2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT) | 979-8-3503-9763-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEEICT56924.2023.10157285

Dysarthria Speech Disorder Classification Using Traditional and Deep Learning Models

Manju Suresh[a,b*], Rajeev Rajan[c] and Joshua Thomas[a]
[a]*Department of Electronics and Communication*, College of Engineering Trivandrum, Thiruvananthapuram, India
[b]*Department of Electronics and Communication*, Mar Baselios College of Engineering and Technology, Thiruvananthapuram
[c]*Department of Electronics and Communication*, Government Engineering College, Barton Hill, Thiruvananthapuram, India
*E-mail: manjusureshvk5@gmail.com

Abstract— **Dysarthria is a motor speech disorder that results in speech difficulties due to the weakness of associated muscles. This unclear speech makes it difficult for dysarthric patients to present himself understood. This neurological limitation is usually occurs due to damages to the brain or central nervous system. Speech therapy can be effectively employed to enhance the range and consistency of voice production and improve intelligibility and communicative effectiveness. Assessing the degree of severity of dysarthria provides vital information on the patient's progress which inturn assists pathologists in arriving at a treatment plan that includes developing automated voice recognition system suitable for dysarthria patients.**

This work performs an exhaustive study on dysarthria severity level classification using deep neural network (DNN) and convolution neural network (CNN) architectures. Mel Frequency Cepstral Coefficients (MFCCs) and their derivatives constitute feature vectors for classification. Using the UA-Speech database, the performance metrics of DNN/CNN based learning models have been compared to baseline classifiers like support vector machine (SVM) and Random Forest (RF). The highest classification accuracy of 97.6\% is reported for DNN under UA speech database. A detailed examination of the performance from the models discussed above reveal that appropriate choice of deep learning architecture ensures better results than traditional classifiers like SVM and Random Forest.

Keywords— **CNN, deep learning, DNN, Dysarthria, motor speech disorder, Random Forest, SVM**

I. INTRODUCTION

In human communication between people, the effective production of speech sound is found to be an important requirement. Sometimes, speech impairment may happen due to the neuromotor disorder which might be due to the weakening of the vital muscles that controls the speech organs. This may be due to neurological damage as in the case of cerebral palsy or any neurodegenerative diseases [1]. Speech quality is affected by lack of articulation intelligibility, poor audibility, unusual prosody, or rhythm and changing speech rate. People with speech impediments may pronounce a few words wrongly or they may have a hard time pronouncing different speech sounds [2].

Dysarthria is a motor speech disorder that occurs due to neurological impairment of motor speech system. Failure of the coordination activity of motor speech system will hinder their control over vocal cords, tongue, throat, lips, and surrounding muscles, for the proper production of speech sounds. Poor articulatory movements results in irregular timing and motion of the lips and jaws which in turn affects accurate delivery of spoken word, thoughsyntactically correct sentences are well formed [3]. Based on the severity of symptoms, speech sessions can be employedto exercise and strengthen muscles of the mouth, ways to slow down speech, strategies to speak louder and techniques to say sounds clearly. People with dysarthria is found to benefit from speech therapy sessions offered by speech – language pathologists to improve communication.

Dysarthria patients are characterized by poor motor skills brought on by lack of muscular coordination. Their shaky hands make it difficult to use interactive devices like keyboard or joystick for communication. This motivated researchers to explore applications based on automatic speech recognition (ASR) for the communication needs of people with dysarthria. Though significant strides have been achieved in automatic speech recognition systems, this technology is still inadequate for people with impaired speech, particularly dysarthria patients.It has been observed that dysarthria speakers are characterizedby abnormal speech patterns within and between individuals and, consequently, difficult to represent in dis-ordered speech ASR systems developed for healthy speakers.

Assessment andclassification of dysarthria severity level is vital for proper diagnosis of disease development, proposing a treatment plan and choosing the right speech therapy sessions to suit the individual needs. Objective (acoustic properties of dysarthric speech) and subjective (perceptual characteristics of dysarthric speech) assessment techniques may be used to gauge the severity of dysarthric subject, with aid of a speech language pathologist [4]. Though perceptive evaluation of dysarthric subject is well documented, such assessments are often described as non-trivial, laborious and highly subjective. Despite these impediments, perceptual method continues to be the de- facto technique in dysarthric

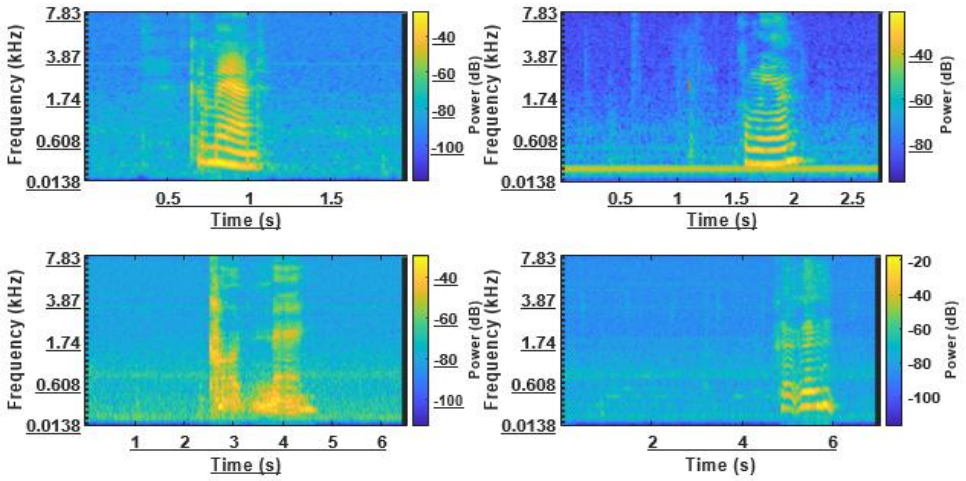

Fig. 1. MelSpectrogram of audio files (a) Very Low, (b) Low (c) Medium (d) High

speech assessment [5]. To counter some of these shortcomings, computer based approach for speech analysis has begun to perform a lead role in dysarthria classification. Computer based ASR systems reduces the costs connected with subjective tests and also offer a repeatable assessment technique. It has also been observed that classi- fication of dysarthria severity enhances the accuracy of ASR systems developed for dysarthria subjects [6]. The proposed work explores traditional and deep learning models for the classification of dysarthric severity.

II. SYSTEM DESCRIPTION

A. Feature Extraction

The purpose of feature extraction is to use a fixed number of signal components to represent a spoken signal. This is because managing all of them would be challenging. some of the acoustic signal's information would be irrelevant to the identification task. The process of feature extraction entails transforming the speech waveform into a parametric represen- tation for additional processing and analysis at a substantially reduced data rate [7].

1) Mel-frequency Cepstral coefficients (MFCC): Vocal muscle coordination affects speech intelligibility, and because of mass tissue changes, MFCCs may detect aberrant vocal fold movements or a lack of vocal fold closure. Frequency-domain features called MFCCs have been demonstrated to enhance speech recognition[8]. With the premise that the speech is wide-sense stationary over brief frames of 10 to 25 ms in length, these are the most widely employed frame-based features. In the mel scale, which simulates the response of the human auditory system, the frequency bands are uniformly spaced. The 13 first static coefficients of the MFCCs, their first derivatives, which indicate velocity (DMFCCs), and their second derivatives, which represent acceleration (DDMFCCs) and produce a vector of 39 MFCCs constitute input vector of each frame.

2) Mel-spectrogram: Many music processing jobs have already successfully used mel-spectrograms. The frequency distribution of a signal over time is shown visually in a spectrogram. A spectrogram where frequencies are scaled to the Mel range is known as a Mel-spectrogram. Fig. 1 shows the mel spectrogram of dyarthria sppeech. The following formula relates the Hertz frequency to the Mel scale, which offers a linear scale for the human auditory system:

$$m = 2595.\ \log_{10}(1 + f/700) \quad (1)$$

where m represents Mels and f represents Hertz.

B. Classifiers

Studying the distinctive features and grouping them into several categories according to their feature set are all part of the classification process [9]. Four classifiers, namely, SVM, Random Forest (RF), DNN and CNN have been used for the proposed task.

1) Support Vector Machine (SVM): SVMs are often used in the categorization of disordered speech because they consistently perform well even with small amounts of speech data, compared to techniques like deep neural networks, which necessitate a substantial quantity of training data. Table [I]

shows SVM parameter setting. N P Narendra and Paavo Alku employed SVM as classifiers in [10], and OpenSMILE, glottal characteristics, and their combinations were used to train several classifiers utilizing reduced and non-reduced feature sets. According to the features that are retrieved from the speech utterance, the SVM classifier can be employed to detect the existence of dysarthria after the completion of training. Experiments show that the glottal characteristics yielded classification accuracies of about 70% for all three categories of speech signal (words, non-words Experiments show that the glottal characteristics yielded classification accuracies of about 70% for all three categories of speech signal (words, non-words, sentences). The results also show that the classification accuracy is increased by including the glottal characteristics in the openSMILE features. Deep learning algorithms use numerous nonlinear processing layers to extract and transform features.

TABLE I. SVM PARAMETER SETTING

Parameter name	Parameter values
SVM type	C-SVM
Regularization parameter C	1
Class number	4
Kernel function	Linear
The degree in kernel function	3
Gamma	Auto
Cache size	200
Tolerance in the termination criteria	0.001

2) Random Forest: An algorithm for supervised learning is random forest. It creates a ”forest” out of a collection of decision trees that were Kim and Chung employed SVM and Random Forest as classifiers. Typically trained using a technique called “bagging”. The main principle of this method is that combining learning models improves the end outcome. Despite the fact that a random forest is merely a collection of decision trees, there are some distinctions. A reliable technique for classification with minimal datasets is the RF algorithm. An RF classifier handles noisy data well and is less affected by outliers. In [11] Hernandez, Kim and Chung employed SVM and Random Forest as classifiers. The depth and quantity of trees are optimized for the RF classifier. The best accuracy was achieved in a forest with 100 trees, with 30 being the ideal depth.

3) Deep Neural Networks (DNN): Complex nonlinear relationships can be modelled by DNNs [12]. A DNN model can simulate feature sets’ high-level abstractions in addition to learning the underlying data structure. DL models perform better than conventional ML networks [13]. Network optimization and loss function minimization are the two main applications of the gradient descent method. DNN models are constructed by stacking n deep layers of Activation functions. The number of neurons in each layer should increase by powers of two as just the model depth increases. Table [II] illustrates the architecture for the experiment. The first layer is composed of the number of nodes scaled by a factor of two which are closest to the input feature vecto*r*.

TABLE II. DNN ARCHITECTURE FOR THE EXPERIMENT

Sl no.	Output shape	Description
1	(40,256)	256 hidden units
2	(40,256)	Drop out (0.25)
3	(256)	256 hidden units
4	(256)	Drop out (0.25)
5	(4)	Time-Distributed Dense

4) Convolutional Neural Networks (CNN): Deep learning is advantageous for use in fields like object classification since CNNs do not require manual feature extraction; instead, they simply learn from the patterns of the training images. Since CNNs do not need to manually extract features; instead, they just learn from the patterns of the training images, deep learning is useful for usage in disciplines like object classification[14][15]. CNN requires significantly less pre-processing as compared to other classification techniques. Additionally, CNN’s layers are three-dimensional with height, width, and depth, in contrast to typical neural networks, and not all of the neurons in one layer are constantly connected to all of the neurons in the following layer. Convolution and pooling layers are alternated for building CNN. Table III shows the proposed CNN architecture. The spectral correlations in the auditory data are captured by the 2D filters employed in CNN’s development. A batch-normalization layer is applied after each of the n stacked 2D convolutional layers with a 3x3 kernel size and a ReLU activation function in order to construct CNN models. The number of feature maps increases by multiples of 2, much like DNN models. Each model uses a 2D max- pooling layer with a pooling size of 2x2 and a dropout layer with a factor of 0.2. In the dense layers, where the number of units reduces as n gets closer to a power of 2, the result is flattened and delivered from here [12].

TABLE III. PROPOSED CNN ARCHITECTURE

Sl no.	Output shape	Description
1	(256,256)	Input layer
2	(256,256,32)	3*3 Convolution, 32 filters. Batch Normalization
3	(128,128,64)	3*3 Convolution, 64 filters. Batch Normalization
4	(64,64,128)	3*3 Convolution, 128 filters. Batch Normalization
5	(32,32,256)	3*3 Convolution, 256 filters. Batch Normalization
6	(16,16,512)	3*3 Convolution, 512 filters. Batch Normalization
7	(8,8,512)	2*2 Max Pooling, Dropout(0.2)
8	(128)	Dense layer, batch Normalization, Dropout(0.2)
9	(64)	Dense layer, batch Normalization, Dropout(0.2)
10	(4)	Softmax

III. PERFORMANCE EVALUATION

A. Dataset

The scarcity of data in the available datasets is one of the main challenges faced by academics studying dysarthric speech. Previous research experiments werespeaker-dependent due to less number of speakers. In order to categorize the severity of dysarthric speech, the study makes use of a number of databases that contain recordings of speakers who have various degrees of dysarthria. We experimented with UA speech database [14]. The details on the dataset is given in Table [IV]. UA-Speech is a compilation of the speech of 19 dysarthric patients and 13 healthy speakers. Only 15 patients' data, nevertheless, are present in this dataset. The Brown corpus contains 155 common words that are repeated three times and correlate to English numerals, computer commands, foreign radio alphabets, and 100 frequent terms. These 465 common terms per speaker make up the 6975 utterances that make up the training data. The corpus additionally includes 300 unique unusual words per speaker that were chosen from children's books that Project Gutenberg has digitally converted [14].

TABLE IV. SEVERITY-WISE SPEAKER DISTRIBUTION

Severity	UA Speech
Very Low	M14, M10, M09, M08, F05
Low	M11, M05, F04
Medium	M16, M07, F02
High	M12, M04, M01, F03

The severity levels for UA-Speech are high, medium, low and very low based on the intelligibility evaluations provided by five untrained listeners.

B. Experimental Framework

The Deep learning network uses an MFCC-based DNN model to precisely classify the severity of dysarthria in patients. The DNN model is created by stacking two dense layers and each such dense layer consisted of 16 units. These dense layers precedes another layer with a dropout of 0.25. A batch size of 32 is used to train the DNN model over a period of 100 epochs. The network is trained using Adam by maximising the category entropy between predictions and targets. The softmax activation function is used to train the model. In Table [V], the DNN configuration is extensively outlined. The entire experiment uses 10% of the corpus for validation. Fig. 3 visualizes the Training and Validation loss and accuracy of DNN model.

A batch-normalization layer comes after each of the four stacked 2D convolutional layers of kernel size 3*3 and ReLU activation function and this make up the CNN model. The model employs a 2D max-pooling layer of size of 2*2, followed by a layer with a dropout of 0.2. With the number of units diminishing in powers of 2, the flattened output of this layer is passed to the dense layers. The 2-D feature maps the 13-dimensional MFCC features that are dispersed along frequency (using the frequency band index) and time (using the frame number). The frame number of the longest word is used to set the frame number of the 2D feature map with zero padding, if necessary[16]. In Table [III], the CNN configuration is extensively outlined. After hyper-parameter tuning, the network is trained using the Adam

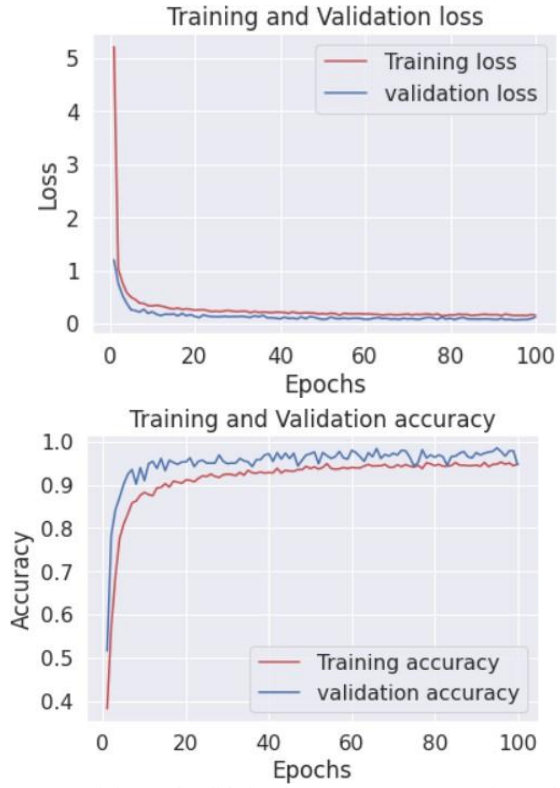

Fig. 2. Training and validation loss (upper pane), Traning and validation accuracy of model (lower pane) of DNN

optimizer witha batch size of 32 across 50 epochs. The softmax activation function is followed by four perceptrons, one for each of the four classes in the output layer.

SVM and random forest (RF) are built as the basic machine learning classifiers. SVM was created for linear kernels, and the best regularization parameter (C) is chosen to be 1, to optimize performance of the classifier on the validation data. While dealing with Radial Basis Functions (RBF) or Poly kernels, the Gamma Hyperparameter is observed to be useful. Since linear kernel has no function, our proposed modeluses the default (auto) gamma value. In Table [I], the SVM configuration is extensively outlined.

Random forest is a meta estimator that employs averagingto increase predicted accuracy and reduce overfitting afterfitting numerous decision tree classifiers to distinct dataset subsamples. In order to achieve optimal accuracy, 100 trees are used in our RF model.

IV. RESULTS AND ANALYSIS

Precision, recall, and F1-score (combines precision and recall scores) are the evaluation or performance metrics

that measures accuracy of deep learning models, which is then compared to that of machine learning models like Support Vector Machine (SVM) and the Random Forest (RF). Overall classification scores obtained for SVM, Random Forest, DNN and CNN are 0.938, 0.934, 0.976, 0.957, respectively. The confusion matrix for the SVM, Random Forest, DNN, and CNN models for the target dataset comprising four severity levels of dysarthria are given in fig. 3.

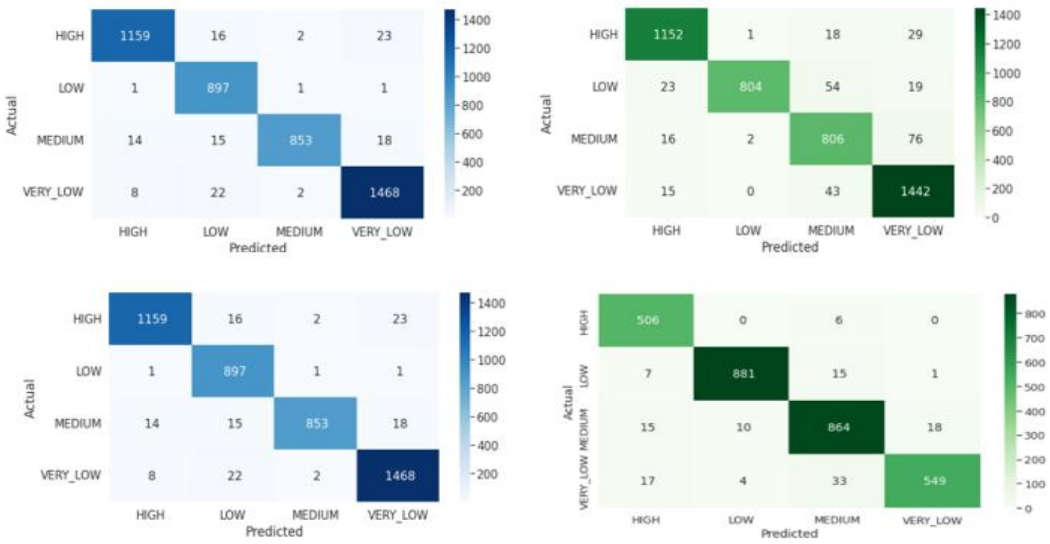

Fig. 3. Confusion Matrix for SVM and RF(upper), DNN and CNN(lower)

TABLE V. PRECISION, RECALL, F1 MEASURE

	MFCC-SVM			MFCC-RF			MFCC-DNN			MEL-SPECTROGRAM-CNN		
SEVERITY	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
VERY LOW	0.954	0.942	0.948	0.920	0.961	0.940	0.972	0.978	0.975	0.966	0.910	0.937
LOW	0.926	0.932	0.929	0.996	0.893	0.942	0.944	0.996	0.969	0.984	0.974	0.979
MEDIUM	0.950	0.935	0.942	0.875	0.895	0.885	0.994	0.947	0.970	0.941	0.952	0.946
HIGH	0.919	0.939	0.929	0.955	0.960	0.957	0.980	0.965	0.973	0.928	0.988	0.957
AVERAGE	**0.937**	**0.937**	**0.937**	**0.936**	**0.927**	**0.931**	**0.972**	**0.971**	**0.971**	**0.954**	**0.956**	**0.954**

It is observed that the proposed deep learning based DNN and CNN models outperforms the machine learning models such as SVM and Random Forest. The proposed experiments shows that the DNN and CNN models are most suited to enhance model accuracy even with complex datasets. Among the experiments, it is seen that DNN model outperforms all other architectures with an accuracy of 97.6% The class-wise accuracy of VERY LOW, LOW, MEDIUM and HIGH severity levels of dysarthria obtained from deep learning approaches outperform the machine learning techniques. The present workreports that DNN and CNN models, yields an accuracy of greater than 95%, across all classes and the proposed models is found to reduce misclassification errors of all classes significantly. It is also observed that the DNN model yielded substantially better results for the severity levels of dysarthria such as Very Low (97%), Low (97%), Medium (97%) and High (97%) in comparison to the corresponding figures 94%, 92%, 94% and 92% obtained for SVM model. The overall results are presented in Table V and is seen that there is an improvement of 4% for the best performing DNN models over the baseline SVM and Random Forest.

From the results it can be inferred that DNN can function as an efficient classifier to provide an unbiased judgement of the dysarthria severity.

V. CONCLUSION

Objective evaluation of the dysarthria speech severity index lends a helping hand to speech pathologist in clinical diagnosis and computer based automatic speech recognition systems developed for dysarthria. Speech severity index classification of dysarthria subjects, employing DNN/CNN based learning models perform well in comparison with baseline classifiers like SVM and Random Forest (RF). In UA-Speech dataset, Mel Frequency Cepstral Coefficiets (MFCCs) are used as feature vectors, and was found effective in distinguishing between speech intelligibility levels of dysarthria individuals. This suggests fast and reliable implementation of automatic dysarthria severity level classification system.

As future work, data augmentation methods can be employed to enhance the accuracy of speaker independent dysarthria severity classification.

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] Bassam Ali Al-Qatab and Mumtaz Begum Mustafa. Classification of dysarthric speech according to the severity of impairment: An analysisof acoustic features. IEEE Access, 9:18183–18194, 2021.

[3] P. A. McRae, K. Tjaden, and B. Schoonings, "Acoustic and perceptual consequences of articulatory rate change in parkinson disease," J. Speech, Lang., Hearing Res., vol. 45, no. 1, pp. 35–50, Feb. 2002.

[4] HM Chandrashekar, Veena Karjigi, and N Sreedevi. Spectro-temporal representation of speech for intelligibility assessment of dysarthria. IEEE Journal of Selected Topics in Signal Processing, 14(2):390–399, 2019.

[5] G. Weismer, R. Martin, and R. Kent, "Acoustic and perceptual ap-proaches to the study of intelligibility," Intell. Speech Disorders, vol.16, pp. 67–118, Apr. 1992.

[6] Myung Jong Kim, Joohong Yoo, and Hoirin Kim. Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models. In Interspeech, pages 3622–3626, 2013.

[7] Sabur Ajibola Alim and N Khair Alang Rashid. Some commonly used speech feature extraction algorithms. IntechOpen London, UK:, 2018.

[8] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimen sionality 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

[9] V. F. M. Mujumdar and R. Kubichek, "Design of a dysarthria classifier using global statistics of speech features," in Int. Conf. Acoust., Speech Signal Process., Dallas, TX, USA, Mar. 2010, pp. 582–585

[10] NP Narendra and Paavo Alku. Dysarthric speech classification using glottal features computed from non-words, words and sentences. In Interspeech, pages 3403–3407.

[11] Hernandez, A.; Kim, S.; Chung, M. Prosody-Based Measures for Auto- matic Severity Assessment of Dysarthric Speech. Appl. Sci. 2020, 10, 6999. https://doi.org/10.3390/app10196999 International Conference on Acoustics, Speech and Signal Processing (ICASSP),

[12] A. A. Joshy, "Automated dysarthria severity classification using deep learning frameworks," in Proc. Eur. Signal Process. Conf., Amsterdam, The Netherlands, 2020, pp. 116–120.

[13] 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.

[14]] H. Kim et al., "Dysarthric speech database for universal access research," in Proc. Interspeech, 2008, pp. 1741–1744. 82:104606, 2023.

[15] Amlu Anna Joshy and Rajeev Rajan. "Automated dysarthria severity classification: A study on acoustic features and deep learning techniques". IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30:1147–1157, 2022.

[16] Amlu Anna Joshy and Rajeev Rajan. Dysarthria severity classification using multi-head attention and multi-task learning, Speech Communication, 147:1–11,2023. 82:104606, 2023.