



Variable STFT Layered CNN Model for Automated Dysarthria Detection and Severity Assessment Using Raw Speech

Kodali Radha¹ · Mohan Bansal² · Venkata Rao Dhulipalla¹

Received: 19 September 2023 / Revised: 10 January 2024 / Accepted: 12 January 2024 /

Published online: 22 February 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024, corrected publication 2024

Abstract

This paper presents a novel approach for automated dysarthria detection and severity assessment using a variable short-time Fourier transform layered convolutional neural networks (CNN) model. Dysarthria is a speech disorder characterized by difficulties in articulation, resulting in unclear speech. The model is evaluated on two datasets, TORGO and UA-Speech, consisting of individuals with dysarthria and healthy controls. Various variations of the CNN's first layer, including spectrogram, log spectrogram, and pre-emphasis filtering (PEF) with and without learnables, are investigated. Notably, the PEF with 5 learnables achieves the highest accuracy in detecting dysarthria and assessing its severity. The study highlights the significance of dataset size, with UA-Speech dataset showing superior performance due to its larger size, enabling better capture of dysarthria severity variations. This research contributes to the advancement of objective dysarthria assessment, aiding in early diagnosis and personalized treatment for individuals with speech disorders.

Keywords Dysarthria severity level assessment · STFT layered CNN · Pre-emphasis filtering · TORGO dataset · UA-Speech dataset

✉ Kodali Radha
radhakodali.vitap@gmail.com

Mohan Bansal
bmohan@iitsonapat.ac.in

Venkata Rao Dhulipalla
dhulipalla0871@gmail.com

¹ Department of Electronics and Communication Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Kanuru, Andhra Pradesh 520007, India

² Electronics and Communication Engineering, Indian Institute of Information Technology Sonapat, IITD Techno Park, Sonapat, Haryana 131001, India

1 Introduction

Effective communication through speech is a fundamental aspect of human interaction that can significantly impact one's quality of life. However, some individuals may experience difficulty with speech production due to neurological causes, resulting in motor speech disorders such as dysarthria [31, 32, 36]. Dysarthria is characterized by weakness, paralysis, or lack of coordination of the motor-speech system, leading to reduced intelligibility, audibility, naturalness, and efficiency of vocal communication [3, 20]. In recent years, there has been an increased interest in developing automated methods for the detection and severity assessment of dysarthria using raw speech, which could potentially improve the diagnosis and treatment of this disorder [39]. Raw waveform models in the classification of dysarthric speech offer several advantages. Raw waveforms capture a more complete representation of the speech signal, including subtle variations in time and frequency, which can help distinguish between healthy and dysarthric speech [35]. By eliminating the need for manual feature extraction, raw waveform models save time and reduce the risk of biased feature selection [25]. Additionally, these models allow for the use of end-to-end approaches, such as convolutional neural networks (CNNs), which can learn complex representations directly from the raw waveform, leading to improved classification accuracy [21]. Raw waveform models provide a practical and straightforward approach for dysarthria assessment by requiring minimal preprocessing of the speech signal [22]. Unlike other methods that involve complex feature extraction [24], these models directly analyze the raw waveform itself, making them easier to implement and interpret [11]. This approach has the potential to enable accurate and efficient dysarthria assessment in both clinical and research settings. A novel model architecture is the short-time Fourier transform (STFT) layered CNN model, which combines the power of CNNs with the STFT representation of the input signal. By extracting meaningful features from both the temporal and spectral domains, the STFT layered CNN model effectively captures essential patterns and variations in dysarthric speech [4, 34]. The model's primary objective is to advance the automation of dysarthria analysis and contribute to the development of more precise and efficient systems for assessing speech disorders.

The subsequent sections of the article are arranged as: Sect. 2 presents the existing literature on dysarthria detection. Section 3 describes the proposed methodology used in the study. Section 4 reports the datasets and experimental results of dysarthria detection and severity assessment using the proposed methodology, showing high accuracy in both tasks. Finally, Sect. 5 concludes the key findings of the suggested work and discusses their implications for future research in the field.

2 Motivation and Related Work

Automated detection and severity assessment of dysarthria using raw speech poses significant challenges in the field of speech processing and healthcare [18]. Dysarthria, characterized by impaired articulation due to muscle weakness or paralysis, is associated with various neurological disorders like Parkinson's disease, stroke, or cerebral palsy [6]. Accurate detection and assessment of dysarthria are crucial for diagnosing

and treating these conditions effectively [19, 33, 38]. Nevertheless, a multitude of methodologies have been employed in recent studies to detect dysarthria and assess its severity across two benchmark datasets: TORGO and UA-Speech.

Raw waveform modeling presents a promising approach for automated dysarthria detection and severity assessment [4, 21]. Upon examining the TORGO dataset, Millet et al. [21] utilized raw waveforms and applied per channel energy normalization (PCEN), long-short-term memory networks (LSTM), and an attention model, achieving an accuracy of 82.4% for dysarthria detection. Similarly, Narendra et al. [22] employed raw glottal flow waveforms, implementing CNN and multi-layered perceptron (MLP), resulting in an accuracy of 81.12%. In contrast, Hernandez et al. [10] employed mel frequency cepstral coefficients (MFCC) coupled with voice quality or prosody, using random forest (RF), support vector machine (SVM), and neural network (NN) models, achieving an accuracy of 75.63% using NN for dysarthria detection. Additionally, Schu et al. [30] used handcrafted features like MFCCs, Mel spectrogram, and sparsity-based features, integrating speech representation learning (SRL) and Mel spectrogram, yielding accuracies ranging from 71.1% to 100.0% for dysarthria detection. Moreover, recent advancements by Joshy et al. [12] using MFCC and constant-Q cepstral coefficients (CQCC) with deep neural network (DNN) models like CNN, gated recurrent unit (GRU), and LSTM achieved a notable accuracy of 96.18% using CNN for dysarthria detection. Radha et al. [23] showcased the effectiveness of employing standard CNN and SincNet on raw waveforms, achieving an accuracy of 95.7% for dysarthria detection. Shifting the focus to the UA-Speech dataset, Gupta et al. [9] utilized onset-offset detection and time-frequency (T-F) representations with CNN and ResNet, achieving an accuracy of 98.9% for severity assessment using short-duration segments. Kachhi et al. [15] employed continuous wavelet transformed (CWT) scalograms with CNN, respectively, achieving accuracies of 87.93% for dysarthria detection and 95.17% for severity assessment.

Compared to traditional feature-based methods, raw waveforms provide a more comprehensive representation of the speech signal, enabling the capture of detailed temporal and spectral characteristics necessary for accurate assessment. Recent technological advancements have facilitated the development of automatic assessment methods utilizing deep learning frameworks. Researchers have introduced innovative approaches for measuring dysarthria severity using sound descriptors [2]. By classifying speech into different severity levels within datasets like the UA-Speech corpus and the TORGO database, they have explored novel avenues in dysarthria assessment. Furthermore, a machine learning-based approach employing bidirectional LSTM has been proposed to classify dysarthric speech as intelligible or non-intelligible [1]. Studies have utilized raw waveform-based CNN models to classify individuals with dysarthria and healthy subjects using the UA-Speech dataset [7, 14, 17, 22]. These models utilize the intricate details found in raw speech waveforms to achieve accurate classification. Through the application of deep learning techniques, the objective is to enhance the recognition of dysarthric individuals within the UA-Speech dataset. Current methods in this field predominantly rely on either feature-based models or complex deep-learning algorithms. To offer straightforward yet effective models that bypass the feature extraction process and enable direct learning by the model, the proposed work

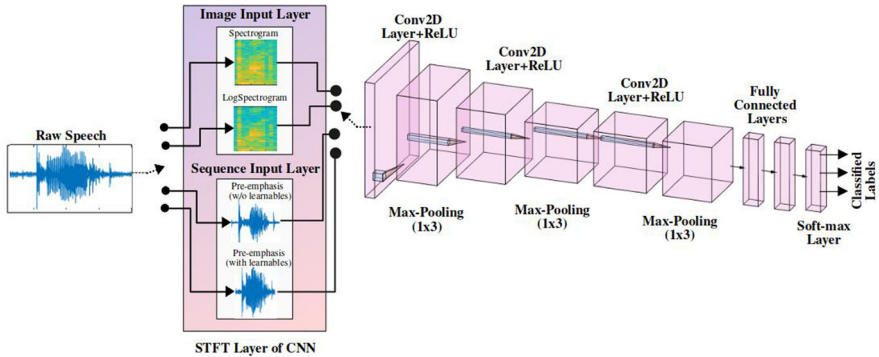


Fig. 1 Flow chart of automated dysarthria detection and dysarthria severity level assessment using STFT layered CNN

introduces a variable STFT layered CNN model. The main contributions of the article are three fold:

- Developed an automated system for dysarthria detection and severity assessment using variable STFT layered CNN model.
- Investigated various variations of the CNN's first layer to enhance accuracy.
- Utilized two datasets, TORGO and UA-Speech, to validate the effectiveness of the proposed approach.

3 Proposed Methodology

3.1 STFT Layered CNN Architecture

This study introduces a novel approach by employing diverse variations in the STFT layer as the initial component of the CNN model. The innovative aspect lies in exploring different configurations of the STFT layer, both with and without learnable parameters. This unique adaptation of the STFT layer within the CNN architecture aims to efficiently capture temporal and spectral information from raw speech waveforms for automated dysarthria detection and severity assessment. The choice of a Hamming window with a length of 1280 and 900 overlap was deliberate to capture specific temporal and spectral features relevant to the raw waveform-based STFT layered CNN. This configuration optimally balances time-frequency resolution while maintaining computational efficiency. The 70% overlap was chosen for a better trade-off between frequency resolution and data redundancy.

3.2 Customized Initial Layers

The representation capabilities of the STFT layered CNN architecture are enhanced by incorporating customized initial layers. These layers are designed to focus on specific aspects of raw waveform modeling, as illustrated in Fig. 1. The architecture includes

image or sequence input layers that have been tailored to suit the task. Additionally, the presence or absence of learnable parameters in these layers allows us to observe the efficiency and performance of the proposed architecture.

3.2.1 Spectrogram-Based STFT Layer

The spectrogram layer is a vital component in converting raw speech waveforms into spectrogram images. It operates without learnable parameters, simplifying the architecture and reducing computational complexity. During the forward pass, the spectrogram layer applies the STFT algorithm to divide the waveform into frames and compute the Fourier transform for each frame [27]. One advantage is its support for automatic differentiation, enabling efficient learning from labeled data. In the prediction phase, the same forward pass is used, seamlessly generating spectrogram images without additional computations. This ensures efficient and accurate predictions on new speech signals.

The equation represents the spectrogram computation using the STFT for a given speech waveform $x(t)$,

$$\text{spec}(t, f) = \text{STFT}(x(t))[f] \quad (1)$$

where, $\text{spec}(t, f)$ denotes the spectrogram value at time index t and frequency bin index f . The STFT takes the speech waveform $x(t)$ and decomposes it into short overlapping frames.

3.2.2 Log Spectrogram-Based STFT Layer

The log spectrogram layer is a useful tool that improves the representation of speech signals. It does this by applying a logarithmic scaling to the spectrogram values, which helps to compress the range of the spectrogram. This compression makes it easier to capture fine details in both soft and loud parts of the speech. It also enhances weak and subtle features in the speech, making them easier to identify [5, 8]. The log spectrogram operation is represented by the equation:

$$\ell_{\text{Spec}}(t, f) = \log(|\text{spec}(t, f)| + \epsilon) \quad (2)$$

Here, $\ell_{\text{Spec}}(t, f)$ represents the log spectrogram value at time frame t and frequency bin f , and $\text{spec}(t, f)$ denotes the corresponding spectrogram value. The term ϵ is a small constant added to ensure numerical stability. Additionally, the logarithmic scaling aids in reducing the impact of background noise and other interfering factors by enhancing the contrast between the desired dysarthric speech components and the noise. Compared to the spectrogram layer, the log spectrogram layer provides a perceptually relevant representation of dysarthric speech [28]. The logarithmic scaling aligns with human auditory perception, as our perception of sound intensity follows a logarithmic nature. By incorporating this perceptual characteristic, the log spectrogram layer effectively captures the important acoustic patterns and structures in dysarthric speech.

3.2.3 Pre-emphasis Filtering (PEF) with and Without Learnables

The STFT layered CNN architecture, when combined with a PEF, incorporates a sequence input layer followed by a 1-D convolution layer. The purpose of the convolution layer is to apply a finite impulse response (FIR) filter with five coefficients. This filtering operation is aimed at extracting relevant features from the input speech signals related to dysarthria. The model leverages these features to automate the detection of dysarthria and assess its severity levels. The inclusion of the pre-emphasis filter is crucial for enhancing the representation of raw audio in the architecture. The inclusion of a PEF in the STFT layered CNN model is motivated by its ability to emphasize higher frequencies and attenuate lower frequencies of speech signals. This adjustment is beneficial because dysarthric speech signals often contain important information in the higher frequency range. By boosting the higher frequencies and reducing the influence of lower frequencies, the PEF enhances the model's ability to capture these significant features and patterns in dysarthric speech [5]. To ensure a linear filtering operation, the bias initializer is set to 'zeros' and the bias learn rate factor is set to 0, keeping the bias fixed during training. The filter weights are initialized using the scaled Kronecker delta sequence [37], resulting in an all-pass filter that does not modify the input signal.

The PEF operation is represented by the equation:

$$y(t) = x(t) - \alpha \cdot x(t - 1) \quad (3)$$

where $y(t)$ is the output signal, $x(t)$ is the input signal, and α is the filter coefficient. In the case of PEF without learnables, a fixed filter coefficient is used throughout training and prediction. This approach improves the performance of dysarthria detection and severity assessment tasks by emphasizing high-frequency components in the input signal. Both the approaches with and without learnables of PEF contribute to the enhanced performance of the STFT layered CNN architecture in these tasks.

3.3 CNN Layers

The CNN layers in the STFT-based architecture play a crucial role in extracting features from the initial layers [26]. These layers consist of a sequence of convolutional operations, batch normalization, and ReLU activation functions. This hierarchical design allows the network to learn intricate patterns and representations from the input spectrogram images or PEF coefficients. The convolutional layers carry out operations that focus on small local regions of the input data. They accomplish this by applying filters that convolve over the data, enabling them to capture spatial relationships and extract important features from the input. Batch normalization is applied after each convolutional layer, normalizing the outputs and improving the stability and speed of training. ReLU activation functions introduce non-linearity, enabling the network to learn complex relationships between the input data and the target output.

In addition to convolution and activation, max-pooling layers are utilized for spatial downsampling. These layers reduce the spatial dimensions of the feature maps, retain-

Table 1 Network configuration of PEF-based CNN model with learnables

Layer index	Layer type	Layer parameters	Learnable parameters
1	Sequence input layer	Name: 'input', Min. length: 8192, Normalization: 'none'	–
2	Conv1d layer	Filter size: 5, No. of filters: 1, Name: 'pre-emphasis-filter', Weights initializer: Kronecker delta, Bias learn rate factor: 0	Filter weights
3	STFT layer	Window: Hamming (1280), Overlap length: 900	–
4 (x1)	Conv2d layer	Filter size: 5, No. of filters: 12	Weights, bias
	Batch norm	–	Offset, scale
	ReLU layer	–	–
	Max pool	Pool size: 3, Stride: 2	–
5 (x2)	Conv2d layer	Filter size: 3, No. of filters: 24	Weights, bias
	Batch norm	–	Offset, scale
	ReLU layer	–	–
	Max pool	Pool size: 3, Stride: 2	–
6 (x4)	Conv2d layer	Filter Size: 3, No. of filters: 48	Weights, bias
	Batch norm	–	Offset, scale
	ReLU layer	–	–
	Max pool	Pool size: 3, Stride: 2	–
–	Drop-out layer	Rate: 20%	–
–	Fully connected layer	Size: Number of output classes	Weights, bias
–	Soft-max layer	–	–

ing the most salient information while discarding redundant details. To mitigate the risk of overfitting, a dropout layer is incorporated into the architecture. Dropout randomly deactivates a fraction of neurons during each training iteration, forcing the network to learn more robust and generalized representations. This regularization technique prevents the network from relying too heavily on specific features and encourages the exploration of different paths in the network. The output of the dropout layer is then connected to a fully connected layer. This layer establishes connections between all neurons and subsequent layers, enabling comprehensive information propagation throughout the network. The details of the network architecture are clearly provided in Table 1.

3.4 Classification Layers

The classification layers in the STFT-based CNN architecture are responsible for making the final predictions. The output of the fully connected layer is passed through a softmax layer, which computes the probabilities of each class. The softmax layer ensures that the predicted probabilities sum up to 1. Finally, the network utilizes a classification layer, which plays a crucial role in identifying the classes within the dataset. This layer enables the network to assign accurate labels to the input samples by leveraging the learned representations and probabilities obtained from the previous layers.

4 Experimental Results

This section presents an evaluation of an automated system designed to detect and assess the severity of dysarthria using raw speech data from two distinct datasets: TORGO and UA-Speech. Two separate experiments were conducted to achieve the objectives of this evaluation. In the first experiment, the system aimed to classify dysarthric speech from healthy control speech, essentially distinguishing between two binary classes. The second experiment focused on classifying the severity level of dysarthria into four distinct categories: very low, low, moderate, and high. This entailed a multi-class classification task. The system utilized the STFT layered CNN. Specifically, four different variations of the CNN's first layer were trained and evaluated for both dysarthria detection experiments. To perform the evaluation, the TORGO and UA-Speech datasets were divided into training sets (80%) and test sets (20%). The trained variants of the STFT layered CNN were then utilized to classify new or unseen speech signals as either dysarthric or healthy control. The performance of the classifiers was assessed using overall accuracy as the evaluation metric.

4.1 Dataset

The study incorporates two datasets to comprehensively evaluate the severity and detection of dysarthria. The inclusion of both moderate and large datasets enables a thorough examination of the performance of the proposed methods. Neural networks, being data-hungry models, benefit greatly from larger datasets as they capture a wider range of patterns and variations within the speech data. Consequently, training the neural networks on larger datasets enhances their ability to provide accurate predictions and robustly generalize to unseen data.

4.1.1 TORGO

The TORGO database [29] is a popular database for investigating dysarthria, a speech disorder that affects articulation. It contains audio recordings and detailed measurements of speech movements from individuals with dysarthria and a control group of healthy individuals. The database includes data from 8 dysarthric speakers and 7

Table 2 Details of TORGO and UA-Speech dataset in the detection of dysarthria and severity level assessment

Dataset	Details of dataset		Dysarthria severity level assessment (multi-class)				
	Dysarthria	Healthy control	Very low	Low	Moderate	High	
TORGO	No. of speakers	7	3	2	3	-	
	Gender Composition	3F/4M	2F/1M	1F/1M	3M	-	
	Speaker IDs	F01, F03, F04, M01, M02, M03, M04, M05	FC01, FC03, FC04, MC01, MC02, MC03, MC04, MC05	F03, F04, M03	F01, M05	M01, M02, M04	-
UA-Speech	No. of speakers	13	5	3	3	4	
	Gender composition	4F/9M	1F/4M	1F/2M	1F/2M	1F/3M	
	Speaker IDs	F02, F03, F04, F05, M01, M04, M05, M07, M08, M09, M10, M11, M12, M14, M16	CF02, CF03, CF04, CF05, CM01, CM04, CM05, CM06, CM08, CM09, CM10, CM12, CM13	F05, M08, M09, M10, M14	F04, M05, M11	F02, M07, M16	F03, M01, M04, M12

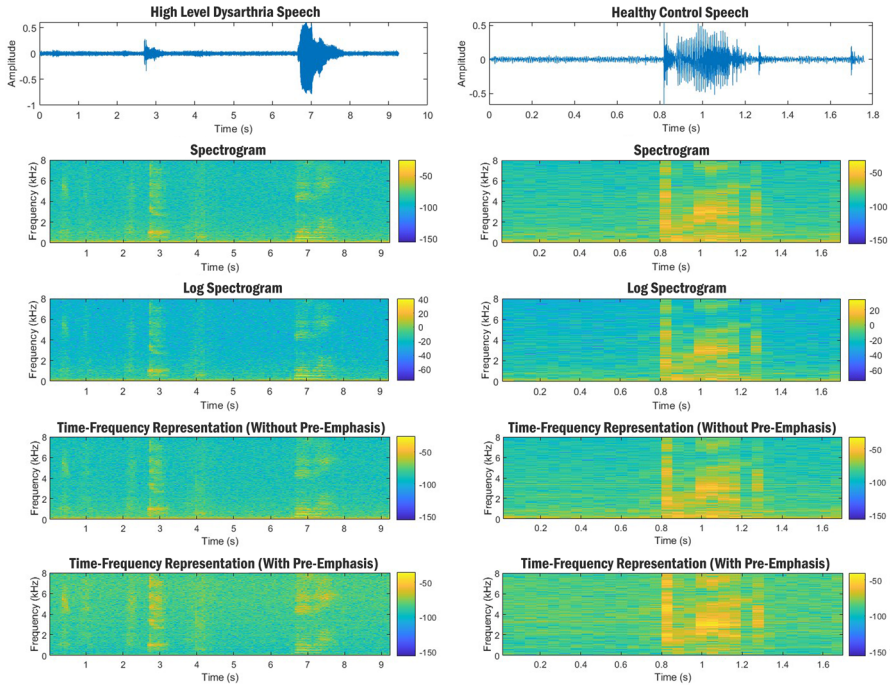
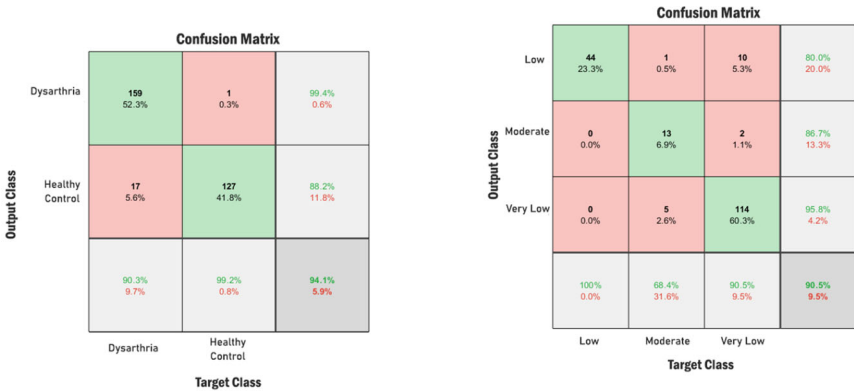


Fig. 2 Investigating STFT layered CNN variants for differentiating high-level dysarthria and healthy control speech on the word “*Command*”

healthy control speakers, all aged between 16 to 50 years as shown in Table 2. The recordings in the database provide rich information on speech production, including both sound data and detailed measurements of speech movements.

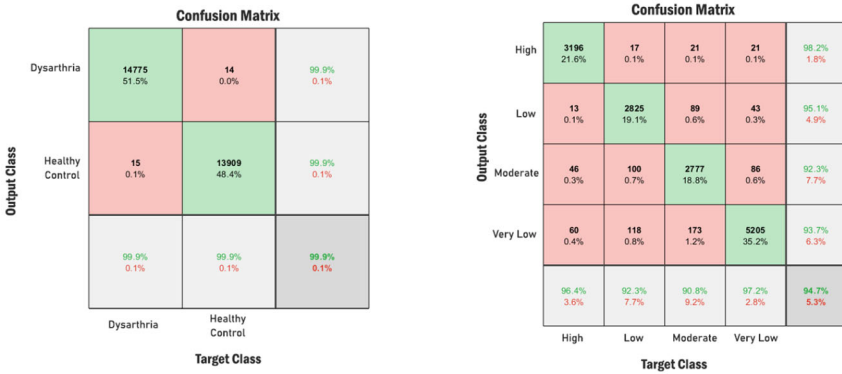
4.1.2 UA-Speech

The UA Speech corpus comprises recordings from 13 healthy control speakers and 15 dysarthric speakers with cerebral palsy [16]. The dataset consists of 455 diverse words, including digits, international radio alphabets, computer commands, common words, and uncommon words. The words were recorded in three blocks, ensuring coverage of all categories for each speaker. In total, each speaker recorded 765 isolated words. The corpus also includes speech intelligibility ratings provided by five impartial listeners for each dysarthric speaker. These ratings serve as an objective assessment of the speakers’ ability to be understood. Based on these ratings, the speakers were classified into four distinct groups: high, moderate, low, and very low, representing varying levels of speech intelligibility as shown in Table 2. Figure 2 shows that individuals with dysarthria struggle to produce speech due to a diminished ability to control their articulatory movements. This is due to existing issues like prosodic abnormalities and resonance changes, which can cause slurred or mumbled speech. On the other hand, healthy individuals do not experience these difficulties and are able to produce clear



(a) Dysarthria detection in TORGO dataset

(b) Dysarthria severity level assessment in TORGO dataset



(c) Dysarthria detection in UA-Speech dataset

(d) Dysarthria severity level assessment in UA-Speech dataset

Fig. 3 Confusion matrices of dysarthria detection and severity assessment using learned PEF-based STFT layered CNN

speech. This valuable information regarding the speakers’ intelligibility and severity levels is utilized to classify and assess the severity of dysarthria.

4.2 Investigating Variations of STFT Layered CNN Model

This study focuses on exploring different variations of STFT layered CNN models for the analysis of raw waveform data, specifically in automatic dysarthria detection and severity level assessment. The goal of this research is to advance the development of automated systems for analyzing speech disorders. To illustrate the differences between dysarthric and healthy speech, Fig. 2 is included. This figure presents T-F representations of the word “command” for both dysarthric and healthy speakers. It is observed that the high-level dysarthric subject took approximately 8 s to utter the word, while a healthy subject took only 1.4 s for the same word. Furthermore,

Table 3 Performance evaluation of STFT layered CNN models for TORGO and UA-Speech datasets in terms of accuracy (%)

Model	First layer	TORGO		UA-Speech	
		Dysarthria detection (binary-class)	Dysarthria severity level assessment (multi-class)	Dysarthria detection (binary-class)	Dysarthria severity level assessment (multi-class)
STFT-layered CNN	Spectrogram layer (w/o learnables)	88.39	84.10	94.97	90.35
	LogSpectrogram layer (w/o learnables)	92.53	85.16	97.12	91.63
	Pre-emphasis filtering (w/o learnables)	92.76	87.34	97.9	93.45
	Pre-emphasis filtering (5 learnables)	94.08	90.50	99.89	94.67

the spectrograms, log spectrograms, and T-F representations with and without PEF were extracted from the first layer of the CNN. These representations highlight the variations between dysarthric and healthy speech, allowing for a visual comparison of the differences in the T-F domain.

4.2.1 Automatic Dysarthria Detection

Initially, the first layer of the CNN was modified as a spectrogram layer without learnables, resulting in an accuracy of 88.39% for the TORGO dataset and 94.97% for the UA-Speech dataset. Subsequently, by utilizing the log spectrogram, there was a noticeable improvement of 92.53% and 97.12% respectively for both datasets in the detection accuracy. This improvement can be attributed to the logarithmic scaling of the spectrogram, which enhances the representation of lower-intensity frequency components and improves the discrimination between dysarthric and healthy speech signals. Furthermore, PEF without learnables in the STFT-based CNN yielded comparable accuracy to the log spectrogram representation. This is because PEF enhances the high-frequency components of the speech signal, enabling the CNN model to capture discriminative features effectively for dysarthria detection. The highest accuracy in dysarthria detection of UA-Speech was achieved using PEF with 5 learnables, specifically with values of 0.4714, -0.0431 , -0.0247 , 0.0015, and -0.0076 . The accuracy achieved for the TORGO dataset was 94.08%, while for the UA-Speech dataset it was 99.89%. Examining the confusion matrix in Figs. 3a and c, we observed that in the TORGO dataset, 17 dysarthria classes were misclassified as healthy subjects. However, in the UA-Speech dataset, due to its larger size, even when 15 dysarthria subjects were misclassified as healthy, it did not significantly affect the overall accuracy of the model. Consequently, with the inclusion of learnable PEF, the final accuracy reached 99.89%, as shown in Table 3. Figure 4 also demonstrates the effectiveness of incorporating learnable parameters in PEF for improving dysarthria detection performance.

4.2.2 Automatic Severity Level Assessment

The accuracy of dysarthria severity level assessment varied across different techniques and datasets. In the TORGO dataset, consisting of three severity levels, the initial spectrogram layer achieved an accuracy of 84.10%, which increased to 85.16% with the log spectrogram layer without learnables. To enhance the model's performance, we incorporated PEF without learnables, which further improved the accuracy to 87.34% for TORGO. In the UA-Speech dataset, with four severity levels, the accuracy improved from 90.35% (spectrogram layer) to 91.63% (log spectrogram layer). With the inclusion of PEF utilizing 5 learnable parameters, the accuracy rates further improved to 90.5% for the TORGO dataset and 94.67% for the UA-Speech dataset. An examination of the results revealed that within the TORGO dataset, a notable proportion of misclassifications occurred specifically at the moderate and very low levels of dysarthria. This trend is visually depicted in Fig. 3b, illustrating the distribution of misclassifications across different severity levels. One possible reason for this observation could be the similarity in acoustic characteristics between these two severity levels, which poses challenges in accurately distinguishing between them.

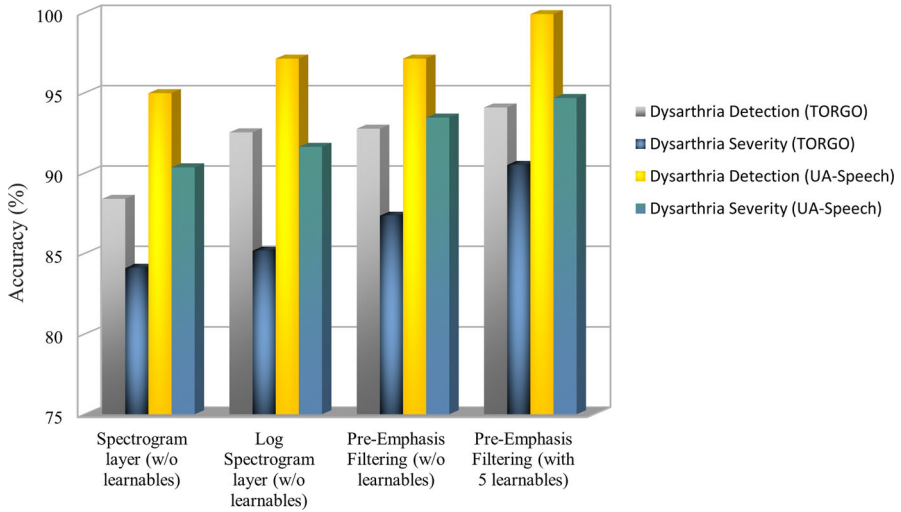


Fig. 4 Performance analysis of STFT layered CNN variants for both datasets

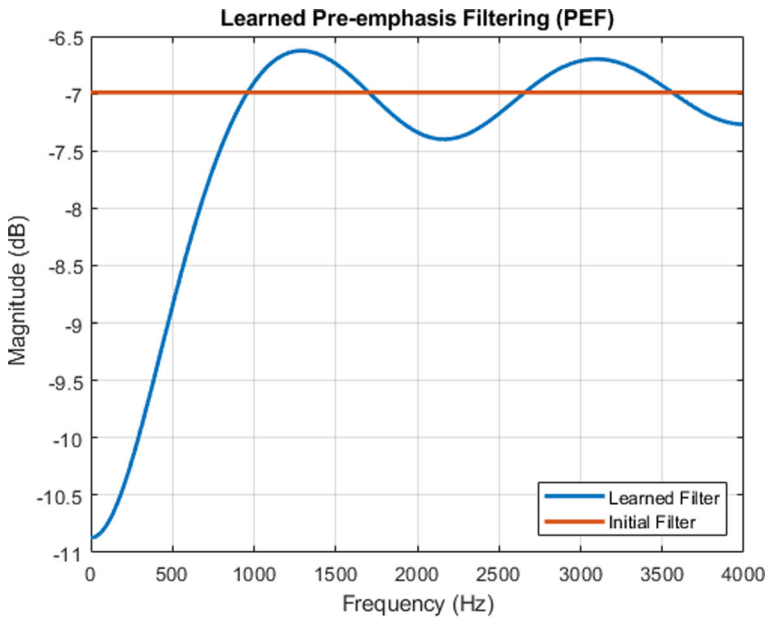


Fig. 5 Frequency response analysis of learned pre-emphasis filtering with 5 learnables

The UA-Speech dataset exhibited a distinct pattern in which the majority of misclassifications were concentrated within specific severity levels. Particularly, misclassifications occurred in the moderate and low levels of dysarthria, accounting for 9.2% and 7.7% respectively. This observation is visually presented in Fig. 3d, which illustrates the distribution of misclassifications across different severity levels in the

Table 4 Comparison of proposed model performance with previous approaches on TORGO dataset

Year/author	Dataset	Front-end approach	Model	Accuracy
2020 Millet et al. [21]	TORGO	Raw waveforms	PCEN, LSTM and attention model	82.4% for dysarthria detection
2020 Narendra et al. [22]	TORGO	Raw glottal flow waveforms	CNN and MLP	81.12% for dysarthria detection
2020 Hernandez et al. [10]	TORGO	MFCCs with voice quality or prosody	RF, SVM, NN	75.63% using NN for dysarthria detection
2022 Schu et al. [30]	TORGO	Handcrafted features like MFCCs, Mel spectrogram, Sparsity-based features	Speech representation learning (SRL) + Mel spectrogram	71.1% to 100.0% for dysarthria detection
2022 Joshy et al. [12]	TORGO	MFCC and CQCC	DNN models like CNN, GRU, LSTM	96.18% using CNN for dysarthria detection
2023 Radha et al. [23]	TORGO	Raw waveforms	Standard CNN and SincNet	95.7% for dysarthria detection
Proposed Method	TORGO	Raw waveforms	Learnable PEF based STFT layered CNN	94.62% for dysarthria detection and 90.15% for severity level assessment

UA-Speech dataset. Additionally, to ensure accurate frequency response, we initialized the filter to a scaled Kronecker delta sequence, acting as an all-pass filter as shown in Fig. 5. The observed higher accuracy in the UA-Speech dataset can be attributed to its larger size, allowing for better capturing of variations in dysarthria severity levels and patterns.

4.3 Comparative Analysis with Previous Approaches

In this section, a comparative analysis is presented, highlighting the proposed system's performance in dysarthria detection and severity level assessment within the domain. The system's evaluation is conducted using two datasets, namely TORGO and UA-Speech, to compare its outcomes against existing approaches. Notably, the proposed system integrates a novel approach, combining a learned PEF technique with STFT layered CNN. This integration showcases substantial advancements in

Table 5 Comparison of proposed model performance with previous approaches on UA-Speech dataset

Year/author	Dataset	Front-end approach	Model	Accuracy
2020 Gupta et al. [9]	UA-Speech	Onset-offset detection, T-F representations	CNN and ResNet	98.9% for severity assessment using short duration segments
2021 Narendra et al. [22]	UA-Speech	Raw glottal flow waveforms	CNN and MLP	87.93% for dysarthria detection
2021 Kachhi et al. [15]	UA-Speech	CWT Scalograms	CNN	95.17% for severity assessment
2023 Joshy et al. [13]	UA-Speech	Mel-spectrograms	Squeeze & Excitation (SE) CNN	97.58% for dysarthria detection
2023 Joshy et al. [14]	UA-Speech	Raw waveform	Multi-head attention based CNN	95.75% for dysarthria detection
Proposed Method	UA-Speech	Raw waveforms	Learnable PEF based STFT layered CNN	99.89% For dysarthria detection and 94.67% For severity level assessment

dysarthria detection and severity level assessment, as indicated in the respective results presented in Tables 4 and 5 for TORGO and UA-Speech datasets, respectively.

5 Conclusion

This article demonstrated the effectiveness of different variants of the STFT layered CNN model for dysarthria detection and severity assessment. The spectrogram layer provided insights into the frequency content of speech, while the log spectrogram layer enhanced detection accuracy by highlighting subtle acoustic details. The inclusion of PEF variants, with and without learnables, amplified relevant speech components and contributed to accurate analysis. Overall, these variants optimized dysarthria detection and severity assessment, with the PEF variant achieving the highest accuracy of 99.89% on the UA-Speech dataset. This research contributes to the development of advanced techniques for objective dysarthria analysis, enabling early diagnosis and personalized treatment for individuals affected by this speech disorder.

Funding This research received no external funding.

Data Availability Statement The open access TORGO data that support the findings of this study is available from the Kaggle repository. The University of Illinois team provided the UA-Speech data upon request. More details about the data are given in Sect. 4.1.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. C. Bhat, H. Strik, Automatic assessment of sentence-level dysarthria intelligibility using BLSTM. *IEEE J. Select. Top. Signal Process.* **14**(2), 322–330 (2020)
2. C. Bhat, B. Vachhani, S.K. Koppurapu, *Automatic assessment of dysarthria severity level using audio descriptors*, in *IEEE International Conference on Acoustics (Speech and Signal Processing (ICASSP)* (IEEE, 2017), pp. 5070–5074
3. M. Carl, E.S. Levy, M. Icht, Speech treatment for Hebrew-speaking adolescents and young adults with developmental dysarthria: a comparison of mSIT and Beataalk. *Int. J. Lang. Commun. Disord.* **57**(3), 660–679 (2022)
4. H. Chandrashekar, V. Karjigi, N. Sreedevi, Spectro-temporal representation of speech for intelligibility assessment of dysarthria. *IEEE J. Sel. Top. Signal Process.* **14**(2), 390–399 (2019)
5. H. Chandrashekar, V. Karjigi, N. Sreedevi, Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech. *IEEE Trans. Neural Syst. Rehabil. Eng.* **28**(12), 2880–2889 (2020)
6. P. Enderby, Disorders of communication: dysarthria. *Handb. Clin. Neurol.* **110**, 273–281 (2013)
7. J. Fritsch, M. Magimai-Doss, Utterance verification-based dysarthric speech intelligibility assessment using phonetic posterior features. *IEEE Signal Process. Lett.* **28**, 224–228 (2021)
8. A. Gallardo-Antolín, J.M. Montero, On combining acoustic and modulation spectrograms in an attention LSTM-based system for speech intelligibility level classification. *Neurocomputing* **456**, 49–60 (2021)
9. S. Gupta, A.T. Patil, M. Purohit et al., Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Netw.* **139**, 105–117 (2021)
10. A. Hernandez, S. Kim, M. Chung, Prosody-based measures for automatic severity assessment of dysarthric speech. *Appl. Sci.* **10**(19), 6999 (2020)
11. A.K. Jardine, D. Lin, D. Banjevic, A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signal Process.* **20**(7), 1483–1510 (2006)
12. A.A. Joshy, R. Rajan, Automated dysarthria severity classification: a study on acoustic features and deep learning techniques. *IEEE Trans. Neural Syst. Rehabil. Eng.* **30**, 1147–1157 (2022)
13. A.A. Joshy, R. Rajan, Dysarthria severity assessment using squeeze-and-excitation networks. *Biomed. Signal Process. Control* **82**, 1–13 (2023)
14. A.A. Joshy, R. Rajan, Dysarthria severity classification using multi-head attention and multi-task learning. *Speech Commun.* **147**, 1–11 (2023)
15. A. Kachhi, A. Therattil, P. Gupta et al, *Continuous wavelet transform for severity-level classification of dysarthria*, in *International Conference on Speech and Computer* (Springer, 2022), pp. 312–324
16. H. Kim, M. Hasegawa-Johnson, A. Perlman et al, *Dysarthric speech database for universal access research*, in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH, 2008)*, pp. 1741–1744
17. D. Korzekwa, R. Barra-Chicote, B. Kostek et al, Interpretable deep learning model for the detection and reconstruction of dysarthric speech. [arXiv:1907.04743](https://arxiv.org/abs/1907.04743) (2019)
18. S. Latif, J. Qadir, A. Qayyum et al., Speech technology for healthcare: opportunities, challenges, and state of the art. *IEEE Rev. Biomed. Eng.* **14**, 342–356 (2020)
19. S.K. Maharana, A. Illa, R. Mannem et al., *Acoustic-to-articulatory inversion for dysarthric speech by using cross-corpus acoustic-articulatory data*, in *IEEE International Conference on Acoustics. (Speech and Signal Processing (ICASSP)* (IEEE, 2021), pp. 6458–6462

20. V. Mendoza Ramos, The added value of speech technology in clinical care of patients with dysarthria. Ph.D. thesis, University of Antwerp (2022)
21. J. Millet, N. Zeghidour, *Learning to detect dysarthria from raw speech*, in *IEEE International Conference on Acoustics. (Speech and Signal Processing (ICASSP))* (IEEE, 2019), pp. 5831–5835
22. N. Narendra, P. Alku, Glottal source information for pathological voice detection. *IEEE Access* **8**, 67745–67755 (2020)
23. K. Radha, M. Bansal, *Automated detection and severity assessment of dysarthria using raw speech*, in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (2023a), pp 1–7. <https://doi.org/10.1109/ICCCNT56998.2023.10307923>
24. K. Radha, M. Bansal, Feature fusion and ablation analysis in gender identification of preschool children from spontaneous speech. *Circuits Syst. Signal Process.* **42**(10), 6228–6252 (2023)
25. K. Radha, M. Bansal, Towards modeling raw speech in gender identification of children using sincNet over ERB scale. *Int. J. Speech Technol.* **26**(3), 641–650 (2023)
26. K. Radha, M. Bansal, R.B. Pachori, Speech and speaker recognition using raw waveform modeling for adult and children’s speech: a comprehensive review. *Eng. Appl. Artif. Intell.* **131**(107), 661 (2024). <https://doi.org/10.1016/j.engappai.2023.107661>
27. S. Reza, M.C. Ferreira, J. Machado et al., A customized residual neural network and bi-directional gated recurrent unit-based automatic speech recognition model. *Expert Syst. Appl.* **215**(119), 293 (2023)
28. P. Roussel, Analysis of cortical activity for the development of brain-computer interfaces for speech. Ph.d. thesis, Université Grenoble Alpes (2021)
29. F. Rudzicz, A.K. Namasivayam, T. Wolff, The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Lang. Resour. Eval.* **46**, 523–541 (2012)
30. G. Schu, P. Janbakhshi, I. Kodrasi, On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches. [arXiv:2211.08833](https://arxiv.org/abs/2211.08833) (2022)
31. S.M. Shabber, M. Bansal, K. Radha, *Machine learning-assisted diagnosis of speech disorders: a review of dysarthric speech*, in *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)* (2023a), pp. 1–6. <https://doi.org/10.1109/ELEXCOM58812.2023.10370116>
32. S.M. Shabber, M. Bansal, K. Radha, *A review and classification of amyotrophic lateral sclerosis with speech as a biomarker*. in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (2023b), pp 1–7. <https://doi.org/10.1109/ICCCNT56998.2023.10308048>
33. B. Suhas, D. Patel, N.R. Koluguri et al, *Comparison of speech tasks and recording devices for voice based automatic classification of healthy subjects and patients with amyotrophic lateral sclerosis*. (INTERSPEECH, 2019), pp. 4564–4568
34. B. Suhas, J. Mallela, A. Illa et al, *Speech task based automatic classification of als and parkinson’s disease and their severity using log mel spectrograms*, in *2020 International Conference on Signal Processing and Communications (SPCOM)* (IEEE, 2020), pp. 1–5
35. N. Tavabi, D. Stück, A. Signorini et al., Cognitive digital biomarkers from automated transcription of spoken language. *J. Prevent. Alzheimer’s Dis.* **9**(4), 791–800 (2022)
36. M.J. Vansteensel, E. Klein, G. van Thiel et al., Towards clinical application of implantable brain-computer interfaces for people with late-stage ALS: medical and ethical considerations. *J. Neurol.* **270**(3), 1323–1336 (2023)
37. P.W. Wong, N. Moayeri, C. Herley, *Optimum pre-and post-filters for robust scalar quantization*, in *Proceedings of Data Compression Conference-DCC’96* (IEEE, 2022), pp. 240–249
38. K.M. Yorkston, Treatment efficacy: dysarthria. *J. Speech Lang. Hear. Res.* **39**(5), S46–S57 (1996)
39. Z. Yue, E. Loweimi, H. Christensen, et al., *Dysarthric speech recognition from raw waveform with parametric CNNs*, in *Proceedings of INTERSPEECH 2022. ISCA-INST SPEECH COMMUNICATION ASSOC* (2022)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.