Wireless Personal Communications (2022) 122:293–307
https://doi.org/10.1007/s11277-021-08899-x

Dysarthric Speech Recognition Using Variational Mode Decomposition and Convolutional Neural Networks

R. Rajeswari[1] · **T. Devi**[1] · **S. Shalini**[1]

Accepted: 8 August 2021 / Published online: 24 August 2021

Abstract
Dysarthric speech recognition requires a learning technique that is able to capture dysarthric speech specific features. Dysarthric speech is considered as speech with source distortion or noisy speech. Hence, as a first step speech enhancement is performed using variational mode decomposition (VMD) and wavelet thresholding. The reconstructed signals are then fed as input to convolutional neural networks. These networks learn dysarthric speech specific features and generate a speech model that supports dysarthric speech recognition. The performance of the proposed method is evaluated using UA-Speech database. The average accuracy values obtained by the proposed method for speakers with different intelligibility levels with VMD based enhancement and without enhancement are 95.95 and 91.80% respectively. The proposed method also provides an increased accuracy value compared to existing methods that are based on generative models and artificial neural networks.

Keywords Automatic speech recognition · Dysarthric speech · Variational mode decomposition · Convolutional neural networks

1 Introduction

Dysarthria represents a speech disorder which is caused by neuromuscular disturbances [1]. It is caused by damage to the brain which occurs at the time of birth, heart attack, tumor or due to other diseases such as cerebral palsy and multiple sclerosis. People suffering from dysarthria exhibit delayed speech, speech with fluctuating speed or articulated speech. They do not have any problems in reading, writing or even in understanding other's speech. Due to the differences in the speech rate of dysarthric people, it is difficult to understand their speech. Hence, automatic speech recognition systems can be helpful in understanding dysarthric speech. Generic automatic speech recognition systems do not perform well in recognizing dysarthric speech [2]. Hence, it is essential to develop automatic speech recognition systems that specifically model the speech of dysarthric people.

✉ R. Rajeswari
rajeswari@buc.edu.in

[1] Department of Computer Applications, Bharathiar University, Coimbatore, Tamil Nadu 641046, India

This paper intends to first enhance the dysarthric speech signals and then develop a model based on convolutional neural networks to automatically recognize single words of dysarthria people.

Due to changes in the vocal fold and glottis muscle the dysarthric signal is different from a normal speech in terms of pitch and amplitude [3]. This difference is caused due to source degradation i.e. due to damage of the muscles in the vocal tract. Experiments have been conducted to find the similarity between speech in noise, where speech is degraded by environmental noise and dysarthric speech, where speech is degraded at source [4]. It has been shown through these experiments that there is a relationship between speech in noise and dysarthric speech. It has also been shown by Yakoub et al. [5] that usage of empirical mode decomposition (EMD) for denoising the dysarthric speech helps in improving the dysarthric speech recognition. Moreover, Ram et al. have shown that variational mode decomposition (VMD) [6] outperforms in speech enhancement [7]. These results have motivated us to make use of VMD to first denoise the dysarthric speech and then build a model using the enhanced signals to recognize the speech. The two main contributions of this paper are (1) application of variational mode decomposition and wavelet thresholding for enhancement of dysarthric speech signals; and (2) development of a convolutional neural network (CNN) based model for automatic recognition of dysarthric speech. The rest of this paper is organized as follows. Section 2 gives an overview of recent work carried out in dysarthric speech recognition. Section 3 describes the methodology proposed in this paper for dysarthric speech recognition. Section 4 describes the dataset used and the results obtained. Section 5 gives the conclusion and directions for further work.

2 Related Work

A lot of research work has been carried out to develop automatic speech recognition systems specifically for dysarthric speech. These works have improved the performance of the dysarthric speech recognition systems by enhancing the dysarthric speech signals and/ or by designing generative or machine learning models to recognize dysarthric speech. Dysarthric speech signals are considered as signals corrupted with noise from source [4] and preprocessing techniques such as enhancement or denoising techniques are applied to improve the dysarthric speech. The generative or machine learning models make use of training data to learn and recognize the dysarthric speech effectively.

Various research works have been carried out that improve dysarthric speech signals by performing signal enhancement. Park et al. mention that initial consonants in dysarthric speech are similar to noise. Hence, they propose a consonant-vowel dependent Weiner filter to remove the noise introduced during pronouncing initial consonants by dysarthric people [8]. Wisler et al. also treat dysarthric speech as speech corrupted with noise [9]. They make use of transfer learning for feature extraction which is robust against noise for classification of dysarthric speech and normal speech. Borrie et al. have also proved that there is a high similarity between speech signals corrupted by environmental distortion and dysarthric speech signals which are distorted at source [4].

Various approaches based on generative models [10–12], artificial neural networks (ANN) [13, 14] and deep neural networks [5, 15–17] are available in the literature for automatic speech recognition of dysarthric speech. These works have predominantly concentrated on designing novel or hybrid features and/ or in building machine learning based systems to effectively recognize dysarthric speech. Generative models are very useful in

modeling spectral sequences which are time varying. One of the widely used generative models is hidden Markov Model (HMM). A HMM based system is developed by Deller et al. [10] to recognize the words uttered by dysarthric speakers. They have performed pre-processing at signal level and latent position (LP) parameter vector to model each word by HMM. A HMM based speech recognizer is used by Lee et al. [11] to assess the level of dysarthria. They have trained word-specific HMM using mel-frequency cepstral coefficients (MFCC). Gaussian mixture models (GMM) are also used to represent the phoneme in every word. Class-specific HMM models are developed by Rajeswari et al. [12] that use MFCC features. The output of these models are given as input to support vector machine (SVM) to enable dysarthric speech recognition. In recent years, artificial neural networks (ANNs) are used to recognize dysarthric speech from acoustic features. An ANN based dysarthric speech recognition system has been built by Shahamiri et al. [13] that makes use of MFCC features. The best set of MFCC features is first identified which is then used by multilayer perceptron (MLP) to recognize the words spoken by dysarthric speakers. Hybrid ANN and HMM speech recognition systems are also developed which utilize the advantages of both ANN and HMM. A study carried out by Polur et al. [14] has shown that a hybrid ANN and HMM model based on MFCC features provides high accuracy in recognizing dysarthric speech. Very recently, deep neural networks are also deployed in automatic recognition of dysarthric speech. A convolutive bottleneck network is proposed by Nakashika et al. [15] to extract features from dysarthric speech signals. The bottleneck layer is used in the network to help extract features that are specific to dysarthric speech. A hybrid DNN and HMM based acoustic model is developed by Joy et al. [16] specifically for the TORGO dysarthric speech database. The system is speaker dependent and utilizes MFCC features to perform speech recognition. Two separate DNN models based on convolutional neural network (CNN) and long short term memory (LSTM) network are developed by Zaidi et al. [17] for dysarthric speech recognition. Acoustic features such as MFCCs, mel-frequency spectral coefficients (MFSCs) or perceptual linear prediction coefficients are extracted from dysarthric speech signals and given as input to CNN and LSTM models. HMM based models have been a standard for automatic speech recognition for decades. Although HMMs are useful in modeling speech signals which vary on time, they are not very effective in discriminating the features for a speech recognition system [12]. ANNs have the advantage of effectively discriminating the features. However, the limitation of ANNs is that handcrafted features have to be extracted from speech signals and given as input to ANNs. The advantage of using deep neural network (DNN) based models is that instead of using handcrafted features, these models learn the features themselves.

Some speech recognition systems improve the performance by first pre-processing the speech signals and then utilizing a machine learning based model for recognition. For instance, Yakoube et al. have utilized empirical mode decomposition (EMD) to enhance the dysarthric speech and later used convolutional neural networks (CNN) to perform phoneme recognition [5]. They have extracted mel-frequency cepstral coefficients (MFCCs) after speech enhancement and given them as input to CNN. In this work also, dysarthric speech signals are considered as noisy signals and hence they are first enhanced and then a CNN based model is developed to recognize the dysarthric speech. Ram et al. [7], have proved that variational mode decomposition (VMD) outperforms EMD in speech enhancement. This has motivated us to use VMD and wavelet based enhancement of dysarthric speech signals. CNNs, which are a type of deep neural networks (DNNs), can automatically learn features from the given input. Hence, in this work a CNN based speech model is developed to automatically learn features from the enhanced dysarthric speech signals for speech recognition.

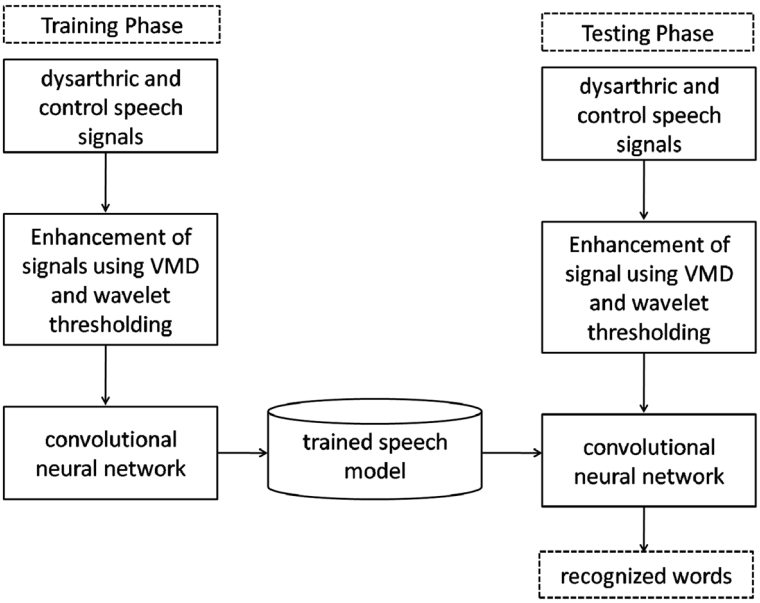


Fig. 1 Block diagram of the proposed system

3 Proposed Method

The CNN based dysarthric speech recognition system proposed in this paper consists of two main phases viz., training phase and the testing phase. The training phase builds and trains the CNN using the dysarthric speech signals. The trained CNN model is evaluated in the testing phase. The training and testing phases consists of two steps. In the first step, the dysarthric speech signals are enhanced using variational mode decomposition. In the second step, convolutional neural networks are used to automatically recognize the word from the speech signal given as input. This section explains these steps involved in automatic speech recognition of dysarthric speech. The block diagram of the proposed method is presented in Fig. 1.

3.1 Enhancement of Dysarthric Speech Using Variational Mode Decomposition and Wavelet Thresholding

VMD takes the advantages of Weiner filtering, Hilbert transform, frequency mixing and heterodyne demodulation to decompose the signal into modes and appropriately reconstruct them. The main objective of variational mode decomposition (VMD) [6] is to decompose the signal into several modes, v_k. These modes are also called as sub signals. Every mode occurs around a centre frequency μ_k. The constrained variational problem for the variational mode decomposition is given by Eq. (1).

$$\begin{aligned} &\min \{w_k\}\{\mu_k\}\left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\Pi t} \right) * v_k(t) \right] e^{-j\mu_k t} \right\|_2^2 \right\} \\ &\text{such that } \sum_k w_k(t) = f(t) \end{aligned} \tag{1}$$

where $f(t)$ is the signal that has to be decomposed, $\left[\left(\delta(t) + \frac{j}{\Pi t} \right) * v_k(t) \right]$ is the Hilbert transformation of w_k, $\delta(t)$ is the Dirac distribution, $*$ is the convolution operation, k ranges from 1 to K (predefined number of modes), t is the time index, j is the imaginary part, ∂_t is the partial derivation, μ_k is the center frequency of the kth mode. $e^{-j\mu_k t}$ shifts the frequency spectrum of every k^{th} mode in its corresponding base mode. The notations $w_k = w_1, w_2, \ldots, w_K$ and $\mu_k = \mu_1, \mu_2, \ldots, \mu_K$ are used to represent the set of all modes and their center frequencies respectively. In the present work, the number of modes, K, is 3. The augmented Lagrangian L can be used to solve the above mentioned constrained variational problem using Eq. (2).

$$\begin{aligned} L(\{w_k\}, \{\mu_k\}, \lambda) =& \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\Pi t} \right) * v_k(t) \right] e^{-j\mu_k t} \right\|_2^2 \\ &+ \left\| f(t) - \sum_k w_k(t) \right\|^2 \\ &+ \left\langle \lambda(t), f(t) - \sum_k w_k(t) \right\rangle \end{aligned} \tag{2}$$

where λ is the Lagrange multiplier and α is the data fidelity constraint parameter. The augmented Lagrangian helps in imposing the constraints. The optimization problem of Eq. (1) can be solved by finding the saddle point of Lagrangian in Eq. (2). The solution of Eq. (2) results in modes, v_k. It has been proved by Donoho et al. [18] that thresholding of detail wavelet coefficients of a signal helps in denoising the signal. Hence, this approach is used in this work to denoise/ enhance the dysarthric speech signals. Firstly, discrete wavelet transform (DWT) is applied to the modes v_k which results in the detail coefficients, D and approximation coefficients, A. Let every level of detail coefficients obtained from DWT be represented by D_L. In ths work, the number of waveletet decomposition levels, L, is 3. For every D_L, soft thresholding of wavelet coefficients is performed using Eq. (3) using a threshold th [18]. In the present work, the threshold value th is computed on a trial and error basis and is set as 0.5.

$$D_L^T = \begin{cases} D_L - th & \text{if } D_L \geq th \\ 0 & \text{if } D_L < th \\ D_L + th & \text{if } D_L \leq -th \end{cases} \tag{3}$$

The enhanced detail coefficients and the approximation coefficients of the modes v_k are used for reconstruction of the modes using inverse discrete wavelet transform. These enhanced modes v_k^e are then summed together for the VMD reconstructed. The reconstructed signal represents the denoised dysarthric speech signal f_d.

3.2 Speech Recognition Using Convolutional Neural Networks

Convolutional neural networks (CNN) belong to a class of feed forward ANN models with a large number of hidden layers that perform feature extraction and classification tasks. The enhanced speech signal is passed as input to a series of convolutional layers and pooling layers which is followed by a fully connected layer to produce class scores. The convolutional layers and pooling layers together act as feature extractors from the input signals, and the fully connected layer acts as a classifier to recognize the spoken word. There are five main operations in the CNN. They are (a) convolution (b) non-linearity (c) pooling (d) flattening and (e) classification. In the convolutional layer, feature maps are extracted from the input signals with the help of kernels. Kernel matrix acts as a sliding window that is moved over the input signal. For every slide, convolution operation is performed with the kernel matrix and the underlying elements of the input. Rectified Linear Unit (ReLU) is the activation function that is applied on the feature maps obtained using convolution operation. ReLU is a piece-wise linear function that examines each element in the feature map and passes the element if it is positive. If the element is negative then the negative element is replaced by zero. It is considered that positive value accumulates knowledge and so it is preserved, but the negative value contributes nothing or minimal and hence it is eliminated. Pooling operation is a subsampling operation that does the dimensionality reduction of each feature map, but retains the knowledge. Subsampling operation is performed by taking the maximum or average over the collection of values under the defined window. In this work, a maximum operation is performed. The flattening layer flattens the output generated by the previous layer to turn them into a single vector that can be used as an input for the next layer. The classification task is performed by a series of fully connected layer which is a flat feed-forward neural network layer that uses non-linear activation function in order to output probabilities of word prediction.

4 Experimental Results

In this section experiments carried out to evaluate the proposed method using a benchmark dataset for dysarthric speech viz., Universal Access (UA)-Speech database and the results obtained are elaborated. The dysarthric speech signals are first enhanced using VMD and wavelet based thresholding. Later, the CNN based model is developed and evaluated for dysarthric speech recognition. All the experiments are carried out using a system with 1.99 GHz processor and 16GB of RAM. Section 4.1 describes the dataset used to evaluate the proposed method and Sect. 4.2 elaborates the results obtained.

4.1 Dataset used for Dysarthric Speech Recognition

The proposed method for dysarthric speech recognition is evaluated using Universal Access (UA)-Speech database [19]. The database consists of 10 digits, 26 radio alphabets, 19 computer commands and 100 common words spoken by 19 dysarthric speakers. The data has been recorded by a 8-microphone array and a digital video camera. The speakers have different speech intelligibility levels, such as ‘very low (0–25%)’, ‘low (25–50%)’, ‘mid (50–75%)’ and ‘high (75–100%)’. The severity of dysarthria is inversely related to the speech intelligibility. For instance, if the speaker is suffering from severe dysarthria, his speech intelligibility will be very low.

The experiments carried out in this work use 10 digits, 19 computer commands and 100 common words spoken by 10 dysarthric speakers (4 male, 6 female) taken from the UA-Speech database. Table 1 shows the list of speakers and the number of utterances made by them for every class of words. To evaluate the performance of the proposed method robustly, the dysarthric speakers belonging to all intelligibility levels are considered in this work. The words uttered by 10 speakers without disabilities from the same database are considered as control words. The dataset is divided into 80% and 20% to form training and test sets respectively. This work uses 2470, 1300 and 130,000 utterances of computer commands, digits and common words respectively. Out of these, 1976, 1040 and 104,000 utterances of computer commands, digits and common words respectively are used for training and the remaining are used for testing. The proposed system is implemented in Python language and Keras Library.

4.2 Results

The dysarthric speech signals are enhanced using VMD with the number of modes as 3. The Daubechies wavelets with 5 levels are used for discrete wavelet transform. Soft wavelet thresholding is used in this work with a threshold, *th*, value of 0.5. The original signal representing the word 'people' spoken by dysarthric speaker 'F02' along with its 3 modes obtained using VMD, wavelet thresholded VMD modes and the enhanced speech signal are shown in Fig. 2. The speech quality of the enhanced dysarthric speech signals is compared with their original signals using the performance measure signal-to-noise-ratio (SNR). In order to calculate the SNR values, the words spoken by corresponding control speakers are used as reference signals. The average SNR values obtained for original dysarthric speech signals and enhanced dysarthric speech signals are presented in Table 2. The SNR values of enhanced dysarthric speech signals using VMD are higher compared to original dysarthric speech signals, which prove that the VMD and wavelet thresholding help in enhancing the dysarthric speech signals.

Table 1 Details of speakers and number of utterances used in this work

Speaker	Intelligibility level	Number of utterances for each word		
		Computer commands	Digits	Common words
F02	Low (29%)	133	70	700
F03	Very low (6%)	133	70	700
F04	Mid (62%)	133	70	700
F05	High (95%)	133	70	700
M01	Very low (19%)	76	40	400
M04	Very low (2%)	95	50	500
M05	Mid (58%)	133	70	700
M07	Low (28%)	133	70	700
M08	High (95%)	133	70	700
M09	High (86%)	133	70	700
Control Speakers: CF02, CF03, CF04, CF05, CM01, CM04, CM05, CM07, CM08, CM09		1235	650	6500
Total utterances		2470	1300	13,000

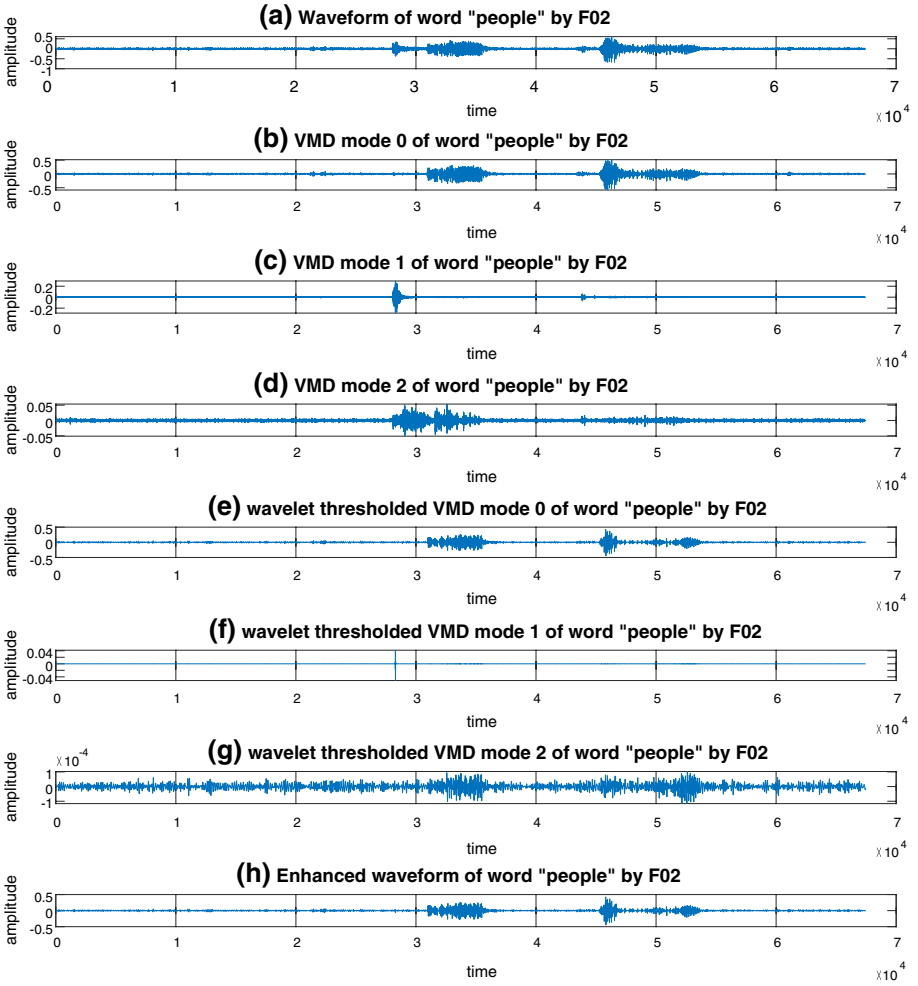


Fig. 2 Original and enhanced signals of word 'people' uttered by F02 **a** original waveform **b** VMD mode 0 **c** VMD mode 1 **d** VMD mode 2 **e** wavelet thresholded VMD mode 0 **f** wavelet thresholded VMD mode 1 **g** wavelet thresholded VMD mode 2 **h** Enhanced waveform

Table 2 Performance of VMD and wavelet thresholding based enhancement of dysarthria speech signals

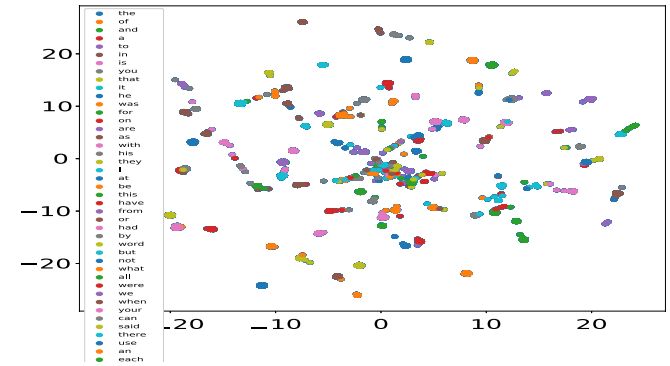
Word category	SNR (dB)	
	Original speech signals	Enhanced speech signals
Common words	3.96	**6.21**
Computer commands	10.25	**15.27**
Digits	3.73	**19.08**

Convolutional neural networks (CNNs) are used in this work to learn the dysarthric speech recognition model from the enhanced speech signals obtained using VMD and wavelet thresholding. CNN has a one-dimensional (1D) input layer which takes the enhanced signal as input. It has four 1D convolutional layers with 8, 16, 32 and 64 output filters respectively. The size of the kernel used in these four convolution layers are 13, 11, 9 and 7 respectively. The rectified linear unit (ReLU) function is used as activation functions in the convolution layers. The max pooling layer helps in performing max pooling operation with a pool size of 3. The max pooling layer is followed by a drop out layer with 20% of neurons being dropped out. The final layer recognizes the word with the number of neurons equal to the number of words in that category and 'softmax' activation function. The scatter plots of features obtained from CNN model for the test instances of common words, computer commands and digits are presented in Fig. 3. The features on 100 classes of common words, 19 classes of computer commands and 10 classes of digits are visualized by t-distributed Stochastic Neighbor Embedding (t-SNE) [20]. t-SNE helps in reducing the dimension of the features so that it is suitable to visualize them. In the present work, t-SNE is used to visualize the features of CNN in a two dimensional space. It can be observed that the features are discriminative.

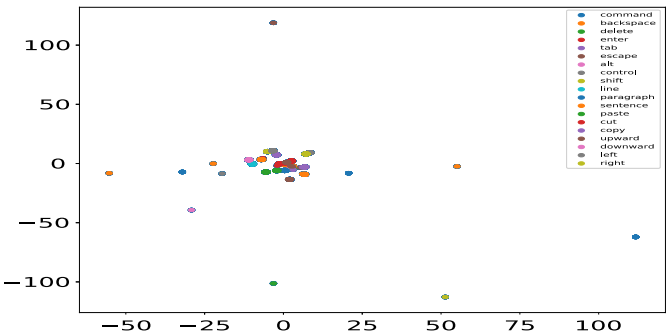
The results obtained for various categories of words, i.e. for common words, computer commands and digits are presented in Table 3. It can be observed from the results that the recognition accuracies for all the categories of words are better in CNN models with VMD based enhancement. Speaker specific results are presented in Table 4. Figures 4, 5, 6 and 7 show the accuracies of speakers with different intelligibility levels 'very low', 'low', 'mid', and 'high' respectively. It can be observed from the results that the average accuracy obtained is 91.80% for the CNN based model. The speech enhancement using VMD and wavelet based thresholding along with CNN has an average accuracy of 95.95%. For every speaker also, the accuracies obtained using VMD based enhancement and CNN provides better results compared to using CNN alone. An exception to this is the results of speaker M01. For this speaker, the results of CNN is better than VMD based enhancement and CNN. The reason for this may be due to inconsistency in words spoken by this speaker. The results obtained using the proposed method are compared with the existing methods available in the literature and are presented in Table 5. It can be observed that the proposed method gives better results compared to the existing methods which are based on generative models and ANN. The results of the proposed method are better than the results obtained using MFCC features and ANN proposed by Shahamiri et al. [13], but they have used all the utterances of all speakers in UA-Speech database.

5 Conclusion

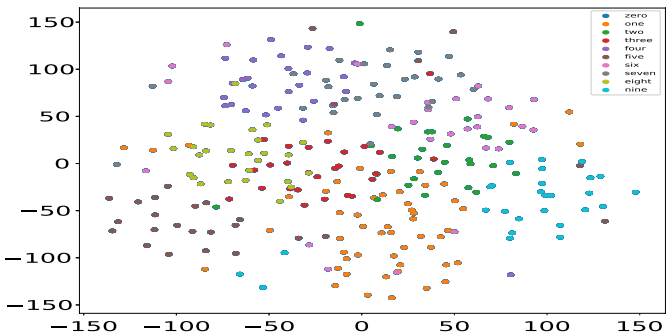
This paper proposed a method for dysarthric speech recognition which is based on two steps. In the first step, variational mode decomposition and wavelet thresholding based speech signal enhancement are performed. In the second step, convolutional neural network is used for learning features from the enhanced signals automatically, rather than providing handcrafted features. The average accuracy obtained from the proposed speech recognition method is 95.95% with VMD based enhancement and 91.80% without speech enhancement. The results of the proposed method are better compared to the results of existing methods. In the present work, only common words, computer commands and digits of 10 speakers from UA-Speech database are used for evaluation. In future, the entire



(a)



(b)



(c)

Fig. 3 Visualization of features for **a** common words **b** computer commands **c** digits

Table 3 Results for various categories of words

Word category	Accuracy (%)	
	CNN	VMD based enhancement and CNN
Common words	91.46	**93.39**
Computer commands	93.76	**97.08**
Digits	97.04	**98.33**

Table 4 Speaker specific results obtained for the proposed method

Intelligibility Level	Speaker	Accuracy (%)	
		CNN	VMD + CNN
very low (0–25%)	F03 (6%)	96.95	**97.51**
	M01 (19%)	**98.48**	94.46
	M04 (2%)	94.74	**95.57**
Average		**96.72**	95.85
Low (26%–50%)	F02 (29%)	91.71	**95.85**
	M07 (28%)	85.17	**88.40**
Average		88.44	**93.37**
Mid (51%–75%)	F04 (62%)	95.43	**97.65**
	M05 (58%)	86.12	**91.71**
Average		90.78	**94.68**
High (76%–100%)	F05 (95%)	97.65	**97.92**
	M08 (95%)	85.17	**98.90**
	M09 (86%)	86.60	**99.04**
Average		89.81	**98.62**
Overall average		91.80	**95.95**

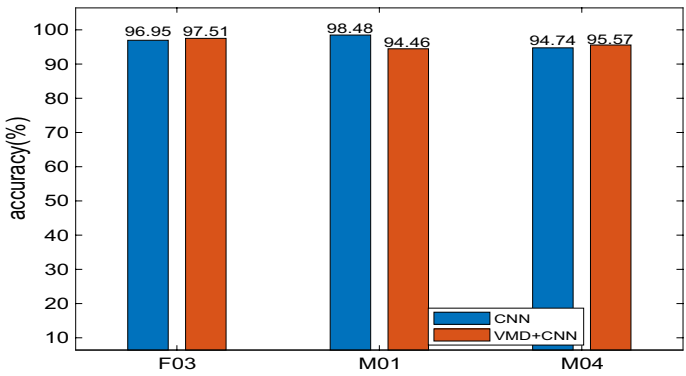


Fig. 4 Accuracies for very low intelligibility speakers

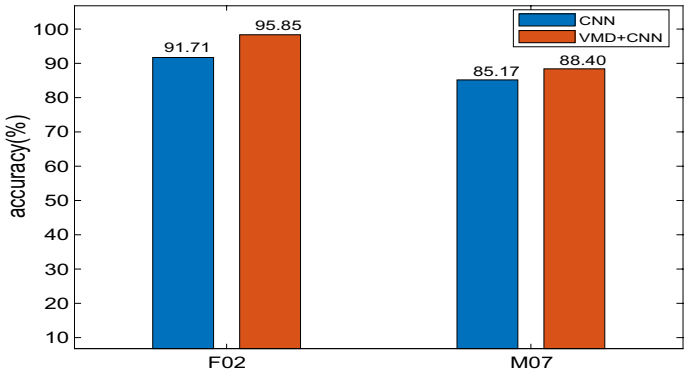


Fig. 5 Accuracies for low intelligibility speakers

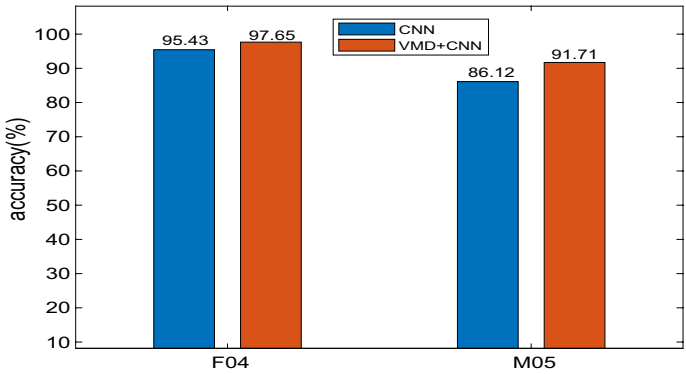


Fig. 6 Accuracies for mid intelligibility speakers

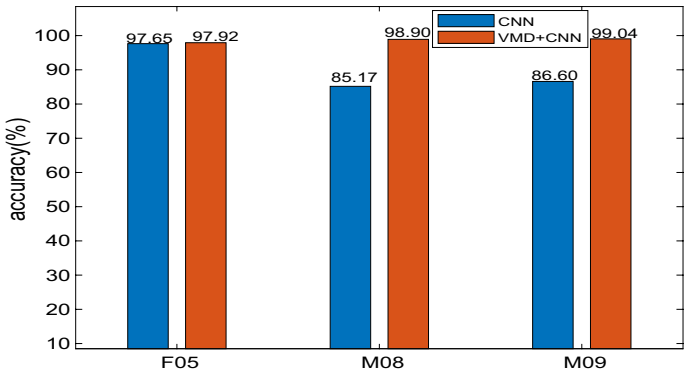


Fig. 7 Accuracies for high intelligibility speakers

Table 5 Comparison of results with existing methods

Method	Dataset used	Pre-processing	Features extracted	Classification method	Accuracy (%)
HMM+SVM [12]	UA-Speech (19 speakers, 19 computer commands and 10 digits)	–	Class-specific HMM using MFCC	SVM	87.91
MLP [13]	UA-Speech (19 speakers, 100 common words, 19 computer commands, 10 digits and 26 radio alphabets)	–	MFCC	MLP (1 hidden layer)	68.88
EMD+CNN [5]	Nemours Corpus	EMD	MFCC	CNN (1 convolutional layer, 1 pooling layer and 2 fully connected layer)	64.86
proposed work (CNN)	UA-Speech (10 speakers, 100 common words, 19 computer commands and 10 digits)	–	–	CNN (4 convolutional layers, 4 pooling layers, 3 fully connected layers and 3,245,124 parameters)	91.80
proposed work (VMD+CNN)	UA-Speech (10 speakers, 100 common words, 19 computer commands and 10 digits)	VMD	–	CNN (4 convolutional layers, 4 pooling layers, 3 fully connected layers and 3,245,124 parameters)	**95.95**

dataset of UA-Speech database will be used for evaluation. Moreover, the layers of the CNNs will be designed to extract features effectively.

Acknowledgements The authors acknowledge the support of the Biomedical Device and Technology Development, Department of Science and Technology, India. The authors would like to thank Professor Mark Hasegawa-Johnson of the University of Illinois for kindly allowing to access the UA-Speech database. The authors would like to thank Bharathiar University for providing the necessary support.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Rampello, L., Rampello, L., Patti, F., & Zappia, M. (2016). When the word doesnt come out: A synthetic overview of dysarthria. *Journal of the Neurological Sciences, 369*, 354–360.
2. Moore, M., Demakethepalli, V. H., & Panchanathan, S. (2018). Whistle-blowing ASRS: Evaluating the need for more inclusive automatic speech recognition systems. *Proceedings of the Annual conference of the International Speech Communication Association INTERSPEECH, 2018*, 466–470.
3. Thoppil, M. G., Kumar, C. S., Kumar, A., & Amose, J. (2017). Speech signal analysis and pattern recognition in diagnosis of dysarthria. *Annals of Indian Academy of Neurology, 20*(4), 302–357.
4. Borrie, S. A., Berk, M. B., Engen, K. V., & Bent, T. (2017). A relationship between processing speech in noise and dysarthric speech. *Journal of Acoustics Society of AMerica, 141*(6), 4460–4467.
5. Yakoub M. S., Selouani S. A., Zaidi B. F., & Bouch A. (2020). Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural networks, EURASIP Journal on Audio, Speech and Music Processing, Article ID: 1. https://doi.org/10.1186/s13636-019-0169-5.
6. Dragomiretskiy, K., & Zosso, D. (2014). Variational mode decomposition. *IEEE Transactions on Signal Processing, 62*(3), 531–544.
7. Ram, R., & Mohanty, M. N. (2017). Comparative analysis of EMD and VMD algorithm in speech enhancement. *International Journal of Natural Computing Research, 6*(1), 17–35.
8. Park, J.H., Seong, W.K., & Kim, H.K. (2011). 'Preprocessing of Dysarthric Speech in Noise Based on CV-Dependent Wiener Filtering', In: Delgado RC., Kobayashi T. (eds) Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop, Springer, New York, pp. 41–47.
9. Wisler, A., Berisha, V., Spanias, A., & Liss, J. (2016). 'Noise robust dysarthric speech classification using domain adaptation', 2016 Digital Media Industry and Academic Forum (DMIAF), pp. 135–138.
10. Deller, J. R., Hsu, D., & Ferrier, L. J. (1991). On the use of hidden Markov modelling for recognition of dysarthric speech. *Computers Methods and Programs in Biomedicine, 35*(2), 125–139.
11. Lee, S. H., Kim, M., Seo, H. G., Oh, B. M., Lee, G., & Leigh, J. H. (2019). Assessment of dysarthria using one word speech recognition with hidden Markov models. *Journal of Korean Medical Science, 34*(13), e108. https://doi.org/10.3346/jkms.2019.34.e108
12. Rajeswari, N., & Chandrakala, S. (2016). Generative model-driven feature learning for dysarthric speech recognition. *Biocybernetics and Biomedical Engineering, 36*, 553–561.
13. Shahamiri, S. R., & Salim, S. S. B. (2014). Artificial networks as speech recognizers for dysarthric speech: Identifying the best performing set of MFCC parameters and studying a speaker independent approach. *Advanced Engineering Informatics, 28*, 102–110.
14. Polur, P. D., & Miller, G. E. (2006). Investigation of an HMM/ ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals. *Medical Engineering and Physics, 28*, 741–748.
15. Nakashika, T., Yoshioka, T., Takiguchi, T., Ariki, Y., Duffner, S., & Garcia, C. (2014). Convolutive bottleneck network with dropout for dysarthric speech recognition. *Transactions on Machine Learning and Artificial Intelligence, 2*(2), 1–15.
16. Joy, N. M., & Umesh, S. (2018). Improving acoustic models in TORGO dysarthric speech database. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26*(3), 637–645.
17. Zaidi, B. F., Selouani, S. A., Boudraa, M., & Yakoub, M. S. (2021). Deep neural network architectures for dysarthric speech analysis and recognition. *Neural Computing and Applications*. https://doi.org/10.1007/S00521-020-05672-2
18. Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory, 41*(3), 613–627.
19. Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., & Frame S. (2008). 'Dysarthric speech database for universal access research', In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 1741–1744.

20. van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.

R. Rajeswari has completed her MCA in 2003 from Madurai Kamaraj University, Madurai, India and Ph.D. in Computer Science in 2012 from Bharathiar University, Coimbatore, India. She joined Department of Computer Applications, Bharathiar University, Coimbatore, India in 2005 as Assistant Professor. She is currently working as Associate Professor in the Department of Computer Applications, Bharathiar University, Coimbatore, India. She has 15 years of experience in teaching/research. Her main research interests include medical image processing, video processing and soft computing. She is currently guiding M.Phil and Ph.D research scholars in Bharathiar University. She is a life member of Computer Society of India.

T. Devi Ph.D.(UK), Dean, Faculty of Research, Professor and Head, Department of Computer Applications, Bharathiar University focuses on state-of-art technology that industries adopt in order to make the students ready for the future world. She is a Gold Medalist (1981-1984) from University of Madras and a CommonWealth Scholar (1994-1998) for her Ph.D from University of Warwick, UK. She has three decades of teaching and research experience from Bharathiar University, Indian Institute of Foreign Trade, New Delhi and University of Warwick, UK. Professor is good in team building and setting goals and achieving. Her research interests include integrated data modeling and framework, meta-modeling, computer assited concurrent engineering and speech processing. Professor had visited UK, Tanzania and Singapore for academic collaborations. She has received various awards including Commonwealth Scholarship, best alumni award and guided 21 Ph.D. Scholars.

S. Shalini is working as Project Assistant for Department of Science and Technology - Biomedical Devices and Technology Development Programme at Bharathiar University, Coimbatore from 2019. She has completed her M.Phil in the year 2019 and MCA in the year 2017. She has the industrial experience more than a year. She also holds one and half years of research experience. Her specialization areas are Speech Processing and Web designing. She was awarded as the 'BEST OUTGOING STUDENT' by Dr. Mahalingam College of Engineering and Technology, Pollachi in the year 2017. Her MCA final year internship project was also awarded as the Best Project of the department by Dr. Mahalingam College of Engineering and Technology, Pollachi in the year 2017. Mrs. S. Shalini has completed Business English Certification from University of Cambridge, UK in the year 2011. She has won various prizes in the elocutions conducted at different occasions.