Journal of Theoretical and Applied Information Technology
15th August 2025. Vol.103. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DYSARTHRIC SPEECH RECOGNITION USING WAVENET CONVOLUTIONAL NEURAL NETWORK WITH PROSODY CONSISTENCY LOSS FUNCTION

ANEETA S ANTONY[1], ROHINI NAGAPADMA[2], AJISH K ABRAHAM[3]

[1]Department of Electronics and Communication, The National Institute of Engineering, Mysuru, India

[2]Department of Electronics and Communication, The National Institute of Engineering, Mysuru, India

[3]Department of Electronics, All India Institute of Speech and Hearing, Mysuru, India

E-mail: [1]aneetaantony123@gmail.com, [2]rohini nagapadma@nie.ac.in, [3]ajish68@aiishmysore.in

ABSTRACT

Dysarthric speech recognition concentrates on understanding speech impairments caused by neurological disorders. Speech recognition increases communication by enhancing adaptability and clarity for individuals. However, dysarthric speech recognition struggles to transcribe speech accurately due to variations in pitch, articulation, and rhythm that significantly differ from typical speech and processing dysarthric across different speakers. In this research, the WaveNet Convolutional Neural Network with Prosody Consistency Loss Function (WNCNN-PCLF) is proposed to recognise and classify dysarthric speech accurately. In CNN, WaveNet is incorporated for capturing local speech features that enhance the model's ability to determine intricate variations and patterns in distorted speech. The PCLF assist in preserving natural speech patterns, which makes for more accurate rhythm and tone representation. Therefore, this integration enables better adaptation to dysarthric speech, which addresses both prosody and articulation issues effectively. Hence, the proposed WNCNN-PCLF achieves a high accuracy of 99.92% and 98.34% using UA-Speech and Kannada datasets compared to existing methods like Densely Squeezed and excitation Attention-gated Network (DySARNet).

Keywords: *Dysarthric Speech Recognition, Local Speech, Prosody Consistency Loss Function, Speakers, Wavenet Convolutional Neural Network.*

1. INTRODUCTION

Dysarthria is a neuro-motor disorder resulting from neurological damage to the motor element of speech production. It is primarily produced by an acquired or congenital neurological problem like brain tumor, cerebral palsy, stroke, brain injury, or neurodegenerative diseases like Huntington disease, amyotrophic lateral sclerosis, or Parkinson's disease [1] [2]. Dysarthric speech is primarily characterised by abnormalities in resonatory, phonatory, prosodic, and articulatory features of speech production, which influence speech intelligibility. Speech-language pathologists perform in clinical settings by utilizing standard intelligibility tests to enhance speech quality and intelligibility [3]. This process involves detaching speech signals and eliminating distortions from noisy speech [4]. It involves numerous origins and a multitude of probable speech patterns from slight modifications to complete incomprehensibility. Individuals with dysarthria have difficulties associated with voice and pronunciation that obstruct their capability to communicate efficiently [5]. The alterations in dysarthric speaker's speech are caused by neurological muscle impairments that affect speech production, which leads to neurological disorders or cognitive disabilities. Such disorders disrupt fluency, pronunciation, minimise human intelligibility and affect the verbal expression of emotions, resulting in social isolation [6]. In the speech subsystem, muscles and muscle groups are effectively coordinated with space and time for speech production, which renders dysarthric speech normally unintelligible. Higher the dysarthria severity, intelligibility of dysarthria speech is lower [7] [8].

The neurological damage affects the function of the speech-motor that impacts physical activities related to motor neurons as well. Human interface with devices and gadgets comprises typing into a keyboard by utilizing hand movement, which is slowed down by a factor of 150 to 300 in dysarthria severe cases in comparison with regular users [9] [10]. Moreover, dysarthric speech is slow by 10 to

17 factors compared to regular speech, around fifteen words per minute in most severe cases. Additionally, it is determined that dysarthric speakers show better prosodic control, which helps to increase communication effectiveness [11]. Over the past decade, researchers have made important strides in severity assessments and dysarthric classification. It determined numerous speech features like pitch variation, speech rate, phonation quality, and articulation precision that help in differentiating various kinds and dysarthria's severity levels. Though, despite these advancements, there are still gaps in understanding which features are most significant for accurately classifying and characterising the severity of dysarthria [12]. However, Deep Learning (DL) provide a promising solution for bridging these communication gaps to determine intricate patterns in speech data [13]. This makes it better to manage speech variations, including background noise and accents. Compared with conventional methods, DL-based methods present significant enhancements in speech performance [14] [15].

1.1 Problem Statement

Despite advancements in dysarthric speech recognition, existing methods struggles with accurately capturing dysarthric speech complexity especially in the variations of articulation, prosody, and rhythm. Dysarthric speech significantly deviates from typical speech patterns which enable challenges for traditional methods to generalize across distant speakers and severity levels. Furthermore, most of the existing methods focus on spectral features without preserving natural prosodic features which are crucial for emotion perception. Therefore, there is a critical need for a more robust and comprehensive model which is effectively process and recognize prosodic features effectively across numerous speakers and languages.

1.2 Objective

The main objective of this research is to enhance recognition accuracy of dysarthric speech which is often impaired because of variations in prosody, articulation, and rhythm. To solve this issue, the research proposes novel DL method as WaveNet Convolutional Neural Network with Prosody Consistency Loss Function (WNCNN-PCLF) for recognising and classifying dysarthric speech accurately by leveraging WaveNet CNN's structure in capturing intricate temporal dependencies in dysarthric speech. The PCLF ensures stability in articulation, pitch, and rhythm variations across various speakers. This improves the model's ability in learning speaker-independent speech patterns while conserving phonetic integrity.

1.3 Scope and Contributions

This research concentrates on development of DL based methods for dysarthric speech recognition by utilizing a combination of WaveNet and CNN models with PCLF. The scope is limited to binary classification using two datasets like UA-Speech and Kannada by considering prosodic and articulatory distortions.

The primary contributions are discussed below in detail:

- In CNN, WaveNet is integrated in modelling long-range dependencies, which enables it to understand intricate speech patterns over time, which is essential for dysarthric speech recognition.
- PCLF is applied to preserve natural speech patterns like rhythm, stress, and intonation which results in better alignment with dysarthric speech variations that enhance intelligibility.
- The Mel-Frequency Cepstral Coefficient (MFCC), Linear prediction cepstral coefficient, spectral flux, spectral centroid, spectral crest, and pitch chroma are used to extract features by capturing both spectral and temporal characteristics of the speech signal. This enhances the model's ability to differentiate between dysarthric and healthy speech variations.
- Compared to existing methods like Densely Squeezed and excitation Attention-gated Network (DySARNet), proposed WNCNN-PCLF achieves a superior accuracy of 99.92% and 98.34% on UA-Speech and Kannada datasets. This improvement is because of integration of WaveNet with CNN that effectively captures spectral features in dysarthric speech. Moreover, PCLF improves model's ability in preserving natural prosodic features that enhance recognition accuracy.

This paper is organised as follows: Section 2 involves literature survey and Section 3 provides proposed methodology. Section 4 analyses experimental results, and conclusion is given in Section 5.

2. LITERATURE SURVEY

Usama Irshad et al. [16] introduced a UTrans encoder-decoder model to analyze Mel-

spectrograms and classify speech as dysarthric or healthy. The introduced approach utilized a transformer encoder based on a hybrid model that contains Vision Transformer (ViT) encoders and Feature Enhancement Block (FEB). This integration extracted local and global pixel information based on localization effectively through optimizing Mel-spectrograms. The consecutive residual connections were included in the system, which minimises feature loss when enhancing spatial data retrieval. However, UTrans face challenges in capturing subtle articulatory distortions due to self-attention does not effectively differentiate between dysarthric and healthy speech patterns.

Francis Jesmar P. Montalbo [17] presented a Densely Squeezed and excitation Attention-gated Network (DySARNet) to diagnose dysarthria and severity estimation. DySARNet was used by integrating Lean Separable Dense Block (LSDB) increase feature reuse by maintaining a large parameter increase via Separable Depthwise Convolution (SDWConv). To enhance awareness and context understanding, DySARNet utilises a Squeeze and Excited Lightweight Residual Attention-gated (SELRA) approach through squeezing extra parameters via SDWConv and depthwise convolution. Nevertheless, DySARNet struggled with generalising across diverse patterns due to attention gating overfitting to dominant acoustic features ignores subtle severity variations.

Kodali Radha et al. [18] developed a variable Short-Time Fourier Transform (STFT) layered Convolutional Neural Network (CNN) to detect dysarthria and analyze severity assessment. STFT layered CNN was applied to extract significant features from both spectral and temporal domains that capture necessary patterns and variations in dysarthric speech. The main goal of the developed method was to automate the assessment of dysarthria and establish more accurate as well as effective systems for evaluating speech disorders. However, STFT with CNN struggled with capturing fine-grained spectral features because of fixed-time frequency resolution, which loses significant frequency information to distinguish speech variations.

Rabbia Mahum et al. [19] suggested a hybrid model with the combination of ensemble deep networks and a transformer encoder scheme to recognize dysarthria speech. Ensemble learning plays a significant role in extracting the features from Mel-spectrograms. Two scenarios were employed that contain VGG16, GoogleNet, DenseNet201, whereas the ensemble comprises Xecption, DenseNet 201, and nception ResNetV2. The transformer model was established using self-attention mechanism that enables the network to focus on significant information with Multilayer Perceptron (MLP) to recognise speech accurately. By using this hybrid method, effective and accurate disease determination was attained. Nevertheless, the hybrid model struggled with high computational complexity and less inference time due to the requirement of processing huge volumes of speech data.

Bhuvaneshwari Jolad and Rajashri Khanai [20] established a Fractional Competitive Crow Search Approach-based Speech Enhancement Generative Adversarial Network (FCCSA-SEGAN) to enhance speech signals. At first, noise from the speech signal was eliminated utilizing the spectral subtraction method. Then, the signal was passed through speech enhancement, where the quality of the signal was enhanced by SEGAN, which was trained by FCCSA. By the inclusion of the Competitive Crow Search Approach (CSSA) and Fractional Calculus (FC), FCCA was attained in that CSSA was a hybrid of CSSA and Competitive Swarm Optimizer (CSO). However, FCCSA-SEGAN's reliance on fractional optimisation results in slower convergence and less efficiency in intricate speech environments.

Shaik Mulla Shabber et al. [21] suggested a fine-tuned DL method to detect dysarthric speech effectively. Pre-processing methods like normalization and noise reduction were used to increase raw speech signal quality and extracted appropriate features. Scalogram images were generated by wavelet transform which capture the characteristics of time-frequency effectively in speech signal that offers visual representation over time. This provide significant insights into speech abnormalities in dysarthria.

Although different models are discussed in recent literature including UTrans encoder-decoder model [16], DySARNet [17], STFT-CNN [18], ensemble model [19], FCCSA-SEGAN [20], and Fine-tuned DL [21]. However, these methods struggled with distant limitations like difficulties in subtle articulatory distortions, overfitting to dominant features, challenges in accurately capturing dysarthric speech complexity especially in the variations, capturing fine-grained spectral features, and high computational resources. To address this issue, WNCNN-PCLF is proposed by capturing local spectral information that makes better modeling of dysarthric speech complexity. PCLF imporves model's ability in retaining natural rhythm which address prosodic variability. This minimize

overfitting and enhance generalization across different speech patterns.

3. PROPOSED METHODOLOGY

In this research, the WNCNN-PCLF is proposed to recognise and classify dysarthric speech in the Kannada and English languages. Initially, UA-Speech and Kannada datasets are considered to evaluate the model's performance. The spectral subtraction is applied to remove the noise from the obtained speech. Subsequently, MFCC, LPCC, spectral crest, spectral flux, pitch chroma, and spectral centroid are the features that are extracted. Finally, the proposed WNCNN-PCLF is used for recognition and classification of dysarthria and healthy. Figure 1 indicates a block diagram for the proposed WNCNN-PCLF.

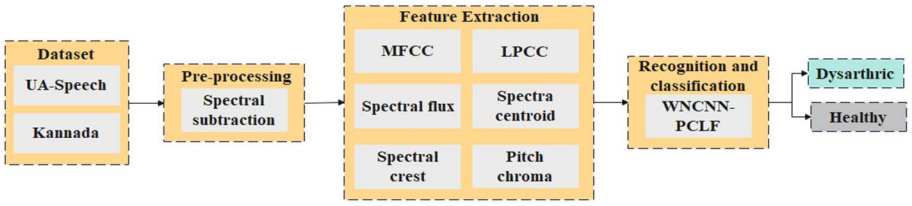

Figure 1: Block diagram for proposed WNCNN-PCLF

3.1 Datasets

This research employs UA-Speech [22] and Kannada datasets to determine the performance of the proposed model in dysarthric speech recognition. The UA-Speech dataset provides a standardised benchmark for determining speech impairments, while the Kannada dataset ensures language-specific analysis. By leveraging both datasets, this research enhances the robustness and accuracy of dysarthric speech recognition systems.

UA-Speech: It involves recordings from fifteen dysarthric speakers with cerebral palsy and 13 without dysarthric. A participant had materials comprising 300 uncommon words, 765 isolated words, and certain had to repeat digits from zero to nine thrice. Other materials have everyday spoken words and radio phonetics with recording frequency samples of 16 KHz. The Mxx represents Male, FXX indicates Female, and C denotes Speakers without dysarthria. Table 1 indicates a detailed dataset description of UA-speech

Kannada dataset: It is gathered from the All India Institute of Speech and Hearing (AIISH) and from native speakers of the Kannada language who articulated a subset of approximately 300 words. For example: ajji, angadi, aspatre, bekul, bele, bekku, chakra, chitte, cycle, badane, aido, amme, aspatre, and so on. These are the words utilised to train the model in Kannada. The obtained information is passed via a pre-processing step for removing noise.

Table 1: Dataset description of UA-Speech

Details of dataset	Binary class	
	Healthy	Dysarthria
No.of. speakers	13	15
Speakers IDs	CF02, CF03, CF05, CF04, CM04, CM01, CM06, CM05, CM09, CMO8, CM12, CM10, CM13	F02, F04, F03, F05, M01, M05, M04, M07, M08, M10, M09, M12, M11, M16, M14
Gender composition	4F/11M	4F/11M

3.2 Pre-processing

After acquiring speech signals, spectral subtraction is applied to remove the noise, which improves dysarthria speech clarity. It enhances the signal-to-noise ratio by eliminating spectral components related to noise. Therefore, this method assists in retaining the significant features of dysarthria speech, which provides better convergence. A noise spectral speech magnitude is removed from a loud speech signal using spectral subtraction. For restoring the magnitude or power signal spectrum by eliminating noise, spectral subtraction is employed. An input speech signal is first buffered and divided into segments of specified length. Each segment is then windowed by utilizing appropriate windowing function which are transformed into spectral components. This assists in isolating primary speech features that make it easier

for recognition models to accurately interpret dysarthic speech. By increasing the speech signal, spectral subtraction contributes to more effective and robust speech recognition systems as well as these signals are passed through feature extraction stage.

3.3 Feature Extraction

A pre-processed output is passed as input to extract features, where features such as MFCC, spectral flux, LPCC, spectral centroid, spectral crest, and chroma are extracted in this stage, which are explained below.

MFCC: It captures speech spectral characteristics effectively, making it effective to recognise dysarthric speech variations. It mimics human auditory perception that enhances recognition accuracy despite dysarthric speech distortions. The sequential processes of MFCC [23] are pre-emphasis, hamming, framing, Fast Fourier Transform (FFT), and Discrete Cosine Transform (DCT).

Spectral Flux: It provides spectral alternations among two consecutive frames. A successive short-term window is considered, and then spectral magnitudes are normasized such that the difference among normalized magnitudes known as spectral flux.

LPCC: It is used to segment the signal and retrieve audio effectively, which works by calculating coefficients of voice samples over time and captures the vocal tract's resonant characteristics. This assists in LPCC [24] speech patterns and enhances recognition accuracy for dysarthric speech.

Spectral centroid: It is calculated depending on spectral shape, with centroid values that are greatly associated with high-frequency brighter textures.

Spectral crest: It determines to computation of signal tonality and differentiates wide as well as narrow-band signals for specifying a subband peak value.

Pitch chroma: It is a significant feature in that chroma represents pitch location in rotary motion by involving pitch rotation angle with $[1 \times 64]$ size.

Therefore, these features capture both temporal and spectral speech characteristics which makes the model more robust to variations. It ensures better discrimination of speech patterns that results in enhanced recognition and intelligibility assessment. Then, the extracted features are fed as input to the recognition and classification process.

3.4 Recognition and Classification

After extracting features, WNCNN-PCLF is used to recognise and classify the dysarthria speech. CNN [25] is effective in capturing local speech patterns and features, which makes it appropriate for dysarthria speech recognition where articulatory distortions occur. WNCNN improves this by modelling long-term speech dependencies by utilizing dilated causal convolutions, which enhance intelligibility in impaired speech. Using PCLF ensures that stress, rhythm, and intonation patterns are preserved and solves prosodic variations in dysarthric speech. Therefore, this combination results in more accurate recognition by refining feature learning and minimizing phoneme misclassification. Overall, the proposed WNCNN-PCLF improves robustness and enables it well-effective for dysarthric speech processing. A detailed description of WNCNN is explained as follows:

WNCNN: It is a deep network that generates waveforms with flexible and large receptive fields, providing better parallelism while capturing long-term dependencies in sequence. An input layer obtains the variables sequence X, auxiliary input A, and input shape as timesteps attributes. Initially, the input is transformed into the residual block's output shape via a convolutional layer for the implementation residual process. The primary component of residual blocks is convolutional layer and Rectified Linear Unit (ReLU). Through these 2 structures, data order is ensured and spatiotemporal nonlinear data mapping is learned. Furthermore, Temporal-Excitation (TE) block depending on the Squeeze and Excitation (SE) is applied for learning long-term dependencies. TE block acquired global temporal data through modelling relationships among convolution channel timesteps U . A transpose function is applied for swapping channel and temporal features' coordinate system. An excitation process is employed in capturing temporal channel dependency and generating a modulation weight set for all channels. Moreover, Fully Connected (FC) with the ratio of dimensionality $r = 2$ and ReLU are utilised for parameterised nonlinearity among time steps. FC restores the coordinate system whereas sigmoid scales a weight. At last, $F_{tr}(.)$ represents the coordinating system and multiplication process $F_{mul}(.)$ indicate integration of outcomes into the backbone network. In WNCNN, the size and number of convolution kernels for each residual block are similar, that makes all residual blocks with uniform shape. In initial residual block, condition and forecast variables are based on ReLU function and convolution process to acquire a channel that contains spatial and temporal features, whereas the

mathematical formula is represented in Equation (1). TE is applied for learning global temporal information to recalibrate channel features, is denoted in Equation (2). TE output block is passed through backbone network via multiplying with U and then the input is incorporated to acquire last output of residual block which is indicated in Equations (3) and (4). Another residual block obtains z_{k-1} and output z_k are demonstrated in Equations (5) and (6)

For the initial residual block k = 1:

$$u_k = \delta(W_{f,k}.X + b_{f,k} + \sum_{i=1}^{p} V_g^i.a^i + b_g \quad (1)$$

$$s_k = F_{tr}(F_{ex}(F_{tr}(u_k))) = (\sigma(W_{k,2}.\delta(W_{k,1}.u_k' + b_{k,1}) + b_{k,2}))' \quad (2)$$

$$e_k = u_k \odot s_k \quad (3)$$

$$z_k = e_k + X + \sum_{i=1}^{p} a^i \quad (4)$$

For other residual block k > 1:

$$u_k = \delta(W_{f,k}.z_{k-1} + b_{f,k}) \quad (5)$$

$$z_k = e_k + z_{k-1} \quad (6)$$

Where $k = range(1, K)$ represents k^{th} residual blocks, K denotes number of residual block, $W_{f,k}$ and $b_{f,k}$ determines weight and bias of convolution filter in k^{th} layer, V_g^i indicates convolutional filter weight of a^i in initial layer, b_g illustrate a convolutional filter's bias in initial layer, δ demonstrates ReLU function, u_k represents output of δ in k^{th} residual blocks. The $W_{k,1}, b_{k,1}, W_{k,2}, b_{k,2}$ presents weight and bias of 1st and 2nd FC layers in TE block of k^{th} residual blocks, σ determines sigmoid function, s_k denotes TE block output in k^{th} residual blocks, e_k and z_k indicates intermediate and last output of k^{th} residual blocks, and $\odot$ illustrates multiplication of associated elements. With linear activation, output layer represents 1×1 a convolutional layer. A final residual block's output z_k performs ReLU calculation and later enters outcome layer using Equation (7)

$$O = W_o.\mathrm{ReLU}(z_k) + b_o \quad (7)$$

Where z_k represents the final residual block's output, W_o and b_o denote output layer's weight and bias, and O illustrates the output layer result.

PCLF: It is responsible to capture prosody feature $H_{Y_0}^P$ from predicted region Y_0 when determining total prosody characteristics $\hat{H}_{\hat{Y}}^P$ represented in original speech. Then, Mean Square Loss (MSE) is used to perform the prosody consistency constraints. The mathematical formula for PCLF is determined using Equation (8)

$$L_{PC} = MSE(H_{Y_0}^P, \hat{H}_{\hat{Y}}^P) \quad (8)$$

The prosody extractor employs a reference encoder of Global Stye Token (GST) model for converting Y_0 and $\hat{Y}$ into high level prosody features with fixed length using Equation (9)

$$H_{Y_0}^P = GST(Y_0), \hat{H}_{\hat{Y}}^P = GST(\hat{Y}) \quad (9)$$

At last, overall loss function is a sum of reconstruction loss and 2 new loss functions, L_{AC} and L_{PC} over all non-contiguous masked regions, therefore, the mask region contains various non-contiguous segments. Thus, the WNCNN-PCLF enhance robustness to speech irregularities by capturing fine-grained temporal dependencies, whereas PCLF increases prosodic feature learning which makes better alignment with natural speech patterns. Therefore, this combination results in enhanced intelligibility and accuracy in recognising dysarthric speech.

4. EXPERIMENTAL RESULTS

The proposed WNCNN-PCLF is simulated utilizing a Python 3.4 environment with Windows 10 operating system, Intel i5 processor, and 64 GB RAM respectively. The selection criteria employed in this research is accuracy, recall, precision, and f1-score. These performance measures are selected depending on relevance in speech recognition and classification field. Accuracy defines overall correctness of recognition whereas precision and recall determine the model's ability to correctly identify speech and avoid false positive/negative. F1-score is the combination of recall and precision respectively. Computational time calculates amount of time takes to complete training or inference tasks while memory usage refers to amount of system required during the execution of model. The mathematical equation for accuracy, f1-score, recall, and precision are represented in equations (10) to (13).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$F1 - Score = \frac{2TP}{2TP+FP+FN} \quad (11)$$

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

$$Precision = \frac{TP}{TP+F} \quad (13)$$

Where TP indicates True Positive, FP determines False Positive, TN illustrates True Negative, and FN denotes False Negative.

4.1 Performance Analysis

Table 2: Evaluation of different DL methods

Methods	Datasets	Accuracy (%)	F1-score (%)	Recall (%)	Precision (%)
InceptionNet-PCLF	UA-Speech	82.21	80.36	81.43	79.32
ResNet-PCLF		85.43	83.33	83.29	83.39
CNN-PCLF		93.21	89.21	89.39	89.04
WNCNN-PCLF		99.92	98.70	98.76	98.65
InceptionNet-PCLF	Kannada	79.01	77.88	76.40	79.43
ResNet-PCLF		82.93	80.65	79.01	82.38
CNN-PCLF		89.43	85.97	82.38	89.90
WNCNN-PCLF		98.34	97.28	97.49	97.08

Table 2 indicates evaluation of different DL methods. The existing methods such as InceptionNet-PCLF, ResNet-PCLF, and CNN-PCLF, are compared with the proposed WNCNN-PCLF. Compared to these methods, WNCNN-PCLF obtains a high accuracy of 99.92% and 98.34% on UA-Speech and Kannada datasets due to its effective capturing of both temporal and spectral speech features. WaveNet performs effectively in complex speech patterns while CNN provides robust phonetic representations. PCLF improves speech intelligibility by applying prosodic consistency, which minimises variability in dysarthric speech. Therefore, this method enhances generalisation by determining acoustic features with typical speech patterns. Moreover, the model minimises distortions and improves pronunciation clarity, which significantly enhances recognition performance for dysarthric speech.

Table 3: Analysis of different loss functions

Loss function	Datasets	Accuracy (%)	F1-score (%)	Recall (%)	Precision (%)
WNCNN-HLF	UA-Speech	87.59	84.28	85.30	83.29
WNCNN-FLF		89.57	87.50	89.40	85.69
WNCNN-BLF		92.38	90.87	92.39	89.40
WNCNN-PCLF		99.92	98.70	98.76	98.65
WNCNN-HLF	Kannada	89.03	86.81	84.29	89.49
WNCNN-FLF		92.48	88.34	89.32	87.39
WNCNN-BLF		95.38	89.19	86.39	92.19
WNCNN-PCLF		98.34	97.28	97.49	97.08

Table 3 represents the performance evaluation of different loss functions. The performance of WNCNN-Hinge LF (HLF), WNCNN-Focal LF (FLF), and WNCNN-Binary LF (BLF) are compared with WNCNN-PCLF. This approach attains a superior accuracy of 99.92% and 98.34% on UA-Speech and Kannada datasets due to its efficient capture both temporal and spectral dependencies in speech signals. Unlike HLF majorly focuses on classification margins whereas PCLF ensures prosodic consistency, which is significant for dysarthric speech variations. While BLF treats all errors equally and PCLF emphasises subtle differences in speech patterns which enhance robustness. Therefore, PCLF improves feature representation by aligning prosodic contours, which minimises misclassification performance and leads to enhanced speech intelligibility as well as high recognition accuracy.

Figure 2 illustrates a graphical representation of k-fold validation. This analysis is used to mitigate certain threats to validity. K-fold validation minimizes the risk of underfitting and overfitting due to selection bias from single train and test split. This enable the model is tested across multiple data subsets by providing robust performance. Compared to k=3,7, and 9, the k=5 achieves an accuracy of 99.92% and 98.34% on UA-Speech and Kannada datasets due to its strike a balance between variance and bias. The model is more sensitive to noise when k=3, which results in lower generalization and

overfitting. With k=7 and 9, the model becomes too smooth, which misclassifies boundary points and increases bias. Furthermore, k=5 shows a better trade-off that minimises sensitivity in outliers while preserving local structure. It manages an optimal decision boundary by using enough neighbours for stable performance, which enhances classification accuracy.

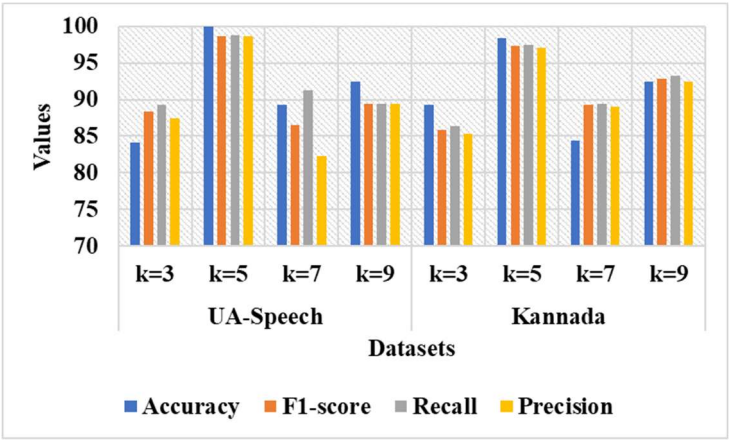

Figure 2: Graphical representation of k-fold validation

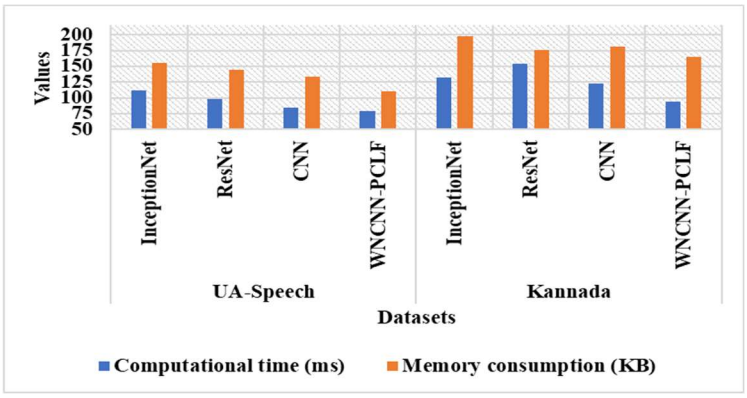

Figure 3: Graphical representation of computational time and memory consumption

Figure 3 shows an evaluation of memory consumption and computational time. A proposed WNCNN-PCLF achieves a less computational time of 79ms and 94ms because of its effective convolutional structure that minimises sequential dependencies compared to existing methods. Unlike traditional loss functions, PCLF concentrate on prosodic consistency without requiring weighting adjustments that assist in minimising additional computations. Moreover, the model can capture speech features effectively, which minimize the requirements for deep network layers and has less computational cost. Therefore, its optimised architecture and loss function increase speed while maintaining high accuracy.

4.2 Comparative Analysis

Table 4 demonstrates comparative analysis of existing methods using UA-Speech datasets. In [20] and [21] the values are presented in decimal form which are converted into percentage as per proposed method values. Compared to existing methods like [16], [17], [18], [19] [20], and [21], the proposed WNCNN-PCLF achieves a high accuracy of 99.92% due to it capture both long-term and short-term dependencies in speech signals, which preserves the prosodic structure. CNN layer captures local spectral features, whereas WaveNet improves sequential modelling to synthesise natural speech. Moreover, PCLF enable consistency in duration, pitch, and energy, which minimize distortions in synthesised speech. Hence, this integration enhances the model's ability with better intelligibility.

Table 4: Comparative Analysis of existing methods using UA-Speech dataset

Methods	Datasets	Accuracy (%)	F1-score (%)	Recall (%)	Precision (%)
FEB 64 UTran-DSR [16]	UA-Speech	N/A	98.7	98.5	N/A
DySARNeT [17]		98.77	98.67	99.04	98.35
STFT-layered CNN [18]		99.89	N/A	N/A	N/A
Hybrid transformer encoder with E1 [19]		98.98	N/A	98.33	97.35
FCCSA-SEGAN [20]		93.0	N/A	93.3	N/A
Fine-tuned DL [21]		96.78	96.78	96.78	96.81
Proposed WNCNN-PCLF		99.92	98.70	98.76	98.65
	Kannada	98.34	97.28	97.49	97.08

4.3 Discussion

A benefit of proposed method and disadvantages of existing methods are presented in detail. A limitation of existing methods like UTrans [16] face challenges in capturing subtle articulatory distortions due to self-attention does not effectively differentiate between dysarthric and healthy speech patterns. DySARNet [17] struggled with generalizing across diverse patterns because attention gating overfitted to dominant acoustic features that ignored subtle variations in severity. STFT with CNN [18] struggled with capturing fine-grained spectral features because of fixed-time frequency resolution, which loses significant frequency information to distinguish speech variations. The hybrid model [19] struggled with high computational complexity and less inference time due to the requirement of processing huge volumes of speech data. The proposed WNCNN-PCLF overcomes these existing method limitations by capturing complex signals effectively. WaveNet's deep hierarchical structure assists in capturing complex temporal dependencies in speech, which makes it efficient to recognise dysarthric speech. The PCLF assist in handling the speech pattern's structural integrity while accounting for inconsistencies established by dysarthria. This enhances the model's accuracy in recognizing subtle phonetic changes. Moreover, the combination of CNN with WaveNet temporal processing enables robust management of incomplete, noisy, or altered speech. Hence, this method results in more accurate and consistent speech recognition for individuals with dysarthria.

4.4 Limitations

This research relies on fixed handcrafted features like MFCC, LPCC, Spectral crest, pitch chroma, and so on for feature extraction. While these features are efficient to capture spectral speech characteristics but does not fully adapt to complex and diverse acoustic patterns determined in dysarthric speech.This limits model's ability to learn tasks specific representation and affects model performance.

5. CONCLUSION

In this research, the WNCNN-PCLF is proposed to recognise and classify dysarthric speech accurately. This method integrates the benefits of WaveNet's ability to model intricate speech patterns with CNN results in enhanced robustness and recognition accuracy in dysarthric speech. The inclusion of PLCF improves the model's capability in preserving natural prosodic features like intonation and rhythm that are distorted in dysarthric speech. The novelty lies in the integration of these components to simultaneously address prosodic and articulatory distortions in dysarthric speech. Compared to existing methods, proposed WNCNN-PCLF ensures better preservation of speech rhythm which significantly enhance recognition accuracy. By considering both temporal and spectral speech features, WNCNN-PCLF makes better speech variation discrimination by enhancing classification and intelligibility. The proposed method's performance is determined via comparison with existing methods, which provides superior outcomes in precision, accuracy, F1-score, and recall, respectively. This method shows significant contributions to improving communication for individuals with speech impairments, which offers an effective solution for dysarthric speech recognition across diverse patterns. The practical implications like the proposed WNCNN-PCLF contributes to improved classification accuracy and robustness in dysarthric speech recognition tasks particularly in prosodic and speaker variability. When compared to existing methods like DySARNeT, the proposed WNCNN-PCLF achieves a better accuracy of 99.92% using the UA-Speech dataset. In the future, the advanced end-to-end DL method will be used which allows to automatically extract features in dysarthric speech. This enhance

robustness across speakers and minimize dependence on manual feature engineering.

REFERENCES

[1] F. Javanmardi, S.R. Kadiri, and P. Alku, "Pre-trained models for detection and severity level classification of dysarthria from speech", *Speech Communication*, Vol. 158, 2024, p. 103047.

[2] M. Mahendran, R. Visalakshi, and S. Balaji, "Dysarthria detection using convolution neural network", *Measurement: Sensors*, Vol. 30, 2023, p. 100913.

[3] S.R. Shahamiri, V. Lal, and D. Shah, "Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 31, 2023, pp. 3407-3416.

[4] Takashima, R., Sawa, Y., Aihara, R., Takiguchi, T. and Imai, Y., 2024. Dysarthric Speech Recognition Using Pseudo-Labeling, Self-Supervised Feature Learning, and a Joint Multi-Task Learning Approach. *IEEE Access,* Vol. 12, 2024, pp. 36990-36999.

[5] C. Yu, X. Su, and Z. Qian, "Multi-stage audio-visual fusion for dysarthric speech recognition with pre-trained models", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 31, 2023, pp. 1912-1921.

[6] G. Alharbi, N. Alamri, and S. Sabbeh, "Automatic Classification of Speech Dysarthric Intelligibility Levels Using Textual Feature", *IEEE Access*, Vol. 13, 2025, pp. 39982-39992.

[7] Y. Lin, L. Wang, Y. Yang, and J. Dang, "CFDRN: A cognition-inspired feature decomposition and recombination network for dysarthric speech recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, 2023, pp. 3824-3836.

[8] S. Sajiha, K. Radha, D. Venkata Rao, N. Sneha, S. Gunnam, and D.P. Bavirisetti, "Automatic dysarthria detection and severity level assessment using CWT-layered CNN model", *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2024, No. 1, 2024, p. 33.

[9] S.M. Shabber and E.P. Sumesh, "AFM signal model for dysarthric speech classification using speech biomarkers", *Frontiers in Human Neuroscience*, Vol. 18, 2024, p. 1346297.

[10] A. Almadhor, R. Irfan, J. Gao, N. Saleem, H.T. Rauf, and S. Kadry, "E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition", *Expert Systems with Applications*, Vol. 222, 2023, p. 119797.

[11] C. Bhat and H. Strik, "Two-stage data augmentation for improved ASR performance for dysarthric speech", *Computers in Biology and Medicine*, Vol. 189, 2025, p. 109954.

[12] A.S. Al-Ali, R.M. Haris, Y. Akbari, M. Saleh, S. Al-Maadeed, and M. Rajesh Kumar, "Integrating binary classification and clustering for multi-class dysarthria severity level classification: a two-stage approach", *Cluster Computing*, Vol. 28, No. 2, 2025, p. 136.

[13] R. Kumar, M. Tripathy, R.S. Anand, and N. Kumar, "Residual Convolutional Neural Network-Based Dysarthric Speech Recognition", *Arabian Journal for Science and Engineering*, Vol. 49, No. 12, 2024, pp. 16241-16251.

[14] R. Vinotha, D. Hepsiba, L.D. Vijay Anand, J. Andrew, and R. Jennifer Eunice, "Enhancing dysarthric speech recognition through SepFormer and hierarchical attention network models with multistage transfer learning", *Scientific Reports*, Vol. 14, No. 1, 2024, p. 29455.

[15] T.A. Mariya Celin, P. Vijayalakshmi, and T. Nagarajan, "Data augmentation techniques for transfer learning-based continuous dysarthric speech recognition", *Circuits, Systems, and Signal Processing*, Vol. 42, No. 1, 2023, pp. 601-622.

[16] U. Irshad, R. Mahum, I. Ganiyu, F.S. Butt, L. Hidri, T.G. Ali, and A.M. El-Sherbeeny, "UTran-DSR: a novel transformer-based model using feature enhancement for dysarthric speech recognition", *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2024, No. 1, 2024, pp. 1-18.

[17] F.J.P. Montalbo, "DySARNet: a lightweight self-attention deep learning model for diagnosing dysarthria from speech recordings", *Multimedia Tools and Applications*, 2024, pp. 1-49.

[18] K. Radha, M. Bansal, and V.R. Dhulipalla, "Variable STFT layered CNN model for automated dysarthria detection and severity assessment using raw speech", *Circuits, Systems, and Signal Processing*, Vol. 43, No. 5, 2024, pp. 3261-3278.

[19] R. Mahum, A.M. El-Sherbeeny, K. Alkhaledi, and H. Hassan, "Tran-DSR: A hybrid model for dysarthric speech recognition using transformer encoder and ensemble

learning", *Applied Acoustics*, Vol. 222, 2024, p. 110019.

[20] B. Jolad and R. Khanai, "An approach for speech enhancement with dysarthric speech recognition using optimization based machine learning frameworks", *International journal of speech technology*, Vol. 26, No. 2, 2023, pp. 287-305.

[21] S.M. Shabber, E.P. Sumesh, and V.L. Ramachandran, "Scalogram based performance comparison of deep learning architectures for dysarthric speech detection," *Artificial Intelligence Review*, Vol. 58, No. 5, p. 128, 2025.

[22] UA-Speech dataset link: https://www.kaggle.com/datasets/aryashah2k/noise-reduced-uaspeech-dysarthria-dataset (Accessed on March 21 2025).

[23] Y.L. Chen, N.C. Wang, J.F. Ciou, and R.Q. Lin, "Combined bidirectional long short-term memory with mel-frequency cepstral coefficients using autoencoder for speaker recognition", *Applied Sciences*, Vol. 13, No. 12, 2023, p. 7008.

[24] H.M. Kadhim, A.H. Ahmed, A.K. Hassan, and S.T. Alfalahi, "Supervised Machine Learning for Speaker Diarization by PNCC with LPCC Audio Coefficients", *Association of Arab Universities Journal of Engineering Sciences*, Vol. 31, No. 3, 2024, pp. 37-45.

[25] G. Bompem and D. Pandluri, "Batch Normalization Based Convolutional Neural Network for Segmentation and Classification of Brain Tumor MRI Images", *International Journal of Intelligent Engineering & Systems*, Vol. 17, No. 2, 2024, pp. 39-49.