Detecting Stuttering with Artificial Intelligence: A Hybrid Method for Brazilian Portuguese

Rubens Perini Buzzeti
GFT Technologies Brasil
Barueri, São Paulo, Brazil
rubensperini@yahoo.com.br

Paula Bianca Meireles de Moura Buzzeti
Mais Fono Fonoaudiologia
Marília, São Paulo, Brazil
paula.moura@paulamourafono.com.br

Roney Lira de Sales Santos
Federal University of Bahia (UFBA)
Camaçari, Bahia, Brazil
roneysantos@ufba.br

Abstract

Speech-language assessment of stuttering is traditionally manual, subjective, and time-consuming. This paper presents the development of software for automatic detection and classification of stuttering-related disfluencies in Brazilian Portuguese, aiming to support clinical assessment. The system follows a two-stage hybrid approach. In the first stage, it applies deterministic algorithms based on automatic speech recognition (ASR) and temporal information to identify simple disfluencies, such as repetitions and pauses. In the second stage, it employs a hierarchical architecture combining a Kohonen network (Self-Organizing Map, SOM) and a Multilayer Perceptron (MLP) to classify complex disfluencies, specifically blocks and prolongations, using acoustic features. Because no publicly available annotated resources exist for this task in Brazilian Portuguese, we built a initial dataset annotated by specialists. The system achieved 89.5% accuracy in classifying complex disfluencies, with a Matthews Correlation Coeficient (MCC) of 0.812. These results indicate the feasibility of the tool as decision support for clinical assessment and establish a baseline for future research.

1 Introduction

Stuttering is a fluency disorder that affects approximately 1% of the world's population (Sheikh et al., 2021). Its diagnosis is a complex clinical process that is traditionally manual, subjective, and demanding in terms of time and professional effort (Alnashwan et al., 2023). Speech-language assessment requires detailed analysis of speech samples that typically contain at least 200 fluent syllables to identify and quantify different types of disfluencies (de Andrade, 2006; Sawyer and Yairi, 2006). In this context, the presence of 3% or more stuttering-like disfluencies (SLDs) is commonly used as a key criterion for clinical diagnosis (Yairi and Ambrose, 1999, 2005).

Although the assessment process is guided by validated clinical protocols, it remains resource-intensive and time-consuming. This reinforces the need for technological tools that can support speech-language pathologists and reduce the time spent analyzing speech samples (Sheikh et al., 2021). This demand is further intensified by the growing adoption of digital health technologies, the potential of machine learning methods in fluency research (Alnashwan et al., 2023), and the expansion of telemedicine, which increases the need for automated methods to support assessment and clinical decision-making.

Within this context, this paper presents an artificial intelligence-based method to automate the detection, organization, and categorization of stuttering disfluencies in Brazilian Portuguese, serving as a tool to support clinical assessment. Due to the scarcity of public datasets focused on stuttering in this language, we built a dataset containing 126 samples of fluent speech, blocks, and prolongations. The proposed method adopts a hybrid approach that combines deterministic algorithms for identifying simple disfluencies, such as repetitions, with a hierarchical architecture composed of a Kohonen network, also known as a Self-Organizing Map (SOM), and a Multilayer Perceptron (MLP) for classifying complex disfluencies, specifically blocks and prolongations, following methodologies previously validated in the literature (Szczurowska et al., 2006; Świetlicka et al., 2013).

The main contributions of this paper are: (1) a hybrid software architecture for the automatic detection of stuttering disfluencies in Brazilian

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 507–515

Portuguese; (2) the creation and validation of a dataset, albeit limited, aimed at classifying complex disfluencies; and (3) the experimental validation of an acoustic feature extraction methodology (Szczurowska et al., 2006) that achieved 89.5% accuracy, establishing a baseline for future research on automatic stuttering detection.

This paper is organized as follows. Section 2 presents related work, contextualizing the main approaches for automatic stuttering detection. Section 3 describes the proposed methodology, including the system's hybrid architecture and the dataset construction process. Section 4 details the experimental protocol and the results obtained across different evaluation iterations. Section 5 presents the developed interface as a clinical assessment support system, demonstrating the practical applicability of the proposed methods. Section 6 discusses the results in light of the literature and the study's limitations. Finally, Section 7 presents the conclusions and directions for future work.

2 Related Works

Automatic stuttering detection is a growing research area at the intersection of Speech-Language Pathology and Artificial Intelligence. Despite recent progress, the inherent complexity of speech fluency phenomena poses significant challenges to reliable automation. This section reviews the main related studies that inform the methodological choices adopted in this work.

Recent systematic reviews, such as those conducted by Sheikh et al. (2021) and Alnashwan et al. (2023), provide a comprehensive overview of the state of the art in automatic stuttering detection. These studies show a consistent trend toward the use of machine learning and deep learning techniques, including artificial neural networks, support vector machines, convolutional neural networks, and recurrent neural networks. They also highlight the predominant use of acoustic features, particularly Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980). However, both reviews identify recurring bottlenecks, especially the lack of methodological standardization and the scarcity of high-quality public datasets, which are generally small and limited in diversity.

The choice of neural network architecture is central to the performance of disfluency detection systems. Although deep learning architectures have gained prominence more recently, seminal work proposed hierarchical approaches that remain relevant. In particular, Szczurowska et al. (2006) presented an architecture that combines an unsupervised Kohonen network with a Multilayer Perceptron (MLP). In this approach, the Kohonen network is used to reduce data dimensionality and extract topological patterns from the speech signal, while the vectors associated with the winning neurons serve as input to the MLP, which performs the final supervised classification. This methodology was later refined by Świetlicka et al. (2013) and demonstrated the ability to capture dynamic aspects of speech, serving as a direct reference for the model implemented in this work.

One of the main obstacles to progress in automatic stuttering detection is the limited availability of public datasets with accurate annotations and adequate acoustic quality. Most studies use widely known datasets such as UCLASS (Howell et al., 2009), one of the oldest and most commonly used resources in the field; FluencyBank (MacWhinney, 2000), primarily focused on language development; and LibriStutter (Cheang et al., 2020), a more recent dataset derived from LibriSpeech (Panayotov et al., 2015). Despite their relevance, these datasets have limitations related to size, diversity, and linguistic focus. Given these constraints and the specific needs of each investigation, building custom datasets, although costly, has become a recurring and often necessary practice, as in the present study.

Overall, the literature indicates that automatic stuttering detection is a promising field but still constrained by structural challenges. Hierarchical architectures combining Kohonen networks and MLPs remain valid and effective (Szczurowska et al., 2006), especially when paired with acoustic features that are sensitive to disfluency phenomena, such as one-third-octave filters. While recent studies highlight the potential of deep learning architectures, including convolutional and recurrent networks, the fundamental challenge remains the scarcity of large, diverse, and well-annotated datasets. In this context, this work is distinguished by its focus on Brazilian Portuguese, since most existing studies and datasets concentrate on English. Thus, the adopted methodology is proposed not only as a classification exercise but as a foundation for developing a tool to support clinical diagnosis in real-world settings.

3 Methodology

The methodology adopted in this work is structured around a hybrid architecture designed to address limited data availability in Brazilian Portuguese and to handle different types of speech disfluencies appropriately. The system operates through two complementary analysis streams: (i) a Natural Language Processing (NLP) approach to detect simple disfluencies, such as repetitions and pauses, and (ii) an Artificial Intelligence-based acoustic analysis to identify complex disfluencies, specifically blocks and prolongations. To enable the machine learning-based approach, an initial step was the construction of a training dataset, given the absence of public resources for this task in Brazilian Portuguese. Additionally, for comparative purposes and to establish a strong baseline, a Support Vector Machine (SVM) with an RBF kernel was implemented and evaluated alongside the proposed MLP architecture.

3.1 Data Collection and Preparation

Building the dataset was identified as one of the most critical factors for the project's success and focused on complex disfluencies, that is, blocks and prolongations, which cannot be reliably detected using textual information alone. The data collection and preparation process was carried out in multiple steps, as described below.

- **Manual collection**: Audio samples of spontaneous and elicited speech were collected from publicly available videos on the web, such as video-sharing platforms.
- **Classification and validation**: All collected material was analyzed by a speech-language pathologist specializing in fluency, who segmented the audio signals and manually labeled disfluency events.
- **Control samples**: For each word exhibiting a complex disfluency, a corresponding control sample of the same word produced fluently was added to the dataset, recorded under similar conditions.

As a result, we built a final dataset composed of 126 audio samples, all standardized in WAV format with a 16 kHz sampling rate. The dataset is imbalanced across classes, reflecting the natural occurrence of the observed phenomena in speech. Table 1 shows the distribution of samples by class.

Class	Number of Samples
Fluent speech (control)	63
Block	45
Prolongation	18
Total	126

Table 1: Distribution of samples in the complex-disfluency dataset

Dispite small, the dataset built in this work fulfills the purpose of of being a starting point for Brazilian Portuguese. It will be publicly released in a future version after the completion of the necessary ethical procedures, including review of privacy, consent, and responsible use of speech data. In addition, we plan continuous expansion and refinement of the dataset by incorporating new samples and increasing speaker diversity, with the goal of improving representativeness and usefulness for future research on automatic stuttering detection in Brazilian Portuguese.

3.2 Hybrid Architecture and Processing Pipeline

With the dataset prepared, we developed the complete solution pipeline, illustrated in Figure 1. The architecture is splitted into the following modules:

1. Main module to process the API requisitions.
2. Module of video to audio convertion.
3. Module of audio transcription
4. Module of disfluencies detection
5. Module of complex disfluencies detection using AI.
6. User Interface

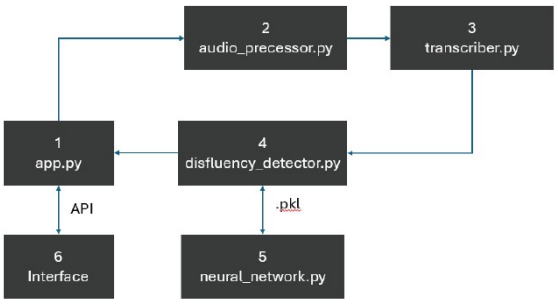

Figure 1: System Architecture.

The pipeline integrates the two proposed analysis streams and consists of the following steps.

1. **Preprocessing and transcription**: The process starts by extracting audio from the input video. The signal is then processed by an Automatic Speech Recognition (ASR) model, specifically a Wav2Vec model (Baevski et al., 2020) trained for Brazilian Portuguese, which generates a text transcription. In addition to the text, this stage provides precise temporal information associated with each word, which is essential for the NLP-based analysis.

2. **Text-pattern detection**: Using the transcription and the timestamps produced by the ASR, a deterministic text-pattern detection algorithm is applied to identify simple disfluencies. This rule-based linguistic and temporal analysis detects word or syllable repetitions, pauses defined by the duration of silence between consecutive segments, and hesitations such as interjections (e.g., "éh" and "aan").

3. **Acoustic analysis**: In parallel with the textual analysis, complex disfluencies, which are purely acoustic phenomena, are handled through audio feature extraction. Following the methodology proposed by Szczurowska et al. (2006), which proved more effective than generic approaches such as MFCCs in preliminary experiments, we extracted 21 spectral features based on one-third-octave filters with A-weighting. These filters cover the 80 Hz to 8000 Hz range, approximating human auditory sensitivity and providing a robust representation of the speech signal.

4. **SOM+MLP architecture**: The extracted acoustic features serve as input to a hierarchical artificial intelligence architecture. First, an unsupervised Kohonen network with a 5×5 topology performs a topological mapping of acoustic patterns. Next, the vectors associated with the winning SOM neurons are used as input to a Multilayer Perceptron (MLP) with 50 hidden-layer neurons, which performs the final supervised classification into three classes: fluent speech, block, and prolongation. The MLP was configured with activation = logistic, solver = lbfgs, maxIter = 200, learningRateInit= 0.01, and alpha = 0.001.

5. **Comparative SVC Model**: For baseline comparison, a Support Vector Machine (SVM) classifier was implemented. The SVM utilized a Radial Basis Function (RBF) kernel with default parameters (C = 1.0, gamma = scale) and probability = True, to ensure fair comparison with the MLP. This model was trained and evaluated on the same feature set derived from the Kohonen network output.

6. **Integration and inference**: After training, the SOM+MLP model is serialized and stored in a `.pkl` file. The backend system, implemented in Python, loads this model to perform acoustic inference. Finally, the outputs of the textual and acoustic analyses are integrated and exposed to the frontend through a REST API, which uses asynchronous processing mechanisms to manage execution time.

4 Experiments and Results

The experimental evaluation of the proposed system was conducted in four successive iterations, each incorporating incremental improvements to the architecture and model parameters. This strategy aimed to analyze, in a controlled manner, the impact of each modification on performance in classifying complex disfluencies.

Across all iterations, the model was trained and evaluated using the dataset described in Section 3.1. We report overall accuracy, Matthews Correlation Coefficient (MCC), and class-level metrics, including precision, recall, and F1-Score for fluent speech and block, as well as precision for prolongation, with particular attention to the effects of class imbalance. Additionally, the standard error (StdErr) of the accuracy is provided to indicate the statistical variability of the results.

First iteration (baseline). The first version of the system used a set of generic acoustic features commonly employed in speech processing. This configuration served as a baseline for comparison with subsequent iterations. The model achieved an overall accuracy of 0.545. Performance for the prolongation class was null, with precision equal to 0.000, highlighting the difficulty of recognizing the minority class under this configuration.

Second iteration (balancing). In the second iteration, we introduced class balancing techniques along with initial adjustments to network parameters. We applied SMOTE oversampling (Chawla et al., 2002) to mitigate class imbalance. This modification improved overall performance, increasing

accuracy to 0.637 and enabling recognition of prolongations with precision of 0.500.

Third iteration (parameter optimization). The third iteration focused on systematic parameter optimization using grid search and cross-validation. We tuned key hyperparameters of the Kohonen network, including learning rate and sigma, as well as MLP parameters such as the number of training epochs and the use of early stopping. These adjustments increased overall accuracy to 0.818 and improved performance on minority classes, reaching prolongation precision of 1.000.

Fourth iteration (literature-based methodology). The fourth iteration introduced structural changes grounded in the methodology proposed by Szczurowska et al. (2006). In this configuration, we used 21 spectral features based on one-third-octave filters, applying A-weighting to approximate human auditory sensitivity and covering the 80 Hz to 8000 Hz range. This approach yielded the best observed performance, with overall accuracy of 89.5%, MCC of 0.812, standard error (StdErr) of 0.0704, and more balanced results across classes, achieving prolongation precision of 0.780. Although the third iteration showed high values on specific metrics, the fourth iteration prioritized greater stability and balance across classes, reducing sensitivity to hyperparameter variation and class imbalance. This behavior suggests a trade-off between peak performance and generalization, with more consistent results in the final configuration.

Table 2 summarizes the evolution of overall accuracy and class-level metrics across the four iterations, enabling a comparative analysis of the impact of each modification introduced in the system.

It is important to note that recall values for the prolongation class are not reported, as this class is represented by a small number of samples ($n = 18$). Preliminary experiments showed that recall estimates for this class were highly unstable and sensitive to minor variations in the test split, which could lead to misleading interpretations.

To further validate the proposed SOM+MLP architecture, a Support Vector Machine (SVM) classifier with an RBF kernel was implemented as a baseline for comparison. Both models were evaluated on the same test set, and their performance is summarized in Table 3.

The results indicate that both the proposed MLP model and the SVM baseline achieved identical overall accuracy and MCC values. This suggests that the feature extraction process via the Kohonen network is highly effective, providing a robust representation of acoustic patterns that allows different classifiers to achieve high performance. The choice of MLP is further justified by its established use in similar hierarchical architectures in the literature.

5 Interface and Clinical Assessment Support System

To demonstrate the practical applicability of the proposed methods, we developed a graphical interface that integrates the entire pipeline described in the previous sections, from audio input to the presentation of fluency analysis results. The interface was designed as a clinical assessment support tool, enabling speech-language pathologists and researchers to inspect the results of automatic disfluency detection in a structured and interpretable manner.

The interface directly implements the hybrid architecture presented in this paper by combining NLP-based analysis for simple disfluencies with AI-based acoustic analysis for complex disfluencies. All processing is performed in the backend, while the interface serves as the visualization and interaction layer, consuming results through a REST API.

Figure 2 presents the initial interface stage[1], where the user uploads a video containing a speech sample. After submission, the system automatically extracts the audio and performs transcription using an ASR model. In addition to the transcribed text, the interface displays temporal information associated with the total audio duration, which supports detection of pauses and repetitions throughout the speech.

Figure 3 shows the output of the integrated textual and acoustic analyses[2]. The original transcript is enriched with explicit labels for automatically detected disfluencies, highlighted in the text. The interface also provides normalized transcript versions, including a disfluency-free form and a syllable-segmented form, facilitating further analyses and clinical interpretation.

Figure 4 presents a consolidated panel of quantitative metrics automatically extracted from the

[1]High-resolution version available at: https://tinyurl.com/mrxtawjw

[2]High-resolution version available at: https://tinyurl.com/mujjbkj3

Iteration	Accuracy	MCC	StdErr	Prec. Fluent	Prec. Block	Prec. Prol.	Rec. Fluent	Rec. Block	Rec. Prol.	F1 Fluent	F1 Block	F1 Prol.
1st	0.545	0.000	0.0000	0.602	0.556	0.000	0.500	0.833	0.000	0.546	0.667	0.000
2nd	0.637	0.000	0.0000	1.000	0.625	0.500	0.500	0.833	0.333	0.667	0.714	0.400
3rd	0.818	0.000	0.0000	1.000	0.750	1.000	0.500	1.000	0.667	0.667	0.857	0.800
4th (Final)	**0.895**	**0.812**	**0.0704**	**0.846**	**1.000**	**1.000**	**1.000**	**0.833**	**0.500**	**0.917**	**0.909**	**0.667**

Table 2: Performance comparison across iterations

Model	Accuracy	MCC	StdErr
MLP (Proposed)	**0.895**	**0.812**	**0.0704**
SVM (Baseline)	0.895	0.812	-

Table 3: Comparative performance of MLP and SVM models

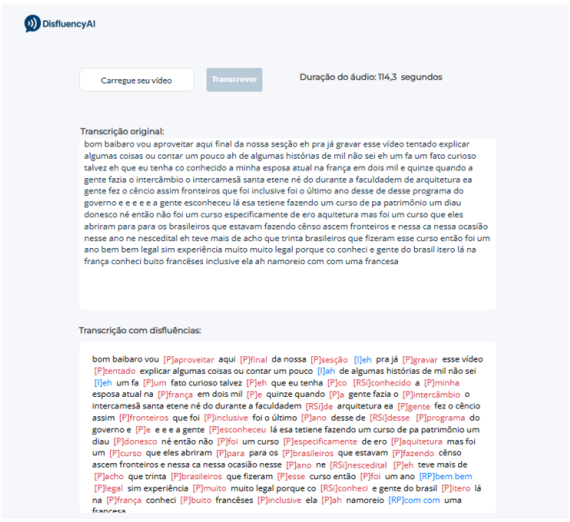

Figure 2: Interface for video upload and automatic speech transcription.

speech analysis[3]. The displayed indicators include the number of fluent words and syllables, production rates per minute, counts of stuttering like disfluencies, and rupture frequency. Based on these indicators, the interface provides an automatic assessment suggestion, which should be interpreted as decision support rather than a replacement for specialized speech-language assessment.

The interface is currently a functional prototype and was used for internal validation of the proposed methods. A public version for testing will be released soon, allowing researchers and practitioners to evaluate the system under controlled scenarios. This release aims to support reproducibility and contribute to the development of computational tools for stuttering assessment in Brazilian Portuguese.

[3]High-resolution version available at: https://tinyurl.com/4m6sxvsk

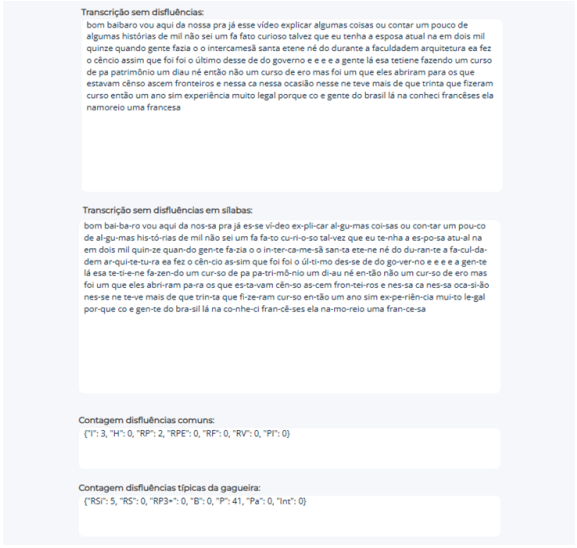

Figure 3: Output of the integrated disfluency analysis in the interface.

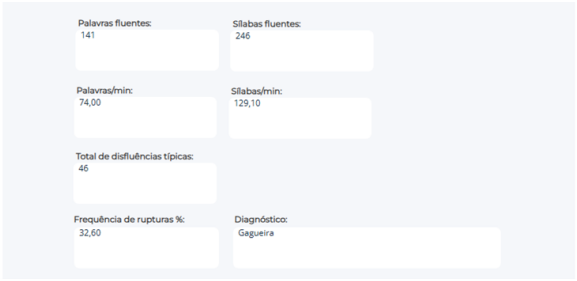

Figure 4: Panel of quantitative metrics automatically extracted from speech analysis.

6 Discussion and Limitations

6.1 Discussion of Results

The iterative evolution of the system highlights the value of an incremental, data-driven approach in developing machine learning models for automatic stuttering detection. The first iteration, which achieved an accuracy of 0.545, served as a baseline by exposing structural limitations, particularly the impact of class imbalance and the inadequacy of generic acoustic features for distinguishing complex disfluencies.

The introduction of balancing techniques in the second iteration empirically confirmed that dataset

imbalance was a key factor affecting classifier performance. Oversampling increased overall accuracy to 0.637 and, for the first time, enabled detection of the minority prolongation class. These results indicate that, even in low-data settings, balancing strategies can substantially improve model sensitivity.

In the third iteration, systematic optimization of Kohonen network and MLP hyperparameters yielded further gains, increasing accuracy to 0.818. These results show that careful tuning of architecture and training procedures can improve discriminative capacity. However, the observed instability in the metrics for minority classes suggested generalization limitations, consistent with potential overfitting under low-data conditions.

The fourth and final iteration represented the most meaningful advancement, achieving an accuracy of 0.895 and more balanced performance across classes. In particular, prolongation precision reached 0.780, reinforcing that adopting literature-grounded acoustic features can decisively improve the model. The high MCC value (0.812) indicates a strong correlation between predictions and ground truth, confirming that the model's performance is not due to chance despite the class imbalance. The identical performance of MLP and SVM (89.5%) suggests that the feature extraction via Kohonen maps is effective in linearizing the acoustic patterns of disfluencies. The standard error (StdErr) of 0.0704 further supports the statistical reliability of these results.

Analyzing the class-wise performance, the system demonstrated excellent performance for the "Fluent" class (F1-Score of 0.917) and "Block" class (F1-Score of 0.909), with the "Block" class achieving a perfect precision of 1.000. The "Prolongation" class, historically the most challenging due to its limited sample size, showed an F1-Score of 0.667, with a precision of 1.000 and a recall of 0.500. This indicates that while the model is highly precise when it identifies a prolongation, it still misses some instances. This behavior suggests a trade-off between peak performance and generalization, with more consistent results in the final configuration, but also highlights the need for further data collection for minority classes.

Beyond quantitative evaluation, we incorporated the results into a functional graphical interface that implements the full pipeline proposed in this paper. This interface enabled verification, in an applied setting, of consistency between model decisions, automatically labeled disfluencies in the transcript, and speech-derived metrics, reinforcing the practical applicability of the approach as a clinical support tool.

In addition, the use of the system through a graphical interface highlighted its potential for real-world contexts by presenting results in an interpretable manner in accordance with existing clinical protocols. This integration of quantitative performance with structured visualization helps bridge the gap between computational methods and speech-language practice, which is essential to adopt automatic tools in assessing stuttering.

6.2 Study Limitations

Despite promising results, some limitations should be considered when interpreting the findings of this work:

- **Dataset size and diversity**: There is no open dataset for Brazilian Portuguese, the only way to accomplish this work was to develop a new one from scratch. Considering the timeline, the most feasiable solution produced a dataset containing 126 audio samples, which is limited for robust training of machine learning models. In addition, collecting data from public sources may introduce biases and may not adequately represent the diversity of the population of people who stutter.

- **Risk of overfitting**: Although strategies such as cross-validation and early stopping were employed, the small dataset size maintains a risk of overfitting, especially for minority classes. In such cases, the model may learn patterns specific to the training set that do not generalize well to new data.

- **Generalization capacity**: The system was validated using a test set drawn from the same distribution as the training data. Additional evaluations on external datasets with greater variability in speakers and acoustic conditions are needed to more rigorously assess generalization.

In addition, during system development we identified technical challenges related to infrastructure and solution architecture. These included frontend response-time limitations due to the computational cost of acoustic inference, as well as incompatibilities across backend library versions. Further-

more, calibrating thresholds for disfluency detection across different speech samples proved to be a complex and iterative process, requiring a careful balance between sensitivity and specificity.

7 Conclusion

This paper presented the first trial of development and evaluation of a hybrid system for automatic detection of speech disfluencies, with a focus on stuttering in Brazilian Portuguese. The proposed approach is a prototype that integrates deterministic text-based methods to identify simple disfluencies, such as repetitions and pauses, with a hierarchical artificial intelligence architecture composed of a Kohonen network and a Multilayer Perceptron, aimed at classifying complex disfluencies, specifically blocks and prolongations.

The experimental results demonstrate the feasibility of the proposed solution, which achieved an overall accuracy of 0.895 and a Matthews Correlation Coefficient (MCC) of 0.812 in the final configuration. The system's iterative evolution, starting from a baseline with limited performance, highlights the importance of class balancing, appropriate acoustic feature selection, and literature-grounded methodologies in developing machine learning models for complex speech phenomena. The F1-Scores of 0.917 for fluent speech, 0.909 for blocks, and 0.667 for prolongations further underscore the model's effectiveness across different disfluency types.

The main contributions of this work include: (i) a low-cost hybrid software architecture designed to support speech-language assessment of stuttering; (ii) the integration of textual and acoustic analyses into a single system, enabling a more comprehensive disfluency assessment; and (iii) the validation of an incremental, data-driven development methodology that can be adapted to other problems in speech processing. Despite limitations related to dataset size and generalization challenges, the results indicate that the system is a meaningful step toward computational tools that support clinical stuttering diagnosis.

As future work, we plan to expand and refine the Brazilian Portuguese stuttering dataset by incorporating more speakers, broader age ranges, and diverse communicative contexts. This expansion will strictly follow ethical protocols and will enable more robust generalization assessments. In addition, we intend to conduct a systematic clinical validation of the system in partnership with speech-language pathologists, comparing tool outputs against human assessments in real diagnostic scenarios.

Another promising direction is investigating deep learning architectures, such as convolutional and recurrent neural networks, which can model the temporal dynamics of the speech signal more directly. Such approaches have shown strong results in speech recognition and analysis tasks (Graves et al., 2013; Hannun et al., 2014), but they require large volumes of annotated data, reinforcing the importance of continued efforts to build suitable datasets. A systematic comparison between these architectures and the proposed hierarchical model may provide further evidence on trade-offs among complexity, interpretability, and performance in clinical applications.

The software implementation developed in this work was created in collaboration with an industry partner and constitutes proprietary intellectual property. For this reason, the source code cannot be publicly released. The dataset used in the experiments is also owned by the company and is therefore not currently available for public distribution. However, the dataset is being continuously expanded and curated as part of an ongoing research and development effort. Future studies may release an extended and properly anonymized version of the dataset, subject to the company's approval and applicable data governance policies.

Acknowledgments

The authors would like to thank the University of São Paulo (USP) and the MBA in Artificial Intelligence and Big Data program for the academic environment and support that contributed to the development of this research. Also we would like to thank the speech therapists at *Mais Fono Fonoaudiologia* for all their support, especially Patrícia França for her contribution to the creation of the dataset.

References

Raghad Alnashwan, Noura Alhakbani, Abeer Al-Nafjan, Abdulaziz Almudhi, and Waleed Al-Nuwaiser. 2023. Computational intelligence-based stuttering detection: A systematic review. *Diagnostics*, 13(23):3537.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Ho Sio Cheang, Themos Kourkounakis, and Tiago H. Falk. 2020. Libristutter: A corpus of read english speech with annotated stuttering. In *Proceedings of Interspeech 2020*, pages 4341–4345.

Steven B. Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Claudia Regina Furquim de Andrade. 2006. *Gagueira Infantil: Risco, Diagnóstico e Programas Terapêuticos*. Pró-Fono, Barueri.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6645–6649.

Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep Speech: Scaling up end-to-end speech recognition. arXiv:1412.5567.

Peter Howell, Stephen Davis, and Jon Bartrip. 2009. The university college london archive of stuttered speech (uclass). *Journal of Speech, Language, and Hearing Research*, 52(2):541–548.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3 edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Jean Sawyer and Ehud Yairi. 2006. The effect of sample size on the assessment of stuttering severity. *American Journal of Speech-Language Pathology*, 15(1):36–44.

Shakeel Ahmad Sheikh, Md Sahidullah, Fabrice Hirsch, and Slim Ouni. 2021. Machine learning for stuttering identification: A comprehensive review. *IEEE Access*, 9:6341–6360.

I. Świetlicka, P. Kowalczyk, and A. Nowicki. 2013. Artificial neural networks in the disabled speech analysis. *Archives of Acoustics*, 38(3):321–328.

I. Szczurowska, M. Kowalski, and J. Nowak. 2006. Application of artificial neural networks for classification of speech disfluencies. *Archives of Acoustics*, 31(2):203–210.

Ehud Yairi and Nicoline G. Ambrose. 1999. Early childhood stuttering i: Persistency and recovery rates. *Journal of Speech, Language, and Hearing Research*, 42(5):1097–1112.

Ehud Yairi and Nicoline G. Ambrose. 2005. *Early Childhood Stuttering: For Clinicians by Clinicians*. Pro-Ed, Austin, TX.