Source and Vocal Tract Cues for Speech-based Classification of Patients with Parkinson's Disease and Healthy Subjects

Tanuka Bhattacharjee[1], *Jhansi Mallela*[1], *Yamini Belur*[2], *Nalini Atchayaram*[2], *Ravi Yadav*[2], *Pradeep Reddy*[2], *Dipanjan Gope*[1], *Prasanta Kumar Ghosh*[1]

[1]Indian Institute of Science, India
[2]National Institute of Mental Health and Neurosciences, India

tanukab@iisc.ac.in, jhansimallela5@gmail.com, prasantg@iisc.ac.in

Abstract

Parkinson's disease (PD) affects both source and vocal tract components of speech. Various speech cues explored in literature for automatic classification of individuals with PD and healthy controls (HC) implicitly carry information about both these components. This work explicitly analyzes the contribution of source and vocal tract attributes toward automatic PD vs. HC classification, which has not been done earlier to the best of our knowledge. Here fundamental frequency (f_o) is used to capture source information. For quantifying vocal tract information, speech waveforms are converted to unvoiced forms and mel-frequency cepstral coefficients (MFCC), denoted by voicing-removed MFCC, are obtained from them. Experimental results suggest that (1) the relative merit of source and vocal tract cues in classifying PD vs. HC largely depends on the speech task being considered, (2) both cues complement each other across all tasks, (3) while MFCC encodes both source and vocal tract features, source information captured by f_o is different and further complements MFCC when the classifiers are trained and tested under clean or matched noise conditions, thereby enabling the feature-level fusion of f_o and MFCC to achieve the best classification accuracy, (4) under unseen noise conditions, f_o alone proves to be a highly noise-robust feature.

Index Terms: Parkinson's disease, source, vocal tract, fundamental frequency, mel-frequency cepstral coefficients

1. Introduction

According to the source-filter model [1], speech is produced by passing a source signal through a filter. The glottal and/or supra-glottal excitation acts as the source and the vocal tract constitutes the filter. Major components of human vocal tract include oral cavity, nasal cavity, pharynx, and larynx. For voiced sounds, the source is predominantly quasi-periodic with minimal aperiodic components which are responsible for the naturalness of speech [2]. In case of unvoiced sounds, the source is essentially a colored noise. This source signal when passed through the vocal tract undergoes changes in its spectral characteristics and results in the speech waveform.

Parkinson's disease (PD) affects speech production in about 90% of the patients even in the early stages [3]. Deficit of the neurotransmitter dopamine in basal ganglia caused by PD hampers coordinated and smooth muscular control. The muscles responsible for speech production are reported to be highly affected by this disease typically leading to hypokinetic dysarthria [4]. Impairments are observed in both source and vocal tract attributes of speech. Signs like monopitch, monoloudness, low voice intensity, and reduced fundamental frequency range [3, 5], as observed in PD, characterize the impact on source component. On the other hand, imprecise articulation, voice nasality and increased acoustic noise [5] primarily account for the dysfunctions associated with the vocal tract. This work attempts to compare the source and vocal tract characteristics in patients affected by PD and analyze how they contribute individually and in combination toward automatic classification of individuals with PD and healthy controls (HC).

Literature Review: Several speech cues have been analyzed in the literature as potential bio-markers for speech-based automatic PD vs. HC (PD/HC) classification. To our knowledge, all the already explored feature sets implicitly encode information about both the source and the vocal tract components of speech. No known work has yet been done to explicitly study these two components in the context of PD/HC classification. Mel-frequency cepstral coefficients (MFCC) and log mel spectrograms, indicative of the spectral characteristics of speech, have been explored in [6] and [7], respectively, for PD/HC classification. Along with MFCC, Khan et al. [8] has employed cepstral separation difference to quantify phonation characteristics and spectral dynamics together with fundamental frequency (f_o) variation to capture respiration and prosodic properties for PD severity assessment. In [9], pitch and MFCC have been compared under the influence of background noise and the constraint of low complexity classifier. Various dysphonia measures including f_o, jitter and shimmer measures, noise-to-harmonics ratio, pitch period entropy, have also been explored for identifying PD and monitoring the disease progression [10, 11]. In [12], the authors have employed a 1D-convolutional neural network (CNN) for learning representations from raw speech waveform itself, whereas, a stacked autoencoder has been used in [13] to extract features from spectrogram and scalogram of speech signals for PD prediction.

Main Contribution: This work aims to explore the utility of source and vocal tract components of speech for PD/HC classification by answering the following four key questions - (1) Which one between source and vocal tract provides more discriminative cues for the classification? (2) Are the cues provided by the two components complementary in nature? (3) Does MFCC carry source specific cues required for the classification or does fusing source features explicitly provide a performance gain? (4) How does the presence of noise impact the classification performance obtained using these feature sets, individually as well as together?

We consider speech recordings for three tasks, e.g. image description (IMAG), diadochokinetic rate (DIDK), and spontaneous speech (SPON). The f_o features are used for quantifying source information whereas vocal tract information is captured exclusively using voicing-removed MFCC (vrMFCC) (described in Section 2). Here we employ CNN-LSTM as a

classifier [6], where LSTM stands for long short term memory. Five-fold cross-validation experiments on 59 PD and 60 HC subjects in both clean and four additive noisy conditions indicate that though the relative merit of source and vocal tract cues varies in different speech tasks, the two components complement each other consistently. Moreover, the source information encoded in f_o further complements that captured by MFCC leading to the highest classification accuracy using both together as features in clean condition. This is found to be true even in noisy cases when the classifier is trained and tested under matched noise and signal-to-noise ratio (SNR) conditions. Furthermore, source cues in the form of f_o are found to be highly robust against unseen noise.

2. PD vs. HC Classification

Figure 1 summarizes various speech features, characterizing source and vocal tract information, which are used in the PD/HC classification analysis pursued in this work. These features are extracted from the raw waveform. MFCC with delta and delta-delta features are computed, which are known to capture cues from both source and vocal tract components of speech simultaneously [14]. To encode the source information exclusively, f_o contours along with their 1^{st} and 2^{nd} order derivatives are utilized. The process of extracting vocal tract characteristics without any influence of the source component involves two steps - voicing removal and MFCC computation. During voicing removal, the speech waveform is modified to eliminate any voicing information present in the signal. To do so, the input speech is decomposed into f_o, spectral envelope and aperiodicity using WORLD analyzer [2]. The obtained f_o estimates are replaced by 0s and the aperiodicity values for all frequency bands are made 1s in order to remove the inherent voicing information, thereby making the source purely noisy throughout the utterance. The speech waveform is then resynthesized by WORLD synthesizer using the modified f_o and aperiodicity values along with the unchanged spectral envelope. MFCC computed from this voicing-removed speech, denoted as *voicing-removed MFCC* (vrMFCC), is thus expected to primarily capture vocal tract cues. Apart from the three sets of features, MFCC, f_o, and vrMFCC, another two sets, namely f_o + vrMFCC and f_o + MFCC, are generated by fusing f_o with vrMFCC and MFCC, respectively, at the feature-level. Each of these five feature sets is then processed by the classifier block shown in Figure 1 to obtain the PD/HC decisions. Here, a CNN-LSTM network, as proposed in [6], is employed as the classifier. However, unlike [6], the general architecture of the network considered here does not involve any max-pooling operation.

3. Dataset

Speech data collection was performed at National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India. 59 PD (45 male and 14 female) and 60 HC (44 male and 16 female) subjects having age in the range 35 - 79 years and 22 - 53 years, respectively, participated in the data collection sessions. The subjects belonged to six different native languages, namely, Bengali, Hindi, Odiya, Tamil, Telugu, and Kannada, with approximately equal proportion of subjects coming from each language. Dysarthria severity of the PD subjects were assessed by five speech-language pathologists (SLP) from the Speech Pathology and Audiology Department, NIMHANS as per the UPDRS-III scale (range: 0 - 4) [15]. The mode of the five severity scores was considered as the final score. The PD subjects had the final scores in the range of 0 - 2. Data for three different speech tasks - IMAG (∼12.83 hours), DIDK (∼4.65 hours), and SPON (∼5.62 hours) are considered. All speech data were recorded at 44.1 kHz and downsampled to 16 kHz. Further details about the speech tasks and data collection protocol can be found in [12]. Besides the speech recordings, high-frequency (HF) channel, pink, and babble noise from NOISEX-92 dataset [16], along with the additive white gaussian noise (AWGN) are considered for simulating noisy speech conditions.

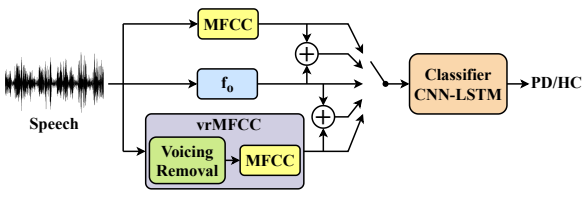

Figure 1: *PD/HC classification scheme*

4. Experimental Setup

4.1. Feature Extraction

The process of estimating the three types of features used in this work is elaborated below. Utterance-level Z-score normalization is applied to each feature dimension independently such that each dimension has zero mean and unit standard deviation (SD) over all frames of a particular utterance.

Fundamental Frequency (f_o): SWIPE (SWIPE' variant) [17] algorithm is used to extract f_o values from speech waveforms every 10 ms. A detailed description of the f_o contour estimation process along with the adapted parameter settings can be found in [9]. The f_o estimates for unvoiced/silence regions are replaced by 0s. Local estimates of the 1^{st} and 2^{nd} order derivatives of the obtained f_o contours are computed over 9 frames using the Librosa package [18], thereby giving rise to 3-dimensional feature vectors for f_o.

MFCC: KALDI speech recognition toolkit [19] is used for computing MFCC with 20 ms frame length and 10 ms overlap. 13-dimensional MFCC along with 1^{st} and 2^{nd} order differences totalling a 39-dimensional feature vector is considered.

Voicing-Removed MFCC (vrMFCC): To compute vrMFCC, voicing removal is first performed on the speech samples using WORLD [2], as stated in Section 2. During this process, f_o estimates of a speech utterance are obtained with a frame period of 5 ms using Harvest algorithm [20]. The floor and ceiling frequencies for the f_o estimation range are set to 70 Hz and 800 Hz, respectively. The spectral envelope and aperiodicity are estimated using CheapTrick [21] and D4C [22] algorithms, respectively. MFCC features of the voicing-removed speech are then extracted following the MFCC computation process described above.

4.2. Classifier Configuration

The CNN-LSTM classifier adapted in this work has an initial 1D-CNN layer with NF filters each of size FS and stride 1. *ReLU* is used as the non-linear activation for this layer. The CNN layer is followed by two LSTM layers each having NC cells with *tanh* activation. The final layer of the model is a dense layer with 2 units and softmax activation. The values of NF, FS, and NC are tuned independently for each feature set by optimizing the validation accuracy while maintaining similar classifier complexity in all cases. We take into account both memory and runtime complexity of the classifiers. The number

Table 1: *Tuned values of CNN-LSTM architecture parameters*

Feature Set	NF	FS	NC	#params	FLOPs
$\mathbf{f_o}$	18	20	64	55500	175.48k
MFCC / vrMFCC	5	20	64	54979	174.46k
$\mathbf{f_o}$ + MFCC / $\mathbf{f_o}$ + vrMFCC	5	20	64	55279	175.06k

of parameters (#params) of the model is used for quantifying memory complexity, whereas runtime complexity is measured by the number of floating point operations (FLOPs) needed by the network. Table 1 summarizes the tuned parameter values for all feature sets. It is to be noted that similar model complexity is preserved in case of all feature sets in order to restrict any particular feature set from receiving added advantages owing to a more complex model in classification.

All classifiers are trained using binary cross entropy loss and Adam optimizer. The learning rate is set to 0.001 and the batch size is kept at 32. The models are trained for a maximum of 60 epochs while early stopping with a patience of 5 based on validation loss is employed to avoid overfitting.

4.3. Noise Conditions

Four different noise conditions, namely, AWGN, HF channel, pink, and babble, are explored in this work. Each noise is added to every speech utterance at 0, 5, 10, and 20 dB SNRs. For HF channel, pink and babble, random segments of required length are extracted from the noise recordings and are added to the speech samples. AWGN, on the other hand, is added to the speech utterances using the *awgn* command of MATLAB (R2017a). We consider two different train-test settings here - matched and mismatched. In matched setup, the data used for training and testing the classifiers cater to the same noise and SNR conditions. On the contrary, in the mismatched case, classifiers are trained using clean speech samples only and are used to evaluate both clean and noisy test utterances. The mismatched case is considered to assess the robustness of the features against unseen noise conditions, which is essential for deployment in practice.

4.4. Evaluation Protocol

All experimental evaluations are performed in the same 5-fold cross-validation setup as used in [9] with each fold containing almost equal number of subjects from PD and HC classes. The mean and SD of the classification accuracies obtained in the 5 folds are reported as the performance metrics. We perform Wilcoxon signed rank test [23] at 10% significance level to determine if the classification accuracies obtained for different feature sets and noise conditions are significantly different across 5 folds. Toward this end, test samples from each fold are divided into 3 random groups of equal sizes. The 15 classification accuracies thus obtained are considered for the test.

5. Results and Discussion

This section attempts to answer the four key questions put forward in Section 1 using supporting results and observations.

5.1. Source ($\mathbf{f_o}$) or Vocal Tract (vrMFCC)?

The classification accuracies obtained using source (f_o) and vocal tract (vrMFCC) features independently in all three speech tasks, e.g. IMAG, DIDK, and SPON, are presented in the first and third rows of Table 2. Wilcoxon signed rank test suggests that source cues significantly outperform vocal tract cues in the SPON task. The spontaneous f_o variations and intonations inherent in SPON task appear to capture discriminative cues which predominate the markers obtained from vocal tract information. For IMAG task on the other hand, vocal tract is observed to outperform source cues by a significant margin. The speech recordings for IMAG task being very short in duration ($\sim$3-5 sec) may not be able to capture f_o variations indicative of PD. Hence the contribution of vocal tract cues become predominant in this case. Lastly, for DIDK task, both source and vocal tract features achieve similar level of classification performance. This might be due to the fact that DIDK, being an artificially crafted speech task, attempts to put similar weightage to both source and vocal tract information. Hence it can be argued that source and vocal tract cues manifest themselves differently in different speech tasks leading to varied relative contribution toward PD/HC classification over tasks.

Table 2: *Mean classification accuracies in % (SD in bracket) obtained using different feature sets; here # indicates that MFCC outperforms vrMFCC at 10% significance level; * and $\triangle$ indicate that f_o + MFCC/vrMFCC outperforms f_o and MFCC/vrMFCC, respectively, at 10% significance level*

Feature Set	Speech Task			
	IMAG	DIDK	SPON	Overall
$\mathbf{f_o}$	74.19 (4.67)	75.33 (2.86)	88.12 (4.44)	79.21
MFCC	85.30 (4.92)	81.23 $^{\#}$ (2.40)	88.04 $^{\#}$ (2.84)	84.86
vrMFCC	83.17 (3.56)	76.45 (4.31)	83.26 (3.41)	80.96
$\mathbf{f_o}$ + vrMFCC	84.74 * (3.69)	83.42 $^{*\triangle}$ (1.29)	90.36 $^{\triangle}$ (4.03)	86.17
$\mathbf{f_o}$ + MFCC	88.65 * (4.21)	83.28 * (4.09)	91.91 $^{*\triangle}$ (1.31)	87.95

5.2. Are They Complementary?

As stated earlier, MFCC incorporates both source and vocal tract information. It can be seen from Table 2 that PD/HC classification accuracies drop on average w.r.t. MFCC when f_o or source information is removed from MFCC to obtain vrMFCC. Significant drops are observed in case of DIDK and SPON tasks (indicated by #). Furthermore, adding the source information back on top of vrMFCC by fusing f_o and vrMFCC (f_o + vrMFCC) improves the average performance over vrMFCC. Significant improvements are observed again in case of DIDK and SPON (indicated by $\triangle$). Since the vocal tract provides predominant cues in the case of IMAG, removal or addition of source information impacts less in this case. Wilcoxon signed rank test also shows that for all speech tasks no significant difference exists between the classification performances obtained using f_o + vrMFCC and MFCC itself. These observations suggest that if vocal tract features are considered as the baseline information, then source indeed provides complementary cues for PD/HC classification. Table 2 also illustrates that the average classification accuracies obtained using f_o + vrMFCC are higher than those obtained using f_o only, with significant improvements in case of IMAG and DIDK tasks (indicated by *) suggesting the complementarity of source and vocal tract features. The lower boost in case of SPON is expected as source itself captures predominant discriminative cues in this case making the contribu-

tion of vocal tract less significant.

5.3. Does Fusion of f_o and MFCC help?

Table 2 reports that the average classification accuracy obtained using f_o + MFCC over all tasks is 8.74% and 3.09% higher than those obtained by f_o and MFCC respectively. The performance enhancement over f_o is significant for all speech tasks (indicated by *), whereas that over MFCC is significant particularly in the case of SPON (indicated by △). This finding indicates that the source information encoded in MFCC is different from that encoded in f_o and both these types of source cues complement each other. For example, shimmer and jitter information could be captured by MFCC, but not by the f_o contour itself. Hence the classification accuracy benefits from this feature fusion over the individual features. It should be noted here that the average classification accuracy obtained by f_o + MFCC over all tasks is higher than that obtained by f_o + vrMFCC as well. Hence we consider only f_o + MFCC for the rest of the analysis.

5.4. What is the Effect of Noise?

Figure 2 shows the classification accuracies obtained using f_o, MFCC, and f_o + MFCC (denoted by fusion in the figure) in all three speech tasks averaged over four different noise conditions - AWGN, HF channel, pink, and babble. It can be observed that under matched train-test condition for IMAG and SPON, average accuracies using all feature sets drop significantly w.r.t. the corresponding clean condition at same level of SNR (10 dB or lower). However, in the case of DIDK, the performance of MFCC drops at higher SNR (20 dB) than the other two feature sets. In this matched condition, fusion significantly outperforms both the individual features at all tasks and SNR conditions except for the 0 dB cases in DIDK and SPON. This suggests that source and vocal tract features are complementary even in the matched train-test noisy conditions. Figure 3 provides a more detailed picture of the performances of all the three feature sets under the influence of each noise separately. It can be observed that fusion significantly outperforms f_o in all cases of IMAG task. For DIDK, the improvement is significant in all but two cases, namely, 10 dB SNR cases for pink and babble noises. Fusion, on average, outperforms f_o for SPON task as well, with significant improvement observed in 50% of the noisy cases. On the other hand, the average classification accuracy obtained

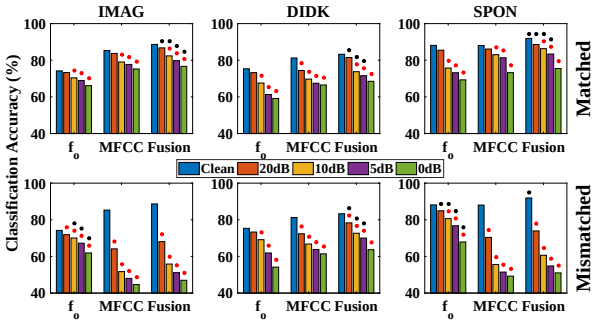

Figure 2: *Mean classification performance over four noise conditions - AWGN, HF channel, pink, and babble; here red dot indicates drop in accuracy w.r.t. clean case at 10% significance level and black dot marks the feature set which outperforms the other two at 10% significance level for a particular SNR condition*

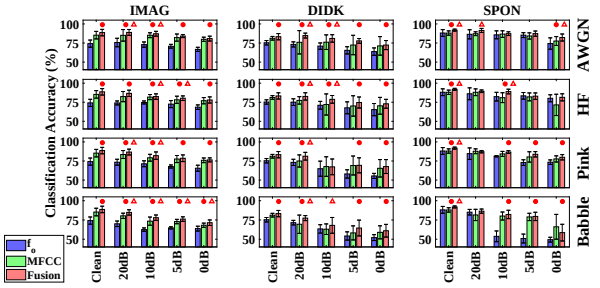

Figure 3: *Classification performance under different noise conditions for matched train-test setting; SD of accuracy is indicated by error bar; here dots and triangles indicate that fusion outperforms f_o and MFCC, respectively, at 10% significance level*

using fusion is higher than that obtained by MFCC in majority of the cases, with significant improvement observed in 41.67% of noisy cases. However, a noticeable decline in the average performance of fusion w.r.t. MFCC is also observed for SPON task under 0 dB babble noise condition. These findings suggest that, under the influence of different types of noises, combining f_o and MFCC helps more in the cases of IMAG and DIDK as compared to SPON.

Contrary to the matched case, in mismatched condition (Figure 2), fusion outperforms the individual features in DIDK only (except 0 dB case). For IMAG and SPON, f_o is found to achieve the best performance at all SNRs except for the 20 dB case of IMAG where all three feature sets achieve similar performances. The inferior performance of fusion in these two tasks relates to the event of MFCC performance getting more severely affected by noise during these two tasks in particular. Moreover, a significant drop in accuracy w.r.t. the clean condition is observed at lower SNR for f_o than MFCC and fusion in two out of the three tasks. These observations indicate f_o to be highly robust against unseen noise and SNR conditions.

6. Conclusions

This work analyzes the speech-based PD/HC classification task from the perspective of source and vocal tract cues. Depending on the speech task, different components are found to capture predominant discriminative features, though the two components complement each other throughout all tasks. Among all the feature sets considered, the fusion of f_o and MFCC is found to attain the highest classification accuracy under both clean and matched train-test conditions. However, robustness against unseen noise is predominantly observed in the case of source features encoded in f_o. Though this paper talks about the behaviour of source component in general, the experiments are restricted to the f_o attribute only. Similar analysis needs to be carried out using other source cues like glottal flow. Experiments involving PD subjects with dysarthria severity spanning the entire UPDRS-III range (0 - 4) are also essential. Another interesting future direction for this work would be to assess the dysarthria severity using source and vocal tract cues.

7. Acknowledgements

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

8. References

[1] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970, no. 2.

[2] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[3] 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.

[4] P. Gómez, J. Mekyska, A. Gómez, D. Palacios, V. Rodellar, and A. Álvarez, "Characterization of Parkinson's disease dysarthria in terms of speech articulation kinematics," *Biomedical Signal Processing and Control*, vol. 52, pp. 312–320, 2019.

[5] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova, "Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation," *Journal of neural transmission*, vol. 124, no. 3, pp. 303–334, 2017.

[6] 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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6784–6788.

[7] S. BN, J. Mallela, A. Illa, Y. BK, 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.

[8] 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.

[9] T. Bhattacharjee, J. Mallela, Y. Belur, N. Atchayaram, 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," Accepted in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021. [Online]. Available: https://tinyurl.com/22r6m8c5

[10] V. Despotovic, T. Skovranek, and C. Schommer, "Speech based estimation of Parkinson's disease using Gaussian processes and automatic relevance determination," *Neurocomputing*, vol. 401, pp. 173–181, 2020.

[11] S. Sharanyaa, P. N. Renjith, and K. Ramesh, "Classification of Parkinson's disease using speech attributes with parametric and nonparametric machine learning techniques," in *3rd International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 2020, pp. 437–442.

[12] J. Mallela, N. A. Yamini Belur, 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. 21st Annual Conference of the International Speech Communication Association, Shanghai, China*, 2020, pp. 4586–4590.

[13] B. Karan, S. S. Sahu, and K. Mahto, "Stacked auto-encoder based time-frequency features of speech signal for Parkinson's disease prediction," in *International Conference on Artificial Intelligence and Signal Processing (AISP)*. IEEE, 2020, pp. 1–4.

[14] P. N. Le, J. Epps, E. H. Choi, and E. Ambikairajah, "A study of voice source and vocal tract filter based features in cognitive load classification," in *20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 4516–4519.

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

[16] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[17] A. Camacho and J. 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.

[18] B. McFee, V. Lostanlen, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, J. Mason, D. Ellis, E. Battenberg, S. Seyfarth, R. Yamamoto, K. Choi, viktorandreevichmorozov, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Hereñú, F.-R. Stöter, P. Friesch, A. Weiss, M. Vollrath, and T. Kim, "Librosa: 0.8.0," Jul. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3955228

[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.

[20] M. Morise *et al.*, "Harvest: A high-performance fundamental frequency estimator from speech signals." in *INTERSPEECH*, 2017, pp. 2321–2325.

[21] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1–7, 2015.

[22] ——, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

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