Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques

Amlu Anna Joshy and Rajeev Rajan, *Member, IEEE*

***Abstract*— Assessing the severity level of dysarthria can provide an insight into the patient's improvement, assist pathologists to plan therapy, and aid automatic dysarthric speech recognition systems. In this article, we present a comparative study on the classification of dysarthria severity levels using different deep learning techniques and acoustic features. First, we evaluate the basic architectural choices such as deep neural network (DNN), convolutional neural network, gated recurrent units and long short-term memory network using the basic speech features, namely, Mel-frequency cepstral coefficients (MFCCs) and constant-Q cepstral coefficients. Next, speech-disorder specific features computed from prosody, articulation, phonation and glottal functioning are evaluated on DNN models. Finally, we explore the utility of low-dimensional feature representation using subspace modeling to give i-vectors, which are then classified using DNN models. Evaluation is done using the standard UA-Speech and TORGO databases. By giving an accuracy of 93.97% under the speaker-dependent scenario and 49.22% under the speaker-independent scenario for the UA-Speech database, the DNN classifier using MFCC-based i-vectors outperforms other systems.**

***Index Terms*— Deep learning, dysarthria, i-vectors, severity assessment.**

I. Introduction

DYSARTHRIA is a motor speech disorder caused due to the poor coordination, or malfunction of speech production subsystems. It is either acquired from a neurological injury such as cerebral palsy, or developed along with any neuro-degenerative disease [1]. It leads to imprecise articulation, low audibility, atypical prosody and variable speech rate, which deteriorate the speech quality. The patients would have hyper-nasality, weak facial reflexes, harsh voice quality, and increased fatigue on speaking. Thus, even though syntactically correct sentences can be formulated, they cannot be phonetically produced, or correctly pronounced by dysarthric patients. This leads to incomprehensible speech and affects their social life. Dysarthric patients have physical incapacities such as trembling hands, due to the weak coordination of muscles, which make the use of a keyboard or joystick-based interactive applications less useful for their communication purposes. Hence, automatic speech recognizer (ASR)-based applications would be helpful to them. However, normal ASRs designed for healthy speakers have high error rates when used by dysarthric speakers, due to the poor articulations.

Manuscript received June 10, 2021; revised November 4, 2021, January 25, 2022, and March 30, 2022; accepted April 17, 2022. Date of publication April 22, 2022; date of current version May 4, 2022. *(Corresponding author: Amlu Anna Joshy.)*

Amlu Anna Joshy is with the Electronics and Communication Engineering Department, College of Engineering Trivandrum, APJ Abdul Kalam Technological University, Thiruvananthapuram 695016, India (e-mail: amluanna02@gmail.com).

Rajeev Rajan was with the Speech and Music Technology Laboratory, Department of Computer Science and Engineering, IIT Madras, Chennai 600036, India. He is now with the Electronics and Communication Engineering Department, College of Engineering Trivandrum, APJ Abdul Kalam Technological University, Thiruvananthapuram 695016, India (e-mail: rajeev@cet.ac.in).

Digital Object Identifier 10.1109/TNSRE.2022.3169814

A. Motivation and Related Work

Dysarthria severity estimation is important in its diagnosis to evaluate the progression of the disease, to identify proper medication and to conduct required speech therapy sessions. Severity assessment can be done using objective (acoustic and physiological measures) and subjective (perceptual) measures by a trained speech-language pathologist (SLP) [2]. Perceptual evaluation would be inconsistent due to the familiarity with the patient and vary across clinicians with experience and listening skills. Also, this could be time-consuming and expensive, hence limited in remote rehabilitation. However, it is important to keep track of the clients with dysarthria during rehabilitation. This paves the way for the need for an automatic dysarthria severity level classification system. This classification can also improve the performance of ASR systems built for dysarthric patients, as evident in [3].

To mimic the human-auditory system, many perceptual features are used in speech processing such as the linear prediction coefficients, linear prediction cepstral coefficients, Mel-frequency cepstral coefficients (MFCCs), constant Q cepstral coefficients (CQCCs), perceptual linear prediction coefficients and the relative spectra coefficients. In literature, MFCCs have proven their usefulness for modeling pathological speech signals in general [4], and for dysarthria severity identification in specific [5], [6], with machine learning classifiers. In [5], MFCCs show their efficiency over log filter banks, and a comparable performance to i-vectors when used for sentence-level dysarthric speech detection from healthy speech. MFCCs are encoded using a deep belief network

(DBN), and employed for dysarthria severity classification using a multilayer perceptron (MLP) in [7]. DBN features have only a marginal improvement over the MFCC based system. The glottal features used in [8] were outperformed by the baseline OpenSMILE-1 feature set, which included MFCCs on considering isolated words. In [9] it is seen that MFCCs together with glottal-to-noise excitation ratio (GNER), and harmonics-to-noise ratio (HNR) outperforms this subset for spastic dysarthria severity classification. These works motivated us to use the basic MFCC features to investigate the performance of various deep learning models for dysarthric severity estimation, to see if significant improvements can be achieved over the machine learning classifiers. CQCCs were introduced in the context of spoofing detection in [10] and have proven to be an excellent choice for speaker recognition in recent years. The authors of [11] have shown that the strength of formants and harmonics in the constant Q transform (CQT) spectrograms decreases as the intelligibility level decreases, and thus demonstrates the efficiency of CQT in dysarthria severity identification. CQCCs have also demonstrated good results in [12], when used as baseline features. We have been motivated by these results to analyse the capability of CQCCs for the proposed task. There are many other existing and novel features explored in literature for improving the accuracy of dysarthria severity identification, such as [13] using the breathiness indices, [12] introducing the perceptually enhanced single frequency filtering based cepstral coefficients (PE-SFCC) and [14] using audio descriptors. However, we wanted to keep our study more focused on analysing the different deep learning classifiers, and hence avoided such complex representations.

Speech disorder specific prosodic features like spectral moments, formants, skewness and MFCCs, are selected by a genetic algorithm and used with a support vector machine (SVM) classifier in [6] for dysarthria speech diagnosis and severity identification. In terms of mean pitch, jitter, shimmer, proportion of the vocalic duration and degree of voiced breaks, prosodic features have been used for dysarthria severity estimation using the linear discriminant analysis combined with Gaussian mixture model (GMM) and SVM in [15] on the Nemours database. Glottal flow patterns have been analysed for studying Parkinson's disease (PD) in [16] and for detecting dysarthric speech from healthy ones in [8] using SVM classifiers. Articulatory features have been used to analyse different motor speech disorders in [17] and [18]. All these works put light on the fact that the speech disorder specific features in terms of prosodic, glottal and articulatory measures are relevant in identifying the dysarthric speech patterns, and hence, we extend our study on dysarthria severity classification using them with deep neural network (DNN) models.

The i-vector subspace modeling has proven to capture many aspects of a person's speech, including gender, age and intelligibility, which makes them efficient in speaker, language and accent recognition [19]. In [20], i-vectors are used with a v-support vector regression predictor for dysarthric speech intelligibility assessment. Perceptual linear prediction features are used for acoustic parameterisation, and a Pearson correlation of 0.9 was obtained. In [12], the i-vectors modelled from PE-SFCC features are used with the probabilistic linear discriminant analysis (PLDA) scoring mechanism for the detection of dysarthric speech from healthy samples, followed by severity estimation. Thus i-vectors are capable of modeling dysarthric severity levels as well, and based on this understanding DNN models are used with them.

DNN has reported excellent results outperforming the conventionally used techniques in speech emotion recognition, speaker recognition and end-to-end speaker verification (SV) system. Deep learning models like DNN, CNN, time delay neural network, and long short term memory (LSTM) network are explored for dysarthric ASR on the TORGO database in [21]. Joint spectro-temporal features from mel-scale spectrogram are used in [2] for dysarthria severity estimation. They showed that the time-frequency CNN which jointly captures spectral and temporal information is superior to the time/frequency CNN which captures either temporal or spectral information and not both. Residual neural networks are used similarly in [22]. LSTM and DNN have been used with lexical and acoustic features to differentiate patients with Huntington disease from healthy ones in [23]. Inspired by these successful implementations, these deep learning architectures have been employed in this work for dysarthric severity classification. The importance of building a speaker-independent (SID) dysarthria intelligibility assessment system is explained by the authors of [24] recently. They have put forth the novel idea of using features obtained from DeepSpeech, an end-to-end Speech-to-Text engine. We have been inspired by these works to do a detailed evaluation of our models under both speaker-dependent (SD) and SID scenarios.

B. Contribution

Our major contributions can be summed up as,

- Performance analysis of the basic deep learning architectures namely, DNN, CNN, gated recurrent units (GRU), and LSTM using MFCCs and CQCCs. Our initial phase of work using MFCCs is reported in [25].
- Assessment of prosodic, glottal, phonetic, and articulatory features on DNN classifiers. Further, dimensionality reduction is performed on the concatenated feature set, and results are analysed.
- Implementation of a 'two-level learning classifier' using i-vector sub-space modeling as the first level, and DNN based classification as the second level.
- Experimentation using the round-robin leave-one-speaker-out (LOSO) cross-validation, to yield SID models.

The rest of the paper is organized as follows. Section II details the databases used in the work, and the experimental design is described in section III. The features and classifiers used are explained in sections IV and V respectively. Results and discussion are given in section VI, and the work is finally concluded in Section VII.

TABLE I
CLASS-WISE PATIENT DESCRIPTION

Severity	*UA-Speech*	*TORGO*
VERY LOW	F05, MO8, M09, M10, M14	F03, F04, M03
LOW	F04, MO5, M11	F01,M05
MEDIUM	F02, M07, M16	M01, M02, M04
HIGH	F03, MO1, M04, M12	-

II. DATABASES

The standard American English dysarthric databases, namely (a) the TORGO [26] database and (b) the Universal Access dysarthric speech corpus (UA-Speech) [27] are used for evaluating the proposed work. The TORGO database comprises of the aligned acoustics and measured 3D articulatory features of utterances from 7 healthy speakers and 8 dysarthric patients. The corpus consists of non-words, words, restricted and unrestricted sentences, of which only the words are used in this study. These consist of the English digits, the international radio alphabets, the twenty most frequent words in the British national corpus, 50 words from the intelligibility section of the Frenchay dysarthria assessment (FDA), and phonetically contrasting pairs of words, as explained in [26]. The recordings of the head-mounted microphone, at a sampling frequency (f_s) of 16 kHz are used. While using this database, 80% of the data is used for training and the rest for testing.

UA-Speech comprises of speech from 13 healthy speakers and 19 dysarthric patients. However, data of only 15 patients are available. There are 155 common words repeated thrice, corresponding to the English digits, computer commands, international radio alphabets, and 100 common words in the Brown corpus. These 465 common words per speaker constitute the training data, which sums up to 6975 utterances. The corpus also has 300 distinct uncommon words per speaker, which are selected from children's novels digitized by Project Gutenberg, using a greedy algorithm to maximize phone-sequence diversity [27]. These are used for testing (4500 unseen words in total) so as to evaluate the robustness of the models. Data from the sixth channel in the microphone array, at $f_s = 16$ kHz, was used, as it had the highest signal-to-noise ratio. The severity levels are assigned as very low, low, medium and high, based on intelligibility as reported by five naive listeners for UA-Speech. For TORGO these are assigned by an SLP in terms of the clinical intelligibility and articulatory functionality, as per FDA. The intelligibility rating of each severity level is as follows: (0-25)%- high, (25-50)%-medium, (50-75)%-low and (75-100)%-very low. Description of the databases is given in Table I.

III. EXPERIMENTAL DESIGN

The elaborated study on dysarthria severity level determination can be explained as three independent experiments.

A. Analysing MFCCs and CQCCs

Vocal muscular coordination influences the speech intelligibility, and MFCCs can capture the irregular vocal fold movements or the lack of vocal-fold closure due to mass tissue changes [4]. CQCCs are obtained as a result of coupling between CQT and the traditional cepstral analysis. They are more closely related to the human perception system, by giving a higher frequency resolution at lower frequencies and higher temporal resolution at higher frequencies. With these understandings, we perform the first experiment (E1), where the basic deep learning strategies, namely DNN, CNN, GRU and LSTM are employed for classification, with MFCCs [25] and CQCCs as features.

B. Analysing Speech Disorder Specific Features

In the second experiment (E2), glottal, articulatory, phonetic and prosodic features are used with DNN. Their efficiency in highlighting the paralinguistic aspects of speech is analysed. Irregular glottal closure pattern and related breathy voice are one of the most evident symptoms observed by clinicians when diagnosing dysarthria. Articulatory features explain the retardation in the lip, tongue, and jaw movements, and model the imprecise articulations in dysarthric speech. The voice quality in terms of stability and periodicity are deteriorated in dysarthric patients and can be explained through the variations in the phonation [28]. Hence, phonetic features relating to perturbation, noise content, and non-linear dynamics [17] are extracted in this regard. The abnormal changes in pitch, loudness and time duration in the dysarthric speech contribute greatly to the detection and analysis of dysarthria. These abnormalities prevent conveying the right emotion and rhythm to the speech, and can be quantified by the prosodic features.

In this work, 36 glottal, 488 articulatory, 28 phonetic and 103 prosodic features are used with DNN. They are concatenated and dimensionality reduction is applied to yield a better interpretable representation. The statistical and unsupervised, factor analysis (FA) technique is used for this. Thus, from the concatenated feature set, factors are created to represent the common variance or correlation among them. A compact and contented description of the multi-variate data can be obtained using this technique. It can be considered an extended and elaborated form of the principal component analysis technique for dimensionality reduction [29]. Thus, the feature vector of dimension 655 is reduced to 200 for the proposed task. Generally, FA is used with machine learning classifiers for the selection of the best features by avoiding redundant representations. This would reduce the complexity imposed on the classifiers and generally improve their performance. However, the deep learning models are inbuilt with the property of feature selection and abstract representations that encapsulates all the non-linearities within. Hence, the dimensionality reduction techniques are not generally used along. But we explore if some improvement can be achieved with this addition.

C. Analysing i-Vectors

I-vector is a new FA front-end approach to SV [30] which maps the high dimensional GMM supervector space to a single total-variability space, unlike the separate speaker and channel dependent subspaces of the conventional FA technique. This total-variability space is a low-dimension subspace which holds the main variabilities describing the data (variabilities

in noise, channel, speaker, age, or the intelligibility characteristics as embedded phonetic contents in the utterance). It is trained with the FA modeling and yields i-vectors which are very suitable to be used with simple predictors [31]. Typically i-vectors are used with the PLDA scoring mechanism or the SVM classifier for SV. DNNs have been used to bring many alterations to the conventional i-vector-PLDA systems for speaker recognition, namely, to provide bottleneck features in the front-end [32], to compute sufficient statistics [33], and to discriminatively model the target and impostor i-vectors [34]. Severity level identification is basically a speaker-group recognition task. We hypothesise that the severity dependent factors encoded in the class-wise computed i-vectors can be efficiently categorised by DNNs. DBNs were trained with MFCCs, and their bottleneck features were used with an MLP for dysarthric speaker recognition in [7], and showed improved results over MFCCs. A similar approach is adopted here, wherein, we do a two-level learning using i-vector subspace modeling as the first level followed by DNN-based classification. Hence, our work moulds severity level dependent i-vectors, and builds DNN classifiers to discriminate them. I-vectors using MFCC (i_{MFCC}) and CQCC (i_{MFCC}) are extracted in this regard, and fed to the DNNs.

IV. Feature Design

13-dimensional MFCCs and their first two derivatives are computed for each 30 ms frame, with a frame-shift of 10 ms. The number of frames is set to an approximate average number of frames over the full utterances in the dataset, which is 180 for TORGO and 400 for UA-Speech. Smaller utterances are zero padded and the larger ones are trimmed. This 2D MFCC array is transposed and arithmetic mean along its horizontal axis is calculated to give the frame-wise averaged 1D features for DNN. For the CNN, GRU and LSTM models, MFCCs are fed frame-wise and derivatives are not used. These networks are capable of learning the temporal information by themselves, so adding the deltas would add redundancy and may highlight the irrelevant speech characteristics, such as emotion, gender, and age. MFCC extraction was done using the librosa python package [35]. Similarly, CQCCs are extracted as explained in [10], using an open-source Matlab implementation as in [36]. The bandwidth is limited to 100Hz - 8kHz ($f_s/2$), with the number of bins per octave set to 48. Re-sampling is done at a sampling period of 16.

For extracting i-vectors, first, the acoustic parameterisation is done using frame-wise 13-dimensional MFCCs and CQCCs, and their first two deltas. Then, the UBM is trained using the expectation-maximization (EM) algorithm in 10 iterations using the auxiliary database comprising healthy audio samples of UA-Speech. The total-variability (TV) matrix is modeled using the sufficient statistics from the databases under study, and the i-vectors are computed. Target (dysarthric) GMM is adapted from the UBM using the eigen-voice adaption method. The target GMM supervector (M) is formulated as a shifted version of UBM, and is given by:

$$M = m + Tw \tag{1}$$

where m represents the UBM supervector, T is a low dimensional rectangular TV matrix, and w is the resulting i-vector. In the E-step of the EM algorithm, w is considered as a latent variable with normal prior distribution $N(0, I)$. Eventually, the i-vectors will be estimated as the mean of the posterior distribution of w, as given in [30]. The extraction was done using the ALIZÉ tool-kit. Performance of the different classifiers is analysed by varying the number of mixtures or Gaussian components (N_g) used in building the UBM (128, 256, 512), and dimension (N_{iv}) of the T matrix used for i-vector extraction (100, 200, 400, 600).

The speech disorder specific features are extracted using the DisVoice python library[1] and the Kaldi toolkit. Only the static (utterance level) features are computed, and the fundamental frequency computation was done using the PRAAT algorithm. 103 prosody features based on duration, fundamental frequency and energy are computed as explained in [18]. They define the pitch and energy contours which discriminate the different severity levels of dysarthria. The articulatory feature vector is extracted by applying the four statistical functions, namely, the mean, standard deviation, skewness, and kurtosis, on the 122 descriptors, which include the bark band energies, formants and MFCCs during the onset and offset transitions, totalling 488 features per utterance. Glottal inverse filtering technique, specifically, iterative and/or adaptive inverse filtering is employed to give the glottal flow patterns, which are then characterised by nine different time-frequency parameters as in [16]. Then the statistical measures are applied to give 36 features per utterance. Seven measures of phonation corresponding to the jitter and shimmer, amplitude and pitch perturbation quotients, GNER, HNR, cepstral harmonics to noise ratio, and the normalized noise energy are calculated, as explained in [28], and the statistical functions are applied, giving 28 features per utterance. Thus, on concatenating we get a feature vector of dimension 655.

V. Classifier Design

Implementation details of the baseline classifiers and the different deep learning architectures used are briefed below.

A. Baseline Classifiers

Machine learning classifiers, SVM and random forest (RF) are built as the baseline classifiers. SVM was built for both linear and radial basis function (RBF) kernels, with the optimal regularisation parameter, c being tuned from 1 to 10. For E1 and E2, $c = 6$ and for E3, $c = 1$ performed the best on the validation data (20% of training data). The RBF-based classifiers reported poorer results, and hence not reported here. The RF classifiers were built for the number of trees (n_{tree}) being tuned on the validation data from 10 to 150. For E1, E2 and E3, $n_{tree} = 50$,125 and 100 performed the best respectively. For E3, the PLDA scoring mechanism is implemented in addition to these classifiers, with eigen channel number=2, eigen voice number=5 and iterations=10.

[1] https://github.com/jcvasquezc/DisVoice

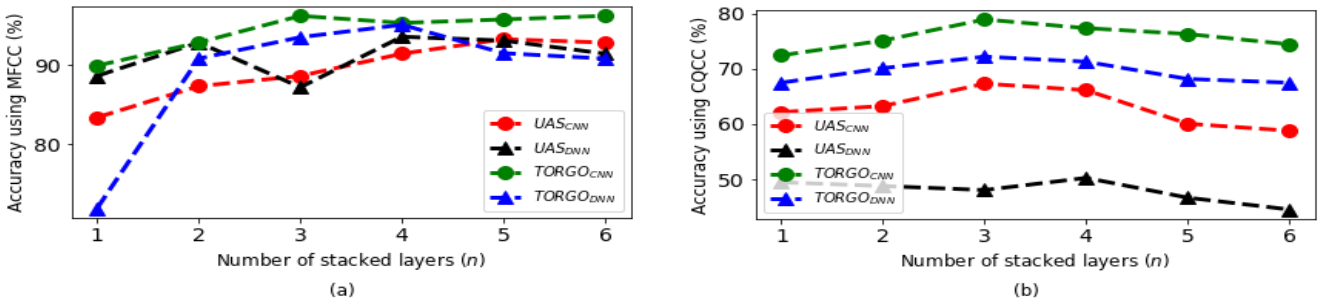

Fig. 1. Variation of the classification accuracy with the number of DNN and CNN layers for (a) MFCC and (b) CQCC.

B. Deep Learning Classifiers

A DNN model can model the high-level abstractions in the feature sets and learn the underlying data structure. DNN models are implemented in Keras by stacking n dense layers of ReLU activation. The number of neurons in each layer is designed to grow up with the model depth, in powers of 2. The first layer has the number of nodes equal to the power of two nearest to the input feature vector. For E1, 39 MFCCs are used, hence the first layer has 32 nodes, the second has 64 and so on. The dense layers are followed by a layer with a dropout factor of 0.4. The output layer has softmax activation. Each DNN is trained with a batch size of 32 and a learning rate of 0.001, for 120 epochs. All these are set after hyper-parameter tuning on the validation data, with Adam optimiser. Final tuning is done with respect to n for all features.

CNN is built using alternating convolution and pooling layers, whose 2D filters capture the spectral correlations in the acoustic features. In the front-end, each speech frame is represented by 13 MFFCs or CQCCs, which when stacked up gives the 2D feature maps for the convolution layers to act upon. Thus, local information can be efficiently extracted from the variabilities embedded in the frame-wise feature representation. CNN models are implemented with n stacked up 2D convolutional layers of 2×2 kernel size and ReLU activation function, each followed by a batch-normalisation layer. The number of feature maps increases in powers of 2 as in the case of DNN models. In all models a 2D max-pooling layer with a pooling size of 2×2 is used, followed by a dropout layer of factor 0.2. The flattened result of this is passed to the dense layers with the number of units decreasing in powers of 2 with n. MFCC features alone are used here, since the temporal information is captured from the frame-wise data provided, and deltas would just add redundancy.

Recurrent neural network (RNN) has proven to be efficient in capturing the temporal dependencies for sequential tasks. LSTM is one of its variations, which can flexibly capture the long-range dependencies by overcoming the vanishing gradient problem in the conventional RNNs. This is achieved by using three gates, namely, the input gate, the forget gate and the output gate, that controls the information flow in the network. They also have memory cells holding past and present information. Thus, by the efficient gating mechanism and usage of memory cells they can mitigate the gradient issues and enable adequate information flow. Three stacked LSTM layers, followed by a dropout layer, and the output dense layer constitute the LSTM models. The number of hidden units (N_h) in the first LSTM layer is taken by the general rule of thumb:

$$N_h = \frac{N_s}{\alpha(N_i + N_o)} \tag{2}$$

where, N_s is the number of training samples used, N_i, the number of input neurons, N_o, the number of output units and α, the scaling factor lying between 2 and 10. Tuning is done for different values of α, as it controls the number of model free parameters, and thus the generalisation capability of the model. The following layers have the number of units tuned and selected to be 600 and 200, respectively.

Proposed for machine translation [37], GRUs are simpler structures of LSTMs, with fewer parameters. Unlike LSTM, GRU has only two gates: an 'update' gate to control the amount of information to be transferred from the previous hidden state to the current hidden state, and a 'reset' gate to effectively drop irrelevant information, leading to better predictions. There are no separate memory cells as well. Thus, they are much simpler and require less computational power. Also, they train faster and require lesser data to generalise. Hence, usable in data-stringent cases like the speech-disorder analysis, where pathologically affected utterances are of limited number. The GRU models are implemented like the LSTM models, and are tuned with respect to α. The frame-wise input is made available as that of CNNs.

VI. Results and Discussion

A. Analysing MFCCs and CQCCs (E1)

The DNN and CNN models are tuned with respect to the increasing number of stacked layers (n). The results are plotted in Fig. 1(a) and (b) for MFCC and CQCC features, respectively. As the model grows in depth with increasing n, the upper layers find efficient feature representations that generalise well across the datasets. Thus, an increase in accuracy was observed up to $n = 4$ for both the databases on using MFCCs with DNNs [25]. While labelling the graphs, UA-Speech is referred to as UAS. As the number of layers increases beyond four, the overall classification accuracy decreases. This is because, the generalisation capability decreased on increasing the model complexity. That is, the network becomes overfit to the training set, and fails to make

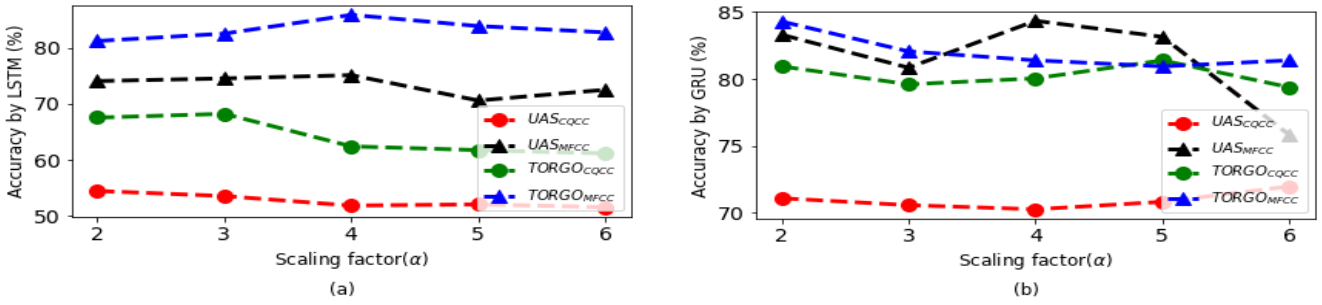

Fig. 2. Variation of the accuracy with parameter α for (a) LSTM and (b) GRU.

the right decision on the unseen test data. For CNN models using MFCCs, $n = 5$ gave the best result in the case of UA-Speech and $n = 3$ for TORGO, as seen in Fig. 1(a). As in the DNN models, a further increase led to a decrease in the accuracy. For models using CQCCs, a similar trend was observed, but with reduced accuracy scores compared to MFCCs. There is about a 20% difference between the accuracies, as evident in the graphs. A possible cause is that, the dysarthric severity is exhibited not only by the speaker characteristics, but also by the speech pattern. The changes in the pattern of monotonicity shown by dysarthric patients are better identified by MFCCs.

The variation of classification accuracy with α, as per (2) for LSTM and GRU models are plotted in Fig. 2(a) and Fig. 2(b), respectively. In the case of LSTM, $\alpha = 4$ gave the best classification accuracy with 85.87% and 75.08% for TORGO and UA-Speech, respectively, on using MFCCs. For models using CQCCs, we find that the maximum accuracy occurs at $\alpha = 2$ itself. This means that an increase in the number of LSTM nodes, and hence, an increase in the model capacity do not favour the models in capturing discriminating features from CQCCs. Again, we can find clear margins between the graphs for MFCCs and CQCCs. The results obtained on TORGO exceed those obtained on UA-Speech, even when the latter has almost five times more the data. This is because the training and testing data of the UA-Speech comprise of completely different words. This again justifies the performance of LSTM in UA-Speech data classification, being worse than the rest. The temporal information identified by the LSTM model from the common words is not sufficient enough to identify the severity level from the uncommon words [25]. This is checked by using a mixed-up data for training and testing, and an accuracy of 88.59% is obtained, which validates the inference. However, GRU could deal with this variability in utterances, and gave a comparable performance on both databases. The advantage of GRU in limited data settings also adds to this.

From Table II we can observe that, there is an improvement of more than 10% for the best performing CNN models over the baseline SVM system on using MFCCs. For models using CQCCs, more than 20% improvement is observed, which shows the incapability of CQCC features in representing the severity level-dependent factors of dysarthric speech. Since SVM models merely work on these features as such for classification, very poor performance is observed. But DNN models do another level of compact and significant feature learning from them, and it is this 'deep learning' that happens on the less efficient CQCCs that brings this difference. The RF classifiers have been able to give better results than SVM on both the features. In our experiments, the performance of DNN is at par with that of CNN, with a clear margin over LSTM. GRU outperforms the rest for CQCCs, showing that the temporal dependencies among CQCC features are relevant in identifying the severity levels. On getting averaged utterance-level features, their contribution is reduced, as seen on SVM and DNN classifiers.

TABLE II
OVERALL CLASSIFICATION RESULTS (%) OF E1

Database	Feature	SVM	RF	DNN	CNN	LSTM	GRU
TORGO	MFCC	82.73	89.69	95.06	**96.18**	85.87	84.30
	CQCC	58.97	71.52	72.19	78.92	67.93	81.39
UAS	MFCC	82.91	87.75	**93.55**	93.24	75.08	84.35
	CQCC	39.25	52.02	50.27	67.31	54.24	71.95

TABLE III
ACCURACY (%) USING SPEECH DISORDER SPECIFIC FEATURES

Database	Classifier	Prosody	Articulation	Glottal	Phonation
TORGO	SVM	60.18	85.87	54.26	63.90
	RF	61.43	85.43	56.86	68.16
	DNN	60.98	**86.77**	56.50	69.73
UA-Speech	SVM	61.68	77.98	55.91	53.33
	RF	62.88	77.71	52.02	60.42
	DNN	62.71	**80.44**	54.47	62.88

B. Analysing Speech Disorder Specific Features (E2)

At first, the severity classification is performed using each of the DisVoice feature sets, namely, prosody, articulation, glottal, and phonation. The results are tabulated in Table III. We can observe that DNN outperforms SVM in all cases, but RF classifiers have managed to give results that are close to those obtained by DNN. For both the databases, the articulatory features give the best results. The articulation deficits and the reduced vowel articulation index have proved to be capable of identifying the stage of PD in [38], which can be related to hypokinetic dysarthria, and hence, our findings adhere to this. From the confusion matrices obtained on using the other DisVoice features we found that, misclassification happens to the nearby classes, and there are no signs of over-fitting from

TABLE IV
CLASSIFICATION ACCURACY (%) ON TUNING N_G AND N_{IV} (BEST OVERALL ACCURACY IN BOLD, BEST AMONG THE TUNING SETTING IN RED)

N_g=128	i_{MFCC}				i_{CQCC}				N_{iv}=200	i_{MFCC}				i_{CQCC}			
N_{iv}	PLDA	DNN	SVM	RF	PLDA	DNN	SVM	RF	N_g	PLDA	DNN	SVM	RF	PLDA	DNN	SVM	RF
100	50.73	92.93	85.64	82.51	55.11	64.28	58.91	51.42	128	53.23	93.73	85.15	80.35	55.47	64.36	59.88	48.71
200	53.23	93.73	85.15	80.35	55.47	64.36	59.88	48.71	256	55.03	93.82	84.87	82.77	58.13	64.99	61.31	50.67
400	52.63	93.06	84.48	75.29	55.41	61.97	59.08	46.60	512	55.05	**93.97**	85.93	82.93	61.48	66.20	61.60	52.02
600	50.65	92.48	83.75	71.22	54.70	62.24	58.26	44.55									

the generalisation curves. This implies that these features, even though capable of detecting dysarthria from healthy speech as seen in literature [16], cannot efficiently embed the differences among the severity levels.

Later, the concatenated feature sets (referred to as 'UAS_{conc}/$TORGO_{conc}$') are used, followed by the FA-based dimensionality reduction being applied, with the number of factors ranging from 100 to 400, in multiples of 100. The best performance is achieved at dimension 200 for both databases, as seen in Fig. 3(a). Hence this dimension is chosen for further analysis, and referred to as FA(200). This is due to the 'curse of dimensionality' problem, and related overfitting of models with increased feature complexity. A decreasing trend in accuracy after a particular n is observed here as well, as visible in Fig. 3(b). This can also be mapped to the reduced generalisation capability of the network with increased model complexity as in E1. The best accuracy of 93.27% can be found for $n = 2$ on TORGO, and on UA-Speech it is 90.80% at $n = 3$. The difference in train and test data has led to decreased classification accuracy on UA-Speech as in E1. The results reported by SVM for TORGO are 82.51% and 86.71% for concatenated and FA(200) feature sets respectively. For UA-Speech this respectively maps to 79.69% and 85.35%. This again validates the usage of FA. Similarly RF gave 82.24% and 73.00% on TORGO, and 89.69% and 82.06% on UA-Speech for the concatenated and FA(200) feature sets respectively. The FA(200) features have not been useful in the case of RF classifiers, as these classifiers themselves do another level of dimensionality reduction while computing the results.

C. Analysing i-Vectors (E3)

A first level tuning of the i-vector feature extraction was done using the UA-Speech database, as it is the largest among the two corpora, and has all four levels of severity. The results of i-vector tuning with respect to N_g and N_{iv} are given in Table IV. At first, N_g is set to 128 and tuning is done with respect to N_{iv}. Classification accuracy by DNN and PLDA was found to be maximum at $N_{iv} = 200$. The most discriminating features embedding information about the intelligibility is contained in a few dimensions, and this leads to the deteriorating performance of the classifier with increasing dimension. The maximum accuracies reported by SVM and RF are 85.64% and 82.51%, respectively. The classifiers show a decrease in accuracy with increased feature dimension, as expected, due to the curse of dimensionality phenomenon. With $N_{iv} = 200$, further tuning was done with respect to N_g. It is observed that maximum accuracy is obtained for $N_g = 512$ by all classifiers. Further increase in mixtures is avoided due to the computational complexity and the marginal improvement observed, which is validated by the results in [39]. Thus, the best performing configuration parameters are $N_g = 512$ and $N_{iv} = 200$. Using this configuration i-vectors for TORGO database are extracted.

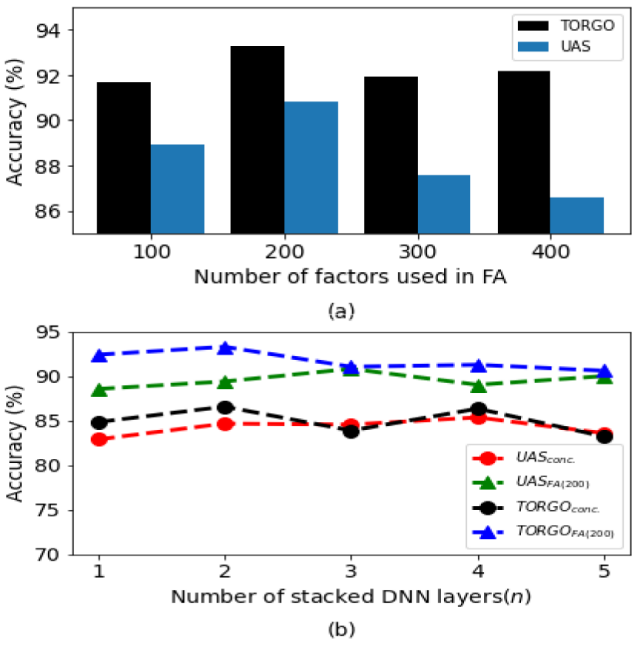

Fig. 3. Variation of the classification accuracy with: (a) number of factors used in FA, and (b) number of DNN layers.

Variation in the DNN classification accuracy with the increasing number of dense layers (n), is given by Fig. 4 (a). For UA-Speech, the network gives the best classification accuracy of 93.97% for $n = 3$, on using i_{MFCC}. A further increase did not help the network, and the generalisation gap is found to be increased. On using i_{CQCC}, the network performs best at $n = 1$, giving 66.20%. This means that the DNN is not learning from i_{CQCC} with growing depth, and is found to be overfitted. We tried using regularisers, but the accuracy of 66.20% was barely uplifted to 66.73%. As a further step, the model complexity was reduced by lowering the number of units in the first dense layer to 32. However, this led to underfitting. Thus, it is inferred that the vocal-tract irregularities of dysarthric speech are not highlighted on i_{CQCC} when used on DNN. This is again validated by the t-SNE plots [40] in Fig. 4 (b) and (c). These are drawn from the output vectors produced by the snippets from various classes for the last dense layer of the trained networks. Good clustering is exhibited by the DNN model using i_{MFCC}s, in contrast to the one using i_{CQCC}s. It can be

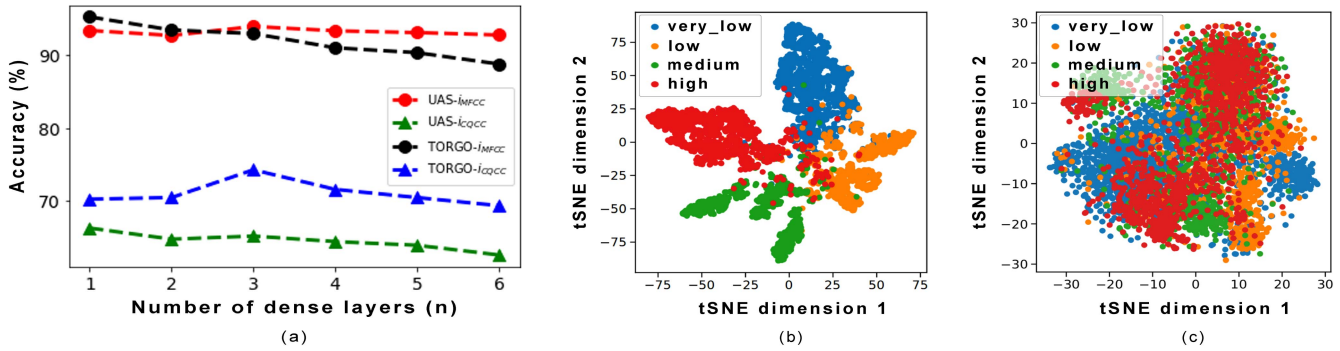

Fig. 4. (a) Variation in DNN performance with increasing *n*, t-SNE plots from the last dense layer of DNN using (b) i_{MFCC} (c) i_{CQCC}.

observed that, the low and the medium class clusters show the speaker variations, or has higher intra-class variability. This is owed to the fact that, the number of speakers, and hence the amount of training data available, is lesser for these classes. Even then, the inter-class variability is well-maintained. For TORGO, the accuracies reported by DNN, SVM and RF classifiers are 95.29%, 87.67% and 84.08% respectively, for i_{MFCC}s. The same reported for i_{CQCC}s are 74.22%, 66.14% and 63.45% respectively. For the DNN models, i_{MFCC}s lead i_{CQCC}s by over 20%, as in the case of UA-Speech, but the tuning results in Fig. 4 (a) are quite different. The accuracy saturated at n=1 for i_{MFCC}s and at n=3 for i_{CQCC}s. But being the largest database among the two, the results of UA-Speech over-rule this trend shown by the DNN classifiers.

D. Evaluating Speaker-Dependency of the Models

To know if the models work well under unseen-speaker scenario, the SID experiments are performed. The different classifiers, with their best tuned settings as obtained from SD experiments, are evaluated using the LOSO cross-validation technique. The UA-Speech database is chosen for this, as it has a larger number of files, and speakers belonging to all four severity levels. The common words from 14 speakers were used for training, and the uncommon words of the left-out speaker were used for testing in each round. Hence, the models are checked on unseen speakers, as well as with unseen utterances. As there are 15 speakers, the average of 15 rounds is reported here. The experiment was performed in two different ways: (a) SID test case with four classes as in previous experiments, (b) SID binary classification of severity levels. Table I shows that the intermediate classes have a lower number of speakers than the border classes, which implies that the UA-Speech database is unbalanced. Hence, the deep learning models may overfit the classes having more speakers. So, we implement this binary classification, merging classes 'low' and 'very low' as one, and 'medium' and 'high as another, as in [24] and [41]. This is helpful in situations wherein a mere understanding of whether the severity level is 'high' or 'low' is enough for the intended medication plan/speech therapy. This also indicates if the system would efficiently handle the problem of detection of dysarthric speech from healthy ones, since the border classes are being differentiated.

TABLE V
AVERAGE LOSO CROSS-VALIDATION ACCURACY (%) OF E1

Type	Feature	SVM	RF	DNN	CNN	LSTM	GRU
4-level	MFCC	29.15	21.91	24.13	30.62	24.02	22.82
	CQCC	28.07	26.24	31.13	**35.69**	28.37	23.18
Binary	MFCC	61.24	61.91	60.42	66.87	61.35	60.18
	CQCC	58.09	54.47	**70.77**	58.87	56.84	58.53

TABLE VI
AVERAGE LOSO CROSS-VALIDATION ACCURACY (%) OF E2 AND E3

Type	Experiment	Feature	SVM	RF	DNN
4-level	E2	Concatenated	26.99	27.84	32.15
		FA(200)	23.91	24.35	36.11
	E3	i_{MFCC}	38.02	38.89	**49.22**
		i_{CQCC}	31.33	28.09	38.25
Binary	E2	Concatenated	54.06	63.53	65.55
		FA(200)	55.47	49.87	59.02
	E3	i_{MFCC}	66.18	68.70	**70.52**
		i_{CQCC}	59.78	51.85	59.16

Table V shows that the trend shown in the SD test case is reversed in the SID case, as CQCCs outperformed MFCCs. This puts light on the fact that the generic characteristics among the same class speakers are better identified by CQCCs. It can be inferred that the dysarthric characteristics specific to the speakers (such as monotonicity) are being highlighted in the MFCC features. This causes the network to perform well even with unseen words of the seen speakers, but not with unseen speakers. CQCC incorporates the minute variations in pitch occurring at lower frequencies due to the higher frequency resolution that CQT provides in low frequency regions. Thus, is found to outperform the STFT-based MFCCs. This is in agreement with the findings reported in [11] using CQT-based and STFT-based spectrograms. The two-class or binary LOSO cross-validation results are given in the later rows. Here, since a better balanced dataset is made available, we find the networks performing well. For all experiments, DNN outperformed the rest, with the best being E1, with CQCC, giving 70.77%. This is at an appreciable gain of over 5% over the best results reported for the same task in [24]. There is also an improvement of over 10% to the other models implemented in E1, which implies that utterance-level CQCC are better than frame-level features here. Another worth noting point is that the performance of the LSTM model has

TABLE VII
COMPARISON WITH PIONEER WORKS IN LITERATURE REPORTING SPEAKER-DEPENDENT (SD) AND SPEAKER-INDEPENDENT (SID) TEST CASES

SI no.	Work	Approach: Feature & Classifier	Data Usage	Results	Remarks
1	A. Tripathi, S. Bhosale, and S. K. Kopparapu [24]	Deep Speech posteriors with SVM	Common words for training and uncommon words for testing for SD case, round-robin LOSO cross-validation for SID case	**53.90%** (SID), 97.40% (SD), 65.20% (binary LOSO)	Best results for SID case using LOSO cross-validation
2	H. Chandrashekar, V. Karjigi, and N. Sreedevi [2]	Spectro-temporal features from mel scale spectrogram with CNN	355 and 100 distinct words used for training and testing respectively in SD case, 100 words of left-out speakers for testing in SID case using round-robin LOSO	98.30% (SD), 49.27% (SID)	Whole database not used. Reported better accuracy on balanced dataset: using only 2 speakers per class for training
3	K. Gurugubelli and A. K. Vuppala [12]	PE-SFCC with i-vector-PLDA modeling	Common words for training and uncommon words for LOSO test	60.78% (SID)	A comprehensive cross-validation not reported
4	Proposed method in E3	i_{MFCC} with DNN	Common words for training and uncommon words for testing for SD case, round-robin LOSO cross-validation for SID case	49.22% (SID), 93.97% (SD), **70.52%** (binary LOSO)	Best results for SID case for binary LOSO cross-validation

improved to be at par with others. Thus, in the seen-speaker scenario when other networks showed very high classification accuracy, there is evident speaker overfitting. However, the RNN models did not show such a trend. Also, we find a comparable performance by the SVM and RF classifiers to that shown by the DNN and RNN networks. This occurs mainly due to the fact that the dataset is unbalanced, and has only few subjects per class. Machine learning classifiers have always shown results better than deep learning models in data stringent situations. This is because while the former learns well with handcrafted features on the available data, the latter aims for consecutive hierarchical identification of complex concepts that represent the underlying data, which demands more amount of data.

Table VI gives the LOSO results of E2 and E3. The FA (200) feature set gave an average cross-validation accuracy of 36.11% using DNN, beating the results of E1. However, the feature set worked poor on SVM and RF. As observed in the SD case, DNN with i_{MFCC} performed best, giving 49.22%. This outperforms the results of E1 and E2, and highlights the efficiency of i-vectors in identifying the severity levels. Although this seems to be too low, it is worth noting that the best SID dysarthria severity assessment system in literature, [24] reported 53.90% only, and ours is nearing the same. This result was reported for a probability-fusion framework on acoustic and textual-derived features extracted using a pre-trained DeepSpeech-1 ASR model, trained with 1000 hours of data. It is important to note that, we achieved a comparable result with the simple MFCC-based i-vectors used on DNN. In binary cross-validation of E2, the concatenated feature set outperformed the FA(200) feature set. We infer that, on working with an unbalanced small dataset a precise and compact feature representation is useful. When the dataset is favoured to include more speakers per class, a detailed feature study would be essential. An average accuracy of 70.52% was given by DNN using i_{MFCC} in E3, which is at par with that obtained in E2 and beats the result in [24] by 5%. Table VII gives a brief comparison with the pioneer works on dysarthria severity classification on UA-Speech. Lexical features such as the fillers, phone rate, error rate, pauses, and goodness of pronunciation as used in [23] could be combined with the i-vectors in a multimodal fusion network to improve the results. Also, the frame-wise computed speech-disorder specific features of E2 can be used for the i-vector generation, as they give good SID results.

E. Discussion

The progression of dysarthria as identified by SLPs using auditory perceptual measures, has been automated in E1 using the perceptual features of speech, namely, MFCCs and CQCCs. The results prove that they can be used with efficient classifiers to provide an unbiased judgement of the dysarthria severity. DNN and CNN classifiers outperform the RNN and machine learning models on using MFCCs. But GRU handled CQCCs better, indicating that their temporal dependencies are important. On the feature side, MFCCs outperformed CQCCs in the SD test case, but CQCCs promise better SID models by showing less speaker-overfitting. On using i-vectors, i_{MFCC}s performed best in the SD case, beating all other features, and achieved an improvement of nearly 20% classification accuracy, compared to raw MFCCs in SID systems. There is an appreciable gain over the conventional i-vector-PLDA paradigm by the proposed second-level DNN learning of the i-vectors. The improvement obtained by DNN over the SVM classifiers is around 8% on using i_{MFCC}s in the SD test case, and 11% in the SID case. The corresponding values over the RF classifiers are 10% and 9% respectively. This gain margin is worth the increased computational time and can be improved with the availability of a larger database. The models suffer from the limitation of being trained with few subjects per class, which is reflected by their poor performance in the SID case. As in the case of deep learning models, building of UBM and modeling of the TV matrix for the i-vector extraction would benefit from the availability of a larger database, resulting in more discriminating i-vectors. However, at present UA-Speech is the largest dysarthric speech database available with all four severity levels, and all pioneer works are done using it. In literature, the speech-disorder specific features have been extensively used in detecting disordered speech from healthy ones, but they have proved to be less useful in modeling the dysarthria severity levels. The 488-dimensional articulation feature set performed the best among these, followed

by the 28-dimensional phonation features which pushed the 103-dimensional prosodic features to the third position. This showed that the dimension of the feature set need not impact accuracy, as in agreement with the findings of [42]. The effectiveness of articulation features compared to the others has been proved in the automatic evaluation of PD patients in [17] and [38]. A detailed statistical analysis must be done to find the potential correlation within each feature class to find the optimum feature descriptors. We would like to do the hypothesis testing using the technique of paraconsistent feature engineering (PFE) [43] as future work. The intra-class similarities and inter-class distinctions exhibited by the different feature sets can be quantified by PFE. This would evaluate the discriminating power of the various features, and rank them for their efficacy in classifying the severity levels. This would also explain why the DNN classifiers cannot give good results with all the features considered for the study.

VII. Conclusion

To the best of our knowledge, the current study is the first detailed investigation on the various deep learning models using different acoustic features for dysarthria severity classification. We have also introduced a second level feature learning on i-vectors using DNNs. Among the different features studied, MFCCs offer the least computational complexity on all the classifiers. However, if accuracy is the prime concern, then the DNN-i_{MFCC} framework has to be used. I-vectors have proved to be the best in speaker-recognition cases, and has now proved to be the best in identifying speaker-groups as well. We would like to explore the recent state-of-the-art features x-vectors as future work. The characteristics of dysarthric speech excitation source and the associated non-linearities in speech production were analysed using Teager energy operator (TEO) profile in [22]. The dysarthric TEO profile was found to be highly irregular, with more high amplitude and noisy bumps noted at high severity levels. Later in [44], it was shown that the energy estimated using the enhanced Teager energy operator (ETEO) [45], is greater than that obtained by using the TEO for higher frequencies for normal speech. Inspired by these results, we would like to explore the usage of ETEO for differentiating the different dysarthria severity levels as future work.

References

[1] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 947–960, May 2011.

[2] H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, "Spectro-temporal representation of speech for intelligibility assessment of dysarthria," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 390–399, Feb. 2020.

[3] M. J. Kim, J. Yoo, and H. Kim, "Dysarthric speech recognition using dysarthria severity-dependent and speaker-adaptive models," in *Proc. Interspeech*, 2013, pp. 3622–3626.

[4] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 10, pp. 1943–1953, Oct. 2006.

[5] C. Bhat and H. Strik, "Automatic assessment of sentence-level dysarthria intelligibility using BLSTM," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 322–330, Feb. 2020.

[6] G. Vyas, M. K. Dutta, J. Prinosil, and P. Harár, "An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features," in *Proc. IEEE 39th Int. Conf. Telecommun. Signal Process.*, vol. 2016, pp. 515–518.

[7] A. Farhadipour, H. Veisi, M. Asgari, and M. A. Keyvanrad, "Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks," *ETRI J.*, vol. 40, no. 5, pp. 643–652, Oct. 2018.

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

[9] M. S. Paja and T. H. Falk, "Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1–4.

[10] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Odyssey*. Bilbao, Spain, 2016, pp. 283–290.

[11] C. H M, V. Karjigi, and N. Sreedevi, "Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2880–2889, Dec. 2020.

[12] K. Gurugubelli and A. K. Vuppala, "Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3403–3407.

[13] H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, "Breathiness indices for classification of dysarthria based on type and speech intelligibility," in *Proc. Int. Conf. Wireless Commun. Signal Process. Netw. (WiSPNET)*, Mar. 2019, pp. 266–270.

[14] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5070–5074.

[15] K. Kadi, S. Selouani, B. Boudraa, and M. Boudraa, "Discriminative prosodic features to assess the dysarthria severity levels," in *Proc. World Congr. Engg.*, vol. 3, 2013, pp. 1–5.

[16] E. A. Belalcázar-Bolanos, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, T. Haderlein, and E. Nöth, "Glottal flow patterns analyses for parkinson's disease detection: Acoustic and nonlinear approaches," in *Proc. Int. Conf. Text, Speech, Dialogue*. Cham, Switzerland: Springer, 2016, pp. 400–407.

[17] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, T. Bocklet, and E. Nöth, "Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease," *J. Commun. Disorders*, vol. 76, pp. 21–36, Nov. 2018.

[18] J. R. Orozco-Arroyave *et al.*, "NeuroSpeech: An open-source software for Parkinson's speech analysis," *Digit. Signal Process.*, vol. 77, pp. 207–221, Jun. 2018.

[19] P. Verma and P. K. Das, "I-vectors in speech processing applications: A survey," *Int. J. Speech Technol.*, vol. 18, no. 4, pp. 529–546, Dec. 2015.

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

[21] C. Espana-Bonet and J. A. Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech," in *Proc. 3rd Int. Conf. Adv. Speech Lang. Technol. Iberian Lang.*, 2016, pp. 97–107.

[22] S. Gupta *et al.*, "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," *Neural Netw.*, vol. 139, pp. 105–117, Jul. 2021.

[23] M. Perez *et al.*, "Classification of Huntington disease using acoustic and lexical features," in *Proc. Interspeech*, 2018, p. 1898.

[24] A. Tripathi, S. Bhosale, and S. K. Kopparapu, "Improved speaker independent dysarthria intelligibility classification using deepspeech posteriors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6114–6118.

[25] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 116–120.

[26] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, Dec. 2012.

[27] H. Kim *et al.*, "Dysarthric speech database for universal access research," in *Proc. Interspeech*, 2008, pp. 1741–1744.

[28] T. Arias-Vergara, J. C. Vásquez-Correa, and J. R. Orozco-Arroyave, "Parkinson's disease and aging: Analysis of their effect in phonation and articulation of speech," *Cognit. Comput.*, vol. 9, no. 6, pp. 731–748, Dec. 2017.

[29] R. A. Reyment and K. Jvreskog, *Applied Factor Analysis in the Natural Sciences*. Cambridge, U.K.: Cambridge Univ. Press, 1996.
[30] N. Dehak, P. J. Kenny, R. Dehak, D. Pierre, and O. Pierre, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.
[31] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "I-vector based speaker recognition on short utterances," in *Proc. Interspeech*, 2011, pp. 2341–2344.
[32] A. Lozano-Diez *et al.*, "Analysis and optimization of bottleneck features for speaker recognition," in *Odyssey*. Bilbao, Spain, 2016, pp. 352–357.
[33] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1695–1699.
[34] O. Ghahabi and J. Hernando, "Deep belief networks for I-vector based speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 1700–1704.
[35] B. McFee *et al.*, "Librosa/librosa: 0.7.0," Zenodo, Tech. Rep., Jul. 2019, doi: 10.5281/zenodo.3270922.
[36] M. Todisco, H. Delgado, and N. W. Evans, "Articulation rate filtering of CQCC features for automatic speaker verification," in *Proc. Interspeech*, 2016, pp. 3628–3632.
[37] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*.
[38] S. Skodda, W. Visser, and U. Schlegel, "Vowel articulation in Parkinson's disease," *J. Voice*, vol. 25, no. 4, pp. 467–472, Jul. 2011.
[39] D. Martínez, P. D. Green, and H. Christensen, "Dysarthria intelligibility assessment in a factor analysis total variability space," in *Proc. Interspeech*, Aug. 2013, pp. 2133–2137.
[40] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.
[41] T. H. Falk, W.-Y. Chan, and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Commun.*, vol. 54, no. 5, pp. 622–631, Jun. 2012.
[42] H. Holmström and V. Zars, "Effect of feature extraction when classifying emotions in speech-an applied study," Ph.D. dissertation, Dept. Comput. Sci., Umeà Univ., Sweden, U.K., 2018.
[43] R. C. Guido, "Paraconsistent feature engineering [lecture notes]," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 154–158, Jan. 2019.
[44] A. T. Patil, R. Acharya, H. A. Patil, and R. C. Guido, "Improving the potential of enhanced teager energy cepstral coefficients (ETECC) for replay attack detection," *Comput. Speech Lang.*, vol. 72, Mar. 2022, Art. no. 101281.
[45] R. C. Guido, "Enhancing teager energy operator based on a novel and appealing concept: Signal mass," *J. Franklin Inst.*, vol. 356, no. 4, pp. 2346–2352, Mar. 2019.