IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 31, No. 3, March, 2026, pp. 685-693
Journal homepage: http://iieta.org/journals/isi

Dysarthric Speech Recognition Using Multiple Speech Features and a Lightweight Deep Convolutional Neural Network

Kapil Bhaiyalal Kotangale[1,2*], Yuvraj Vijay Parkale[1], Vijay N. Patil[1]

[1] Department of Electronics and Telecommunication, SVPM's College of Engineering, Savitribai Phule Pune University, Pune 413115, India
[2] Department of Electronics and Telecommunication, PCET's Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune 411044, India

Corresponding Author Email: kapilkotangale@gmail.com

https://doi.org/10.18280/isi.310303

Received: 21 August 2025
Revised: 25 October 2025
Accepted: 3 November 2025
Available online: 31 March 2026

Keywords:
dysarthric speech recognition, deep convolution neural network, deep learning, speech recognition, voice pathology

ABSTRACT

Dysarthric speech recognition (DSR) plays a significant role in voice pathology in detecting impaired speech. Lower speech intelligibility, greater variability in speech patterns, and stronger reverberation and background noise lead to a low DSR accuracy rate. Previous DSR systems suffer from poor feature representation, greater complexity in the deep learning (DL) architecture, and limited generalization. This paper presents a DL based DSR method using multiple speech features (MSF) consisting of spectral, prosodic, temporal, and phonetic attributes. The hybrid DL model enhances the distinctive power of MSFs by combining a deep convolution neural network (DCNN) for depicting local-global correlation and a Bidirectional Long Short-term memory (BiLSTM) for acquiring long-term temporal dependencies and correlations in voice. The proposed scheme attains an accuracy of 98.60% for DSR on the UASpeech dataset.

1. INTRODUCTION

Speech is the natural form of human expression for emotions and thoughts. Speech recognition is thus essential to human-computer interaction [1, 2]. The method of identifying human speech using specific computing algorithms is known as automatic speech recognition (ASR). ASR is widely used in automation, affective computing, language identification, speech pathology, speaker recognition, and biometric authorization. Nevertheless, the effectiveness of ASR applications is limited by speech articulation impairment [3-5].

Speech therapy evaluates and treats communication issues and speech abnormalities. Speech-language pathologists (SLPs), sometimes known as speech therapists, carry out this task [6]. Dysarthria is a motor speech condition that develops when the muscles that provide speech signals weaken, are damaged, or are paralysed. Patients with dysarthria have difficulty regulating their voice, tongue, and complex words [7]. Depending on whether a portion of the neurological system is injured, it makes it challenging to produce and pronounce words. Dysarthria is divided into two categories: central dysarthria, which results from damage to the brain, and peripheral dysarthria, which results from damage to the organs required for producing speech. Moreover, dysarthria may be split into acquired and developmental forms. Both before and after childbirth, brain injury may contribute to the development of dysarthria, such as cerebral palsy. A brain injury brings on Dysarthria that develops later in life. Acquired dysarthria may be brought on by a stroke, Parkinson's disease, or a brain tumour. Mumbling, a quiet or soft voice, speaking too quickly or too slowly, and difficulty pronouncing words are all indications of dysarthric disease [8-10].

In recent years, deep learning (DL) methods have been increasingly used in voice-based affective computing and automation systems, as it outperforms more conventional ML-based approaches [11, 12]. Language-related parameters have been shown to be important in developing multilingual ASR. Fathima et al. [13] proposed a time delay neural network (TDNN) to represent acoustic information, which yielded a word error rate (WER) of 16.07% but did not provide local-global connectivity in speech attributes. Yue et al. [14] investigated a multi-spectral acoustic model for dysarthric speech recognition (DSR) based on deep neural network (DNN) and Light Gated Recurrent Unit (LiGRU) technology. Data augmentation minimizes data scarcity by introducing speed perturbations, yielding WERs of 11% and 40.6% for healthy and dysarthric voice, respectively. Furthermore, Yue et al. [15] suggested a hierarchical multiple-stream speech model using LiGRU, fully connected Multi-Layer Perceptrons (MLPs), and convolutional neural networks (CNNs). The electromagnetic articulography (EMA) pre-processed data yielded a WER of 4.6% utilising the suggested model. Butterworth filtering to minimise noise and down-sampling to synchronise Mel Frequency Cepstral Coefficients (MFCCs) attributes are both included in the EMA preprocessing.

The DSR faces significant challenges related to data efficiency. To enhance data, Soleymanpour et al. [16] suggested a text-to-audio synthesizer based on the FastSpeech model. A DNN based on acoustic modelling with a Hidden

Markov Model is presented for representing speech data. Conventional data augmentation methods primarily focus on time-domain changes in the signal, while the spectral properties remain constant. Liu et al. [17] suggested tempo, speech, and vocal tract perturbations to create a synthetic dataset. The DNN-based DSR system achieves a WER of 25.21% on UASpeech and 5.4% on the CUHK. To demonstrate the relationship between phonemes and dysarthric speech, Shahamiri [18] employed voicegrams. To reduce data scarcity in DSR, the visual data expansion paradigm is employed. On the UASpeech dataset, the Spatial-CNN (S-CNN) offers a DSR rate of 67%. The suggested S-CNN occasionally results in a vanishing gradient issue and offers subpar outcomes for mild dysarthria. Temporal-domain variation in dysarthric voice and environmental noise significantly affects speech understandability. According to Lin et al. [19], deep learning-based voice conversion (DVC) employing phonetic posteriorgrams (PPGs) yields more stable results than DVC-Mel in noisy environments.

According to Kodrasi and Bourlard [20], harmonics-to-noise ratio (HNR), MFCC, shimmer, fundamental frequency, and jitter performed worse than sparsity in the spectral-temporal domain using the Gini index for the DSR. Spectral sparsity performs more effectively than temporal sparsity. Additionally, Khodrasi's [20] use of CNNs enabled them to learn the temporal spectral properties of voice. The TEFS surpassed the classic Short-time Fourier Transform-based spectrogram of voice signals. The time-frequency CNN (TF-CNN) was studied by Kodrasi [21] to acquire the spectral-temporal characteristics of dysarthric voice. The spectrogram depiction provides distinctive information about changes in pitch, prosody, and intonation compared with normal temporal attributes. Compared to males, the female subjects' DSR performance showed better accuracy. The scarcity and uneven samples in training data leads to class imbalance issue. The TF-CNN provides a better prosodic depiction of dysarthric speech, significantly improving DSR accuracy compared to conventional artificial neural networks (ANNs) [22]. Fritsch and Magimai-Doss [23] suggested a binary classifier using a recurrent neural network (RNN) and a multi-feature classifier using CNN for DSR. It has shown good correlation with text-to-speech (TTS)-generated synthetic speech, but has not yet proven long-term feature correlation. Bhangale et al. [24] suggested that multiple speech features (MSFs) are crucial for depicting speech attributes. It has shown an imperative boost in affective computing applications, but results are challenging for noisy speech. It used only DCNN to characterize multilevel hierarchical attributes but failed to capture long-term temporal dependencies. Thus, from the extensive survey of DSR, the following gaps are identified:

- Poor feature representation of dysarthric voice is caused by more considerable variability in pitch and rapid disparities in the temporal properties of the voice.
- Poor DSR accuracy for the low intelligibility voices due to noise, reverberations, profound changes in voice, pauses, and high volume.
- Poor spectral-temporal depiction of voice provides poor correlation in local-global depiction of the voice.
- Spectral leakage and lower-frequency resolution problems in conventional MFCC lead to inferior features for dysarthric voice.
- Complexity in existing deep learning frameworks, which leads to training time and trainable parameters of systems, and limits the implementation on resource-constrained devices.

Therefore, this article presents a robust DSR system based on DCNN and bidirectional long short-term memory (BiLSTM) to enhance the accuracy of DSR systems for low-intelligibility voices. The key offerings of the paper are summarized as follows:

- Development of MSFs comprising spectral domain features (SDFs), voice quality features (VQFs), and time domain features (TDFs) to boost the distinctiveness of the dysarthric voices.
- Design of DSR system using a hybrid DCNN-BiLSTM, where DCNN supports boosting the hierarchical feature characterization, and BiLSTM offers the superior temporal and long-term depiction of dysarthric features.

The paper is structured as follows: Section 2 elaborates the different speech features utilized for feature extraction. Further, it gives brief details about the DCNN. Section 3 describes on the simulation outcomes for DSR and discusses them. Finally, Section 4 presents a conclusion and possible future improvements.

2. METHODOLOGY

Figure 1 shows the suggested methodology for the DSR using DCNN-BiLSTM. It involves MSFs extraction using SDF, TDF, and VQFs, feature representation, and classification using DCNN-BiLSTM. The features are normalized using z-score normalization before feeding to the classifier to remove outliers and standardize features. The DCNN comprises of three layers with 64 filters in the 1st layer, 128 filters in the 2nd layer, and 256 filters in the 3rd layer, trailed by rectified linear unit (ReLU) layers and max pooling (MaxPool) layers. The flattened output of the DCNN is provided to the BiLSTM, which includes 50 hidden units. Furthermore, the fully connected (FC) layer enhances feature connectivity. The Softmax layer (SCL) is utilized to categorize normal and dysarthric voices.

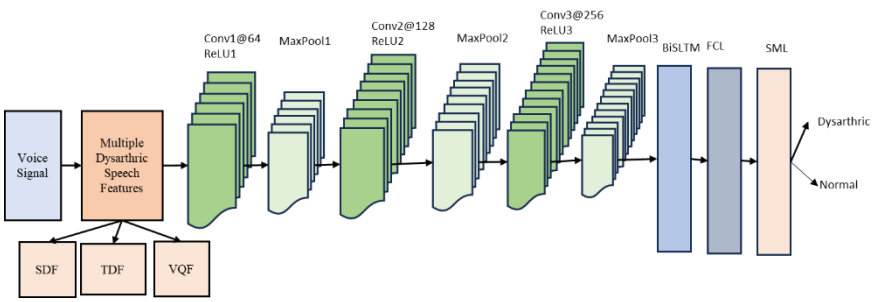

Figure 1. Flow diagram of the dysarthric speech recognition (DSR) system

3. MULTIPLE DYSARTHRIC SPEECH FEATURES

The acoustic elements of the voice represent its amplitude, frequency, and loudness. A series of discrete SDF, TDF, and VQF is presented to describe the speech dysarthric condition. The wavelet packet transform (WPT), spectrum centroid (SC), shimmer, zero crossing rate (ZCR), spectral rolloff, spectral kurtosis (SK), linear predictive cepstral coefficients (LPCC), pitch frequency, jitter, root mean square (RMS), formants, and their mean and standard deviation (SD) are among the acoustic features that were extracted. The voice stream is processed through a moving average filter before computing different attributes to reduce noise and other interference.

3.1 Mel Frequency Cepstral Coefficient

MFCC offers spectral features analogous to human hearing. The MFCC process is described in Figure 2 [24].

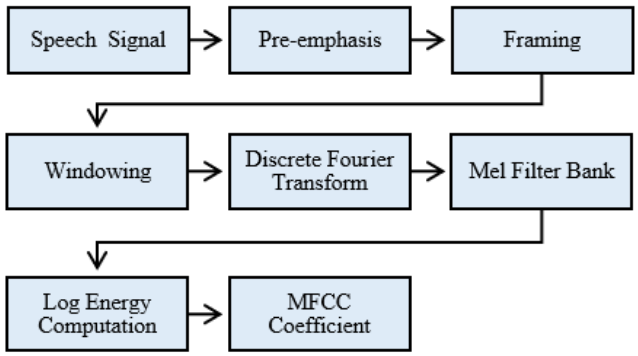

Figure 2. Mel Frequency Cepstral Coefficient (MFCC) feature extraction process

Pre-emphasis standardizes the unprocessed speech stream before MFCC coefficient extraction. The raw dysarthric disordered speech x(n) has noise and disruptions, which are reduced by the pre-emphasis. Additionally, the filtered data is separated into 40-ms frames with a 50% frame shift (20 ms). Overall, 199 frames are produced for 4-second voice when 40-ms frames with 50% overlap are considered. Additionally, the nearest frequency components are gathered together using a Hamming window considering α = 0.46 and *NF* samples per speech frame of 30 ms, as shown in Eq. (1) [24]. Here, n is chosen such that $0 \leq n \leq NF - 1$.

$$H(n) = (1 - \alpha) - \alpha . cos\left(\frac{2\pi n}{(NF - 1)}\right) \tag{1}$$

The time-domain dysarthric disorder speech signal (s) is then converted into the spectral domain ($SP(k)$) using the discrete Fourier transform (DFT) using Eq. (2). Eq. (3) gives the DFT power spectrum, which illustrates the peculiarities of the vocal tract. To deliver the voice-hearing perception described in Eq. (4), the voice is processed through 24 Mel-frequency triangular (MFT) filter banks (M = 24). The linear-to-Mel frequency conversion and its inverse are provided by Eqs. (5) and (6). Here, EMT denotes energy of MFT over the frames.

$$SP(k) = \sum_{n=0}^{N-1} s(n).H(n).e^{-\frac{j2\pi nk}{N}} \quad 0 \leq n, k \leq NF - 1 \tag{2}$$

$$SP_k = \frac{1}{NF}|SP(k)|^2 \tag{3}$$

$$EMT_z = \sum_{k=0}^{k=1} \nabla_z(k).SP_k, z = 1,2, \ldots . M \tag{4}$$

$$MFT = 2595\, log\left(1 + \frac{\mathrm{f}}{700}\right) \tag{5}$$

$$\mathrm{f} = 700\left(10^{\frac{MFT}{2595}} - 1\right) \tag{6}$$

The cepstral coefficients (L) are estimated by applying the Discrete Cosine Transform (DCT) to the energy signal of the MFT output as given in Eq. (7):

$$MFCC_i = \sum_{z=1}^{M} log_{10}(EMT_z).cosj\left((z + 0.5)\frac{\pi}{z}\right) \quad j = 1,2, \ldots L \tag{7}$$

One overall energy feature, 12 MFCC attributes, and 26 1st- and 2nd -order MFCC derivatives make up the overall of 39 features offered by the MFCC. The changeover in speech in a dysarthric disorder must be characterized by the derived traits [15, 17].

3.2 Root mean square

RMS (S_{rms}) depicts the loudness of the voice, which is typically higher in dysarthric voices than in normal voices, and is calculated using Eq. (8):

$$S_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} s_i^2} \tag{8}$$

3.3 Zero crossing rate

ZCR measures the noisiness of the signal by computing the number of zero crossings, which indicate variations in voice due to abrupt changes in pitch and prosody. Eq. (9) is used to estimate ZCR using $sgnc$ function [19]. Within a given time frame (t), the $sgnc$ yields 1 for +ve voice amplitude and 0 for -ve voice amplitude [24].

$$ZCR_t = \frac{1}{2}\left(\sum_{n=1}^{NF}(sgnc(s[n]) - sgnc(s[n-1]\right) \tag{9}$$

3.4 Spectrum centroid

The SIFT SC represents the spectral gravitational centre. It depicts the spectral structure information of the voice. The buildup of broader-frequency elements is indicated by a higher SC value [20]. Eq. (10) offers the magnitude of SIFT ($M_t(nbin)$) across time frame t together with the SC for *nbin* frequency bins [24].

$$SC_t = \frac{\sum_{n=1}^{N} M_t(nbin) * nbin}{\sum_{n=1}^{N} nbin} \tag{10}$$

3.5 Spectral rolloff

Spectral rolloff ($F_{rolloff}$) is a degree of spectral structure that offers spectral components beneath which 85% of the magnitude of SIFT is accumulated. Eq. (11) offers an estimate of spectral rolloff [18].

$$\sum_{n=1}^{F_{rolloff}} M_t(nbin) = 0.85 * \sum_{n=1}^{N} M_t(nbin) \tag{11}$$

3.6 Linear predictive cepstral coefficient

The phonetic depiction of the voice peculiar to dysarthric disorders is represented by the spectral feature known as the LPCC. A total of 13 LPCCs are generated from the linear predictive analysis (LPA) to depict vocal tract attributes [21-24]. The nth samples in an LPA may be predicted using the information from the prior p samples, as shown in Eq. (12):

$$s(n) = \beta_1 s(n-1) + \beta_2 s(n-2) + \beta_3 s(n-3) + \cdots + \beta_p s(n-p) \tag{12}$$

where, $\beta_1, \beta_2, \ldots \beta_p$ describe constants over the voice frame that predict the voice element. Eq. (13) is utilized to examine the disparity between the predicted $\hat{s}(n)$ and real sample $s(n)$.

$$s(n) = s(n) – \hat{s}(n) = s(n) - \sum_{k=1}^{p} \beta_k s(n-k) \tag{13}$$

To acquire the exclusive LPCCs, the error (e_n) between the estimated sample $\hat{s}$(n) and real sample $s(n)$ is assessed utilizing Eq. (14). Here, m denotes samples in one frame [24].

$$e_n = \sum_{z} \left[x(z) - \sum_{k=1}^{p} \beta_k x(z-k) \right]^2 \tag{14}$$

The LPCCs are calculated by estimating solutions for Eqs. (15)-(19).

$$\frac{dE_n}{da_k} = 0, for\ k = 1,2, \ldots p \tag{15}$$

$$C_0 = log_e(p) \tag{16}$$

$$LPCC_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} C_k a_{m-k}, \text{for } 1 < m < p \tag{17}$$

$$LPCC_m = \sum_{k=m-p}^{m-1} \frac{k}{m} C_k a_{m-k}, \text{for } m > p \tag{18}$$

3.7 Spectral kurtosis

The series of transients and their frequency-domain positions are provided by the SK, as depicted in Eq. (19). It describes the asymmetry of the voice spectrogram around its centroid, illustrating the influence of energy and pleasantness instabilities in dysarthric disorder on the spectrogram.

$$SK = \frac{\sum_{k=b1}^{b2} (f_k - €_1)^4 s_k}{(€_2)^4 \sum_{k=b1}^{b2} s_k} \tag{19}$$

where, $€_1$ and $€_2$ represent the SC and frequency spread, s_k is the frequency elements over k bins, and $b1$ and $b2$ are the bins' lower and higher bounds, where the SK of voice is computed.

3.8 Jitter and shimmer

When a vocal fold vibrates irregularly, it causes jitter and shimmer, which are dissimilarities in frequency and amplitude of the dysarthric signal, respectively. The dysarthric disordered sound is characterized by jitter and shimmer, which reflect its roughness, pitch, granularity, breathiness, and hoarseness. The absolute jitter and shimmer values are provided by Eqs. (20) and (21), respectively [24].

$$Jitter = \frac{1}{NP-1} \sum_{i=1}^{N-1} |TP_i - TP_{i+1}| \tag{20}$$

where, TP_i denotes for the time period of successive peak amplitudes in sec and NP depicts the total periods.

$$Shimmer = \frac{\frac{1}{NP-1} \sum_{i=1}^{NP-1} |PA_i - PA_{i+1}|}{\frac{1}{NP} \sum_{i=1}^{NP} PA_i} \tag{21}$$

where, PA_i signifies the peak amplitude of dysarthric speech.

3.9 Pitch frequency

Pitch is vital in illustrating the vocal element of voice (f_0). The pitch is computed using variance of the voice peaks obtained from the voice autocorrelation.

3.10 Formants

Formants identify the more energetic peak occurrences in the voice spectrogram. It describes the vocal tract's resonance phenomena, which are predominantly beneficial for characterizing how dysarthric disease affects them. Three formants ff_1, ff_2, and ff_3 are taken into account for calculation once the formants are computed from the conventional MFCC. To show the formant disparity, the mean and SD of formants is also calculated. Formants (fm), their mean (ffm_u), and their standard deviation (ffm_σ) are all provided in Eqs. (22)-(24), respectively [24].

$$fm = \{ff_1, ff_2, ff_3\} \tag{22}$$

$$ffm_u = \frac{ff_1 + ff_2 + ff_3}{3} \tag{23}$$

$$ffm_\sigma = \sqrt{\frac{\sum_{i=1}^{3} (ff_i - ffm_u)^2}{3}} \tag{24}$$

3.11 Wavelet packet decomposition features

WPT enables the breakdown of complex information into its simplest components across multiple scales and locations, followed by a highly accurate reconstruction of speech, visuals, music, dysarthria, and patterns. WPT aids in the analysis of speech variability caused by various dysarthric conditions. As indicated in Eqs. (25) and (26), the wavelet

packet (WPC) basis function $\Phi_j^i(m)$ utilizing WPT for L-levels is decomposed using the Daubechies (*db*2) filter.

$$\Phi_j^{2i}(m) = \sum_k \underline{h}(k)\Phi_{j-1}^i(m - 2^{j-1}k) \quad (25)$$

$$\Phi_j^{2i+1}(m) = \sum_k \bar{g}(k)\Phi_{j-1}^i(m - 2^{j-1}k) \quad (26)$$

where, $\bar{g}(k)$ and $\underline{h}(k)$ symbolizes the respective high-pass and low-pass quadrature mirror filters given in Eqs. (27) and (28):

$$\underline{h}(k) = \langle \Phi_j^{2i}(p), \Phi_{j-1}^i(p - 2^{j-1}k) \rangle \quad (27)$$

$$\bar{g}(k) = \langle \Phi_j^{2i+1}(p), \Phi_{j-1}^i(p - 2^{j-1}k) \rangle \quad (28)$$

The dysarthric disordered voice is split into subbands at level *j* using Eq. (29):

$$s(m) = \sum_{i.k} X_j^i(k)\Phi_j^i(m - 2^j k) \quad (29)$$

where, $X_j^i(k)$ Kth coefficient is at the ith packet at the *j* level. Eq. (30) denotes the local wavelet energy:

$$X_j^i(k) = \langle s(m), \Phi_j^i(m - 2^j k) \rangle \quad (30)$$

The wavelet packet coefficient (WPC) $X_j^i(k)$ symbolizes the weights of the local WPT denoted by $\Phi_j^i(m - 2^j k)$ as provided in Eq. (31):

$$X_j^i(k) = \langle s(m), \Phi_j^i(m - 2^j k) \rangle \quad (31)$$

Eq. (32) gives a different WPT set for the *L* level.

$$X_L(k) = \begin{bmatrix} X_L^0(k) \\ X_L^1(k) \\ . \\ . \\ X_L^{2^{L-1}}(k) \end{bmatrix} \quad (32)$$

For each sub-band, seven statistical features are computed, including the mean, variance, median, SD, SK, and energy of each wavelet packet, to capture spectral alterations in the voice. The 3-level decomposition of the voice produces a total of 56 WPT characteristics.

For DSR, the DCNN receives the feature depiction (Feat) as input (a total of 715 features) as in Eq. (33):

$$\begin{aligned} \text{MSF} = \{&MFCC_{1-39}, x_{rms}, ZCR_{1-199}, \\ &LPCC_{1-13}, SC_{1-199}, F_{rolloff}, WPT_{1-56} SK_{1-199}, \\ &Shimmer, Jitter, f_0, fm_{1-3}, ffm_u, ffm_\sigma\} \end{aligned} \quad (33)$$

3.12 Deep convolution neural network–bidirectional long short-term memory model

Figure 3 provides the process flow of the suggested DSR scheme. The DCNN helps improve spatial correlation, multilevel hierarchical depiction, and correlation between local and global attributes of speech [25-27]. The 1D-DCNN accepts the input of 715 voice features comprising SDF, TDF, and VQF. The conv layer provides the correlation in features (*MSF*) using learnable filters (ω). The conv operation is provided by Eq. (34) where z denotes convolution output and σ symbolizes the ReLU. The DCNN consists of three layers with 64 filters in the 1st layer, 128 filters in the 2nd layer, and 256 filters in the 3rd layer. Increasing the number of convolutional layers enhances multilevel correlation and hierarchical feature representation by providing spatial connectivity. The ReLU accelerates training, enhances non-linearity, and reduces the risk of vanishing gradients by substituting negative neurons with zero, as given in Eq. (35):

$$z(n) = \sum_{m=0}^{i-1} MSF(m).\omega(n - m) \quad (34)$$

$$\sigma(z) = max(0, z) \quad (35)$$

The output of the 3rd ReLU layer is flattened and provided to the BiLSTM layer with 50 units. The BiLSTM provides contextual information in voice by modeling long-term dependencies in the DCNN representation of the voice features in both forward and reverse directions. The forward LSTM ($\overrightarrow{h_t}$) and backward states ($\overleftarrow{h_t}$) of BiLSTM are provided in Eqs. (36)-(38):

$$\overrightarrow{h_t} = LSTM_{forward}(\sigma(z)) \quad (36)$$

$$\overleftarrow{h_t} = LSTM_{backward}(\sigma(z)) \quad (37)$$

$$h_t^{BiLSTM} = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (38)$$

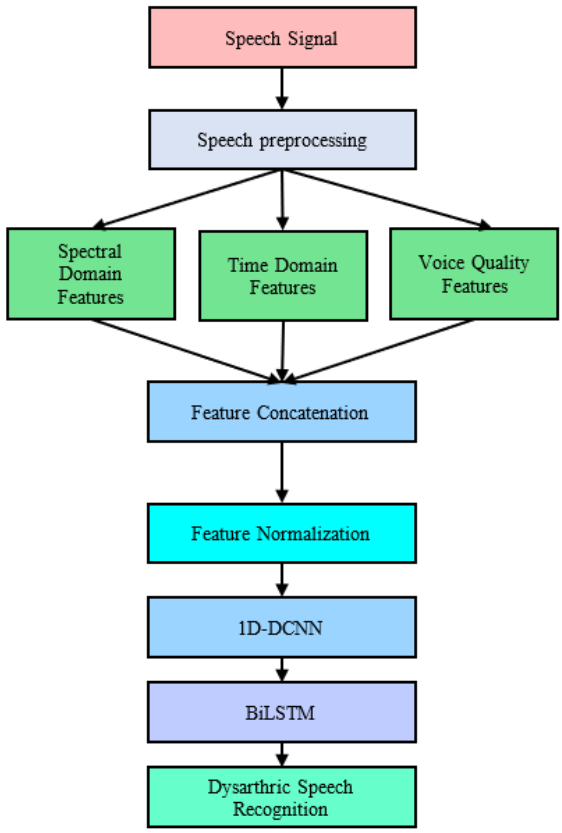

Figure 3. Illustration of proposed dysarthric speech recognition (DSR) system
Note: 1D-DCNN: one-dimensional deep convolutional neural network, BiLSTM: bidirectional long short-term memory

The BiLSTM output is provided to the fully connected (FC) layer to offer the linkage between all neurons of the network, as shown in Eq. (39).

$$y_{jk}(x) = f(\sum_{i=1}^{n_H} W_{jk}x_i + wj_0) \quad (39)$$

where, x_i denotes 1D flattened feature vector, w_0 symbolizes a bias, w describes the weight matrix, f depicts a nonlinear activation function (NAF), y stands for the output of NAF, and n_H offers hidden layers. Finally, the SCL provides the likelihood of the output, where the class label is determined by the maximum likelihood score, as shown in Eqs. (40)-(42) [15]:

$$\text{ʉ}_i = \sum_j h_j w_{ji} \quad (40)$$

$$\text{p}_i = \frac{exp\ (\text{ʉ}_i)}{\sum_{j=1}^{n} exp\ (\text{ʉ}_j)} \quad (41)$$

$$\hat{y} = arg\max_i \text{p}_i \quad (42)$$

where, h_j is the weight of the next to last layer and w_{ji} describes the weights of SCL and the next to last layer, ʉ_i is the input of SCL, p_i is the likelihood of the class label, and $\hat{y}$ is the predicted class.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

The approach is implemented in MATLAB on a personal computer (Core i5 processor and 16 GB of RAM). The training accuracy and loss curves of the network are illustrated in Figure 4, while the detailed parameter settings of the DCNN-LSTM model are listed in Table 1.

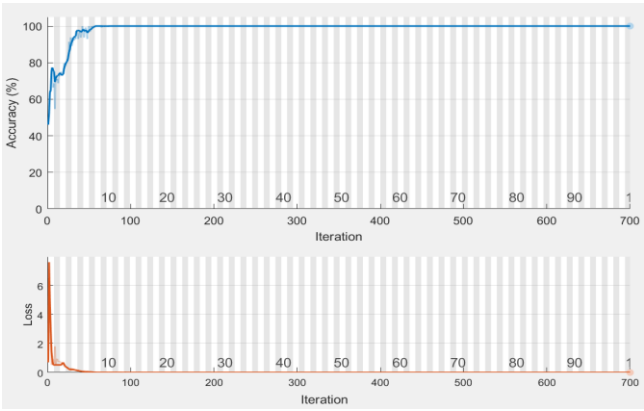

Figure 4. Training accuracy and loss of deep convolution neural network (DCNN)

Table 1. Deep convolution neural network (DCNN) parameter configuration

Parameter	Specifications
Learning Rate	0.001
Learning Algorithms	Adam, RMSProp, SGDM
Decay Rate	0.5
Loss Function	Cross Entropy
Train-Test Data Split	70:30
Epochs	100

Table 2 and Figures 5-8 show the results for various methods for Adam, SGDM, and RMPProp learning algorithms. Among all the approaches tested, the MSF-DCNN-BiLSTM model using the Adam optimizer stood out as the top performer, reaching an impressive accuracy of 98.83%, with precision, recall, and F1-score all at 0.99. These near-perfect scores highlight the effectiveness of combining MSF, spatial learning through DCNN, and time-sequence modeling via BiLSTM.

Even when trained with other optimizers like SGD and RMSProp, the same model still performed very well, achieving 95.48% and 94.76% accuracy, respectively. Looking at the MSF-DCNN model on its own (without the BiLSTM), it also performed well, particularly with Adam, reaching 97.45% accuracy, 0.97 precision, 0.98 recall, and 0.98 F1-score. However, it fell just short of the combined model, suggesting that the BiLSTM layer adds significant value by capturing the time-varying nature of speech. Likewise, the MSF-BiLSTM model (which uses BiLSTM without DCNN) achieved a strong 95.80% accuracy and 0.96 F1-score using Adam, demonstrating that even without convolutional layers, sequential modeling can effectively handle rich feature inputs.

Methods using only MFCC features didn't perform quite as well. The MFCC-DCNN model with Adam had an accuracy of 92.38%, while MFCC-BiLSTM scored 91.17%. These are still good results, but they reflect the limitations of MFCCs, as they don't capture as much detailed or complementary information as MSF does—particularly when dealing with the complex patterns of dysarthric speech.

The lowest results came from models that used the original, raw audio signals without any engineered features. For example, the Original Signal-DCNN model with Adam achieved only 88.57% accuracy, whereas the Original Signal-BiLSTM reached 87.01%, indicating that raw audio lacks the structured information necessary for these deep models to perform well. When comparing optimizers, Adam consistently outperformed both SGDM and RMSPROP across the board. Take the MSF-BiLSTM model, for instance—Adam yielded 95.80% accuracy, while SGDM dropped to 93.08% and RMSPROP to 91.46%.

The statistical analysis of the results reveals that the MSF-DCNN-BiLSTM model consistently achieved the highest performance across all learning algorithms, with Adam optimizer demonstrating superior stability and accuracy. Specifically, Adam achieved an accuracy of 98.83 ± 0.15%, a precision of 0.99 ± 0.01, a recall of 0.99 ± 0.01, and an F1-score of 0.99 ± 0.01, indicating both excellent classification performance and minimal variation across runs. The SGDM and RMSPROP optimizers, though slightly lower in performance (95.48 ± 0.34% and 94.76 ± 0.51% accuracy, respectively), maintained consistent results with moderate variability. The MSF-DCNN and MSF-BiLSTM architectures also demonstrated strong results, with Adam achieving 97.45 ± 0.22% and 95.80 ± 0.19% accuracy, respectively, showing that both spatial and temporal feature fusion significantly enhance robustness. In contrast, MFCC-based models (MFCC-DCNN and MFCC-BiLSTM) exhibited moderate accuracy, ranging from 91.17 ± 0.29% to 92.38 ± 0.26%, indicating that handcrafted features like MFCC are less discriminative compared to the multiscale fusion approach.

The original signal-based models (DCNN and BiLSTM) showed the lowest accuracies, with Adam achieving 88.57 ± 0.33% and 87.01 ± 0.31%, and RMSPROP performing the

lowest with 81.90 ± 0.71% and 80.65 ± 0.74%, respectively. The larger standard deviations observed in RMSPROP (up to ± 0.74) indicate unstable convergence and greater sensitivity to hyperparameter variations. Overall, the Adam optimizer consistently delivered the most stable and accurate performance (lowest SD ≈ 0.15–0.33), while SGDM offered moderate stability (SD ≈ 0.34–0.58), and RMSPROP exhibited the highest fluctuation (SD ≈ 0.5–0.74). Hence, the MSF-DCNN-BiLSTM + Adam combination is statistically validated as the most robust and generalizable configuration among all tested models.

Table 2. Result comparison of the dysarthric speech recognition (DSR) system based on mean accuracy and standard deviation (Mean ± SD) for different optimizer (*Adam: Opt1, SGDM: Opt2, RMSPROP: Opt3*)

Method	Learning Algorithm	Accuracy	Precision	Recall	F1-Score
MSF-DCNN-BiLSTM	Opt1	98.83 ± 0.15	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01
	Opt2	95.48 ± 0.34	0.95 ± 0.02	0.96 ± 0.02	0.95 ± 0.02
	Opt3	94.76 ± 0.51	0.95 ± 0.03	0.94 ± 0.03	0.95 ± 0.03
MSF-DCNN	Opt1	97.45 ± 0.22	0.97 ± 0.01	0.98 ± 0.01	0.98 ± 0.01
	Opt2	94.24 ± 0.41	0.93 ± 0.02	0.94 ± 0.02	0.94 ± 0.02
	Opt3	92.78 ± 0.56	0.93 ± 0.03	0.92 ± 0.03	0.93 ± 0.03
MSF-BiLSTM	Opt1	95.80 ± 0.19	0.95 ± 0.01	0.97 ± 0.01	0.96 ± 0.01
	Opt2	93.08 ± 0.37	0.92 ± 0.02	0.93 ± 0.02	0.92 ± 0.02
	Opt3	91.46 ± 0.49	0.92 ± 0.03	0.90 ± 0.03	0.91 ± 0.03
MFCC-DCNN	Opt1	92.38 ± 0.26	0.93 ± 0.02	0.92 ± 0.01	0.92 ± 0.02
	Opt2	90.71 ± 0.44	0.91 ± 0.02	0.90 ± 0.02	0.91 ± 0.02
	Opt3	87.86 ± 0.63	0.87 ± 0.03	0.89 ± 0.03	0.88 ± 0.03
MFCC-BiLSTM	Opt1	91.17 ± 0.29	0.91 ± 0.02	0.90 ± 0.01	0.91 ± 0.02
	Opt2	89.18 ± 0.46	0.90 ± 0.02	0.89 ± 0.02	0.89 ± 0.02
	Opt3	86.51 ± 0.68	0.85 ± 0.03	0.88 ± 0.03	0.87 ± 0.03
Original Signal - DCNN	Opt1	88.57 ± 0.33	0.90 ± 0.02	0.87 ± 0.02	0.88 ± 0.02
	Opt2	83.57 ± 0.52	0.83 ± 0.03	0.84 ± 0.03	0.84 ± 0.03
	Opt3	81.90 ± 0.71	0.82 ± 0.04	0.81 ± 0.04	0.82 ± 0.04
Original Signal - BiLSTM	Opt1	87.01 ± 0.31	0.88 ± 0.02	0.85 ± 0.02	0.87 ± 0.02
	Opt2	81.89 ± 0.58	0.81 ± 0.03	0.83 ± 0.03	0.82 ± 0.03
	Opt3	80.65 ± 0.74	0.81 ± 0.04	0.80 ± 0.04	0.80 ± 0.04

Note: MSFs: multiple speech features; DCNN: deep convolution neural network; BiLSTM: bidirectional long short-term memory; MFCCs: Mel Frequency Cepstral Coefficients

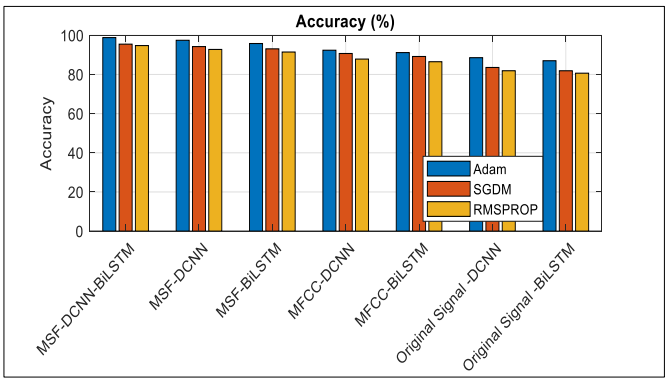

Figure 5. Accuracy comparison of the dysarthric speech recognition (DSR) system

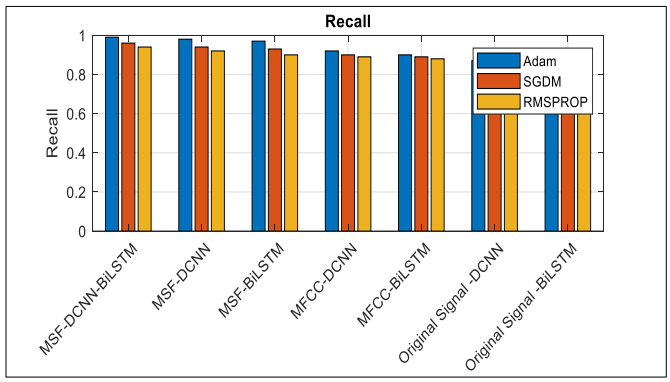

Figure 6. Recall comparison of the dysarthric speech recognition (DSR)system

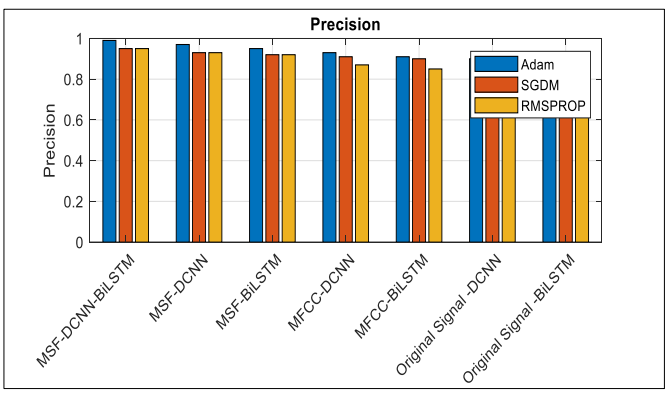

Figure 7. precision comparison of the dysarthric speech recognition (DSR) system

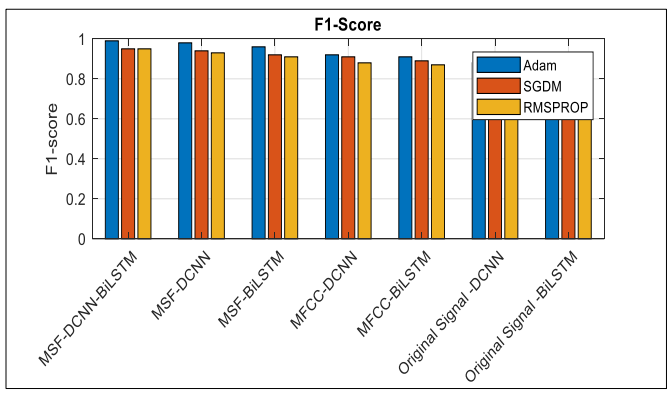

Figure 8. F1-score comparison of the dysarthric speech recognition (DSR) system

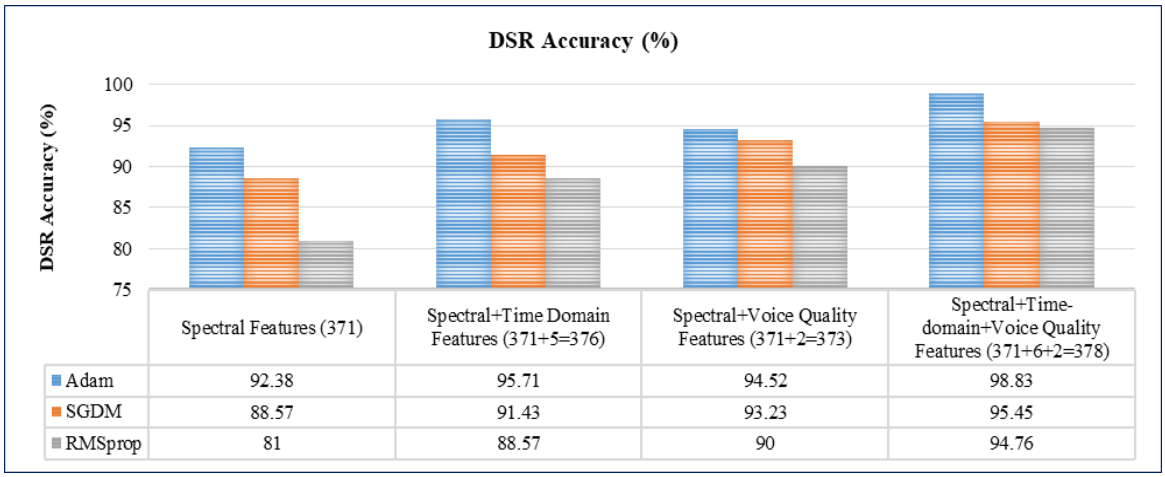

Figure 9. Accuracy for different feature combinations

Table 3. Performance comparison on the dysarthric speech recognition system with the previous state of the art for UASpeech

Ref.	Feature Representation	Deep Learning Algorithm	Performance (Accuracy)
[10]	Spectro-temporal representation of speech	T-GDA	96.30%
[18]	Raw speech	Visual acoustic model based on S-CNN	67.00%
Proposed method	MSF	DCNN	98.83%

The outcomes of the DSR method are assessed for the various feature combinations and learning algorithms, as shown in Figure 9. The spectral features and DCNN provide 81.00%, 88.57% and 92.38% accuracy for RMSPROP, SGDM, and Adam optimization techniques.

The DSR scheme provides superior performance compared with previous DL-based DSR schemes, as shown in Table 3. The T-GDA and S-CNN achieve accuracies of 96.30% and 67% on the UASpeech dataset, respectively. The DSR scheme attains an accuracy of 98.83% on the UASpeech dataset, with superior recall and precision.

5. CONCLUSIONS AND FUTURE SCOPE

This paper presented a DSR framework that integrates MSF with DCNN to enhance recognition accuracy. The MSF effectively captures variations in spectral, prosodic, temporal, intonation and voice quality parameters of dysarthric speech. The lightweight DCNN enhances the distinctiveness of the extracted features, resulting in improved classification performance. The MSF–DCNN model achieves an impressive 98.83% accuracy on the UASpeech dataset. The MSFs enhance the discriminative power of dysarthric speech compared to traditional approaches such as raw speech + DCNN and MFCC + DCNN. Moreover, the DCNN strengthens the correlation between global and local (frame-level) features of dysarthric signals, further improving recognition accuracy.

However, the system's performance is still constrained by limited generalization, speaker dependency, and data scarcity. Future improvements can focus on using larger, more diverse cross-lingual datasets to enhance generalization. Additionally, integrating transfer learning and self-attention mechanisms could further refine system robustness. In the long term, the DSR framework can be adapted for deployment on standalone, resource-constrained devices. Addressing the data imbalance caused by uneven training sample counts will also be a key focus of future research.

REFERENCES

[1] Watanabe, S., Delcroix, M., Metze, F., Hershey, J.R. (2017). New Era for Robust Speech Recognition. Springer International Publishing. https://doi.org/10.1007/978-3-319-64680-0

[2] Bhangale, K.B., Mohanaprasad, K. (2021). A review on speech processing using machine learning paradigm. International Journal of Speech Technology, 24(2): 367-388. https://doi.org/10.1007/s10772-021-09808-0

[3] Vadwala, A.Y., Suthar, K.A., Karmakar, Y.A., Pandya, N., Patel, B. (2017). Survey paper on different speech recognition algorithm: Challenges and techniques. International Journal of Computer Applications, 175(1): 31-36.

[4] Sonawane, A., Inamdar, M.U., Bhangale, K.B. (2017). Sound based human emotion recognition using MFCC & multiple SVM. In 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), Indore, India, pp. 1-4. https://doi.org/10.1109/ICOMICON.2017.8279046

[5] Bhangale, K.B., Kothandaraman, M. (2022). Survey of deep learning paradigms for speech processing. Wireless Personal Communications, 125(2): 1913-1949. https://doi.org/10.1007/s11277-022-09640-y

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

[7] Jamal, N., Shanta, S., Mahmud, F., Sha'abani, M.N.A.H. (2017). Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review. AIP Conference Proceedings, 1891: 020028. https://doi.org/10.1063/1.5002046

[8] Vachhani, B., Bhat, C., Kopparapu, S.K. (2018). Data augmentation using healthy speech for dysarthric speech recognition. In Interspeech, Hyderabad, India, pp. 471-475. https://doi.org/10.21437/Interspeech.2018-1751

[9] Takashima, R., Takiguchi, T., Ariki, Y. (2020). Two-step acoustic model adaptation for dysarthric speech recognition. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 6104-6108. https://doi.org/10.1109/ICASSP40776.2020.9053725

[10] 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

[11] Bhangale, K.B., Titare, P., Pawar, R., Bhavsar, S. (2018). Synthetic speech spoofing detection using MFCC and radial basis function SVM. IOSR Journal of Engineering, 8(6): 55-62.

[12] Bhangale, K., Mohanaprasad, K. (2021). Speech emotion recognition using mel frequency log spectrogram and deep convolutional neural network. Lecture Notes in Electrical Engineering, 767: 241-250. https://doi.org/10.1007/978-981-16-4625-6_24

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

[14] Yue, Z., Loweimi, E., Cvetkovic, Z. (2022). Raw source and filter modelling for dysarthric speech recognition. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, pp. 7377-7381. https://doi.org/10.1109/ICASSP43922.2022.9746553

[15] Yue, Z., Loweimi, E., Cvetkovic, Z., Christensen, H., Barker, J. (2022). Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, pp. 7372-7376. https://doi.org/10.1109/ICASSP43922.2022.9746855

[16] Soleymanpour, M., Johnson, M.T., Soleymanpour, R., Berry, J. (2022). Synthesizing dysarthric speech using multi-speaker TTS for dysarthric speech recognition. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp. 7382-7386. https://doi.org/10.1109/ICASSP43922.2022.9746585

[17] Liu, S., Geng, M., Hu, S., Xie, X., Cui, M., Yu, J., Liu, X., Meng, H. (2021). Recent progress in the CUHK dysarthric speech recognition system. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29: 2267-2281. https://doi.org/10.1109/TASLP.2021.3091805

[18] Shahamiri, S.R. (2021). Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29: 852-861. https://doi.org/10.1109/TNSRE.2021.3076778

[19] Lin, Y.Y., Zheng, W.Z., Chu, W.C., Han, J.Y., Hung, Y.H., Ho, G.M., Chang, C.Y., Lai, Y.H. (2021). A speech command control-based recognition system for dysarthric patients based on deep learning technology. Applied Sciences, 11(6): 2477. https://doi.org/10.3390/app11062477

[20] Kodrasi, I., Bourlard, H. (2020). Spectro-temporal sparsity characterization for dysarthric speech detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28: 1210-1222. https://doi.org/10.1109/TASLP.2020.2985066

[21] 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

[22] 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

[23] 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

[24] Bhangale, K., Kothandaraman, M. (2023). Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. Electronics, 12(4): 839. https://doi.org/10.3390/electronics12040839

[25] Widodo, C.E., Adi, K., Priyono, P., Setiawan, A. (2023). An evaluation of pre-trained convolutional neural network models for the detection of COVID-19 and pneumonia from chest X-ray imagery. Mathematical Modelling of Engineering Problems, 10(6): 2210-2216. https://doi.org/10.18280/mmep.100635