Low Complexity Model with Single Dimensional Feature for Speech Based Classification of Amyotrophic Lateral Sclerosis Patients and Healthy Individuals

Anjali Jayakumar, Dept. of Electrical Engg., Indian Institute of Science, Bengaluru, India
Email: anjalij@iisc.ac.in
Seena Vengalil, National Institute of Mental Health and Neurosciences, Bengaluru, India
Nalini Atchayaram, National Institute of Mental Health and Neurosciences, Bengaluru, India
Tanuka Bhattacharjee, Dept. of Electrical Engg., Indian Institute of Science, Bengaluru, India
Email: tanukab@iisc.ac.in
Yamini Belur, National Institute of Mental Health and Neurosciences, Bengaluru, India
Prasanta Kumar Ghosh, Dept. of Electrical Engg., Indian Institute of Science, Bengaluru, India

2024 International Conference on Signal Processing and Communications (SPCOM) | 979-8-3503-5045-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/SPCOM60851.2024.10631600

***Abstract*—Lightweight automatic diagnostic tools for Amyotrophic Lateral Sclerosis (ALS) and the associated dysarthria are essential for deployment in resource-limited platforms like mobile phones or general purpose computers. This study performs speech-based low-complexity classification of ALS and healthy subjects by cutting down (1) model complexity and (2) input feature dimensionality. Low complexity Dense Neural Network (DNN) models with 2 or less hidden layers are explored in comparison with the highly complex state-of-the-art Convolutional Neural Network (CNN) with Bidirectional Long Short-Term Memory (BiLSTM) architecture. On the other hand, various temporal statistics (standard deviation, autocorrelation at varying lags) obtained from the commonly used Mel-Frequency Cepstral Coefficients (MFCC) or its individual coefficients are investigated as the low dimensional features. Experiments with 72 ALS and 55 healthy subjects using Spontaneous Speech (SPON) and Diadochokinetic Rate (DIDK) tasks indicate the following. Model complexity reduction with DNN architectures gives comparable, or in some cases better performance, w.r.t. the CNN-BiLSTM model. DNN architectures, with lag 1 autocorrelation of MFCC (along with its delta and double delta coefficients) as the input feature vector for SPON task and standard deviation of the same for DIDK task, can respectively achieve 5.67% and 6.59% higher mean classification accuracies than the CNN-BiLSTM model with entire MFCC sequence as input while causing 99.99% reduction in the model parameter count. Moreover, using single dimensional standard deviation feature of the first delta coefficient for SPON and that of the second delta coefficient for DIDK, together with the DNN models, achieve 94.59% further reduction in the model parameter count while incurring only 1.76% and 5.17% further decrease, respectively, in the classification performance.**

I. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disorder characterized by the progressive degeneration of motor neurons in the brain and spinal cord, leading to muscle weakness, atrophy, and eventual paralysis. In almost 30% of patients with ALS, one of the early signs experienced is dysarthria - a motor speech disorder resulting from the impaired control of the muscles responsible for speech production [1]. As the disease advances, individuals encounter challenges in producing sounds, modulating pitch, and maintaining proper vocal quality. ALS has no known cure, and its diagnosis involves ruling out other similar conditions by clinical assessments, genetic tests, Electromyography (EMG) and Magnetic Resonance Imaging (MRI) scanning, making it tedious and time expensive [2]. Thus, automatic diagnosis of ALS is the need of the hour, considering the disease's progressive nature and the potential of timely intervention to enhance patients' lifespan and the quality of life.

Machine learning models, specifically deep learning algorithms, can play a crucial role in the automatic diagnosis of ALS [3]. Low complexity models ensure accurate results with minimal computational resources, enabling clinical use on platforms like a mobile phone or a general-purpose computer. This is especially useful in remote areas, promoting patient engagement and autonomy in healthcare management [4].

Recent research has focused on using Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) architecture for ALS vs healthy control (HC) classification. The model implemented in [5] outperforms Dense Neural Networks (DNN) and Support Vector Machines (SVM) in 2-class (ALS/PD vs HC) and 3-class (ALS vs PD vs HC) classification. Also, in [6] high accuracy is achieved using CNN with Bidirectional LSTM (BiLSTM) on raw-speech waveform. However, these models have high computational complexity and resource requirement because of the integration of the recurrent layers.

Mel-Frequency Cepstral Coefficients (MFCCs) have been widely used as speech representations for different dysarthric speech applications [5], [7], [8], [9], [10], [11][1]. Joshy et al. [3] reported that MFCC demonstrates low complexity among various features and classifiers for dysarthia severity

[1]A few plots depicting the difference in mel-spectrogram for healthy speech and ALS-induced dysarthic speech is available at https://sites.google.com/view/anjalijayakumar/publications/als-vs-hc

classification. MFCC and pitch are compared as input features for low complexity models under noisy conditions for ALS/PD vs HC classification in [12], and single-dimensional pitch achieved similar performance to multi-dimensional MFCC with greater noise robustness. Illa et al. [7] utilize temporal mean and standard deviation (SD) of MFCC for ALS vs HC classification, whereas Bhattacharjee et al. [8] incorporate mean, median, SD, and Root Mean Square value of MFCC for ALS-induced dysarthria severity classification.

Dimensionality reduction of MFCC has also been studied for other speech applications. According to Shahamiri et al. [13], using 12-D MFCC features alone gives the best accuracy for Artificial Neural Network based ASR system. Sharma et al. [14] have performed MFCC feature dimensionality reduction by exploring individual coefficients of 12-D MFCC for ASR. In [15], the first 3 MFCC coefficients have been found to be sufficient for EMG based classification of ALS vs HC. However, to the best of our knowledge, such an analysis focusing on the individual components of the MFCC feature set has not been conducted in the context of speech based ALS vs HC classification.

This study proposes two distinct methods for achieving low complexity in ALS and HC classification - by model complexity reduction and by feature dimensionality reduction. A CNN-BiLSTM model from [6] is used as the reference model, and three different DNN models of varying complexities as the low complexity classifiers. DNN models have shown promising performances in different tasks involving ALS-induced dysarthic speech, like dysarthria severity prediction [8] and dysarthric speech recognition [16]. Feature dimensionality reduction involves analyzing 12-D MFCCs, their derivatives and individual coefficients as well as their temporal statistics such as SD and auto-correlation with lag 1 and 2 (AC(1) and AC(2)) aiming to capture the temporal relationship within the MFCC frames.

Experiments involving 72 ALS and 55 HC subjects, using Spontaneous Speech (SPON) and Diadochokinetic Rate (DIDK) tasks, demonstrate that reducing model complexity with DNN architectures yields comparable, and in certain instances, better performance than the CNN-BiLSTM model, emphasizing the potential for achieving resource-efficient classification. Notably, using the entire feature set of MFCCs, deltas and double deltas, we achieve a substantial reduction of 99.99% in model parameters, corresponding to a 5.67% increase in mean accuracy for the SPON task using AC(1), and a 6.59% increase for the DIDK task using SD as the temporal statistic. Additionally, feature dimensionality reduction to a single dimensional feature results in an extra 94.59% reduction in model parameters, leading to a 1.76% decrease in SPON task performance with the SD of first delta coefficient, and a 5.17% decrease for the DIDK task with the SD of the second delta coefficient.

II. Dataset

We utilize a dataset including speech samples from 72 (46M + 26F) ALS and 55 (40M + 15F) HC subjects, collected from the National Institute of Mental Health and Neurosciences, Bengaluru, India. The mean (SD) of age in years was 55.36 (10.80) for the ALS and 46.62 (6.85) for the HC speakers. The subjects had three different native languages - Bengali, Kannada, and Telugu. Dysarthria severity of the ALS subjects were rated by three Speech-Language Pathologists as per the 5-point speech component of ALSFRS-R scale [0 (Loss of useful speech) to 4 (Normal speech)] [17]. The mode of these three ratings was regarded as the final dysarthria severity score. There were 10, 13, 17, 17 and 15 subjects with severity score 0, 1, 2, 3, and 4, respectively. We recorded two different speech tasks, SPON and DIDK, from each subject. In SPON task, the subjects were asked to speak about *a festival they celebrate* and *a place they had recently visited* in their respective native languages for about 1 minute each. The mean (SD) duration of the recordings in seconds for SPON task was 60.75 (18.15) for ALS and 59.06 (20.78) for HC. In DIDK task, they were asked to take a deep breath and rapidly repeat a mono-syllabic or a tri-syllabic sequence like '*pa-pa-pa*', '*ta-ta-ta*', '*ka-ka-ka*', '*pataka*' and '*badaga*' [5]. Upto 3 trials were recorded for each sequence depending on a subject's level of comfort. The mean (SD) duration in seconds, combining the three repetitions was 15.92 (9.44) for ALS and 18.80 (8.33) for HC. For the SPON task, recordings of durations 138.7 minutes and 107.28 minutes were obtained from the ALS and the HC classes, respectively, whereas for the DIDK task, those were 93.65 minutes and 85.85 minutes, respectively. More information on data collection procedures and recording setting can be found in [5].

III. Method

The overall pipeline of the ALS vs HC classification methodology used in this study is given in Fig. 1. We extract the features from the speech recordings, segment them into 2-second chunks with 1-second overlap to maintain fixed input contexts, and feed these feature chunks or their various temporal statistics to the classification models. These models are trained using chunk level features and labels. During testing, majority voting is performed over the predictions of all chunks of an utterance to determine the final class label. Two methods are explored for reducing the complexity of the classification process - Model complexity reduction and feature dimensionality reduction.

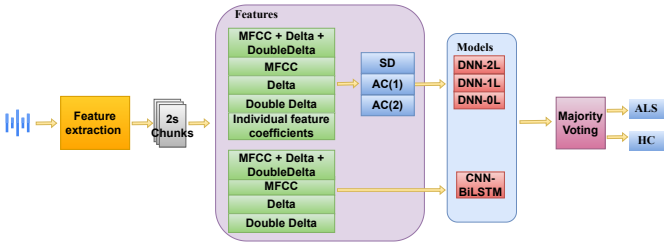


Fig. 1: ALS vs HC classification methodology

A. Model complexity reduction

We start with the state-of-the-art CNN-BiLSTM architecture proposed in [6] and go down to less complex DNN models with decreasing number of dense layers. While CNN-BiLSTM captures both spatial and temporal characters of speech and can perform hierarchical feature extraction, its recurrent layers increase model complexity in terms of number of parameters (#params) and number of floating point operations (FLOPs) required, making it expensive in terms of memory and run-time. A DNN has a much simpler architecture, and further reducing the number of dense layers makes it a more concise model for classification. Our objective is to find a suitable trade-off between model complexity and classification accuracy.

B. Feature dimensionality reduction

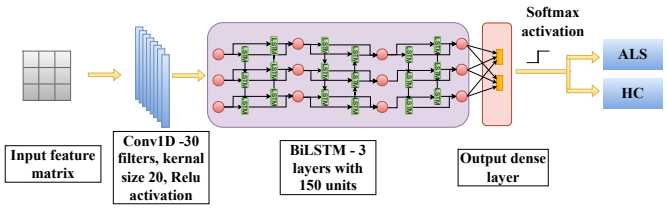
We use MFCCs along with their deltas and double deltas derived from speech as the input features. MFCCs capture the spectral characteristics of speech making them suitable for different speech based classification tasks. The delta and double delta coefficients represent the rate of change and acceleration of MFCCs over time respectively, thus capturing the dynamic aspects of speech spectrum along with the static characteristics [18]. In this work, we explore the potential of reducing the dimensionality of the MFCC feature set in order to enhance the simplicity of the ALS vs HC classification by minimizing unnecessary computational load while retaining the essential information needed for accurate classification. We analyze the performance of MFCC, delta and double delta components separately, as well as their individual coefficients, as compared to the entire feature set comprising all the three components together. We extract the temporal statistics, namely, SD, AC1 and AC2, of different components/coefficients of MFCC[2]. These statistics enable the models to gain more compact insights into the feature variability. The SD calculated over 2s frames of speech utterances reflects the fluctuations in the spectral characteristics and speech dynamics, capturing the potential difference between ALS and healthy speech. These statistics are explored as the input feature vectors for the DNN classifiers, whereas, the MFCC + delta + double delta feature chunks are fed directly to the baseline CNN-BiLSTM classifier.

IV. Experimental Setup

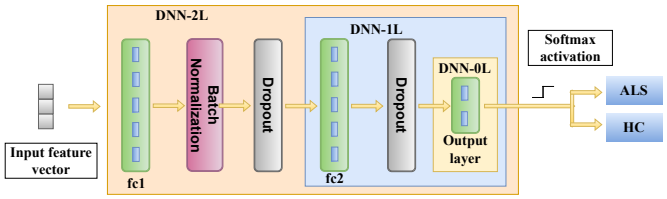
A. Feature extraction

13D MFCCs with their corresponding delta and double delta measures are used as the reference feature. These are computed from each 20 ms speech frame with 10 ms overlap using the Kaldi speech recognition toolkit [19]. The energy coefficients are removed and a 36D feature vector is obtained. To reduce the feature dimensionality, each component including MFCCs, deltas, and double deltas are analyzed individually, and subsequently their individual coefficients are examined. For the DNN model, we incorporate the temporal statistics SD, AC(1) and AC(2) derived from the MFCCs and their deltas, as the features.

[2]The distribution of the 1D features of the two classes are available at https://sites.google.com/view/anjalijayakumar/publications/als-vs-hc



(a) CNN-BiLSTM model



(b) DNN models of varying complexity; here, fc1 and fc2 are fully connected layers with 128 units each, and output layer is a dense layer with 2 units

Fig. 2: Model configurations for ALS vs HC classification

B. Model description

This study explores four distinct models. The reference model utilizes a CNN-BiLSTM architecture from [6]. DNN-2L is a DNN model which employs the architecture proposed in [8]. It has two dense layers of 128 units and ReLU activation functions, along with batch-normalization and dropout layers to prevent overfitting, and an output dense layer of 2 units using softmax activation function. DNN-1L contains a single dense layer with 128 units and ReLU activation with dropout layer, and the output dense layer. DNN-0L merely comprises of the output dense layer. The hyper-parameters are tuned using the validation accuracy. The model configurations are given in Fig. 2. The model complexity across various input features for the four models is given in Table I.

C. Training and Evaluation

For training and evaluating the models, 5-fold cross validation methodology is used. The dataset is split into five folds, with 14-15 subjects from ALS class and 10-13 subjects from

TABLE I: Model complexity for different models and feature dimensions

Model	Feature dimension	#params	FLOPs
CNN-BiLSTM	36	1321832	2400000
	12	1307462	2380000
DNN-2L	36	22,018	21500
	12	18,946	18430
	1	17538	17020
DNN-1L	36	4,994	9730
	12	1992	3580
	1	514	768
DNN-0L	36	74	9540
	12	26	3350
	1	4	514

HC class in each fold. The subjects are distributed evenly in terms of age, gender, language and ALSFRS-R score across the folds. Out of these, three folds are used for training, one for validation and one for testing in each iteration. All models are trained using the Adam optimizer with a learning rate of 0.001 and binary cross entropy loss function. A batch size of 32 is utilized for training. The CNN-BiLSTM model is trained for a maximum of 20 epochs following [6], while the DNN model is trained for a maximum of 100 epochs. Early stopping, based on validation loss with a patience of 8, is implemented during training to prevent overfitting. The mean and SD of balanced accuracy scores obtained on the test sets over the 5 folds are reported as the performance metrics. We perform Wilcoxon signed-rank test [20] at 1% significance level to identify if the classification accuracies obtained for different feature and model configurations are significantly different across 5 folds. For this purpose, we randomly divide the test set of each fold into 3 sub-groups of equal sizes. The 15 accuracy values thus obtained for each feature and model combination are then used for the significance test.

V. Results and Discussion

The classification accuracies of the different models for different input features are given in Fig. 3[3]. It is evident from the plots that the DNN models exhibit performances comparable to the reference CNN-BiLSTM model while utilizing significantly fewer resources. The standard deviation and auto-correlations give similar performances, providing the flexibility to choose between them. Reducing the complexity of the DNN models further aids in less resource utilization while maintaining the level of performance. DNN-0L, which merely consist of a softmax function, works as good as the CNN-BiLSTM for lower delta and double delta coefficients. Fig. 4 demonstrates the accuracies obtained on the individual coefficients of MFCCs, deltas and double deltas, compared to the CNN-BiLSTM performances[3]. It is noteworthy that the first delta and double delta coefficients for the SPON task, and the second delta and double delta coefficients for the DIDK task, achieve higher performances among other coefficients, and is close to the CNN-BiLSTM performance.

In general, the SPON task obtains higher mean accuracy than the DIDK task, with an average difference of 4.84% in case of all coefficients, and 4.24% in case of individual coefficients for DNN model, and 0.75% for CNN-BiLSTM model. For the SPON task, DNN-0L employing AC(1) of the entire feature set achieves a remarkable reduction in #params by 99.99% and FLOPs by 99.60% with a 5.67% increase in accuracy compared to the CNN-BiLSTM model. DNN-0L employing the SD of the first delta coefficient gives a further reduction in complexity of 94.59% in terms of #params and 94.61% in terms of FLOPs, with a mere decrease in accuracy of 1.76%. For DIDK task, DNN-0L employing SD over the entire feature set gives 6.69% increase in accuracy as compared to CNN-BiLSTM, while DNN-0L employing only the SD of the second delta coefficient incurs a decrease in accuracy of only 5.17%. Among the 36 combinations of DNN models and input features, 16 for SPON and 10 for DIDK are statistically similar to, while 6 for SPON and 3 for DIDK outperform the CNN-BiLSTM model at 1% significance level according to Wilcoxon signed rank test.

The findings suggest that, DNN models prove to be a promising alternative to the reference CNN-BiLSTM model for ALS vs HC classification when computational resource management is a concern. Further, the lower coefficients of delta and double delta seem to contain the most discriminative features for the classification enabling a single dimensional feature to be used for accurate and efficient classification.

VI. Conclusion

In this work, we present two approaches for achieving low complexity classification of ALS and HC speech. DNN models can replace the reference CNN-BiLSTM model significantly reducing model complexity. Also, single dimensional feature vectors can capture the informative characteristics for effective classification. This is a promising result for low-resource diagnostic tools for ALS suitable for practical implementations. This work lays a foundation for ALS vs HC classification using single dimensional feature and very low complexity

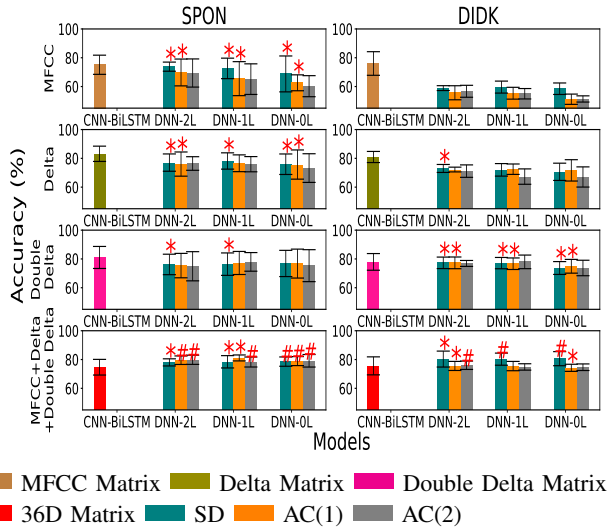
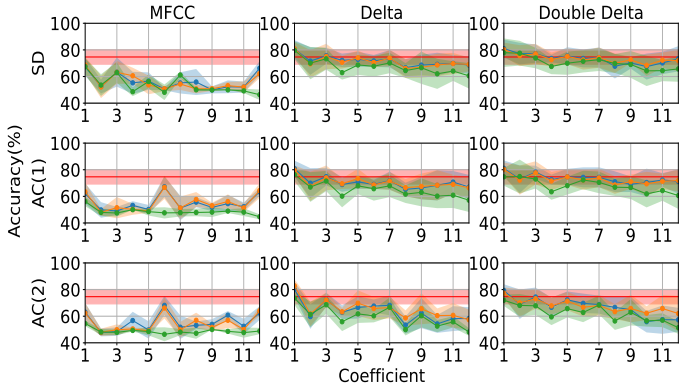
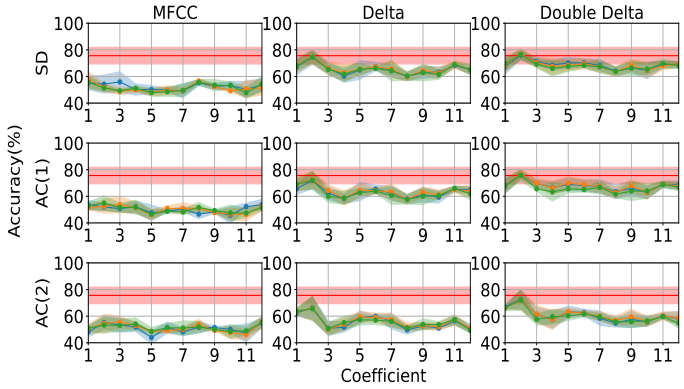


Fig. 3: Mean classification accuracies (in %) over 5-fold cross-validation obtained using different models and input features; here, MFCC matrix, delta matrix and double delta matrix contains 12D MFCC, 12D delta and 12D double delta, respectively, and 36D matrix contains 36 coefficients combining MFCC, delta and double delta; the error bars indicate the SD of the 5-fold classification accuracies ; here, * indicates that the model performance is statistically similar to, and # indicates that the model outperformed the corresponding CNN-BiLSTM model performance at 1% significance level according to Wilcoxon signed rank test

[3]The plots for True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR) are available at https://sites.google.com/view/anjalijayakumar/publications/als-vs-hc



(a) SPON



(b) DIDK

CNN-BiLSTM(MFCC+Delta+DoubleDelta)(#params=1321832)
DNN-2L(#params=17538) DNN-1L(#params=514)
DNN-0L(#params=4)

Fig. 4: Mean classification accuracies (in %) over 5-fold cross-validation obtained using temporal statistics of individual MFCC, delta and double delta coefficients for SPON and DIDK tasks; the shaded regions indicate the SD of the 5-fold classification accuracies

models. Future works may include extending the study to various datasets, and the exploration of additional methodology to further enhance the accuracy of the classification task.

Acknowledgement - We express our sincere gratitude to the Department of Science and Technology (DST), Government of India for supporting this work.

REFERENCES

[1] B. Tomik and R. J. Guiloff, "Dysarthria in amyotrophic lateral sclerosis: A review," *Amyotrophic Lateral Sclerosis*, vol. 11, no. 1-2, pp. 4–15, 2010.

[2] O. Hardiman, L. H. Van Den Berg, and M. C. Kiernan, "Clinical diagnosis and management of amyotrophic lateral sclerosis," *Nature reviews neurology*, vol. 7, no. 11, pp. 639–649, 2011.

[3] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification: A study on acoustic features and deep learning techniques," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1147–1157, 2022.

[4] J. Joe and G. Demiris, "Older adults and mobile phones for health: A review," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 947–954, 2013.

[5] J. Mallela, A. Illa, S. BN, S. Udupa, Y. Belur, N. Atchayaram, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson's disease and healthy controls with CNN-LSTM using transfer learning," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6784–6788.

[6] J. Mallela, Y. Belur, N. Atchayaram, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Raw speech waveform based classification of patients with ALS, Parkinson's disease and healthy controls using CNN-BLSTM," in *Proc. 21^{st} Annual Conference of the International Speech Communication Association, Shanghai, China*, 2020, pp. 4586–4590.

[7] A. Illa, D. Patel, B. Yamini, M. SS, N. Shivashankar, P.-K. Veeramani, S. Vengalil, K. Polavarapui, S. Nashi, N. Atchayaram, and P. K. Ghosh, "Comparison of speech tasks for automatic classification of patients with Amyotrophic Lateral Sclerosis and healthy subjects," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6014–6018.

[8] T. Bhattacharjee, A. Jayakumar, Y. Belur, N. Atchayaram, R. Yadav, and P. K. Ghosh, "Transfer Learning to Aid Dysarthria Severity Classification for Patients with Amyotrophic Lateral Sclerosis," in *Proc. INTERSPEECH*, 2023, pp. 1543–1547.

[9] B. Suhas, J. Mallela, A. Illa, B. Yamini, N. Atchayaram, R. Yadav, D. Gope, and P. K. Ghosh, "Speech task based automatic classification of ALS and Parkinson's Disease and their severity using log Mel spectrograms," in *International conference on signal processing and communications (SPCOM)*. IEEE, 2020, pp. 1–5.

[10] A. Benba, A. Jilbab, and A. Hammouch, "Discriminating between patients with Parkinson's and neurological diseases using cepstral analysis," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 24, no. 10, pp. 1100–1108, 2016.

[11] T. Khan, L. E. Lundgren, D. G. Anderson, I. Nowak, M. Dougherty, A. Verikas, M. Pavel, H. Jimison, S. Nowaczyk, and V. Aharonson, "Assessing Parkinson's disease severity using speech analysis in non-native speakers," *Computer Speech & Language*, vol. 61, p. 101047, 2020.

[12] T. Bhattacharjee, J. Mallela, Y. Belur, N. Atchayarcmf, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Effect of noise and model complexity on detection of Amyotrophic Lateral Sclerosis and Parkinson's disease using pitch and MFCC," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7313–7317.

[13] S. R. Shahamiri and S. S. Binti Salim, "Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach," *Advanced Engineering Informatics*, vol. 28, no. 1, pp. 102–110, 2014.

[14] S. Sharma, M. Kumar, and P. K. Das, "A technique for dimension reduction of MFCC spectral features for speech recognition," in *International Conference on Industrial Instrumentation and Control (ICIC)*. IEEE, 2015, pp. 99–104.

[15] A. B. M. S. U. Doulah and S. A. Fattah, "Neuromuscular disease classification based on mel frequency cepstrum of motor unit action potential," in *International Conference on Electrical Engineering and Information Communication Technology*, 2014, pp. 1–4.

[16] S. Tejaswi and S. Umesh, "DNN acoustic models for dysarthric speech," in *Twenty-third National Conference on Communications (NCC)*. IEEE, 2017, pp. 1–4.

[17] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, B. A. S. Group, and A. complete listing of the BDNF Study Group, "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the neurological sciences*, vol. 169, no. 1-2, pp. 13–21, 1999.

[18] T. Kathiresan and V. Dellwo, "Cepstral derivatives in MFCCs for emotion recognition," in 4^{th} *International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2019, pp. 56–60.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, iEEE Catalog No.: CFP11SRW-USB.

[20] R. Woolson, "Wilcoxon signed-rank test," *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007.