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/ICASSP49357.2023.10094959

STATIC AND DYNAMIC SOURCE AND FILTER CUES FOR CLASSIFICATION OF AMYOTROPHIC LATERAL SCLEROSIS PATIENTS AND HEALTHY SUBJECTS

Tanuka Bhattacharjee[1*], *Chowdam Venkata Thirumala Kumar*[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) affects speech production. Even the elementary sustained vowel utterances get impaired. For these, the impairments can be in achieving vowel-specific articulatory configurations, reflected in *static* acoustic cues, and/or in sustaining a configuration for a prolonged duration, reflected in *dynamic* cues. Such cues can further be attributed to the *vocal cord (source)* and *vocal tract (filter)* involved in speech production. This paper analyzes the relative contributions of these static (captured through average spectral characteristics) and dynamic (captured through spectral variations over time) source and filter cues toward automatic classification of ALS patients and healthy subjects using sustained utterances of /a/, /i/, /o/ and /u/. Experiments with 80 ALS patients and 80 healthy subjects suggest that the source cues (static/dynamic) are not the primary discriminators. For /i/, the static filter cues achieve the highest mean classification accuracy of 76.66%, whereas, for /a/, /o/ and /u/, the dynamic filter attributes contribute the most attaining average accuracies of 66.29%, 73.03% and 70.27%, respectively. Hence, ALS patients seem to face difficulties in forming the front closed vocal tract structure of /i/, whereas, holding the target vocal tract shape for long appears to be the primary challenge in case of /a/, /o/ and /u/.

Index Terms— Amyotrophic Lateral Sclerosis, vowel, static, dynamic, source-filter

1. INTRODUCTION

The neuro-degenerative Amyotrophic Lateral Sclerosis (ALS) disease impairs the speech musculature, among others, leading to dysarthria. The speed and/or range of movements of lips, jaw, tongue and velum get severely restricted [1, 2]. Poor laryngeal control during phonation leads to erroneous voicing and abnormal prosodic patterns like reduced pitch range [2]. Dysfunctions in the respiratory and resonatory [2] sub-systems of speech are also evident.

Sustained vowel (SV) productions get critically affected in dysarthria due to ALS. According to the source-filter model [3], during the production of a vowel sound, airflow from the lungs passes through the vibrating vocal folds generating a quasi-periodic (voiced) source signal (S) with minimal aperiodic components [4]. This signal then passes through the vocal tract which acts as a filter (F) and produces the vowel sound. Specific vocal tract configurations give rise to specific vowels. During an SV production, a subject is supposed to prolong a vowel with correct pronunciation while maintaining uniform pitch and loudness. Thus, it not only calls for achieving the target S and F configurations specific to a vowel, but also for uniformly sustaining that designated structure for a prolonged duration. Due to restricted muscular control, ALS patients might face difficulties in accomplishing either/both of these goals. They are often reported to make compensatory articulatory movements to mimic targeted sounds [5]. This paper captures the deformities in the gross S and F configurations through *static cues* (ST) and the unusual temporal variations in these configurations through *dynamic cues* (DY), as elaborated in Table 1. We aim to analyze the relative discriminative capabilities of source-static (S-ST), source-dynamic (S-DY), filter-static (F-ST) and filter-dynamic (F-DY) cues for SV-based ALS vs healthy control (HC) classification.

Several acoustic analyses of vowels pronounced by ALS patients are present in the literature. Lee et al. [6] have observed the vowel /i/ to undergo the highest decline in intelligibility with increase in dysarthria severity. Reduced acoustic vowel contrast owing to impaired tongue movements [7, 8] and frequent mis-identifications in the height dimension of vowels due to limited tongue height control [9] are also reported in case of ALS subjects as compared to HCs. Further, researchers have performed automatic ALS vs HC classification using various acoustic cues derived from SVs, particularly /a/, /e/, /i/, /o/, /u/ and /æ/. Mel frequency cepstral coefficients (MFCC) and log mel spectrograms have been explored in [10, 11, 12, 13]. Along with MFCC, Vashkevich et al. [11] have analyzed a wide variety of other features like jitter, shimmer, harmonic structure etc. Tena et al. [14] have examined phonatory-subsystem and time-frequency features. In [15], a 1D-convolutional neural network (CNN) has been used for learning representations from raw speech waveforms. Though all of S-ST, S-DY, F-ST and F-DY cues have been implicitly incorporated in these approaches, none of these works attempts to identify the relative utilities of these four types of cues for the classification task at hand. There lies the contribution of this paper. Thus, our aim is not to outperform the state-of-the-art classification approaches, but to understand which cues play role towards the discrimination.

We analyze four SVs - /a/, /i/, /o/ and /u/, for automatic ALS vs HC classification. Mean of MFCC with delta coefficients (M_m) and standard deviation (SD) of spectral amplitudes over time at the first 8 harmonic frequencies (H_d) are used as the ST and DY features, respectively. Experimental validations using linear discriminant analysis (LDA) classifier confirm that these two features, when extracted from the original SV utterances, can together encode the major discriminative information present in the SVs, thereby helping the simple LDA classifier achieve similar level of classification accuracy as state-of-the-art feature engineering based approaches [11] as well as CNN-LSTM algorithms [10] (LSTM stands for long short term memory). This substantiates the use of these two particular features

[*]These authors contributed equally to this work.

Table 1. Description of static (ST) and dynamic (DY) cues present in source (S) and filter (F) components of dysarthric SVs

		Description	Potential reason	Clinical sign	Acoustic cues
S	**ST**	Unusual average characteristics of source excitation	Impaired respiratory and laryngeal function [16]	Weakened or strained voice, hoarseness [17]	Mean harmonic-to-noise ratio, average loudness
	DY	Unusual temporal variations in source excitation	Impaired laryngeal control [2]	Difficulties in controlling pitch [17]	Jitter, pitch period entropy [11]
F	**ST**	Impaired vocal tract configuration	Restricted articulatory mobility [2]	Poor articulation [16]	Mean spectral envelope, mean log-area ratio
	DY	Unusual temporal fluctuations in vocal tract configuration	Articulatory muscle weakening [16]	Irregular articulation [16]	Temporal variations in spectral envelope

as the representative ST and DY cues. To further capture these cues specific to S and F components of SVs, we propose to manipulate the SVs using the WORLD vocoder [4] in such ways that only the required components are preserved in the modified utterances while suppressing the redundant attributes. It is observed that the relative discriminative capabilities of S-ST, S-DY, F-ST and F-DY cues are vowel dependent. None of S-ST and S-DY cues of any SV is found to be a major discriminator between ALS and HC groups. Among the F cues, F-ST turns out to be the best discriminator in case of /i/, while F-DY attributes perform the best for /a/, /o/ and /u/. Thus, ALS patients seem to find it difficult to position the tongue in close proximity of palate while uttering the front close vowel /i/, whereas, maintaining the vocal tract structure for a long duration seems to be the primary challenge in case of the other three vowels.

2. DATASET

Sustained utterances of /a/, /i/, /o/ and /u/ were collected from 80 ALS (50M, 30F) and 80 HC (62M, 18F) subjects at National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India. The ALS and HC groups had ages in the ranges of 28 - 77 and 22 - 65 years, respectively. Three speech-language pathologists from NIMHANS rated the dysarthria severity of the ALS patients following the 5 point speech component of the ALSFRS-R scale [18]. The mode of these three ratings was taken as the final severity. Equal number of patients were recruited from each severity level. Upto 3 sustained utterances of a vowel were recorded from each subject totalling 858 and 842 utterances from the ALS and HC groups, respectively. For both groups, nearly equal number of utterances belonged to each vowel. The mean (SD) of durations of the utterances were 4.05 (2.29) and 5.71 (1.98) sec, respectively, for ALS and HC subjects. All utterances were recorded at a sampling frequency of 44.1 kHz and then downsampled to 16 kHz. More details about the data collection protocol and the recording setup are present in [15].

3. METHOD

The proposed method of extracting ST and DY cues associated with the S and F components of SVs comprises four steps, namely, decomposition, modification, synthesis and feature extraction, as illustrated in Figure 1. The first three steps are explained next, followed by the choice of the specific ST and DY features to be considered.

First, an SV utterance is decomposed into fundamental frequency (F_0), spectral envelope (SP) and aperiodicity (AP) components using the WORLD analyzer [4]. Different types of modifications, as listed next, are then applied to these components to retain only the required attributes in the signal synthesized subsequently.

1. To remove the effect of F from an SV utterance, the estimated spectral envelope is modified to 1s in all frequency bands. Speech is then synthesized by WORLD synthesizer using the modified spectral envelope along with the unchanged F_0 and aperiodicity. This makes F allpass in nature, and hence, the modified signal captures only S without any influence of the vocal tract. ST and DY features obtained from this modified utterance serve as S-ST and S-DY cues.
2. To extract F information without any influence of S, we devoice an SV utterance by replacing the F_0 estimates obtained from WORLD analyzer with 0s and the aperiodicity for all frequency bands with 1s [19]. Speech is then synthesized by WORLD synthesizer with the modified F_0 and aperiodicity along with the unchanged spectral envelope. This makes S white throughout the utterance while retaining the F characteristics. ST measures computed from this modified signal serve as F-ST cues.
3. Lastly, we replace the F_0 and aperiodicity estimates obtained from WORLD analyzer with their respective mean values over the middle 1 sec of the utterance. Speech is then synthesized by WORLD synthesizer with the modified F_0 and aperiodicity along with the unchanged spectral envelope. Thus, S profile is made constant throughout this modified speech, while variations in F are preserved. DY cues extracted from this signal capture F-DY attributes.

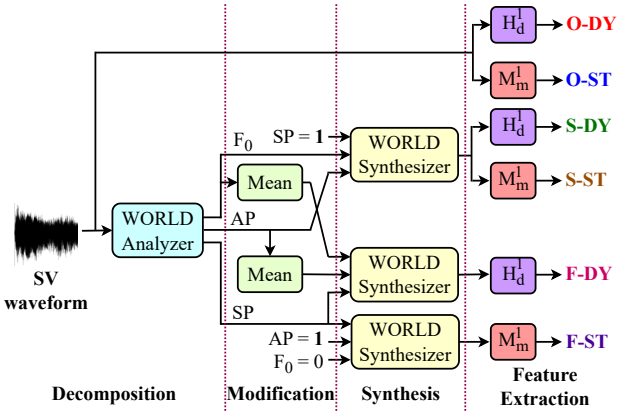

Fig. 1. Proposed method for extracting ST and DY cues from SV utterances and their S and F components; here, AP and SP refer to aperiodicity and spectral envelope, M_m^1 and H_d^1 refer to M_m and H_d computed from the middle 1 sec of an utterance and **1** denotes a matrix with all entries as 1

We also extract ST and DY features from the original SV utterances, referred as O-ST and O-DY, respectively, which capture cues associated to S and F components together. All the cues described above are used individually to perform ALS vs HC classification in a vowel-specific manner. LDA is used as the classifier in all cases.

Choice of static and dynamic cues: Vashkevich et al. [11] have explored an extensive set of SV features for ALS vs HC classification which includes cues of both ST and DY natures. We consider the same set excluding two features - distance of spectral envelopes

and convergence of second formants of /a/ and /i/, as these cannot be mapped to individual vowels. Among this adopted 64D feature pool, we group mean harmonic-to-noise ratio (H/N_m) (1D), mean glottal-to noise excitation ratio (G_m) (1D), mean spectral amplitudes over time at the first 8 harmonic frequencies (H_m) (8D) and M_m (24D) as ST cues because they capture the average characteristics of the speech signals. H/N_m and G_m quantify average noise excitation properties, whereas, H_m and M_m record the average spectral profile. The remaining features, namely, jitter (JT) (4D), shimmer (SH) (5D), directional perturbation factor (DPF) (1D), SD of glottal-to noise excitation ratio (G_s) (1D), phonatory frequency range (PFR) (1D), pitch period entropy (PPE) (1D), pathological vibrato index (PVI) (1D), H_d (8D) and RelH (H_r) (8D), are clubbed in the DY group as these are descriptive of the temporal variations in speech attributes. JT and DPF capture perturbations in F_0. SH estimates amplitude perturbations in the speech signal. PFR, PPE and PVI encode modulation properties of F_0. G_s measures the variations in the noise excitation, whereas, spectral variations are quantified by H_d and H_r. We perform vowel-specific ALS vs HC classification using each of these features separately. The feature having the highest average classification accuracy over the four vowels is then chosen from each of ST and DY groups. M_m and H_d emerge as the best performing features in their respective groups (refer Section 5), and hence, are considered as the representative ST and DY cues for all experiments. Following [11], we consider the complete SVs for feature computation while selecting the best ST and DY cues. However, the two best features derived from only the middle 1 sec of the utterances (denoted as M_m^1 and H_d^1) are considered subsequently. This is because the most stable articulatory configuration, without transient variations, is expected to be attained during the middle portion. If an SV utterance lasts ≤ 1 sec, then we consider the entire utterance.

4. EXPERIMENTAL SETUP

Feature Extraction

To compute the features taken from [11], we use the implementations given by the authors (https://github.com/Mak-Sim/Troparion). However, the implementations for H/N_m and the harmonic measures are not available. So, we calculate H/N_m in the PRAAT software [20] with pitch range set to 50-450 Hz and all other parameters fixed at their default values. For computing the harmonic features, we follow the steps mentioned in [11] with the only exception that we sample the spectral amplitudes at $(p \times 16)$ bins, where $p = 1, 2, ..., 8$. These bins approximately correspond to the first 8 harmonics of F_0.

To estimate S-ST, S-DY, F-ST and F-DY cues, an SV utterance is decomposed, modified and synthesized using WORLD, as elaborated in Section 3. During decomposition, F_0 estimates of speech are obtained with a frame period of 5 ms using IRAPT algorithm [21]. The floor and ceiling frequencies of the F_0 estimation range are set to 50 Hz and 450 Hz, respectively, to match with the settings in [11]. Spectral envelope and aperiodicity are estimated using CheapTrick [22] and D4C [23] algorithms, respectively. M_m^1 and H_d^1 are finally extracted from the modified utterances.

Evaluation Protocol

All experiments are performed in the 5-fold cross-validation setup. Each disjoint fold contains equal number of subjects, and hence nearly equal number of utterances of each vowel, from ALS and HC classes. The distributions of age, gender and dysarthria severity are similar across the folds. We report the mean and SD of classification accuracies obtained in the 5 folds as the performance metrics. Moreover, Wilcoxon signed-rank test [24] at 1% significance level is carried out to determine if the classification accuracies obtained using different feature sets are significantly different. For that, subjects from the test fold in each iteration are divided into 4 random groups of equal sizes. The 20 classification accuracies thus obtained are considered for the signed-rank test.

5. RESULTS AND DISCUSSION

Determining ST and DY cues: Figure 2 illustrates the ALS vs HC classification accuracies obtained in the cases of the four vowels while using individual ST and DY features adopted from [11]. Here, the features are computed from the complete durations of the original SV utterances. Among the ST group, M_m is found to achieve the highest average accuracy of 68.13% over the four vowels. H_d attains the highest mean accuracy of 70.81% over all vowels among the DY group. Hence, these two features are selected as the representative ST and DY cues. As mentioned in Section 3, the middle 1 sec of the sustained utterances are expected to better capture the stable ST and DY patterns without any transient effect. It can be observed from the first four rows of Table 2 that the classification accuracies obtained using M_m^1 and H_d^1 extracted from the middle 1 sec of the original SVs are statistically similar to those achieved using M_m and H_d computed from the complete durations of the same utterances. However, the SD of accuracies are lower in most cases while considering only the middle 1 sec of SVs as compared to the entire utterances. That is, considering only the stable segments of speech enhances the consistency in the performance as the transient artifacts are minimized in this case. However, even after considering only the middle 1 sec of the utterances, high SD of classification accuracies are obtained in some cases during the current and subsequent analyses. This is due to the small size of the dataset considered in this work.

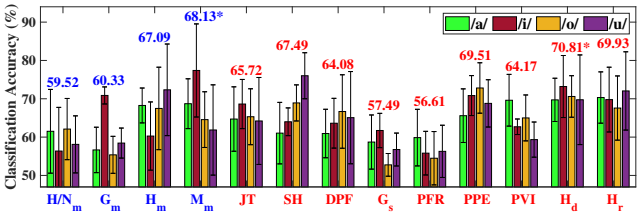

Fig. 2. Mean ALS vs HC classification accuracies (SD in error bar) obtained using different ST (blue) and DY (red) cues extracted from complete durations of SVs; accuracies averaged over all vowels are shown on top of each group of bars; * indicates the features having the highest average accuracy over all vowels among each of ST and DY groups

Table 2. Mean ALS vs HC classification accuracies in % (SD in bracket) obtained using representative ST and DY cues of original SVs as compared to baseline feature sets; here * indicates that H_d^1 outperforms M_m^1 as per signed-rank test

Features	Vowels			
	/a/	/i/	/o/	/u/
M_m	68.72 (6.50)	77.39 (12.14)	64.56 (7.27)	61.86 (11.78)
H_d	69.71 (5.64)	73.20 (8.09)	70.60 (5.42)	69.74 (11.70)
M_m^1 (O-ST)	62.24 (7.35)	75.75 (10.92)	64.12 (7.41)	58.80 (6.55)
H_d^1 (O-DY)	73.92 (3.20)*	71.69 (4.50)	75.57 (2.44)*	68.49 (3.28)
M_m^1 + H_d^1	70.80 (5.20)	79.37 (9.70)	74.28 (7.29)	71.62 (8.29)
Baseline-64D (from entire utterance)	73.76 (8.36)	81.00 (5.63)	73.22 (6.33)	73.24 (3.28)
Baseline-64D (from middle 1.5 sec)	73.85 (5.09)	80.74 (4.97)	70.81 (9.78)	71.36 (6.87)

Comparison with baseline: We compare the performances of M_m^l + H_d^l and the baseline 64D feature set computed from the entire SV utterances. However, for fair comparison, the baseline should also be extracted from the middle 1 sec of the utterances. But in doing so, PVI computation with the implementation given by the authors of [11] leads to 0 values for all utterances. To avoid this issue, we proceed to compute all features from the middle 1.5 sec of the utterances to form a second baseline. As shown in the last three rows of Table 2, M_m^l + H_d^l can achieve classification accuracies which are statistically equivalent to those obtained using the baseline feature sets computed from both entire utterances and middle 1.5 sec of the utterances. That means, only these two features together can capture discriminative cues to the same extent as the much larger 64D feature set. The performance of M_m^l + H_d^l is also found to be comparable with the state-of-the-art MFCC + CNN-LSTM approach [10] which achieves the mean (SD) of classification accuracies (in %) as 77.82 (6.12), 68.62 (5.13), 74.19 (4.80) and 64.96 (8.87) for /a/, /i/, /o/ and /u/, respectively. In fact, as per the signed-rank test, M_m^l + H_d^l significantly outperforms MFCC + CNN-LSTM in case of /i/. Hence it is justified to use only M_m^l and H_d^l as the representative ST and DY features, respectively, for further experimentations.

Comparison of O-ST and O-DY features: The two features M_m^l and H_d^l extracted from the middle 1 sec of the original SV utterances serve as O-ST and O-DY cues, respectively, as shown in Figure 1. Table 2 tells that the relative contribution of O-ST and O-DY cues towards ALS vs HC classification is vowel dependent. This is expected as pronunciations of different vowels require different articulatory involvements. O-ST attribute achieves higher average classification accuracy than O-DY in case of the front close vowel /i/ indicating that the gross articulatory configuration for /i/ differ predominantly between ALS and HC subjects. However, the performance of O-DY is also not statistically inferior. Hence, the ALS patients also seem to introduce some degrees of unwanted fluctuations in the articulatory configuration while sustaining /i/. In the cases of /a/, /o/ and /u/, O-DY outperforms O-ST. For /a/ and /o/, the superiority of O-DY is statistically significant. Thus in these cases, the major differences between ALS and HC utterances seem to lie in the extent of variations in the target configuration over the course of an utterance.

Comparison of ST and DY cues of S and F: Next we proceed to accredit the discriminative capabilities of O-ST and O-DY cues to those of S-ST, F-ST and S-DY, F-DY cues, respectively. Vowel-wise classification performances of these four types of cues are listed in Table 3. For /i/, the highest average classification accuracy is achieved using F-ST, which significantly outperforms S-ST. Though F-DY also significantly outperforms S-DY for /i/, the mean performance of F-DY is lower than that of F-ST. On the other hand, F-DY features attain the best average classification accuracies for the remaining vowels. For /o/ and /u/, the superiority of F-DY over S-DY is statistically significant as well. These observations might signify that the ALS subjects find it difficult to achieve the target front closed vocal tract configuration of /i/, possibly due to the impaired tongue height control as reported in [9]. For /a/, /o/ and /u/, maintaining the required vocal tract structure all along a prolonged utterance might be most difficult, possibly due to muscle weakening. The inferior performances of S-ST and S-DY features might suggest that the source excitation is less discriminative between ALS and HC utterances. Nonetheless, some impairments indeed exist in the source excitation of ALS SVs which lead to ALS vs HC classification accuracies above the chance level while employing S-ST and S-DY cues.

Table 3. Mean ALS vs HC classification accuracies in % (SD in bracket) obtained using representative ST and DY cues of S and F components of SVs; here # and † indicate respectively that F-ST significantly outperforms S-ST and F-DY significantly outperforms S-DY as per signed-rank test

Features	Vowels			
	/a/	/i/	/o/	/u/
S-ST	55.27 (2.82)	61.85 (7.83)	56.32 (5.33)	55.82 (8.26)
S-DY	62.11 (2.68)	57.90 (5.86)	60.00 (4.59)	57.18 (5.16)
F-ST	60.25 (6.57)	76.66 (12.90)$^{\#}$	64.27 (6.55)	63.51 (6.60)
F-DY	66.29 (8.43)	68.86 (1.91)†	73.03 (3.49)†	70.27 (5.27)†

Table 4. Mean ALS vs HC classification accuracies in % (SD in bracket) obtained using F-DY cues of mismatched utterances

		F_0 + aperiodicity			
		/a/	/i/	/o/	/u/
spectral envelope	/a/	-	66.85 (6.03)	75.13 (3.85)	76.93 (2.82)
	/i/	73.08 (2.49)	-	69.57 (6.47)	70.75 (3.23)
	/o/	74.37 (4.38)	66.55 (3.73)	-	73.07 (4.22)
	/u/	71.22 (4.70)	69.70 (4.43)	74.40 (6.49)	-

Effect of harmonic locations: The harmonic locations (attribute of S) are the frequencies at which the speech spectrum is sampled to compute the DY H_d^l cue. These harmonics are kept constant throughout the utterance during F-DY computation, thereby nullifying the DY effects of S. Since F-DY cues turn out to be the primary discriminator between ALS and HC groups in case of three out of four vowels, we proceed to investigate further if the locations of the harmonics (though constant) used for obtaining F-DY cues play any role towards the discriminative capabilities inherent in the feature. For this purpose, we decompose the original SV utterances using WORLD, as described in Section 3, and synthesize *mismatched* utterances by replacing the obtained F_0 and aperiodicity estimates with the average F_0 and aperiodicity over the middle 1 sec of some random utterance of a different vowel while keeping the spectral envelope unchanged. H_d^l is then extracted from these mismatched utterances. Table 4 shows that F_0 and aperiodicity of /a/, /o/ and /u/ when used with the spectral envelope of any vowel for F-DY computation lead to mostly similar levels of ALS vs HC classification accuracies, while F_0 and aperiodicity of /i/ always have inferior performance. Hence, the locations of the harmonics, or in other words, the frequencies at which the spectrum is sampled, are indeed important for capturing F-DY attributes.

6. CONCLUSION

This paper analyzes the SV-based ALS vs HC classification from the perspective of static and dynamic cues of source and filter components of speech. Depending on the vowel at hand, different cues are found to capture predominant discriminative information. In case of /i/, static filter cues are observed to be the best discriminator among the four types of features. However, for /a/, /o/ and /u/, dynamic filter cues achieve the highest mean classification accuracies. Achieving the vocal tract configuration involving proximal placement of the tongue and palate, specific to the front close vowel /i/, seems to get difficult for the patients having ALS-induced dysarthria. On the other hand, maintaining a constant vocal tract shape seems to become the primary hurdle in the cases of the other three vowels - /a/, /o/ and /u/. An interesting future direction for this work would be to analyze the effect of increasing dysarthria severity on the ST and DY cues under consideration.

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.

7. REFERENCES

[1] 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.

[2] Panying Rong, Yana Yunusova, Jun Wang, Lorne Zinman, Gary L Pattee, James D Berry, Bridget Perry, and Jordan R Green, "Predicting speech intelligibility decline in Amyotrophic Lateral Sclerosis based on the deterioration of individual speech subsystems," *PloS one*, vol. 11, no. 5, pp. e0154971, 2016.

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

[4] Masanori Morise, Fumiya Yokomori, and Kenji 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.

[5] Sanjana Shellikeri, Yana Yunusova, Danielle Thomas, Jordan R Green, and Lorne Zinman, "Compensatory articulation in Amyotrophic Lateral Sclerosis: Tongue and jaw in speech," in *Proceedings of Meetings on Acoustics*. Acoustical Society of America, 2013, vol. 19, p. 060061.

[6] Jimin Lee, Emily Dickey, and Zachary Simmons, "Vowel-specific intelligibility and acoustic patterns in individuals with dysarthria secondary to Amyotrophic Lateral Sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 1, pp. 34–59, 2019.

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

[8] Panying Rong, Evan Usler, Linda M Rowe, Kristen Allison, Jonghye Woo, Georges El Fakhri, and Jordan R Green, "Speech intelligibility loss due to Amyotrophic Lateral Sclerosis: The effect of tongue movement reduction on vowel and consonant acoustic features," *Clinical Linguistics & Phonetics*, vol. 35, no. 11, pp. 1091–1112, 2021.

[9] Jimin Lee, Heejin Kim, and Yong Jung, "Patterns of misidentified vowels in individuals with dysarthria secondary to Amyotrophic Lateral Sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 8, pp. 2649–2666, 2020.

[10] 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.

[11] 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.

[12] 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.

[13] BN Suhas, 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.

[14] 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.

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

[16] 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.

[17] "How ALS affects speech," https://www.targetals.org/2022/03/28/how-als-affects-speech/, [Online; accessed 25-Oct-2022].

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

[19] Tanuka Bhattacharjee, Jhansi Mallela, Yamini Belur, Nalini Atchayaram, Ravi Yadav, Pradeep Reddy, Dipanjan Gope, and Prasanta Kumar Ghosh, "Source and vocal tract cues for speech-based classification of patients with Parkinson's disease and healthy subjects.," in *INTERSPEECH*, 2021, pp. 2961–2965.

[20] Paul Boersma and David Weenink, "Praat: doing phonetics by computer [computer program], version 6.2.06," *retrieved 23 January 2022 from https://www.praat.org*, 2022.

[21] Elias Azarov, Maxim Vashkevich, and Alexander Petrovsky, "Instantaneous pitch estimation based on RAPT framework," in *Proceedings of the* 20^{th} *European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 2787–2791.

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

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

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