Automated Dysarthria Severity Classification Using Deep Learning Frameworks

Amlu Anna Joshy, Rajeev Rajan
College of Engineering Trivandrum,
APJ Abdul Kalam Technological University,
Thiruvananthapuram, Kerala, India.

***Abstract*—Dysarthria is a neuro-motor speech disorder that renders speech unintelligible, in proportional to its severity. Assessing the severity level of dysarthria, apart from being a diagnostic step to evaluate the patient's improvement, is also capable of aiding automatic dysarthric speech recognition systems. In this paper, a detailed study on dysarthia severity classification using various deep learning architectural choices, namely deep neural network (DNN), convolutional neural network (CNN) and long short-term memory network (LSTM) is carried out. Mel frequency cepstral coefficients (MFCCs) and its derivatives are used as features. Performance of these models are compared with a baseline support vector machine (SVM) classifier using the UA-Speech corpus and the TORGO database. The highest classification accuracy of 96.18% and 93.24% are reported for TORGO and UA-Speech respectively. Detailed analysis on performance of these models shows that a proper choice of a deep learning architecture can ensure better performance than the conventionally used SVM classifier.**

***Index Terms*—dysarthria, intelligibility, automatic assessment, deep learning**

I. Introduction

Dysarthria is a motor speech impairment, often characterized by speech that is indiscernible by human listeners [1]. It results from a neurological injury such as cerebral palsy or any neuro-degenerative disease such as Parkinsons disease [2]. The motor speech sub-systems get impaired, leading to imprecise articulation, low audibility, atypical speech prosody and variable speech rate, which deteriorate the speech quality [3]. The speech intelligibilty can be analysed to determine the dysarthria severity level, and this can help in monitoring the patient's progression and planning speech therapy sessions [1]. Subjective assessment by a trained speech language pathologist would be expensive and inconsistent, which paves way to the need for an automatic dysarthria severity level classification system. The dysarthric patients suffer from poor co-ordination of muscles and related physical disabilities that make the use of keyboard or touch-screen based interactive applications difficult for them. This demands the need for automatic speech recognizers (ASR), and dysarthria severity level classification can improve the performance of an ASR as evident in [4].

In literature, dysarthric speech intelligibility assessment has been done either by ASR based methods [5] or by blind intelligibility assessment [6]. In [7], glottal parameters, extracted from the voice source signal using quasi-closed phase glottal inverse filtering method, are explored for dysarthric speech classification into healthy or dysarthric. Intelligibility assessment from i-vectors is done in [8], using a v-support vector regression (vSVR) predictor. A non-linguistic method of severity assessment is proposed in [1] using audio descriptors and artificial neural network (ANN). In [9], perceptually enhanced single frequency filtering based cepstral coefficients (PE-SFCC) are proposed for intelligibility assessment, and i-vectors with probabilistic linear discriminant analysis (PLDA) scoring mechanism is used for classification. Deep learning models like DNN, CNN, time delay neural network (TDNN), and LSTM are explored for dysarthric ASR on TORGO database in [10].

Due to their capacity to capture "global" spectral envelope properties, MFCCs are employed in numerous perceptually motivated audio classification tasks and speech recognition systems, in addition to their widespread use in automatic monophonic or polyphonic timbre recognition [11]. MFCCs are encoded using a deep belief network (DBN), and employed for dysarthia severity classification using a multilayer perceptron (MLP) in [12]. DBN features have only a marginal improvement over MFCC based system. This motivated us to use the basic MFCC features alone to investigate the performance of various deep learning models for dysarthric severity classification.

The rest of the paper is organized as follows. A brief overview of the system and classification frameworks is described in Section II. Performance evaluation is done in Section III, followed by the analysis of results in Section IV. Finally the paper is concluded in Section V.

II. System Description

In feature extraction phase, MFCCs are computed framewise. Three deep learning strategies, namely DNN, CNN and LSTM are employed in classification phase and their performance is compared with a baseline SVM-based classifier. The detailed description is given in the following subsections.

A. Feature Extraction

Speech intelligibility is an indicator of dysarthria severity [13], which is influenced by the vocal muscular coordination. MFCCs have the capability to capture the irregular vocal fold movements or the lack of vocal-fold closure due to mass/tissue changes [14]. 13-dimensional MFCCs and their

first two derivatives (a total of 39 features) are computed framewise for DNN and SVM models with frame-length of 25 ms and frame-shift of 5 ms. The derivatives are not used for CNN and LSTM models to avoid redundancy, as the networks are capable of learning the temporal information.

B. Classification Phase

As a baseline, an SVM classifier with a linear kernel is initially experimented with MFCCs. Later, deep learning methodologies, namely DNN, CNN and LSTM are implemented in successive phases. In addition to learning the nonlinear mapping between the inputs and outputs, DNN is also capable of understanding the underlying data structure. Hence, it can effectively handle the different variabilities in a speech signal. CNN uses alternating convolution and pooling layers instead of stacked dense layers to give the advantage of local information extraction and spatial invariance. In the front-end, each speech frame is represented by MFFCs, which when stacked up gives the 2D feature map for the convolution layers to act upon [15, 16]. Finally LSTM-based recurrent neural network (RNN) is implemented, as it can capture long-range temporal dependencies by overcoming the vanishing gradient problem in conventional RNNs.

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	-

III. Performance evaluation

A. Dataset

The proposed technique is validated using two different dysarthric databases, namely (a) Universal Access dysarthric speech (UA-Speech) Corpus [17] and (b) TORGO database [18]. UA-Speech corpus comprises data from 13 control speakers and 19 dysarthric speakers. There are 765 word-utterances per speaker, corresponding to 300 distinct uncommon words and 3 repetitions of the 10 digits, 19 computer commands, 26 international radio alphabets and 100 common words. While using this database, the uncommon words are used for testing and the rest for training, which accounts for 300 and 465 utterances per speaker respectively. Only the data of 15 patients are available and thus a total of 6975 training files and 4500 test files are used in this work. Testing with uncommon words ensures that the network is evaluated on unseen words. The severity levels are very low, low, medium and high.

The TORGO database consists of aligned acoustics and measured 3D articulatory features from speakers with either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). It has dysarthric utterances of 8 speakers (3 females and 5 males) and utterances of 7 non-dysarthric/healthy speakers (3 females and 4 males). The utterances are categorised as (1) non-words, (2) short words such as digits, international radio alphabets, (3) restricted sentences and (4) unrestricted sentences. In this work, only words are used and there are 2227 such utterances in total. The severity levels are very low, low and medium. In the performance evaluation, 60% of the data is used for training the network, 20% for validation and another 20% for testing. The description of databases is given in Table 1.

TABLE II: CNN (upper pane) and LSTM (bottom pane) architectures used in the experiment

Output size	Description
(13,180)	2D MFCC
(12, 179, 16)	2*2 Convolution, 16 filters,Batch Normalization
(11, 178, 32)	2*2 Convolution, 32 filters,Batch Normalization
(10, 177, 64)	2*2 Convolution, 64 filters, batch normalization
(10, 177, 64)	2*2 Convolution, 64 filters, batch normalization
(9, 176, 128)	2*2 Convolution, 128 filters, batch normalization
(8, 175, 256)	2*2 Convolution, 256 filters,batch normalization
(4, 87, 256)	2*2 MaxPooling, Dropout(0.2), Followed by Flattened
128	Dense layer,batch normalization,Dropout(0.2)
64	Dense layer,batch normalization,Dropout(0.2)
4	Softmax

Output size	Description
(13,180)	2D MFCC
(13, 102)	LSTM, 102 hidden units
(13, 600)	LSTM, 600 hidden units
200	LSTM, 200 hidden units
200	Dropout(0.2)
4	Softmax

B. Experimental Framework

Implementation of the DNN models are done by stacking n dense layers, with the number of neurons growing with model depth, in powers of 2. Model with n=1 had a single dense layer of 16 units, n=2 had a dense layer of 16 units followed by one with 32 units and so on. The dense layers are followed by a layer with dropout of 0.25.

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. 2D max-pooling layer with pooling size of 2*2 is used in all models, followed by a dropout layer with dropout of 0.2. The flattened result of this is passed to the dense layers with number of units decreasing in powers of 2 with n. 13 dimensional MFCC features are arranged as 2-D feature maps, distributed along both frequency (using the frequency band index) and time (using the frame number). The frame number of 2D feature map is set to the frame number of the longest-uttered word with zero padding if needed. The detailed description on CNN configuration is given in Table II.

The LSTM-RNN models are implemented with 3 stacked LSTM layers, followed by a dropout layer and the output dense layer. The configuration is given in Table II. The number of hidden units (N_h) in the first LSTM layer is given by :

$$N_h = \frac{N_s}{\alpha(N_i + N_o)} \tag{1}$$

where, N_s is the number of training samples used, N_i, the

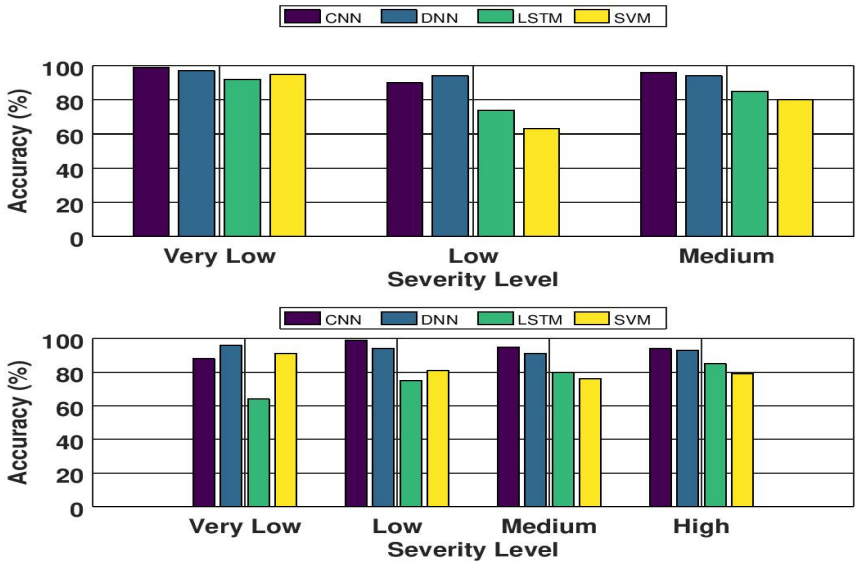

Fig. 1: Class-wise accuracies for different models

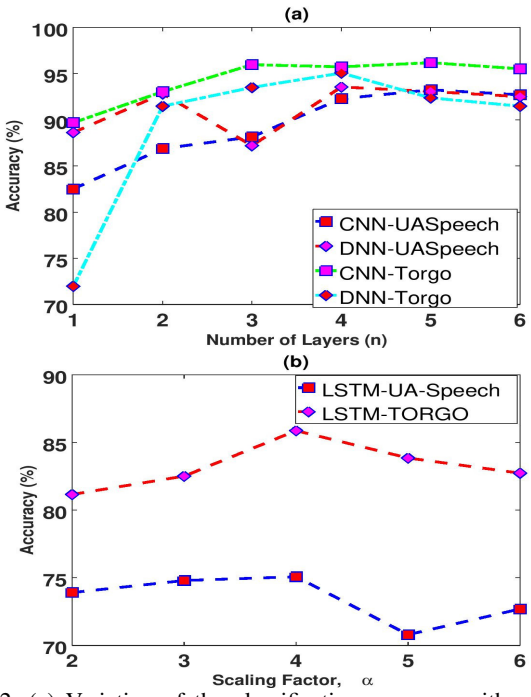

Fig. 2: (a) Variation of the classification accuracy with number of layers for DNN and CNN (b) (a) Variation of the classification accuracy with parameter α for LSTM

number of input neurons, N_o, the number of output units and a scaling factor, α, lying between 2 and 10. Tuning is done for different values of N_h by varying α.

IV. Results and Analysis

Severity classification is carried out using the experimental setup described in Section III. All the experiments are performed with batch size of 32 and learning rate of 0.001 for 120 epochs, after hyperparameter tuning. The overall results are shown in Table III. It is observed that there is an improvement of 14% and 10% for the best performing CNN models over the

TABLE III: Overall classification accuracy

	SVM	DNN	CNN	LSTM
TORGO	82.73	95.06	96.18	85.87
UA-Speech	82.91	93.55	93.24	75.08

baseline SVM system. As per the experiments reported in [10], DNN performed the best, as compared to CNN and LSTM for impaired speech. In our experiments, the performance of CNN is at par with that of DNN for UA-Speech, with a clear margin over LSTM. For TORGO database, a slight mileage in the performance is observed for CNN over DNN.

TABLE IV: Confusion matrix of SVM and CNN for TORGO

SVM	Very Low	low	Medium
Very Low	**177**	3	6
Low	11	**61**	25
Medium	16	16	**131**

CNN	Very Low	low	Medium
Very Low	**184**	0	2
Low	2	**87**	8
Medium	2	4	**157**

CNNs have been demonstrated effective in extracting useful features in spectral, temporal and spectro-temporal domains. This experimental results validate the claim that DNN and CNN-based acoustic models help to better discriminate between phonemes for the high and mid speech intelligibility groups in speech recognition [19]. From literature, it is seen that, in the context of dysarthic speech recognition, LSTM-RNN performs better than DNN for mildly affected dysarthric speakers, while giving worst performance for severely affected patients [10]. The confusion matrices for baseline SVM model

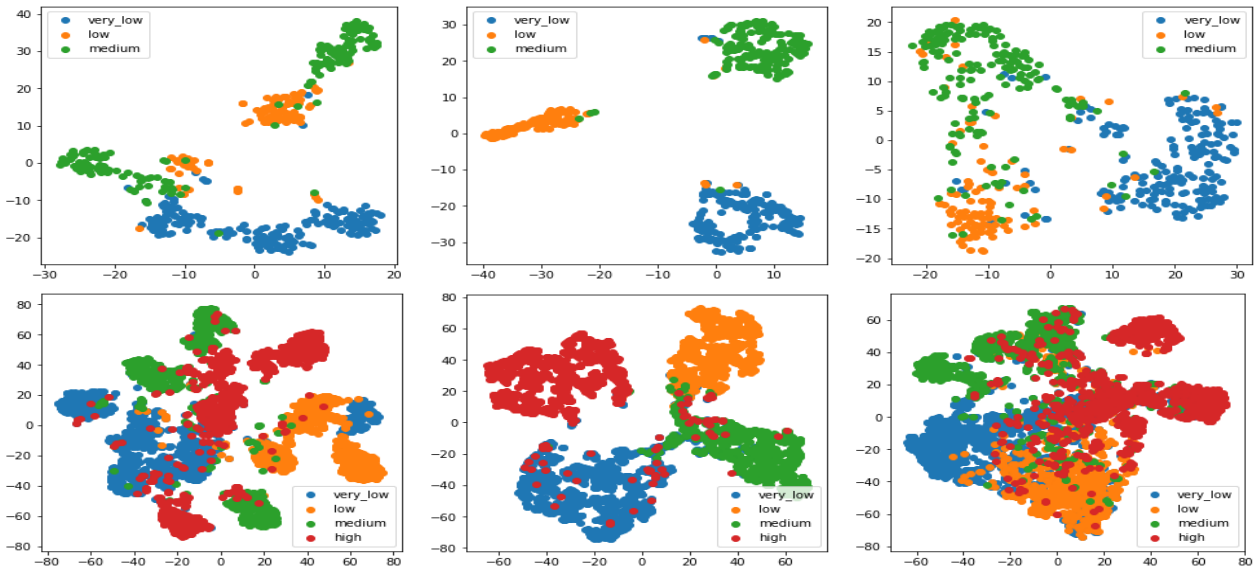

Fig. 3: t-SNE plot with output vectors produced by the snippets from various classes from the last dense layer for DNN, CNN and LSTM for TORGO (Upper pane) and UA Speech (Lower pane)

and best performing CNN model for TORGO and UA-Speech database are shown in Table IV and Table V, respectively. It is evident that many mis-classification errors among classes are reduced considerably by the CNN models. The class-wise accuracy measure is given in the bar plot shown in Fig. 1. As expected, the classification accuracy for less severe (very low and low) is high on the corpus for DNN and CNN, and both ensure minimum 80% accuracy for all classes. It is worth noting that the overall classification accuracy outperforms the accuracy cited in many recent works [9, 7] for these dysarthric speech corpora.

TABLE V: Confusion matrix of SVM and CNN for UA-Speech

SVM	Very Low	Low	Medium	High
Very Low	**1368**	10	45	77
Low	91	**731**	52	26
Medium	114	15	**683**	88
High	101	30	120	**949**

CNN	Very Low	Low	Medium	High
Very Low	**1317**	42	128	13
Low	0	**894**	4	2
Medium	5	34	**852**	9
High	33	11	23	**1133**

DNN is tuned with different number of stacked dense layers (n) and the results are plotted in Fig. 2(a). For n=4, the network showed its best performance for both the databases, giving 95.06% for TORGO and 93.55% for UA-Speech. As the number of layers increased beyond four, the overall classification accuracy decreased. A possible cause for this is that, the network becomes overfit to the training set and the network fails to make right decision on the new unseen data. In CNN, tuning is done with respect to n, and n=5 gave the best result (93.24%) in case of UA-Speech and for TORGO model with n=6 performed best (96.18%), as seen in Fig. 2(a). This is due to the fact that, as n increases, the model grows in depth, and the upper layers find efficient feature representations that are invariant to small perturbations leading to better model generalization. The accuracy of the LSTM system is also evaluated for various α, as given in equation 1. The variation of classification accuracy with α is plotted in Fig. 2(b). As seen in Fig. 2(b), α = 4 gave the best classification accuracy with 85.87% and 75.08% for TORGO and UA-Speech, respectively.

Fig. 3 visualizes the output vectors produced by the snippets from various classes for the last dense layer of the trained network using t-SNE. Note that, there is a good clustering (as represented with colour) and a general separation of different classes for CNN compared to DNN and LSTM. We used precision, recall and F1 measures as the performance metrics for various frameworks and these are given in the Table VI and VII. Average F1 measures of 0.95 and 0.96 are obtained for TORGO database by DNN and CNN, respectively, and F1 measure of 0.93 is reported for UA-Speech by both CNN and DNN. Although UA-Speech database had almost five times more data than TORGO, there is a reduction in accuracy due to the fact that the test set had words all different from those in the training set. This again justifies the performance of LSTM in UA-Speech data classification, being worse than the SVM base-model. The temporal information learned by the LSTM model from the common words are not sufficient enough in identifying the severity level, when it dealt with uncommon words. The performance of LSTM is re-investigated by using a mixed-up data, and an accuracy of 88.59% is obtained, with an average F1 score of 0.88.

To the best of our knowledge, the current study is the first detailed investigation on the various deep learning models for dysarthria severity classification. In [1], an average

TABLE VI: Precision (P), recall (R), and F1 measure for TORGO

	SVM			DNN			CNN			$LSTM$		
Severity	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Very Low	0.87	0.95	0.91	0.97	0.97	0.97	0.98	0.99	0.98	0.92	0.92	0.92
Low	0.76	0.63	0.69	0.91	0.94	0.92	0.96	0.90	0.95	0.73	0.74	0.73
Medium	0.81	0.80	0.81	0.96	0.94	0.95	0.94	0.96	0.95	0.86	0.85	0.86

TABLE VII: Precision (P), recall (R), and F1 measure for UA-Speech

	SVM			DNN			CNN			$LSTM$		
Severity	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Very Low	0.82	0.91	0.86	0.92	0.96	0.94	0.97	0.88	0.92	0.88	0.64	0.74
Low	0.93	0.81	0.87	0.95	0.94	0.94	0.91	0.99	0.95	0.65	0.75	0.70
Medium	0.76	0.76	0.76	0.95	0.91	0.93	0.85	0.95	0.89	0.70	0.80	0.78
High	0.83	0.79	0.81	0.94	0.93	0.94	0.98	0.94	0.96	0.72	0.85	0.78

classification accuracy of more than 95% is reported using audio descriptors. However, it is noted that the work is done on a subset of both datasets. Our results for both DNN and CNN models using MFCCs outperform those obtained using i-vectors for UA-Speech [9]. Thus, it is reasonable to conclude that proper choice of classification framework can ensure better performance even with baseline MFCC features as compared to other recent proposed features.

V. Conclusion

Objective assessment of dysarthria severity can aid clinical diagnostics and dysarthric speech recognition systems. This paper describes a comparative study on dysarthria severity level classification using different deep learning techniques, namely DNN, CNN and LSTM. MFCCs are used as features and analysis has been done with respect to SVM-based classifier for UA-Speech and TORGO datasets. The results indicate that both CNN and DNN outperform LSTM based systems, and proves to be far better than conventionally used SVM-based classifier. Also, the efficiency of MFCC features in discriminating the different intelligibility levels promises fast and reliable implementation of an automatic dysarthria severity classification system. As future work, the recent state-of-the art features such as x-vectors and i-vectors can be explored.

References

[1] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," *Proc. of International Conference on Acoustics, Speech and Signal Processing*, pp. 5070–5074, 2017.

[2] J. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio Speech Language Processing*, pp. 947–960, 2011.

[3] R. Palmer and P. Enderby, "Methods of speech therapy treatment for stable dysarthria: A review," *International Journal of Speech, Language and Pathology*, pp. 140–153, 2007.

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

[5] R. Hummel, W.-Y. Chan, and T. H. Falk, "Spectral features for automatic blind intelligibility estimation of spastic dysarthric speech," in *Proc. of Interspeech*, 2011.

[6] V. Berisha, R. Utianski, and J. Liss, "Towards a clinical tool for automatic intelligibility assessment," *Proc. of International Conference on Acoustics, Speech and Signal Processing*, pp. 2825–2828, 2013.

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

[8] 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 Transactions on Accessible Computing*, vol. 6, no. 3, pp. 1–21, 2015.

[9] K. Gurugubelli and A. K. Vuppala, "Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment," *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 3403–3407, 2019.

[10] C. Espana-Bonet and J. A. Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech,," *Proc. of Third International Conference on Advances in Speech and Language Technologies for Iberian Languages*, pp. 97–107, 2016.

[11] G. Richard, S. Sundaram, and S. Narayanan, "An overview on perceptually motivated audio indexing and classification," *Proc. of the IEEE*, vol. 101, pp. 1939–1954, 2013.

[12] A. Farhadipour, H. Veisi, M. Asgari, and M. A. Keyvanrad, "Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks," *ETRI Journal*, pp. 643–652, 2018.

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

[14] J. I. Godino-Llorente, P. Gómez-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 Transactions on Biomedical Engineering*, vol. 53, pp. 1943–1953, 2006.

[15] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," *Proc. of Advances in Neural Information Processing Systems*, pp. 1096–1104, 2009.

[16] D. Hau and K. Chen, "Exploring hierarchical speech representations using a deep convolutional neural network," *Proc. of 11th UK Workshop Computation Intelligence*, 2011.

[17] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," *Proc. of Interspeech*, pp. 1741–1744, 2008.

[18] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 6, no. 4, pp. 523–541, 2012.

[19] M. Kim, B. Cao, K. An, and J. Wang, "Dysarthric speech recognition using convolutional lstm neural network," *Proc. of Interspeech*, pp. 2948–2952, 2018.