EFFECT OF NOISE AND MODEL COMPLEXITY ON DETECTION OF AMYOTROPHIC LATERAL SCLEROSIS AND PARKINSON'S DISEASE USING PITCH AND MFCC

Tanuka Bhattacharjee[⋆], *Jhansi Mallela*[⋆], *Yamini Belur*[†], *Nalini Atchayaram*[‡], *Ravi Yadav*[‡], *Pradeep Reddy*[‡], *Dipanjan Gope*[⋆⋆], *Prasanta Kumar Ghosh*[⋆]

[⋆]EE Department, [⋆⋆]ECE Department, Indian Institute of Science, Bengaluru 560012, India
[†]Department of Speech Pathology and Audiology , [‡]Department of Neurology,
National Institute of Mental Health and Neurosciences, Bengaluru 560029, India

ABSTRACT

Dysarthria due to Amyotrophic Lateral Sclerosis (ALS) and Parkinson's disease (PD) impacts both articulation and prosody in an individual's speech. Complex deep neural networks exploit these cues for detection of ALS and PD. These are typically done using recordings in laboratory condition. This study aims to examine the robustness of these cues against background noise and model complexity, which has not been investigated before. We perform classification experiments with pitch and Mel-frequency cepstral coefficients (MFCC) using models of three different complexities and additive white Gaussian noise in four signal-to-noise-ratio (SNR) conditions. The findings are as follows: 1) In clean condition, pitch performs similar to MFCC across most model complexities considered, suggesting that one-dimensional pitch pattern provides discriminative cues for the classification to an extent equal to that of multi-dimensional MFCC, 2) Similar trend is observed in noisy cases when classifiers are trained and tested in matched noise and SNR conditions, 3) When the classifiers trained on clean data are applied in noisy cases, pitch based average classification accuracies are found to be 20.09% and 24.73% higher than those using MFCC for ALS vs. healthy and PD vs. healthy, respectively, suggesting robustness of pitch based classifier against noise and model complexity.

Index Terms— Amyotrophic Lateral Sclerosis, Parkinson's disease, Pitch, Mel-frequency cepstral coefficients, Model complexity, Noise.

1. INTRODUCTION

Amyotrophic Lateral Sclerosis (ALS) [1] and Parkinson's disease (PD) [2] are incurable neuro-degenerative disorders which affect muscle movements. Early detection is critical in both cases for timely commencement of therapeutic measures which can prolong the life expectancy of the patients and enhance the quality of their lives. Unfortunately, there exists no single blood or laboratory test that can confirm ALS or PD. Diagnosis is done based on subjective assessment of symptoms and medical histories, along with various neurological and physical examinations. Thus the process is highly time expensive. Diagnosis of ALS based on El Escorial criteria [3] requires a median diagnosis time of 14 months [4], where the average life expectancy of these patients is only 2-5 years from the time of disease onset [5]. Moreover, the clinicians' subjectivity and perception being involved, the diagnosis process may be susceptible to various sources of errors and biases. Thus, accurate automated diagnostic tool is a need of the hour.

Dysarthria is experienced by almost all individuals suffering from ALS and PD, with it being the early sign of ALS in about 30% of the patients [6, 7]. Different aspects of speech functions including articulation, respiration, phonation and prosody are reported to get affected in these diseases [8, 9]. Various cues descriptive of these speech components have been studied in the literature for classifying healthy controls (HC) and patients with ALS/PD. Deep neural network (DNN) based classifiers can exploit the information present in these cues to perform the classification with high degree of accuracy. Mel frequency cepstral coefficients (MFCC), representative of spectral characteristics and articulation, has been widely used for this purpose [10, 11, 12]. Suhas et al. [10] employed dense neural network to perform the classification, whereas Mallela et al. [11] explored 1D-convolutional neural network (CNN) and long short term memory (LSTM) based classifier using transfer learning approach. Log Mel spectrograms have been found to perform better than MFCC in the context of 2D-CNN based automatic classification and severity prediction of ALS and PD [13]. Cepstral separation difference (CSD) indicative of phonation characteristics and spectral dynamics together with fundamental frequency variation as markers of respiration and prosody have been used in [12] for PD classification and its severity prediction. The authors employed random forest classifier in this work. Vashkevich et al. [14] have proposed novel features based on analysis of the envelope and formant structures of vowels for automatic diagnosis of bulbar ALS. In a recent work, Mallela et al. [15] have achieved very high classification performance by directly using raw speech waveform in a CNN-Bidirectional LSTM based framework.

Although the DNN based algorithms described above are reported to achieve high degree of classification accuracy, these models are very expensive in terms of both run-time and memory requirement. Hence powerful computing resources are crucial for evaluating these models, which imposes restrictions on the deployment of such models in practice. Low complexity classification models suitable for running on-device in mobile phones or general purpose computers might be more appropriate in order for it to be useful to the majority of the population. Behaviour of different speech cues under the constraint of low complexity classifiers is not well analyzed yet.

Experiments related to the existing classification methods have been mostly carried out on clean speech recorded in controlled and noise-free laboratory or hospital environments. However, presence of background noise in the speech data is inevitable while deploying these systems in practical scenarios like home-based monitoring. Noise often buries or alters the distinctive information present in the signal, thereby leading to mis-classification which may prove to be fatal in cases. Robustness of the speech cues against different vari-

ants of noise is yet another unexplored area in this field of research.

This work aims to explore the behaviour of different speech cues under the influence of background noise and the constraint of low complexity classifier. We particularly focus on pitch and MFCC for this purpose. The suitability of MFCC for classification of HCs and patients with ALS/PD is already well established. Pitch, on the other hand, has not been explicitly used as a feature of speech in this context, though it has been reported to get affected as a prosodic component of speech in these diseases [16, 17]. We experiment with these two speech cues using CNN-LSTM classifier [11] of three different levels of complexity at four different signal-to-noise ratio (SNR) conditions. Though MFCC is proven to perform well in literature, it is high dimensional and prone to alter in additive noise conditions. Pitch, on the contrary, is single dimensional and noise robust pitch estimation algorithms like PEFAC [18] and SWIPE [19] exist. Hence, pitch could be better suited under the constraints of noise and model complexity. We do not consider fusion of multiple features in this work as that increases the computational complexity and, hence, is less suitable for resource constrained applications.

2. DETAILS OF THE STUDY

We consider two different classification tasks in this study, namely, ALS vs. HC and PD vs. HC classifications, the features and classifiers for which are summarized in the subsections below along with various noise conditions and classifier complexities.

2.1. Features

Pitch, associated with prosody, and MFCC, representative of spectral properties, are considered as features for classification in this paper. Pitch pattern captures speaking rate along with other prosodic features. Since both ALS and PD lower the speaking rate of an individual, cues related to these diseases can be learned from pitch patterns using data-driven DNN models. Another typical symptom of these two diseases is muscle weakening. This affects the precise movements of articulators like lips and tongue, resulting in improper vocal tract shape, thereby altering spectral characteristics. MFCC is used to learn cues indicative of this aspect of the impairment.

2.2. Classifier

A CNN-LSTM based deep neural network classifier, following [11], is adapted in this work. ALS and PD affect the paralinguistic characteristics of speech in a suprasegmental level. The CNN-LSTM model can extract these suprasegmental features from frame level acoustic features in a data-driven manner. The classifier, as shown in Fig. 1, operates on feature chunks obtained from speech segments with overlapping frames. The initial 1D-CNN layer, followed by maxpooling, extracts local and time-invariant patterns from frame level acoustic features using temporal convolution. The temporal dynamics of the feature sequence is then captured by the LSTM layer. The hidden state outputs of the LSTM at the last frame index are then passed through a dense layer with softmax activation to obtain the decisions (class labels). During inferencing, majority voting is performed over all segments of test speech utterance.

2.3. Classifier Complexity

High model complexity in terms of both memory and runtime becomes a major concern while deploying complex deep neural network models in systems with limited computational resource like mobile phones. The memory complexity is quantified by the number of parameters (#params) of the network, while the time complexity is measured as the number of floating point operations (FLOPs) needed by the network. In this work, we analyze CNN-LSTM models of three different levels of complexity - low, medium and high. These are obtained by varying number of filters in the CNN layer, number of LSTM layers as well as the number of units in each LSTM layer.

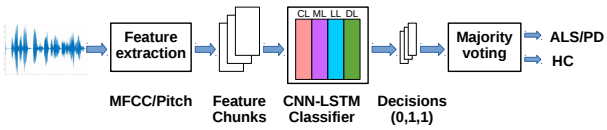


Fig. 1. ALS/PD vs HC classification system; here CL: CNN layer, ML: Maxpooling layer, LL: LSTM layer, DL: Dense layer

2.4. Noise Conditions

Speech recordings obtained outside controlled environment are prone to various background noises, thus requiring the ALS/PD prediction system to be noise-robust. We study the relative suitability of pitch and MFCC features for such noise-robust classification. We consider two different settings in this context, matched and mismatched train-test conditions. In matched case, the noise and SNR of the data used in training and testing the classifier are matched. In case of mismatched condition, classifier trained with clean data is used to test both clean and noisy test samples. Although the matched condition is expected to produce better classification accuracy, we consider the mismatched case to examine the generalization ability of the classifier to unseen environmental conditions. This is particularly useful as it is not practically feasible to train classifiers in numerous possible noise types and SNR conditions.

3. DATASET

Speech data used in this work were collected at National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India. The data collection protocol was reviewed and approved by the Ethics committee of NIMHANS. Each subject signed a consent form prior to the data collection. Table 1 summarizes the gender and age statistics of the subjects of three classes, namely ALS, PD and HC, who participated in the data collection sessions. Diagnosis of ALS (El Escorial criteria) or PD for the subjects considered in this work was made by Neurologists at NIMHANS. The subjects had six different native languages, namely, Bengali, Hindi, Odiya, Tamil, Telugu and Kannada, with approximately equal proportion of subjects belonging to each language. Spontaneous speech recordings were collected from each subject as they talked about *a festival they celebrate* and *a place that they have recently visited* for approximately one minute each in their native language. The subjects were given a few minutes of preparation time before they started. Four different speech tasks namely spontaneous speech, sustained phoneme production, diadochokinetic task and image description task have been considered in literature [10, 11, 15]. However, in this work we consider only spontaneous speech since pitch modulation during this

Table 1. Gender and age details of subjects

Condition	Gender		Age range
	#Male	#Female	(years)
ALS	38	21	36 - 75
PD	45	14	35 - 79
HC	44	16	22 - 53

task happens naturally as the subject speaks unlike other tasks where the pitch pattern is primarily influenced by the target stimulus of the task. The speech data were recorded at 44.1 kHz using Zoom H-6 recorder with XYH-6 stereo X/Y capsule high quality unidirectional microphone [20] from a distance of 2 feet from the subject. The data were then downsampled to 16 kHz for all further processing. A total of 5.62 hours of recordings were obtained.

4. EXPERIMENTAL SETUP

4.1. Pitch Estimation

Two well known pitch estimation algorithms - SWIPE [19] and PEFAC [18], are used to estimate pitch values every 10 ms. We particularly use the SWIPE' variant [19] in this work, though it is referred to as SWIPE in this work. SWIPE is set to estimate pitch by searching in the range of 75-500 Hz with a resolution of 48 steps per octave. The spectrum is computed using Hann window with 50% overlap and sampled every $1/20^{th}$ of equivalent rectangular bandwidth (ERB). Pitch estimates with strength lower than 0.2 for clean speech and 0.15 for noisy speech are treated as undefined and correspond to unvoiced/silence regions. In case of PEFAC, a pitch estimate is considered to correspond to unvoiced/silence region if the probability of that frame being voiced is less than 0.3 for clean speech and 0.35 for noisy speech. In case of both algorithms, the pitch estimates for unvoiced/silence regions are replaced by 0s. The pitch contours in the voiced regions are further smoothed using 5-point moving average.

4.2. MFCC Computation

MFCC features are computed using 20 ms frame length with 10 ms overlap. We use 13-dimensional MFCC along with first and second order differences resulting into a 39-dimensional feature vector. KALDI speech recognition toolkit [21] is used for this purpose.

4.3. Model Configurations and Noise Conditions

We consider CNN-LSTM classifiers having three different levels of complexity - low, medium and high. At each level, the model architectures for both pitch and MFCC are tuned by optimizing the validation accuracy, while at the same time maintaining the #params and FLOPs close in case of both features. Fig. 2 shows the exact configuration adapted in each case. For instance, the medium complexity model for pitch has an initial 1D-CNN layer (CL) with 35 filters (NF) each of size 20 (FS). ReLU is used as the non-linear activation here. This layer thus corresponds to a total of 735 weights and bias parameters. This is followed by a maxpooling layer (ML) that performs the pooling operation over a window size of 4 (PS). The subsequent LSTM layer (LL) has 32 cells, thus accounting for 8704 parameters. The final dense layer (DL) with 2 units and softmax activation adds another 66 parameters. Hence this architecture involves a total of 9505 parameters, all of which are estimated during training. It is to be noted here that LL contributes the most towards the total parameter count. FLOPs count for each model configuration is also mentioned in the figure.

In order to simulate noise conditions, the white Gaussian noise is added to each speech utterance at SNRs of 0, 5, 10 and 20dB.

4.4. Train-Test Configuration

Experimental evaluation is done in a 5-fold cross validation setup. All subjects are evenly divided into 5 groups, each comprising almost equal number of subjects from ALS/PD and HC classes. Each group is balanced w.r.t. age, language and disease severity. In every fold of cross validation, data of 3 groups of subjects are used in training while data of 1 group are used in validation. The remaining group is used for testing. Thus, testing is done in unseen subject condition. Speech features (pitch/MFCC) are chunked into 2 sec segments with 1 sec overlap. The chunked data are then used for training and testing the CNN-LSTM models. The models are trained using binary cross entropy loss function and Adam optimizer with a learning rate of 0.001. Training is done for a maximum of 60 epochs. Early stopping based on validation loss with a patience of 5 is employed for regularization purpose. Batch size is kept at 32. Mean and standard deviation (SD) of classification accuracy over the 5 folds of evaluation are used as the performance metrics. Wilcoxon signed rank test [22] is performed to examine if the classification accuracies obtained using pitch and MFCC are significantly different across 5 folds. For this purpose, test cases from all 5 folds are divided into 15 random groups of equal sizes and classification accuracies in all of these groups are considered. The classification accuracies using pitch and MFCC are reported to be significantly different if the obtained p-value from the Wilcoxon signed rank test is less than 0.05.

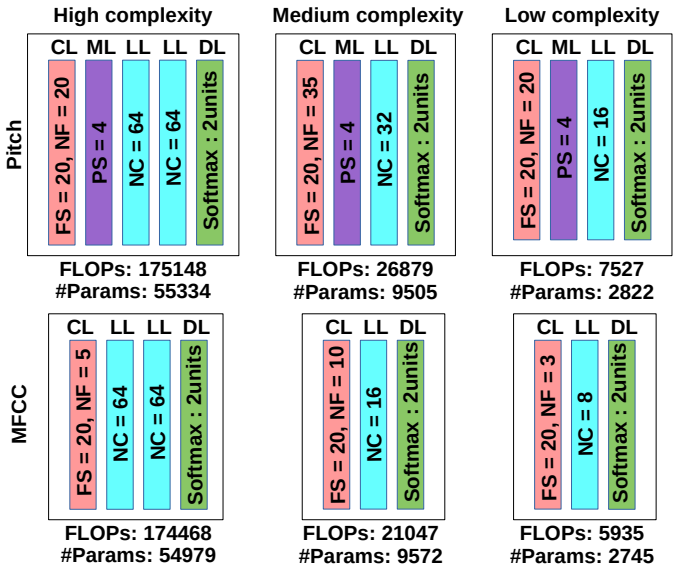
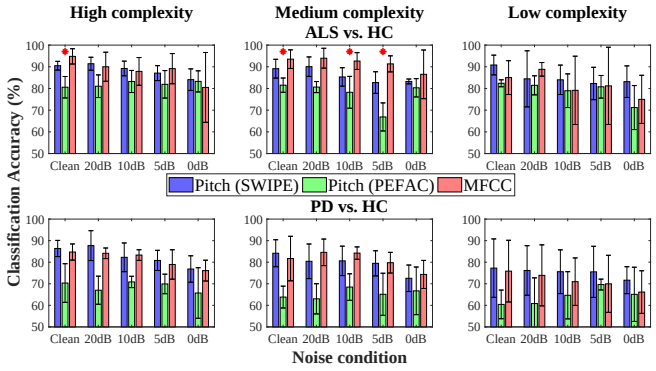


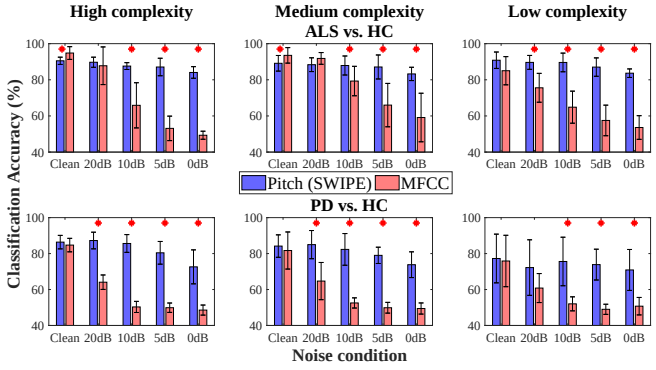
Fig. 2. Configurations of CNN-LSTM models used with pitch and MFCC for different levels of complexity; here CL: CNN layer, ML: Maxpooling layer, LL: LSTM layer, DL: Dense layer, FS: Filter size, NF: Number of filters, PS: Pooling window size, NC: Number of LSTM cells

5. RESULTS & DISCUSSION

Fig. 3 shows the classification accuracies for two classification tasks considered, e.g. ALS vs. HC and PD vs. HC, under different settings of model complexity and noise conditions. Fig. 3(a) illustrates the results for matched train-test condition. It can be observed that the average accuracies obtained using the pitch estimated by SWIPE are higher than those obtained for the pitch estimated by PEFAC under all cases of model complexity and noise conditions in both classification tasks. This suggests SWIPE to be more appropriate for pitch estimation for the classification task at hand. It has been reported in [23] that the voicing decision error made by PEFAC is higher than that of SWIPE under both clean and additive white Gaussian noise conditions. Since speaking pattern is expected to carry important



(a) Matched train-test condition



(b) Mismatched train-test condition

Fig. 3. Classification performance for varying model complexity and noise condition; SD of accuracy is indicated by error bar; here red stars (⋆) indicate the cases where performance of pitch (SWIPE) and MFCC differ at 5% significance level (i.e., $p < 0.05$ in Wilcoxon signed rank test)

information for ALS/PD vs. HC classification, the erroneous voicing decisions might lead to an inferior classification performance using PEFAC than SWIPE. Hence, we consider SWIPE only for the rest of the experiments and discussion. Wilcoxon signed rank test is also performed between accuracies obtained using MFCC and pitch computed by SWIPE. It can be observed from the figure that under clean condition, MFCC outperforms pitch in case of high and medium complexity models for ALS vs. HC classification. In all other cases, both features achieve similar performances. This further indicates that though pitch is a one-dimensional feature, it is as informative as MFCC for these particular classification tasks in most cases, especially when low complexity classifiers are considered. With decrease in SNR level, mean accuracies obtained by both pitch and MFCC drop for both ALS vs. HC and PD vs. HC classifications in most of the cases. Although the difference in the performances obtained by the two features under the influence of noise is not statistically significant here, in case of low complexity models, drop in the performance of pitch is less than that in case of MFCC. Moreover the SD of accuracy is much higher for MFCC, which indicates pitch based classifiers to be more consistent under noise and complexity constraints.

Fig. 3(b) presents the classification accuracies obtained under mismatched train-test configuration. It can be observed that the average performance obtained using pitch in this case is similar to that obtained in case of matched train-test setting. However, performance of MFCC drops w.r.t. those achieved in matched case. It can further

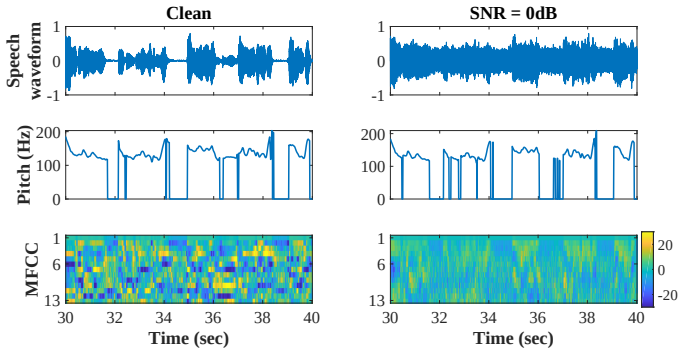


Fig. 4. Illustration of pitch and MFCC obtained from a 10 sec speech segment of an ALS patient under clean and 0dB SNR conditions

be noticed that the average classification accuracy using pitch remains mostly unchanged with decreasing SNR levels, whereas performance using MFCC deteriorates drastically as the level of noise present in the signal increases. This trend is present in the case of all model complexities considered for both ALS vs. HC and PD vs. HC classifications. This indicates the inherent robustness of the pitch estimates to the background noise. This is attributed to the noise robust nature of the pitch extraction algorithm considered (SWIPE). The pitch trajectories obtained under noisy and clean conditions are reasonably similar, leading to similar test performances when the classifier models trained with clean data are used. However, MFCC gets severely affected by additive noise, leading to significantly different feature estimates obtained during clean and noisy speech conditions. Due to the change in the feature estimates under noisy condition, the classifier model trained with clean data fails to make correct classifications in noisy case. Fig. 4 provides an evidence for the reasoning above. Figure illustrates a 10 sec segment of speech data recorded from an ALS patient. The data is examined under clean and 0dB SNR conditions. It can be observed that the pitch trajectories obtained in both cases appear similar. On the contrary, changes in the MFCC values are visually noticeable. These observations suggest that pitch equips the classifiers with better generalization abilities to different SNR conditions in case of both ALS vs. HC and PD vs. HC classifications.

6. CONCLUSION

This work presents a comparative study of the performance of two different speech features - pitch and MFCC for the purpose of classifying HC and patients with ALS and PD under two limiting conditions - low complexity classifiers and presence of background noise. Pitch is observed to provide similar level of distinctive information as MFCC in clean and matched train-test conditions. In case of mismatched train-test, pitch is found to be more noise robust and provides the classifiers with better generalization ability to unseen SNR conditions. Since no language specific bias is introduced in the experimental design, the results are expected to hold for all native languages. In future, we would like to examine the noise robustness of different speech features in various additive noise conditions as well as real noisy recordings. Future work for this study would also involve experimentation using denoising algorithms in both matched and mismatched cases.

7. ACKNOWLEDGEMENT

Authors thank the department of science and technology (DST), Govt of India for their support in this work.

8. REFERENCES

[1] "What is ALS?," https://www.als.org/understanding-als/what-is-als/, [Online; accessed 21-Oct-2020].

[2] "Typical symptoms of Parkinson's disease," https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/, [Online; accessed 21-Oct-2020].

[3] Benjamin Rix Brooks, Robert G Miller, Michael Swash, and Theodore 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.

[4] Matthew C Kiernan, Steve Vucic, Benjamin C Cheah, Martin R Turner, Andrew Eisen, Orla Hardiman, James R Burrell, and Margaret C Zoing, "Amyotrophic Lateral Sclerosis," *The lancet*, vol. 377, no. 9769, pp. 942–955, 2011.

[5] Adriano Chio, Giancarlo Logroscino, Orla Hardiman, Robert Swingler, Douglas Mitchell, Ettore Beghi, and Bryan G Traynor, "Prognostic factors in ALS: a critical review," *Amyotrophic Lateral Sclerosis*, vol. 10, no. 5-6, pp. 310–323, 2009.

[6] Barbara Tomik and Roberto J Guiloff, "Dysarthria in Amyotrophic Lateral Sclerosis: a review," *Amyotrophic Lateral Sclerosis*, vol. 11, no. 1-2, pp. 4–15, 2010.

[7] "Parkinson's disease," https://www.asha.org/, [Online; accessed 21-Oct-2020].

[8] Lavoisier Leite Neto and Ana Carolina Constantini, "Dysarthria and quality of life in patients with Amyotrophic Lateral Sclerosis," *Revista CEFAC*, vol. 19, no. 5, pp. 664–673, 2017.

[9] Serge Pinto, Canan Ozsancak, Elina Tripoliti, Stéphane Thobois, Patricia Limousin-Dowsey, and Pascal Auzou, "Treatments for dysarthria in Parkinson's disease," *The Lancet Neurology*, vol. 3, no. 9, pp. 547–556, 2004.

[10] Suhas BN, Deep Patel, Nithin Rao Koluguri, Yamini Belur, Pradeep Reddy, Nalini Atchayaram, Ravi Yadav, Dipanjan Gope, and Prasanta Kumar Ghosh, "Comparison of speech tasks and recording devices for voice based automatic classification of healthy subjects and patients with Amyotrophic Lateral Sclerosis," in *INTERSPEECH*, 2019, pp. 4564–4568.

[11] Jhansi Mallela, Aravind Illa, Suhas BN, Sathvik Udupa, Yamini Belur, Nalini Atchayaram, Ravi Yadav, Pradeep Reddy, Dipanjan Gope, and Prasanta Kumar 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.

[12] Taha Khan, Lina E Lundgren, David G Anderson, Irena Nowak, Mark Dougherty, Antanas Verikas, Misha Pavel, Holly Jimison, Slawomir Nowaczyk, and Vered Aharonson, "Assessing Parkinson's disease severity using speech analysis in non-native speakers," *Computer Speech & Language*, vol. 61, pp. 101047, 2020.

[13] Suhas BN, Jhansi Mallela, Aravind Illa, Yamini BK, Nalini Atchayaram, Ravi Yadav, Dipanjan Gope, and Prasanta Kumar 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.

[14] Maxim Vashkevich, Elias Azarov, Alexander Petrovsky, and Yuliya Rushkevich, "Features extraction for the automatic detection of ALS disease from acoustic speech signals," in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE, 2018, pp. 321–326.

[15] Jhansi Mallela, Yamini Belur, Nalini Atchayaram, Ravi Yadav, Pradeep Reddy, Dipanjan Gope, and Prasanta Kumar Ghosh, "Raw speech waveform based classification of patients with ALS, Parkinson's Disease and healthy controls using CNN-BLSTM," accepted in *INTERSPEECH*, 2020.

[16] Jordan R Green, Yana Yunusova, Mili S Kuruvilla, Jun Wang, Gary L Pattee, Lori Synhorst, Lorne Zinman, and James D Berry, "Bulbar and speech motor assessment in ALS: challenges and future directions," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, no. 7-8, pp. 494–500, 2013.

[17] Leah K Bowen, Gabrielle L Hands, Sujata Pradhan, and Cara E Stepp, "Effects of Parkinson's disease on fundamental frequency variability in running speech," *Journal of medical speech-language pathology*, vol. 21, no. 3, pp. 235, 2013.

[18] Sira Gonzalez and Mike Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *19th European Signal Processing Conference*. IEEE, 2011, pp. 451–455.

[19] Arturo Camacho and John G Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[20] "Zoom XYH-6 adjustable stereo microphone capsule," https://www.zoom.co.jp/products/product-accessories/xyh-6-xy-stereo-microphone-capsule/, [Online; accessed 05-Feb-2021].

[21] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý, "The Kaldi speech recognition toolkit," in *Workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

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

[23] Lyudmila Sukhostat and Yadigar Imamverdiyev, "A comparative analysis of pitch detection methods under the influence of different noise conditions," *Journal of voice*, vol. 29, no. 4, pp. 410–417, 2015.