ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095664

EXPLORING THE ROLE OF FRICATIVES IN CLASSIFYING HEALTHY SUBJECTS AND PATIENTS WITH AMYOTROPHIC LATERAL SCLEROSIS AND PARKINSON'S DISEASE

Tanuka Bhattacharjee[1], *Yamini Belur*[2], *Atchayaram Nalini*[2], *Ravi Yadav*[2], *Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering Department, Indian Institute of Science, Bengaluru, India
[2]National Institute of Mental Health and Neurosciences, Bengaluru, India

ABSTRACT

Dysarthria due to Amyotrophic Lateral Sclerosis (ALS) and Parkinson's Disease (PD) impairs sustained phoneme productions. Vowels and fricatives get affected differently owing to the differences in their production mechanisms. This paper examines three sustained voiceless fricatives - /s/, /sh/ and /f/, as compared to three sustained vowels - /a/, /i/ and /o/, for classifying patients with ALS/PD and Healthy Controls (HC). Fricatives are found to achieve higher classification accuracies than /a/ and /o/, though /i/ outperforms all. Patients seem to find it difficult to form constrictions while producing fricatives, or to proximally position the tongue and palate while uttering /i/, due to dysarthria. Unwanted voicing added to voiceless fricatives by the patient population further contributes towards the discrimination. Both source (related to vocal cord) and filter (related to vocal tract) cues of fricatives, on average, outperform those of vowels. Lastly, decision-level fusion of /i/-/s/-/sh/, with a pooled classifier for these three phonemes, achieves the highest mean ALS vs. HC classification accuracy of 83.35%, although in PD vs. HC case, fusion of multiple /i/ utterances performs the best with an accuracy of 80.03%.

Index Terms— Dysarthria, Fricatives, Vowels, Constriction, Voicing.

1. INTRODUCTION

Dysarthria, prevalent in Amyotrophic Lateral Sclerosis (ALS) and Parkinson's Disease (PD), affects various aspects of speech function, particularly, phonation, articulation and respiration [1, 2]. All of these three sub-systems of speech can be thoroughly assessed by Sustained Phoneme Production (SPP) tasks [3]. SPP is commonplace in clinical speech assessment routines due to its simplicity and ease of use. Thus, SPP can be a potential task for speech-based automatic diagnosis of ALS and PD. This paper analyzes the relative utility of different fricatives as compared to vowels in SPP task based classification of ALS/PD patients and Healthy Controls (HCs).

The physiological mechanisms of uttering vowels and fricatives differ. Hence, the impact of dysarthria on their productions may also vary significantly. Fricatives are produced by forcing the air to flow turbulently through a narrow constriction in the vocal tract resulting into frication. A vowel sound, on the other hand, is produced when the airflow from the lungs passes through the vibrating vocal folds (voicing) followed by a relatively open vocal tract which acts as a resonance chamber. Though the tongue can be placed close to the roof of the mouth, as in the case of close vowels, no constriction is formed in the vocal tract and air can flow without generating any audible frication. Fricatives can be voiced or voiceless; however, we limit our analysis to voiceless fricatives only. Restricted movements of articulators like lips, jaw, tongue, and velum, as observed in ALS [4] and PD [3], lead to altered configurations of vocal tract during phoneme production. Among others, the place and manner of constriction formation might be significantly altered in case of fricatives. This can further result into varied nature of air turbulence at the constriction. Also, dysarthric subjects often add voicing to voiceless fricatives [5] which can provide additional discrimination between ALS/PD and HC classes. Thus, it would be worthwhile to investigate if these factors provide sustained fricatives (SFs) any edge over sustained vowels (SVs) for ALS/PD vs. HC classification.

Multiple works present in the literature have analyzed SVs, particularly /a/, /e/, /i/, /o/ and /u/, for ALS/PD vs. HC classification. Quan et al. [6] have utilized dynamic articulation transition features and bidirectional Long Short Term Memory (LSTM) network for this purpose. Different time-frequency features including short-time Fourier transform [7], Mel-Frequency Cepstral Coefficients (MFCC) [8], tunable Q-factor wavelet coefficients [9] and Hilbert spectrum based cepstral features [10] have also been explored together with Support Vector Machine (SVM) [7, 9, 10] and Random Forest (RF) [9] classifiers. Moreover, features related to the phonatory subsystem, like pitch, jitter, shimmer, harmonic-to-noise ratio [8, 11, 12], have been considered as well in this context.

Although a few studies have investigated the spectral properties of fricatives /s/ and /sh/ produced by ALS and PD patients at word initial positions [13, 14, 15], limited analysis on SFs exist for automatic ALS/PD vs. HC classification. The only works present in the literature have considered three SFs (/s/, /sh/, /f/) and five SVs (/a/, /i/, /o/, /u/, /æ/) together to train or test the classifiers [3, 16, 17, 18]. In [3] and [18], MFCC has been considered as the feature, whereas log mel spectrograms have been extracted in [17]. SVM [18], Dense Neural Network [18], 2D-Convolutional Neural Network (CNN) [17] and CNN-LSTM [3] have been explored as the classifiers. Mallela et al. [16] have fed raw speech waveforms directly to a CNN-bidirectional LSTM framework.

To the best of our knowledge, this paper, for the first time, extensively analyzes the discriminative power of solely the SFs and assesses their relative utility w.r.t. the SVs for ALS/PD vs. HC classification. Particularly, we aim to answer three key questions - (1) How do different SFs contribute towards the classification tasks at hand and how do their performances compare with various SVs? (2) How do the discriminative power inherent in the source (associated with vocal cord) and filter (associated with vocal tract) [19] cues of SFs compare with those of SVs? (3) Do multiple utterances of the same or different phoneme(s), considering both SFs and SVs, contain complementary information such that their decision-level fusion would provide a performance gain over the individual phoneme utterances? Thus, our contribution does not lie in proposing novel classifiers or new speech tasks; it is in identifying these key questions and designing experiments to answer the same.

We consider three voiceless SFs - /s/, /sh/, /f/ and three SVs - /a/, /i/, /o/. The experimental observations are as follows. (1)

Table 1. Number and duration of utterances of different phonemes obtained from ALS, PD and HC groups; each cell entry is in the form of x / y (z), where, x is the number of utterances, y is mean duration (in sec) of the utterances and z is SD of the durations (in sec) of the utterances

Condition	/a/	/i/	/o/	/s/	/sh/	/f/
ALS	88 / 4.81 (2.51)	89 / 4.21 (2.66)	88 / 4.07 (2.42)	86 / 2.58 (1.74)	88 / 2.21 (1.50)	87 / 1.88 (1.29)
PD	86 / 5.65 (2.62)	85 / 5.47 (2.40)	85 / 5.24 (2.33)	90 / 3.48 (1.92)	92 / 2.85 (1.88)	90 / 2.11 (1.42)
HC	88 / 6.13 (1.75)	85 / 6.06 (2.17)	82 / 6.03 (1.82)	84 / 4.87 (1.73)	84 / 3.97 (1.41)	84 / 3.23 (1.22)

With MFCC as feature and CNN-LSTM [3] as classifier, fricatives achieve at least 4.04% and 8.73% higher (absolute) mean accuracies than /a/ and /o/ for ALS vs. HC and PD vs. HC classifications, respectively. However, /i/ outperforms the fricatives by (absolute) at least 0.95% and 6.19%, respectively, in the two classification tasks. Dysarthria seems to affect constriction formations during utterances of fricatives, as well as, the proximal placement of tongue and palate during the production of /i/, thus embedding discriminative cues in these phonemes. Altered nature of air turbulence at the constriction sites for dysarthric SFs, along with the frequently observed unwanted voicing of voiceless fricatives by dysarthric subjects, further contributes towards the differentiating abilities of SFs. (2) Fricatives commonly attain higher mean accuracies than vowels while using MFCC associated with either source or filter estimates of the utterances. (3) Decision-level fusion of /i/, /s/ and /sh/ is found to achieve the highest ALS vs. HC classification accuracy (83.35%), indicating the complementary nature of the cues present in these phonemes. However, the best PD vs. HC classification accuracy (80.03%) is obtained by decision-level fusion of multiple /i/ utterances.

2. DATASET

Data collection was carried out at National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India. 35 subjects (25M + 10F) from each of ALS, PD and HC categories, totalling 105 subjects (75M + 30F), contributed their speech samples. The subjects of the three groups had ages in the ranges of 36-70, 45-73 and 35-62 years, respectively. Dysarthria severity of the ALS and PD subjects were rated by three Speech-Language Pathologists from NIMHANS following the 5-point speech component of ALSFRS-R scale [0 (Loss of useful speech) to 4 (Normal speech)] [20] and that of UPDRS-III scale [0 (Normal speech) to 4 (Unintelligible speech)] [21], respectively. The mode of the three ratings was considered as the final severity score. Approximately equal number of ALS subjects were selected from each of the five dysarthria severity levels. In case of PD, participants had severity scores in the range of 0-2 with approximately equal proportion coming from each level.

Sustained utterances of three voiceless fricatives - /s/, /sh/, /f/, and three vowels - /a/, /i/, /o/ were recorded. The subjects were asked to take a deep breath and prolong a phoneme at a comfortable pitch and loudness level. The process was repeated 1-3 times for each phoneme depending on a subject's level of comfort. The number of utterances of each phoneme obtained from ALS, PD and HC subjects, along with the mean and Standard Deviation (SD) of the duration of the utterances, is mentioned in Table 1. All speech data were recorded at 44.1 kHz sampling frequency and downsampled to 16 kHz. More details about the data collection protocol and the recording setup can be found in [16].

3. EXPERIMENTAL SETUP

This section describes the experiments conducted to answer the key questions listed in Section 1. The associated features and classifiers, along with the evaluation scheme are also summarized.

3.1. Experimental Design

1. SFs vs. SVs: We perform ALS/PD vs. HC classification using MFCC of different SFs and SVs in order to examine their relative utilities for these classification tasks.

2. Source - filter analysis: To analyze the discriminative power of source and filter cues of SFs, as compared to those of SVs, MFCC associated with estimates of these components of the utterances are used to classify ALS/PD vs. HC subjects.

3. Effect of fusion: To exploit the complementary information that might be present in different sustained utterances of the same or different phoneme(s) performed by a subject, we conduct intra- and inter-phoneme decision-level fusion over utterances of /i/, /s/ and /sh/. The choice of the three phonemes is due to their superior performances observed in the first experiment (refer Section 4). In the intra-phoneme fusion case, we make use of the predictions obtained in the first experiment and perform majority voting over the predictions of three repetitions of a phoneme recorded by a subject. On the other hand, inter-phoneme fusion is performed through majority voting over predictions of one utterance each of /i/, /s/ and /sh/ recorded by a subject. We consider two different classifier training schemes in case of inter-phoneme fusion - (1) three distinct classifier models are trained corresponding to the three phonemes (same as the first experiment), and (2) a single pooled classifier is trained using utterances of all the three phonemes taken together. Both approaches utilize MFCC of the original utterances as the features. Given sufficient representation ability, a single pooled model might be able to learn the entire space constituted by the three phonemes being considered.

All the experiments elaborated above use the CNN-LSTM network proposed in [3] as the classifier.

3.2. Feature Extraction

A 2-step feature extraction process is adopted in this work.

1. Source-Filter Estimation: N^{th} order Linear Prediction (LP) is performed on every 20 ms frame of a sustained utterance with 10 ms overlap using the autocorrelation method. The sequence of LP Coefficient (LPC) vectors thus obtained characterize the *time-varying filter* component of the speech, whereas, the LP residual, combined over frames using the overlap-add method, represents the *source*. 2^{nd} order pre-emphasis ($\alpha = 0.99$) is applied before performing LP for spectral equalization purpose following [22]. The computations are done in MATLAB R2021a. We experiment with $N = \{8, 16, 32, 64, 128\}$.

2. MFCC Computation: 12D MFCC (excluding energy coefficient) with delta and double-delta measures constituting a 36D feature vector are computed for every 20 ms frame with 10 ms overlap. KALDI speech recognition toolkit [23] is used to compute these for the original utterances (referred to as O-MFCC) and their source estimates (referred to as S-MFCC). MFCC features for the filter estimates, referred to as F-MFCC, are computed in MATLAB R2021a. For this purpose, LPC vector of each 20 ms frame of the original utterance is first mapped to complex frequency response of the filter, which is then converted to MFCC.

3.3. Classifier

A CNN-LSTM based classifier architecture, as proposed in [3], is employed for all experiments. The first layer of the network is a 1D-CNN having 30 filters each with kernel size 20, stride 1 and *ReLU* activation. It takes MFCC features chunked into 0.5 sec segments with 0.25 sec overlap as input. This layer is followed by a Maxpooling layer of window size 4. These two layers together extract local and time-invariant patterns from frame level MFCC vectors. The temporal dynamics of the MFCC vector sequences is then captured by two LSTM layers each with 64 cells and *tanh* activation. The hidden state outputs of the last LSTM layer at the last frame index are finally fed to a 2-unit dense layer with *softmax* activation to obtain the predicted class labels.

The classifiers are trained using Adam optimizer with binary cross entropy loss function. The learning rate is kept at 0.001 and the batch size is set to 16. The models are trained for a maximum of 100 epochs. To avoid overfitting, early stopping criteria with a patience of 8 based on validation loss is imposed. During testing, mode of the predictions corresponding to all chunks of a test utterance is considered as the final decision.

3.4. Evaluation Protocol

Five-fold cross-validation procedure is implemented with each fold comprising equal number of ALS, PD and HC subjects. All folds have similar distributions of age, gender and dysarthria severity scores. In every iteration of cross validation, data from three folds are used in training while data from one fold each are used in validation and testing. Mean and SD of classification accuracies obtained in five folds of evaluation are used as the performance metrics. It is to be noted here that, since all subjects could not perform three utterances of each phoneme or equal number of utterances of all phonemes, we can not include all utterances during the testing phases of the fusion experiments. However, all data are used for training in the respective folds. Table 2 reports the number of phoneme sets used for testing the fusion approaches. A set contains three utterances of the same phoneme in case of intra-phoneme fusion, whereas, one utterance each of /i/, /s/ and /sh/ comprises a phoneme set in inter-phoneme fusion case.

4. RESULTS AND DISCUSSION

4.1. SFs vs. SVs

Table 3 presents the ALS/PD vs. HC classification accuracies obtained using O-MFCC of the six phonemes under consideration. The average accuracies obtained over all fricatives are observed to be 5.70% and 5.38% higher than those achieved over all vowels in case of ALS vs. HC and PD vs. HC classifications, respectively. While comparing individual phonemes, all fricatives are found to outperform /a/ and /o/ w.r.t. mean accuracies in both classification tasks. /sh/ achieves the highest mean performance among the fricatives, followed by /s/. However, /i/ attains the best average accuracy among all the phonemes being studied. Productions of different fricatives require constrictions to be formed between different pairs of articulators, namely, tongue against the alveolar ridge in case of /s/, tongue behind the alveolar ridge in case of /sh/ and between lower lip and upper teeth in case of /f/. Though no constriction is formed while uttering the high vowel /i/, tongue is placed in close proximity of the palate. In contrast to these phonemes, /a/ and /o/ require the vocal tract to be relatively more open. Better performances of the fricatives and /i/ might indicate that ALS and PD population face difficulties

Table 2. Number of phoneme sets used during the testing phases of intra and inter-phoneme fusion approaches; for intra-phoneme case number of phoneme sets is equal to the number of subjects considered, whereas, in inter-phoneme fusion case the number of subjects is given in parentheses

Condition	Intra-phoneme fusion			Inter-phoneme fusion
	/i/	/s/	/sh/	
ALS	26	24	24	77 (33)
PD	24	27	28	84 (35)
HC	24	24	24	83 (35)

in forming constrictions in case of fricatives, or even, proximally placing tongue and palate in case of /i/. These lead to altered or compensatory configurations of the vocal tract and altered nature of airflow. However, it can be observed from Table 3 that the SD values of 5-fold classification accuracies are quite high in all cases. This might be due to the small size of the dataset considered in this work.

Figure 1 illustrates some specimen spectrograms of vowel /i/ and fricative /sh/ uttered by ALS, PD and HC subjects. In case of /i/, clear harmonics of fundamental frequency as well as prominent formant bars can be observed in the HC utterance. The harmonic structure becomes less evident in higher frequency regions of ALS and PD spectrograms, where the representations become more noise-like. The resonant energies corresponding to formants are also lower in these utterances as compared to the HC case. On the other hand, some harmonic structure indicating the presence of voicing can be observed in the /sh/ spectrogram for ALS, whereas that for HC comprises only high frequency components confirming voiceless nature of the utterance. Lack of high frequency component is observed in case of PD /sh/. All these factors presumably contribute towards differentiating ALS/PD and HC utterances.

In order to further validate the presence of unwanted voicing in SFs performed by the dysarthric subjects as opposed to HCs, we ex-

Table 3. Mean classification accuracies in % (SD in bracket) obtained using O-MFCC of different sustained phonemes

Phonemes		ALS vs. HC	PD vs. HC
Vowels	/a/	62.88 (7.91)	55.97 (9.89)
	/i/	**78.42 (10.03)**	**72.85 (12.04)**
	/o/	68.40 (5.47)	51.78 (8.73)
	Overall	**69.90**	**60.20**
Fricatives	/s/	76.90 (7.86)	65.37 (7.84)
	/sh/	77.47 (7.56)	66.66 (9.40)
	/f/	72.44 (6.24)	64.70 (10.43)
	Overall	**75.60**	**65.58**

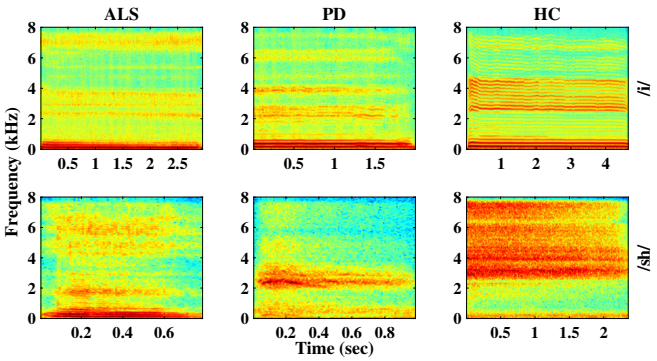


Fig. 1. Illustrative narrowband spectrograms of sustained /i/ and /sh/ utterances performed by ALS, PD and HC subjects

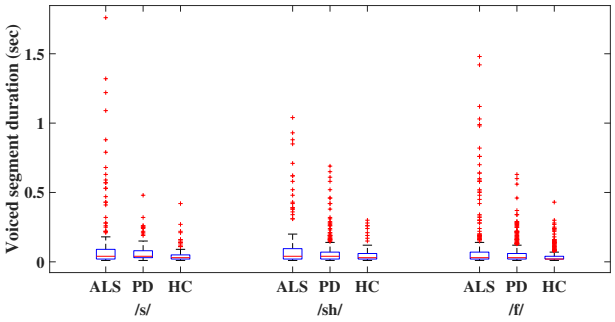
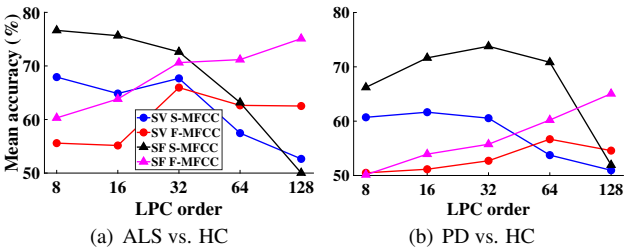


Fig. 2. Distributions of durations of voiced segments detected in different SFs produced by ALS, PD and HC subjects; durations in ALS and PD cases are longer than those of HC at 1% significance level as per Wilcoxon ranksum test

tract the pitch estimates of SFs at 100 Hz using the PRAAT software [24] with default pitch settings. The voiced segments present in the SFs are then identified as the segments of contiguous finite non-zero pitch values which are surrounded by zero pitch frames. Figure 2 shows the distributions of the durations of these voiced segments for ALS, PD and HC classes. Ideally no voicing should be present in the utterances of voiceless fricatives performed by HCs. However, voiced segments were actually detected in some of the HC SFs. Manual inspection confirmed those as algorithmic errors, which certainly would occur in the cases of ALS and PD also. To maintain the fairness of comparison, we report durations of all segments which were detected as voiced. Wilcoxon ranksum test [25] at 1% significance level on the voiced segment durations suggests that ALS and PD SFs have significantly longer voiced segments, and, hence, higher degree of unwanted voicing, than HC SFs.

4.2. Source - Filter Analysis

Figure 3 plots the ALS/PD vs. HC classification accuracies, averaged over all SVs and over all SFs, obtained using S-MFCC and F-MFCC estimated with varying LPC orders. It can be observed that, at lower LPC orders, S-MFCC outperforms F-MFCC for both SV and SF in case of both classification tasks, while F-MFCC achieves better performance at higher LPC orders. This is expected because at high LPC orders, more detailed structures are captured in the filter estimate and the source estimate becomes nearly white. Figure 3 further shows that, with S-MFCC, the obtained accuracies averaged over all SFs (plotted in black) are higher than those averaged over all SVs (plotted in blue) at most LPC orders for both classification tasks (except LPC order 128 in ALS vs. HC case). The trend is same for F-MFCC as well at all LPC orders except 8 in PD vs. HC case. The altered vocal tract shape due to restricted articulatory mobility and the impaired constriction formation might be responsible for the



(a) ALS vs. HC (b) PD vs. HC

Fig. 3. Mean classification accuracies (in %) over all SVs and those over all SFs obtained using S-MFCC and F-MFCC estimated with varying LPC orders

Table 4. Mean classification accuracies in % (SD in bracket) obtained using intra- and inter-phoneme decision-level fusion

	Fusion scheme	ALS vs. HC	PD vs. HC
Intra	**/i/+/i/+/i/**	81.83 (13.35)	**80.03 (11.96)**
	/s/+/s/+/s/	80.04 (8.58)	70.05 (13.19)
	/sh/+/sh/+/sh/	79.95 (8.90)	66.15 (11.36)
Inter	**/i/+/s/+/sh/ (Distinct model)**	82.02 (8.31)	75.67 (7.58)
	/i/+/s/+/sh/ (Pooled model)	**83.35 (5.93)**	72.65 (9.63)

discriminative abilities of SF F-MFCC. The disturbed constrictions might affect the air turbulence created at that site, thus influencing source estimation. This, combined with the unwanted voicing added to the voiceless fricatives by ALS and PD patients, may contribute towards the discriminative cues embedded in S-MFCC of SFs.

4.3. Effect of Fusion

Comparison of Tables 3 and 4 reveals that intra-phoneme decision-level fusion for /i/, /s/ or /sh/ achieves higher classification accuracy than the corresponding single utterances in all cases except /sh/ in PD vs. HC task where the performance remains nearly the same. This highlights the varied nature of cues captured in different utterances of a single phoneme. Inter-phoneme fusion provides further improvement in accuracy than intra-phoneme case for ALS vs. HC classification task. The pooled classifier model is observed to perform better than distinct models for different phonemes. Pooled model increases mean accuracy and reduces SD as compared to distinct model case, thus making the system more robust and efficient at the same time. The superior performance of inter-phoneme fusion over intra-phoneme case further emphasizes the complementary nature of the cues present in different phoneme utterances. This is evident because productions of different phonemes involve different courses of movements of the articulators. Thus, combining them broadens the scope of assessment of the articulators. However, no performance gain over intra-phoneme fusion case of /i/ is observed while using inter-phoneme fusion for PD vs. HC classification.

5. CONCLUSIONS

This work analyzes SFs in comparison with SVs for automatic ALS/PD vs. HC classification. Phonemes involving constrictions in the vocal tract (fricatives) or even close placement of tongue and palate (/i/) are found to be better differentiators than the relatively open ones. Presence of unwanted voicing in the utterances of voiceless fricatives performed by the patient population further contributes towards the classification capabilities of the fricatives. Moreover, different phonemes are observed to capture complementary cues making inter-phoneme fusion the best choice for ALS vs. HC classification. However, the same is not empirically true for PD vs. HC case. Though we claim proximity of certain articulators to be a prime factor, further verification is needed. Thus, an important future direction for this work would be to derive some quantifying measures of such proximity from the speech signals and to use those directly for performing the classifications.

Acknowledgement: We thank Navaneetha G and Agniv Chatterjee for their valuable assistance in data preparation. We also thank the Department of Science and Technology (DST), Govt. of India for supporting this work.

6. REFERENCES

[1] Lavoisier Leite and Ana Carolina Constantini, "Dysarthria and quality of life in patients with Amyotrophic Lateral Sclerosis," *Revista CEFAC*, vol. 19, pp. 664–673, 2017.

[2] 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.

[3] Jhansi Mallela, Aravind Illa, BN Suhas, 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.

[4] Aravind Illa, Deep Patel, BK Yamini, Meera SS, N Shivashankar, Preethish-Kumar Veeramani, Seena Vengalii, Kiran Polavarapui, Saraswati Nashi, Nalini Atchayaram, and Prasanta Kumar Ghosh, "Comparison of speech tasks for automatic classification of patients with Amyotrophic Lateral Sclerosis and healthy subjects," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6014–6018.

[5] Frank Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163–1177, 2013.

[6] Changqin Quan, Kang Ren, and Zhiwei Luo, "A deep learning based method for Parkinson's disease detection using dynamic features of speech," *IEEE Access*, vol. 9, pp. 10239–10252, 2021.

[7] Biswajit Karan, Sitanshu Sekhar Sahu, Juan Rafael Orozco-Arroyave, and Kartik Mahto, "Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction," *Computer Speech & Language*, vol. 69, pp. 101216, 2021.

[8] Maxim Vashkevich and Yu Rushkevich, "Classification of ALS patients based on acoustic analysis of sustained vowel phonations," *Biomedical Signal Processing and Control*, vol. 65, pp. 102350, 2021.

[9] C Okan Sakar, Gorkem Serbes, Aysegul Gunduz, Hunkar C Tunc, Hatice Nizam, Betul Erdogdu Sakar, Melih Tutuncu, Tarkan Aydin, M Erdem Isenkul, and Hulya Apaydin, "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform," *Applied Soft Computing*, vol. 74, pp. 255–263, 2019.

[10] Biswajit Karan, Sitanshu Sekhar Sahu, Juan Rafael Orozco-Arroyave, and Kartik Mahto, "Hilbert spectrum analysis for automatic detection and evaluation of Parkinson's speech," *Biomedical Signal Processing and Control*, vol. 61, pp. 102050, 2020.

[11] Alberto Tena, Francesc Clarià, Francesc Solsona, and Mònica Povedano, "Detecting bulbar involvement in patients with Amyotrophic Lateral Sclerosis based on phonatory and time-frequency features," *Sensors*, vol. 22, no. 3, pp. 1137, 2022.

[12] Diogo Braga, Ana M Madureira, Luis Coelho, and Reuel Ajith, "Automatic detection of Parkinson's disease based on acoustic analysis of speech," *Engineering Applications of Artificial Intelligence*, vol. 77, pp. 148–158, 2019.

[13] Kris Tjaden and Greg S Turner, "Spectral properties of fricatives in Amyotrophic Lateral Sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 6, pp. 1358–1372, 1997.

[14] Yunjung Kim, "Acoustic characteristics of fricatives /s/ and /ʃ/ produced by speakers with Parkinson's disease," *Clinical archives of communication disorders*, vol. 2, no. 1, pp. 7, 2017.

[15] Paul McRae and Kris Tjaden, "Spectral properties of fricatives in Parkinson's disease," *The Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 1854–1854, 1998.

[16] 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," in *Proc. 21st Annual Conference of the International Speech Communication Association, Shanghai, China*, 2020, pp. 4586–4590.

[17] BN Suhas, Jhansi Mallela, Aravind Illa, BK Yamini, 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.

[18] BN Suhas, Deep Patel, Nithin Rao Koluguri, Yamini Belur, Pradeep Reddy, Atchayaram Nalini, 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.

[19] Gunnar Fant, *Acoustic theory of speech production*, Walter de Gruyter, 1970.

[20] Jesse M Cedarbaum, Nancy Stambler, Errol Malta, Cynthia Fuller, Dana Hilt, Barbara Thurmond, Arline Nakanishi, BDNF ALS Study Group, and 1A complete listing of the BDNF Study Group, "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the neurological sciences*, vol. 169, no. 1-2, pp. 13–21, 1999.

[21] Douglas J Gelb, Eugene Oliver, and Sid Gilman, "Diagnostic criteria for Parkinson disease," *Archives of neurology*, vol. 56, no. 1, pp. 33–39, 1999.

[22] Andreas I Koutrouvelis, George P Kafentzis, Nikolay D Gaubitch, and Richard Heusdens, "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 316–328, 2015.

[23] 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, IEEE Catalog No.: CFP11SRW-USB.

[24] Paul Boersma and David Weenink, "Praat: doing phonetics by computer [computer program]," 2022, Version 6.2.06, retrieved 24 August 2022 from https://www.praat.org.

[25] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50 – 60, 1947.