StamFree: A Gamified AI System for Speech Disfluency Detection and Therapy in Children

Adithya P Binu
Dept. of Computer Science and Engineering
St. Joseph's College Of Engineering And Technology, Palai
Kerala, India
adithyapbinu2026@cs.sjcetpalai.ac.in

Devika Rajeev
Dept. of Computer Science and Engineering
St. Joseph's College Of Engineering And Technology, Palai
Kerala, India
devikarajeev2026@cs.sjcetpalai.ac.in

Doney Siby
Dept. of Computer Science and Engineering
St. Joseph's College Of Engineering And Technology, Palai
Kerala, India
doneysiby2026@cs.sjcetpalai.ac.in

Emitta Mathew
Dept. of Computer Science and Engineering
St. Joseph's College Of Engineering And Technology, Palai
Kerala, India
emittamathew2026@cs.sjcetpalai.ac.in

Dr. Joby P P
Dept. of Computer Science and Engineering
St. Joseph's College Of Engineering And Technology, Palai
Kerala, India
hodcs@sjcetpalai.ac.in

***Abstract*—Speech disfluency, commonly referred to as stammering, is a multifaceted communication disorder that predominantly affects children between 6 to 12, with substantial consequences for social confidence, academic achievement, and psychosocial development. Although traditional speech therapy demonstrates efficacy, it frequently necessitates intensive clinical supervision and repetitive exercises, which may be stressful and monotonous for pediatric patients, resulting in low adherence. This study presents StamFree, an innovative, child-focused gamified therapy system utilizing advanced artificial intelligence.**

In contrast to earlier systems that utilize basic signal processing, StamFree employs WavLM, a state-of-the-art self-supervised deep learning model, to analyze speech directly from raw audio waveforms. This architecture enables robust multi-class classification of disfluencies, accurately distinguishing between *repetitions*, *prolongations*, and *blocks*. The system incorporates a novel Stress-Based Progression Strategy, which organizes phonemes into hierarchical tiers according to articulatory stress levels: low, medium, and high. By integrating this progression with an adaptive unlocking mechanism, StamFree ensures that users achieve proficiency with lower-stress sounds prior to advancing, thereby minimizing cognitive overload. Interactive mini-games that reinforce breathing control and pacing further contribute to a low-anxiety, engaging therapeutic environment, promoting sustained practice beyond the clinical context.

***Index Terms*—Speech disfluency, Stammering, Gamified therapy, WavLM, Child-centric UI, Multi-class Speech Analysis.**

I. Introduction

Speech disfluency, or stammering, is a speech disorder that mostly shows up in childhood. It interrupts the smooth flow of speech and is often marked by sound repetitions, prolonged sounds, or sudden pauses. For many children, stammering makes communication tough. It can also affect their confidence, participation in class, and social relationships. Getting early and consistent help is important. This support can help children improve their speech fluency and lessen long-term emotional effects. [9].

Traditional speech therapy from Speech-Language Pathologists effectively treats stammering. However, this therapy often requires frequent visits to clinics, making it expensive and difficult for many families to access. Children also need to practice repetitive speech exercises at home, which can feel boring or stressful. As a result, children may lose interest, leading to inconsistent practice and slower progress.

Recent progress in artificial intelligence has made it possible to analyze speech and automatically recognize patterns of disfluency. Many of the current systems use standard signal processing methods or neural networks trained on selected features like spectrograms [11]. While these methods can identify speech issues, they often fail to capture the natural flow and context of speech. Additionally, most systems only focus on detection and do not provide engaging therapy support, especially for children.

To tackle these challenges, this paper presents StamFree, a speech disfluency detection and therapy system for kids that uses AI. StamFree employs WavLM, a self-supervised deep learning model, to examine raw speech audio and identify different types of disfluencies, like repetitions, prolongations, and blocks. Unlike traditional methods, StamFree combines speech analysis with interactive games aimed at children.

The system uses a stress-based progression model. Children begin with simple, low-stress sounds. They only move on to more difficult speech tasks after they show improvement.

StamFree combines AI-driven analysis with enjoyable therapy activities. This approach creates a supportive and engaging environment that encourages children to practice regularly and boosts their confidence.

II. Related Work

The field of automated speech pathology detection has changed a lot. It has moved from traditional signal processing techniques to deep learning frameworks. This section looks at how stutter detection methods have developed and points out the specific gaps that StamFree seeks to fill.

A. Traditional Approaches and Spectrogram Analysis

Early research in speech dysfluency mainly depended on manual feature engineering. Studies used signal processing techniques like Mel-Frequency Cepstral Coefficients (MFCCs) and Zero Crossing Rates along with traditional classifiers such as Support Vector Machines (SVMs) [10]. Although these methods set a baseline, they did not have the ability to capture complex temporal dependencies.

As deep learning advanced, the focus shifted to Convolutional Neural Networks (CNNs). A typical method involved turning audio signals into Mel-spectrogram images and treating dysfluency detection as a computer vision problem [6], [7]. Although these CNN-based systems improved detection accuracy, they mainly emphasized local spatial features rather than the sequential context of speech. Additionally, many of these architectures were restricted to binary classification (fluent vs. stuttered) and could not identify specific types of dysfluency, such as repetitions, prolongations, and blocks. [9].

B. Self-Supervised Learning: The Shift to WavLM

Recent advancements have focused on self-supervised learning (SSL) models that learn directly from raw audio waveforms. While early SSL models like Wav2Vec 2.0 showed better performance than handcrafted features, they often have difficulty with background noise and speaker variability. These are critical factors in pediatric speech therapy, where acoustic environments are unpredictable.

To tackle this issue, WavLM was introduced as a strong successor. It was pre-trained on large datasets with a denoising modeling task [3]. Studies show that WavLM does much better than earlier BERT-based audio models in tasks like separating speech from noise and managing various speaker traits. However, most current uses of WavLM in speech pathology are strictly diagnostic. They give clinicians data but do not provide any interactive tools for patients. [11].

C. Gamification in Pediatric Therapy

Parallel to AI advancements, studies have looked at gamification in healthcare. Research shows that game-based therapy greatly improves motivation and adherence in children [4]. However, many gamified speech tools currently depend on strict, rule-based feedback, such as basic volume detection, instead of smart semantic analysis.

D. The Gap: Intelligent Interaction

The literature shows a clear difference. Powerful AI models, such as WavLM, do not engage users well. On the other hand, current gamified apps do not provide smart, real-time analysis. StamFree fills this gap by combining a refined WavLM model with a Stress-Based Progression Framework. Unlike earlier efforts, our system adjusts the difficulty of exercises according to specific phonemic stress levels. This creates a flexible, child-focused therapeutic environment that is both clinically based and engaging.

III. Methodology

The StamFree system is a child-focused framework designed to detect speech disfluency and support therapy through gamified digital interaction. The system combines AI-based speech analysis with engaging therapeutic activities to assist children in improving speech fluency. The overall workflow includes capturing speech input, preprocessing audio, detecting disfluency using a deep learning model, categorizing stress levels, and delivering appropriate therapy exercises.

A. Speech Datasets

To train and evaluate the speech disfluency detection model, StamFree utilizes publicly available speech pathology datasets, including FluencyBank and SEP-28k. These datasets contain recordings of both fluent and disfluent speech collected from individuals who stutter.

FluencyBank is a clinically curated dataset containing annotated speech samples with detailed labels for different types of disfluency events. SEP-28k is a large-scale dataset designed specifically for speech disfluency research and contains a wide variety of speech recordings collected in real-world conditions.

The combined dataset contains more than 32,000 annotated audio clips, covering multiple disfluency categories such as repetitions, prolongations, and speech blocks. The use of multiple datasets increases diversity in speaker characteristics, recording environments, and speech patterns, improving the robustness and generalization capability of the model.

B. Overall System Workflow

The methodology begins with gathering speech input from the child through a mobile app. The recorded audio goes to the backend processing unit, where it is preprocessed with steps like normalization and resampling to fit the speech model's input needs. Unlike traditional systems that rely on manually created acoustic features [10], StamFree processes raw audio signals directly.

The main analytical part of the system is based on WavLM [3], a self-supervised deep learning model. WavLM sets itself apart from earlier models like Wav2Vec 2.0 with its pre-training objective. This objective includes masked speech prediction along with denoising modeling. This ability makes it very effective for analyzing children's speech, which often varies in pitch, pace, and background noise. The model is fine-tuned using labeled speech disfluency datasets to carry

out multi-class classification. It identifies whether the speech segment is a repetition, prolongation, block, or fluent speech.

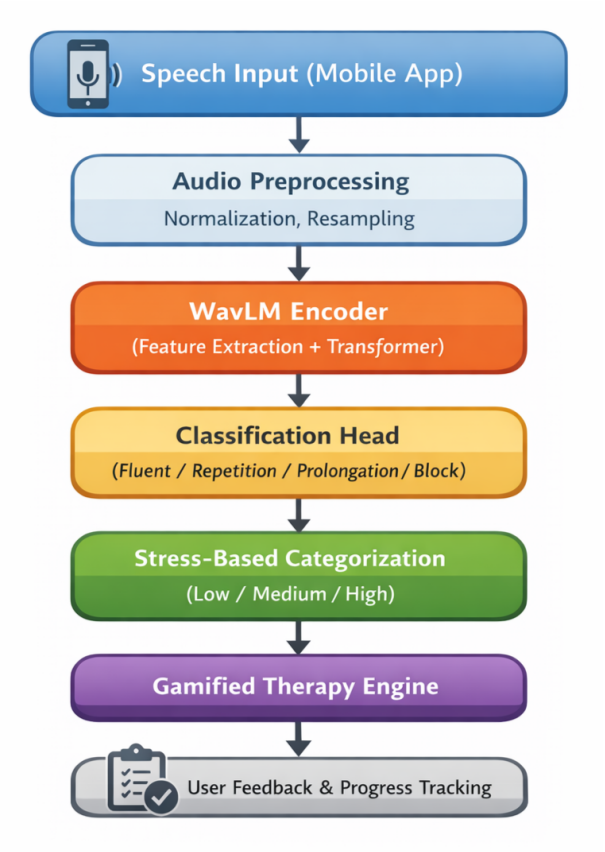

Fig. 1. Overall architecture of the StamFree speech disfluency detection and therapy system.

C. Disfluency Classification Logic

Once the WavLM encoder extracts the speech embeddings, they go to a task-specific classification head. This classification layer provides probability scores for each disfluency category. The system selects the predicted disfluency type based on the highest confidence score. This multi-class classification method allows the system to distinguish between different stammering patterns instead of just offering a basic fluent or non-fluent decision. This distinction is crucial for personalized therapy.

D. Evaluation Metrics

To evaluate the performance of the speech disfluency detection model, standard classification metrics are used, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's ability to correctly identify different types of speech disfluencies.

Accuracy measures the overall proportion of correctly classified samples among all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision measures the proportion of correctly predicted positive observations among all predicted positives.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall (also known as sensitivity) measures the proportion of correctly predicted positive observations among all actual positives.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-score represents the harmonic mean of precision and recall, providing a balance between the two metrics.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

E. Stress-Based Sound Categorization

An important contribution of StamFree is the use of stress-based sound categorization. Speech sounds are sorted into low-stress, medium-stress, and high-stress categories based on how complex they are to produce and the effort needed for their production. Low-stress sounds include simple vowels and nasals. Medium- and high-stress sounds include plosives, fricatives, phoneme clusters, and continuous speech patterns. This categorization serves as the basis for a structured therapy progression.

F. Adaptive Gamified Therapy Mapping

Based on the type of speech difficulty and the related stress level, the system picks a fitting gamified activity. Each game targets a specific speech challenge [4]. For instance, breathing games help reduce speech blocks, slow-paced speech games promote controlled speech, sound-tracking games manage prolongation, and word-tapping games reduce repetition. The games come in different difficulty levels to support gradual improvement without overwhelming the child.

G. Progression and Unlocking Mechanism

The system uses a flexible unlocking method to manage therapy progression. Children start with low-stress levels and can move to higher levels only after showing consistent fluency, as confirmed by the AI model. This way of progressing based on mastery helps children gain confidence at each stage before tackling more complex speech tasks. The child's performance history is continuously updated and saved, allowing for personalized therapy paths.

H. Child-Centric Design Considerations

Throughout the methodology, special emphasis is placed on creating a low-anxiety and engaging environment for children. Clinical feedback is shown through positive game responses, such as unlocking levels or offering visual rewards, instead of using numerical scores or error messages. This design choice helps support motivation, keeps kids engaged, and promotes emotional well-being while maintaining therapeutic effectiveness.

IV. Experimental Setup

The experimental setup of the StamFree system aimed to assess how well AI-based speech disfluency detection works and how it fits into a therapy framework focused on children and gamification. The setup involves preparing the dataset, fine-tuning the model, implementing the system, and evaluating its performance.

A. Dataset Description

To train and evaluate the speech disfluency detection model, we used a hybrid dataset that combined samples from publicly available speech disfluency datasets like UCLASS and SEP-28k [8]. These datasets include recordings of fluent and disfluent speech, featuring various stammering patterns such as repetition, prolongation, and blocks. Using a hybrid dataset increases robustness by capturing both clinical and real-world speech variations.

All audio samples were set to a fixed sampling rate of 16kHz to ensure compatibility with the WavLM architecture. Basic preprocessing steps like silence trimming and amplitude normalization were used to reduce noise and improve model stability. No manual feature extraction was done; the system directly processes raw audio signals..

B. Model Training and Configuration

The main speech analysis part uses the WavLM Base model [3], which was fine-tuned for classifying different types of speech disfluency. The pretrained weights let the model take advantage of detailed speech patterns learned from a large amount of unlabeled audio data (94,000 hours). A classification layer specific to the task was added on top of the encoder to predict repetition, prolongation, block, or fluent speech.

Model training used the PyTorch framework along with the HuggingFace Transformers library. The dataset split into training and validation sets to track performance and avoid overfitting. We applied standard optimization methods like learning rate scheduling and early stopping to improve convergence.

C. System Implementation

The StamFree system uses a client-server setup. A mobile app was created to capture speech input from children while they play. The recorded audio is sent securely to a backend server. There, the trained WavLM model processes the audio. The system then identifies the type of disfluency and maps it to the related gamified therapy module. This information is sent back to the app almost instantly.

The backend services handle audio processing, inference execution, user progress tracking, and adaptive level unlocking. This modular design ensures scalability and allows the system to support multiple users simultaneously.

D. Evaluation Metrics

To assess model performance, we used standard classification metrics like accuracy, precision, recall, and F1-score. These metrics help evaluate how well the system identifies different types of speech disfluency. We also made qualitative observations on user engagement and progression through game levels to evaluate the practical effectiveness of the gamified therapy approach.

V. Results and Discussion

This section presents the experimental results from the StamFree system and discusses how well the AI-based speech disfluency detection and gamified therapy framework works. The evaluation looks at the performance of the speech classification model and the practical usability of the child-focused therapy design.

A. Model Performance Results

The fine-tuned WavLM model was tested on a validation subset of the hybrid dataset to see how well it identified different types of speech disfluency. Standard classification metrics, including accuracy, precision, recall, and F1-score, were used for this evaluation. The model performed well in distinguishing between repetition, prolongation, blocks, and fluent speech.

B. Comparative Discussion

When you compare StamFree to traditional speech disfluency detection systems that rely on handcrafted features or CNN-based spectrogram analysis, StamFree shows clear improvements. Existing systems usually perform binary classification and do not understand the context of speech [7]. Additionally, standard Wav2Vec 2.0 implementations often perform poorly in the presence of variable background noise. In contrast, WavLM, which has a specialized denoising pretraining task, allows the proposed system to capture long-term dependencies in raw audio, even in challenging recording environments [3].

Moreover, the use of a self-supervised learning framework enables the model to learn richer speech representations without requiring extensive manual feature engineering. This improves the robustness of disfluency detection across different speakers, accents, and recording conditions.

Most current methods focus only on detection and offer little or no therapy support. StamFree goes beyond detection by combining real-time classification results with engaging gamified therapy. Additionally, the adaptive progression mechanism ensures that therapy activities are tailored to the child's speech difficulty level, promoting gradual improvement and sustained engagement during practice sessions.

TABLE I
COMPARISON OF STAMFREE WITH EXISTING SYSTEMS

System	Approach	Key Capability
MFCC + SVM Systems	Traditional ML (SVM)	Binary classification of speech disfluency
CNN Spectrogram Models	CNN-based models	Spectrogram-based speech analysis
StutterAI [11]	ML + NLP	Speech analysis without therapy support
StamFree (Proposed)	WavLM Transformer Model	Multi-class disfluency detection with gamified therapy

C. Gamified Therapy Effectiveness

The researchers evaluated how well the gamified therapy framework worked by looking at how users interacted and progressed through game levels. Children were more engaged when therapy came in the form of interactive games instead of repetitive drills [4]. The stress-based progression strategy made sure that users did not face complex speech tasks too soon, which helped lower their frustration and anxiety.

The adaptive unlocking mechanism was crucial for keeping children motivated. They could see their progress through level advancement and rewards. This design encourages regular practice, which is important for effective speech therapy.

D. Discussion

The experimental results show that StamFree effectively combines precise speech disfluency detection with an engaging therapeutic experience designed for children. The use of self-supervised learning (WavLM) and gamification tackles significant issues found in current systems. While this evaluation mainly highlights technical performance and qualitative observations, the results suggest that StamFree has great potential as an additional resource for pediatric speech therapy.

VI. CONCLUSION AND FUTURE SCOPE

This paper introduced StamFree, a child-focused AI-powered game system designed to assist with speech disfluency detection and therapy. The system uses self-supervised speech analysis with WavLM alongside interactive, stress-aware gaming. This mix provides an engaging and low-anxiety setting for therapy aimed at children. Unlike traditional systems that focus only on speech detection, StamFree combines classification with adaptable therapy. This enables personalized and gradual speech improvement.

The method showed how to analyze raw audio to spot different types of speech disfluency, such as repetition, prolongation, and blocks. The stress-based progression strategy helped children start with simple speech tasks and move on only after they have gained enough fluency. Experimental observations and results show that the system reliably identifies disfluency patterns while keeping users engaged through gamified activities.

The key strength of StamFree is its ability to connect precise speech analysis with fun therapy delivery for kids. By turning speech exercises into enjoyable games, the system motivates regular practice, which is crucial for effective speech therapy in children. The client-server design also allows for scalability and real-time feedback, making the system suitable for practical use.

In future work, the system can be improved by adding real-time feedback during gameplay and increasing support for multiple languages to reach different user groups [2]. Connecting with a therapist or parent monitoring dashboard could help professionals track progress from afar and tailor therapy plans. Additionally, larger clinical evaluations and long-term user studies can be done to further confirm the effectiveness of StamFree in real-world pediatric therapy settings.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to their institution for their continuous guidance and support throughout the development of this work.

REFERENCES

[1] A. Abedal-Kareem et al., "ClearFlow: Empowering fluent communication through RAG-based text generation for people who stutter," in *Proc. International Conference on Artificial Intelligence and Communication Systems*, 2025.

[2] J. Snehith et al., "Advanced multimodal emotion recognition using integrative analysis of multilingual text, EEG, voice signals, and facial expressions," in *Proc. International Conference on Intelligent Systems*, 2025.

[3] S. Chen et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505-1518, 2022.

[4] N. E. Syam et al., "Speech therapy assistance through gamification," in *Proc. International Conference on Healthcare Informatics*, 2024.

[5] N. Nabilah et al., "Implementation of facial expression recognition using convolutional neural networks," in *Proc. International Conference on Computer Vision and Pattern Recognition*, 2024.

[6] K. Mittal et al., "Evolving diagnostic techniques for speech disorders: Investigating dysarthria classification through DenseNet201 CNN framework," *IEEE Access*, vol. 12, pp. 11234–11245, 2024.

[7] A. Kheterpal et al., "CNN-driven innovations in dysarthria classification and speech disorder management," *Journal of Speech Technology*, vol. 9, no. 2, pp. 85–96, 2024.

[8] V. N. Narasinga et al., "Enhancing stuttering detection: A syllable-level stutter dataset," in *Proc. IEEE International Conference on Signal Processing*, 2024.

[9] I. Sindhu et al., "Automatic speech and voice disorder detection using deep learning: A systematic literature review," *IEEE Access*, vol. 11, pp. 54321–54338, 2024.

[10] P. H. Keerthi et al., "Analysis of features for dysarthria severity classification from speech," in *Proc. International Conference on Speech Processing*, 2024.

[11] B. Arachchi et al., "StutterAI: A virtual assistant for stutter detection and analysis using machine learning and NLP," in *Proc. International Conference on Artificial Intelligence Applications*, 2023.