JASA Express Letters

DECEMBER 16 2024

Inter-speaker acoustic differences of sustained vowels at varied dysarthria severities for amyotrophic lateral sclerosis

Tanuka Bhattacharjee; Seena Vengalil; Yamini Belur; Nalini Atchayaram; Prasanta Kumar Ghosh

JASA Express Lett. 4, 125203 (2024)
https://doi.org/10.1121/10.0034613

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Inter-speaker acoustic differences of sustained vowels at varied dysarthria severities for amyotrophic lateral sclerosis

Tanuka Bhattacharjee,[1] Seena Vengalil,[2] Yamini Belur,[2,a)] Nalini Atchayaram,[2] and Prasanta Kumar Ghosh[1]

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

tanuka1111@gmail.com, seenavengalil@gmail.com, yamiini.bk@gmail.com, atchayaramnalini@yahoo.co.in, prasantg@iisc.ac.in

Abstract: We study inter-speaker acoustic differences during sustained vowel utterances at varied severities of Amyotrophic Lateral Sclerosis-induced dysarthria. Among source attributes, jitter and standard deviation of fundamental frequency exhibit enhanced inter-speaker differences among patients than healthy controls (HCs) at all severity levels. Though inter-speaker differences in vocal tract filter attributes at most severity levels are higher than those among HCs for close vowels /i/ and /u/, these are comparable with or lower than those among HCs for the relatively more open vowels /a/ and /o/. The differences typically increase with severity except for a few parameters for /a/ and /i/.

[Editor: Douglas D O'Shaughnessy] https://doi.org/10.1121/10.0034613

Received: 26 June 2024 **Accepted:** 23 November 2024 **Published Online:** 16 December 2024

1. Introduction

Inter-speaker differences in speech acoustics is a well known phenomenon. According to the source-filter model of speech production,[1] the glottal or supra-glottal excitation, denoted as the source signal, passes through the vocal tract, which acts as a filter to generate the speech utterances. Hence, the inter-speaker differences can be inherent to either or both of the source and filter components. We aim to compare the degree of these differences existing among amyotrophic lateral sclerosis (ALS) patients at varied dysarthria severity levels with those prevailing among the healthy control (HC) subjects. We also perform similar comparisons between different severity levels of ALS-induced dysarthria.

Dysarthria due to ALS impairs both source and filter components of speech utterances. Impairments in the source component are primarily caused by poor laryngeal control and compromised respiratory functionality.[2] Erroneous voicing, abnormal prosodic patterns, and poor voice quality characterize these impairments.[2] Features capturing these aspects like fundamental frequency (f_0), jitter, shimmer, and harmonic-to-noise ratio (HNR) have been widely used for automatic ALS vs HC classifications.[3] Restricted articulatory mobility and dysfunctions in the resonatory sub-system of speech lead to impaired filter functions.[2] Velocities of movements of articulators like lips, jaw, tongue, and velum decrease in ALS.[2] Patients often make compensatory articulatory configurations to mimic some target sounds,[4] e.g., they often exaggerate lip protrusion to compensate for impaired tongue retraction.[5] Impairments in the filter component lead to imprecise and irregular articulations[6] as well as atypical spectral characteristics of speech utterances. Formants and spectral envelopes of sustained vowels capturing such impairments have also been used for the automatic detection of ALS.[3] The timing and degree of involvement of different speech sub-systems vary across individuals due to the heterogeneous phenotypic expression of ALS.[7] These variations can impact the degree of inter-speaker differences in the source and filter attributes of speech.

A few studies exploring the inter-speaker differences in voice acoustics for HC subjects and ALS patients have been reported in the literature. Hallin *et al.*[8] have observed large inter-speaker variations of speech range profile area during running speech and voice range profile area during sustained phonations and glissandi on /a:/ among HCs. Significant differences have been observed by Ternström and Pabon[9] in how the spectrum balance varies over the voice range among HCs. On the other hand, voice dysfunction in ALS is reported to result in varying acoustic changes across individual speakers.[10] According to Strand *et al.*,[11] the effect of ALS on f_0 during sustained utterance of /a/ is not universal for all patients. These inter-speaker differences could be due to the differences in multiple laryngeal parameters, e.g., muscle

a)Author to whom correspondence should be addressed.

mass, cricothyroid function, and degree of spasticity or flaccidity of laryngeal muscles. Kent *et al.*[12] have reported inter-speaker variability among the ALS population with respect to jitter and shimmer during sustained utterances of /a/.

Though the previously-mentioned few studies have analysed the inter-speaker differences in a few aspects of the source component of ALS speech, no study has reported a systematic statistical comparison between the degree of such differences existing among the ALS subjects of a certain dysarthria severity level and that prevailing among the HCs. No such systematic comparison between different dysarthria severity levels has also been reported. Moreover, the filter level inter-speaker differences existing among the ALS population remain relatively unexplored to date. We aim to perform a statistical analysis to understand if and to what extent the inter-speaker differences in the source and the filter components of speech utterances get altered at different severity levels of ALS-induced dysarthria. This study can help us understand how speech production is affected at different severity levels of this disease. These insights can further help in designing robust speaker normalization strategies required to develop speaker-agnostic models for speech recognition, language identification, speech enhancement, etc., for the ALS subjects of different dysarthria severity levels. Such normalization strategies can also be useful in acoustic cohort studies for these patients.

2. Data

We consider sustained utterances of four vowels, namely, /a/, /i/, /o/, and /u/. Sustained vowel utterances are chosen as these are relatively time invariant, easy to produce/elicit, and less susceptible to influences related to language, dialect, etc., as compared to continuous speech.[13] Since different vowel productions have different articulatory and acoustic targets, all vowels are not equally affected by the impairments in a particular speech subsystem. As ALS affects different speech subsystems or different parts of a speech sub-system at different times for different individuals during the disease progression,[7] we consider multiple vowels to capture the effects of a larger spectrum of impairments.

Data collection was performed at the National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India. We recruited 35 ALS (18 M + 17 F; age range: 36–70 years) and 40 HC (22 M + 18 F; age range: 34–65 years) subjects. The native languages of the subjects included Bengali, Gujarati, Hindi, Kannada, Malayalam, Tamil, Telugu, and Urdu. Three speech-language pathologists (SLPs) rated the dysarthria severity of the ALS subjects based on the perception formed by listening to pre-recorded spontaneous speech samples in the subjects' respective native languages. The 5-point [0 (loss of useful speech) – 4 (normal speech)] rating scale as used in the speech function item of ALSFRS-R[14] was adopted for this purpose. The mode of the three SLPs' ratings was considered as the final severity score. We did not include any subject who received three different ratings from three SLPs. Since the severity score of 4 indicates normal speech, we also did not include the ALS subjects who received a score of 4 from even one SLP. We selected 8 (4 M + 4 F; age range: 41–70 years), 7 (4 M + 3 F; age range: 43–68 years), 10 (5 M + 5 F; age range: 36–70 years), and 10 (5 M + 5 F; age range: 42–62 years) ALS subjects with final severity scores 0, 1, 2, and 3, respectively. We grouped the subjects with severity scores of 0 and 1 together as the *severe dysarthric group* (ALS_s) and those with severity scores of 2 and 3 together as the *mild dysarthric group* (ALS_m).

Each subject recorded sustained vowel utterances in chronological order of /a/, /i/, /o/, and /u/. They were asked to take a deep breath and perform a sustained utterance of a vowel at comfortable f_0 and loudness levels. They were given demonstrations of sustained utterances of each vowel. Example words from the subjects' respective native languages containing the target vowel were mentioned and explained. Human instructors gave all instructions and explanations in the subjects' respective native languages. Up to three utterances of each vowel were recorded from a subject depending on his/her level of comfort. All data were recorded at 44.1 kHz using a Zoom H-6 recorder (ZOOM, Hauppauge, NY)[15] placed at a distance of 2 ft from the subject. Recordings were then downsampled to 16 kHz. The number of utterances of each vowel obtained from different subject groups, along with the mean and standard deviation (SD) of the durations of the utterances, is given in Table 1.

3. Method

3.1 Measures of inter-speaker acoustic differences

Inter-speaker acoustic differences are estimated individually for the source and the filter components of the utterances.

Table 1. Number and duration of the vowel utterances obtained from different subject groups.

	Number of utterances				Mean (SD) of durations (in sec) of utterances			
Group	/a/	/i/	/o/	/u/	/a/	/i/	/o/	/u/
ALS_s	38	41	41	41	3.89 (2.40)	2.91 (2.22)	3.12 (2.55)	2.95 (2.48)
ALS_m	55	55	55	55	5.07 (3.42)	4.71 (3.57)	4.42 (2.90)	4.42 (2.98)
HC	105	103	102	103	6.04 (1.98)	6.12 (2.31)	5.95 (2.18)	5.78 (2.19)

Source level inter-speaker differences: These are quantified as the absolute differences between inter-speaker pairs of different source parameters computed from the sustained utterances of a vowel. In particular, eight source parameters are considered, namely, jitter (local), jitter (rap), jitter (ppq5), shimmer (local), shimmer (apq3), shimmer (apq5), SD of f_0, and mean HNR. These parameters, characterizing phonatory stability and f_0 variations, have been used previously for analysing inter- and intra-speaker variability of vocal characteristics for different subject populations.[11,16] We first extract the f_0 estimates of the utterances at 100 Hz using the cross correlation method in the PRAAT software.[17] We set the frequency range for f_0 candidate search from 50 to 450 Hz,[3] keeping all other settings to their respective default values. No pitch halving or doubling effect is observed with this setting in our dataset. Using the estimated f_0 values, we compute the eight source parameters listed previously at 200 ms frames with 50% overlap. Successful computation of jitter (ppq5) at the lowest possible f_0 of 50 Hz is not ensured if a frame length less than 120 ms is used, as it requires five complete f_0 cycles to be present in a frame. As long as the frame length is $\geq$ 120 ms (i.e., all source parameters can be computed), the obtained estimates of the parameters do not vary with varying frame length. This is because the utterances at hand are sustained vowels. So we arbitrarily choose a frame length of 200 ms for the process of source parameter estimation which satisfies the $\geq$ 120 ms criteria. Each source parameter is averaged over all frames of an utterance. Descriptive statistics of the source parameter values for different vowels and different subject groups are given in Appendix A. For every vowel of every subject group, we consider every possible inter-speaker pair of utterances of that vowel belonging to that group. We compute the absolute differences of each individual source parameter between the two utterances of each such pair. Let D^s_{hc}, D^s_m, and D^s_s be the random variables (RVs) indicating the source level inter-speaker differences thus obtained within the HC, ALS_m, and ALS_s groups, respectively. For computing the cross-severity differences between ALS_s and ALS_m, denoted by the RV D^s_{sm}, we consider all possible utterance pairs of a vowel such that one utterance of the pair is from the ALS_s group and another from the ALS_m group. We compute the absolute differences of each individual source parameter between the two utterances of each such pair. All samples of D^s_{hc}, D^s_m, D^s_s, and D^s_{sm} thus obtained are considered for further comparison.

Filter level inter-speaker differences: These are measured in terms of the cosine hyperbolic (cosh) spectral distances[18] between inter-speaker pairs of average filter (vocal tract) power spectra of a vowel estimated through the iterative adaptive inverse filtering (IAIF) algorithm.[19] Cosh distance is widely used to quantify the difference between two speech spectra for various speech processing applications.[18,20] We first normalize the speech samples of a sustained vowel utterance to zero mean and unit variance. The normalized speech signal is then divided into 200 ms frames with 50% overlap. We perform inverse filtering on each frame using the IAIF method to estimate the linear prediction coefficients (LPC) for the inverse vocal tract filter. The IAIF algorithm has two hyperparameters — (1) the order of LPC analysis for vocal tract, and (2) the order of LPC analysis for glottal source. An LPC order of 12 has been used by Alku[19] for vocal tract analysis of natural sustained vowel utterances using the IAIF algorithm. Hall[21] has stated that a maximum LPC order of 30 is sufficient for practical vocal tract modelling purposes. In dysarthric speech based applications of IAIF as well, similar LPC orders (e.g., 18[22] and 24[23]) have been used for vocal tract analysis. Based on these, in this work, the order of LPC analysis for vocal tract, or in other words, the LPC order for the inverse filter is varied from 12 to 30 at a step of 6. The LPC order for glottal source analysis is set to 8, which is the default value of this parameter in the IAIF algorithm for a speech signal having 16 kHz sampling frequency. The obtained LPC for the inverse filter at each 200 ms frame is converted to a 1024-point filter power spectrum. Last, the filter power spectra obtained from all frames of an utterance are averaged (averaging the LPC for the inverse filters obtained from all frames of an utterance first and then converting the average LPC to the filter power spectrum also do not change the observed trends). Similar to the source parameters, the estimated filter power spectra also do not vary significantly with the frame length. For example, at any LPC order, the average filter power spectra obtained from an utterance using 200 ms and 300 ms frame lengths with 50% overlaps have a median cosh distance of only 0.002–0.003. Similar to the source case, here also, we consider every possible inter-speaker pair of utterances for every vowel of every subject group. We consider the cosh spectral distances computed between the filter power spectra of the two utterances of each such pair. Let D^f_{hc}, D^f_m, and D^f_s be the RVs symbolizing the filter level inter-speaker differences thus obtained within the HC, ALS_m, and ALS_s groups, respectively. For computing the cross-severity differences between ALS_s and ALS_m, denoted by the RV D^f_{sm}, we consider all possible utterance pairs of a vowel such that one utterance of the pair is from the ALS_s group and another from the ALS_m group. We compute the cosh spectral distance between the filter power spectra of the two utterances of each such pair. All samples of D^f_{hc}, D^f_m, D^f_s, and D^f_{sm} thus obtained are considered for further comparison.

Choice of suitable utterance segment for inter-speaker acoustic difference computations: We first decide to consider the middle 1 s segments of the utterances for computing the source and filter level inter-speaker differences. The middle portion is selected because the most stable articulatory and phonatory configurations, with least transient variations, are expected to be attained during this portion. If any utterance lasts $\leq$ 1 s, then the complete duration of that utterance is considered.[24] In our dataset, 3.18% and 21.12% utterances obtained from the ALS_m and ALS_s groups, respectively, are $\leq$ 1 s long, whereas, all utterances produced by the HC subjects are >1 s long. The mean and SD of the durations of the utterances having $\leq$ 1 s duration are 0.76 (0.23) s and 0.77 (0.13) s for ALS_m and ALS_s, respectively. However, computation of the source parameters, as described previously, requires the speech segments under consideration to be completely voiced. Moreover, the IAIF algorithm used for inverse vocal tract filter estimation is primarily designed for

vowels having purely voiced glottal excitations. Since dysarthria can affect the voicing characteristics of the vowel utterances, we first proceed to check if the decided segments of the utterances are completely voiced. For this purpose, we use the aforementioned f_0 estimates obtained from the utterances. The estimates are obtained at 100 Hz frequency, i.e., an estimate is obtained for every 10 ms audio frame. Such 10 ms frames having finite positive f_0 estimates are identified as voiced, while the others as unvoiced. Though sustained vowel utterances are typically expected to be completely voiced, in our dataset, the utterances obtained from HC, ALS_m and ALS_s groups have 0.21%–5.47%, 0.20%–14.57% and 0.28%–33.92% of all frames as unvoiced, respectively. In the case of HC, the unvoiced frames are present only at the very beginning and end of the utterances, whereas in the dysarthric speech, the unvoiced frames are encountered in the middle portions of the utterances as well. Hence, for all source and filter parameter computations, we use the middle 1 s segment of an utterance only if that segment is completely voiced. This condition is satisfied for all HC utterances and 89.09% and 70.81% utterances produced by ALS_m and ALS_s, respectively. For the rest of the utterances, we identify all voiced segments present in an utterance as the segments of contiguous voiced frames surrounded by unvoiced frames and select the longest voiced segment overlapping with the middle 1 s segment of that utterance. Among these utterances, in 47.06% and 38.46% cases of the ALS_m and ALS_s groups, respectively, the length of the selected longest voiced segment is >1 s. In these cases, we consider the middle 1 s subsegment of the selected voiced segment for the analysis. Otherwise, if the duration of the selected longest voiced segment is ≤ 1 s, that complete segment is considered. In the case of an utterance having ≤ 1 s duration, the longest voiced segment present in that utterance is considered. As a result of these selection criteria, the shortest utterance segments selected for subsequent analyses from the HC, ALS_m, and ALS_s groups are found to be 1, 0.31, and 0.24 s long, respectively. The mean and SD of the durations of the selected utterance segments having <1 s duration are 0.70 (0.23) s and 0.71 (0.16) s for the ALS_m and ALS_s groups, respectively.

3.2 Statistical analysis

Kolmogorov-Smirnov tests[25] at alpha level 0.01 confirm that the source/filter level inter-speaker differences or their logarithms do not follow normal distribution except the logarithm of D_s^f for /o/. So, we perform the non-parametric Wilcoxon Ranksum (WR) test[26] at an alpha level of 0.01 to determine if two RVs denoting the source/filter level inter-speaker differences corresponding to two different subject group configurations have continuous distributions with significantly different medians. Since we have a widely varying number of utterances from different subject groups, the number of samples of D_m^*, D_s^*, D_{sm}^*, and D_{hc}^* also vary significantly (exact sample counts can be found in Appendix B). So, while comparing any two of these RVs, we randomly choose as many samples of the RV having a larger number of samples as the other RV with a smaller number of samples. This random selection is performed in a gender- and age-matched fashion. We perform this random selection 20 times to cover varied sample subsets of the RV with a larger number of samples. We report the fraction of times (out of 20) a significant statistical difference is observed between the two RVs being compared and denote this metric as the confidence score. While comparing D_m^*, D_s^* or D_{sm}^* with D_{hc}^*, the confidence score is prefixed by a "+"/"−" sign if the samples of the ALS difference category at hand have a significantly higher/lower median than that of the HCs. For D_m^* vs D_s^* comparison, the confidence score is prefixed by a "+"/"−" sign if D_s^* has a significantly higher/lower median than D_m^*. No case is observed where out of the 20 trials some correspond to the "+" and some to the "−" sign. If no significant statistical difference is observed in any of the 20 trials, we report the confidence score as 0.

4. Results

4.1 Source level inter-speaker differences

Figure 1 plots the median values of D_{hc}^s, D_m^s, D_s^s, and D_{sm}^s, for the four vowels with regard to the eight source parameters.

Comparison between ALS and HC: Jitter (local), jitter (ppq5), and SD of f_0 exhibit significantly higher median inter-speaker differences with +1 confidence scores within and across the two severity groups of ALS, as compared to those of HCs, for all the four vowels. Jitter (rap) also follows a similar trend in all cases except in the case of D_m^s for /o/ and /u/. D_m^s is statistically not different from D_{hc}^s for /o/, whereas, it has a significantly higher median value than D_{hc}^s with a very low confidence score of +0.1 for /u/. The trends largely vary across different vowels in the case of shimmer. For /a/, with respect to all three shimmer parameters, D_s^s and D_{sm}^s have significantly higher median values than D_{hc}^s with high confidence scores of +1. D_m^s, however, exhibits significantly higher median values than D_{hc}^s with the low confidence scores of +0.1 and +0.05 in the cases of shimmer (local) and shimmer (apq3), respectively. D_m^s is statistically not different from D_{hc}^s with respect to shimmer (apq5). In the case of /i/, for all shimmer parameters, all of D_m^s, D_s^s, and D_{sm}^s exhibit significantly lower median values than D_{hc}^s with high confidence scores of −1. Similar trends are observed for /o/ and /u/ as well in the case of D_m^s for all shimmer measures and D_{sm}^s for shimmer (apq3). Here, the confidence scores vary in [−0.8, −1]. For shimmer (local) and shimmer (apq5) of /o/ and /u/, the median values of D_s^s are significantly higher than those of D_{hc}^s with a confidence score of at least +0.75. D_{sm}^s is statistically not different from D_{hc}^s in these cases except for shimmer (apq5) of /u/, where D_{sm}^s has a significantly lower median value than D_{hc}^s with a low confidence of −0.25. Moreover, for shimmer (apq3), D_s^s remains statistically not different from D_{hc}^s in the case of /u/, whereas it exhibits a significantly higher median value than D_{hc}^s with a low confidence score of +0.35 in the case of /o/. Last, with regard to mean HNR, /a/ and /i/ exhibit significantly higher median values of D_m^s and D_{sm}^s, than D_{hc}^s, with +1 confidence scores, while D_s^s has a

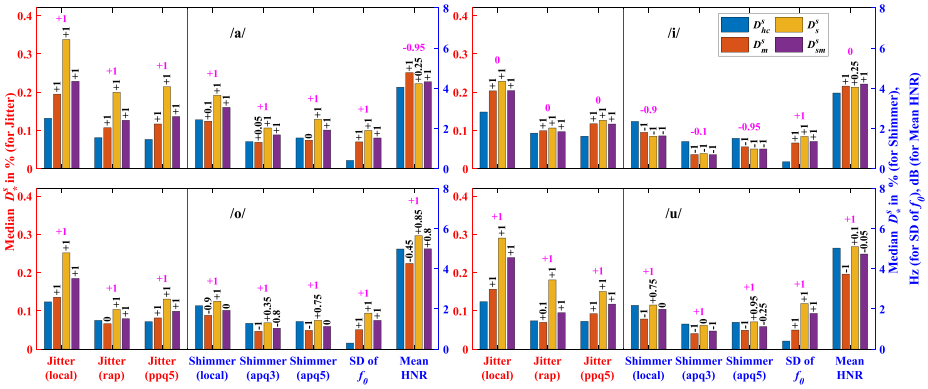

Fig. 1. Median source level inter-speaker differences for various subject groups during sustained utterances of the four vowels; inter-speaker differences with respect to jitter parameters follow the left *y* axes (in red) and those with respect to the other source parameters follow the right *y* axes (in blue); plotted median D^s_{hc} values are obtained from the entire HC population; confidence scores for comparison of D^s_m, D^s_s, and D^s_{sm} with D^s_{hc} are mentioned in black on top of the respective bars; confidence score for comparison between D^s_m and D^s_s is given in magenta on top of each group of bars.

significantly higher median than D^s_{hc} with a confidence score of only +0.25 for both the vowels. D^s_m, in the cases of /o/ and /u/, has significantly lower median values than D^s_{hc} with confidence scores of –0.45 and −1, respectively. However, D^s_s and D^s_{sm} for /o/ have higher median values than D^s_{hc} with confidence scores of +0.85 and +0.8, respectively. Moreover, for /u/, D^s_s has a higher median than D^s_{hc} with a confidence score of only +0.1, while D^s_{sm} has a lower median than D^s_{hc} with a confidence score of only −0.05.

Comparison between mild and severe dysarthria: SD of f_0 exhibits significantly higher median inter-speaker differences with +1 confidence scores for ALS_s, as compared to ALS_m, for all four vowels. The trend is the same in the cases of all jitter and shimmer parameters for the vowels /a/, /o/, and /u/. For /i/, however, D^s_s is statistically not different from D^s_m with respect to any jitter measure. For shimmer (local) and shimmer (apq5), D^s_s of /i/ exhibits significantly lower median values than D^s_m with high confidence scores of −0.9 and −0.95, respectively. Though shimmer (apq3) of /i/ shows a similar trend as well, the confidence score is very low (−0.1). Last, with regard to mean HNR, /o/ and /u/ exhibit significantly higher median values of D^s_s, than D^s_m, with high confidence scores of +1, while /a/ exhibits a significantly lower median value of D^s_s, as compared to D^s_m, with a high confidence score of –0.95. However, for /i/, D^s_s and D^s_m are statistically not different.

4.2 Filter level inter-speaker differences

Figure 2 plots the median D^f_{hc}, D^f_m, D^f_s, and D^f_{sm} values for the four vowels under consideration against increasing LPC order for the inverse filter. For every vowel, the median value of each of D^f_{hc}, D^f_m, D^f_s, and D^f_{sm} increases with the increase in the LPC order. At higher LPC orders, the filter power spectra undergo less smoothing and hence contain finer details. Lack of alignment of such fine information along the frequency axis between a pair of filter power spectra increases the corresponding cosh distance, thus leading to higher median filter level inter-speaker differences at higher LPC orders.

Comparison between ALS and HC: Vowel /i/ is the only one for which the median values of all of D^f_m, D^f_s, and D^f_{sm} are significantly higher than those of D^f_{hc} with high confidence scores of +0.9 or +1 at all LPC orders except only one case. At the order of 12, D^f_m is statistically not different from D^f_{hc}. For /u/ as well, D^f_s and D^f_{sm} have significantly higher median values than D^f_{hc} with +1 confidence scores at all LPC orders. D^f_m in this case, however, is statistically not different from D^f_{hc} at the LPC orders of 18 and 30, whereas D^f_m has significantly lower median values than D^f_{hc} with low confidence scores of −0.1 at the other LPC orders. Vowel /a/ also has statistically not different D^f_m and D^f_{hc} at all LPC

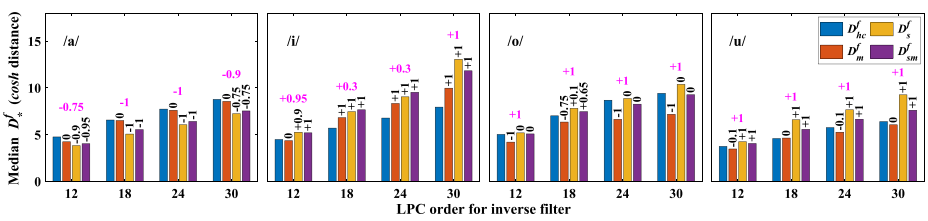

Fig. 2. Median filter level inter-speaker differences for different subject groups during sustained utterances of the four vowels; plotted median D^f_{hc} values are obtained from the entire HC population; confidence scores for comparison of D^f_m, D^f_s, and D^f_{sm} with D^f_{hc} are mentioned in black on top of the respective bars; confidence score for comparison of D^f_m and D^f_s is given in magenta on top of each group of bars.

orders. Moreover, D_s^f and D_{sm}^f for /a/ exhibit significantly lower median values than D_{hc}^f with high confidence scores in $[-0.75, -1]$ at all LPC orders. Last, for /o/, D_m^f has a significantly lower median than D_{hc}^f with high confidence scores of -0.75 or -1 at all LPC orders. However, D_s^f and D_{sm}^f for /o/ are statistically not different from D_{hc}^f at all LPC orders except 18 where the median values of D_s^f and D_{sm}^f are significantly higher than that of D_{hc}^f with confidence scores of $+0.1$ and $+0.65$, respectively.

Comparison between mild and severe dysarthria: Vowels /o/ and /u/ exhibit significantly higher median values of D_s^f than those of D_m^f with $+1$ confidence scores at all LPC orders. Though for vowel /i/, D_s^f has significantly higher median values than D_m^f with high confidence scores of $+0.95$ and $+1$, respectively, at the LPC orders of 12 and 30, the confidence scores are low ($+0.3$) at other LPC orders. Vowel /a/ is the only one for which D_s^f exhibits significantly lower median values than D_m^f with high confidence scores of -0.75, -0.9, or -1 at all LPC orders.

5. Discussion

Different ALS subjects, belonging to the same or different dysarthria severity groups, experience different degrees of deterioration in the control over vocal cord vibrations. This may lead to higher inter-speaker differences of jitter parameters[27] within and across the two dysarthria severity groups of the ALS population than those persisting among the HCs [except for jitter (rap) in the case of D_m^s for /o/ and /u/]. The higher inter-speaker differences observed among the SD of f_0 values for these patients as compared to the HCs, might be linked to the variable degrees of laryngeal impairments, and hence laryngeal tension, experienced by these patients.[27] However, similar trends of increased inter-speaker differences are not manifested consistently across all vowels in the case of shimmer and HNR measures. Vowel /a/ is the only one that exhibits enhanced inter-speaker differences within ALS_s group as well as between ALS_s and ALS_m groups, as compared to those existing among HCs, with respect to all source parameters. The same is also true in the case of D_m^s for all source parameters except shimmer measures. Thus, /a/ seems to be more sensitive to the varieties of source impairments occurring in ALS-induced dysarthria.

Vowels /i/ and /u/ are close vowels. Individuals with ALS-induced dysarthria seem to experience difficulties in forming close vocal tract configurations which require the tongue to be brought in close proximity to the palate.[24] This is possibly due to the impairments in controlling the tongue height.[28] Thus, different patients may achieve the target vocal tract structures of close vowels to different degrees of accuracy. Moreover, the patients often try to compensate for the impairments of one articulator with another to mimic the target sound.[4] These factors may lead to higher filter level inter-speaker differences among the patients with dysarthria than those existing among the HC population. Though such a pattern is observed in the cases of D_s^f and D_{sm}^f for both /i/ and /u/, the patterns for D_m^f differ between the two vowels. D_m^f has significantly higher median values than D_{hc}^f in the case of the close front vowel /i/, whereas D_m^f and D_{hc}^f are statistically not different in the case of the close back vowel /u/. This might indicate that forming close front configurations becomes more difficult from the early stages of the disease as compared to the close back configurations. On the other hand, /a/ and /o/ require the vocal tract to be relatively more open. Thus, it might be relatively easier for the patients to achieve the target vocal tract configurations of these vowels, thereby resulting in statistically not different or even significantly less filter level inter-speaker differences among these subjects as compared to the HCs during sustained utterances of /a/ and /o/.

Subjects having a variety of native languages are considered in this work. This might have some impacts on the observed inter-speaker acoustic differences as native language can influence the accent of a speaker. However, every subject was given demonstrations of the intended pronunciations of the vowels to ensure uniformity across all utterances. Moreover, statistical comparisons of the first and second formant values of different vowels across different native languages suggest that the formant values are not statistically different between the majority of the language pairs under consideration ($\geq 92.86\%$) for every vowel at hand. This indicates that the effect of the native languages on our observations is minimal. The detailed formant analysis can be found in Appendix C. Another factor that can impact our analysis is the variations in the durations of the utterance segments considered. All utterance segments obtained from HCs are 1 s long, whereas 7.27% and 26.09% utterance segments obtained from ALS_m and ALS_s groups, respectively, have <1 s duration. However, even on excluding all segments having <1 s duration, the observed patterns of inter-speaker differences do not change significantly. Only the magnitude of confidence scores changes in some cases, though the change is >0.2 in only 17.50% of cases. The sign of the confidence score does not change in any case. This might be because we average the estimates of source parameters and filter power spectra over all frames of a selected utterance segment prior to obtaining the inter-speaker differences. Averaging the estimates over all frames of a 1 s utterance segment or a <1 s segment may not make much difference. In addition to these aspects, the relative degree of inter-speaker acoustic differences observed among different subject groups is also influenced by the spectrum of severity included in the different groups with dysarthria.

6. Conclusion

We analyze the degree of inter-speaker acoustic differences within and across ALS patients with mild and severe dysarthria during different sustained vowel utterances. We study the differences in the individual source and filter components of the utterances. The degree of inter-speaker differences, with regard to the different parameters considered, are found to vary

at different dysarthria severity levels. Phoneme specific trends are also observed. In the future, we plan to analyze the intra-speaker speech acoustic variations for these patients, as compared to those of the HCs.

Supplementary Material

See the supplementary material for Appendixes A, B, and C.

Acknowledgments

We thank the Department of Science and Technology (DST), Government of India for supporting this work.

Author Declarations

Conflict of Interest

The authors have no conflict of interest to be disclosed.

Ethics Approval

The data collection protocol followed in this work was reviewed and approved by the ethics committee of NIMHANS, Bengaluru, India. Also, informed consent was obtained from each subject in his/her native language prior to data collection.

Data Availability

The data that support the findings of this study are available on request from the corresponding author.

References

[1]G. Fant, *Acoustic Theory of Speech Production* (Walter de Gruyter, Berlin, 1970).
[2]P. Rong, Y. Yunusova, J. Wang, L. Zinman, G. L. Pattee, J. D. Berry, B. Perry, and J. R. Green, "Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems," PLoS One **11**(5), e0154971 (2016).
[3]M. Vashkevich and Y. Rushkevich, "Classification of ALS patients based on acoustic analysis of sustained vowel phonations," Biomed. Signal Process. Control **65**, 102350 (2021).
[4]S. Shellikeri, Y. Yunusova, D. Thomas, J. R. Green, and L. Zinman, "Compensatory articulation in Amyotrophic Lateral Sclerosis: Tongue and jaw in speech," Proc. Mtgs. Acoust. **19**(1), 060061 (2013).
[5]R. Johnson, N. Kim, K. Tjaden, and A. Mefferd, "Articulatory strategies and their acoustic consequences: Investigating tongue retraction and lip protrusion tradeoffs in talkers with amyotrophic lateral sclerosis," J. Acoust. Soc. Am. **148**(4_Supplement), 2583 (2020).
[6]B. Tomik and R. J. Guiloff, "Dysarthria in amyotrophic lateral sclerosis: A review," Amyotrophic Lateral Sclerosis **11**(1-2), 4–15 (2010).
[7]M. Eshghi, K. P. Connaghan, S. E. Gutz, J. D. Berry, Y. Yunusova, and J. R. Green, "Co-occurrence of hypernasality and voice impairment in amyotrophic lateral sclerosis: Acoustic quantification," J. Speech. Lang. Hear. Res. **64**(12), 4772–4783 (2021).
[8]A. E. Hallin, K. Fröst, E. B. Holmberg, and M. Södersten, "Voice and speech range profiles and voice handicap index for males - methodological issues and data," Logoped. Phoniatr. Vocol. **37**(2), 47–61 (2012).
[9]S. Ternström and P. Pabon, "Accounting for variability over the voice range," in *Proceedings of the 23rd International Congress on Acoustics (ICA) Integrating 4th EAA Euroregio*, Aachen, Germany (September 9–13, 2019), pp. 7775 – 7780.
[10]J. Green, K. Connaghan, Y. Yunusova, K. Stipancic, S. Gutz, and J. Berry, "Vocal changes across disease progression in amyotrophic lateral sclerosis (ALS)," J. Acoust. Soc. Am. **144**(3_Supplement), 1966 (2018).
[11]E. A. Strand, E. H. Buder, K. M. Yorkston, and L. O. Ramig, "Differential phonatory characteristics of four women with amyotrophic lateral sclerosis," J. Voice **8**(4), 327–339 (1994).
[12]J. F. Kent, R. D. Kent, J. C. Rosenbek, G. Weismer, R. Martin, R. Sufit, and B. R. Brooks, "Quantitative description of the dysarthria in women with amyotrophic lateral sclerosis," J. Speech. Lang. Hear. Res. **35**(4), 723–733 (1992).
[13]B. R. Gerratt, J. Kreiman, and M. Garellek, "Comparing measures of voice quality from sustained phonation and continuous speech," J. Speech. Lang. Hear. Res. **59**(5), 994–1001 (2016).
[14]J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, and A. Nakanishi, "The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function," J. Neurol. Sci. **169**(1–2), 13–21 (1999).
[15]Information on the Zoom microphone is available at https://tinyurl.com/23mhat4w (Last viewed June 1, 2024).
[16]A. Remacle, M. Garnier, S. Gerber, C. David, and C. Petillon, "Vocal change patterns during a teaching day: Inter- and intra-subject variability," J. Voice **32**(1), 57–63 (2018).
[17]P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 6.4.01) [computer program]," https://www.praat.org (Last viewed November 30, 2023).
[18]A. Gray and J. Markel, "Distance measures for speech processing," IEEE Trans. Acoust. Speech. Signal Process. **24**(5), 380–391 (1976).
[19]P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," Speech Commun. **11**(2-3), 109–118 (1992).
[20]B. J. Borgström and A. Alwan, "Log-spectral amplitude estimation with generalized gamma distributions for speech enhancement," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic (May 22–27, 2011), pp. 4756–4759.
[21]M. C. Hall, "Objective quality evaluation of parallel-formant synthesised speech," in *Proceedings of the First European Conference on Speech Communication and Technology (Eurospeech)*, Paris, France (September 27–29, 1989), pp. 2629–2632.

[22]A. A. Joshy, P. Parameswaran, S. R. Nair, and R. Rajan, "Statistical analysis of speech disorder specific features to characterise dysarthria severity level," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece (June 4–10, 2023), pp. 1–5.
[23]N. Narendra, B. Schuller, and P. Alku, "The detection of Parkinson's disease from speech using voice source information," IEEE/ACM Trans. Audio. Speech Lang. Process. **29**, 1925–1936 (2021).
[24]T. Bhattacharjee, C. V. T. Kumar, Y. Belur, A. Nalini, R. Yadav, and P. K. Ghosh, "Static and dynamic source and filter cues for classification of Amyotrophic Lateral Sclerosis patients and healthy subjects," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece (June 4–10, 2023), pp. 1–5.
[25]F. J. Massey, Jr., "The Kolmogorov-Smirnov test for goodness of fit," J. Am. Stat. Assoc. **46**(253), 68–78 (1951).
[26]J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference* (CRC Press, Boca Raton, FL, 2014).
[27]P. Suphinnapong, O. Phokaewvarangkul, N. Thubthong, A. Teeramongkonrasmee, P. Mahattanasakul, P. Lorwattanapongsa, and R. Bhidayasiri, "Objective vowel sound characteristics and their relationship with motor dysfunction in Asian Parkinson's disease patients," J. Neurol. Sci. **426**, 117487 (2021).
[28]J. Lee, H. Kim, and Y. Jung, "Patterns of misidentified vowels in individuals with dysarthria secondary to Amyotrophic Lateral Sclerosis," J. Speech. Lang. Hear. Res. **63**(8), 2649–2666 (2020).