímulo para o músculo que ele inerva. Essa medida pode ser feita tanto para os nervos motores quanto para os sensoriais. Neste último caso, a

resposta do nervo ao estímulo é chamada potencial de ação do nervo sensorial (SNAP), e sua amplitude é da ordem de 5-10 mV. Por sua vez, é reconhecido que a VCN é influenciada pela amplitude e características do estímulo, e depende também do nervo específico. Os valores normais de referência estão na faixa de 40 a 70 m/s em adultos e reduzem na presença de neuropatias.

De maneira geral, existem muitas aplicações para o uso de EMG. A EMG é usada clinicamente para o diagnóstico de problemas neurológicos e neuromusculares. É usado para diagnóstico por laboratórios de marcha e por médicos treinados no uso de avaliações biomédica ou ergonômica. A EMG também é usada em muitos tipos de laboratórios de pesquisa, incluindo aqueles envolvidos em biomecânica, controle motor, fisiologia neuromuscular, distúrbios do movimento, controle postural e fisioterapia, além de também ser amplamente usada para o reconhecimento da fala desde a década de 80 (WANG et al., 2021).

APÊNDICE B – Eletrodos

A energia produzida pelos músculos é a fonte do sinal EMG, que é primeiramente detectada pelos eletrodos. Existem vários modelos de eletrodos, que geralmente podem ser divididos em dois tipos: eletrodos de superfície e eletrodos intramusculares. Ambos os eletrodos são igualmente adequados para aquisição de sinal. O fator que determina a escolha de um eletrodo ou outro é a profundidade do músculo a ser avaliado. No caso de músculos superficiais, eletrodos de superfície devem ser utilizados, pois não causam desconforto durante a coleta de dados. No entanto, no caso de músculos profundos, eletrodos intramusculares devem ser selecionados para avaliação muscular para evitar interferência (*cross-talk*) de sinais de músculos mais rasos.

Os eletrodos de superfície podem ser passivos ou ativos. No caso dos eletrodos passivos, a pele constitui uma barreira entre o potencial de ação da unidade motora e os eletrodos. Portanto, a impedância da pele deve ser considerada. Essa impedância pode variar dependendo da umidade, nível de óleo e densidade do estrato córneo da pele. Para obter medidas EMG adequadas, devem ser obtidos valores de impedância da pele entre 5.000 e 10.000 ohms. Para atingir esses valores, é necessário um processo de preparação da pele, que inclui depilação, limpeza da pele com água e sabão e esfoliação suave com álcool. Os eletrodos tensoativos possuem um pré-amplificador que amplifica os sinais EMG assim que chegam aos eletrodos, minimizando as interferências externas. Portanto, os cuidados com o controle da impedância da pele podem se limitar à limpeza com álcool.

Existe também a classificação de eletrodos entre úmidos e secos. Eletrodos úmidos comumente usados requerem gel eletrólito condutivo ou esponja entre o eletrodo e a pele, mas podem fornecer sinais sEMG de alta qualidade. Os eletrodos úmidos geralmente requerem preparação para redução da impedância pele-eletrodo e artefatos de movimento. Além disso, os eletrodos úmidos podem não ser ideais para uso em interfaces sEMG uma vez que o gel condutor pode secar, pode causar irritação e desconforto, e é causa potencial de alergia na pele e inflamação. Eletrodos secos modernos não requerem gel condutivo e preparação de pele, e ainda pode atingir a qualidade do sinal comparável a eletrodos úmidos. Por esta razão, a eletrodos secos podem ser mais aplicáveis para interfaces sEMG.

O material do eletrodo também afeta seu comportamento eletroquímico. Eletrodos polarizáveis (por exemplo, eletrodos de ouro, platina e irídio) são caracterizados pelo comportamento capacitivo porque apenas a corrente de deslocamento passa entre a pele e o eletrodo, enquanto eletrodos não polarizáveis (por exemplo, eletrodos galvanizados e de prata/cloreto de prata) se comportam como resistores, pois permitem um livre fluxo de carga através da interface eletrodo-pele. Nenhum eletrodo é perfeitamente não polarizável

ou polarizável, mas podem ser definidos assim por se aproximarem de tais características.

Eletrodos polarizáveis não são recomendados para medições sEMG devido à sua alta sensibilidade ao movimento. Geralmente são usados eletrodos não polarizáveis, de prata/cloreto de prata (Ag/AgCl) por serem altamente estáveis. Tal tipo de eletrodo consiste em uma superfície de metal prateado revestida com uma fina camada de cloreto de prata. Alguns polímeros e tecidos de fios revestidos por uma camada condutora também se mostraram promissores como materiais para eletrodos. Esses eletrodos são ideais para integração têxtil e podem produzir qualidade de sinal sEMG comparável ao de eletrodos Ag/AgCl (HAKONEN; PIITULAINEN; VISALA, 2015).

Em termos de tamanho ideal de eletrodos para interfaces sEMG, foi realizado um estudo considerando o contexto de estratégia de controle com base em reconhecimento da postura, obtidos de múltiplos sinais EMG. Os tamanhos dos eletrodos (1 cm $\times$ 1 cm, 2 cm $\times$ 2 cm e 3 cm $\times$ 3 cm) mostraram não afetar significativamente a precisão de classificação e taxas de sucesso do controle, no entanto, o benefício de eletrodos maiores era que os sinais sEMG adquiridos com eles foram significativamente menos sensíveis às mudanças do local de gravação do sEMG quando deslocados até 2 cm perpendiculares às fibras musculares. Isso porque o eletrodo com o maior volume de captação possivelmente estava registrando uma parte do sinal mesmo após o deslocamento. Uma estratégia alternativa sugerida para reduzir o efeito de mudanças de eletrodo em precisão de classificação é incluir amostras de possíveis mudanças durante o treinamento do classificador. No entanto, esta abordagem é demorada e problemática porque os eletrodos precisam ser movido para locais de deslocamento esperados durante o treinamento (YOUNG; HARGROVE; KUIKEN, 2011).