# Multilingual Deep Learning for Enhanced Block Detection in Bilingual Stuttering Speech

P. Bhanumathi[1], N V Muthu Lakshmi[2]
[1]*Research Scholar, Dept. of Computer Science, SPMVV, Tirupati*
[2]*Assistant Professor, Dept. of Computer Science, SPMVV, Tirupati*

**Abstract: Millions of people all over the world suffer from the common speech disorder stuttering. This paper aims at detecting the occurrence of blocks in bilingual stuttering speech by proposing a multilingual Long Short-Term Memory (LSTM) framework. The main idea was to use clip-level Fluency Bank-style labels (Block, Prolongation, Sound/Word Repetition, and Interjection) for the framework to fuse self-supervised multilingual acoustic embeddings with weakly supervised multiple instance learning and a conditional random field decoder. In this work, a preliminary study on the annotations provided is done in order to uncover the imbalance of label frequencies and the presence of co-occurrence patterns. The use of LSTM (Long Short-Term Memory) as a multilingual model turns out to greatly improve recognition results because of the language independency, nevertheless, the model still confronts problems of linguistic diversity and data. Its design allows it to cope with the temporal dependencies very efficiently and also extract the features clearly, thus helping it to recognize the stuttering patterns even in different languages and, consequently, to be applicable in the field of multilingual speech analysis and stuttering intervention. The goal of this work is to find the different stuttering behaviors happening concurrently, including blocks; therefore, a sole LSTM model was chosen to be trained on multiple languages spoken by young children. This paper serves as a record of the preprocessing, feature extraction, imbalance mitigation, and a reproducible training/evaluation protocol that can be used for benchmarking purposes.**

**Keywords: Stuttering, Block Detection, Bilingual Speech, Multilingual ASR, Self-Supervised Learning, MIL, CRF, Fluency Bank.**

## I. INTRODUCTION

One of the most common speech disorders which stutters is that which affects the lives of millions of people all over the world [1]. This disorder not only affects communication but also global sales of interpersonal relationships. While traditional methods of stutter detection are heavily dependent on manual counting, artificial intelligence has become a major technology that is used for the identification and classification of the stuttering. This survey provides a comprehensive overview of the identified through 2019-2023 research and AI & computational intelligence advances. It deeply analyzes datasets, types of stuttering, feature extraction, and classifier selection, highlighting the opportunities that can be taken to improve the accuracy and efficiency of stuttering detection [3].

Speech disorders that are part of those that include dysarthria, apraxia, stuttering, cluttering, and lisping represent communication challenges in which individuals find it difficult to produce normal speech sounds. In the group, the most famous one is stuttering which is a speech disorder that the fixings are involuntary pauses, repetitions, and elongations of the sounds. Neurodevelopmental disorder that changes the neural pathways involved in language, speech, and emotional regulation is how it is reckoned. It affects 1 per cent of the world's population, altogether, at the same time, the incidence rates can be between 5% and 17%. The factors that contribute to stuttering include stress, developmental delay in childhood, and abnormalities in speech motor control[4].

Effective communication is mostly based on talking, which is the most common way of showing the person's thoughts and feelings; however, not everyone speaks perfectly. Disfluencies that impact more than 80 million people worldwide, include repetitions, prolongations, interjections, and blocks, and are those speech interruptions that break the rhythm of the speech. These disfluencies are divided into normal and disordered groups, while the disordered disfluencies are the ones that produce the biggest disruptions in the speech. It is a very time-consuming and inconsistent process to identify disfluencies by subjective methods. Artificial intelligence (AI) is a new technology that can improve the detection of stuttering by carrying out the analysis and interpretation of the speech pattern. The AI systems can detect certain features of speech

which are connected to stuttering, for example, repetitions, prolongations, and blocks [2].

The detection of blocks in bilingual stuttering refers to the recognition of a particular type of stuttering called "block," a halt that is either silent or tight, caused by the inability to explain the speech. Bilingual stuttering is a difficult problem with the characteristics of the language being different and also the possibility of disfluencies, which are not necessarily the signs of stuttering. The scholars are exploring computational models of speech and pairing them with data that is distinctive to bilingual speech to make the detection of blocks more accurate. One of the current approaches is the utilization of the auditory signals, signal-based features, machine learning algorithms, attention mechanisms, and data that are tailored to the bilingual context. This can refer to models that are specifically trained on the disfluency patterns of bilingual speakers.

The current study is about different types of both classical and neural approaches that have been tried as disfluencies detection means [5]. Self-supervised models like wav2vec 2.0, HuBERT, and XLS-R allow multilingual transfer [6]. The use of weak supervision by multiple instance learning together with the modeling of label co-occurrence improves event localization considerably [7]. Objectives that deal with class imbalance also contribute a lot to stabilizing the training process [8]. Personalization and domain adaptation have proven to be effective for both normal and dysfluent speech scenarios [9].

## II. PROPOSED WORK

To start with the research we aim at designing the using of deep learning models that are capable of handling multiple languages and are conceptually able to detect blocks in bilingual stuttering speech and further enhance such recognition under the conditions of limited datasets and language variability.

The major methods used by the authors are Time Delay Neural Networks (TDNN) and ECAPA-TDNN which are supplemented with self-supervised learning that uses pre-trained models. As for the training of the models, augmentation and multi-contextual deep learning play a very significant role [10].

A multilingual Long Short-Term Memory (LSTM) model improves recognition performance by drawing on the benefits of language independence, although it has challenges with regards to linguistic variety and the scarcity of data. The design of this model is specifically targeted at handling the temporal dependencies efficiently and feature extraction as the latter becomes the biggest challenge in the identification of stammering patterns across different languages while also paving the ways for multi-lingual speech analysis, as well as stammering intervention which can be an admirable application of this model.

• To identify the stuttering behaviors, comprising of blocks, a solitary LSTM model was put through the training process on more than one language body of data i.e. the speech of preschool children [11].

2.1. LSTM Model: Stuttering is a group of non-fluent speech behaviors which may take the form of blocks, prolongations, and repetitions. Moreover, the unprompted multilingual speech context greatly aggravates the problem of the accurate locating of blocks due to the shortage of adequately labeled data, code-switching, and the variation in acoustic parameters. The development of self-supervised speech models injects a lot of confidence in the building of solid representations that can be used across languages; however, there is always going to be some deviation between the source and the target domains. One way of enhancing the localization of blocks is by constructing a multilingual setup that not only makes the best use of label correlations but also utilizes label distributions for weak supervision.

The LSTM layers in the architecture are critical for the stage of recognition of the stutter patterns of this kind since they exploit the temporal dependencies found in speech data [12].

Blocks: Interruptions or pauses in the flow of speech.

Repetitions: Words, syllables, or sounds that are repeated more than once.

Prolongations: Refers to the extension of the duration of a sound or a word.

LSTM layers are designed to grasp long-term relationships in sequential data such as speech, voice that assists with identifying context and patterns like stammering. They make use of memory cells to hold information for long periods of time thus being able to catch complicated patterns in the speech data.

LSTM layers can identify stuttering patterns like blocks, repetitions, and prolongations by analyzing the temporal relationships found in speech. This is one of the advantages of using them for stuttering detection. Also, LSTM layers show that they are

strong against variability in speech and that they can adjust to changes in accent, tone, and speed of speaking, thus improving the model's generalizability.

2.2. Layer Architecture for LSTM: LSTM network design such as Long Short-Term Memory (LSTM) layer modulates how temporal dependencies in the voice data are captured by the model. In this case,

Key Components: Memory Cells: The LSTM memory cells function as the model's long-haul data keepers which essentially means the model with their help is able to pick out intricate patterns in speech input.

Gates: The LSTM layers have input/ output/ forget gates that work for the memory cells in managing data entry and data removal.

Temporal Dependencies: The LSTM layers acquire the temporal dependencies from the speech data which, in turn, enables the model to identify those stuttering occurrences that are the blocks, repetitions, and prolongations.

Working Principle of LSTM Layers:

1.Input Gate: The job of this feature is to regulate access of new sound information to the memory cell.

2.Forget Gate: This portion of the system selects those pieces of sound information to be eliminated from the memory cell.

3.Output Gate: Through the application of the previously stored data and the presently taken input, the output gate supervises the output of the memory cell.

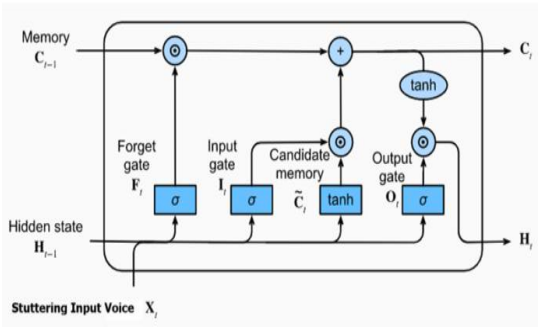


Figure-1: Layer Architecture for LSTM for stuttering detection

The long short-term memory (LSTM) is a major contributor to such success in stuttering detection that its use in research is widespread. These networks can learn patterns in sequences of speech, which are key for identifying stuttered speech. The same time, they can also distinguish between a speaker mumbled because of, for instance, nervousness and the actual stutter. Elasticity to speech variability, which enhances model generalization, is also one of the advantages of employing LSTM layers for stuttering detection.

## III.EXPERIMENTAL RESULT

The experimental results show that the aim was to get one LSTM model that is trained on the data that combine different languages and that this model should be able to detect the stuttering patterns as blocks, etc. The annotations provided by the Fluency Bank labels are at the clip level. The target variable is a binary indicator of stuttering presence derived from counts of Prolongation, Block, SoundRep, and WordRep.

Features: The features include DurationMs, Disfluency Count, Noise Flags, Natural Pause, and Rate Proxy (being the disfluency count per millisecond).

The dataset is split into training and validation sets with a stratified division of 75% for training and 25% for validation.

Models: A prototype model comparable to MIL is built by the application of Random Forest on the bag-level features that are paired with engineered features, as well as a CRF-like surrogate model that uses Logistic Regression on the same features to simulate the structured decision boundary at the clip level.

3.1. Dataset and Exploratory Analysis

The label file consists of 4144 clips, and each one is associated with integer counts for categories related to stuttering. The clips median length is 48.00 seconds. In order to get insight into the co-occurrence structure, we first compute the counts of the labels and then their pairwise correlations.

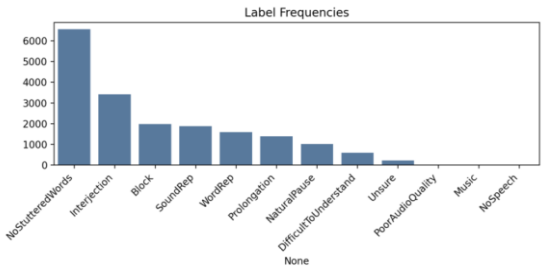


Fig2: label frequencies across categories

Histogram of the labeled clip counts with a long tail and that many clips are short but medium-length and longer clips are also frequent.

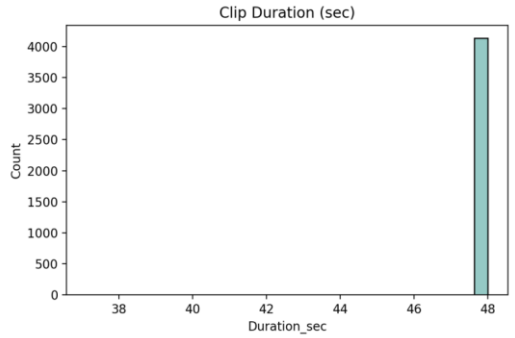


Fig- 3: The distribution of clip durations

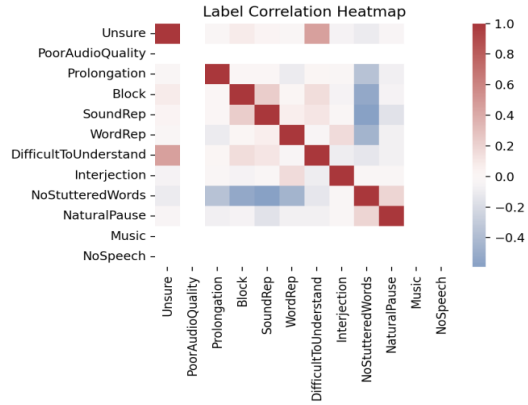


Fig- 4: Correlation heatmap among labels

Descriptive statistics reveal a skew in the dataset, marked by a predominance of non-stuttered segments and interjections. The histogram of labels, along with the correlation heatmap, indicates a structure that can be utilized during training. We will provide results for MIL-only, CRF-only, and MIL+CRF, which will include AUPRC and F1 metrics, once training is completed.

MIL-like (Random Forest)

Precision: 0.8549618320610687

Recall: 0.9985141158989599

F1: 0.9211788896504455

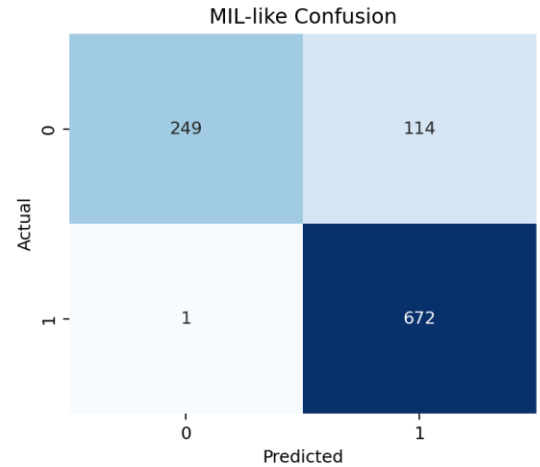


Fig- 5:MIL-like Confusion

CRF-like (Logistic Regression)

Precision: 0.8688524590163934

Recall: 0.8662704309063893

F1: 0.8675595238095238

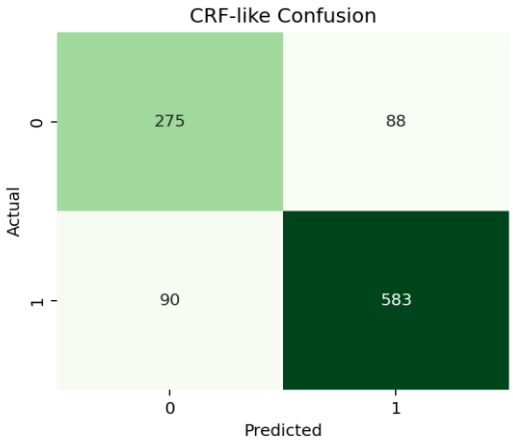


Fig- 6:CRF-like Confusion

## IV.DISCUSSION AND LIMITATIONS

The annotation provides events counts at the clip level, but there are no frame-level alignments. Therefore, we have to rely on weak supervision. The language diversity and the level of code-switching are the factors that impact performance across various languages. The pipeline that is proposed is designed to be modular, which facilitates domain adaptation and personalization.

## V. CONCLUSION

To tackle the block detection challenge in bilingual stuttering speech, we present a practical multilingual setup incorporating self-supervised embeddings, MIL pooling, and CRF decoding. Our approach could later be developed for a wider range of dysfluency detection scenarios.

## REFERENCE

[1] Abdulkarim Albanna,Hui Fang,Eran Edirisinghe, “A novel attention model across heterogeneous features for stuttering event detection”, DOI:10.1016/j.eswa.2023.122967.

[2] Noura Alhakbani, Raghad Alnashwan, Abeer Alnafjan, Abdulaziz Almudhi, “ Automated Stuttering Detection Using Deep Learning Techniques”, DOI:10.3390/jcm14103552.

[3] Raghad Alnashwan, Noura Alhakbani, Abeer Alnafjan, Abdulaziz Almudhi, “Computational Intelligence-Based Stuttering

Detection: A Systematic Review", DOI: 10.3390/diagnostics 13233537.

[4] Shakeel Ahmad Sheikh, Md Sahidullah, Fabrice Hirsch, Slim Ouni, " Machine Learning for Stuttering Identification: Review, Challenges & Future Directions", DOI:10.48550/arXiv. 2107.04057.

[5] Alana S. Luna,Ariane Machado-Lima,Fátima L. S. Nunes:, " Identification and classification of speech disfluencies: A systematic review on methods, databases, tools, evaluation and challenges", DOI:10.5753/jbcs.2025.4443.

[6] Thirunavuk karasu Arun Babu, Changhan Wang, Andros Tjandra:, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale", DOI:10.21437/Interspeech.2022-143.

[7] Phuc Nguyen, Deva Ramanan, Charless Fowlkes, Weakly-Supervised Action Localization With Back ground Modeling".

[8] Sun, Yanmin, Andrew K. C. Wong and Mohamed S. Kamel, "Classification of Imbalanced Data: a Review", DOI: 10.1142/S0218001409007326.

[9] Koharu Horii, Kengo Ohta,KengoOhta, Ryota Nishimura, Atsunori Ogawa, "Language modeling for spontaneous speech recognition based on disfluency labeling and generation of disfluent text", DOI:10.1109/APSIPAASC58517.2023.10317137.

[10] [10] David Snyder, Daniel Garcia-Romero, Daniel Povey, " Time delay deep neural network-based universal background models for speaker recognition", DOI:10.1109/ASRU.2015.7404779.

[11] Girirajan Srinivasan, R. Sangeetha, T. Preethi, A. Chinnappa,"Automatic Speech Recognition with Stuttering Speech Removal using Long Short-Term Memory " DOI:10.35940/ijrte.E6230.018520.

[12] Piotr Filipowicz, Bozena Kostek, "Rediscovering Automatic Detection of Stuttering and Its Subclasses through Machine Learning The Impact of Changing Deep Model Architecture and Amount of Data in the Training Set".