LSTM-Based Speech Analysis for Automatic Classification of Dysarthria Severity Levels

Guilherme B. F. Santos
Escola Politécnica de Pernambuco
Universidade de Pernambuco
Recife, Brazil
gbfs@poli.br

Pedro L. S. G. Camara
Escola Politécnica de Pernambuco
Universidade de Pernambuco
Recife, Brazil
plsgc@poli.br

Diego M. P. F. Silva
Escola de Tecnologia e Comunicação
Universidade Católica de Pernambuco
Recife, Brazil
diego.silva@unicap.br

Andrea M. N. C. Ribeiro
Departamento de Engenharia Eletrônica e Sistemas
Universidade Federal de Pernambuco
Recife, Brazil
andrea.marianogueira@ufpe.br

Sérgio M. M. Fernandes
Escola Politécnica de Pernambuco
Universidade de Pernambuco
Recife, Brazil
sergio.fernandes@upe.br

Rodrigo de P. Monteiro
Escola Politécnica de Pernambuco
Universidade de Pernambuco
Recife, Brazil
rodrigo.monteiro@poli.br

***Abstract*—Dysarthria is a motor speech disorder that affects the articulation and pronunciation of words, typically caused by damages to the neurological system responsible for speech. Classifying the severity levels of dysarthria is a clinically relevant task that can assist healthcare professionals in determining the most appropriate treatment strategies based on the degree of impairment. This study investigates the ability of Long Short-Term Memory (LSTM) models with Mel Frequency Cepstral Coefficients (MFCC) to classify speech samples of individuals with different levels of dysarthria using the UASpeech and Torgo datasets. Results showed good model performance with consistent accuracy across different levels of dysarthria severity, achieving values above 98% regarding the four severity levels considered in this work.**

***Index Terms*—Dysarthria, Severity Classification, Automatic Speech Recognition (ASR), Deep Learning, MFCC.**

I. Introduction

Dysarthria is a speech disorder that affects the functions of the tongue, jaw, soft palate, and vocal cords, causing communication difficulties for individuals. The symptoms of dysarthria can vary from mild to severe, and may include unclear articulation, slurred speech, monotone voice, difficulty in controlling the voice volume, and breathing challenges during speech. These symptoms can lead to significant communication challenges and negatively impact individuals' quality of life and self-esteem [1].

Dysarthria can be classified into different levels of severity, determined through clinical assessments and standardized tests, such as the Frenchay Dysarthria Assessment (FDA-2) [2], which evaluates oral motor function, articulation, and intonation, and the Assessment of Intelligibility of Dysarthric Speech (AIDS) [2], which measures speech intelligibility and efficiency. The severity classification often includes:

- **Mild dysarthria:** characterized by slight deviations in word articulation, with speech remaining intelligible although less clear.
- **Moderate dysarthria:** characterized by greater difficulty in articulating words, with speech often interrupted by pauses and efforts to produce correct sounds, impairing comprehension.
- **Severe dysarthria:** marked by extremely impaired articulation, with largely incomprehensible words, making verbal communication extremely difficult.

The treatment of dysarthria typically involves a multidisciplinary approach, which may include speech therapy to improve articulation and speech comprehension, occupational therapy to enhance control of the muscles related to speech [1], and the use of assistive technologies, such as Automatic Speech Recognition (ASR), to facilitate communication [4]. Automatic Speech Recognition for individuals with dysarthria takes into account the challenges posed by the condition and employs recognition techniques to improve their quality of life.

Research in ASR for dysarthric speech has explored various techniques based on Deep Neural Networks (DNNs) [3]. Among these, Convolutional Neural Networks (CNNs) have been widely adopted to extract spatial and temporal features from speech signals. Additionally, Recurrent Neural Networks (RNNs) and Transformers have demonstrated strong performance across a range of natural language processing tasks [4]. Among the RNNs, the Long Short-Term Memory (LSTM) is frequently employed in speech recognition applications, including within the context of dysarthric speech. LSTM networks are specifically designed to manage long-range temporal dependencies in sequential data such as speech, making them particularly suitable for capturing the complex and sequential patterns characteristic of dysarthric speech [6].

The Long Short-Term Memory (LSTM) is a neural network architecture known for its performance with sequential data. Its ability to model temporal dependencies and capture sequential patterns makes it particularly well-suited to address the unique characteristics of dysarthric speech. The LSTM's capacity to retain long-term information and learn from data sequences

makes it a suitable choice for processing dysarthric speech, enabling more accurate analysis and a better understanding of the variations and patterns in this type of discourse [4].

Despite their potential to handle the complexities of dysarthric speech, it remains unclear how LSTMs perform across different levels of dysarthria severity. This study addresses that gap by analyzing LSTM's ability to recognize various severity levels using the UASpeech dataset [5]. We also apply severity classification to the TORGO dataset [12], a clinically relevant resource lacking explicit severity labels, which comprises speech recordings from individuals with dysarthria primarily caused by cerebral palsy or amyotrophic lateral sclerosis, alongside control speakers. Through this analysis, we aim to determine if LSTM's performance varies significantly according to dysarthria severity, revealing potential model limitations and suggesting directions for future research. Our objective is to improve the accuracy of Automatic Speech Recognition (ASR) for individuals with dysarthria, regardless of their degree of speech impairment, and contribute to the development of more robust and inclusive ASR systems.

This study aims to evaluate the effectiveness of LSTM models in recognizing speech from individuals with varying levels of dysarthria severity. The analysis focuses on key aspects, such as model accuracy across severity levels and the ability to generalize. The objective is to contribute to the development of more robust and inclusive systems capable of meeting the needs of individuals with different degrees of speech impairment, thereby advancing the field of assistive speech recognition.

The remainder of this paper is organized as follows: Section II presents related work, Section III describes the proposed methodology, Section IV discusses the results and analysis, and Section V provides conclusions and outlines directions for future research.

II. Related Works

This section presents a review of ASR techniques applied to dysarthric speech, highlighting relevant research in the field. The review is organized into subsections that address various techniques and specific studies. Each subsection discusses an article, describing the methodologies employed, the results obtained, and the implications for ASR in the context of dysarthria.

The reviewed studies highlight promising approaches for recognizing dysarthric speech, particularly the use of LSTM networks and feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC). Based on these findings, the present work adopts LSTM and MFCC to investigate model performance in classifying different levels of dysarthria severity, aligning its methodology with the most relevant evidence in the literature.

A. A Survey of Automatic Speech Recognition for Dysarthric Speech [4]

This study provides a review of ASR techniques applied to dysarthria. Dysarthria, a neuromuscular disorder that impairs the ability to speak clearly, presents significant challenges for ASR, as symptoms and severity can vary greatly among individuals. Deep learning methods, including convolutional neural networks, deep neural networks, and recurrent neural networks, have emerged as powerful tools for improving the performance of ASR systems targeting this population. Additionally, the scarcity of dysarthric speech data is identified as a major obstacle, underscoring the need for more robust and representative speech databases.

Regarding evaluation metrics, this study highlights the importance of Word Error Rate (WER) and Phone Error Rate (PER) in assessing the effectiveness of ASR models in accurately recognizing and processing dysarthric speech. WER measures the proportion of incorrectly predicted words, reflecting the overall intelligibility of the system, while PER focuses on errors at the phoneme level, providing a better analysis of pronunciation accuracy. Together, these metrics offer a comprehensive evaluation of model performance in challenging speech scenarios.

B. Dysarthric Speech Recognition Using Convolutional LSTM Neural Network [6]

The article explores the use of Convolutional neural networks combined with Long short-term memory networks (CLSTM-RNNs, from Convolutional Long Short-Term Memory Recurrent Neural Networks) for recognizing dysarthric speech. The research utilizes a speech database comprising recordings from nine dysarthric patients to evaluate the effectiveness of this approach. The authors argue that integrating the local feature extraction capabilities of CNNs with the temporal modeling strengths of LSTMs can lead to more accurate and effective speech recognition for individuals with dysarthria.

The proposed method applies this combination, a relatively underexplored approach in the context of dysarthric speech to enhance recognition performance. The results suggest that the combined model outperforms systems based solely on CNNs or LSTMs, representing a significant improvement over conventional ASR techniques. Specifically, a gain of approximately 15% was observed compared to standard LSTM models, as measured by the Phone Error Rate, which quantifies the proportion of incorrectly identified phonemes relative to the total number of phonemes. This metric is critical for assessing ASR system accuracy, as it directly captures phonetic-level errors and offers detailed insights into the model's performance in recognizing speech sounds.

These findings demonstrate substantial improvements over traditional methods and highlight the potential of the proposed approach for future ASR applications in clinical and rehabilitation contexts, where accuracy and efficiency are essential for facilitating effective communication for individuals with dysarthria.

C. Diagnosing Dysarthria with Long Short-Term Memory Networks [7]

This article investigates the use of Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units

to diagnose dysarthria in Mandarin speakers based on syllable pronunciation samples. The study explores various LSTM network architectures within a binary classification task, demonstrating a significant performance advantage over a baseline fully connected network. Notably, it achieves over 90% of Area Under the Curve (AUC) in Receiver Operating Characteristic (ROC) curves when classifying new speakers, provided that a sufficient number of cepstral coefficients are used. The article emphasizes the need for accessible diagnostic methods for dysarthria, particularly in light of the shortage of professional speech therapists in China and the country's growing elderly population.

Current dysarthria assessment methods largely rely on subjective auditory perceptions and/or physical oropharyngeal and electroglottography examinations, which often suffer from low patient compliance. By contrast, the system proposed in the study offers a less invasive and more accessible alternative, using speech alone as input for preliminary screening or as a complement to traditional diagnostic methods.

The article details the model architecture, which processes speech recordings by converting them into Mel-Frequency Cepstral Coefficients that are fed into an LSTM, which in turn connects to a logistic regression layer to produce the final classification. Variants of the LSTM model, including bidirectional LSTMs and multi-layer LSTMs with regularization techniques such as dropout, were employed to enhance diagnostic accuracy.

The evaluation methodology involved testing on a dataset of 69 adult Mandarin speakers, utilizing metrics such as accuracy, recall, F-score, and AUC to assess model performance. The results demonstrate that LSTM networks can effectively capture and leverage temporal information from speech input to improve dysarthria diagnosis accuracy.

D. Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques [8]

The article presents a detailed investigation into the automated classification of dysarthria severity using various deep learning techniques and acoustic features. It examines the effectiveness of different neural network architectures, including Deep Neural Networks, Convolutional Neural Networks, Gated Recurrent Units (GRUs), and Long Short-Term Memory Networks, in assessing dysarthria severity levels.

The analysis focuses on acoustic features such as Mel-Frequency Cepstral Coefficients and Constant-Q Cepstral Coefficients (CQCCs), evaluating their ability to capture speech nuances associated with dysarthria. Additionally, i-vectors, which compactly represent speech variations, are studied in conjunction with DNN models to assess their effectiveness in discriminating between dysarthria severity levels. The models were trained and evaluated using established databases in the field, including TORGO [12] and UASpeech [5].

The results indicate that DNNs when combined with MFCC-based i-vectors, outperform other models and feature combinations in classification accuracy. Furthermore, the study highlights that while MFCCs provide effective classification in speaker-dependent scenarios, i-vectors demonstrate superior performance in speaker-independent contexts.

III. Methodology

The development of the present study was divided into stages that include a review of related works, configuration of the computational environment, data processing, the construction of the deep learning model, among others. First, a theoretical framework was established on automatic speech recognition techniques applied to dysarthric speech. A literature review was conducted using reputable databases such as IEEE Xplore, ResearchGate, SciELO, and MDPI. Relevant studies addressing technologies applied to this specific field were reviewed, allowing the identification of methodologies used and results obtained. The review included techniques such as convolutional neural networks, deep neural networks, and recurrent neural networks, with a special focus on the application of LSTM networks.

This section presents the proposed methodology for classifying different levels of severity. Figure 1 shows the proposed methodology for recognizing the level of dysarthria. The process begins with the collection of audio data from the UASpeech and Torgo datasets, followed by preprocessing steps such as standardization and padding. Acoustic features are then extracted using MFCCs. These features are fed into an LSTM-based model designed to learn temporal patterns in speech data. Finally, the model outputs the predicted severity level of dysarthria. The Long Short-Term Memory (LSTM) architecture was specifically chosen for its inherent capability to process sequential data and capture long-range dependencies, making it highly effective for analyzing the temporal nuances present in dysarthric speech patterns.

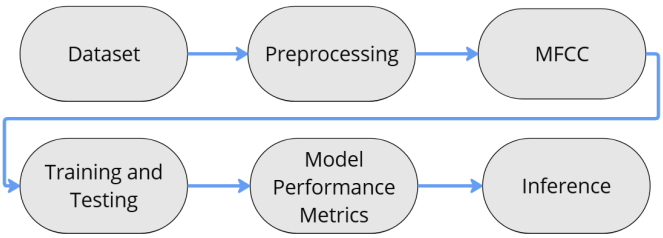

Fig. 1. Proposed methodology

A. Datasets

The UASpeech dataset [5] is a dataset for studies on dysarthric speech. This dataset was developed to provide a robust base of speech recordings, covering significant variations in dysarthria severity and speech intelligibility. It includes more than 400 words spoken by a total of 32 speakers, of which 13 are individuals without dysarthria (control) and 19 are speakers with dysarthria at different levels of severity labeled. The audio files in the dataset have a sampling rate of 16 kHz and are in the ".wav" format.

The TORGO dataset [12], is another widely used dataset for the study of dysarthric speech. It contains recordings

of speakers with cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS), as well as control speakers without speech impairments. The dataset includes both scripted and spontaneous speech, recorded in controlled environments. Contains 23 hours of speech data, from 8 speakers with dysarthria and 7 speakers without dysarthria (control). The audio recordings were made using two channels: a head-mounted microphone and a throat microphone, providing multiple acoustic perspectives. The sampling rate is 16 kHz and the files are in the ".wav" format.

B. Pre-Processing

The data used came from the previously mentioned datasets, which contained audio samples from the ".wav" format. The samples were resized so that all had the same duration, using the zero-padding technique, in which shorter sequences were padded with zeros until the length of the longest sequence was approximately 53 seconds. This standardization ensured uniform input dimensions for the subsequent LSTM model processing. The severity levels were converted into numerical values using Label Encoder [10]. This transformation ensured the compatibility of the labels with the LSTM model. Subsequently, the labels were converted to a one-hot encoding format for use with categorical loss functions. The one-hot encoding technique transforms numerical values into binary vectors, allowing the model to efficiently handle multiple severity classes. The audio signals, stored as bytes, were converted into numerical arrays using the Soundfile library [11], a crucial step for their efficient interpretation and subsequent numerical processing by the model.

C. MFCC

Representative acoustic features were extracted from the audio, with an emphasis on MFCCs. MFCCs were chosen for their ability to capture relevant information about the frequency spectrum, enabling the model to identify important patterns for the classification and recognition of dysarthric speech. The extraction process involved computing 40 Mel-Frequency Cepstral Coefficients (MFCCs) from each audio sample, using a sampling rate of 16 kHz. These coefficients were calculated by applying a short-time Fourier transform (STFT), followed by mapping the powers of the spectrum to the Mel scale, taking the logarithm, and finally computing the discrete cosine transform (DCT) of the result. The choice of 40 coefficients was based on empirical testing to balance computational efficiency and feature richness for the classification task.

D. Training and Testing

The LSTM model architecture was designed for speech recognition. The included layers were: **1.** a Masking Layer to ignore the padded values during zero-padding, ensuring that the model does not consider irrelevant information inserted during the standardization of the sequences, **2.** an LSTM Layer with 128 hidden units, configured to capture the temporal and sequential dependencies of the data, **3.** a Dense Layer responsible for mapping the latent representations of the LSTM to the severity class space, and **4.** the Output Layer configured with softmax activation, providing normalized probabilities for each severity class.

The dataset was divided into two subsets: 70% for training, used to adjust the model parameters, and 30% for testing, to evaluate the final performance of the model. The split was performed randomly, ensuring class balance across subsets.

The training was performed using the categorical cross-entropy loss function, appropriate for multi-class classification tasks. A batch size of 32 was used, processing 32 samples at a time during weight updates, and the maximum number of epochs was set to 100. The data was provided in batches, and weight updates were carried out iteratively to minimize the loss function. The process was optimized using Early Stopping configured with the patience of 10, a technique that monitors the validation metric and automatically stops training when no significant improvements are observed, reducing the risk of overfitting.

The model was implemented using the TensorFlow library, which provided support for GPU training, significantly reducing execution time. The architecture was defined using Keras, which offers an interface for building deep learning models. To evaluate the model's performance, metrics such as accuracy, precision, and recall were used. These metrics are defined by Equations (1), (2) and (3), respectively. Although WER and PER are commonly used in speech recognition tasks, they were not applied in this study because the objective is not to evaluate the correctness of word-level transcription, but rather to classify the severity of speech impairment. Therefore, classification metrics such as accuracy, precision, and recall are more suitable and aligned with the task of identifying dysarthria severity levels.

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

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

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

in which TP (true positive) represents the cases in which the model correctly classified as positive, TN (true negative) refers to the cases correctly identified as negative. FP (false positive) occurs when the model incorrectly classifies a negative case as positive, while FN (false negative) refers to positive cases that were erroneously classified as negative.

The results were analyzed to verify the model's ability to recognize different levels of dysarthric speech, evaluating the LSTM model's capacity to handle the acoustic variability present in the UASpeech dataset, considering different speakers and variations in speech characteristics.

We used the Google Colab virtual environment with 51.0 GB of RAM and disk capacity of up to 235.7 GB, along with a T4 GPU hardware accelerator.

IV. Results

Table 1 shows the results for accuracy, precision, and recall in classifying different levels of dysarthria using the LSTM model with MFCC in UASpeech [5]. Those results seem competitive when compared to other approaches that achieve, for instance, 93% accuracy using DNNs and CNNs [8]

TABLE I
LSTM Model Performance Across Severity Levels in UASpeech [5]

Severity Levels	Very Low	Low	Medium	High
Accuracy	99.39%	99.21%	99.49%	99.60%
Precision	99.30%	98.10%	99.31%	99.28%
Recall	99.15%	98.68%	98.12%	98.23%

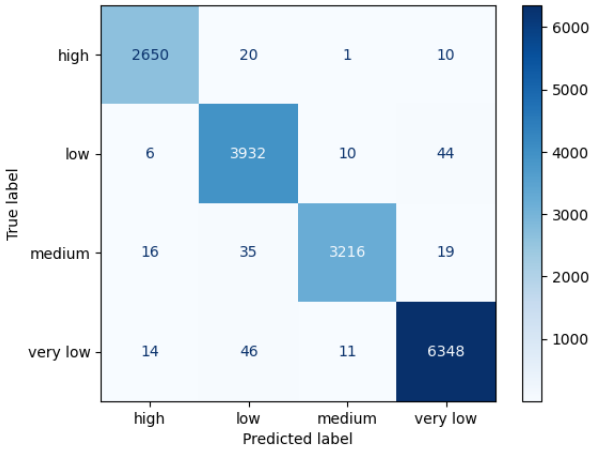

Fig. 2. Confusion Matrix for Severity Classification in UASpeech

Figure 2 shows the results from the LSTM for severity classification in UASpeech [5] using the confusion matrix metric. In the matrix, we observe along the main diagonal the high number of correct predictions for the different levels of dysarthria severity. This demonstrates that the LSTM model is a competitive solution for classifying the different severity levels.

Despite the promising results, some confusion between classes can be observed in Figure 2. Most of the errors in the "High" severity class were predicted as "Low" (20 cases). The "Medium" class had the highest proportion of errors, with 70 misclassified samples, which is 2.1% of its total. These results suggest that while the model is very accurate, it has more difficulty distinguishing between some levels, especially when they are close.

Figure 3 presents a Boxplot graph of the metrics in UASpeech [5], which shows how consistent and effective the model was in performing the proposed task of classifying different severity levels, achieving average values above 98% for all metrics. Accuracy shows low variability and remains consistently high, reflecting the model's overall correctness in predictions. The precision also remains high, though a small outlier is observed, indicating that, in most cases, when the model assigns a severity level, it is likely correct, highlighting its reliability in positive predictions. Recall, while still high, displays greater variation, suggesting that the model may miss a few true instances of certain severity levels. This implies that while the model is precise in its classifications, future improvements could focus on increasing sensitivity to ensure even fewer missed cases. Together, these results indicate a strong and robust model performance across different severity categories.

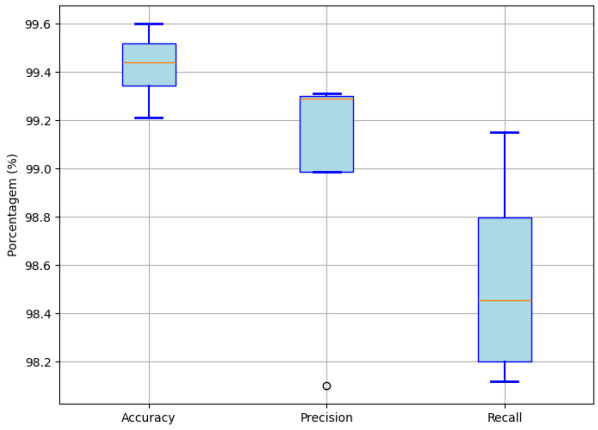

Fig. 3. Boxplot of Model Performance in UASpeech

TABLE II
Distribution of Predicted Severity Levels in the Torgo Dataset

Severity Levels	Very Low	Low	Medium	High
Number of Samples	3335	11414	522	1281

Table 2 presents the distribution of predicted severity levels for the Torgo dataset [12], obtained using an LSTM model previously trained on UASpeech [5], which contains severity labels. Since Torgo does not include severity annotations, they represent an initial automatic categorization aimed at exploring potential stratification of the dataset by severity levels, which could be useful for subsequent analyses. This stratification also contributes to a better understanding of the dataset's composition and can support the development of more efficient and specialized models targeted at specific severity levels of dysarthria.

V. Conclusion

This study explored the performance of the LSTM model in recognizing dysarthria severity levels, using the UASpeech and TORGO datasets. The analysis of the results suggest that LSTM is a promising architecture for capturing the temporal dependencies inherent in dysarthric speech. The model demonstrated strong performance in speech transcription tasks, successfully distinguishing between different severity levels of dysarthria. Accurately determining dysarthria severity is essential for enabling more effective and personalized treatment strategies. Severity plays a critical role in shaping therapeutic

approaches, assisting healthcare professionals in selecting the most appropriate interventions for each patient.

Furthermore, the adoption of hybrid architectures or more advanced models holds potential for further enhancing performance. As future work, one promising direction is the development of specialist models tailored to each severity level of dysarthria, which could further improve classification accuracy and provide more targeted support in clinical applications. Overall, this work contributes to advancing research in the field of Automatic Speech Recognition (ASR) for dysarthric speech, providing valuable insights and representing an initial step towards a broader and more comprehensive study.

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001

REFERENCES

[1] A. R. Busanello, S. A. F. de N. Castro, and A. A. A. Rosa, "Disartria e doença de Machado-Joseph: relato de caso," *Revista da Sociedade Brasileira de Fonoaudiologia*, vol. 12, no. 3, São Paulo, 2007.

[2] N. Hijikata, M. Kawakami, A. Wada, M. Ikezawa, K. Kaji, Y. Chiba, M

[3] K. M. Yorkston, D. R. Beukelman, and C. Traynor, *Assessment of Intelligibility of Dysarthric Speech*. Austin, TX: Pro-Ed., 1984.

[4] B. F. Zaidi, S. A. Selouani, M. Boudraa, and M. S. Yakoub, "Deep neural network architectures for dysarthric speech analysis and recognition," *Neural Comput. Appl.*, 2021.

[5] Z. Qian and K. Xiao, "A survey of automatic speech recognition for dysarthric speech," *Electronics*, vol. 12, no. 20, p. 4278, 2023.

[6] H. Kim *et al.*, "UASpeech," 2023.

[7] M. Kim, B. Cao, K. An, and J. Wang, "Dysarthric speech recognition using convolutional LSTM neural network," in *Proc. Interspeech*, Hyderabad, pp. 2948–2952, Sept. 2018.

[8] A. Mayle, Z. Mou, R. Bunescu, S. Mirshekarian, L. Xu, and C. Liu, "Diagnosing dysarthria with long short-term memory networks," in *Proc. Interspeech*, Graz, Sept. 2019.

[9] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification: A study on acoustic features and deep learning techniques," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, 2022.

[10] Scikit-learn, "LabelEncoder. Version 0.12," 2024. [Online]. Available: https://scikit-learn.org/dev/modules/generated/sklearn.preprocessing.LabelEncoder.html

[11] Python Software Foundation, "Soundfile 0.12.1," 2023. [Online]. Available: https://pypi.org/project/soundfile/

[12] Department of Computer Science, University of Toronto, "TORGO Database," 2010. [Online]. Available: https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html