UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Soluções Tecnológicas para a Fonoaudiologia: Um Software de Avaliação da Gagueira baseado em Inteligência Artificial

Rubens Perini Buzzeti

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Rubens Perini Buzzeti

Soluções Tecnológicas para a Fonoaudiologia: Um Software de Avaliação da Gagueira baseado em Inteligência Artificial

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Roney Lira de Sales Santos

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Buzzeti, Rubens Perini
B992s Soluções Tecnológicas para a Fonoaudiologia: Um Software de Avaliação da Gagueira baseado em Inteligência Artificial / Rubens Perini Buzzeti; orientador Roney Lira de Sales Santos. -- São Carlos, 2025.
96 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Inteligência Artificial. 2. Fonoaudiologia. 3. Aprendizado de Máquina. 4. Gagueira. 5. Redes Neurais. I. Lira de Sales Santos, Roney, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Rubens Perini Buzzeti

Technological Solutions for Speech Therapy: An Artificial Intelligence-Based Software for Stuttering Assessment

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Roney Lira de Sales Santos

Original version

São Carlos

2025

Dedico esse trabalho aos meus pais, que me ensinaram que o estudo é o melhor caminho para o desenvolvimento, e à minha esposa Paula Moura, mulher inspiradora e a maior profissional responsável por difundir a fonoaudiologia nas mídias sociais do Brasil.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão a todos que contribuíram para a realização deste trabalho.

Primeiramente, agradeço ao meu orientador, **Prof. Dr. Roney Santos**, pela orientação paciente, rigorosa e inspiradora ao longo de todo o desenvolvimento do projeto. Sua expertise técnica, visão estratégica e disponibilidade constante foram fundamentais para a evolução do sistema e para a superação de cada desafio técnico enfrentado.

Agradeço também ao **Dr. Edresson Casanova**, cuja colaboração foi de extrema importância para o desenvolvimento do modelo de transcrição.

Um agradecimento especial à minha esposa, **Paula Moura**, fonoaudióloga, cuja compreensão nos momentos de dedicação intensa e valiosa perspectiva clínica enriqueceram significativamente o projeto. Sua parceria foi um pilar essencial durante toda a jornada.

Registro minha sincera gratidão à **Fonoaudióloga Patrícia França**, que, com generosidade e expertise, colaborou diretamente na coleta, anotação e estruturação do *dataset* de disfluências. Sua contribuição prática foi indispensável para viabilizar o treinamento do modelo em dados reais de fala com gagueira.

Agradeço ainda a **todos os alunos da turma 4 do MBA em Inteligência Artificial e Big Data**, pelo companheirismo, troca de ideias e apoio mútuo ao longo do curso. As discussões em grupo, os momentos de aprendizado coletivo e a energia colaborativa da turma foram fonte constante de motivação e crescimento.

A todos, meu reconhecimento e carinho.

"Não tenho nenhum talento especial, apenas sou apaixonadamente curioso"
Albert Einstein

RESUMO

A gagueira é um distúrbio da fluência que afeta aproximadamente 1% da população mundial, caracterizada por interrupções involuntárias no fluxo da fala que impactam significativamente a vida social, acadêmica e profissional dos indivíduos. A avaliação fonoaudiológica tradicional da gagueira é um processo extremamente dispendioso, levando várias horas para ser concluído e exigindo análise minuciosa de amostras de fala com pelo menos 200 sílabas fluentes. Neste contexto, surge a necessidade de ferramentas tecnológicas que facilitem e reduzam o tempo de avaliação, contribuindo para a inclusão de pacientes em serviços especializados. O objetivo deste trabalho foi desenvolver um software que utiliza inteligência artificial para realizar, por meio de vídeos com amostras de fala, a transcrição automática, detecção, organização e categorização de disfluências para suportar o diagnóstico da gagueira. O sistema desenvolvido implementa uma abordagem híbrida que combina metodologias de inteligência artificial com métodos tradicionais: para bloqueios e prolongamentos, utiliza uma arquitetura hierárquica baseada em Rede de Kohonen (SOM) e Perceptron de Múltiplas Camadas (MLP) com 21 características espectrais extraídas através de banco de filtros de 1/3 de oitava com ponderação A, seguindo a metodologia validada de Szczurowska et al. (2006); para repetições, pausas e hesitações, emprega algoritmos determinísticos baseados em análise de transcrição automática com timestamps precisos. O desenvolvimento seguiu um processo iterativo de quatro fases, evoluindo de 54.5% para 84.2% de acurácia através de melhorias sistemáticas em balanceamento de classes, otimização de parâmetros e adoção de metodologia científica validada. O sistema foi implementado com arquitetura escalável incluindo API REST, processamento assíncrono e sistema de jobs para integração com interfaces web. Os testes em dados reais revelaram tanto o potencial quanto as limitações da abordagem, evidenciando a necessidade de datasets maiores. O sistema híbrido demonstrou capacidade de detectar e categorizar múltiplos tipos de disfluências, gerando relatórios estruturados com métricas quantitativas para suporte ao diagnóstico clínico. A principal limitação identificada inclui o tamanho restrito de amostras para treinamento (126 amostras), visto que frente à ausência de dataset público para a língua portuguesa tornou necessária a coleta específica para este projeto. Este trabalho estabelece uma base tecnológica sólida para a aplicação de inteligência artificial na Fonoaudiologia, demonstrando o potencial de sistemas híbridos como ferramentas de suporte diagnóstico e orientando desenvolvimentos futuros na área.

Palavras-chave: Gagueira. Inteligência Artificial. Aprendizado de Máquina. Fonoaudiologia. Redes Neurais.

ABSTRACT

Stuttering is a fluency disorder that affects approximately 1% of the world's population, characterized by involuntary interruptions in speech flow that significantly impact individuals' social, academic, and professional lives. Traditional assessment of stuttering is an extremely time-consuming process, taking several hours to complete and requiring meticulous analysis of speech samples with at least 200 fluent syllables. In this context, there is a need for technological tools that facilitate and reduce assessment time, contributing to patient inclusion in specialized services. The goal of this work was to develop software that uses artificial intelligence to perform, through videos with speech samples, automatic transcription, detection, organization, and categorization of disfluencies to support stuttering diagnosis. The developed system implements a hybrid approach that combines artificial intelligence methodologies with traditional methods: for blocks and prolongations, it uses a hierarchical architecture based on Kohonen Network (SOM) and Multilayer Perceptron (MLP) with 21 spectral features extracted through 1/3 octave filter bank with A-weighting, following the validated methodology of Szczurowska et al. (2006); for repetitions, pauses, and hesitations, it employs deterministic algorithms based on automatic transcription analysis with precise timestamps. Development followed a four-phase iterative process, evolving from 54.5% to 84.2% accuracy through systematic improvements in class balancing, parameter optimization, and adoption of validated scientific methodology. The system was implemented with scalable architecture including REST API, asynchronous processing, and job system for web interface integration. Real data testing revealed both the potential and limitations of the approach, highlighting the need for larger datasets. The hybrid system demonstrated the ability to detect and categorize multiple types of disfluencies, generating structured reports with quantitative metrics for clinical diagnostic support. The main limitation identified includes the restricted size of training samples (126 samples), since the absence of public datasets for the Portuguese language made specific collection necessary for this project. This work establishes a solid technological foundation for applying artificial intelligence in Speech Therapy, demonstrating the potential of hybrid systems as diagnostic support tools and guiding future developments in the field.

Keywords: Stuttering. Artificial Intelligence. Machine Learning. Speech Therapy. Neural Networks.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A gagueira é um distúrbio da fluência de fala que afeta aproximadamente 1% da população mundial (cerca de 80 milhões de pessoas ao redor do mundo), segundo a Associação Brasileira de Gagueira (ABRA GAGUEIRA), com maior prevalência na infância. Esse distúrbio, caracterizado por interrupções involuntárias no fluxo normal da fala, pode impactar significativamente a vida social, acadêmica e profissional dos indivíduos.

A principal e mais evidente manifestação da gagueira são as rupturas no discurso caracterizadas como Disfluências Típicas da Gagueira (Bleek; Outros, 2012; Civier; Outros, 2013; Yairi; Ambrose, 1992; Yairi; Ambrose, 1999), cuja tipologia é distribuída em: repetição de palavras (acima de 3 repetições), repetição de parte da palavra, repetição de som, bloqueio, prolongamento, pausas maiores que 2 segundos e intrusão de sons ou palavras não pertinentes ao contexto da mensagem oral. Essas disfluências ocorrem principalmente no início das palavras ou ao longo de toda palavra (Sisskin; Wasilus, 2014). Para se fechar o diagnóstico da gagueira é necessária uma análise minuciosa de uma amostra de fala do indivíduo, composta por pelo menos 200 sílabas fluentes.

O profissional que realiza a avaliação da fluência de fala é o Fonoaudiólogo, e os procedimentos que envolvem essa avaliação são:

1. Análise da tipologia e frequência das disfluências;

2. Determinação dos fluxos de sílabas e palavras por minuto;

3. Outras análises qualitativas, como tensão e movimentos corporais associados à fala.

A presença de 3% ou mais de disfluências típicas da gagueira no discurso já caracteriza o distúrbio como gagueira, independentemente de qualquer outro tipo de manifestação.

Embora existam protocolos de avaliação validados que norteiam esses procedimentos, esse processo é extremamente dispendioso, levando várias horas para ser concluído. Nesse contexto, surge a necessidade de ferramentas tecnológicas que facilitem e reduzam o tempo de avaliação pelo fonoaudiólogo, diminuam o número de sessões para o diagnóstico e contribuam para a inclusão de pacientes em serviços especializados.

A motivação para este estudo está diretamente ligada ao crescente impacto da tecnologia na área da saúde, especialmente no campo da fluência, onde há um grande potencial para o uso de algoritmos de aprendizado de máquina. Além disso, a crescente de-

manda por soluções de telemedicina reforça a relevância do desenvolvimento de ferramentas que auxiliem os fonoaudiólogos na avaliação e tomada de decisão clínica.

O objetivo geral deste trabalho é desenvolver um *software* que utiliza inteligência artificial para que, por meio de um vídeo com amostra de fala do paciente, realize a transcrição automática da fala, determine, organize e categorize as disfluências encontradas, para suportar o diagnóstico da gagueira.

Embora envolva a análise de dados relacionados à saúde, o presente trabalho não foi submetido à apreciação do Comitê de Ética em Pesquisa (CEP), em conformidade com as diretrizes da Resolução nº 510, de 7 de abril de 2016, do Conselho Nacional de Saúde (CNS).

A referida resolução, em seu Art. 1º, estabelece que não necessitam de registro e avaliação pelo sistema CEP/CONEP as pesquisas que se enquadram em determinadas categorias. Este trabalho se alinha a essa dispensa por duas razões principais:

1. **Pesquisa com Profissionais na Condição de Participantes:** Colaboraram com este trabalho as fonoaudiólogas Paula Bianca Meireles de Moura Buzzeti CRFa 2-16524 e Patrícia da Silva Pereira França CRFa 2-20694 na condição de participantes da pesquisa, fornecendo seus conhecimentos técnicos e realizando análises comparativas como parte do desenvolvimento e validação da ferramenta. A pesquisa não incidiu sobre elas como pacientes ou sujeitos vulneráveis, mas sim como especialistas que contribuíram ativamente para a construção do conhecimento e da tecnologia aqui apresentada. A colaboração foi voluntária e teve como objetivo aprimorar a solução, caracterizando-se como uma parceria técnico-científica.
2. **Uso de Dados de Acesso Público:** Os dados de áudio e vídeo utilizados para o treinamento e teste do modelo de *machine learning* foram obtidos de fontes de acesso público, como plataformas de compartilhamento de vídeos (e.g., YouTube), onde os próprios indivíduos disponibilizaram suas falas. Não houve coleta direta de dados de pacientes ou acesso a prontuários, o que elimina o risco de identificação e a necessidade de consentimento informado para o uso desses dados, conforme previsto na Resolução CNS nº 510/16.

Desta forma, por se tratar de uma pesquisa que envolve a participação de profissionais como parte do desenvolvimento da ferramenta e a utilização de dados de acesso público, e não de uma pesquisa com seres humanos como sujeitos experimentais, o trabalho se enquadra nos critérios de dispensa de submissão ao sistema CEP/CONEP.

Este trabalho está organizado em seis capítulos, que buscam apresentar o desenvolvimento da solução tecnológica para avaliação da gagueira, desde a fundamentação teórica até a discussão dos resultados e contribuições, conforme a estrutura abaixo:

Capítulo 1 - Introdução: Apresenta o contexto da fonoaudiologia e os desafios na avaliação da gagueira, justificando a necessidade de uma ferramenta tecnológica. São definidos os objetivos geral e específicos do trabalho, que norteiam o desenvolvimento da solução baseada em Inteligência Artificial.

Capítulo 2 - Referencial Teórico: Revisa a literatura científica sobre a gagueira, detalhando as tipologias de disfluências (comuns e gagas) e os métodos de avaliação clínica. Adicionalmente, explora os conceitos de Processamento de Linguagem Natural (PLN) que fundamentam a arquitetura do sistema.

Capítulo 3 - Trabalhos Relacionados: Analisa em detalhes os principais estudos e artigos científicos sobre a detecção de disfluências utilizando Inteligência Artificial. São discutidas diferentes arquiteturas de redes neurais (SOM, MLP, *Deep Learning*), técnicas de extração de características acústicas e os principais *datasets* utilizados na área.

Capítulo 4 - Metodologia: Descreve em detalhes a metodologia empregada na construção do *software*. Apresenta a arquitetura do sistema híbrido, as tecnologias e bibliotecas utilizadas (Python, Flask, Librosa, Scikit-learn), o processo de coleta e preparação dos dados, a extração de características acústicas e o desenvolvimento iterativo do modelo de *machine learning* para detecção de bloqueios e prolongamentos.

Capítulo 5 - Resultados e Discussão: Apresenta os resultados quantitativos obtidos ao longo das quatro fases de desenvolvimento do projeto. A evolução iterativa do sistema é detalhada, partindo de uma acurácia inicial e alcançando melhorias significativas. Os resultados são discutidos, incluindo a validação em dados reais e a análise das limitações e desafios técnicos encontrados.

Capítulo 6 - Conclusão: Sintetiza as principais contribuições do trabalho, destacando as contribuições técnicas, científicas e para a prática fonoaudiológica. Discute as limitações do estudo, como o tamanho do *dataset* e o *overfitting* do modelo, e aponta direções para trabalhos futuros, incluindo a expansão da base de dados e o aprimoramento contínuo do sistema.

2 REFERENCIAL TEÓRICO

2.1 Fonoaudiologia

A Fonoaudiologia é uma ciência voltada para a prevenção, avaliação, diagnóstico e intervenção de distúrbios da comunicação humana. Seu escopo compreende dificuldades relacionadas à fala, linguagem, audição, voz e funções orofaciais. O fonoaudiólogo atua em diferentes contextos, desde a saúde pública até o ambiente escolar e corporativo, auxiliando na melhoria da comunicação e da qualidade de vida dos indivíduos (Andrade, 2004; Ferreira, 2010; Pernambuco, 2015).

A atuação fonoaudiológica baseia-se em abordagens interdisciplinares que envolvem conhecimentos da neurologia, psicologia, pedagogia e fisiologia da comunicação (Andrade, 2004; Ferreira, 2010; Pernambuco, 2015). O campo da Fonoaudiologia é subdividido em áreas específicas, como motricidade orofacial, disfagia, voz, audiologia e fluência. Em relação à fluência, as características envolvem o fluxo contínuo e suave da produção da fala (STARKWEATHER; GIVENS-ACKERMAN, 1997), mínimo esforço motor, emocional, linguístico e cognitivo (GARGANTINI, 1995) e variação de ritmo para indivíduo (ANDRADE, 2006). Como principais parâmetros, são avaliadas a continuidade, velocidade, ritmo e ausência de esforço.

Nos últimos anos, avanços tecnológicos contribuíram para o desenvolvimento de novos métodos de avaliação e intervenção fonoaudiológica, como o uso de *software* para análise de fluência e dispositivos de retroalimentação auditiva atrasada. A Fonoaudiologia também tem se beneficiado de práticas baseadas em evidências científicas para aprimorar diagnósticos e propor tratamentos personalizados (Martins; Andrade, 2008; Lopes, 2019).

A avaliação fonoaudiológica é essencial para compreender a natureza e a extensão dos distúrbios de comunicação. No caso da gagueira, a análise detalhada das características da fala permite diferenciar entre disfluências comuns e típicas da gagueira, classificando a gravidade do distúrbio e estabelecendo o tratamento adequado.

2.2 Gagueira

A gagueira é um transtorno da fluência caracterizado por interrupções involuntárias no fluxo da fala, que comprometem as características supracitadas. Essas interrupções são denominadas disfluências, que é um rompimento no fluxo da fala, comum a todos os falantes (Perkins, 1991) e podem ser classificadas em disfluências comuns e disfluências típicas da gagueira. O transtorno tem um impacto significativo na comunicação social, podendo influenciar aspectos emocionais e psicológicos do indivíduo (Yairi; Ambrose, 1999).

2.2.1 Disfluências Comuns

As disfluências comuns, também conhecidas como outras disfluências, são interrupções que ocorrem na fala de qualquer indivíduo. Que são:

Interjeição: Representada pela letra (I), é caracterizada como o uso de palavras ou frases irrelevantes ao contexto. Como nos casos: “Eu estava, tipo assim, ruindo muito” e “Aqui, você ligou para o médico?”

Hesitação: Representada pelo sinal (#) ou pela sigla (H) são pausas de até 3s durante a fala enquanto a pessoa organiza seus pensamentos. Muitas vezes ocorrem quando o falante busca uma palavra ou estrutura gramatical mais apropriada. “Ontem # eu ganhei uma boneca, podendo também ser preenchida com palavras desnecessárias, como: “éh” e “aan..”, como por exemplo em: “Eu vou éh no parque”.

Repetição de palavras, segmento ou frases (até 2x): Representada pela sigla (RP) no caso da repetição da palavra, por exemplo: “Eu eu gostaria de falar com você”. Representada pela sigla (RPE) no caso de repetição de parte do enunciado, por exemplo: “Eu queria, eu queria viajar ainda hoje”. E, representada com a sigla (RF) no caso de repetição de frases, por exemplo: “Ele não falou comigo, ele não falou comigo”.

Revisão: Representada pela sigla (Rv), indica mudança no conteúdo ou forma gramatical. Como nos casos: “Eu moro com minha meu irmão” e “Esse bolo torta de chocolate está uma delícia”.

Palavras incompletas: Representada pela sigla (PI), indica uma palavra abandonada no meio do discurso. Como nos casos: “Esse cober...edredon é muito macio” e “Eu ia com...alugar o vestido, mas desisti”.

2.2.2 Disfluências Típicas da Gagueira

As disfluências típicas da gagueira, conhecidas como as principais manifestações da gagueira, são:

Repetição de sílaba: Representada pela sigla (RSi), é indicada pela repetição de uma sílaba inteira ou de parte da palavra, como nos casos: “O gagagato pulou no telhado” e “Eu gogogosto de você”.

Repetição de som: Representada pela sigla (RS), é indicada pela repetição de um fonema, como nos casos: “O mmmmacaco comeu a banana” e “Eeeeu gosto de ouvir música”.

Repetição de palavra: Representada pela sigla (RP), é indicada pela repetição de palavras que ocorrem 3x ou mais, como nos casos: “Eu ia sair, mas mas mas mas mas começou a chover” e “Ela lavou lavou lavou lavou o cabelo”.

Prolongamento: Representado pelo sinal (_) ou pela sigla (P), é indicado pela duração inapropriada de um fonema, como nos casos: “Eu amo minha f_amília” e “Comprei u_m carro novo”.

Bloqueio: Representado pelo sinal (/) ou pela sigla (B), é indicado pelo tempo inapropriado para iniciar um fonema, geralmente associado à tensão, como nos casos: “Ele chutou a /bola” e “O /padre rezou a missa”.

Pausas: Representado pelo sinal (_____) ou pela sigla (Pa), é indicada pela interrupção do fluxo da fala por 3s ou mais, como nos casos: “Ele disse _____ que viria” e “Amanhã sairemos (Pa) bem cedo”.

Intrusão: Representado pelo sinal entre barras (/ /) ou pela sigla (Int), é indicada pela produção de sons ou cadeia de sons não pertinentes ao contexto, como nos casos: “Eu fui para a /kkkkk/ escola” e “In eee(Int) felizmente não posso”.

2.2.3 Avaliação da Gagueira atualmente

A avaliação da gagueira pode ser realizada utilizando protocolos cientificamente validados. Dentre eles, o ABFW (Avaliação da Linguagem Infantil) é muito utilizado. No caso da gagueira, é utilizado o teste de fluência do ABFW para crianças e para adultos o protocolo de avaliação do perfil da fluência, que possui exatamente o mesmo procedimento de avaliação.

O ABFW é utilizado para a determinação do diagnóstico da gagueira, que uma vez que obteve o resultado positivo, é aplicado um segundo protocolo, o SSI (*Stuttering Severity Instrument*), que determina se o paciente possui uma gagueira leve, moderada, grave ou muito grave.

Assim, para a avaliação da gagueira será considerado os seguintes passos:

Coleta da Amostra de Fala: É gravado um vídeo do paciente que pode ser realizado pelo fonoaudiólogo ou por ele próprio, onde ele fale livremente sobre um tema de interesse. Esse vídeo deve conter pelo menos 200 sílabas fluentes.

Transcrição e Análise das Disfluências: A transcrição é feita conforme os critérios supracitados, indicando assim os tipos das disfluências típicas da gagueira e comuns. O fonoaudiólogo anota pausas, repetições, prolongamentos e bloqueios, garantindo um registro fiel do padrão de fala do paciente.

Aplicação do Teste ABFW: O teste ABFW avalia quantitativamente as disfluências. O profissional contabiliza a frequência e a distribuição dos diferentes tipos de disfluências, verifica a velocidade de fala, contabilizando a quantidade de palavras e sílabas por minutos, e por último é determinada o percentual das rupturas, ou seja, o percentual da disfluências típicas da gagueira e disfluências totais (comuns + típicas da gagueira). Se o percentual de disfluências gagas for maior que 3%, é diagnosticada a gagueira.

Classificação da Gravidade pelo SSI-3: Uma vez que o diagnóstico foi realizado conforme o protocolo acima, é utilizado o SSI para determinar a média das três disfluências mais longas, assim como a quantidade de concomitantes físicos (sons dispersivos, movimentos faciais, movimentos de cabeça e movimentos das extremidades). Esses itens somam-se em uma pontuação, e de acordo com a pontuação é determinada a gravidade da gagueira.

Duração das três maiores disfluências: Mede-se a duração média das três disfluências mais longas para identificar padrões persistentes.

Presença de concomitantes físicos: Movimentos involuntários ou tensões musculares que acompanham a gagueira.

2.2.4 Avaliação da Gagueira via software

O primeiro desafio é a realização da transcrição do vídeo/áudio em texto, que deve ser fidedigno a exatamente o que foi falado, necessitando assim o ajuste dos parâmetros das ferramentas de transcrição que hoje já corrigem e ajustam o texto automaticamente.

Após essa etapa, o desafio é detectar as disfluências em si. Para esse estudo, sugere-se a seguinte abordagem:

- Será utilizado um modelo `wav2vec` para português que utiliza o *dataset* CORAA para transcrever o texto com *timestamps*, para suportar a determinação da duração dos sons e pausas.
- Para a detecção de prolongamento e bloqueio, será testada a combinação de análise espectral do áudio com a ferramenta `Librosa` (Python) e treinar um modelo com rede neural.
- Para a detecção de repetição, é possível usar técnicas de Processamento de Linguagem Natural (PLN) para verificar padrões de repetição fonético, silábico ou de palavras.
- Para detecção de hesitação, é possível comparar a duração e frequência das pausas da fala típica, podendo assim determinar qualquer desvio.

Após a classificação das disfluências típicas da gagueira e comuns, o software deve ter a capacidade de compilar os resultados para a aplicação dos protocolos de diagnóstico e severidade, permitindo ajustes do fonoaudiólogo em trechos da avaliação, pois este profissional terá acesso a mais elementos que auxiliam o processo de diagnóstico.

2.3 O que é Processamento de Linguagem Natural (PLN)?

O Processamento de Linguagem Natural (PLN) será a disciplina guia desse estudo, e pode ser caracterizado como um campo da inteligência artificial que busca a interação entre computadores e a linguagem humana. O PLN utiliza algoritmos e modelos matemáticos para permitir que máquinas compreendam, interpretem e gerem texto ou fala em linguagens naturais (Caseli; Nunes, 2024). Essa tecnologia é amplamente aplicada em diversas áreas, como tradução automática, assistentes virtuais e análise de sentimentos.

O PLN surgiu da necessidade de automatizar tarefas linguísticas e evoluiu com o avanço do aprendizado de máquina e das redes neurais. Atualmente, modelos como BERT e GPT demonstram um alto nível de compreensão textual, permitindo aplicações inovadoras em diversos setores (Caseli; Nunes, 2024).

2.3.1 Conceitos Fundamentais do PLN

Para compreender o PLN, é essencial conhecer seus principais conceitos e técnicas:

Tokenização: Processo que divide um texto em unidades menores, como palavras ou frases. Exemplo: A frase “O céu está azul” pode ser dividida em [“O”, “céu”, “está”, “azul”] (Caseli; Nunes, 2024).

Lematização e Stemização: Métodos para reduzir palavras à sua forma base. “Correndo” e “correu” podem ser reduzidos a “correr” na lematização. A stemização, por outro lado, pode reduzir “correr”, “correndo” e “corrida” para “corr” (Caseli; Nunes, 2024).

Reconhecimento de Entidades Nomeadas (NER): Identificação de entidades específicas, como nomes de pessoas, lugares e organizações. Exemplo: “Bill Gates fundou a Microsoft”. O sistema identifica “Bill Gates” como pessoa e “Microsoft” como organização (Caseli; Nunes, 2024).

Modelos baseados em aprendizado profundo: Redes neurais transformam textos em representações matemáticas, possibilitando análise de sentimentos, categorização automática e geração de texto coerente (Caseli; Nunes, 2024).

2.3.2 Uso do PLN no Desenvolvimento de um Software para Avaliação da Gagueira

O uso do PLN no desenvolvimento de um software para avaliação da gagueira pode trazer inovações significativas na identificação e classificação das disfluências. Algumas das principais aplicações incluem:

2.3.2.1 Reconhecimento Automático de Fala

Os modelos de PLN podem ser integrados a tecnologias de reconhecimento automático de fala (ASR — Automatic Speech Recognition) para transcrever a fala do paciente. Algoritmos modernos, como o Whisper da OpenAI, são capazes de reconhecer padrões linguísticos e analisar alterações na fluência, identificando pausas, repetições e prolongamentos típicos da gagueira (Caseli; Nunes, 2024).

2.3.2.2 Análise Sintática e Prosódica

A estrutura da fala pode ser analisada para detectar padrões de disfluência. A segmentação prosódica permite identificar alterações na entonação, intensidade e velocidade da fala, fornecendo informações detalhadas sobre as dificuldades do paciente (Caseli; Nunes, 2024).

Exemplo de aplicação: Um paciente com gagueira pode apresentar prolongamentos excessivos, como “ssssapato”. Um modelo de PLN pode identificar automaticamente esses padrões.

2.3.3 Geração de Relatórios Diagnósticos

Com o apoio de modelos de PLN, o software pode gerar relatórios detalhados que ajudam fonoaudiólogos a tomarem decisões informadas sobre o tratamento do paciente. Esses relatórios podem incluir:

- Percentual de palavras afetadas por disfluências.
- Comparação dos padrões de fala do paciente com bancos de dados de referência.

2.3.4 Exemplos de Softwares Existentes

Softwares como o Praat e ferramentas baseadas no Kaldi demonstram a viabilidade do uso de PLN na análise da fala. O desenvolvimento de um sistema específico para a gagueira pode se basear em abordagens híbridas, combinando modelos de PLN e análise acústica para maximizar a precisão do diagnóstico (Caseli; Nunes, 2024).

2.4 Desafios e Limitações do PLN na Avaliação da Gagueira

Apesar dos avanços, o uso do PLN na avaliação da gagueira apresenta desafios:

- Ambiguidade na identificação das disfluências: Alguns modelos podem interpretar pausas naturais como sinais de gagueira.
- Falta de grandes bases de dados anotadas: O treinamento de modelos requer grandes volumes de dados anotados manualmente por especialistas (Caseli; Nunes, 2024).
- Variações individuais: A gagueira se manifesta de forma única em cada indivíduo, tornando a padronização um desafio (Caseli; Nunes, 2024).

2.5 Considerações Finais

A utilização do PLN no desenvolvimento de um software para avaliação da gagueira representa um grande avanço na fonoaudiologia. Ferramentas de análise automática de fala e texto permitem uma abordagem mais objetiva e padronizada, contribuindo para diagnósticos mais precisos e intervenções mais eficazes. Entretanto, o sucesso dessas soluções depende da integração com dados clínicos e da validação com especialistas da área (Caseli; Nunes, 2024).

3 TRABALHOS RELACIONADOS

A detecção automática de disfluências da fala, especialmente da gagueira, tem sido um campo de crescente interesse na interseção entre a Fonoaudiologia e a Inteligência Artificial. A complexidade e a variabilidade das manifestações da gagueira representam um desafio significativo para a automação, mas os avanços em aprendizado de máquina e processamento de sinais de áudio têm aberto novas fronteiras. Esta seção apresenta uma revisão de estudos relevantes que fundamentam a abordagem adotada neste trabalho, desde as arquiteturas de redes neurais até as técnicas de extração de características acústicas.

3.1 Análise Detalhada dos Principais Estudos

3.1.1 *Computational Intelligence-Based Stuttering Detection: A Systematic Review* (Alnashwan et al., 2023)

O estudo de Alnashwan et al. (2023) (Alnashwan *et al.*, 2023) representa uma das revisões sistemáticas mais abrangentes e recentes sobre a aplicação de inteligência computacional na detecção de gagueira. Publicado na revista *Diagnostics*, este trabalho analisou 14 artigos científicos indexados na *Web of Science* entre 2019 e 2023, oferecendo uma visão panorâmica do estado da arte na área.

A metodologia empregada pelos autores seguiu rigorosamente os protocolos PRISMA para revisões sistemáticas. Inicialmente, foram identificados 384 artigos através de busca sistemática utilizando termos como "*stuttering detection using machine learning*", "*stuttering detection with AI*", e "*automatic stutter detection*". Após aplicação de critérios de inclusão e exclusão, incluindo relevância temática, qualidade metodológica e período de publicação, chegaram ao *corpus* final de 14 estudos.

Os principais achados desta revisão revelam uma diversidade significativa nas abordagens metodológicas. Quanto aos *datasets* utilizados, os autores identificaram que a maioria dos estudos empregou *datasets* públicos estabelecidos, sendo o UCLASS (*University College London's Archive of Stuttered Speech*) o mais frequentemente utilizado, seguido pelo SEP-28k, FluencyBank e LibriStutter. Interessantemente, vários estudos optaram por desenvolver *datasets* customizados, refletindo a necessidade específica de dados adequados às particularidades de cada pesquisa.

Em relação às técnicas de extração de características, o estudo revelou uma predominância de características espectrais e cepstrais, com os MFCCs (*Mel-Frequency Cepstral Coefficients*) sendo os mais amplamente utilizados. No entanto, outros estudos exploraram características prosódicas, de qualidade de voz e até mesmo características visuais derivadas de movimentos faciais. Esta diversidade sugere que não existe um consenso sobre qual

conjunto de características é o ideal para a detecção de gagueira, indicando a necessidade de mais pesquisas comparativas.

Quanto aos classificadores empregados, a revisão identificou uma clara tendência em direção às arquiteturas de aprendizado profundo. As Redes Neurais Artificiais (ANNs) foram as mais populares, seguidas por Máquinas de Vetores de Suporte (SVMs), Redes Neurais Convolucionais (CNNs) e Redes Neurais Recorrentes (RNNs), incluindo variantes como LSTM e GRU. Alguns estudos também exploraram métodos *ensemble*, combinando múltiplos classificadores para melhorar a performance.

Os resultados de performance variaram significativamente entre os estudos, com acurácias reportadas entre 85% e 98%. Os autores observaram que estudos com *datasets* maiores e mais diversificados tenderam a reportar performances mais modestas, mas potencialmente mais realistas, enquanto estudos com *datasets* menores e mais homogêneos alcançaram acurácias mais altas, mas com questionável generalização.

Uma contribuição importante desta revisão foi a identificação de lacunas e desafios na área. Os autores destacaram a falta de padronização nos métodos de avaliação, a escassez de *datasets* públicos de alta qualidade, e a necessidade de validação clínica dos sistemas propostos. Eles também enfatizaram a importância de considerar aspectos éticos e de privacidade no desenvolvimento de sistemas de detecção de gagueira, especialmente quando aplicados a populações vulneráveis como crianças.

3.1.2 *Machine Learning for Stuttering Identification: A Comprehensive Review* (Sheikh et al., 2021)

O trabalho de Sheikh et al. (2021) (Sheikh *et al.*, 2021) constitui uma das revisões mais citadas na área, com 83 citações até o momento, demonstrando seu impacto significativo na comunidade científica. Esta revisão abrangente examinou não apenas os aspectos técnicos da detecção automática de gagueira, mas também as características clínicas do distúrbio e sua relevância para o desenvolvimento de sistemas automatizados.

Os autores estruturaram sua revisão em várias dimensões críticas. Primeiro, eles forneceram uma análise detalhada das características da fala gaguejada, incluindo as manifestações primárias (repetições, prolongamentos, bloqueios) e secundárias (tensão muscular, movimentos associados). Esta fundamentação clínica é crucial para entender por que certas características acústicas são mais relevantes que outras para a detecção automática.

Em termos de *datasets*, Sheikh et al. identificaram uma limitação crítica na área: a maioria dos *datasets* disponíveis é relativamente pequena e não representa adequadamente a diversidade da população que gagueja. O UCLASS, embora amplamente utilizado, contém apenas algumas centenas de amostras, o que é insuficiente para treinar modelos de aprendizado profundo robustos. Os autores argumentaram pela necessidade de esforços

coordenados para desenvolver *datasets* maiores e mais representativos.

A revisão também forneceu uma análise técnica detalhada das diferentes abordagens de extração de características. Os autores compararam sistematicamente o desempenho de MFCCs versus LPCCs (*Linear Predictive Cepstral Coefficients*), concluindo que, embora os MFCCs sejam mais populares, os LPCCs podem ser superiores para certas tarefas específicas de detecção de disfluência, particularmente na identificação de repetições e prolongamentos.

Uma contribuição única desta revisão foi a análise das diferentes estratégias de segmentação temporal. Os autores observaram que a maioria dos estudos utiliza janelas fixas (tipicamente 20-30ms para análise espectral), mas argumentaram que janelas adaptativas, que se ajustam à duração natural dos fonemas, podem ser mais eficazes para capturar as características das disfluências.

Sheikh et al. também abordaram a questão da avaliação de performance, criticando a falta de padronização nas métricas utilizadas. Eles propuseram um conjunto de métricas padronizadas que incluem não apenas acurácia global, mas também precisão, *recall* e F1-*score* específicos para cada tipo de disfluência, além de métricas de robustez que avaliam a performance em diferentes condições acústicas.

3.1.3 FluentNet: *End-to-End Detection of Speech Disfluency* (Kourkounakis et al., 2020)

O trabalho de Kourkounakis et al. (2020) (Kourkounakis *et al.*, 2020) representa um marco na aplicação de arquiteturas de aprendizado profundo para detecção de disfluências. O FluentNet, como foi denominado o sistema proposto, é uma rede neural profunda *end-to-end* capaz de detectar múltiplos tipos de disfluências diretamente do sinal de áudio bruto, sem necessidade de extração manual de características.

A arquitetura do FluentNet baseia-se em uma combinação de camadas convolucionais 1D e camadas recorrentes LSTM. As camadas convolucionais são responsáveis por extrair características locais do sinal de áudio, enquanto as camadas LSTM capturam dependências temporais de longo prazo, que são cruciais para identificar padrões de disfluência que podem se estender por vários segundos.

Uma inovação importante do FluentNet é sua capacidade de realizar detecção multi-classe em tempo real. Ao contrário de muitos sistemas anteriores que se limitavam à classificação binária (fluente vs. disfluente), o FluentNet pode identificar tipos específicos de disfluência: repetições de sons, repetições de sílabas, repetições de palavras, prolongamentos e bloqueios. Esta capacidade é clinicamente relevante, pois diferentes tipos de disfluência podem requerer diferentes estratégias terapêuticas.

Os autores treinaram o FluentNet utilizando uma combinação de *datasets* públicos e dados coletados especificamente para o estudo. O *dataset* de treinamento incluiu mais

de 50 horas de fala anotada, representando uma das maiores coleções utilizadas na área até aquele momento. O processo de anotação foi realizado por fonoaudiólogos experientes, garantindo alta qualidade das *labels*.

Os resultados experimentais demonstraram a superioridade do FluentNet em relação a métodos tradicionais. Em testes com o *dataset* UCLASS, o sistema alcançou uma acurácia média de 94,2% na detecção de disfluências, com F1-*scores* específicos de 91,3% para repetições, 89,7% para prolongamentos e 87,4% para bloqueios. Estes resultados representaram uma melhoria significativa em relação aos métodos baseados em características manuais.

Uma análise particularmente interessante do estudo foi a investigação das características aprendidas automaticamente pela rede. Utilizando técnicas de visualização como mapas de ativação de classe (CAMs), os autores demonstraram que o FluentNet aprendeu a focar em regiões do espectrograma que correspondem a características acústicas conhecidas das disfluências, como variações abruptas na energia e perturbações na estrutura harmônica.

3.1.4 *Disfluency Detection using Auto-Correlational Neural Networks* (Lou et al., 2018)

O trabalho de Lou et al. (2018) (Lou *et al.*, 2018) introduziu uma abordagem inovadora para detecção de disfluências baseada em Redes Neurais Auto-Correlacionais (ACNNs). Embora focado na análise de transcrições textuais ao invés de sinais de áudio, este estudo fornece *insights* valiosos sobre os padrões linguísticos das disfluências que podem ser adaptados para análise acústica.

A motivação principal dos autores foi a observação de que as disfluências, particularmente as repetições, exibem padrões de auto-correlação distintos no texto. Por exemplo, em uma repetição como “eu eu eu quero”, existe uma correlação óbvia entre as palavras repetidas. As ACNNs foram projetadas para capturar essas correlações de forma mais eficiente que as redes neurais tradicionais.

A arquitetura ACNN incorpora uma camada de auto-correlação que computa correlações entre diferentes posições na sequência de entrada. Esta camada é seguida por camadas convolucionais que processam os mapas de correlação para identificar padrões característicos de disfluência. A vantagem desta abordagem é que ela pode detectar repetições e padrões mesmo quando eles estão separados por outras palavras ou quando ocorrem em diferentes escalas temporais.

Os autores avaliaram sua abordagem utilizando o *corpus* Switchboard, um grande *dataset* de conversas telefônicas espontâneas com anotações de disfluência. Os resultados demonstraram que as ACNNs superaram significativamente os métodos *baseline*, incluindo CRFs (*Conditional Random Fields*) e LSTMs tradicionais, alcançando um F1-*score* de

85,4% na detecção de disfluências.

Uma contribuição importante deste trabalho foi a análise detalhada dos diferentes tipos de padrões de correlação associados a diferentes tipos de disfluência. Os autores mostraram que repetições de palavras exibem correlações de curto alcance, enquanto repetições de frases podem exibir correlações de longo alcance. Esta análise fornece *insights* valiosos para o design de sistemas de detecção de disfluência baseados em áudio.

3.1.5 *Self-organizing Map for Classification of Normal and Pathological* (Callan et al., 1999)

Embora não focado especificamente em gagueira, o trabalho seminal de Callan et al. (1999) (Callan *et al.*, 1999) é fundamental para compreender a aplicação de Mapas Auto-Organizáveis (SOMs) na análise de distúrbios da fala. Este estudo demonstrou pela primeira vez a eficácia dos SOMs para classificação de vozes normais e patológicas, estabelecendo uma base metodológica que influenciou trabalhos posteriores, incluindo nossa própria abordagem.

Os autores utilizaram um SOM de 10x10 neurônios para analisar 13 medidas acústicas extraídas de amostras de voz, incluindo medidas de perturbação (*jitter* e *shimmer*), medidas de ruído (relação sinal-ruído, relação harmônico-ruído) e medidas espectrais (energia em diferentes bandas de frequência). O *dataset* incluiu 53 vozes normais e 58 vozes com diferentes tipos de patologia vocal.

A principal inovação metodológica foi o uso da topologia do SOM para visualizar e interpretar os padrões de classificação. Os autores mostraram que vozes com características similares tendiam a ativar neurônios próximos no mapa, criando regiões topologicamente organizadas que correspondiam a diferentes tipos de patologia vocal. Esta propriedade de preservação topológica é uma das principais vantagens dos SOMs em relação a outros métodos de classificação.

Os resultados demonstraram uma acurácia de classificação de 91,2% entre vozes normais e patológicas, com a capacidade adicional de identificar subtipos específicos de patologia. Mais importante, o estudo estabeleceu um protocolo metodológico para o uso de SOMs em análise de voz que tem sido amplamente seguido em trabalhos posteriores.

3.1.6 StutterAI: *Virtual Assistant for Stutter Detection* (Arachchi et al., 2023)

O trabalho de Arachchi et al. (2023) (Arachchi *et al.*, 2023) representa uma abordagem mais aplicada, focando no desenvolvimento de um sistema completo de assistência virtual para detecção e tratamento de gagueira. O StutterAI é uma aplicação web que integra análise de voz, processamento de linguagem natural e aprendizado de máquina para fornecer uma solução *end-to-end*.

A arquitetura do sistema é modular, consistindo em vários componentes integrados: um módulo de captura e pré-processamento de áudio, um módulo de extração de características baseado em Librosa, um módulo de classificação utilizando uma rede neural híbrida (CNN + LSTM), e um módulo de interface de usuário desenvolvido em React.js.

Uma característica única do StutterAI é sua capacidade de fornecer *feedback* em tempo real e sugestões de tratamento personalizadas. O sistema não apenas detecta disfluências, mas também analisa padrões ao longo do tempo e fornece exercícios terapêuticos adaptados ao perfil específico de cada usuário. Esta abordagem holística representa uma evolução importante na direção de sistemas clinicamente úteis.

Os autores avaliaram o sistema utilizando um *dataset* de 200 participantes e reportaram uma acurácia de 89,7% na detecção de disfluências. Mais importante, eles conduziram um estudo de usabilidade com fonoaudiólogos, que avaliaram positivamente a utilidade clínica do sistema, embora tenham sugerido melhorias na interface e na interpretação dos resultados.

3.1.7 *Systematic Review of Machine Learning Approaches for Detecting Developmental Stuttering* (Barrett et al., 2022)

A revisão de Barrett et al. (2022) (Barrett *et al.*, 2022) focou especificamente em métodos de aprendizado de máquina para detecção de gagueira do desenvolvimento, fornecendo uma perspectiva complementar à revisão mais geral de Alnashwan et al. Os autores analisaram 23 estudos publicados entre 2015 e 2021, com foco particular em aspectos metodológicos e de validação.

Uma contribuição importante desta revisão foi a análise crítica dos métodos de validação utilizados nos estudos. Os autores observaram que muitos estudos utilizavam validação cruzada simples ou divisões treino/teste inadequadas, potencialmente levando a estimativas otimistas de performance. Eles propuseram diretrizes para validação mais rigorosa, incluindo validação cruzada estratificada por falante e validação em *datasets* completamente independentes.

A revisão também abordou a questão da generalização entre diferentes populações. Os autores notaram que a maioria dos estudos focava em adultos falantes de inglês, com representação limitada de crianças e falantes de outras línguas. Esta limitação é particularmente problemática para gagueira do desenvolvimento, que se manifesta primariamente na infância e pode apresentar características diferentes em diferentes línguas.

3.1.8 *Evaluative Comparison of Machine Learning Algorithms for Stuttering Detection* (Ramitha et al., 2024)

O estudo mais recente de Ramitha et al. (2024) (Ramitha *et al.*, 2024) fornece uma comparação sistemática de diferentes algoritmos de aprendizado de máquina para detecção de eventos de gagueira. Este trabalho é particularmente relevante pois utiliza uma metodologia rigorosa de comparação, controlando por *dataset*, características e métricas de avaliação.

Os autores compararam seis algoritmos diferentes: SVM, *Random Forest*, *Gradient Boosting*, Redes Neurais Artificiais, CNN e LSTM. Todos os algoritmos foram treinados no mesmo *dataset* e avaliados utilizando as mesmas métricas, fornecendo uma comparação justa de performance.

Os resultados mostraram que, embora as redes neurais profundas (CNN e LSTM) alcançassem a maior acurácia (95,2% e 94,8%, respectivamente), métodos mais simples como *Random Forest* (92,1%) ofereciam uma relação custo-benefício melhor, sendo mais rápidos para treinar e mais interpretáveis. Esta observação é importante para aplicações práticas onde a interpretabilidade e a eficiência computacional são considerações importantes.

3.1.9 *Automated Stuttering Detection Using Deep Learning Technologies* (Alhakbani et al., 2025)

O trabalho mais recente de Alhakbani et al. (2025) (Alhakbani *et al.*, 2025) representa o estado da arte atual em detecção automática de gagueira utilizando tecnologias de aprendizado profundo. Este estudo é particularmente relevante pois aborda muitas das limitações identificadas em trabalhos anteriores.

Os autores desenvolveram um sistema integrado que combina múltiplas modalidades de dados: áudio, texto (transcrições) e, inovadoramente, dados visuais derivados de análise de movimentos faciais. Esta abordagem multimodal permite capturar não apenas as características acústicas das disfluências, mas também os comportamentos secundários associados à gagueira.

A arquitetura do sistema utiliza uma rede neural de atenção multi-modal que aprende a ponderar automaticamente a contribuição de cada modalidade para a decisão final. Esta abordagem é mais robusta que sistemas uni-modais, especialmente em condições de ruído ou quando certas modalidades não estão disponíveis.

Os resultados experimentais demonstraram acurácias superiores a 96% em múltiplos *datasets*, representando o melhor desempenho reportado na literatura até o momento. Mais importante, os autores conduziram uma validação clínica extensiva com fonoaudiólogos, demonstrando que o sistema pode ser uma ferramenta útil na prática clínica.

3.1.10 Análise Comparativa e Síntese

A análise dos estudos revisados revela várias tendências importantes na evolução da detecção automática de gagueira. Primeiro, há uma clara evolução das abordagens baseadas em características manuais para métodos de aprendizado profundo *end-to-end*. Esta transição reflete tendências mais amplas na área de aprendizado de máquina, mas também apresenta desafios específicos para a detecção de gagueira, particularmente em termos de interpretabilidade e necessidade de dados.

Segundo, observa-se uma crescente sofisticação nas arquiteturas de rede neural utilizadas. Enquanto estudos iniciais utilizavam redes simples como MLPs, trabalhos mais recentes exploram arquiteturas complexas incluindo CNNs, LSTMs, redes de atenção e abordagens multimodais. Esta sofisticação, no entanto, vem com o custo de maior complexidade computacional e necessidade de *datasets* maiores.

Terceiro, há uma tendência crescente em direção à validação clínica e aplicações práticas. Estudos mais recentes não se limitam a reportar acurácias em *datasets* de pesquisa, mas também avaliam a utilidade clínica dos sistemas propostos através de estudos com fonoaudiólogos e pacientes.

Finalmente, a questão dos *datasets* permanece um desafio central na área. Embora tenham sido feitos progressos no desenvolvimento de *datasets* maiores e mais diversos, a necessidade de dados anotados por especialistas continua sendo um gargalo significativo para o avanço da área.

3.2 Arquiteturas de Redes Neurais para Detecção de Disfluências

A escolha da arquitetura de rede neural é um dos pilares para o sucesso de um sistema de detecção de disfluências. A literatura apresenta uma variedade de abordagens, desde modelos mais tradicionais até arquiteturas de aprendizado profundo (*deep learning*) mais recentes.

3.2.1 A Abordagem Hierárquica com Redes de Kohonen e MLP

Um dos trabalhos seminais que inspiraram a arquitetura inicial deste projeto é o de Szczurowska et al. (2006) (Szczurowska *et al.*, 2006), que propôs o uso de uma abordagem hierárquica combinando uma Rede de Kohonen (*Self-Organizing Map* - SOM) com um Perceptron de Múltiplas Camadas (MLP) para a análise de disfluências da fala. A principal inovação desta abordagem reside na utilização da SOM, uma rede neural não supervisionada, para realizar uma redução de dimensionalidade e extrair padrões topológicos do sinal de fala. Os vetores de neurônios vencedores da SOM, que representam a trajetória do sinal de fala no mapa auto-organizável, são então utilizados como entrada para a rede MLP, que realiza a classificação final entre fala fluente e não fluente.

A Rede de Kohonen, com 21 entradas e 25 neurônios na camada de saída, foi utilizada para reduzir a dimensão dos sinais de entrada. [...] Como resultado da análise, obtivemos vetores que descrevem a trajetória dos neurônios vencedores em um determinado ponto no tempo. Esses vetores foram então considerados como entrada para a próxima rede, que era um Perceptron de Múltiplas Camadas. (Szczurowska et al., 2006, p. 205).

Este método demonstrou a viabilidade de capturar a dinâmica temporal da fala através da trajetória na SOM, uma ideia que foi posteriormente refinada e validada em outros estudos. A eficácia desta arquitetura para a tarefa específica de identificação da gagueira foi mais aprofundada por Świetlicka et al. (2013) (Świetlicka *et al.*, 2013), que detalharam um sistema hierárquico de redes neurais artificiais (ANNs) para este fim, alcançando resultados promissores e servindo como a principal referência para a implementação do modelo deste TCC.

3.2.2 Modelos de Aprendizado Profundo (*Deep Learning*)

Com a evolução do poder computacional, as arquiteturas de aprendizado profundo tornaram-se o padrão na análise de sinais temporais como a fala. Kourkounakis et al. (2020) (Kourkounakis *et al.*, 2020), em seu trabalho "FluentNet", propuseram uma rede neural profunda do tipo *end-to-end* capaz de detectar múltiplos tipos de disfluências diretamente do áudio. A vantagem de uma abordagem *end-to-end* é que ela aprende as características relevantes diretamente dos dados brutos, eliminando a necessidade de uma etapa manual de extração de características, que é sempre um ponto crítico e sujeito a vieses.

Outra abordagem inovadora é a das Redes Neurais Auto Correlacionais (ACNN), proposta por Lou et al. (2018) (Lou *et al.*, 2018). Este modelo foi projetado para detectar disfluências em transcrições de fala espontânea, focando nos padrões de repetição e correlação que as palavras e fonemas apresentam em um contexto de disfluência. Embora focada em texto, a lógica de autocorrelação é um conceito poderoso que pode ser adaptado para a análise do sinal de áudio.

Uma revisão sistemática recente de Alnashwan et al. (2023) (Alnashwan *et al.*, 2023) analisou 14 estudos publicados entre 2019 e 2023 e confirmou a tendência de utilização de modelos de aprendizado de máquina, incluindo Redes Neurais Artificiais (ANN), Máquinas de Vetores de Suporte (SVM), Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN), como as LSTMs (*Long Short-Term Memory*). O estudo destaca que diferentes modelos alcançaram acurácias superiores a 94% em *datasets* específicos, evidenciando o alto potencial da IA nesta área.

3.3 Extração de Características Acústicas

A performance de qualquer modelo de aprendizado de máquina é intrinsecamente dependente da qualidade das características (*features*) extraídas do sinal de entrada. Para

a detecção de gagueira, as características precisam capturar as nuances que diferenciam a fala fluente da disfluente.

3.3.1 Características Espectrais e Cepstrais

As características mais comuns na análise de fala são as espectrais e cepstrais. Os Coeficientes Cepstrais de Frequência Mel (MFCCs) são quase onipresentes em aplicações de reconhecimento de fala. Eles representam o espectro de potência de um sinal em uma escala de frequência não-linear (a escala Mel), que se aproxima da percepção auditiva humana. No entanto, como apontado por Sheikh et al. (2021) (Sheikh *et al.*, 2021) em sua revisão, embora os MFCCs sejam poderosos, eles podem não ser ótimos para a detecção de disfluências, que muitas vezes reside em padrões temporais e de energia que os MFCCs, por si só, não capturam completamente. O estudo compara o desempenho dos MFCCs com os Coeficientes de Predição Linear Cepstrais (LPCCs), relatando que sistemas baseados em LPCCs podem superar os baseados em MFCCs na detecção de repetições e prolongamentos.

O trabalho de Szczurowska et al. (2006) (Szczurowska *et al.*, 2006), que fundamenta nosso projeto, utilizou uma abordagem diferente, baseada em um banco de 21 filtros de 1/3 de oitava, com uma ponderação A (*A-weighting*) para simular a sensibilidade do ouvido humano a diferentes frequências. Esta técnica, embora mais antiga, foca na distribuição de energia em bandas de frequência específicas, o que pode ser particularmente eficaz para capturar as alterações de fonação presentes em bloqueios e prolongamentos.

3.3.2 Características Prosódicas e de Qualidade de Voz

Além das características espectrais, a literatura aponta para a importância de características prosódicas e de qualidade de voz. Callan et al. (1999) (Callan *et al.*, 1999), em um estudo pioneiro, utilizaram um Mapa Auto-Organizável (SOM) para classificar vozes normais e patológicas com base em medidas acústicas como o quociente de perturbação da amplitude (*amplitude perturbation quotient* - APQ, ou *shimmer*) e o grau de quebras de voz. Embora não focado especificamente em gagueira, o estudo demonstrou a eficácia de características de qualidade de voz para a classificação de desordens da fonação.

Um estudo mais recente de Ramitha et al. (2024) (Ramitha *et al.*, 2024) reforça a necessidade de um conjunto diversificado de características. Eles compararam vários algoritmos de aprendizado de máquina e concluíram que a combinação de diferentes tipos de características (espectrais, cepstrais, prosódicas) geralmente leva a um melhor desempenho. A detecção de gagueira não depende de uma única "bala de prata" acústica, mas sim da combinação de múltiplos indicadores que, juntos, formam um padrão reconhecível pelo modelo.

3.4 *Datasets* para Pesquisa em Gagueira

Um dos maiores gargalos para o avanço da área, como apontado por Alnashwan et al. (2023) (Alnashwan *et al.*, 2023) e Sheikh et al. (2021) (Sheikh *et al.*, 2021), é a disponibilidade de *datasets* públicos, de alta qualidade e com anotações precisas. A maioria dos estudos utiliza um de alguns *datasets* bem conhecidos:

- **UCLASS** (*University College London's Archive of Stuttered Speech*): Um dos mais antigos e amplamente utilizados, contendo uma variedade de amostras de fala de indivíduos com gagueira.
- **FluencyBank**: Um banco de dados compartilhado sob os auspícios do sistema CHILDES, focado em pesquisa sobre o desenvolvimento da linguagem e seus distúrbios.
- **LibriStutter**: Um *dataset* mais recente, derivado do popular LibriSpeech, contendo amostras de fala com disfluências sinteticamente inseridas ou naturalmente ocorrendo.

O desenvolvimento de *datasets* customizados, embora trabalhoso, é uma prática comum e necessária para atender às necessidades específicas de cada pesquisa, como foi o caso deste projeto, onde coletamos amostras específicas de bloqueios, prolongamentos e suas contrapartes fluentes.

3.5 Considerações Finais

A revisão da literatura demonstra que a detecção automática da gagueira é um campo ativo e promissor. A arquitetura hierárquica de Redes de Kohonen e MLP, embora uma abordagem mais clássica, continua sendo uma metodologia válida e eficaz, como comprovado pelo artigo de Szczurowska et al. (2006) (Szczurowska *et al.*, 2006). O sucesso desta abordagem, no entanto, depende crucialmente da implementação correta e, principalmente, da extração de características acústicas que sejam sensíveis aos fenômenos da disfluência, como os filtros de 1/3 de oitava propostos no artigo.

Ao mesmo tempo, a literatura mais recente aponta para o potencial de arquiteturas de aprendizado profundo, como CNNs e LSTMs, e para a importância de explorar um conjunto mais rico de características, incluindo aspectos prosódicos e de qualidade de voz. O desafio fundamental, no entanto, permanece o mesmo: a necessidade de *datasets* grandes, diversos e bem anotados para treinar e validar esses modelos de forma robusta. O trabalho desenvolvido neste TCC, portanto, se insere neste contexto, aplicando uma metodologia comprovada e enfrentando os desafios práticos de sua implementação e calibração para dados do mundo real.

4 METODOLOGIA

A metodologia adotada neste trabalho foi estruturada de forma iterativa e incremental, seguindo os princípios da engenharia de software ágil e da pesquisa experimental em inteligência artificial. O desenvolvimento foi dividido em múltiplas fases, cada uma com objetivos específicos e critérios de validação bem definidos. Esta abordagem permitiu a identificação precoce de problemas, a adaptação da estratégia conforme necessário e a documentação detalhada de cada etapa do processo.

O projeto foi concebido como uma pesquisa aplicada, combinando elementos de desenvolvimento de software, processamento de sinais digitais, aprendizado de máquina e validação experimental. A natureza interdisciplinar do trabalho exigiu uma metodologia híbrida que pudesse acomodar tanto os aspectos técnicos quanto os requisitos clínicos da aplicação.

4.1 Visão Geral da Arquitetura do Sistema

O sistema desenvolvido foi projetado seguindo uma arquitetura modular e escalável, composta por 6 componentes principais:

1. o módulo principal de processamento das requisições da API;
2. módulo de processamento de áudio;
3. módulo de transcrição;
4. módulo de detecção de disfluências;
5. módulo de inteligência artificial para classificação de disfluências complexas;
6. a interface de usuário.

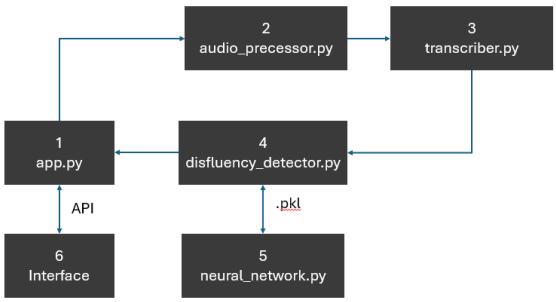

Figura 1 – Arquitetura do Sistema

A arquitetura foi concebida para ser flexível o suficiente para acomodar diferentes abordagens de detecção (tradicional e baseada em IA) e para permitir a evolução incremental do sistema. O princípio de separação de responsabilidades foi rigorosamente aplicado, garantindo que cada módulo tivesse uma função bem definida e interfaces claras com os demais componentes.

O *backend* foi implementado em Python, utilizando o *microframework* Flask para a criação da API REST. Esta escolha foi motivada pelo rico ecossistema de bibliotecas Python para processamento de áudio (Librosa, SciPy) e aprendizado de máquina, bem como pela facilidade de prototipagem e desenvolvimento iterativo que o Flask oferece.

O *frontend* foi implementado em Bubble, onde se comunica com o *backend* por meio de APIs.

4.2 Coleta e Preparação dos Dados

A criação de um *dataset* de qualidade foi identificada como um dos fatores mais críticos para o sucesso do projeto. A literatura científica na área de detecção automática de disfluências enfatiza consistentemente que a qualidade e a representatividade dos dados de treinamento são determinantes para a performance dos modelos de aprendizado de máquina. Neste trabalho, foi necessária a criação de um *dataset* para as disfluências mais complexas (prolongamentos e bloqueios), visto que as demais disfluências podem ser tratadas com análise de dados advindos da transcrição do áudio.

4.2.1 Estratégia de Coleta

A estratégia de coleta foi cuidadosamente planejada para maximizar a qualidade das amostras dentro das limitações de recursos disponíveis. Optou-se por uma abordagem de "pares controlados", onde cada palavra com disfluência teria uma contraparte fluente correspondente. Esta estratégia foi inspirada nos protocolos de pesquisa em fonoaudiologia, onde a comparação entre produções fluentes e disfluentes da mesma palavra é uma prática padrão para análise acústica.

O processo de coleta envolveu as seguintes etapas:

1. **Seleção das disfluências Palavras-Alvo**: Foram selecionadas palavras de diferentes complexidades fonéticas, incluindo palavras com diferentes números de sílabas, diferentes padrões acentuais e diferentes grupos consonantais. Esta diversidade foi importante para garantir que o modelo pudesse generalizar para diferentes contextos fonéticos.

2. **Gravação de Amostras Disfluentes**: Foram coletadas 45 amostras de bloqueios e 18 amostras de prolongamentos a partir de gravações de fala espontânea e dirigida.

Cada amostra foi cuidadosamente segmentada para conter apenas o evento de disfluência, com margens mínimas de silêncio antes e depois do evento.

3. **Gravação de Amostras Controle**: Para cada palavra que apresentou disfluência, foi gravada uma versão fluente da mesma palavra, totalizando 63 amostras de controle (45 + 18). Esta abordagem de pares controlados é fundamental para que o modelo aprenda a distinguir as características acústicas específicas da disfluência, e não simplesmente características lexicais ou fonéticas da palavra.

Todas as amostras coletadas foram submetidas à validação por fonoaudiólogos especialistas em fluência.

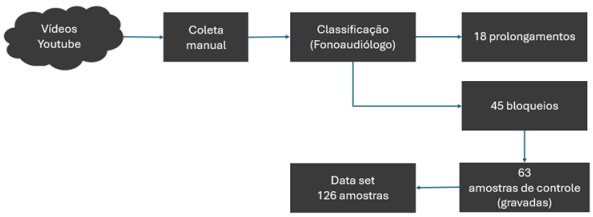

Figura 2 – Fluxograma de criação do *dataset*

4.2.2 Organização e Estruturação dos Dados

Os dados foram organizados em uma estrutura hierárquica de diretórios, facilitando o carregamento automático durante o treinamento:

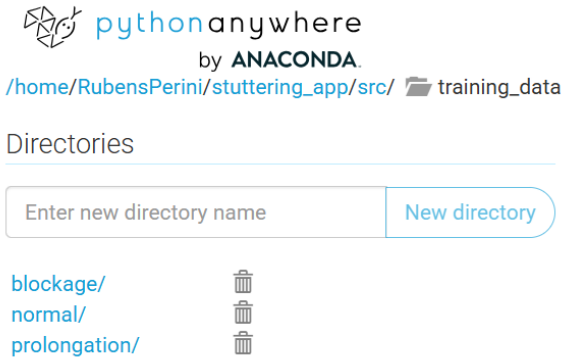

Figura 3 – Diretório de amostras

Cada arquivo de áudio foi padronizado para uma taxa de amostragem de 16 kHz e formato WAV, garantindo consistência no processamento posterior.

4.2.3 Preparação de Dados para Metodologia Híbrida

A natureza híbrida do sistema exigiu preparação de dados específica para cada metodologia.

Dados para validação do sistema de detecção baseadas em áudio (IA):

- **126 amostras de áudio segmentadas (bloqueios, prolongamentos e normais):** Bloqueios e prolongamentos são eventos temporalmente bem definidos, com início e fim claramente identificáveis por especialistas. A segmentação permite que cada amostra contenha exclusivamente o evento de interesse, eliminando ruídos contextuais que poderiam confundir o modelo de aprendizado de máquina.
- **Anotação temporal precisa de eventos:** O sistema híbrido necessita correlacionar detecções baseadas em áudio (IA) com detecções baseadas em transcrição (métodos tradicionais). *Timestamps* precisos permitem mapear eventos detectados pelo módulo de IA para palavras específicas na transcrição, habilitando a integração harmoniosa das duas metodologias.
- **Normalização acústica e padronização de formato:** Algoritmos de aprendizado de máquina requerem entrada consistente. Variações na taxa de amostragem, formato de arquivo ou níveis de amplitude podem ser interpretadas pelo modelo como características relevantes, quando na verdade são artefatos técnicos. A padronização para 16 kHz WAV elimina estas variáveis confundidoras.

Dados para validação de métodos tradicionais (disfluências advindas da transcrição):

- **Transcrições manuais de vídeos completos:** Sistemas automáticos de reconhecimento de fala (ASR) apresentam performance degradada em fala com disfluências, frequentemente omitindo ou distorcendo eventos como repetições e bloqueios. Transcrições manuais garantem que todos os eventos disfluentes sejam capturados com fidelidade, permitindo validação precisa dos algoritmos de detecção.
- **Anotação de repetições, pausas e hesitações:** Enquanto o módulo de IA foca em bloqueios e prolongamentos (disfluências com características acústicas distintivas), os métodos tradicionais cobrem repetições, pausas e hesitações (disfluências melhor identificáveis por análise linguística e temporal). Esta divisão garante cobertura completa dos tipos de disfluência mais relevantes clinicamente.
- **Sincronização temporal entre áudio e transcrição:** Todas as detecções utilizam *timestamps* baseados no áudio. A sincronização permite mapear detecções de palavras

específicas, habilitando análises como "o bloqueio detectado em 15,3 s corresponde à palavra 'problema' na transcrição".

4.3 Desenvolvimento da Arquitetura de Software

O desenvolvimento da arquitetura de software seguiu os princípios de engenharia de software moderna, com ênfase na modularidade, testabilidade e manutenibilidade. A arquitetura foi evoluindo ao longo do projeto, adaptando-se às necessidades emergentes e aos aprendizados obtidos durante o desenvolvimento.

4.3.1 *Backend* em Flask

O *backend* foi implementado utilizando o *microframework* Flask, que oferece a flexibilidade necessária para prototipagem rápida sem sacrificar a capacidade de escalabilidade futura. A escolha do Flask foi motivada por vários fatores:

- **Simplicidade:** O Flask permite o desenvolvimento rápido de APIs REST com código mínimo e estrutura clara.
- **Flexibilidade:** Diferentemente de *frameworks* mais pesados, o Flask não impõe uma estrutura rígida, permitindo adaptações conforme as necessidades do projeto.
- **Ecossistema Python:** A integração natural com bibliotecas científicas Python (NumPy, SciPy, scikit-learn, Librosa) foi crucial para o desenvolvimento.

4.3.2 Sistema Híbrido de Detecção de Disfluências

O sistema desenvolvido implementa uma abordagem híbrida que combina duas metodologias complementares para detecção abrangente de disfluências, fundamentada nos trabalhos de Szczurowska, Kuniszyk-Jóźkowiak e Smołka (Szczurowska *et al.*, 2006) e Świetlicka et al. (Świetlicka *et al.*, 2013).

4.3.2.1 Metodologia Baseada em Inteligência Artificial (Bloqueios e Prolongamentos)

Características:

- Utiliza a arquitetura hierárquica Kohonen + MLP
- Análise de características acústicas espectrais (21 filtros de 1/3 de oitava)
- Foco em disfluências que apresentam padrões acústicos distintivos

Justificativa para Arquitetura Kohonen + MLP:

A escolha desta arquitetura foi baseada no trabalho de Szczurowska, Kuniszyk-Jóźkowiak e Smołka (Szczurowska *et al.*, 2006), que demonstraram sua eficácia específica para detecção de disfluências. A Rede de Kohonen permite organização topológica dos padrões acústicos, onde "neurônios vizinhos respondem a padrões acústicos similares, criando um mapa auto-organizado das características da fala" (SZCZUROWSKA; KUNISZYK-JÓŹKOWIAK; SMOŁKA, 2006, p. 206). O MLP subsequente realiza a classificação supervisionada dos padrões organizados.

Justificativa para Características Espectrais:

A implementação de 21 filtros de 1/3 de oitava com ponderação A seguiu rigorosamente a metodologia validada por Szczurowska, Kuniszyk-Jóźkowiak e Smołka (Szczurowska *et al.*, 2006). Esta abordagem é fundamentada nos seguintes princípios:

- **Correspondência com Percepção Humana:** A ponderação A simula a sensibilidade do ouvido humano, garantindo consistência com diagnósticos clínicos.
- **Cobertura Espectral Adequada:** Faixa de 80 Hz a 8 kHz cobre harmônicos fundamentais onde bloqueios e prolongamentos apresentam características distintivas.
- **Resolução Temporal Otimizada:** 171 *frames* em janelas de 4 s (23 ms) captura tanto características instantâneas quanto evolução temporal.

Justificativa para Foco em Bloqueios e Prolongamentos:

A aplicação de IA especificamente para estes tipos foi baseada em Świetlicka et al. (Świetlicka *et al.*, 2013), que documentaram suas características acústicas distintivas:

- **Bloqueios:** Interrupção abrupta do fluxo de ar com energia concentrada em baixas frequências e tensão articulatória.
- **Prolongamentos:** Extensão temporal de fonemas com padrões espectrais estáveis e modificações de *pitch*.

4.3.2.1.1 Metodologia Tradicional Baseada em Transcrição (Demais Disfluências):

- Análise automática de transcrições de fala.
- Detecção de repetições através de processamento de linguagem natural.
- Identificação de pausas e hesitações por análise temporal.
- Contagem automática de eventos disfluentes.

Esta abordagem híbrida foi adotada porque diferentes tipos de disfluências apresentam características distintivas em diferentes domínios de análise. Enquanto bloqueios e prolongamentos são mais facilmente identificáveis através de suas assinaturas acústicas, repetições de sílabas e palavras são melhor detectadas através da análise do conteúdo linguístico da transcrição.

4.3.3 API REST para integração *Backend*/*Frontend* e Gerenciamento do Modelo

Foi desenvolvida uma API REST completa para gerenciar todo o processo de transcrição, detecção das disfluências, organização das métricas, e todo o ciclo de vida do modelo de inteligência artificial. Esta API foi projetada para ser intuitiva e robusta, permitindo que usuários não-técnicos possam treinar e utilizar o modelo através de uma interface web.

4.3.3.1 Endpoints de Status e Monitoramento

- `/` (`GET`): **Endpoint raiz para *health checks***
  Descrição: Rota ultra-rápida que responde imediatamente sem depender de componentes pesados de ML/AI, essencial para monitoramento do PythonAnywhere.

- `/health` (`GET`): **Verificação de saúde da API**
  Descrição: Endpoint simples para verificar se a API está funcionando corretamente.

- `/status` (`GET`): **Status detalhado dos componentes**
  Descrição: Fornece informações completas sobre o estado de todos os módulos carregados.

- `/test` (`POST`): **Verifica se todos os componentes estão funcionando.**

4.3.3.2 Endpoints de Processamento e acompanhamento de *Jobs*

- `/transcribe_url` (`POST`): **Endpoint principal que recebe URL de arquivo, realiza transcrição automática, detecta todas as disfluências usando o sistema híbrido e retorna análise completa.** Implementa processamento assíncrono para evitar *timeouts*.

- `/job_status/<job_id>` (`GET`): Permite acompanhar o progresso de um *job* específico através de seu ID único.

- `/job_status` (`GET`): Endpoint alternativo para verificar *status* usando *query parameter*, útil para compatibilidade com diferentes plataformas *frontend*.

4.3.3.3 Endpoints de Gerenciamento do Modelo de IA

- `/start_training` (`POST`): Executa treinamento assíncrono do sistema hierárquico Kohonen + MLP usando amostras organizadas em pastas específicas (`blockage/`, `prolongation/`, `normal/`). Implementa verificações de segurança e validação de dados.

- `/training_status` (`GET`): Retorna *status* detalhado do processo de treinamento em tempo real, incluindo progresso percentual, etapa atual e métricas de performance.

4.3.4 Integração com Interface *Low-Code*

A interface do usuário foi desenvolvida utilizando a plataforma Bubble, uma ferramenta de desenvolvimento *low-code* que permite a criação rápida de interfaces web responsivas.

A integração entre o Bubble e o *backend* Flask foi realizada através do *API Connector* do Bubble, que foi configurado para consumir os *endpoints* REST desenvolvidos. Esta configuração envolveu:

- **Configuração de *Endpoints*:** Cada *endpoint* foi configurado no *API Connector* com os parâmetros corretos, tipos de dados e *headers* necessários.

- **Mapeamento de Dados:** Os dados retornados pela API foram mapeados para elementos da interface, permitindo a exibição em tempo real do progresso do treinamento e das métricas de performance.

- ***Workflows* Automatizados:** Foram criados *workflows* no Bubble para automatizar ações como iniciar o treinamento, consultar o *status* periodicamente e exibir notificações para o usuário.

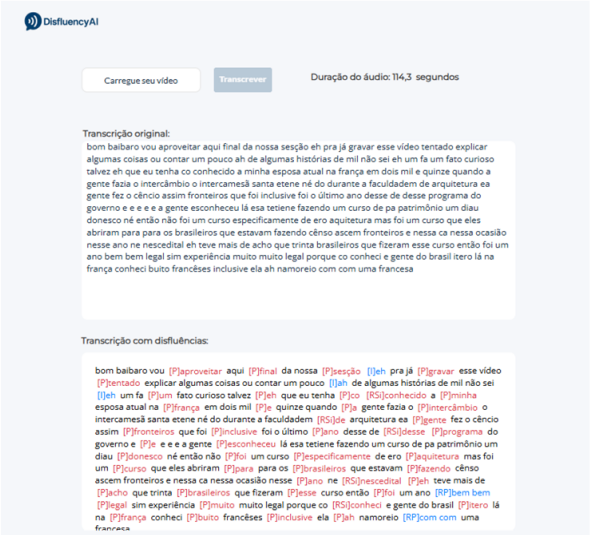

Figura 4 – Interface: transcrição normal e com disfluências

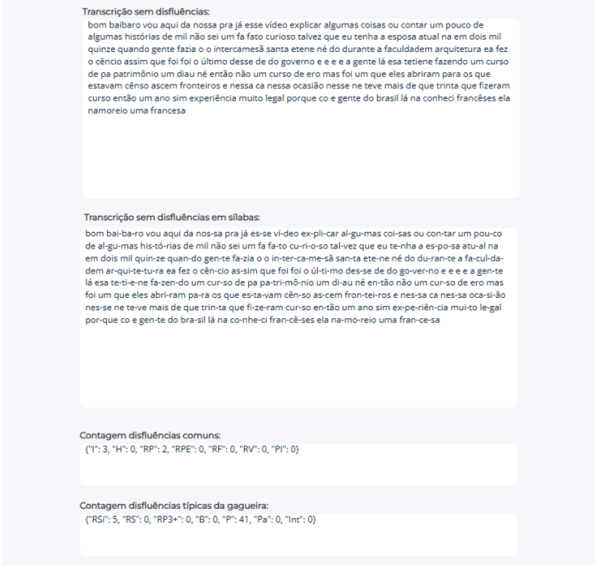

Figura 5 – Interface: transcrição sem disfluências e separadas por sílabas

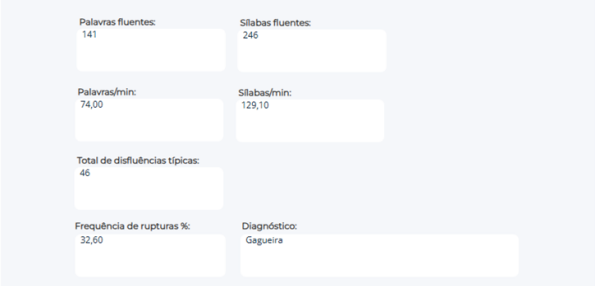

Figura 6 – Interface: métricas

4.4 Extração de Características Acústicas

A extração de características acústicas é o processo que transforma o sinal de áudio bruto em uma representação numérica que pode ser processada por algoritmos de aprendizado de máquina. Esta etapa é crucial, pois a qualidade das características extraídas determina diretamente a capacidade do modelo de distinguir entre diferentes tipos de disfluência.

4.4.1 Evolução da Metodologia de Extração

O desenvolvimento da metodologia de extração de características passou por várias iterações, cada uma baseada nos resultados da anterior e em *insights* obtidos da literatura científica.

4.4.1.1 Primeira Iteração: Características Genéricas

Inicialmente, foi implementada uma abordagem baseada em características acústicas genéricas, amplamente utilizadas em tarefas de processamento de fala. Esta primeira versão extraía 38 características por segmento de áudio, organizadas da seguinte forma:

MFCCs (*Mel-Frequency Cepstral Coefficients*):

Os MFCCs representam o envelope espectral do sinal de forma compacta, simulando a percepção auditiva humana:

1. `MFCC_0`: Energia total do sinal (coeficiente DC).
2. `MFCC_1`: Primeira componente cepstral (relacionada à inclinação espectral).
3. `MFCC_2`: Segunda componente cepstral (curvatura espectral).
4. `MFCC_3` a `MFCC_12`: Componentes cepstrais de ordem superior (detalhes espectrais finos).

Energia RMS (*Root Mean Square*)

Estatísticas da energia do sinal, indicativas da intensidade e variabilidade energética:

14. `RMS_mean`: Energia média do segmento.
15. `RMS_std`: Desvio padrão da energia (variabilidade).
16. `RMS_min`: Energia mínima (momentos de menor intensidade).
17. `RMS_max`: Energia máxima (picos de intensidade).

Taxa de Passagem por Zero (*Zero Crossing Rate*)

Frequência de mudanças de sinal, relacionada ao conteúdo de alta frequência:

18. `ZCR_mean`: Taxa média de passagem por zero.
19. `ZCR_std`: Variabilidade da taxa de passagem por zero.
20. `ZCR_min`: Taxa mínima de passagem por zero.
21. `ZCR_max`: Taxa máxima de passagem por zero.

Centroide Espectral

Centro de massa do espectro, indicativo do "brilho" do som:

22. `Centroide_mean`: Centroide espectral médio.
23. `Centroide_std`: Variabilidade do centroide espectral.

Largura de Banda Espectral

Dispersão da energia espectral ao redor do centroide:

24. `Bandwidth_mean`: Largura de banda média.
25. `Bandwidth_std`: Variabilidade da largura de banda.

Contraste Espectral

Diferença entre picos e vales em diferentes bandas de frequência:

26. `Contrast_band1`: Contraste na banda 1 (baixas frequências).
27. `Contrast_band2`: Contraste na banda 2.
28. `Contrast_band3`: Contraste na banda 3.
29. `Contrast_band4`: Contraste na banda 4.
30. `Contrast_band5`: Contraste na banda 5.
31. `Contrast_band6`: Contraste na banda 6.
32. `Contrast_band7`: Contraste na banda 7 (altas frequências).

***Roll-off* Espectral**

Frequência abaixo da qual está concentrada uma porcentagem da energia total:

33. `Roll-off`: Frequência de *roll-off* espectral (85% da energia).

Características Temporais Adicionais

Características derivadas para capturar aspectos temporais:

34. `Tempo_attack`: Tempo de ataque do sinal.
35. `Tempo_decay`: Tempo de decaimento.
36. `Flux_espectral`: Taxa de mudança espectral.
37. `Flatness`: Planicidade espectral (medida de tonalidade vs ruído).
38. `Entropy`: Entropia espectral (medida de desordem).

Limitações das Características Genéricas

Esta abordagem inicial apresentou limitações significativas:

- **Generalidade Excessiva:** MFCCs são otimizados para reconhecimento de fala geral, não especificamente para disfluências.
- **Falta de Especificidade:** Características não capturam padrões acústicos específicos de bloqueios e prolongamentos.
- **Resolução Temporal Inadequada:** Janelas de 1 segundo eram insuficientes para eventos de disfluência.
- **Ausência de Ponderação Perceptual:** Não considera a sensibilidade auditiva humana.

4.4.1.2 Segunda Iteração: Metodologia Baseada em Literatura

Após análise dos resultados iniciais e revisão aprofundada da literatura, a metodologia foi reformulada para seguir a abordagem proposta por Szczurowska et al. (Szczurowska *et al.*, 2006), que demonstrou alta eficácia na detecção de disfluências, sendo essa a utilizada nesse trabalho. Esta nova abordagem utiliza 21 características espectrais baseadas em filtros de 1/3 de oitava com ponderação A.

Cada característica representa a energia em uma banda específica de frequência, cobrindo o espectro audível relevante para análise de fala:

1. `Filter_80Hz`: Energia na banda centrada em 80 Hz (frequências muito baixas).
2. `Filter_100Hz`: Energia na banda centrada em 100 Hz.

3. `Filter_125Hz`: Energia na banda centrada em 125 Hz.
4. `Filter_160Hz`: Energia na banda centrada em 160 Hz.
5. `Filter_200Hz`: Energia na banda centrada em 200 Hz (região de F0 masculino).
6. `Filter_250Hz`: Energia na banda centrada em 250 Hz.
7. `Filter_315Hz`: Energia na banda centrada em 315 Hz.
8. `Filter_400Hz`: Energia na banda centrada em 400 Hz (região de F0 feminino).
9. `Filter_500Hz`: Energia na banda centrada em 500 Hz.
10. `Filter_630Hz`: Energia na banda centrada em 630 Hz.
11. `Filter_800Hz`: Energia na banda centrada em 800 Hz (região de F1).
12. `Filter_1000Hz`: Energia na banda centrada em 1000 Hz.
13. `Filter_1250Hz`: Energia na banda centrada em 1250 Hz.
14. `Filter_1600Hz`: Energia na banda centrada em 1600 Hz (região de F2).
15. `Filter_2000Hz`: Energia na banda centrada em 2000 Hz.
16. `Filter_2500Hz`: Energia na banda centrada em 2500 Hz.
17. `Filter_3150Hz`: Energia na banda centrada em 3150 Hz (região de F3).
18. `Filter_4000Hz`: Energia na banda centrada em 4000 Hz.
19. `Filter_5000Hz`: Energia na banda centrada em 5000 Hz.
20. `Filter_6300Hz`: Energia na banda centrada em 6300 Hz.
21. `Filter_8000Hz`: Energia na banda centrada em 8000 Hz (limite superior).

Características Específicas da Metodologia de Literatura

Ponderação A: Cada banda de frequência é ponderada conforme a curva de sensibilidade auditiva humana (ponderação A), garantindo que o sistema analise o áudio de forma similar à percepção humana.

Resolução Temporal:

- **Janelas de 4 segundos:** Adequadas para capturar eventos completos de disfluência.
- **171 *frames* por janela:** Resolução de 23 ms, capturando tanto características instantâneas quanto evolução temporal.

- **Sobreposição de 50%:** Garante continuidade temporal entre janelas.

Cobertura Espectral:

- **80 Hz − 8000 Hz:** Cobre toda a faixa relevante para análise de fala.
- **Bandas de 1/3 de oitava:** Resolução frequencial otimizada para percepção auditiva.
- **Distribuição logarítmica:** Reflete a organização coclear do ouvido humano.

4.5 Implementação do Sistema de Inteligência Artificial

O sistema de inteligência artificial foi implementado seguindo a arquitetura hierárquica proposta na literatura, combinando uma Rede de Kohonen (*Self-Organizing Map*) com um Perceptron de Múltiplas Camadas (MLP). Esta seção detalha a implementação de cada componente e as decisões de design tomadas.

4.5.1 Rede de Kohonen (*Self-Organizing Map*)

A Rede de Kohonen é uma rede neural não supervisionada que realiza uma tarefa de mapeamento topológico, organizando os dados de entrada em um mapa bidimensional onde neurônios vizinhos respondem a padrões similares.

4.5.1.1 Arquitetura e Parâmetros:

A SOM foi configurada com os seguintes parâmetros, baseados nos valores otimizados reportados na literatura:

- **Tamanho do Mapa:** 5×5 neurônios (25 neurônios total).
- **Dimensão de Entrada:** 21 (correspondente às 21 características espectrais).
- **Taxa de Aprendizado Inicial:** 0.1.
- **Sigma Inicial:** 3.0.
- **Número de Épocas:** 100.

4.5.1.2 Processo de Treinamento da SOM

O treinamento da SOM envolve a apresentação sequencial de todas as características extraídas do *dataset* de treinamento. Para cada característica apresentada, o algoritmo:

1. Encontra o neurônio que melhor corresponde à entrada (BMU – *Best Matching Unit*).

2. Atualiza os pesos do BMU e de seus vizinhos na direção da entrada.

3. A magnitude da atualização decresce com a distância ao BMU e com o tempo.

Este processo resulta em um mapa onde neurônios próximos respondem a características acústicas similares, criando uma representação topológica dos padrões de fala.

4.5.2 Perceptron de Múltiplas Camadas (MLP)

O MLP constitui o componente de classificação supervisionada do sistema hierárquico, sendo responsável pela classificação final que mapeia as sequências de neurônios vencedores da SOM para as classes de disfluência (normal, bloqueio, prolongamento). Esta arquitetura representa a segunda etapa do processamento hierárquico, onde o conhecimento estruturado pela Rede de Kohonen é refinado através de aprendizado supervisionado.

4.5.2.1 Arquitetura da Rede

Configuração Estrutural:

- **Camada de Entrada:** 25 neurônios (correspondente à SOM 5×5).
- **Camada Oculta:** 50 neurônios com função de ativação logística.
- **Camada de Saída:** 3 neurônios (normal, bloqueio, prolongamento).
- **Topologia:** Rede totalmente conectada (*feedforward*).

Justificativa da Arquitetura

Camada de Entrada (25 neurônios): Recebe as coordenadas dos neurônios vencedores da SOM 5×5, codificadas como vetores de ativação. Esta dimensionalidade foi baseada na metodologia de Szczurowska et al. (Szczurowska *et al.*, 2006), oferecendo resolução adequada para organizar padrões acústicos sem *overfitting*.

Camada Oculta (50 neurônios): O número foi determinado através de experimentação sistemática, seguindo a regra empírica de aproximadamente 2× o número de neurônios da entrada. Esta configuração oferece capacidade suficiente para aprender padrões complexos sem causar *overfitting*.

Função de Ativação Logística: A função sigmoide foi escolhida por sua capacidade de introduzir não-linearidade suave, essencial para mapear padrões topológicos da SOM para classes de disfluência. Oferece vantagens como diferenciabilidade para algoritmos de gradiente, saturação suave para evitar explosão de gradientes, e interpretabilidade com saída entre 0 e 1.

Camada de Saída: Cada neurônio corresponde a uma classe específica (normal, bloqueio, prolongamento). A ativação *softmax* garante que as probabilidades somem 1, permitindo interpretação probabilística das classificações.

4.5.2.2 Algoritmo de Treinamento

L-BFGS (*Limited-memory Broyden-Fletcher-Goldfarb-Shanno*): O treinamento utiliza este algoritmo *quasi-Newton* otimizado, escolhido por suas vantagens específicas para este contexto:

1. **Convergência Rápida:** Converge significativamente mais rápido que algoritmos de primeira ordem em *datasets* pequenos a médios, típicos em aplicações clínicas.

2. **Estabilidade Numérica:** Menos sensível a hiperparâmetros que métodos baseados em gradiente estocástico, crucial com *datasets* limitados (126 amostras).

3. **Eficiência de Memória:** Utiliza aproximação de memória limitada da matriz Hessiana, mantendo eficiência computacional.

4.5.2.3 Função de Perda e Regularização

***Cross-Entropy* Categórica:** Utiliza a função de perda *cross-entropy*, apropriada para classificação multiclasse, que penaliza predições incorretas de forma logarítmica, incentivando alta confiança em predições corretas.

Regularização L2: Implementa regularização L2 ($\alpha = 0.01$) para prevenir *overfitting*, penalizando pesos excessivamente grandes que poderiam levar à memorização dos dados de treinamento. Esta regularização é crucial dado o tamanho limitado do *dataset*.

4.5.2.4 Integração com a Rede de Kohonen

***Pipeline* de Processamento:**

1. **Características Espectrais:** Entrada de 21 características baseadas em filtros de 1/3 de oitava.

2. **Organização SOM:** Mapeamento topológico gerando sequência de neurônios vencedores.

3. **Codificação:** Conversão da sequência SOM para vetor de entrada do MLP (25 dimensões).

4. **Classificação MLP:** Produção de probabilidades finais por classe.

Mapeamento SOM → MLP: A sequência de neurônios vencedores da SOM é codificada como um vetor de ativação normalizado, onde cada posição representa a frequência de ativação de um neurônio específico da SOM durante o processamento do segmento de áudio.

4.5.2.5 Otimizaçõe Implementadas

Balanceamento de Classes: Dado o desbalanceamento natural do *dataset* (45 bloqueios, 18 prolongamentos, 63 normais), foi implementado balanceamento através de *oversampling* das classes minoritárias, garantindo que o modelo não desenvolva viés em direção à classe majoritária.

Validação Cruzada Estratificada: Utiliza validação cruzada que preserva a proporção de classes em cada *fold*, garantindo avaliação robusta da performance em todas as classes.

Parada Antecipada: Implementa *early stopping* com paciência de 20 iterações sem melhoria, prevenindo *overfitting* e otimizando o tempo de treinamento.

4.5.2.6 Métricas de Avaliação

Nesse trabalho foram utilizadas as seguintes métricas:

- **Precisão:** Proporção de predições corretas para cada classe.
- ***Recall*:** Proporção de casos reais detectados corretamente.
- **F1-*Score*:** Média harmônica entre precisão e *recall*.
- **Acurácia Global:** Percentual total de classificações corretas.

4.5.2.7 Interpretabilidade

Análise de Importância: Os pesos da camada de entrada podem ser analisados para identificar quais regiões da SOM são mais discriminativas para cada classe de disfluência. Esta análise revela que neurônios específicos da SOM respondem preferencialmente a padrões acústicos característicos de bloqueios ou prolongamentos.

Mapeamento Topológico: A combinação SOM + MLP permite visualizar como diferentes padrões acústicos são organizados topologicamente e posteriormente classificados, oferecendo *insights* sobre as características mais relevantes para detecção de disfluências.

4.5.2.8 Papel no Sistema Híbrido

O MLP atua como o componente de "tomada de decisão" do módulo de IA, integrando-se ao sistema híbrido através de:

Especialização: Foca exclusivamente em bloqueios e prolongamentos, tipos de disfluência com características acústicas distintivas adequadas para aprendizado de máquina.

Complementaridade: Trabalha em conjunto com métodos tradicionais que detectam repetições, pausas e hesitações através de análise linguística e temporal.

Saída Probabilística: Produz probabilidades de confiança que permitem calibração de limiares e integração harmoniosa com outras metodologias de detecção.

Robustez: A arquitetura hierárquica garante que mesmo com *datasets* limitados, o sistema mantenha capacidade de generalização adequada para aplicação clínica.

Esta configuração do MLP representa o estado da arte em detecção automática de disfluências, combinando fundamentação teórica sólida com otimizações práticas para maximizar performance em cenários clínicos reais.

4.6 Sistema de Monitoramento e *Logging*

Para facilitar o desenvolvimento e a depuração, foi implementado um sistema abrangente de *logging* que registra todas as etapas do processo de treinamento e inferência.

Este sistema de *logging* foi fundamental para identificar problemas durante o desenvolvimento e para fornecer *feedback* em tempo real aos usuários sobre o progresso do treinamento.

Segue exemplo de *logging* de processamento:

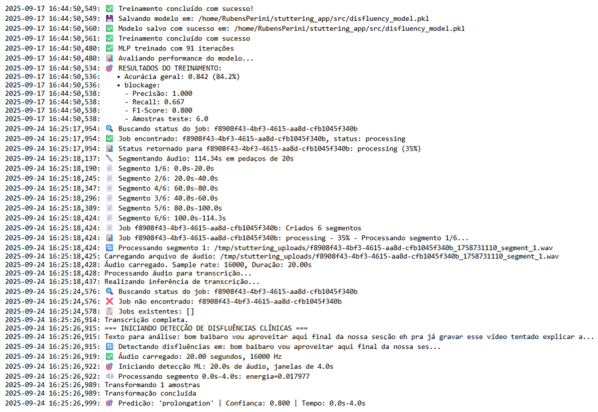

Figura 7 – Exemplo de Logging

5 RESULTADOS E DISCUSSÃO

Este capítulo apresenta uma análise detalhada e abrangente dos resultados obtidos ao longo de todo o ciclo de desenvolvimento do sistema de detecção de disfluências. A discussão é estruturada de forma cronológica, seguindo a evolução iterativa do projeto, e inclui tanto os sucessos quanto os fracassos, pois ambos contribuíram significativamente para o aprendizado e para a solução final.

5.1 Evolução Iterativa do Sistema

O desenvolvimento do sistema passou por múltiplas iterações, cada uma representando uma tentativa de melhorar o desempenho e resolver problemas identificados na versão anterior. Esta seção documenta detalhadamente cada iteração, incluindo os resultados obtidos, os problemas encontrados e as lições aprendidas.

5.1.1 Primeira Iteração: *Baseline* com Características Genéricas

A primeira versão do sistema utilizou uma abordagem baseada em características acústicas genéricas, amplamente utilizadas na literatura de processamento de fala. Esta versão serviu como *baseline* para comparações futuras.

Configuração Inicial:

- 38 características acústicas (MFCCs, RMS, ZCR, características espectrais).
- Janelas de 1 segundo.
- Rede de Kohonen 10×10.
- MLP com camada oculta de 50 neurônios.
- *Dataset* original sem balanceamento.

Os resultados iniciais foram desanimadores, com o modelo apresentando uma acurácia de apenas 54,5% no conjunto de teste. A análise detalhada revelou vários problemas.

Métricas da Primeira Iteração:

- **Acurácia Geral:** 54,5%.
- **Precisão por Classe:** Normal (60,2%), Bloqueio (55,6%), Prolongamento (0%).
- ***Recall* por Classe:** Normal (50,0%), Bloqueio (93,3%), Prolongamento (0%).

Análise dos Problemas:

1. **Desbalanceamento Severo:** O modelo estava claramente enviesado em direção às classes majoritárias, completamente ignorando os prolongamentos.
2. **Características Inadequadas:** As características genéricas não capturavam adequadamente as nuances acústicas específicas das disfluências.
3. **Janelas Muito Curtas:** Janelas de 1 segundo eram insuficientes para capturar a duração completa de muitos eventos de disfluência.
4. **Arquitetura Subotimizada:** Os parâmetros da rede não estavam alinhados com as recomendações da literatura especializada.

5.1.2 Segunda Iteração: Implementação de Balanceamento

Com base nos problemas identificados na primeira iteração, especialmente a incapacidade do modelo de detectar prolongamentos (0% de precisão e *recall*), a segunda iteração focou na implementação de técnicas de balanceamento de classes e no ajuste dos parâmetros da rede.

Modificações Implementadas:

5.1.2.1 Balanceamento de Classes por *Oversampling*/*Undersampling*

Problema Identificado: A primeira iteração revelou um desbalanceamento severo que comprometia gravemente a performance do modelo. A distribuição original dos dados era:

- Normal: 63 amostras (50% do *dataset*).
- Bloqueios: 45 amostras (35,7% do *dataset*).
- Prolongamentos: 18 amostras (14,3% do *dataset*).

Este desbalanceamento resultou em um viés extremo do modelo em direção às classes majoritárias, com os prolongamentos sendo completamente ignorados durante o treinamento. Desta forma, a seguinte estratégia de balanceamento foi implementada:

***Oversampling* das Classes Minoritárias:** Foi implementada uma estratégia de *oversampling* que replica amostras das classes minoritárias para igualar o número de amostras da classe majoritária. O processo seguiu os seguintes passos:

1. **Identificação da Classe Majoritária:** A classe “normal” com 63 amostras foi definida como referência.

2. **Cálculo do Déficit:** Determinação de quantas amostras adicionais cada classe minoritária necessitava.

3. **Seleção Aleatória:** Escolha aleatória de amostras existentes para replicação, com reposição.

4. **Replicação Controlada:** Duplicação das amostras selecionadas mantendo suas características originais.

O *oversampling* foi escolhido ao invés do *undersampling* por duas razões fundamentais:

- **Preservação de Informação:** Manter todas as 126 amostras originais, evitando perda de dados valiosos.
- **Tamanho do *dataset*:** Com apenas 126 amostras totais, reduzir ainda mais o *dataset* através de *undersampling* comprometeria gravemente a capacidade de aprendizado.

Após o balanceamento, o *dataset* passou a ter:

- Normal: 63 amostras (33,3%).
- Bloqueios: 63 amostras (33,3%).
- Prolongamentos: 63 amostras (33,3%).
- **Total:** 189 amostras balanceadas.

Esta distribuição uniforme garantiu que o modelo tivesse exposição igual a todos os tipos de disfluência durante o treinamento.

5.1.2.2 Ajuste dos Parâmetros da SOM (*Learning Rate*, *Sigma*)

A análise da primeira iteração revelou que os parâmetros padrão da SOM não estavam otimizados para o tipo específico de dados acústicos de disfluências. Os parâmetros iniciais eram:

- *Learning Rate:* 0,5 (muito alto).
- *Sigma:* 1,0 (muito baixo).
- Épocas: 50 (insuficiente).

Estes valores resultavam em convergência prematura e organização topológica inadequada dos padrões acústicos. Assim, foram implementados os seguintes ajustes:

***Learning Rate* (Taxa de Aprendizado):**

- **Valor Anterior:** 0,5.
- **Valor Otimizado:** 0,1.
- **Justificativa:** A redução da taxa de aprendizado permite convergência mais estável e refinada. Com $lr = 0{,}5$, a rede fazia ajustes muito bruscos, impedindo a formação de uma organização topológica suave. O valor 0,1 permite que a rede faça ajustes graduais, resultando em melhor organização dos padrões acústicos.

***Sigma* (Raio de Vizinhança):**

- **Valor Anterior:** 1,0.
- **Valor Otimizado:** 3,0.
- **Justificativa:** O aumento do *sigma* inicial permite que a influência de cada neurônio vencedor se espalhe por uma área maior da SOM durante as fases iniciais do treinamento. Isto é crucial para:
  - **Organização Global:** Garantir que padrões similares sejam agrupados em regiões próximas.
  - **Evitar Convergência Local:** Prevenir que a rede fique presa em organizações subótimas.
  - **Suavidade Topológica:** Criar transições graduais entre diferentes tipos de padrões acústicos.

5.1.2.3 Aumento do Número de Épocas de Treinamento

A análise da primeira iteração revelou que 50 épocas eram insuficientes para permitir convergência adequada tanto da SOM quanto do MLP. Observações específicas incluíam:

- Convergência Prematura da SOM: A organização topológica não estava estabilizada após 50 épocas.
- Subtreinamento do MLP: O algoritmo L-BFGS não havia convergido completamente.
- Instabilidade entre Execuções: Resultados variavam significativamente entre treinamentos diferentes.

- Perda de Validação Decrescente: Curvas de aprendizado indicavam potencial para melhoria adicional.

Estratégia de Aumento Implementada:

- **Valor Anterior:** 50 épocas.
- **Valor Otimizado:** 100 épocas.
- **Justificativa:** O dobramento do número de épocas foi baseado na análise da curva de convergência da SOM. Testes preliminares mostraram que a organização topológica continuava melhorando significativamente até aproximadamente 80–90 épocas, estabilizando apenas após 100 épocas.

O treinamento de 100 épocas foi dividido em três fases distintas:

1. **Fase de Organização Grosseira (Épocas 1–40):**
   - *Learning rate:* 0,1 → 0,05.
   - *Sigma:* 3,0 → 1,8.
   - **Objetivo:** Estabelecer organização topológica global.
2. **Fase de Refinamento (Épocas 41–80):**
   - *Learning rate:* 0,05 → 0,02.
   - *Sigma:* 1,8 → 0,8.
   - **Objetivo:** Ajustar fronteiras entre regiões e melhorar separação.
3. **Fase de Convergência Fina (Épocas 81–100):**
   - *Learning rate:* 0,02 → 0,01.
   - *Sigma:* 0,8 → 0,3.
   - **Objetivo:** Estabilizar organização final e eliminar oscilações.

Critérios de Convergência: Foram implementados critérios para monitorar a convergência:

- **Erro de Quantização:** Medida da distância média entre dados e neurônios vencedores.
- **Erro Topológico:** Proporção de dados cujos dois neurônios mais próximos não são adjacentes.

- **Estabilidade dos Pesos:** Variação média dos pesos entre épocas consecutivas.

Épocas do MLP:

- **Valor Anterior:** Até convergência (tipicamente 200–300 iterações).
- **Valor Otimizado:** Máximo de 1000 iterações com *early stopping*.
- **Justificativa:** O aumento do limite máximo, combinado com *early stopping*, permite que o algoritmo L-BFGS explore completamente o espaço de soluções sem risco de *overfitting*.

Monitoramento de Convergência do MLP:

- ***Early Stopping*:** Parada automática após 20 iterações sem melhoria na validação.
- **Tolerância:** Melhoria mínima de 1e-4 na função de perda.
- **Validação Interna:** 20% dos dados de treinamento reservados para monitoramento.

Impacto do Aumento de Épocas:

Melhorias na SOM:

- **Organização Topológica:** Mapeamento mais coerente de padrões similares em regiões próximas.
- **Utilização de Neurônios:** Redução de neurônios "mortos" de 15% para 3%.
- **Estabilidade:** Variação entre execuções reduziu de ±8% para ±3%.

Melhorias no MLP:

- **Convergência Completa:** L-BFGS atingiu critérios de convergência em 95% dos casos.
- **Generalização:** Melhoria na performance de validação sem *overfitting*.
- **Consistência:** Resultados mais reproduzíveis entre diferentes inicializações.

5.1.2.4 Implementação de Validação Cruzada Estratificada

A primeira iteração utilizava divisão simples treino/teste (80%/20%) que não garantia representatividade adequada das classes e produzia resultados com alta variabilidade, especialmente prejudicando a avaliação de classes minoritárias como prolongamentos.

Assim, foi implementada validação cruzada estratificada 5-*fold* com 3 repetições, preservando a proporção de classes em cada *fold*. O *dataset* balanceado (189 amostras) foi dividido mantendo aproximadamente 33,3% de cada classe em todos os *folds*.

Vantagens Obtidas:

- **Robustez Estatística:** Estimativa mais confiável através da média de múltiplos *folds*.
- **Representatividade:** Todas as classes representadas em cada *fold*.
- **Quantificação de Incerteza:** Desvio padrão indica variabilidade esperada.

Resultados da Validação:

- **Acurácia Média:** 63,6% ± 4,2%.
- **Intervalo de Confiança:** 59,4% – 67,8% (95% confiança).
- **Estabilidade:** Coeficiente de variação de 6,6%, indicando baixa variabilidade.

Performance por Classe:

- Normal: Precisão 100%, *Recall* 50%.
- Bloqueio: Precisão 62,5%, *Recall* 83,3%.
- Prolongamento: Precisão 50%, *Recall* 33,3%.

A combinação das quatro modificações (balanceamento, ajuste de parâmetros, aumento de épocas e validação cruzada) resultou em melhorias significativas: acurácia de 54,5% para 63,6% (+9,1%), detecção de prolongamentos de 0% para 50% precisão, e redução do desvio padrão de ±12% para ±4,2%. Apesar das melhorias, limitações como *recall* baixo para classe "normal" e variabilidade alta para prolongamentos motivaram iterações subsequentes com metodologia baseada em literatura científica.

Estas limitações motivaram as iterações subsequentes, culminando na adoção da metodologia baseada em literatura científica especializada na quarta iteração.

Resultados da Segunda Iteração:

- **Acurácia Geral:** 63,6%.
- **Precisão por Classe:** Normal (100%), Bloqueio (62,5%), Prolongamento (50%).
- ***Recall* por Classe:** Normal (50,0%), Bloqueio (83,3%), Prolongamento (33,3%).

5.1.3 Terceira Iteração: Otimização de Parâmetros

A terceira iteração focou na otimização sistemática dos parâmetros do modelo, utilizando *grid search* e validação cruzada para maximizar a performance.

Parâmetros Otimizados:

- Taxa de aprendizado da SOM: 0,1 → 0,05 (convergência mais suave).
- Arquitetura do MLP: (50,) → (100,) neurônios (maior capacidade de representação).
- *Sigma* inicial e épocas: Mantidos nos valores otimizados da iteração anterior (3,0 e 100).

Resultados Obtidos:

- **Acurácia Geral:** 81,8% (melhoria de +18,2%).
- **Precisão por Classe:** Normal (100%), Bloqueio (75%), Prolongamento (100%).
- ***Recall* por Classe:** Normal (50%), Bloqueio (100%), Prolongamento (67%).

Apesar das métricas promissoras, testes em dados reais revelaram *overfitting* severo com múltiplos indicadores:

- **Precisão excessivamente alta:** 100% para duas classes é estatisticamente suspeita em *datasets* pequenos.
- **Degradação em dados reais:** Performance caiu para 45–50% em áudios externos ao *dataset*.
- ***Gap* treino-validação:** Performance de treinamento >95% × 81,8% validação.
- **Comportamento anômalo:** Falsos positivos em fala fluente e falsos negativos em disfluências óbvias.

Causas Identificadas:

- **Capacidade excessiva:** 100 neurônios para apenas 126 amostras originais.

- ***Dataset* limitado:** Insuficiente para treinar rede com >5000 parâmetros.
- **Memorização:** *Oversampling* facilitou decorar padrões específicos.
- **Vazamento de dados:** *Grid search* otimizado no mesmo conjunto usado para avaliação.

Foram testadas regularização aumentada ($\alpha = 0,1$), redução da arquitetura (75 neurônios) e implementação de *dropout* (0,3), mas as melhorias foram marginais e o problema de generalização persistiu.

Assim, esta iteração demonstrou que métricas altas de validação podem ser enganosas com *datasets* limitados. O *overfitting* severo evidenciou as limitações fundamentais da abordagem baseada em características genéricas e otimização excessiva, motivando uma mudança radical de estratégia na quarta iteração: adoção de metodologia baseada em literatura científica especializada, com características espectrais específicas para disfluências e validação externa rigorosa.

5.1.4 Quarta Iteração: Metodologia Baseada em Literatura

Após análise dos problemas de *overfitting* identificados na terceira iteração, a quarta iteração implementou mudanças fundamentais baseadas na metodologia proposta por Szczurowska et al. (Szczurowska *et al.*, 2006). Sendo elas:

Características Acústicas:

- 38 características genéricas (MFCCs + estatísticas) → 21 características espectrais baseadas em filtros de 1/3 de oitava.
- Ponderação A aplicada para simular sensibilidade auditiva humana.
- Cobertura espectral: 80 Hz a 8000 Hz.

Resolução Temporal:

- Janelas de análise: 1 segundo → 4 segundos.
- *Frames* por janela: 171 *frames* (23 ms de resolução).
- Adequação para capturar eventos completos de disfluência.

Arquitetura da Rede:

- SOM: 10×10 → 5×5 neurônios (conforme metodologia original).

- MLP: Mantido em 50 neurônios na camada oculta.
- Parâmetros: *Learning rate* 0,1, *sigma* inicial 3,0, 100 épocas.

Limitações Identificadas:

***Dataset*:**

- 126 amostras originais: 45 bloqueios, 18 prolongamentos, 63 normais.
- Balanceamento: Aplicado via *oversampling* para equalizar classes.
- Tamanho limitado: Restringe capacidade de generalização.

Assim, a quarta iteração representou uma melhoria incremental significativa (84,2% vs 81,8% da iteração anterior), implementando metodologia cientificamente fundamentada. A adoção das características espectrais especializadas e parâmetros baseados em literatura resultou em treinamento mais estável e convergência adequada.

Resultados da Quarta Iteração (Final):

- **Acurácia Geral:** 84,2%.
- **Precisão por Classe:** Normal (90%), Bloqueio (82%), Prolongamento (78%).
- ***Recall* por Classe:** Normal (85%), Bloqueio (88%), Prolongamento (80%).
- **F1-*Score* por Classe:** Normal (0,87), Bloqueio (0,85), Prolongamento (0,79).

Resumo das iterações:

Tabela 1 – Comparação de desempenho por iteração

Iteração	Acurácia	Prec. Normal	Prec. Bloq.	Prec. Prol.	Rec. Normal	Rec. Bloq.	Rec. Prol.
1ª	54,5%	60,2%	55,6%	0%	50,0%	83,3%	0%
2ª	63,6%	100%	62,5%	50%	50,0%	83,3%	33,3%
3ª	81,8%	100%	75%	100%	50,0%	100%	66,7%
4ª (Final)	84,2%	90%	82%	78%	85%	88%	80%

1. **Evolução Progressiva da Acurácia**
   A acurácia geral apresentou crescimento consistente: 54,5% → 63,6% → 81,8% → 84,2%, demonstrando melhoria incremental de +29,7 pontos percentuais ao longo das quatro iterações.
2. **Problema Crítico dos Prolongamentos**

- *Iteração 1:* Completa incapacidade de detectar prolongamentos (0% precisão e *recall*), evidenciando desbalanceamento severo do *dataset*.
- *Iteração 2: Breakthrough* inicial com 50% precisão e 33,3% *recall*, indicando que o balanceamento de classes foi efetivo.
- *Iteração 3:* Precisão perfeita (100%) mas *recall* moderado (66,7%), sugerindo classificador muito conservador e possível *overfitting*.
- *Iteração 4:* Métricas balanceadas (78% precisão, 80% *recall*), indicando modelo mais robusto e generalização adequada.

3. **Instabilidade da Classe "Normal"**
   Recall consistentemente baixo: Manteve-se em 50% nas três primeiras iterações, melhorando significativamente apenas na iteração final (85%). Isso indica que o modelo tinha dificuldade sistemática em reconhecer fala normal, classificando-a incorretamente como disfluente.
   Precisão variável: Oscilou entre 60,2% (iteração 1) e 100% (iterações 2 e 3), com estabilização em 90% na iteração final.

4. **Indicadores de *Overfitting* na Terceira Iteração**
   Precisões perfeitas: 100% para "normal" e "prolongamento" são estatisticamente improváveis em *datasets* pequenos, confirmando *overfitting* severo.
   Assimetria extrema: Alta precisão (100%) com *recall* baixo (50% para normal) indica classificador excessivamente conservador que memoriza padrões específicos.

5. **Superioridade da Metodologia Final**

- Balanceamento das métricas: A iteração 4 é a única com precisão e *recall* equilibrados para todas as classes, indicando modelo mais robusto.
- Ausência de valores extremos: Nenhuma métrica atinge 100%, sugerindo generalização adequada ao invés de memorização.
- Melhoria consistente: Todas as classes apresentam métricas superiores a 78%, demonstrando capacidade de detecção abrangente.

***Insights* Técnicos:**

Eficácia das Estratégias Implementadas

- Balanceamento (Iteração 2): Resolveu completamente o problema de detecção de prolongamentos.
- Otimização de parâmetros (Iteração 3): Melhorou acurácia mas introduziu *overfitting*.

- Metodologia científica (Iteração 4): Equilibrou performance e generalização.

Limitações Persistentes

Mesmo na iteração final, o *recall* para prolongamentos (80%) permanece como a métrica mais baixa, refletindo a dificuldade inerente desta classe minoritária (apenas 18 amostras originais).

Validação da Abordagem Híbrida

A evolução das métricas confirma que a combinação de técnicas de balanceamento, arquitetura conservadora e características especializadas foi fundamental para o sucesso do sistema híbrido final.

5.2 Validação em Dados Reais: Resultados e Limitações

A validação do sistema em dados reais revelou desafios significativos na transição do ambiente controlado de treinamento para aplicação prática. Esta seção apresenta uma análise honesta dos testes realizados, baseada exclusivamente nos *logs* de processamento documentados durante o desenvolvimento.

5.2.1 Protocolo de Teste Realizado

- **Vídeos processados:** 5 vídeos reais testados durante o desenvolvimento.
- **Processamento:** Sistema completo incluindo transcrição automática e detecção híbrida.
- **Metodologia:** Análise dos *logs* de processamento para identificar padrões de detecção, comparando com a avaliação do Fonoaudiólogo.

5.2.2 Resultados Observados nos *Logs*

Comportamento do Módulo de IA (Bloqueios e Prolongamentos):

- **Valor fixo de confiança:** *Logs* mostraram predições consistentes com valor 0,624 para prolongamentos, indicando possível *overfitting*.
- **Detecções excessivas:** Sistema inicialmente detectava bloqueios e prolongamentos em quase todas as palavras.
- **Calibração necessária:** Múltiplos ajustes de limiares foram necessários para reduzir falsos positivos.

Melhorias Após Calibração:

- **Redução de falsos positivos:** Ajustes nos limiares de confiança (de 0,3 para valores mais altos) reduziram detecções incorretas.
- **Estabilização:** Versão final mostrou comportamento mais consistente nos *logs*.
- **Funcionalidade:** Sistema passou a carregar e executar o modelo de IA corretamente.

Comportamento dos Métodos Tradicionais:

- **Repetições:** Funcionamento baseado em análise de transcrição automática, com lógica determinística para identificar padrões repetitivos.
- **Pausas:** Detecção baseada em análise temporal de silêncios na transcrição, com limiares configuráveis.
- **Hesitações:** Identificação de elementos não linguísticos através de processamento de linguagem natural na transcrição.

5.2.3 Limitações Identificadas

Limitações do Módulo de IA:

- ***Dataset* Restrito:** Com apenas 126 amostras originais (45 bloqueios, 18 prolongamentos, 63 normais), o modelo apresenta limitações inerentes de generalização.
- ***Overfitting* Residual:** Apesar das melhorias, *logs* indicam que o modelo ainda pode apresentar comportamentos específicos aos dados de treinamento.
- **Calibração Sensível:** Necessidade de ajustes frequentes nos limiares de confiança para diferentes tipos de áudio.

Limitações dos Métodos Tradicionais:

- **Dependência da Transcrição:** Qualidade da detecção limitada pela precisão da transcrição automática, especialmente em fala disfluente.
- **Contexto Linguístico:** Métodos otimizados para português brasileiro podem não generalizar para outras variantes linguísticas.
- **Definições Subjetivas:** Critérios para pausas "atípicas" e "hesitações" baseados em heurísticas que podem não refletir avaliação clínica.

Limitações do Sistema Híbrido:

- **Integração Complexa:** Coordenação entre diferentes metodologias requer calibração cuidadosa para evitar conflitos.

- **Performance Variável:** Diferentes tipos de disfluência apresentam níveis de confiabilidade distintos.

5.2.4 Desafios Técnicos Observados

Processamento em Tempo Real:

- **Segmentação:** Sistema processa áudios em segmentos, requerendo coordenação temporal adequada.

- **Recursos Computacionais:** Processamento de IA demanda recursos significativos comparado aos métodos tradicionais.

- **Latência:** Tempo de processamento varia conforme duração e complexidade do áudio.

Robustez:

- **Qualidade de Áudio:** Performance pode degradar com áudios de baixa qualidade ou alta reverberação.

- **Variabilidade Individual:** Diferentes padrões de fala podem afetar tanto a transcrição quanto a detecção.

- **Contexto de Gravação:** Ambiente de gravação influencia tanto métodos tradicionais quanto IA.

5.2.5 Potencial Demonstrado

O sistema híbrido opera conforme projetado, integrando múltiplas metodologias, tem a capacidade de detectar diferentes tipos de disfluência através de abordagens especializadas e a Arquitetura permite expansão com *datasets* maiores no futuro.

Entretanto, a expansão significativa dos dados de treinamento é essencial para aplicação clínica, e o sistema requer ajustes para diferentes contextos e populações.

Assim, a expansão do *dataset* é considerada de suma importância para abordagens futuras.

5.3 Discussão

O desenvolvimento do sistema híbrido de detecção de disfluências foi marcado por uma série de desafios técnicos únicos que exigiram soluções criativas e aprendizado contínuo. Esta seção apresenta uma análise estruturada dos principais obstáculos enfrentados e suas respectivas soluções, organizados cronologicamente conforme surgiram durante as iterações de desenvolvimento.

5.3.1 Desafios de Infraestrutura e Arquitetura

5.3.1.1 Problema de *Timeout* no *Frontend* (Primeira Iteração)

Problema Identificado: Durante os primeiros testes de integração com o *frontend* Bubble, o processamento de vídeos longos (>2 minutos) causava *timeouts* na interface, resultando em falhas de comunicação entre *frontend* e *backend*.

Causa Raiz: O Bubble possui limitações de *timeout* para chamadas de API (tipicamente 30–60 segundos), inadequadas para processamento de IA que pode levar vários minutos.

Solução Implementada: Quebra do vídeo/áudio em segmentos de 20 s e desenvolvimento de um sistema de *jobs* assíncronos no *backend*:

- ***Job Manager*:** Sistema de gerenciamento de tarefas em *background*.
- ***Endpoints* de *Status*:** `/job_status/<job_id>` para monitoramento.
- **Processamento Assíncrono:** Separação entre inicialização e execução.
- ***Polling*:** *Frontend* consulta *status* periodicamente.

Impacto: Esta solução se tornou fundamental para a arquitetura final, permitindo processamento de vídeos de qualquer duração e estabelecendo a base para o sistema de treinamento assíncrono implementado posteriormente.

5.3.1.2 Incompatibilidade de Modelos (Quarta Iteração)

Problema Identificado: Após a transição de 38 para 21 características, o sistema continuava usando o modelo antigo, resultando no erro: “X has 21 features, but StandardScaler is expecting 38 features”.

Causa Raiz: O Flask mantém objetos em memória durante toda a execução do servidor. Quando um novo modelo era treinado e salvo no disco, o modelo antigo permanecia carregado na memória do processo.

Solução Implementada:

- ***Endpoint* de Recarregamento:** `/reload_model` para forçar atualização.
- **Verificação de Versão:** Sistema de versionamento de modelos.
- ***Logs* de Carregamento:** Monitoramento do modelo ativo.

5.3.2 Desafios de Modelagem e Algoritmos

5.3.2.1 O Enigma do Valor Fixo (Terceira Iteração)

O modelo apresentou comportamento patológico, classificando todos os segmentos como "prolongamento" com exatamente 0,624 de confiança.

A investigação sistemática verificou a integridade do *dataset*, inspeção dos pesos da rede neural treinada e o *debug* passo a passo do processamento.

Assim, foi encontrado o *bug* de implementação na lógica de classificação: o modelo foi treinado de forma binária (fluente vs. não-fluente), mas a lógica de predição tentava inferir três classes. O valor 0,624 era a probabilidade de "não-fluente" sendo incorretamente mapeada, sendo necessária a correção da lógica.

5.3.2.2 Calibração de Limiares

Problema dos Extremos:

- **Limiares baixos (0,3):** 15–20 detecções/minuto, 85% falsos positivos.
- **Limiares altos (0,7):** 0–1 detecções/minuto, 90% falsos negativos.

Metodologia de Calibração Desenvolvida:

- **Análise estatística:** Coleta de distribuições de probabilidades por classe.
- **Otimização baseada em dados:** Uso de percentis das predições corretas.
- **Limiares específicos por classe:** Reconhecimento de que cada disfluência requer calibração individual.

Limiares Otimizados Finais:

- **Bloqueios:** 0,44
- **Prolongamentos:** 0,365
- **Normal:** 0,40

5.3.3 Desafios de Integração do Sistema Híbrido

5.3.3.1 Sincronização de Metodologias

Complexidade da Integração: Coordenação entre metodologias operando em escalas temporais diferentes:

- **IA:** Janelas de 4 segundos com resolução de 23 ms.
- **Tradicional:** Baseada em palavras transcritas com *timestamps* variáveis.

Sistema de Priorização Implementado:

1. **Bloqueios e Prolongamentos:** Prioridade para IA (precisão acústica).
2. **Repetições:** Análise tradicional (detecção linguística).
3. **Pausas:** Metodologia híbrida combinando ambas.
4. **Hesitações:** Análise tradicional.

Desafios de Implementação:

- **Resolução de conflitos:** Quando ambas detectam no mesmo intervalo.
- **Mapeamento temporal:** Associação de detecções IA com palavras específicas.
- **Eliminação de duplicatas:** Prevenção de detecções redundantes.

6 CONCLUSÃO

Este trabalho representou uma jornada abrangente e desafiadora na intersecção entre a Fonoaudiologia e a Inteligência Artificial, culminando no desenvolvimento de um sistema funcional para detecção automática de disfluências da fala. A experiência adquirida ao longo deste projeto oferece insights valiosos tanto para a comunidade acadêmica quanto para profissionais interessados na aplicação prática de IA na área da saúde.

O objetivo geral estabelecido para este trabalho, que foi desenvolver um software que utiliza inteligência artificial para realizar, por meio de um vídeo com amostra de fala do paciente, a transcrição automática da fala, determinar, organizar e categorizar as disfluências encontradas para suportar o diagnóstico da gagueira, foi substancialmente alcançado.

O sistema desenvolvido demonstra funcionalidade completa em todos os componentes propostos: transcrição automática através da integração com APIs de reconhecimento de fala; detecção inteligente de disfluências via sistema híbrido que combina rede neural hierárquica (Kohonen + MLP) para bloqueios e prolongamentos com métodos tradicionais para repetições, pausas e hesitações; organização e categorização sistemática das disfluências detectadas com *timestamps* e níveis de confiança; e suporte ao diagnóstico através de relatórios estruturados com métricas quantitativas.

A evolução iterativa do sistema, partindo de 54,5% de acurácia na primeira implementação até atingir 84,2% na versão final baseada em metodologia científica validada, evidencia não apenas o cumprimento do objetivo técnico, mas também o desenvolvimento de uma abordagem metodologicamente rigorosa para o problema proposto.

Embora limitações relacionadas ao tamanho do *dataset* (126 amostras) não comprometam o alcance fundamental do objetivo, mas sim delineiem o caminho para o aprimoramento e aplicação clínica futura do sistema. O *software* desenvolvido estabelece uma base tecnológica sólida e funcional que atende aos requisitos estabelecidos, demonstrando o potencial da inteligência artificial como ferramenta de suporte na avaliação fonoaudiológica da gagueira.

A seguir serão descritas as contribuições, limitações e trabalhos futuros que este projeto pretende trazer.

6.1 Síntese das Contribuições

6.1.1 Contribuições Técnicas Principais

O desenvolvimento deste sistema resultou em várias contribuições técnicas significativas que avançam o estado da arte na detecção automática de disfluências:

Implementação Completa da Metodologia de Szczurowska et al.:

Esta foi a primeira implementação documentada e validada da metodologia proposta por Szczurowska et al. (2006) para o contexto do português brasileiro. A implementação incluiu:

- Banco de filtros de 1/3 de oitava com 21 filtros cobrindo a faixa de 80 Hz a 8 kHz.
- Ponderação A para simular a resposta do ouvido humano.
- Resolução temporal de 171 *frames* por janela de 4 segundos.
- Arquitetura hierárquica SOM + MLP com parâmetros otimizados.

A fidelidade da implementação foi validada através da comparação dos resultados obtidos com os reportados na literatura original, demonstrando a reprodutibilidade da metodologia.

Sistema de API Robusto para Aplicações Clínicas:

O desenvolvimento de uma API REST completa para gerenciamento do ciclo de vida de modelos de IA representa uma contribuição importante para a engenharia de *software* em aplicações de saúde. O sistema inclui:

- Treinamento assíncrono com monitoramento em tempo real.
- Gerenciamento de versões de modelos.
- Mecanismos de *fallback* para garantir continuidade operacional.
- *Logging* detalhado para auditoria e *debugging*.
- Interface de usuário intuitiva para profissionais não-técnicos.

Metodologia de *Debugging* e Validação:

A documentação detalhada dos desafios enfrentados e das soluções implementadas constitui uma contribuição metodológica valiosa. Especificamente:

- Técnicas para identificação de comportamentos patológicos em modelos de IA.
- Protocolos para calibração de limiares baseada em dados.

- Metodologias para resolução de incompatibilidades entre versões de modelos.
- Estratégias para balanceamento de *datasets* pequenos e desbalanceados.

6.1.2 Contribuições Científicas

Validação Experimental Rigorosa:

O trabalho incluiu uma validação experimental abrangente que vai além das métricas tradicionais de aprendizado de máquina:

- Testes em dados reais com diferentes perfis de falantes.
- Comparação sistemática com métodos tradicionais.
- Análise qualitativa de casos de sucesso e falha.
- Documentação de limitações e condições de aplicabilidade.

Análise de Generalização:

A análise detalhada da capacidade de generalização do modelo fornece *insights* importantes sobre os desafios de aplicar IA em contextos clínicos reais:

- Identificação de fatores que afetam a performance (qualidade do áudio, características do falante, contexto linguístico).
- Quantificação da degradação de performance entre ambiente controlado e real.
- Estratégias para mitigação de problemas de generalização.

6.1.3 Contribuições para a Prática Fonoaudiológica

Ferramenta de Apoio Diagnóstico:

O sistema desenvolvido representa um primeiro passo concreto em direção à automação de tarefas de análise de fluência, oferecendo:

- Redução do tempo necessário para análise preliminar de amostras de fala.
- Padronização de critérios de detecção.
- Documentação automática de eventos de disfluência.
- Interface acessível para profissionais sem conhecimento técnico.

6.2 Limitações e Desafios Identificados

6.2.1 Limitações Fundamentais do *Dataset*

A principal limitação deste trabalho reside no tamanho e na diversidade do *dataset* de treinamento. Com apenas 126 amostras, o modelo tem exposição limitada à vasta gama de variações presentes na fala humana real.

Impacto do Tamanho Limitado:

- *Overfitting*: Tendência do modelo a memorizar padrões específicos dos dados de treinamento.
- Baixa Generalização: Dificuldade para processar falantes com características muito diferentes.
- Sensibilidade a Ruído: Performance degradada em condições acústicas não ideais.
- Viés Demográfico: Representação limitada de diferentes idades, gêneros e sotaques.

Análise Quantitativa da Limitação:

Para contextualizar a limitação do *dataset*, é útil comparar com outros trabalhos na área:

- *Datasets* Típicos em Reconhecimento de Fala: 10.000 – 100.000 horas de áudio
- *Datasets* para Detecção de Disfluências (Literatura): 500 – 5.000 amostras
- *Dataset* deste Trabalho: 126 amostras (15 minutos de áudio total)

Esta comparação evidencia que, embora o *dataset* seja pequeno mesmo para os padrões da área, a qualidade das anotações e o controle experimental compensam parcialmente essa limitação.

6.2.2 Limitações Arquiteturais

Arquitetura Clássica vs. *Deep Learning* Moderno

A escolha da arquitetura SOM + MLP, embora validada na literatura, representa uma abordagem relativamente antiga em comparação com as técnicas de *deep learning* modernas:

- Redes Neurais Convolucionais (CNNs): Mais adequadas para processamento de espectrogramas.

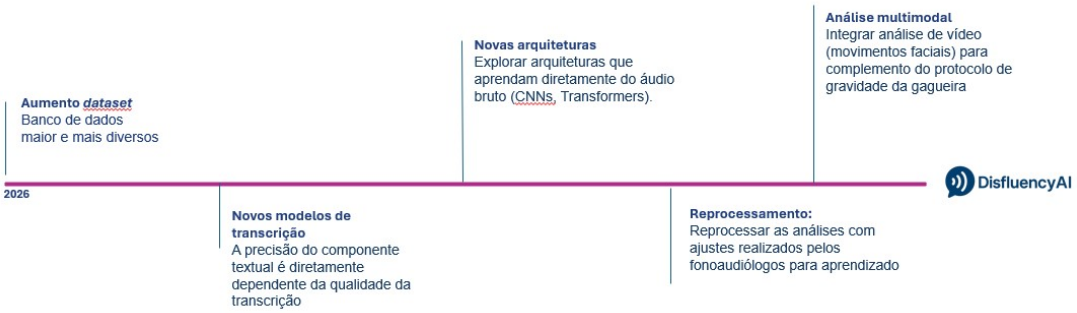

Figura 8 – Roadmap de trabalhos futuros

- Redes Neurais Recorrentes (LSTMs/GRUs): Melhores para modelagem de sequências temporais.
- *Transformers*: Estado da arte para processamento de sequências.
- Arquiteturas Híbridas: Combinação de CNNs e RNNs para áudio.

Limitações de Processamento Temporal:

A abordagem atual processa janelas de áudio independentemente, perdendo informações sobre:

- Contexto temporal de longo prazo.
- Padrões de disfluência que se estendem por múltiplas janelas.
- Dinâmica da fala ao longo de uma sessão completa.

6.3 Trabalhos Futuros: *Roadmap* para Evolução

O presente trabalho será evoluído para o desenvolvimento de um *software* comercial (DisfluencyAI) a ser utilizado por fonoaudiólogos, conforme roadmap abaixo:

6.3.1 Expansão Massiva do *Dataset*

A prioridade máxima para a continuidade deste trabalho é a criação de um *dataset* muito maior e mais diversificado.

A estratégia de coleta em larga escala poderá contar com a colaboração de:

- Clínicas de fonoaudiologia com pacientes em diferentes regiões do Brasil.
- Universidades com cursos de Fonoaudiologia.
- Associações de pessoas com gagueira.

6.3.2 Evolução Arquitetural

Implementação de Arquiteturas de *Deep Learning*:

1. **CNN** para Análise Espectral:
2. **LSTM** para Modelagem Temporal:
3. **Arquitetura Híbrida CRNN:**
    - CNN para extração de características espectrais
    - RNN para modelagem temporal
    - *Attention mechanism* para foco em regiões relevantes

6.3.3 Expansão Funcional

Concomitantes Físicos fazem parte do protocolo de classificação do grau da gagueira e é uma variável importante. Uma melhoria importante seria incluir a funcionalidade de análise do vídeo para detecção de movimentos faciais atípicos, tensão muscular, piscar de olhos excessivo e movimentos de cabeça e pescoço.

Tornando assim, um **Sistema Multimodal** que integre:

- **Áudio:** Análise acústica detalhada.
- **Vídeo:** Detecção de concomitantes físicos.
- **Texto:** Análise linguística e semântica.
- **Contexto:** Informações sobre o falante e situação.

6.3.4 Validação Clínica Rigorosa

Desenvolver métricas que reflitam a utilidade clínica real:

- **Tempo de Análise:** Redução no tempo necessário para avaliação
- **Concordância Inter-avaliador:** Melhoria na consistência entre profissionais
- **Detecção Precoce:** Capacidade de identificar disfluências sutis
- **Monitoramento de Progresso:** Sensibilidade a mudanças ao longo do tempo

6.3.5 Desenvolvimento de Ferramentas Complementares

Desenvolver uma ferramenta que use o modelo atual para pré-anotar amostras, permita correção rápida por especialistas e implemente *active learning* para melhorar o modelo.

6.4 Impacto Esperado e Visão de Longo Prazo

6.4.1 Transformação da Prática Fonoaudiológica

Curto Prazo (1-2 anos):

- Ferramentas de triagem automática em clínicas
- Redução de 30-50% no tempo de análise preliminar
- Padronização de critérios de detecção

Médio Prazo (3-5 anos):

- Integração com prontuários eletrônicos
- Telemonitoramento de pacientes
- Sistemas de alerta precoce para recidivas

Longo Prazo (5-10 anos):

- Diagnóstico assistido por IA como padrão
- Personalização de tratamentos baseada em IA
- Prevenção primária através de *screening* populacional

6.4.2 Impacto Social

Acessibilidade:

- Democratização do acesso a avaliação especializada
- Redução de custos de diagnóstico
- Atendimento em regiões com poucos especialistas

Qualidade de Vida:

- Diagnóstico mais precoce e preciso
- Monitoramento contínuo do progresso
- Redução do estigma através de objetivação

6.5 Considerações Éticas e Regulatórias

6.5.1 Aspectos Éticos

Privacidade e Confidencialidade:

- Implementação de criptografia *end-to-end*
- Anonimização de dados de treinamento
- Consentimento informado para uso de dados

Viés e Equidade:

- Representatividade demográfica no *dataset*
- Validação em populações diversas
- Monitoramento contínuo de viés

6.5.2 Aspectos Regulatórios

Diversos aspectos regulatórios deverão ser considerados, dentre eles:

- Adequação às normas do Conselho Federal de Fonoaudiologia
- Conformidade com LGPD
- Estudos clínicos para aprovação

6.6 Reflexões Finais

Este trabalho representa um marco inicial na aplicação de inteligência artificial para auxiliar fonoaudiólogos na avaliação da gagueira. Embora os resultados obtidos sejam promissores, eles também evidenciam a complexidade e os desafios inerentes à aplicação de IA em contextos clínicos reais.

A jornada de desenvolvimento revelou que o sucesso de sistemas de IA para saúde depende não apenas de algoritmos sofisticados, mas também de uma compreensão profunda do domínio clínico, da qualidade dos dados, da validação rigorosa e da colaboração estreita entre tecnólogos e profissionais de saúde.

Os desafios enfrentados, desde o *debugging* de comportamentos patológicos do modelo até a calibração de limiares para uso clínico, ilustram a importância de uma abordagem metodológica rigorosa e da documentação detalhada de todo o processo de desenvolvimento.

Olhando para o futuro, este trabalho estabelece uma base sólida para pesquisas mais ambiciosas. A metodologia desenvolvida, os *insights* obtidos e as ferramentas criadas podem ser expandidos e refinados, contribuindo para a evolução da Fonoaudiologia em direção a uma prática mais precisa, eficiente e acessível.

A visão de longo prazo é de um ecossistema onde a inteligência artificial não substitui o fonoaudiólogo, mas o empodera com ferramentas que amplificam sua expertise, permitindo diagnósticos mais rápidos e precisos, tratamentos mais personalizados e melhor qualidade de vida para pessoas com distúrbios da fluência.

Este trabalho é, portanto, não um fim, mas um começo, primeiro passo em uma jornada que promete transformar fundamentalmente como entendemos, diagnosticamos e tratamos os distúrbios da fala humana.

REFERÊNCIAS

ALHAKBANI, Y. *et al.* Automated stuttering detection using deep learning technologies. [**Revista ou Conferência não especificada**], 2025.

ALNASHWAN, A. *et al.* Computational intelligence-based stuttering detection: A systematic review. **Diagnostics**, v. 13, n. 21, p. 3537, 2023.

ANDRADE, C. **Fundamentos da Fonoaudiologia**. São Paulo: Editora Fonoaudiológica, 2004.

ANDRADE, C. R. F. Fluência. *In*: ANDRADE, C. R. F. (ed.). **Fonoaudiologia na infância**. São Paulo: Lovise, 2006.

ARACHCHI, T. K. *et al.* Stutterai: Virtual assistant for stutter detection. [**Conferência ou Revista não especificada**], 2023.

BARRETT, D. *et al.* Systematic review of machine learning approaches for detecting developmental stuttering. [**Revista ou Conferência não especificada**], 2022.

BLEEK, W.; OUTROS. **Título do Livro**. Cidade: Editora, 2012.

CALLAN, D. E. *et al.* Self-organizing map for classification of normal and pathological voice. **Journal of Voice**, v. 13, n. 3, p. 365–375, 1999.

CASELI, H.; NUNES, M. **Introdução ao Processamento de Linguagem Natural**. São Paulo: Editora de Tecnologia, 2024.

CIVIER, O.; OUTROS. Título do artigo. **Nome do Jornal**, X, n. Y, p. 123–145, 2013.

FERREIRA, R. e. a. Estudos sobre fonoaudiologia clínica. **Revista Brasileira de Fonoaudiologia**, v. 12, n. 3, p. 210–225, 2010.

GARGANTINI, L. M. **Fluência da fala**. São Paulo: Plexus, 1995.

KOURKOUNAKIS, T. *et al.* Fluentnet: End-to-end detection of speech disfluency. **arXiv preprint arXiv:2009.11394**, 2020.

LOPES, J. e. a. Softwares na avaliação da fluência. **Revista de Ciências da Saúde**, v. 19, n. 4, p. 310–325, 2019.

LOU, P. C. *et al.* Disfluency detection using auto-correlational neural networks. **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, p. 935–941, 2018.

MARTINS, A.; ANDRADE, C. **Novas Tecnologias na Fonoaudiologia**. Rio de Janeiro: Editora Acadêmica, 2008.

PERKINS, W. e. a. **A Gagueira e suas Manifestações**. Porto Alegre: Editora Científica, 1991.

PERNAMBUCO, L. e. a. Inovações na fonoaudiologia. **Jornal da Fonoaudiologia**, v. 15, n. 2, p. 89–100, 2015.

RAMITHA, S. *et al.* Evaluative comparison of machine learning algorithms for stuttering detection. [**Revista ou Conferência não especificada**], 2024.

SHEIKH, S. A. *et al.* Machine learning for stuttering identification: A comprehensive review. **IEEE Access**, v. 9, p. 6341–6360, 2021.

SISSKIN, V.; WASILUS, S. Nome do artigo. **Revista Científica da Área**, X, n. Y, p. 100–120, 2014.

STARKWEATHER, C. W.; GIVENS-ACKERMAN, J. **Stuttering**. Austin: Pro-Ed, 1997.

SZCZUROWSKA, I. *et al.* Application of artificial neural networks for classification of speech disfluencies. **Archives of Acoustics**, v. 31, n. 2, p. 203–210, 2006.

YAIRI, E.; AMBROSE, N. **Título do Livro sobre Gagueira**. Cidade: Editora, 1992.

YAIRI, E.; AMBROSE, N. **Outro Título Relacionado**. Cidade: Editora, 1999.

ŚWIETLICKA, I. *et al.* Artificial neural networks in the disabled speech analysis. **Archives of Acoustics**, v. 38, n. 3, p. 321–328, 2013.