Ingénierie des Systèmes d'Information
Vol. 31, No. 6, June, 2026, pp. 2023-2033
Journal homepage: http://iieta.org/journals/isi

A Lightweight Hybrid Deep Learning Framework Integrating Multi-Acoustic Features and Convolutional Neural Network and Gated Recurrent Unit Networks for Dysarthric Speech Recognition

Namita Kure[1,2*], Somnath B. Dhonde[3]

[1] Department of Electronics and Telecommunication Engineering, AISSMS Institute of Information Technology, Pune 411001, India
[2] Department of Electronics and Telecommunication Engineering, Dr. D. Y. Patil Institute of Technology, Pune 411018, India
[3] Department of Electronics and Telecommunication Engineering, AISSMS College of Engineering, Pune 411001, India

Corresponding Author Email: npachling@gmail.com

https://doi.org/10.18280/isi.310623

Received: 27 March 2026
Revised: 25 May 2026
Accepted: 6 June 2026
Available online: 30 June 2026

Keywords:
dysarthric speech recognition, deep learning, convolutional neural network, gated recurrent unit, acoustic feature extraction, feature selection, assistive speech technology

ABSTRACT

Dysarthric speech recognition remains challenging due to reduced intelligibility, large speaker variability, and complex acoustic characteristics caused by impaired motor control. Existing recognition methods often suffer from insufficient feature representation and limited modeling of temporal dependencies in dysarthric speech signals. This study proposes a lightweight hybrid deep learning (DL) framework that integrates multiple acoustic feature (MAE) representation with a convolutional neural network and gated recurrent unit (CNN-GRU) architecture for dysarthric speech recognition. A comprehensive acoustic feature set, including spectral-domain features, time-domain features, and voice-quality features, is extracted to characterize dysarthric speech patterns. Cross-correlation-based feature selection is further introduced to reduce feature redundancy and improve computational efficiency. The proposed framework is evaluated on the UASpeech dataset using speaker-independent and speaker-dependent evaluation strategies. Experimental results demonstrate that the proposed CNN-GRU model achieves an accuracy of 98.11%, with an F1-score of 98.12% when using the selected multi-acoustic features. Compared with individual CNN and GRU models, the proposed hybrid architecture provides improved recognition capability by jointly capturing local feature representations and long-term temporal dependencies. Additional ablation experiments confirm the effectiveness of the feature selection strategy and model configuration. The proposed framework provides a computationally efficient approach for automatic dysarthric speech recognition and may support future development of assistive speech technologies.

1. INTRODUCTION

Speech is one of the most basic ways individuals express their thoughts and ideas. Therefore, speech recognition [1, 2] is needed for human-computer interaction. An automatic speech recognition is detection of human speech by use of particular computer algorithm. Speech recognition is the basis for automation, language identification, emotion detection, speaker recognition, voice pathology, and biometric authentication. However, speech articulation inadequacy affects the effectiveness of speech recognition-based applications [3]. Speech therapy analyzes and treats speech disorders and communication difficulties. Speech-language pathologists (SLPs) execute this role [4].

Dysarthria is a motor speech disorder caused by weakness, damage, or paralysis of muscles involved in speaking. Patients cannot control their tongue, voice, and articulation [5]. The production and pronunciation of words may be challenging, depending on which area of the nervous system is impacted. Dysarthria comes in two forms: peripheral dysarthria and central dysarthria. Peripheral dysarthria is caused by damage to organs utilized in production of speech. Brain damage leads to central dysarthria. Dysarthria may also be characterized as acquired or developing. Developmental dysarthria may develop later in life as a consequence of brain injury. Brain injury causes cerebral palsy. Brain injury occurred before and after birth may lead to dysarthria [6, 7]. Acquired dysarthria can result after a stroke, Parkinson's disease, or brain tumor. Dysarthria is characterized by mumbling, slurring, soft or quiet speech, speaking too rapidly or slowly, and problems pronouncing words [8].

Various traditional machine learning (ML) and deep learning (DL)-based dysarthric speech recognition (DSR) systems have shown significant improvements in DSR performance. However, efficiency and reliability of the system are limited by poor feature representation, feature redundancy, high computational complexity, and inferior long-term and temporal representations of the dysarthric voice. The existing work lacks reliability for low-intelligibility speech. Thus, this work aims to improve the efficiency, reliability, temporal

fidelity, and long-term dependencies of the DSR system [9, 10].

To improve feature representation and DSR accuracy, this paper presents a hybrid DL framework. This DL framework incorporates handcrafted features for DSR. An outline of the article's main contributions is summarized as follows:

•The representation of dysarthric voice based on spectral-domain features (SDF), time-domain features (TDF) and voice-quality features (VQF) to enhance feature depiction capability. Further, cross-correlation-based feature selection (CBFS) is employed to choose important features. CBFS minimizes the model's computational complexity.

•To present a lightweight deep convolutional neural network-gated recurrent unit (DCNN-GRU) for improving hierarchical feature depiction capability and distinctiveness for DSR on the UASpeech dataset. The DCNN captures correlations and high-level correlations in complex features. The GRU captures long-term dependencies and temporal dynamics in handcrafted features to improve feature distinctiveness.

The paper is structured as follows: Section 2 provides a literature review of various recent DSR techniques. Section 3 presents various speech features for extraction and details the methodology. Section 4 presents experimental results and discussion. Section 5 presents conclusions and potential future developments.

2. LITERATURE REVIEW

DL techniques have recently been extensively adopted in several voice-based automation mechanisms. DL techniques have demonstrated superior performance to conventional ML-based methods. Janbakhshi et al. [11] used a spectro-temporal representation (STR) combined with T-GDA to jointly capture spectral and temporal characteristics. Their approach achieved 96.30% accuracy on the UASpeech dataset. It showed that carefully engineered time-frequency representations can dramatically enhance DSR performance. Bhangale et al. [12] suggested that Mel spectrograms can capture spectro-temporal variations in speech. However, the spectral depiction using Mel filters lacks temporal and long-term correlation in the speech that fails to characterize the voice quality.

Various time frequency depiction schemes have been employed along with DL techniques to boost the DSR accuracy for low voice intelligibility. Fritsch and Magimai-Doss [13] further demonstrated that both recurrent neural network (RNN)-based binary classifiers and CNN-based multi-feature classifiers correlate strongly with synthetic TTS-generated dysarthric speech. According to Chandrashekar et al. [14], time-frequency CNNs yield a more accurate representation of dysarthric speech patterns than traditional neural networks, leading to improved recognition accuracy. Similarly, an article [15] proposed a time-frequency CNN trained using SIFT, SFF, and CQT representations. Although the model performed better on female speakers—likely due to more consistent articulation—training data imbalance was identified as a limitation. Kodrasi et al. [16] used CNNs to capture temporal and spectral structure using TEFS features, outperforming conventional spectrogram techniques. TEFS-CNN achieved accuracy 85.72%, whereas the SIFT-based CNN reached only 69.76%.

Mahum et al. [17] suggested the Swin transformer captures complex multilevel dysarthric features from the voice spectrograms. The multi-view and multi-block modelling help to achieve 98.66% accuracy for DSR. Its effectiveness is challenging due to complex architecture and lower deployment flexibility on resource limited healthcare devices. Yue et al. [18] suggested multimodal voice modelling that collaborates the acoustic features with articulatory features to enhance the feature depiction. It resulted in WER of 7.6 for TORGO dataset for DSR but the generalization ability of the system is challenging. He et al. [19] suggested cycle GAN model combined with Inception-fusion blocks for generating synthetic sample to reduce the class imbalance issue. It uses error correction scheme to ensure sematic information which resulted in overall WER of 48.69% for low intelligibility voices. Choi et al. [20] proposed that WhisperX-medium provides most satisfactory results for human transcription accuracy for low intelligibility dysarthric voice. The WhisperX models offers better perceptual rating and stronger correlations in ease of understanding and WER. Recording longer voice samples from the dysarthric patient is challenging due to disability and ethical issues which creates data scarcity issue. To tackle this issue, Vijayalakshmi et al. [21] presented data synthesis using text-to-speech models for data augmentation. The FastSpeech2 based TTS has shown WER of 3.49% for Tamil dysarthric voice samples with mild dysarthria. However, the prosodic variation over languages limits the generalization capacity of DSR systems. Vinet et al. [22] suggested Whisper-v3 Model for estimating the intelligibility of dysarthric voices based on phonetic attributes. It offered web based remote applications for DSR what helps to speech language therapists to support the decision making. Chung et al. [23] suggested that Whisper based DSR systems show adaptability to the speaker and needs higher dataset for training. It shows noise robustness, robustness to speech variability, context aware transcription and multilingual support. Its efficiency is challenging due high training and inference time, computational intricacy and need of fine-tuning. To improve ASR effectiveness, Fathima et al. [24] adopted a multilingual TDNN-based model that integrates acoustic modeling with language-dependent cues. With a 16.07% word error rate (WER), their method proved the value of adding language-specific information to feature learning.

Farhadipour et al. [25] employed Gammatonegram spectrograms, a perceptually motivated auditory representation, as input features. Using AlexNet which is well-known CNN architecture. This method achieved 92.3% accuracy, demonstrating the benefits of combining biological features with DL frameworks. Similarly, Kumar et al. [26] utilized raw speech as input to a Residual CNN (ResCNN) framework. This model achieved a modest 75.91% accuracy, indicating better results than simple CNNs with raw audio but still limited compared to feature-rich approaches.

From the survey, various research gaps are identified, such as the high complexity of existing DL algorithms, which limits their feasibility for real-time diagnosis. Existing techniques suffer from poor feature representation, feature redundancy, and inadequate temporal modeling of voice, leading to low DSR accuracy. It has limited accuracy due to low-intelligibility voice and limited modeling of complex dysarthric features. Many methods show limited long-term depiction of the voice signals. These methods cause lower generalization capability.

3. METHODS

Figure 1 illustrates the DCNN-GRU-based DSR system. It includes preprocessing, acoustic feature extraction, DCNN-GRU-based feature representation, and dysarthria classification. It comprises multiple acoustic feature extraction methods, including SDF, TDF, and VQF, to capture frequency-time-, and prosodic-related voice attributes. Furthermore, a CBFS method is employed to identify important features. The DCNN-GRU is used to classify the input speech as normal or dysarthric. The results are verified across different feature counts using various performance metrics.

3.1 Multiple dysarthric speech features

Acoustic features capture the quantifiable characteristics of a speech signal. It includes intensity, frequency distribution, and amplitude fluctuations. The speech signal is first passed through a moving average filter to minimize signal irregularities and suppress background noise. Following this preprocessing step, a comprehensive set of acoustic features are extracted to represent dysarthric speech patterns effectively as given in Figure 2. The selected feature set includes SDF, TDF, VQF. Spectral features include MFCC, Linear Prediction Cepstral Coefficients (LPCC), Wavelet Packet Transform (WPT), spectral centroid, spectral roll off, and spectral kurtosis. Time-domain Feature measures root-mean-square (RMS) and Zero Crossing Rate (ZCR) energy. Pitch measures fundamental frequency of vocal fold vibrations. Voice quality feature like jitter and shimmer are computed to capture irregularities in vocal fold vibration. Formant-based statistics, including the standard deviation and mean of formant frequencies, are also extracted to describe articulatory modifications. These 715 features provide a comprehensive representation of the speech characteristics associated with dysarthria. Feature details are presented in Table 1.

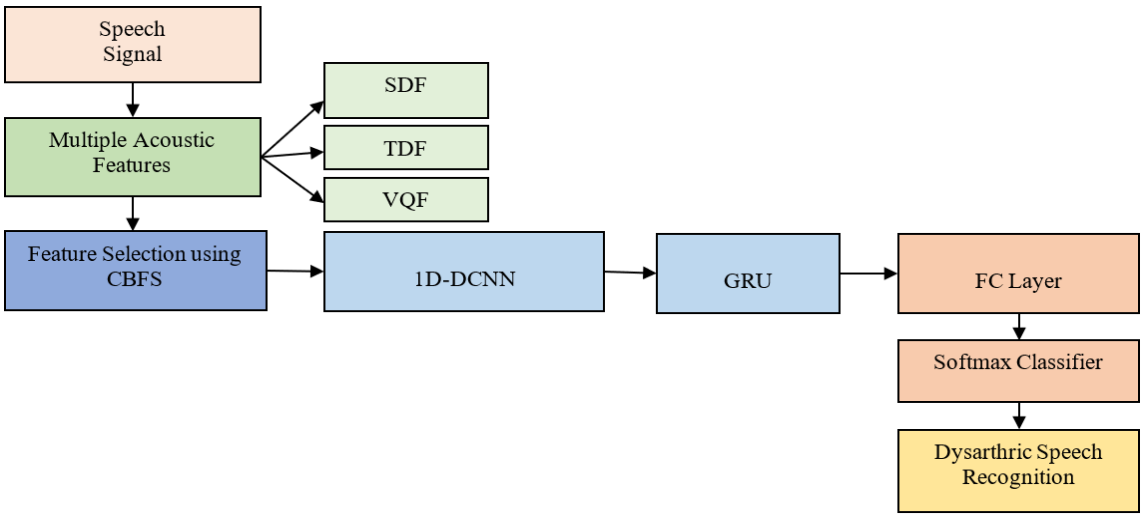


Figure 1. The suggested dysarthric speech recognition (DSR) system's flow diagram

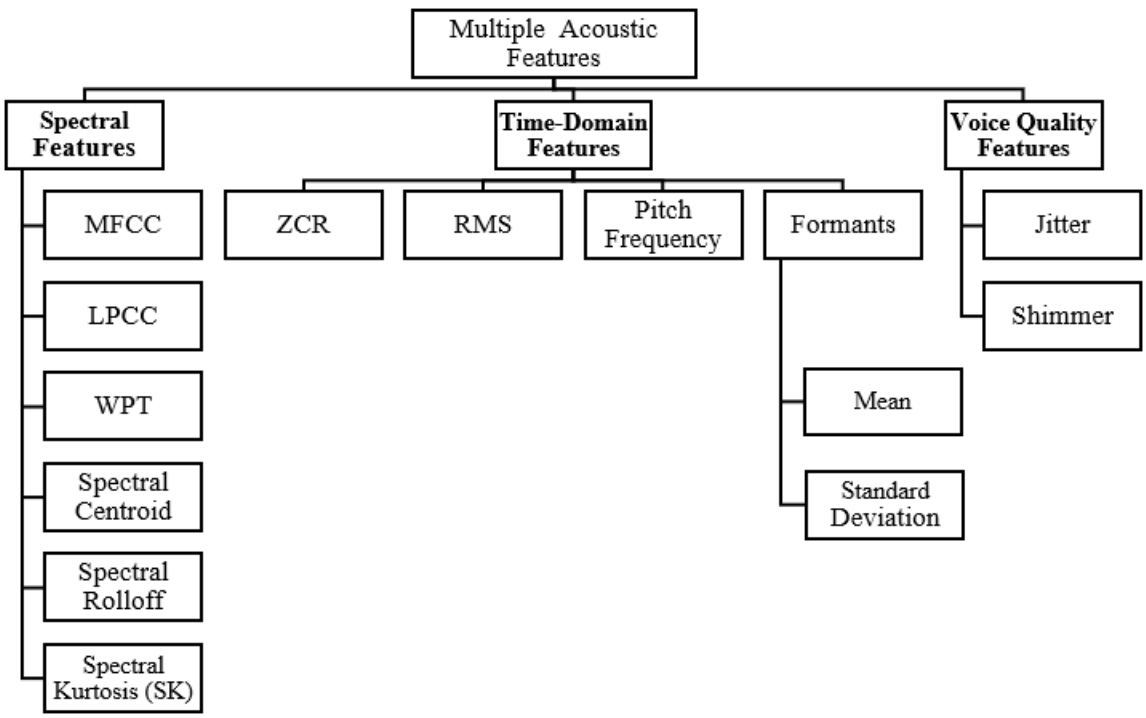


Figure 2. Different multiple acoustic features (MAFs) for depicting dysarthric voice

Table 1. Overview of MAFs utilized for CNN-GRU-based DSR

Features	Significance to Dysarthric Speech	No. of Features
MFCC	Captures spectral envelope and formant structure changes; dysarthric speech often shows spectral distortion.	39
LPCC	Represents vocal tract configuration; useful for detecting articulatory imprecision in dysarthria.	13
SC	Frequency accumulation	1
Fo	Changes in formant trajectories indicate articulatory dysfunction and vowel distortion.	3
Std Dev of Fo	Measures pitch stability; dysarthric speech often shows irregular pitch patterns.	1
Pitch Freq	Reflects prosody and intonation; monotone or unstable pitch is common in dysarthria.	1
Jitter	Indicates cycle-to-cycle variation in pitch; higher jitter values are linked with impaired phonation.	1
Shimmer	Reflects amplitude variation; increased shimmer can be due to poor vocal fold control.	1
Mean of Fo	Average pitch level; abnormal mean pitch may signal motor control issues in phonation.	1
Kurtosis	Measures peakedness of spectral distribution; deviations can reflect abnormal vocal tract dynamics.	257
Skewness	Asymmetry in signal distribution; may indicate non-uniform articulatory effort.	1
ZCR	Zero-Crossing Rate; sensitive to unvoiced phonemes, often altered in dysarthric articulation.	169
Activity	Measures signal energy/activity; may be reduced in hypokinetic dysarthria.	1
Mobility	Reflects signal variability; lower mobility might suggest restricted articulatory motion.	1
Complexity	Quantifies signal dynamics; reduced complexity can reflect motor speech deficits.	1
WWPT (5 level decomposition, features for each subband like energy, entropy, median, kurtosis, mean, variance, standard deviation)	Wavelet Packet Transform captures time-frequency information; useful for identifying subtle speech anomalies.	224
Total Features		**715**

Note: MAFs = multiple acoustic features, CNN = convolutional neural network, GRU = gated recurrent unit, MFCC = Mel Frequency Cepstral Coefficients, LPCC = Linear Prediction Cepstral Coefficients, Fo = Formant freq, ZCR = Zero Crossing Rate, WPT = Wavelet Packet Transform.

3.2 Correlation based feature selection

Important features are chosen using the CBFS. The CBFS lowers DCNN-GRU's computational complexity and increases the method's deployment flexibility on devices with limited resources. Features with higher correlation values show greater dependence. Highly correlated features can be dropped to enhance feature distinctiveness. Cross-correlation of two feature sequences $x(n)$ and $y(n)$, is specified by Eq. (1).

$$\text{CrossCorr}\big(x(n); y(n)\big) = \sum_{k=-\infty}^{\infty} x(k) * y(n+k) \qquad (1)$$

Two sequences with equal values result in a higher cross-correlation value. The CBFS helps to select crucial features by minimizing redundancy. The features are selected based on the maximum CrossCorr score, selecting 50, 100, 200, ..., 700 features to validate their importance.

3.3 Deep convolutional neural network-gated recurrent unit architecture

Figure 3 shows the DCNN-GRU framework. It consists of a 1-D DCNN followed by two GRU layers with 50 units each. Proposed compact one-dimensional DCNN takes 715 acoustic features (Feat) for input. The system uses 5 CNN layers and 2 GRU layers to minimize computational complexity, thereby increasing the method's feasibility for deployment on standalone healthcare devices. The GRU increases the temporal depiction and long-term dependency of dysarthric voice. The first convolutional layer consists of a convolution operation followed by a Rectified Linear Unit (ReLU) activation (Conv1 → ReLU1), producing an output of size 715 × 1 × 32. The second layer (Conv2 → ReLU2) generates feature maps of size 715 × 1 × 64. The third layer (Conv3 → ReLU3) outputs feature maps of size 715 × 1 × 96. Fourth layer (Conv4 → ReLU4) outputs feature maps of size 715 × 1 × 128.Fifth (Conv5 → ReLU5) outputs feature maps of size 715 × 1 × 256. In each convolutional layer, one-dimensional input is convolved with a filter to extract high-level representations of dysarthric speech.

The Eq. (2) shows convolution of input features which is Feat(n) with a filter denoted as w(n) of length l, and the resulting feature map for the l^{th} layer is given by Eq. (3), where the z_i^l denotes i^{th} feature map, z_j^{l-1} represents j^{th} feature from the previous layer and w_{ij}^l is filter kernel, the bias is represented as b_i^l, and σ represents ReLU activation function. The ReLU layer is computationally efficient and helps mitigate the vanishing gradient problem, shown in Eq. (4).

$$z(n) = Feat(n) \times w(n) = \sum_{m=0}^{i-1} Feat(m).w(n-m) \qquad (2)$$

$$z_i^l = \sigma\left(b_i^l + \sum_j z_j^{l-1} \times w_{ij}^l\right) \qquad (3)$$

$$\sigma(z) = max(0, z) \qquad (4)$$

Flattened output of last MaxPool layer is provided to GRU to improve temporal modeling and long-term dependencies in the acoustic features. The GRU plays a significant role in DSR.GRU can model temporal dependencies in speech while

being computationally more efficient than traditional LSTMs. Dysarthric speech often contains irregular timing, slurred articulation, and unpredictable acoustic patterns. This can be challenging for standard models to capture. GRUs are particularly effective in this domain because they adaptively control the flow of information. It allows the model to retain relevant historical context while filtering out noise or variability typical in impaired speech. GRU calculates reset gate (r_t), update gate (z_t) and output gate (h_t) at time step t, given input is x_t and previous hidden state is h_{t-1}.

The term z_t is update gate in a GRU. It performs a crucial contribution in controlling how much of the previous memory h_{t-1} is retained at the current time step as given by Eq. (5).

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{5}$$

The sigmoid function σ in the update gate outputs values between 0 and 1. It enables the GRU to control memory retention. Here, x_t is input at the current step, and h_{t-1} the previous hidden state. The weights W_z, U_z, and bias b_z are learnable parameters. The gate output decides how much past information to retain versus how much new input to incorporate. Values near 1 favor memory retention, while values near 0 prioritize new data crucial for handling the variability in dysarthric speech.

The reset gate (r_t) decides how much information has to be dropped using Eq. (6). Where W_r is weight of input to reset gates, U_r is weights of reset gate and b_r is bias for reset gate.

$$r_t = \sigma(W_r x_t + U_r h_t^{-1} + b_r) \tag{6}$$

The candidate hidden state ($\hat{h}_t$) combines input and reset gate and stores new information as given in Eq. (7).

$$\hat{h}_t = tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{7}$$

where, h_t is hidden state shown in Eq. (8). The sigmoid activation function is denoted by σ, $tanh$ signifies hyperbolic tangent activation, symbolizes element-wise multiplication, and the current hidden state is denoted as h_t.

$$h_t = (1\text{-}z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \tag{8}$$

Following the GRU layer, the network employs one FC layer. FC layer consist of 10 hidden layers. Within FC layer, input feature vector is linearly transformed by weight matrix. To present non-linearity and enable network to model complex patterns, non-linear activation function is applied, as described into Eq. (9).

$$y_{jk}(x) = f(\sum_{i=1}^{n_H} W_{jk} x_i + wj_0) \tag{9}$$

Here, x_i depict value taken from flattening the feature vector, the w_0 denotes term bias, w represents the weight matrix, here f represents non-linear activation function, the y is non-linear transformation output, no hidden layers are denoted as n_H. At end, a probability distribution over the class labels is provided by Softmax classifier, whereas the highest probability yields output class label, as mentioned in Eqs. (10)-(12) and (16).

$$z_i = \sum_j h_j w_{ji} \tag{10}$$

$$p_i = \frac{exp(z_i)}{\sum_{j=1}^{n} exp(z_j)} \tag{11}$$

$$\hat{y} = arg \max_i p_i \tag{12}$$

In Eq. (10), h_j is the penultimate layer weight, here the weights of Softmax and the penultimate layer is represented by w_{ji} and input of the Softmax layer is denoted by z_i, probability of the class label is represented by p_i and predicted class label is $\hat{y}$.

The parameter configurations of the DCNN-GRU are provided in Table 2, which provides filter size, stride, activation functions, and total trainable parameters of network. The DCNN-GRU has a total of 609,270 trainable parameters (~0.6 M), which increases model's deployment flexibility on standalone devices with limited computational resources.

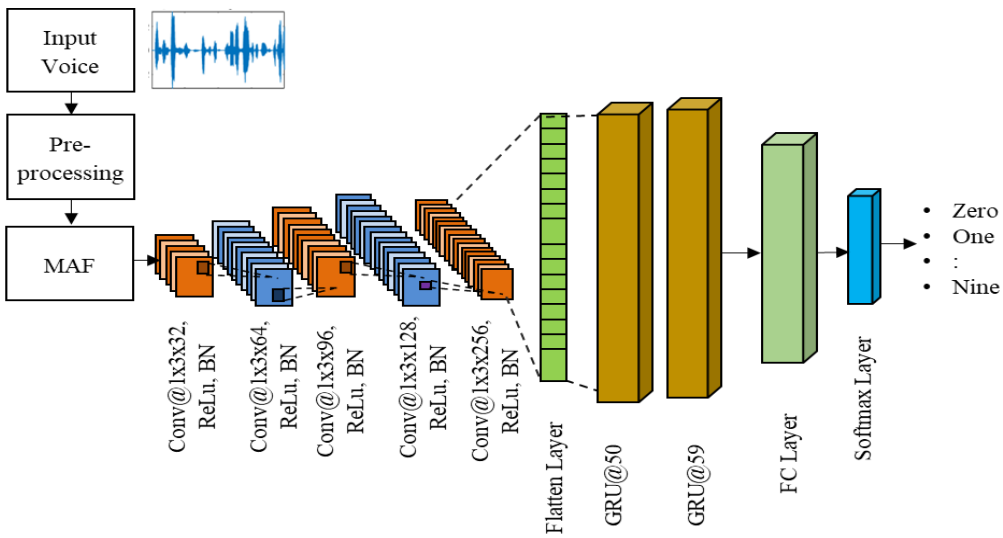


Figure 3. Framework of deep convolutional neural network-gated recurrent unit (DCNN-GRU) for dysarthric speech recognition (DSR)

Table 2. Parameter configurations of the DCNN-GRU for 300 features chosen using CBFS

Layers	Sublayers	Input Feature Dimensions	Filter Size	No. of Filters / Hidden Units	Output Feature Dimensions	Stride	Learnable Parameters
Input Layer	MAF	[300 × 1]	–	–	–	–	–
DCNN	Conv-1	[300 × 1]	1 × 3	32	300 × 1 × 32	1	128
	BN-1	300 × 1 × 32	–	–	300 × 1 × 32	1	64
	ReLU-1	300 × 1 × 32	–	–	300 × 1 × 32	1	–
	Conv-2	300 × 1 × 32	1 × 3	64	300 × 1 × 64	1	6208
	BN-2	300 × 1 × 64	–	–	300 × 1 × 64	1	128
	ReLU-2	300 × 1 × 64	–	–	300 × 1 × 64	1	–
	Conv-3	300 × 1 × 64	1 × 3	96	300 × 1 × 96	1	18528
	BN-3	300 × 1 × 96	–	–	300 × 1 × 96	1	192
	ReLU-3	300 × 1 × 96	–	–	300 × 1 × 96	1	–
	Conv-4	300 × 1 × 96	1 × 3	128	300 × 1 × 128	1	36992
	BN-4	300 × 1 × 128	–	–	300 × 1 × 128	1	256
	ReLU-4	300 × 1 × 128	–	–	300 × 1 × 128	1	–
	Conv-5	300 × 1 × 128	1 × 3	256	300 × 1 × 256	1	98560
	BN-5	300 × 1 × 256	–	–	300 × 1 × 256	1	512
	ReLU-5	300 × 1 × 256	–	–	300 × 1 × 256	1	–
	Flatten	300 × 1 × 256	–	–	76800 × 1	–	–
GRU	GRU-1	76800 × 1	–	50	76800 × 50	–	7800
	GRU-2	76800 × 50	–	50	50 × 1	–	15150
	Flatten	50 × 1	–	–	50 × 1	–	–
FC Layer	FC-2	50 × 1	–	2	50 × 2	–	102
Total Trainable Parameters							**609270**

Note: MAFs = multiple acoustic features; DCNN = deep convolutional neural network, GRU = gated recurrent unit, CBFS = cross-correlation-based feature selection, FC = fully connected.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 System details

Simulation of the proposed method is performed in MATLAB on a personal computer running Windows, equipped with a Core i5 CPU and 4 GB of RAM. Figure 4 shows the proposed network's training accuracy and training loss. DCNN is trained with the Adam optimizer for 100 epochs, a batch size of 64, and an initial learning rate of 0.001. The parameters of the DCNN and different feature extraction algorithms are shown in Table 3.

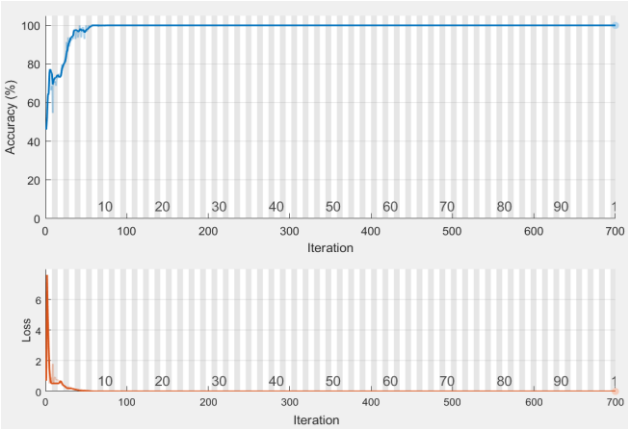


Figure 4. Training accuracy and training loss graph of proposed deep convolutional neural network-gated recurrent unit (DCNN-GRU)

4.2 Dataset

The results are assessed on the UASpeech dataset. We have considered the digit speech samples of 15 speakers with dysarthria and 13 normal speakers. Age of speakers ranges from 18 to 58 years. Dataset consists of speech samples with four intelligibility levels: high, medium, low, and very low intelligibility. Samples with high intelligibility are considered normal, and those with low, medium, or very low intelligibility are considered dysarthric. We have considered 3,000 normal and 3,000 dysarthric samples. We have considered 70% of the samples for training and 30% for testing [27]. Stability of the experimental results is evaluated using 5-fold and 10-fold cross-validation in both speaker dependent mode (SDM) and speaker independent mode (SIM).

Table 3. Method parameters and specifications

Method	Parameter	Specification
DCNN-GRU	Batch Size	64
	Epochs	100
	Dropout Rate	0.5
	Loss Function	Cross-entropy
	Momentum	0.8
	Initial Learning Rate	0.001
	Learning Method	Adam
MFCC	Frame size	40 ms
	Frame Overlapping	20 ms
	Mel Filters	32
	Cepstral lifters	13
	Sampling Rate	16000 Hz
WPT	Filter	Db2
	Decomposition level	5

Note: DCNN = deep convolutional neural network, GRU = gated recurrent unit, WPT = Wavelet Packet Transform, MFCC = Mel Frequency Cepstral Coefficients.

4.3 Performance matrices

The results of the DSR system are evaluated using various qualitative and quantitative measures, including recall, precision, selectivity, negative predictive value (NPV), F1-

score, and accuracy, as given in Eqs. (13)–(18), respectively.

$$Recall = \frac{True\ Positive}{True\ positive + False\ Negative} \quad (13)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (14)$$

$$Selectivity = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (15)$$

$$NPV = \frac{TN}{TN + FN} \quad (16)$$

$$F1 - score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (17)$$

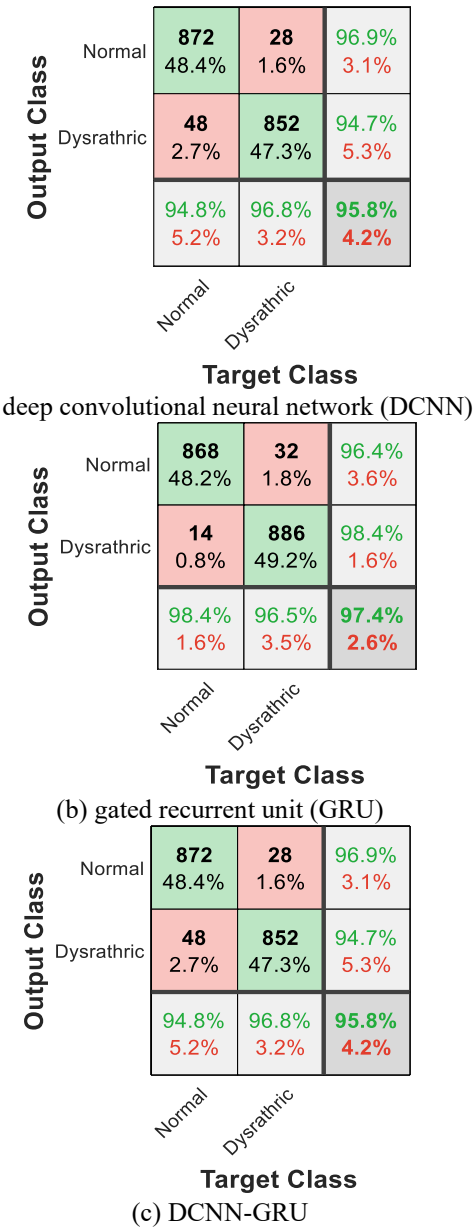
$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (18)$$

4.4 Discussion

The confusion matrices for the DCNN, GRU, and DCNN + GRU with multiple acoustic features (MAFs) are shown in Figure 5, using 300 features selected by CBFS for SIM.

Table 4 and Figure 6 present performance of the DSR system utilizing different feature sets—SDF, SDF + TDF, SDF + VQF, and MAF (SDF + TDF + VQF)—with three classifiers: DCNN, GRU, and hybrid DCNN-GRU. When only SDF features are used, the DCNN model achieves 86.22% accuracy, 81.00% recall, 90.45% precision, and 85.46% F1-score. GRU slightly improves accuracy to 87.22% and F1-score to 86.71%. The hybrid DCNN-GRU further enhances performance to 88.56% accuracy and 88.08% F1-score. With the addition of temporal information (SDF + TDF), all metrics improve significantly. The DCNN reaches 89.89% accuracy and 90.27% F1-score. GRU attains 91.67% accuracy and 91.83% F1-score. The hybrid DCNN-GRU achieves even higher results with 93.11% accuracy, 93.67% recall, 92.64% precision, and 93.15% F1-score. This confirms that temporal descriptors help to capture speech dynamics more effectively. Further improvement is observed when voice quality features (VQF) are added to Spectral Domain Feature (SDF + VQF). The DCNN records 93.83% accuracy and 93.88% F1-score, GRU increases to 94.50% accuracy and 94.60% F1-score, while the hybrid DCNN-GRU delivers 96.28% accuracy, 96.33% recall, 96.23% precision, and an F1-score 96.28%. These results indicate that VQ-based representations are highly discriminative for dysarthric speech. The best overall performance is achieved using Multiple Acoustic Feature (MAF) fusion (SDF + TDF + VQF). With a DCNN, the system achieves 95.78% accuracy and 95.73% F1-score, while a GRU achieves 97.44% accuracy and 97.47% F1-score. The proposed DCNN-GRU hybrid model achieves the highest scores across all metrics: 98.11% accuracy, 98.33% recall, 97.90% precision, 97.89% selectivity, 98.33% NPV, and 98.12% F1-score. The multi-acoustic feature combined with a hybrid DL architecture provides superior recognition capability for dysarthric speech.

Table 5 and Figure 7 compares three optimization algorithms—Stochastic Gadient Descent Momentum (SGDM), Root Mean Square Propagation (RMSProp), and Adaptive Moment Estimation (Adam)—for training the proposed DSR model using standard performance metrics. The SGDM optimizer achieves 97.22% accuracy, 97.00% recall, 97.43% precision, 97.44% selectivity, 97.01% NPV, and 97.22% F1-score. This indicates stable, balanced performance. The RMSProp slightly improves overall accuracy to 97.50% and F1-score to 97.48%, while showing very high precision (98.20%) and selectivity (98.22%), which means it is more effective in reducing false positives. However, its recall (96.78%) and NPV (96.82%) are marginally lower than SGDM. This indicates a small trade-off in detecting all true dysarthric samples. The Adam optimizer achieves the highest, 98.11% accuracy, 98.89% recall, 98.87% NPV, and 98.13% F1-score. Although its precision (97.37%) and selectivity (97.33%) are slightly lower than RMSProp, Adam maintains a superior balance between sensitivity and specificity. These results demonstrate that Adam offers quicker convergence and improved generalization, making it the highly suitable optimization method for proposed Dysarthric Speech Recognition framework.



(a) deep convolutional neural network (DCNN)

(b) gated recurrent unit (GRU)

(c) DCNN-GRU

Figure 5. Confusion matrix for multiple acoustic feature (MAF)-based dysarthric speech recognition (DSR) for speaker independent mode (SIM)

Table 4. Performance of DSR system for different features and classifiers (SIM)

Features	Classifier	Accuracy	Recall	Precision	Selectivity	NPV	F1-Score
SDF	DCNN	86.22	81.00	90.45	91.44	82.80	85.46
	GRU	87.22	83.33	90.36	91.11	84.54	86.71
	DCNN-GRU	88.56	84.56	91.91	92.56	85.70	88.08
SDF + TDF	DCNN	89.89	93.78	87.01	86.00	93.25	90.27
	GRU	91.67	93.67	90.06	89.67	93.40	91.83
	DCNN-GRU	93.11	93.67	92.64	92.56	93.60	93.15
SDF + VQF	DCNN	93.83	94.56	93.21	93.11	94.48	93.88
	GRU	94.50	96.33	92.93	92.67	96.19	94.60
	DCNN-GRU	96.28	96.33	96.23	96.22	96.33	96.28
MAF (SDF + TDF + VQF)	DCNN	95.78	94.67	96.82	96.89	94.78	95.73
	GRU	97.44	98.44	96.51	96.44	98.41	97.47
	DCNN-GRU	98.11	98.33	97.90	97.89	98.33	98.12

Note: DSR = dysarthric speech recognition, SDF = spectral-domain features, TDF = time-domain features, VQF = voice-quality features, DCNN = deep convolutional neural network, GRU = gated recurrent unit, NPV = negative predictive value, SIM = speaker independent mode; MAFs = multiple acoustic features.

Table 5. Comparison of DCNN-GRU-based DSR for different optimization techniques

Optimization Method	Accuracy	Recall	Precision	Selectivity	NPV	F1-Score
SGDM	97.22	97.00	97.43	97.44	97.01	97.22
RMSProp	97.50	96.78	98.20	98.22	96.82	97.48
Adam	98.11	98.89	97.37	97.33	98.87	98.13

Note: SGDM = Stochastic Gadient Descent Momentum, DSR = dysarthric speech recognition, DCNN = deep convolutional neural network, GRU = gated recurrent unit, NPV = negative predictive value.

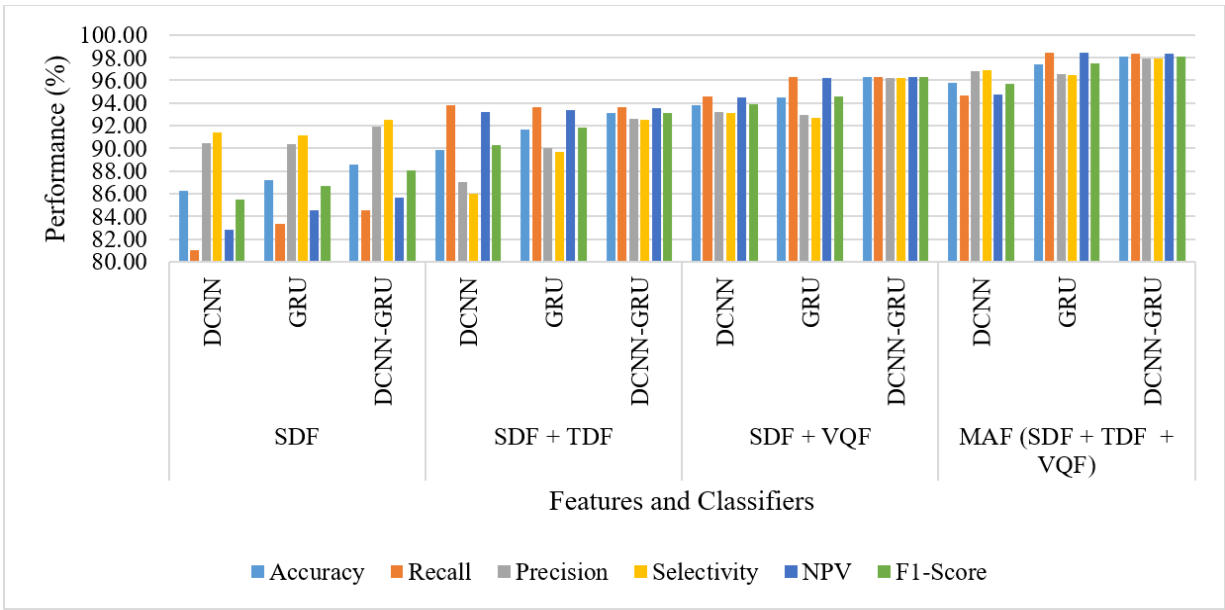


Figure 6. Visualizations of the dysarthric speech recognition (DSR) results for various features and classification methods

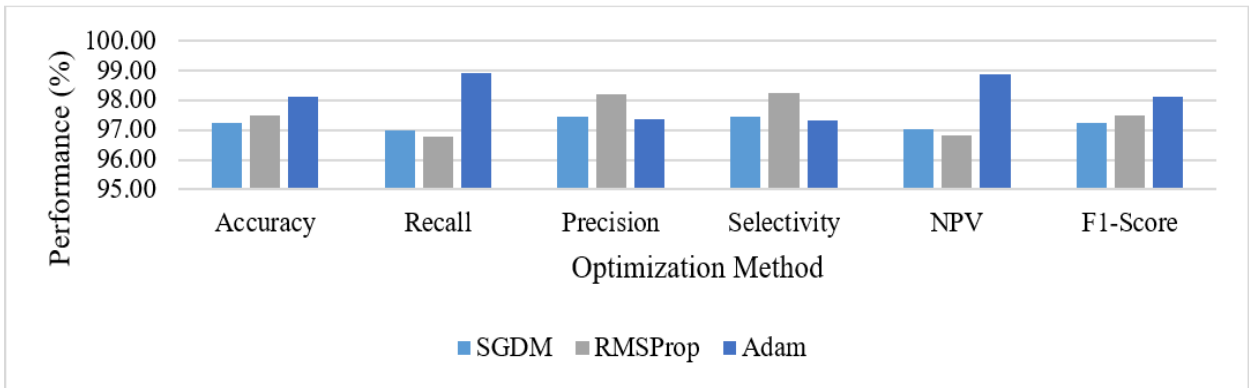


Figure 7. Visualizations of results of deep convolutional neural network-gated recurrent unit (DCNN-GRU)-based dysarthric speech recognition (DSR) for different optimization techniques

Table 6. DSR accuracy comparison for different features chosen using CBFS

Features Selected	DCNN	GRU	DCNN-GRU
50	81	84.2	87.25
100	85.75	90.45	93.1
200	92.4	93.56	96.25
300	95.5	97.3	98.1
400	94.25	96.35	97.9
500	93.5	94.75	97.55
600	92.1	93.2	96.1
700	91.45	92.65	95.25

Note: DSR = dysarthric speech recognition, DCNN = deep convolutional neural network, GRU = gated recurrent unit, CBFS = cross-correlation-based feature selection.

Table 6 presents the accuracy for the different features selected using CBFS. Effectiveness of CBFS-based feature selection is analyzed in the DCNN, GRU, and DCNN-GRU models for DSR. Increasing the number of features (up to 300) improves feature distinctiveness and accuracy. The DCNN-GRU achieves mean overall accuracies of 98.1% with 300 features, 93.1% with 100 features, 87.25% with 50 features, and 96.25% with 200 features. However, increasing the number of features beyond 300 increases redundancy and slightly decreases feature discriminability. The DCNN-GRU achieves accuracies of 97.9%, 97.55%, 96.1%, and 95.25% for 400, 500, 600, and 700 features, respectively. The DCNN-GRU achieves improvements of 0.82% and 2.73% over the GRU and DCNN respectively. The above improvements are for 300 features selected using CBFS. CBFS offers features with lower intra-class correlation and higher inter-class correlation to enhance feature depiction and minimizes redundancy in features.

Table 7 presents the DSR results from 10-fold cross-validation. This shows that all models perform strongly in both SDM and SIM, with consistent improvements as the architectures become more advanced. In SDM, the DCNN model achieves 93.25% mean accuracy ,2.1 standard deviation and a 95% of CI (confidence interval) of [91.45, 95.05], indicating relatively higher variability. The GRU model improves performance to 95.65%, with a reduced standard deviation of 1.6 and a CI of [94.25, 97.05], indicating greater stability. The hybrid DCNN-GRU model achieves the best results with a mean accuracy of 97.15%, the lowest standard deviation of 1.1, and a tighter confidence interval of [96.25, 98.05]. This reflects superior accuracy and robustness. Robust feature selection using CBFS improves model stability, and DCNN-GRU enhances feature distinctiveness, improving the accuracy of the DSR system.

Table 7. Statistical results analysis for SDM and SIM for 10-fold cross-validation

Method	SDM			SIM		
	Mean Accuracy	Standard Deviation	95% of CI	Mean Accuracy	Standard Deviation	95% of CI
DCNN	93.25	2.1	[91.45, 95.05]	95.48	1.85	[94.18, 96.78]
GRU	95.65	1.6	[94.25, 97.05]	97.31	1.3	[96.21, 98.41]
DCNN-GRU	97.15	1.1	[96.25, 98.05]	98.1	0.85	[97.4, 98.8]

Note: SDM = Speaker Dependent Mode, SIM = Speaker Independent Mode, DCNN = deep convolutional neural network, GRU = gated recurrent unit.

Table 8. Ablation study for DCNN-GRU-based DSR for UASpeech dataset

Parameter	Specification	Average Accuracy (%)
Feature Selection Scheme	PCA	95.65%
	ICA	96.20%
	PSO	96.45%
	CBFS	98.11%
Optimization Algorithm	SGDM	97.22
	RMSProp	97.50
	Adam	98.11
Learning Rate	0.0001	97.20
	0.001	98.11
	0.01	96.40
	0.1	93.10
Batch Size	16	96.50
	32	96.85
	64	98.11
	128	97.50

Note: DCNN = deep convolutional neural network, GRU = gated recurrent unit, DSR = dysarthric speech recognition, CBFS = cross-correlation-based feature selection, SGDM = Stochastic Gadient Descent Momentum, PCA = Principal Component Analysis, PSO = Particle Swarm Optimization, ICA = Independent Component Analysis.

4.5 Ablation study

Table 8 provides ablation study of the system. Principal Component Analysis (PCA) obtained an average accuracy of 95.65% for feature selection. Independent Component Analysis (ICA) improved the performance to 96.20% and Particle Swarm Optimization (PSO) to 96.45%. The proposed CBFS method gave the best result of 98.11% accuracy, which is 2.46% better than PCA and about 1.7-1.9% better than ICA and PSO. When it comes to optimization algorithms, SGDM achieved 97.22% accuracy, RMSProp 97.50%, and Adam the highest, 98.11%. This indicates that Adam has the best convergence capability.

The learning-rate effect was also significant. The accuracy was 97.20% with a learning rate of 0.0001, and it was increased to 98.11% with a learning rate of 0.001. However, the performance drops to 96.40% at 0.01, and further drops to 93.10% at 0.1. It shows that excessively high learning rates will negatively affect model stability and generalization. The accuracy of recognition was also influenced by the batch size, 96.50%, 96.85%, and 97.50% were achieved by 16, 32 and 128, respectively. The best performance of 98.11% is achieved with a batch size of 64, indicating that a moderate batch size provides a better trade-off between learning efficiency and model generalization.

4.6 Results comparison with traditional Dysarthric Speech Recognition techniques

The DCNN-GRU-based DSR achieves higher accuracy than earlier DL-based DSR methods, as illustrated in Table 9. Effectiveness of proposed model is compared for various popular ML models with and without feature selection.

The 715 MAFs provide 82.35% accuracy for support vector machine with radial basis function (SVM-RBF), 92.20% for logistic regression (LR), 92.45% for multi-layer perceptron (MLP), 93.105% for DCNN, 93.505% for GRU, and 95.25% for DCNN-GRU. Feature selection using CBFS yields

significant boosts in overall accuracy: 98.11% for DCNN-GRU-CBFS, 97.315% for GRU-CBFS, and 95.45% for DCNN-CBFS with 300 features. The proposed method, which integrates MAFs with a DCNN-GRU, achieves a remarkable 98.83% accuracy on the UASpeech dataset. Compared with Shahamiri [18], who achieved 67.00% accuracy with raw speech and a Spatial CNN, the DCNN-GRU shows a 47.47% improvement. According to Kumar et al. [26], who achieved 75.91% accuracy with a Residual CNN on raw speech, the proposed method shows a 30.22% improvement. In contrast to Farhadipour et al. [25], who achieved 92.30% accuracy with AlexNet on gammatonegram spectrograms, the Proposed Method demonstrates a 7.07% improvement. Even compared to the high-performing method of Janbakhshi et al. [11], which used spectro-temporal features and T-GDA to achieve 96.30%, the Proposed Method still yields a 2.63% improvement. These results highlight the superiority of using a diverse set of features and a deeper DCNN model, making the Proposed Method the most effective approach for DSR in this comparison.

Table 9. DSR accuracy comparison of DCNN-GRU-based DSR with existing methods for UASpeech dataset

Ref.	Feature Representation	DL Algorithm	Performance (Accuracy)
Janbakhshi et al. [11]	STF	T-GDA	96.30%
Shahamiri [19]	Raw speech	S-CNN	67.00%
Farhadipour et al. [25]	Gammatonegram spectrogram	AlexNet	92.3%
Kumar et al. [26]	Raw Speech	Residual CNN	75.91%
Kodrasi et al. [17]	Spectro-temporal representation	CNN	85.72%
Chandrashekar et al. [15]	CQT	CNN	95.8% (Male) 98% (Female)
Proposed Method	Multiple Acoustic Features	SVM-RBF	82.35%
		LR	92.20%
		MLP	92.45%
		DCNN	93.10%
		GRU	93.50%
		DCNN-GRU	95.25%
		DCNN-CBFS	95.48%
		GRU-CBFS	97.31%
		DCNN-GRU-CBFS	98.11%

Note: DSR = dysarthric speech recognition, DL = Deep Learning, SVM-RBF = support vector machine with radial basis function, LR = logistic regression, MLP = multi-layer perceptron, DCNN = deep convolutional neural network, GRU = gated recurrent unit, CBFS = cross-correlation-based feature selection.

5. CONCLUSION AND FUTURE SCOPE

Thus, this article presents DSR using multiple speech features and DCNN-GRU to improve DSR accuracy. The multiple speech features help capture the various changes in the prosodic, SDF, TDF, and VQF parameters of dysarthric speech. Proposed lightweight DCNN improves distinctiveness of dysarthric speech features, thereby enhancing DSR accuracy. The GRU helps achieve temporal modeling and long-term correlations in speech features. The proposed multiple speech features and DCNN-GRU achieve 98.11% accuracy on the UASpeech dataset for DSR. The combination of SDF, TDF, and VQF improves dysarthric speech features and DSR accuracy compared to traditional raw speech + DCNN and MFCC + DCNN. DCNN enhances the uniqueness of traditional features and establishes a good correlation between frame-level, local-level, and global-level features of dysarthric speech.

In the future, more attention can be given to the class imbalance problem caused by uneven sample sizes in training datasets. In the future, the system's generalization capability can be analyzed by conducting experiments on datasets with diverse languages, dialects, genders, and intelligibility levels. The model's explainability and interpretability at the system and user levels can be enhanced to improve trust in the system for potential healthcare applications.

ACKNOWLEDGEMENTS

For continuous guidance, support and for providing research related facilities, authors would like to thank AISSMS Institute of Information Technology, Pune and Dr. D. Y. Patil Institute of Technology, Pune.

REFERENCES

[1] Bhat, C., Strik, H. (2025). Speech technology for automatic recognition and assessment of dysarthric speech: An overview. Journal of Speech, Language, and Hearing Research, 68(2): 547-577. https://doi.org/10.1044/2024_JSLHR-23-00740

[2] Nassif, A.B., Shahin, I., Attili, I., Azzeh, M., Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. IEEE Access, 7: 19143-19165. https://doi.org/10.1109/ACCESS.2019.2896880

[3] Fatehi, K., Torres, M.T., Kucukyilmaz, A. (2025). An overview of high-resource automatic speech recognition methods and their empirical evaluation in low-resource environments. Speech Communication, 167: 103151. https://doi.org/10.1016/j.specom.2024.103151

[4] Chongfuengparinya, T., Piromsopa, K., Chandrachai, A., Kitisomprayoonkul, W., Phuthong, T. (2026). Evolution and knowledge structure of assistive technology adoption for dysarthria: A 20-year bibliometric study. Computers in Human Behavior Reports, 22: 101134. https://doi.org/10.1016/j.chbr.2026.101134

[5] Qian, Z., Xiao, K. (2023). A survey of automatic speech recognition for dysarthric speech. Electronics, 12(20): 4278. https://doi.org/10.3390/electronics12204278

[6] Kotangale, K.B., Parkale, Y.V., Patil, V.N. (2026). Dysarthric speech recognition using multiple speech features and a lightweight deep convolutional neural

network. Ingénierie des Systèmes d'Information, 31(3): 685-693. https://doi.org/10.18280/isi.310303

[7] Pennington, L., Parker, N.K., Kelly, H., Miller, N. (2016). Speech therapy for children with dysarthria acquired before three years of age. The Cochrane Database of Systematic Reviews, 7: CD006937-CD006937. https://doi.org/10.1002/14651858.cd006937.pub3

[8] Vachhani, B., Bhat, C., Kopparapu, S.K. (2018). Data augmentation using healthy speech for dysarthric speech recognition. In Interspeech, pp. 471-475. https://doi.org/110.21437/Interspeech.2018-1751

[9] Yuan, H., Song, Y., Duan, X., Tao, Q., Zhang, N., Yu, Y. (2026). PPFR-conformer for dysarthria speech recognition: from phoneme perception to feature refinement. Neurocomputing, 671: 13268. https://doi.org/10.1016/j.neucom.2026.132684

[10] Sapkota, P., Kathania, H.K., Kutum, S. (2026). Role of SSL models: Finetuning and feature optimization for dysarthric speech recognition and keyword spotting. Computers and Electrical Engineering, 131: 110921. https://doi.org/10.1016/j.compeleceng.2025.110921

[11] Janbakhshi, P., Kodrasi, I., Bourlard, H. (2021). Subspace-based learning for automatic dysarthric speech detection. IEEE Signal Processing Letters, 28: 96-100. https://doi.org/10.1109/LSP.2020.3044503

[12] Bhangale, K., Mohanaprasad, K. (2022). Speech emotion recognition using mel frequency log spectrogram and deep convolutional neural network. In Futuristic Communication and Network Technologies, pp. 241-250. https://doi.org/10.1007/978-981-16-4625-6_24

[13] Fritsch, J., Magimai-Doss, M. (2021). Utterance verification-based dysarthric speech intelligibility assessment using phonetic posterior features. IEEE Signal Processing Letters, 28: 224-228. https://doi.org/10.1109/LSP.2021.3050362

[14] Chandrashekar, H.M., Karjigi, V., Sreedevi, N. (2019). Spectro-temporal representation of speech for intelligibility assessment of dysarthria. IEEE Journal of Selected Topics in Signal Processing, 14(2): 390-399. https://doi.org/10.1109/JSTSP.2019.2949912

[15] Chandrashekar, H.M., Karjigi, V., Sreedevi, N. (2020). Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28(12): 2880-2889. https://doi.org/10.1109/TNSRE.2020.3035392

[16] Kodrasi, I. (2021). Temporal envelope and fine structure cues for dysarthric speech detection using CNNs. IEEE Signal Processing Letters, 28: 1853-1857. https://doi.org/10.1109/LSP.2021.3108509

[17] Mahum, R., Ganiyu, I., Hidri, L., El-Sherbeeny, A.M., Hassan, H. (2025). A novel Swin transformer-based framework for speech recognition for dysarthria. Scientific Reports, 15(1): 20070. https://doi.org/10.1038/s41598-025-02042-7

[18] Yue, Z., Loweimi, E., Cvetkovic, Z., Barker, J., Christensen, H. (2026). Raw acoustic-articulatory multimodal dysarthric speech recognition. Computer Speech & Language, 95: 101839. https://doi.org/10.1016/j.csl.2025.101839

[19] He, Y., Seng, K.P., Lim, C.S., Ang, L.M. (2026). Robust dysarthric speech recognition with GAN enhancement and LLM correction. Advanced Intelligent Systems, 8(2): e202500873. https://doi.org/10.1002/aisy.202500873

[20] Choi, J., Moya-Galé, G., Hwang, K., Hirschberg, J., Levy, E.S. (2026). Automatic speech recognition for intelligibility assessment in children with dysarthria. Journal of Speech, Language, and Hearing Research, 69(4): 1438-1454. https://doi.org/10.23641/asha.31397457

[21] Vijayalakshmi, P., Gladston, A.R., Ramani, B., et al. (2026). Leveraging synthetic speech: TTS-driven data augmentation for effective dysarthric speech recognition. Computer Speech & Language, 100: 101961. https://doi.org/10.1016/j.csl.2026.101961

[22] Vinet, P., Dillenbourg, P., Slot, A., et al. (2026). Automatic Speech recognition and acoustic analysis for dysarthria assessment in telerehabilitation: User-centered design and usability study. JMIR Formative Research, 10(1): e85230. https://doi.org/10.2196/85230

[23] Chung, Y., Hong, J., Lee, J., Kim, E. (2026). Diagnosis-aware multitask fine-tuning of whisper for dysarthric speech recognition. Speech Communication, 103393. https://doi.org/10.1016/j.specom.2026.103393

[24] Fathima, N., Patel, T., Mahima, C., Iyengar, A. (2018). TDNN-based multilingual speech recognition system for low resource Indian languages. In Interspeech, pp. 3197-3201. https://doi.org/10.21437/Interspeech.2018-2117

[25] Farhadipour, A., Veisi, H. (2024). Gammatonegram representation for end-to-end dysarthric speech processing tasks: Speech recognition, speaker identification, and intelligibility assessment. Iran Journal of Computer Science, 7(2): 311-324. https://doi.org/10.1007/s42044-024-00175-y

[26] Kumar, R., Tripathy, M., Anand, R.S., Kumar, N. (2024). Residual convolutional neural network-based dysarthric speech recognition. Arabian Journal for Science and Engineering, 49(12): 16241-16251. https://doi.org/10.1007/s13369-024-08919-5

[27] Kim, H., Hasegawa-Johnson, M., Gfunderson, J., et al. (2023). UASpeech. IEEE DataPort. https://doi.org/10.21227/F9TC-AB45