International Journal of Informatics and Communication Technology (IJ-ICT)
Vol. 12, No. 3, December 2023, pp. 242~250
ISSN: 2252-8776, DOI: 10.11591/ijict.v12i3.pp242-250

A comprehensive survey of automatic dysarthric speech recognition

Shailaja Yadav[1], Dinkar Manik Yadav[2], Kamalakar Ravindra Desai[3]

[1]Department of Electronics and Telecommunication, G.H. Raisoni College of Engineering and Management, Pune, India
[2]Department of Electronics and Telecommunication, Faculty of Engineering, S.N.D College of Engineering and Research Center, Yeola, India
[3]Department of Electronics and Telecommunication, Faculty of Engineering, Bharati Vidyapeeth College of Engineering, Kolhapur, India

Article Info

Article history:

Received Nov 19, 2022
Revised Mar 9, 2023
Accepted Apr 24, 2023

Keywords:

Dysarthric speech recognition
Speech intelligibility
Speech recognition
Voice pathology

ABSTRACT

The need for automated speech recognition has expanded as a result of significant industrial expansion for a variety of automation and human-machine interface applications. The speech impairment brought on by communication disorders, neurogenic speech disorders, or psychological speech disorders limits the performance of different artificial intelligence-based systems. The dysarthric condition is a neurogenic speech disease that restricts the capacity of the human voice to articulate. This article presents a comprehensive survey of the recent advances in the automatic dysarthric speech recognition (DSR) using machine learning (ML) and deep learning (DL) paradigms. It focuses on the methodology, database, evaluation metrics, and major findings from the study of previous approaches. From the literature survey it provides the gaps between exiting work and previous work on DSR and provides the future direction for improvement of DSR. The performance of the various machine and DL schemes is evaluated for the DSR on UASpeech dataset based on accuracy, precision, recall, and F1-score. It is observed that the DL based DSR schems outperforms the ML based DSR schemes.

Corresponding Author:

Shailaja Yadav
Department of Electronics and Telecommunication, Faculty of Engineering
G.H. Raisoni College of Engineering and Management
Flat No.302, Aditya Malhar, Pipeline Road, Ravet-Pune, Maharashtra, India
Email: shailaja.yadav123@gmail.com

1. INTRODUCTION

Dysarthria is a speech disorder generated due to weakness in speed production muscle or when an individual is unable to control them. It frequently causes slow or slurred speech which is difficult to understand. Dysarthria can be caused due to neural disorder, troat or tongue muscle weakness, or facial paralysis [1], [2]. The muscle used for speed production is controlled by the nervous system and brain. Mostly dysarthria is caused due to damage to these muscles. Dysarthria is grouped into developmental and acquired dysarthria. The developmental dysarthria normally found in children is occurred due to brain damage during or before birth. The acquired dysarthria generally occurred due to brain damage in adulthood or later in life such as brain tumors, stroke, head injury, motor neuron disease, or Parkinson's disease [3]–[5].

The term "dysarthria" refers to a variety of neurological speech abnormalities caused by injury to the central or peripheral nerve systems. Reduced stress, sluggish speech pace, hyper-nasality, muscular stiffness, spasticity, monopitch, and a limited range of speech motions are all signs of dysarthric speech. It can impact the subglottal, laryngeal, and articulatory systems, which can make speech production difficult. Stroke, Parkinson's disease, and cerebral palsys are the most common roots of motor speech difficulties. According to reports, improving human-machine interaction for persons with dysarthria is becoming increasingly important in order to boost overall wellness and independence. Physical impairments are common in people with dysarthria, making common input methods (typing and touch screen) difficult to use [6], [7].

Traditionally, the language or speech therapist diagnosed dysarthria disorder by asking people to read passages loudly, recite numbers or weekdays, make various sounds or talk about any familiar topic. The traditional techniques performance is limited due to various factors such as inadequate knowledge of experts, tiredness, and fatigue. Dysarthria may affect phonation, breathing, prosody, articulation, resonance, and lip movement. It shows a larger variation in speech intelligibility. The scope of intelligibility is huge and may depend upon the extent of nervous system damage. The typical symptoms of the dysarthria are listed in Figure 1. Because of articulatory difficulties, there is no uniformity in articulation. Pronunciation changes and speaking pace slows as a result of exhaustion. All of these distinctiveness impair the dysarthric speaker's intelligibility (the degree to which others can understand their speech) and limit verbal interactions, reducing their quality of life [8], [9].

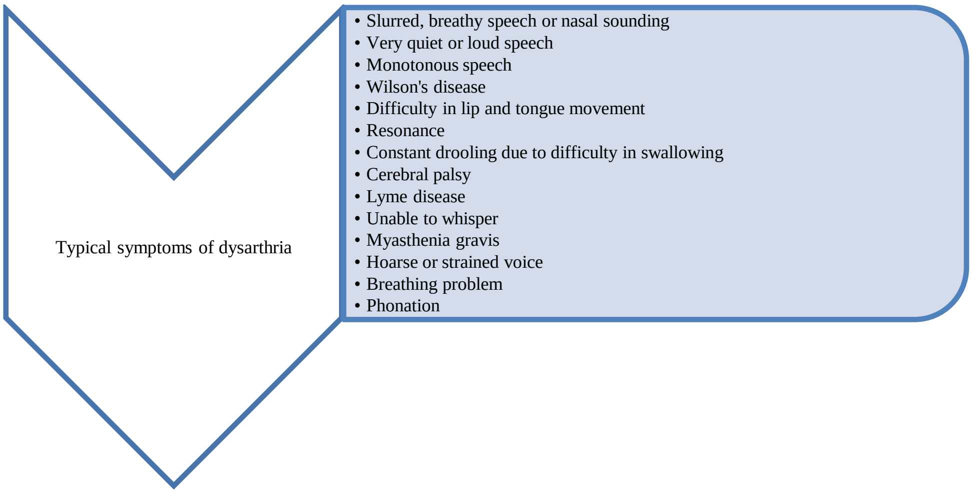

Typical symptoms of dysarthria
• Slurred, breathy speech or nasal sounding
• Very quiet or loud speech
• Monotonous speech
• Wilson's disease
• Difficulty in lip and tongue movement
• Resonance
• Constant drooling due to difficulty in swallowing
• Cerebral palsy
• Lyme disease
• Unable to whisper
• Myasthenia gravis
• Hoarse or strained voice
• Breathing problem
• Phonation

Figure 1. Typical symptoms of dysarthria

The classification system helps to narrow down the dimension of perceptual analysis of dysarthric speech. The classification of dysarthric speech is given in Figure 2. Most clinicians find this useful to correct or reduce the deficit found in dysarthric speech production. Normal speakers typically communicate at rates between 150 to 200 words per minute. The speech is clear, timely, and contextually relevant. Speakers with severe impairments communicate at a rate of fewer than 15 words per minute. This reduction in the rate of communications has implications in the quantity and the quality. People suffering from dysarthria are generally physically challenged. It is difficult for them to handle the conventional keyboard or mouse interfaces. Dysarthric speakers experience difficulty to contribute enough samples of speech data. Some dysarthric speakers get tired soon which may lead to distress. They often fall short to utter certain sounds, which results in phonetic variation [10], [11].

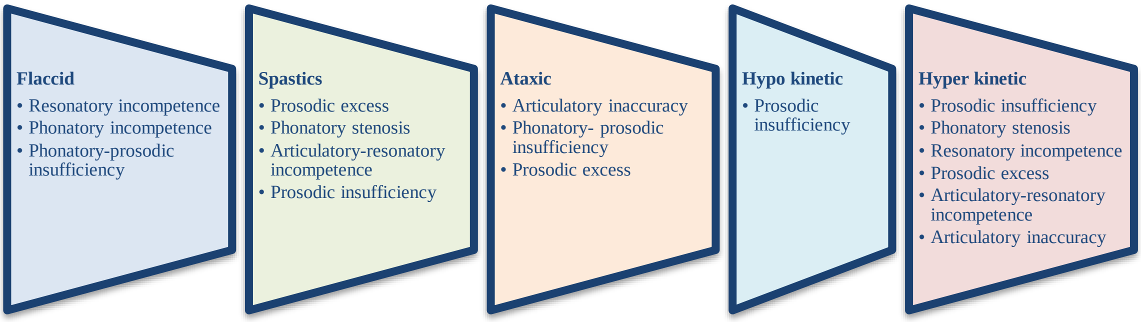

Flaccid
• Resonatory incompetence
• Phonatory incompetence
• Phonatory-prosodic insufficiency
Spastics
• Prosodic excess
• Phonatory stenosis
• Articulatory-resonatory incompetence
• Prosodic insufficiency
Ataxic
• Articulatory inaccuracy
• Phonatory- prosodic insufficiency
• Prosodic excess
Hypo kinetic
• Prosodic insufficiency
Hyper kinetic
• Prosodic insufficiency
• Phonatory stenosis
• Resonatory incompetence
• Prosodic excess
• Articulatory-resonatory incompetence
• Articulatory inaccuracy

Figure 2. Types and features of dysarthria

The generalized process of dysarthric speech recognition (DSR) is shown in Figure 3 that encompasses the pre-processing, feature representation, classification, and DSR. The pre-processing phase deals with the primary processing on the dysarthric speech to improve the quality of features and performance of the classifiers. It encompasses framing, cropping, speech separation, noise suppression, windowing, normalization, speech enhancement, and data augmentation. The dysarthric speech contains different types of the reverberations, silent regions, stops, wide variety in pitch, and energy of the signal which tends to use speech enhancement to enhance DSR effectiveness. The feature extraction is important phase to collect the distinctive and unique characteristics of the normal and dysarthric speech. The features are generally grouped into spectral, prosodic, voice quality, and teager-energy operator features. Traditional machine learning (ML) based DSR includes feature extraction followed by classification whereas in deep learning (DL) the feature extraction may not be used as DL techniques often refers to combination of hidden feature extraction layers and classification layer. However, many hybrid DL algorithms uses the traditional features as the input to boost the speech intelligibility, feature representation, and DSR accuracy.

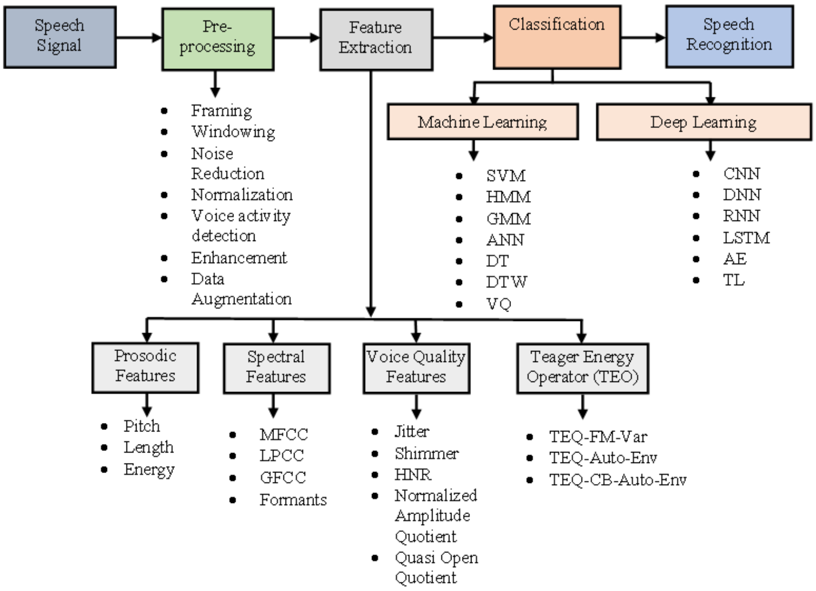


Figure 3. Generalized process of DSR

Various DSR strategies have been presented in last two decades. This section gives a quick overview of recent DSR approaches. Voice tremor has been quantified using phonation parameters that define disordered voice, such as jitter and fundamental frequency [9], [12]. To avoid the gender and acoustic environment dependence of these parameters, a pitch period entropy-based evaluation was developed [13]. Hypophonia has also been described using fluctuation of energy and short-time energy [14]. The Teager-Kaiser energy operator which provides the speech intensity measure is utilized to adjust for signal frequency [15]. To explore the influence on articulatory dynamics and speech intelligibility, acoustic cues based on the first three formants and their respective bandwidths can be studied [16]. Vowel space area (VSA) has been investigated for assessing speech intelligibility [17]. A support vector machine (SVM) classifier was used to investigate a method for distinguishing dysarthric speech from healthy speech using a collection of glottal and openSMILE characteristics [18]. Gurugubelli and Vuppala [19] investigated analytic phase characteristics generated from voice signals using the single frequency filtering (SFF) approach. Audio descriptor information used for determining musical instrument timbre were combined with an artificial neural network (ANN) model to classify dysarthric speech severity levels [20]. For dysarthria classification, multi-tapered spectral estimation was used to extract audio descriptor features.

Research by Johnson *et al.* [21] evaluate recognition performance for dysarthric speech compared with automatic speech recognition (ASR) systems based on Gaussian mixture model (GMM) hidden Markov models (HMMs) and SVMs [22]. The experimental results showed that the HMM-based model may provide robustness against large-scale word-length variances. Meanwhile, the SVM-based model can alleviate the effect of deletion of or reduction in consonants. Rudzicz [23] investigated acoustic models of GMM–HMM, conditional random field, SVM, and ANNs [24]. The results showed that the ANNs provided higher accuracy than other models. Revathi *et al.* [25] presented multiple such as Gamma tone energy (GFE), modified group delay function cepstrum (MGDFC), and stock well features for isolated DSR. It used decision level fusion with the help of vector quantization (VQ) classifier. It used speech enhancement scheme to minimize the distortions and improve the speech intelligibility. It resulted in word error rate (WER) of 4% for the dysarthric subjects with 6% intelligibility. Qatab and Mustafa [26] used four types of features such as spectral, cepstral, voice quality, prosodic, and overall speech features along with SVM, ANN, linear discriminent analysis (LDA), classification and regression tree (CART), Naïve Bayes (NB), and random forest (RF) classifier for DSR. Seven feature selection algorithms have been presented for the feature selection to select the dominant features such as conditional information feature extraction (CIFE), double input symmetrical relevance (DISR), interaction capping (ICAP), conditional mutual information maximization (CMIM), conditional redundancy (Condred), joint mutual information (JMI), and relief. It provided average ranking score of 4.88 for RF and relief feature selection. Janbakhshi *et al.* [27] presented singular value decomposition (SVD) for the spectro-temporal representation of the dysarthric speech and temporal grassmann discriminant analysis (T-GDA) for the DSR. It outperformed the traditional mel frequency cepstral coefficient (MFCC)-SVM based DSR. The subspace based learning shows superior discrimination between normal and dysarthric speech. The temporal subspace gives enhanced performance compared with spectral subspace.

Recently, DL technology has been widely used in many voiced based automation systems and has proven it can provide better performance than conventional ML based methods [28], [29]. Fathima *et al.* [30] applied a multilingual time delay neural network (TDNN) system that combined acoustic modeling and language specific information to increase ASR performance. The experimental results showed that the TDNN-based ASR system achieved suitable performance, as the WER was 16.07% in this study. Yue *et al.* [31] investigated convolutional and light gated recurrent unit (LiGRU) based multi-spectra acoustic model for DSR. It used data augmentation to minimize the data scarcity problem using speed perturbation which has given 11% and 40.6% WER for normal and dysarthric speech. Yue *et al.* [32] developed multi-stream acoustic model based on convolutional neural network (CNN), LiGRU, and fully connected multi layer perceptron (MLP) and optimal fusion technique for DSR. The proposed model provided a WER of 4.6% for the pre-processed data using electromagnetic articulography (EMA). The EMA pre-processing includes Butterworth filter for measurement noise minimization and down-sampling for synchronization of MFCC features.

The data efficiency is major obstacle in the DSR. Soleymanpour *et al*. [33] proposed text to speech (TTS) synthesizer for the data augmentation based on FastSpeech model. The augmented data provided to deep neural network (DNN)-HMM with light bidirectional GRU that has given a WER improvement of 12.2% over the baseline model. Traditional data augmentation approaches majorly focuses on the temporal variations of the signal however spectral envelope remains same. Liu *et al.* [34] presented vocal tract length perturbation (VTLP), tempo perturbation and speed perturbation for the data augmentation that concentrates on temporal as well as spectral transformations of the dysarthric speech signal. The DNN and Neural architecture search (NAS) based DSR provides WER of 25.21 % and 5.4% for UASpeech and CUHK dataset respectively. Shahamiri [35] used voicegram to provide the correlation between phonemes and the dysarthric speech. The visual data augmentation model is used for the data augmentation to minimize data scarcity problem in DSR. The spatial-convolutional neural network (S-CNN) provides an accuracy of 67% on UASpeech dataset. The proposed S-CNN some time causes vanishing gradient problem and provides poor results for the moderate dysarthria. The intelligibility of the speech is hugely affected due to time domain variance of dysarthric speech and background noise. Lin *et al.* [36] suggested that the DL based voice conversion (DVC) using phonetic posteriorgram (PPG) provides stable performance compared with DVC-mel under noisy condition.

Kodrasi and Bourlard [37] suggested that spectro-temporal sparsity using the Gini index provided better performance than shimmer, jitter, fundamental frequency, harmonics to noise ratio (HNR), and MFCC for the DSR. It is observed that spectral sparsity has proven better performance than temporal sparsity. Kodrasi [38] used CNN for learning the temporal spectral characteristics obtained using temporal envelope and fine structure (TEFS). The TEFS outperformed the traditional short-time fourier transform (SIFT) based speech signal spectrogram. The TEFS-CNN provides 85.72% accuracy for DSR whereas SIFT-CNN provides 69.76% accuracy for DSR. Chandrashekar *et al.* [39] investigated the time–frequency CNN for capturing the temporal as well as spectral properties of the dysarthric speech. The spectro-temporal properties of the speech signals are obtained using SIFT, spectrograms using SFF, and constant Q-transform (CQT). The DSR performance has shown higher accuracy for the female subjects compared with the male subject. The training data deficiency resulted in class imbalance problem. The time-frequency based CNN provides better spectro-temporal variation of the dysarthric speech which has shown significant improvement in DSR accuracy over the traditional ANNs [40]. Fritsch and Doss [41] presented recurrent neural network (RNN) based binary and CNN based multi-feature classifier. It provided high correlation for synthesized speech generated using TTS. Table 1 provides the summary of various DSR techniques based on ML and DL approaches.

Table 1. Summary of ML and DL based DSR

Ref.	Speech enhancement	Data augmentation	Feature extraction	Classifier	Database	Performance metrics	Remark
[31]	Cepstral processing to separate filter and speech element	Speed perturbation	CNN-LiGRU	Softmax	TORGO	WER -40.6% (dysarthric), 11% (normal)	Combination of excitation and vocal tract component can be used for speaking stylemodelling
[32]	EMA	-	CNN-LiGRU-FCMLP	Softmax	TORGO	WER -4.6%	Over-fitting problem for high level articulatory feature fusion
[33]	-	TTS	DNN-HMM-BLiGRU	Softmax	TORGO	WER -41.6%	The severity of dysarthric speech depeds upon energy, duration and pitch of the signal.
[34]	-	VTLP, tempo perturbation and speed perturbation	Model based speaker adaptation and cross-domain generation of visual features	DNN-NAS	UASpeech and Chinese University of Hong Kong (CUHK)	- WER=25.21% (UASpeech) - WER=5.4% (CUHK)	High WER for low intelligibility speaker
[35]	-	Visual data augmentation	Voicegram	S-CNN	UASpeech	Accuracy= 67%	- Provides less temporal representation of speech - May cause vanishing gradient problem
[36]	-	-	-	CNN with a PPG	10 samples of 19 Chinese commands for 3 users	CNN–PPG-93.49%, CNN-MFFC-65.67%, ASR based system-89.59%	Class imbalance problem issue due to uneven dataset size
[37]	-	-	Spectro-temporal sparsity using the Gini index	SVM	Spanish database (PC-GITA database)	Accuracy= 83.30% (GST), 76.7% (MFCC), 60% (HNR), 57% (Shimmer), 52% (Jitter), 54.40% (Fo)	- Less recognition rate due to less number of features - Not suitable for larger dataset
[38]	-	-	TEFS	CNN	PC-GITA database	Accuracy =85.75, AUC=0.93	Less feature discrimination due to higher intra-class and lower interclass variability - Can not handle complex auditory models
[39]	-	-	SIFT, spectrograms using SFF, CQT	Time-Frequency CNN	Universal Access and TORGO	Accuracy= 98.00% (female), 95.80% (male)	- Class imbalance problem - Complexity of network - High computation time
[26]	-	-	Spectral, cepstral, voice quality, prosodic, overall speech features	LDA, CART, NB, ANN, SVM, and RF	NEMOURS database	Average ranking score for RF and relief feature selection (4.88)	- Ability to classify speech based on severity level - Feature selection is important for DSR - Not applicable for larger dataset - Less performance than DL approaches
[27]	-	-	SVD	T-GDA	PC-GITA, MoSpeeDi, UASpeech	Accuracy-82.0±3.5% (PC-GITA), 80.5±4.7% (MoSpeeDi), 96.30% (UA)	Temporal subspaces provide better representation of normal and dysarthric speech compared with spectral subspaces
[41]	-	-	Pearson’s correlation coefficient and Spearman’s correlation coefficient	RNN	UASpeech database	PCC (0.950), SCC (0.957)	Provides high correlation for synthesized speech generated using TTS

This paper presents a comprehensive survey of distinct ML-based and DL-based DSR systems. It focuses on the DSR methodology that comprises enhancement, data augmentation, feature extraction, feature selection, and classification techniques. It analyses the dataset, experimental results, and performance metrics to depict the merits, demerits, and challenges of the present DSR systems. Additionally the performance of the various ML and DL based DSR schems is evaluated on the UASpeech dataset and results are analyzed using accuracy, recall, precision, and F1-score. The rest of paper is structured as follow: section 2 depicts the generalized process of the automatic DSR and gives the succinct survey of recent ML and DL based speech emotion recognition (SER) systems, section 2 elaborates the detailed description of the method, section 3 gives detailed results and its findings, and section 4 concludes the paper and paves the way for future enhancement through future scope.

2. RESEARCH METHOD

The process of the proposed analysis of different feature extraction and classification techniques for the DSR is illustrated in the Figure 4. The proposed system used pre-emphasis filtering which uses the moving average filter for minimizing noise and normalizing the speech. It diminish the irregularities present in the speech signal.

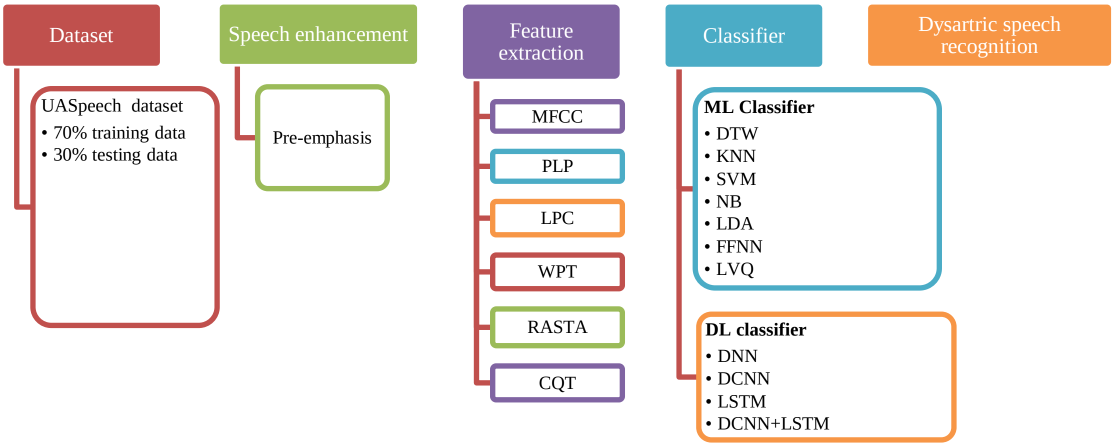


Figure 4. Proposed research method of DSR

The proposed system accepts the speech samples from the UASpeech dataset. The samples are cropped or appended to 10 second duration to make all data uniform. Out of total UASpeech data 70% and 30% samples are taken for training and testing purpose. It considers various features for the MFCC, perceptual linear prediction (PLP) coding, linear predictive coding (LPC), wavelet packet transform (WPT) 3 levels, relative spectra (RASTA), and CQT. The features are used to train various ML classifiers such as dynamic time warping (DTW), K-nearest neighbour (KNN), SVM, NB, LDA, feedforward neural network (FFNN), and linear vector quaintization (LVQ). The feature extracton stage consists of different features using traditional algorithms such as MFCC (13 MFCC features, 13 delta feature, and 13 delta-delta features), PLP features, LPC features (13 features), WPT features (3 level features), RASTA features, and CQT cepstogram features. Futher, it utilize the different ML classifiers for dysarthric voice recognition such as KNN, NB, SVM, DTW, LDA, LVQ, and FFNN. It considers the spectrogram representation of the signal for the two dimensional DL algorithms. It utilizes the deep convolutional neural network (DCNN), DNN, long short-term memory (LSTM), and DCNN-LSTM for the one analysis of the DSR for one dimensional siganla nd two dimensional speech signal. The performance of the proposed system is evaluated based on DSR accuracy.

3. RESULTS AND DISCUSSION

This section provides the experimental results of the various machine and DL based schemes for the DSR. It considers various features for the MFCC, PLP coding, LPC, WPT (3 levels), RASTA, and CQT. The features are used to train various ML classifiers such as DTW, KNN, SVM, NB, LDA, FFNN, and LVQ. It used UASpeech dataset for the experimentation as given in Table 2. It is noted that the MFCC+SVM provides highest 83.26% accuracy for the DSR compared with other algorithms such as DTW, KNN, NB, LDA, FFNN, and LVQ. It is observed that the MFCC spectrogram provides better spectral characteristics of the dysarthric speech signal that helps to capture the changes occurred on the speech due to dysarthria. The experimentations are carried out on UASpeech dataset which is cropped for 5 sec duration. Total 1,000 samples of normal and dysarthric speech are considered for the evaluation.

Table 2. Performance of ML based DSR

Feature extraction techniques	Classifiers						
	DTW	KNN	SVM	NB	LDA	FFNN	LVQ
PLP	54.79	60.63	62.00	57.38	54.79	52.78	55.00
RASTA	62.09	63.83	65.26	51.33	62.09	56.65	57.50
LPC	46.23	59.09	63.00	45.33	46.23	43.45	56.45
WPT	46.34	68.00	72.50	69.56	62.23	60.86	53.56
CQT	65.87	72.35	78.00	71.45	68.54	63.50	59.00
MFCC	62.23	75.54	83.26	73.35	67.00	64.00	61.23

Various DL based DSR schemes such as DNN, DCNN, LSTM, and DCNN-LSTM are utilized to evaluate the performance of DSR on UASpeech dataset as given in Figure 5. It used five layered 1-D DNN that gives 85% accuracy for raw speech and 87.5% accuracy for 39 MFCC coefficients that encompasses 13 MFCC coefficients, 13 delta coefficients, and 13 delta-delta coefficients that represents the spectral variation over the frames of the speech. It provides 89.45% and 90.56% accuracy for 2-D representation of the speech signal using CQT and MFCC spectrogram. It is noted that 2D representation of the speech signal provides better spectral and spatial representation of the speech signal and helps to improve the accuracy over 1-D representation of the signal. Further, it used 5 layered DCNN which encpasses convolution, batch normalization, and maximum pooling layer at every layer. It uses 32, 64, 96, 128, and 256 filters for first to fifth layer of the DCNN. The DCNN provides gives 86.60% accuracy for raw speech and 88.80% accuracy for 39 MFCC coefficients. It provides 90.10% and 91% accuracy for CQT and MFCC spectrogram. Afterward, LSTM with five layers is employed for representing the temporal characterstics of the dysarthric signal which has given 85%, 86.20%, 87%, and 88.50% accuracy for the raw speech+LSTM, MFCC coefficients+LSTM, CQT spectrogram+LSTM, and MFCC spectrogram+LSTM respectively. DCNN helps to achieve best spectral representation however lacks in time domain representation of the signal. To improve the time domain characteristics LSTM is collaborated with the DCNN which combines the frequency domain and time domain characteristics of the speech sigal for DSR. The DCNN-LSTM provides gives 88.20% accuracy for raw speech and 89.20% accuracy for 39 MFCC coefficients that encompasses 13 MFCC coefficients, 13 delta coefficients, and 13 delta-delta coefficients that represents the spectral variation over the frames of the speech. It provides 91.5% and 93% accuracy for 2-D representation of the speech signal using CQT and MFCC spectrogram.

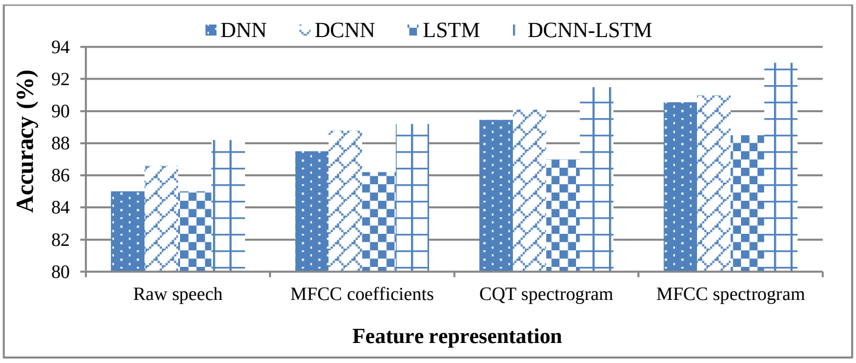


Figure 5. Performance of DL based DSR

4. CONCLUSION

Thus, this article presents the DSR based on various ML and DL approaches that covers the methodology, database, evaluation metrics, advantages, disadvantages, and finding from the study. It is observed that the DL techniques outperformed the traditional ML techniques because of its superior feature representation. The DL approaches are less dependent on the hand crafted features unlike traditional ML based approaches. The experimental results shows that the DL based DSR schems outperforms the ML based DSR schemes and provides better feature representation compared with traditional handcrafted features. The performance of DL framework is better for 2-D representation of the speech signal compared with 1-D signal because of higher representation capability in spectral and spatial domin. Also, combination DCNN and LSTM provides superiority over DNN, DCBB, and LSTM which has better feature representation capability in spectral and temporal domain. Database generation is challenging task because of unavailability of theproper resources and proper ground truth. The DSR is very challenging due to variability in the speech intelligibility because of various attributes such as language, age, gender, region, and noise.

REFERENCES

[1] F. Abakarim and A. Abenaou, "Comparative study to realize an automatic speaker recognition system," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 1, pp. 376–382, 2022, doi: 10.11591/ijece.v12i1.pp376-382.

[2] K. B. Bhangale and K. Mohanaprasad, "A review on speech processing using machine learning paradigm," *International Journal of Speech Technology*, vol. 24, no. 2, pp. 367–388, 2021, doi: 10.1007/s10772-021-09808-0.

[3] Z. J. M. Ameen and A. A. Kadhim, "Machine learning for Arabic phonemes recognition using electrolarynx speech," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 1, pp. 400–412, 2023, doi: 10.11591/ijece.v13i1.pp400-412.

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

[5] K. B. Bhangale and M. Kothandaraman, "Survey of Deep Learning Paradigms for Speech Processing," *Wireless Personal Communications*, vol. 125, no. 2, pp. 1913–1949, 2022, doi: 10.1007/s11277-022-09640-y.

[6] K. L. Ong, C. P. Lee, H. S. Lim, and K. M. Lim, "Speech emotion recognition with light gradient boosting decision trees machine," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 4, pp. 4020–4028, 2023, doi: 10.11591/ijece.v13i4.pp4020-4028.

[7] K. Bhangale and M. Kothandaraman, "Speech emotion recognition based on multiple acoustic features and deep convolutional neural network," *Electronics*, vol. 12, no. 4, pp. 1–17, 2023, doi: 10.3390/electronics12040839.

[8] M. Kim, B. Cao, K. An, and J. Wang, "Dysarthric speech recognition using convolutional LSTM neural network," in *Interspeech 2018*, 2018, pp. 2948–2952, doi: 10.21437/Interspeech.2018-2250.

[9] M. Vasilakis and Y. Stylianou, "Voice pathology detection based eon short-term jitter estimations in running speech," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 3, pp. 153–170, 2009, doi: 10.1159/000219951.

[10] K. B. Bhangale, P. Desai, S. Banne, and U. Rajput, "Neural Style Transfer: Reliving art through Artificial Intelligence," in *2022 3rd International Conference for Emerging Technology (INCET)*, 2022, pp. 1–6, doi: 10.1109/INCET54531.2022.9825254.

[11] R. Takashima, T. Takiguchi, and Y. Ariki, "Two-step acoustic model adaptation for dysarthric speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustic s, Speech and Signal Processing (ICASSP)*, 2020, pp. 6104–6108, doi: 10.1109/ICASSP40776.2020.9053725.

[12] S. Skodda, W. Visser, and U. Schlegel, "Short- and long-term dopaminergic effects on dysarthria in early Parkinson's disease," *Journal of Neural Transmission*, vol. 117, no. 2, pp. 197–205, 2010, doi: 10.1007/s00702-009-0351-5.

[13] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, 2009, doi: 10.1109/TBME.2008.2005954.

[14] J. Shao, J. K. MacCallum, Y. Zhang, A. Sprecher, and J. J. Jiang, "Acoustic analysis of the tremulous voice: assessing the utility of the correlation dimension and perturbation parameters," *Journal of Communication Disorders*, vol. 43, no. 1, pp. 35–44, 2010, doi: 10.1016/j.jcomdis.2009.09.001.

[15] D. Dimitriadis, A. Potamianos, and P. Maragos, "A comparison of the squared energy and teager-kaiser operators for short-term energy estimation in additive noise," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2569–2581, 2009, doi: 10.1109/TSP.2009.2019299.

[16] K. M. Allison, L. Annear, M. Policicchio, and K. C. Hustad, "Range and precision of formant movement in pediatric dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 7, pp. 1864–1876, 2017, doi: 10.1044/2017_JSLHR-S-15-0438.

[17] K. L. Lansford and J. M. Liss, "Vowel acoustics in dysarthria: Speech disorder diagnosis and classification," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 1, pp. 57–67, 2014, doi: 10.1044/1092-4388(2013/12-0262).

[18] N. P. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67745–67755, 2020, doi: 10.1109/ACCESS.2020.2986171.

[19] K. Gurugubelli and A. K. Vuppala, "Analytic phase features for dysarthric speech detection and intelligibility assessment," *Speech Communication*, vol. 121, pp. 1–15, 2020, doi: 10.1016/j.specom.2020.04.006.

[20] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5070–5074, doi: 10.1109/ICASSP.2017.7953122.

[21] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, "HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria," in *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, 2006, vol. 3, pp. 1060–1063, doi: 10.1109/ICASSP.2006.1660840.

[22] K. B. Bhangale, "Human Body Detection in Static Images Using HOG & Piecewise Linear SVM," *International Journal of Innovative Research and Development*, vol. 3, no. 6, pp. 179–184, 2014.

[23] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4605–4608, doi: 10.1109/ICASSP.2009.4960656.

[24] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 947–960, 2011, doi: 10.1109/TASL.2010.2072499.

[25] A. Revathi, R. Nagakrishnan, and N. Sasikaladevi, "Comparative analysis of Dysarthric speech recognition: multiple features and robust templates," *Multimedia Tools and Applications*, vol. 81, no. 22, pp. 31245–31259, 2022, doi: 10.1007/s11042-022-12937-6.

[26] B. A. A.-Qatab and M. B. Mustafa, "Classification of Dysarthric Speech According to the Severity of Impairment: An Analysis of Acoustic Features," *IEEE Access*, vol. 9, pp. 18183–18194, 2021, doi: 10.1109/ACCESS.2021.3053335.

[27] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Subspace-Based Learning for Automatic Dysarthric Speech Detection," *IEEE Signal Processing Letters*, vol. 28, pp. 96–100, 2021, doi: 10.1109/LSP.2020.3044503.

[28] K. B. Bhangale, P. Titare, R. Pawar, and S. Bhavsar, "Synthetic Speech Spoofing Detection Using MFCC and Radial Basis Function SVM," *IOSR Journal of Engineering (IOSRJEN)*, vol. 8, no. 6, pp. 55–62, 2018.

[29] K. Bhangale and K. Mohanaprasad, "Speech Emotion Recognition Using Mel Frequency Log Spectrogram and Deep Convolutional

Neural Network," *Lecture Notes in Electrical Engineering*, vol. 792, pp. 241–250, 2022, doi: 10.1007/978-981-16-4625-6_24.

[30] N. Fathima, T. Patel, M. C, and A. Iyengar, "TDNN-based multilingual speech recognition system for low resource Indian languages," in *Interspeech 2018*, 2018, pp. 3197–3201, doi: 10.21437/Interspeech.2018-2117.

[31] Z. Yue, E. Loweimi, and Z. Cvetkovic, "Raw Source and Filter Modelling for Dysarthric Speech Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7377–7381, doi: 10.1109/ICASSP43922.2022.9746553.

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

[33] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "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)*, 2022, pp. 7382–7386, doi: 10.1109/ICASSP43922.2022.9746585.

[34] S. Liu *et al.*, "Recent Progress in the CUHK Dysarthric Speech Recognition System," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 2267–2281, 2021, doi: 10.1109/TASLP.2021.3091805.

[35] S. R. Shahamiri, "Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 852–861, 2021, doi: 10.1109/TNSRE.2021.3076778.

[36] Y. Y. Lin *et al.*, "A speech command control-based recognition system for dysarthric patients based on deep learning technology," *Applied Sciences*, vol. 11, no. 6, pp. 1–16, 2021, doi: 10.3390/app11062477.

[37] I. Kodrasi and H. Bourlard, "Spectro-Temporal Sparsity Characterization for Dysarthric Speech Detection," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 28, pp. 1210–1222, 2020, doi: 10.1109/TASLP.2020.2985066.

[38] I. Kodrasi, "Temporal Envelope and Fine Structure Cues for Dysarthric Speech Detection Using CNNs," *IEEE Signal Processing Letters*, vol. 28, pp. 1853–1857, 2021, doi: 10.1109/LSP.2021.3108509.

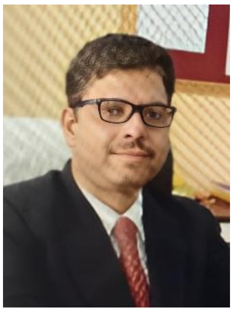
[39] H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, "Investigation of Different Time-Frequency Representations for Intelligibility Assessment of Dysarthric Speech," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 12, pp. 2880–2889, 2020, doi: 10.1109/TNSRE.2020.3035392.

[40] H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, "Spectro-Temporal Representation of Speech for Intelligibility Assessment of Dysarthria," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 390–399, 2020, doi: 10.1109/JSTSP.2019.2949912.

[41] J. Fritsch and M. Magimai-Doss, "Utterance Verification-Based Dysarthric Speech Intelligibility Assessment Using Phonetic Posterior Features," *IEEE Signal Processing Letters*, vol. 28, pp. 224–228, 2021, doi: 10.1109/LSP.2021.3050362.

BIOGRAPHIES OF AUTHORS

Shailaja Yadav received M.E. in digital system from Savitribai Phule Pune University, Pune, Maharashtra, India in 2015. She is currently pursuing the Ph.D. degree with G.H. Raisoni College of Engineering and Management, Wagholi-Pune, Maharashtra, India. She is currently working as an assistant professor in the Department of Electronics and Telecommunication at D.Y. Patil College of Engineering, Akurdi-Pune, Maharashtra, India since 2018. She has 16 years of teaching experience in various reputed colleges in India. Her main research interests focus on speech processing, machine learning, artificial intelligence, and deep learning. She is a life member of the Institution of Electronics and Telecommunication Engineers (IETE), India. She can be contacted at email: shailaja.yadav123@gmail.com.



Dinkar Manik Yadav received his bachelor degree from Dr. B.A.M University, Aurangabad, India in 1994. The Ph.D. degree from Bharati Vidyapeeth, Pune, Maharashtra, India in 2009. Under his guidance 11 research scholars were awarded Ph.D. degree. His areas of research interests are image processing, signal processing and medical imaging. He is currently working as principal at S.N.D. College of Engineering and Research Center, Yeola, Nasik-Maharashtra, India. He can be contacted at email: dineshyadav800@gmail.com.

Dr. Kamalakar Ravindra Desai received his first degree from Shivaji University, Electronics, Kolhapur in June 1998. He also has a master's degree from Shivaji University, Electronics and Telecommunication, Kolhapur in June 2006. Received the Ph.D. from Shivaji University for "Error computation in GPS signals" in 2016. He is currently a Professor in Bharati Vidyapeeth's College of Engineering Kolhapur. His main interest focuses on satellite and telecomunication, networking, embedded system. He can be contacted at email: krdesai2013@gmail.com.