INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Raw speech waveform based classification of patients with ALS, Parkinson's Disease and healthy controls using CNN-BLSTM

Jhansi Mallela[1], *Aravind Illa*[1], *Yamini Belur*[2], *Nalini Atchayaram*[3], *Ravi yadav*[3], *Pradeep Reddy*[3], *Dipanjan Gope*[4], *Prasanta Kumar Ghosh*[1]

[1]EE Department, Indian Institute of Science, India
[2]Department of Speech Pathology and Audiology,
National Institute of Mental Health and Neurosciences, India
[3]Department of Neurology, National Institute of Mental Health and Neurosciences, India
[4]ECE Department, Indian Institute of Science, India

jhansimallela5@gmail.com, aravindece77@gmail.com, prasantag@gmail.com

Abstract

Analysis of speech waveform through automated methods in patients with Amyotrophic Lateral Sclerosis (ALS), and Parkinson's disease (PD) can be used for early diagnosis and monitoring disease progression. Many works in the past have used different acoustic features for the classification of patients with ALS and PD with healthy controls (HC). In this work, we propose a data-driven approach to learn representations from raw speech waveform. Our model comprises of 1-D CNN layer to extract representations from raw speech followed by BLSTM layers for the classification tasks. We consider 3 different classification tasks (ALS vs HC), (PD vs HC), and (ALS vs PD). We perform each classification task using four different speech stimuli in two scenarios: i) trained and tested in a stimulus-specific manner, ii) trained on data pooled from all stimuli, and test on each stimulus separately. Experiments with 60 ALS, 60 PD, and 60 HC show that the frequency responses of the learned 1-D CNN filters are low pass in nature, and the center frequencies lie below 1kHz. The learned representations form raw speech perform better than MFCC which is considered as baseline. Experiments with pooled models yield a better result compared to the task-specific models.

Index Terms: Amyotrophic Lateral Sclerosis, Parkinson's disease, Bidirectional LSTM, Convolutional neural network

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is one of the most prevalent motor neuron diseases that progressively damage both upper (UMN), and lower motor neurons (LMN) in the brain, and spinal cord, respectively. Symptoms appear in the muscles that control speech (dysarthria), and swallowing (dysphagia) or in the limbs occurring in different orders in different patients as the disease progresses [1]. Nearly 25 to 30% of ALS patients have dysarthria [2] as the first or predominant sign in the early stage of the disease resulting in slow, slurred, strained, or whispered speech. Men are affected more frequently than women with a ratio of 3:1 [3]. The life expectancy of a person with ALS, on average, lies in the range of 2 to 5 years from the time of diagnosis and about 5-10% of patients live more than 10 years [4]. The worldwide annual incidence of ALS is 1.9/100,000 [5] whereas in India, it is 1/100,000 with a prevalence rate of 4/100,000 [3]. Revised El Escorial criteria is used for the diagnosis of ALS [6]. The ALS Functional Rating Scale-Revised (ALSFRS-R) is used to monitor the progression of ALS [7]. Similar to ALS, Parkinson's disease (PD) is also a progressive nervous system disorder that affect voluntary movements [8]. It is caused due to the loss of nerve cells in the part of the brain called the substantia nigra which are responsible for producing a chemical called dopamine that acts as a messenger between the parts of the brain and nervous system. Reduced levels of dopamine result in slow and abnormal movements and show symptoms like bradykinesia, tremors in hands and arms, muscle stiffness, balance problems, and dysphagia. About 90% of the individuals with PD experience voice disorders in the early stages, while 45% experience articulation problems, and 20% experience fluency disorders as the disease progresses [9]. It is estimated that about 7-10 million people are affected with PD worldwide. Many people with PD live for 10 to 20 years after being diagnosed.

Currently, there are no specific laboratory tests to diagnose ALS or PD. Tests such as EMG, nerve conduction study, MRI, muscle biopsy have been used to rule out other diseases rather than confirming ALS. ALS and PD are difficult to diagnose. It is also difficult to distinguish the two in the early stages due to some common symptoms. Diagnosis of both ALS and PD is based on clinical observation of about 14 months [10]. By that time, the disease condition may reach its final stages and the survival rate may fall drastically. Presently, there is no treatment to cure both ALS and PD but treatment with the early diagnosis can prevent complications, and slower the progression. Experienced physicians have to diagnose ALS and PD which becomes impossible as the patient population increases. Hence, there is a need for automated methods for early prediction, differentiation of the diseases and monitoring the progression of the diseases which reduces the risk of false or late diagnosis and also avoids clinicians' subjectivity in the diagnosis of the diseases.

The speech production in an individual involves multiple nerves which control the movement of the lips, jaw, tongue, facial muscles, and vocal cords. ALS causes a reduction in the stimulation of these muscles resulting in a slow, effortful, slurred speech, and breathy or hoarse voice. With the reduced levels of dopamine, PD makes patients experience freezing of the jaw, tongue, and lips, repetition of the same words or phrases over and over again and slurred speech. As speech is affected in both the cases, it is often used as a biomarker for automated methods. It is observed that there is reduction in the vowel space area in bulbar ALS patients compared to Healthy control (HC) [11]. The rate of articulatory movement of ALS patients is found to be lower than those with HC [12]. There have been works on speech for automatic classification of ALS and HC with syllable rate and maximum phonation duration based on fractal analysis using diadochokinetic (DIDK) rates as speech stimuli in ALS patients [13]. Bandini et al. [14, 15] used both speech and non-speech tasks from a video based analysis of facial movements and kinematic features of the jaw, and lips for

http://dx.doi.org/10.21437/Interspeech.2020-2221

automatic detection of ALS patients. Aravind et al. [12] used acoustic and articulatory features for classification of ALS patients with HC using support vector machines (SVM) and deep neural networks (DNN). Suhas B N et al. [16] studied the performance of three different speech stimuli namely, spontaneous speech, diadochokinetic rate, and sustained phoneme production in the classification of ALS vs HC using SVM and DNN.

In all the above mentioned approaches (SVM and DNN), they have used hand crafted features like MFCCs, which are computed at a frame level and further statistics are computed from them at a supra-segmental level. MFCCs are cepstral coefficients derived based on human auditory perception, which performs well in speech recognition, speaker verification and several other speech related problems. However, these handcrafted features may not be optimal for classifying patients and healthy controls. Recent advancements in machine learning techniques enable learning task-specific features from the raw waveform using an end-to-end network [17, 18]. We hypothesize that such representation learning discriminates speech of ALS and PD from those of the healthy subjects in a better way compared to handcrafted features. In this work, we consider four speech stimuli, namely, image description (IMAG), spontaneous speech (SPON), diadochokinetic rate (DIDK), and sustained phoneme production (PHON). Following the works in [19, 20], we deploy a 1-D convolutional neural network (CNN) layer to learn representations from raw waveform. It is known that, LSTM networks are well-suited to capture the temporal dynamics on time series data. So, in this work, following 1-D CNN layer, we use a bidirectional long short term memory (BLSTM) network.

2. Data collection

Data was collected from the patients recruited at the National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India following the approval of the NIMHANS ethics committee. The data was collected once the patients were diagnosed to have ALS (using El Escorial criteria) or PD by neurologists at NIMHANS. Severity ratings for ALS and PD are given by 5 speech-language pathologists (SLP) from the Speech Pathology and Audiology Department, NIMHANS as per ALSFRS-R (0:Loss of useful speech to 4:Normal) [7], UPDRS-III (0:Normal to 4:Unintelligible speech) [21] for ALS and PD patients, respectively. Here ALSFRS(4), and UPDRS-III(0) indicate that there is no human intelligible loss of function in speech but patients with these scores have other symptoms of the diseases. Due to the lack of PD subjects with severities above UPDRS-III of 2, we considered a UPDRS-III range of 0-2. We considered the majority of severity scores by five SLPs as the final score.

For the experiments in this study, we used 60 ALS, 60 PD, and 60 HC subjects' speech data. For ALS, the mean age of 30 men is 58.60 (range of 33-76) and that of 30 women is 56 (range of 38-75). For the case of the PD, the mean age of 34 men is 58.22 (range of 34-78) and that of the 26 women is 56.99 (range of 36-74). For HC, the mean age of 30 men is 44.21 (range of 26-48) and that of 30 women is 46.93 (range of 31-65). The recruited subjects in this work come from six different native languages, namely, Bengali, Kannada, Tamil, Hindi, Telugu, and Odiya in an approximately equal proportion. None of the subjects in the HC group had any history of symptoms related to ALS or PD. More details of the dataset used in this study can be found in [22]. Zoom H-6 recorder with XYH-6 X/Y capsule high-quality unidirectional microphone [23] was used to record the speech data from a distance of 2 feet from the subject at a sampling rate of 44.1 kHz.

As mentioned earlier, we considered 4 different speech stimuli, namely IMAG, PHON, DIDK, and SPON. In IMAG, the subjects were asked to describe images (ranging from 30 to 70 depending on the subjects' comfort level) shown to them on a computer screen in front of them. Subjects described in their own native language. In PHON, sustained phonemes are produced corresponding to 5 vowels (*/a/, /i/, /o/, /u/, /æ/*) and 3 fricatives (*/s/, /sh/, /f/*). DIDK consists of recordings of a sequence of monosyllabic targets (*"pa-pa-pa", "ta-ta-ta", "ka-ka-ka"*) and their combinations (*"pataka"* and *"badaga"*). In SPON, the subjects were asked to spontaneously talk about two events (*"a place that they have recently visited"*, *"a festival they celebrate"*), each for one minute in their native language. In PHON and DIDK, each phoneme (or fricative or syllable sequence) was repeated thrice in succession. For the SPON task, preparation time was given to the subject before recording until they were ready to speak. The choice and the significance of all the above speech stimuli are explained in [16, 22]. The total duration of the recordings from all the subjects for IMAG, PHON, DIDK, and SPON is 12.83, 5.79, 4.65, and 5.62 hours, respectively.

3. Proposed Approach

Fig. 1 and Fig. 2 illustrate the 1D-CNN based representation learning and BLSTM based classification tasks (ALS vs HC, PD vs HC, and ALS vs PD), respectively.

Learning representations from the raw speech waveform:
Given a speech utterance, we slice it into overlapping frames of length f_l and shift f_s and remove the silence frames using voice activity detection (VAD) [24]. As VAD provides variable length speech segments, we further chunk them using a window length of c_l and shift c_s. The resulted chunks are given as input to 1-D CNN layer which has n_m number of filters with a filter length of l_m. The n_m number of filters and the corresponding bias vector are denoted as $\mathbf{F}=\{\mathbf{F}_k\}_{k=1}^{n_m}$ (where, $\mathbf{F}_k \in \mathbb{R}^{1\times l_m}$), and $\mathbf{b}\in \mathbb{R}^{n_m}$, respectively. The output of each filter in the CNN layer for a speech chunk $\mathbf{s}_n \in \mathbb{R}^{1\times f_l}$ with index n, is computed by

$$\mathbf{O}_n = \sigma(\log(|\mathbf{F} * \mathbf{s}_n + \mathbf{b}|)) \quad (1)$$

where, $\mathbf{O}_n \in \mathbb{R}^{(f_l - l_m + 1)\times n_m}$, σ denotes non-linear activation function and * indicates the convolution operation. With reference to the work in [25], log operation is applied on the absolute value of CNN filter before performing non linear activation. Max-pooling of size ($f_l - l_m$+1) is applied over time to reduce the output size of the CNN layer $\mathbf{O}_n$ which could help in discarding the short term phase information and results in $1\times n_m$ dimensional output $\mathbf{o}_n$. As shown in Fig. 1, for a given speech chunk, the representations are learned using a 1-D CNN and max-pooling layers with a batch normalization layer in between.

Classification using BLSTM network:
In this work, we propose a cascaded 1-D CNN and BLSTM network for the classification of patients with ALS, PD, and HC using raw speech waveform as shown in Fig. 2. The BLSTM network comprises of 3 BLSTM layers with *tanh* as activation function. The output of the BLSTM network is fed to a dense layer and the softmax layer generates a 2 dimensional output with probabilities using the dense layer output. Binary cross-entropy loss function is used to optimize the weights of the BLSTM network. Fig. 2 illustrates the block diagram of the proposed approach. For the given speech waveform, the representations are learned using 1-D CNN and given as input to the BLSTM layers which are used to capture the temporal dynam-

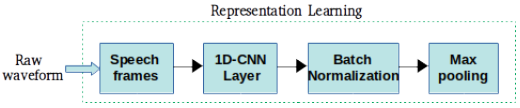

Figure 1: *Illustration of the proposed approach of learning representations from the raw waveform.*

ics of the sequence. The last hidden state output of the BLSTM layer is given to the dense layer. Finally, we use softmax as an activation function at the output of the dense layer to classify patients of ALS or PD or HC. The classification tasks using BLSTM are performed at chunk level. Decisions from all the speech chunks are combined to perform majority voting and obtain a decision on the entire speech utterance.

4. Experimental setup

For experiments with raw speech waveform, we first downsample the data from 44.1kHz to 8kHz. The 1-D time-series data is then framed with f_l=20ms (160 samples) and f_s=10ms (80 samples). Following VAD, we chunk speech segment with duration c_l=2sec (200 frames), and shift c_s=1 sec (100 frames). We perform experiments with a 5-fold cross-validation setup, each fold comprising 12 subjects from each of the three classes, namely ALS, PD, and HC. Among these five folds, three folds are used for training, one for development, and one for the test. In all the five folds, the data is equally distributed in terms of age, gender, severity, and language. The optimal number of CNN filters n_m is experimentally determined in each fold separately by maximizing the performance on the validation set. For this purpose, we experimented with different choices of the number of CNN filters n_m=256, 128, and 64, with filter length l_m= 120. For BLSTM network, we choose three BLSTM layers each with 150 units, and *tanh* as activation function.

To compare the performance of the proposed approach, we perform classification using MFCCs with the proposed BLSTM network as a baseline. Similar to the proposed approach using learned representations, for MFCCs also, we perform chunk level (2 sec) classification and obtain utterance level decision using majority voting. We use classification accuracy and p value from sign-rank test [26] as evaluation metrics. The sign-rank test is performed using the five folds' classification accuracies of the baseline and the proposed approach where each test fold is again split into three sub folds since a minimum of five variables are required for this sign-rank test.

5. Results and Discussion

We present the results for three different classification tasks (ALS/HC, PD/HC, ALS/PD) each with four different speech stimuli (IMAG, DIDK, PHON, and SPON) in terms of average classification accuracy along with standard deviation (SD). In each classification task, we present the results for all speech stimuli in two scenarios: (i) stimuli-specific model, (ii) pooled model. In the first scenario, we train the classifier in a stimulus-specific manner (e.g., ALS/HC, for IMAG, train with IMAG data and test on IMAG data) whereas in the second scenario, we train a classifier with data of all the speech stimuli (pooled model) and test on each stimulus separately. We also present an analysis of the use of pre-emphasis (PE) [27, 20] in the proposed approach and compare it with baseline. Table 1 consists of three major columns where each column presents the results of different classification tasks (column1- ALS/HC, column2- PD/HC, column3- ALS/PD). As mentioned above, the results in each column divide into two scenarios, namely stimulus-specific and pooled.

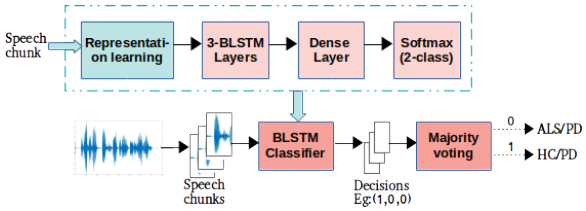

Figure 2: *Illustration of proposed classifier with raw waveform using BLSTM.*

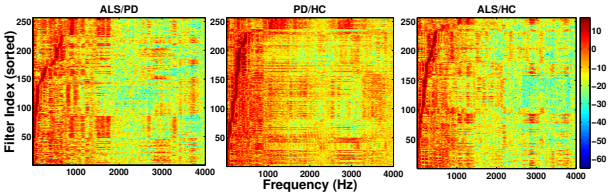

Figure 3: *Magnitude response of 256 filters for ALS/PD, PD/HC, and ALS/HC without using pre-emphasis.*

Classification between ALS and HC (ALS/HC): In the case of ALS/HC (first column in Table 1), we observe that, the proposed approach using the task-specific model shows significant improvement in all speech stimuli compared to using MFCCs. Though MFCCs with the pooled model perform better than MFCCs with the stimulus-specific model, the proposed method with the pooled model performs even better in all cases except for SPON. The proposed approach provides relative improvements of 7.88%, 4.67%, 3.11%, and 1.76% in IMAG, PHON, DIDK, and SPON stimuli, respectively compared to the baseline (MFCCs with stimulus specific model).

Classification between PD and HC (PD/HC): In PD/HC (second column in Table 1), compared to the baseline (MFCCs), the proposed approach using a stimulus-specific model show significant improvement in all speech stimuli using both stimulus-specific and pooled model. Interestingly, in PHON, the proposed approach with the pooled model significantly improves about 8.61% and 16.79% compare to the proposed approach with the stimulus-specific model and baseline, respectively. The proposed approach provides relative improvements of 12.37%, 16.79%, 11.19%, and 8.13% in IMAG, PHON, DIDK, and SPON stimuli, respectively compared to the baseline (MFCCs with stimulus specific model).

Classification between ALS and PD (ALS/PD): In ALS/PD classification (third column in Table 1), compared to baseline (MFCCs), the proposed approach using the stimulus-specific model shows significant improvement in all speech stimuli except in PHON. Though there is no improvement in PHON, it improves in SPON by 14.37% and 16.85% using the stimulus-specific model and pooled model, respectively. Overall there is an improvement of 6.40%, 0.03%, 2.95%, and 16.85% in IMAG, PHON, DIDK, and SPON, respectively, by using the proposed representation learning over the baseline. The reason behind the minute improvements in ALS/PD compared to ALS/HC, and PD/HC could be because of some similarities between the speech during the progression of these two diseases. Those similarities pose a challenge to the classifier to learn disease-specific characteristics from speech.

In Table 1 the values in blue color represent the cases where p <0.05 from the statistical test indicating significant improvement over baseline MFCC features. From the classification accuracies of the pooled models, it is shown that the classifier is learning the features of ALS and PD unlike learning the stimulus-specific features. We also observe that in most cases

Table 1: *Average classification accuracy (SD in brackets) of the proposed approach and baseline (MFCC) for three classification tasks, with four different speech stimuli. Blue entries indicate the cases where the proposed approach performs significantly ($p < 0.05$) better than MFCC.*

Classifiers	**ALS/HC**				**PD/HC**				**ALS/PD**			
Speech stimuli	IMAG	PHON	DIDK	SPON	IMAG	PHON	DIDK	SPON	IMAG	PHON	DIDK	SPON
	Stimulus-specific model				Stimulus-specific model				Stimulus-specific model			
MFCC	90.14 (4.85)	87.67 (1.10)	93.63 (3.35)	96.51 (3.37)	84.91 (2.13)	65.39 (2.92)	81.89 (5.32)	90.14 (2.92)	71.88 (3.08)	72.72 (2.91)	79.73 (3.90)	68.25 (10.06)
Raw- PE(0.97)	96.98 (1.72)	87.81 (2.69)	94.86 (1.65)	98.27 (1.84)	96.11 (1.57)	59.45 (7.01)	90.26 (3.30)	96.20 (2.67)	74.73 (5.90)	72.75 (4.61)	81.22 (4.96)	82.62 (9.85)
Raw- No PE	97.31 (1.75)	89.47 (3.82)	96.39 (1.94)	96.16 (2.80)	95.01 (2.17)	73.57 (5.05)	89.51 (2.20)	95.71 (4.10)	76.87 (8.20)	72.67 (5.76)	76.43 (2.75)	78.00 (7.84)
	Pooled model				Pooled model				Pooled model			
MFCC	93.08 (1.2)	85.56 (5.0)	95.88 (2.2)	98.69 (1.9)	91.51 (2.06)	70.25 (8.93)	89.87 (3.48)	95.34 (2.73)	74.56 (4.83)	66.54 (3.81)	78.71 (2.08)	73.47 (10.09)
Raw- PE(0.97)	97.97 (1.68)	90.98 (1.91)	96.74 (2.31)	98.27 (2.36)	96.95 (2.92)	78.61 (6.46)	92.59 (2.63)	98.27 (3.78)	75.08 (4.85)	72.40 (6.36)	82.68 (6.09)	85.10 (8.61)
Raw- No PE	98.02 (1.75)	92.34 (1.43)	96.73 (2.45)	97.86 (2.38)	97.28 (2.01)	82.18 (9.86)	93.08 (4.92)	97.41 (3.88)	78.28 (7.32)	69.33 (4.67)	82.09 (3.16)	82.36 (7.84)

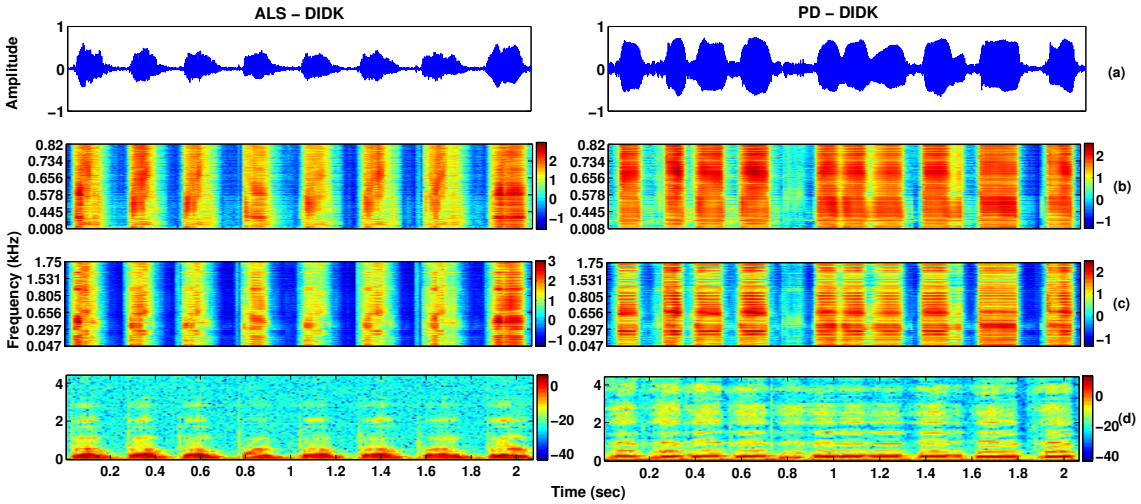

Figure 4: *Illustration of /pa/ sequence spoken by ALS and PD patients using (a) speech waveform, (b) 1-D CNN output without PE, (c) 1-D CNN output with PE, and (d) spectrogram*

the results without PE are performing better than using PE on the raw waveform. The six pooled CNN-BLSTM results (3 classifications tasks with and without PE) reported in Table 1 are obtained from different choices of n_m=64, 128, and 256, out of which 37.5%, 33.3%, and 29.16% of models perform the best with 64, 128, and 256 filters, respectively. We plot the magnitude response of 256 learned filters without PE to understand the frequency of the 1-D CNN filters in the representation learning. In Fig. 3 the center frequencies of the filters and the corresponding filter indices are indicated in x-axis, and y-axis, respectively. The filter indices are sorted according to the center frequencies and the colour intensity variation indicates the magnitude response of the filters. From Fig. 3, it is observed that the frequency responses are low pass in nature and are centered at a frequency below 800Hz, 500Hz, and 400Hz for ALS/PD, PD/HC, and ALS/HC, respectively. In Fig. 4, we plot the output of the 1-D CNN layer with PE, without PE, and spectrogram for a sample of DIDK sequence, for both ALS and PD patients saying repetitions of phoneme sequence /pa/. From Fig. 4, it is observed that the CNN filters of 1-D CNN layer are learning to identify vowel pattern /a/ and similar observations are found in other speech tasks too. This could allow the classifier to extract features related to speech rate cues by enhancing vowels in low-frequency regions. It is known that speech rate tends to decline in dysarthric speech compared to HC. Hence, the representations learned by 1-D CNN from raw waveform help to achieve better classification accuracies compared to the baseline (MFCC).

6. Conclusion

In this work, we propose a cascaded architecture comprising 1-D CNN layer and BLSTM layers for the classification of patients with ALS, PD, and HC using raw speech waveform. From the analysis of the learned CNN filter response, it is revealed that the filters are low pass in nature and the center frequencies lie below 800Hz, 500Hz, and 400Hz for ALS/PD, PD/HC, and ALS/HC, respectively. Experiments with stimulus-specific and pooled models revealed that the pooled model performs better than stimulus-specific ones. A comparison of the proposed approach with baseline acoustic features (MFCC) revealed that the proposed approach significantly performs better than baseline.

7. Acknowledgements

We thank Ms Renuka and Ms Hima Jyothi for their valuble technical assistance. This work is supported by the Department of Science and Technology (DST), Govt. of India.

8. References

[1] S. Meshinchi and R. J. Arceci, "Prognostic factors and risk-based therapy in pediatric acute myeloid leukemia," *The Oncologist*, vol. 12, no. 3, pp. 341–355, 2007.

[2] R. D. Kent, R. L. Sufit, J. C. Rosenbek, J. F. Kent, G. Weismer, R. E. Martin, and B. R. Brooks, "Speech deterioration in amyotrophic lateral sclerosis: A case study," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 6, pp. 1269–1275, 1991.

[3] A. Nalini, K. Thennarasu, M. Gourie-Devi, S. Shenoy, and D. Kulshreshtha, "Clinical characteristics and survival pattern of 1153 patients with amyotrophic lateral sclerosis: experience over 30 years from India," *Journal of the neurological sciences*, vol. 272, no. 1-2, pp. 60–70, 2008.

[4] A. Chiò, G. Logroscino, O. Hardiman, R. Swingler, D. Mitchell, E. Beghi, B. G. Traynor, E. Consortium *et al.*, "Prognostic factors in ALS: a critical review," *Amyotrophic Lateral Sclerosis*, vol. 10, no. 5-6, pp. 310–323, 2009.

[5] K. C. Arthur, A. Calvo, T. R. Price, J. T. Geiger, A. Chio, and B. J. Traynor, "Projected increase in amyotrophic lateral sclerosis from 2015 to 2040," *Nature communications*, vol. 7, no. 1, pp. 1–6, 2016.

[6] M. Gourie-Devi, A. Nalini, and S. Sandhya, "Early or late appearance of "dropped head syndrome" in amyotrophic lateral sclerosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 74, no. 5, pp. 683–686, 2003.

[7] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, B. A. S. Group, A. complete listing of the BDNF Study Group *et al.*, "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.

[8] L. M. De Lau and M. M. Breteler, "Epidemiology of parkinson's disease," *The Lancet Neurology*, vol. 5, no. 6, pp. 525–535, 2006.

[9] G. Moya-Galé and E. S. Levy, "Parkinson's disease-associated dysarthria: prevalence, impact and management strategies," *Research and Reviews in Parkinsonism*, vol. 9, pp. 9–16, 2019.

[10] B. R. Brooks, R. G. Miller, M. Swash, and T. L. Munsat, "El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis," *Amyotrophic lateral sclerosis and other motor neuron disorders*, vol. 1, no. 5, pp. 293–299, 2000.

[11] B. Yamini, N. Shivashankar, and A. Nalini, "Vowel space area in patients with Amyotrophic Lateral Sclerosis," *Amyotrophic Lateral Sclerosis*, vol. 9, no. 1, pp. 118–119, 2008.

[12] A. Illa, D. Patel, B. Yamini, N. Shivashankar, P. K. Veeramani, K. Polavarapui, S. Nashi, A. Nalini, P. K. Ghosh *et al.*, "Comparison of speech tasks for automatic classification of patients with amyotrophic lateral sclerosis and healthy subjects," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6014–6018.

[13] A. S. Mefferd, G. L. Pattee, and J. R. Green, "Speaking rate effects on articulatory pattern consistency in talkers with mild ALS," *Clinical linguistics & phonetics*, vol. 28, no. 11, pp. 799–811, 2014.

[14] A. Bandini, J. R. Green, B. Taati, S. Orlandi, L. Zinman, and Y. Yunusova, "Automatic detection of amyotrophic lateral sclerosis (ALS) from video-based analysis of facial movements: speech and non-speech tasks," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 150–157.

[15] A. Bandini, J. R. Green, L. Zinman, and Y. Yunusova, "Classification of Bulbar ALS from Kinematic features of the Jaw and Lips: Towards computer-mediated assessment," in *INTERSPEECH*, 2017, pp. 1819–1823.

[16] B. Suhas, D. Patel, N. Rao, Y. Belur, P. Reddy, N. Atchayaram, R. Yadav, D. Gope, and P. K. Ghosh, "Comparison of speech tasks and recording devices for voice based automatic classification of healthy subjects and patients with amyotrophic lateral sclerosis," *Proc. Interspeech 2019*, pp. 4564–4568.

[17] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *arXiv preprint arXiv:1304.1018*, 2013.

[18] H. Muckenhirn, M. M. Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4884–4888.

[19] H. Dinkel, N. Chen, Y. Qian, and K. Yu, "End-to-end spoofing detection with raw waveform CLDNNS," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4860–4864.

[20] A. Illa and P. K. Ghosh, "Representation learning using convolution neural network for acoustic-to-articulatory inversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5931–5935.

[21] D. J. Gelb, E. Oliver, and S. Gilman, "Diagnostic criteria for Parkinson's disease," *Archives of neurology*, vol. 56, no. 1, pp. 33–39, 1999.

[22] J. Mallela, A. Illa, B. Suhas, 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6784–6788.

[23] "Zoom high-quality unidirectional microphone, available online: https://www.zoom-na.com/, last accessed:10/15/2019." [Online]. Available: https://www.zoom-na.com/products/product-accessories/zoom-xyh-6-xy-stereo-microphone-capsule

[24] V. Panayotov, M. Maciejewski, and D. Povey, "Voice activity detection." [Online]. Available: https://github.com/kaldi-asr/kaldi/blob/master/src/ivector/voice-activity-detection.h

[25] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs." in *Interspeech*, 2016, pp. 3434–3438.

[26] D. Rey and M. Neuhäuser, *Wilcoxon-Signed-Rank Test*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1658–1659. [Online]. Available: https://doi.org/10.1007/978-3-642-04898-2_616

[27] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, "End-to-end speech recognition from the raw waveform," *arXiv preprint arXiv:1806.07098*, 2018.