Florescer entre Sons e Silêncios

Ferramenta de Apoio à Classificação de Vocalizações Não Verbais com Inteligência Artificial

Fernanda Floriano Silva
fernanda.floriano@usp.br
ICMC-USP
São Carlos, SP

Alessandra Alaniz Macedo*
ale.alaniz@usp.br
FFCLRP-USP
Ribeirão Preto, SP

ABSTRACT

Communication, a fundamental human right, can be compromised in neurodevelopmental disorders, such as in nonverbal children with Autism Spectrum Disorder, whose vocalizations often lack intelligibility. This study explores how artificial intelligence can support phonological analysis in this context. A Brazilian Portuguese dataset was built, combining reference phonemes — processed with acoustic feature extraction and data augmentation — and vocalizations from a nonverbal child after preprocessing. Unsupervised methods revealed consistent phonological approximations, particularly in nasal categories. In the supervised analysis, samples were represented through Bag-of-Audio-Words (BoAW) combined with acoustic features, and class imbalance was addressed using SMOTE. The evaluated models included KNN, RF, MLP, SVM, and CNN. Results showed that SVM achieved the best performance in terms of phonetic/articulatory equivalences, RF demonstrated robustness in unbalanced scenarios, and CNN reached high accuracy on the validation set. Comparison with perceptual-auditory analyses by speech therapists confirmed relevant convergences. These findings highlight the feasibility of computational models as complementary resources to clinical listening, supporting therapeutic interventions and the development of child speech.

KEYWORDS

artificial intelligence, audio cluster, bag of audio words, convolutional neural networks, support vector machine, assistive tool

1 INTRODUÇÃO

A comunicação é essencial para o desenvolvimento infantil, mas pode ser comprometida por transtornos do neurodesenvolvimento, como o Transtorno do Espectro Autista (TEA), que afeta a interação social e a linguagem. Estima-se que 70 milhões de pessoas tenham TEA no mundo e, no Brasil, o Censo 2022 identificou 2,4 milhões de casos (1,2% da população) [9]. Aproximadamente 20–30% dessas crianças apresentam comprometimento na fala, abrangendo quadros não verbais ou com dificuldades significativas [5, 17, 20].

Nesse contexto, surgem produções classificadas como *fala atípica*, isto é, emissões vocais que se afastam dos padrões típicos de fluência e articulação [2], entre as quais se destaca a *fala não verbal* — sons que não formam palavras compreensíveis [1].

Sistemas de reconhecimento de fala, como o *Whisper*, da *OpenAI* [16], e o *Google Speech-to-Text* [8], embora treinados para ampla cobertura de domínios, mostram-se pouco eficazes diante da fala atípica. Sem adaptação de dados, frequentemente classificam vocalizações de crianças não verbais como inaudíveis ou incompreensíveis [18]. Essa limitação reduz seu potencial como recurso terapêutico.

Este artigo apresenta um estudo exploratório sobre o uso de métodos de inteligência artificial (IA) na análise de vocalizações de crianças não verbais. O trabalho contemplou o desenvolvimento do fluxo de processamento (*backend*), incluindo a criação do conjunto de dados, a aplicação de técnicas de aprendizado de máquina (AM) e experimentos de classificação, que constituem a base funcional do sistema. Como resultado preliminar, elaborou-se também um protótipo de interface gráfica, ainda em versão inicial, destinado a ilustrar como esses recursos poderão ser integrados em uma futura ferramenta computacional de apoio à prática clínica.

Busca-se, assim, oferecer suporte às práticas terapêuticas, com foco na detecção de sons pouco perceptíveis em terapia. Além disso, pretende-se contribuir para o avanço acadêmico e social, em direção a estratégias acessíveis que ampliem a autonomia comunicativa e favoreçam a inclusão de crianças com dificuldades de fala.

2 TRABALHOS RELACIONADOS

A comunicação é um direito humano fundamental, garantido pela Convenção sobre os Direitos das Pessoas com Deficiência, da **ONU** [6], e pela **Lei Brasileira de Inclusão** (Lei nº 13.146/2015) [4], que asseguram acesso à educação inclusiva e às tecnologias assistivas. Também a Lei nº 10.098/2000 [3] prevê a eliminação de barreiras comunicacionais por meio de sistemas alternativos e aumentativos, reforçando o papel central das tecnologias assistivas nesse processo.

Com os avanços digitais, a Comunicação Aumentativa e Alternativa (CAA) incorporou soluções de IA para reconhecimento de padrões acústicos e gestuais [10, 13, 14]. O *EcoScript* fornece *feedback* em tempo real e pode se adaptar a vocalizações atípicas, enquanto o *SofiaFala*, no Brasil, combina AM e princípios da fonoaudiologia para respostas multimodais personalizadas [11].

No campo do reconhecimento de fala atípica, um dos estudos, de Deller et al. [7], analisou fala disártrica associada à paralisia cerebral e propôs abordagens específicas baseadas na consistência de vogais, evidenciando a necessidade de adaptar métodos de reconhecimento às particularidades acústicas desse tipo de produção. Em crianças com TEA, Mohanta e Mittal [13] alcançaram 97% de acurácia com SVM e KNN a partir de formantes e frequência fundamental, demonstrando o potencial de algoritmos clássicos quando aplicados a atributos acústicos cuidadosamente selecionados. Modelos neurais personalizados também têm mostrado resultados promissores em bases reduzidas, como no estudo de Mulfari et al. [15],

*Advisor.

In: XXIV Workshop de Ferramentas e Aplicações (WFA 2025). Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web (WFA'2025). Rio de Janeiro/RJ, Brasil. Porto Alegre: Brazilian Computer Society, 2025.
© 2025 SBC – Sociedade Brasileira de Computação.
ISSN 2596-1683

que explorou métodos dependentes do falante para fala disártrica, destacando a relevância de soluções adaptativas. Já McGonigle et al. [12] avaliaram o desempenho do *Whisper* no reconhecimento da fala de crianças com e sem atrasos de desenvolvimento, evidenciando limitações de sistemas genéricos e reforçando a necessidade de ajustes específicos para o contexto clínico.

Apesar dos avanços, ainda são raras as soluções para o reconhecimento de vocalizações pré-linguísticas, especialmente em crianças não verbais e no contexto do português brasileiro. É necessário desenvolver modelos que conciliem a singularidade de cada criança com padrões acústicos fundamentais da produção de fala, criando bases mais sólidas para aplicações clínicas e tecnológicas.

3 PROVA DE CONCEITO EXPERIMENTAL

3.1 Configuração

Nesta pesquisa, foi conduzido um estudo exploratório para avaliar como técnicas de AM podem apoiar a identificação de padrões fonológicos em vocalizações não verbais. O corpus analisado reuniu 29 trechos de uma criança não verbal, previamente segmentados, além de um conjunto de fonemas do português brasileiro com características acústicas extraídas e aumentadas.

A arquitetura (Figura 1) foi organizada em módulos sequenciais, sendo eles: (a) extração de atributos acústicos e aumento de dados; (b) pré-processamento e organização das amostras infantis; (c) exploração e visualização não supervisionada; (d) classificação supervisionada, incluindo o uso de SMOTE e, em alguns casos, representações avançadas com *Bag of Audio Words* (BoAW); e uma CNN baseada diretamente em atributos acústicos. Essa arquitetura modular permite comparar diferentes abordagens, favorece a reprodutibilidade e pode ser reutilizada em pesquisas futuras no reconhecimento de fala atípica e em outros cenários clínicos.

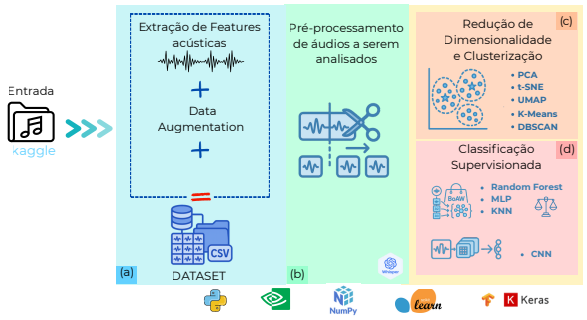

Figure 1: Arquitetura geral da proposta.

3.2 Dataset de Fonemas

O estudo utilizou como base o *Brazilian Portuguese Phonemes – Audio*[1], composto por 31 fonemas organizados em classes fonológicas segundo o ponto de articulação (consoantes oclusivas, fricativas, laterais, nasais e vogais). O fonema foi adotado como unidade de análise por ser a menor forma distintiva da língua [19].

De forma complementar, foram incorporadas vocalizações reais da criança não verbal, processadas em um módulo desenvolvido neste estudo. Os áudios foram isolados por meio de *voice activity detection* (VAD), seguidos de ajuste manual de limites e rótulos (`Criança`, `Terapeuta`). Esse processo resultou em 29 trechos rotulados como `Criança`, que passaram a compor o conjunto de dados para as análises comparativas.[2]

As amostras — tanto os fonemas quanto as vocalizações infantis — foram representadas por atributos acústicos que registram propriedades objetivas do sinal: medidas gerais (duração, intensidade média, frequência dominante) e parâmetros específicos de análise (13 coeficientes MFCC, taxa de cruzamento por zero, energia RMS, centróide e largura de banda espectrais, além dos formantes F1–F3). Esses atributos foram organizados em tabelas `.csv` e normalizados para garantir comparabilidade entre classes.

Para ampliar a representatividade e reduzir desbalanceamentos, aplicaram-se técnicas de aumento de dados (*pitch shifting*, *time-stretching* e inserção de ruído controlado). Com isso, o conjunto inicial de 59 arquivos foi expandido para aproximadamente 600 registros, distribuídos entre as diferentes classes fonológicas. O módulo de construção foi projetado para atualização contínua, de modo que novos áudios podem ser incorporados, processados e rotulados automaticamente, favorecendo a personalização conforme o desenvolvimento comunicativo de cada criança.

3.3 Métodos de Aprendizado de Máquina (AM)

3.3.1 ***Métodos Não Supervisionados.*** Para exploração inicial do espaço fonológico, foram empregados métodos de aprendizado não supervisionado. A etapa de visualização contou com técnicas de redução de dimensionalidade — PCA, t-SNE e UMAP —, que projetaram os vetores de características acústicas em duas dimensões, permitindo inspecionar proximidades entre fonemas e vocalizações. Na etapa de clusterização, aplicaram-se K-Means e DBSCAN, com o objetivo de identificar agrupamentos formados a partir de similaridades acústicas.

3.3.2 ***Métodos Supervisionados.*** Foram conduzidos experimentos supervisionados para avaliar a classificação automática de vocalizações em diferentes categorias fonológicas. Os classificadores foram treinados com amostras rotuladas do conjunto de fonemas, aplicando-se a técnica *SMOTE* ao conjunto de treino para mitigar o desbalanceamento entre classes. As representações foram geradas pelo modelo *Bag-of-Audio-Words* (*BoAW*), construído a partir de coeficientes MFCC quantizados em 40 palavras acústicas e combinados às demais características do conjunto. O resultado foi um vetor fixo por amostra, utilizado como entrada nos classificadores.

Foram avaliados diferentes modelos supervisionados: - SVM : adequado para lidar com fronteiras não lineares; - KNN: abordagem simples e interpretável, baseada em proximidade; - Random Forest (RF): robusto a heterogeneidades do espaço acústico; - MLP: capaz de capturar padrões não lineares em representações tabulares; - CNN: estruturada diretamente sobre atributos acústicos, com camadas convolucionais e regularização por *dropout* e *early stopping*.

[1] https://www.kaggle.com/datasets/jonascarvalho/brazilian-portuguese-phonemes-audio Os dados da plataforma Kaggle tiveram o uso previamente autorizado por e-mail pelos responsáveis pela disponibilização do conteúdo.

[2] Os áudios da criança não verbal foram produzidos e autorizados pelos pais da criança, os quais estão conduzindo a pesquisa e providenciando o cadastro da pesquisa na Plataforma Brasil.

3.4 Ambiente e Tecnologias

Os experimentos foram conduzidos em Python, no ambiente Google Colab com suporte a GPU. Utilizaram-se bibliotecas especializadas para diferentes etapas do fluxo: processamento de áudio (*Librosa*, *SoundFile*, *FFmpeg*), manipulação e análise de dados (*NumPy*, *Pandas*, *SciPy*) e modelagem de AM (*scikit-learn*, *TensorFlow/Keras*, *umap-learn*). A reprodutibilidade foi assegurada pela definição de *seeds* fixas e pela implementação de `pipelines` que integraram normalização dos atributos, balanceamento de classes e execução padronizada dos modelos.

4 AVALIAÇÃO EXPERIMENTAL

Foram conduzidos experimentos exploratórios para avaliar em que medida métodos computacionais poderiam apoiar a escuta clínica de vocalizações atípicas. Utilizaram-se o conjunto de fonemas descrito na metodologia e 29 trechos de uma criança não verbal, processados por algoritmos de agrupamento e aprendizado supervisionado. Os resultados foram então comparados com a análise perceptivo-auditiva conduzida por fonoaudiólogas experientes[3].

Nos métodos não supervisionados, cada técnica ofereceu uma perspectiva distinta. O PCA e o t-SNE aproximaram a vocalização da criança de vogais nasais. O UMAP indicou maior proximidade com consoantes nasais. O K-Means formou agrupamentos regulares, também com destaque para vogais nasais. O DBSCAN, embora tenha classificado alguns pontos como ruído, manteve a aproximação com nasais. No conjunto, os métodos mostraram que as vocalizações tenderam a se organizar próximas a categorias fonológicas específicas, em especial as nasais, correspondentes às produções mais consistentes da criança (Figura 2).

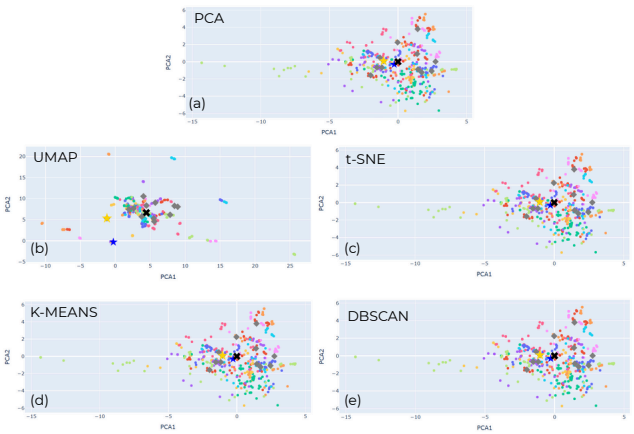

Figure 2: Projeções do espaço fonológico obtidas por diferentes métodos: (a) PCA, (b) UMAP, (c) t-SNE (visualização) e (d) K-Means, (e) DBSCAN (clusterização).

Na etapa supervisionada, aplicaram-se SVM, KNN, Random Forest, MLP e CNN com representações BoAW e atributos acústicos. O KNN mostrou maior sensibilidade a ruídos, mas sua revocação aumentou de 0,71 para 0,83 após a aplicação do SMOTE. O MLP apresentou desempenho intermediário, alcançando 0,89 de precisão e 0,88 de F1-score. O Random Forest manteve métricas estáveis, em torno de 0,89–0,90 de acurácia, mostrando-se robusto mesmo em bases reduzidas. Esses resultados estão resumidos no gráfico radar (Figura 3).

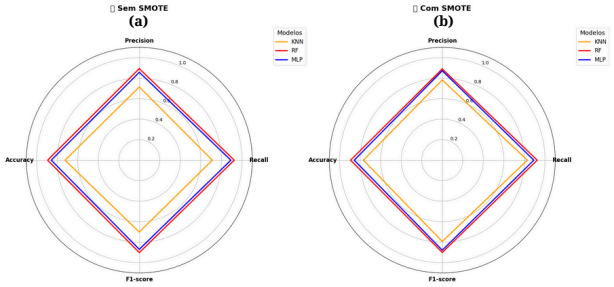

Figure 3: Gráfico radar comparando métricas de desempenho (acurácia, precisão, revocação e F1-score) dos modelos supervisionados, com e sem SMOTE.

Entre os modelos testados, a CNN destacou-se como a abordagem mais promissora, atingindo acurácia entre 0,87 e 0,90 no conjunto de validação. Suas curvas de treino indicaram crescimento contínuo até estabilizar após a 20ª época, enquanto a perda decresceu de forma consistente; como a acurácia se manteve estável e a perda não voltou a subir, não houve sinais de *overfitting*. A CNN mostrou-se bem ajustada e capaz de generalizar adequadamente, reforçando seu potencial como solução eficaz para o reconhecimento de vocalizações atípicas (Figura 4).

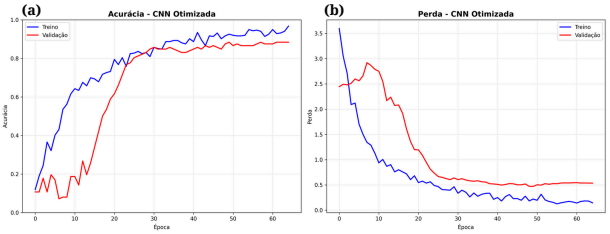

Figure 4: Evolução da acurácia e da perda durante o treinamento da CNN.

Na comparação com a análise perceptivo-auditiva, as fonoaudiólogas relataram convergência em classes como consoantes oclusivas e nasais, correspondentes às produções mais consistentes da criança. Por outro lado, observaram divergências em sons pouco inteligíveis, pois consideraram coarticulação e sons não linguísticos — aspectos não capturados pelos modelos computacionais, o que evidencia limitações na comparação entre análise humana e automática.

A Tabela 1 sintetiza, de forma quantitativa, os percentuais de acertos exatos e de equivalências fonético-articulatórias observadas.

[4]. Observa-se que o Random Forest obteve melhor desempenho em acertos exatos, enquanto o SVM foi superior em equivalências. A CNN, embora limitada pela quantidade de dados, demonstrou potencial de generalização semelhante aos demais.

[3]A análise perceptivo-auditiva consiste na avaliação de vocalizações realizada por meio da escuta clínica especializada, permitindo identificar aspectos como inteligibilidade, precisão articulatória, qualidade vocal e aproximações com categorias fonéticas [2].

[4]Equivalências fonético-articulatórias correspondem a casos em que a produção da criança, embora não coincida exatamente com o fonema esperado, é próxima em termos de ponto ou modo de articulação [2].

Table 1: Percentuais de acerto por método, considerando a análise perceptivo-auditiva como padrão-ouro.

Método	Acerto Exato (%)	Equivalência fonética/articulatória (%)
Cluster	6,7	40,0
SVM	6,7	46,0
RF (sem SMOTE)	20,0	43,3
RF (com SMOTE)	20,0	43,3
CNN	3,0	43,3

5 CONCLUSÕES E TRABALHOS FUTUROS

Este estudo exploratório evidenciou a aplicabilidade de técnicas de IA para apoiar a análise de vocalizações atípicas em contextos de dados escassos. O fluxo integrado proposto, combinando extração de características, visualização e classificação, mostrou-se capaz de identificar aproximações fonético-articulatórias relevantes e de complementar a escuta clínica especializada. Em conjunto, os resultados quantitativos e as projeções não supervisionadas (Tabela 1; Figuras 3 e 4) confirmam que métodos de IA podem oferecer perspectivas complementares à análise clínica. Observou-se melhor desempenho em categorias associadas a funções articulatórias mais estáveis, sugerindo que, embora o desenvolvimento vocal inicial seja limitado em sons, a especialização progressiva do *dataset* tende a aprimorar a precisão dos modelos com a evolução infantil.

As principais contribuições incluem a demonstração da viabilidade de distinguir categorias fonéticas em vocalizações reais e a aproximação entre métodos computacionais e a prática clínica. Destaca-se o desenvolvimento de uma arquitetura modular, reutilizável e expansível, além da constatação de que cada criança apresenta padrões vocais próprios, reforçando a necessidade de conjuntos de dados especializados (*datasets*). Também se observou que fonoaudiólogas identificaram nuances de ressonância pouco perceptíveis sem apoio do sistema, evidenciando vantagens no uso combinado de análises humanas e computacionais.

As limitações concentram-se no número reduzido de participantes, restrito a uma única criança, na baixa inteligibilidade das vocalizações e nas incertezas de rotulação. Observa-se ainda a dependência do pré-processamento dos áudios infantis e as diferenças acústicas entre a base de referência (voz masculina adulta) e os áudios analisados (voz infantil feminina). Destaca-se que classificações incorretas não oferecem risco às crianças, já que o sistema não tem caráter diagnóstico, mas atua apenas como apoio clínico.

Como trabalhos futuros, prevê-se a ampliação e especialização da base de dados, com maior diversidade de vozes, coletas longitudinais e rotulação colaborativa junto a terapeutas, além da otimização de modelos promissores, como CNN e Random Forest. Pretende-se ainda desenvolver uma ferramenta capaz de capturar vocalizações em tempo real durante as sessões de terapia da fala e gerar relatórios automáticos com hipóteses de produções vocais, acompanhadas dos trechos correspondentes para escuta direta do terapeuta, bem como disponibilizar publicamente o *dataset*.

Por fim, este trabalho destaca que limitar a comunicação constitui uma forma de exclusão social. A tecnologia assistiva deve ser compreendida como instrumento de inclusão e como expressão do direito humano à comunicação, essencial para a participação plena na sociedade. Ao propor soluções que ampliam as possibilidades de expressão, o estudo contribui para enfrentar barreiras e promover avanços em tecnologias inclusivas.

REFERENCES

[1] American Psychiatric Association. 2023. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR* (5ª, texto revisado ed.). American Psychiatric Publishing, Washington, DC.

[2] Cristina R. F. Andrade. 2002. *Fonoaudiologia: uma abordagem educacional.* Lovise, São Paulo.

[3] Brasil. 2000. Lei n° 10.098, de 19 de dezembro de 2000. http://www.planalto.gov.br/ccivil_03/leis/L10098.htm. Estabelece normas gerais e critérios básicos para a promoção da acessibilidade das pessoas com deficiência ou com mobilidade reduzida.

[4] Brasil. 2015. Lei nº 13.146, de 6 de julho de 2015: Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência). http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13146.htm. Acessado em 9 jul. 2025.

[5] Centers for Disease Control and Prevention. 2025. CDC - Centers for Disease Control and Prevention. https://www.cdc.gov/ Acesso em: 05-07-2025.

[6] Organização das Nações Unidas. 2006. Convenção sobre os Direitos das Pessoas com Deficiência. https://www.ohchr.org/pt/instruments-mechanisms/instruments/convention-rights-persons-disabilities. Acessado em 9 jul. 2025.

[7] J. Deller, D. Hsu, and L. Ferrier. 1987. Recognition of Cerebral Palsy Speech: Technical Method and a Study of Vowel Consistency. In *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 12. 1461–1464. https://doi.org/10.1109/ICASSP.1987.1169507

[8] Google Cloud. 2024. Speech-to-Text Documentation. https://cloud.google.com/speech-to-text Acesso em: 05-07-2025.

[9] Instituto Brasileiro de Geografia e Estatística (IBGE). 2025. *Censo Demográfico 2022 identifica 2,4 milhões de pessoas diagnosticadas com autismo no Brasil.* IBGE. https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/43464-censo-2022-identifica-2-4-milhoes-de-pessoas-diagnosticadas-com-autismo-no-brasil Acessado em 7 de julho de 2025.

[10] Eunyeoul Lee, Eunseo Yang, Jinyoung Huh, and Uran Oh. 2024. EcoScript: A Real-Time Presentation Supporting Tool using a Speech Recognition Model. In *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI).* 96–101. https://doi.org/10.1109/IRI62200.2024.00031

[11] Alessandra Alaniz Macedo, Vinícius de S. Gonçalves, Patrícia P. Mandrá, Vivian Motti, Renato F. Bulcão-Neto, and Kamila Rios da Hora Rodrigues. 2024. A mobile application and system architecture for online speech training in Portuguese: design, development, and evaluation of SofiaFala. *Multimedia Tools and Applications* (aug 2024). https://doi.org/10.1007/s11042-024-19980-5 Acesso em: 05-07-2025.

[12] Michelle McGonigle et al. 2024. Evaluating Whisper ASR on the Speech of Children With and Without Developmental Delays. *Journal of Speech, Language, and Hearing Research* (2024).

[13] Abhijit Mohanta and Vinay Kumar Mittal. 2022. Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features. *Computer Speech & Language* 72 (2022), 101287. https://doi.org/10.1016/j.csl.2021.101287

[14] Davide Mulfari, Antonio Celesti, and Massimo Villari. 2021. Deep Learning Applications in Telerehabilitation Speech Therapy Scenarios. *Applied Sciences* 11, 3 (2021), 1177.

[15] Davide Mulfari, Antonio Celesti, and Massimo Villari. 2022. Exploring AI-based Speaker Dependent Methods in Dysarthric Speech Recognition. In *Proceedings of the 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid).* 958–964. https://doi.org/10.1109/CCGrid54584.2022.00117

[16] OpenAI. 2022. Introducing Whisper. https://openai.com/index/whisper/ Acesso em: 13-03-2025.

[17] Organização Pan-Americana da Saúde. 2020. Transtorno do espectro autista – OPAS/OMS. https://www.paho.org/pt/topicos/transtorno-do-espectro-autista Acesso em: 04-03-2025.

[18] Helen Tager-Flusberg. 2005. Language and Communication in Autism. In *Handbook of Autism and Pervasive Developmental Disorders*, Fred R. Volkmar (Ed.). Wiley. Disponível em: https://www.academia.edu/72826966/Language_and_communication_in_autism.

[19] N. S. Trubetzkoy. 1969. *Principles of Phonology.* University of California Press, Berkeley. Translation of *Grundzüge der Phonologie* (1939) by C. A. M. Baltaxe.

[20] World Health Organization. 2023. Autism spectrum disorders. https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders Accessed: 04-2025.