



# Deep Learning-Based Acoustic Feature Representations for Dysarthric Speech Recognition

M. Latha<sup>1,2</sup> · M. Shivakumar<sup>1,2</sup> · G. Manjula<sup>1,2</sup> · M. Hemakumar<sup>2,3</sup> · M. Keerthi Kumar<sup>1,2</sup>

Received: 28 September 2022 / Accepted: 17 December 2022 / Published online: 20 March 2023  
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2023

## Abstract

Dysarthria is a motor speech disorder and the most common neurodegenerative disease characterized by low volume in precise articulation, poor coordination of respiratory and pulmonary subsystems, and irregular pauses. The key challenge with Dysarthria is reduced intelligibility in speech that forces an individual to undergo numerous socio-professional adversities due to the lack of self-expression, constrained presentation, limited career opportunities, and accessibility towards the state-of-art advanced technologies, especially voice-controlled activities or human machine interface. Though, numerous efforts have been made to exploit acoustic features from dysarthric speeches to improve intelligibility; however, most of the existing approaches are limited due to change in acoustic features across languages. The approaches suggesting machine learning-driven automatic speech recognition too have failed in retaining most suitable set of acoustic features and learning environment to yield precise word recognition. Considering it as motivation, in this paper, a Hybrid Acoustic Feature-Driven Deep Learning model for Dysarthric Speech Recognition and perceptible speech generation was developed. It exploited Short-Term Fourier Transform driven Mel-Frequency Cepstral Coefficient (MFCC), Pitch, LPC, Gammatone Frequency Cepstral Coefficient (GFCC), Short-Time Energy (STE), Zero-Crossing Rate (ZCR), and short-time ZRC (ST-ZRC) as composite acoustic features to perform LSTM-CNN-based learning for dysarthric speech recognition (DSR) and perceptible speech generation. The statistical performance revealed that the proposed model achieves the highest known accuracy for DSR and ASR systems.

**Keywords** Dysarthric speech · Acoustic features · Deep learning · Human–machine interface · Intelligible communication

## Introduction

Speech, being the primary mode of inter-personal communication, represents an acoustic signal which is employed to exchange thoughts and feelings. The articulation of any speech signal primarily depends on physiology and

neurological factors that often vary from one individual to another giving rise to the normal and abnormal speech generations. When an individual is subsisting a couple of speech sounds and is unable to pronounce the intended or specific sound, he or she is stated to be suffering from speech disorder. Typically, the speech disorders caused due to the impaired motor speech planning, programming, control, or neuromuscular execution are stated to be the motor speech disorders, which are broadly classified into three types; stuttering, apraxia, and dysarthria [2].

Dysarthria is stated to be a complex problem caused in the speech generation systems that often takes place due to certain injury or disease to the brain, cranial nerves or nervous system malfunctions. It is a motor speech disorder and the most common neurodegenerative disease characterized by low volume in precise articulation, poor coordination of respiratory and pulmonary subsystems, irregular pauses, etc. Undeniably, the key challenge in enhancing communication with dysarthric speech is low

---

This article is part of the topical collection “Advances in Computational Intelligence for Artificial Intelligence, Machine Learning, Internet of Things and Data Analytics” guest edited by S. Meenakshi Sundaram, Young Lee, and Gururaj K S.

---

✉ M. Latha  
latha@gsss.edu.in

<sup>1</sup> Department of Electronics and Communication Engineering, GSSS Institute of Engineering and Technology for Women, Mysuru, Karnataka, India

<sup>2</sup> Affiliated to VTU, Belagavi, Karnataka, India

<sup>3</sup> Govt. College for Women (Autonomous), Mandya, Karnataka, India

intelligibility with disrupted acoustic characteristics and hence lower perceptibility. This as a result impacts communication with dysarthric patients [1, 2]. The articulatory errors during the utterance of speech segments hinder the speech articulation and hence limit communication efficacy. Such dysarthric limitations severely vary with change in symptoms such as minimized volumed vocal tract, atypical speech prosody, inaccurate articulation, varying or non-uniform speech rate, reduced tongue flexibility, strained voice quality, etc. Based on the extent of neurological malfunction and the type of dysarthria, speech-intelligibility too varies significantly [2].

Authors have identified numerous of acoustic signals and allied coefficients to characterize dysarthric speech(es) and its severity [2]. Most of these approaches merely address either acoustic feature extraction or acoustic feature-driven dysarthric speech classification and allied automatic speech recognition [1–3]. The existing methods primarily target to serve varied commercial applications. In fact, automatic speech recognition system serves as an interface to generate more significant, intelligible, perceptual, and pervasive. Considering above facts as research gap and motivation, in this paper, a first of its kind speech processing environment is designed that converts the original dysarthric speech(es) to the perceptible sounds for optimal peer communication. Unlike classical approaches, this research exploits multiple acoustic features including MFCC, Pitch, LPC, Gammatone Frequency Cepstral Coefficient (GFCC), Short-Time Energy (STE), Zero-Crossing Rate (ZCR), and short-time ZRC (ST-ZRC) to perform LSTM-CNN-based learning for acoustic classification. Unlike classical wavelet-based MFCC features, in this paper, the focus was made to exploit optimal spatio-temporal features so as to alleviate the noise components and retain optimal feature coefficients for training. To achieve it, short-term Fourier transform with Hann windowing concept was taken into consideration. Eventually, obtaining the set of features from both normal speeches as well as dysarthric speeches in Kannada bi-syllabic words, feature fusion was performed. Thus, the fused feature set was projected as input to the Long- and Short-Term Convolutional Neural Network (LSTM-CNN) model to perform learning. Unlike major existing machine learning or deep learning-based approaches which merely classify the input speeches as normal or dysarthric or identifies the word(s), our proposed model reconstructs or transforms the input dysarthric speech into perceptible sound signal. This as a result not only achieves words recognition but also generates perceptible or intelligible acoustics (sound speeches) for the dysarthric input. In this manner, it can be vital towards inter-personal communication systems for dysarthric person. Moreover, it can also be significant for numerous commercial applications, such as HMI, speech operated smart acts, etc. The simulation results revealed that the proposed model

exhibits more efficient and reliable performance than any known DSR system.

The other sections of this manuscript are divided as follows. Section two discusses some of the key related works pertaining to ASR/DSR and acoustic modelling for dysarthric speech analysis, which is followed by the research questions in Section three. The overall research problem formulation is given in section three, while the proposed model implementation detail is given in section four. Simulation results and allied inferences are given in section five, while the research conclusion is detailed in section six. References used in this study are given at the end of the manuscript.

## Related Work

The survey is conducted on the existing methods on the dysarthric speech analysis and also addresses on the recent deep learning algorithms so as to obtain the desired research outcome are presented below.

Most of the classical dysarthric speech recognition methods apply structured methods, such as HMM–GMM [4]. These structure-based approaches employ HMMs to design the sequential architecture of the speech signal, while GMM is employed to model the spectral representation of the acoustic waveform. However, such approaches require significantly large volume of training data that turn out to be complex and very difficult for dysarthric speeches. Therefore, the conventional HMM–GMM-based methods cannot be suggested for dysarthric speech classification [5]. Though, as an alternative deep neural network (DNNs) has been found robust and effective towards different pathological speech processing tasks [6]. In this research, authors [6] applied the convolutional long short-term memory recurrent neural networks to perform dysarthric speech detection in a speaker-independent manner. The use of CNN helped in extracting significant local features for learning; however, limited features could not achieve the higher accuracy which is must in real-world applications. To exploit the efficacy of both DNN and HMM methods for automatic speech recognition, authors [7] designed DNN-HMM model; however, its computational complexity and cost cannot be ignored. In [8], authors designed a time-delay NN-assisted denoising autoencoder to perform dysarthric speech learning. Here, authors applied denoising autoencoder merely to perform feature augmentation, which was followed by learning using DNN-HMM for speaker recognition. Authors in [9] stated that pitch can be a viable acoustic cue to perform dysarthric speech recognition. As classifier authors applied gated neural network and Bayesian gated neural network models. Authors [10] applied a multilayer perceptron (MLP) ANN model for speech analysis of the Parkinson patients. Training over a large dataset obtained for both the normal speech

as well as Parkinson's disease patients authors performed two-class classification. However, the acoustic features of both Parkinson's disease and dysarthria differs significantly. Authors in [11] applied gated ANN to learn over a composite feature containing both acoustic features as well as visual features. Authors applied prosody features based on pitch to perform two-class classification. In [12], authors assessed the correlation between dysarthric speech and noisy speech in the form of intelligibility. Authors assessed whether there can be any association in between the ability to deal with noisy speech uttered in degraded background and the ability to understand dysarthric speech. Interestingly, authors [13] concluded that the listeners having the ability to understand certain speech in noisy environment can also understand dysarthric speeches. Though, it cannot be generalized neither for human being not for an HMI. Authors in [2] performed deep belief network (DBN)-based feature extraction followed by multiple perceptron ANN-based two-class classification for dysarthric speaker identification. Authors claimed that DBN-based features are more effective than classical MFCC features to perform dysarthric speech classification. Despite the numerous works, generalizing a set of optimal features and learning environment has remained a challenge for researchers. On the other hand, no significant effort is made towards dysarthric speech recognition in Kannada language. Noticeably, as per our knowledge, there is no model available that could convert the raw dysarthric speeches into perceptible speech format.

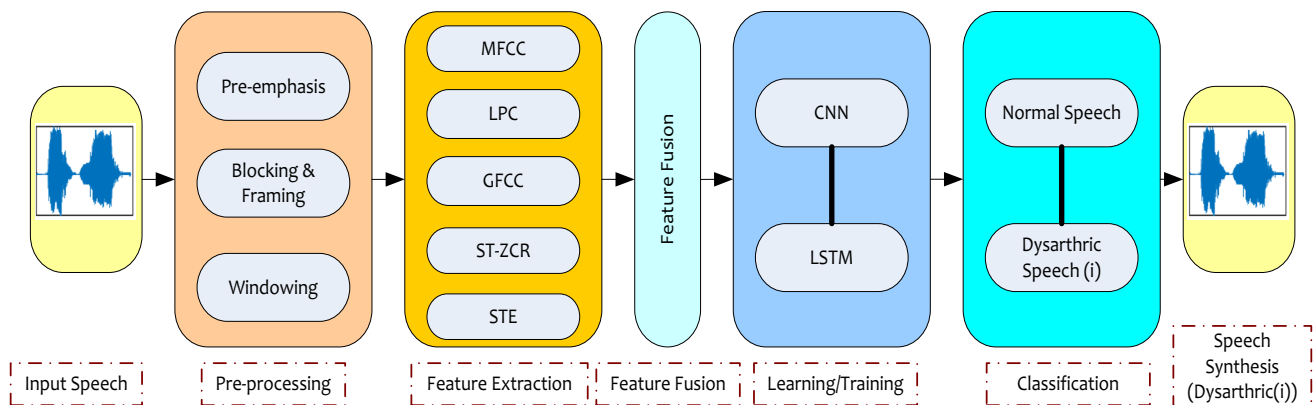
This current work primarily focused on achieving an optimal set of acoustic features and learning environment to perform dysarthric speech recognition and reconstruction. To achieve it, the proposed model intends to exploits multiple acoustic features, including MFCC, LPC, GFCC, ZCR, pitch, etc. from the Kannada bi-syllabic words obtained from both normal as well as dysarthric persons. Subsequently, it fuses the different acoustic features to derive a composite feature vector which is trained using LSTM-CNN model that

recognizes the dysarthric words and reconstructs a perceptible or intelligible sound speech. In this manner, the overall proposed model transforms the raw dysarthric speech into the equivalent perceptible speech output to make communication effective.

## Proposed Model

This current research predominantly emphasizes on extracting an optimal set of acoustic cues or features to be trained with a lightweight deep learning environment to perform raw dysarthric speech into perceptible (or high-intelligible) normal speech. Unlike major existing studies where authors have merely employed standalone feature-based dysarthric speaker or speech recognition, this research focused not only speech recognition but reconstruction of the same to ensure optimal communication with peers, including commercial applications or HMIs. To achieve it, the proposed model amalgamated a set of distinct acoustic features including MFCC, LPC, GFCC, Pitch, STE, ZCR, and ST-ZCR to derive an optimal (acoustic) feature vector for training. Noticeably, unlike classical MFCC features, to retain maximum possible spatial-temporal features, also applied Short-Term Fourier Transform (STFT) with Hann windowing to extract MFCC. Finally, the amalgamated or fused feature vector was learnt over LSTM-CNN environment to perform two-class classification. Subsequently, based on the classified output, the proposed model applies speech synthesis toolbox to reconstruct the dysarthric speech into perceptible speech output. The overall proposed model is depicted in Fig. 1.

In this present work, dictionary-based concept that helps understanding or matching the raw acoustic features, and thus generates the perceptible (probable) speech as output to complete communication. The overall proposed model is developed for the dysarthric speech reconstruction in



**Fig. 1** Proposed dysarthric speech recognition system

Kannada language, where Kannada Bi-syllabic words are used to perform model training. To characterize efficacy, performance characterization is done in terms of accuracy, precision, recall, F-Measure, true-positive rate (TPR), false-positive rate (FPR), recognition error, etc. The overall proposed model is accomplished in different phases which includes Data Acquisition, Acoustic Feature Extraction, Feature Fusion, LSTM-CNN-based two-class classification, and Dictionary-assisted feature-sensitive normal speech synthesis.

### Data Acquisition

The voice samples were obtained from both male as well as female respondents, from the normal as well as dysarthric speaker category. The speech subjects were in the age range of 16–25 years. The speech samples from both normal as well as dysarthric speakers were recorded and saved in \*.wav format for further feature extraction. The pre-recorded samples were obtained for the following (Table 1) Kannada Bi-syllabic words (for both normal as well as dysarthric speakers). A total of 200 samples each with the considered Bi-syllabic (Table 1) were obtained that resulted a set of 1600 samples for further feature extraction.

Once, obtaining the speech samples from the subjects, each sample was processed for multiple feature extraction.

### Acoustic Feature Extraction

Once collecting the speech samples from the participants, it is further processed individual speech sample for feature extraction. Unlike major existing approaches where standalone feature has been applied for learning and speech recognition, later further performed multiple feature extraction for each speech sample. More specifically, this work extracted MFCC, LPC, GFCC, Pitch, STE, ZCR, and ST-ZCR features to perform further learning

and classification. However, the proposed model intended to exploit maximum possible spatio-temporal features to make feature learning better and robust.

In this paper, this work is applied with STFT-based MFCC extraction, where unlike the conventional FFT-based methods, STFT is applied to convert spatial domain into frequency-domain value. The proposed MFCC extraction method employs the key processing elements of framing or blocking, windowing, and STFT-driven MFCC coefficient estimation.

In framing or blocking phase, the continuous input speech samples were blocked into smaller frames containing  $N$  samples. Here, these frames are generated in such manner that the subsequent frames separated by means of  $M$  samples ( $M < N$ ) enable overlapping of the adjacent frame by  $N - M$  samples. This is done so as to retain the sufficient samples with optimal information. In case, the size of frame is smaller in comparison to overlapping signals, the samples contained in the frame would not have sufficient information to make further decision. Thus, following above approaches, the complete input speech samples (for both normal as well as dysarthric subjects) were converted into multiple small frames. Once obtaining the frames, we executed a function called windowing that helped minimizing the disruptions at the start as well as at the end of the frame. To achieve it, Hamming window concept is used, such that the window function (i.e., Hamm windowing) is multiplied with each frame. For a window function being defined as  $W_n(m)$ ,  $0 \leq m \leq N_m - 1$ , where  $N_m$  be the sample quality within each retrieved frame (also called frame quality), the resulting output after windowing the input speech is (1)

$$Y(m) = X(m)W_n(m), 0 \leq m \leq N_m - 1. \quad (1)$$

In (1),  $Y(m)$  states the resulting windowed signal. We applied Hamming window, defined as (2)

$$W_n(m) = 0.54 - 0.46\cos\left(\frac{2\pi m}{(N_m - 1)}\right), 0 \leq m \leq N_m - 1. \quad (2)$$

It has been found that the human perception towards sound frequency does not follow a linear scale, and therefore, for each tone containing the actual frequency of  $f$  (in Hz) represents a subjective pitch which is estimated on a scale, often called “Mel-scale” [14]. Mathematically, it is defined as (3)

$$f_{mel} = 2595\log_{10}\left(1 + \frac{f}{700}\right). \quad (3)$$

In (3),  $f_{mel}$  represents the subjective pitch in Mels in conjunction with a frequency in Hz. This as a result puts

**Table 1** Kannada bi-syllabic words

SN.	Normal Subject	Dysarthric Subject
1	ಪದೆ pada-(/p^d^/)	ಪದೆ pada-(/p^d^/)
2	ತಪೆ tapa-(/t^p^/)	ತಪೆ tapa-(/t^p^/)
3	ಪಟೆpaTa-(/p^t^/)	ಪಟೆpaTa-(/p^t^/)
4	ದಡೆdaDa-(/d^d^/)	ದಡೆdaDa-(/d^d^/)
5	ದಷೆpa-(/d^p^/)	ದಷೆpa-(/d^p^/)
6	ತಡೆaDa-(/t^d^/)	ತಡೆaDa-(/t^d^/)
7	ತಲೆಟೆaTa-(/t^t^/)	ತಲೆಟೆaTa-(/t^t^/)
8	ಬಡೆbaDa-(/b^d^/)	ಬಡೆbaDa-(/b^d^/)

foundation for MFCC definition, signifying a baseline acoustic feature (set) for speech recognition [13, 14].

### Zero-Crossing Rate

The zero crossing rate (ZCR) represents a viable acoustic feature having the ability to discriminate a fricative sound from the others. It represents the acoustic features reflecting voiced and unvoiced cues or characteristics that can easily discriminate the fricative and affricative sounds. To detect the significant acoustic features from the subject's sample, we applied ZCR along with the Short-Time Energy (STE). Typically, higher ZCR value with low STE signifies the unvoiced speech. Thus, to estimate the values for ZCR and STE, respectively

$$Z_n(x) = \sum_m^M \text{sign}(x[j] - \text{sign}(x[j - 1])).w[n - m]. \quad (4)$$

In (4), the function  $\text{sign}$  is considered in between the two subsequent samples  $x[j]$  and  $x[j - 1]$ . It applied scaled rectangular window with the scale value of  $\frac{1}{2 \cdot |x|}$

$$E_n(x) = \sum_m x^2[m].h[n - m]. \quad (5)$$

In Eqs. (4–5),  $x$  states the input signal (i.e., subjects' speech input), while  $h$  states the Hamming window as defined in (2). Moreover, the parameter  $N$  states the length of the window, which is selected same as that of the analysis frame length.

### Pitch

In this study, pitch is considered as one of the acoustic features to perform dysarthric speech recognition. In general, pitch is defined as the measure of the sound frequency in Hz. In other words, higher the frequency results higher pitch. For instance, male speech subjects would have low pitch in comparison to the females who have the higher pitch. In our proposed work, we estimated the pitch values on the recorded speech samples for both normal as well as dysarthric subjects. Here, we applied MATLAB function `pitch` to estimate the measures for the speech samples.

### Feature Fusion

In majority of the existing systems, authors have applied standalone feature to perform classification towards dysarthric speech recognition or speaker recognition. However, most of these approaches have shown limited performance, which could be hypothesized to be mainly due to inefficient features or somewhat the learning environment. Considering this fact, in this paper, unlike conventional standalone

feature-based DSR, this current work proposed with hybrid feature model where multiple acoustic features, including MFCC, LPC, GFCC, Pitch, ZCR, and STE, were concatenated together to generate a composite feature vector (6)

$$\begin{aligned} Feat_{COMP} \\ = [Con(Feat_{MFCC}, Feat_{GFCC}, Feat_{LPC}, Feat_{ZCR}, Feat_{STE}, Feat_{Pitch})]. \end{aligned} \quad (6)$$

Thus, horizontally concatenating the different features, it is able to generate a composite feature vector which was later used for classification. Considering feature non-linearity and diversity in this paper, a hybrid DNN model named CNN–LSTM was applied to perform DSR.

### Hybrid CNN–LSTM Model-Based Classification

In the past, numerous machine learning models along with deep learning models have been applied for speaker recognition as well as DSR; however, their ability to learn over composite feature vector as discussed above can be limited. This can be because of the high non-linearity and diversity amongst feature sets. Considering this fact, it is able to design a lightweight but robust hybrid deep learning environment using CNN and LSTM models. Noticeably, CNN being a multilayer perceptron neuro-computing environment helps learning over the complex (non-linear) feature, while LSTM retains optimal performance than any complex recurrent CNN (RCNN) models. Moreover, the CNN-driven local feature helps the combined deep learning environment to achieve significant feature to achieve higher accuracy. In our proposed model, the key motive behind CNN was to obtain the significant local features from the high-layer inputs and transfer them to the lower layer for more complex but significant features. Thus, the proposed hybrid CNN–LSTM model intends to achieve better performance as well as computational efficacy.

### Results and Discussion

This research primarily focused on developing a robust dysarthric speech recognition and intelligible or perceptible normal speech generation system to support effective peer communication. In sync with this objective, the focus was made on exploiting different acoustic features from the normal as well as dysarthric person's speeches. Recalling the fact that the classical acoustic cues, such as MFCC, pitch, etc., which does not address noise or background interference (say, cluster present), this research exploited Short-Term Fourier Transform (STFT) with Hamming windowing method for key feature extraction. In this study, the use of STFT enabled retention of the optimal spatio-temporal



acoustic (significant) cues, for better learning. The model was designed with 1024 size. Moreover, this study hypothesized that the amalgamation of the different acoustic features can yield better feature learning and hence can perform precise speech recognition. In this relation, multiple acoustic features, including MFCC, GFCC, Pitch, LPC, ZCR, and STE, were obtained from each sample (encompassing both normal as well as dysarthric speech samples).

Current work not only focused on the assessments of pathological subjects but also on the transformation of the unintelligible dysarthric speech into intelligible speech (Perceptible speech output). To achieve it, the supporting features have been exploited including Chroma Short-Term Fourier Transform (STFT), Spectral Centroid (SC), Roll off Factor (RoF), Spectral Bandwidth (SB), Root-Mean-Square Error (RMSE), Zero-Crossing Rate (ZCR), MFCC Features, and GTCC Features. These extracted features were processed for fusion followed by training using deep learning model, where a modified LSTM-CNN deep learning model with ADAM optimization algorithm with 4480 epochs was designed. Unlike Stochastic Gradient Descent with Momentum (SGDM) Learning, ADAM (adaptive learning rate optimization algorithm) has been found performing better in terms of accuracy and error rate.

In this study, it is mainly focused on performing dysarthric speech recognition for the Kannada bi-syllabic words, and therefore characterized the performance in terms of two-class classification efficiency. To achieve it, the following confusion matrix parameters, such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are computed. Current research work is also validated for the 80% training data and the 20% testing data. The proposed robust model first classifies the input as normal speech or dysarthric speech, and once identifying it to be a dysarthric speech, it transforms the dysarthric speech into intelligible or perceptible normal speech (audio) output. The performance analysis has been done in terms of accuracy, sensitivity, and specificity. A test simulation with random input exhibited that the proposed model exhibits accuracy of 93.3%, while retaining error or false-positive ratio of 0.06.

### Intra-model Assessment

In this intra-model assessment, the performance is stated by the proposed DSR method in terms of true-positive rate (TPR), false-positive rate (FPR), DSR recognition error rate (%), area under curve (AUC), accuracy (%), precision, recall (also called sensitivity), and F-measure (say, specificity). However, our prime objective was to analyze efficacy of the proposed DSR system with the different test conditions. Table 2 presents the TPR, FPR, error rate (%), and AUC performance by the proposed model. To assess efficacy under different test conditions. It is able to

**Table 2** DSR statistical performance analysis-1

Test sample	TPR	FPR	Error rate (%)	AUC
1	0.943	0.019	4.4	0.992
2	0.956	0.016	3.2	0.999
3	0.997	0.008	4.3	1
4	1	0.006	2.6	0.999
5	1	0	5.1	1
Avg.	0.98	0.0098	3.92	0.998

**Table 3** DSR statistical performance analysis-2

Test sample	Accuracy (%)	Precision	Recall	F-measure
1	95.6	0.94	0.98	0.96
2	96.8	0.95	0.98	0.94
3	95.7	0.96	0.98	0.97
4	97.4	0.94	0.95	0.94
5	94.9	0.96	0.99	0.95
Avg.	96.0	0.95	0.976	0.95

simulate proposed model with ten random acoustic signals or samples belong to the normal as well as dysarthric speech subjects. However, for ease of representation in this work, presented results with reference to the five random test conditions.

### Inter-model Assessment

In this paper, the performance of the proposed model is examined with other state-of-art dysarthric speech or speaker recognition systems. Noticeably, since the proposed model addresses dysarthric speech classification, only those methods performing either dysarthric speech recognition or speaker recognition are compared in this section (Table 3).

The above results presented the DSR performance by the proposed model; however, to assess its relative performance efficacy. Despite the training accuracy of 100%, authors could achieve the testing accuracy of merely 40.6%, which is significantly lower than our propose model. This result clearly indicates that the use of multiple acoustic features for training can yield higher accuracy towards DSR purposes. The above stated performance clearly indicates that the standalone features cannot yield the higher efficacy. On the contrary, the amalgamation of the different features including MFCC and LPC can result better performance. The highest classification accuracy with conventional MFCC, delta-MFCC,

delta–delta–MFCC, and distributed DTC-based MFCC were 90.36%, 90.68%, 91.35%, and 96.72%, respectively.

## Conclusion

In this paper, a first of its kind solution was designed that focuses not only on improving the (dysarthric) word recognition but also intends to reconstruct perceptible speech output from the dysarthric raw speech. The classical standalone feature-driven ASR or DSR systems, the proposed model exploited multiple acoustic features including Short-Term Fourier Transform driven Mel-Frequency Cepstral Coefficient (MFCC), Pitch, LPC, Gammatone Frequency Cepstral Coefficient (GFCC), Short-Time Energy (STE), Zero-Crossing Rate (ZCR), short-time ZRC (ST-ZRC) to generate a combined (acoustic) feature vector for learning. Thus, the estimated fused feature vector was learnt using LSTM-CNN model that at first identified the dysarthric words and transformed it into equivalent perceptible sound signal or speeches. The statistical performance characterization revealed that the proposed model delivers accuracy of (96%), Precision (0.95), Recall (0.976), and F-Measure of (0.95), which is the highest amongst the major known approaches so far. It exhibits robustness of the proposed model to perform dysarthric speech recognition and reconstruction to serve optimal communication amongst peers. The proposed model can be vital for numerous commercial applications, including HMI, speech-controlled smart tasks, etc. The statistical performance revealed that the proposed model achieves the highest known accuracy for DSR and ASR systems. Since, the proposed model was examined over Kannada bi-syllabic words only, and hence, in future, it can be assessed with other languages as well. Moreover, the model contributed in this paper was examined for word-level dysarthric speech (intelligible) transformation. In future, authors can explore its efficacy for sentence-level transformation.

**Funding** No funding has been received for this research work.

**Data availability** Real time data has been collected from the premier institute.

## Declarations

**Conflict of Interest** The authors of this paper hereby declare that there is no conflict of interest.

## References

1. Latha M, Shivakumar M, Manjula R. A study of acoustic characteristics, prosodic and distinctive features of dysarthric speech. *Grenze Int J Comput Theory Eng (Spec Issue)*. 2018;2018:228–35.
2. Farhadipour A, Veisi H, Asgari M, Keyvanrad MA. Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks. *ETRI J*. 2018;40(5):643–52.
3. Mohammed SY, Ahmed SS, Brahim ZF, AsmaBouchair B. Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. *EURASIP J Audio Speech Music Process*. 2020;1:1–7.
4. Young S, Evermann G, Gales M, Hain T, Kershaw D, Moore G, Odell J, Ollason D, Povey D, Valtchev V, et al. *The htk book (for htk version 3.3)*. Cambridge University Engineering Department, 2005. 2006.
5. Oue S, Marxer R, Rudzicz F. Automatic dysfluency detection in dysarthric speech using deep belief networks. In: *Proceedings of SLPAT 2015: 6th workshop on speech and language processing for assistive technologies*. 2015. p. 60–4.
6. Kim MJ, Cao B, An K, Wang J. Dysarthric speech recognition using convolutional LSTM neural network. In: *Interspeech*. 2018. p. 2948–52.
7. Vachhani B, Bhat C, Kopparapu SK. Data augmentation using healthy speech for dysarthric speech recognition. In: *Interspeech*. 2018. p. 471–5.
8. Bhat C, Das C, Vachhani B, Kopparapu SK. Dysarthric speech recognition using time-delay neural network based denoising autoencoder. In: *Interspeech*. 2018. p. 451–5.
9. Liu S, Hu S, Liu X, Meng H. On the use of pitch features for disordered speech recognition. 2019. p. 4130–4.
10. Wu J. Application of artificial neural network on speech signal features for Parkinson's disease classification. 2019.
11. Hu S, Liu S, Chang HF, Geng M, Chen J, Chung TKH, Yu J, Wong KH, Liu X, Meng H. The cuhk dysarthric speech recognition systems for English and Cantonese. 2019. p. 3669–70.
12. Borrie SA, Baese-Berk M, Van Engen K, Bent T. A relationship between processing speech in noise and dysarthric speech. *J Acoust Soc Am*. 2017;141(6):4660–7.
13. Memon S, Gregory MA. A novel approach for MFCC feature extraction. In: *IEEE Conference*. 2011. p. 1–5.
14. Albaqshi H, Sagheer A. Dysarthric speech recognition using convolutional recurrent neural networks. *Int J Intell Eng Syst*. 2020;13(6):384–92.

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

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