Aprendizado de máquina no apoio à transcrição e classificação da fala gaguejada: uma revisão sistemática da literatura

Rodrigo José S. de Almeida[1], Damires Yluska Souza[1], Luciana Pereira Oliveira[1], Débora Vasconcelos Correia[2], Samara Ruth Neves B. Pinheiro[2] e Estevão S. da Silva Sousa[2]

[1]Programa de Pós-Graduação em Tecnologia da Informação (PPGTI) – Instituto Federal da Paraíba – IFPB, Campus João Pessoa, João Pessoa-PB, Brasil

[2]Departamento de Fonoaudiologia – Universidade Federal da Paraíba – UFPB, Campus I, João Pessoa - PB - Brasil

rodrigo-almeida.ra@academico.ifpb.edu.br, {damires,luciana.oliveira}@ifpb.edu.br, debora@ccs.ufpb.br, {samaraneves00, silvestreestevao18}@gmail.com

Abstract. *In the healthcare domain, stuttering identification is manually performed by speech therapists for diagnostic purposes. In this scenario, Machine Learning (ML) can be a valuable tool to support this activity, for example, by automating the transcription of stuttered speech and the classification of disfluencies. This work presents a systematic literature review aiming to investigate how studies have provided or utilized ML methods for transcription and classification of stuttered speech. It also seeks to identify to what extent these studies are applied to effectively support the clinical practice of speech therapists. This work also includes a survey of datasets, languages, diagnostic criteria, and challenges faced in stuttering identification.*

Resumo. *Na área da Saúde, a identificação da gagueira é realizada manualmente por fonoaudiólogos para fins diagnósticos. Neste contexto, o Aprendizado de Máquina (AM) pode ser uma ferramenta valiosa para apoiar esta atividade por meio, por exemplo, da automatização da transcrição de falas gaguejadas e da classificação de disfluências. Este trabalho apresenta uma revisão sistemática da literatura que busca investigar como os trabalhos têm provido ou utilizado métodos de AM para transcrição e classificação da fala gaguejada. Busca-se também identificar até que ponto os trabalhos têm sido aplicados no apoio efetivo à prática clínica do fonoaudiólogo. A análise inclui um levantamento de conjuntos de dados, idiomas, critérios diagnósticos e desafios enfrentados na identificação da gagueira.*

1. Introdução

A gagueira é um transtorno da fluência que inicia na infância e pertence ao grupo dos transtornos da comunicação e do neurodesenvolvimento [APA 2022]. Ela acomete cerca de 5% das crianças pré-escolares e 1% da população adulta, e o seu surgimento ocorre mais frequentemente entre os 2 e os 4 anos de idade, de forma gradual ou abrupta [Yairi e Ambrose 2013]. Entre os principais sinais e sintomas da gagueira encontra-se a produção da fala com excessivas disfluências. As disfluências são categorizadas em: Disfluências Típicas da Gagueira (DTG), classificadas como prolongamento (inicial ou medial), bloqueio, repetição (de som, de sílaba e de palavra monossilábica), pausa e intrusão; e Outras Disfluências (OD), classificadas em revisão, hesitação, interjeição, repetição (de palavra não monossilábica, de segmento e de frase) e palavra incompleta [Oliveira et al. 2023].

O instrumento utilizado no Brasil para avaliar quem gagueja é o Protocolo de Avaliação da Fluência do Teste de Linguagem Infantil ABFW [Andrade 2023]. A avaliação, basicamente, envolve a transcrição manual da amostra de fala gaguejada em ferramenta de edição de texto e a classificação das disfluências pelo fonoaudiólogo. Neste contexto, a interface entre a Computação e a Fonoaudiologia para identificação da gagueira em práticas clínicas tem se mostrado necessária e promissora [Almeida et al. 2023]. Soluções computacionais são fundamentais para a otimização do processo de trabalho do fonoaudiólogo e diagnóstico da gagueira. Neste âmbito, destaca-se o emprego da inteligência artificial, em particular, de métodos de Aprendizado de Máquina (AM) e Aprendizado Profundo (AP).

O AM capacita sistemas a aprenderem comportamentos ou padrões específicos automaticamente, utilizando exemplos disponíveis [Mitchell 1997]. Essa abordagem permite a criação de modelos preditivos capazes de identificar padrões, como as disfluências, em conjuntos de dados, como amostras de fala. Por outro lado, o AP, uma subárea do AM, emprega conceitos de redes neurais artificiais para realizar tarefas complexas de aprendizado e reconhecimento de padrões [Oliveira et al. 2021]. Vale pontuar que, neste trabalho, será utilizado o termo mais amplo AM para referir-se tanto ao AM quanto ao AP, para fins de simplificação.

Sistemas de Reconhecimento Automático de Fala (ASR) têm sido utilizados para converter a fala em texto escrito. Porém, costumam acusar erro e falhar em reconhecer falas gaguejadas [Fox et al. 2021]. Isso significa que devido à presença das disfluências, como repetições, bloqueios e prolongamentos, por exemplo, a transcrição das palavras gaguejadas não é realizada adequadamente pelos sistemas de ASR. Tal circunstância tem levado à necessidade de realização de transcrições manuais. No tocante à classificação das disfluências, estudos com AM têm sido aplicados para identificar padrões [Kourkounakis et al. 2021]. As pesquisas sobre a classificação automática das disfluências na fala gaguejada tiveram início em 1995 [Howell e Sackin 1995]. Posteriormente, um avanço ocorreu com a disponibilização de dados de áudio *online* gratuitos, juntamente com transcrições de amostras de fala de pessoas que gaguejam [Bloodstein et al. 2021].

Apesar dos avanços na área, os métodos de AM para transcrição e classificação da fala gaguejada ainda se limitam ao âmbito da pesquisa e, geralmente, não estão disponíveis para apoio à prática clínica do fonoaudiólogo. Portanto, este trabalho busca responder à seguinte questão de pesquisa: Qual o estado da arte sobre os métodos de transcrição e classificação da fala gaguejada por meio do AM?

Poucos trabalhos de revisão da literatura foram encontrados no contexto do presente estudo [Barret et al. 2022; Sheikh et al. 2022]. Este trabalho inclui questões de pesquisa não consideradas anteriormente como sobre métodos de transcrição de fala e aplicação clínica como desfecho. Assim, a revisão sistemática da literatura (RSL) apresentada neste trabalho busca fornecer uma visão atualizada dos métodos de AM utilizados para transcrever e classificar a fala gaguejada no período de 2019 a 2023, contribuindo para a identificação de lacunas e para o desenvolvimento de pesquisas futuras, com foco na prática clínica.

Este artigo está organizado da seguinte forma: A Seção 2 introduz conceitos e descreve alguns trabalhos relacionados. Na Seção 3, são apresentados os procedimentos metodológicos adotados. A Seção 4 dispõe os resultados e discussões sobre os achados da revisão. Por fim, a Seção 5 discorre sobre as considerações finais e indica pesquisas futuras.

2. Fundamentação Teórica e Trabalhos Relacionados

A capacidade de máquinas entenderem textos ou áudios por meio de linguagem natural é ainda um desafio [Oliveira et al. 2022]. O Processamento de Linguagem Natural (PLN) envolve não

somente este desafio de entendimento mas também o contexto de transcrições de fala. Neste panorama, os ASR, que transcrevem automaticamente a fala em texto escrito, têm sido utilizados para lidar com dados de entrada na forma de áudio [Schneider et al. 2019]. Juntamente com estratégias de PLN, métodos de AM e de AP têm sido empregados não somente para transcrições mas, principalmente, para classificação de disfluências. São diversas as opções de métodos que podem ser usados para classificação [Barret et al. 2022; Sheikh et al. 2022]. Como ilustração, destacam-se aqui os métodos SVM, RNN, LSTM e CNN.

O SVM, do inglês *Support Vector Machine*, ou Máquina de Vetores de Suporte, é um classificador linear que separa as amostras de dados em suas classes correspondentes, criando uma linha ou hiperplano [Cervantes et al. 2020]. O SVM tem se destacado em problemas de classificação em diversos campos, devido à sua ótima capacidade em lidar com conjuntos de dados complexos e em identificar padrões em grandes volumes de dados [Sheik et al. 2022]. Já as Redes Neurais Recorrentes (RNNs) possuem conexões entre neurônios/nós que formam um grafo direcionado ao longo de uma sequência temporal, permitindo assim que a rede exiba um comportamento dinâmico ao longo do tempo [Oliveira et al. 2022]. As redes LSTM (*Longest Shortest Term Memory*) têm sido exploradas pela sua capacidade de compreender a linguagem, destacando-se em aplicações como Classificação de Texto e Modelagem de Linguagem [Oliveira et al. 2022]. Modelos baseados em Redes Neurais Convolucionais (CNNs) têm se destacado por utilizarem estruturas de redes neurais profundas para aprender, por exemplo, representações de texto. As CNNs são redes neurais projetadas para trabalhar com dados estruturados em grade, como imagens, áudio, espectrogramas, quadros de vídeo, entre outros [Chollet 2021]. Uma CNN é composta por várias camadas em um pipeline, incluindo camadas de convolução, pooling e camadas totalmente conectadas. Por meio de múltiplos mapas de características, as CNNs conseguem capturar as dependências espaciais e temporais dos dados de entrada [Chollet 2021]. Mais recentemente, arquiteturas baseadas em Transformer têm sido usadas também em tarefas de PLN, devido principalmente a seus mecanismos de atenção, permitindo compreensão contextual e extração de informações do texto [Oliveira et al. 2022].

Neste cenário, uma revisão sistemática com meta-análise sobre métodos de AM e AP para identificar a gagueira [Barret et al. 2022] avaliou modelos para a classificação automática de disfluências na fala. A revisão aponta que ainda não se sabe quais características do sinal de áudio são mais relevantes para distinguir a fala fluente da disfluente, nem quais métodos de AM são mais eficazes para classificar as disfluências. Sheikh et al. [2022] realizaram outra revisão ampla da literatura, porém não sistemática, com foco na aplicação de métodos de AM para a identificação da gagueira. A revisão citada expõe uma análise detalhada e comparativa dos diferentes métodos utilizados para classificar as disfluências na fala de pessoas que gaguejam, bem como das características acústicas analisadas. Entre os desafios observados, os pesquisadores elencam a escassez de conjuntos de dados para pesquisas sobre identificação da gagueira. Além disso, destacam que os conjuntos de dados de falas gaguejadas frequentemente sofrem desequilíbrios, nos quais o número de amostras disponíveis para as diferentes categorias das disfluências não é uniforme.

Apesar desses trabalhos relacionados [Barret et al. 2022; Sheikh et al. 2022] investigarem métodos de AM para a identificação da gagueira, eles não incluíram nas suas questões de pesquisa como é realizado o procedimento de transcrição das amostras de fala, atividade que precede a classificação das disfluências. O presente trabalho busca identificar métodos de transcrição de fala. Outro diferencial da presente revisão é investigar se os estudos que utilizaram métodos de AM apresentaram a aplicação clínica como desfecho. Logo, a presente revisão torna-se relevante por oferecer uma visão atualizada sobre os métodos de AM em procedimentos fundamentais para o processo de diagnóstico da gagueira, no tocante à

transcrição da fala gaguejada e à classificação das disfluências e na possível integração destes procedimentos baseados em AM nas práticas clínicas.

3. Metodologia

Esta revisão foi conduzida a partir das recomendações de Kitchenham e seu grupo [Kitchenham et al. 2007]. Para isso, previamente, o protocolo de pesquisa foi elaborado contendo a proposta da revisão, objetivo, questões de pesquisa, *strings* de busca, recorte temporal, critérios de inclusão e exclusão, fontes de informação, trabalhos relacionados e cronograma. Para responder à questão de pesquisa mais geral, elegeu-se, as seguintes sub-questões de pesquisa: (1) Quais são os métodos de AM utilizados para a transcrição e classificação da fala gaguejada?; (2) Quais conjuntos de dados foram empregados e em que idiomas e países?; (3) Quais critérios diagnósticos fonoaudiológicos foram adotados para identificar uma amostra de fala como gaguejada e quais disfluências foram classificadas?; (4) Quais estudos foram aplicados à prática clínica?; (5) Quais desafios observados?

Os critérios de inclusão para a seleção de artigos foram: trabalhos primários e completos, publicados em inglês ou português brasileiro, no período de 2019 a 2023, que respondessem a pelo menos uma das questões de pesquisa. Foram excluídos artigos secundários ou terciários, assim como resumos de conferências e trabalhos não revisados por pares. Utilizou-se as seguintes bases de dados: *Institute of Electrical and Electronic Engineers* (IEEE), *Association for Computing Machinery* (ACM), *Science Direct*, *Web of Science* (WoS), *Medical Literature Analysis and Retrievel System Online* e *National Library of Medicine* (MEDLINE/PubMed) e Biblioteca Digital da Sociedade Brasileira de Computação (SBC-OpenLib).

Para a busca dos artigos, foram definidas as seguintes *strings* de busca em inglês: ("*stuttering*" OR "*stammering*" OR "*fluency disorder*") para identificar a população de interesse; ("*machine learning*" OR "*deep learning*") para especificar os métodos; e ("*transcription*" OR "*detection*" OR "*speech recognition*") para definir as aplicações dos métodos de AM utilizadas para o diagnóstico da gagueira. As *strings* foram combinadas em sequência (população, métodos e aplicação), por meio do operador booleano "AND". A busca dos artigos foi realizada por único revisor.

Os artigos coletados foram inseridos no *software Rayyan*, uma plataforma *web* gratuita que oferece suporte à seleção de referências para revisões sistemáticas [Ouzzani et al. 2016]. Após a inserção no *Rayyan*, realizou-se a remoção dos artigos duplicados, em seguida, para a seleção dos artigos que foi realizada por 3 revisores blindados, de forma manual e em três fases. A primeira fase ocorreu a partir da leitura do título e resumo dos artigos. Na segunda fase, realizou-se a leitura da introdução e conclusão dos artigos. E na terceira fase, a leitura do texto completo para aplicação dos critérios de elegibilidade. Ao término de cada fase, uma reunião para resolução dos conflitos foi realizada com outros 3 revisores, sendo 2 especialistas em Computação e 1 especialista em Fonoaudiologia.

Os estudos incluídos foram organizados por ano de publicação em planilhas no *software Excel*, conforme as seguintes variáveis: autor e ano, base de dados, título, métodos para transcrição, métodos para classificação das disfluências, disfluências classificadas, conjuntos de dados utilizados, país, idioma, aplicação à prática clínica e desafios observados. Os 3 revisores que participaram da seleção dos artigos, também, de forma blindada, extraíram os dados de todos os estudos selecionados. Ao término, realizou-se a consolidação dos achados dispostos a seguir.

4. Resultados e Discussão

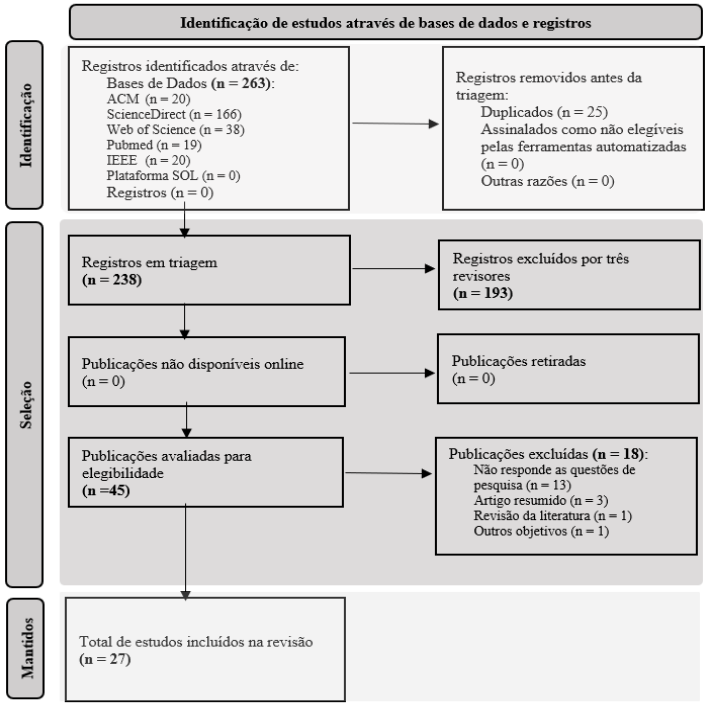

Figura 1. Fluxograma do processo de seleção de trabalhos

Conforme a Figura 1, inicialmente foram identificados 263 artigos nas bases de dados e, destes, 25 estudos duplicados foram removidos. Na primeira fase, 45 estudos foram selecionados. Na segunda fase, 27 estudos permaneceram na amostra e seguiram mantidos na terceira fase da seleção.

Tabela 1. Trabalhos primários selecionados na RSL

ID	Trabalho	Base	Título
A1	(Manjutha et. al., 2019)	IEEE	An Optimized Cepstral Feature Selection method for Dysfluencies Classification using Tamil Speech Dataset
A2	(Su et. al., 2019)	IEEE	Cross-Domain Deep Visual Feature Generation for Mandarin Audio\Visual Speech Recognition
A3	(Arbajian et. al., 2019)	Web of Science	Effect of speech segment samples selection in stutter block detection and remediation
A4	(Alharbi et. al., 2020)	Science Direct	Sequence labeling to detect stuttering events in read speech
A5	(Gupta et. al., 2020)	Web of Science	Deep Learning Bidirectional LSTM based Detection of Prolongation and Repetition in Stuttered Speech using Weighted MFCC
A6	(Kourkounakis et. al., 2020)	IEEE	Detecting Multiple Speech Disfluencies Using a Deep Residual Network with Bidirectional Long Short-Term Memory
A7	(Adepu et. al., 2020)	IEE	Interviewee Performance Analyzer Using Facial Emotion Recognition and Speech Fluency Recognition
A8	(Mishra et. al., 2021)	Web of Science	Optimization of stammering in speech recognition applications
A9	(Kourkounakis et. al., 2021)	ACM	FluentNet: End-to-End Detection of Stuttered Speech Disfluencies With Deep Learning
A10	(Sheikh et. al., 2021)	IEEE	StutterNet: Stuttering Detection Using Time Delay Neural Network
A11	(Al-Banna et. al., 2022)	Web of Science	Stuttering Disfluency Detection Using Machine Learning Approaches
A12	(Prabhu & Seliya, 2022)	IEEE	A CNN-Based Automated Stuttering Identification System
A13	(Bayerl et. al., 2022)	Web of Science	Detecting Dysfluencies in Stuttering Therapy Using wav2vec 2.0
A14	(Jouaiti & Dautenhahn, 2022)	IEEE	Dysfluency Classification in Stuttered Speech Using Deep Learning for Real-Time Applications
A15	(Jegan & Jayagowri, 2022)	Web of Science	MFCC and Texture Descriptors based Stuttering Dysfluencies Classification using Extreme Learning Machine
A16	(Mohapatra et. al., 2022)	ACM	Speech Disfluency Detection with Contextual Representation and Data Distillation
A17	(Murugan, Cherukuri & Donthu, 2022)	IEEE	Efficient Recognition and Classification of Stuttered Word from Speech Signal using Deep Learning Technique
A18	(Sheikh et. al., 2022)	IEEE	Robust Stuttering Detection via Multi-task and Adversarial Learning
A19	(Deepak et. al., 2022)	Science Direct	An artificially intelligent approach for automatic speech processing based on triune ontology and adaptive tribonacci deep neural networks
A20	(Sharma et. al., 2023)	Science Direct	Comparative analysis of various feature extraction techniques for classification of speech disfluencies
A21	(Filipowicz & Kostek, 2023)	Web of Science	Rediscovering Automatic Detection of Stuttering and Its Subclasses through Machine Learning-The Impact of Changing Deep Model Architecture and Amount of Data in the Training Set
A22	(Sheikh et. al., 2023)	IEEE	Advancing Stuttering Detection via Data Augmentation, Class-Balanced Loss and Multi-Contextual Deep Learning.
A23	(Asci et. al., 2023)	Web of Science	Acoustic analysis in stuttering: a machine-learning study.
A24	(Mohapatra et. al., 2023)	IEEE	Efficient Stuttering Event Detection Using Siamese Networks
A25	(Liao et al. 2023)	ACM	Improving Readability for Automatic Speech Recognition Transcription
A26	(Deng et. al., 2023)	ACM	Confidence Score Based Speaker Adaptation of Conformer Speech Recognition Systems
A27	(Bayerl et. al., 2023)	Web of Science	Classification of stuttering The ComParE challenge and beyond

Nas próximas seções, são apresentados e discutidos os resultados da revisão conforme as questões de pesquisa definidas.

4.1. Quais são os métodos de AM utilizados para a transcrição e classificação da fala gaguejada?

A transcrição de amostras de falas gaguejadas foi objeto de estudo de apenas um artigo (A25). O artigo utiliza o método APR (*Post-Processing for Readability*) como etapa de pós-processamento do ASR, para melhorar a qualidade das transcrições. O método realiza correções gramaticais, melhora a fluidez e a formatação, bem como converte a linguagem informal em linguagem escrita formal. Dessa forma, o trabalho A25 realizou a testabilidade do APR em modelos pré-treinados Transformer, com destaque para o desempenho superior do modelo RoBERTa, quando comparados aos modelos MASSA e UniLM. Os pesquisadores concluem que a escassez de pesquisas na área reflete uma lacuna significativa, o que indica um horizonte promissor para novos estudos, uma vez que a transcrição da fala gaguejada é a primeira etapa para o diagnóstico da gagueira.

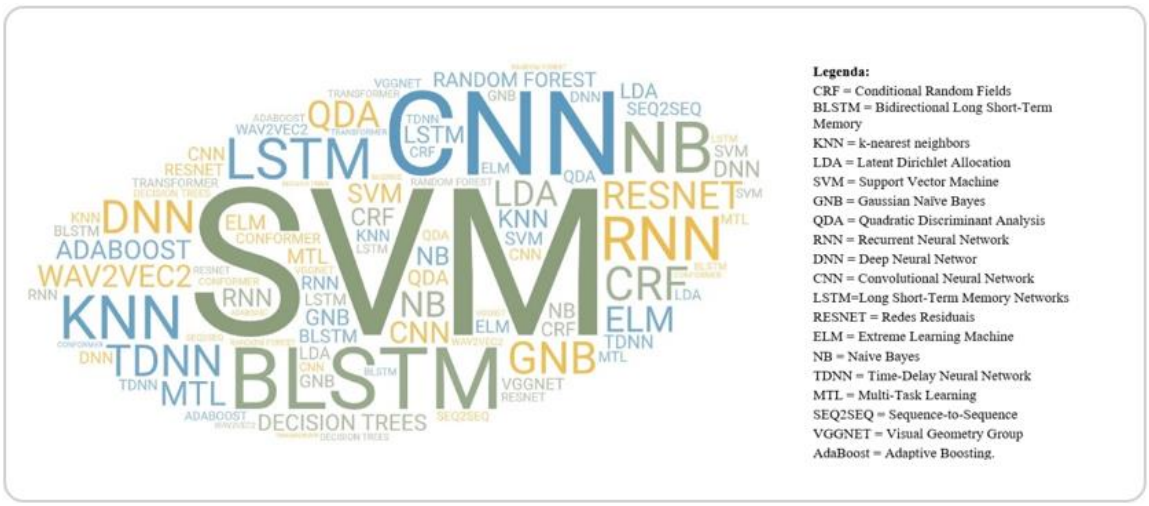

Figura 2 - Métodos de AM utilizados para classificação das disfluências

A Figura 2 apresenta os métodos de AM utilizados nos trabalhos selecionados na presente RSL para classificação das disfluências. Destacam-se quantitativamente o SVM (n=8), CNN (n=8), BLSTM (n=6) e LSTM (n=5) como os métodos mais utilizados. Cabe aqui esclarecer que, apesar de utilizar ontologia (TriNNOnto), o artigo A19 foi incluído por apresentar um modelo híbrido associado ao AM, por meio da rede neural profunda Tribonacci [Deepak et al. 2022]. Assim como observado em Barret et al. [2022], os métodos SVM e CNN foram os mais utilizados pelos estudos, por sua eficácia em lidar com problemas de classificação, como é necessário para a classificação das disfluências. Eles possuem capacidade para lidar com alta dimensionalidade de dados e separação não linear das classes. Os métodos BLSTM e LSTM também foram frequentemente utilizados por lidarem com sequências de dados e serem capazes de capturar dependências temporais, como é o caso da fala, caracterizada por uma sequência de sons ao longo do tempo.

4.2. Quais conjuntos de dados foram utilizados e em idiomas e países?

A Tabela 2 apresenta os conjuntos de dados de áudio com amostras de falas gaguejadas utilizados nas pesquisas encontradas sobre identificação da gagueira. Em termos quantitativos de uso, destaca-se o conjunto de dados UCLASS (*University College London Archive of*

Stuttered Speech). Os demais conjuntos mais utilizados foram o FluencyBank, o Sep-28k e o LibriStutter.

Conjunto de Dados	Acesso Web	País	Idioma	Quantidade de Trabalhos
Tamil Speech Dataset	https://data.ldcil.org/tamil-raw-speechh-corpus	Índia	Tamil	1
DAPRA GALE	https://catalog.ldc.upenn.edu/LDC2017S25	China	Mandarim	1
UCLASS	https://www.uclass.psychol.ucl.ac.uk/	Reino Unido	Inglês	10
Speech Accent Archive	https://accent.gmu.edu/	Estados Unidos	Inglês	2
LibriSpeech	https://www.openslr.org/12	Estados Unidos,	Inglês	2
LibriStutter	https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP3/NKVOGQ	Canadá	Inglês	2
FluencyBank	https://fluency.talkbank.org/	Estados Unidos	Inglês	5
SEP-28k	https://paperswithcode.com/dataset/sep-28k	Estados Unidos	Inglês	4
KSoF-C	https://paperswithcode.com/dataset/ksof	Alemanha	Alemão	1
Dados Próprios	https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2023.1169707/full#supplementary-material	Itália	Italiano	1
TORGO	https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html	Canadá	Inglês	1
GEC data	https://gotutiyan.github.io/GEC-Info/	Estados Unidos	Inglês	1
AMI Corpus	https://groups.inf.ed.ac.uk/ami/corpus/	Escócia	Inglês	1

Tabela 2 - Conjuntos de dados distribuídos por país, idioma e quantidade de trabalhos

Possivelmente o UCLASS foi mais utilizado por disponibilizar arquivos de áudio com transcrições de falas gaguejadas alinhadas temporalmente com os sinais de áudio, conforme requerido pelos modelos de treinamento. Destaca-se a ausência de pesquisas brasileiras e de conjuntos de dados com falas gaguejadas em português brasileiro. A maior parte das pesquisas foi realizada nos Estados Unidos, sendo o inglês o idioma mais frequente nos conjuntos de dados. Cabe pontuar que nem todas as pesquisas utilizaram os conjuntos de dados dispostos na Tabela 2. O A23 coletou 53 amostras de fala de pessoas que gaguejam e 71 de falantes fluentes, com o objetivo de fazer uso do método SVM para distinguir os grupos por meio da análise da qualidade vocal, e não das disfluências. Portanto, criar e disponibilizar publicamente conjuntos de dados com amostras de falas gaguejadas nos diversos idiomas é hoje uma necessidade para o avanço das pesquisas.

4.3. Quais critérios diagnósticos fonoaudiológicos foram adotados para identificar uma amostra de fala como gaguejada e quais disfluências foram classificadas?

A Tabela 3 apresenta as disfluências classificadas por meio do AM. As disfluências mais estudadas foram: prolongamento, interjeição, bloqueio, repetição de palavras e repetição de som. Observa-se que a classificação utilizada não esclareceu a tipologia do prolongamento ou da repetição, o que compromete a classificação das disfluências para fins diagnósticos. Por exemplo, a depender do tipo da palavra na qual a repetição ocorre, se monossilábica ou não, a disfluência pode ser considerada DTG ou OD. Apesar da possibilidade de ocorrência de OD na amostra de fala de quem gagueja, para diagnóstico da gagueira, as DTG devem ser mais frequentes. Caso contrário, pode-se considerar a possibilidade de ocorrência da taquifemia, outro transtorno da fluência [Oliveira et al. 2023]. Logo, observa-se pouca clareza no tocante à tipologia da disfluência estudada e critérios fonoaudiológicos adotados para a classificação das amostras de fala como gaguejadas.

Tabela 3 - Disfluências Identificadas

Disfluências	Quantidade de Trabalhos
Repetição de Som	10
Repetição de Sílaba	3
Repetição de Palavra (sem especificação do tipo de palavra)	11
Repetição (sem especificação da tipologia)	5
Repetição de Frase	4
Bloqueio	11
Prolongamento (sem especificação da posição na palavra)	17
Interjeição	13
Revisão	2

Interessante pontuar que apenas cinco trabalhos não se propuseram a analisar disfluências específicas. Além do A23, mencionado na seção anterior, encontram-se o A2, A7, A19 e A26. O A2 propôs um método de AM para geração de características visuais com a finalidade de reconhecimento audiovisual da fala gaguejada em Mandarim. Já no artigo A7 objetivou-se desenvolver um sistema automatizado para investigar o desempenho de pessoas entrevistadas para vagas de emprego, para isso, analisou-se o reconhecimento de emoções faciais e da fluência na fala, classificando os discursos dos entrevistados como fluentes ou não fluentes, sem identificação específica das disfluências. Em A19 o estudo teve como objetivo elaborar um modelo ASR multilíngue para aprimorar o processo de detecção do reconhecimento automático da fala, utilizando ontologia e modelagem de características, formando assim as entradas de uma rede neural. Por fim, em A26 o objetivo foi desenvolver técnicas de adaptação do próprio falante baseadas em escores de confiança para sistemas de reconhecimento de fala, visando personalizar os ASR para usuários individuais, utilizando representações de parâmetros dependentes do próprio falante.

4.4. Quais estudos foram aplicados à prática clínica?

Quanto à aplicação dos métodos de AM, na transcrição de amostras de falas gaguejadas e classificação das disfluências, observou-se que nenhum dos artigos apresentou como desfecho um produto que pudesse ser utilizado clinicamente para auxiliar no diagnóstico da gagueira. Tal observação evidencia a necessidade de se desenvolver pesquisas que também possuam este objetivo.

4.5. Quais os desafios observados?

No tocante à transcrição da fala gaguejada, um dos desafios mais significativos reside na ampla variação da fala humana. Essa variação inclui diferenças na pronúncia, entonação, ritmo e demais aspectos da fala, o que torna a sua análise e interpretação uma tarefa complexa para os sistemas de ASR. Tal variabilidade apresenta-se como obstáculo para o desenvolvimento de modelos robustos de reconhecimento de fala e para identificação precisa dos padrões das disfluências [Manjutha et al. 2019].

No que diz respeito à classificação das disfluências, a disponibilidade limitada de conjuntos de dados representativos de falas gaguejadas tem sido um desafio importante. Especialmente para o idioma português brasileiro, onde não foram encontrados estudos nem conjuntos de dados. A obtenção de dados de treinamento e teste de qualidade é necessária para o desenvolvimento e validação de modelos para identificação da fala gaguejada, pois a falta desses dados pode limitar a capacidade de avanços na área [Kourkounakis et al. 2020].

5. Considerações Finais

Esta RSL objetivou conhecer o estado da arte dos métodos de AM para transcrição e classificação da fala gaguejada no período de 2019 a 2023. Através desta pesquisa, foi possível

observar a escassez de estudos sobre transcrição automática da fala gaguejada, e identificar que os métodos de AM mais frequentemente utilizados para a detecção das disfluências foram SVM, CNN, BLSTM e LSTM, pela eficácia em lidar com problemas de classificação e sequências de dados, respectivamente. O principal conjunto de dados utilizado foi o UCLASS, do Reino Unido, disponibilizado publicamente com amostras de fala transcritas em inglês e alinhadas temporalmente aos arquivos de áudio.

Prolongamentos, interjeições, bloqueios, repetições de palavras e de sons foram as disfluências mais frequentemente estudadas. Porém, observou-se pouca clareza sobre quais disfluências estavam sendo consideradas DTG e OD, possivelmente, em virtude da distância entre o âmbito das pesquisas sobre identificação da gagueira e a prática clínica diagnóstica. Esta limitada aproximação mostra-se evidente desde a exposição inespecífica dos critérios diagnósticos adotados para classificar as amostras de fala como gaguejadas, até a ausência de desfechos aplicáveis ao processo de trabalho do fonoaudiólogo. Como principais desafios destacaram-se a própria natureza singular da fala humana e suas variações, bem como a limitada oferta de conjuntos de dados em diferentes idiomas.

Como limitação deste estudo destaca-se a ausência de análise das características acústicas da fala como, por exemplo, transformada de Fourier, MFCCs (*Coeficientes Cepstrais de Frequência Mel*) e características espectrais. Essas técnicas são utilizadas para entender as propriedades acústicas da gagueira e melhorar a precisão dos modelos de AM. Sugere-se, portanto, que em próximos estudos essas análises sejam consideradas a fim de favorecer o entendimento acerca das variações e especificidades das disfluências.

Observou-se que as pesquisas sobre métodos de AM na área da transcrição e classificação da fala gaguejada ainda estão em seus primeiros passos. Neste sentido, almeja-se que pesquisas futuras possam criar e disponibilizar conjuntos de dados em diversos idiomas, especialmente em português brasileiro. Busca-se, adicionalmente, o desenvolvimento de abordagens de AM para fins de diagnóstico da gagueira, com métodos para transcrição e detecção da fala gaguejada de modo aplicado à prática clínica.

6. Referências

Adepu, Y., Boga, V. R., & Sairam, U. (2020, November). Interviewee performance analyzer using facial emotion recognition and speech fluency recognition. In *2020 IEEE International Conference for Innovation in Technology (INOCON)* (pp. 1-5). IEEE.

Al-Banna, A. K., Edirisinghe, E., Fang, H., & Hadi, W. (2022). Stuttering disfluency detection using machine learning approaches. *Journal of Information & Knowledge Management*, *21*(02), 2250020.

Alharbi, S., Hasan, M., Simons, A. J., Brumfitt, S., & Green, P. (2020). Sequence labeling to detect stuttering events in read speech. *Computer Speech & Language*, *62*, 101052.

Almeida, R. J. S., Fernandes, D. Y. S., Oliveira, L. P., & Correia, D. V. (2023). Desafios e oportunidades na integração do ambiente clínico e digital para apoio ao diagnóstico da gagueira. *Computação Brasil*, (51), 37-41.

Ambrose, N. G., & Yairi, E. (1999). Normative disfluency data for early childhood stuttering. *Journal of Speech, Language, and Hearing Research*, *42*(4), 895-909.

American Psychiatric Association. (2022). Childhood-Onset Fluency Disorder (Stuttering). In Diagnostic and statistical manual of mental disorders (5th ed.).

Andrade, C. D., Befi-Lopes, D. M., Fernandes, F. D. M., & Wertzner, H. F. (2004). ABFW: teste de linguagem infantil nas áreas de fonologia, vocabulário, fluência e pragmática. *São Paulo: Pró-Fono*.

Arbajian, P., Hajja, A., Raś, Z. W., & Wieczorkowska, A. A. (2019). Effect of speech segment samples selection in stutter block detection and remediation. *Journal of Intelligent Information Systems*, *53*, 241-264.

Asci, F., Marsili, L., Suppa, A., Saggio, G., Michetti, E., Di Leo, P., & Costantini, G. (2023). Acoustic analysis in stuttering: a machine-learning study. *Frontiers in Neurology*, *14*, 1169707.

Barrett, L., Hu, J., & Howell, P. (2022). Systematic review of machine learning approaches for detecting developmental stuttering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 1160-1172.

Bayerl, S. P., Wagner, D., Nöth, E., & Riedhammer, K. (2022). Detecting dysfluencies in stuttering therapy using wav2vec 2.0. *arXiv preprint arXiv:2204.03417*.

Bayerl, S. P., Gerczuk, M., Batliner, A., Bergler, C., Amiriparian, S., Schuller, B., ... & Riedhammer, K. (2023). Classification of stuttering–The ComParE challenge and beyond. *Computer Speech & Language*, *81*, 101519.

Bloodstein, O., Ratner, N. B., & Brundage, S. B. (2021). *A handbook on stuttering*. Plural Publishing.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189-215.

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Deepak, G., Surya, D., Trivedi, I., Kumar, A., & Lingampalli, A. (2022). An artificially intelligent approach for automatic speech processing based on triune ontology and adaptive tribonacci deep neural networks. *Computers & Electrical Engineering*, *98*, 107736.

Deng, J., Xie, X., Wang, T., Cui, M., Xue, B., Jin, Z., ... & Meng, H. (2022). Confidence score based conformer speaker adaptation for speech recognition. *arXiv preprint arXiv:2206.12045*.

Filipowicz, P., & Kostek, B. (2023). Rediscovering Automatic Detection of Stuttering and Its Subclasses through Machine Learning—The Impact of Changing Deep Model Architecture and Amount of Data in the Training Set. *Applied Sciences*, *13*(10), 6192.

Fox, C. B., Israelsen-Augenstein, M., Jones, S., & Gillam, S. L. (2021). An evaluation of expedited transcription methods for school-age children's narrative language: automatic speech recognition and real-time transcription. *Journal of Speech, Language, and Hearing Research*, *64*(9), 3533-3548.

Gupta, S., Shukla, R. S., Shukla, R. K., & Verma, R. (2020). Deep learning bidirectional LSTM based detection of prolongation and repetition in stuttered speech using weighted MFCC. *International Journal of Advanced Computer Science and Applications*, *11*(9).

Howell, P., & Sackin, S. (1995, August). Automatic recognition of repetitions and prolongations in stuttered speech. In *Proceedings of the first World Congress on fluency disorders* (Vol. 2, pp. 372-374). Nijmegen, The Netherlands: University Press Nijmegen.

Howell, Peter & Davis, Stephen & Bartrip, Jon. (2009). The University College London Archive of Stuttered Speech (UCLASS). *Journal of speech, language, and hearing research*: JSLHR. 52. 556-69. 10.1044/1092-4388(07-0129).

Jegan, R., & Jayagowri, R. (2022). MFCC and texture descriptors based stuttering dysfluencies classification using extreme learning machine. *International Journal of Advanced Computer Science and Applications*, *13*(8).

Jouaiti, M., & Dautenhahn, K. (2022, May). Dysfluency classification in stuttered speech using deep learning for real-time applications. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6482-6486). IEEE.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. *Technical Report EBSE* 2007-001, Keele University and Durham University Joint Report.

Kourkounakis, T., Hajavi, A., & Etemad, A. (2020, May). Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6089-6093). IEEE.

Kourkounakis, T., Hajavi, A., & Etemad, A. (2021). Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 2986-2999.

Lea, C., Mitra, V., Joshi, A., Kajarekar, S., & Bigham, J. P. (2021, June). Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6798-6802). IEEE.

Liao, J., Eskimez, S., Lu, L., Shi, Y., Gong, M., Shou, L., ... & Zeng, M. (2023). Improving readability for automatic speech recognition transcription. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(5), 1-23.

Manjutha, M., Subashini, P., Krishnaveni, M., & Narmadha, V. (2019, October). An optimized cepstral feature selection method for dysfluencies classification using Tamil speech dataset. In *2019 IEEE International Smart Cities Conference (ISC2)* (pp. 671-677). IEEE.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mishra, N., Gupta, A., & Vathana, D. (2021). Optimization of stammering in speech recognition applications. *International Journal of Speech Technology*, *24*(3), 679-685.

Mitchell, T. (1997). Machine learning.-New York, NY, USA: McGraw Hill. *Inc. isbn*, *70428077*.

Mohapatra, P., Islam, B., Islam, M. T., Jiao, R., & Zhu, Q. (2023, June). Efficient stuttering event detection using siamese networks. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

Mohapatra, P., Pandey, A., Islam, B., & Zhu, Q. (2022, July). Speech disfluency detection with contextual representation and data distillation. In *Proceedings of the 1st ACM international workshop on intelligent acoustic systems and applications* (pp. 19-24).

Murugan, K., Cherukuri, N. K., & Donthu, S. S. (2022, June). Efficient Recognition and Classification of Stuttered Word from Speech Signal using Deep Learning Technique. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)* (pp. 774-781). IEEE.

Oliveira, B. S. N., do Rêgo, L. G. C., Peres, L., da Silva, T. L. C., & de Macêdo, J. A. F. (2022). Processamento de linguagem natural via aprendizagem profunda. *Sociedade Brasileira de Computação*.

Oliveira, C. M. C., Correia, D. V., & Di Ninno, C. Q. M. S. (2023). Avaliação da Fluência. In C. A. S. Azoni, J. O. de Lira, D. A. C. Lamônica, D. B. de Oliveira e Britto (Orgs.), *Tratado de Linguagem: perspectivas contemporâneas.* (2ª ed., pp. 109-117). Ribeirão Preto, SP: Book Toy.

Oliveira, L. P., Santos, J. H. D. S., de Almeida, E. L., Barbosa, J. R., da Silva, A. W., de Azevedo, L. P., & da Silva, M. V. (2021, April). Deep learning library performance analysis on raspberry (IoT device). In *International Conference on Advanced Information Networking and Applications* (pp. 383-392). Cham: Springer International Publishing.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, *5*, 1-10.

Prabhu, Y., & Seliya, N. (2022, December). A CNN-based automated stuttering identification system. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1601-1605). IEEE.

Ratner, N. B., & MacWhinney, B. (2018). Fluency Bank: A new resource for fluency research and practice. *Journal of fluency disorders*, *56*, 69-80.

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Sharma, N. M., Kumar, V., Mahapatra, P. K., & Gandhi, V. (2023). Comparative analysis of various feature extraction techniques for classification of speech disfluencies. *Speech Communication*, *150*, 23-31.

Sheikh, S. A., Sahidullah, M., Hirsch, F., & Ouni, S. (2021, August). Stutternet: Stuttering detection using time delay neural network. In *2021 29th European Signal Processing Conference (EUSIPCO)* (pp. 426-430). IEEE.

Sheikh, S. A., Sahidullah, M., Hirsch, F., & Ouni, S. (2022, August). Robust stuttering detection via multi-task and adversarial learning. In *2022 30th European Signal Processing Conference (EUSIPCO)* (pp. 190-194). IEEE.

Sheikh, S. A., Sahidullah, M., Hirsch, F., & Ouni, S. (2023). Advancing stuttering detection via data augmentation, class-balanced loss and multi-contextual deep learning. *IEEE Journal of Biomedical and Health Informatics*.

Sheikh, S. A., Sahidullah, M., Hirsch, F., & Ouni, S. (2022). Machine learning for stuttering identification: Review, challenges and future directions. *Neurocomputing*, *514*, 385-402.

Su, R., Liu, X., Wang, L., & Yang, J. (2019). Cross-domain deep visual feature generation for mandarin audio–visual speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 185-197.