Modeling Sub-Band Information Through Discrete Wavelet Transform to Improve Intelligibility Assessment of Dysarthric Speech

Laxmi Priya Sahu[1], Gayadhar Pradhan[1], Jyoti Prakash Singh[2] *

[1] Department of Electronics and Communication Engineering, National Institute of Technology Patna, (India)

[2] Department of Computer Science and Engineering, National Institute of Technology Patna (India)

Received 26 January 2022 | Accepted 25 March 2022 | Published 3 October 2022

Abstract

The speech signal within a sub-band varies at a fine level depending on the type, and level of dysarthria. The Mel-frequency filterbank used in the computation process of cepstral coefficients smoothed out this fine level information in the higher frequency regions due to the larger bandwidth of filters. To capture the sub-band information, in this paper, four-level discrete wavelet transform (DWT) decomposition is firstly performed to decompose the input speech signal into approximation and detail coefficients, respectively, at each level. For a particular input speech signal, five speech signals representing different sub-bands are then reconstructed using inverse DWT (IDWT). The log filterbank energies are computed by analyzing the short-term discrete Fourier transform magnitude spectra of each reconstructed speech using a 30-channel Mel-filterbank. For each analysis frame, the log filterbank energies obtained across all reconstructed speech signals are pooled together, and discrete cosine transform is performed to represent the cepstral feature, here termed as discrete wavelet transform reconstructed (DWTR)- Mel frequency cepstral coefficient (MFCC). The *i-vector* based dysarthric level assessment system developed on the universal access speech corpus shows that the proposed DTWR-MFCC feature outperforms the conventional MFCC and several other cepstral features reported for a similar task. The usages of DWTR-MFCC improve the detection accuracy rate (DAR) of the dysarthric level assessment system in the text and the speaker-independent test case to 60.094 % from 56.646 % MFCC baseline. Further analysis of the confusion matrices shows that confusion among different dysarthric classes is quite different for MFCC and DWTR-MFCC features. Motivated by this observation, a two-stage classification approach employing discriminating power of both kinds of features is proposed to improve the overall performance of the developed dysarthric level assessment system. The two-stage classification scheme further improves the DAR to 65.813 % in the text and speaker-independent test case.

Keywords

Approximation Coefficient, Cepstral Coefficients, Detail Coefficient, DWT, Dysarthria, Dysarthric Level, IDWT, Sub-Band Signal.

DOI: 10.9781/ijimai.2022.10.003

I. Introduction

Dysarthria reduces the speech intelligibility of a person by affecting the speech production system [1]. Parkinson's disease, amyotrophic lateral sclerosis, cerebral palsy, brain tumor, and brain injury are some of the causes for developing dysarthria in a person [2]–[4]. The speech intelligibility of a dysarthric person varies from near-normal to unintelligible depending on the level of severity. From the mid-level of severity, it is difficult to understand the spoken utterances of a dysarthric person by unfamiliar listeners [5]. In the conventional approach, the intelligibility level of a spoken utterance is measured by subjective assessment in clinical applications. The subjective assessment approach is costlier in time and money with a possibility of biasness towards the previous knowledge of experts about the type of disease [6]. Due to easy accessibility and consistent performance, recently, automated objective assessment methods are explored for the diagnosis of dysarthria in the primary stages [7].

The speech quality of a person suffering from dysarthria differs from a normal speaker due to change in loudness level, fundamental frequency, voice instability, voice breaks, and speaking rate [8], [9]. Consequently, the performance of the automatic speech recognition (ASR) system for a dysarthric speaker degrades compared to the normal speaker [10]. Using this aspect, some of the reported works used the ASR system for evaluating the level of dysarthria [10]–[14]. The word recognition rate (WRR), state-level log-likelihood ratios (SLLRs), and log-likelihoods (LLs) are used as the measuring parameter for the evaluation of the intelligibility level of the dysarthric speech. Such approaches are more suitable when a fixed set of words are used for testing different speakers. The performance of the ASR system also varies depending on the linguistic context of the speaker. The scarcity

* Corresponding author.

E-mail addresses: laxmipriya.ec16@nitp.ac.in (L. P. Sahu), gdp@nitp.ac.in (G. Pradhan), jps@nitp.ac.in (J. P. Singh).

of dysarthric speech data and limited availability of reference language models make the blind intelligibility assessment methods more useful [15]. The blind intelligibility assessment approaches use a classifier for differentiating the healthy speaker and dysarthric speaker and further separate according to their severity level. Some of the blind modeling approaches include the classical modeling methods like support vector machine (SVM) [16], Gaussian mixture model(GMM) [17], and recently reported neural network (NN) based modeling methods [18]–[22]. The NN based methods such as artificial neural network (ANN) [18], deep neural network (DNN) [19], convolutional neural network (CNN) [23], [24], long short-term memory network (LSTM) [25], bidirectional LSTM (BLSTM) and recurrent neural network (RNN) [26] have been explored for the intelligibility assessment of dysarthric speech. The *i-vector* representation of input speech data is also explored for assessment of dysarthria [27], [28]. The *i-vector* based representation maps the varying length of input speech utterance into a fixed dimension. Various feature projection and scoring schemes in combination with *i-vector* improve the performance of dysarthric level assessment system [27], [28]. Some of the reported works also used the combination of various statistical modeling and NN-based approaches [16], [17], [29]–[31]. Despite the use of sophisticated acoustic modeling methods the performances reported for dysarthric level assessment are less.

The aforementioned modeling approaches mostly use the spectral domain features for acoustic representation of the input speech data [18], [32]. The spectral representation of the input speech data such as spectrogram [33], log filterbank energy [26], Mel-frequency cepstral coefficients (MFCCs) [32], multitaper MFCC [18], perceptual linear prediction cepstral coefficients (PLPCCs) [34], [35], linear prediction cepstral coefficients (LPCC) [36], constant Q cepstral coefficients (CQCCs) [37] and line spectral frequencies have been explored for the assessment of dysarthric level. Several voice quality features [38]–[40], prosodic features [17], and excitation source features [41]–[43] are also explored for the assessment of dysarthria. The extraction of voice quality and prosodic feature from a speech signal is difficult and performance is highly dependent on the employed feature extraction approach. On the other hand, the cepstral feature can be easily extracted from the input speech data and very frequently used in the development of speech based applications.

The speech signal within a sub-band varies at a fine level depending on the type and level of dysarthria. Following the human perception of the speech signal [32], most of the cepstral features are extracted by analyzing the short-term magnitude spectra using a Mel-filterbank. Consequently, the fine level information present in the higher frequency regions is smoothed out due to the larger bandwidth of Mel-filters. The discriminating information present at the fine level in the short-term magnitude spectra can be captured up to a certain level by increasing the size of the filterbank. The increase in filterbank size may also capture the redundancy present in the lower frequency regions. Alternatively, the fine level information can be captured by decomposing the speech signal into different sub-band signals and analyzing the magnitude spectra of each sub-band signal [14]. Motivated by these observations, in this paper, we have firstly decomposed the speech signal using discrete wavelet transform (DWT) [44] into approximation and detail coefficients, respectively, at each level. The speech signals representing different sub-bands are then reconstructed using inverse DWT (IDWT) [44], [45]. In the process of IDWT, the speech signals are reconstructed by using the detail coefficient obtained at each level of decomposition and making all other coefficients to zero vector. Finally, at the last level of decomposition, the speech signal representing the lower frequency region is reconstructed by using only the approximation coefficients and making all detail coefficients to zero vectors. The log filterbank energies are computed by analyzing the short-term discrete Fourier transform (DFT) magnitude spectra of each reconstructed speech using the Mel-filterbank. For each analysis frame, the log filterbank energies obtained across all reconstructed speech are pooled together, and discrete cosine transform (DCT) is applied to represent the 13-dimensional base cepstral feature, here termed as discrete wavelet transform reconstructed - Mel frequency cepstral coefficient (DWTR-MFCC). The experimental results presented in this study show that DWTR-MFCC enhances discrimination among the overlapping classes compared to the conventional MFCC feature [46]. It also carries additional information to MFCC features. Motivated by this observation finally, a two-stage classification scheme is proposed to improve the overall performance of the dysarthric assessment system.

The remainder of the paper is organized as follows: Section II describes the proposed feature extraction method using DWT for the assessment of dysarthric level. Section III presents experimental setup for the development of *i-vector* based dysarthric level assessment system. The experimental results are presented in Section IV. Finally, Section V concludes this study.

II. Proposed Feature for Assessment of Dysarthric Level

The wavelet transform-based approaches are most preferred for time-frequency analysis of different types of signals [44], [47]–[50]. In the following section, decomposition and reconstruction of input speech using DWT followed by the proposed approach for the computation of DWTR-MFCC feature are presented.

A. Decomposition and Reconstruction of Speech Signal Using DWT

Using DWT, the speech signal $s(n)$ can be decomposed into high-frequency detail coefficients and low-frequency approximation coefficients by passing through a series of high-pass and low-pass filters, respectively. At a particular level of decomposition, the detail coefficient ($D_{i,j}$), and approximation coefficient ($A_{i,j}$) can be obtained as given in [44]. Equation (1) and Equation (2) represents the detail coefficients and approximation coefficients of the input signal $s(n)$, respectively.

$$D_{i,j} = \sum_{n} s(n)\psi_{i,j}(n) \tag{1}$$

$$A_{i,j} = \sum_{n} s(n)\phi_{i,j}(n) \tag{2}$$

where integer i and j provide the information about the amount of scaling and shifting of the wavelet function, respectively. The mother wavelet function ($\psi_{i,j}(n)$) and the father wavelet function ($\phi_{i,j}(n)$) are extracted from the continuous wavelet transform (CWT) using most commonly used dyadic grid arrangement [44], [47]. Equation (3) represents the mother wavelet function ($\psi_{i,j}(n)$).

$$\psi_{i,j}(n) = 2^{-i/2}\psi(2^{-i}n - j) \tag{3}$$

Equation (4) represents the father wavelet function ($\phi_{i,j}(n)$).

$$\phi_{i,j}(n) = 2^{-i/2}\phi(2^{-i}n - j) \tag{4}$$

Equation (5) refers to the representation of the signal $s(n)$ as the combination of detail and approximation coefficients.

$$s(n) = \sum_{j=-\infty}^{\infty} A_{i_0,j}\phi_{i_0,j}(n) + \sum_{i=-\infty}^{i_0}\sum_{j=-\infty}^{\infty} D_{i,j}\psi_{i,j}(n) \tag{5}$$

where, $A_{i_0,j}$ represents the approximation coefficient at level i_0. The approximation coefficient at i^{th} level of decomposition can be obtained by combining the detail and approximation coefficients obtained at

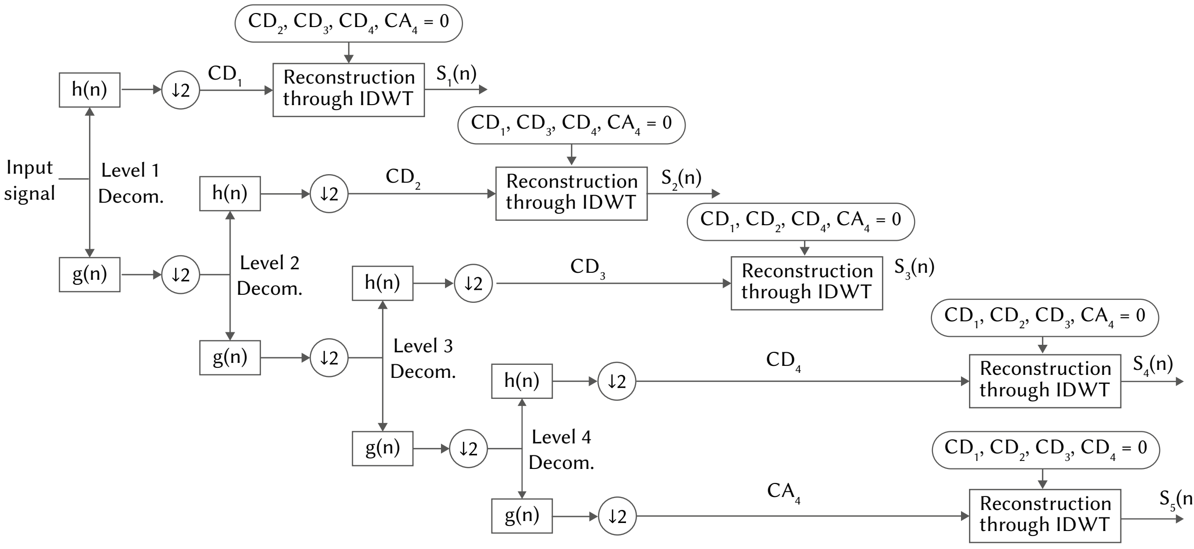

Fig. 1. The block diagram illustrating the DWT based four level multiresolution decomposition of the speech signal. The $h(n)$ and $g(n)$ represents the high pass and lowpass filter, respectively. CD_1, CD_2, CD_3, CD_4 and CA_4 represents the detail coefficient at level1, level2, level3, level4 and approximation coefficient at level4 decomposition, respectively. $s_l(n)$, $l = 1; 2; ...5$ represents the reconstructed signals.

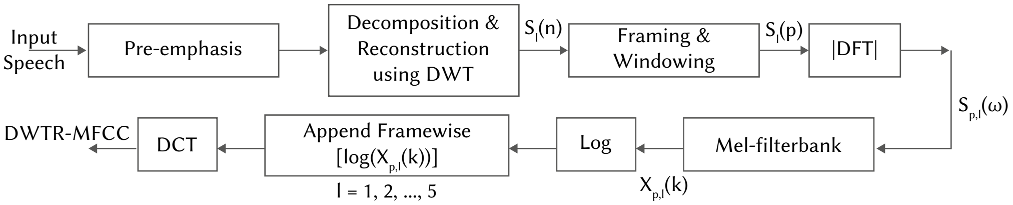

Fig. 2. The block diagram representation of proposed DWTR-MFCC feature extraction process.

$(i+1)^{th}$ decomposition level [44]. The block diagram representing four-level of wavelet decomposition and reconstruction of five sub-band speech signals is illustrated in Fig. 1. Let, the original speech signal $s(n)$ be the approximation coefficient at level zero. As shown in the block diagram, in each level, the approximation coefficient is decomposed into detail and approximation coefficients of the next higher level. The decomposition is performed by processing the approximation coefficient through a pair of high-pass and low-pass filters having impulse response $h(n)$ and $g(n)$, respectively. At each level, decomposed signals are downsampled to half of the original sampled signal ($\downarrow 2$) to remove the redundant samples while satisfying Nyquist's criteria [44], [47]. As shown in the block diagram, a four-level DWT-based decomposition finally results in four detail coefficients (CD_1, CD_2, CD_3 and CD_4) and one approximation coefficient (CA_4). The five sub-band signals are then reconstructed through inverse DWT (IDWT) [44],[50]. During the reconstruction process, one coefficient is preserved while the other coefficients are made to be a zero vector. On the use of one coefficient vector in the IDWT reconstruction process, the resulting signal contains frequency components only in that region. Therefore, the reconstructed sub-band speech $s_l(n), l = 1, 2, \dots 5$ represents 4–8 kHz, 2–4 kHz, 1–2 kHz, 0.5–1 kHz, and 0–0.5 Hz frequency band, respectively for a 16kHz sampled speech data.

B. Computation of Proposed DWTR-MFCC Feature

The proposed method for the computation of the DWTR- MFCC feature is depicted in Fig. 2. As shown in the block diagram, the input speech signal $s(n)$ is processed through the following sequence of steps to extract the DWTR-MFCC feature.

1. The speech signal is subject to a pre-emphasis filter with a filter coefficient of 0.97 to boost the high-frequency component. As explained in the previous section, the decomposition and reconstruction of five sub-band speech signals are then performed using the Daubechies (db) wavelet function [45]. Let, the reconstructed sub-band signals are represented by $s_l(n), l = 1, 2, \dots 5$.
2. The short-term analysis of each reconstructed signal $s_l(p)$ is performed by processing with a fixed-length Hamming window of size 20 ms with a frame-shift of 5 ms. The short-term magnitude spectra are then computed by performing DFT on the short-term analysis frames $s_l(p)$, where $p = 1, 2, \dots P$. The total number of analysis frames is represented by P. The short-term DFT magnitude spectra for p^{th} analysis frame of l^{th} sub-band signal is denoted by $S_{p,l}(\omega)$. The nature of short-term magnitude spectra for original and reconstructed sub-band signals are depicted in Fig. 3. This analysis is performed for a center frame of vowel /a/ taken from the dysarthric speaker. The logarithmically compressed magnitude spectrum of the original frame is given in Fig. 3 (a). The logarithmically compressed magnitude spectra obtained for sub-band signals reconstructed using CA_4, CD_4, CD_3, CD_2 and CD_1 are given in Fig. 3(b)-Fig. 3(f), respectively. The magnitude spectra for each sub-band signal are different. Therefore, fine level information can be extracted by analyzing the spectra separately.
3. The Mel-frequency warping of each short-term magnitude spectra $S_{p,l}(\omega)$ is performed using a 30 channel Mel-filters. The size of the Mel-filterbank has remained the same for each sub-band signal $s_l(n)$. The filterbank energies are then computed by following the standard procedure of the MFCC feature extraction. Here, the Mel-filterbank energy for the p^{th} analysis frame of l^{th} sub-band signal is represented by $X_{p,l}(k)$.
4. The Mel-filterbank energies are logarithmically compressed to reduce the dynamic range. For each analysis frame, log compressed filterbank energies obtained across all the sub-band signals are pooled together and discrete cosine transform (DCT) is performed to compute the 13-dimensional base DWTR-MFCC feature.

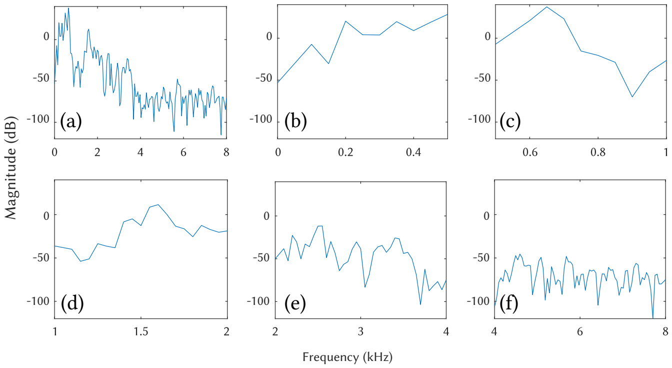

Fig. 3. Plot illustrating the nature of DFT magnitude spectra for sub-band signals. This analysis is performed for a center frame of vowel {a{ taken from the dysarthric speaker M11 of UA-speech corpus. (a) Logarthically compressed magnitude spectrum of the original frame. (b)-(f) Logarithmically compressed magnitude spectra obtained for sub-band signals reconstructed using CA_4, CD_4, CD_3, CD_2 and CD_1, respectively.

III. Experimental Setup

The automatic assessment of dysarthric level is performed using universal access speech (UA-Speech) corpus [51]. This database contains the speech data from 15 dysarthric speakers, which includes 4 female and 11 male speakers. The speech data of each speaker contains isolated words in three different blocks (B_1, B_2, and B_3). Each block contains a total of 255 words out of which 155 words are repeated and the rest 100 words are uncommon for all blocks. The phoneme level diversity is present in the database due to the availability of monosyllables, bisyllables, and polysyllables with the combination of various words. The subjective intelligibility of the speakers varies from 2% to 95%, which is classified into four intelligibility levels, namely very-low (0%-25%), low (26%-50%), mid (51%-75%) and high (76%-100%).

In this study, four datasets namely, the training dataset (*Train_Data*), development dataset (*Dev_Data*), speaker-dependent test dataset (*SD_Test*), and speaker-independent test dataset (*SI_Test*) are derived from the UA-Speech corpus. These datasets are prepared by balancing the recording microphone to remove the sensor effect. The selection of speaker and utterances for each dataset is done as follows:

1. *Train_Data*: This dataset contains the common words from all the blocks and uncommon words from two blocks (B_1 and B_3) of 8 speakers.
2. *Dev_Data*: This dataset contains a part of *Train_Data* equally distributed among dysarthric levels, speakers, and microphones, which is used for the development of universal background model (UBM), total variability matrix (T-matrix) and learning matrices for projection of *i-vectors* to lower dimensional subspace.
3. *SD_Test*: This dataset contains the speech utterances of the 100 uncommon words from block B_2 of 8 speakers present in the *Train_Data*. This dataset is speaker-dependent and text-independent.
4. *SI_Test*: This dataset contains speech data of 100 uncommon words from block B_2 of the remaining 7 speakers. Therefore, this dataset is speaker and text-independent compared to *Train_Data*.

The speakers selected for each dataset is summarized in Table I. All the experimental studies are performed using 16 kHz sampled speech data.

A. Front-End Speech Parameterization

In all the experimental studies, the speech data is firstly pre-emphasized using a high-pass filter with a filter coefficient of 0.97. The short-term analysis of the pre-emphasized speech signal is performed by using an overlapping Hamming window of 20 ms duration at a frame-shift of 5 ms. The performance of dysarthric level assessment system employing proposed DWTR-MFCC feature is compared with MFCC [46], [52], linear prediction cepstral coefficient (LPCC) [53], constant Q cepstral coefficients (CQCCs) [37] and spectral moment time-frequency distribution augmented by low-order cepstra (SMAC) [54] acoustic features. The dimension of base DWTR-MFCC, MFCC, LPCC and CQCC is fixed at 13. As presented in [54], [55], the dimension for the base SMAC feature is fixed at 18, which contains 16 first-order spectral moments extracted using 16-channel Mel-spaced Gabor filterbank and two first-order coefficients of Garbor filtered spectra. To further analyze the impact of shape and size of filterbank on MFCC, the features are extracted using the varied size of Mel and linear filterbank. However, for each case, the 12 dimensional cepstral coefficients along with the energy coefficients are used as the feature vector [28], [46]. For each acoustic feature, the delta (Δ) and delta-delta ($\Delta - \Delta$) coefficients are computed using two preceding and two succeeding feature vectors from the current feature vector. The base feature is then appended to delta (Δ) and delta-delta ($\Delta - \Delta$) features. The feature vectors corresponding to the non-speech regions are removed by processing the speech signal through an energy-based voice activity detection (VAD) [56]. The cepstral mean-variance normalization (CMVN) [57] are then applied to the selected feature vectors to follow a zero mean unit variance distribution.

TABLE I. Speakers Selected for Training Dataset (*Train_Data*), Development Dataset (*Dev_Data*), Speaker-Dependent Test Dataset (*SD_Test*) and Speaker-Independent Test Dataset (*SI_Test*). The Abbreviations F and M Refer to the Female and Male Speakers, Respectively

Intelligibility Level	Train_Data, Dev_Data and SD_Test	SI_Test
Very low(0%-25%)	F03, M04	M12, M01
Low(26%-50%)	F02, M07	M16
Mid(51%-75%)	F04, M05	M11
High(76%-100%)	F05, M08	M09, M10, M14

B. Development of I-Vector Based Dysarthric Assessment System

In the *i-vector* based approach, for the given set of acoustic feature vectors, a lower-dimensional vector of fixed size is created to represent the input speech data [58]–[60]. In this approach, firstly, the class-dependent Gaussian mixture model (GMM) mean supervector is created by adapting a class independent universal background model (UBM). The GMM mean supervector is then projected into a lower-dimensional subspace for mapping the given utterances to a fixed dimension, as proposed in [58]. Equation (6) refers to the *i-vector* representation of a given speech utterance.

$$M = m + Tw \tag{6}$$

where M is the adapted GMM mean supervector, m is the UBM mean supervector, T is the total variability matrix and w is the *i-vector*.

The *i-vectors* extracted from a given speech utterance contains speaker and sensor information along with the information of dysarthria. Therefore, for improving the performance of the dysarthric assessment system, the speaker and channel information need to be normalized. In this study, we have explored the linear discriminant analysis (LDA) [61] and within-class covariance normalization (WCCN) [62] for reducing the session and channel variabilities. The performance of the developed dysarthric assessment system is also evaluated by applying WCCN to the dimensionality reduced *i-vectors* obtained through LDA. The four levels dysarthric assessment is performed by comparing the *i-vectors* of the test speech with the trained *i-vectors* of each dysarthric level. We have performed both cosine kernel [63] and probabilistic linear discriminant analysis (PLDA) [64] based scoring mechanisms. The class representative *i-vectors* for a particular dysarthric level is created by pooling all the speech data corresponding to that class. During testing, the *i-vectors* are extracted from each test data and compared with the trained *i-vector* of each class. The assignment of the test data to a particular class is done based on the maximum score. In this study, UBM model contains 512 Gaussian components, the rank of *i-vector* is fixed at 100 and the LDA dimension is fixed at 10.

IV. Experimental Results and Discussion

The performance of dysarthric level assessment systems is measured using detection accuracy rate (DAR) and also analyzed using confusion matrices.

A. Performance of the Baseline System

In the MFCC feature extraction process [46], the triangular filters are placed in a nonlinear scale to map the frequency bins in Hz to Mel-scale following the human speech perception. The Mel-scale warping emphasizes the lower frequency bins than the higher frequency bins [46]. Consequently, the fine level features those lie in the higher frequency range may not be captured by the MFCC feature. Alternatively, the linearly spaced filterbank provides equal emphasis to all the frequency bins [52], [65]. To study the impact of Mel and linearly spaced filterbank on the cepstral coefficients for the task of dysarthric level assessment, we have extracted the cepstral coefficients by replacing the Mel-frequency triangular filterbank with linearly spaced triangular filterbank in the feature extraction process. The cepstral coefficients extracted using linear filterbank are termed as linear frequency cepstral coefficient (LFCC) [46], [66].

The performances of the dysarthric level assessment system for MFCC and LFCC features on the SD_Test and SI_Test datasets are summarized in Table II. In this study, MFCC and LFCC features are extracted using 40 channel filterbank [28]. For both kinds of features, the DAR observed for SI_Test is less compared to the SD_Test. As mentioned earlier, the *i-vectors* captures the speaker factors along with the information of dysarthria. Since the speakers present in Train_Data and SD_Test are the same, the speaker factor present in the *i-vectors* is normalized up to a great extent. From these preliminary experimental results, it is evident that the cosine kernel-based scoring provides better DAR than the PLDA-based scoring for the SI_Test dataset. The WCCN followed by LDA provides improved DAR than WCCN and LDA-based projection. For both the test datasets, the performance observed for the MFCC feature is better than the LFCC. Therefore, further studies are performed using the Mel-frequency filterbank.

TABLE II. The Performance of the Dysarthric Assessment System Using MFCC and LFCC Features Extracted Using 40 Channel Filterbank. The Performance Is Given in Terms of DAR (in %) for Cosine Kernel and PLDA Based Scoring Schemes

Test Dataset	Feature Type	Cosine Kernel			PLDA
		LDA	WCCN	LDA-WCCN	LDA
SD_Test	MFCC LFCC	87.529 86.933	88.280 87.123	91.436 88.883	91.747 89.209
SI_Test	MFCC LFCC	46.562 43.948	47.719 45.089	48.886 46.969	47.104 45.854

1. Impact of Mel-Filterbank Size on Dysarthria Discrimination of MFCC Feature

As discussed earlier, the information about dysarthria is present at a fine level in the short-term magnitude spectra. Consequently, the dysarthria discrimination of the MFCC feature may be enhanced up to an extent by increasing the size of the Mel-filterbank during the feature computation process. The performance of the dysarthric level assessment system employing the MFCC feature for the varied size of Mel-filterbank is given in Table III. It is evident that the DAR for the SI_Test dataset employing the MFCC feature improves with an increase in the size of the Mel-filterbank and the best DAR is observed when the Mel-filterbank size is 160. On further improving the size of the filterbank, it captures the redundancy present in the magnitude spectra. On the other hand, by increasing the filterbank size, DAR reduces for the SD_Test dataset. *Most of the practical applications demand assessment of dysarthric levels in speaker and text-independent mode*. Therefore, this study emphasizes the SI_Test mode of operation. Further studies on the MFCC feature is presented using 160 channel Mel-filterbank.

TABLE III. The Performance of the Dysarthric Assessment System Employing MFCC Feature for Different Sizes of Mel- Filterbank. The Performance Is Given in Terms of DAR (in %) for Cosine Kernel and PLDA Based Scoring Schemes

Filterbank Size	SD_Test		SI_Test	
	Cosine Kernel	PLDA	Cosine Kernel	PLDA
	LDA-WCCN	LDA	LDA-WCCN	LDA
40	91.436	91.747	48.886	47.104
80	90.247	90.934	52.448	50.489
120	89.622	90.121	54.500	51.875
160	88.090	88.933	56.646	54.563
200	88.895	89.214	55.812	52.677
240	89.064	89.838	53.636	51.864

2. Impact of Cepstral Liftering on Dysarthria Discrimination of MFCC Feature

In the MFCC feature extraction process, cepstral liftering is used to estimate the spectral envelope that represents the resonance structure of the vocal tract system [46], [53]. On the other hand, the cepstral liftering operation smooths out the pitch harmonics [67], [68]. The pitch harmonics may contain information about dysarthria. To capture the effect of pitch harmonics in the MFCC feature, instead of using fixed 13 dimensional base cepstral coefficients, we have studied the performance of the dysarthric level assessment system employing different sizes of base MFCC. In this study, the MFCC feature is extracted using 160 channel Mel-filterbank. The performance of the dysarthric level assessment system for varied dimensions of base MFCC feature is given in Table IV. For both cosine kernel and PLDA based scoring schemes, on increasing the dimension of cepstral coefficients the performance of the dysarthric level assessment system is improved for the SD_Test dataset. On the other hand, for the SI_Test dataset, the best performance is observed for 13 dimensional base MFCC feature. It may be due to modeling speaker information instead of dysarthria. These experimental results show that the optimal performance for the SI_Test dataset is observed for the 13 dimensional base MFCC feature extracted using 160 channel Mel-filterbank.

TABLE IV. The Performance of the Dysarthric Assessment System Employing Varied Dimensions of MFCC Feature Extracted Using 160 Mel-Filterbank. The Performance is Given in Terms of DAR (in %) for Cosine Kernel and PLDA Based Scoring Schemes

Feature Dim.	SD_Test		SI_Test	
	Cosine Kernel	PLDA	Cosine Kernel	PLDA
	LDA-WCCN	LDA	LDA-WCCN	LDA
8	84.995	85.106	51.625	50.048
13	88.090	88.933	56.646	54.563
18	89.967	90.529	48.614	44.761
23	90.935	91.841	44.083	41.281

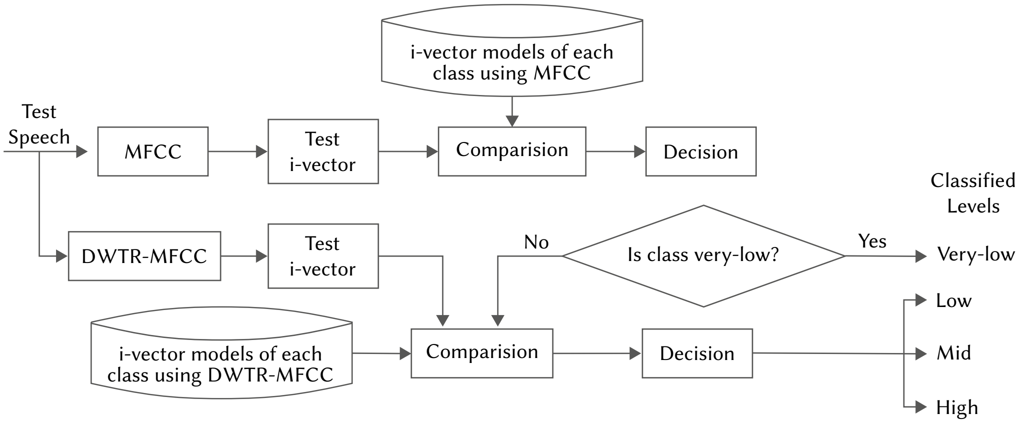

Fig. 4. Block diagram illustrating the proposed scheme for combining the efficacy of the MFCC and DWTR-MFCC feature to improve the performance of the dysarthric level assessment system.

B. Performance of Dysarthric Level Assessment System Using DWTR-MFCC Feature

The performance of the developed dysarthric level assessment system employing the DTWR-MFCC feature is given in Table V. For performance comparison, the DAR obtained for the MFCC with optimal feature parameter selection, LPCC [53], CQCC [37], SMAC [54] features is also given in the Table. 5. Similar to the MFCC, the explored features also provide improved performance for the cosine kernel-based scoring than the PLDA-based scoring scheme. Furthermore, for all the explored features, a reduced DAR is observed for the SI_Test dataset compared to the SD_Test dataset. In the case of the SI_Test dataset, CQCC provides improve performance than LPCC and SMAC features. On the other hand, for the SD_Test dataset, SMAC provides better performance than LPCC and CQCC features. However, the MFCC feature extracted using optimal parameter selection provides better performance than all the explored features for the SI_Test dataset. In the case of the SI_Test dataset, the usage of the proposed DTWR-MFCC feature improves performance compared to the MFCC feature. As discussed earlier, in most of the practical applications, the dysarthric level assessment needs to be done in speaker and text-independent mode. Therefore, the proposed way of computing the cepstral coefficient is more effective than the conventional MFCC feature.

TABLE V. The Performance of the Dysarthric Assessment System is Given for Proposed DWTR-MFCC and Explored Features. The Performance is Given in Terms of DAR (in %) for Cosine Kernel and PLDA Based Scoring Schemes

Filterbank Size	SD_Test		SI_Test	
	Cosine Kernel	PLDA	Cosine Kernel	PLDA
	LDA-WCCN	LDA	LDA-WCCN	LDA
DWTR-MFCC	85.620	86.589	60.094	57.271
MFCC	88.090	88.933	56.646	54.563
LPCC	92.904	93.436	50.323	48.531
CQCC	88.367	88.972	53.989	49.875
SMAC	93.823	94.017	47.458	44.271

For further analysis, we have computed the confusion matrices obtained using the DWTR-MFCC and MFCC features. The confusion matrices obtained for the MFCC and DWTR-MFCC features on the SI_Test dataset are given in Table VI. Table VI (a) and Table VI (b) represents the confusion matrices obtained for cosine kernel scoring on LDA-WCCN projected *i-vectors* for the MFCC and DWTR-MFCC, respectively. Table VI (c) and Table VI (d) represents the PLDA scoring on LDA projected *i-vectors* for MFCC and DWTR-MFCC, respectively.

By comparing the confusion matrices obtained by the DWTR-MFCC feature with the MFCC features given in Table VI, it can be observed that the DWTR-MFCC feature is more discriminating for mid and low dysarthric classes. On the other hand, the MFCC feature is more discriminating for the very-low dysarthric level. Motivated by this observation, next we have studied the possibilities of improving the overall performance of the developed dysarthric level assessment system by employing discriminating power of both kinds of features.

TABLE VI. Confusion Matrices are Given for the *i-vector* Based Dysarthric Level Assessment Systems Using the Proposed DWTR-MFCC Feature. This Analysis is Given on the SI_Test Dataset. (A) and (B) Cosine Kernel Scoring on LDA-WCCN Projected *i-Vectors* for MFCC and DWTR-MFCC, Respectively. (C) and (D) PLDA Scoring on LDA Projected *i-Vectors* for MFCC and DWTR-MFCC, Respectively. The Abbreviation H, M, L, and VL Refer to the High, Mid, Low, and Very-Low Speech Intelligibility Groups, Respectively

(a)

		Predicted class			
		H	M	L	VL
True class	H	83.083	10.084	4.500	2.333
	M	23.750	50.500	19.500	6.250
	L	7.750	43.500	39.500	9.250
	VL	20.750	8.875	16.875	53.500

DAR: 56.646 %

(c)

		Predicted class			
		H	M	L	VL
True class	H	75.500	18.500	4.333	1.667
	M	11.250	68.750	16.500	3.500
	L	2.750	59.750	34.500	3.000
	VL	14.250	26.250	20.000	39.500

DAR: 54.563 %

(b)

		Predicted class			
		H	M	L	VL
True class	H	83.750	9.833	3.250	3.167
	M	18.500	58.250	12.250	11.000
	L	5.500	26.250	61.750	6.500
	VL	23.750	4.750	34.875	36.625

DAR: 60.094 %

(d)

		Predicted class			
		H	M	L	VL
True class	H	70.833	18.083	7.167	3.917
	M	7.750	68.250	16.250	7.750
	L	2.000	34.250	60.000	3.750
	VL	13.000	10.500	46.500	30.000

DAR: 57.271 %

C. Performance of Dysarthric Level Assessment Systems Using Proposed Two-Stage Classification Scheme

To utilize the classification efficacy of the MFCC and DWTR-MFCC features, in this study at the first level, the classification scores obtained for both kinds of features are combined prior to decision. Next, a two-stage classification scheme is proposed to fully utilize the classification efficacy of both kinds of features. The block diagram representation of the proposed two-stage classification scheme is shown in Fig. 4. As shown in the figure, at the first stage a given test utterance is classified employing the MFCC feature. Depending on the maximum score, if the assigned class is a very-low intelligible group then the test data is assigned to that class and the classification process is terminated. On the other hand, if the assigned class is any other intelligible group (high, mid, and low), further classification among these groups is performed employing the dysarthric level assessment system developed using the DWTR-MFCC feature, and the final class assignment is done depending on the maximum score obtained for test data.

TABLE VII. The Performance Dysarthric Level Assessment System is Given for Score Level Combination and the Proposed Combination Scheme. The Performance is Given in Terms of DAR (in %) for Cosine Kernel and PLDA Based Scoring Schemes

Feature Type	SD_Test		SI_Test	
	Cosine Kernel	PLDA	Cosine Kernel	PLDA
	LDA-WCCN	LDA	LDA-WCCN	LDA
MFCC	88.090	88.933	56.646	54.563
DWTR-MFCC	85.620	86.589	60.094	57.271
Score-comb	89.560	90.465	61.261	58.823
Proposed	87.122	88.026	65.813	60.750

TABLE VIII. Confusion Matrices are Given for the *I-Vector* Based Dysarthric Level Assessment Systems Using Score Combination and Proposed Approach, Respectively. This Analysis is Given on the SI_Test Dataset. (A) and (B) Cosine Kernel Scoring on LDA-WCCN Projected *I-Vectors* for Score Combination and Proposed Method, Respectively. (C) and (D) PLDA Scoring on LDA Projected I-Vectors for Score Combination and Proposed Method, Respectively. The Abbreviation H, M, L, and VL Refer to the High, Mid, Low, and Very-Low Speech Intelligibility Groups, Respectively

(a)

		Predicted class			
		H	M	L	VL
True class	H	88.417	6.833	2.750	2.000
	M	19.750	57.000	15.750	7.500
	L	5.000	34.500	52.250	8.250
	VL	22.375	6.250	24.000	47.375

DAR: 61.261 %

(c)

		Predicted class			
		H	M	L	VL
True class	H	76.917	16.167	5.083	1.833
	M	7.500	75.000	12.750	4.750
	L	2.250	47.250	46.000	4.500
	VL	14.125	18.750	29.750	37.375

DAR: 58.823 %

(b)

		Predicted class			
		H	M	L	VL
True class	H	86.000	8.167	3.500	2.333
	M	16.500	63.000	14.250	6.250
	L	4.250	25.750	60.750	9.250
	VL	18.625	3.250	24.625	53.500

DAR: 65.813%

(d)

		Predicted class			
		H	M	L	VL
True class	H	73.250	17.250	7.833	1.667
	M	7.750	70.250	18.500	3.500
	L	2.000	35.000	60.000	3.000
	VL	11.000	9.250	40.250	39.500

DAR: 60.750%

The DAR obtained using the score level combination and proposed two-stage classification scheme is summarized in Table VII. The score level combination improves the DAR compared to individual features for both cosine kernel and PLDA based scoring methods. The DAR is further improved on the use of the proposed two-stage classification scheme. For cosine kernel-based scoring on LDA-WCCN projected *i-vectors*, the DAR improved to 65.813% in case of the SI_Test dataset. To further analyze the merits of the proposed two-stage classification scheme, the confusion matrices obtained for the proposed classification scheme are compared with the score level combination. Table VIII (a) and Table VIII (b) represents the confusion matrices obtained for cosine kernel scoring on LDA-WCCN projected *i-vectors* for score level combination and proposed approach, respectively. Table VIII (c) and Table VIII (d) represents the PLDA scoring on LDA projected *i-vectors* for score level combination and proposed method, respectively. By comparing the confusion matrices obtained for both approaches, it can be observed that in case of the proposed two-stage classification method the classification accuracy of each class is improved. This is mainly due to the complete utilization of the discrimination power of both kinds of the feature. This experimental result also shows that the proposed DWTR-MFCC carries additional information than MFCC.

V. Conclusion

The work presented in this paper aims at improving the performance of an automatic dysarthric level assessment system by capturing the fine level information present in different sub-bands of the speech signal. To capture the fine level information, firstly, the performance of the dysarthric level assessment system employing the MFCC feature is studied by varying the shape and size of the triangular filterbank. Next, for a better representation of sub-band information, the input speech signal is decomposed into four levels using DWT decomposition. For each input speech signal, five speech signals representing different sub-bands are then reconstructed using IDWT. The log filterbank energies are computed by analyzing the DFT magnitude spectra of each reconstructed speech using a 30-channel Mel-filterbank. For each analysis frame, the log filterbank energies obtained across all reconstructed speech are combined, and DCT is performed to represent the cepstral feature, termed as DWTR-MFCC in this study. The performance of *i-vector* based four-level dysarthric assessment system on the UA-Speech corpus shows that the overall performance of the system employing the MFCC feature improves by increasing the size of Mel-filterbank. However, a large overlapping between mid and low dysarthric levels is observed. On the use of the DWTR-MFCC feature, performance of the developed dysarthric level assessment is further improved by reducing the overlapping between mid and low dysarthric levels. But, reduced classification accuracy is observed for very-low dysarthric levels due to miss classification of very-low dysarthric level to low dysarthric level. Motivated by these observations, finally, a two-stage classification approach is proposed by employing the efficacy of MFCC and DWTR-MFCC features. The proposed approach improves the classification accuracy of the developed dysarthric level assessment system by reducing the overlapping between any two classes without loss of performance for individual features (MFCC or DWTR-MFCC).

References

[1] J. R. Duffy, *Motor speech disorders e-book: Substrates, differential diagnosis, and management.* Elsevier Health Sciences, 2019.

[2] R. Sandyk, "Resolution of dysarthria in multiple sclerosis by treatment with weak electromagnetic fields," *International Journal of Neuroscience*, vol. 83, no. 1-2, pp. 81–92, 1995.

[3] J. Müller, G. K. Wenning, M. Verny, A. McKee, K. R. Chaudhuri, K. Jellinger, W. Poewe, I. Litvan, "Progression of dysarthria and dysphagia in postmortem-confirmed parkinsonian disorders," *Archives of neurology*, vol. 58, no. 2, pp. 259–264, 2001.

[4] S. Skodda, H. Rinsche, U. Schlegel, "Progression of dysprosody in Parkinson's disease over time—a longitudinal study," *Movement disorders: official journal of the Movement Disorder Society*, vol. 24, no. 5, pp. 716–722, 2009.

[5] J. B. Polikoff, H. T. Bunnell, "The nemours database of dysarthric speech: A perceptual analysis," in *Proc. ICPS*, 1999, pp. 783–786.

[6] R. D. Kent, "Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders," *American Journal of Speech-Language Pathology*, vol. 5, no. 3, pp. 7–23, 1996.

[7] G. Constantinescu, D. Theodoros, T. Russell, E. Ward, S. Wilson, R. Wootton, "Assessing disordered speech and voice in Parkinson's disease: a telerehabilitation application," *International journal of language & communication disorders*, vol. 45, no. 6, pp. 630–644, 2010.

[8] K. K. Baker, L. O. Ramig, E. S. Luschei, M. E. Smith, "Thyroarytenoid muscle activity associated with hypophonia in parkinson disease and

aging," *Neurology*, vol. 51, no. 6, pp. 1592–1598, 1998.

[9] S. Skodda, W. Visser, U. Schlegel, "Vowel articulation in Parkinson's disease," *Journal of Voice*, vol. 25, no. 4, pp. 467–472, 2011.

[10] A. Maier, T. Haderlein, F. Stelzle, E. Nöth, E. Nkenke, F. Rosanowski, A. Schützenberger, M. Schuster, "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer,"*EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–7, 2009.

[11] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, E. Nöth, "Peaks–a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[12] K. Riedhammer, G. Stemmer, T. Haderlein, M. Schuster, F. Rosanowski, E. Noth, A. Maier, "Towards robust automatic evaluation of pathologic telephone speech," in *Proc. ASRU (Workshop)*, 2007, pp. 717–722.

[13] M. J. Kim, Y. Kim, H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically- structured sparse linear model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 694–704, 2015.

[14] C. Deng, G. Lai, H. Deng, "Improving word vector model with part-of-speech and dependency grammar information," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 4, pp. 276–282, 2020.

[15] F. Rudzicz, "Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech," in *Proc. International ACM SIGACCESS on Computers and Accessibility*, 2007, pp. 255–256.

[16] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *Proc. ICASSP*, 2009, pp. 4605–4608.

[17] K. L. Kadi, S. A. Selouani, B. Boudraa, M. Boudraa, "Automated diagnosis and assessment of dysarthric speech using relevant prosodic features," in *Transactions on Engineering Technologies*, 2014, pp. 529– 542.

[18] C. Bhat, B. Vachhani, S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *Proc. ICASSP*, 2017, pp. 5070–5074.

[19] M. Perez, W. Jin, D. Le, N. Carlozzi, P. Dayalu, A. Roberts, E. M. Provost, "Classification of huntington disease using acoustic and lexical features," in *Proc. INTERSPEECH*, 2018, pp. 1898–1902.

[20] N. Saleem, M. I. Khattak, "Deep neural networks for speech enhancement in complex-noisy environments," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 84–90, 2020.

[21] N. Saleem, M. I. Khattak, E. Verdú, "On improvement of speech intelligibility and quality: A survey of unsupervised single channel speech enhancement algorithms," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 2, pp. 78–89, 2020.

[22] B. Yang, R. Ding, Y. Ban, X. Li, H. Liu, "Enhancing direct-path relative transfer function using deep neural network for robust sound source localization," *CAAI Transactions on Intelligence Technology*, pp. 1–9, 2021.

[23] M. J. Kim, B. Cao, K. An, J. Wang, "Dysarthric speech recognition using convolutional LSTM neural network," in *Proc. INTERSPEECH*, 2018, pp. 2948–2952.

[24] H. Liu, P. Yuan, B. Yang, G. Yang, Y. Chen, "Head-related transfer function–reserved time- frequency masking for robust binaural sound source localization," *CAAI Transactions on Intelligence Technology*, pp. 26–33, 2021.

[25] A. A. Joshy, R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in *Proc. European Signal Processing Conference*, 2021, pp. 116–120.

[26] C. Bhat, H. Strik, "Automatic assessment of sentence- level dysarthria intelligibility using BLSTM," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 322–330, 2020.

[27] K. Gurugubelli, A. K. Vuppala, "Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment," in *Proc. ICASSP*, 2019, pp. 6410–6414.

[28] K. Gurugubelli, A. K. Vuppala, "Analytic phase features for dysarthric speech detection and intelligibility assessment," *Speech Communication*, vol. 121, pp. 1–15, 2020.

[29] S.-A. Selouani, H. Dahmani, R. Amami, H. Hamam, "Using speech rhythm knowledge to improve dysarthric speech recognition," *International Journal of Speech Technology*, vol. 15, no. 1, pp. 57–64, 2012.

[30] J. Kim, N. Kumar, A. Tsiartas, M. Li, S. S. Narayanan, "Automatic intelligibility classification of sentence- level pathological speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 132–144, 2015.

[31] K. Kadi, S. A. Selouani, B. Boudraa, M. Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge," *Biocybernetics and Biomedical Engineering*, vol. 36, no. 1, pp. 233–247, 2016.

[32] A. Benba, A. Jilbab, A. Hammouch, "Detecting patients with Parkinson's disease using Mel frequency cepstral coefficients and support vector machines," *International Journal on Electrical Engineering and Informatics*, vol. 7, no. 2, pp. 297–307, 2015.

[33] H. Chandrashekar, V. Karjigi, 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.

[34] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, A. Miguel, "Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," *ACM Transactions on Accessible Computing*, vol. 6, no. 3, pp. 1–21, 2015.

[35] A. Benba, A. Jilbab, A. Hammouch, "Discriminating between patients with Parkinson's and neurological diseases using cepstral analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 10, pp. 1100–1108, 2016.

[36] S. Oue, R. Marxer, F. Rudzicz, "Automatic dysfluency detection in dysarthric speech using deep belief networks," in *Proc. SLPAT*, 2015, pp. 60–64.

[37] J. C. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[38] M. Little, P. McSharry, E. Hunter, J. Spielman, L. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *Nature Precedings*, pp. 1–27, 2008.

[39] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[40] I. Kodrasi, H. Bourlard, "Super-gaussianity of speech spectral coefficients as a potential biomarker for dysarthric speech detection," in *Proc. ICASSP*, 2019, pp. 6400–6404.

[41] N. Narendra, P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences," in *Proc. INTERSPEECH*, 2018, pp. 3403–3407.

[42] N. Narendra, P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Communication*, vol. 110, pp. 47–55, 2019.

[43] N. Narendra, P. Alku, "Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features," *Computer Speech & Language*, vol. 65, pp. 1–14, 2021.

[44] S. G. Mallat, "A theory for multiresolution signal decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–693, 1989.

[45] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Transactions on information theory*, vol. 36, no. 5, pp. 961–1005, 1990.

[46] S. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[47] P. Singh, G. Pradhan, S. Shahnawazuddin, "Denoising of ecg signal by non-local estimation of approximation coefficients in dwt," *Biocybernetics and Biomedical Engineering*, vol. 37, no. 3, pp. 599–610, 2017.

[48] H. Ayad, M. Khalil, "Qam-dwt-svd based watermarking scheme for medical images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 3, pp. 81–89, 2018.

[49] R. Abbasi, M. Esmaeilpour, "Selecting statistical characteristics of brain signals to detect epileptic seizures using discrete wavelet transform and perceptron neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 5, pp. 33–38, 2017.

[50] A. K. H. Al-Ali, D. Dean, B. Senadji, V. Chandran, G. R. Naik, "Enhanced forensic speaker verification using a combination of dwt and mfcc feature warping in the presence of noise and reverberation conditions," *IEEE Access*, vol. 5, pp. 15400–15413, 2017.

[51] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang,

K. Watkin, S. Frame, "Dysarthric speech database for universal access research," in *Proc. INTERSPEECH*, 2008, pp. 1741– 1744.

[52] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy- Wilson, S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Proc. ASRU (Workshop)*, 2011, pp. 559–564.

[53] L. R. Rabiner, R. W. Schafer, "Introduction to digital speech processing," *Foundations and Trends in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.

[54] P. Tsiakoulis, A. Potamianos, D. Dimitriadis, "Spectral moment features augmented by low order cepstral coefficients for robust asr," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 551–554, 2010.

[55] K. Maity, G. Pradhan, J. P. Singh, "A pitch and noise robust keyword spotting system using smac features with prosody modification," *Circuits, Systems, and Signal Processing*, vol. 40, no. 4, pp. 1892–1904, 2021.

[56] J. G. Wilpon, L. R. Rabiner, T. Martin, "An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints," *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 3, pp. 479–498, 1984.

[57] O. Viikki, K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[58] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[59] D. Garcia-Romero, C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. INTERSPEECH*, 2011, pp. 249–252.

[60] G. Pradhan, S. R. M. Prasanna, "Speaker verification by vowel and nonvowel like segmentation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 854–867, 2013.

[61] S. Balakrishnama, A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, pp. 1–8, 1998.

[62] A. O. Hatch, S. S. Kajarekar, A. Stolcke, "Within- class covariance normalization for svm-based speaker recognition.," in *Proc. INTERSPEECH*, 2006, pp. 1471– 1474.

[63] N. Dehak, R. Dehak, J. R. Glass, D. A. Reynolds, P. Kenny, *et al.*, "Cosine similarity scoring without score normalization techniques.," in *Proc. Odyssey*, 2010, pp. 1–5.

[64] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Proc. ICASSP*, 2013, pp. 7649–7653.

[65] Y. Jung, Y. Kim, H. Lim, H. Kim, "Linear-scale filterbank for deep neural network-based voice activity detection," in *Proc. O-COCOSDA*, 2017, pp. 1–5.

[66] S. Debnath, P. Roy, "Audio-visual automatic speech recognition using pzm, mfcc and statistical analysis," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 121–133, 2021.

[67] B. Pattanayak, G. Pradhan, "Pitch-robust acoustic feature using single frequency filtering for children's kws," *Pattern Recognition Letters*, vol. 150, pp. 183–188, 2021.

[68] J. G. Wilpon, C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. ICASSP*, vol. 1, 1996, pp. 349–352.

Laxmi Priya Sahu

Laxmi Priya Sahu received her B.Tech. degree in Electronics and Telecommunication Engineering from Biju Patnaik University of Technology (BPUT), Odisha, India, in 2012 and M.Tech. degree in Communication System Engineering from KIIT University, Odisha, India, in 2015. She is pursuing Ph.D. in the Department of Electronics and Communication Engineering at National Institute of Technology Patna, India. Her research interests are speech signal analysis, and biomedical signal processing.

Gayadhar Pradhan

Gayadhar Pradhan received his M.Tech. and Ph.D. degrees in Electronics and Electrical Engineering from Indian Institute of Technology Guwahati, India, in 2009 and 2013, respectively. He is currently working as Associate professor in the Department of Electronics and Communication Engineering at National Institute of Technology Patna, India. His research interests are speech signal processing, speaker recognition and speech recognition.

Jyoti Prakash Singh

Jyoti Prakash Singh is an assistant professor and the Department Head of Computer Science and Engineering at the National Institute of Technology Patna in India. He has co-authored seven textbooks and one edited book with McGraw Hill, Elsevier, and Springer, among others. He has over 45 international journal publications in leading publishers, as well as over 50 international conference proceedings. He was a co-investigator in an MietY-sponsored project to develop algorithms for spam/fake calls in telephone conversations. His research interests include social media mining, deep learning, data security, and speech processing. He is an Associate Editor for the International Journal of Electronic Government Research. In 2020, he received the S4DS Data Scientist (Academia) Award from the Society for Data Science.